Multi-species biclustering: An integrative method to identify functional gene conservation between multiple species

by

Peter Waltman

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Program in Computational Biology

New York University

January, 2012

Dr. Richard Bonneau

Adviser

DEDICATION

I would like to dedicate this to my mother, who despite the odds, continues to struggle in her fight with Acute Myeloid Leukemia – I wish I had the strength you've shown. Also, to Dr. Jim Schmolze, who was the one person who managed to convince me that I wasn't too old to start my Ph.D. at 34 – and who was also taken from this Earth at far too young an age when he lost his own fight with pancreatic cancer.

ACKNOWLEDGMENT

I would simply like to thank all the friends who managed to keep me sane during this entire process, and especially during my first year in the program – we may not see each other as often as we used to, but you all know who you are. I would also especially like to thank Ashley Rose Bate who had a critical role in re-energizing our analysis of both triplets, and the Gram-positive triplet in particular. When Thadeous and I were stuck with 450 biclusters that we had no idea what to do with, she provided invaluable help in identifying key biclusters of interest. The pipetting accident you had was truly an accident of fate. In addition, I should also thank Thadeous Kacmarczyk for helping Ashley narrow down and focus the stories we came up with; for being the conduit between the Eichenberger & Bonneau labs; and for the amazing visualization you put up. Last, I want to thank Patrick Eichenberger for all his contributions with our analysis as well; Kris Gunsalus, for providing much needed advice and a sympathetic ear this past winter; and, of course, Rich Bonneau for your direction, and for giving me a chance when I first joined the lab.

ABSTRACT

Background: Several recent comparative functional genomics projects have indicated that the co-regulation of many genes is conserved across species, at least in part. This suggests that comparative analysis of functional genomics data-sets could prove powerful in identifying co-regulated groups that are conserved across multiple species.

Results: We present recent work to extend our cMonkey algorithm to simultaneously bicluster heterogeneous data from multiple species to identify conserved modules of orthologous genes, which can yield evolutionary insights into the formation of regulatory modules. We also present results from the multi-species analysis to two triplets of bacteria. The first of these is a triplet of Gram-positive bacteria consisting of *Bacillus subtilis*, *Bacillus anthracis*, and *Listeria monocytogenes*, while the second is a triplet of Gram-negative bacteria that includes *Escherichia coli*, *Salmonella typhimurium* and *Vibrio cholerae*. Finally, we will present initial results from the multi-species biclustering analysis of human and mouse hematopoietic differentiation data.

Conclusion: Analysis of biclusters obtained revealed a surprising number of gene groups with conserved modularity and high biological significance as judged by several measures of cluster quality. We also highlight cases of interest from the Gram-positive triplet, including one that suggests a temporal difference in the expression of genes governing sporulation in the two *Bacillus* species. While analysis of the mouse and human hematopoietic differentiation is preliminary, it indicates the applicability of

this analysis to eukaryotic systems, including comparison of cancer model systems. Finally, we suggest ways in which this analysis could be extended to identify divergent modules that may exist between normal and disease tissue.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION – PROKARYOTIC FUNCTIONAL GENOMICS

Co-written with: Thadeous Kacmarczyk[1☯] and Richard Bonneau[1,2,*]

1. Center for Comparative Functional Genomics, New York University Dept. of Biology, New York, NY, 10003, USA.

2. Courant Institute for Mathematical Sciences, Computer Science Dept., New York University, New York, NY, 10003, USA.

☯ Equal Authorship.

* To whom correspondence should be addressed.

Keywords: Systems Biology, Network inference, *Halobacterium NRC-1, Halobacterium salinarum, Caulobacter crescentus, Escherichia coli*, microarray, proteomics.

Original Publication: Waltman, P., T. Kacmarczyk, et al. (2009). Prokaryotic Systems Biology. Plant systems biology, Annual plant reviews. G. Coruzzi and R. A. Gutierrez. Ames, Iowa, Blackwell Pub.**:** 67-136.

Author contributions:

- Waltman, P wrote the sections covering *Caulobacter crescentus*, *Escherichia coli* and the first 5 sub-sections of the section on *Halobacterium salinarium NRC-1* (sections 1.3, 1.5-1.6.5). He also updated both the introductory and conclusion sections that were original written by Bonneau, R (sections 1.1 and 1.7-1.8, respectively).
- Kacmarczyk, T wrote the section providing the review of core technologies as well as the section on *Baciulls subtilis* (sections 1.2 and 1.4).
- Bonneau, R wrote the second half of section describing *Halobacterium salinarium NRC-1*, as well the original introductory and conclusion sections (sections 1.6.6-9, 1.1 and 1.7-1.8, respectively).

1.1 Overview of Systems Biology

Recent advances in systems biology have dramatically accelerated the rate at which biologists can acquire data on all informational levels of the cell (genome sequence, RNA, protein, protein modification, metabolites, etc.). Concurrent advances in computational biology have begun to allow for large multi-group efforts that integrate these diverse data sources in order to generate predictive dynamical models of whole cells (Bonneau, Facciotti et al. 2007). In addition, several groups such as the ENCODE Project Consortium (Birney, Stamatoyannopoulos et al. 2007) and the modENCODE Consortium (Celniker, Dillon et al. 2009) have recently produced first drafts of global models of the functional elements of the genomes of several eukaryotic organisms, including *Homo sapiens* and the model organisms *Drosophila*

melanogaster (Negre, Brown et al. 2011) and *Caenorhabditis elegans* (Gerstein, Lu et al. 2010). In this introduction, which is heavily based upon the chapter "Prokaryotic Systems Biology" in *Plant Systems Biology* (Waltman, Kacmarczyk et al. 2009), we'll discuss in detail several prokaryotic functional genomics projects, with the dual goals of 1) illustrating how recent advances in computational techniques have advanced and aided these projects; and 2) motivating the research that is presented in this thesis dissertation.

We will show that, although many challenges remain, we are beginning to cross critical milestones in our efforts to learn systems-wide quantitative models of prokaryotic cells and their interactions with their environments. To do this, we will provide a very brief explanation for the novice of the significance and utility of prokaryotic systems biology. This will be followed by a review of some of the technologies that are used to generate the high-throughput experimental data which systems biologists analyze, after which we will provide examples of how systems biology approaches have been used with four (4) prokaryotic organisms. The first of these describes a multi-year, multi-team effort that used primarily non-computational and non-systems-level approaches to map regulatory circuit governing the cell cycle of *Caulobacter crescentus*. The second section discusses a number of systems biology efforts to characterize various aspects of the regulatory network governing *Bacillus subtilis* under various conditions, with many of these using the different technologies that are available. This will be followed by a section describing three projects that

explored various aspects of the regulatory network governing *Escherichia coli*; one of which was a project that used primarily non-systems techniques to map the acid shock response of *E. coli*; another that used a novel computational method to infer a global regulatory network based on systems-level expression data; and a third that integrated known metabolic and regulatory interactions to generate an *in silico* model of *E. coli*. The third section describes a multi-year project to map the complete regulatory network of the archael organism, *Halobacterium salinarium NRC-1* that was performed by a single group which combined novel computational method development and wet-bench verification of the predictions from these in an iterative manner. Finally, in the remaining sections of this chapter, we motivate the research that we present in following chapters of this thesis.

1.1.1 The importance of microbes:

Bacteria and archaea are abundant, diverse and important organisms. Many currently relevant human pathogens are prokaryotic. Microbes have been used for fermentation of foodstuffs for eons and more recently have been used in engineering and synthesis applications spanning the full range of human activities (e.g. bacteria can serve as platforms for the synthesis of drugs, vitamins, food additives). Prokaryotes play critical roles in our environment and are central to efforts to mitigate the human impact associated with solid waste/sewage, industrial toxic waste, and agriculture. Prokaryotic biology is critical to our understanding the history of our

environment. Prokaryotes have traditionally provided biologists useful tools for molecular and cell biology across all systems.

1.1.2 Experimental advantages of prokaryotic systems biology:

Archaea and bacterial systems offer a distinct advantage in complexity. Although they have all the properties of life that warrant our awestruck admiration, such as self-assembly, robustness, reproducible autonomous decision making, they are orders of magnitude less complex than Eukaryotes, they often allow for collection of larger amounts of material in the lab. Prokaryotes are often synchronizable (as are many eukaryotic systems) and often amenable/robust to the manipulations needed for single celled measurements (Alon 2007). Often the genetics of a given prokaryotic system will allow for rapid construction of knock out and/or over-expression strains that can be used to directly query the global result of specific genetic perturbation (this is the case for all organisms described herein). Unfortunately these experimental advantages do not extend to all organisms and several prokaryotes participate in complex communities that currently elude even laboratory culture, and are thus only now coming into focus via metagenomic sequencing directly from the environment (Handelsman 2004). In this review we will focus on organisms that are amenable to genetics, culture and have full genome sequence.

1.1.3 Types of questions

Before we begin our discussion we need to discuss the types of questions one might answer with prokaryotic systems biology.

1.1.3.1 Core biology:

The first and most fundamental question one might ask is “how do all systems components interact to form core aspects of biology with components and/or strategies common to many systems.” For example we might study the cell cycle in several organisms and compare common themes in an attempt to reveal the functional requirements or ancestral progenitor of cell cycle control in different niches/organisms. Systems biology becomes essential in answering this type of question due to the sheer number of genes involved in many core processes. So the fact that much of the cell is involved makes techniques based on global measurements a natural fit to the question. So-called master regulators (hubs) are prevalent in biology and determining the targets and control of such master regulators is more directly accomplished via global techniques (such as ChIP-chip, yeast one hybrid, microarray measurement following a genetic perturbation to the gene, etc.).

1.1.3.2 Environmental:

Another case where global measurements are key is in the deciphering of an organisms response to its environment. A typical structure for such a study involves the use of genomics techniques to identify key players in a physiological response to a given cell environment, followed by more focused studies to investigate/validate the role or necessity of the discovered proteins/genes. Many of the earliest studies employing microarrays in prokaryotic systems were designed to characterize a cells genome-wide/transcriptome response to environmental stress. In these studies we look

for novel regulation of known processes that have been discovered, novel associations between proteins of unknown function with known environmental responses.

1.1.3.3 Disease related pathogens:

In cases where the prokaryote of interest is also a human pathogen, our question is: “how to maximally disrupt the pathogen, disrupt its interaction with the human host or vector, or otherwise mitigate its effect on human health”. In this study we will focus less on this type of study, as the interaction with the human host often requires as much study as the internal workings of the pathogen of interest. This prokaryote-host interaction is, although currently the focus of several systems biology efforts, beyond the scope of this review.

1.1.3.4 Engineering

Genome-wide models will inevitably be required if we are to rationally engineer microbial systems. Reasons for engineering microbial systems span human efforts and include: bioenergy, remediation of industrial waste sites, production of difficult to synthesize compounds.

1.1.4 Global models

1.1.4.1 Emergent properties:

Emergent properties are properties of a system that cannot be trivially traced back to properties of any single component of the system. Simple examples of emergent properties abound in nature such as flock behavior, the decisions and

patterns of ant and termite colonies, dramatic trends in human economies, a tabby cat's stripes, spiral waves in heart defibrillation, etc. When we refer to the meaningful properties of highly complex systems as emergent in this review it is simply a compact way of describing the simple notion that if large complex systems have many inter-component interactions then only by modeling the global system can we hope to recapitulate or model the overall system behavior. Systems that involve interactions on multiple scales, interactions between components that involve loops (such as feedback loops), and non-linear effects such as saturation, recovery and auto-excitation all contribute to the degree to which systems are likely to have difficult to predict emergent properties.

Nearly all biological systems exhibit complex phenotypes and physiologies that are not attributable to single subsystems or genes, and all biological systems are large, complex systems involving all of the interaction types typically leading to systems dominated by emergent behavior. Thus we must view important properties of living systems as interdependent, emergent, or at least highly epigenetic phenomena. Regardless of our diction we rapidly arrive at the conclusion that highly interconnected phenomena like metabolism, signaling and regulation require modeling at the global, genome-wide, scale if we are to construct predictive models of cellular behavior.

1.1.4.2 Global models require the new approaches to experimental design, technologies, and analysis:

This motivation for global measurement and modeling of biology has led to prokaryotic biologists, working on several systems, to adopt some aspect of genomic (genome-wide) experimentation and analysis. In the end this has led to many successes and many mistakes, as the field wrestles with technical and computational challenges generated by high throughput methods. After a decade of systems biology, many biologists feel a bit unclear, pedagogically, as to the state of modern biology. Many people incorrectly feel that biology is currently a disjoint field, with labs that perform systems-biology/global studies existing in a sub-field separate from those biologists that perform one-gene-at-a-time studies. One point we hope to convey by reviewing several functional genomics projects below is that many of the most interesting results are from work where more focused studies of subsystems and small numbers of genes are embedded in or guided by global analysis.

1.2 Review of core technologies for prokaryotic systems biology.

Here we will briefly review the core technologies found in a typical prokaryotic functional genomics pipeline as discussed throughout the paper. We will place emphasis on these techniques in our discussion of four specific functional genomics projects below. This section illustrates that many of these technologies, although found throughout studies of Eukaryotes as well, were first developed in prokaryotic systems.

1.2.1 Genomics

The sequenced genome is an essential prerequisite to determining the parts-list for an organism, encoding its RNA transcripts, proteins, as well as several patterns and properties beyond our current understanding. The field of genomics has expanded during the past decade from the static study of DNA sequences, annotation, and structure to dynamic studies of functional and comparative genomics, but all rests squarely on our ability to determine complete genome sequences for organisms in a cost effective manner. The process and capabilities of genome sequencing has dramatically changed since the first complete genome in 1995 of *Haemophilus influenzae Rd.* (Fleischmann, Adams et al. 1995) with innovations in cloning and high throughput DNA sequencing technology. The Sanger (Sanger, Nicklen et al. 1977), or chain termination method, is still the primary method for sequencing (although new technologies are most certainly poised to overtake it as the most commonly used method). Sanger sequencing has seen many optimizations and improvements since the laboratory of Leroy Hood first automated the process in the mid 1980's. These advancements include advances in fluorescent labels and detection, capillary electrophoresis and microfluidics, automation, informatics and computational power, and now typically produce ~100kbp per run of a typical capillary sequencing machine. There are two new promising technologies: 454 Life Sciences sequencing can produce 30Mbp per run by utilizing a sequence-by-synthesis (SBS) approach which integrates pyrosequencing, massively parallel sequencing and microfabricated picoliter reactors

(Margulies, Egholm et al. 2005). This technology has already resulted in economically sequenced genomes (in combination with Sanger for longer reads). Solexa sequencing technology also uses SBS and massively parallel technology on a clonal single molecule array, and is working towards 1Gbp per run. Emerging methods based on other technologies such as, sequence-by-hybridization, mass spectroscopy, and single molecule nanopore sequencing are also being investigated. Regardless of which technique wins the race, it is clear that sequencing 100s of prokaryotic genomes by single groups with modest funding is on the horizon. Even without these new technologies high throughput sequencing is pouring out raw data at a fantastic rate. This along with new techniques for protein annotation has allowed us to compile a very large compendium of gene and protein families that greatly facilitate our management of the complexity of any given proteome (Tatusov, Koonin et al. 1997; Finn, Mistry et al. 2006). Comparative genomics can illustrate genetic programs that are global properties of organisms as well as properties specific to a species. This sequencing power offers opportunities into the natural microbial world.

Much of the earth's biomass is comprised of microorganisms that participate in tightly interconnected microbial communities. In many cases these communities are too complex or adapted to a very particular microenvironment to culture. This inability to culture a large number of microbes important to the environment under standard laboratory conditions has motivated the development of metagenomics (sequencing microbial communities directly from the environment, for example host tissue or soil,

to study dynamic species interactions and diversity is being called environmental genomics or metagenomics (Tringe and Rubin 2005)). Recent studies emphasize the insights to be gained from metagenomic studies. Assembly of environmental microbial sequences from acid mine drainage biofilms is one of several recent metagenomic projects that illustrates that microbial community genomes can be reconstructed to high completeness given sufficient coverage (Tyson, Chapman et al. 2004). Sogin et al, surveyed the deep sea to show that current sequence databases represent only as small fraction of global microbial diversity (Sogin, Morrison et al. 2006). The Sargasso Sea metagenomics survey revealed a substantial amount of phylogenetic diversity and complexity, identified 1.2 million genes and sampled from an estimated 1,800 bacterial species (Venter, Remington et al. 2004). Three new investigations from the Sorcerer II Global Ocean Sampling expedition have enhanced this dataset, which now includes 6.3 billion base pairs (Rusch, Halpern et al. 2007; Yooseph, Sutton et al. 2007). Metagenomics shows us environmentally relevant protein frequency of occurrence and diversity and that, when we consider the planet-wide diversity of microbial ecology, we have just scratched the surface, with respect to diversity, of microbial genomes and proteomes.

1.2.2 Proteome Annotation

Give the genome the next step is to predict proteins, functional RNAs and other transcribed regions; we will only discuss annotation briefly. Methods for annotating proteomes are still evolving, but generally rely on a mix of sequence-

similarity, protein-domain or protein family searches (such as COG and Pfam). Structure prediction based methods for genome annotation are emerging (Bonneau, Baliga et al. 2004; Malmstrom, Riffle et al. 2007) and rely on fold recognition and de novo structure prediction to extend the reach of our ability to detect distant homology (structure similarity is conserved across a greater evolutionary distance than sequence similarity). Methods for solving protein structures experimentally remain costly and a mix of experimental structural biology and computational structure biology are likely going to lead to prokaryotic genomes characterized at the protein 3D structure level to high levels of completeness. Another promising note is that as more sequences are added to the databases our ability to find sequence based homology via intervening sequences (e.g. via multiple iterations of PSI-BLAST) also increases.

1.2.3 Transcriptomics

Transcriptomics is the measurement and study of the properties and dynamics of all mRNA transcripts in the cell (the transcriptome). There are a variety of tools used to measure transcriptomes, the most common being the microarray. All such tools are high throughput methods for detecting and measuring the expression level, or relative abundance of mRNA transcripts, for every gene within the cell, and results in a snapshot of all the genes present at one time in the cell for a given condition. The methods measure the abundance of RNAs, which is a convolution of the rate of synthesis, transport, and degradation. Two common goals of transcriptomics are to identify genes that are differentially expressed and recognize patterns in gene

expression that correlate with the phenotype. The main technologies used to explore this are DNA microarrays and serial analysis of gene expression (SAGE). SAGE quantifies transcript levels by sequencing and counting cDNAs converted from small unique tags of samples of RNA (Velculescu, Zhang et al. 1995). Parallel gene expression analysis is typically done by either one-color oligonucleotide arrays from Affymetrix (GeneChip) or NimbleGen, or by two-color spotted/printed arrays that can be oligonucleotides, cDNAs, or ESTs printed onto a glass slide. The physical microarray consists of probes (complementary to the RNA being measured), the oligonucleotide or cDNA, printed (in the case of cDNA arrays) or built (for oligo arrays) onto a glass slide or silicon chip. The array is perfused with the cell extracts of RNA tagged with a fluorescent dye (Cy3, Cy5); labeled RNAs thus hybridize to the DNA probes. Ideally it is the specificity guaranteed (excepting cross-hybridization) by reverse complementarity that is core to all microarray technologies (alas, nothing similar to reverse complementarity exists for proteins). Lastly, fluorescent intensities are read to measure the relative abundance. It is therefore important to design the experiment correctly for the comparison to be made. Data is collected by exciting the fluorescent dye tagged RNA and scanning the image. Array scanners usually have software that automatically scans the image, locates the spots, and computes the intensities. The intensity data is converted into numerical data that can then be further analyzed statistically to identify differentially expressed genes. RNA-seq – a more recent technique that leverages next-generation deep-sequencing to generate highly

precise assays that are more accurate than previous technologies - is also quickly emerging as a transformative technology (Wang, Gerstein et al. 2009). Expression profiles can be compared among different cells or tissues (e.g. cancerous versus non-cancerous), time points, and perturbations. Clustering microarray data was an important development (hierarchical clustering, k-means, self-organizing maps) for identifying patterns of co-expressed genes.

1.2.4 Proteomics

Proteomics is a very large and expanding field with a large diversity of aims and corresponding techniques. Recent advances have allowed identification and quantification of all of the proteins that exist in a cell, their abundance, post-translational modifications, interactions, localization, and modifications. However, determination of an organism's proteome is difficult due to the complexity of the large number of proteins and their modifications (Bray 1995). We focus on studies that aim to complete the characterization of the proteome by identifying and quantifying all of the proteins encoded by the genome. The main methods for quantifying, characterizing and profiling proteins and complexes are: two-dimensional gel electrophoresis (2DE), mass spectroscopy (MS), matrix-assisted laser desorption-ionization time-of-flight (MALDI-TOF/MS) and other combinations of MS (LC-MS, GC-MS, etc.)(Aggarwal and Lee 2003). Advanced protein array technology can assay protein activity as well as identifying protein-protein and protein-DNA interactions, here we focus on MS based proteomics (Poetz, Schwenk et al. 2005; Vemuri and Aristidou 2005).

Recent developments have included the development of methods for measuring relative protein levels (proteome wide) by incorporating stable isotopically labeled reagents into multiple samples (by cell culture, in SILAC, and by labeling with reagents in ITRAQ, ICAT) (Gygi, Rist et al. 1999; Zhang, Spellman et al. 2006). In these experiments each sample is labeled with a reagent containing different numbers of stably incorporated heavy isotopes and MS is simultaneously performed on multiple samples. These methods (e.g. SILAC, ICAT, ITRAQ) promise to provide proteome wide measurements analogous to multi-color microarrays. Many technical challenges remain, but mass-spec based proteomics is currently central to many functional genomics projects, and with inbound improvements in resolution, reliability, cost (as well as improvements in surrounding methods such as reagents and fractionation steps) we will only see the importance of these technologies increase.

1.2.5 Techniques for measuring protein-DNA and protein-protein interactions.

Proteins function as networks of interconnected components, involving networks composed of protein-protein and protein-DNA and protein-RNA interactions for the cell overlaid to form an overall network for a given organism (Ge, Walhout et al. 2003). Techniques for measuring such interactions are thus highly relevant to prokaryotic functional genomics projects.

High-throughput interaction mapping methods have been developed for measuring all three of these interaction networks. For example, yeast 2-hybrid (Y2H) (Walhout and Vidal 2001) and chromatin immunoprecipitation (ChIP-chip) assays are

methods for identifying protein-DNA interactions and co-immunoprecipitation (co-IP) is used to identify protein complexes from cell extracts. Chromatin immunoprecipitation (ChIP-chip) assays aim to identify the specific regions of the genome a given protein binds. Proteins that interact with DNA will, by this procedure, enrich segments containing high affinity binding sites for these proteins. Introduced in 2000 and 2001 by 3 papers that reported its first successful use, the general goal of ChIP-chip is to use chromatin immunoprecipitation to help identify the upstream binding sites for a given transcription factor. To accomplish this, the general strategy is as follows. Once the transcription factor protein under consideration has been bound either *in vivo* or *in vitro* to its DNA target, it is cross-linked to the DNA target, often with formaldehyde, which can easily be unlinked with heat. After cross-linking, the DNA is lysed, usually by sonication and the protein-DNA complex is then immunoprecipitated using an anti-body specific to the transcription factor being studied, allowing the cross-linked protein-DNA complex to be isolated. After unlinking the transcription factor from the DNA, the DNA fragments are PCR amplified and labeled before finally being evaluated with microarrays to identify enriched regions of the genome that correspond to binding regions for the transcription factor. Similar to ChIP-chip, a newer technique called chromatin immunoprecipitation, followed by sequencing, or ChIP-seq, improves upon these previous techniques by leveraging next-generation sequencing technologies to allow

for high-throughput assays of binding sites, while providing far greater resolution than ChIP-chip (Park 2009).

1.3 *Caulobacter crescentus*

The non-pathogenic oligotroph *Caulobacter crescentus* is a Gram-negative [alpha]-proteobacterium that lives in aquatic environments; for the remainder of this section, we will refer to it as *Caulobacter*. Morphologically, *Caulobacter* exhibits 3 distinct phenotypes. The first, referred to as swarmer cells (SW cells for the remainder), are motile, rod-like cells that have in one pole both a flagellum, as well as two type IV pili adjacent to the flagellum. Due to an as of yet unknown signal, an SW cell will metamorphose into a stalked, or ST, cell, during which the pili are retracted and the flagellum is ejected, replaced by a 'stalk' that is formed from a thin extension of the cell wall that can help serve as an anchor for the new ST cell (Ausmees and Jacobs-Wagner 2003; Skerker and Laub 2004; Holtzendorff, Reinhardt et al. 2006).

At the same time as this SW to ST cell transformation, chromosomal replication, which had been repressed in the SW cell, is initiated from a single origin of replication and the cell enters S phase. Following the completion of the chromosomal replication, the two copies are sequestered to the two polar halves of the pre-divisional (PD) cell, a new flagellum and pili are generated on the pole opposite of the stalk, and a diffusion barrier develops separating the two polar halves. Once cell division is complete, yielding both an SW and an ST cell, chromosomal replication is reinitiated in the ST cell, while the new SW cell will relocate via chemotaxis, with

chromosomal replication inhibited in it until it differentiates into an ST cell and the entire process reinitiates.

The key observations to draw from the *Caulobacter* cell cycle are: 1) its asymmetrical nature – as it yields 2 morphologically different daughter cells, and 2) the replication process yields exactly 2 daughter cells (Skerker and Laub 2004). In contrast, E. coli cell division in logarithmic phase can replicate the genome up to 4 times before cell division occurs (Skerker and Laub 2004). As cell division in *Caulobacter* yields exactly 2 daughter cells, it exhibits a periodicity that lends itself well to the examination of the bacterial cell cycle. In addition, the asymmetric nature of its cell cycle allows researchers the opportunity to study bacterial cell differentiation – an aspect shared with many other bacteria such as *Bacillus subtilis* (Skerker and Laub 2004). However, as this asymmetry is accomplished via asymmetric localization of proteins, a good portion of the current research directed at deciphering this process uses lab techniques directed at study of single genes (such as localization studies using green fluorescent protein, GFP). We will outline/review both systems-wide studies employing microarrays, proteomics and ChIP-chip (as is the mandate of this chapter) alongside studies aimed at determining the function of small numbers/single genes. Thus, our goal is to illustrate how systems-level techniques have been used alongside these more focused studies to successfully identify the regulation of the cell cycle in *Caulobacter*.

1.3.1 A first application of genome-wide expression profiling to *Caulobacter*.

The first systems-level examination of *Caulobacter's* RNA expression during its cell cycle was reported by Laub *et al.* in late 2000 (Laub, McAdams et al. 2000). Interestingly, this was reported in advance of the publishing of *Caulobacter's* complete genome which was published 3 months later by Nierman *et al.* in 2001 (Nierman, Feldblyum et al. 2001). As such, the cDNA microarrays they used did not cover the entire set of ORF's in the *Caulobacter* genome, however they did represent 2966 predicted ORF's, corresponding to nearly 80% of the 3767 that would be reported by Nieman *et al.* Sampling every 15 minutes over the complete 150 minute cell cycle progression from SW cell to ST cell and final asymmetric cell division, Laub *et al.* identified 553 cell cycle regulated genes, 72 of which had been previously identified using earlier genetic techniques. Clustering these cell cycle-regulated genes using self-organizing maps (SOM's), Laub *et al.* discovered that these were organized into sets of functionally associated genes that were induced in synchronization with the various events of the cell cycle. These included coordinated sets of genes involved in DNA replication and cell division, protein synthesis and polar morphogenesis. Significant among these included homologs of the E. coli cell division genes *ftsI*, *ftsW*, *ftsQ*, *ftsA*, and *ftsZ*, the gene for the tubulin-like GTPase FtsZ, an essential protein for cell division. Additionally, 16 histidine kinases were among these cell cycle regulated genes, of which only 4 at the time had been characterized, these being CheA, DivJ, CckA, and PleC.

1.3.2 Laub, McAdams *et al.*, 2000 – probing the CtrA regulon.

In addition to this time course expression profile, Laub *et al.* also explored the regulon of CtrA, a member of the two-component response regulators that had already been identified using earlier genetic techniques to be a master regulator of the *Caulobacter* cell cycle (Ausmees and Jacobs-Wagner 2003). This was accomplished by comparing the expression profiles of wild-type *Caulobacter* with those of a temperature sensitive mutant, revealing 144 differentially expressed gene transcripts as a result of CtrA expression. To identify which of these were directly regulated by CtrA, Laub *et al.* used MEME (Bailey and Elkan 1994) to construct a consensus profile of known CtrA binding sites and then used this profile in conjunction with the expression data to identify several previously unknown genes under direct CtrA regulation, including *divK*, a single domain response regulator. Finally, Laub *et al.* compared the mRNA expression of wild-type *Caulobacter* with another that contained an allele that produces a form of CtrA that is both proteolysis-resistant and constitutively active, resulting in cell cycle to arrest at the G_1 (SW) stage. From these assays, they were able to identify a nearly 70% overlap with those genes differentially expressed in the temperature sensitive mutant.

These findings were partially validated in a 2002 paper where CtrA targets were identified by performing chromatin immunoprecipitation with microarrays (aka ChIP-chip or ChIP-on-chip). Interestingly, this was one of the first papers to use ChIP-chip data and for this reason, see § 1 for further discussion of ChIP-chip. Using

this, then new, ChIP-chip method, Laub *et al.* (Laub, Chen et al. 2002) identified 138 regions enriched for CtrA binding; the 196 genes flanking these regions were then considered likely targets of CtrA. Of these, 116 had been assayed by the microarray expression profiling reported by their earlier paper, as well as new expression profiling they performed of a ctrA temperature-sensitive mutant over a 4 hour time period (longer than the 2.5 hour cell-cycle) that was aimed at identifying CtrA-dependent genes (including those not involved in the cell-cycle). Combining these three data sets together allowed Laub *et al.* to identify 55 CtrA binding sites that corresponded to 34 individual genes and 21 putative operons yielding a total of 95 genes. Among these included five genes involved in cell division and cell wall metabolism, 14 regulatory genes, and 29 polar morphogenesis genes, with the remaining 47 either unknown (25) or not discussed (22). Notably, these also included *ccrM*, a methyltransferase previously known to be under CtrA regulation, as well the gene responsible for producing S-adenosylmethionine (SAM), the substrate used by CcrM for methylation. In addition, they also confirmed other prior results including those that showed CtrA had multiple binding sites in the origin of replication, as well as directly regulated a number of the main genes responsible for cell division, including *ftsA, ftsQ ftsW* and *ftsZ*.

1.3.3 DivK:

Soon after these global characterizations of CtrA effects, Hung and Shapiro (Hung and Shapiro 2002) described the impact of the single-domain response

regulator, DivK, on the *Caulobacter* cell cycle by using a cold sensitive, *divK-cs*, strain. They discovered that when grown at the restrictive temperature, the *divK-cs* strain developed into long, filamentous stalk-like cells. A return to the permissive temperature allowed these cells to recover morphologically, as cell division was permitted to proceed, indicating the cell cycle of the *divK-cs* strain had been halted at the G_1-S stage by the restrictive temperature. To further explore this behavior Hung and Shapiro next used cDNA microarrays to characterize the mRNA expression profiles of the *divK-cs* strain during growth in both the restrictive and permissive temperatures. From these, they discovered that many of the *Caulobacter* cell cycle genes, including those involved in DNA replication, as well as pili and flagellar synthesis, were repressed during growth in the restrictive temperature, but became induced following the return to the permissive temperature. Combining these new results with the prior understanding that CtrA must be proteolyzed in order for DNA replication to initiate, they next performed a series of immunoblot and pulse-chase analyses to examine CtrA quantities in the *divK-cs* strain. From these experiments, they discovered that at the restrictive temperature, the *divK-cs* strain failed to proteolyze CtrA, thus preventing the initiation of DNA replication, leading Hung and Shapiro to conclude that DivK is requisite for CtrA proteolysis. While the exact mechanism by which DivK mediated CtrA proteolysis was still unclear, Hung and Shapiro, noting that *divK* had been shown to be part of the CtrA regulon, further concluded that the two participate in a regulatory circuit with each other.

1.3.4 Dissection of CckA's global effect

Shortly following these reports on the role of DivK in the *Caulobacter* cell cycle, Jacobs *et al.* (Jacobs, Ausmees et al. 2003) described the results of a series of experiments performed to elucidate the effects of CckA, a histidine kinase, upon the phosphorylation of the CtrA response regulator. As phosphorylation of CtrA is one of the mechanisms by which CtrA activity is regulated and earlier studies had indicated CckA has a role in phosphorylating CtrA, the goal of their study was to explore CckA's role in regulating CtrA activity. As their initial step, Jacobs *et al.* used microarrays and gel electrophoresis to compare the RNA and protein expression profiles of a *ctrA* temperature sensitive mutant strain with those of a temperature sensitive mutant strain for *cckA*. Discovering that RNA and protein expression was virtually identical in both strains, Jacobs *et al.* next used ^{32}P radiolabelling and immunoprecipitation with a *Caulobacter* wild-type strain to illustrate that phosphorylated CtrA and CckA (CtrA~P and CckA~P) possessed nearly matching patterns of expression during the cell cycle. Subsequent viability studies illustrated that while a Δ*cckA* mutant strain was unviable, it could be rescued via a phosphorylation-independent *ctrA* mutation, providing evidence that suggested CckA was crucial for providing CckA~P mediated phosphorylation of CtrA. A final test comparing RNA expression of a Δ*ctrA*Δ*cckA* double mutant strain with that from a Δ*cckA* strain, revealed nearly identical expression of the cell cycle-regulated genes for both strains. As such, Jacobs *et al.* concluded from all these tests that CckA is a

required regulator for CtrA phosphorylation and subsequent activation, though they were unsure of what the exact mechanism for this regulation is.

1.3.5 The cell cycle circuit circa 2004:

Thus, from the results of these systems-level experiments, along with those from other non-systems level studies of *Caulobacter* proteomic localization, a regulatory circuit centered on CtrA that governed *Caulobacter's* cell cycle gradually began to emerge by early 2004. For example, it was understood that CtrA was expressed at high levels during the SW cell (or G_1) stage, but was quickly proteolyzed by a ClpXP-dependant process during the transition to an ST cell. As a result of the decrease in CtrA in the cell, it was understood that the CtrA-controlled inhibition of DNA replication is released, allowing for replication to begin. Additionally, it was also understood that expression of *ctrA* was induced shortly following the initiation of replication, however, there was still confusion about the transcription machinery driving this (Skerker and Laub 2004).

Specifically, by 2004 it was understood that as the levels of CtrA increase in the cell, CtrA acts to repress transcription from a weak upstream promoter, CtrAP1, while also activating expression from a stronger upstream promoter CtrAP2. It was still unclear, though, what the exact mechanism was behind the expression of either of these two promoters. For example, it was understood that *ctrAP1* could only be expressed during the short window of replication when the new daughter strand is unmethylated. Furthermore, it had been discovered that newly expressed CtrA is

quickly phosphorylated into its active form, CtrA~P which subsequently induces expression of the CcrM methyltransferase that methylates the daughter strand. In so doing, CtrA~P inhibits further activity from the *ctrAP1* promoter. However, it was not yet clear what transcription factor induces the transcription from *ctrAP1* (Skerker and Laub 2004).

It was also understood that the newly produced CtrA was phosphorylated (CtrA~P) and that in the stalked portion of the PD cell, CtrA was again proteolyzed by a ClpXP-dependant process, allowing DNA replication to continue. However, while the phosphorylation of CtrA was understood to be related to CckA phosphorylation, as described above, it was still unclear how the two were related. Furthermore, the mechanism that allowed for the localized degradation of CtrA within the stalked end of the predivisional cell was still unknown, though, it was suspected that it was related to the localization to the stalked end of DivJ, a DivK kinase, which as described above, will induce CtrA proteolysis (Skerker and Laub 2004).

1.3.6 Holtzendorff's GcrA - model

The next major step in the exploration of *Caulobacter's* cell cycle was provided by Holtzendorff *et al.* (Holtzendorff, Hung et al. 2004) who reported in 2004 that they had identified GcrA as a second master regulator of the *Caulobacter* cell cycle. In their findings, Holtzedndorff *et al.* discovered that GcrA participates in a regulatory circuit with CtrA where in the first step of this circuit, *gcrA* is transcriptionally repressed by CtrA. However, in the next step of the circuit, the

proteolysis of CtrA upon entry into S phase releases both the CtrA-mediated inhibition of DNA replication, as well as CtrA's repression of *gcrA* expression. This subsequently, allows GcrA to induce *ctrA* expression from the CtrA P1 promoter during the short period while *ctrAP1* is still in its hemi-methylated state on the daughter strand. The circuit is closed when the resulting CtrA~P expression from the activation of the CtrA P1 promoter consequently re-represses *gcrA* transcription, thereby indirectly repressing the activation of the P1 promoter.

While the majority of the methods Holtzendorff *et al.* (Holtzendorff, Hung et al. 2004) applied to identify the role of GcrA were not systems level techniques, such as β-galactosidase assays and immunoblotting, they also performed expression profiling to characterize its regulon once its role had been identified. Using oligo-arrays that contained probe sets for 3761 predicted ORF's, Holtzendorff examined the expression profile of a Δ*gcrA* mutant strain in which a copy of *gcrA* was added under the control of a xylose-inducible promoter. From the expression profile of this strain, Holtzendorff discovered 125 known cell cycle genes that were GcrA dependent. Of these 125 genes, however, only 8 overlapped with the CtrA regulon that had been identified previously by Laub *et al.* (Laub, McAdams et al. 2000; Laub, Chen et al. 2002). Moreover, the fact that the two regulons for CtrA and GcrA consisted of only 30% of the 553 cell cycle-regulated genes Laub *et al.* identified (Laub, McAdams et al. 2000) led Holtzendorff *et al.* to conclude that there were likely to exist additional proteins regulating *Caulobacter's* cell cycle.

1.3.7 Global exploration of the effects of DnaA

The next such cell cycle-regulating protein to be identified was DnaA, the DNA replication initiation factor. At the time, it was already well-established that DnaA played a major role in the initiation of DNA replication whereby binding to specific binding motifs within the origin of replication, called DnaA boxes, it 'melts' the hydrogen bonds holding together the double-stranded DNA, allowing polymerases to access the individual strands. However, in 2005 Hottes *et al.* (Hottes, Shapiro et al. 2005) published results that indicated, similar to both *E. coli* and *B. subtilis*, DnaA also functioned as a transcription factor in *Caulobacter*. Using a *dnaA*-inducible strain (*dnaA* under control of a xylose-inducible promoter), Hottes *et al.* performed expression profiling to identify 40 genes that were DnaA-dependent, 10 of which were known to be GcrA induced. They next used the *in silico* motif-prediction tool, MEME, to identify DnaA boxes within the upstream regions of 13 of these, including *gcrA*, *ftsZ*, and *podJ* which Hottes *et al.* verified by using electrophoretic mobility shift assays. Given these results, Hottes *et al.* concluded that these 13 genes comprised a regulon under the direct transcriptional control of DnaA, with DnaA serving as a promoter for GcrA, FtsZ and PodJ.

1.3.8 Holtzendorff's model of the cell-cycle control circuit

Thus, by this point, we had an emerging model involving 3 master regulators. Starting with active CtrA~P, the dephosphorylation and proteolysis of CtrA releases its repression of DNA replication as well as both the *gcrA* promoter (P_{gcrA}) and its own

weak P1 promoter (*ctrA*P1). The release of this CtrA mediated repression consequently allows DnaA to induce expression of GcrA. In kind, GcrA induces expression of *ctrA* via expression of CtrA's weak P1 promoter, see figure 1. However, as illustrated in figure 1, this newly expressed and phosphorylized CtrA (CtrA~P) subsequently further accelerates its own induction by simultaneously repressing its P1 promoter, while inducing expression of its stronger P2 promoter (*ctrA*P2), with this repression of its P1 promoter occurring via two mechanisms. The first of these being direct repression of *ctrA*P1 by the binding CtrA~P upstream of the P1 promoter. The second occurring when CtrA-induced expression of the CcrM methyltransferase methylates the newly generated daughter strand, and thereby completely suppresses further expression of the P1 promoter by GcrA (Holtzendorff, Reinhardt et al. 2006). However, still left unanswered by this model are questions such as what is the mechanism by which phosphorylated CckA (CckA~P) controls the phosphorylation (and, thus activity) of CtrA. Another is the question of what is the mechanism by which phosphorylated DivK (DivK~P) induces the dephosphorylation and proteolysis of CtrA. A recently work by Biondi *et al.* addresses many of these questions; however, before discussing this paper, we need to make a brief detour to describe the underlying methods and motivation of the work.

1.3.9 Skerker *et al.*'s phosphotransfer method

In their paper, Biondi *et al.* utilized a biochemical phosphotransfer mapping method that had been developed in their lab and described by Skerker *et al.* in 2005

(Skerker, Prasol et al. 2005) which they named phosphotranfer profiling. In this phosphotransfer profiling technique, a soluble kinase domain of a histidine kinase is autophosphorylated with radiolabelled ATP (**[**γ^{-32}]ATP**)** and then incubated in separate *in vitro* experiments with each individual full-length response regulator. Using an added autophosphorylated histidine kinase as a reference, phosphotranfer reactions between the kinase domain and their specific response regulators can be identified when the radiolabel is either depleted from the histidine kinase band or is transferred to the response regulator (which can be identified as a band that corresponds to its molecular weight). Therefore with this method, researchers can systematically examine the complete compliment of response regulators of a given genome for phosphotransfer reactions with a given kinase.

With this phosphotransfer method, Skerker *et al.* (Skerker, Prasol et al. 2005) identified a signaling pathway between the cell envelop proteins CenK and CenR, and soon after, Biondi working with Skerker and others used the method to identify a signaling pathway involved in stalk biogenesis between ShkA and TacA (Biondi, Skerker et al. 2006). Later, noting these open questions regarding CckA and DivK and their relationships with CtrA, Biondi *et al.* (Biondi, Reisinger et al. 2006) set out to determine their roles in *Caulobacter's* cell cycle. Their first step was to definitively determine whether or not CckA had the capacity to phosphorylize CtrA, which they accomplished by using phosphotransfer profiling. From these tests, Biondi *et al.* determined that while CckA could autophosphorylate via the phosphorylation of its

receiver domain (CckA-RD) by its histidine kinase domain (CckA-HK), CckA had no direct role in the phosphorylation of CtrA. Given these results, they suspected there existed an histidine phosphotransferase (HPT) which served as an intermediary between CckA~P and CtrA, as Jacobs *et al.* (Jacobs, Ausmees et al. 2003) had speculated in their initial exploration of CckA's relationship with CtrA.

1.3.10 Identifying the key histidine phosphotransferase

However, none of the predicted genes in the *Caulobacter* genome were annotated as being an HPT. Therefore, using common characteristics of HPT's as criteria, along with the requirement that any such gene must have an ortholog in another genome that also contained orthologs for CckA and CtrA as well, Biondi *et al.* identified a single candidate that they subsequently named ChpT. To validate this hypothesis, they next performed viability as well as expression profiling experiments of a *chpT* deletion strain containing a plasmid with a xylose-inducible copy of *chpT*. From these tests, Biondi *et al.* discovered that in a glucose-only environment this strain was virtually identical to the $ctrA^{ts}$ and $cckA^{ts}$ strains that Jacobs *et al.* had used when grown at the restrictive temperature, strongly indicating a connection between the three genes. Given these results, Biondi *et al.* next returned to the phosphotranspher profiling method to examine the relationship between these three genes. From this method, Biondi *et al.* ascertained that, indeed, ChpT serves as the histidine phosphotransferase bridge between CckA and CtrA.

Furthermore, they also discovered that while CckA is ChpT's only input, ChpT can phosphorylate both CtrA as well as the single-domain response regulator, CpdR, which had only just recently been shown by Iniesta *et al.* to be critical to the localization of CtrA's protease, ClpXP, to the stalked cell pole (Iniesta, McGrath et al. 2006) during the SW to ST transition. Though not discussed in detail here as it was primarily a non-systems level study, this earlier work had demonstrated that CpdR while in its un-phosphorylated state controls the localization of ClpXP to the stalked cell pole, thereby facilitating CtrA proteolysis by ClpXP. Moreover, they too had demonstrated that Cck~P was responsible for CpdR phosphorylation, resulting in ClpXP de-localization from the pole. Thus, by demonstrating that ChpT served as the histidine phosphotransferase between both CtrA and CpdR, Biondi *et al.* had shown the mechanism by which CckA both activated and prevented its proteolysis.

1.3.11 DivK's role in CtrA regulation

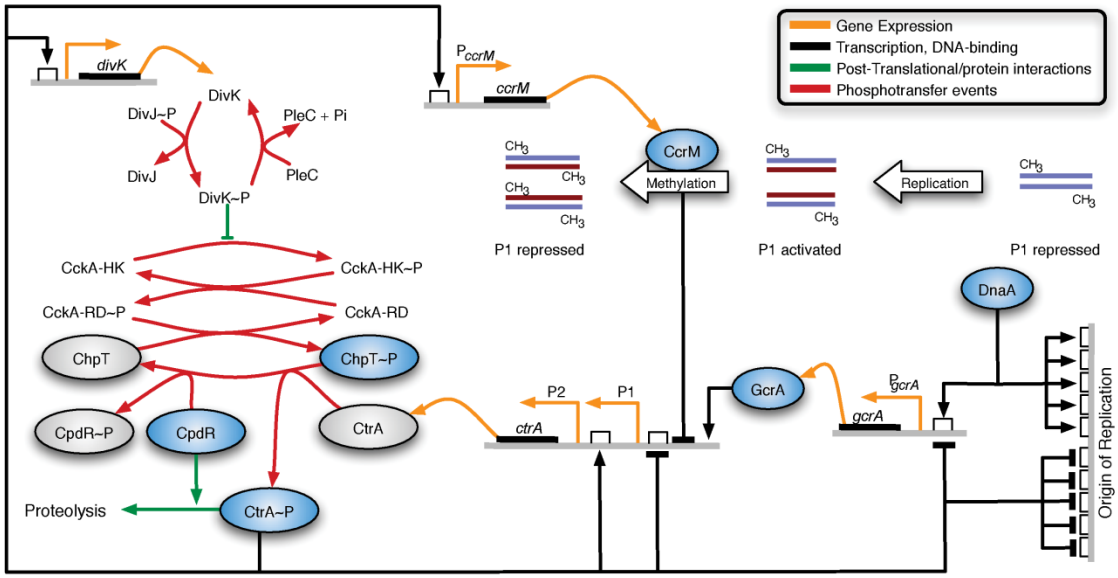
With these results indicating a clear phosphotranspher CckA-ChpT-CtrA pathway, Biondi *et al.* turned their attention to DivK and its role in the dephosphorylation and proteolysis of CtrA. Combining their results along with those of Hung and Shapiro who had shown that a *divK*^{ts} mutant strain was phenotypically similar to a constitutive expressing CtrA strain led Biondi *et al.* to hypothesize that phosphorylated DivK (DivK~P) inhibited CtrA activity by inhibiting activity of CckA. To verify their theory, they compared the CckA~P levels within a *divK*^{ts} mutant strain

with those of a wild type strain, finding a 4-fold increase of CckA~P in the *divK*^{ts} mutant strain, giving evidence that DivK~P inhibited CckA~P.

1.3.12 divK localization impacts cckA

However, as Jacobs *et al.* (Jacobs, Ausmees et al. 2003) had illustrated that CckA~P was also dynamically localized during the cell cycle, Biondi *et al.* performed a long series of GFP localization experiments to determine the mechanisms driving this. In their earlier work, Jacobs *et al.* had shown that CckA was localized to the swarmer pole during G_1 phase, but was subsequently delocalized during the G_1-S phase transition before becoming localized to both poles of the predivisional cell and then later delocalized from in the new stalked cell. While not discussed in detail, Biondi *et al.* used GFP localization experiments to illustrate that indeed, DivK~P triggers CckA to delocalize and inactivate, resulting in a consequent inactivation of CtrA. Furthermore, as previous studies had shown that DivJ, a DivK kinase, localized to the stalked pole, while PleC, a DivK~P phosphatase, localized to the swarmer pole, Biondi *et al.* hypothesized that cell division was crucial for DivK~P induced delocalization of CckA which they also verified using GFP localization experiments. Finally, using a constitutively expression DivK strain, they also demonstrated that the timing of DivK expression, normally mediated by CtrA, was necessary for normal or wild-type cell cycle progression.

Figure 1.1: Caulobacter cell-cycle circuit. Overview of the cell-circuit controlling the Caulobacter cell cycle. Biochemical interactions are as indicated by the key in the figure. Proteins in their activated state are shaded in blue, while those in their deactivated state are shaded in grey. Adapted from (Biondi, Reisinger et al. 2006; Holtzendorff, Reinhardt et al. 2006)

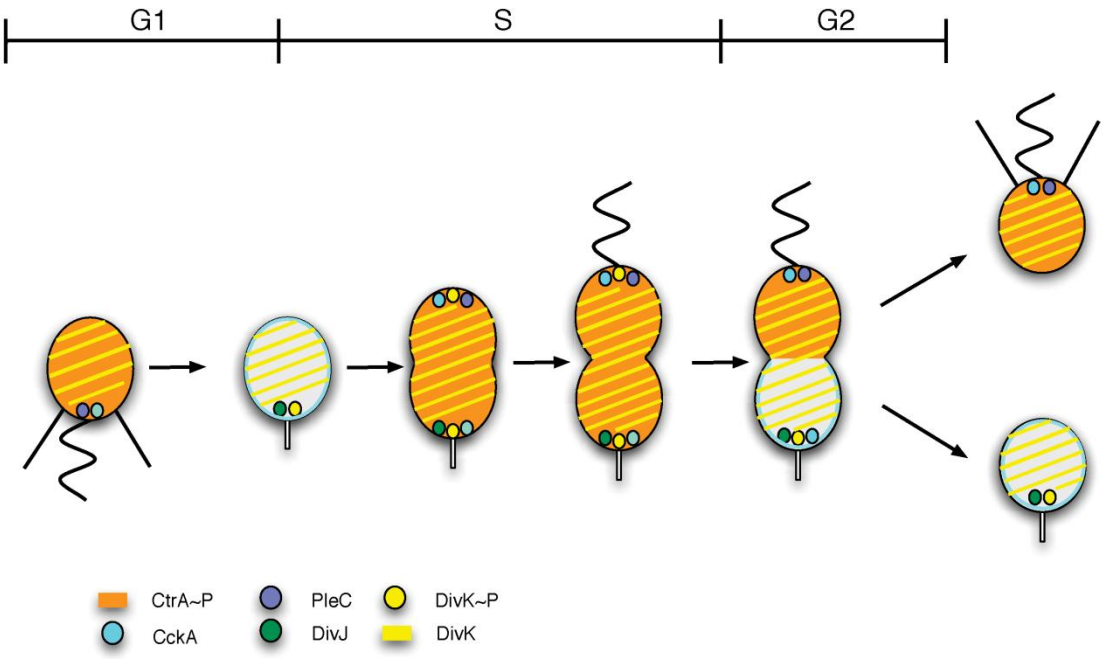


1.3.13 The current model

Thus, from the results of this work has emerged yet two more new feedback loops that drive the *Caulobacter* cell cycle, both of which involve a phosphotransfer cascade that starts with CckA and its activated form, CckA~P, and are determined by the proteomic localization within the cell. In this circuit, as is illustrated in figure 1, CtrA~P induces expression of DivK, which when phosphorylated by its kinase, DivJ, will cause delocalization and proteolysis of CckA, preventing CckA from initiating this cascade. In contrast, when DivK~P is inactivated by its phosphatase, PleC, into its inactive form DivK, this repression of CckA is lifted, allowing CckA to initiate a

phosphotransfer cascade that passes through the histidine phosphotransferase, ChpT. In turn, ChpT~P both deactivates CpdR-mediated proteolysis of CtrA by phophorylating it into its inactive form, CpdR~P, as well as phosphorylating CtrA into its active form, CtrA~P, thereby completing the loop. Significant to understanding this regulatory circuit is to recognize the role that localization plays in determining the activity and inactivity of DivK. Specifically, as illustrated in figure 2, DivK's kinase and phosphatase, DivJ and PleC, respectively are located in the two opposing poles of a late predivisional cell, with the PleC phosphatase in the swarmer pole and DivJ in the stalked pole. As such, with PleC in the swarmer pole inhibiting DivK~P activity, CtrA~P is left unencumbered to repress further DNA replication, while the opposite is the case in the stalked pole, where DivJ induced phosphorylation of DivK and subsequent proteolysis of CtrA~P allows replication to reinitiate.

Figure 1.2: Caulobacter localization. Schematic of the proteomic localization during the cell cycle progression of CtrA~P, CckA, DivJ, PleC, DivK and DivK~P. Adapted from (Biondi, Reisinger et al. 2006).



1.3.14 Future Caulobacter work:

Thus far, the bulk of the current research has focused on the regulatory relationships of the *Caulobacter* cell cycle. However, as Biondi *et al.* identified, the temporal dynamics of expression will need to be an area of further study. Additionally, given its crucial role in the organism, localization and the mechanisms driving this deserve further attention. On this last point, effort has focused on the polar organelle development protein, PodJ, which has been associated with PleC localization, though, neither the exact relationship and mechanism is known at this time (Jacobs-Wagner 2004), nor is that which determines DivJ localization. Further

mapping of *Caulobacter's* stress response and metabolism also present areas for further research as well.

1.4 *Bacillus subtilis*

Bacillus subtilis is one of the best studied model organisms in biology today. *B. subtilis* is a robust, non-pathogenic, aerobic, rod-shaped bacterium in the division Firmicutes; it's a member of the class Bacilli that includes other gram positive genera such as *Staphylococcus*, *Streptococcus*, *Enterococcus*, and *Clostridium*. As a model organism, *B. subtilis* has been studied for over a century (happily predating the earliest pub med article), it was chosen as the best representative of the Gram-positive bacteria, and studying it can help us understand the biology of these organisms. The importance of the Bacillus genus spans biomedicine (w/ several pathogenic spore-forming closely related species), industry (with several economically critical syntheses carried out in Bacillus species) and agricultural (members of the genus are insect pathogens that are used as a bio-insecticide). Bacilli are commonly found in soil, water sources and in association with plants (Kunst, Ogasawara et al. 1997). *B. subtilis* can be manipulated with relative ease since much of its genetics, biochemistry, and physiology are well established. Other important properties that make *B. subtilis* useful to study are: it is naturally competent, can form endospores, contains systems for motility, has a highly diversified set of two-component signal transduction pathways, quorum sensing, and a protein secretion system useful for expression of engineered proteins.

B. subtilis plays an important role in industrial and medical fields and has been used as a platform for the biosynthesis of small molecules and proteins because it is

one of several bacteria that can secrete enzymes at gram per liter concentrations directly into medium (Kunst, Ogasawara et al. 1997). It is known specifically for producing proteases and amylases and is currently being developed as a vaccine development platform (Kunst, Ogasawara et al. 1997; Ferreira, Ferreira et al. 2005). Importantly, its secretion system is more compact (has fewer components) than that of *E. coli* (Yamane, Bunai et al. 2004).

Bacteria commonly use a two-component signal transduction mechanism to respond to changing environmental conditions (Fabret, Feher et al. 1999). These phosphotransfer systems contain two components, a histidine protein kinase that autophosphorylates, and a response regulator protein that elicits a specific response (as described above) (Stock, Robinson et al. 2000; Mascher, Helmann et al. 2006). Homologous versions of this system in several organisms have been shown to initiate and direct various processes such as sporulation, chemotaxis, aerobic and anaerobic respiration, and competence (Fabret, Feher et al. 1999; Ogura and Fujita 2007).

Several species of Bacillus also produce and release chemical signals, called autoinducers or pheromones, which act as cell-cell signaling molecules between bacteria (Miller and Bassler 2001). As population density increases so do these signals, until a threshold is reached and gene expression is modulated. This process is called quorum sensing and controls responses such as competence, sporulation, motility, biofilm formation, and others (Miller and Bassler 2001). Quorum sensing is

an active area of research, as biofilm formation is critical to several biomedical and bio-industrial applications.

1.4.1 Genome sequence and annotation

The complete genome sequence of *B. subtilis* became available in 1997 revealing a sequence of 4.21Mbp containing about 4,106 protein coding genes (Kunst, Ogasawara et al. 1997). Bioinformatics approaches revealed other properties of the genome such as, a large family of putative ABC transporters, a variable G+C ratio of 43.5%, repetitive elements, and an average predicted protein size of 890bp (Kunst, Ogasawara et al. 1997). The *B. subtilis* genome is similar in size to *E. coli* (4.6Mbp) and share roughly 1000 orthologous genes. Comparing these two genomes, which diverged about one billion years ago, will facilitate evolutionary studies of core genes, while comparisons of *B. subtilis* to other more closely related genomes, such as *B. anthrasis*, may provide information about conserved promoter structure and aid in diverse bioinformatics techniques from biclustering to gene finding.

1.4.2 Initial forays into transcriptomics

Exploration of whole genome expression profiles in *B. subtilis* began in 2000 by Fawcett et al, who were able to assign a number of genes to the sporulation process by using nylon-substrate macroarrays, covering ~96% of predicted ORFs, and Hidden Markov models to study the transcriptional profile of early to middle stages of sporulation (Fawcett, Eichenberger et al. 2000). Ye and colleagues, using two-color glass slide arrays, compared mRNA levels from aerobic and anaerobic conditions (Ye,

Tao et al. 2000). The results of these initial genome wide investigations revealed complex expression patterns, including many genes of unknown function with highly different expression under the measured conditions, indicating that much still remained to be learned about the control of spore formations and spore induction/control.

1.4.3 Bacillus stress responses

A number of investigations have focused on the cellular response to stress at the transcriptome level in *B. subtilis* (this so-called stress response is a key focus of several prokaryotic functional genomics projects). Yoshida et al, studied glucose repression by a combined approach of microarray and 2D gel electrophoresis, with a focus on the genes dependent on catabolite control protein, CcpA (Yoshida, Kobayashi et al. 2001). Helmann et al investigated the general stress response to heat shock in order to establish its profile thus allowing it to be compared to other stress response profiles (Helmann, Wu et al. 2001). Nakano et al described the role of Spx as a global transcriptional regulator of disulfide stress conditions (Nakano, Kuster-Schock et al. 2003). Ren et al observed the induction of stress response genes by investigating the growth inhibition mechanism of a natural brominated furanone (Ren, Bedzyk et al. 2004). Also in the search of new antibiotics, Lin et al., determined B. subtilis expression profiles in response to treatment with subinhibitory amounts of chloramphenicol, erythromycin, and gentamicim (Lin, Connelly et al. 2005). Hayashi and colleagues determined that there is a direct interaction, during H_2O_2 oxidative

stress, between PerR, a stress response regulator, and srfA, an operon involved in surfactin biosynthesis (Hayashi, Ohsawa et al. 2005). Allenby et al characterized the phosphate starvation, PhoP, regulon, identifying some new members and a connection to the sigB general stress regulon (Allenby, O'Connor et al. 2005). Ogura et al, investigated the role of RapD, one of 11 Rap proteins that typically inhibit response regulators, and found it to be a negative regulator, in conjuntion with SigX and RghR, of the ComA regulon (Ogura, Tsukahara et al. 2007).

Overall these genomic studies helped to bring in many key proteins that would have been missed, including several proteins never before linked to a known process. Once these proteins are discovered by genomic techniques they are quickly validated and integrated into the aggregate picture of stress response.Furthermore, identification of key genes and proteins has enabled the construction of networks between the various pathways and processes within the cell.

1.4.4 Exploration of Bacillus two-component regulatory systems

As described above, two-component regulatory systems are characterized by a sensor protein (e.g. kinase) and a response regulator protein (e.g. DNA-binding protein). Ogura et al began using whole genome microarray analysis in order to identify the target genes of the response regulators DegU, ComA, and PhoP (Ogura, Yamaguchi et al. 2001). Using the same strategy as Ogura et al, overexpressing the response regulator in mutants for their sensor kinase, Kobayashi et al, further analyzed

24 different two-component regulatory systems (Kobayashi, Ogura et al. 2001). These studies greatly expanded our knowledge of kinase -> target-gene specificity, and interestingly, the role of cross-talk between these sensory systems. For example, they identified many new genes regulated by ComK along with some previously known genes and identified a cellular state they called, the K-state, as a time for the cell to rest and recover from stress that is separate from sporulation (Berka, Hahn et al. 2002). This work was quickly followed up by Ogura et al, who then explored the roles of many ComK regulated genes, in order to better understand competence (Ogura, Yamaguchi et al. 2002). Britton et al, performed a genome wide analysis of sigmaH, which is involved mainly in transitioning from growth to stationary phase, but is also involved in initiation into sporulation and competence (Britton, Eichenberger et al. 2002). Hamon et al., investigated genes involved in biofilm formation that are regulated by AbrB (Hamon, Stanley et al. 2004) whose results led to the discovery of two non-transcription factor gene products, a signal peptidase and a secreted protein, that play an essential role in biofilm formation. Serizawa et al., studied the YvrGHb two-component system and found it to control the maintenance of the cell surface and its proteins, as well as being involved in preventing autolysis (Serizawa, Kodama et al. 2005). Keijser et al investigated the regulatory process and outlined key events of spore germination and outgrowth by microscopy, genome wide expression profiles, and metabolite analysis (Keijser, Beek et al. 2007).

1.4.5 Other uses for microarrays

Several reports have focused on RNAs other than mRNA, such as tRNA, untranslated RNAs, and RNAs involved in the processing of other RNAs. Ohashi et al, examined the modulation of the translation machinery during sporulation, finding in accordance with previous reports that there tends to be a dramatic global decrease in RNA, but that certain ribosomal rRNA and mRNA genes either remain the same or can increase (Ohashi, Inaoka et al. 2003). Dittmar et al aimed to quantify tRNA transcription, processing, and degradation levels on a genomic scale and developed specifically for tRNAs, a microarray and method of selectively labeling them (Dittmar, Mobley et al. 2004). Silvaggi et al., Investigated the small non-translated RNAs involved in sporulation by microarray analysis with a microarray of intergenic regions as probes and a comparative computational analysis that predicts conserved RNA secondary structures (Silvaggi, Perkins et al. 2006).

Earl et al examined 17 *B. subtilis* strains in order to quantify their diversity and identify regions of variability by microarray-based comparative genomic hybridization (M-CGH) (Earl, Losick et al. 2007). M-CGH results in a measure of gene presence or absence by quantifying the relative hybridization efficiencies from two differently labeled bacterial strains. AS, bacterial genomes are dynamic, they found the gene content of their collection of strains to have at least 28% variability, meaning the genes could either have diverged or are missing.

1.4.6 Probing Bacillus with ChIP-chip

ChIP-chip (described above) in combination with transcriptional profiling and gel electrophoretic mobility shift assays has been performed to identify 103 additional genes regulated by Spo0A, the master regulator for entry into sporulation (Molle, Fujita et al. 2003) and many new targets of CodY, a GTP-activated repressor of early stationary genes in *B. subtilis* (Molle, Fujita et al. 2003). Also, a centromere-like element in *B. subtilis* was defined by mapping the binding sites for RacA, a chromosome remodeling and anchoring gene, and identifying 25 high selectivity binding sites (Ben-Yehuda, Fujita et al. 2005).

1.4.7 The *B. subtilis* proteome

The global study of proteomes (e.g. using mass-spectroscopy coupled with multiple seperation strategies) lags behind transcriptome studies in reproducibility, cost and accuracy. Studying the dynamic proteome is confounded by several factors, for example: 1) there is a lack cost effective methods for designing high affinity, high specificity, capture agents for all proteins in a given genome, and 2) several post-translational modifications of a protein can complicate its identification and quantification. The genome of *B. subtilis* contains more than 4100 genes and therefore we expect at least on the order of 4100 gene products. The proteome of *B. subtilis* has been studied for more than 20 years starting with explorations of heat shock proteins (Streips and Polio 1985). Then with the sequencing of the genome, establishment of online databanks, and advances in MS and 2D-PAGE technology, proteome wide

characterizations became possible. In the cytosol of vegetatively growing cells, Buttner et al., first identified over 300 proteins (Buttner, Bernhardt et al. 2001), then Eymann identified 876 proteins (Eymann, Dreisbach et al. 2004). Tam et al., identified over 200 proteins in cells under stress or starvation conditions (Tam le, Antelmann et al. 2006). Finally, Wolff et al, has increased the number of identified proteins to 1395, thus covering over one third of the *B. subtilis* proteome (Wolff, Otto et al. 2006; Wolff, Antelmann et al. 2007). Clearly with slightly more than a third of the *B. subtilis* proteome identified, dynamical characterization of the proteome (both levels of proteins and protein modifications) will reveal a great deal of novel biological information (sequence specific degradation and translational control, specificity and dynamics of modification, etc).

1.4.8 Yeast 2-hybrid investigation of the Bacillus protein interaction network

As described above, yeast 2-hybrid (Y2H) analysis is a widely used method for detecting protein-protein interactions and screens can scale to test whole genomes (Fields and Song 1989). Noirot-Gros et al, made an initial Y2H analysis of DNA replication components in *B. subtilis* identifying 69 proteins with 91 interactions (Noirot-Gros, Dervyn et al. 2002). Their investigation yielded several interesting results that connect DNA replication to diverse cellular processes, including membrane and signaling pathways.

Predictions from the work of Noirot-Gros et al, influenced Meile et al, to perform a larger scale semi-systematic protein localization study for over 100 proteins in *B. subtilis* (Meile, Wu et al. 2006). To accomplish this, they developed a new approach for the rapid construction of GFP fusion constructs. In their study, 110 ORFs were selected, 50 chosen from known DNA replication components identified by previous Y2H screens. The remaining 60 selections were from various functional categories, including some of unknown function, from different functional categories based on annotations from Subtilist, Swiss-Prot, and NCBI. Overall, 90% of the proteins they studied were tagged with GFP with 78% tagged on both the N- and C- ends. In summary, they were able to identify interesting localization patterns for 85 previously un-localized proteins, and thus identified new proteins associated with DNA-replication machinery. The locations of all proteins in the cell, under various conditions, will need to be compiled before there can be a clear picture of the organism at the systems level.

1.4.9 Investigating metabolome changes during sporulation

Clearly the levels of metabolites are important to microbial biology, but methods for measuring the metabolome are much less widely adopted than methods for measuring the transcriptome and proteome. Capillary electrophoresis mass spectroscopy (CE-MS) is a powerful, quantitative tool for the direct and sensitive global analysis of metabolites. Soga et al., were able to determine a total of 1692 metabolites by splitting sample using three purification schemes (one each for cationic

metabolites, anionic metabolites, and nucleotides/coenzyme A compounds) in parallel to separate and subsequently identify metabolites (Soga, Ohashi et al. 2003). To detect as many metabolites as possible they used an instrument wide range of approximately 70 to 1000 m/z. Their novel strategy was lengthy, 16 hours per run, with several runs required, but is highly automated. Soga et al used their metabolomic approach to profile metabolites before and during sporulation. They characterized unknown peaks by combining CE-MS results with bioinformatics and made headway into determining the (partially characterized prior) link between sporulation in *B. subtilis* is and the metabolic network. Thus, revealing possible functional links from some uncharacterized metabolites. The power of their approach was nicely demonstrated by the ability to simultaneously monitor glycolytic, pentose phosphate, and TCA pathway sporulation metabolite responses consistent with previous data. The study showed that metabolite concentrations cannot be accurately resolved by transcriptome analysis and revealed significant changes in metabolites during *B. subtilis* sporulation important for deciphering this important process.

1.4.10 A systems approach to reconstruction of the sporulation control circuit.

As is commonly known, multicellular organisms contain many different types of cells. The mechanism of cellular differentiation is a fundamental problem in biology. Various developmental processes such as cell growth, morphogenesis, cell death, etc, occur in bacteria, with sporulation being a prime example. Sporulation can be considered a developmental process, albeit a simple one, as it is the process by

which an organism differentiates from a vegetative cell type into a completely different cell type, the spore. The fate of each cell type is due to both its particular developmental gene expression program, as well as its interaction with the cell's environment. *B. subtilis*, like many gram-positive, low G+C content, bacteria is known to undergo this transformation, and is among the best studied in this area. Inhospitable environmental conditions cause *B. subtilis* to begin the sporulation process, but it is typically induced in the laboratory by low nutrient conditions, e.g. the removal a carbon, nitrogen or phosphorus source (Piggot and Hilbert 2004). In the beginning of sporulation, a septum forms asymmetrically, near one end of the cell, dividing it into two cells, the larger mother cell and the smaller forespore; the forespore is to become the mature spore. Immediately following septum formation, the two cells have identical genomes but asymetric gene expression programs. In the next stage, the forespore is completely engulfed by the mother cell in a phagocytic-like process. The mother cell then nurtures the endospore surrounding it with proteins that form a spore cortex, and a spore coat. Finally, the mother cell lyses to release the fully developed and remarkably resilient spore (the spore is resistant to heat, UV and γ radiation, and various chemicals and enzymes). When nutrients are again sensed in the environment, the spore can germinate and flourish as a vegetative cell (Setlow 2003).

Various independent transcriptome analyses have elucidated, on a genome-wide level, many of the relationships between genes, including a catalog for sporulation the process of at least 600 genes (Fawcett, Eichenberger et al. 2000;

Britton, Eichenberger et al. 2002; Eichenberger, Jensen et al. 2003; Feucht, Evans et al. 2003; Molle, Fujita et al. 2003). Eichenberger and colleagues utilized an elegant microarray strategy in conjunction with computational, biochemical, and *in vivo* analyses attempting to take transcriptome analysis a step further (Eichenberger, Fujita et al. 2004). Their systems level investigation comprehensibly illustrated a regulatory circuit by integrating data from transcriptomics and genomics approaches thus characterizing the mechanism controlling the cell's decision to sporulate, and the timing of the process by which the spore is assembled.

Transcription in bacteria is mediated by sigma (σ) factors (general transcription factors involved in a large fraction of bacterial transcription initiations). Sigma factors bind to specific promoter regions, and in Bacillus have been shown to be master regulators with sequence specific affinity for separate promoters. There are at least 17 sigma factors in *B. subtilis* but only 6 have a notable role in sporulation (Moszer 1998; Moszer, Jones et al. 2002). Gene expression during sporulation is coordinated by 4 sigma factors σ^E, σ^F, σ^G, and σ^K. The regulatory cascade in the forespore is initiated by σ^F; it includes 48 genes organized in 36 transcription units whose products govern spore morphogenesis and germination properties (Wang, Setlow et al. 2006). After engulfment, σ^G regulates transcription of genes involved in chromosome condensation and equipping the spore for germination. In the mother cell, σ^E begins the cascade and turns on 262 genes (Zheng and Losick 1990; Eichenberger, Jensen et al. 2003; Eichenberger, Fujita et al. 2004). Two of the targets

of σ^E are DNA binding proteins, SpoIIID and GerR (Kunkel, Kroos et al. 1989; Stevens and Errington 1990; Tatti, Jones et al. 1991; Errington 2003; Eichenberger, Fujita et al. 2004). The function of GerR was previously unknown and now has a role as a negative regulator, switching off genes in the σ^E regulon. SpoIIID is interesting in that it acts as a repressor for some genes activated by σ^E and activates additional genes in conjunction with σ^E. SpoIIID is important for activating many coat proteins and especially the genes for an inactive proprotein, pro-σ^K, that ultimately converts to mature σ^K upon reception of an intercellular signal governed by forespore specific σ^G. This signal is important for keeping the separate mother cell and forespore programs coordinated during the morphogenesis (Errington 2003; Hilbert and Piggot 2004). The σ^K regulon includes sets of genes for the spore cortex, structural components of the spore coat and germination (Steil, Serrano et al. 2005), and importantly GerE. Last in the mother cell line hierarchy, GerE, a DNA binding protein, activates a final set of 36 genes and represses about half of the genes activated by σ^K. For example, two cell wall hydrolases are activated that play a role in lysis of the mother cell when spore morphogenesis is complete.

Eichenberger et al compared RNA from mutants in transcriptional regulators suspected/known to control sporulation; using prior knowledge of the sporulation process they were able to construct near-optimal experimental designs for measuring the effects of these perturbed transcription factors. As a result of their transcriptional profiling strategy, two DNA-binding proteins, SpoIIID and GerR, turned on by σ^E

were found to have significant effects on the σ^E regulon. SpoIIID extensively affects the σ^E regulated transcription pattern, influencing over half of the σ^E regulon. This seems to be accomplished by direct interaction, as evidenced by assaying the promoter regions of the modulated genes. Evidence for direct interaction with the promoter regions was obtained first by identifying SpoIIID binding sites with gel electrophoresis mobility-shift assays and DNAse I footprinting. Their application of *in vivo* ChIP-chip revealed many regions on the chromosome that SpoIIID bound that did not include genes not known to be under under its control, and some sites were located within protein coding regions, possibly indicating an architectural role for SpoIIID. Finally, computational binding site sequence analysis was used to find putative conserved motifs in the upstream region of genes regulated by SpoIIID. Analysis of GerR by transcriptional profiling found that no genes that were dependent upon GerR for activation, but many genes were inhibited by GerR. Following SpoIIID in the cascade, the σ^K regulon was delineated by transcriptional profiling and further resolved by computational sequence analysis to identify a conserved motif in the promoters of the σ^K regulated genes. The last regulator in this cascade, another DNA-binding protein, GerE, was found to inhibit the expression of slightly over half of the σ^K regulon and activate at least 36 additional genes at the end of the mother-cell line of gene expression.

A comprehensive program of the mother-cell line of gene expression can be drawn from these results together, see figure 3. The resulting model consits of a

hierarchical regulatory cascade of three DNA-binding proteins (SpoIIID, GerR, and GerE) and two general transcription factors (sigma factors σ^E and σ^K); σ^E begins the cascade by activating transcription of 262 genes. SpoIIID and GerR repress many genes of the σ^E regulon and SpoIIID and σ^E activate 10 additional genes. σ^K activates 75 more genes, and finally, GerE, represses over half of the σ^K regulon and activates 36 more genes. Eichenberger et al., compiled these results into a transcriptional network composed of a linked series of five type-1 feed forward loops (FFLs) (Milo, Shen-Orr et al. 2002; Shen-Orr, Milo et al. 2002; Mangan and Alon 2003). Two of the FFLs are coherent and have the property of being persistence detectors (low pass filters), these may be used to minimized the effect of high frequency noise (Mangan and Alon 2003). Three of the FFLs are incoherent and have the property of producing pulses of gene transcription (Mangan and Alon 2003).

Finally, they performed comparative analyses to determine possible conservation of this spore formation curcuit in other endospore forming bacteria. There are differences in the presence of certain regulatory proteins, for example, *Bacillus* and *Clostridium* contain orthologs for σ^E, σ^K and SpoIIID including conserved sequence recognition domains, but *Clostridium* is missing GerE and GerR. Also, there is variation in the composition of each individual regulon among species, for example: 75% of the *B. subtilis* σ^E regulon have orthologs in *B. anthracis* and *B. cereus* whereas only 40% have orthologs in *Clostridium*, and 50% of the *B. subtilis* σ^K regulon have orthologs in *B. anthracis* and *B. cereus* compared to 20% that have

orthologs in *Clostridium*. They show that this pattern of conservation is consistent with the fact that the σ^K regulon contains many components of the spore's outer surface and that spore surfaces of *B. subtilis*, *B. anthracis* and *B. cereus* are known to be quite different quite different (Chada, Sanstad et al. 2003), the low level of conservation among σ^K regulons may be due to adaptation to an ecological niche. Thus, the sporulation circuit (the regulatory control of the decision to sporulate and the subsequent control of spore assembly) is more conserved than the target protein components (the spore coat proteins). Finally, Wang et al, extended this work by investigating the forespore line of gene expression and synthesized a single model summarized in figure 3 (Wang, Setlow et al. 2006).

1.5 *Escherichia coli*

Discovered in 1886 by Theodore Escherich, *Escherichia coli* is a Gram-negative species of bacteria that inhabit the mammalian gut, specifically the colon or lower intestines. As one of the best studied organisms of the pre-genomic era, *E. coli,* like *B. subtilis* was an early target for sequencing and in 1997, the complete sequence for the K-12 (MG1655) strain, consisting of 4,639,221 base pairs, was completed and reported by Blattner *et al.* (Blattner, Plunkett et al. 1997).

1.5.1 Early systems-wide studies:

Shortly following the completed sequencing of the *E. coli* genome in 1997, later that same year, the first two genome-wide microarray studies of *Sacromyces*

cerevisiae were reported. The first, by DeRisi *et al.* (DeRisi, Iyer et al. 1997), used spotted cDNA arrays to profile the expression changes of yeast during diauxic shift, and then later Wodicka *et al.* (Wodicka, Dong et al. 1997) used 25-mer oligonucleotide arrays from Affymetrix to profile the expression differences of yeast grown on rich versus minimal media. Closely following these initial studies, two early projects were performed to develop microarrays for *E. coli*. The first of these, described by Tao, Busch *et al.* (Tao, Bausch et al. 1999) was a microarray that used nylon membranes and radio-labels for the cDNA; making the experiment essentially a genome-wide northern blot. In contrast, the second project by Wei, Lee *et al.* (Wei, Lee et al. 2001) used the technique developed by Pat Brown to develop a two-color, spotted cDNA microarray on glass slides. Shortly following these initial studies, Richmond, Glasner *et al.* (Richmond, Glasner et al. 1999) compared these two microarray technologies by comparing the expression profiles reported for two well-studied environmental responses. Specifically, in their comparison, they used both technologies to explore the RNA expression profiles of *E. coli's* heat shock response, as well as exposure to the *lac* operon inducer, isopropyl-b-D-thiogalactopyranoside (IPTG). In their results, the authors reported that both microarray varieties indicated expression differences for genes in both the *lac* and melibiose operons for the IPTG tests, both of which were expected given previous published experimental work. A sizeable intersection between the genes that the two technologies reported as being induced during the heat shock response was found; 62 of the 77 genes reported by the

nylon membrane microarrays were also identified as being induced by the glass cDNA arrays. In contrast, the authors reported little overlap between the genes the two technologies identified as being down regulated. Despite this discrepancy, the authors concluded that glass microarrays were more reproducible and therefore recommended it as the preferred method.

1.5.2 Overview of early *E. coli* microarray studies

Shortly after these initial projects, Selinger, Cheung *et al.* (Selinger, Cheung et al. 2000) introduced the first Affymetrix chips designed for *E. coli* in a paper that compared the expression profiles of *E. coli* during logarithmic growth and stationary phases (on a rich medium). In addition to probes for the 4290 predicted ORF's in the E. coli genome, these new chips also contained probe sets for non-coding RNA's such as tRNA's and ribosomal rRNA's. While there was some discussion of results of the biological findings of their experiment, the focus of the paper, not surprisingly, was on the technology and the advantages offered by using short oligos, rather than whole cDNA's. The primary of which being lower cross-hybridization. However, it's important to also note that as these were still the early days of microarray design, these chips had the design flaw of failing to randomize the location of the probes on the chip. For example, the top half of the array contained all the probe sets for ORF's and untranslated RNA's, while all the tRNA and rRNA probe sets were all located along the bottom edge of the chip. As described by Qian and Kluger(Qian, Kluger et al. 2003), chips that manifest such a linearity in probe location are prone to biasing the

expression levels reported when there is an uneven distribution of RNA in the solution that is hybridized to the chip.

Following the announcements of these new *E. coli*-specific genome-wide microarrays, they were quickly adopted by researchers who began applying them in systems-wide studies of various environmental and metabolic responses. While many of these responses had already been the subject of earlier studies using previously existing genetic techniques, for most this was the first time they had been studied at a genome, or systems-wide, level. Early examples include explorations of the SOS response (Courcelle, Khodursky et al. 2001), metal-ion tolerance (Brocklehurst and Morby 2000), osmostress (Weber and Jung 2002), and adaptation to acetate and propionate (Polen, Rittmann et al. 2003). More recent examples of stress-response examinations include inhibition of cell division (Arends and Weiss 2004), anti-microbial peptides (Hong, Shchepetov et al. 2003; Tomasinsig, Scocchi et al. 2004), and cadmium toxicity (Wang and Crowley 2005).

These early studies were primarily descriptive in nature but were also key in motivating the development of several analysis techniques suited to these genome-wide measurement technologies. In this sense, they can be viewed as foundational as they reported systems wide expression differences, from which new hypotheses could be drawn that could be validated and further explored in later studies. For example, it was shown by Barbosa and Levy (Barbosa and Levy 2000), and later partially validated by Pomposiello, Bennik and Demple (Pomposiello, Bennik et al. 2001) that

there was a previously unknown overlap between the multiple antibiotic resistance and oxidative stress regulons (MarRA and SoxRS, respectively), a finding that would not have been easily identifiable using previous experimental methodologies. Another example would be the results reported by Zheng, Wang *et al.* (Zheng, Wang et al. 2001) who discovered an additional overlap for the SoxRS response regulon with that of the OxyR response regulon.

1.5.3 System level studies of regulatory interactions governing the glutamate dependent acid response (AR):

One example of how systems level biology assisted in the study of *E. coli* focused on and helped elucidate a complex network of regulatory interactions governing its glutamate-dependent acid resistance or response (AR). While the ability of *E. coli* to develop acid resistance was first observed over 50 years ago, it was not until 1995, during the pre-genomic era, that it was discovered that there exist 4 distinct systems within *E. coli* for acquiring AR (Lin, Lee et al. 1995; Foster 2004). These include one system that is repressed by glucose (and only functions in its absence), another that is dependent on arginine, as well as one more that is dependent upon lysine, and finally a fourth that is glutamate dependent (with the last three functioning in environments that include glucose). Of these, the glutamate-dependent system is the most effective, the best studied, and the one upon which systems biology has had the most impact.

The first systems-level foray into the understanding of *E. coli*'s glutamate-dependent AR was performed by Hommais *et al.* (Hommais, Krin et al. 2001), though the original goal of the work was to explore the role of *E. coli*'s nucleoid-associated protein, H-NS. Using nylon membrane microarrays to compare the RNA expression profiles of wild type and an *hns* mutant strain, Hommais *et al.* identified expression differences for genes involved in processes including those that were then known to be involved in osmolarity and acid resistance. Note, for the majority of the observed gene expression differences, the expression was induced or elevated in the Δ*hns* strain, leading them to conclude that H-NS was a repressor of gene regulation. Among the genes up-regulated in the Δ*hns* strain included *evgA*, the regulator from the EvgAS two-component system, as well as *gadA* and *gadB*, the two glutamate decarboxlyases known to be required for acid resistance, as well as *gadC*, the GABA/glutamate antiporter required by AR. Noting the induction of the genes involved in acid resistance, Hommais *et al.* next explored the impact of the Δ*hns* upon acid resistance. Comparing the effects of arginine, lysine and glutamate acid stress upon both the Δ*hns* and the wild-type strains, Hommais *et al.* discovered that the Δ*hns* strain only conferred a resistance when in the presence of glutamate. Based on these results, Hommais *et al.* used plasmid-induced overexpression strains to identify *yhiX* (later renamed to *gadX*) as a gene whose overexpression will impart acid resistance, leading them to conclude that it was likely to be a transcription factor necessary for glutamate-dependent AR.

A year after Hommais *et al.* published their results, Masuda and Church (Masuda and Church 2002) set out to explore the regulon of the EvgA response regulator protein in the EvgAS two-component signaling system, with the hope that characterizing the response would help identify EvgA's functional role. To accomplish this, they used *E. coli* specific chips from Affymetrix to compare the expression profiles of EvgA knockout and overexpressing (via a transfected plasmid) strains to identify potential target genes of the EvgA regulon. Now, as EvgA's functional role was still unclear at the time, they also developed a similar set of strains from an *acrAB* knockout strain, as it had been reported by Nishino and Yamaguchi (Nishino and Yamaguchi 2001) that EvgA overexpression would bestow antibiotic resistance to this strain. Comparing the expression profiles of all these strains, Masuda and Church were able to identify 79 genes with induced expression as well as another 24 that were repressed or reduced.

1.5.3.1 Exploring the genes necessary for acid resistance

Of these, they noted that several were genes known to be involved in conferring acid resistance, motivating their exploration of the effect of EvgA overexpression upon the organism's response to acid stress. Thus, to verify their hypothesis, they performed survivability tests for *E. coli* in a low pH environment and discovered that, as they suspected, EvgA overexpressing strains were, indeed, acid resistant. Given this validation, the authors then performed another series of survivability experiments using knockout strains for each of the genes most strongly

induced by EvgA overexpression. From these tests, Masuda and Church were able to identify 3 genes, *ydeO*, *ydeP*, and *yhiE* (later renamed to *gadE*) that were required for the acid response of *E. coli* in logarithmic growth, while also discovering that *gadE* is key to the organism's acid response while in stationary phase.

1.5.3.2 Identifying EvgA's role in acid resistance

Along with their findings for the acid stress response, Masuda and Church also performed a similar set of experiments to explore the drug resistance that was induced by EvgA overexpression in the Δ*acr* strains. In so doing, they were able to identify the YhiUV efflux pump and the TolC outer membrane channel proteins as being key to the Δ*acr* strain's drug resistance during EvgA overexpression. However, in later tests, they also observed that EvgA overexpression could not confer drug resistance for strains without this Δ*acr* deletion. For this reason, combining their observations about both the drug and acid shock response, Masuda and Church concluded that EvgA's primary role is not in coordinating the organism's drug response, but instead its acid shock response.

1.5.3.3 Expanding the list of AR regulators

In a similar project, performed nearly concurrently with that done by Masuda and Church, Nishino *et al.* (Nishino, Inazumi et al. 2003) partially validated Masuda and Church's findings. For example, they too recognized the induced expression of genes known to be involved in the organism's acid response and thus tested the effects of EvgA overexpression on the survivability of the organism. While they also

observed an increased resistance to acid shock, they however did not pursue this further and thus did not identify the critical roles of *ydeO*, *ydeP*, and *gadE* in its acid response.

In contrast to the Masuda *et al.* and Nishino *et al.* investigations, the goal of Tucker *et al.* (Tucker, Tucker et al. 2002) was specifically to explore *E. coli*'s glutamate-dependent AR. To accomplish this, they used nylon membrane chips to compare the expression profiles of *E. coli* during logarithmic growth in glucose-rich media of varying pH, with pH's of 7.4, 5.5 and 4.5. Of the genes they identified as being induced were included 6 genes that were either known or suspected of being transcription factors, including 4 in the *hdeA-gadA* region with these being *yhiF*, *gadE*, *gadX*, and *gadW*[1]. Similar to Masuda and Church, to further explore the roles of the induced genes, Tucker *et al.* generated gene knockout strains and performed survivability tests on these. Focusing on 7 genes in the *hdeA-gadA* region, they discovered that only one, *gadE*, was critical for the organism to become acid resistant and for this reason, they concluded that it likely was an AR transcription factor.

Following up their initial study, Masuda and Church (Masuda and Church 2003) developed a set of deletion and overexpression *E. coli* strains for each of the *ydeO*, *ydeP*, and *gadE* genes they identified in their earlier study. From the results of a series of susceptibility tests for these strains, they hypothesized that there exists a set

[1] Note, in the text the last 3 genes are referred to as *yhiE*, *yhiX*, and *yhiW*, but were later renamed using the *gad* prefix once they were recognized as being members of the glutamate AR regulon. For the sake of clarity and consistency, we use their current naming scheme rather than those used in the original text.

of cascading regulatory interactions where EvgA induces YdeO which subsequently induces GadE. To validate this, they used a combination of *in vitro* and *in silico* systems-level methods. Specifically, via the expression profiles of a new set of deletion mutants for the *ydeO* and *evgA* genes, individually and in combination, Masuda and Church identified 2 distinct regulons. One of these being induced directly by EvgA expression (including YdeO), while the other was indirectly induced by EvgA via YdeO. To further validate EvgA induction of YdeO, Masuda and Church used the *in silico* motif discovery tool, ALIGNACE, (Roth, Hughes et al. 1998) to identify a putative 18bp binding motif in the upstream regions of the genes that they predicted to be induced directly by EvgA. Next, the putative binding sites in the upstream regions of *ydeP* and *b1500* (a gene upstream of *ydeO* that they suspected formed an operon with it) were mutated in a new set of *E. coli* strains that were subsequently subjected to acid resistance tests. The results from these experiments indicated that the putative EvgA binding sites were, as they suspected, necessary for acid resistance. Combining their latest results with those of Hommais *et al.*, Masuda and Church postulated a regulatory cascade with H-NS repressing EvgA, while EvgA induces YdeO. As described above, YdeO was proposed to induce GadE, the transcription factor responsible for inducing acid resistance, with their complete model summarized in figure 3.

It is important to note, however, that in the network they proposed, Masuda and Church argued that GadX – an AraC-like protein identified by Hommais *et al.* as

being induced in *hns* mutants - did not induce GadE. In contrast, earlier studies had concluded GadX was part of a complex regulatory circuit involving another AraC-like protein GadW, the stress sigma factor, σ^S, and CRP (cAMP receptor protein). Masuda and Church based their argument on a comparison of the expression profiles of *gadX* deletion and *gadX* overexpression strains during exponential growth, which did not show *gadE* to be differentially expressed. In contrast, in a nearly concurrent study, Tucker, Tucker *et al.* (Tucker, Tucker et al. 2003) compared the expression profiles of wild-type and deletion strains for *gadX* and *gadW* during stationary phase which they believed indicated a regulatory relationship between GadX and GadE. However, Tucker *et al.* argued that GadX works with GadW to integrate signals from other sources, though the exact mechanism of this is as yet unknown. Regardless, this inconsistency reminds us that co-expression is one facet of a highly interconnected system (one on many informational levels including protein, protein modification, etc.), but do not diminish the pioneering contributions made by these first global studies.

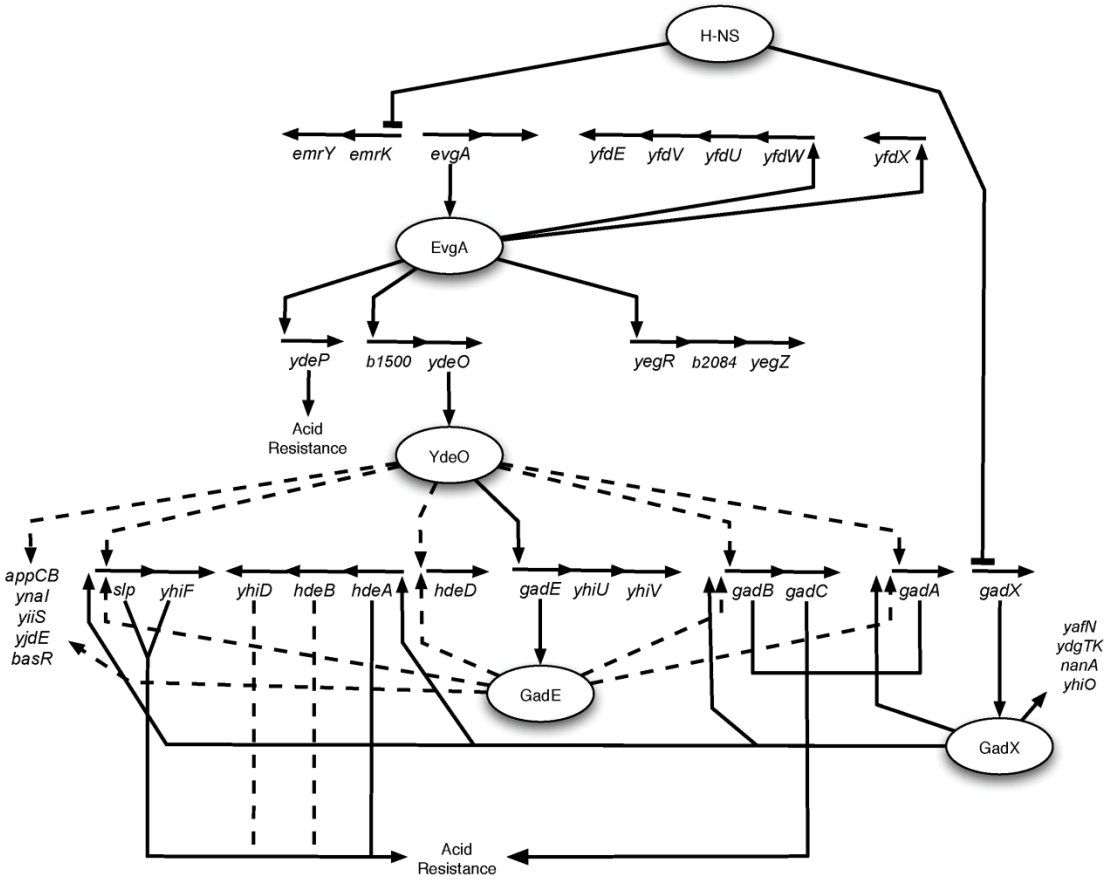


Figure 1.3: *E. coli* glutamate dependent acid resistance circuit. Masuda and Church's model for *E. coli's* glutamate dependent acid response, adapted from (Masuda and Church 2003). Solid lines represent confirmed regulatory relationships, while dotted lines represent relationships that were unclear. In this model, H-NS serves to repress evgA and gadX expression, while EvgA expression induces expression of ydeO and ydeP, both necessary for acid resistance, with YdeO expression inducing gadE expression. Additionally, dotted lines are used to connect both YdeO and GadE with gadA, hdeD and the slp-yhiF, hdeAB-yhiD and gadBC operons to reflect uncertainty as to whether these were under YdeO's direct control or via GadE. Reflecting their conclusion that GadX did not induce GadE expression, GadX is shown to induce some of the acid response genes and operons, but not gadE-yhiUV.

1.5.4 Global computational models of *E. coli* metabolism and regulation.

In addition to allowing researchers to study *E. coli* on a genome-wide scale, the completion of the sequencing of the *E. coli* genome also opened the door for the first genome-scale *in silico* models. In fact, *in silico* models of *E. coli* have been around since as early as 1990 (R. A. Majewski and Domach 1990), however, these were limited in both scale and complexity, usually comprising a small set of genes and modeling only a few processes. In contrast, the more recent models of regulation and metabolism contain thousands of genes involved in a nearly comprehensive number of processes (Covert, Knight et al. 2004).

We'll describe these models and the ways they're being used in greater detail below, but first we need to cover a few basics.

1.5.4.1 Data driven models of the *E. coli* regulatory network

Generally speaking, the full spectrum of *in silico* research can be divided into 2 distinct classes consisting of: 1) regulatory network inference and 2) modeling of the full metabolic network and its interactions with a subset of the regulatory network. While there are networks that have been generated via manual collation and collection of experimentally validated interactions from published literature (Ogata, Goto et al. 1998; Karp, Riley et al. 2000; Salgado, Gama-Castro et al. 2006), it is expensive and time-consuming to create and maintain these networks as they require expert knowledge and extensive experimentation (full field X 50 years) to be generated. As a result, there have been a number of *in silico* methods developed that attempt to infer

regulatory relationships from genome-wide experimental data such as microarray expression and ChIP-chip data. Often, these methods use computational learning algorithms that have been adapted to work specifically with biological data (D'Haeseleer, Wen et al. 1999; Weaver, Workman et al. 1999; Friedman, Linial et al. 2000; van Someren, Wessels et al. 2000; Vanet, Marsan et al. 2000; Segal, Taskar et al. 2001; van Someren, Wessels et al. 2002; Bar-Joseph, Gerber et al. 2003; Segal, Shapira et al. 2003; Stuart, Segal et al. 2003; Hashimoto, Kim et al. 2004; Bonneau, Reiss et al. 2006; Slonim, Friedman et al. 2006; Faith, Hayete et al. 2007). In addition to offering the prospect of a cheaper and less costly solution, these automatic methods also have the possibility of identifying previously unknown protein interactions, providing quantitative means for experimental design, and a means for inferring the roles of genes of unknown function. The development of these algorithms is still in the nascent stage and currently there does not yet exist a "gold standard" data set or known interaction map that can be used to gauge their performance. We will return to the second class (models including flux through metabolic networks) shortly, but first lets discuss one network inference project that has recently been applied to *E. coli*.

While, as mentioned above, there have been a number of efforts reported in recent years to infer the regulatory networks of various organisms and other systems the first such effort for *E. coli* has only recently been reported in early 2007 by Faith *et al.* (Faith, Hayete et al. 2007). Faith *et al.* used an unsupervised network inference method, the context likelihood relatedness algorithm (or CLR), of their own

construction on a data set consisting of 445 *E. coli* Affymetrix microarray expression measurements coming from both published sources as well as new experiments (> 1/2 of the collated data was new). Once generated, they then validated their inferred network using a combination of *in vitro* and *in silico* methods. We'll discuss the overall work and its findings shortly, but first let's describe the inference method they used.

1.5.4.2 The CLR algorithm

The context likelihood relatedness (or CLR) algorithm compares expression profiles of the genes by utilizing mutual information (MI), a commonly used information theoretic similarity measure. Mutual information is defined as the *relative entropy* between the joint distribution and the product distribution of 2 random variables, *X* and *Y*, defined mathematically as:

$$I(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)} \right)$$

where I(X:Y) is the MI for two variables (in this case the levels of two genes under a large number of conditions, p(x) is the probability of seeing a value for x in the distribution, p(y) is the probability of seeing a value for y in the distribution, and p(x,y) is the probability of seeing a given value for x and y in a single observation or sample.

Generally speaking MI can be understood to be a measure of the coupling between the distributions of two random variables, or in the case of the CLR

algorithm, the similarity between the distributions of 2 genes. So, to get an intuition of how this measure operates, consider an example of 2 genes that are completely independent of each other such that $p(gene_1, gene_2) = p(gene_1)\ p(gene_2)$. As this situation would give us a fractional value (the fraction inside of the log function, that is) equal to 1, we would get a mutual information of 0 as $log(1) = 0$. Additionally, another key aspect of MI is that we are guaranteed that the mutual information between any two variables will be greater than or equal to 0. Thus, mutual information is a measure of the non-independence of two variables (or genes in our case). Importantly the measure can detect relationships that would not be detected by a metric such as the Pearson correlation.

The CLR algorithm first calculates the background distribution of mutual information scores for each gene, estimated for each gene by determining the pairwise mutual information between it and the rest of the genes in the data set. Then, using this background distribution of pairwise MI scores, the CLR algorithm calculates the likelihood of their score. In so doing, this allows the CLR algorithm to filter out those genes that have spurious similarities with large numbers of other genes.

To improve the likelihood that high scoring gene pairs are causal and improve the stability and run time of their algorithm, Faith *et al.* selected a subset containing 328 known or putative transcription factors and used these as the centroids or mediods of their clustering scheme. In so doing, they correctly reduced both the overall search

space of their algorithm, improved the stability of the result with respect to small changes in the data, and reduced the cost of the requisite computation.

To validate the results from their CLR algorithm, Faith *et al.* used the RegulonDB database (Salgado, Gama-Castro et al. 2006) for its set of known interactions for *E. coli*. Using these known interactions (culled from the literature) to calculate precision and recall (percent true positives and percent true positives found), Faith *et al.* found that at a 60% precision rate, CLR identified 1079 interactions, of which 338 were known and 741 putative. Additionally, Faith *et al.* further explored all of the discovered putative regulons containing 5 or more genes using the *in silico* motif analysis tool MEME, discovering significant motifs in 28 of the 61 regulons examined, with 13 of these corresponding to known motifs. As yet another validation method, Faith *et al.* also performed *in vivo* validation using Chip-qPCR for 3 of the transcription factors they considered significant, identifying 21 previously unknown interactions. Finally, the regulatory network identified a potential combinatorial transcriptional control of iron transport by the central metabolism of *E. coli*, which Faith *et al.* validated using real time quantitative PCR.

While these are clearly impressive and interesting results, one should also note a few limitations of their approach; many of these limitations represent limitations for all methods given current datasets and thus future directions for the field of regulatory network inference. By limiting their search space to that of the known transcription factors, as many other techniques do, the CLR algorithm cannot detect auto-regulated

proteins such as the CtrA master control regulator in *C. caulobacter* (nor can any other method we are aware of). Potentially, this could be resolved if the upstream region of the transcription factor corresponding to a particular regulon was included in upstream sequences that was validated using either of the *in silico* or *in vivo* methods they employed.

1.5.4.3 Dynamic models of regulation and metabolism

In contrast, dynamic cellular models, as their name would imply, attempt to simulate the internal physiology of a cell. A number of different approaches have been created to do this including thermodynamic (Loew and Schaff 2001; Beard, Liang et al. 2002; Edwards, Ramakrishna et al. 2002; Moraru, Schaff et al. 2002), stochastic (Arkin, Ross et al. 1998), cybernetic (Varner and Ramkrishna 1998; Varner and Ramkrishna 1999; Guardia, Gambhir et al. 2000) and constraint-based models (R. A. Majewski and Domach 1990; Edwards and Palsson 2000; Covert, Schilling et al. 2001; Edwards, Ibarra et al. 2001; Covert and Palsson 2002; Edwards, Ramakrishna et al. 2002; Reed, Vo et al. 2003; Covert, Knight et al. 2004; Barrett, Herring et al. 2005). However, of these, constraint-based models are the only approach that has been shown to be scalable to genome-wide models as the others depend upon highly specified parameterizations of attributes such as polymerase availability and quantity, as well as other environmental factors such as temperature. For this reason, they don't scale well to full genome-wide models, and we will focus on constraint-based methods

below. (Covert, Schilling et al. 2001; Price, Papin et al. 2003; Reed and Palsson 2003).

1.5.4.3.1 Constraint-based overview & basic stoichiometric matrix

The constraint-based approach described by Price *et al.* uses a matrix of pre-specified constraints as the central model (Price, Papin et al. 2003). As such, rather than a single solution, a constraints based model may have multiple valid solutions provided they don't violate these constraints. The earliest constraint-based models were designed to model the metabolism of a cell in steady state by using a matrix representation of the metabolic network for a given cell, denoted as ***S***, that encodes the stoichiometry of each of the biochemical reactions within that cell. To find the allowable rates of each reaction (generally not known for more than a minority of reactions in any cell) we find the null space of S by setting $\boldsymbol{Sv=0}$, where $\boldsymbol{v}$ is vector of the fluxes in the reactions described in *S* (and the unknown we are searching for). For readers less familiar with this type of modeling we expand this discussion below. For a more extensive discussion of this type of modeling we refer interested readers to Palsson's recent book, aptly named "Systems Biology" (Palsson 2006).

1.5.4.3.1.1 The S matrix explained

Let us first examine the matrix ***S***. Each row of ***S*** represents a single metabolic compound or metabolite, while each column represents an individual reaction that reflects the stoichiometry of that reaction. So, for example, the following hypothetical reaction involving 4 reactants:

$$A + B \rightarrow C + D$$

would be represented by the vector $(-1, -1, 1, 1)^T$, where the symbol, ^{T}, is used to represent the transpose of the vector. As such, the first two values of the vector (the -1's) correspond to the compounds that are consumed in the reaction, namely A and B, while the latter two values correspond to the compounds produced by the reaction, C and D.

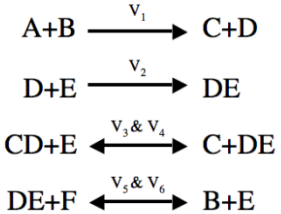
Continuing our example, a reversible reaction such as:

$$CD + E \leftrightarrow C + DE$$

would require two vectors to represent the two possible reactions, i.e. $(-1,-1,1, 1)^T$ and $(1,1, -1,-1)^T$. A more complex system involving 4 reactions (2 reversible and 2 irreversible) and 8 reactants is illustrated in figure 4.

Figure 1.4: A stoichiometric matrix, S, for a system of 4 reactions involving 8 reactants. Reaction 1 corresponds to column one, reaction 2 to column 2, reaction 3 to columns 3 and 4, and reaction 4 corresponds to columns 5 and 6.

-1	0	0	0	0	0	A
-1	0	0	0	1	-1	B
1	0	1	-1	0	0	C
1	-1	0	0	0	0	D
0	-1	-1	1	1	-1	E
0	0	0	0	-1	1	F
0	0	-1	1	0	0	CD
0	1	1	-1	-1	1	DE



Given a stoichiometry matrix we still need to know the relative rate constants corresponding to each reaction, a vector of rates **v**. One assumption is that the cell is at homeostasis (or will reach homeostasis following any perturbation. This assumption allows us to set **Sv**=0, this equality combined with other assumptions about allowable rates (which impose only very broad constraints on rates, such as rates must be > 0) allows us to find sets of allowable rates, **v**, which in turn allow us to predict the outcome of changes in metabolic flux following perturbations. This approach, forcing the solution to exist in the null space of ***S***, centers on the simplifying assumption that the organism and/or cell operates in perfect homeostasis. Other uses of this encoding of metabolism do not require this assumption and are discussed briefly below, for example Palsson's group has also performed analysis that do not require the assumption **Sv**=0, and have also carried out analysis that couple the metabolic and regulatory networks to successfully predict systems wide properties.

1.5.4.3.1.2 A general approach for using stoichiometric models in simulation

Shortly following the formulation of this framework, in 1994 Varma and Palsson (Varma and Palsson 1994) illustrated how these metabolic models could be applied in simulation. As an iterative approach, their algorithm divides the simulation into equal sized time slices or time points. Provided an initial condition for the first time point t_i, the solution ***v*** that produces optimal growth is chosen. Following this,

any perturbations to the external media by this flux state are calculated, and then fed back into model to produce the state for the next time point t_{i+1}.

1.5.4.3.2 The first metabolic-only models

While, the first stoichiometric models of *E. coli* appeared as early as 1990 (R. A. Majewski and Domach 1990), the first genome wide model was the *i*JE660 *in silico* model that Edwards and Palsson developed in 2000 (Edwards and Palsson 2000). To build their model, Edwards and Palsson relied on established databases such as EcoCyc, MPW and KEGG (Ogata, Goto et al. 1998; Selkov, Grechkin et al. 1998; Karp, Riley et al. 2000) which contain massive collections of experimentally discovered enzymatic and metabolic reactions that have been manually culled from the available literature. Using these resources as the basis for their model, it contained 705 genes, as well as 436 metabolites involved in 720 reactions. Once completed, Edwards and Palsson used the model as a platform to perform a series of *in silico* gene knockout simulations, accomplished by removing the enzyme under consideration *in silico* by setting all relevant reaction rates (those involving that enzyme) to zero. To gauge the performance of their model, Edwards and Palsson next compared their *in silico* results with those from known experiments and found that their model had a predictive accuracy of 86%. In subsequent studies, their model was also used to predict optimal growth rates and evolutionary adaptation (Edwards, Ibarra et al. 2001; Ibarra, Edwards et al. 2002).

1.5.4.3.3 Incorporating regulatory networks into constraints based metabolic models

So, what was missing from these purely stoichiometric models? A careful reader will likely have noticed that regulatory interactions between transcription factors and enzymes were not part of the initial models. Originally, this stemmed from the assumption that an organism would regulate protein expression so as to optimize the metabolism of the compounds available to it; therefore, by focusing on metabolic rates, one could argue the model was implicitly taking these regulatory relationships into account (rolling regulatory influences on flux into the allowable rates found during the calculation of **v)**. This initial lack of regulatory information was also is a result of the fact that regulatory networks are less well determined than metabolic networks. However, to more realistically reflect the underlying biology, in 2001, Covert *et al.* (Covert, Schilling et al. 2001) introduced into the model the use of Boolean logic to represent the various regulatory relationships between genes.

As an example of such Boolean logic, consider the hypothetical case of a microbe having thriving happily on its preferred carbon source, *Carbon1*, while also having the capacity to utilize a secondary carbon source, *Carbon2*, when *Carbon1* is unavailable. Continuing the example, imagine that there exists a regulatory relationship such that the transcription of a protein to transport *Carbon2* into the cell is repressed when the microbe is in the presence of *Carbon1*. If we use *RPc1* to represent an external cell sensor protein for *Carbon1* and *tTc2* to represent a

transcription factor that induces transcription of the transporter protein, this relationship can easily be encoded using the following Boolean logic rules:

RPc1 = IF (Carbon1)

tTc2 = IF NOT (RPc1).[2]

In their initial description of these Boolean rules, Covert *et al.* applied these to a simplified model that, as a proof of principle, covered only a few growth conditions. However, in 2002 Covert and Palsson extended this approach to generate a model of the central metabolism of *E. coli* (Covert and Palsson 2002). Using a literature based approach similar to that used to build the *i*JE660 model, Covert and Palsson generated a regulatory network consisting of 149 genes that regulated 73 enzymes and 16 other regulatory proteins. To produce their final model, this regulatory network was combined with the *i*JE660 metabolic model, with the final product containing 45 reactions whose availability was impacted by the regulatory relationships represented in the regulatory network. In a new set of *in silico* gene deletion simulations, using both the new regulatory network as well as the original metabolic network, Covert and Palsson discovered that the regulatory model improved the overall performance from 83% for the metabolic model to 91% correctly predicted growth responses.

Following their early success, in 2004 Covert *et al.* (Covert, Knight et al. 2004) reported that they had extended this approach to create *i*MC1010^{v1}, the first genome-

[2] Example taken from Covert, Schilling and Palsson, 2001 Covert, M. W., C. H. Schilling, et al. (2001). "Regulation of gene expression in flux balance models of metabolism." J Theor Biol **213**(1): 73-88.

wide metabolic and transcriptional model or *in silico* strain. The *i*MC1010^{v1} strain was actually an extension of an earlier metabolic model, *i*JR904, that had been reported the year before by Reed *et al.* (Reed, Vo et al. 2003), who themselves had extended the earlier *i*JE660 model to include 904 genes following the release of the updated *E. coli* genome in 2001 (Serres, Gopal et al. 2001). This latest model was extended to include 1010 genes, 104 of which transcription factors that regulated 479 of the remaining 906 genes in the model. To validate their new model, Covert *et al.* compared the predicted growth responses with 13,750 known experimentally-derived growth phenotypes available from the ASAP database (Glasner, Liss et al. 2003), discovering that their model correctly predicted the growth response in 78.7% of the cases.

To improve on their model, Covert *et al.* analyzed those cases where the known response disagreed with those that the model predicted, in the process identifying several suspected cases of missing or unknown enzymes and transcriptional interactions. Furthermore, focusing on the organism's response to oxygen deprivation, Covert *et al.* also performed microarray expression profiling of several gene knockout strains they created to explore this response. From the results of the analysis of this expression data, a number of updates to the model's regulatory network were made, resulting in their next *in silico* model, *i*MC1010^{v2}, which too was then tested using the same growth response cases that had been applied to *i*MC1010^{v1}. Unfortunately, the improvement in the number of correctly identified growth responses was negligible (+5 cases called correctly). Despite these disappointing

results, Cover *et al.* observed, however, that the new model was far more successful in predicting the expression differences of genes that had been revealed to be differentially expressed by the microarray data.

Following this announcement of the *i*MC1010^{vX} *in silico* strains, in 2005 Barret *et al.* (Barrett, Herring et al. 2005) reported the result of an interesting experiment where they compared the simulations of the *i*MC1010^{v1} strain grown in various media. For their experiment, the media chosen was selected such that it would cover the full range of growth media that could be used by the *i*MC1010^{v1} strain. Enumerating all possible combinations of carbon, nitrogen, phosphate, sulfur and electron-acceptor sources resulted in 108,723 combinations, 15,580 of which induced sufficient predicted growth by the *i*MC1010^{v1} strain to be used in their comparison. Note, that rather than comparing the resulting growth phenotype, as was done by previous studies, they instead compared the predicted gene expression and activities *during* these simulations against one another. Using an agglomerative clustering algorithm in combination with principal components analysis, Barret *et al.* discovered that most of the simulations grouped together into a relatively small number clusters – 36 or 13, depending upon whether gene expression or gene activity was compared. Moreover, for either type of comparison, Barret *et al.* discovered that these clusters were characterized by the terminal electron acceptor available in the *in silico* growth environment. These results led them to conclude that despite the multitude of possible environments *E. coli* could be subjected to its genetic system is designed to function in

a few dominant modes of response. Or, as they succinctly summarized it, their results were consistent with the hypothesis that "system complexity is built in to robustly provide for simple behavior".

Though it would be interesting to see how these results would compare with a similar experiment using the MC1010^{v2} model, if we focus on just the technological aspects for the moment, the ability to perform simulations on nearly 110,000 different media is impressive, in and of itself. While acknowledging the current limitations of the existing models, it is clear that they still have the capacity to provide some important observations about the underlying nature of these organisms. Considering the fact that nearly 80% of their phenotype predictions (growth or not-growth) were accurate, this is clearly a milestone for global *in silico* modeling of global dynamics.

E. coli **metabolomics**

Metabolomics, briefly, is the study of all metabolites (small molecules), and their dynamics, for various conditions in an organism. The metabolome is crucial to our understanding of phenotype and fitness outcomes of different cell states (Fiehn 2002) and the number of metabolites accessible is on the order of hundreds to thousands. There is evidence from comparing multiple complete genomes of a common core of enzymes that are fundamental for metabolism (Jardine, Gough et al. 2002). Metabolism may be conserved to some degree at the enzyme level, but the processes and networks by which the various organisms convert metabolites varies

significantly (Peregrin-Alvarez, Tsoka et al. 2003). The field of metabolomics is advancing quickly. One example, important for industry and medicine, is the improvement of bacterial strains by metabolic engineering.

Nobeli and colleagues attempted to characterize the *E. coli* metabolome using two-dimensional NMR to classify and identify metabolites systems-wide from living cells (Nobeli, Ponstingl et al. 2003). They compiled their dataset of 745 metabolites, a subset of the complete metabolome, from publicly available, experimentally verified data from the EcoCyc (Keseler, Collado-Vides et al. 2005) and KEGG (Kanehisa, Goto et al. 2006) databases. Clustering of the metabolites revealed a continuum with significant overlap of clusters and no clearly defined classes of metabolites (with respect to presence of absence under varying conditions). This early study demonstrated a novel systems level perspective of the metabolome. Much 'omic' data is available and its integration is fundamental to understanding the complexities and robustness of a living system in its environment.

1.6 *Halobacterium salinarium NRC-1*

The archaeal *Halobacterium salinarum NRC-1* is a halophillic (salt loving) organism that can not only survive, but requires highly saline environments, flourishing in environments such as the Great Salt Lake in Utah with ~4.5M salinity (or roughly 5-10 times the salinity of sea water). Halobacterium can also withstand a surprising variety of other stresses, such as oxidative stress, DNA-damaging chemicals, heavy metals, UV and gamma radiation, low oxygen, and desiccation. To

withstand high salt, it maintains an isoosmotic cytoplasm by eliminating some Na+ ions and maintaining a high intracellular K, Mg (and also Na) ion concentration. As such, its genome possesses multiple ion transporters such as active K+ transporters (KdpABC), Na+ / H+ antiporters (NhaC proteins), low affinity ion transporters driven by membrane potential (Trk proteins), and heavy metal (arsenic and cadmium) transporters. More importantly, *Halobacterium* flourishes in these environments by adjusting its physiology appropriately in response to numerous external stimuli. For example, it can relocate, in search of favorable environments, using sensors that can discriminate beneficial and detrimental spectra of light (Bogomolni and Spudich 1982; Spudich and Bogomolni 1984; Spudich, Takahashi et al. 1989; Spudich 1993), an aerotaxis transducer (HtrVIII) (Brooun, Bell et al. 1998) and buoyant gas-filled vesicles (DasSarma 1993). One of the hallmarks of *Halobacterium* is its ability to survive anaerobically using light and/or arginine as energy sources and aerobically as a chemoheterotroph. *Halobacterium* generates energy from light by its retinal-containing light-driven ion transporters, bacteriorhodopsin and halorhodopsin (Kolbe, Besir et al. 2000; Luecke, Schobert et al. 2000). Additionally, *Halobacterium* can also ferment arginine via the arginine deiminase pathway with each mole of arginine fermented yielding one mole of ATP (Ruepp and Soppa 1996). As such an extremeophile, it represents an interesting, yet still poorly understood class of organisms. Moreover, from a systems biology perspective archaea present an interesting opportunity as while they are prokaryotic organisms, they share many

attributes with eukaryotes such as eukaryotic-like transcription, translation and TATA-boxes. Though they have been the subjects of study since the 1960s, in 2000 the first *Halobacterium salinarum* genome was sequenced, opening the door for further systems-level study of the organism (Ng, Kennedy et al. 2000; Dassarma, Berquist et al. 2006).

Below, we go through how some of these efforts have been applied to *Halobacterium*. We will illustrate a systematic process consisting of the following steps:

1. Define all of the elements in the cell (or organism). Develop an initial model of the cell using existing knowledge, i.e. literature review.
2. Perturb the system environmentally and/or genetically (knockouts, over expressions, etc.) and globally assay the relationships of the elements one to another (e.g., levels of mRNA and protein, protein/protein interactions, etc.). Integration of data from different sources is critical to a complete understanding.
3. Compare the model with the experimental results to formulate new hypotheses which explain the discrepancies.
4. Test these hypotheses with a new series of perturbations and update the model to more accurately reflect the experimental results.
5. Iterate steps 2 - 4.

1.6.1 Sequencing of Halobacterium

As mentioned above sequencing of the *H. salinarium NRC-1* genome was completed by Ng *et al.* in 2000 (Ng, Kennedy et al. 2000), who used a whole-genome shotgun strategy to sequence the genome which consists of one large replicon and two, relatively smaller replicons. The larger of these contains ~2Mbp (2,571,010 bp, exactly), while the two smaller replicons, pNRC100 and pNRC200 each contain roughly 200 and 350 Kpb, respectively (191,346 and 365,425 exactly). Using the *in silico* gene prediction program, GLIMMER (Salzberg, Delcher et al. 1998; Delcher, Harmon et al. 1999), Ng and colleagues identified 2682 putative genes, of which 2111 were located on the large replicon, while 197 and 374 were found on the 2 smaller replicons, pNRC100 and pNRC200, respectively. To assign function to these, the putative genes were translated and then submitted to NETBLAST (Altschul, Madden et al. 1997) to query for homologues in the nonredundant database of proteins hosted on the National Center for Biotechnology Information (NCBI). The results from this search revealed that 1658 had significant matches, though of these matches, only 1067 had known function while the remainder were hypothetical proteins. Of these matches to genes with known function were genes involved in metabolism, cellular envelop maintenance, photobiology, DNA replication, transcription and translation. Interestingly, Ng *et al.* also identified 91 transposable insertion elements, with the majority of these (62) located on the 2 smaller replicons or minichromosomes, leading

them to conclude that these play a significant role in *Halobacterium* evolution by allowing the organism to gain new genes.

1.6.2 Baliga *et al.*, 2002 – systems wide exploration of energy production in differing environments.

Following the sequencing of *Halobacterium*, the first system-level analysis was reported by Baliga *et al.* in 2002 (Baliga, Pan et al. 2002), who explored the combined RNA and protein expression of *Halobacterium* during anaerobic energy production. As mentioned above, from earlier studies, it was already known that in anaerobic conditions, *Halobacterium* could generate energy from either arginine fermentation or photosynthesis. Additionally, from earlier studies, it was known that during phototrophic growth the organism generates numerous copies of a light-driven proton pump called bacteriorhodopsin (bR), which is a protein complex composed of the 2 proteins bacterioopsin (Bop), and retinal. During phototrophic growth these proton pumps are organized in a two-dimensional lattice called the purple membrane. Furthermore, it was also known that another protein, Bat, regulated the expression of itself, as well as 3 others involved in bR synthesis, *bop*, *brp*, and *crtB1*.

Thus, to explore the regulatory network driving phototrophic growth, Baliga *et al.* performed RNA and protein expression analyses of 4 different strains, including the NRC-1 wild-type, a *bop* knockout strain (*bop*-), as well as both a *bat* overexpression (*bat*+) and knockout (*bat*-) strain. Using cDNA microarrays, they discovered that the *bop*- strain exhibited little expression difference from the *bat*+

strain. However, as would be expected of a transcription factor, the *bat*+ and *bat*- exhibited significant numbers of differentially expressed genes, with 151 and 157 differentially expressed genes, respectively. What was not expected, though, was that functionally, their expression profiles were inverted, as those genes involved in photosynthesis were induced in the *bat*+ strain, but repressed in the *bat*- strain, while the opposite was the case for those involved in arginine fermentation (repressed in the *bat*+ strain, but induced in the *bat*- strain). While the exact mechanism for this was unclear, Baliga *et al.* hypothesized that this inversion represents a strategy to maintain a steady level of ATP within the cell. Additionally, subsequent proteomics studies using the ICAT technique (Gygi, Rist et al. 1999) found a number of differentially expressed proteins had no corresponding change in mRNA (33/50), indicating posttranslational effects upon protein expression. Furthermore, *in silico* promoter analysis of the genes induced in the *bat*+ strain found only one additional gene containing the Bat binding site, indicating that most of these were subject to indirect regulation by Bat. However, promoter analysis using MEME was able to identify a likely binding motif among five genes involved with arginine fermentation. In so doing, their study lead to new hypotheses later verified with future genetic modifications and later iterations of the group's systems-level analyses.

1.6.3 The functional annotation of Halobacterium proteome

In addition to these findings, Baliga *et al.* discovered that they had also been able to verify the existence (at the protein and transcript level) of 496 of the 971

hypothetical genes in the *Halobacterium* genome – those that had been predicted by gene finders, but had no homologues with other known genes. This annotation was further expanded by Bonneau *et al.* (Bonneau, Baliga et al. 2004) who in a paper from 2004 reported both a new functional, structure-based annotation of the *Halobacterium* genome, as well as a new contextual annotation of the genome that linked proteins by associations such as shared operon membership.

To update this proteome annotation, Bonneau *et al.* used a method which they'd used previously in the critical assessment of structure prediction (CASP3,4 & 5) (Bonneau, Strauss et al. 2001; Bonneau, Tsai et al. 2001; Chivian, Kim et al. 2003) which used two algorithms, Ginzu and Rosetta (Bonneau, Tsai et al. 2001; Aloy, Stark et al. 2003; Bradley, Chivian et al. 2003; Fischer, Rychlewski et al. 2003; Kinch, Wrabl et al. 2003) to predict protein domain boundaries and protein structure. The method is a hierarchical workflow that utilizes a protein domain-centric approach to identify function and structure starting only with the primary sequence of a predicted protein. As an initial, pre-processing step, each query sequence is filtered for regions that are likely to be either transmembrane, coiled coils, signal peptides or a disordered region. These regions are removed from further analysis, with the remainder submitted to their protein-domain parsing program, Ginzu, which attempts to parse the primary sequence into likely domains and identify their functions by using a hierarchical workflow (with more accurate methods placed at the top of this hierarchy). The first step of this process is to use PSI-BLAST to search for sequence

matches to the PDB, resulting in high-quality, high-likelihood domains of known function. For those regions of the protein not identified by this PSI-BLAST search, they are next queried using HMMER for matches in Pfam. If any regions still have not been identified by these previous searches, as a third step Ginzu next attempts to identify matches to protein structures using Fold Recognition. As the fourth and final step of Ginzu, any regions not recognized by the previous 3 methods are aligned to all known sequences using PSI-BLAST; multiple sequence alignments are parsed for block patterns indicative of domain structure. Finally, all domains not matched by a known structure using these methods are then passed to the Rosetta algorithm, a *de novo* structure prediction algorithm that uses information from the PDB to identify likely local structure confirmations.

With their functional annotation process, Bonneau *et al.* found 1077 of the 2596 protein coding genes in the *Halobacterium* genome had significant matches found by the initial PSI-BLAST search of the PDB. Additionally, 610 domains were identified by querying the Pfam database, with an additional 670 domains identified using the two *de novo* structure prediction methods (Rosetta). While 1234 protein domains could not be annotated by this method, this still translates into a nearly 30% improvement over the collection of sequence-based methods which had initially been used.

1.6.3.1 Protein associations and structure prediction to derive putative annotations for proteins

To generate their contextual annotation of associations, Bonneau *et al.* considered 4 possible association types, including protein-protein interactions, fusions of *Halobacterium* protein domains found in other genomes, proteins grouped into operons, and phylogenetic profile edges (Tatusov, Natale et al. 2001). To identify putative protein-protein interactions, they used the COG (Clusters of Orthologous Genes) database, along with other databases of known interactions to infer 1143 likely interactions. For the fusions of *Halobacterium* domains, a method described by Enright *et al.* (Enright, Iliopoulos et al. 1999) was utilized to identify 2460 suspected associations. To identify operons, two methods were used, one which considered clusters of genes with shared directionality, while the other considered nearby pairs of genes which had orthologs in other genomes that were similarly co-located (Mellor, Yanai et al. 2002; Moreno-Hagelsieb and Collado-Vides 2002). With these two methods, 1335 total putative operon associations were identified. Finally, 525 association links were added using the phylogenetic profile method of Marcotte *et al.* to identify collections of genes which often co-occur in different genomes (Marcotte, Pellegrini et al. 1999; Eisenberg, Marcotte et al. 2000). These associations and the prior proteome annotation effort provided a rich environment in which to explore protein function that was much greater than the sum of the individual parts.

1.6.4 *Halobacterium's* stress response following exposure to ultraviolet radiation

We now further review *Halobacterium's* stress response following exposure to ultraviolet (UV) radiation (Baliga, Bjork et al. 2004). Damage to DNA as a result of exposure to shortwave UV light (UV-C) falls into two categories, one being pyrimidine and pyrimidone phosphoproducts that are created between the C6 and C4 carbons of neighboring pyrimidine nucleotides (i.e. T-C or C-C), while the other are cyclobutane pyrimidine dimers (CPD) that are created between the C4 and C5 positions of neighboring pyrimidines of the same type (i.e. C-C or T-T). Similarly, there exists two repair mechanisms in most organisms, one of which is the nucleotide excision repair (NER) system that can occur at any time, but is better with repairing phosphoproducts. The second is a photolyase-catalyzed phosphoreaction that can only occur in the presence of light, and is more effective at repairing CPD's. Note, however, both repair pathways can repair both types of DNA lesions. Prior to Baliga *et al.*'s exploration of the UV response, it had been known that *Halobacterium* had homologs for proteins in both systems, including homologs for both bacterial and eukaryotic NER proteins, though there were still questions regarding the exact machinery of these repair mechanism within the organism.

As an initial foray, Baliga *et al.* explored the UV-C resistance of the *Halobacterium*, by exposing *Halobacterium* in a thin liquid culture to UV-C radiation, finding that up to 110 J/m^2 there was no loss of viability and 37% survivability

following 280 J/m^2. However, these initial tests also indicated that photoreactivation was a major UV repair mechanism (growth in light following exposure was 16 times more likely (16-fold) than growth in dark conditions). For this reason, they next focused on two photolyase homologs *phr1* and *phr2* within the *Halobacterium* genome. While it was already known that *phr2* was a photolyase, the role of *phr1* was still unknown. Using 3 strains, consisting of a *phr1* knockout (*phr1*-), a *phr2* knockout (*phr2*-), and a *phr1* and *phr2* double knockout (*phr1-/phr2*-), they found that their results clearly revealed that only *phr2* functioned as a CPD photolyase, as the *phr1*- strain exhibited no difference from the wild-type following UV exposure. As they also found that both the *phr2*- and *phr1-*/ *phr2*- strains exhibited ~3.5 fold increased survivability when grown in the presence of light versus dark following UV exposure, they next explored the processes occurring during what they termed light versus dark repair following exposure to UV light.

To accomplish this, they used an experimental procedure where they examined the organism, grown in either light or dark conditions, at 30 and 60 minutes post UV-exposure, as well as a control (no UV exposure) after 60 minutes growth in light. Thus, five separate assays were performed (L30, L60, D30, D60, and C60). Using new 70-mer oligonucleotide microarrays to assay the RNA expression at these time points, they found that a total of 420 genes who's mRNA was differentially expressed, with 273 of these only occurring during the repair tests, 40 of which occurred in both repair conditions and 61 that occurred in both the control and repair assays. One of the more

interesting findings from these assays was the difference in number of genes that were repressed after 60 minutes repair growth in light (L60) assay versus those that were differentially expressed in the other repair assays. Specifically, while <2% of *Halobacterium's* genes were differentially expressed in any of the other repair assays, roughly 12% of the genome was found to be down regulated in the L60 assay, including nearly all the ribosomal and RNA polymerase genes. This massive down-regulation has also been found to be a general stress response in other conditions, as well as other organisms.

Based on the structure-based reannotation of the genome, Baliga *et al.* were able to identify at least two transcription factors, genes VNG1318H and VNG0019H, who's function were unknown previously. In addition, using the association annotation that Bonneau *et al.* described, along with their own expression results and information from the Kyoto Encyclopedia of Genes and Genomes, Baliga *et al.* were able to identify and visualize the response of biomodules using Cytoscape (Shannon, Markiel et al. 2003), a genomic data visualization tool. We will discuss Cytoscape in greater detail below. However, all these combined tools and newly acquired information allowed Baliga *et al.* to formulate a number of new conclusions and hypotheses. Among these was the conclusion that *phr2,* and not *phr1,* was clearly a photolyase and the major mechanism of UV-C damage repair. Another conclusion, based on the number of genes downregulated in the L60 sample, was that the major response to UV-C damage is a halt in transcription and translation to allow the

organism or cell to recover from the UV-C induced damage before regular cell activity and division restarts (a result also seen in other organisms stress response). Furthermore, they identified 3 new putative transcriptional regulators involved in repair damage, including the VNG1218H gene that we mentioned above. Finally, the new experimental data and computational analyses techniques also allowed Baliga *et al.* to speculate on 2 parallel mechanisms involving Cobalamin (B-12) biosynthesis.

1.6.5 Data Visualization: Cytoscape and the Gaggle

Cytoscape is a computer program that Shannon *et al.* (Shannon, Markiel et al. 2003) first reported in 2003, which displays the genes and associations of a given organism as a network where the genes represent nodes, and the associations represented as edges between the genes/nodes. Furthermore, attributes such as function and mRNA and protein expression data can then be assigned to each gene in the network. With this setup, Boolean networks and active transcriptional paths calculated using mRNA expression data can then be explored in context of the other data types integrated into the network to gain systems level insights and formulate hypotheses for further testing. See cytoscape.org for details, code and Cytoscape compatible tools (plugins).

1.6.6 The quest for the global Halobacterium regulatory network: Philosophy.

Distilling regulatory networks from large genomic, proteomic and expression datasets is one of the most important mathematical problems in biology today (Yuh, Bolouri et al. 1998; Friedman, Linial et al. 2000; Wahde and Hertz 2001; Ideker, Ozier

et al. 2002; Lee, Rinaldi et al. 2002; Shmulevich, Lahdesmaki et al. 2003; Hashimoto, Kim et al. 2004; Bonneau, Reiss et al. 2006). The development of accurate models of global regulatory networks is key to the understanding of a cell's dynamic behavior and its response to internal and external stimuli. A major goal of the Halobacterium project was thus to combine all data (including the data generated by the focused studies above) to generate a global regulatory network.

Methods for inferring and modeling regulatory networks must strike a balance between model complexity - a model must be sufficiently complex to describe the system accurately - and the limitations of the available data - in spite of dramatic advances in our ability to measure mRNA and protein levels in cells, nearly all biological systems are underdetermined with respect to the problem of regulatory network inference. We focus on further development of our algorithms for learning co-regulated modules and regulatory networks. Our aim is to learn models of regulation from data that include units of time, concentration (or at least relative concentration) and to explicitly model regulator binding-sites.

1.6.7 Halobacterium global regulatory network inference. Methods, motivations, challenges and current progress.

1.6.7.1 Challenges:

A major challenge is to distill, from large genome-wide data sets, a reduced set of factors describing the behavior of the system. The number of potential regulators is often on the same order as the number of observations in current genome-wide

expression and proteomics datasets. A further challenge in regulatory network modeling is the complexity of accounting for transcription factor interactions and the interactions of transcription factors with environmental factors (*e.g.* it is known that many transcription regulators form heterodimers, or are structurally altered by an environmental stimulus such as light, thereby altering their regulatory influence on certain genes). A third challenge and practical consideration in network inference is that biology data sets are often heterogeneous mixes of equilibrium and kinetic (time-series) measurements; both types of measurements can provide important supporting evidence for a given regulatory model if they are analyzed simultaneously. Last, but not least, is the challenge that data-derived network models be predictive, and not just descriptive: can one predict the system-wide response in differing genetic backgrounds, or when the system is confronted with novel stimulatory factors or novel combinations of perturbations?

We describe the methods we used to predict the global network from The *Halobacterium* Data compendium as a two-part process (step 1, cMonkey, step2, the Inferelator). We follow this discussion with a brief discussion of the tools that are used to explore this data, the resulting networks and associated annotation data (the Gaggle).

1.6.7.2 Step 1: cMonkey, the need for integrative biclustering:

Learning and modeling of regulatory networks can be greatly aided by reducing the dimensionality of the search space prior to network inference. Two ways

to approach this are 1) limiting the number of regulators under consideration, and 2) grouping genes that are co-regulated into clusters. In the first case, candidates can be prioritized based on their functional role, *e.g.* limiting the set of potential predictors to include only transcription factors, and by grouping together regulators that are in some way similar. In the second case, gene-expression clustering, or unsupervised learning of gene-expression classes, is commonly applied. It is often incorrectly assumed that co-expressed genes correspond to co-regulated genes. However, for the purposes of learning regulatory networks it is desirable to classify genes on the basis of *co-regulation* (shared transcriptional control) as opposed to simple *co-expression*. Furthermore, many standard clustering procedures assume that co-regulated genes are co-expressed across all observed experimental conditions. Since genes are often regulated differently under different conditions, this assumption is likely to break down as the size and variety of data grows. *Biclustering* was developed to better address the full complexity of finding co-regulated genes under multifactor control by grouping genes on the basis of coherence under *subsets* of observed conditions (Cheng and Church 2000; Tanay, Sharan et al. 2002; Yu 2002; Kluger, Basri et al. 2003; Segal, Shapira et al. 2003; Sheng, Moreau et al. 2003; Yu 2003; Tanay, Sharan et al. 2004).

Co-regulated genes are often functionally (physically, spatially, genetically, and/or evolutionarily) linked (Moreno-Hagelsieb and Collado-Vides 2002; Harbison, Gordon et al. 2004). For example, genes whose products form a protein complex are

likely to be co-regulated. Other types of associations among genes, or their protein products, that can imply functional couplings include (a) presence of common cis-regulatory motifs; (b) co-occurrence in the same metabolic pathway(s); (c) cis-binding to common regulator(s); (d) physical interaction; (e) common ontology; (f) paired evolutionary conservation among many organisms; (g) common synthetic phenotypes upon joint deletion with a third gene; (h) sub-cellular co-location; and (i) proximity in the genome, or in bacteria and archaea, operon co-occurrence. These associations can be either derived experimentally or computationally (either pre-computed ahead-of-time, or on-the-fly during the clustering process); indeed it is common practice to use one or more of these associations as a post-facto measure of the biological quality of a gene cluster. However, it is important to note that these data types, to varying degrees, can contain a high rate of false positives, or may imply relationships that have no direct implication for co-regulation. Therefore in their consideration as evidence for co-regulation, these different sources of evidence should be treated as priors, with different amounts of influence on the overall procedure based upon prior knowledge of (or assumptions about) their quality and/or relevance.

Because a biological system's interaction with its environment is complex and gene regulation is multi-factorial, genes might not be co-regulated across all experimental conditions observed in any comprehensive set of transcript or protein levels. Also, genes can be involved in multiple different processes, depending upon the state of the organism during a given experiment. Therefore, a biologically

motivated clustering method should be able to detect patterns of co-expression across subsets of the observed experiments, and to place genes into multiple clusters. So-called biclustering, clustering both genes and experimental conditions, is a widely studied problem and many different approaches to it have been published (Cheng and Church 2000; D'Haeseleer, Liang et al. 2000; Tanay, Sharan et al. 2002; Yu 2002; Kluger, Basri et al. 2003; Segal, Shapira et al. 2003; Sheng, Moreau et al. 2003; Yu 2003; Balasubramanian, LaFramboise et al. 2004; Tanay, Sharan et al. 2004). Unlike standard clustering methods, most biclustering algorithms place genes into more than one cluster (genes can play more than one functional role in the cell). Because biclustering is an NP-hard problem (D'Haeseleer, Liang et al. 2000), no solution is guaranteed to find the optimal set of biclusters. However, many of these procedures have successfully demonstrated the value of biclustering when applied to real-world biological data (Balasubramanian, LaFramboise et al. 2004; Reiss, Baliga et al. 2006).

We compared the method to several other methods including but not limited to: Order Preserving Sub-matrix (OPSM(Ben-Dor, Chor et al. 2003)), Iterative Signature (ISA(Bergmann, Ihmels et al. 2003)), Bimax (Prelic, Bleuler et al. 2006), and SAMBA(Shamir, Maron-Katz et al. 2005). We also compared our method to hierarchical clustering and k-means clustering. We used multiple parameterizations of each competing method. In addition, we performed these analyses on cMonkey runs with various model parameters up- and down-weighted to demonstrate tolerance of the cMonkey method to different parameterizations of free parameters. Additional details

on the analysis are provided previously (Reiss, Baliga et al. 2006). All biclusters generated by the cMonkey as well as the other algorithms we tested are available for interactive exploration via Cytoscape and the Gaggle (Shannon, Markiel et al. 2003; Shannon, Reiss et al. 2006) at (http://labs.systemsbiology.net/baliga/cmonkey/).

1.6.7.3 Comparison in the context of regulatory network inference:

A major motivation of cMonkey is to provide a method for deriving co-regulated groups of genes for use in subsequent regulatory network inference procedures. Thus, we wish to find coherent groups of genes over those conditions with a large amount of variation. In other words, we are hoping to detect sub-matrices in the expression data matrix which are coherent and simultaneously have high information content or overall variance (and probability given the network and motif components). In addition, we need to find biclusters with many conditions/observations included, as this increases the significance of each bicluster and also of the subsequently inferred regulatory influences for that bicluster. In general we see that cMonkey generates biclusters with a significantly greater number of experiments than the other methods (higher coverage). Even with this additional constraint (i.e. including a greater number of experiments in the clusters) and further constraints that cMonkey imposes with the association network and motif priors, the algorithm in general generates biclusters with a "tighter" profile, as measured by mean bicluster residual. Thus, we find that biclusters generated by cMonkey are generally better suited for inference algorithms such as the Inferelator (and potentially other

methods as well). We tested this by running the Inferelator on biclusters generated by SAMBA for Halobacterium and then comparing the predictive performance of the resultant regulatory network models on newly-collected data, relative to those generated for cMonkey generated biclusters. We found that, largely due to the smaller number of experiments included in SAMBA biclusters, the inferred network was significantly less able to predict new experiments (an increase in the predictive error from 0.368 to 0.470; p-value of difference by t-test $< 1 \times 10^{-22}$) (Kanehisa, Goto et al. 2004). We find that cMonkey performs well in comparison to all other methods when the trade-off between sensitivity, specificity, and coverage is considered, particularly in context of the other bulk characteristics (cluster size, residual, etc.). Most importantly, cMonkey significantly improves the performance of downstream network inference procedures. cMonkey biclusters do a better job at regenerating the expression data than other methods, and a similar job at recapitulating the external (as well as internal) measures of bicluster quality.

1.6.7.4 Step 2: The Inferelator:

Given modules from a clustering/biclustering algorithm, for example cMonkey, we are then faced with the task of learning which genes and environmental conditions influence/control each module/cluster/bicluster/gene. We have described an algorithm for doing this, the Inferelator, which infers regulatory influences for genes and/or gene clusters from mRNA and/or protein expression levels. The method uses standard regression and model shrinkage (L1-shrinkage) techniques to select

parsimonious, predictive models for the expression of a gene or cluster of genes as a function of the levels of transcription factors, environmental influences and interactions between these factors (Thorsson, Hornquist et al. 2005). The procedure can simultaneously model equilibrium and time-course expression levels, such that both kinetic and equilibrium expression levels may be predicted by the resulting models. Through the explicit inclusion of time, and gene-knockout information, the method is capable of learning causal relationships. It also includes a novel solution to the problem of encoding interactions between predictors into the regression. We discuss the results from an initial run of this method on a set of microarray observations from the halophilic archaeon, *Halobacterium NRC-1*. We have found the network to be predictive of newly measured data and have also validated parts of the network using ChIP-chip.

1.6.7.4.1 Model formulation:

We assume that the expression level of a gene, or the mean expression level of a group of co-regulated genes, y, is influenced by the level of N other factors in the system: $\boldsymbol{X}=\{x_1, x_2, \ldots, x_N\}$. We consider factors for which we have measured levels under a wide range of conditions; in our work on *Halobacterium* we use transcription factor transcript levels and the levels of external/environmental conditions as predictors and gene and bicluster transcript levels as the response. The relation between y and $\boldsymbol{X}$ is given by the kinetic equation:

$$\tau \frac{dy}{dt} = -y + g(\beta \bullet Z) \qquad (1)$$

Here, $\boldsymbol{Z}$ = { $z_1(\boldsymbol{X})$, $z_2(\boldsymbol{X})$, …, $z_P(\boldsymbol{X})$ } represents a set of functions of the regulatory factors $\boldsymbol{X}$. The coefficients *beta* describe the influence of each element of $\boldsymbol{Z}$, with positive coefficients corresponding to inducers of transcription, and negative coefficients to transcriptional repressors (Wahde and Hertz 2001). The constant *tau* is the time constant of the level *y* in the absence of external determinants. We use a novel encoding of interactions by allowing functions in $\boldsymbol{Z}$ to be either: 1) the identity function of a single variable or 2) the minimum of two variables (Jürgen Richter-Gebert 2003). For example, the inner product of the design matrix and linear coefficients for two predictors that are participating in an interaction is:

$$\beta \boldsymbol{Z} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 \min(x_1, x_2) \qquad (2)$$

Using this encoding, for example, if x_1 and x_2 represent the levels of components forming an obligate dimer that activates *y* (x_1 AND x_2 required for expression of *y*), we would expect to fit the model such that $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 1$. This encoding results in a linear interpolation of (linearly smoothed approximation to) the desired Boolean function. This and other interactions (OR, XOR, AND), as well as interactions involving more than two components, can be fit by this encoding. In regression terminology, the influencing factors, $\boldsymbol{X}$, are referred to as regressors or predictors, while the functions $\boldsymbol{Z}$ specify what is often referred to as the "design matrix".

With this scheme for encoding interactions in the design matrix, we expect to capture many of the interactions between predictors necessary for modeling realistic regulatory networks, in a readily interpretable form. To date we have limited the procedure to binary interactions, as it is unlikely that the quantity of data used would support learning beyond these pair-wise interactions. Many other methods for capturing transcription factor cooperatively exist as well (Das, Banerjee et al. 2004). We have shown that removal of the capability to model interactions in this way reduces the predictive power of the Inferelator over the newly collected validation data set.

Various functional forms can be adopted for the function *g*, called the "nonlinearity" or "activation" function for artificial neural networks, and the "link" function in statistical modeling. The function *g* often takes the form of a sigmoidal, or logistic, activation function. This form has been used successfully in models of developmental biology (von Dassow, Meir et al. 2000). The function is compatible with L1-shrinkage (the method for enforcing model parsimony) (van Someren, Wessels et al. 2000; van Someren, Wessels et al. 2002; Efron 2003).

The simplified kinetic description of equation (1) encompasses essential elements to describe gene transcription, such as control by specific transcriptional activators (or repressors), activation kinetics, and transcript decay, while at the same time facilitating access to computationally efficient methods for searching among a combinatorially large number of possible regulators. To better understand specific

details of regulation, it will almost certainly be required to follow up on specific regulatory hypotheses using more mechanistically detailed descriptions. Although this method (explicit time component) does not lessen the need for correct experimental design it does: 1) facilitate using data with reasonable variation in sampling structure and 2) allow for the simultaneous combination of data from equilibrium and time-series data.

1.6.7.5 Predictive power of the *Halobacterium* network over new data (performance on novel combinations of environmental and genetic perturbations):

Our initial application of the method to *Halobacterium* resulted in a statistically learned regulatory network that can predict, with reasonable accuracy, mRNA levels of ~1,900 out of the total ~2,400 genes found in the genome, using relative concentrations of transcription regulators and environmental factors as predictors. We find that applying cMonkey to our expression compendium, the metabolic network, comparative genomics edges and upstream sequences gives us a set of ~300 biclusters spanning ~2000 of the 2400 genes in this organism. This set of biclusters is also linked to a set of putative cis-acting regulatory motifs (some validated by prior experiments). The learned network controlling the 300 biclusters and 159 individual genes contained 1431 regulatory influences (network edges) of varying strength. Of these regulatory influences, 495 represent interactions between two TFs or between a TF and an environmental factor. We selected the null model for

21 biclusters (no influences or only weak regulatory influences found), indicating that we are stringently excluding under-determined genes and biclusters from our network model. The ratio of data points to estimated parameters is approximately 67 (one time constant plus three regulatory influences, on average, from 268 conditions). The explicit time component and interaction component (which distinguish this method from other such shrinkage methods) were essential for predictive performance over the validation data and the new data.

In order to test predictive performance we chose to test the network model (trained prior on the 268 conditions available at the time) over 130 additional new measurements, collected after model fitting. We found that the prediction error over the training set was essentially the same as that over the new dataset. This is encouraging as the new data included environmental perturbations, new combination of environmental and genetic perturbations and time series measurements after novel entrainments of the cell. This predictive power is a prerequisite to further interpretation of organization of key processes in the network. The ability of the same network to predict transcriptional control in novel environments (>130 new experiments) verifies that, irrespective of the nature of the environmental perturbation, Halobacterium utilizes a core set of regulatory mechanisms to maintain homeostasis under extreme conditions. The resultant network (as well as biclusters and supporting tools) for *Halobacterium NRC-1* in Cytoscape, available as a Cytoscape/Gaggle web

start at: http://halo.systemsbiology.net/inferelator (Bonneau 2006; Shannon, Reiss et al. 2006).

1.7 The relationship between systems biology and traditional molecular biology

During our review of these systems-biology prokaryotic projects, our aim was also to illustrate that these systems biology projects were also well integrated with countless other more traditional molecular biology studies. Thus, if one looks at any single group, one might incorrectly see a divide between systems biology and biology as a whole. However, looking across all studies for a single organism, one sees that hypotheses generated by global studies have permeated field-wide and, in a corresponding manner, high-confidence single-gene results from traditional reductionist biology commonly guide the design of global studies. Therefore, rather than the two branches being in competition with each other, we argue that they are involved in a complex and mutually beneficial exchange. In this sense then, any effort that improves the accuracy of the hypotheses that are being generated by systems biologists, especially those that are *in silico* will be of benefit to the entire field, regardless of whether one is a systems or traditional molecular biologist.

1.8 ADDENDUM: Comparative functional genomics of prokaryotes and other subsequent projects

Before continuing, we remind the reader that this introductory chapter is based heavily upon a review of prokaryotic system biology that was published in 2009 (Waltman, Kacmarczyk et al. 2009). Neither then, nor now, were the reviews of the systems biology projects that it presents meant to be comprehensive, field-wide reviews for each organism. Nor are they now intended to provide definitive reviews that include all the ongoing research that has taken place since the original publication for the four (4) organisms. Rather, the intention of this chapter is to provide recent examples of how genomics has been applied to study each of these organisms in order to illustrate the advantages offered by systems biology approaches.

While a comprehensive review of subsequent research for all four (4) organisms is not provided, we instead will present a brief review of some of the most recent work that has been applied to *B. subtilis* and *E. coli*, the two model organisms described above. In addition, in subsequent chapters, we will provide detailed descriptions of a novel, comparative method that was recently developed and used to analyze both of these model organisms (as well as several closely related species to both). The layout of the subsequent chapters will be presented below.

Subsequent to the publishing of the original chapter that this introduction is based upon, numerous efforts have taken place to expand upon the systems biology projects described above. While these include both computational and experimental

systems biology approaches, we limit the further discussion only to those computational efforts that have taken place for both *E.coli* and *B. subtilis*. For example, work to develop combined regulatory and metabolic networks for *E. coli* has been ongoing (Lewis, Cho et al. 2009; Chandrasekaran and Price 2010). In addition, since the project by Faith *et al* to infer the transcriptional regulatory network inference of *E. coli* (Faith, Hayete et al. 2007), multiple subsequent projects have since taken place in an attempt to improve the accuracy of the inferred networks that are generated (Babu, Musso et al. 2009; Lemmens, De Bie et al. 2009; Zare, Sangurdekar et al. 2009; Kaleta, Gohler et al. 2010). Notable amongst these more recent projects is one by Lemmens *et al* (Lemmens, De Bie et al. 2009) that utilizes a novel, integrative, condition-dependent module network inference method (Lemmens, De Bie et al. 2009) called DISTILLER, that in many ways is similar to the cMonkey and Inferelator module network pipeline described in section 1.6.7. For example, DISTILLER also aims to identify genes with correlated expression profiles that share common binding motifs in their upstream binding regions. However, the motifs that DISTILLER incorporates must be specified prior to runtime, for example, including those from RegulonDB (Gama-Castro, Salgado et al. 2011) or Transfac (Matys, Fricke et al. 2003), thus limiting its capacity to identify novel putative regulatory modules. Despite this limitation, DISTILLER was also recently used by the same group to infer a transcriptional regulatory network for *B. subtilis* (Fadda, Fierro et al. 2009) as well. In addition to this regulatory network for *B. subtilis*, Goezler *et al* developed a

network of metabolic interactions for *B. subtilis* as well, that was built via manual curation (Goelzer, Bekkal Brikci et al. 2008). Finally, Vazquez *et al* (Vazquez, Freyre-Gonzalez et al. 2009) performed a system-wide expression analysis of both *B. subtilis* and *E. coli* to identify and compare the global network governing the response to glucose for each organism.

This last project being an example of how the comparison of the results from multiple functional genomics projects devoted to different organisms offers a look into the evolution of not just sequences but sub-networks, networks and biomodules across bacterial and archaeal clades. This possibility is made particularly exciting by recent advances in the reconstruction of phylogenetic histories of microbes that explicitly model lateral gene transfer**.** Uncovering these relationships at the module and network level (in addition to the sequence level) is possible given the scale of prokaryotic systems; in fact several meta-genomics projects, such as the Human Microbiome Project (Turnbaugh, Ley et al. 2007)**,** already exist and have begun to show results such as the characterization of the community differences between obese and lean individuals (Turnbaugh, Ley et al. 2006; Turnbaugh, Hamady et al. 2009).

Given the large number of prokaryotic functional genomics projects, multi-species analysis (inferring networks and modules over multiple species datasets) is one of the next major challenges, as prokaryotic systems rarely exist in clonal isolation (consortia of microbes inhabiting ecological niches are the relevant system to study in many cases). To prevent any misunderstanding, we should clarify that in the context

of metagenomics that multi-species can mean one of two things. In one sense, "multi-species" can mean the mapping and modeling of the complex interactions between the members of a given microbial community, some of which are known to be dependent on other community members, and cannot survive – or be cultured - on their own. While this is an interesting topic and will be exciting an area of research, current methods are not yet quite ready to provide the level of granularity such an analysis will require.

In the other meaning, "multi-species" is used to refer the leveraging of comparative biological analysis to identify modules and sub-networks that are putatively conserved between organisms. As such, we argue that a multi-species approach like this offers the possibility of identifying more biologically relevant modules than those which a traditional single-species method might find. In chapter 2 of this thesis, we present a novel algorithm that will detect putatively conserved modules by simultaneously considering data from multiple organisms by extending the integrative framework utilized by cMonkey by allowing it to integrate data from multiple organisms. In addition, in chapter 2, we will also present some of the biological highlights that were found when it was applied to a triplet of Gram-positive prokaryotes. In chapter 3, we will present the rigorous validation that was performed to evaluate the results from this triplet of Gram-positive prokaryotes, as well as a second triplet of Gram-negative prokaryotes. Finally, in chapter 4, we will present initial results from a recent "multi-platform" extension of this multi-species method

which has been used to perform a comparative analysis of mouse and human hematopoietic differentiation.

1.9 References

Aggarwal, K. and K. H. Lee (2003). "Functional genomics and proteomics as a foundation for systems biology." Brief Funct Genomic Proteomic **2**(3): 175-184.

Allenby, N. E., N. O'Connor, et al. (2005). "Genome-wide transcriptional analysis of the phosphate starvation stimulon of Bacillus subtilis." J Bacteriol **187**(23): 8063-8080.

Alon, U. (2007). An introduction to systems biology : design principles of biological circuits. Boca Raton, FL, Chapman & Hall/CRC.

Aloy, P., A. Stark, et al. (2003). "Predictions without templates: new folds, secondary structure, and contacts in CASP5." Proteins **53 Suppl 6**: 436-456.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Arends, S. J. and D. S. Weiss (2004). "Inhibiting cell division in Escherichia coli has little if any effect on gene expression." J Bacteriol **186**(3): 880-884.

Arkin, A., J. Ross, et al. (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." Genetics **149**(4): 1633-1648.

Ausmees, N. and C. Jacobs-Wagner (2003). "Spatial and temporal control of differentiation and cell cycle progression in Caulobacter crescentus." Annu Rev Microbiol **57**: 225-247.

Babu, M., G. Musso, et al. (2009). "Systems-level approaches for identifying and analyzing genetic interaction networks in Escherichia coli and extensions to other prokaryotes." Mol Biosyst **5**(12): 1439-1455.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Balasubramanian, R., T. LaFramboise, et al. (2004). "A graph-theoretic approach to testing associations between disparate sources of functional genomics data." Bioinformatics **20**(18): 3353-3362.

Baliga, N. S., S. J. Bjork, et al. (2004). "Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1." Genome Res **14**(6): 1025-1035.

Baliga, N. S., M. Pan, et al. (2002). "Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach." Proc Natl Acad Sci U S A **99**(23): 14913-14918.

Bar-Joseph, Z., G. K. Gerber, et al. (2003). "Computational discovery of gene modules and regulatory networks." Nat Biotechnol **21**(11): 1337-1342.

Barbosa, T. M. and S. B. Levy (2000). "Differential expression of over 60 chromosomal genes in Escherichia coli by constitutive expression of MarA." J Bacteriol **182**(12): 3467-3474.

Barrett, C. L., C. D. Herring, et al. (2005). "The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states." Proc Natl Acad Sci U S A **102**(52): 19103-19108.

Beard, D. A., S. D. Liang, et al. (2002). "Energy balance for analysis of complex metabolic networks." Biophys J **83**(1): 79-86.

Ben-Dor, A., B. Chor, et al. (2003). "Discovering local structure in gene expression data: the order-preserving submatrix problem." J Comput Biol **10**(3-4): 373-384.

Ben-Yehuda, S., M. Fujita, et al. (2005). "Defining a centromere-like element in Bacillus subtilis by Identifying the binding sites for the chromosome-anchoring protein RacA." Mol Cell **17**(6): 773-782.

Bergmann, S., J. Ihmels, et al. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." Phys Rev E Stat Nonlin Soft Matter Phys **67**(3 Pt 1): 031902.

Berka, R. M., J. Hahn, et al. (2002). "Microarray analysis of the Bacillus subtilis K-state: genome-wide expression changes dependent on ComK." Mol Microbiol **43**(5): 1331-1345.

Biondi, E. G., S. J. Reisinger, et al. (2006). "Regulation of the bacterial cell cycle by an integrated genetic circuit." Nature **444**(7121): 899-904.

Biondi, E. G., J. M. Skerker, et al. (2006). "A phosphorelay system controls stalk biogenesis during cell cycle progression in Caulobacter crescentus." Mol Microbiol **59**(2): 386-401.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-1474.

Bogomolni, R. A. and J. L. Spudich (1982). "Identification of a third rhodopsin-like pigment in phototactic Halobacterium halobium." Proc Natl Acad Sci U S A **79**(20): 6250-6254.

Bonneau, R. (2006). The Inferelator Cytoscape Web Start.

Bonneau, R., N. S. Baliga, et al. (2004). "Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1." Genome Biol **5**(8): R52.

Bonneau, R., M. T. Facciotti, et al. (2007). "A predictive model for transcriptional control of physiology in a free living cell." Cell **131**(7): 1354-1365.

Bonneau, R., D. J. Reiss, et al. (2006). "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo." Genome Biol **7**(5): R36.

Bonneau, R., C. E. Strauss, et al. (2001). "Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation." Proteins **43**(1): 1-11.

Bonneau, R., J. Tsai, et al. (2001). "Rosetta in CASP4: progress in ab initio protein structure prediction." Proteins **Suppl 5**: 119-126.

Bradley, P., D. Chivian, et al. (2003). "Rosetta predictions in CASP5: successes, failures, and prospects for complete automation." Proteins **53 Suppl 6**: 457-468.

Bray, D. (1995). "Protein molecules as computational elements in living cells." Nature **376**(6538): 307-312.

Britton, R. A., P. Eichenberger, et al. (2002). "Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of Bacillus subtilis." J Bacteriol **184**(17): 4881-4890.

Brocklehurst, K. R. and A. P. Morby (2000). "Metal-ion tolerance in Escherichia coli: analysis of transcriptional profiles by gene-array technology." Microbiology **146 (Pt 9)**: 2277-2282.

Brooun, A., J. Bell, et al. (1998). "An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins." J Bacteriol **180**(7): 1642-1646.

Buttner, K., J. Bernhardt, et al. (2001). "A comprehensive two-dimensional map of cytosolic proteins of Bacillus subtilis." Electrophoresis **22**(14): 2908-2935.

Celniker, S. E., L. A. Dillon, et al. (2009). "Unlocking the secrets of the genome." Nature **459**(7249): 927-930.

Chada, V. G., E. A. Sanstad, et al. (2003). "Morphogenesis of bacillus spore surfaces." J Bacteriol **185**(21): 6255-6261.

Chandrasekaran, S. and N. D. Price (2010). "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis." Proc Natl Acad Sci U S A **107**(41): 17845-17850.

Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.

Chivian, D., D. E. Kim, et al. (2003). "Automated prediction of CASP-5 structures using the Robetta server." Proteins **53 Suppl 6**: 524-533.

Courcelle, J., A. Khodursky, et al. (2001). "Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli." Genetics **158**(1): 41-64.

Covert, M. W., E. M. Knight, et al. (2004). "Integrating high-throughput and computational data elucidates bacterial networks." Nature **429**(6987): 92-96.

Covert, M. W. and B. O. Palsson (2002). "Transcriptional regulation in constraints-based metabolic models of Escherichia coli." J Biol Chem **277**(31): 28058-28064.

Covert, M. W., C. H. Schilling, et al. (2001). "Regulation of gene expression in flux balance models of metabolism." J Theor Biol **213**(1): 73-88.

D'Haeseleer, P., S. Liang, et al. (2000). "Genetic network inference: from co-expression clustering to reverse engineering." Bioinformatics **16**(8): 707-726.

D'Haeseleer, P., X. Wen, et al. (1999). "Linear modeling of mRNA expression levels during CNS development and injury." Pac Symp Biocomput: 41-52.

Das, D., N. Banerjee, et al. (2004). "Interacting models of cooperative gene regulation." Proc Natl Acad Sci U S A **101**(46): 16234-16239.

DasSarma, S. (1993). "Identification and analysis of the gas vesicle gene cluster on an unstable plasmid of Halobacterium halobium." Experientia **49**(6-7): 482-486.

Dassarma, S., B. R. Berquist, et al. (2006). "Post-genomics of the model haloarchaeon Halobacterium sp. NRC-1." Saline Systems **2**: 3.

Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res **27**(23): 4636-4641.

DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-686.

Dittmar, K. A., E. M. Mobley, et al. (2004). "Exploring the regulation of tRNA distribution on the genomic scale." J Mol Biol **337**(1): 31-47.

Earl, A. M., R. Losick, et al. (2007). "Bacillus subtilis genome diversity." J Bacteriol **189**(3): 1163-1170.

Edwards, J. S., R. U. Ibarra, et al. (2001). "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data." Nat Biotechnol **19**(2): 125-130.

Edwards, J. S. and B. O. Palsson (2000). "The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities." Proc Natl Acad Sci U S A **97**(10): 5528-5533.

Edwards, J. S., R. Ramakrishna, et al. (2002). "Characterizing the metabolic phenotype: a phenotype phase plane analysis." Biotechnol Bioeng **77**(1): 27-36.

Efron, B. (2003). "Robbins, Empirical Bayes and Microarrays."

Eichenberger, P., M. Fujita, et al. (2004). "The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis." PLoS Biol **2**(10): e328.

Eichenberger, P., S. T. Jensen, et al. (2003). "The sigmaE regulon and the identification of additional sporulation genes in Bacillus subtilis." J Mol Biol **327**(5): 945-972.

Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-826.

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.

Errington, J. (2003). "Regulation of endospore formation in Bacillus subtilis." Nat Rev Microbiol **1**(2): 117-126.

Eymann, C., A. Dreisbach, et al. (2004). "A comprehensive proteome map of growing Bacillus subtilis cells." Proteomics **4**(10): 2849-2876.

Fabret, C., V. A. Feher, et al. (1999). "Two-component signal transduction in Bacillus subtilis: how one organism sees its world." J Bacteriol **181**(7): 1975-1983.

Fadda, A., A. C. Fierro, et al. (2009). "Inferring the transcriptional network of Bacillus subtilis." Mol Biosyst **5**(12): 1840-1852.

Faith, J. J., B. Hayete, et al. (2007). "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles." PLoS Biology **5**(1): e8.

Fawcett, P., P. Eichenberger, et al. (2000). "The transcriptional profile of early to middle sporulation in Bacillus subtilis." Proc Natl Acad Sci U S A **97**(14): 8063-8068.

Ferreira, L. C., R. C. Ferreira, et al. (2005). "Bacillus subtilis as a tool for vaccine development: from antigen factories to delivery vectors." An Acad Bras Cienc **77**(1): 113-124.

Feucht, A., L. Evans, et al. (2003). "Identification of sporulation genes by genome-wide analysis of the sigmaE regulon of Bacillus subtilis." Microbiology **149**(Pt 10): 3023-3034.

Fiehn, O. (2002). "Metabolomics--the link between genotypes and phenotypes." Plant Mol Biol **48**(1-2): 155-171.

Fields, S. and O.-k. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-246.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-251.

Fischer, D., L. Rychlewski, et al. (2003). "CAFASP3: the third critical assessment of fully automated structure prediction methods." Proteins **53 Suppl 6**: 503-516.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Foster, J. W. (2004). "Escherichia coli acid resistance: tales of an amateur acidophile." Nat Rev Microbiol **2**(11): 898-907.

Friedman, N., M. Linial, et al. (2000). "Using Bayesian networks to analyze expression data." J Comput Biol **7**(3-4): 601-620.

Gama-Castro, S., H. Salgado, et al. (2011). "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)." Nucleic Acids Res **39**(Database issue): D98-105.

Ge, H., A. J. Walhout, et al. (2003). "Integrating 'omic' information: a bridge between genomics and systems biology." Trends Genet **19**(10): 551-560.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Glasner, J. D., P. Liss, et al. (2003). "ASAP, a systematic annotation package for community analysis of genomes." Nucleic Acids Res **31**(1): 147-151.

Goelzer, A., F. Bekkal Brikci, et al. (2008). "Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis." BMC Systems Biology **2**(1): 20.

Guardia, M. J., A. Gambhir, et al. (2000). "Cybernetic modeling and regulation of metabolic pathways in multiple steady states of hybridoma cells." Biotechnol Prog **16**(5): 847-853.

Gygi, S. P., B. Rist, et al. (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**(10): 994-999.

Hamon, M. A., N. R. Stanley, et al. (2004). "Identification of AbrB-regulated genes involved in biofilm formation by Bacillus subtilis." Mol Microbiol **52**(3): 847-860.

Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." Microbiol Mol Biol Rev **68**(4): 669-685.

Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.

Hashimoto, R. F., S. Kim, et al. (2004). "Growing genetic regulatory networks from seed genes." Bioinformatics **20**(8): 1241-1247.

Hayashi, K., T. Ohsawa, et al. (2005). "The H2O2 stress-responsive regulator PerR positively regulates srfA expression in Bacillus subtilis." J Bacteriol **187**(19): 6659-6667.

Helmann, J. D., M. F. Wu, et al. (2001). "Global transcriptional response of Bacillus subtilis to heat shock." J Bacteriol **183**(24): 7318-7328.

Hilbert, D. W. and P. J. Piggot (2004). "Compartmentalization of gene expression during Bacillus subtilis spore formation." Microbiol Mol Biol Rev **68**(2): 234-262.

Holtzendorff, J., D. Hung, et al. (2004). "Oscillating global regulators control the genetic circuit driving a bacterial cell cycle." Science **304**(5673): 983-987.

Holtzendorff, J., J. Reinhardt, et al. (2006). "Cell cycle control by oscillating regulatory proteins in Caulobacter crescentus." Bioessays **28**(4): 355-361.

Hommais, F., E. Krin, et al. (2001). "Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS." Mol Microbiol **40**(1): 20-36.

Hong, R. W., M. Shchepetov, et al. (2003). "Transcriptional profile of the Escherichia coli response to the antimicrobial insect peptide cecropin A." Antimicrob Agents Chemother **47**(1): 1-6.

Hottes, A. K., L. Shapiro, et al. (2005). "DnaA coordinates replication initiation and cell cycle transcription in Caulobacter crescentus." Mol Microbiol **58**(5): 1340-1353.

Hung, D. Y. and L. Shapiro (2002). "A signal transduction protein cues proteolytic events critical to Caulobacter cell cycle progression." Proc Natl Acad Sci U S A **99**(20): 13160-13165.

Ibarra, R. U., J. S. Edwards, et al. (2002). "Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth." Nature **420**(6912): 186-189.

Ideker, T., O. Ozier, et al. (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics **18 Suppl 1**: S233-240.

Iniesta, A. A., P. T. McGrath, et al. (2006). "A phospho-signaling pathway controls the localization and activity of a protease complex critical for bacterial cell cycle progression." Proc Natl Acad Sci U S A **103**(29): 10935-10940.

Jacobs-Wagner, C. (2004). "Regulatory proteins with a sense of direction: cell cycle signalling network in Caulobacter." Mol Microbiol **51**(1): 7-13.

Jacobs, C., N. Ausmees, et al. (2003). "Functions of the CckA histidine kinase in Caulobacter cell cycle control." Mol Microbiol **47**(5): 1279-1290.

Jardine, O., J. Gough, et al. (2002). "Comparison of the small molecule metabolic enzymes of Escherichia coli and Saccharomyces cerevisiae." Genome Res **12**(6): 916-929.

Jürgen Richter-Gebert, B. S., Thorsten Theobald (2003). "First steps in tropical geometry." Proc. Conference on Idempotent Mathematics and Mathematical Physics.

Kaleta, C., A. Gohler, et al. (2010). "Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis." BMC Syst Biol **4**: 116.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Res **34**(Database issue): D354-357.

Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-280.

Karp, P. D., M. Riley, et al. (2000). "The EcoCyc and MetaCyc databases." Nucleic Acids Res **28**(1): 56-59.

Keijser, B. J., A. T. Beek, et al. (2007). "Analysis of temporal gene expression during Bacillus subtilis spore germination and outgrowth." J Bacteriol **23**: 23.

Keseler, I. M., J. Collado-Vides, et al. (2005). "EcoCyc: a comprehensive database resource for Escherichia coli." Nucleic Acids Res **33**(Database issue): D334-337.

Kinch, L. N., J. O. Wrabl, et al. (2003). "CASP5 assessment of fold recognition target predictions." Proteins **53 Suppl 6**: 395-409.

Kluger, Y., R. Basri, et al. (2003). "Spectral biclustering of microarray data: coclustering genes and conditions." Genome Res **13**(4): 703-716.

Kobayashi, K., M. Ogura, et al. (2001). "Comprehensive DNA microarray analysis of Bacillus subtilis two-component regulatory systems." J Bacteriol **183**(24): 7365-7370.

Kolbe, M., H. Besir, et al. (2000). "Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution." Science **288**(5470): 1390-1396.

Kunkel, B., L. Kroos, et al. (1989). "Temporal and spatial control of the mother-cell regulatory gene spoIIID of Bacillus subtilis." Genes Dev **3**(11): 1735-1744.

Kunst, F., N. Ogasawara, et al. (1997). "The complete genome sequence of the Gram-positive bacterium Bacillus subtilis." Nature **390**(6657): 249-256.

Laub, M. T., S. L. Chen, et al. (2002). "Genes directly controlled by CtrA, a master regulator of the Caulobacter cell cycle." Proc Natl Acad Sci U S A **99**(7): 4632-4637.

Laub, M. T., H. H. McAdams, et al. (2000). "Global analysis of the genetic network controlling a bacterial cell cycle." Science **290**(5499): 2144-2148.

Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.

Lemmens, K., T. De Bie, et al. (2009). "DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli." Genome Biol **10**(3): R27.

Lemmens, K., T. De Bie, et al. (2009). "The condition-dependent transcriptional network in Escherichia coli." Ann N Y Acad Sci **1158**: 29-35.

Lewis, N. E., B. K. Cho, et al. (2009). "Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content." J Bacteriol **191**(11): 3437-3444.

Lin, J., I. S. Lee, et al. (1995). "Comparative analysis of extreme acid survival in Salmonella typhimurium, Shigella flexneri, and Escherichia coli." J Bacteriol **177**(14): 4097-4104.

Lin, J. T., M. B. Connelly, et al. (2005). "Global transcriptional response of Bacillus subtilis to treatment with subinhibitory concentrations of antibiotics that inhibit protein synthesis." Antimicrob Agents Chemother **49**(5): 1915-1926.

Loew, L. M. and J. C. Schaff (2001). "The Virtual Cell: a software environment for computational cell biology." Trends Biotechnol **19**(10): 401-406.

Luecke, H., B. Schobert, et al. (2000). "Coupling photoisomerization of retinal to directional transport in bacteriorhodopsin." J Mol Biol **300**(5): 1237-1255.

Malmstrom, L., M. Riffle, et al. (2007). "Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology." PLoS Biol **5**(4): e76.

Mangan, S. and U. Alon (2003). "Structure and function of the feed-forward loop network motif." Proc Natl Acad Sci U S A **100**(21): 11980-11985.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-753.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Mascher, T., J. D. Helmann, et al. (2006). "Stimulus perception in bacterial signal-transducing histidine kinases." Microbiol Mol Biol Rev **70**(4): 910-938.

Masuda, N. and G. M. Church (2002). "Escherichia coli gene expression responsive to levels of the response regulator EvgA." J Bacteriol **184**(22): 6225-6234.

Masuda, N. and G. M. Church (2003). "Regulatory network of acid resistance genes in Escherichia coli." Mol Microbiol **48**(3): 699-712.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Research **31**(1): 374-378.

Meile, J. C., L. J. Wu, et al. (2006). "Systematic localisation of proteins fused to the green fluorescent protein in Bacillus subtilis: identification of new proteins at the DNA replication factory." Proteomics **6**(7): 2135-2146.

Mellor, J. C., I. Yanai, et al. (2002). "Predictome: a database of putative functional links between proteins." Nucleic Acids Res **30**(1): 306-309.

Miller, M. B. and B. L. Bassler (2001). "Quorum sensing in bacteria." Annu Rev Microbiol **55**: 165-199.

Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-827.

Molle, V., M. Fujita, et al. (2003). "The Spo0A regulon of Bacillus subtilis." Mol Microbiol **50**(5): 1683-1701.

Moraru, II, J. C. Schaff, et al. (2002). "The virtual cell: an integrated modeling environment for experimental and computational cell biology." Ann N Y Acad Sci **971**: 595-596.

Moreno-Hagelsieb, G. and J. Collado-Vides (2002). "A powerful non-homology method for the prediction of operons in prokaryotes." Bioinformatics **18 Suppl 1**: S329-336.

Moszer, I. (1998). "The complete genome of Bacillus subtilis: from sequence annotation to data management and analysis." FEBS Lett **430**(1-2): 28-36.

Moszer, I., L. M. Jones, et al. (2002). "SubtiList: the reference database for the Bacillus subtilis genome." Nucleic Acids Res **30**(1): 62-65.

Nakano, S., E. Kuster-Schock, et al. (2003). "Spx-dependent global transcriptional control is induced by thiol-specific oxidative stress in Bacillus subtilis." Proc Natl Acad Sci U S A **100**(23): 13603-13608.

Negre, N., C. D. Brown, et al. (2011). "A cis-regulatory map of the Drosophila genome." Nature **471**(7339): 527-531.

Ng, W. V., S. P. Kennedy, et al. (2000). "Genome sequence of Halobacterium species NRC-1." Proc Natl Acad Sci U S A **97**(22): 12176-12181.

Nierman, W. C., T. V. Feldblyum, et al. (2001). "Complete genome sequence of Caulobacter crescentus." Proc Natl Acad Sci U S A **98**(7): 4136-4141.

Nishino, K., Y. Inazumi, et al. (2003). "Global analysis of genes regulated by EvgA of the two-component regulatory system in Escherichia coli." J Bacteriol **185**(8): 2667-2672.

Nishino, K. and A. Yamaguchi (2001). "Analysis of a complete library of putative drug transporter genes in Escherichia coli." J Bacteriol **183**(20): 5803-5812.

Nobeli, I., H. Ponstingl, et al. (2003). "A structure-based anatomy of the E.coli metabolome." J Mol Biol **334**(4): 697-719.

Noirot-Gros, M. F., E. Dervyn, et al. (2002). "An expanded view of bacterial DNA replication." Proc Natl Acad Sci U S A **99**(12): 8342-8347.

Ogata, H., S. Goto, et al. (1998). "Computation with the KEGG pathway database." Biosystems **47**(1-2): 119-128.

Ogura, M. and Y. Fujita (2007). "Bacillus subtilis rapD, a direct target of transcription repression by RghR, negatively regulates srfA expression." FEMS Microbiol Lett **268**(1): 73-80.

Ogura, M., K. Tsukahara, et al. (2007). "The Bacillus subtilis NatK-NatR two-component system regulates expression of the natAB operon encoding an ABC transporter for sodium ion extrusion." Microbiology **153**(Pt 3): 667-675.

Ogura, M., H. Yamaguchi, et al. (2002). "Whole-genome analysis of genes regulated by the Bacillus subtilis competence transcription factor ComK." J Bacteriol **184**(9): 2344-2351.

Ogura, M., H. Yamaguchi, et al. (2001). "DNA microarray analysis of Bacillus subtilis DegU, ComA and PhoP regulons: an approach to comprehensive analysis of B.subtilis two-component regulatory systems." Nucleic Acids Res **29**(18): 3804-3813.

Ohashi, Y., T. Inaoka, et al. (2003). "Expression profiling of translation-associated genes in sporulating Bacillus subtilis and consequence of sporulation by gene inactivation." Biosci Biotechnol Biochem **67**(10): 2245-2253.

Palsson, B. (2006). Systems biology : properties of reconstructed networks. Cambridge ; New York, Cambridge University Press.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature reviews. Genetics **10**(10): 669-680.

Peregrin-Alvarez, J. M., S. Tsoka, et al. (2003). "The phylogenetic extent of metabolic enzymes and pathways." Genome Res **13**(3): 422-427.

Piggot, P. J. and D. W. Hilbert (2004). "Sporulation of Bacillus subtilis." Curr Opin Microbiol **7**(6): 579-586.

Poetz, O., J. M. Schwenk, et al. (2005). "Protein microarrays: catching the proteome." Mech Ageing Dev **126**(1): 161-170.

Polen, T., D. Rittmann, et al. (2003). "DNA microarray analyses of the long-term adaptive response of Escherichia coli to acetate and propionate." Appl Environ Microbiol **69**(3): 1759-1774.

Pomposiello, P. J., M. H. Bennik, et al. (2001). "Genome-wide transcriptional profiling of the Escherichia coli responses to superoxide stress and sodium salicylate." J Bacteriol **183**(13): 3890-3902.

Prelic, A., S. Bleuler, et al. (2006). "A systematic comparison and evaluation of biclustering methods for gene expression data." Bioinformatics **22**(9): 1122-1129.

Price, N. D., J. A. Papin, et al. (2003). "Genome-scale microbial in silico models: the constraints-based approach." Trends Biotechnol **21**(4): 162-169.

Qian, J., Y. Kluger, et al. (2003). "Identification and correction of spurious spatial correlations in microarray data." Biotechniques **35**(1): 42-44, 46, 48.

R. A. Majewski and M. M. Domach (1990). "Simple constrained-optimization view of acetate overflow in *E. coli*." Biotechnology and Bioengineering **35**(7): 732-738.

Reed, J. L. and B. O. Palsson (2003). "Thirteen years of building constraint-based in silico models of Escherichia coli." J Bacteriol **185**(9): 2692-2699.

Reed, J. L., T. D. Vo, et al. (2003). "An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)." Genome Biol **4**(9): R54.

Reiss, D. J., N. S. Baliga, et al. (2006). "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." BMC Bioinformatics **7**: 280.

Ren, D., L. A. Bedzyk, et al. (2004). "Differential gene expression to investigate the effect of (5Z)-4-bromo- 5-(bromomethylene)-3-butyl-2(5H)-furanone on Bacillus subtilis." Appl Environ Microbiol **70**(8): 4941-4949.

Richmond, C. S., J. D. Glasner, et al. (1999). "Genome-wide expression profiling in Escherichia coli K-12." Nucleic Acids Res **27**(19): 3821-3835.

Roth, F. P., J. D. Hughes, et al. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol **16**(10): 939-945.

Ruepp, A. and J. Soppa (1996). "Fermentative arginine degradation in Halobacterium salinarium (formerly Halobacterium halobium): genes, gene products, and transcripts of the arcRACB gene cluster." J Bacteriol **178**(16): 4942-4947.

Rusch, D. B., A. L. Halpern, et al. (2007). "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific." PLoS Biol **5**(3): e77.

Salgado, H., S. Gama-Castro, et al. (2006). "RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions." Nucleic Acids Res **34**(Database issue): D394-397.

Salzberg, S. L., A. L. Delcher, et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-548.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-176.

Segal, E., B. Taskar, et al. (2001). "Rich probabilistic models for gene expression." Bioinformatics **17 Suppl 1**: S243-252.

Selinger, D. W., K. J. Cheung, et al. (2000). "RNA expression analysis using a 30 base pair resolution Escherichia coli genome array." Nat Biotechnol **18**(12): 1262-1268.

Selkov, E., Jr., Y. Grechkin, et al. (1998). "MPW: the Metabolic Pathways Database." Nucleic Acids Res **26**(1): 43-45.

Serizawa, M., K. Kodama, et al. (2005). "Functional analysis of the YvrGHb two-component system of Bacillus subtilis: identification of the regulated genes by DNA microarray and northern blot analyses." Biosci Biotechnol Biochem **69**(11): 2155-2169.

Serres, M. H., S. Gopal, et al. (2001). "A functional update of the Escherichia coli K-12 genome." Genome Biol **2**(9): RESEARCH0035.

Setlow, P. (2003). "Spore germination." Curr Opin Microbiol **6**(6): 550-556.

Shamir, R., A. Maron-Katz, et al. (2005). "EXPANDER--an integrative program suite for microarray data analysis." BMC Bioinformatics **6**: 232.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Shannon, P. T., D. J. Reiss, et al. (2006). "The Gaggle: an open-source software system for integrating bioinformatics software and data sources." BMC Bioinformatics **7**: 176.

Shen-Orr, S. S., R. Milo, et al. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-68.

Sheng, Q., Y. Moreau, et al. (2003). "Biclustering microarray data by Gibbs sampling." Bioinformatics **19 Suppl 2**: ii196-205.

Shmulevich, I., H. Lahdesmaki, et al. (2003). "The role of certain Post classes in Boolean network models of genetic networks." Proc Natl Acad Sci U S A **100**(19): 10734-10739.

Silvaggi, J. M., J. B. Perkins, et al. (2006). "Genes for small, noncoding RNAs under sporulation control in Bacillus subtilis." J Bacteriol **188**(2): 532-541.

Skerker, J. M. and M. T. Laub (2004). "Cell-cycle progression and the generation of asymmetry in Caulobacter crescentus." Nat Rev Microbiol **2**(4): 325-337.

Skerker, J. M., M. S. Prasol, et al. (2005). "Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis." PLoS Biol **3**(10): e334.

Slonim, N., N. Friedman, et al. (2006). "Multivariate information bottleneck." Neural Comput **18**(8): 1739-1789.

Soga, T., Y. Ohashi, et al. (2003). "Quantitative metabolome analysis using capillary electrophoresis mass spectrometry." J Proteome Res **2**(5): 488-494.

Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proc Natl Acad Sci U S A **103**(32): 12115-12120.

Spudich, E. N., T. Takahashi, et al. (1989). "Sensory rhodopsins I and II modulate a methylation/demethylation system in Halobacterium halobium phototaxis." Proc Natl Acad Sci U S A **86**(20): 7746-7750.

Spudich, J. L. (1993). "Color sensing in the Archaea: a eukaryotic-like receptor coupled to a prokaryotic transducer." J Bacteriol **175**(24): 7755-7761.

Spudich, J. L. and R. A. Bogomolni (1984). "Mechanism of colour discrimination by a bacterial sensory rhodopsin." Nature **312**(5994): 509-513.

Steil, L., M. Serrano, et al. (2005). "Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of Bacillus subtilis." Microbiology **151**(Pt 2): 399-420.

Stevens, C. M. and J. Errington (1990). "Differential gene expression during sporulation in Bacillus subtilis: structure and regulation of the spoIIID gene." Mol Microbiol **4**(4): 543-551.

Stock, A. M., V. L. Robinson, et al. (2000). "Two-component signal transduction." Annu Rev Biochem **69**: 183-215.

Streips, U. N. and F. W. Polio (1985). "Heat shock proteins in bacilli." J Bacteriol **162**(1): 434-437.

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-255.

Tam le, T., H. Antelmann, et al. (2006). "Proteome signatures for stress and starvation in Bacillus subtilis as revealed by a 2-D gel image color coding approach." Proteomics **6**(16): 4565-4585.

Tanay, A., R. Sharan, et al. (2004). "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proc Natl Acad Sci U S A **101**(9): 2981-2986.

Tanay, A., R. Sharan, et al. (2002). "Discovering statistically significant biclusters in gene expression data." Bioinformatics **18 Suppl 1**: S136-144.

Tao, H., C. Bausch, et al. (1999). "Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media." J Bacteriol **181**(20): 6425-6440.

Tatti, K. M., C. H. Jones, et al. (1991). "Genetic evidence for interaction of sigma E with the spoIIID promoter in Bacillus subtilis." J Bacteriol **173**(24): 7828-7833.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.

Tatusov, R. L., D. A. Natale, et al. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1): 22-28.

Thorsson, V., M. Hornquist, et al. (2005). "Reverse engineering galactose regulation in yeast through model selection." Stat Appl Genet Mol Biol **4**: Article28.

Tomasinsig, L., M. Scocchi, et al. (2004). "Genome-wide transcriptional profiling of the Escherichia coli response to a proline-rich antimicrobial peptide." Antimicrob Agents Chemother **48**(9): 3260-3267.

Tringe, S. G. and E. M. Rubin (2005). "Metagenomics: DNA sequencing of environmental samples." Nat Rev Genet **6**(11): 805-814.

Tucker, D. L., N. Tucker, et al. (2002). "Gene expression profiling of the pH response in Escherichia coli." J Bacteriol **184**(23): 6551-6558.

Tucker, D. L., N. Tucker, et al. (2003). "Genes of the GadX-GadW regulon in Escherichia coli." J Bacteriol **185**(10): 3190-3201.

Turnbaugh, P. J., M. Hamady, et al. (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484.

Turnbaugh, P. J., R. E. Ley, et al. (2007). "The human microbiome project." Nature **449**(7164): 804-810.

Turnbaugh, P. J., R. E. Ley, et al. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**(7122): 1027-1031.

Tyson, G. W., J. Chapman, et al. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature **428**(6978): 37-43.

van Someren, E. P., L. F. Wessels, et al. (2002). "Genetic network modeling." Pharmacogenomics **3**(4): 507-525.

van Someren, E. P., L. F. Wessels, et al. (2000). "Linear modeling of genetic networks from experimental data." Proc Int Conf Intell Syst Mol Biol **8**: 355-366.

Vanet, A., L. Marsan, et al. (2000). "Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals." J Mol Biol **297**(2): 335-353.

Varma, A. and B. O. Palsson (1994). "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110." Appl Environ Microbiol **60**(10): 3724-3731.

Varner, J. and D. Ramkrishna (1998). "Application of cybernetic models to metabolic engineering: investigation of storage pathways." Biotechnol Bioeng **58**(2-3): 282-291.

Varner, J. and D. Ramkrishna (1999). "Metabolic engineering from a cybernetic perspective. 1. Theoretical preliminaries." Biotechnol Prog **15**(3): 407-425.

Vazquez, C. D., J. A. Freyre-Gonzalez, et al. (2009). "Identification of network topological units coordinating the global expression response to glucose in Bacillus subtilis and its comparison to Escherichia coli." BMC Microbiol **9**: 176.

Velculescu, V. E., L. Zhang, et al. (1995). "Serial analysis of gene expression." Science **270**(5235): 484-487.

Vemuri, G. N. and A. A. Aristidou (2005). "Metabolic engineering in the -omics era: elucidating and modulating regulatory networks." Microbiol Mol Biol Rev **69**(2): 197-216.

Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

von Dassow, G., E. Meir, et al. (2000). "The segment polarity network is a robust developmental module." Nature **406**(6792): 188-192.

Wahde, M. and J. Hertz (2001). "Modeling genetic regulatory dynamics in neural development." J Comput Biol **8**(4): 429-442.

Walhout, A. J. and M. Vidal (2001). "High-throughput yeast two-hybrid assays for large-scale protein interaction mapping." Methods **24**(3): 297-306.

Waltman, P., T. Kacmarczyk, et al. (2009). Prokaryotic Systems Biology. Plant systems biology, Annual plant reviews. G. Coruzzi and R. A. Gutierrez. Ames, Iowa, Blackwell Pub.**:** 67-136.

Wang, A. and D. E. Crowley (2005). "Global gene expression responses to cadmium toxicity in Escherichia coli." J Bacteriol **187**(9): 3259-3266.

Wang, S. T., B. Setlow, et al. (2006). "The forespore line of gene expression in Bacillus subtilis." J Mol Biol **358**(1): 16-37.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews. Genetics **10**(1): 57-63.

Weaver, D. C., C. T. Workman, et al. (1999). "Modeling regulatory networks with weight matrices." Pac Symp Biocomput: 112-123.

Weber, A. and K. Jung (2002). "Profiling early osmostress-dependent gene expression in Escherichia coli using DNA macroarrays." J Bacteriol **184**(19): 5502-5507.

Wei, Y., J. M. Lee, et al. (2001). "High-density microarray-mediated gene expression profiling of Escherichia coli." J Bacteriol **183**(2): 545-556.

Wodicka, L., H. Dong, et al. (1997). "Genome-wide expression monitoring in Saccharomyces cerevisiae." Nat Biotechnol **15**(13): 1359-1367.

Wolff, S., H. Antelmann, et al. (2007). "Towards the entire proteome of the model bacterium Bacillus subtilis by gel-based and gel-free approaches." J Chromatogr B Analyt Technol Biomed Life Sci **849**(1-2): 129-140.

Wolff, S., A. Otto, et al. (2006). "Gel-free and gel-based proteomics in Bacillus subtilis: a comparative study." Mol Cell Proteomics **5**(7): 1183-1192.

Yamane, K., K. Bunai, et al. (2004). "Protein traffic for secretion and related machinery of Bacillus subtilis." Biosci Biotechnol Biochem **68**(10): 2007-2023.

Ye, R. W., W. Tao, et al. (2000). "Global gene expression profiles of Bacillus subtilis grown under anaerobic conditions." J Bacteriol **182**(16): 4458-4465.

Yooseph, S., G. Sutton, et al. (2007). "The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families." PLoS Biol **5**(3): e16.

Yoshida, K., K. Kobayashi, et al. (2001). "Combined transcriptome and proteome analysis as a powerful approach to study genes under glucose repression in Bacillus subtilis." Nucleic Acids Res **29**(3): 683-692.

Yu, J. Y. a. H. W. a. W. W. a. P. (2003). "Enhanced biclustering on expression data." Yang,J., Wang,H., Wang,W., and Yu,P. 2003. Enhanced biclustering on expression data. In Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering (BIBE), pp. 321-327.

Yu, J. Y. a. W. W. a. H. W. a. P. S. (2002). delta-cluster: Capturing Subspace Correlation in a Large Data Set. {Icde}.

Yuh, C. H., H. Bolouri, et al. (1998). "Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene." Science **279**(5358): 1896-1902.

Zare, H., D. Sangurdekar, et al. (2009). "Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations." BMC Syst Biol **3**: 39.

Zhang, G., D. S. Spellman, et al. (2006). "Quantitative phosphotyrosine proteomics of EphB2 signaling by stable isotope labeling with amino acids in cell culture (SILAC)." J Proteome Res **5**(3): 581-588.

Zheng, L. B. and R. Losick (1990). "Cascade regulation of spore coat gene expression in Bacillus subtilis." J Mol Biol **212**(4): 645-660.

Zheng, M., X. Wang, et al. (2001). "DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide." J Bacteriol **183**(15): 4562-4570.

2. MULTI-SPECIES INTEGRATIVE BICLUSTERING

Co-written with: Thadeous Kacmarczyk[2*], Ashley R Bate[2], Daniel B. Kearns[5], David J. Reiss[4], Patrick Eichenberger[2+], Richard Bonneau[1, 2, 3+]

1 – Computer Science Department; Warren Weaver Hall (Room 305); 251 Mercer Street; New York, NY 10012; USA

2 – Center for Genomics and Systems Biology; Department of Biology; New York University; Silver Building (Room 1009); 100 Washington Square East; New York, NY 10003; USA

3 – Computational Biology Program; New York University; Warren Weaver Hall (Room 1105); 251 Mercer Street; New York, NY 10012; USA

4 – Institute for Systems Biology; 1441 North 34th Street; Seattle, WA 98103; USA

5 – Department of Biology; Indiana University; 1001 East 3rd Street; Jordan Hall 142; Bloomington, IN 47405; USA

***** This authors contributed equally to this work.

Keywords: motility, sporulation, clustering, data-integration, *Bacillus*, regulation

Original article: Waltman, P., T. Kacmarczyk, et al. (2010). "Multi-species integrative biclustering." Genome Biololgy **11**(R96).

NOTE: The sections of the original article that this chapter is based upon that describe the quantitative analysis which was performed have been combined with the relevant method sections of the original supplementary material to form the basis of chapter 3. The gene lists and images that were also contained in the supplement of the original article can now be found in Appendix 1, while the additional plots from the original supplementary material can now be found in Appendix 2.

Author contributions: Provided below, in section 2.7.

Abstract

We describe an algorithm, multi-species cMonkey, for the simultaneous biclustering of heterogeneous multiple-species data collections and apply the algorithm to two triplets of bacteria. The first of these is a triplet of Gram-positive bacteria consisting of *Bacillus subtilis*, *Bacillus anthracis*, and *Listeria monocytogenes*, while the second is a triplet of Gram-negative bacteria that includes *Escherichia coli*, *Salmonella typhimurium* and *Vibrio cholerae*. The algorithm reveals evolutionary insights into the surprisingly high degree of conservation of regulatory modules across these three species and allows data and insights from well-studied organisms to complement the analysis of related but less well studied organisms. This chapter is heavily based upon our article which was published in Genome Biology (Waltman, Kacmarczyk et al. 2010).

2.1 Introduction

The rapidly increasing volume of genome scale data has enabled global regulatory network inference and genome-wide prediction of gene function within single organisms. In this work, we exploit another advantage of the growing quantity of genomics data: by comparing genome-wide datasets for closely related organisms, we can add a critical evolutionary component to systems biology data analysis. Whereas several well-developed tools exist for identifying orthologous genes on the basis of sequence similarity, the identification of conserved co-regulated gene groups (modules) is a relatively recent problem requiring development of new methods. Here, we present an algorithm that performs integrative biclustering for multiple-species datasets in order to identify conserved modules and the conditions under which these modules are active. The advantages of this method are that 1) conserved modules are more likely to be biologically significant than co-regulated gene groups lacking detectable conservation, and 2) the identification of these conserved modules can provide a basis for investigating the evolution of gene regulatory networks.

Clustering has long been a popular tool in analyzing systems biology data types (e.g. the clustering of microarray data to generate putative co-regulated gene groups). The majority of genomics studies employ clustering methods that require genes to participate in mutually exclusive clusters, such as hierarchical agglomerative clustering (HAC) (McQuitty 1966), k-means clustering (MacQueen 1967) and singular value decomposition derived methods (Golub and Kahan 1965; Alter, Brown

et al. 2000; Alter, Brown et al. 2003). Because most genes are unlikely to be co-regulated under every possible condition (for instance, bacterial genes can have more than one transcription start site and, in that case, each site will be regulated by a different set of transcription factors depending on the cell's state), defining mutually exclusive gene clusters cannot capture the complexity of transcriptional regulatory networks. Clearly, sophisticated integrative methods are needed to arrive at the identification of more mechanistically meaningful condition-dependent conserved modules.

Biclustering refers to the simultaneous clustering of both genes and conditions (Lazzeroni and Owen 1999; Cheng and Church 2000). Early works (Morgan and Sonquist 1963) introduced the idea of biclustering as "direct clustering" (Hartigan 1972), node deletion problems on graphs (Yannakakis 1981), and biclustering (Mirkin 1996). More recently, biclustering has been used in several studies to address the biologically relevant condition dependence of co-expression patterns (Cheng and Church 2000; Ben-Dor, Chor et al. 2003; Bergmann, Ihmels et al. 2003; Kluger, Basri et al. 2003; Tanay, Sharan et al. 2004; Supper, Strauch et al. 2007; DiMaggio, McAllister et al. 2008; Gan, Liew et al. 2008; Lu, Huggins et al. 2009). Additional genome-wide data (such as association networks and transcription factor binding sites) greatly improves the performance of these approaches (Tanay, Sharan et al. 2004; Elemento and Tavazoie 2005; Reiss, Baliga et al. 2006; Huttenhower, Mutungu et al. 2009). Examples include the most recent version of SAMBA, which incorporates

experimentally validated protein-protein and protein-DNA associations into a Bayesian framework (Tanay, Sharan et al. 2004), and *cMonkey* (Reiss, Baliga et al. 2006), an algorithm we recently introduced.

cMonkey integrates expression and sequence data, metabolic and signaling pathways (Kanehisa, Goto et al. 2002), protein-protein interactions, and comparative genomics networks (Mellor, Yanai et al. 2002; Bowers, Pellegrini et al. 2004; Price, Huang et al. 2005) to estimate condition dependent co-regulated modules. We have previously shown that *cMonkey* can be used to "pre-cluster" genes prior to learning global regulatory networks (Bonneau, Facciotti et al. 2007). Biclusters are iteratively optimized, starting with a random or semi-random seed, via a Monte Carlo Markov chain (MCMC) process. At each iteration, each bicluster's state is updated based upon conditional probability distributions computed using the bicluster's previous state. This enables *cMonkey* to determine the probability that a given gene or condition belongs in the bicluster, *dependent upon* the current state of the bicluster. The components of this conditional probability (one for each of the different data types) are modeled independently as *p*-values based upon individual data likelihoods, which are combined to determine the full conditional probability of a given gene or condition belonging to a given bicluster.

Previous multi-species clustering methods generally fall into two classes (for reviews see (Tirosh, Bilu et al. 2007; Lu, Huggins et al. 2009)). The first class attempts to match conditions between species in order to identify similarities and

differences for a given cell process (McCarroll, Murphy et al. 2004; Khaitovich, Hellmann et al. 2005; Gilad, Oshlack et al. 2006; Tirosh, Weinberger et al. 2006). By requiring matched conditions, this approach is not well suited to large sets of public experiments, as it is limited to only the conditions that have direct analogs for both species. The second class of multi-species clustering methods employs a strategy where the datasets for each organism are reduced to a unit-less measure of co-expression (for example Pearson's correlation) that are then used to compare co-expression patterns in multiple species (Stuart, Segal et al. 2003; Bergmann, Ihmels et al. 2004; Ihmels, Bergmann et al. 2005; Tanay, Regev et al. 2005; Dutilh, Huynen et al. 2006; Tirosh and Barkai 2007). This second class includes methods analyzing the conservation of individual orthologous pairs (Dutilh, Huynen et al. 2006; Tirosh and Barkai 2007) and those seeking to identify larger conserved modules (Stuart, Segal et al. 2003; Bergmann, Ihmels et al. 2004; Tanay, Regev et al. 2005). The common objective is to gain insight into the evolution of related species; including the role of duplication in regulatory network evolution and the occurrence of convergent evolution vs. conserved co-expression (Ihmels, Bergmann et al. 2005; Tirosh and Barkai 2007). However, none of these studies can be considered a true multi-species biclustering algorithm; for example both (Bergmann, Ihmels et al. 2004) and (Tanay, Regev et al. 2005) perform the analyses of the different species sequentially. Furthermore, with the exception of (Tanay, Regev et al. 2005), the methods were limited to considering only expression data.

Below, we present multi-species *cMonkey*, a biclustering framework that enables us to integrate data across multiple species and multiple data-types simultaneously. Our approach maintains the independence of the organism-specific data while still allowing for true biclustering. Specifically, gene membership in multiple clusters is possible and integration of a variety of data types remains an integral part of the approach. Once the conserved modules have been identified, our method further allows the discovery of species-specific modifications (which we term *elaborations,* i.e. the addition of species-specific genes that fit well with the conserved core of the bicluster according to the multi-data score). The ability to find species specific elaborations of conserved co-regulated core sets of genes is a unique strength of the method and is critical to understanding the evolution and function of conserved modules.

Our multi-species biclustering method was applied to two triplets of bacteria, one a Gram-positive triplet and the other a Gram-negative, with the method used to analyze all the possible pairings between the three species of a given triplet. Each triplet consisted of a model organism for the class of bacteria that the triplet represented, as well as two pathogens, where one of the pathogens was closely related species to the model organism, and the second was an outgroup. For example, in the case of the Gram-positive triplet, this triplet contained three closely related species of Firmicutes: *Bacillus subtilis*, *Bacillus anthracis* and *Listeria monocytogenes*. As one of the best-studied bacterial model organisms, *B. subtilis* was selected due to the

wealth of publicly available genomic data and the large amount of knowledge accumulated on this organism over the years. Additionally, *B. subtilis* and *B. anthracis* have similar life cycles, alternating between vegetative cell and dormant spore states (Piggot and Coote 1976; Stragier and Losick 1996; Errington 2003; Waltman, Kacmarczyk et al. 2009; de Hoon, Eichenberger et al. 2010). The third member of the triplet, *L. monocytogenes*, was selected as it shares similar morphology and physiology with *B. subtilis* and *B. anthracis*, but lacks the ability to form spores. In addition, *B. anthracis* and *L. monocytogenes* are pathogenic species, while *B. subtilis* is non-pathogenic. Evolutionarily, the *Bacillus* and *Listeria* genera are estimated to have separated more than one billion years ago (Battistuzzi, Feijao et al. 2004). Analysis of the biclusters obtained as a result of the procedure revealed several gene groups of interest and led us to formulate new hypotheses about the biology of these organisms. Specifically, we were able to detect a temporal difference between the two *Bacillus* species in the expression of a group of metabolic genes involved in spore formation. Furthermore, the unexpected identification of a bicluster for genes required for flagellum formation in *B. anthracis* prompted us to re-examine the capacity for flagellar-based motility in that species.

Similar to the Gram-positive triplet, the Gram-negative triplet contained three [gamma]-proteobacteria *Escherichia coli, Salmonella typhimurium* and *Vibrio cholera*. In this triplet, *S. typhimurium* is the more closely related of the two pathogens to *E. coli*, estimated to have evolutionarily separated within the last 150

million years, while *V. cholera* is estimated to have separated nearly 750 million years ago(Battistuzzi, Feijao et al. 2004).

2.2 Results: Examples of conserved modules detected by the multi-species analysis: Application to the Gram-positive triplet

There are two ways in which we will demonstrate the strengths of our novel method. In the first of these, appearing below, we will provide clear examples of conserved modules that correspond to conserved biological processes. In the second of these two methods, which appears in Chapter 3, we will provide a detailed comparison of several genome-wide metrics that were used to evaluate our method with six others. In some cases, these alternate methods are multi-stage - as is our novel multi-species method – thus there were a total of fifteen (15) methods that we compared. We direct the reader to the methods section (2.5) for a detailed presentation of our multi-species algorithm. In order to reduce complexity, we limit the discussion below to the results from the analyses performed on the Gram-positive triplet, and direct the reader to Appendix 4 for an example that was found in the analyses of the Gram-negative triplet.

To illustrate the strength of our method's ability to identify conserved modules and also to highlight species specific elaboration of these modules, we focus on two processes – endospore formation (sporulation) and flagellum synthesis. In the case of sporulation, both *B. anthracis* and *B. subtilis* can sporulate, while *L. monocytogenes*

cannot (Stragier 2002). Similarly, both *B. subtilis* (Kearns and Losick 2003) and *L. monocytogenes* (Grundling, Burrack et al. 2004) possess flagella and are motile, while *B. anthracis* is a non-motile species (Sterne and Proom 1957).

2.2.1 Biclusters involved in sporulation shared between B. subtilis and B. anthracis:

Sporulation is a cellular differentiation process that *B. subtilis* and *B. anthracis* undergo as a response to resource depletion (Piggot and Coote 1976; Stragier and Losick 1996; Errington 2003; de Hoon, Eichenberger et al. 2010). Sporulating cells divide asymmetrically near one cell pole to produce a smaller cell, the forespore and a larger cell, the mother cell. The forespore will differentiate into a highly resistant dormant cell type called an endospore (hereafter: spore). The mature spore is surrounded by two membranes and a thick proteinaceous layer (the coat). A modified peptidoglycan layer (the cortex) is synthesized in the intermembrane space.

As expected, the multi-species method identified several sporulation modules from the *B. subtilis*-*B. anthracis* pairing and no sporulation modules from the pairings involving *L. monocytogenes*. Here, we focus on three biclusters (32, 82 and 84), whose orthologous cores contained largely non-overlapping sets of genes. Analysis of the gene content indicated that each bicluster was involved in distinct biological functions during sporulation. Bicluster 84 primarily contained genes involved in metabolic functions (**Figures S43 and S44**). Bicluster 32 contained genes involved in activation of late sporulation σ factors (σ^G and σ^K) and cortex synthesis (**Figures S39**

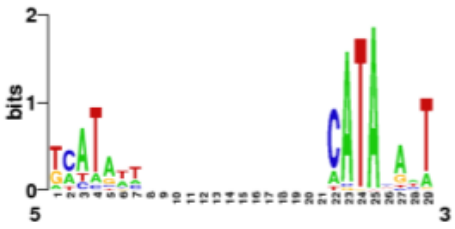
and S40). Bicluster 82 contained a majority of spore coat genes (**Figures S41 and S42**).

Most of the genes found in those three biclusters had been previously identified as members of the mother cell transcriptome in *B. subtilis* (Feucht, Evans et al. 2003; Eichenberger, Fujita et al. 2004; Steil, Serrano et al. 2005). Specifically, 16 of the 26 core genes from the metabolism bicluster, 36 of the 38 core genes from the cortex bicluster and the 24 core genes from the coat bicluster are expressed under the control of the early mother-cell σ factor, σ^E. Nevertheless, the metabolism bicluster contained five previously unrecognized sporulation genes (*ykwC, ctaC, ctaD, ctaE* and *ctaF*). The *ykwC* gene encodes a protein from the 3-hydroxyisobutyrate dehydrogenase family, which is consistent with the function of several other genes found in that bicluster (e.g. the *mmg* and *yngJ* operons (Hsiao, Revelles et al.)). The *cta* operon encodes the four subunits of cytochrome C oxidase. These genes are subject to catabolite repression by glucose, therefore their expression is prevented during exponential growth in glucose-containing medium (Liu and Taber 1998). During sporulation initiation, the *cta* operon is activated by Spo0A~P (the master regulator of sporulation) (Fawcett, Eichenberger et al. 2000). The neighboring *ctaA* gene, which is transcribed in the divergent direction, has been previously reported to be controlled by RNA polymerase containing σ^E (Paul, Zhang et al. 2001). Examination of the *ctaC* upstream region reveals a possible σ^E binding site with a reasonable match to the consensus (Figure 2.1). Protracted expression of these genes

(first under the control of Spo0A~P in the pre-divisional cell and then of σ^E after asymmetric division) is consistent with the conclusions of previous studies indicating that tricarboxylic acid cycle function (and by extension the electron transport chain) is required during sporulation (Ireton, Jin et al. 1995; Jin, Levin et al. 1997; Matsuno, Blais et al. 1999).

Figure 2.1: Putative σ^E binding site in the regulatory upstream sequence of the ctaC operon. Four genes from Bicluster 84, *ctaC ctaD ctaE ctaF*, encode the subunits for cytochrome C oxidase. These genes have not been shown to be regulated by σ^E, however, the region upstream of *ctaC* contains a possible σ^E binding site. **(A)** Consensus binding sequence for σ^E (Eichenberger, Fujita et al. 2004; Sierro, Makita et al. 2008). **(B)** Portion of the intercistronic region from *B. subtilis* between *ctaB* and *ctaC* showing the respective -35 and -10 regions of the potential σ^E binging site (blue**,** bold, uppercase letters). Also indicated are the consensus sequences for the two regions, the putative ribosomal binding site (underlined, uppercase letters), and the translation start site (green, uppercase letters).

A



B

ctaC acatagatg**GCATTTT**gagtaaaaaaact**CATAATTT**aag- 78 bp -acc<u>AGAAAAGGGG</u>ttgggtgaaATG

Consensus: DYMTRWW CATAHAWT

D=A/G/T, H=A/C/T, M=A/C, R=A/G, W=A/T, Y=C/T

Unexpectedly, we uncovered a key species-specific difference in the timing of expression of one conserved sporulation module (the metabolism bicluster). The expression data we used for *B. anthracis* is a time series transcriptional profile of the entire life-cycle, from germination through sporulation (Bergman, Anderson et al. 2006). Expression of genes from the metabolism bicluster reaches its maximal level at t=180 minutes (Figure 2.2a), 2 hours before the expression peak of genes from the cortex and coat biclusters at t=270 minutes. No such temporal difference exists between the metabolism bicluster and the other two biclusters during *B. subtilis* sporulation (Figure 2.2b), because most of these genes are directly controlled by σ^E in *B. subtilis*. We propose that the observed timing difference between the two species is caused by transcriptional re-wiring. In support of this interpretation, examination of the regulatory sequence upstream of the genes from the metabolism bicluster did not reveal obvious σ^E binding sites in *B. anthracis*, while putative σ^E promoters were present upstream of genes from the cortex and coat biclusters in both species. Thus, in *B. anthracis*, the metabolism bicluster may be under the control of a transcription factor active prior to σ^E activation. This is further supported by the fact that in *B. anthracis sigE* itself is expressed after the expression peak of the metabolism bicluster (Ihmels, Bergmann et al. 2005).

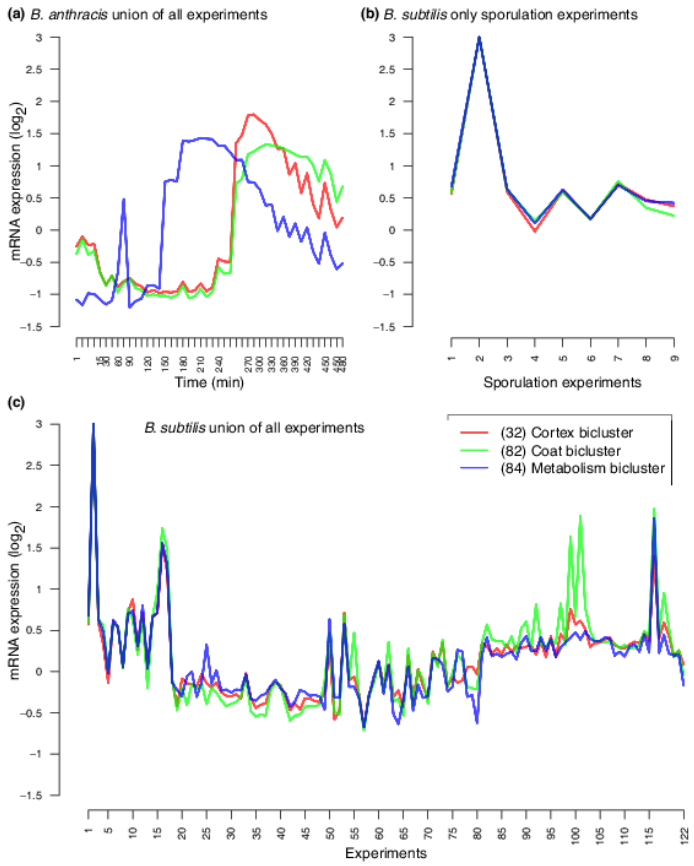


Figure 2.2: Expression profiles of 3 partially conserved sporulation biclusters, identified by the multi-species analysis of *B. subtilis* and *B. anthracis*. Bicluster 84 (blue line) is composed primarily of genes involved in metabolic functions during sporulation, bicluster 82 (green line) includes primarily genes encoding spore coat proteins, and bicluster 32 (red line) contains genes involved in spore cortex formation and activation of the σ factors required for the latest stages of sporulation. **(A)** The *B. anthracis* biclusters display distinct profiles, revealing a temporal aspect not present in the *B. subtilis* dataset. The *B. subtilis* biclusters all follow the same expression profile (i.e. similar expression over nearly every experimental condition included in the dataset), as shown in **(B)** only sporulation experimental conditions (with abscissa corresponding to: (1) Hour 2 *sigF*, (2) Hour 2.5 *sigE*, (3) Hour

3.5 *gerR*, (4) Hour 3.5 *spoIIID*, (5) Hour 4 *sigG*, (6) Hour 4.5 *sigK*, (7) Hour 5 *spoVT*, (8) Hour 5.5 *gerE*, (9) Hour 6.5 *gerE*) and (C) all experimental conditions within the three biclusters.

2.2.2 Flagellar assembly biclusters *shared between* B. subtilis, B. anthracis *and* L. monocytogenes:

Assembly of the bacterial flagellum is a well-known pathway (Figure 2.3A) that has been studied over a wide range of prokaryotes (Macnab 2003; Liu and Ochman 2007; Liu and Ochman 2007). It contains approximately 25 proteins conserved across numerous species, though not all these species are motile (Liu and Ochman 2007). Here we use the expression of flagellar genes as another benchmark of the multi-species method. We expected that multi-species integrative biclustering with any pairing including *B. anthracis* would be unable to recover modules enriched with flagellar genes. Nonetheless, we discovered that flagellar modules were retrieved with all possible pairings (Figure 2.3 and Table 7.13-Table 7.24). Furthermore, recovery was well supported by the *B. anthracis* portion of the multi-data score. This result was unexpected as it was assumed that the loss of motility would be rapidly followed by loss of coordinated expression of flagellar genes.

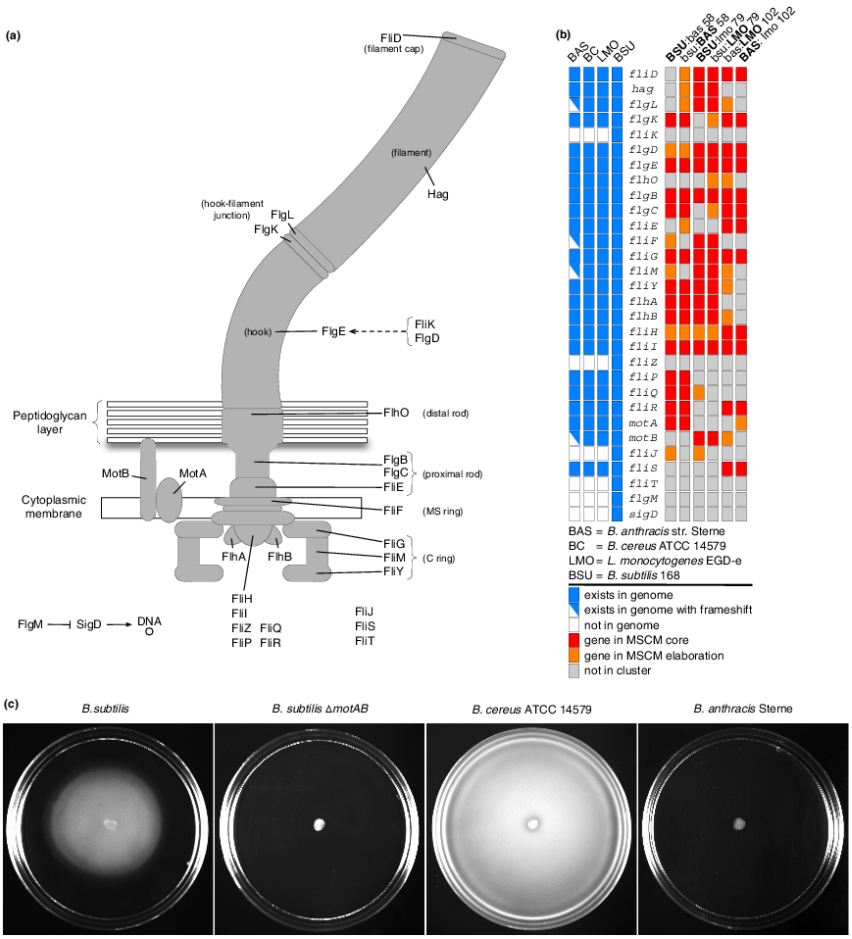


Figure 2.3: Conserved motility modules active in all three organisms and motility assays. (A) We show a schematic of the flagellar apparatus for *B. subtilis* showing the location of 26 flagellar proteins, two motor proteins (MotA and MotB) and two transcriptional regulators (FlgM and SigD) (using gene names from *B. subtilis)*. **(B)** The left panel shows the presence (blue)/absence (white) of the corresponding genes in the genomes of *B. anthracis* Sterne (BAS), *B. cereus* ATCC 14579 (BC), *L. monocytogenes* EGD-e (LMO) and *B. subtilis* 168 (BSU). In *B. anthracis* Sterne, *motB*, *fliM*, *fliF*, and *flgL* are represented by two colors indicating a gene coding for a truncated protein due to a frameshift mutation that introduces a premature stop codon. The right panel shows the gene presence in the main flagellar bicluster resulting from each of the three pairwise multi-species biclusterings. Indicated are

genes of the flagellar apparatus - included in the bicluster core (red), in the elaboration of the bicluster (orange), and not included in the bicluster (gray). *B. subtilis* and *L. monocytogenes* are both known to be flagellated and motile. *B. anthracis* Sterne is non-motile, but our results indicate a bicluster enriched for genes involved in flagellar biosynthesis. **(C)** Swimming motility was assayed on 0.3% agar plates for *B. cereus* ATCC 14579, *B. anthracis* Sterne, *B. subtilis* PY79, and *B. subtilis* PY79 Δ*motAB*::*tet* (strain DS219). *B. cereus* and *B. subtilis* are motile (Kearns and Losick 2003; Salvetti, Ghelardi et al. 2007). A deletion of *motAB* in *B. subtilis* impairs motility (Mirel, Lustre et al. 1992; Blair, Turner et al. 2008). The assay shows that *B. anthracis* Sterne is not motile under the conditions tested.

One simple explanation of the conservation of the *B. anthracis* motility bicluster would be that the strain is, in fact, still motile or able to recover motility through a common reversion mutation. To explore and partially rule out this possibility we confirmed experimentally that *B. anthracis* Sterne was non-motile at 37˚C by performing swimming motility assays on 0.3% agar plates (Figure 2.3c). We used *B. cereus* ATCC 14579 and *B. subtilis* PY79 as positive controls for swimming (Kearns and Losick 2003; Salvetti, Ghelardi et al. 2007) and *B. subtilis* PY79 Δ*motAB*::*tet* as a negative control (Mirel, Lustre et al. 1992). Even after prolonged incubation of those plates at 37˚C for several days, we were unable to observe motile *B. anthracis* cells.

B. anthracis Sterne lacks six flagellar genes present in *B. subtilis* (*fliK, fliO, fliJ, fliT, flgM* and *sigD*) (Kanehisa 2009). Although most of these genes are likely to be essential for flagellum function in *B. subtilis* (Table 2.1), they are absent in several

motile species, including *L. monocytogenes* and *B. cereus*. These genes may in fact be dispensable for motility if a different gene provides a corresponding function. For example, while σ^D and FlgM (the anti-σ^D factor) regulate flagellar gene expression in *B. subtilis*, they are not found in *L. monocytogenes*, where flagellar gene expression is regulated by the transcription factor MogR, which is absent in *B. subtilis* (Grundling, Burrack et al. 2004) (Table 2.2). We performed a BLAST search-based analysis of the presence or absence of flagellar assembly and chemotaxis genes for *L. monocytogenes* and various *Bacillus* species (Table 2.3). Since *B. cereus* is the closest motile relative to *B. anthracis* (Rasko, Ravel et al. 2004), we focused on cases where a flagellar gene was present in *B. cereus* and absent in *B. anthracis*. Specifically, BLAST searches were performed against the genomes of various *B. anthracis* strains using *B. cereus* ATCC 14579 protein sequences as a reference. In *B. anthracis* Sterne two strong hits were retrieved for MotB; each of which covered a different half of the *B. cereus* MotB sequence. Upon closer inspection, it was found that both these coding sequences derived from the same gene which had undergone a frameshift mutation via a one base-pair deletion. The frameshift resulted in an in-frame stop codon shortly following the deletion (Figure 2.4). In *B. subtilis*, *motB* has been shown to be essential for motility ((Mirel, Lustre et al. 1992) and Figure 2.3c).

Table 2.1: *B. subtilis* flagellar assembly genes that are missing in *B. anthracis*, and their associated function. The genes in the table are present in the *B. subtilis* flagellar assembly pathway as indicated by KEGG, but missing in *B. anthracis*.

Gene	Function
flgM	anti-σ^D factor
fliJ	Part of the type III secretion chaperone-usher complex
fliK	hook length regulator
fliO	Part of the Type III secretion apparatus
fliT	Chaperone
sigD	Sigma factor responsible for the expression of motility and chemotaxis genes

Table 2.2: Major Regulators of Motility in B. subtilis and L. monocytogenes

Regulator	Organism	Function	Reference
sigD	*B. subtilis*	Sigma factor responsible for the expression of motility and chemotaxis genes	Marques-Magana and Chamberlin, 1994 (Marquez-Magana and Chamberlin 1994)

mogR	*L. monocytogenes*	Transcriptional repressor of flagellar genes when at temperatures > 37°C	Grundling *et al*., 2004 (Grundling, Burrack et al. 2004)
degU	*L. monocytogenes*	Response regulator which controls temperature-responsive expression of *gmaR*	Shen *et al*., 2006 (Shen and Higgins 2006)
gmaR	*L. monocytogenes*	Antirepressor of MogR	Shen *et al*., 2006 (Shen and Higgins 2006)

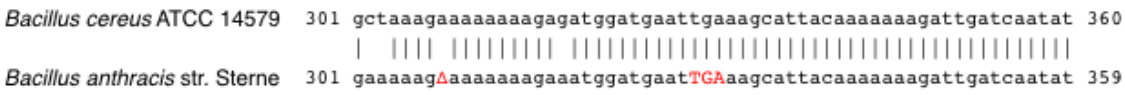


Figure 2.4: *B. anthracis* Sterne frameshift mutations. *B. anthracis* Sterne contains frameshift mutations in five motility genes (motB, flgL, fliF, fliM, and cheV). The frameshift mutation in motB was discovered in B. anthracis Sterne by first performing protein BLAST searches using B. cereus ATCC 14579 MotB as the query sequence. Investigation of the *B. anthracis* Sterne gene sequence revealed a single base deletion (red delta, Δ) that causes a frameshift mutation resulting in a stop codon (red, uppercase letters) thus truncating the MotB protein. The other *B. anthracis* frameshift mutations (flgL, fliF, fliM, cheV) were discovered using the same method.

Table 2.3: Genetic compositions of various *Bacillus* and other closely related species. As shown in the table, the genetic composition of the *B. anthracis* strains Sterne, A2012 and CDC 684 are almost identical to the motile species *B. cereus*, *B. thuringiensis*, *B. weihenstephanensis and L. monocytogenes*. In contrast, the *B. anthracis* strains Ames, Ames 0581 and A0248 are lacking multiple genes present in the other motile organisms.

Genes	*B.subtilis* *B. halodurans* *B. licheniformis* (all)	*B. amyloliquefaciens* *B. pumilus*	*B. clausii*	*B. anthracis* (Sterne, A2012, CDC 684) *B. cereus* (all) *B. thuringiensis* *B. weihenstephanensis* *L. monocytogenes*	*B. anthracis* (Ames, Ames 0581, A0248)
flgL	X	X	X	X	
flgM	X	X	X		
fliF	X	X	X	X	
fliJ	X	X	X		
fliK	X	*			
fliM	X	X	X	X	
fliO	X	X	X		
FliT	X	X			
cheC	X	X	X		
cheD	X	X	X		
cheV	X	X	X	X’	

Chew	X	X	X	X''

X gene is present in the KEGG flagellar assembly pathway

X' gene is not present in B. anthracis Sterne or A2012

X'' gene is present in B. anthracis Sterne and A2012 but not the other organisms in the column

** gene is not present in KEGG but is recognized by NCBI*

We then examined the protein sequences of all the flagellar proteins in *B. anthracis* Sterne by performing multiple alignments with other related *Bacilli* and discovered that three additional proteins appeared truncated: FliM, FliF and FlgL. Investigation of the gene sequences for these proteins in *B. anthracis* Sterne revealed that they all contained a frameshift mutation, which resulted in the introduction of an in-frame stop codon. In *B. subtilis*, *fliM* mutations result in a non-flagellated phenotype (Zuberi, Ying et al. 1990), while *fliF* and *flgL* are essential for flagellar assembly in *L. monocytogenes* (Bigot, Pagniez et al. 2005; Todhanakasem and Young 2008). In addition, we found a similar frameshift in *cheV*, a gene required for chemotaxis in *B. subtilis*.

The presence of the frameshift mutations for these key motility genes most likely explains why *B. anthracis* Sterne is non-motile and does not readily revert back to a motile phenotype. Importantly, this observation indicates that a conserved module can persist for some time even after the loss of the associated phenotype.

2.3 Discussion

Any attempt to detect conserved modules across multiple species data collections needs to simultaneously address the following non-trivial challenges: 1) modules may be active or coherent in subsets of the conditions for each species; 2) in most cases there is little or no correspondence between the experimental conditions and experimental designs across different species datasets; 3) the amount and quality of data available often varies dramatically across species of interest; 4) modules may not be conserved in their regulation or function; 5) conserved modules may have extensive species specific elaborations that complicate their detection; 6) in many cases, the sequence-based orthology is not a one-to-one mapping; and 7) integration of additional data-types needs to be robust to the differences in the available data and annotation completeness of the species considered. In this investigation, we have introduced a new algorithm, multi-species cMonkey (MScM) that allows us to address all of these challenges in a unified analysis. We tested 6 other biclustering and clustering methods in various combinations (13 clustering and biclustering formulations were tested) and found no other method capable of balancing all of these challenges. We have shown that MScM provides better or comparable coverage, functional enrichment scores, bicluster coherence, and conservation than other tested methods, with all other methods failing in one of the main categories of assessment. Furthermore, our method effectively balances the influence of each organism, preventing organisms with more complete datasets from dominating the analysis,

while also integrating other supporting data types, enabling the method to identify more biologically relevant modules and delimit the conditions over which the modules are active. The fact that the MScM biclusters have many fold higher conservation scores than several of the tested methods suggests that they have a higher level of biological significance than equally co-expressed (and/or equally functionally enriched) non-conserved alternate biclusters. An analysis that takes into account several validation metrics supports the idea that MScM is the top performing method for comparative biclustering.

In the single species setting, single-species-*cMonkey* and other biclustering methods, particularly COALESCE, are comparable in performance (when one considers score, enrichment and coverage *but not conservation*). Our analysis suggests that multi-species extensions of other top performing algorithms (particularly COALESCE) will also perform well at detecting conserved modules (assuming that such extensions are possible). For all the organisms pairings, there was a consistent increase in the percentage of GO and KEGG enrichments from the shared to elaboration steps of the MScM method. This results from shared biclusters that contain enrichments that are insignificant until genes from outside of the orthologous core are added during the elaboration step. We argue that this improved functional coherence illustrates the necessity of a species-specific elaboration step in any type of multi-species analysis similar to the one described here. Future work will include

development of methods for adding no-obvious homologs, and perhaps phenologs (McGary, Park et al.), to the comparative phase of our analysis.

2.4 Conclusion

A careful examination of several of the conserved biclusters generated as part of the MScM analysis indicates that our method can reveal important new biology. For instance, we found two cases where conserved biclusters function differently in the species analyzed. The recovery of a flagellar module in the non-motile *B. anthracis* species shows that it is possible to identify conserved modules, even in cases where phenotypic divergence suggests none should exist. In addition, a key temporal difference in the sporulation programs for *B. subtilis* and *B. anthracis* emerged that led us to propose that a rewiring event took place during the evolution of the expression of a group of metabolic genes involved in sporulation. Our biclustering approach also appears useful in generating functional hypotheses for genes that are grouped with other genes of previously established functions, considering that many of the unannotated genes contained in biclusters with GO or KEGG enrichments are well supported across six or more datasets (2 organisms x 3 or more data-types). Our method also reveals new links between functions that were previously considered to be separate, such as the association of the *cta* operon and the *ykwC* gene with several other *B. subtilis* metabolism genes.

2.5 Materials and Methods

Here, we describe the main steps in the multi-species *cMonkey* algorithm, which is implemented in the R programming language and freely available (Waltman, Kuppusamy et al. 2010). We emphasize the novel modifications to the algorithm that allow for identifying biclusters in a multi-species context; for a more detailed description of the individual cMonkey scoring function components see (Reiss, Baliga et al. 2006). Methods used for global assessment and comparison of our methods to other biclustering and clustering methods, experimental validation of results, and code release as well as two simple multi-species clustering methods of our own construction (multi-species k-means and balanced multi-species k-means), are also described. A complete description of the data used for each organism is provided in the supplemental section.

2.5.1 Multi-species cMonkey method overview

Briefly, the MScM algorithm is composed of three steps, with an optional fourth (Figure 2.5). (1) the identification of orthologous genes between closely related species (2) an iterative, *Monte Carlo* optimization within the space of shared orthologs (involving pairs of orthologous genes) (3) an iterative, *Monte Carlo* optimization in the space of each organism's genome that *elaborates* the biclusters found in step 2 by adding non-orthologous genes and (4, optional) an application of the original, single-species method to the remainder of each organism's genome (that was not added to the

conserved biclusters found in steps 2 and 3) to identify completely species-specific biclusters.

2.5.1.1 Algorithm overview:

1. *Identification of orthologous genes*
2. *Identification of shared biclusters by optimizing multi-species cMonkey score (orthologous gene space)*
3. *Single-species Elaboration of shared biclusters from step 2 (single-species full genome space)*
4. *Identification of non-shared biclusters (single species full genome space) (optional)*

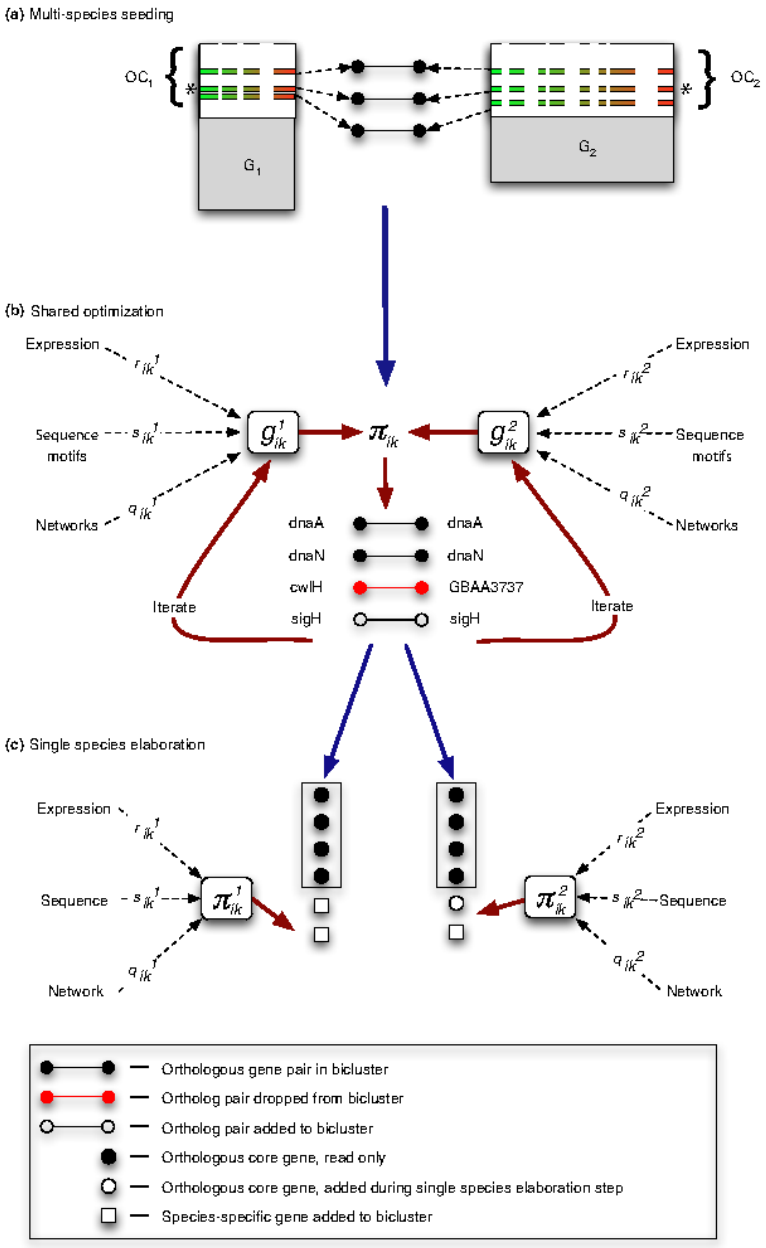


Figure 2.5: Schematic overview of multiple-species method. **(a)** shared-space bicluster seeds are generated by calculating the pairwise correlation of the gene-pairs to a randomly selected gene-pair. **(b)** The shared-space multi-species optimization, where orthologous gene pairs are iteratively added or dropped from the bicluster according to the multi-species multi-data score. **(c)** When completed, shared-

space biclusters are separated into their respective species, and further optimized during the elaboration step. During this step the genes from the original shared-space bicluster are prevented from being dropped, as indicated by the boxes surrounding these genes (represented as black circles).

2.5.1.2 Determining putative orthologs spanning relevant genomes (step 1)

Our analysis requires the identification of putative orthologs between each pair of organisms as input. As identification of ortholog sets between species is not a primary focus of this investigation, we rely on publicly available tools and resources to define our starting set of putative orthologs between two or more species. Dependent upon the organisms used, there may be databases that can provide these ortholog sets, such as the well-annotated list of orthologs from the Mouse Genomics Informatics database (Bult, Eppig et al. 2008). In cases where a pre-existing curated list of orthologs is unavailable, we use the InParanoid algorithm (Remm, Storm et al. 2001) as two recent benchmarks (Hulsen, Huynen et al. 2006; Chen, Mackey et al. 2007) determined it to be among the most accurate when identifying pairwise orthology. InParanoid allows for the identification of families of orthologous and paralogous genes that are shared by 2 genomes, rather just single pair matches (for example, as the *cotZ* gene in *B. subtilis* has two possible orthologs in *B. anthracis*, *cotZ1* and *cotZ2*, both the *cotZ-cotZ1* and *cotZ-cotZ2* pairs will be considered by our algorithm). This feature of the InParanoid algorithm is useful in the context of this work as it allows for more permissive supersets of putative orthology from which we

can sample using cMonkey (thus letting the data select amongst orthologous super-sets).

2.5.1.3 Defining the multi-species data-space

In the first phase of our algorithm, biclustering is performed on groups of orthologous genes (in this study we limit the algorithm to pairs, but the algorithm is easily extendable to larger groups). For any two genomes, $\boldsymbol{G_U}$ and $\boldsymbol{G_V}$, we use $\boldsymbol{OC_U}$ and $\boldsymbol{OC_V}$ to refer to the portions of these genomes with one or more orthologs in the other genome, which we term the '*Orthologous Cores*' of these genomes. Furthermore, we will use $\boldsymbol{OC_{UV}}$ to refer to the list of *all possible* pairings of orthologs between the species, which for convenience we will refer to as '*orthologous pairs*'. In the case of gene families, where genes from one genome have several putative orthologs in the other, we allow the algorithm to separately consider gene pairs for each of the possible pairwise relationships. Thus, if we have a family, *f*, that has 4 members in genome *U*, $OC_U^f = \left\{g_U^1, \ldots g_U^4\right\}$, and 3 in genome *V*, $OC_V^f = \left\{g_V^1, \ldots g_V^3\right\}$, this will result in 12 possible pairs for this family, i.e. $OC_{UV}^f = \left\{g_U^1 g_V^1, g_U^1 g_V^2, \ldots, g_U^4 g_V^2, g_U^4 g_V^3\right\}$.

2.5.1.4 Seeding the biclustering

The first step in building multiple-species biclusters out of ortholog pairs, as defined above, consists of seeding a bicluster (selecting a starting subset of

orthologous pairs to define as the starting bicluster). For example, this can be done via selection of a random subset of orthologous pairs as a 'seed' which is then optimized. For this study we choose a semi-random seeding (Figure 2.5a). We choose a random orthologous gene pair and then 1) define the bicluster seed to be the 70% of conditions in each organism's dataset where the ortholog pair has the highest variance, and 2) add the most correlated 5-10 ortholog pairs (where the correlation is calculated as the average for each gene in the ortholog pair over only the conditions in the bicluster). We refer to this simple procedure for seeding the bicluster optimization as semi-random seeding. The main motivation behind this scheme (described in supplement and prior (Reiss, Baliga et al. 2006)) is to improve the convergence rate by jump-starting the optimization, though MScM can also be used to refine randomly generated seeds.

2.5.1.5 Finding biclusters in the multi-species data-space (step 2)

Given a bicluster seed (semi-random, random or a seed generated via a different method) we begin the multi-species optimization by iteratively adding and dropping genes and conditions as part of a simulated annealing optimization of the multi-species integrative score (Figure 2.5c). Letting $\boldsymbol{X_U}$ and $\boldsymbol{X_V}$ represent the expression datasets for the two genomes considered, a single-species bicluster is defined as a set of genes and a set of conditions in $\boldsymbol{X_U}$ and $\boldsymbol{X_V}$. In the single-species biclustering case, we calculate a combined score for every gene in the genome (given the supporting data) to determine the likelihood of it being added or dropped from the

bicluster. Extending this idea to the multi-species space requires that for every orthologous pair, we can determine the likelihood of that ortholog pair being added or dropped from a given shared-space bicluster. We do this by combining the single-species gene scores (calculated separately for each organism within its independent data space) for the genes in an orthologous pair to compute the multi-species score π_{ik}:

$$\pi_{ik} = p\left(y_{ik} = 1 \mid g_{ik}^{U}, g_{ik}^{V}\right) \propto \exp\left(\beta_0 + \beta_1\left(g_{ik}^{U} + g_{ik}^{V}\right)\right)$$

where g_{ik}^{U} and g_{ik}^{V} are the species-specific likelihoods for the members of pair *i* for bicluster *k*, and β_0 and β_1 are the parameters of the logistic regression. Note, this framework can easily be extended to more than 2 organisms, where the likelihood of the orthologous N-tuples for the ***N*** organisms in would be defined as:

$$\pi_{ik} = p\left(y_{ik}^{1} = 1,\ldots,y_{ik}^{|N|} = 1 \mid g_{ik}^{1},\ldots,g_{ik}^{|N|}\right) \propto exp\left(\beta_0 + \beta_1 \sum_{n \in N} g_{ik}^{n}\right)$$

The parameters in this regression determine a decision boundary between genes in and out of the bicluster (fitted to the combined single-species scores for the pairs in $\boldsymbol{OC_{UV}}$ at the previous iteration). It is important to note that individual data-types from each species are not concatenated or combined through any other lossy or unbalanced transformation. The multiple species integration occurs solely via the computation of this decision boundary at this final step in computing the score. We believe that this imparts significant flexibility to the algorithm that will allow extension to other data-types and larger collections of species in the future within this

framework. For each organism j $\left(j \in \{U,V\}\right)$, g_{ik}^{j} is defined as in the single-species *cMonkey* algorithm, as:

$$g_{ik}^{j} = r_0 \log\left(\tilde{r}_{ik}^{j}\right) + s_0 \log\left(\tilde{s}_{ik}^{j}\right) + \sum_{n \in N} q_0^n \log\left(\tilde{q}_{nik}^{j}\right)$$

where $\tilde{r}_{ik}^{j}$, $\tilde{s}_{ik}^{j}$, and $\tilde{q}_{nik}^{j}$ are the individual likelihoods for the expression, sequence and networks, as defined by our earlier work and r_0, s_0 and q_0 are mixing parameters set to roughly equalize the influence of each data type in this work (these mixing parameters can also be used to increase the influence of single datatypes if desired). For this work these mixing parameters were set such that each data-type would have equal aggregate effect on the biclustering. Each of these individual score components, $\tilde{r}_{ik}^{j}$, $\tilde{s}_{ik}^{j}$, and $\tilde{q}_{nik}^{j}$, are described previously (Reiss, Baliga et al. 2006) and in the supplemental section. The probability that any gene pair in $\boldsymbol{OC_{UV}}$ is added to the growing bicluster is a well-balanced function of the evidence derived from the integrated dataset for each species, formulated as the two multi-data scores, g_{ik}^{1} and g_{ik}^{2}, that represent the individual species support values for each gene in an orthologous pair (g_{ik}^{1} and g_{ik}^{2} represent the multi-data-type integration for each organism separately and π_{ik} effects the multi-species integration). Once this coupled version of the *cMonkey* score is obtained, the algorithm progresses in a manner similar to SSCM, but adding and removing pairs from the bicluster during each iteration instead, and stopping when convergence criteria are met (Reiss, Baliga et al. 2006; Bonneau, Facciotti et al. 2007). At this

stage, the formation of ortholog-pair biclusters, we limit any given bicluster to including only a single pair from any one particular ortholog family. Multiple members of an orthologous core can be included in different biclusters, and additional members of any given family of orthologs can be added in the following species specific elaboration stage.

2.5.1.6 Identification of species-specific elaborations of conserved-core biclusters (step 3)

In this step, we identify species-specific modifications to the biclusters that are discovered during the orthologous-pair biclustering described above (Figure 2.5c). To do this, we decouple the orthologous pairs from the shared-space modules to generate two biclusters, one for each organism, which represent the conserved cores of a putative, conserved, co-regulated module. These effectively serve as 'super-seeds' for this step, which are each separately optimized in a manner similar to the original single-species cMonkey method, but now considering the full genomes of each respective organism (genes without clear orthologs in the other organism can now be added if supported by the integrative score). Unlike the original method, though, in this step, we anchor these searches by preventing the genes from the original shared-space orthologous cores from being dropped. In so doing, we maintain the original putative, conserved module, while allowing the addition of species-specific or non-conserved orthologous core genes that fit well to the bicluster in a species-specific manner. During this stage, we also remove the constraint that only one gene from a given orthologous group can be selected by a given shared bicluster to permit

detection of bona fide co-regulation of multiple members of paralogous gene families (e.g. enabling the potential identification of dosage selection of paralogous genes). Finally, unlike either the shared-space or single-species optimizations previously described, where the mixing parameters, r_0, p_0 and q_0, follow a structured annealing schedule during the optimization, in this optimization step we hold these mixing parameters constant, using the final values from the shared optimization for these.

2.5.1.7 Identification of species-specific biclusters (optional, step 4)

Once the multi-species analysis has been completed, as an optional final step, any species-specific modules that are completely unique to each organism can be identified by running single-species cMonkey on the remaining un-biclustered genes. We direct the reader to the supplementary material for a more detailed description and discussion of this step as it is not a main focus of this first demonstration of our method. These last two species-specific steps provide our method the strength and flexibility to simultaneously identify both conserved, partially conserved and species-specific modules, giving a correct limiting behavior across a wide range of possible species pairings.

2.5.1.8 Algorithm pseudocode:

Figure 2.6: Multi-species cMonkey algorithm pseudocode.

Algorithm 1 MSCM.shared(organisms, orthologs, num.biclust, iter.max)

for $i = 1$ to *num.biclust* **do**
  $bicluster \leftarrow$ seed.bicluster($organisms, orthologs, conditions$)
  $iter \leftarrow 1$
  repeat
    { calculate the shared gain for each ortholog pair }
    for *ortholog.pair* in *orthologs* **do**
      for *org* in *organisms* **do**
        { compute motif likelihods in promoter regions of genes }
        $s \leftarrow$ detect.motifs($orthologs, upstream.sequences$)
        $gain_{shared}[ortholog.pair, org]$ += G($bicluster[org], conditions[org], ortholog.pair[org], org, r, s, q$)
      end for
    end for
    $model \leftarrow$ logit($gain_{shared}, bicluster$)
    { calculate probability drop and add genes }
    for *ortholog.pair* in *orthologs* **do**
      if $ortholog.pair \in bicluster$ **then**
        $prob_{membership}[ortholog.pair] \leftarrow \mathrm{P}^{drop}(gain_{shared}[ortholog.pair], model)$
      else
        $prob_{membership}[ortholog.pair] \leftarrow \mathrm{P}^{add}(gain_{shared}[ortholog.pair], model)$
      end if
    end for
    { calculate probability drop and add conditions in each organism }
    for *condition* in *conditions* **do**
      if $condition \in bicluster$ **then**
        $prob_{membership}[condition] \leftarrow \mathrm{P}^{drop}(gain_{shared}[condition], model)$
      else
        $prob_{membership}[condition] \leftarrow \mathrm{P}^{add}(gain_{shared}[condition], model)$
      end if
    end for
    update *bicluster* based on $prob_{membership}$ sample distribution
    $iter$ += 1
  until convergence **or** $iter == iter.max$
  $bicluster.list[i] \leftarrow bicluster$
end for
return *bicluster.list*

2.5.2 Data set analyzed

In the following sections, we provide for each of the Gram-positive and Gram-negative triplet a detailed description of the data sets that were analyzed. In most cases, the expression data was collected from the GEO omnibus database (Edgar,

Domrachev et al. 2002; Barrett, Troup et al. 2007), though additional *B. subtilis* data also came from the KEGG Expression Database (Goto, Kawashima et al. 2000). In addition to these expression data sets, we also included upstream sequence data (200 bases upstream of the start codon), retrieved from RSA Tools (van Helden 2003) as well as network associations from KEGG (Kanehisa and Goto 2000; Kanehisa, Goto et al. 2002; Kanehisa, Goto et al. 2006; Kanehisa, Araki et al. 2008), Prolinks (Bowers, Pellegrini et al. 2004) and Predictome (Mellor, Yanai et al. 2002). We used InParanoid (Remm, Storm et al. 2001; Alexeyenko, Tamas et al. 2006) to identify putative sets of orthologs between these three species of each triplet.

2.5.2.1 Gram-positive triplet

For *B. subtilis*, we compiled an expression data matrix that consisted of 314 conditions from 15 studies that examine the regulons of over 40 known transcriptional regulators and sigma factors (Kobayashi, Ogura et al. 2001; Ogura, Yamaguchi et al. 2001; Yoshida, Kobayashi et al. 2001; Ogura, Yamaguchi et al. 2002; Asai, Yamaguchi et al. 2003; Doan, Servant et al. 2003; Eichenberger, Jensen et al. 2003; Molle, Nakaura et al. 2003; Tojo, Matsunaga et al. 2003; Watanabe, Hamano et al. 2003; Yoshida, Yamaguchi et al. 2003; Bunai, Ariga et al. 2004; Eichenberger, Fujita et al. 2004; Serizawa, Yamamoto et al. 2004; Yoshida, Ohki et al. 2004; Hayashi, Ohsawa et al. 2005; Hayashi, Kensuke et al. 2006; Wang, Wu et al. 2006). For the two pathogens, the *L. monocytogenes* expression matrix contained 56 conditions that were compiled from 8 studies covering early stationary phase, salt, alkali, and cold shocks

(Marr, Joseph et al. 2006; Hu, Oliver et al. 2007; Hu, Raengpradub et al. 2007; Severino, Dussurget et al. 2007; Bowman, Bittencourt et al. 2008; Giotis, Muthaiyan et al. 2008; Raengpradub, Wiedmann et al. 2008); while the *B. anthracis* matrix contained 51 conditions from a single study by Bergman et al (Bergman, Anderson et al. 2006) covering the full life-cycle of the *B. anthracis* Sterne strain. As mentioned previously, most data was collected from the GEO omnibus database (Edgar, Domrachev et al. 2002; Barrett, Troup et al. 2007), though additional *B. subtilis* data also came from the KEGG Expression Database (Goto, Kawashima et al. 2000). (Kanehisa and Goto 2000; Remm, Storm et al. 2001; Kanehisa, Goto et al. 2002; Mellor, Yanai et al. 2002; Bon, Casaregola et al. 2003; Bowers, Pellegrini et al. 2004; Alexeyenko, Tamas et al. 2006; Kanehisa, Goto et al. 2006; Kanehisa, Araki et al. 2008) To generate the list of orthologous pairs for each pairing, we used InParanoid with the default settings (BLOSSUM45 substitution matrix), to identify 2225 orthologous groups between *B. subtilis* and *B. anthracis*, 1439 between *B. subtilis* and *L. monocytogenes*, and 1494 between *B. anthracis* and *L. monocytogenes*. Note, that while these are the total number of groups, the total number of genes and orthologous pairs is larger as we also include non-best-matching orthologs in our analysis. Table 2.4 and Table 2.5 provide full listings of the number of genes, conditions and edges (by network association) in our database for each organism, as well as the total number of genes, orthologs and ortholog families for each pairing between the Gram-positive triplet.

Table 2.4: Size of the data sets used for the Gram-positive triplet, by organism.

Number of:	*Bacillus subtilis*	*Bacillus anthracis*	*Listeria monocytogenes*
genes	3928	5861	2795
conditions	314	51	56
association edges:			
operon	839	997	494
metabolic (KEGG)	49630	73981	36825
Gene Neighbor (Prolinks)	6105	7338	1982
Phylogenetic Profile (Prolinks)	6036	7703	1970
Gene Cluster (Prolinks)	839	997	494
COG-code	227096	370354	110489

Table 2.5: Total number of orthologs, orthologous families, and ortholog pairs generated by InParanoid for the Gram-positive triplet, by organism pairing.

Number of:	*B. subtilis* – *B. anthracis*	*B. subtilis* – *L. monocytogenes*	*B. anthracis* – *L. monocytogenes*
orthologous groups	2225	1439	1494
orthologous pairs	2443	1564	1690
multi-member groups	118	95	129
Remaining unique genes (per organism)	*B. subtilis*: 2279 *B. anthracis*: 2339	*B. subtilis*: 1519 *L. monocytogenes*: 1478	*B. anthracis*: 1634 *L. monocytogenes*: 1537

2.5.2.2 Gram-negative triplet

The *E. coli* expression data matrix consisted of 507 conditions from 16 projects acquired from the Many Microbe Microarrays Database (M3D) (Faith, Driscoll et al. 2008) covering various conditions including: genetic perturbations, changes in oxygen concentration and pH, growth phases, antibiotic treatment, heat shock, and different media. The *S. typhimurium* expression data matrix consisted of 138 conditions from 8 studies acquired from the Stanford Microarray Database (SMD) (Sherlock, Hernandez-Boussard et al. 2001) covering various conditions including: chemical effects, nutrient limitation, library verification, strain comparison, media comparisons, time course, and mutants (Chan, Baker et al. 2003; Detweiler, Monack et al. 2003; Kim and Falkow 2003; Kim and Falkow 2004; Prouty, Brodsky et al. 2004; Chan, Kim et al. 2005; Lawley, Chan et al. 2006; Halbleib, Saaf et al. 2007). Finally, the *V. cholerae* expression data contained 441 conditions that were also collected from the SMD from 10 studies that explored host response(Merrell, Butler et al. 2002), chitin utilization (Meibom, Li et al. 2004), competence(Meibom, Blokesch et al. 2005; Blokesch and Schoolnik 2007), mucosal escape response (Nielsen, Dolganov et al. 2006), metabolism (Shi, Romero et al. 2006), comparisons with non-pathogenic strains (Keymer, Miller et al. 2007; Miller, Keymer et al. 2007), pigment (Valeru, Rompikuntal et al. 2009) and virulence (Nielsen, Dolganov et al. 2010). Table 2.6 and

Table 2.7 provide full listings of the number of genes, conditions and edges (by network association) in our database for each organism, as well as the total number of

genes, orthologs and ortholog families for each pairing between the Gram-negative triplet.

Table 2.6: Size of the data sets used for the Gram-negative triplet, by organism.

Number of:	*Escherichia coli*	*Salmonella typimurium*	*Vibrio cholerae*
genes	4264	3745	3335
conditions	507	138	441
association edges:			
operon	3414	2104	1920
metabolic (KEGG)	96931	75363	106530
Gene Neighbor (Prolinks)	29228	29942	19996
Phylogenetic Profile (Prolinks)	20058	20094	17460
Gene Cluster (Prolinks)	6048	6476	1920
COG-code	644856	379484	525328

Table 2.7: Total number of orthologs, orthologous families, and ortholog pairs generated by InParanoid for the Gram-negative triplet, listed by organism pairing.

Number of:	*E. coli – S. typhimurium*	*E. coli – V. cholera*	*S. typhimurium – V. cholerae*
orthologous groups	2827	1834	1700
orthologous pairs	2856	1965	1831
multi-member groups	22	86	77
Remaining unique genes	*E. coli:* 1428	*E. coli:* 1961	*S. typhimurium:* 1972

(per organism)	*S. typhimurium:* 900	*V. cholera:* 1467	*V. cholera:* 1594

2.5.3 External tools used

Ortholog analysis and identification was performed using InParanoid version 2.0 on protein sequences in Fasta format that were retrieved from NCBI Bacterial Genomes (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/), and using BLAST version 2.2.10. During the cMonkey optimizations, MEME & MAST version 3.5.7 was used as part of the iterative search for new motifs. Upstream sequences were retrieved from Regulatory Sequence Analysis Tools (RSAT) (van Helden 2003; Thomas-Chollier, Sand et al. 2008). All GO term enrichments were calculated using the GO-TermFinder library (Sherlock 2009), using a Bonferonni false discovery correction. All KEGG pathway enrichments were calculated using a utility built in-house for this purpose; also Bonferonni corrected.

2.5.4 Visualization and exploration of multi species biclusters

The Comparative Microbial Module Resource (Kacmarczyk and Bonneau 2010) is an integrated collection of diverse functional genomics datasets and software tools that facilitate the visualization and analysis of conserved cMonkey biclusters, or putatively co-regulated gene modules, across species. The interface allows the visualization and exploration of a bicluster's properties (such as, coupled multi-species biclusters, conserved orthologous core gene members, species-specific gene members,

experimental conditions, gene co-expression pattern, sequence motif logos, and significant functional annotations). Integration with the Gaggle allows access to additional biological information from online databases and further analysis (e.g. integrated tools include but are not limited to: the FireGoose plugin, cytoscape, the Data Matrix Viewer, and an R goose for using the R language and environment for statistical computing and graphics). A comprehensive description of the CMMR is available in Appendix 3 (section 5).

2.5.5 Multi-species cMonkey code release, maintenance and documentation:

Both the multiple-species cMonkey and a re-factored single-species cMonkey are freely available for download and use (Waltman, Kuppusamy et al. 2010). This website includes functionality for bug tracking, tutorials on use, example datasets and runs of the algorithm, links to required packages, and python code developed to aid in data-import. MS-cMonkey is written in R (R Development Core Team 2009) with a data-import module written in Python and has three main modules:

1. Reader: cMonkey is given gene expression matrices and ortholog pairs, along with optional protein association networks and upstream sequences. The user may request cMonkey to automatically find the required and optional datatypes for each organism.
2. The main code: written in R, contains bicluster seeding, bicluster overall optimization, scoring functions, and methods for output and visualization of results.

3. Validation and visualization codes: codes that implement the bicluster and biclustering assessment described, code to facilitate connection to network and cluster visualization tools such as the Gaggle.

All code (cMonkey, the reader, and validation code) are freely available. We have attempted to make several of the steps required for assembling and integrated dataset automatic in this code release, in the hope that this will extend the usefulness of the algorithm to a greater number of biologists. The biologist needs to only prepare simple gene expression matrices and pairs of orthologs. The rest of the datatypes will be queried from biological databases (networks, sequences, annotations for validation scripts, etc.). All input and output will also be stored in a portable, standard relational database that will readily permit use of the integrated dataset and cMonkey results by other tools. These key changes to how data is imported and stored in cMonkey's database and the core data-object for cMonkey allow for multi-species integration. The biologist may use the Reader in two modes: automatic or manual. In automatic mode, the biologist need prepare only gene expression matrices and pairs of orthologs, while protein association networks and upstream sequences are queried from biological databases such as BioNetBuilder (Avila-Campillo, Drew et al. 2007), MicrobesOnline (Dehal, Joachimiak et al. 2009), Prolinks (Bowers, Pellegrini et al. 2004), STRING (Snel, Lehmann et al. 2000; Jensen, Kuhn et al. 2009) and RSAT (van Helden 2003; Thomas-Chollier, Sand et al. 2008).

2.5.6 Swimming motility assays

Individual colonies of *B. subtilis* PY79 (Youngman, Perkins et al. 1984) and DS219 (Blair, Turner et al. 2008), *Bacillus cereus* ATCC 14579 (obtained from Daniel Ziegler, Bacillus Genetic Stock Center, Ohio State University) and *Bacillus anthracis* Sterne (a gift from Adam Driks, Loyola University Chicago) were picked with a wooden stick and inoculated into Luria-Bertaini (LB) 10 g tryptone, 5 g yeast extract, 5 g NaCL per L broth. Cultures were grown to log phase and 3 μl of the broth culture was centrally inoculated on LB Agar plates containing 0.3% Agar. Motility was scored after ~20 hour incubation at 30˚C. Plates were photographed against a dark background such that areas of bacterial colonization appear light.

2.6 Abbreviations Used

OC: Orthologous Core (the set of actively expressed orthologous genes shared between a group of organisms on which we run our multi-species biclustering)

MS: Multiple-species

SS: Single species

SSCM: Single-species *cMonkey*

MScM: Multi-species *cMonkey*

MSISA: Multi-species ISA

MSKM: Multi-species K-Means

BMSKM: *Balanced* Multi-species K-Means

EO: Expression Only

FD: Full Data (EO and FD are used to distinguish between expression only tests and full data runs of integrative methods)

SH: shared biclusters, biclusters generated only from orthologous pairs (MScM, MSKM, BMSKM)

EL: elaborated biclusters, multi-species biclusters that have additional genes unique to each organism added (MScM, MSKM, BMSKM)

P: purified biclusters, applies only to the ISA algorithm (MSISA-P)

R: refined biclusters, applies only to the ISA algorithm (MSISA-R)

2.7 Author's Contributions

PHW implemented and tested the method, and prepared the manuscript. TJK collated the datasets analyzed, generated the visualizations of the results, and aided in validation of results. ARB did extensive studies of the sporulation and flagellar biclusters, performed the motility assays and helped in the writing of the manuscript. DJR helped conceive of the project, and provided initial guidance. DBK provided guidance on the biological interpretation and analysis of results and provided strains for motility assays. PE oversaw all biological aspects of the project, contributed to the validation and visualization of the results, and aided in the writing of the manuscript. RB conceived and oversaw all aspects of the project, and aided in the writing of the manuscript.

2.8 Acknowledgements

The authors would like to thank Kris Gunsalus and the NBrowse team, and Nitin Baliga and the Gaggle team for helpful discussions, Daniel Ziegler from the Bacillus Genetic Stock Center and Adam Driks for strains. We would like to thank Trishank Kuppusamy for his work on methods for automatically importing data into the multiple-species cMonkey framework. We would like to thank Keith Keller of the Lawrence Berkeley National Lab Microbes Online team for work to enable automatic import of large numbers of microbial data-sets. This work was in part supported by NSF DBI-0820757, the Roadmap for Medical Research PN2 EY016586, and an NIH Physical Sciences Oncology Center (U54 CA143907). We acknowledge the financial support of NIH grant GM081571 to PE, GM092616 to RB and PE and Department of the Army award number W81XWH-04-1-0307 to RB and PE. The content of this material does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

2.9 References

Alexeyenko, A., I. Tamas, et al. (2006). "Automatic clustering of orthologs and inparalogs shared by multiple proteomes." Bioinformatics **22**(14): e9-15.

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proc Natl Acad Sci U S A **97**(18): 10101-10106.

Alter, O., P. O. Brown, et al. (2003). "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms." Proc Natl Acad Sci U S A **100**(6): 3351-3356.

Asai, K., H. Yamaguchi, et al. (2003). "DNA microarray analysis of Bacillus subtilis sigma factors of extracytoplasmic function family." FEMS Microbiol Lett **220**(1): 155-160.

Avila-Campillo, I., K. Drew, et al. (2007). "BioNetBuilder: automatic integration of biological networks." Bioinformatics **23**(3): 392-393.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res **35**(Database issue): D760-765.

Battistuzzi, F. U., A. Feijao, et al. (2004). "A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land." BMC Evol Biol **4**: 44.

Ben-Dor, A., B. Chor, et al. (2003). "Discovering local structure in gene expression data: the order-preserving submatrix problem." J Comput Biol **10**(3-4): 373-384.

Bergman, N. H., E. C. Anderson, et al. (2006). "Transcriptional profiling of the Bacillus anthracis life cycle in vitro and an implied model for regulation of spore formation." J Bacteriol **188**(17): 6092-6100.

Bergmann, S., J. Ihmels, et al. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." Phys Rev E Stat Nonlin Soft Matter Phys **67**(3 Pt 1): 031902.

Bergmann, S., J. Ihmels, et al. (2004). "Similarities and differences in genome-wide expression data of six organisms." PLoS Biol **2**(1): E9.

Bigot, A., H. Pagniez, et al. (2005). "Role of FliF and FliI of Listeria monocytogenes in Flagellar Assembly and Pathogenicity." Infect. Immun. **73**(9): 5530-5539.

Blair, K. M., L. Turner, et al. (2008). "A molecular clutch disables flagella in the Bacillus subtilis biofilm." Science **320**(5883): 1636-1638.

Blokesch, M. and G. K. Schoolnik (2007). "Serogroup conversion of Vibrio cholerae in aquatic reservoirs." PLoS Pathog **3**(6): e81.

Bon, E., S. Casaregola, et al. (2003). "Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns." Nucleic Acids Res **31**(4): 1121-1135.

Bonneau, R., M. T. Facciotti, et al. (2007). "A predictive model for transcriptional control of physiology in a free living cell." Cell **131**(7): 1354-1365.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biol **5**(5): R35.

Bowman, J. P., C. R. Bittencourt, et al. (2008). "Differential gene expression of Listeria monocytogenes during high hydrostatic pressure processing." Microbiology **154**(Pt 2): 462-475.

Bult, C. J., J. T. Eppig, et al. (2008). "The Mouse Genome Database (MGD): mouse biology and model systems." Nucleic Acids Res **36**(Database issue): D724-728.

Bunai, K., M. Ariga, et al. (2004). "Profiling and comprehensive expression analysis of ABC transporter solute-binding proteins of Bacillus subtilis membrane based on a proteomic approach." Electrophoresis **25**(1): 141-155.

Chan, K., S. Baker, et al. (2003). "Genomic Comparison of Salmonella enterica Serovars and Salmonella bongori by Use of an S. enterica Serovar Typhimurium DNA Microarray." J. Bacteriol. **185**(2): 553-563.

Chan, K., C. C. Kim, et al. (2005). "Microarray-Based Detection of Salmonella enterica Serovar Typhimurium Transposon Mutants That Cannot Survive in Macrophages and Mice." Infect. Immun. **73**(9): 5438-5449.

Chen, F., A. J. Mackey, et al. (2007). "Assessing performance of orthology detection strategies applied to eukaryotic genomes." PLoS ONE **2**(4): e383.

Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.

de Hoon, M. J., P. Eichenberger, et al. (2010). "Hierarchical evolution of the bacterial sporulation network." Curr Biol **20**(17): R735-745.

Dehal, P. S., M. P. Joachimiak, et al. (2009). "MicrobesOnline: an integrated portal for comparative and functional genomics." Nucl. Acids Res.: gkp919.

Detweiler, C. S., D. M. Monack, et al. (2003). "virK, somA and rcsC are important for systemic Salmonella enterica serovar Typhimurium infection and cationic peptide resistance." Mol Microbiol **48**(2): 385-400.

DiMaggio, P. A., Jr., S. R. McAllister, et al. (2008). "Biclustering via optimal re-ordering of data matrices in systems biology: rigorous methods and comparative studies." BMC Bioinformatics **9**: 458.

Doan, T., P. Servant, et al. (2003). "The Bacillus subtilis ywkA gene encodes a malic enzyme and its transcription is activated by the YufL/YufM two-component system in response to malate." Microbiology **149**(Pt 9): 2331-2343.

Dutilh, B. E., M. A. Huynen, et al. (2006). "A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation." BMC Genomics **7**: 10.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

Eichenberger, P., M. Fujita, et al. (2004). "The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis." PLoS Biol **2**(10): e328.

Eichenberger, P., S. T. Jensen, et al. (2003). "The sigmaE regulon and the identification of additional sporulation genes in Bacillus subtilis." J Mol Biol **327**(5): 945-972.

Elemento, O. and S. Tavazoie (2005). "Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach." Genome Biol **6**(2): R18.

Errington, J. (2003). "Regulation of endospore formation in Bacillus subtilis." Nat Rev Microbiol **1**(2): 117-126.

Faith, J. J., M. E. Driscoll, et al. (2008). "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata." Nucleic Acids Res **36**(Database issue): D866-870.

Fawcett, P., P. Eichenberger, et al. (2000). "The transcriptional profile of early to middle sporulation in Bacillus subtilis." Proc Natl Acad Sci U S A **97**(14): 8063-8068.

Feucht, A., L. Evans, et al. (2003). "Identification of sporulation genes by genome-wide analysis of the sigmaE regulon of Bacillus subtilis." Microbiology **149**(Pt 10): 3023-3034.

Gan, X., A. W. Liew, et al. (2008). "Discovering biclusters in gene expression data based on high-dimensional linear geometries." BMC Bioinformatics **9**: 209.

Gilad, Y., A. Oshlack, et al. (2006). "Expression profiling in primates reveals a rapid evolution of human transcription factors." Nature **440**(7081): 242-245.

Giotis, E. S., A. Muthaiyan, et al. (2008). "Genomic and proteomic analysis of the Alkali-Tolerance Response (AlTR) in Listeria monocytogenes 10403S." BMC Microbiol **8**: 102.

Golub, G. and W. Kahan (1965). "Calculating the Singular Values and Pseudo-Inverse of a Matrix." Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis **2**(2): 205-224.

Goto, S., S. Kawashima, et al. (2000). "KEGG/EXPRESSION: A Database for Browsing and Analysing Microarray Expression Data."

Grundling, A., L. S. Burrack, et al. (2004). "Listeria monocytogenes regulates flagellar motility gene expression through MogR, a transcriptional repressor required for virulence." Proc Natl Acad Sci U S A **101**(33): 12318-12323.

Halbleib, J. M., A. M. Saaf, et al. (2007). "Transcriptional modulation of genes encoding structural characteristics of differentiating enterocytes during development of a polarized epithelium in vitro." Mol Biol Cell **18**(11): 4261-4278.

Hartigan, J. A. (1972). "Direct Clustering of a Data Matrix." Journal of the American Statistical Association **67**(337): 123-129.

Hayashi, K., T. Kensuke, et al. (2006). "Bacillus subtilis RghR (YvaN) represses rapG and rapH, which encode inhibitors of expression of the srfA operon." Mol Microbiol **59**(6): 1714-1729.

Hayashi, K., T. Ohsawa, et al. (2005). "The H2O2 stress-responsive regulator PerR positively regulates srfA expression in Bacillus subtilis." J Bacteriol **187**(19): 6659-6667.

Hsiao, T. L., O. Revelles, et al. "Automatic policing of biochemical annotations using genomic correlations." Nat Chem Biol **6**(1): 34-40.

Hu, Y., H. F. Oliver, et al. (2007). "Transcriptomic and phenotypic analyses suggest a network between the transcriptional regulators HrcA and sigmaB in Listeria monocytogenes." Appl Environ Microbiol **73**(24): 7981-7991.

Hu, Y., S. Raengpradub, et al. (2007). "Phenotypic and transcriptomic analyses demonstrate interactions between the transcriptional regulators CtsR and Sigma B in Listeria monocytogenes." Appl Environ Microbiol **73**(24): 7967-7980.

Hulsen, T., M. A. Huynen, et al. (2006). "Benchmarking ortholog identification methods using functional genomics data." Genome Biol **7**(4): R31.

Huttenhower, C., K. T. Mutungu, et al. (2009). "Detailing regulatory networks through large scale data integration." Bioinformatics **25**(24): 3267-3274.

Ihmels, J., S. Bergmann, et al. (2005). "Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program." PLoS Genet **1**(3): e39.

Ireton, K., S. Jin, et al. (1995). "Krebs cycle function is required for activation of the Spo0A transcription factor in Bacillus subtilis." Proc Natl Acad Sci U S A **92**(7): 2845-2849.

Jensen, L. J., M. Kuhn, et al. (2009). "STRING 8--a global view on proteins and their functional interactions in 630 organisms." Nucl. Acids Res. **37**(suppl_1): D412-416.

Jin, S., P. A. Levin, et al. (1997). "Deletion of the Bacillus subtilis isocitrate dehydrogenase gene causes a block at stage I of sporulation." J Bacteriol **179**(15): 4725-4732.

Kacmarczyk, T. and R. Bonneau. (2010). "Comparative Microbial Module Resource." from http://meatwad.bio.nyu.edu/.

Kanehisa, M. (2009). "Bacterial chemotaxis - Bacillus anthracis Sterne." Retrieved September 24, 2009, from http://www.genome.jp/dbget-bin/show_pathway?bat02030.

Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-484.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Res **34**(Database issue): D354-357.

Kanehisa, M., S. Goto, et al. (2002). "The KEGG databases at GenomeNet." Nucleic Acids Res **30**(1): 42-46.

Kearns, D. B. and R. Losick (2003). "Swarming motility in undomesticated Bacillus subtilis." Mol Microbiol **49**(3): 581-590.

Keymer, D. P., M. C. Miller, et al. (2007). "Genomic and phenotypic diversity of coastal Vibrio cholerae strains is linked to environmental factors." Appl Environ Microbiol **73**(11): 3705-3714.

Khaitovich, P., I. Hellmann, et al. (2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees." Science **309**(5742): 1850-1854.

Kim, C. C. and S. Falkow (2003). "Significance analysis of lexical bias in microarray data." BMC Bioinformatics **4**: 12.

Kim, C. C. and S. Falkow (2004). "Delineation of upstream signaling events in the salmonella pathogenicity island 2 transcriptional activation pathway." J Bacteriol **186**(14): 4694-4704.

Kluger, Y., R. Basri, et al. (2003). "Spectral biclustering of microarray data: coclustering genes and conditions." Genome Res **13**(4): 703-716.

Kobayashi, K., M. Ogura, et al. (2001). "Comprehensive DNA microarray analysis of Bacillus subtilis two-component regulatory systems." J Bacteriol **183**(24): 7365-7370.

Lawley, T. D., K. Chan, et al. (2006). "Genome-wide screen for Salmonella genes required for long-term systemic infection of the mouse." PLoS Pathog **2**(2): e11.

Lazzeroni, L. and A. Owen (1999). Plaid models for gene expression data.

Liu, R. and H. Ochman (2007). "Origins of flagellar gene operons and secondary flagellar systems." J Bacteriol **189**(19): 7098-7104.

Liu, R. and H. Ochman (2007). "Stepwise formation of the bacterial flagellar system." Proc Natl Acad Sci U S A **104**(17): 7116-7121.

Liu, X. and H. W. Taber (1998). "Catabolite regulation of the Bacillus subtilis ctaBCDEF gene cluster." J Bacteriol **180**(23): 6154-6163.

Lu, Y., P. Huggins, et al. (2009). "Cross species analysis of microarray expression data." Bioinformatics **25**(12): 1476-1483.

Macnab, R. M. (2003). "How bacteria assemble flagella." Annu Rev Microbiol **57**: 77-100.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

Marquez-Magana, L. M. and M. J. Chamberlin (1994). "Characterization of the sigD transcription unit of Bacillus subtilis." J Bacteriol **176**(8): 2427-2434.

Marr, A. K., B. Joseph, et al. (2006). "Overexpression of PrfA leads to growth inhibition of Listeria monocytogenes in glucose-containing culture media by interfering with glucose uptake." J Bacteriol **188**(11): 3887-3901.

Matsuno, K., T. Blais, et al. (1999). "Metabolic imbalance and sporulation in an isocitrate dehydrogenase mutant of Bacillus subtilis." J Bacteriol **181**(11): 3382-3391.

McCarroll, S. A., C. T. Murphy, et al. (2004). "Comparing genomic expression patterns across species identifies shared transcriptional profile in aging." Nat Genet **36**(2): 197-204.

McGary, K. L., T. J. Park, et al. (2010). "Systematic discovery of nonobvious human disease models through orthologous phenotypes." Proc Natl Acad Sci U S A **107**(14): 6544-6549.

McQuitty, L. L. (1966). "Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data." Educational and Psychological Measurement **26**(4): 825-831.

Meibom, K. L., M. Blokesch, et al. (2005). "Chitin induces natural competence in Vibrio cholerae." Science **310**(5755): 1824-1827.

Meibom, K. L., X. B. Li, et al. (2004). "The Vibrio cholerae chitin utilization program." Proc Natl Acad Sci U S A **101**(8): 2524-2529.

Mellor, J. C., I. Yanai, et al. (2002). "Predictome: a database of putative functional links between proteins." Nucleic Acids Res **30**(1): 306-309.

Merrell, D. S., S. M. Butler, et al. (2002). "Host-induced epidemic spread of the cholera bacterium." Nature **417**(6889): 642-645.

Miller, M. C., D. P. Keymer, et al. (2007). "Detection and transformation of genome segments that differ within a coastal population of Vibrio cholerae strains." Appl Environ Microbiol **73**(11): 3695-3704.

Mirel, D. B., V. M. Lustre, et al. (1992). "An operon of Bacillus subtilis motility genes transcribed by the sigma D form of RNA polymerase." J Bacteriol **174**(13): 4197-4204.

Mirkin, B. G. (1996). Mathematical classification and clustering. Dordrecht ; Boston, Kluwer Academic Publishers.

Molle, V., Y. Nakaura, et al. (2003). "Additional targets of the Bacillus subtilis global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis." J Bacteriol **185**(6): 1911-1922.

Morgan, J. N. and J. A. Sonquist (1963). "Problems in the analysis of survey data, and a proposal." Journal of the American Statistical Association(58): 415-434.

Nielsen, A. T., N. A. Dolganov, et al. (2006). "RpoS controls the Vibrio cholerae mucosal escape response." PLoS Pathog **2**(10): e109.

Nielsen, A. T., N. A. Dolganov, et al. (2010). "A bistable switch and anatomical site control Vibrio cholerae virulence gene expression in the intestine." PLoS Pathog **6**(9).

Ogura, M., H. Yamaguchi, et al. (2002). "Whole-genome analysis of genes regulated by the Bacillus subtilis competence transcription factor ComK." J Bacteriol **184**(9): 2344-2351.

Ogura, M., H. Yamaguchi, et al. (2001). "DNA microarray analysis of Bacillus subtilis DegU, ComA and PhoP regulons: an approach to comprehensive analysis of B.subtilis two-component regulatory systems." Nucleic Acids Res **29**(18): 3804-3813.

Paul, S., X. Zhang, et al. (2001). "Two ResD-controlled promoters regulate ctaA expression in Bacillus subtilis." J Bacteriol **183**(10): 3237-3246.

Piggot, P. J. and J. G. Coote (1976). "Genetic aspects of bacterial endospore formation." Bacteriol Rev **40**(4): 908-962.

Price, M. N., K. H. Huang, et al. (2005). "A novel method for accurate operon predictions in all sequenced prokaryotes." Nucleic Acids Res **33**(3): 880-892.

Prouty, A. M., I. E. Brodsky, et al. (2004). "Bile-salt-mediated induction of antimicrobial and bile resistance in Salmonella typhimurium." Microbiology **150**(Pt 4): 775-783.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Raengpradub, S., M. Wiedmann, et al. (2008). "Comparative analysis of the sigma B-dependent stress responses in Listeria monocytogenes and Listeria innocua strains exposed to selected stress conditions." Appl Environ Microbiol **74**(1): 158-171.

Rasko, D. A., J. Ravel, et al. (2004). "The genome sequence of Bacillus cereus ATCC 10987 reveals metabolic adaptations and a large plasmid related to Bacillus anthracis pXO1." Nucleic Acids Res **32**(3): 977-988.

Reiss, D. J., N. S. Baliga, et al. (2006). "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." BMC Bioinformatics **7**: 280.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-1052.

Salvetti, S., E. Ghelardi, et al. (2007). "FlhF, a signal recognition particle-like GTPase, is involved in the regulation of flagellar arrangement, motility behaviour and protein secretion in Bacillus cereus." Microbiology **153**(8): 2541-2552.

Serizawa, M., H. Yamamoto, et al. (2004). "Systematic analysis of SigD-regulated genes in Bacillus subtilis by DNA microarray and Northern blotting analyses." Gene **329**: 125-136.

Severino, P., O. Dussurget, et al. (2007). "Comparative transcriptome analysis of Listeria monocytogenes strains of the two major lineages reveals differences in virulence, cell wall, and stress response." Appl Environ Microbiol **73**(19): 6078-6088.

Shen, A. and D. E. Higgins (2006). "The MogR transcriptional repressor regulates nonhierarchal expression of flagellar motility genes and virulence in Listeria monocytogenes." PLoS Pathog **2**(4): e30.

Sherlock, G. (2009). "GO-TermFinder." from http://search.cpan.org/dist/GO-TermFinder/.

Sherlock, G., T. Hernandez-Boussard, et al. (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-155.

Shi, J., P. R. Romero, et al. (2006). "Evidence supporting predicted metabolic pathways for Vibrio cholerae: gene expression data and clinical tests." Nucleic Acids Res **34**(8): 2438-2444.

Sierro, N., Y. Makita, et al. (2008). "DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information." Nucleic Acids Res **36**(Database issue): D93-96.

Snel, B., G. Lehmann, et al. (2000). "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." Nucl. Acids Res. **28**(18): 3442-3444.

Steil, L., M. Serrano, et al. (2005). "Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of Bacillus subtilis." Microbiology **151**(Pt 2): 399-420.

Sterne, M. and H. Proom (1957). "Induction of motility and capsulation in Bacillus anthracis." J Bacteriol **74**(4): 541-542.

Stragier, P. (2002). A Gene Odyssey: Exploring the Genomes of Endospore-Forming Bacteria. Bacillus subtilis and its closest relatives : from genes to cells. A. L. Sonenshein, J. A. Hoch and R. Losick. Washington, D.C., ASM Press**:** pp. 519-525.

Stragier, P. and R. Losick (1996). "Molecular genetics of sporulation in Bacillus subtilis." Annu Rev Genet **30**: 297-241.

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-255.

Supper, J., M. Strauch, et al. (2007). "EDISA: extracting biclusters from multiple time-series of gene expression profiles." BMC Bioinformatics **8**: 334.

Tanay, A., A. Regev, et al. (2005). "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast." Proc Natl Acad Sci U S A **102**(20): 7203-7208.

Tanay, A., R. Sharan, et al. (2004). "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proc Natl Acad Sci U S A **101**(9): 2981-2986.

Thomas-Chollier, M., O. Sand, et al. (2008). "RSAT: regulatory sequence analysis tools." Nucl. Acids Res. **36**(suppl_2): W119-127.

Thomas-Chollier, M., O. Sand, et al. (2008). "RSAT: regulatory sequence analysis tools." Nucleic Acids Res **36**(Web Server issue): W119-127.

Tirosh, I. and N. Barkai (2007). "Comparative analysis indicates regulatory neofunctionalization of yeast duplicates." Genome Biol **8**(4): R50.

Tirosh, I., Y. Bilu, et al. (2007). "Comparative biology: beyond sequence analysis." Curr Opin Biotechnol **18**(4): 371-377.

Tirosh, I., A. Weinberger, et al. (2006). "A genetic signature of interspecies variations in gene expression." Nat Genet **38**(7): 830-834.

Todhanakasem, T. and G. M. Young (2008). "Loss of flagellum-based motility by Listeria monocytogenes results in formation of hyperbiofilms." J Bacteriol **190**(17): 6030-6034.

Tojo, S., M. Matsunaga, et al. (2003). "Organization and expression of the Bacillus subtilis sigY operon." J Biochem **134**(6): 935-946.

Valeru, S. P., P. K. Rompikuntal, et al. (2009). "Role of melanin pigment in expression of Vibrio cholerae virulence factors." Infect Immun **77**(3): 935-942.

van Helden, J. (2003). "Regulatory sequence analysis tools." Nucleic Acids Res **31**(13): 3593-3596.

Waltman, P., T. Kacmarczyk, et al. (2010). "Multi-species integrative biclustering." Genome Biololgy **11**(R96).

Waltman, P., T. Kacmarczyk, et al. (2009). Prokaryotic Systems Biology. Plant systems biology, Annual plant reviews. G. Coruzzi and R. A. Gutierrez. Ames, Iowa, Blackwell Pub.**:** 67-136.

Waltman, P., T. K. Kuppusamy, et al. (2010). "cMonkey2." from http://ms2.bio.nyu.edu/cMonkey2-trac/.

Wang, Q. Z., C. Y. Wu, et al. (2006). "Integrating metabolomics into a systems biology framework to exploit metabolic complexity: strategies and applications in microorganisms." Appl Microbiol Biotechnol **70**(2): 151-161.

Watanabe, S., M. Hamano, et al. (2003). "Mannitol-1-phosphate dehydrogenase (MtlD) is required for mannitol and glucitol assimilation in Bacillus subtilis: possible cooperation of mtl and gut operons." J Bacteriol **185**(16): 4816-4824.

Yannakakis, M. (1981). "Node-Deletion Problems on Bipartite Graphs." SIAM J. Comput. **10**(2): 310-327.

Yoshida, K., K. Kobayashi, et al. (2001). "Combined transcriptome and proteome analysis as a powerful approach to study genes under glucose repression in Bacillus subtilis." Nucleic Acids Res **29**(3): 683-692.

Yoshida, K., Y. H. Ohki, et al. (2004). "Bacillus subtilis LmrA is a repressor of the lmrAB and yxaGH operons: identification of its binding site and functional analysis of lmrB and yxaGH." J Bacteriol **186**(17): 5640-5648.

Yoshida, K., H. Yamaguchi, et al. (2003). "Identification of additional TnrA-regulated genes of Bacillus subtilis associated with a TnrA box." Mol Microbiol **49**(1): 157-165.

Youngman, P., J. B. Perkins, et al. (1984). "Construction of a cloning site near one end of Tn917 into which foreign DNA may be inserted without affecting transposition in Bacillus subtilis or expression of the transposon-borne erm gene." Plasmid **12**(1): 1-9.

Zuberi, A. R., C. W. Ying, et al. (1990). "Transcriptional organization of a cloned chemotaxis locus of Bacillus subtilis." J. Bacteriol. **172**(4): 1870-1876.

3. QUANTITIVE VALIDATION OF MULTI-SPECIES CMONKEY

Original article: Waltman, P., T. Kacmarczyk, et al. (2010). "Multi-species integrative biclustering." Genome Biololgy **11**(R96).

NOTE: This chapter contains sections from the original article that this chapter is based upon which describe the quantitative analysis that was performed in combination with the relevant method sections of the original supplementary material. The majority of the main text of the original article now serves as Chapter 2. The gene lists and images that were also contained in the supplement of the original article can now be found in Appendix 1, while the additional plots from the original supplementary material can now be found in Appendix 2.

Author contributions: Provided above, in section 2.7.

In this chapter we provide a description and genome-wide benchmarking of the multispecies integrative biclustering method (or FD-MScM for full-data multi-species *cMonkey*). We compare our method to the original single-species *cMonkey* algorithm, a simple k-means clustering method that has been adapted to multi-species analysis and to several other single- and multi-species biclustering algorithms. We will refer only to analysis of pairs of organisms here and focus primarily on the *B. subtilis-B.*

anthracis pair, though we performed the validation on all the pairings that were possible from the two organism triplets that were analyzed (yielding six (6) total pairings, with three each from both the Gram-positive and Gram-negative triplet). We note that the method scales linearly with the number of species being analyzed and can be extended to larger numbers of organisms. The difficulties in validating biclustering performance and the need to compare the algorithm to primarily single species methods required that we initially limit the scope of this work to the simpler pairwise case. This chapter is based heavily upon the global validation section of our article published in Genome Biology (Waltman, Kacmarczyk et al. 2010)

As described in chapter 2, our method is composed of two sequential phases: an initial step where conserved cores are learned in an integrated multiple-species fashion and a later step where species-specific features are added to the conserved core (called the elaboration step). The algorithm takes as input a matrix of normalized expression data for each organism (where each organism's data matrix may be normalized separately), upstream sequences for all genes, and one or more networks for each organism (in this case we used metabolic and signaling pathways from KEGG, predicted co-membership in an operon and phylogenetic profile networks). The experimental datasets collected for both triplets are described fully in chapter 2 (Table 2.4–

Table 2.7).

As described in chapter 2, the method begins by randomly selecting a single orthologous pair (e.g. *dnaA*) around which to build a seed bicluster. For the randomly selected orthologous pair, conditions are chosen in each organism's expression matrix where the orthologous gene from that organism is most significantly differentially expressed. The semi-random seed is completed by adding the 5 to 10 most correlated orthologous pairs (e.g. *dnaN*) to the randomly selected seed pair (over the conditions defined in each species). This heuristic seeding is required as most of the MScM score terms demand that a bicluster have three or more genes in each organism to compute the scores required for further iterations. Once seeded, orthologous gene pairs are then iteratively added to (e.g. *sigH*) or dropped from (e.g. *cwlH*) the growing bicluster using the multi-data/multi-species score until no improvements can be made (convergence). After a bicluster converges, new biclusters are seeded and built from additional random seeds until no significant biclusters can be found or a maximum number of biclusters is reached.

Biclusters are generated sequentially and the number of biclusters to be optimized is chosen by the user. Considering that initially optimized biclusters will be unaffected by later biclusters, the number of biclusters is set higher than the expected number of true co-regulated modules. For each of the three possible species pairs, we generated 150 biclusters in the shared (multi-species) data-space that were then elaborated in the single-species data-space. Thus, each bicluster contains a conserved core (orthologous pairs that were added based on the entire integrated dataset), and 0

or more genes that were added during the elaboration step (performed separately for each organism, based on each single species dataset).

3.1 Genome-wide assessment of multi-species biclustering performance

To validate MScM, we compared it to several multi-species and single-species methods (Table 3.1). Among the single-species methods, we included the single-species version of *cMonkey* (SSCM), which was previously shown to be competitive with other biclustering methods (Reiss, Baliga et al. 2006); as well as two recent single-species biclustering methods, QUBIC (Li, Ma et al. 2009) and Coalesce (Huttenhower, Mutungu et al. 2009) (COAL). In addition, we compared our method to a multi-species version of the ISA biclustering algorithm (MSISA) (Bergmann, Ihmels et al. 2003); and two multi-species clustering methods, a simple multi-species K-means algorithm (MSKM) (Herschkowitz, Simin et al. 2007) and a *balanced* multi-species K-means clustering method (BMSKM). We constructed the BMSKM version to balance the disproportionate size of expression datasets between the two species and thereby perform a more meaningful comparison to MScM. We refer to the results as "shared" (SH) if we restrict our analysis to orthologous pairs between the two species and "elaborated" (EL) if a second step is used to add species-specific genes, i.e. MScM-EL. When possible, we evaluate both SH and EL results. In order to remain consistent with the MSISA nomenclature (Bergmann, Ihmels et al. 2003) we also use the terms *purified* (MSISA-P) and *refined* (MSISA-R), as these terms were used in the

original work describing these methods. Descriptions of the multi-species methods can be found in the methods section. When evaluating integrative methods that take into account more than just expression data (FD: full data) we also compare to expression-only (EO) runs of each method. Our evaluation of the various methods is based on two criteria: 1) the ability to detect statistically significant modules, and more importantly to this work, 2) the ability to identify **conserved** modules. We show that MScM produces biclusters that are a good balance of coverage, functional significance, and conservation, suggesting that the biclusters obtained by this procedure are of greater biological significance.

Table 3.1: Key to abbreviations used for methods tested. Tested methods are shown organized by main method (leftmost column) data-types used, and whether the analysis was performed over the full genome or restricted to only genes with orthologs across the species analyzed. For each formulation (method, data and multi-species mode) we provide the short name or abbreviation that is used in tables, figures and throughout the text.

	Expression Only		**Full Data**	
Multi-Species	**shared space**	**full genome (elaboration)**	**shared space**	**full genome (elaboration)**
cMonkey	EO-MScM-SH	EO-MScM-EL	FD-MScM-SH	FD-MScM-EL
ISA*	MSISA-P	MSISA-R	NA	NA
K-Means*	MSKM-SH	MSKM-EL	NA	NA
(Balanced) K-Means*	BMSKM-SH	BMSKM-EL	NA	NA

Single-Species	Expression Only	Full Data
cMonkey	EO-SSCM	FD-SSCM
Coalesce	EO-COAL	FD-COAL
Qubic*	QUBIC	NA

*** Expression only method by method definition - no distinction between "expression only" or "full data" is necessary.**

We also note that the validation was originally performed only for the Gram-positive triplet (*B. subtilis*, *B. anthracis*, and *L. monocytogenes*). A subsequent, partial validation was later performed on the Gram-negative triplet (*E. coli*, *S. typhimurium* and *V. cholerae*), though, the validation on this triplet did not include the permutation tests we describe below, as these proved to largely uninformative because the results of nearly all the methods compared were significantly better than random.

3.1.1 Using multiple metrics for validating multi-species biclustering:

Validation and comparison of clustering methods remains a difficult problem (Prelic, Bleuler et al. 2006; Reiss, Baliga et al. 2006). There is, as of yet, no "solved" organism (i.e., an organism whose full regulatory network is known and experimentally validated) that can be used as a benchmark. Artificial datasets are also of limited value due to the complexity of generating reasonable synthetic datasets (one

would have to generate sequences, expression data and networks, and make assumptions about the evolution of these data-types). In the face of these challenges, several criteria for judging the biological significance of gene clusters have been implemented. We will focus on five metric classes: 1) bicluster coherence; 2) functional enrichment; 3) coverage; 4) overlap between biclusters and 5) conservation. We evaluate bicluster coherence with five metrics that gauge the support of the three data-types cMonkey integrates, described further below and in the supplement. We also assess the number of biclusters that have a significant enrichment, considering that enrichment metrics imply that co-functional and interacting genes (by protein-protein or regulatory interaction) should have a higher probability of clustering. Expression matrix coverage and overlap between biclusters were calculated as the percentage of data-matrix elements that can be in one or more biclusters (as opposed to just genes). Gene-wise comparisons can be found in the supplementary information.

The last metric we consider, unique to multi-species datasets, is the conservation of (bi)clustered genes between the two species. Although we cannot know *a priori* what percentage of co-regulated genes will be preserved, we can state for two closely related organisms that: 1) if two biclustering methods are equivalent (according to all other metrics), then the more conserved method is likely to be of higher biological significance; 2) the conserved score between biclustering methods should be well separated from a random background, but still lower than one. In addition, more distantly related organisms should have less conserved co-regulation.

By strictly enforcing a perfect conservation between the species, the two k-means variants (B/MSKM) are good examples of methods that over-estimate the degree of conservation between two species.

Figure 3.1–Figure **3.2** and Table 3.2 present this multiple-metric comparison for the *B. subtilis – B. anthracis* pairing; the summary of this multi-metric comparison for the results of the other organism pairings from the 2 triplets can be found in Table **3.3**–Table 3.7, and the associated figures can be found in Appendix 2. Given the above metrics and evolutionary considerations our assessment of methods attempts to balance the 5 metric classes above:

bicluster-quality =

[data support: (1) coherence, (2) functional enrichment] X

[completeness: (3) coverage, (4) overlap] X

[conservation: (5) conservation score]

3.1.2 Comparing the degree of conserved co-regulation detected by each method:

A bicluster is considered to be perfectly conserved when all of the orthologous genes from that bicluster are found in a single bicluster in the related species. We evaluated the ability of all the tested methods to identify conserved biclusters using a metric similar to the F-statistic (Stein, Eissen et al. 2003), which gauges the degree of recovery between a bicluster in one species with that of the closest bicluster in the other species. For the multi-species methods, we calculated the metric using the shared

bicluster for one organism with its bicluster counterpart in the other. Details of the procedure can be found in the methods section.

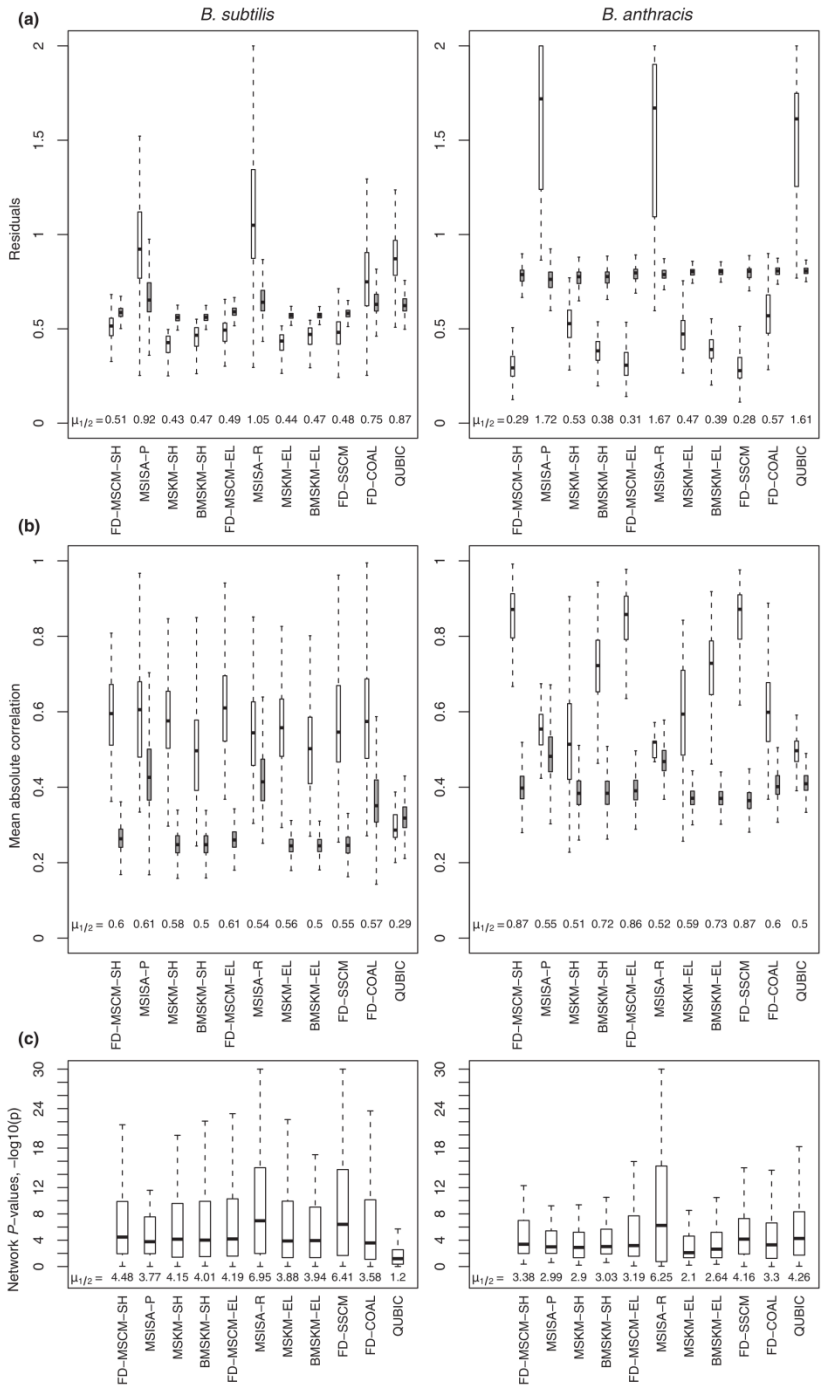


Figure 3.1: Comparing the distribution of expression and network coherence for single and multi-species methods for *B. subtilis* – *B. anthracis*. A comparison of the expression and network coherence

for the different MS and SS methods. For brevity, we only present here the results from full data methods (FD) from the *B. subtilis-B. anthracis* pairing (the results for the other pairings and expression only (EO) methods can be found in the supplementary material). Abbreviations are given for each method, a key to these abbreviations can be found in Table 3.1. Across the three comparisons, no method outperformed all other methods as judged by all three metrics, with the MScM results performing competitively with the others. **(a)** The distributions of the residuals from each method for the pairing of *B. subtilis* and *B. anthracis*. We also show, next to each distribution (in gray), the residuals from randomly shuffled (bi)clusters that match the size distribution for each method with n=1000 for the number of copies of the original set of (bi)clusters (same number of genes, conditions and (bi)clusters). Most methods tested were significantly better than random for both organisms; the exceptions being MSISA, Qubic, and Coalesce. In addition, this plot illustrates the tendency of MSKM to allow an organism with a considerably larger expression dataset to dominate the analysis. **(b)** The distributions of the average absolute correlation from each method for the pairing of *B. subtilis* and *B. anthracis* are displayed to allow comparison between methods that identify inversely correlated biclusters (MSISA, Qubic) and those that do not. As in (A), we also display the results from a randomly shuffled distribution next to each method in gray (n=1000). In all cases, with the exception of Qubic for *B. subtilis*, the method was significantly higher than random. **(c)** The distributions of the association p-values (-log10) from each method compared.

Using this simple measure of conservation, we evaluated the results from all the multi-species (MS) methods with those from several single-species (SS) methods (Table 3.2 displays the results for the *B. subtilis-B. anthracis* pairing;

Table **3.3**–Table 3.7 for the others). With the exception of MSISA-R, the MS methods displayed a far greater degree of conservation than any of the SS methods,

with the shared (SH) steps (and the equivalent MSISA-P step) having perfect conservation, and the elaboration (EL) steps having conservation scores >0.85. As they overestimate the conservation between the two species by assuming perfect conservation for all orthologous pairs during their shared steps, both B/MSKM-EL results display a greater degree of conservation than the MScM-EL results. In contrast, none of the SS methods possessed a conservation score > 0.125 (although it is likely that this score underestimates the degree of conserved co-regulation conservation scores for many of these methods were still significantly greater than random (*unpublished results*)).

The low conservation score for closely related organisms obtained when running SS methods on individual datasets was surprising. We expected that the truly conserved co-regulated gene groups would be detected individually by the SS methods and thus contribute to higher conservation scores. We attribute the low conservation scores in part to biologically relevant differences in co-regulation, but also to the fact that SS biclusters are supported by smaller datasets that contain systematic errors that likely differ between species (and thus, correctly cancel out in the multi-species analysis). Importantly, the greater conservation scores for MScM had little or no negative impact on the other commonly used evaluation metrics we employed.

3.1.3 Coherence of biclusters, coverage and bicluster overlap:

In this section we evaluate the ability of each method to simultaneously find coherent biclusters (Figure 3.1), cover the input dataset, and minimize the overlap

between biclusters (Figure 3.2). We assess bicluster expression coherence by 1) residual, the mean error when the average expression value over the bicluster is used to predict gene expression levels, (Figure 7.14–Figure 7.18); and 2) mean correlation, the average pairwise correlation between all (bi)cluster members, taking the absolute value of the correlation to allow unbiased comparison between methods that identify inversely correlated patterns (QUBIC and MSISA) and those that do not (Figure 7.19–Figure 7.24). These two measures are dependent on the number of conditions and rows in the bicluster and overall coverage of the data-matrix. Therefore, in all cases we compare co-expression values to a randomized background generated specifically for that biclustering (see methods). We assess bicluster network coherence by 3) association network p-values, a measure of the significance of the sub-networks within biclusters compared to the full network (Figure 7.25–Figure 7.30). We assess bicluster sequence coherence by 4) upstream motif E-values, a measure of the quality/significance of the upstream binding site motifs detected for each bicluster (Figure 7.31–Figure 7.36); and 5) sequence p-values, representing the preferential partitioning of the discovered motifs to genes in the bicluster over the remainder of the genome (Figure 7.37–Figure 7.42). We direct the reader to the supplementary material and prior work (Reiss, Baliga et al. 2006) for detailed descriptions of these metrics, along with the individual comparisons. Note, in the case of the non-integrative methods, sequence and network based metrics or scores were calculated *post hoc* for the (bi)clusters they produced.

3.1.3.1 Summary of bicluster coherence metric evaluations

We found that for all 5 coherence metrics, FD-MScM performed as well or better than the other methods (Table 7.25–Table 7.34). Specifically, in the case of the Gram-positive triplet, FD-MScM performed as well or better than the other methods in 71 of the 92 individual comparisons of the expression residual distributions, in all 92 of the mean correlation comparisons, in 77 of the 92 comparisons for the network association p-values, in 69 of the 92 comparisons for the motif E-values, and in 72 of the 92 comparisons for the sequence p-values. Note, the large number of comparisons (92) results from the fact that we have three organism pairings and that for each run we must separate the multi-species run into a set of biclusters for each species to calculate these validation metrics (thus each species pair results in 2 x the number. Similar results were observed for FD-MScM on the Gram-negative triplet as well; and, likewise, comparisons with EO-MScM on the Gram-positive triplet (Table 7.35–Table **7.39**) indicated that for four of the five metrics, it did as well or better than the other methods tested – the sole exception being motif E-values. Note, we re-iterate that during the generation of the EO-MScM results, the MScM optimization was run solely on expression data, with the scores for the other supporting data types (sequence and association networks) calculated *apriori*; thus, it is interesting that it'd do so well on these other data types. In contrast, in the comparisons with the EO-MScM results for the Gram-negative triplet, EO-MScM fared *worse* than its competitors in three of the 5 metrics, though, ironically, the metrics in which it did better were those associated

with the sequence data. It is unclear why this was the case. However, given that four (4) of the competitors that consistently outperformed it were other cMonkey versions (FD-SSCM, EO-SSCM, and FD-MScM-SH/EL), the most likely explanation is that in this particular instance, the optimization settled into a sub-optimal local minima, as is possible with Monte Carlo methods.

In the comparisons with the random permutation results for the expression metrics (Table 7.40–Table 7.41), expression residuals for the MScM and SSCM were all significantly better than random distributions generated for each method (differing cluster and bicluster sizes required a separate calculation of the random background for these expression coherence metrics for each method and for each data-set), for all organisms and pairing combinations, as were those for the two MS k-means variants (B/MSKM). In contrast, the residuals from both QUBIC and the two MSISA steps were all significantly worse than random; while the residuals from COAL were significantly better for *B. anthracis*, but somewhat worse for *B. subtilis* and *L. monocytogenes*. However, when considering the mean correlation results, nearly all methods were better than random; the sole exception to this being the MSISA results for *L. monocytogenes* in the pairing with *B. subtilis*.

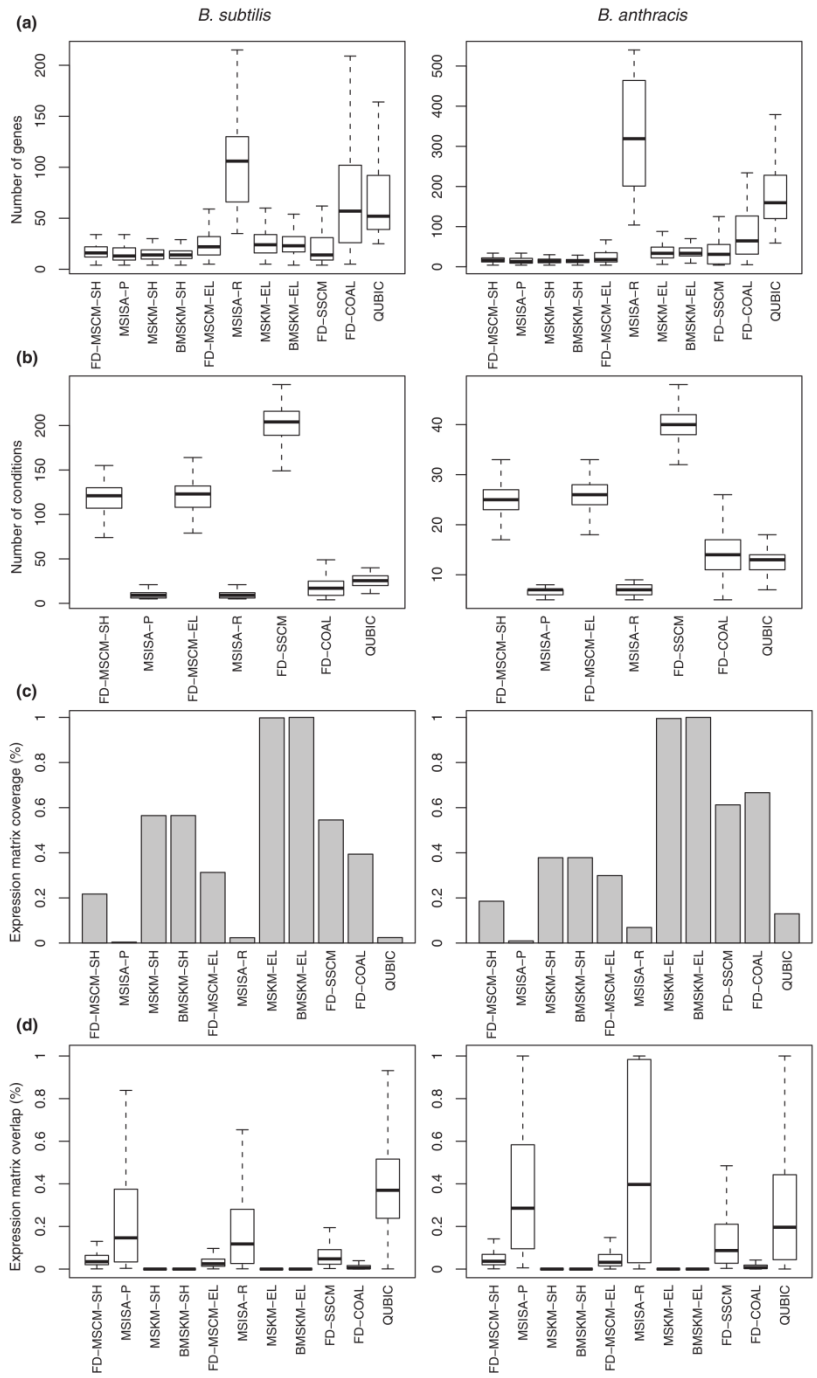


Figure 3.2: Comparison of the size, coverage and overlap for single and multi-species methods for the *B. subtilis – B. anthracis* pairing (full data results only, where applicable). For brevity, we only present here the results from full data methods (FD) from the *B. subtilis-B. anthracis* pairing (results for

the other pairings and EO methods can be found in the supplementary material) **(A)** The distribution of the number of genes in the (bi)clusters from the different methods. There is a consistent increase in the median size between the shared and elaboration steps (this is most extreme in the case of the MSISA method). For both organisms, Coalesce and Qubic produced the next largest biclusters, in terms of the number of genes. **(B)** The distribution of the number of conditions in the biclusters from the different biclustering methods only. We do not show this for the MSKM and BMSKM results as these methods use all conditions. For both organisms, the MS/SS cMonkey methods produced the biclusters with the most conditions. The MSISA method produced the biclusters with the least number of conditions. **(C)** The coverage of the total expression data matrix by the (bi)clusters from the different methods is displayed. The elaborated results of the MSKM and BMSKM methods achieve perfect coverage, by definition. The MSISA and Qubic biclusters had the smallest coverage of any of the methods, while the Coalesce biclusters achieved coverages comparable with the SSCM biclusters. **(D)** The distribution of all pairwise, non-zero overlaps between the (bi)clusters from the different methods; overlap in terms of the overlap of expression matrix elements, rather than genes. By definition, the MSKM and BMSKM clusters have no overlap, while the MSISA and Qubic biclusters had the greatest. Of the biclustering methods, Coalesce had the least overlap. Coalesce identifies more distinct biclusters with greater numbers of genes, but fewer conditions; and the SS/MS cMonkey methods identify biclusters that are slightly more overlapped than does Coalesce, with fewer genes, but covering more conditions.

3.1.3.2 Summary of bicluster coverage and overlap

Regardless of the pairing, both QUBIC and MSISA produced biclusters with the most genes (Figure 7.43-Figure 7.48) and fewest conditions (Figure 7.49-Figure 7.54), while also simultaneously having the least coverage (Figure 7.55-Figure 7.66) and most redundant set of biclusters (Figure 7.67-Figure 7.78). The sole exception to

this are the QUBIC results from the Gram-negative triplet, which while still having the fewest conditions, least coverage and greatest overlap, produced the biclusters with the fewest genes. This difference between the two triplets is most likely attributable to a different parameterization that was used for QUBIC, as described in section 3.4.4.

We exclude QUBIC and MSISA from further consideration for this reason. By contrast, the two B/MSKM variants display complete coverage of the data space. Although it is not possible to say what the optimal value for coverage should be, it is clear that: 1) numbers approaching 100% include several false positives (with respect to conserved co-regulation) as one cannot reasonably expect every gene to be a member of a conserved regulatory module; 2) methods that cover 2% or less of the data space are likely missing the majority of conserved co-regulation. We note that the coverage of both the genome and expression dataset for MScM is considerably smaller in comparison to SSCM and COAL. This is not unexpected because the search spaces are constrained by the orthologous core, with the search space of the elaboration step indirectly constrained by results of the shared step. The SS methods typically had better coverage, reflecting that a significant fraction of co-expressed gene-groups are not conserved across the species investigated.

3.1.4 Estimating functional coherence via enrichment of function annotations:

We compared the percentages of biclusters that were significantly enriched (p-value < 0.01) for both GO terms and co-presence in KEGG pathways. Again, we limit the discussion of these below to the pairing of *B. subtilis* with *B. anthracis* (Figure 3.3

and Figure 7.79, in greater detail), though similar patterns were observed with the other pairings as well (Figure 7.80-Figure 7.84). For all of the multi-species methods, there was a consistent increase between the shared and elaboration optimizations, indicating the importance of adding species-specific genes to conserved co-regulated cores. For example, for FD-MScM, the percentage of biclusters with GO term enrichments increases from 51.3% to 56.0% for *B. subtilis* (from 51.3% to 72.7% for *B. anthracis*) between the shared and elaboration optimizations (similarly, for MSKM, the increase is from 50.7% to 63.5% for *B. subtilis*; 39.2% to 75.7% for *B. anthracis*). The large increase observed for the MSISA results (53.7% to 95.1% for *B. subtilis*; 75.7% to 100% for *B. anthracis*) is a reflection of the small number of large and highly redundant biclusters it identifies. When a filter is applied that allows a GO term to be enriched for only a single bicluster, these percentages drop considerably (70.1% for *B. subtilis*, 39% for *B. anthracis*, MSISA-R biclusters).

The percentage of biclusters with enriched KEGG pathways is much higher for the MS methods than for SSCM. For example, the percentage of the FD-MScM-EL for *B. subtilis* was enriched 15.3%, while the percentage of the FD-SSCM results was 11.5% (21.3% vs. 9.4% for *B. anthracis*). We observed a pattern similar to what was observed with the GO terms, in the sense that there was also a consistent increase between the shared and elaboration runs. For example, with FD-MScM, the percentages increase from 12.7% to 15.3% for *B. subtilis* (12.7% to 21.3% for *B. anthracis*).

We also compared the performance of different species-species pairings (see supplement for data). We observed that for both of the pairings involving *B. subtilis*, the residuals of the clusters generated by MSKM were significantly better for the *B. subtilis* clusters, but significantly worse for the other organisms. As the *B. subtilis* expression dataset contained nearly six times more conditions than the other organisms, a key limitation of this and other similarly constructed methods is the dominance of a single species in the results. This effect was muted by the 'balancing' procedure (i.e. the BMSKM method). However, while the performance for the organism with the smaller dataset improved, the performance for the organism with the larger dataset decreased significantly. A similar effect was observed with MSISA.

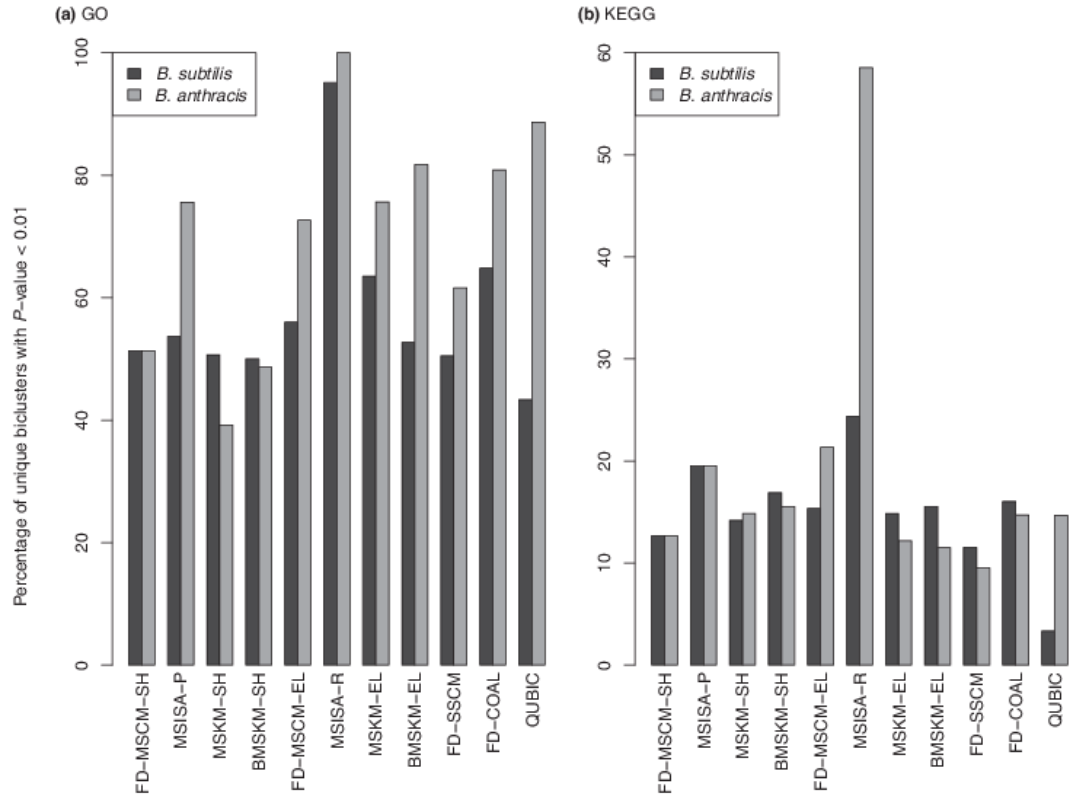


Figure 3.3: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments for the single and multi-species methods for the *B. subtilis* – *B. anthracis* pairing. (A) GO Terms. For all multi-species methods there is a consistent increase from the shared to elaboration step, with the percentage of elaborated biclusters with significant GO term enrichments consistently greater than those from the single species optimization. **(B) KEGG Pathways.** For both of the multi-species biclustering methods (MScM and MSISA), there is a consistent increase in percentage from the shared to elaborated optimizations, similar to the GO term enrichments, with a similarly large increase for the refined MSISA biclusters for *B. anthracis*. The two k-means clustering variants showed either negligible increase or even a decrease between the shared and elaboration steps.

Finally, we noticed that there was a consistent increase in the quality of the motifs associated with the biclusters returned by the elaboration step of both MS methods. One possible explanation for this behavior is simply algorithmic, namely, that MEME (Bailey and Elkan 1994), the motif inference tool we use, is able to infer more significant motifs from the larger pool of sequences accessible to the elaborated biclusters. Another reason may be that this behavior indicates a significant species-specific change at the level of binding sites, even when the gene membership in a module is conserved (an example of this is provided below). Our methodology for modeling and detecting binding sites as part of the multi-species procedure can likely be improved substantially and should prove a promising area for future work.

3.2 Overview of the (bi)cluster comparison metrics

We compared the relative performances of the four multi-species methods (MScM, MSISA, MSKM and BMSKM), and the three single species methods (SSCM, Coalesce and Qubic) compared in this study using 5 metric classes: 1) bicluster coherence; 2) functional enrichment; 3) coverage; 4) overlap between biclusters; and 5) conservation, described in the main text (Tables 2-7). We gauge bicluster coherence with five commonly used metrics that gauge the degree of support that is provided to each bicluster by the three data types that cMonkey integrates (expression, sequence and association networks). For comparison of SSCM to other biclustering algorithms, and comparison between single species biclustering and clustering algorithms, see (Prelic, Bleuler et al. 2006; Reiss, Baliga et al. 2006). Our

coherence metrics are: 1) expression residuals, a measure of the coherence of expression across the two species datasets for conditions within the bicluster; 2) mean correlation, the average pairwise correlation between members of a (bi)cluster (taking the absolute value to allow fair comparison between methods that identify inversely correlated patterns (QUBIC and MSISA) and those that do not; 3) network p-values, a measure of the significance of the sub-networks within biclusters compared to the full network; 4) motif E-values, a measure of the quality/significance of the upstream binding site motifs detected for each bicluster; and 5) sequence p-values, an estimate of a sequence's match to the motifs associated with a (bi)cluster. Each of the coherence metrics will be described in greater detail below as we discuss the relative performance of MScM to the other methods.

3.3 Quick-glance tables for all pairings

In the tables below, we compare several metrics of bicluster conservation, coverage, and functional enrichment. In all cases metrics are averaged over all biclusters produced by that method for each species. Abbreviations are given for each method, see Table 1 for a key to their abbreviations. In each column, the results for *B. subtilis* are listed first, with those for *B. anthracis* listed in parentheses. **Conservation Score** provides an estimate of the conservation identified between biclusters of the different organisms as defined in the methods; **Mean Correlation** measure the coherence of the biclusters given the expression; **Mean Net p-value** measures the enrichment of network edges within biclusters; **Mean Number of Genes** and

Conditions and **Number of Biclusters** summarize the size distributions of the (bi)clusters identified; **Coverage** is the percentage of the total expression data that is found in one or more (bi)cluster; **Overlap** estimates the redundancy of the (bi)clusters, overlap is calculated as the mean of the max % overlap for each bicluster in the full set of biclusters for a given method; **Percent (bi)clusters enriched (pval < 0.01)** for **GO/KEGG** provides an estimate of the functional significance of the (bi)clusters identified; and **Number of Unique Enriched Terms** for **GO/KEGG** are the number of unique terms across all biclusters for that method, this number of enriched terms provides an estimate of the redundancy of the biological functions enriched in one or more biclusters across the full set of biclusters for any given method. Further explanations of these metrics can be found within the text and supplement.

Table 3.2: Summary of evaluation criteria for the single and multi-species methods for the *B. subtilis – B. anthracis* pairingglffdlkjsajffladsfdf.

	Conservation Score	Mean Correlation: absolute val	Net p-value: -log10	Mean Number of Genes	Mean Number of Conditions	Number of Biclusters
EO MScM-SH	1	0.52 (0.69)	8.21 (6.45)	16.78 (16.78)	125.74 (25.86)	148 (148)
FD MScM-SH	**1**	**0.59 (0.85)**	**9.10 (8.57)**	**21.82 (21.82)**	**116.97 (24.87)**	**150 (150)**
MSISA-P	1	0.60 (0.56)	5.92 (5.63)	16.90 (16.90)	10.22 (6.85)	41 (41)
MSKM-SH	1	0.58 (0.52)	11.49 (11.62)	14.99 (14.99)	314 (51)	148 (148)
BMSKM-SH	**1**	**0.49 (0.72)**	**9.89 (12.19)**	**15.00 (15.00)**	**314 (51)**	**148 (148)**
EO MScM-EL	0.907	0.54 (0.69)	7.41 (6.35)	22.74 (23.60)	129.69 (27.07)	148 (148)
FD MScM-EL	**0.852**	**0.61 (0.84)**	**7.64 (8.65)**	**33.75 (34.63)**	**119.87 (26.26)**	**150 (150)**
MSISA-R	0.093	0.55 (0.51)	3.54 (8.87)	106.05 (335.71)	10.22 (6.93)	41 (41)
MSKM-EL	0.956	0.56 (0.58)	10.27 (6.65)	26.49 (39.44)	314 (51)	148 (148)
BMSKM-EL	**0.959**	**0.50 (0.71)**	**8.58 (7.93)**	**26.54 (39.63)**	**314 (51)**	**148 (148)**
EO SSCM	0.098	0.70 (0.91)	8.58 (7.43)	26.19 (34.11)	193.40 (38.66)	161 (210)
FD SSCM	**0.124**	**0.56 (0.82)**	**10.14 (7.31)**	**23.06 (40.65)**	**200.76 (39.81)**	**295 (315)**
EO COAL	0.107	0.58 (0.64)	5.21 (5.06)	86.65 (115.71)	20.09 (13.13)	300 (158)
FD COAL	**0.101**	**0.59 (0.62)**	**5.27 (5.69)**	**88.16 (131.12)**	**20.24 (14.24)**	**287 (136)**

QUBIC	0.054	0.36 (0.49)	1.38 (5.90)	71.59 (188.25)	25.45 (12.63)	150 (150)
			GO		**KEGG**	
	Coverage element-wise	**Mean Overlap element-wise**	**Percent (bi)clusters enriched (pval < 0.01)**	**Number Unique Enriched Terms**	**Percent (bi)clusters enriched (pval < 0.01)**	**Number Unique Enriched Pathways**
EO MScM-SH	18.69% (15.73%)	4.76% (5.20%)	33.78% (37.16%)	378 (338)	4.05% (6.76%)	10 (16)
FD MScM-SH	**21.71% (18.53%)**	**5.33% (5.93%)**	**51.33% (51.33%)**	**575 (500)**	**12.67% (12.67%)**	**24 (28)**
MSISA-P	0.41% (0.95%)	22.24% (34.64%)	53.66% (75.61%)	160 (164)	19.51% (19.51%)	12 (15)
MSKM-SH	56.49% (37.83%)	0% (0%)	50.68% (39.19%)	617 (559)	14.19% (14.86%)	22 (25)
BMSKM-SH	**56.52% (37.85%)**	**0% (0%)**	**50.00% (48.65%)**	**658 (578)**	**16.89% (15.54%)**	**29 (34)**
EO MScM-EL	25.03% (21.68%)	4.38% (5.06%)	40.54% (60.81%)	449 (485)	11.49% (10.81%)	18 (18)
FD MScM-EL	**31.29% (29.90%)**	**4.00% (5.72%)**	**56.00% (72.67%)**	**649 (664)**	**15.33% (21.33%)**	**30 (37)**
MSISA-R	2.36% (6.90%)	18.34% (46.28%)	95.12% (100.00%)	287 (235)	24.39% (58.54%)	10 (20)
MSKM-EL	99.80% (99.52%)	0% (0%)	63.51% (75.68%)	732 (675)	14.86% (12.16%)	31 (30)
BMSKM-EL	**100% (100%)**	**0% (0%)**	**52.70% (81.76%)**	**743 (710)**	**15.54% (11.49%)**	**35 (25)**
EO SSCM	39.48% (46.81%)	9.44% (14.10%)	42.24% (66.19%)	499 (629)	10.56% (17.62%)	19 (29)

FD SSCM	**54.55% (61.24%)**	**7.53% (15.46%)**	**50.51% (61.59%)**	**746 (712)**	**11.53% (9.52%)**	**32 (31)**
EO COAL	40.21% (66.40%)	1.94% (2.12%)	63.67% (76.58%)	744 (659)	17.67% (9.49%)	32 (24)
FD COAL	**39.39% (66.63%)**	**2.06% (2.16%)**	**64.81% (80.88%)**	**776 (686)**	**16.03% (14.71%)**	**24 (24)**
QUBIC	2.43% (12.95%)	38.34% (26.49%)	43.33% (88.67%)	227 (331)	3.33% (14.67%)	5 (13)

Table 3.3: Summary of evaluation criteria for the single and multi-species methods for the *B. subtilis – L. monocytogenes* pairing.

	Conservation Score	**Mean Correlation: absolute val**	**Net p-value: -log10**	**Mean Number of Genes**	**Mean Number of Conditions**	**Number of Biclusters**
EO MScM-SH	1	0.52 (0.64)	15.18 (8.20)	14.51 (14.51)	127.45 (27.31)	150 (150)
FD MScM-SH	**1**	**0.59 (0.80)**	**10.73 (8.79)**	**16.09 (16.09)**	**121.36 (25.96)**	**147 (147)**
MSISA-P	1	0.60 (0.47)	6.82 (0.00)	5.88 (5.88)	10.85 (4.97)	33 (33)
MSKM-SH	1	0.59 (0.51)	12.14 (12.66)	9.83 (9.83)	314 (56)	145 (145)
BMSKM-SH	**1**	**0.52 (0.63)**	**11.96 (12.39)**	**9.78 (9.78)**	**314 (56)**	**146 (146)**
EO MScM-EL	0.951	0.54 (0.64)	13.59 (8.49)	20.05 (18.92)	132.92 (30.29)	150 (150)
FD MScM-EL	**0.884**	**0.61 (0.81)**	**9.13 (7.41)**	**26.44 (25.73)**	**123.17 (28.84)**	**147 (147)**
MSISA-R	0.060	0.55 (0.50)	3.15 (3.12)	106.39 (113.05)	10.37 (6.42)	38 (38)
MSKM-EL	0.963	0.56 (0.55)	7.52 (8.37)	26.85 (18.66)	314 (56)	145 (145)
BMSKM-EL	**0.949**	**0.53 (0.64)**	**7.55 (9.43)**	**26.90 (19.14)**	**314 (56)**	**146 (146)**

EO SSCM	0.096	0.70 (0.86)	8.58 (4.98)	26.19 (30.95)	193.40 (40.99)	161 (83)
FD SSCM	**0.147**	**0.56 (0.71)**	**10.14 (6.70)**	**23.06 (19.79)**	**200.76 (42.32)**	**295 (300)**
EO COAL	0.088	0.58 (0.81)	5.21 (5.60)	86.65 (78.81)	20.09 (12.04)	300 (81)
FD COAL	**0.095**	**0.59 (0.80)**	**5.27 (5.46)**	**88.16 (84.15)**	**20.24 (12.73)**	**287 (78)**
QUBIC	0.048	0.36 (0.45)	1.38 (5.26)	71.59 (182.92)	25.45 (19.91)	150 (150)

			GO		KEGG	
	Coverage element-wise	**Mean Overlap element-wise**	**Percent (bi)clusters enriched (pval < 0.01)**	**Number Unique Enriched Terms**	**Percent (bi)clusters enriched (pval < 0.01)**	**Number Unique Enriched Pathways**
EO MScM-SH	15.65% (26.16%)	6.46% (6.17%)	22.67% (20.67%)	339 (303)	4.67% (3.33%)	11 (13)
FD MScM-SH	**15.85% (26.08%)**	**5.95% (5.87%)**	**37.41% (35.37%)**	**427 (371)**	**10.20% (6.12%)**	**19 (19)**
MSISA-P	0.14% (0.42%)	29.03% (45.95%)	48.15% (37.04%)	109 (86)	18.18% (21.21%)	8 (9)
MSKM-SH	36.28% (50.98%)	0% (0%)	37.93% (34.48%)	500 (398)	10.34% (8.97%)	24 (27)
BMSKM-SH	**36.35% (51.09%)**	**0% (0%)**	**30.82% (31.51%)**	**479 (411)**	**11.64% (11.64%)**	**18 (24)**
EO MScM-EL	21.98% (32.52%)	5.36% (6.47%)	37.33% (30.67%)	449 (386)	8.67% (8.00%)	16 (15)
FD MScM-EL	**25.95% (40.29%)**	**4.65% (6.12%)**	**56.46% (53.74%)**	**542 (468)**	**16.33% (10.20%)**	**23 (19)**

MSISA-R	2.27% (5.44%)	17.80% (57.90%)	97.37% (92.11%)	285 (179)	31.58% (57.89%)	13 (14)
MSKM-EL	99.11% (96.78%)	0.00% (0.00%)	59.31% (44.14%)	640 (476)	15.17% (11.03%)	28 (20)
BMSKM-EL	**100% (100%)**	**0.00% (0.00%)**	**51.37% (46.58%)**	**669 (480)**	**12.33% (11.64%)**	**25 (24)**
EO SSCM	39.48% (37.34%)	9.44% (15.76%)	42.24% (55.42%)	499 (298)	10.56% (19.28%)	19 (15)
FD SSCM	**54.55% (61.27%)**	**7.53% (13.19%)**	**50.51% (36.91%)**	**746 (451)**	**11.53% (5.67%)**	**32 (17)**
EO COAL	40.21% (41.73%)	1.94% (8.65%)	63.67% (53.09%)	744 (319)	17.67% (11.11%)	32 (12)
FD COAL	**39.39% (43.07%)**	**2.06% (9.69%)**	**64.81% (56.41%)**	**776 (294)**	**16.03% (11.54%)**	**24 (11)**
QUBIC	2.43% (9.14%)	38.34% (62.22%)	43.33% (100.00%)	227 (175)	3.33% (62.00%)	5 (4)

Table 3.4: Summary of evaluation criteria for the single and multi-species methods for the B. anthracis – L. monocytogenes pairing. Note, results for MSISA and BMSKM are not reported as these methods were not performed for this pairing.

	Conservation Score	**Mean Correlation (absolute value)**	**Net p-value (-log10)**	**Number of Genes**	**Number of Conditions**	**Number of Biclusters**
EO MScM-SH	1	0.63 (0.63)	5.90 (5.92)	15.78 (15.78)	25.60 (27.51)	141 (141)
FD MScM-SH	**1**	**0.82 (0.77)**	**8.82 (6.28)**	**16.81 (16.81)**	**24.82 (26.05)**	**148 (148)**
MSKM-SH	1	0.69 (0.60)	9.95 (13.62)	10.20 (10.20)	51.00 (56.00)	145 (145)
EO MScM-EL	0.963	0.63 (0.63)	6.79 (6.51)	20.69 (19.79)	26.96 (30.59)	141 (141)

FD MScM-EL	**0.906**	**0.80 (0.78)**	**8.15 (5.59)**	**25.26 (23.90)**	**26.43 (28.97)**	**148 (148)**
MSKM-EL	0.943	0.70 (0.63)	5.95 (9.19)	39.63 (19.11)	51.00 (56.00)	145 (145)
EO SSCM	0.090	0.91 (0.86)	7.43 (4.98)	34.11 (30.95)	38.66 (40.99)	210 (83)
FD SSCM	**0.126**	**0.82 (0.71)**	**7.31 (6.70)**	**42.02 (19.79)**	**39.87 (42.32)**	**300 (300)**
EO COAL	0.102	0.64 (0.81)	5.06 (5.60)	115.71 (78.81)	13.13 (12.04)	158 (81)
FD COAL	**0.101**	**0.62 (0.80)**	**5.69 (5.46)**	**131.12 (84.15)**	**14.24 (12.73)**	**136 (78)**
QUBIC	0.045	0.49 (0.45)	5.90 (5.26)	188.25 (182.92)	12.63 (19.91)	150 (150)

			GO		**KEGG**	
	Coverage (element-wise	**Overlap (element-wise)**	**Percent Significant (bi)clusters**	**Number of Significant Terms**	**Percent Significant (bi)clusters**	**Num. Unique Pathways**
EO MScM-SH	12.99% (26.51%)	6.03% (5.51%)	20.57% (21.28%)	281 (286)	7.09% (6.38%)	10 (11)
FD MScM-SH	**13.97% (27.33%)**	**6.71% (5.87%)**	**40.54% (43.24%)**	**432 (423)**	**10.14% (10.81%)**	**18 (17)**
MSKM-SH	25.22% (52.92%)	0.00% (0.00%)	38.62% (37.24%)	454 (443)	9.66% (9.66%)	20 (21)
EO MScM-EL	16.89% (32.56%)	5.60% (6.10%)	52.48% (33.33%)	466 (359)	9.22% (7.09%)	21 (13)
FD MScM-EL	**20.80% (38.82%)**	**5.99% (5.61%)**	**79.73% (56.76%)**	**590 (479)**	**16.22% (14.86%)**	**30 (23)**
MSKM-EL	97.97% (99.14%)	0.00% (0.00%)	79.31% (53.79%)	742 (525)	11.72% (11.03%)	24 (24)

EO SSCM	46.81% (37.34%)	14.10% (15.76%)	66.19% (55.42%)	629 (298)	17.62% (19.28%)	29 (15)
FD SSCM	**60.29% (61.27%)**	**15.72% (13.19%)**	**62.33% (36.91%)**	**707 (451)**	**10.00% (5.67%)**	**32 (17)**
EO COAL	66.40% (41.73%)	2.12% (8.65%)	76.58% (53.09%)	659 (319)	9.49% (11.11%)	24 (12)
FD COAL	**66.63% (43.07%)**	**2.16% (9.69%)**	**80.88% (56.41%)**	**686 (294)**	**14.71% (11.54%)**	**24 (11)**
QUBIC	12.95% (9.14%)	26.49% (62.22%)	88.67% (100.00%)	331 (175)	14.67% (62.00%)	13 (4)

Table 3.5: Summary of evaluation criteria for the single and multi-species methods for the *E. coli – S. typhimurium* pairing.

	Conservation Score	**Mean Correlation: absolute val**	**Mean Net p-value: -log10**	**Mean Number of Genes**	**Mean Number of Conditions**	**Number of Biclusters**
EO MScM-SH	1	0.52 (0.45)	7.51 (3.56)	20.95 (20.95)	230.65 (58.93)	150 (150)
FD MScM-SH	**1**	**0.68 (0.55)**	**16.40 (13.65)**	**26.28 (26.28)**	**227.58 (56.08)**	**149 (149)**
MSISA-P	1	0.56 (0.60)	3.78 (8.43)	7.78 (7.78)	25.72 (11.88)	60 (60)
MSKM-SH	1	0.59 (0.29)	9.64 (5.65)	19.07 (19.07)	507.00 (138.00)	148 (148)
BMSKM-SH	**1**	**0.54 (0.37)**	**11.77 (4.40)**	**18.85 (18.85)**	**507.00 (138.00)**	**150 (150)**
EO MScM-EL	0.894	0.54 (0.47)	4.71 (3.35)	29.13 (27.72)	231.82 (62.19)	150 (150)
FD MScM-EL	**0.764**	**0.66 (0.50)**	**19.92 (16.81)**	**39.65 (36.64)**	**227.23 (56.15)**	**149 (149)**
MSISA-R	0.022	0.52 (0.46)	6.13 (3.97)	38.85 (189.47)	25.72 (13.50)	60 (60)

MSKM-EL	0.994	0.57 (0.31)	8.84 (4.98)	28.81 (25.30)	507.00 (138.00)	148 (148)
BMSKM-EL	**0.995**	**0.54 (0.38)**	**9.61 (3.87)**	**28.43 (24.97)**	**507.00 (138.00)**	**150 (150)**
EO SSCM	0.106	0.76 (0.66)	6.73 (3.58)	26.31 (27.54)	346.84 (91.96)	204 (155)
FD SSCM	**0.1**	**0.59 (0.58)**	**19.50 (5.00)**	**19.40 (29.97)**	**354.48 (94.13)**	**425 (157)**
EO COAL	0.097	0.64 (0.57)	6.28 (3.18)	70.53 (100.58)	39.71 (14.89)	239 (159)
FD COAL	**0.095**	**0.63 (0.57)**	**6.18 (3.16)**	**70.43 (100.67)**	**38.96 (14.88)**	**247 (159)**
QUBIC	0.038	0.91 (0.86)	27.73 (6.33)	6.67 (6.88)	27.45 (5.41)	139 (113)

			GO		**KEGG**	
	Coverage element-wise	**Mean Overlap element-wise**	**Percent (bi)clusters enriched (pval < 0.01)**	**Num. of Unique Enriched Terms**	**Percent (bi)clusters enriched (pval < 0.01)**	**Num. Unique Pathways**
EO MScM-SH	24.63% (25.89%)	4.31% (4.50%)	33.33% (36.67%)	479 (453)	9.33% (10.67%)	19 (18)
FD MScM-SH	**25.88% (26.83%)**	**6.42% (6.60%)**	**65.10% (67.11%)**	**806 (656)**	**23.49% (21.48%)**	**33 (38)**
MSISA-P	0.42% (0.64%)	9.57% (25.72%)	45.00% (33.33%)	228 (175)	23.33% (20.00%)	17 (12)
MSKM-SH	66.18% (75.35%)	0.00% (0.00%)	62.84% (61.49%)	918 (742)	15.54% (18.24%)	33 (39)
BMSKM-SH	**66.30% (75.49%)**	**0.00% (0.00%)**	**58.00% (54.67%)**	**885 (739)**	**12.00% (12.00%)**	**32 (32)**
EO MScM-EL	31.13% (32.29%)	4.38% (4.10%)	46.67% (35.33%)	617 (424)	11.33% (10.00%)	25 (19)
FD MScM-EL	**33.88% (33.27%)**	**5.79% (5.08%)**	**89.93% (81.21%)**	**999 (720)**	**48.32% (40.94%)**	**58 (53)**

MSISA-R	2.10% (2.69%)	6.00% (90.99%)	90.00% (18.33%)	570 (63)	31.67% (5.00%)	37 (2)
MSKM-EL	100.00% (100.00%)	0.00% (0.00%)	69.59% (58.78%)	1037 (721)	16.22% (18.24%)	40 (37)
BMSKM-EL	**100.00% (100.00%)**	**0.00% (0.00%)**	**71.33% (51.33%)**	**1054 (728)**	**13.33% (12.00%)**	**32 (32)**
EO SSCM	45.92% (40.74%)	13.18% (10.90%)	59.80% (29.03%)	926 (355)	17.16% (5.16%)	44 (12)
FD SSCM	**69.12% (38.08%)**	**10.53% (14.14%)**	**64.24% (28.66%)**	**1221 (316)**	**12.71% (5.73%)**	**47 (9)**
EO COAL	25.18% (50.91%)	2.75% (1.87%)	77.41% (32.08%)	986 (388)	33.47% (3.77%)	49 (14)
FD COAL	**25.79% (50.90%)**	**2.70% (1.87%)**	**79.76% (33.33%)**	**984 (391)**	**35.22% (5.03%)**	**48 (11)**
QUBIC	0.53% (0.51%)	24.35% (8.43%)	76.26% (14.16%)	437 (84)	38.85% (3.54%)	21 (2)

Table 3.6: Summary of evaluation criteria for the single and multi-species methods for the *E. coli – V. cholerae* pairing. Note, results for BMSKM are not reported as this method was not performed on this pairing.

	Conservation Score	**Mean Correlation: absolute value**	**Mean Net p-value: -log10**	**Mean Number of Genes**	**Mean Number of Conditions**	**Number of Biclusters**
EO MScM-SH	1	0.52 (0.41)	7.46 (5.99)	20.97 (20.97)	229.32 (176.89)	150 (150)
FD MScM-SH	**1**	**0.70 (0.55)**	**19.25 (15.14)**	**18.49 (18.49)**	**226.87 (168.06)**	**150 (150)**
MSISA-P	1	0.56 (0.69)	6.49 (13.01)	6.70 (6.70)	26.27 (55.32)	37 (37)
MSKM-SH	1	0.56 (0.43)	15.65 (6.75)	12.35 (12.35)	507.00 (441.00)	148 (148)

EO MScM-EL	0.944	0.54 (0.42)	7.93 (5.99)	29.01 (28.54)	231.03 (191.19)	150 (150)
FD MScM-EL	**0.748**	**0.66 (0.50)**	**20.03 (18.41)**	**31.34 (27.84)**	**226.27 (168.87)**	**150 (150)**
MSISA-R	0.022	0.51 (0.48)	6.21 (17.80)	46.38 (318.38)	26.27 (38.00)	37 (37)
MSKM-EL	0.961	0.55 (0.44)	14.62 (6.38)	28.81 (22.53)	507.00 (441.00)	148 (148)
EO SSCM	0.147	0.76 (0.68)	6.73 (8.82)	26.31 (21.86)	346.84 (289.72)	204 (202)
FD SSCM	**0.196**	**0.59 (0.60)**	**19.50 (9.04)**	**19.40 (24.41)**	**354.48 (266.31)**	**425 (274)**
EO COAL	0.141	0.64 (0.59)	6.28 (5.88)	70.53 (49.69)	39.71 (28.27)	239 (247)
FD COAL	**0.139**	**0.63 (0.59)**	**6.18 (6.04)**	**70.43 (50.65)**	**38.96 (28.49)**	**247 (248)**
QUBIC	0.003	0.91 (0.92)	27.73 (14.44)	6.67 (4.52)	27.45 (25.17)	139 (148)

			GO		KEGG	
	Coverage element-wise	**Mean Overlap element-wise**	**Percent (bi)clusters enriched (pval < 0.01)**	**Num. of Unique Enriched Terms**	**Percent (bi)clusters enriched (pval < 0.01)**	**Num. Unique Pathways**
EO MScM-SH	21.36% (25.17%)	5.90% (5.20%)	38.67% (34.67%)	612 (559)	14.00% (16.00%)	31 (28)
FD MScM-SH	**18.38% (21.28%)**	**7.40% (6.38%)**	**72.00% (66.00%)**	**830 (671)**	**19.33% (24.67%)**	**51 (45)**
MSISA-P	0.23% (0.56%)	12.25% (27.87%)	45.95% (45.95%)	209 (170)	32.43% (35.14%)	23 (27)
MSKM-SH	42.87% (54.81%)	0.00% (0.00%)	55.41% (50.00%)	851 (719)	16.89% (15.54%)	42 (37)

EO MScM-EL	27.53% (33.98%)	5.54% (4.51%)	50.67% (47.33%)	773 (652)	15.33% (17.33%)	43 (35)
FD MScM-EL	**26.86% (29.26%)**	**6.18% (5.03%)**	**94.67% (86.00%)**	**1093 (831)**	**44.00% (38.00%)**	**69 (59)**
MSISA-R	1.58% (3.57%)	6.02% (81.10%)	94.59% (100.00%)	480 (148)	48.65% (97.30%)	30 (20)
MSKM-EL	100.00% (100.00%)	0.00% (0.00%)	69.59% (60.81%)	1035 (845)	18.92% (16.89%)	43 (45)
EO SSCM	45.92% (48.68%)	13.18% (11.05%)	59.80% (55.94%)	926 (638)	17.16% (20.30%)	44 (35)
FD SSCM	**69.12% (50.56%)**	**10.53% (11.10%)**	**64.24% (62.77%)**	**1221 (717)**	**12.71% (22.99%)**	**47 (37)**
EO COAL	25.18% (20.43%)	2.75% (3.23%)	77.41% (55.87%)	986 (531)	33.47% (17.81%)	49 (27)
FD COAL	**25.79% (20.90%)**	**2.70% (3.16%)**	**79.76% (57.66%)**	**984 (545)**	**35.22% (16.13%)**	**48 (28)**
QUBIC	0.53% (0.68%)	24.35% (12.74%)	76.26% (30.41%)	437 (132)	38.85% (9.46%)	21 (7)

Table 3.7: Summary of evaluation criteria for the single and multi-species methods for the *S. typhimurium – V. cholerae* pairing. Note, results for MSISA are not reported as it was not performed on this pairing.

	Conservation Score	Mean Correlation: absolute value	Mean Net p-value: -log10	Mean Number of Genes	Mean Number of Conditions	Number of Biclusters
EO MScM-SH	1	0.44 (0.37)	4.31 (4.57)	19.96 (19.96)	56.58 (173.34)	150 (150)
FD MScM-SH	**1**	**0.55 (0.51)**	**11.68 (13.37)**	**17.81 (17.81)**	**54.57 (162.05)**	**150 (150)**
MSKM-SH	1	0.31 (0.49)	4.76 (4.90)	11.45 (11.45)	138.00 (441.00)	148 (148)

BMSKM-SH	**1**	**0.39 (0.43)**	**3.91 (4.27)**	**11.45 (11.45)**	**138.00 (441.00)**	**148 (148)**
EO MScM-EL	0.939	0.45 (0.39)	3.45 (5.00)	27.29 (26.01)	60.67 (188.61)	150 (150)
FD MScM-EL	**0.819**	**0.50 (0.48)**	**15.78 (19.96)**	**26.24 (26.07)**	**54.81 (163.84)**	**150 (150)**
MSKM-EL	0.965	0.35 (0.47)	3.43 (5.28)	25.30 (22.53)	138.00 (441.00)	148 (148)
BMSKM-EL	**0.966**	**0.41 (0.44)**	**2.68 (4.62)**	**25.30 (22.53)**	**138.00 (441.00)**	**148 (148)**
EO SSCM	0.126	0.66 (0.68)	3.58 (8.82)	27.54 (21.86)	91.96 (289.72)	155 (202)
FD SSCM	**0.1**	**0.58 (0.60)**	**5.00 (9.04)**	**29.97 (24.41)**	**94.13 (266.31)**	**157 (274)**
EO COAL	0.104	0.57 (0.59)	3.18 (5.88)	100.58 (49.69)	14.89 (28.27)	159 (247)
FD COAL	**0.104**	**0.57 (0.59)**	**3.16 (6.04)**	**100.67 (50.65)**	**14.88 (28.49)**	**159 (248)**
QUBIC	0.023	0.86 (0.92)	6.33 (14.44)	6.88 (4.52)	5.41 (25.17)	113 (148)

			GO		**KEGG**	
	Coverage element-wise	**Mean Overlap element-wise**	**Percent (bi)clusters enriched (pval < 0.01)**	**Number of Unique Enriched Terms**	**Percent (bi)clusters enriched (pval < 0.01)**	**Num. Unique Pathways**
EO MScM-SH	21.26% (24.31%)	5.61% (4.74%)	29.33% (30.00%)	368 (368)	6.67% (6.00%)	11 (9)
FD MScM-SH	**18.26% (19.75%)**	**6.05% (5.17%)**	**60.00% (60.67%)**	**571 (570)**	**16.67% (16.00%)**	**32 (38)**
MSKM-SH	45.26% (50.82%)	0.00% (0.00%)	45.27% (43.24%)	590 (587)	11.49% (13.51%)	21 (30)
EO MScM-EL	27.65% (31.83%)	4.79% (4.31%)	32.67% (36.00%)	395 (483)	9.33% (11.33%)	12 (15)

228

FD MScM-EL	**24.48% (26.53%)**	**5.03% (4.56%)**	**88.00% (93.33%)**	**690 (761)**	**37.33% (40.00%)**	**50 (60)**
MSKM-EL	100.00% (100.00%)	0.00% (0.00%)	42.57% (53.38%)	625 (719)	10.81% (18.24%)	29 (34)
BMSKM-EL	**100.00% (100.00%)**	**0.00% (0.00%)**	**45.27% (52.03%)**	**615 (742)**	**6.76% (13.51%)**	**22 (32)**
EO SSCM	40.74% (48.68%)	10.90% (11.05%)	29.03% (55.94%)	355 (638)	5.16% (20.30%)	12 (35)
FD SSCM	**38.08% (50.56%)**	**14.14% (11.10%)**	**28.66% (62.77%)**	**316 (717)**	**5.73% (22.99%)**	**9 (37)**
EO COAL	50.91% (20.43%)	1.87% (3.23%)	32.08% (55.87%)	388 (531)	3.77% (17.81%)	14 (27)
FD COAL	**50.90% (20.90%)**	**1.87% (3.16%)**	**33.33% (57.66%)**	**391 (545)**	**5.03% (16.13%)**	**11 (28)**
QUBIC	0.51% (0.68%)	8.43% (12.74%)	14.16% (30.41%)	84 (132)	3.54% (9.46%)	2 (7)

3.4 Methods

3.4.1 Explanation of the (bi)cluster coherence metrics

3.4.1.1 Residuals

Cheng and Church (Cheng and Church 2000) originally introduced residuals as a measure of bicluster coherence. For our purposes, we use a modified version of the residual measure used that takes into account gene-wise expression variance. Thus, if we let x_{ij} be the expression value for gene g in condition c, these are defined for any bicluster containing a set of G genes over C conditions as:

$$resid\left(G,C\right)=\left(\frac{\frac{1}{\left|G\right|\left|C\right|}\sum_{g\in G,c\in C}abs\left(x_{gc}-x_{gC}-x_{Gc}+x_{GC}\right)}{\frac{1}{\left|G\right|}\sum_{g\in G}var_{C}\left(x_{g}\right)}\right)$$

where

$$x_{gC}=\frac{1}{\left|C\right|}\sum_{c\in C}x_{gc},\ x_{Gc}=\frac{1}{\left|G\right|}\sum_{g\in G}x_{gc},\ x_{GC}\frac{1}{\left|G\right|\left|C\right|}\sum_{g\in G,c\in C}x_{gc}$$

and

$$var_{C}\left(x_{g}\right)=\frac{1}{\left|C\right|}\sum_{c\in C}\left(x_{gc}-x_{gC}\right)^{2}$$

As such, they can be understood to be a measure of the average deviation from the signal present within the bicluster, normalized by the average variance of the genes in G for the conditions in C. As a simple comparison, the residuals from the (bi)clusters

produced by each method were pooled and compared with each other using two-sided Wilcoxon's non-parametric rank tests. We direct the reader to Figure 7.13-Figure 7.18 as well as Table 7.25 (FD-MScM Table 7.35 (EO-MScM) and Table 7.40 (randomized tests) for the results of these comparisons, for each pairing of organisms.

3.4.1.2 Mean correlations

We also evaluated (bi)cluster expression coherence using the average pairwise correlation between genes in a (bi)cluster, over the conditions in the (bi)cluster. Because some of the methods we evaluated in this study could identify biclusters with inversely correlated patterns of expression, we took the absolute values of these correlations. Thus, if we let x_{ij} be the expression value for gene g in condition c, the mean correlations are defined for any bicluster containing a set of G genes over C conditions as:

$$mean.cor(G,C) = \left(\frac{|G|(|G|-1)}{2}\right)^{-1}\left(\sum_{x,y\in G;\, x\neq y} abs\left(\frac{\sum_{c\in C}(x_c-\bar{x})(y_c-\bar{y})}{\sqrt{\sum_{c\in C}(x_c-\bar{x})^2\sum_{c\in C}(y_c-\bar{y})^2}}\right)\right)$$

As a simple comparison, the average pairwise correlations from the (bi)clusters produced by each method were pooled and compared with each other using two-sided Wilcoxon's non-parametric rank tests. We direct the reader to Figure 7.19-Figure 7.24, as well as Table 7.27 (FD-MScM), Table 7.36 (EO-MScM) and Table 7.41 (randomized tests) for the results of these comparisons.

3.4.1.3 Network Association p-values

Briefly, the association p-values for a bicluster are modeled using a hypergeometric distribution, where for a given bicluster b_k for genome *G*, the association p-value for an individual network, *N*, is calculated as:

$$pvalue\left(b_k, N\right) = \frac{\begin{pmatrix} |N| \\ n_{b_k \to b_k} \end{pmatrix} \begin{pmatrix} poss\left(G\right) - |N| \\ poss\left(b_k\right) - n_{b_k \to b_k} \end{pmatrix}}{\begin{pmatrix} poss\left(G\right) \\ poss\left(b_k\right) \end{pmatrix}}$$

Where $n_{b_k \to b_k}$ is the number of edges in *N* shared between the genes in b_k; and for any given set of vertices, *X*, *poss(X)* is the number of edges if *X* were completely connected, i.e.

$$poss\left(X\right) = \frac{|X|\left(|X|-1\right)}{2}$$

As a simple comparison, the association p-values for all network types were pooled together and compared using two-sided Wilcoxon's non-parametric rank tests. We direct the reader to Figure 7.25Figure 7.30, as well as Table 7.29 (FD-MScM) and Table 7.37 (EO-MScM) for the results of these comparisons.

3.4.1.4 Motif E-values

Motif E-values were generated by MEME, the motif discovery tool used by cMonkey (Bailey and Elkan 1994). MEME uses a metric, called an E-value, which was first described by Hertz and Stormo (Hertz and Stormo 1999) with the aim to assess the statistical significance of the information content (or relative entropy) of a

sequence motif, defined as (Stormo 2000). Thus, for a given motif, an E-value is an estimate of the expected number of motifs of the same length that have the same or greater information content as the motif being considered. The E-value can be interpreted as the score for a one-sided p-value for the null distribution of information content for motifs of a given length. Therefore, the larger the E-value of a motif, the less significant it is; the smaller the E-value, the more significant it is. As a simple comparison, the E-values from the (bi)clusters produced by each method were pooled and compared with each other using two-sided Wilcoxon's non-parametric rank tests. In this case, we selected the first motif identified by MEME for the (bi)clusters (as these are generally the most reliable). We direct the reader to Figure 7.31-Figure 7.36, as well as Table 7.31 (FD-MScM) and Table 7.38 (EO-MScM) for the results of these comparisons.

3.4.1.5 Sequence p-values

In addition to the motif E-values, we also compared the distributions of the sequence p-values that were returned by MAST, the motif search utility used by cMonkey (Bailey and Gribskov 1998). Briefly, sequence p-values are an estimate of the significance of a sequence's match to one or more motifs, and can be understood to be a measure of the likelihood of a random sequence having as good or better match or matches. For a given sequence and motif, the motif's PSSM is used to score the degree of the match (likelihood of a match) to a sliding window across the length of the sequence, with the maximal match selected as sequence's score for that motif. The

p-value reported by MAST, then, is simply this score if working with a single motif, and in the case of multiple motifs, it is the multiplication (or addition if using log-likelihoods) of the individual motifs match score to the sequence. To compare each optimization, then, we calculated the average p-value for the genes in each bi(cluster) with respect to the bi(cluster's) associated motifs, and compared the distributions of these. We direct the reader to Figure 7.37-Figure 7.39, as well as Table 7.33 (FD-MScM) and Table 7.39 (EO-MScM) for these comparisons.

3.4.2 Multi-species k-means and balanced multi-species k-means

For comparison we also re-implemented a simple multi-species k-means method (MSKM) similar to the method used in Herschkowitz *et al.* to compare human and mouse microarray data (Herschkowitz, Simin et al. 2007). In this simple method, only the reciprocal best Blast matches are selected as orthologous pairs. These one-to-one pairwise relationships are first used to form a concatenated expression matrix, so that a row in this matrix corresponds to the concatenation of the expression data for 2 orthologous genes. This concatenated expression matrix is next clustered using k-means, using the Euclidean distance metric and with k=150 (as this was the same size used for the test of the multi-species cMonkey method) to generate what we will call shared k-means clusters. Next, as a modification to Herschkowitz's shared k-means algorithm, we added a subsequent step, similar to the elaboration step of the multi-species cMonkey algorithm. In this step, the components of the shared k-means centroids are separated by organism (into the components that correspond to the

organism-specific conditions of the concatenated expression dataset). For each organism, then, the organism-specific shared k-means (sub-)centroids are used to perform a Voronoi partitioning of that organism's non-orthologous core expression data. Thus, in this step, the orthologous genes that belonged to the original shared k-means clusters remain in their original cluster.

As our comparisons indicated that MSKM is prone to allowing an organism to dominate the analysis if its expression data has far more conditions than the other, we also implemented a *balanced* version of the multi-species k-means algorithm (BMSKM). There are a number of ways this balancing could be implemented. One would be to use individual weights for the different conditions from the different species. Another, even simpler implementation, which we used, is to concatenate the smaller dataset to itself so that it has roughly an equivalent number of conditions as the larger dataset, and use this in the MSKM analysis instead. For example, when *B. anthracis*, with 51 conditions in its expression data, was paired with *B. subtilis*, which has >300 conditions, a new dataset for *B. anthracis* was generated that contained the original *B. anthracis* dataset concatenated it to itself 5 times, so that there were 6 copies of each condition. This analysis was not applied to this pairing of *B. anthracis* and *L. monocytogenes* as their expression datasets are roughly equivalent in size. In the case of the Gram-negative triplet, BMSKM was not performed for the pairing involving *E. coli* and *V. cholerae* as they had nearly the same number of conditions, but was applied to the other 2 pairings.

3.4.3 Multi-species Iterative Signature Algorithm

We re-implemented a multi-species version of the Iterative Signature Algorithm (ISA) described by Bergmann et al (Bergmann, Ihmels et al. 2003), using the isa2 package for R (Bergmann, Ihmels et al. 2003; Csardi 2010), available from CRAN. A more thorough discussion of the MSISA method can be found in the supplement, but as a quick review of the method, MSISA contains five main steps:

1. A well-characterized organism is used as a 'reference' organism, with a less characterized organism as the 'target' organism (note, we use the terminology of a later paper from the same group (Ihmels, Bergmann et al. 2005) which employs a similar strategy for multi-species comparisons).
2. Using a pre-generated set of biclusters from the reference organism, biclusters containing genes that have putative orthologs in the target organism are selected and used to generate 'homologous' biclusters for the target organism that contain these putative orthologs such that there is a direct one-to-one mapping between the biclusters for both organisms.
3. Standard, single-species ISA is performed on the target organism, using only these homologous biclusters as seeds.
4. The intersection of the input to and results from step 3 are selected to generate a set of 'purified' biclusters in order to select only the conserved genes in the reference organism.

5. In the final step, single-species ISA is run again on each organism, but using the purified biclusters to generate a set of 'refined' biclusters for each organism. As such, this step is similar to the elaboration step of MScM as it is allows species-specific modifications to be added to the purified bicluster.

For combinatoric reasons, MSISA was only applied to the pairings involving the respective model organism of each triplet. For example, with the gram-positive triplet, MSISA was only applied to the pairings that involved *B. subtilis*, using *B. subtilis* as the reference organism as it is the best studied organism of the three we consider in this study. Hence there are no MSISA results to report for the pairing of *B. anthracis* with *L. monocytogenes*, nor any for the *S. typhimurium* and *V. cholerae* pairing.

3.4.4 External tools used

Coalesce was downloaded and compiled from the Sleipnir library that is available from the published website (Huttenhower, Mutungu et al. 2009). In all cases, Coalesce was run with the default parameters. Similarly, QUBIC was retrieved and compiled from source code, which is available from (Li, Ma et al. 2009). QUBIC was run with the default parameters for continuous data in the case of the Gram-positive triplet, and for the organisms in the Gram-negative triplet, it was run, using 10 for the number of ranks (i.e. "-r 5").

3.5 References

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bailey, T. L. and M. Gribskov (1998). "Combining evidence using p-values: application to sequence homology searches." Bioinformatics **14**(1): 48-54.

Bergmann, S., J. Ihmels, et al. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." Phys Rev E Stat Nonlin Soft Matter Phys **67**(3 Pt 1): 031902.

Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.

Csardi, G. (2010). "isa2: The Iterative Signature Algorithm. R package version 0.2.1." from http://cran.r-project.org/web/packages/isa2/index.html.

Herschkowitz, J. I., K. Simin, et al. (2007). "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors." Genome Biol **8**(5): R76.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7-8): 563-577.

Huttenhower, C., K. T. Mutungu, et al. (2009). "Sleipnir Library." from http://function.princeton.edu/sleipnir.

Ihmels, J., S. Bergmann, et al. (2005). "Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program." PLoS Genet **1**(3): e39.

Li, G., Q. Ma, et al. (2009). "CSBL Biclustering." from http://csbl.bmb.uga.edu/~maqin/bicluster/.

Li, G., Q. Ma, et al. (2009). "QUBIC: a qualitative biclustering algorithm for analyses of gene expression data." Nucleic Acids Res **37**(15): e101.

Prelic, A., S. Bleuler, et al. (2006). "A systematic comparison and evaluation of biclustering methods for gene expression data." Bioinformatics **22**(9): 1122-1129.

Reiss, D. J., N. S. Baliga, et al. (2006). "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." BMC Bioinformatics **7**: 280.

Stein, B., S. M. Eissen, et al. (2003). On Cluster Validity and the Information Need of Users. Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03), Benalmádena, Spain. M. H. Hanza, ACTA Press**:** 216-221.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

4. MULTI-PLATFORM, MULTI-SPECIES BICLUSTERING OF HUMAN AND MOUSE HEMATOPOIETIC CELL DATA

In this chapter we present initial results from a comparative analysis of human and mouse hematopoietic cell expression data (with a focus on immune system cells) that was performed using an early, experimental version of a new *multi-platform* version of the multi-species cMonkey algorithm. The results presented below are intended to be a part of two larger, multi-lab collaborations, one to infer the global regulatory network governing the T helper 17 (Th17) cell lineage; the other the global regulatory network governing Burkitt's lymphoma.

4.1 Introduction

Many of the same reasons that made leukemia an attractive target for Dr. Sidney Farber's earliest forays in the 1940s into cancer and the development of chemotherapy also make it well-suited for systems' biology analysis today (Mukherjee 2010). Chief amongst these are the relative ease-of-access one has to hematopoietic cell samples in comparison to those from other tissue types. In addition, flow cytometry also allows one to more easily isolate specific cell sub-types or lineages. At the broadest level, hematopoietic cell lineages are classified into being either a myeloid or lymphoid cell lineage. The myeloid cell lineages includes erythrocytes and other cell lineages that are primarily involved in the innate immune response (i.e. neutrophils, monocytes, basophils, etc.); and the lymphoid cell lineages primarily contains cell lineages involved in the adaptive immune response (T- and B-

cells), though, it also contains natural killer (NK) cells, which are part of the innate immune system. All these various cell lineages stem from a common class of multipotent hematopoietic stem cells (or HSCs) that undergo a complex, non-reversible differentiation process that is driven by a combination of external and transcriptional signaling (for more complete reviews of this process, see (Iwasaki and Akashi 2007; Orkin and Zon 2008; Kaushansky 2010)).

This complexity of the differentiation process is yet other reason for why the immune system is well-suited for systems biology analysis. As this brief introduction is not intended to be a comprehensive review of this prior work, we direct the reader to (Gardy, Lynn et al. 2009; Germain, Meier-Schellersheim et al. 2011) for discussions of prior systems biology analyses of the immune system. Most recently, an extensive analysis of 38 different cell lineages from varying steps in the human hematopoietic differentiation process was recently published (Novershtern, Subramanian et al. 2011). This analysis was able to characterize a number of different modules of genes that were differentially expressed by the different cell lineages they considered. It also inferred two putative regulatory networks, with one based solely on expression data, while the other was based on sequence data alone, though, it did use the modules from the expression analysis for gene sets during the analysis.

Another recent study (Painter, Davis et al. 2011) performed a more restricted analysis that compared expression data from mouse T and B cell data to identify

differentially expressed genes between the 2 cell lineages and also develop differential signatures for each cell type. From a technical perspective, one of the most interesting aspects of the data set that was analyzed for this project is that it was a multi-platform data set that was collected using multiple platforms, including whole-genome arrays from Agilent, Affymetrix and Nimblegen. Also integrated into this data set was data that is publicly available from the Immunological Genome Project, (hereafter ImmGen) (Heng and Painter 2008). Finally, while the focus of this study was upon the B and T cell lineages, the signatures that were developed for each were then compared to a larger compendium of expression data from other mouse immune cell lineages that are available in the ImmGen data set.

However, neither of these studies employed a comparative approach, nor did they effectively integrate multiple data types, except for the *a posteriori* manner in which the gene sets from the expression analysis were used during the sequence-based network inference of the first project (Novershtern, Subramanian et al. 2011). Furthermore, in the case of the multi-platform analysis of the mouse T and B cell lineages (Painter, Davis et al. 2011), the analysis required that the integrated data set only contain those genes in the intersection of the 4 platforms that were included, thus only 12000 genes were contained in the final data set. In contrast, below, we will present preliminary results from a prototype of a multi-platform, multi-species version of cMonkey that was used to perform a comparative analysis of human and mouse

immune system cell data from multiple platforms without loss of data during the integration.

4.2 Materials and Methods

4.2.1 Data sets analyzed

Below, we provide detailed descriptions of the data sets that were generated for both human and mouse. Summary information can be found in Table 4.1.

4.2.1.1 Expression data

The primary source for the mouse immune cell expression data was the public repository of expression data that is provided by the ImmGen project (Heng and Painter 2008), which consists of 508 different samples from 14 different terminal lineages and numerous intermediate lineages that were measured using the Affymetrix Mouse Gene 1.0 ST Array. An additional 61 RNAseq conditions examining the Th17 cell lineage were generated from various time series and knock down or knock out studies of key transcription factors during the differentiation from naïve CD4+ cell to Th17.

The human expression data is composed of a heterogeneous collection 532 samples from nearly 20 studies that can be classified with 3 different, general categories, including explicit hematopoietic differentiation studies (Lee, Hanspers et al. 2004; Dybkaer, Iqbal et al. 2007; Elo, Jarvenpaa et al. 2010; Filen, Ylikoski et al. 2010; Novershtern, Subramanian et al. 2011; Prots, Skapenko et al. 2011), immune

response (Buzzeo, Yang et al. 2007; Martinez-Llordella, Puig-Pey et al. 2007; Dower, Ellis et al. 2008; Grumann, Scharf et al. 2008; Radom-Aizik, Zaldivar et al. 2008; Woszczek, Chen et al. 2008; Li, Sze et al. 2010; Yu, Hu et al. 2010), and disease and other general profiling studies (Kim, Tchernyshyov et al. 2006; Piccaluga, Agostinelli et al. 2007; Mosig, Rennert et al. 2008; Abbas, Wolslegel et al. 2009; Longo, Lugar et al. 2009). In total, the human expression data contained 136 conditions that were assayed using the Affymetrix Human Genome U133 Plus 2.0 Array, 185 that were assayed with the Affymetrix Human Genome U133A Array, and 211 samples that were assayed with the Affymetrix GeneChip HT-HG_U133A Early Access Array (hereafter referred to as the U133+2, U133A and U133AofA arrays, respectively).

All microarray data was downloaded from the NCBI Gene Expression Omnibus (GEO) database (Edgar, Domrachev et al. 2002; Barrett, Troup et al. 2007) as raw .CEL files and normalized with RMA, using the Bioconductor suite of bioinformatics tools (Gentleman, Carey et al. 2004). We also emphasized that in all cases, with the exception of the specialized U133AofA array, the latest custom CDF probe mappings that were generated by (Dai, Wang et al. 2005) were used when processing the raw .CEL data as two recent reviews have indicated these are more accurate than the original probe mapping provided by the manufacturer (Sandberg and Larsson 2007; Mieczkowski, Tyburczy et al. 2010). This custom CDF also has the added advantage of providing probe sets that have a strict one-to-one mapping between genes and probe sets. In so doing, this avoids the need to merge probe sets which map to a single gene

as was necessary with the specialized HT-HG_U133A Early Access Array for which no custom CDF was available. We acknowledge that a strict one-to-one mapping is an over-simplification which ignores potential protein isoforms, but this is a problem common to all oligo-based arrays. Finally, we should clarify that RMA was applied to only those samples that belong to a common platform, independent of those that were generated with the others (i.e. all samples from the U133A arrays were normalized together with RMA). Integration of these different platform-specific data sets is explained below.

In order to generate the RNAseq data, all RNAseq samples were sequenced using Illumina sequencer (Illumina Hiseq-2000), 36bp single ends with fragment size of 225bp for library preparation. All reads were aligned to the mouse genome, version v.mm9, using Bowtie (Langmead, Trapnell et al. 2009). RNAseq reads per gene were quantified using Cufflinks (Roberts, Trapnell et al. 2011) to determine the expression levels, measured by reads per kilobase per million (RPKM). All genes with a median expression over the RNAseq samples that were less than 5 were excluded from consideration, and the final set was log-transformed to allow it to be compared with the data from the microarray platforms.

For both organisms, samples from the various platforms that were included in this study were integrated into a single "meta-expression" matrix using the following simple, two-step strategy. In the first step, each platform-specific data set was row (gene) normalized to have a mean of 0 with a standard deviation of 1 in order to

prevent any platform-specific biases from impacting the other platform-specific data sets. In the second, each platform-specific data set was merged into a "meta-matrix" whose dimensions were determined by the union of both the genes and conditions. In the cases where a particular platform-specific data set lacked a given gene, NA's (null values in R) were inserted into the matrix. As the genomic coverage between some of the platforms was considerable this had the effect of generating a data matrix with large blocks of NA's. For example, in the human data, the U133+2 array has a coverage of greater than 18,000 genes, while the U133A has a coverage of just over 12,000 genes, and the U133AofA array has a coverage of over 13,000 genes.

While the intersection of these gene sets is considerable, with nearly 11,000 genes, we wanted to avoid the loss of information by filtering the matrix to consider only these genes. There are additional considerations within a comparative environment as well as we will elaborate upon further when describing our new multi-platform version of multi-species cMonkey. In addition to not limiting the analysis to only those genes in the intersection, we also did not attempt to impute values for these large blocks of NA's as the size of these blocks was too large for any such effort to be meaningful. Nor did we try to inject random values as we did not want to risk allowing these to skew the analysis one way or the other.

4.2.1.2 Association data

Networks for both organisms were retrieved from multiple databases including Bind (Bader, Donaldson et al. 2001), BioGRID (Stark, Breitkreutz et al. 2006;

Breitkreutz, Stark et al. 2008; Stark, Breitkreutz et al. 2011), DIP (Xenarios, Rice et al. 2000; Xenarios, Fernandez et al. 2001; Xenarios, Salwinski et al. 2002; Salwinski, Miller et al. 2004), HPRD , InnateDB (Lynn, Chan et al. 2010), IntAct (Hermjakob, Montecchi-Palazzi et al. 2004; Kerrien, Alam-Faruque et al. 2007; Aranda, Achuthan et al. 2010) , InteroPORC (Michaut, Kerrien et al. 2008), MatrixDB (Chautard, Ballut et al. 2009), MINT (Zanzoni, Montecchi-Palazzi et al. 2002; Chatr-aryamontri, Ceol et al. 2007; Ceol, Chatr Aryamontri et al. 2010), Reactome (Matthews, Gopinath et al. 2009; Croft, O'Kelly et al. 2011), STRING (Jensen, Kuhn et al. 2009), iRefIndex (Razick, Magklaras et al. 2008), and the Pathway Commons (Cerami, Gross et al. 2011). We note that several of these are "meta pathway databases" as they include associations from other databases, many of which we downloaded associations from directly. The other source databases that contributed edges via these meta databases which we have not already been listed include CORUM (Ruepp, Brauner et al. 2008), MPact (Guldener, Munsterkotter et al. 2006), MPPI (Pagel, Kovac et al. 2005) and OPHID (Brown and Jurisica 2005).

As cMonkey does not use weighted graphs with its association data, all associations from these various databases were classified as being either high or low confidence associations, and split into separate networks that were assigned weights in the scoring function that reflected these respective confidences. When making these high or low confidence assignments, the method by which a given interaction was determined was the primary determinant, with electronically inferred associations (i.e.

interologs, text mining) compromising the entirety of the low-confidence associations. To avoid double-counting when calculating the network score, all duplicate edges were removed so that no two genes could share more than one association.

Table 4.1: Size of the data sets used for the human and mouse immune system analysis, by organism.

Number of:	Human			Mouse	
	U133+2	U133A	U133 AofA	ImmGen	RNA-seq
genes	18107	12060	13276	21124	8966
intersection		10069		8514	
total genes		21096		21634	
conditions	136	185	211	508	61
total conditions		532		569	
association edges:					
high-confidence:		261641		9521	
low-confidence:		838743		373604	
Ortholog pairs:			15737		

4.2.1.3 Putative Orthology Predictions

All putative orthology predictions between human and mouse were retrieved from the Mouse Genome Database (MGD) (Blake, Bult et al. 2011), which provides a comprehensive list of nearly 17850 orthology predictions that are produced via both

manual curation and electronic inference from the HomoloGene database provided by the NCBI (Wheeler, Barrett et al. 2006). While HomoloGene allows for the identification of paralogous relationships, this list only contains one-to-one matches between the two genomes. As some of this list of nearly 17850 orthology relationships that MGD provides includes genes from outside of our data sets, this yielded a total of 15737 orthologous pairs for our analysis.

4.2.2 Multi-platform, multi-species cMonkey

4.2.2.1 Motivation

As described in section 4.2.1.1, both the mouse and human data have a similar "blocky" nature of the matrices, meaning that both contain large blocks of unobserved values that are the result of the merging of data sets with highly different gene coverage. As mentioned above, a simple filtering strategy where only those genes in the intersection of the different platforms included in this study are considered would not be appropriate in a comparative context, as we illustrate in Figure 4.1. In this simplified example, there are two organisms, each with a multi-platform "meta-expression" matrix similar to the one being analyzed, where one of the platforms has considerably greater genomic coverage than the other. In this figure, then, it is easy to see that in each organism's respective expression data, there are 3 possible classes of genes – 1) those in the intersection, 2) those that are only represented by the platform with larger genomic coverage, and a 3) far smaller set of genes that are only represented by the platform with the smaller coverage. The addition of other

platforms to a matrix only further complicates the number of sets of genes that are possible.

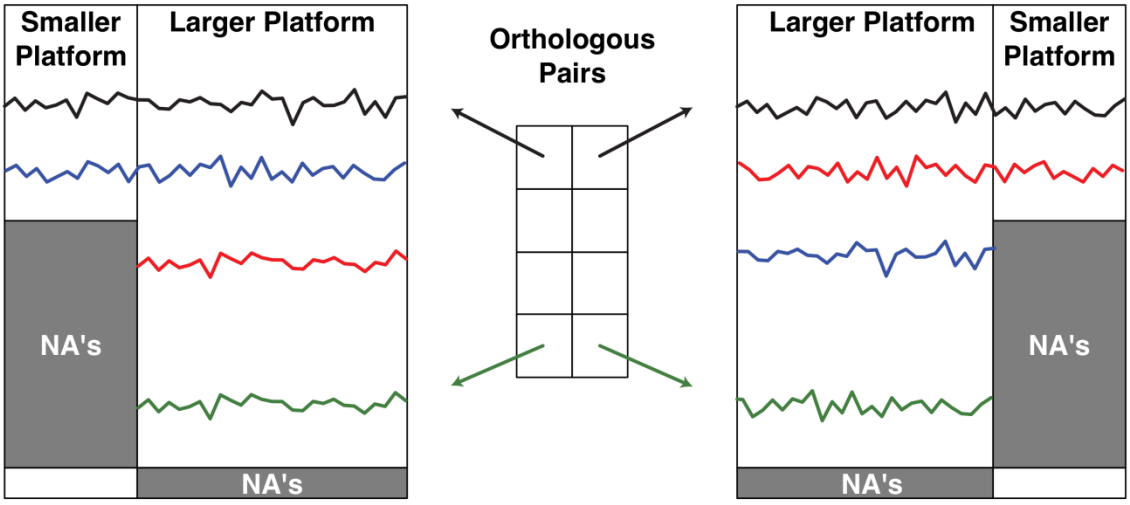


Figure 4.1: Demonstration of the 4 major classes of orthologous pairs that are possible in a multi-platform comparative analysis. The first class, represented by the black orthologous pairs, corresponds to those genes that are represented by all platforms in the matrix. Similarly, the green orthologs correspond to those genes that are only represented in the platforms with larger genomic coverage. Finally, the blue and red orthologs correspond to the other 2 combinations of these that are possible. **Table 4.2** displays the number of orthologs which correspond to these sets in the human and mouse immune system expression data.

Continuing the example, with 3 sets of genes in each organism's expression matrix possible, this translates into a total of 9 possible classes of orthologous pairs that are possible – with the majority of these being contained in the 4 classes that correspond to the largest sets of genes in the respective expression matrices. Thus, a simple strategy that includes only those genes that are in the intersection of the different

platforms will likely translate into a considerable loss of biological information. In the case of the human and mouse immune system data sets, this filtering strategy results in fewer than 5000 ortholog pairs that are available – out of the nearly 16000 that are possible (Table 4.2). Given the significant loss of information that is possible when one only includes those genes which are represented on all platforms, we elected to merge the data using the strategy described in section 4.2.1.1.

Table 4.2: Number of orthologs that corresponds to the four major classes of genes in the human and mouse immune system expression data.

Human gene class	number of orthologs	Mouse gene class
Intersection:	4997	**Intersection:**
ImmGen only:	4110	**Intersection:**
ImmGen only:	2890	**U133+2 only:**
Intersection:	2629	**U133+2 only:**
Others:	1111	**Others:**

We were surprised to discover, however, that such a simple change in the expression matrix (the allowance of large "blocks" of unobserved values) in the data posed a number of challenges to MScM, which – like many other methods – was designed with the assumption of a single, common platform for all samples. Figure 4.2 provides an example of one such challenge which can occur in the single- and multi-species analyses of a multi-platform expression matrix. In this example, we display a snapshot after several iterations of the optimization of a bicluster that

contains conditions in both the low- and high-coverage platforms that at one point contained 8 genes, where 4 are represented in both platforms, while 4 are not. At this iteration, we see that 3 of the 4 genes that are represented in both platforms have been removed from the bicluster as a result of the Monte Carlo optimization, leaving only a single gene represented in those conditions from the low-coverage platform. Having so few genes in these conditions causes a number of both technical and statistical issues, and is best avoided.

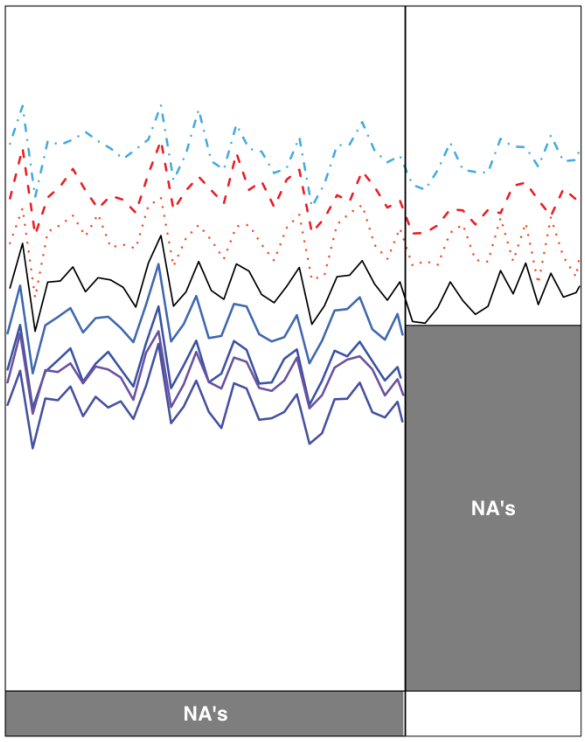


Figure 4.2: Example of the error states that are possible with a Monte Carlo search strategy with a multi-platform data set.

To address this issue, there are several possible approaches. The simplest of these is to prune out during the optimization any conditions from the bicluster that have an insufficient number of genes that are represented in them. Experiments with this approach indicated that this often causes a sudden and dramatic reduction in the number of conditions that are included in a bicluster, resulting in a considerable loss of coverage. A second option would be to use a different scoring function for the expression data. While this still remains a valid option, it is unclear whether this will introduce other issues. Instead, we opted to update the search strategy that MScM employs to avoid these issues entirely, which we present below.

4.2.2.2 Algorithm overview

To work in this space, we first define the concept of a set of "orthologous basis pairs," which is simply one of the classes of orthologous pairs described above – for example, the class of orthologous pairs which correspond to those genes that are represented in all platforms of both organisms' expression matrices. Using these *basis pairs*, we split the shared step of the multi-species method into 2 sub-steps. During the first step, we limit the analysis to a single set of basis pairs such that we only seed and optimize a bicluster using the orthologous pairs in that basis pair set. This allows us to avoid those cases where after several steps of the iteration, there are conditions with fewer than 2 genes. We call this part of the optimization the "basis-pair step" - which is analogous to the shared step of the multi-species optimization.

The goal of the basis-pair step is to establish a bicluster with sufficient enough data support that it can be used to anchor a search within the complete set of orthologous pairs. We call this second optimization the "augment" step - which is analogous to the elaboration step of the multi-species method. To ensure that we avoid those cases where the bicluster conditions lack a sufficient number of bicluster genes that have valid values for them, we require that a minimum of 3 genes and 10 conditions from the original basis-pair step bicluster remain in the bicluster as it is being augmented.

4.3 Results

Results are currently preliminary as the multi-platform extension of cMonkey is still experimental. Currently we have 500 shared biclusters generated, with 250 generated from the two basis pair sets that correspond to genes in the mouse RNAseq data (Table 4.2). Note there are no elaborated biclusters at this point.

Evaluation is ongoing, however encouraging as the new multi-platform, multi-species cMonkey method (MPMScM hereafter) has identified several biclusters of interest. For example, one of these, bicluster 31, includes a number of genes involved in hematopoietic and lymphoid organ development as well as B cell receptor signaling (Figure 7.85 and Table 7.42). It's interesting as well because MPMScM also identified a major histocompatibility complex, class II (MHC, class II) module, bicluster 87 (Figure 7.86 and Table 7.43), which contains genes that are all also part of bicluster just mentioned (bicluster 31), and which we suspect is a sub-process of that

larger module. Additional modules of interest that MPMScM identified include an innate immune response module that contains genes involved in Gram-positive bacterial response, phagocytosis, and inflammatory response several of which are Toll-like receptor genes (bicluster 2, see Figure 7.87 and Table 7.44). Last, it also identified an adaptive immune response module that contains genes involved in T-cell differentiation and T-cell immune response, bicluster 480 (Figure 7.88 and Table 7.45).

4.4 Future Directions

As future steps that may be explored when working in this space, there are several ideas that may help with the analysis in this space. The first would make a small, but possibly significant modification during the augment step where we further pre-seed the basis pair biclusters with those orthologous pairs outside of the basis pair set that correlate best with the mean expression of the bicluster. In the second, we would consider each component of the meta-expression matrix separately in a scheme where each component would have its' own expression-component-specific weight within the joint log-likelihood. In so doing, this would allow researchers to specify weights for each platform that reflected their confidence in its accuracy, as is currently allowed by cMonkey with the different sources of association data.

4.5 References

Abbas, A. R., K. Wolslegel, et al. (2009). "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus." PLoS ONE **4**(7): e6098.

Aranda, B., P. Achuthan, et al. (2010). "The IntAct molecular interaction database in 2010." Nucleic Acids Res **38**(Database issue): D525-531.

Bader, G. D., I. Donaldson, et al. (2001). "BIND--The Biomolecular Interaction Network Database." Nucleic acids research **29**(1): 242-245.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: Mining tens of millions of expression profiles - Database and tools update." Nucleic Acids Research **35**(SUPPL. 1): D760-D765.

Blake, J. A., C. J. Bult, et al. (2011). "The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics." Nucleic Acids Research **39**(Database issue): D842-848.

Breitkreutz, B. J., C. Stark, et al. (2008). "The BioGRID Interaction Database: 2008 update." Nucleic acids research **36**(Database issue): D637-640.

Brown, K. R. and I. Jurisica (2005). "Online predicted human interaction database." Bioinformatics **21**(9): 2076-2082.

Buzzeo, M. P., J. Yang, et al. (2007). "Hematopoietic stem cell mobilization with G-CSF induces innate inflammation yet suppresses adaptive immune gene expression as revealed by microarray analysis." Experimental hematology **35**(9): 1456-1465.

Ceol, A., A. Chatr Aryamontri, et al. (2010). "MINT, the molecular interaction database: 2009 update." Nucleic Acids Res **38**(Database issue): D532-539.

Cerami, E. G., B. E. Gross, et al. (2011). "Pathway Commons, a web resource for biological pathway data." Nucleic Acids Res **39**(Database issue): D685-690.

Chatr-aryamontri, A., A. Ceol, et al. (2007). "MINT: the Molecular INTeraction database." Nucleic Acids Res **35**(Database issue): D572-574.

Chautard, E., L. Ballut, et al. (2009). "MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions." Bioinformatics **25**(5): 690-691.

Croft, D., G. O'Kelly, et al. (2011). "Reactome: a database of reactions, pathways and biological processes." Nucleic Acids Res **39**(Database issue): D691-697.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Research **33**(20): e175.171-e175.179.

Dower, K., D. K. Ellis, et al. (2008). "Innate immune responses to TREM-1 activation: overlap, divergence, and positive and negative cross-talk with bacterial lipopolysaccharide." Journal of immunology **180**(5): 3520-3534.

Dybkaer, K., J. Iqbal, et al. (2007). "Genome wide transcriptional analysis of resting and IL2 activated human natural killer cells: gene expression signatures indicative of novel molecular signaling pathways." BMC Genomics **8**: 230.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Research **30**(1): 207-210.

Elo, L. L., H. Jarvenpaa, et al. (2010). "Genome-wide profiling of interleukin-4 and STAT6 transcription factor regulation of human Th2 cell programming." Immunity **32**(6): 852-862.

Filen, S., E. Ylikoski, et al. (2010). "Activating transcription factor 3 is a positive regulator of human IFNG gene expression." Journal of immunology **184**(9): 4990-4999.

Gardy, J. L., D. J. Lynn, et al. (2009). "Enabling a systems biology approach to immunology: focus on innate immunity." Trends in Immunology **30**(6): 249-262.

Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Germain, R. N., M. Meier-Schellersheim, et al. (2011). "Systems biology in immunology: a computational modeling perspective." Annual review of immunology **29**: 527-585.

Grumann, D., S. S. Scharf, et al. (2008). "Immune cell activation by enterotoxin gene cluster (egc)-encoded and non-egc superantigens from Staphylococcus aureus." Journal of immunology **181**(7): 5054-5061.

Guldener, U., M. Munsterkotter, et al. (2006). "MPact: the MIPS protein interaction resource on yeast." Nucleic Acids Res **34**(Database issue): D436-441.

Heng, T. S. and M. W. Painter (2008). "The Immunological Genome Project: networks of gene expression in immune cells." Nature immunology **9**(10): 1091-1094.

Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "IntAct: an open source molecular interaction database." Nucleic Acids Res **32**(Database issue): D452-455.

Iwasaki, H. and K. Akashi (2007). "Hematopoietic developmental pathways: on cellular basis." Oncogene **26**(47): 6687-6696.

Jensen, L. J., M. Kuhn, et al. (2009). "STRING 8--a global view on proteins and their functional interactions in 630 organisms." Nucleic Acids Res **37**(Database issue): D412-416.

Kaushansky, K. (2010). Hematopoietic Stem Cells, Progenitors, and Cytokines. Williams hematology. L. MA, K. TJ, S. U, Kaushansky K and P. JT. New York, McGraw-Hill Medical**:** xxiii, 2439 p.

Kerrien, S., Y. Alam-Faruque, et al. (2007). "IntAct--open source resource for molecular interaction data." Nucleic Acids Res **35**(Database issue): D561-565.

Kim, J. W., I. Tchernyshyov, et al. (2006). "HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia." Cell metabolism **3**(3): 177-185.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome biology **10**(3): R25.

Lee, M. S., K. Hanspers, et al. (2004). "Gene expression profiles during human CD4+ T cell differentiation." International immunology **16**(8): 1109-1124.

Li, J., D. M. Sze, et al. (2010). "Clonal expansions of cytotoxic T cells exist in the blood of patients with Waldenstrom macroglobulinemia but exhibit anergic properties and are eliminated by nucleoside analogue therapy." Blood **115**(17): 3580-3588.

Longo, N. S., P. L. Lugar, et al. (2009). "Analysis of somatic hypermutation in X-linked hyper-IgM syndrome shows specific deficiencies in mutational targeting." Blood **113**(16): 3706-3715.

Lynn, D. J., C. Chan, et al. (2010). "Curating the innate immunity interactome." BMC Syst Biol **4**: 117.

Martinez-Llordella, M., I. Puig-Pey, et al. (2007). "Multiparameter immune profiling of operational tolerance in liver transplantation." American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons **7**(2): 309-319.

Matthews, L., G. Gopinath, et al. (2009). "Reactome knowledgebase of human biological pathways and processes." Nucleic Acids Res **37**(Database issue): D619-622.

Michaut, M., S. Kerrien, et al. (2008). "InteroPORC: automated inference of highly conserved protein interaction networks." Bioinformatics **24**(14): 1625-1631.

Mieczkowski, J., M. E. Tyburczy, et al. (2010). "Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements." BMC Bioinformatics **11**: 104.

Mosig, S., K. Rennert, et al. (2008). "Monocytes of patients with familial hypercholesterolemia show alterations in cholesterol metabolism." BMC medical genomics **1**: 60.

Mukherjee, S. (2010). The emperor of all maladies : a biography of cancer. New York, Scribner.

Novershtern, N., A. Subramanian, et al. (2011). "Densely interconnected transcriptional circuits control cell states in human hematopoiesis." Cell **144**(2): 296-309.

Orkin, S. H. and L. I. Zon (2008). "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology." Cell **132**(4): 631-644.

Pagel, P., S. Kovac, et al. (2005). "The MIPS mammalian protein-protein interaction database." Bioinformatics **21**(6): 832-834.

Painter, M. W., S. Davis, et al. (2011). "Transcriptomes of the B and T lineages compared by multiplatform microarray profiling." Journal of immunology **186**(5): 3047-3057.

Piccaluga, P. P., C. Agostinelli, et al. (2007). "Gene expression analysis of peripheral T cell lymphoma, unspecified, reveals distinct profiles and new potential therapeutic targets." The Journal of clinical investigation **117**(3): 823-834.

Prots, I., A. Skapenko, et al. (2011). "Analysis of the transcriptional program of developing induced regulatory T cells." PLoS ONE **6**(2): e16913.

Radom-Aizik, S., F. Zaldivar, Jr., et al. (2008). "Effects of 30 min of aerobic exercise on gene expression in human neutrophils." Journal of applied physiology **104**(1): 236-243.

Razick, S., G. Magklaras, et al. (2008). "iRefIndex: a consolidated protein interaction database with provenance." BMC Bioinformatics **9**: 405.

Roberts, A., C. Trapnell, et al. (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome Biology **12**(3): R22.

Ruepp, A., B. Brauner, et al. (2008). "CORUM: the comprehensive resource of mammalian protein complexes." Nucleic Acids Res **36**(Database issue): D646-650.

Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-451.

Sandberg, R. and O. Larsson (2007). "Improved precision and accuracy for microarrays using updated probe set definitions." BMC Bioinformatics **8**: 48.

Stark, C., B. J. Breitkreutz, et al. (2011). "The BioGRID Interaction Database: 2011 update." Nucleic acids research **39**(Database issue): D698-704.

Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic acids research **34**(Database issue): D535-539.

Wheeler, D. L., T. Barrett, et al. (2006). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **34**(Database issue).

Woszczek, G., L. Y. Chen, et al. (2008). "Leukotriene D(4) induces gene expression in human monocytes through cysteinyl leukotriene type I receptor." The Journal of allergy and clinical immunology **121**(1): 215-221 e211.

Xenarios, I., E. Fernandez, et al. (2001). "DIP: The Database of Interacting Proteins: 2001 update." Nucleic Acids Res **29**(1): 239-241.

Xenarios, I., D. W. Rice, et al. (2000). "DIP: the database of interacting proteins." Nucleic Acids Res **28**(1): 289-291.

Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucleic Acids Res **30**(1): 303-305.

Yu, W. H., H. Hu, et al. (2010). "Bioinformatics analysis of macrophages exposed to Porphyromonas gingivalis: implications in acute vs. chronic infections." PLoS ONE **5**(12): e15613.

Zanzoni, A., L. Montecchi-Palazzi, et al. (2002). "MINT: a Molecular INTeraction database." FEBS Lett **513**(1): 135-140.

5. COMPARATIVE MICROBIAL MODULES RESOURCE: GENERATION AND VISUALIZATION OF MULTI-SPECIES BICLUSTERS

Co-written with: Thadeous Kacmarczyk[1], Ashley R Bate[1], Patrick Eichenberger[1], Richard Bonneau[1,2,3]

1 – Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, 10003, USA

2 – Computer Science Department, Courant Institute for Mathematical Sciences, New York University, New York, NY, 10003, USA

3 – Computational Biology Program, New York University, New York, NY, 10003, USA

* This author contributed equally to this work.

5.1 Abstract

The increasing abundance of large-scale, high-throughput datasets for many closely related organisms provides opportunities for comparative analysis via the simultaneous (bi)clustering of datasets from multiple species. These analyses require a reformulation of how to organize multi-species datasets and visualize comparative genomics data analyses results. Recently, we developed a method, multi-species cMonkey, which integrates heterogeneous high-throughput datatypes from multiple species to identify conserved regulatory modules (biclusters). Here we present an integrated data visualization system, built upon the Gaggle, enabling exploration of

our method's results (available at http://meatwad.bio.nyu.edu/cmmr.html). The system can also be used to explore other comparative genomics datasets and outputs from other data analysis procedures (e.g. results from other multiple-species clustering programs or from independent clustering of different single-species datasets). We provide an example use of our system for two bacteria, *Escherichia coli* and *Salmonella typhimurium*. We illustrate the use of our system by exploring conserved biclusters involved in nitrogen metabolism, uncovering a putative function for *yjjI*, a currently uncharacterized gene that we predict to be involved in nitrogen assimilation.

5.2 Author Summary

Advancing high-throughput experimental technologies are providing access to genome-wide measurements of multiple information levels (e.g. mRNA, protein, interactions, functional assays, etc.) for multiple related species. We present a biclustering algorithm and an associated visualization system, for generating and exploring regulatory modules derived from analysis of integrated multi-species genomics datasets. We use multi-species-cMonkey, an algorithm of our own construction that can integrate diverse systems-biology datatypes from multiple species to form biclusters (condition-dependent regulatory modules) that are both conserved across the multiple species analyzed and biclusters that are specific to subsets of the processed species. Our resource (an integrated web and java based system) allows biologists to explore both conserved and species-specific biclusters in the context of the data, associated networks for both species, and existing annotations

for both species. Our focus in this work is on the use of the integrated system with examples drawn from exploring modules associated with nitrogen metabolism in two, gram negative bacteria (E. coli and S. typhimurium) for which sufficient genomics data is available.

5.3 Introduction

It is now routine to have genomics data for multiple organisms of interest. For example, data may be available for both an organism of primary relevance to a specific study (perhaps a recently sequenced pathogen), as well as data for related model species (that offer advantages such as having better explored genetics, larger and more complete genomics datasets or ease of use in the lab). Tools and algorithms for comparative analysis of multi-species datasets are therefore in high demand. Comparative analysis of gene sequences is a mainstay in computational biology (Altschul, Madden et al. 1997), but comparative methods for genomics data analysis are relatively new, primarily due to the fact that only recently have researchers had access to large-scale datasets from multiple species (Stuart, Segal et al. 2003; Ihmels, Bergmann et al. 2005; Tanay, Regev et al. 2005; Tirosh, Bilu et al. 2007; Lu, Huggins et al. 2009; Chikina and Troyanskaya 2011). Several recent studies have shown that comparative genomics analysis improves our ability to learn regulatory interactions, co-regulated groups, and to delineate the conserved components of fundamental pathways and modules (Bergmann, Ihmels et al. 2004; Tanay, Sharan et al. 2004; Berg and Lassig 2006; Reiss, Baliga et al. 2006; Waltman, Kacmarczyk et al. 2010; Chikina

and Troyanskaya 2011). In particular, multiple-species clustering and biclustering can be used to detect conserved co-regulated gene groups and serve as a foundation to begin characterizing key differences in the regulatory programs of related species. In this work we present a data visualization system that enables the visualization and exploration of integrative multi-species biclustering analysis. We aim to both present our system and provide a general example of how multi-species datasets can be integrated by coupling new multiple-species biclustering algorithms with a system of visualization tools coordinated across organisms by predicted orthology relationships. Our interface is built on a loosely coupled system architecture that connects multiple tools and databases using the Gaggle (Shannon, Reiss et al. 2006), Sungear (Poultney, Gutierrez et al. 2007), and Cytoscape (Cline, Smoot et al. 2007). This interface provides coordinated access to multiple-species clusters, biclusters and networks derived from comparative genomics analysis tools such as multi-species cMonkey (MScM) (Waltman, Kacmarczyk et al. 2010).

5.3.1 The challenges of visualizing multiple species data

The analysis of multiple species datasets presents several challenges not encountered when analyzing single species datasets. In addition to the display and exploration of multiple datatypes (networks, cis-regulatory sequences and genomic context, transcriptome and proteome data) we add the challenge of tracking connections between orthologous groups of genes. In this work we focus on exploring sets of

multi-species biclusters generated with MScM. A typical multi-species biclustering (set of biclusters) will consist of:

1. The source data used to:
    a. Compute the biclustering (for each species, its protein association networks, upstream sequences and expression data)
    b. Perform post-analytic evaluations, such as enrichment of ontology terms (i.e. GO functions and KEGG pathways)
2. A set of conserved biclusters, i.e. composed of pairs of orthologous genes spanning both species
3. Species-specific elaborations of the conserved biclusters – genes added to conserved biclusters based on evidence in a single species (including genes lacking putative orthologs in the other species) following the initial generation of the conserved core of the bicluster
4. Species-specific biclusters (biclusters composed entirely of genes lacking detectable orthology relationships between the two species)

Our system to navigate this analysis enables exploration of both the conserved biclusters (in the context of both species) and the elaborated portion of biclusters (in the context of each individual species dataset) and illustrates general strategies for building loosely coupled systems for exploring other multi-species genomics analysis.

5.3.2 Data integration across multiple species

High-throughput data exists for many microbial organisms on multiple information levels (i.e. genome sequences, transcriptomics, proteomics, metabolomics, networks of pathways and interactions). Collecting and integrating diverse and heterogeneous datasets from disparate databases is not trivial and poses a number of barriers to automating the process. One of the most significant barriers to automation of data-import is the inconsistency between the naming schemes for loci, mRNA and protein products that are employed by the major public repositories such as NCBI, Uniprot and EMBL. Versioning can also be an issue if a given data source is delayed in updating their annotations. Our resource (described below) integrates diverse data (listed in full detail below) from microarray experiments, genomic sequences, and various functional associations, and uses a database (linked to the Gaggle) to translate gene names (across datatypes and disparate resources) and ortholog names (across species). We will focus our examples on two closely related γ-Proteobacteria: *E. coli* and *S. typhimurium*.

5.3.3 Multi-species Integrated Biclustering

Clustering and biclustering are typically used to identify groups of co-expressed genes that, ideally, represent true regulatory modules and co-functional groups such as pathways and complexes. Biclustering groups genes into condition-specific gene clusters, and can allow genes to participate in more than one bicluster. Many biclustering methods have been previously described, for example, SAMBA

(Tanay, Sharan et al. 2002), QUBIC (Li, Ma et al. 2009), ISA (Ihmels, Bergmann et al. 2004), BIMAX (Prelic, Bleuler et al. 2006), and NNN (Huttenhower, Flamholz et al. 2007), and other algorithms (Cheng and Church 2000; Ben-Dor, Chor et al. 2003; Kluger, Basri et al. 2003; Supper, Strauch et al. 2007; Lu, Huggins et al. 2009). Recent integrative biclustering methods, such as MATISSE (Ulitsky and Shamir 2007), the recent version of SAMBA (Tanay, Sharan et al. 2004), and cMonkey (Reiss, Baliga et al. 2006; Waltman, Kacmarczyk et al. 2010) have shown that incorporating additional datatypes (e.g. protein interactions, cis-acting transcription factor binding sites) improves performance (with respect to the identification of co-functional putative co-regulated modules). There are many benefits to comparing elements among species considering that a high fraction of co-regulated modules are conserved, in whole or in part, across species (Ihmels, Bergmann et al. 2005; Tirosh and Barkai 2007). Recent access to multiple genomics datasets from multiple species has allowed for new comparative analyses of genomics data, for example discovering regulatory elements (Elemento, Slonim et al. 2007) and the MScM algorithm (Waltman, Kacmarczyk et al. 2010) used here. MScM learns coregulated modules by integrating expression data across subsets of experimental conditions, co-occurrence of putative cis-acting regulatory motifs in the regulatory regions of bicluster members, functional associations and physical interactions. The output consists of condition dependent conserved modules of orthologous gene groups as well as species-specific elaborations of these conserved groups. The method is a true biclustering method: a typical

conserved bicluster is typically supported by a subset of the input data for each species.

5.3.4 Component tools of our system

To enable exploration of a multi-species integrative biclustering result, we have constructed a system using the Gaggle and MScM (Figure 5.1). The Gaggle is a Java program that integrates tools by broadcasting gene, network and data selections between tools (for example nodes selected in Cytoscape are sent to the Gaggle, which then sends the selections to all tools which then automatically mirror those selections). The Gaggle has been shown to enable efficient creation of multi-tool systems to explore complex datasets and associated analysis (Bonneau, Facciotti et al. 2007). Also, the loosely coupled visualization systems the Gaggle enables have several advantages including: systems-performance advantages (one tool crashing does not disable the whole system), development advantages (existing tools need not be reengineered and can be incorporated with small development costs), and maintenance advantages (due to the modularity of the resulting systems). We have extended the gaggle tools (and built a corresponding database) to give the user the ability to mirror gene selections in tools populated with results for one organism with the corresponding selection of the correct orthologs in the network, data, and bicluster views of the other organism. Several component tools and databases are compatible (or have been made compatible as part of this work) with the Gaggle, including: Sungear, Cytoscape, Cytoscape plugins such as BioNetBuilder (Avila-Campillo, Drew

et al. 2007), several online public databases containing annotations and genomic sequence via the FireGoose (Bare, Shannon et al. 2007), a Global gene-synonym Translator, and several tools designed to enable exploration of the genomics data available for each species (such as the data matrix viewer (DMV) and annotations viewer). Selections in any tool are sent to the Gaggle which broadcasts both those gene selections to all tools for the organism in which the original selection was made and the orthologs in the other species of the selected genes. We show that this simple strategy enables effective exploration of this multi-datatype, multi-species integrative analysis.

5.4 Materials and Methods

We present an overview of the MScM algorithm, and the system we have constructed for visualizing the resulting multiple-species biclusters. Further methodological detail, additional validation of our method, and a full description of the dataset used to demonstrate our resource can be found in the supplemental section (section 5.8).

5.4.1 Data sets acquisition, integration and import to our system

Microarray data was acquired from several large, public repositories such as the Gene Expression Omnibus (GEO) (Edgar, Domrachev et al. 2002; Barrett and Edgar 2006), ArrayExpress (Brazma, Parkinson et al. 2003; Parkinson, Kapushesky et

al. 2009), Stanford Microarray Database (SMD) (Sherlock, Hernandez-Boussard et al. 2001; Hubble, Demeter et al. 2009), Many Microbes Microarray database (M3D) (Faith, Driscoll et al. 2008), and KEGG Expression (Kanehisa, Goto et al. 2002), with newer datasets manually obtained from individual publications. Genomic sequences corresponding to the upstream promoter regions of each predicted gene in each genome were retrieved from Regulatory Sequence Analysis Tools (RSAT) (van Helden 2003; Thomas-Chollier, Sand et al. 2008). Lastly, functional associations, in the form interaction networks, were automatically acquired from multiple sources including Prolinks (Bowers, Pellegrini et al. 2004), Predictome (Mellor, Yanai et al. 2002), STRING (Snel, Lehmann et al. 2000; Jensen, Kuhn et al. 2009), and MicrobesOnline (Dehal, Joachimiak et al. 2009). We have created a data compendium containing all publicly available data for a number of microbial species including several Gram negative species *Escherichia coli*, *Salmonella typhimurium*, *Vibrio cholerae*, *Helicobacter pylori*, *Desulfovibrio vulgaris*; three related Gram positive species *Bacillus subtilis*, *Bacillus anthracis*, *Listeria monocytogenes*, and the archeon *Halobacterium salinarum*; within this compendium all name translations have been curated to minimize error due to incorrect translation of gene synonyms. In selecting this group of microbial species, we decided to start with the two most extensively studied bacterial model organisms, *E. coli* and *B. subtilis*, include several closely related species and some representatives from important clades of the microbial tree of life. Additional species will be included in future versions of the database, as a

sufficient amount of large-scale data becomes available for those species. A full listing of all datasets used in this study for both species (including references to papers describing both original collection, and several data-bases that aided the import and curation of the datasets) are provided in the supplemental materials (section 5.8).

5.4.2 Multi-species cMonkey

The MScM algorithm consists of four main steps. Beginning with step 1, putative orthologous relationships between genes in each species are identified using InParanoid (Remm, Storm et al. 2001). InParanoid identifies not only single gene pair relationships (one-to-one) but also families of homologous genes (one-to-many, many-to-many). This allows for flexibility when considering which orthologous gene pairs to cluster (i.e. in many-to-many groups the selection of orthologous pairs is driven by the genomics data, see Text S1 for details). After defining the set of gene pairs (pairs of genes spanning the two species, one pair per putative orthology relationship; genes are often in several putative orthology relationships following step one), or conserved core, step 2 identifies the conserved biclusters via an iterative Monte Carlo optimization of the MScM score (a score that judges biclusters composed of multiple orthologous gene pairs by a simultaneous scoring of expression, networks and upstream binding site support for the bicluster in each species). To determine the likelihood of an orthologous gene pair belonging to a bicluster, we compute a single, multi-species score based on the combined single-species scores for each gene supported by each organism's individual data space (expression, common sequence

motif, and connected subnetwork). The putative-orthology based gene coupling between species is removed in step 3, where each detected conserved bicluster is split into two single-species biclusters and species-specific additions are made separately for each species using the single species cMonkey score. The conserved core of the bicluster detected in step 2 is preserved (treated as read only) while species-specific additions to the conserved biclusters (including both non-orthologous and orthologous genes) are discovered via this iterative optimization. An optional step 4 (not carried out in this study) identifies purely species-specific biclusters for each organism using the original cMonkey algorithm applied to genes not yet in any conserved (multi-species) bicluster.

We have made the cMonkey and MScM code available including tools for automating many of the data acquisition and processing steps required for assembling an integrated dataset (Waltman, Kuppusamy et al. 2010). These tools facilitate automatic queries to online biological databases such as BioNetBuilder, MicrobesOnline (Dehal, Joachimiak et al. 2009), Prolinks (Bowers, Pellegrini et al. 2004), STRING (Snel, Lehmann et al. 2000; Jensen, Kuhn et al. 2009) and RSAT (van Helden 2003; Thomas-Chollier, Sand et al. 2008) (for network and upstream data). All input and output are stored in a MySQL database to facilitate use of the integrated dataset and MScM results by other tools. We also include example inputs for the algorithm both as flat files and as R data objects for those wishing to use data not in public databases (requiring manual mode). These key changes to how data is imported

and stored in the MScM database and the core data-object for cMonkey and MScM are critical novel changes to the code that are required for multi-species integration and scaling of the code to much larger datasets and organisms.

5.4.3 Visualizing multi-species clustering and biclusters

We created a database containing the MScM biclustering analysis data compendium for a number of microbial species. Our pipeline begins with several post-processing steps to convert cMonkey output to Gaggle compatible formats. Enrichment of functional annotations within biclusters is determined for each bicluster and the bicluster is assigned any significant annotations (p-values < 0.05). From the statistical components of each bicluster (e.g. residual, functional enrichment significance values, etc) a score is computed. Specifically, the bicluster score is computed using Stouffer's z-score method for meta-analysis from a collection of bicluster statistics. Data files are generated for the complete bicluster network and the subnetwork of related biclusters before the website for a result is generated. Lists of orthologous genes between each species are generated as part of the analysis and loaded into the synonym/ortholog database.

5.4.4 Multi-species extension of the Gaggle

To mirror selections simultaneously in several tools that visualize different aspects of the data, the results and the comparison between species we utilize the Gaggle, a loosely coupled system of web applications (geese) (Shannon, Reiss et al.

2006). The Gaggle is a software framework that integrates independent application tools and biological data into an environment that allows the exchange of data among tools. All of the tools employed in our resource are Java web-starts or directly integrated into the web interface, thus removing any barrier to use based on tool compatibility, installation or data-transfer. The Gaggle also serves to coordinate the deployment and interoperation of these Java Web Start tools. Each individual application, or goose, can be launched with the click of a button on the BiclusterCard. The geese included in the resource are: a Global synonym Translator, BioNetBuilder (Cytoscape plug-in), the FireGoose, Data Matrix Viewer, Annotations viewer, Cytoscape (bicluster network and gene network viewers), and Sungear. All the tools are connected through a communication hub called the Gaggle Boss, which passes simple messages among the geese (called broadcasting), summarized in Figure 5.1. When a broadcast is received, the goose will display the relevant information for that data. BiclusterCards and online databases (e.g. STRING, KEGG, etc.) connect to the tools through the FireGoose, a browser plug-in for Firefox adding the capability to communicate with the Gaggle. Embedded in each BiclusterCard is microformat code containing metadata (e.g. gene names, bicluster nodes, condition names) that can be broadcasted to other geese. The Bicluster Network viewer is a Cytoscape goose that displays a network of bicluster interactions, where nodes are biclusters, and edges are any shared properties (e.g. functional annotation, gene overlap, etc). Similarly the Gene Associations viewer is a Cytoscape goose that displays the gene associations

from the data compendium. A Data Matrix Viewer goose acts as a spreadsheet program that can display and plot gene expression values. The Annotations goose displays a table of the genes and their various annotations (e.g. locus tag, gene name, protein id, gene id accession, etc.)—this is specific to a single organism. There is a Global Translator that, given a list of genes from one species, can display the orthologous genes from another species. Lastly, the MScM output showing gene expression, gene subnetwork, sequence motifs, and motif locations in promoter sequence, can be displayed in the ClusterInfo Viewer.

5.4.5 The Web and Gaggle interface to our multi-species biclustering

A web interface was implemented to facilitate exploration of the multi-species biclusters. The starting page allows users to create several types of queries and contains a text box to input a gene name or group of genes, select boxes to choose bicluster sets from single and, core or elaborated multi-species cMonkey analyses, and a submit button to begin the search for biclusters containing the gene or genes of interest from the selected biclustering analyses (Figure 5.2A). Any biclusters returned from a search are presented as a list ranked by bicluster score. A first step in organizing the diverse information contained in (supporting) each bicluster was to create a system for generating bicluster summaries that link to online tools and source data. To this end, for each bicluster, our system creates a 'BiclusterCard'. Each BiclusterCard provides the following information in the form of expandable/collapsible tabs, Figure 5.2B:

- Gaggle tools: Embedded links to integrated software tools
- Statistics: The number of genes and conditions in the bicluster, score, residual, mean motifs p-value, motif E-values
- Enrichment Summary: based on the most significant annotations from COG, KEGG and GO enrichment analysis
- Core Genes: Genes table for conserved core members of the bicluster– including GO, KEGG, and COG gene annotations
- Elaborated Genes: Same as above, but for elaborated members of the bicluster
- Experiments: Table with links to the meta-data and primary articles
- Bicluster Motifs: if any motifs were found, the sequence logo is displayed here along with matches to any known motifs
- Enrichment Analysis: Tables for GO, KEGG, and COG annotation enrichment – with description and significance values
- Related Biclusters: Table with links to biclusters with similar functional/pathway annotations, similar motifs, or overlapping gene members
- Plots: Bicluster plots for gene expression, mean gene expression, expression heatmap, and motif locations in gene promoter regions

Each element of the bicluster card is generated automatically by our system, is compatible with outputs from other widely used biclustering tools, and provides links to descriptions/tutorials for using the linked tools or databases.

5.5 Results/Discussion

To demonstrate our resource's capabilities, we explore nitrogen metabolism associated multi-species biclusters with the specific biological goal of identifying new genes functionally associated with nitrogen metabolism in *E.coli* and *S. typhimurium*. For a global validation of our multi-species biclustering method and a detailed comparison of our method to several other methods, as well as a detailed description of the complete dataset used in this study see the supplemental section (section 5.8) provided in the electronic version of this article. The CMMR is available at http://meatwad.bio.nyu.edu/cmmr.html.

5.5.1 Exploring nitrogen metabolism in an *E. coli* and *S. typhimurium* integrated genomics dataset

Nitrogen is an essential input into several metabolic pathways including amino acid and nucleotide biosynthesis, and can act as a terminal electron acceptor in dissimilatory nitrate reactions (Stanley, Gunsalus et al. 2007). It is common for some microbes including *E. coli* to use nitrogen for energy-harvesting purposes in anaerobic and nutrient depleted conditions (Stanley, Gunsalus et al. 2007). A central component of nitrogen assimilation and metabolism is nitrate reductase, a membrane bound enzyme that catalyzes the conversion of nitrate to nitrite. The *narGHJI* operon encodes the multiple subunits of nitrate reductase A in *E. coli*. The following section sequentially guides the reader through using our system to explore biclusters containing genes in the *nar* operon and other nitrogen metabolism associated genes. A

web tutorial for the use of our system can also be found at: http://meatwad.bio.nyu.edu/psbr/index.php/Tutorials

5.5.1.1 Identifying a potential role for unknown genes in biclusters containing nar genes

We begin our exploration of identifying conserved biclusters containing *nar* genes by searching for "narG" in the core set of genes from an *E. coli* and *S. typhimurium* MScM bicluster set (Figure 5.2A; typing 'narG' into the gene-name textbox, selecting the core checkbox and clicking 'submit' on the CMMR start page, will retrieve any biclusters containing *narG* in the core set of genes). The results page returned following our "narG" query includes a header with links to the CMMR wiki, links to tutorials, a description of the search query and a list of any retrieved biclusters, in this case 3 biclusters were found (Figure 5.2B). There is a button for each bicluster that will display its BiclusterCard (see materials and methods). Looking at the first BiclusterCard for *E. coli* bicluster-57 (eco57), we will click on the 'Coupled Bicluster' button to open the BiclusterCard for *S. typhimurium* bicluster 57 (stm57). Expanding the 'Statistics' tab shows that eco57 contains 75 genes (51 core genes, 24 elaborated genes), 226 experiments, whereas stm57 contains 66 genes (51 core genes, 15 elaborated genes) and 43 experiments (Figure 5.3A). This first table highlights differences in gene membership of the two biclusters. The 'Enrichment Summary' shows similar but not identical annotations involved in various metabolic activities related to anaerobic respiration and energy production from nitrogen for both

biclusters (Figure 5.3B). The ‘Experiments’ tab shows that expression of these genes changes under a variety of conditions including: stress, growth on minimal media, anaerobic metabolism, and DNA damage. Expanding the ‘Enrichment Analysis’ tab displays tables containing significant COG, GO and KEGG annotations. We can see that eco57 and stm57 differ in the ranking of the KEGG pathway annotations and stm57 includes an additional pathway (Figure 5.3C). This could reflect slightly different uses of these modules in these organisms or discrepancies in the gene annotations.

Then, looking at the gene GO, KEGG and COG annotations by expanding the ‘Core Genes’ tab we see many genes have the same or similar annotations and some have either none or different annotations such as *narG* and *yjjI* (Figure 5.3D). Finally, under the ‘Plots’ tab we can view plots for gene expression profiles, bicluster mean expression, and an expression heatmap – to visualize differences in clustering bicluster gene members (Figure 5.4A).

Expanding the ‘Bicluster Motifs’ tab displays the motifs detected in the bicluster. Two of the detected motifs for eco57 show similarity to known nitrate/nitrite response transcriptional regulator binding motifs (Figure 5.4B). Motif #1 matches the *E. coli* FNR (fumarate and nitrate reduction) binding consensus sequence (TTGAT N4 ATCAA) (Winteler and Haas 1996) and eco57 motif #3 corresponds to the NarP binding sequence (Kazakov, Cipriano et al. 2007; Gama-Castro, Salgado et al. 2011). The sequence motifs of stm57 show no notable similarity to known motifs. The FNR

homolog in *S. typhimurium*, *oxaR*, has a similar but less defined consensus sequence (Fink, Evans et al. 2007), which could account for the lack of association with stm57 motif #1. The promoter motif patterns display which gene members share common motifs and the location in the gene's upstream sequence. Identical motif patterns indicate they are an operon, such as operon *narGHJI* (Figure 5.4C). MScM and MicrobesOnline (Alm, Huang et al. 2005; Price, Huang et al. 2005) predict *yjjI* and *yjjW* to be in an operon, which is reflected in eco57 (*yjjW* is present in the elaborated gene set) but not stm57 (Figure 5.4C). Exploring the correspondence of the MScM detected motifs with known nitrogen metabolism motifs increases our level of confidence that this bicluster is truly coregulated in both organisms.

Among the core gene list for this bicluster, *yjjI* is described only as encoding a conserved protein with no functional annotation (Figure 5.3D). To examine this gene in the context of multiple network-types, the original data, and the biclustering, we now open several Gaggle tools, including the bicluster and gene network Cytoscape geese, Data Matrix Viewer, BioNetBuilder, and the Global Translator. First, we explore associations between core gene members of eco57 and stm57. For the 51 genes in the core gene member subnetworks, eco57 has 518 associations and stm57 has 420 edges, with no associations for *yjjI* (Figure 5.5A; associations shown are operon edges, metabolic pathway edges, phylogenetic profile edges, and protein interaction edges between genes in different biclusters). Next, we explore the expression profiles of the bicluster gene members and conditions by broadcasting

them to the Data Matrix Viewer. Selecting *yjjI*, we can see that it has similar expression to other bicluster gene members (Figure 5.5). Thus, the data (sequence motifs, associations, expression) supports eco57 and stm57 as coherent, putatively coregulated gene groups, and gene *yjjI*, while lacking associations, is supported by common motifs and correlated expression. We can use more Gaggle tools to search for additional information characterizing the bicluster gene members, particularly *yjjI*. For example, broadcasting the gene members to BioNetBuilder, we can browse protein structure and functional predictions. YjjI is predicted to have a domain structure that matches a "Class III anaerobic ribonucleotide reductase NRDD subunit" (Fontecave, Eliasson et al. 1989) and a function prediction of oxidoreductase activity (Riffle, Malmstrom et al. 2005; Malmstrom, Riffle et al. 2007). If we broadcast *yjjI* to other online databases such as Entrez Gene (Maglott, Ostell et al. 2005), we find that *yjjI* is adjacent to *yjjW*, but no information that they are in an operon. As mentioned above, both MScM and MicrobesOnline have predicted them to be in an operon. There is further information from EcoGene (Rudd 2000) reporting *yjjI* as an ortholog of *H. influenzae hi0521*, which is a *pflB* homolog and coding for a formate acetyltransferase (Kolker, Makarova et al. 2004). Taken together, this information suggests a role for YjjI in nitrogen metabolism. It is important to note that a corresponding single-species bicluster in *E. coli* was not found (in the *E. coli* single species cMonkey run we find no bicluster with significant gene overlap to this significant conserved bicluster), further illustrating the importance of the MScM method. However, the species-

specific elaborations of the bicluster may display additional information, such as, individual adaptations to this metabolic process.

Another possible use of our system is the exploration of collections of biclusters to identify novel interactions among modules. In the context of this example we can extract the subnetworks of biclusters related to the nar bicluster described above from a network that displays associations between biclusters by broadcasting the list of related biclusters from the BiclusterCard to the Bicluster Network Viewer (Figure 5.5C). Biclusters are nodes with width and height proportional to the number of genes and conditions, respectively, and shared significant KEGG pathway, COG function, and GO function annotations are edges). The subnetwork shows 27 related biclusters for *E. coli* and 17 biclusters for *S. typhimurium*; in this subnetwork there are several biclusters containing gene modules highlighting complementary interactions such as: amino acid biosynthesis/metabolism pathways and glutamate metabolism (bicluster-61); NADH dehydrogenase, succinate dehydrogenase (bicluster-43), citrate fermentation (bicluster-147), and amino acid ABC-type transporters (bicluster-148). This highlights the presence of conserved core interactions among eco57 and stm57 with other modules and independent species-specific modifications within these modules.

Using the CMMR, much knowledge was uncovered from the search of just a single gene, *narG*. In one case, for a currently uncharacterized gene, *yjjI*, the gathering of diverse information such as: putative orthology between two species, co-expression

and common putative regulatory motifs with other bicluster genes, and a prediction for the protein's structure and function, was facilitated by the various BiclusterCards and Gaggle tools. In another case, we could explore modules (biclusters) that included *narG* and their interactions.

5.5.2 Conclusions

We have developed a publicly accessible web resource for comparative genomics studies of several prokaryotic organisms, with plans to expand this resource over time. As described above, in our example with coupled *E. coli – S. typhimurium* bicluster 57, the combination of our method for simultaneously biclustering multiple datasets from multiple species and easy to use exploration system quickly led to novel biological insights and generate an informed hypothesis about the involvement of gene *yjjI*, a currently uncharacterized gene, in nitrogen metabolism. The complexity and richness of the results of comparative genomics data analysis requires a system like the one presented here. We present specific examples of the use of our system in the hopes of sparking discussion about what the next generations of comparative genomics analysis and visualization systems should look like. Our paper focuses on the combined, multi-tool interface required by biologists wishing to explore the biological significance and function of multi-species, multi-datatype biclusters and their species-specific elaborations and deletions. An important aspect of our system is the ability to submit new data for analysis and integrate the results into the resource for public access. We provide multiple avenues for researchers wishing to build this

system for their species of interest (tools and code are publicly available) and/or we will run our analysis and build this system for researchers without computational resources.

5.6 Acknowledgments

The authors would like to thank Aviv Madar, Alex Greenfield, Chris Poultney, and Kieran Mace for their helpful comments and discussions.

5.7 Figures

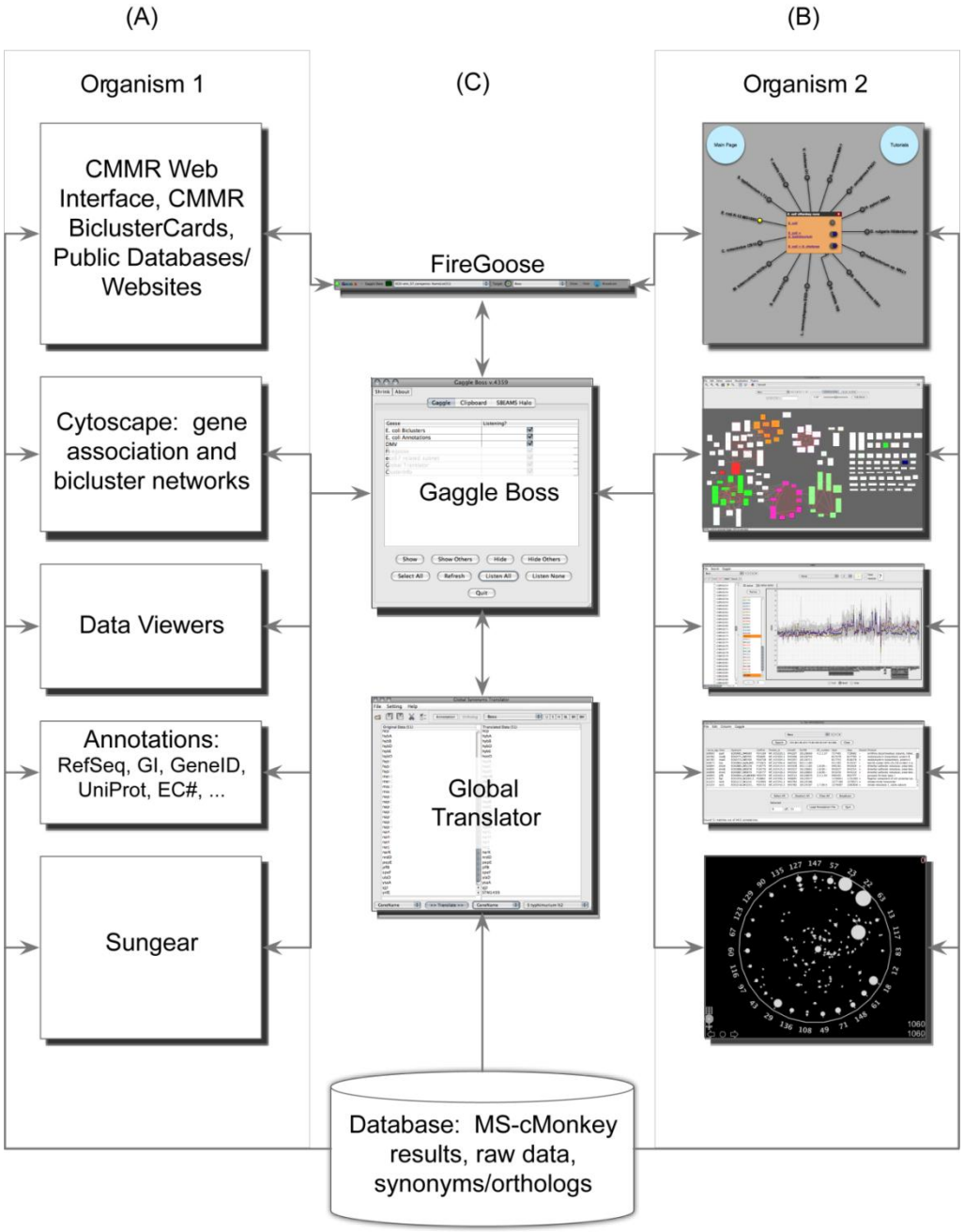


Figure 5.1: Overview of the Comparative Microbial Module Resource components (CMMR). The CMMR consists of an integrated suite of web components for visualizing the diverse aspects of the

multi-species, multi-datatype analysis; facilitating access to each organism's dataset. (A) Written descriptions of the individual components (for hypothetical Organism 1). (B) The corresponding graphics of each component goose displaying example data (for hypothetical Organism 2). Each of the components fetches information from the data compendium (MScM results, and raw data). (C) The CMMR integrative components: the FireGoose allows transfer of data between web pages and gaggled software, the Gaggle Boss acts as a hub for passing communications among the geese, and the Global Translator converts among gene annotations, accessions and translates orthologous genes between organisms. The arrows represent information flow between tools, primarily as broadcasts between tools and the Gaggle boss.

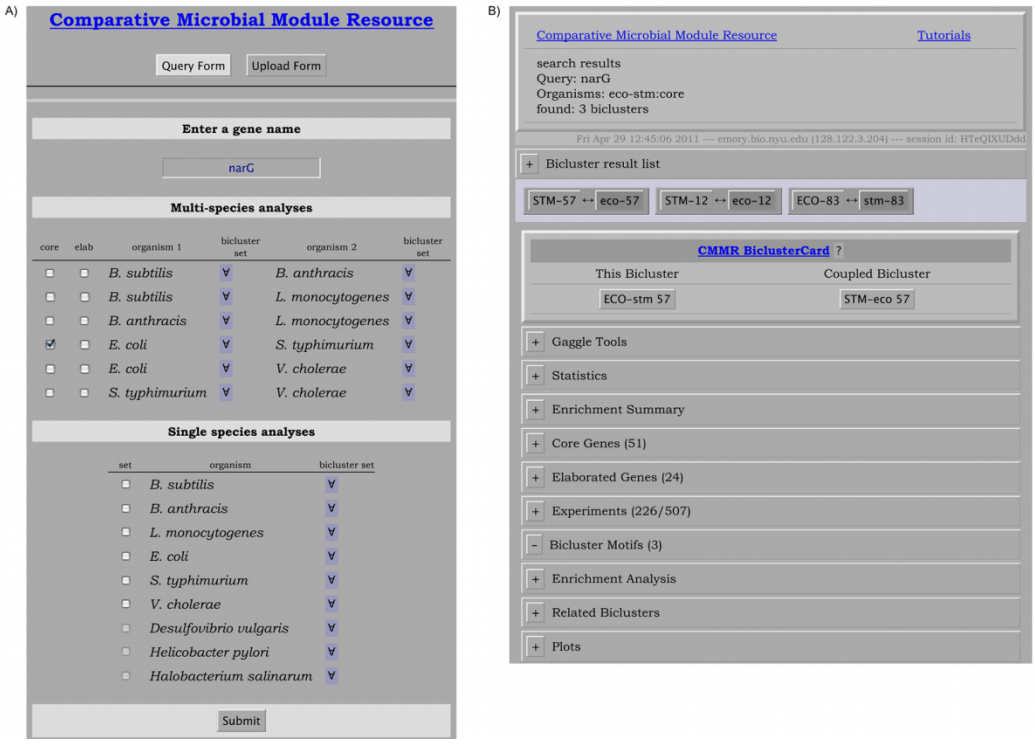


Figure 5.2: CMMR Query Page and BiclusterCard. The CMMR web interface allows users to search for biclusters of interest, with each resulting bicluster displayed in a BiclusterCard format. (A)

The CMMR search page showing the title link to the CMMR wiki, query form button, upload form button, and input fields. Shown is the query form with an example search for *narG* in the core set (check box) of bicluster gene members for a MScM run of *E. coli – S. typhimurium*. (B) The result page from this search – a user has access to the CMMR wiki, tutorials, a brief description of the search query, the resulting bicluster list and BiclusterCards. The BiclusterCard contains links to Gaggle tools, and expandable/collapsible tabs to display the bicluster's diverse supporting information.

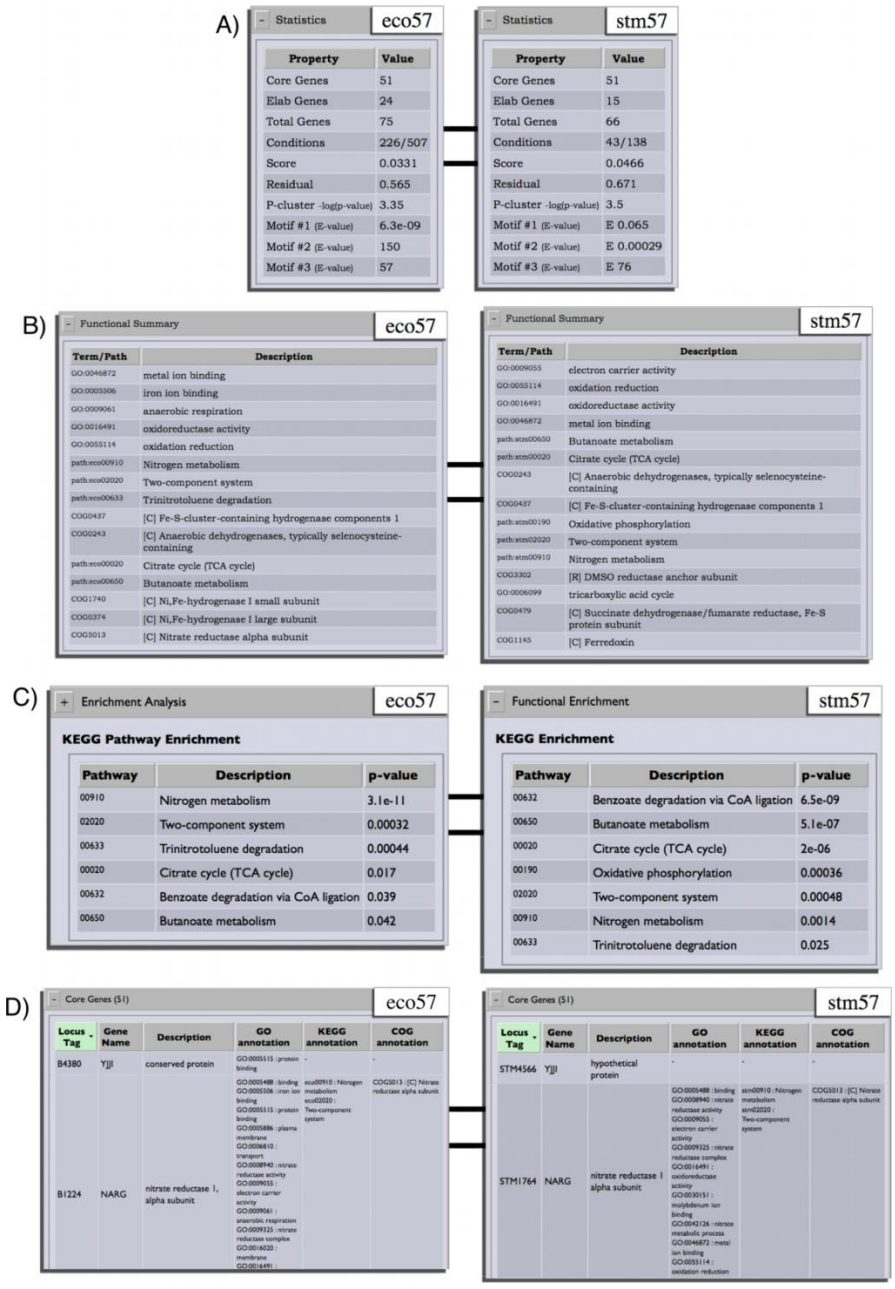


Figure 5.3: BiclusterCard components I: Statistics, Enrichment Summary, Core Gene Table, KEGG Pathway Enrichment. The BiclusterCard is a summary of the information supporting a bicluster, including links to online tools and source data. Shown in the figure are the expanded tabs for: statistics, enrichment summary from COG, GO and KEGG enrichment analysis, KEGG pathway

enrichment, and core gene table for multi-species bicluster *E. coli – S. typhimurium* bicluster 57. (A) Statistics tab for eco57 (left) and stm57 (right) displays a table with the following columns: Property and Value. The information contained in this table includes: the number of core and elaborated genes, fraction of conditions in the bicluster, the bicluster score, bicluster residual, bicluster mean p-value (mean of all motifs found in the promoter sequences), and the E-value for each motif found in the bicluster. (B) Enrichment Summary tab for eco57 (top) and stm57 (bottom) displays a table with the following columns: Term/Pathway and Description. This table lists the most significant annotations from ontological enrichment tests of COG, KEGG pathway, and GO annotations. (C) The Functional Enrichment tab displays tables listing the significant annotations from the COG, GO and KEGG enrichment analyses. Shown is the KEGG pathway enrichment table for eco57 (top) and stm57 (bottom). The table consists of the following columns: Pathway, Description, and p-value. Each column can be sorted. (D) Core Gene tab for eco57 (top) and stm57 (bottom), showing the number of core genes (51), and a table containing the following columns: Locus Tag, Gene Name, Description, GO annotations, KEGG annotations, and COG annotations. Locus Tag, Gene Name and Description columns can be sorted.

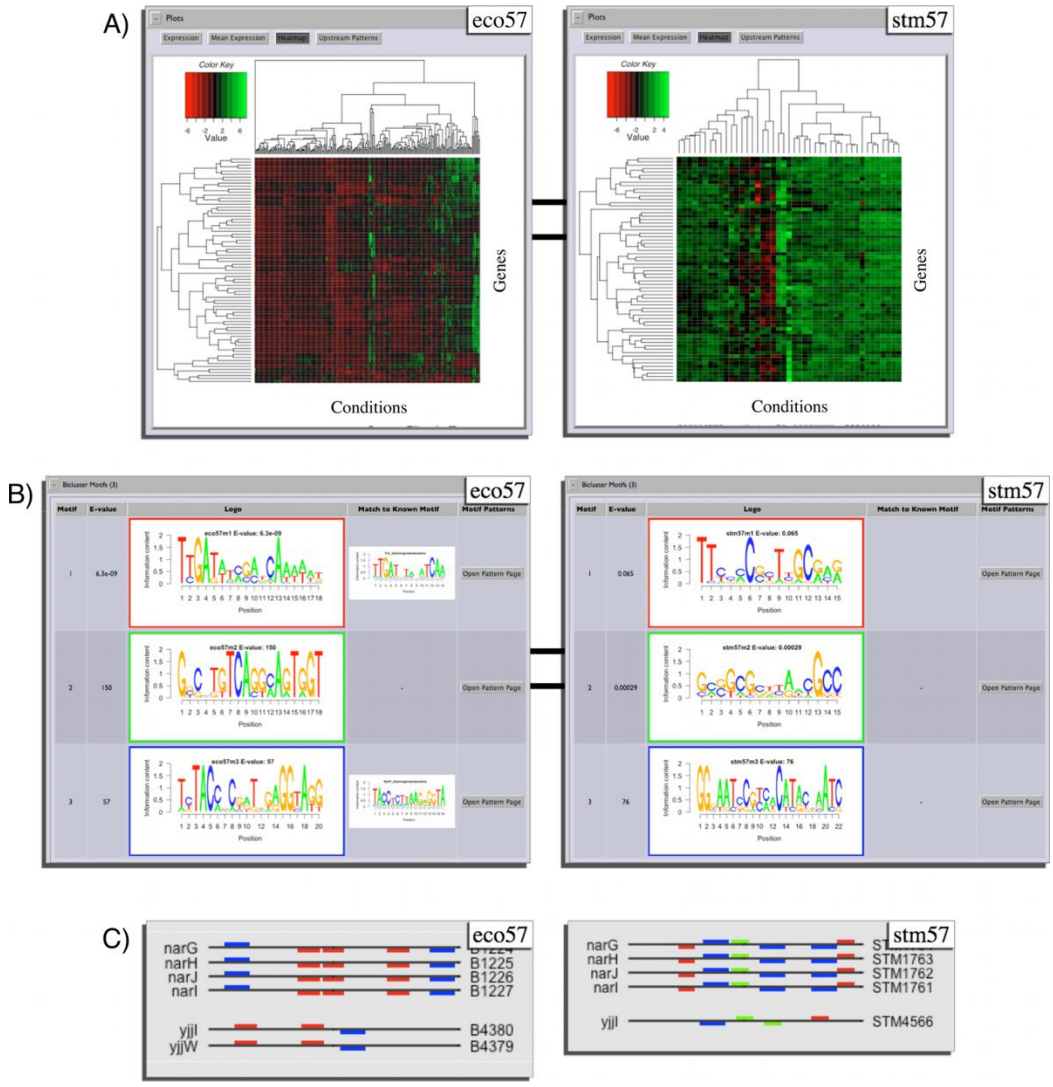


Figure 5.4: BiclusterCard components II: Bicluster Motifs, Upstream Patterns, Plots. Shown in the figure are the expanded tab for Plots displaying a gene expression heatmap, the expanded tab for Bicluster Motifs, and an example of the upstream motif patterns for multi-species bicluster *E. coli – S. typhimurium* bicluster 57. (A) Example plot of a gene expression heatmap for the bicluster genes and conditions in eco57 (left) and stm57 (right); upregulated expression (green) and downregulated expression (red). (B) Putative regulatory sequence motifs found in bicluster gene member promoters for eco57 (left) and stm57 (right). The table displays a row for each motif found and columns for the motif

number, E-value, sequence logo, matches to any known motifs, and a link to motif pattern page. Eco57 motif #1 matches the known FNR binding sequence and motif #3 matches the known NarP binding sequence. (C) The promoter motif patterns for the motifs shown in (B) for eco57 (left) and stm57 (right). The location of the motifs are represented by colored rectangles on the promoter sequence (black line) and the colors correspond to the logo border colors seen in (B); motif #1 (red), motif #2 (green) motif #3 (blue). For the bicluster gene members shown, bicluster motifs #1 and #3 appear in the promoter regions of the eco57 members, whereas all three bicluster motifs appear in the promoters for the stm57 members. The identical motif pattern indicates MScM has determined them to be in an operon. It is known that *narGHJI* exist as an operon, but MScM has determined that *yjjI* is in an operon with *yjjW* (this is also predicted by (Price, Huang et al. 2005)). However, *yjjW* is found only in the elaborated gene set of eco57 and it is not found in stm57.

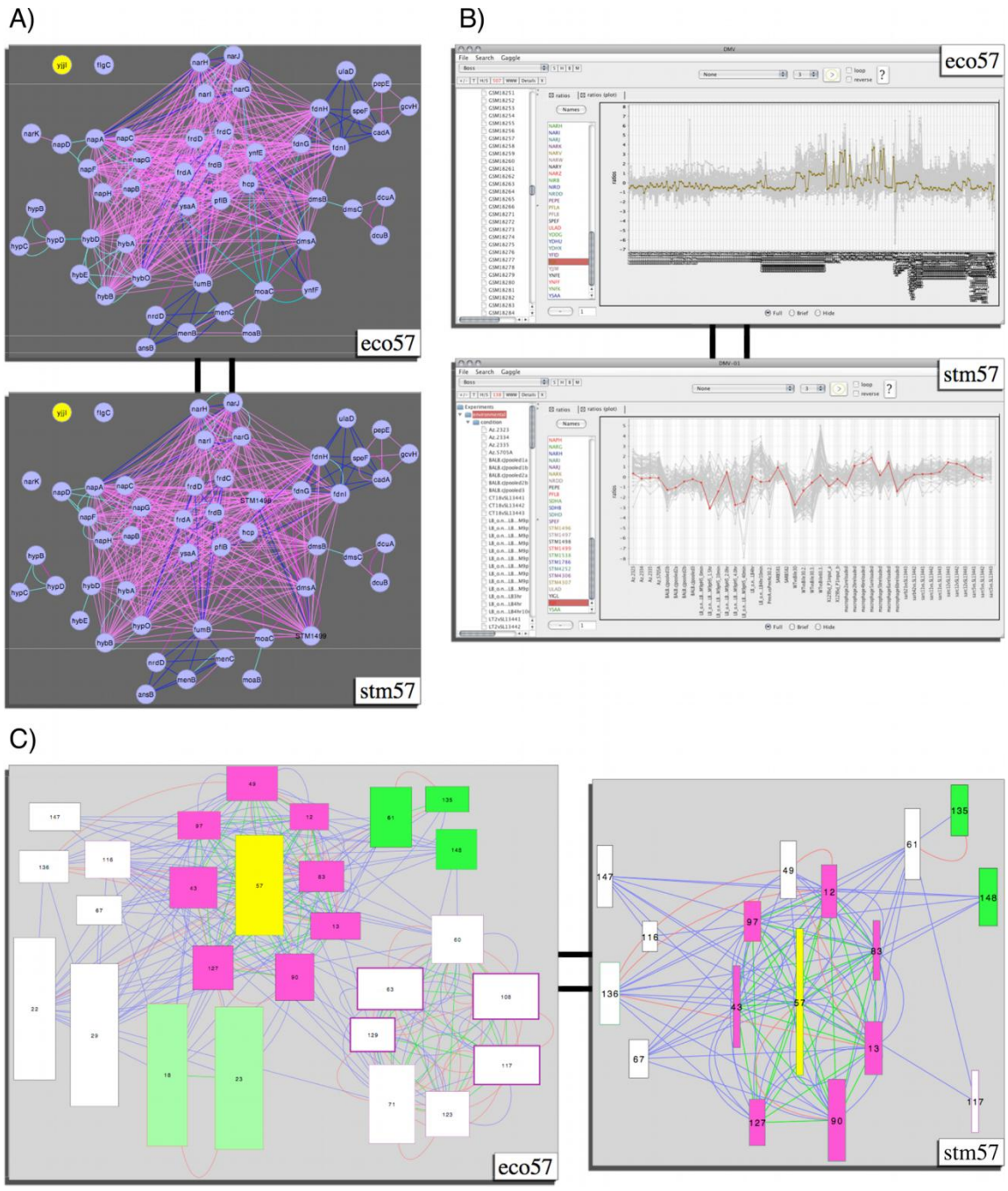


Figure 5.5: CMMR linked Gaggled tools. Expanding the Gaggle tools tab on the BiclusterCard for multi-species bicluster *E. coli – S. typhimurium* bicluster 57, reveals a list of links (buttons) to the various Gaggle tools. (A) The Gene Associations button opens a Cytoscape goose that displays the core

genes subnetwork for eco57 (top) and stm57 (bottom). The nodes represent genes and edges represent associations based on data from the compendium, indicated in yellow is gene *yjjI*. Edges are shared annotations: COG code (pink), Prolinks phylogenetic profile (purple), metabolic pathway (blue), operon (light cyan), and Predictome phylogenetic pattern (dark cyan). (B) The expression profiles for the genes and conditions from eco57 (top) and stm57 (bottom) can be explored by opening the Data Matrix Viewer. Using the FireGoose, the bicluster's genes and conditions can be broadcast from the BiclusterCard. We can see how the expression profile of gene *yjjI* (indicated by the colored line) matches other profiles in the bicluster. (C) The Bicluster Network button opens a Cytoscape goose to display the complete bicluster network where each node is a bicluster (width and height proportional to number of genes and conditions, respectively) and edges represent any shared properties and annotations. We can explore the related bicluster subnetwork for bicluster 57 (yellow), eco57 (left) and stm57 (right), by broadcasting the list of related biclusters (using the FireGoose) from the BiclusterCard to select those biclusters and display them in a new window. There are 10 additional biclusters in the eco57 subnetwork. Node fill color represents significant COG annotation, border color represents significant GO annotation, node border thickness represents residual, and edge color represents shared COG (green) KEGG (red), or GO (blue) annotations.

5.8 Supplementary text

5.8.1 Materials

5.8.1.1 Dataset analyzed

The *E. coli* expression data matrix consisted of 507 conditions from 16 projects acquired from the Many Microbe Microarrays Database (M3D) (Faith, Driscoll et al. 2008) covering various conditions including: genetic perturbations, changes in

oxygen concentration and pH, growth phases, antibiotic treatment, heat shock, and different media.

The *S. typhimurium* expression data matrix consisted of 138 conditions from 8 studies acquired from the Stanford Microarray Database (SMD) (Sherlock, Hernandez-Boussard et al. 2001; Hubble, Demeter et al. 2009) covering various conditions including: chemical effects, nutrient limitation, library verification, strain comparison, media comparisons, time course, and mutants.

Table 5.1: Total number of genes, conditions, and association edges in each dataset used for the multi-species analysis, by organism.

Number of:		***E. coli***	***S. typhimurium***
	Genes	4264	3745
	Conditions	507	138
Association edges			
Source	Egde type		
	Operon	3414	2104
KEGG	**Metabolic**	96931	75363
Prolinks	**Gene Neighbor**	29228	29942
Prolinks	**Phylogenetic Profile**	20058	20094
Prolinks	**Gene Cluster**	6048	6476
COG	**COG-code**	644856	379484

Table 5.2: Total number of orthologs, orthologous families, and ortholog pairs generated by InParanoid.

Number of:	*E. coli*	*S. typhimurium*
orthologous groups	2827	
orthologous pairs	2856	
multi-member groups	22	
Remaining unique genes	2836	2845

5.8.2 Methods

5.8.3 MScM Algorithm Pseudocode Overview

Define *organisms*, *orthologs*, *num.biclust*, and *iter.max* to be each organism's dataset (expression, genomic sequence, network associations), putative orthologs between the organisms, the number of biclusters to search for, and the maximum number of iterations for the procedure, respectively. The method is a Monte Carlo optimization that, given a bicluster seed, optimizes a bicluster by iteratively adding or dropping genes and conditions according to the multi-species score (*gain*). The individual likelihoods for the *gain* for expression, sequence, and association networks, are represented by *r*, *s*, and *q*, respectively. The membership probability ($prob_{membership}$) of becoming part of the bicluster is based on the *gain* and the decision boundary formed using logistic regession (model). See (Waltman, Kacmarczyk et al. 2010) for the complete description of the method.

Algorithm 1 MSCM.shared(organisms, orthologs, num.biclust, iter.max)

for $i = 1$ to $num.biclust$ **do**
  $bicluster \leftarrow$ seed.bicluster($organisms, orthologs, conditions$)
  $iter \leftarrow 1$
  repeat
    { calculate the shared gain for each ortholog pair }
    for $ortholog.pair$ in $orthologs$ **do**
      for org in $organisms$ **do**
        { compute motif likelihods in promoter regions of genes }
        $s \leftarrow$ detect.motifs($orthologs, upstream.sequences$)
        $gain_{shared}[ortholog.pair, org]$ += G($bicluster[org], conditions[org], ortholog.pair[org], org, r, s, q$)
      end for
    end for
    $model \leftarrow$ logit($gain_{shared}, bicluster$)
    { calculate probability drop and add genes }
    for $ortholog.pair$ in $orthologs$ **do**
      if $ortholog.pair \in bicluster$ **then**
        $prob_{membership}[ortholog.pair] \leftarrow \mathrm{P}^{drop}(gain_{shared}[ortholog.pair], model)$
      else
        $prob_{membership}[ortholog.pair] \leftarrow \mathrm{P}^{add}(gain_{shared}[ortholog.pair], model)$
      end if
    end for
    { calculate probability drop and add conditions in each organism }
    for $condition$ in $conditions$ **do**
      if $condition \in bicluster$ **then**
        $prob_{membership}[condition] \leftarrow \mathrm{P}^{drop}(gain_{shared}[condition], model)$
      else
        $prob_{membership}[condition] \leftarrow \mathrm{P}^{add}(gain_{shared}[condition], model)$
      end if
    end for
    update $bicluster$ based on $prob_{membership}$ sample distribution
    $iter$ += 1
  until convergence **or** $iter == iter.max$
  $bicluster.list[i] \leftarrow bicluster$
end for
return $bicluster.list$

5.8.4 Validation

Table 5.3: Quick lookup table for methods considered by this study.

Multi-Species	Expression Only		Full Data	
	Shared space	**full genome (elaboration)**	**Shared space**	**full genome (elaboration)**
cMonkey	EO-MScM-SH	EO-MScM-EL	FD-MScM-SH	FD-MScM-EL
ISA*	MSISA-P	MSISA-R	NA	NA
K-Means*	MSKM-SH	MSKM-EL	NA	NA
(Balanced) K-Means*	BMSKM-SH	BMSKM-EL	NA	NA

Single-Species	Expression Only	Full Data
cMonkey	EO-SSCM	FD-SSCM
Coalesce	EO-COAL	FD-COAL
Qubic*	QUBIC	NA

*** Expression only method by method definition - no distinction between "expression only" or "full data" is necessary.**

5.8.5 Overview of the bicluster comparison metrics

A comparison of the relative performances of four multi-species methods (MScM, MSISA, MSKM and BMSKM), and three single species methods (SSCM, Coalesce and Qubic) in this study are based on 5 metric classes: 1) bicluster

coherence; 2) functional enrichment; 3) coverage; 4) overlap between biclusters; and 5) conservation. Bicluster coherence is determined by the combination of five commonly used metrics that gauge the degree of support provided to each bicluster by the three data types that MScM integrates (expression, sequence and association networks). See [10] for comparisons of SSCM to other biclustering algorithms, and [19] comparisons between single species biclustering and clustering algorithms. Our coherence metrics are: 1) expression residuals – a measure of the coherence of expression across the two species datasets for conditions within the bicluster; 2) mean correlation – the average pairwise correlation between members of a (bi)cluster (taking the absolute value to allow fair comparison between methods that identify inversely correlated patterns (QUBIC and MSISA) and those that do not; 3) network p-values – a measure of the significance of the sub-networks within biclusters compared to the full network; 4) motif E-values – a measure of the quality/significance of the upstream binding site motifs detected for each (bi)cluster; and 5) sequence p-values – an estimate of a sequence's match to the motifs associated with a (bi)cluster. Each of the coherence metrics is described in greater detail in (Waltman, Kacmarczyk et al. 2010).

5.8.6 Quick-glance table & Additional figures for the *E. coli – S. typhimurium* pairing

The Quick-glance table for *E. coli* and *S. typhimurium* is available in Appendix 1 (Table 3.5), and the appropriate figures can be found in sections 7.2.1 and 7.2.2.

5.8.7 Description of highlighted biclusters

5.8.7.1 *E. coli* bicluster 57

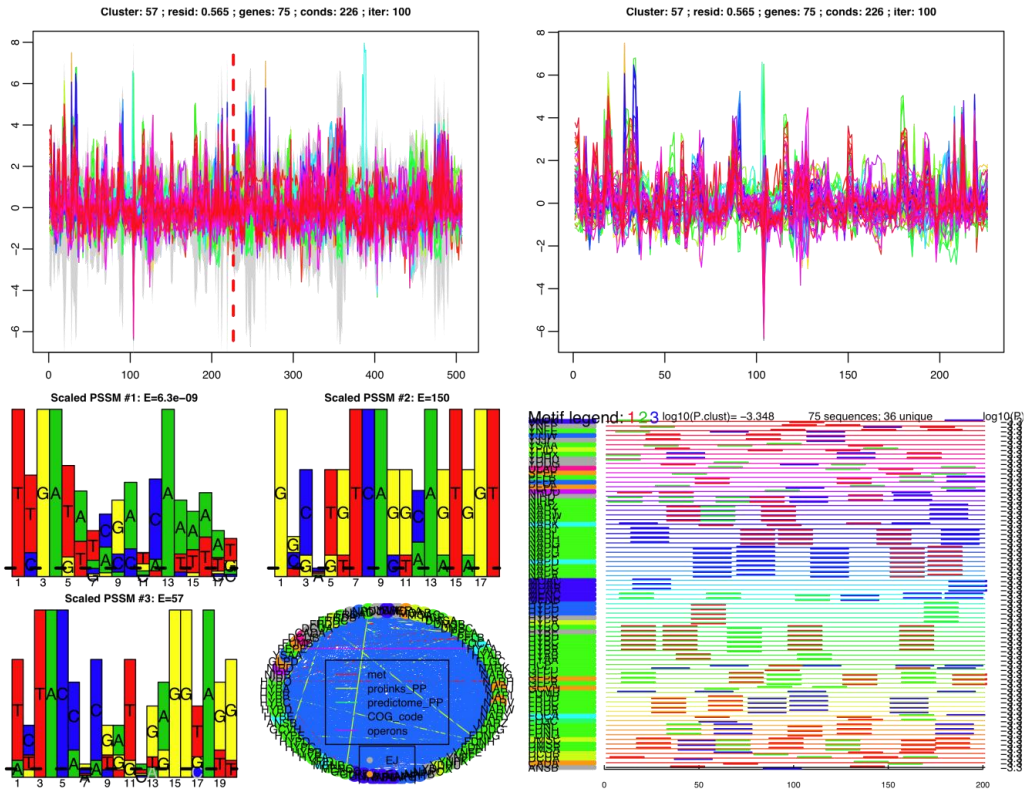


Figure 5.6: *E. coli* bicluster 57 MScM output image

5.8.7.2 *E. coli* bicluster 57 core gene list

Locus	Name	Description
B0693	SPEF	ornithine decarboxylase isozyme, inducible

Locus	Name	Description
B0782	MOAB	molybdopterin biosynthesis protein B
B0783	MOAC	molybdopterin biosynthesis, protein C
B0873	HCP	hybrid-cluster [4Fe-2S-2O] protein in anaerobic
B0894	DMSA	dimethyl sulfoxide reductase, anaerobic, subunit
B0895	DMSB	dimethyl sulfoxide reductase, anaerobic, subunit
B0896	DMSC	dimethyl sulfoxide reductase, anaerobic, subunit
B0903	PFLB	pyruvate formate lyase I
B1074	FLGC	flagellar component of cell-proximal portion of
B1587	YNFE	oxidoreductase subunit
B1588	YNFF	oxidoreductase subunit
B1476	FDNI	formate dehydrogenase-N, cytochrome B556 (gamma)
B1475	FDNH	formate dehydrogenase-N, Fe-S (beta) subunit, nitrate-inducible
B1474	FDNG	formate dehydrogenase-N, alpha subunit, nitrate-inducible
B1227	NARI	nitrate reductase 1, gamma (cytochrome b(NR))
B1226	NARJ	molybdenum-cofactor-assembly chaperone subunit
B1225	NARH	nitrate reductase 1, beta (Fe-S) subunit

Locus	Name	Description
B1224	NARG	nitrate reductase 1, alpha subunit
B1223	NARK	nitrate/nitrite transporter
B2202	NAPC	nitrate reductase, cytochrome c-type, periplasmic
B2203	NAPB	nitrate reductase, small, cytochrome C550
B2204	NAPH	ferredoxin-type protein essential for electron
B2205	NAPG	ferredoxin-type protein essential for electron
B2206	NAPA	nitrate reductase, periplasmic, large subunit
B2207	NAPD	assembly protein for periplasmic nitrate
B2208	NAPF	ferredoxin-type protein, predicted role in
B2261	MENC	o-succinylbenzoyl-CoA synthase
B2262	MENB	dihydroxynaphthoic acid synthetase
B2997	HYBO	hydrogenase 2, small subunit
B4131	CADA	lysine decarboxylase 1
B2727	HYPB	GTP hydrolase involved in nickel liganding into
B2728	HYPC	protein required for maturation of hydrogenases
B2729	HYPD	protein required for maturation of hydrogenases

Locus	Name	Description
B2904	GCVH	glycine cleavage complex lipoylprotein
B2957	ANSB	periplasmic L-asparaginase II
B2992	HYBE	hydrogenase 2-specific chaperone
B2993	HYBD	predicted maturation element for hydrogenase 2
B2995	HYBB	predicted hydrogenase 2 cytochrome b type
B2996	HYBA	hydrogenase 2 4Fe-4S ferredoxin-type component
B3573	YSAA	predicted hydrogenase, 4Fe-4S ferredoxin-type
B4021	PEPE	(alpha)-aspartyl dipeptidase
B4122	FUMB	anaerobic class I fumarate hydratase (fumarase
B4123	DCUB	C4-dicarboxylate antiporter
B4138	DCUA	C4-dicarboxylate antiporter
B4151	FRDD	fumarate reductase (anaerobic), membrane anchor
B4152	FRDC	fumarate reductase (anaerobic), membrane anchor
B4153	FRDB	fumarate reductase (anaerobic), Fe-S subunit
B4154	FRDA	fumarate reductase (anaerobic) catalytic and
B4196	ULAD	3-keto-L-gulonate 6-phosphate decarboxylase

Locus	Name	Description
B4238	NRDD	anaerobic ribonucleoside-triphosphate reductase
B4380	YJJI	conserved protein

5.8.7.3 *E. coli* bicluster 57 elaborated gene list

Locus	Name	Description
B0781	MOAA	molybdopterin biosynthesis protein A
B0902	PFLA	pyruvate formate lyase activating enzyme 1
B0904	FOCA	formate channel
B0972	HYAA	hydrogenase 1, small subunit
B0973	HYAB	hydrogenase 1, large subunit
B1465	NARV	nitrate reductase 2 (NRZ), gamma subunit
B1466	NARW	nitrate reductase 2 (NRZ), delta subunit
B1467	NARY	nitrate reductase 2 (NRZ), beta subunit
B1468	NARZ	nitrate reductase 2 (NRZ), alpha subunit
B1473	YDDG	aromatic amino acid exporter
B1593	YNFK	predicted dethiobiotin synthetase

B1670	YDHU	predicted cytochrome
B1671	YDHX	predicted 4Fe-4S ferridoxin-type protein
B2241	GLPA	sn-glycerol-3-phosphate dehydrogenase
B2242	GLPB	sn-glycerol-3-phosphate dehydrogenase
B2243	GLPC	sn-glycerol-3-phosphate dehydrogenase
B2579	YFID	autonomous glycyl radical cofactor
B2726	HYPA	protein involved in nickel insertion into
B2730	HYPE	carbamoyl dehydratase, hydrogenases 1,2,3
B2994	HYBC	hydrogenase 2, large subunit
B3365	NIRB	nitrite reductase, large subunit, NAD(P)H-binding
B3366	NIRD	nitrite reductase, NAD(P)H-binding, small
B3426	GLPD	sn-glycerol-3-phosphate dehydrogenase, aerobic, FAD/NAD(P)-binding
B4379	YJJW	predicted pyruvate formate lyase activating

5.8.8 S. typhimurium bicluster 57

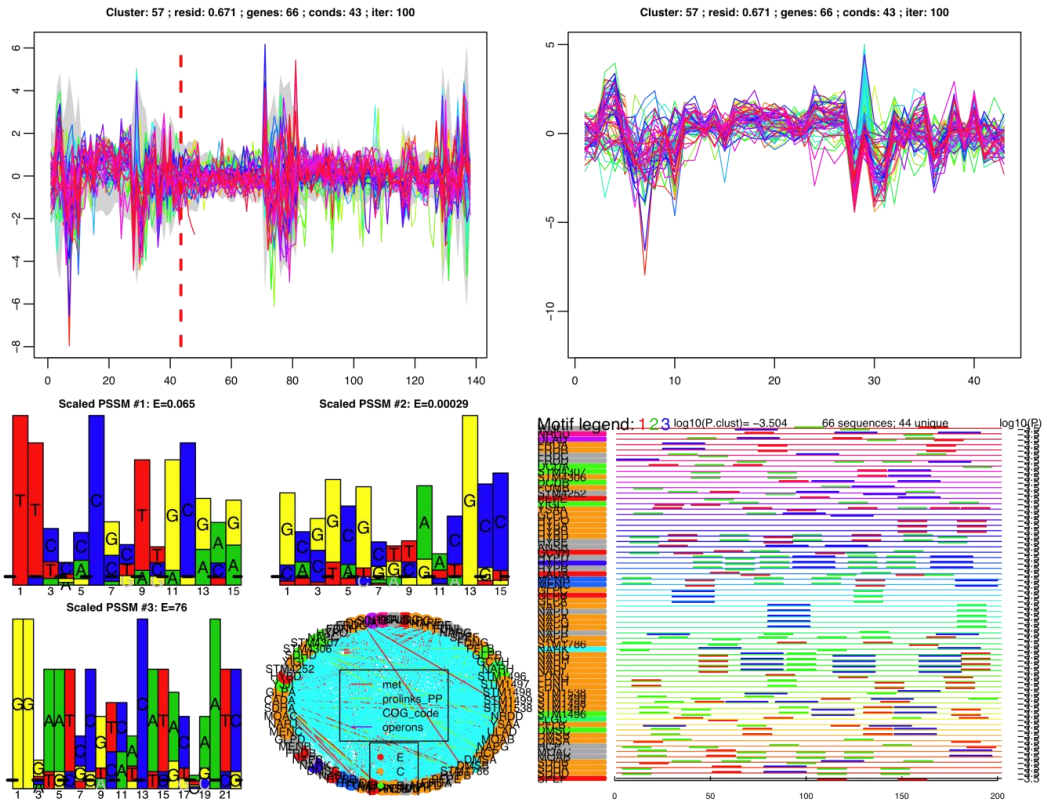


Figure 5.7: *S. typhimurium* bicluster 57 MScM output image

5.8.8.1 *S. typhimurium* bicluster 57 core gene list

Locus	Name	description
STM0701	SPEF	ornithine decarboxylase isozyme
STM0803	MOAB	molybdopterin biosynthetic protein B
STM0804	MOAC	molybdenum cofactor biosynthesis protein C

STM0937	HCP	hydroxylamine reductase
STM0964	DMSA	anaerobic dimethyl sulfoxide reductase subunit
STM0965	DMSB	anaerobic dimethyl sulfoxide reductase subunit
STM0966	DMSC	anaerobic dimethyl sulfoxide reductase subunit
STM0973	PFLB	pyruvate formate lyase I
STM1175	FLGC	flagellar basal body rod protein FlgC
STM1498		putative dimethyl sulphoxide reductase
STM1499		putative dimethyl sulphoxide reductase chain A1
STM1568	FDNI	formate dehydrogenase-N subunit gamma
STM1569	FDNH	formate dehydrogenase-N beta subunit
STM1570	FDNG	formate dehydorgenase-N alpha subunit
STM1761	NARI	nitrate reductase 1 gamma subunit
STM1762	NARJ	nitrate reductase 1 delta subunit
STM1763	NARH	nitrate reductase 1 beta subunit
STM1764	NARG	nitrate reductase 1 alpha subunit
STM1765	NARK	nitrite extrusion protein
STM2255	NAPC	cytochrome c-type protein NapC

STM2256	NAPB	diheme cytochrome c550
STM2257	NAPH	quinol dehydrogenase membrane component
STM2258	NAPG	quinol dehydrogenase periplasmic component
STM2259	NAPA	periplasmic nitrate reductase
STM2260	NAPD	assembly protein for periplasmic nitrate
STM2261	NAPF	ferredoxin-type protein
STM2306	MENC	O-succinylbenzoate synthase
STM2307	MENB	naphthoate synthase
STM3150	HYPO	hydrogenase 2 small subunit
STM2559	CADA	lysine decarboxylase 1
STM2855	HYPB	hydrogenase nickel incorporation protein HypB
STM2856	HYPC	hydrogenase isoenzymes formation protein
STM2857	HYPD	putative hydrogenase formation protein
STM3054	GCVH	glycine cleavage system protein H
STM3106	ANSB	L-asparaginase II
STM3145	HYBE	hydrogenase 2-specific chaperone
STM3146	HYBD	predicted maturation element for hydrogenase 2

STM3148	HYBB	predicted hydrogenase 2 cytochrome b type
STM3149	HYBA	hydrogenase 2 protein HybA
STM3666	YSAA	putative oxidoreductase
STM4190	PEPE	peptidase E
STM4300	FUMB	fumarase B
STM4301	DCUB	anaerobic C4-dicarboxylate transporter
STM4325	DCUA	anaerobic C4-dicarboxylate transporter
STM4340	FRDD	fumarate reductase subunit D
STM4341	FRDC	fumarate reductase subunit C
STM4342	FRDB	fumarate reductase iron-sulfur subunit
STM4343	FRDA	fumarate reductase flavoprotein subunit
STM4386	ULAD	3-keto-L-gulonate-6-phosphate decarboxylase
STM4452	NRDD	anaerobic ribonucleoside triphosphate reductase
STM4566	YJJI	hypothetical protein

5.8.8.2 *S. typhimurium* bicluster 57 elaborated gene list

Locus	Name	description

STM0733	SDHD	succinate dehydrogenase cytochrome b556 small
STM0734	SDHA	succinate dehydrogenase flavoprotein subunit
STM0735	SDHB	succinate dehydrogenase iron-sulfur subunit
STM1496	STM1496	putative dimethylsulfoxide reductase
STM1497	STM1497	putative dimethyl sulphoxide reductase
STM1538	STM1538	putative hydrogenase-1 large subunit
STM1786	STM1786	hydrogenase-1 small subunit
STM2284	GLPA	sn-glycerol-3-phosphate dehydrogenase subunit A
STM2285	GLPB	anaerobic glycerol-3-phosphate dehydrogenase
STM2286	GLPC	sn-glycerol-3-phosphate dehydrogenase subunit C
STM3526	GLPD	glycerol-3-phosphate dehydrogenase
STM3962	YIGL	predicted hydrolase
STM4252	STM4252	putative inner membrane protein
STM4306	STM4306	putative anaerobic dimethylsulfoxide reductase
STM4307	STM4307	putative anaerobic dimethylsulfoxide reductase

5.9 References

Alm, E. J., K. H. Huang, et al. (2005). "The MicrobesOnline Web site for comparative genomics." Genome Res **15**(7): 1015-1022.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Avila-Campillo, I., K. Drew, et al. (2007). "BioNetBuilder: automatic integration of biological networks." Bioinformatics **23**(3): 392-393.

Bare, J. C., P. T. Shannon, et al. (2007). "The Firegoose: two-way integration of diverse data from different bioinformatics web resources with desktop applications." BMC Bioinformatics **8**: 456.

Barrett, T. and R. Edgar (2006). "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis." Methods Enzymol **411**: 352-369.

Ben-Dor, A., B. Chor, et al. (2003). "Discovering local structure in gene expression data: the order-preserving submatrix problem." J Comput Biol **10**(3-4): 373-384.

Berg, J. and M. Lassig (2006). "Cross-species analysis of biological networks by Bayesian alignment." Proc Natl Acad Sci U S A **103**(29): 10967-10972.

Bergmann, S., J. Ihmels, et al. (2004). "Similarities and differences in genome-wide expression data of six organisms." PLoS Biol **2**(1): E9.

Bonneau, R., M. T. Facciotti, et al. (2007). "A predictive model for transcriptional control of physiology in a free living cell." Cell **131**(7): 1354-1365.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biol **5**(5): R35.

Brazma, A., H. Parkinson, et al. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res **31**(1): 68-71.

Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.

Chikina, M. D. and O. G. Troyanskaya (2011). "Accurate quantification of functional analogy among close homologs." PLoS Computational Biology **7**(2): e1001074.

Cline, M. S., M. Smoot, et al. (2007). "Integration of biological networks and gene expression data using Cytoscape." Nat Protoc **2**(10): 2366-2382.

Dehal, P. S., M. P. Joachimiak, et al. (2009). "MicrobesOnline: an integrated portal for comparative and functional genomics." Nucl. Acids Res.: gkp919.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

Elemento, O., N. Slonim, et al. (2007). "A universal framework for regulatory element discovery across all genomes and data types." Mol Cell **28**(2): 337-350.

Faith, J. J., M. E. Driscoll, et al. (2008). "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata." Nucleic Acids Res **36**(Database issue): D866-870.

Fink, R. C., M. R. Evans, et al. (2007). "FNR is a global regulator of virulence and anaerobic metabolism in Salmonella enterica serovar Typhimurium (ATCC 14028s)." J Bacteriol **189**(6): 2262-2273.

Fontecave, M., R. Eliasson, et al. (1989). "Oxygen-sensitive ribonucleoside triphosphate reductase is present in anaerobic Escherichia coli." Proc Natl Acad Sci U S A **86**(7): 2147-2151.

Gama-Castro, S., H. Salgado, et al. (2011). "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)." Nucleic Acids Res **39**(Database issue): D98-105.

Hubble, J., J. Demeter, et al. (2009). "Implementation of GenePattern within the Stanford Microarray Database." Nucleic Acids Res **37**(Database issue): D898-901.

Huttenhower, C., A. I. Flamholz, et al. (2007). "Nearest Neighbor Networks: clustering expression data based on gene neighborhoods." BMC Bioinformatics **8**: 250.

Ihmels, J., S. Bergmann, et al. (2004). "Defining transcription modules using large-scale gene expression data." Bioinformatics **20**(13): 1993-2003.

Ihmels, J., S. Bergmann, et al. (2005). "Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program." PLoS Genet **1**(3): e39.

Jensen, L. J., M. Kuhn, et al. (2009). "STRING 8--a global view on proteins and their functional interactions in 630 organisms." Nucl. Acids Res. **37**(suppl_1): D412-416.

Kanehisa, M., S. Goto, et al. (2002). "The KEGG databases at GenomeNet." Nucleic Acids Res **30**(1): 42-46.

Kazakov, A. E., M. J. Cipriano, et al. (2007). "RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes." Nucleic Acids Res **35**(Database issue): D407-412.

Kluger, Y., R. Basri, et al. (2003). "Spectral biclustering of microarray data: coclustering genes and conditions." Genome Res **13**(4): 703-716.

Kolker, E., K. S. Makarova, et al. (2004). "Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae." Nucleic Acids Res **32**(8): 2353-2361.

Li, G., Q. Ma, et al. (2009). "QUBIC: a qualitative biclustering algorithm for analyses of gene expression data." Nucleic Acids Res **37**(15): e101.

Lu, Y., P. Huggins, et al. (2009). "Cross species analysis of microarray expression data." Bioinformatics **25**(12): 1476-1483.

Maglott, D., J. Ostell, et al. (2005). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **33**(Database issue): D54-58.

Malmstrom, L., M. Riffle, et al. (2007). "Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology." PLoS Biol **5**(4): e76.

Mellor, J. C., I. Yanai, et al. (2002). "Predictome: a database of putative functional links between proteins." Nucleic Acids Res **30**(1): 306-309.

Parkinson, H., M. Kapushesky, et al. (2009). "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression." Nucleic Acids Res **37**(Database issue): D868-872.

Poultney, C. S., R. A. Gutierrez, et al. (2007). "Sungear: interactive visualization and functional analysis of genomic datasets." Bioinformatics **23**(2): 259-261.

Prelic, A., S. Bleuler, et al. (2006). "A systematic comparison and evaluation of biclustering methods for gene expression data." Bioinformatics **22**(9): 1122-1129.

Price, M. N., K. H. Huang, et al. (2005). "MicrobesOnline Operon Predictions for Escherichia coli str. K-12 substr. MG1655." from http://www.microbesonline.org/operons/gnc511145.html.

Reiss, D. J., N. S. Baliga, et al. (2006). "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." BMC Bioinformatics **7**: 280.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-1052.

Riffle, M., L. Malmstrom, et al. (2005). "The Yeast Resource Center Public Data Repository." Nucleic Acids Res **33**(Database issue): D378-382.

Rudd, K. E. (2000). "EcoGene: a genome sequence database for Escherichia coli K-12." Nucleic Acids Res **28**(1): 60-64.

Shannon, P. T., D. J. Reiss, et al. (2006). "The Gaggle: an open-source software system for integrating bioinformatics software and data sources." BMC Bioinformatics **7**: 176.

Sherlock, G., T. Hernandez-Boussard, et al. (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-155.

Snel, B., G. Lehmann, et al. (2000). "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." Nucl. Acids Res. **28**(18): 3442-3444.

Stanley, J. T., R. P. Gunsalus, et al. (2007). Biosynthesis of Monomers, Nitrogen Assimilation. Microbial Life. J. T. Stanley. Sunderland, MA, Sinauer Associates Inc.**:** ???

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-255.

Supper, J., M. Strauch, et al. (2007). "EDISA: extracting biclusters from multiple time-series of gene expression profiles." BMC Bioinformatics **8**: 334.

Tanay, A., A. Regev, et al. (2005). "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast." Proc Natl Acad Sci U S A **102**(20): 7203-7208.

Tanay, A., R. Sharan, et al. (2004). "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proc Natl Acad Sci U S A **101**(9): 2981-2986.

Tanay, A., R. Sharan, et al. (2002). "Discovering statistically significant biclusters in gene expression data." Bioinformatics **18 Suppl 1**: S136-144.

Thomas-Chollier, M., O. Sand, et al. (2008). "RSAT: regulatory sequence analysis tools." Nucleic Acids Res **36**(Web Server issue): W119-127.

Tirosh, I. and N. Barkai (2007). "Comparative analysis indicates regulatory neofunctionalization of yeast duplicates." Genome Biol **8**(4): R50.

Tirosh, I., Y. Bilu, et al. (2007). "Comparative biology: beyond sequence analysis." Curr Opin Biotechnol **18**(4): 371-377.

Ulitsky, I. and R. Shamir (2007). "Identification of functional modules using network topology and high-throughput data." BMC Syst Biol **1**: 8.

van Helden, J. (2003). "Regulatory sequence analysis tools." Nucleic Acids Res **31**(13): 3593-3596.

Waltman, P., T. Kacmarczyk, et al. (2010). "Multi-species integrative biclustering." Genome Biololgy **11**(R96).

Waltman, P., T. K. Kuppusamy, et al. (2010). "cMonkey2." from http://ms2.bio.nyu.edu/cMonkey2-trac/.

Winteler, H. V. and D. Haas (1996). "The homologous regulators ANR of Pseudomonas aeruginosa and FNR of Escherichia coli have overlapping but distinct specificities for anaerobically inducible promoters." Microbiology **142 (Pt 3)**: 685-693.

6. DISCUSSION AND FUTURE DIRECTIONS

We present above a novel method for identifying modules of functionally conserved genes that identifies conserved modules by integrating multiple sources of data from multiple organisms simultaneously. This is in comparison to sequence-based comparison methods such as the Clusters of Orthologous Groups (COG) database (Tatusov, Galperin et al. 2000) and InParanoid (Remm, Storm et al. 2001), as well as several other more recent methods (Chen, Mackey et al. 2006; DeLuca, Wu et al. 2006; Schneider, Dessimoz et al. 2007). However, the analysis provided by these sequence-based methods aims to identify clusters of orthology, based upon homology, not co-regulation. Even more recent methods augment these sequence-based methods by seeking to identify functional orthology via the comparison of conserved protein-protein interaction (PPI) networks – so-called network homology (Flannick, Novak et al. 2006; Kalaev, Smoot et al. 2008; Singh, Xu et al. 2008; Liao, Lu et al. 2009; Park, Singh et al. 2011; Zinman, Zhong et al. 2011). Of these, only one (Zinman, Zhong et al. 2011) explicitly includes co-expression data in its analysis by including edges based on correlation, though, it would be relatively easy for the other methods to include these as well. In any such analysis, obviously, great care must be taken to avoid circularity of reasoning, which would be possible for example by including interolog edges (Yu, Luscombe et al. 2004). As described above in section 2.1, methods that are based primarily upon gene expression data can be segmented into two classes – those that seek to find matching conditions between the organisms and

those that do not. We direct the reader to the section 2.1 for a more thorough discussion of this.

Both of the newer approaches (those based primarily upon PPI networks, versus those based primarily upon expression data) have their strengths and weaknesses. Interaction-based function orthology methods provide the confidence of clustering orthologous genes based on known or putative interactions, though known interaction provide far higher degree of confidence. However, the sparsity of these interactions for most organisms, especially those that are high-confidence, is a limiting factor on the genes that can be analyzed. Correlation-based edges provide one possible workaround to this limit for interaction-based methods, however, the calculations made when generating these do not easily allow for condition-dependence, as is possible with biclustering methods. Similarly, most expression-based comparative, multi-species methods allow for comprehensive, genome-wide analyses. However, they too often fail to allow for condition-dependent expression patterns. Finally, all these methods focus on the identification of conserved modules, but don't provide mechanisms for identifying species-specific changes or modification to these.

In comparison to these other methods, multi-species cMonkey (MScM) allows for the integration of interactions or associations with sequence and co-expression data, while allowing for condition-dependence with the expression data. In addition, MScM is explicitly written to identify species-specific difference to these conserved modules. However, MScM is not without its areas for potential improvement as well.

For example, the heuristic, correlation-based seeding strategy that MScM currently uses (described in section 2.5.1.4) could be replaced by one of these network-homology based methods. An alternate strategy would be to use the balanced, multi-species k-means algorithm to generate the initial seeds. While this would provide complete coverage, at least initially, this will also require a decision over how to handle the initial k clusters that are generated. For example, would the search strategy need to be modified to process the initial k biclusters sequentially – as is currently done – or concurrently? This also raises the question of how or whether to prune some of the initial k biclusters from consideration. One possible idea would be to use a consensus approach that employs both the network homology and k-means clustering approaches together to determine an initial set of high-confidence seeds that can be further optimized, using the MCMC search that is used by MScM.

While a change to the seeding strategy could be fruitful, there are several aspects to MScM that definitely deserve further attention, with two that should take priority. The first of these two aspects that deserve the most scrutiny are the integration or 'mixing' parameters that are used in the joint-likelihood function which is used to estimate the likelihood of a gene belonging to a bicluster. Currently these are set by a schedule that gives the sequence support little weight in the beginning, and slowly increases it during the optimization of a single bicluster. Similarly, the weight given to the network/interaction support starts off relatively high, and is progressively reduced during the optimization of a single cluster. However, despite these general

strategies, there are many details that are not well-established yet, e.g. the exact weights at the beginning or end are not well-established, nor the rate at which the increase or decrease should occur. It is unclear at the time of this writing what would be an optimal method to determine these, as any such method to determine these would need to allow users to specify the maximum weight they want given to a particular data type, while still allowing MScM to determine the optimal weighting of the different data types.

However, this discussion naturally raises the question of what is optimal in an unsupervised learning environment such as the one with co-expression data. While we provide in Chapter 3 a through comparison of MScM with several other methods, this is time-consuming and for this reason, does not lend itself well to an optimization problem. The reason for this, as we discuss in Chapter 3 is that there is no complete 'gold-standard' clustering result that can be used to benchmark the results.

This also becomes an issue in terms of what should be considered to be the 'correct' number of biclusters that are generated by MScM, which would be the second major aspect of MScM that deserves further attention. However, without knowing *a priori* how many functional modules are active in an expression data set, it is difficult to determine the appropriate number of biclusters to generate. Currently, MScM estimates the number of biclusters that will be generated based on the number of genes in the expression data set, divided by the estimated mean size of each bicluster. One possible heuristic that could be used to address this is to have MScM

make a decision based on the latest bicluster that is generated to determine if it will continue to search for new biclusters. In this case, it would consider the number of previously unclustered genes or percentage of the expression data that is added by the latest results. If the number of new genes or areas of the expression matrix that are added by MScM falls below a certain threshold, then MScM could then decide to stop adding new biclusters to the set of those that it has identified.

An alternate strategy would be to have MScM identify far more biclusters that are expected, and in a post-processing step, MScM would prune these down to a minimal set. In any solution such as this, MScM would need to make a decision amongst numerous, overlapping bicluster/modules. This, in turn, raises the question of what one should consider to be meaningful overlap between the modules, i.e. which biclusters are sub-modules of others?; which biclusters reflect pleiotropy?; which are simply the method settling into the same optima multiple times?

One naïve solution to this question would be to use a threshold to determine sufficient similarity or difference between biclusters (a threshold on the percentage of the sub-matrices that overlap). In addition, the non-expression data types such as sequence motifs and association edges could also be helpful in determining similarity and difference. At this point, it is unclear what would be the optimal solution, but this could be a fruitful area for further research. GO term and/or KEGG pathway annotations of the non-overlapping genes in a bicluster could also provide some

guidance as well; as would comparisons of the conditions that are included in the different bicluster.

This question of how to determine similarity and difference between biclusters is essential when evaluating the stability of the method, where by stability we mean the consistency and reproducibility of the biclustering results between different analyses performed by the same method, on the same data set – an issue for all Monte Carlo methods, including MScM. As of yet, no attempt has been made to quantify the stability of MScM (nor SScM), though, anecdotally, in our experience, the modules with clearest signal are retrieved consistently between different runs, though, it is unknown at this point how reproducible are those modules with more subtle signals.

Ultimately, this question of stability should also need to be considered when determining the optimality of the parameterization of any given MScM run. However, as we state in section 2.3 "[w]e have shown that MScM provides better or comparable coverage, functional enrichment scores, bicluster coherence, and conservation than other tested methods, with all other methods failing in one of the main categories of assessment. Furthermore, our method effectively balances the influence of each organism, preventing organisms with more complete datasets from dominating the analysis, while also integrating other supporting data types, enabling the method to identify more biologically relevant modules and delimit the conditions over which the modules are active. The fact that the MScM biclusters have many fold higher conservation scores than several of the tested methods suggests that they have a higher

level of biological significance than equally co-expressed (and/or equally functionally enriched) non-conserved alternate biclusters. An analysis that takes into account several validation metrics supports the idea that MScM is the top performing method for comparative biclustering." Thus, despite these open questions and limitations, MScM provides a robust and novel solution to the comparative analyses of multiple-species, by providing analytical aspects that no other method yet provides.

6.1 References

Chen, F., A. J. Mackey, et al. (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." Nucleic Acids Research **34**(suppl 1): D363-D368.

DeLuca, T. F., I.-H. Wu, et al. (2006). "Roundup: a multi-genome repository of orthologs and evolutionary distances." Bioinformatics **22**(16): 2044-2046.

Flannick, J., A. Novak, et al. (2006). "Graemlin: general and robust alignment of multiple large interaction networks." Genome Research **16**(9): 1169-1181.

Kalaev, M., M. Smoot, et al. (2008). "NetworkBLAST: comparative analysis of protein networks." Bioinformatics **24**(4): 594-596.

Liao, C. S., K. Lu, et al. (2009). "IsoRankN: spectral methods for global alignment of multiple protein networks." Bioinformatics **25**(12): i253-258.

Park, D., R. Singh, et al. (2011). "IsoBase: a database of functionally related proteins across PPI networks." Nucleic Acids Research **39**(Database issue): D295-300.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-1052.

Schneider, A., C. Dessimoz, et al. (2007). "OMA Browser—Exploring orthologous relations across 352 complete genomes." Bioinformatics **23**(16): 2180-2182.

Singh, R., J. Xu, et al. (2008). "Global alignment of multiple protein interaction networks with application to functional orthology detection." Proc Natl Acad Sci U S A **105**(35): 12763-12768.

Tatusov, R. L., M. Y. Galperin, et al. (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Research **28**(1): 33-36.

Yu, H., N. M. Luscombe, et al. (2004). "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs." Genome Research **14**(6): 1107-1118.

Zinman, G., S. Zhong, et al. (2011). "Biological interaction networks are conserved at the module level." BMC Systems Biology **5**(1): 134.

7. SUPPLEMENTARY INFORMATION

7.1 Gene lists and bicluster images of the biological highlights for the Gram-positive triplet

7.1.1 Full descriptions of highlighted biclusters

7.1.1.1 Gene lists for *B. subtilis, B. anthracis* Sterne sporulation clusters 32, 82, and 84.

7.1.1.1.1 B. subtilis - B. anthracis cluster 32

7.1.1.1.1.1 *B. subtilis* cluster 32

Figure 7.1: *B. subtilis* cluster 32 image (post-elaboration)

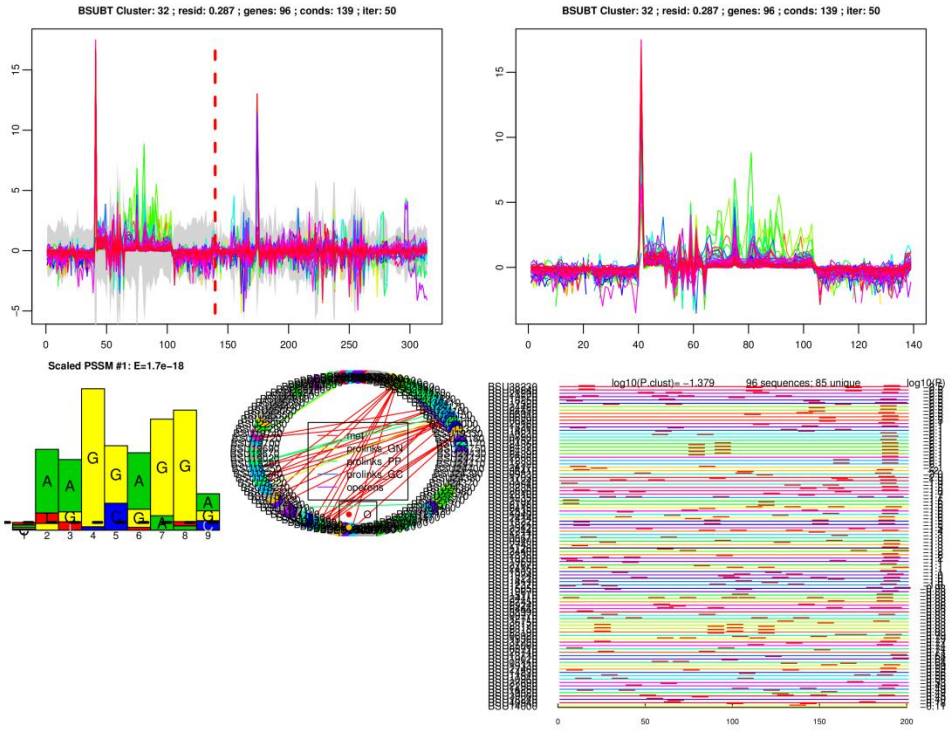


Table 7.1: *B. subtilis* cluster 32 core genes

Locus	Name	Function
BSU06890	*cotJA*	polypeptide composition of the spore coat; required for the assembly of CotJC
BSU06900	*cotJB*	polypeptide composition of the spore coat
BSU06910	*cotJC*	polypeptide composition of the spore coat
BSU23190	*dacB*	D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 5*)
BSU28380	*gerM*	germination (cortex hydrolysis) and sporulation (stage II, multiple polar septa)
BSU27440	*glnH*	glutamine ABC transporter (glutamine-binding protein)
BSU27450	*glnM*	glutamine ABC transporter (integral membrane protein)
BSU27460	*glnP*	glutamine ABC transporter (integral membrane protein)
BSU27430	*glnQ*	glutamine ABC transporter (ATP-binding protein)
BSU15320	*sigE*	sporulation sigma factor SigE

BSU23180	*spmA*	spore maturation protein
BSU23170	*spmB*	spore maturation protein
BSU24430	*spoIIIAA*	mutants block sporulation after engulfment (stage III sporulation)
BSU24420	*spoIIIAB*	stage III sporulation protein SpoAB
BSU24410	*spoIIIAC*	mutants block sporulation after engulfment (stage III sporulation)
BSU24400	*spoIIIAD*	mutants block sporulation after engulfment (stage III sporulation)
BSU24390	*spoIIIAE*	mutants block sporulation after engulfment (stage III sporulation)
BSU24380	*spoIIIAF*	mutants block sporulation after engulfment (stage III sporulation)
BSU24370	*spoIIIAG*	mutants block sporulation after engulfment (stage III sporulation)
BSU24360	*spoIIIAH*	mutants block sporulation after engulfment (stage III sporulation)
BSU27980	*spoIVFA*	inhibition of SpoIVFB (negative regulation) and hypothesised to stabilize the thermolabile SpoIVFB product (positive regulation) (stage IV sporulation)
BSU27970	*spoIVFB*	membrane metalloprotease
BSU27670	*spoVB*	involved in spore cortex synthesis (stage V sporulation)
BSU01570	*ybaN*	hypothetical protein
BSU09940	*yhaL*	hypothetical protein
BSU11510	*yjbE*	hypothetical protein
BSU14110	*ykuK*	hypothetical protein
BSU13710	*ykvI*	hypothetical protein
BSU15030	*ylbJ*	hypothetical protein
BSU15650	*yloB*	hypothetical protein
BSU25350	*yqfD*	hypothetical protein
BSU25060	*yqfZ*	hypothetical protein
BSU24440	*yqhV*	hypothetical protein
BSU27690	*yrzE*	hypothetical protein

BSU28100	*ysxE*	hypothetical protein
BSU29240	*ytrI*	hypothetical protein
BSU28960	*ytxC*	hypothetical protein
BSU32350	*yunB*	hypothetical protein

Table 7.2: *B. subtilis* cluster 32 elaboration genes

Locus	Name	Function
BSU17260	*aprX*	alkaline serine protease
BSU17030	*cotE*	morphogenic protein
BSU26740	*cypA*	cytochrome P450-like enzyme
BSU12370	*exuR*	transcriptional regulator (LacI family)
BSU19690	*kamA*	lysine 2,3-aminomutase
BSU36410	*mbl*	MreB-like protein
BSU24170	*mmgA*	acetyl-CoA acetyltransferase
BSU24160	*mmgB*	3-hydroxybutyryl-CoA dehydrogenase
BSU24150	*mmgC*	acyl-CoA dehydrogenase
BSU24140	*mmgD*	citrate synthase 3
BSU14000	*patA*	aminotransferase A
BSU38990	*scoA*	succinyl CoA:3-oxoacid CoA-transferase (subunit A)
BSU19330	*sodF*	superoxide dismutase
BSU36750	*spoIID*	required for complete dissolution of the asymmetric septum (stage II sporulation)
BSU15170	*spoVD*	penicillin-binding protein
BSU09400	*spoVR*	involved in spore cortex synthesis (stage V sporulation)
BSU37830	*spsJ*	spore coat polysaccharide synthesis

BSU19320	*sqhC*	squalene-hopene cyclase
BSU12350	*yjmF*	D-mannonate oxidoreductase
BSU14830	*ylaM*	glutaminase
BSU18220	*yngF*	enoyl-CoA hydratase
BSU18230	*yngG*	hydroxymethylglutaryl-CoA lyase
BSU18240	*yngH*	acetyl-CoA carboxylase biotin carboxylase subunit
BSU18250	*yngI*	acyl-CoA synthetase
BSU12710	*xkdR*	hypothetical protein
BSU00160	*yaaH*	hypothetical protein
BSU03110	*ycgH*	hypothetical protein
BSU03670	*yclF*	hypothetical protein
BSU05710	*ydhD*	hypothetical protein
BSU06920	*yesJ*	hypothetical protein
BSU09830	*yhaX*	hypothetical protein
BSU08980	*yhbH*	hypothetical protein
BSU09770	*yheD*	hypothetical protein
BSU10230	*yhfH*	hypothetical protein
BSU10400	*yhxC*	hypothetical protein
BSU10960	*yitE*	hypothetical protein
BSU12110	*yjfA*	hypothetical protein
BSU12320	*yjmC*	hypothetical protein
BSU12330	*yjmD*	hypothetical protein
BSU14250	*yknT*	hypothetical protein
BSU14810	*ylaK*	hypothetical protein
BSU17320	*ymaF*	hypothetical protein

BSU18210	*yngE*	hypothetical protein
BSU18260	*yngJ*	hypothetical protein
BSU19020	*yobN*	hypothetical protein
BSU19670	*yodN*	hypothetical protein
BSU19700	*yodP*	hypothetical protein
BSU19720	*yodR*	hypothetical protein
BSU19740	*yodT*	hypothetical protein
BSU21290	*yomN*	hypothetical protein
BSU22980	*ypbG*	hypothetical protein
BSU25420	*yqeW*	hypothetical protein
BSU26660	*yrdN*	hypothetical protein
BSU29160	*ytvI*	hypothetical protein
BSU31740	*yuxH*	hypothetical protein
BSU31730	*yuzC*	hypothetical protein
BSU38240	*ywcA*	hypothetical protein
BSU38230	*ywcB*	hypothetical protein
BSU39000	*yxjC*	hypothetical protein
BSU40940	*yyaD*	hypothetical protein

7.1.1.1.1.2 B. anthracis cluster 32

Figure 7.2: *B. anthracis* cluster 32 image (post-elaboration)

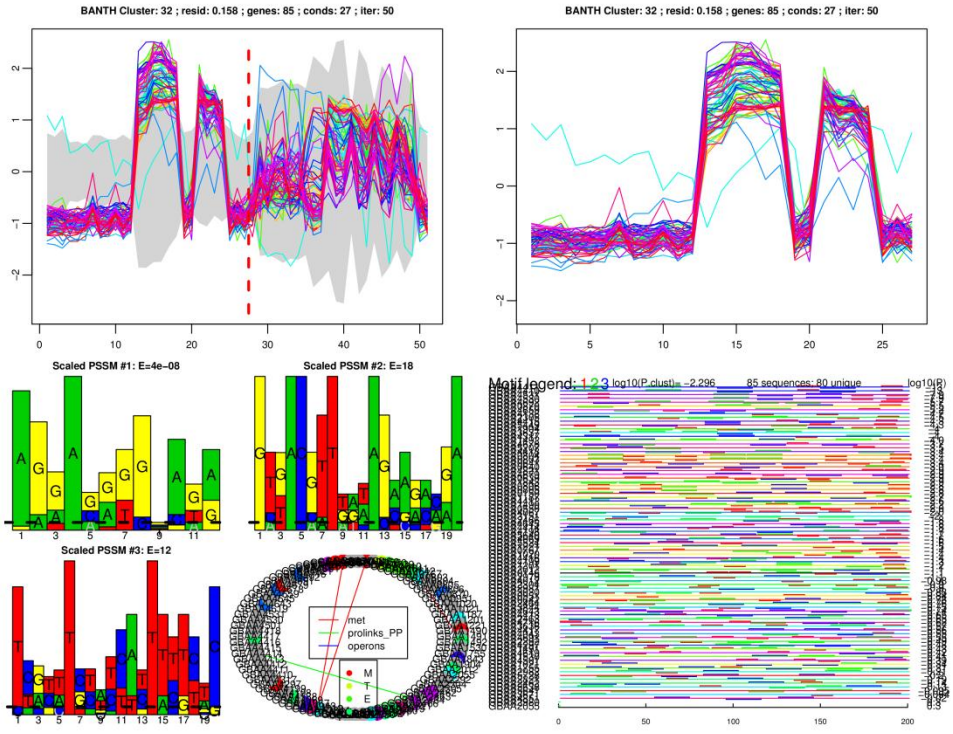


Table 7.3: *B. anthracis* cluster 32 core genes

Locus	Name	Function
GBAA0805	*cotJA*	cotja protein
GBAA0804	*cotJB*	cotjb protein
GBAA0803	*cotJC*	cotjc protein
GBAA4716	*gerM*	germination protein gerM
GBAA4043	*sigE*	sporulation sigma factor SigE
GBAA1491	*spmA*	spore maturation protein a
GBAA4417	*spoIIIAA*	stage iii sporulation protein aa

GBAA4416	*spoIIIAB*	stage III sporulation protein SpoAB
GBAA4415	*spoIIIAC*	stage iii sporulation protein ac
GBAA4414	*spoIIIAD*	stage iii sporulation protein ad
GBAA4413	*spoIIIAE*	stage iii sporulation protein ae
GBAA4412	*spoIIIAF*	stage iii sporulation protein af
GBAA4411	*spoIIIAG*	stage iii sporulation protein ag
GBAA4679	*spoIVFA*	stage iv sporulation protein fa
GBAA4678	*spoIVFB*	stage iv sporulation protein fb
GBAA4643	*spoVB*	stage v sporulation protein b
GBAA0640	-	amino acid abc transporter, amino acid-binding protein
GBAA0639	-	amino acid abc transporter, atp-binding protein
GBAA0641	-	amino acid abc transporter, permease protein
GBAA0642	-	amino acid abc transporter, permease protein
GBAA4012	-	cation-transporting atpase, e1-e2 family
GBAA1490	-	d-alanyl-d-alanine carboxypeptidase family protein
GBAA0150	-	polysaccharide deacetylase, putative
GBAA1492	-	spore maturation protein
GBAA4530	-	sporulation protein
GBAA4410	-	stage iii sporulation protein ah
GBAA1020	-	hypothetical protein
GBAA1201	-	hypothetical protein
GBAA2012	-	hypothetical protein
GBAA4138	-	hypothetical protein
GBAA4198	-	hypothetical protein
GBAA4418	-	hypothetical protein

GBAA4501	-	hypothetical protein
GBAA4645	-	hypothetical protein
GBAA4691	-	hypothetical protein
GBAA4821	-	hypothetical protein
GBAA4851	-	hypothetical protein
GBAA5207	-	hypothetical protein

Table 7.4: *B. anthracis* cluster 32 elaboration genes

Locus	Name	Function
GBAA5449	*celA-3*	pts system, cellobiose-specific iib component
GBAA0146	*cwlD*	germination-specific n-acetylmuramoyl-l-alanine amidase
GBAA5640	*cwlJ-2*	cell wall hydrolase
GBAA4297	*dacF*	d-alanyl-d-alanine carboxypeptidase
GBAA1530	*spoIVA*	stage iv sporulation protein a
GBAA0767	*spoVR*	stage v sporulation protein r
GBAA1221	-	bacteriocin o-metyltransferase, putative
GBAA1755	-	bnr repeat domain protein
GBAA3030	-	catalase
GBAA3668	-	glycosyl hydrolase, family 18
GBAA0870	-	hydrolase, haloacid dehalogenase-like family
GBAA4659	-	lysm domain protein
GBAA2055	-	magnesium transporter, cora family
GBAA2980	-	polyketide synthesis domain protein
GBAA2981	-	polyketide synthesis domain protein
GBAA4067	-	prophage lambdaba02, ftsk/spoiiie family protein

GBAA2462	-	pts system, cellobiose-specific iib component, putative
GBAA2463	-	pts system, cellobiose-specific iic component, putative
GBAA5524	-	stage ii sporulation protein
GBAA4692	-	stage vi sporulation protein d, putative
GBAA2979	-	transcriptional regulator, putative
GBAA0550	-	hypothetical protein
GBAA0806	-	hypothetical protein
GBAA0951	-	hypothetical protein
GBAA1021	-	hypothetical protein
GBAA1187	-	hypothetical protein
GBAA1843	-	hypothetical protein
GBAA1904	-	hypothetical protein
GBAA2292	-	hypothetical protein
GBAA2304	-	hypothetical protein
GBAA2305	-	hypothetical protein
GBAA2464	-	hypothetical protein
GBAA2466	-	hypothetical protein
GBAA2821	-	hypothetical protein
GBAA2982	-	hypothetical protein
GBAA3151	-	hypothetical protein
GBAA3636	-	hypothetical protein
GBAA3637	-	hypothetical protein
GBAA3638	-	hypothetical protein
GBAA3671	-	hypothetical protein
GBAA3844	-	hypothetical protein

GBAA4069	-	hypothetical protein
GBAA4199	-	hypothetical protein
GBAA4317	-	hypothetical protein
GBAA4531	-	hypothetical protein
GBAA4619	-	hypothetical protein
GBAA5641	-	hypothetical protein
GBAA5728	-	hypothetical protein

7.1.1.1.2 B. subtilis - B. anthracis cluster 82

7.1.1.1.2.1 *B. subtilis* cluster 82

Figure 7.3: *B. subtilis* cluster 82 image (post-elaboration)

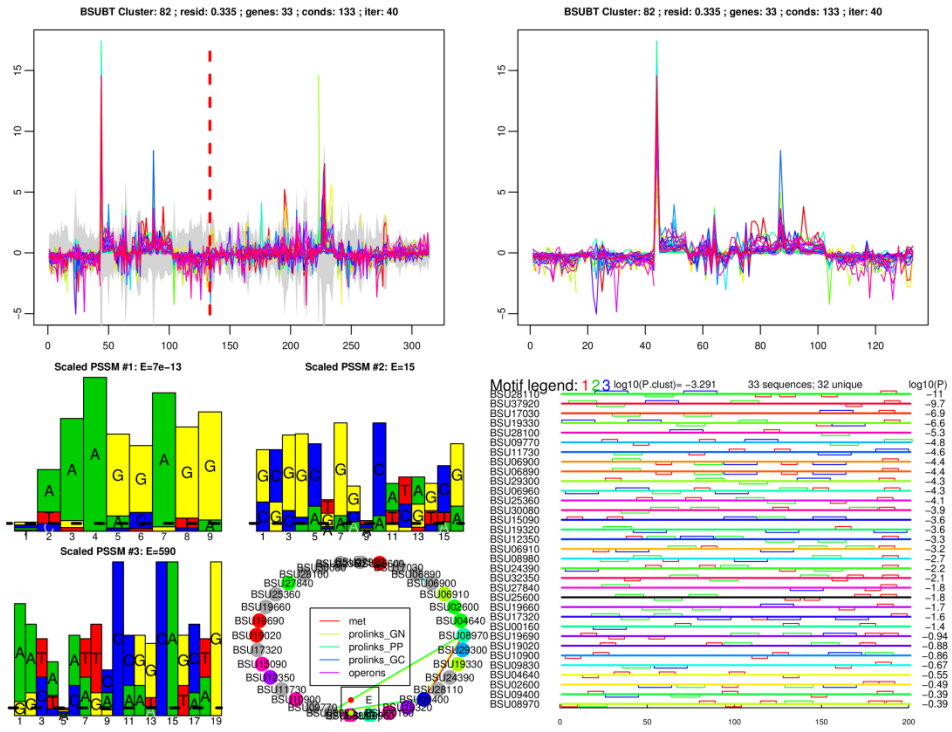


Table 7.5: B. subtilis cluster 82 core genes

Locus	Name	Function
BSU04640	*alr*	D-alanine racemase
BSU25600	*comER*	late competence protein ComER
BSU17030	*cotE*	morphogenic protein (spore coat protein)
BSU06890	*cotJA*	polypeptide composition of the spore coat; required for the assembly of CotJC
BSU06900	*cotJB*	polypeptide composition of the spore coat
BSU06910	*cotJC*	polypeptide composition of the spore coat
BSU02600	*cwlJ*	cell wall hydrolase (stored in the spore coat)
BSU08970	*prkA*	serine protein kinase
BSU27840	*safA*	morphogenetic protein associated with SpoVID (spore coat protein)

BSU19330	*sodF*	superoxide dismutase
BSU28110	*spoVID*	required for assembly of the spore coat (stage VI sporulation)
BSU09400	*spoVR*	involved in spore cortex synthesis (stage V sporulation)
BSU19320	*sqhC*	squalene-hopene cyclase
BSU00160	*yaaH*	hypothetical protein (spore coat protein)
BSU09830	*yhaX*	hypothetical protein (spore coat protein)
BSU08980	*yhbH*	hypothetical protein
BSU10900	*yisY*	hypothetical protein (spore coat protein)
BSU11730	*yjbX*	hypothetical protein (spore coat protein, cotO)
BSU19020	*yobN*	hypothetical protein
BSU19660	*yozD*	hypothetical protein
BSU25360	*yqfC*	hypothetical protein
BSU28100	*ysxE*	hypothetical protein (spore coat protein)
BSU30080	*yteV*	hypothetical protein
BSU32350	*yunB*	hypothetical protein
BSU37920	*ywdL*	hypothetical protein (spore coat protein, gerQ)

Table 7.6: *B. subtilis* cluster 82 elaboration genes

Locus	Name	Function
BSU19690	*kamA*	lysine 2,3-aminomutase
BSU29300	*ribR*	riboflavin kinase
BSU24390	*spoIIIAE*	mutants block sporulation after engulfment (stage III sporulation)
BSU12350	*yjmF*	D-mannonate oxidoreductase
BSU06960	*yesN*	hypothetical protein
BSU09770	*yheD*	hypothetical protein

BSU15090	*ylbO*	hypothetical protein
BSU17320	*ymaF*	hypothetical protein

Note: spore coat protein assignments are from Henriques and Moran (Henriques and Moran 2007).

7.1.1.1.2.2 B. anthracis cluster 82

Figure 7.4: *B. anthracis* cluster 82 image (post-elaboration)

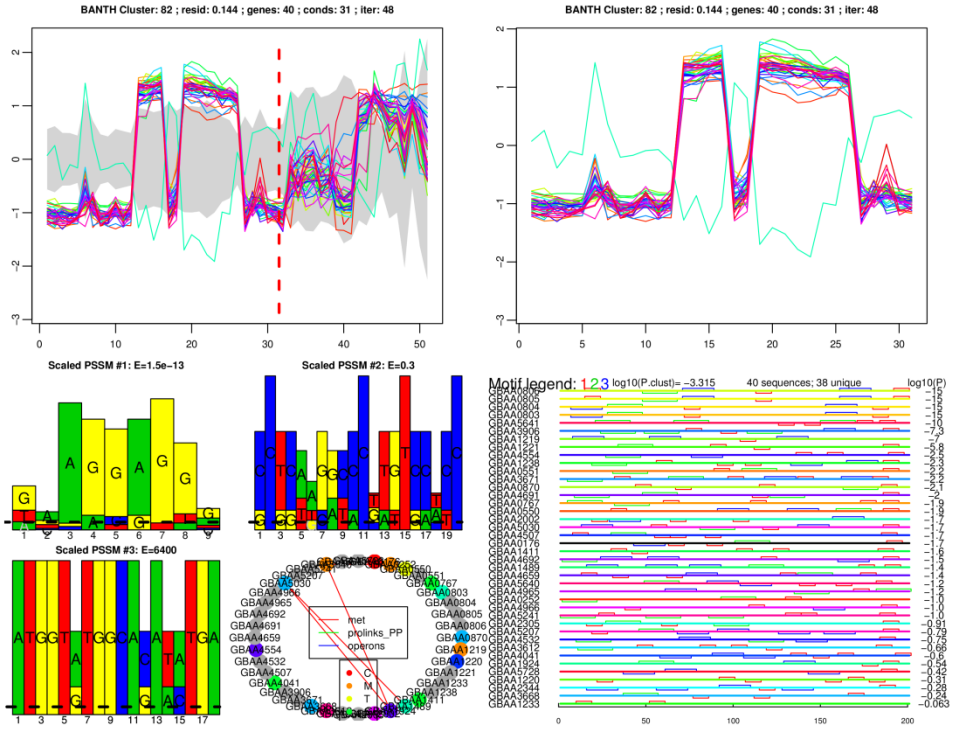


Table 7.7: *B. anthracis* cluster 82 core genes

Locus	Name	Function

GBAA3906	*cotE*	spore coat protein e
GBAA0805	*cotJA*	cotja protein
GBAA0804	*cotJB*	cotjb protein
GBAA0803	*cotJC*	cotjc protein
GBAA5640	*cwlJ-2*	cell wall hydrolase
GBAA0252	*dal-1*	alanine racemase
GBAA0767	*spoVR*	stage v sporulation protein r
GBAA1924	-	amine oxidase, flavin-containing
GBAA3668	-	glycosyl hydrolase, family 18
GBAA5030	-	hydrolase, alpha/beta fold family
GBAA0870	-	hydrolase, haloacid dehalogenase-like family
GBAA4554	-	late competence protein ComER
GBAA4659	-	lysm domain protein
GBAA3612	-	squalene-hopene cyclase
GBAA4692	-	stage vi sporulation protein d, putative
GBAA1489	-	superoxide dismutase
GBAA0550	-	hypothetical protein
GBAA0551	-	hypothetical protein
GBAA1233	-	hypothetical protein
GBAA2305	-	hypothetical protein
GBAA4532	-	hypothetical protein
GBAA4691	-	hypothetical protein
GBAA4965	-	hypothetical protein
GBAA5207	-	hypothetical protein
GBAA5641	-	hypothetical protein

Table 7.8: *B. anthracis* cluster 82 elaboration genes

Locus	Name	Function
GBAA1238	*cotZ-2*	spore coat protein z
GBAA0176	-	alcohol dehydrogenase, zinc-containing
GBAA1221	-	bacteriocin o-metyltransferase, putative
GBAA1219	-	glycosyl transferase, group 2 family protein
GBAA5241	-	spore coat protein f-related protein
GBAA2002	-	transcriptional regulator, arsr family
GBAA0806	-	hypothetical protein
GBAA1220	-	hypothetical protein
GBAA1411	-	hypothetical protein
GBAA2344	-	hypothetical protein
GBAA3671	-	hypothetical protein
GBAA4041	-	hypothetical protein
GBAA4507	-	hypothetical protein
GBAA4966	-	hypothetical protein
GBAA5728	-	hypothetical protein

7.1.1.1.3 B. subtilis - B. anthracis cluster 84

7.1.1.1.3.1 *B. subtilis* cluster 84

Figure 7.5: *B. subtilis* cluster 84 image (post-elaboration)

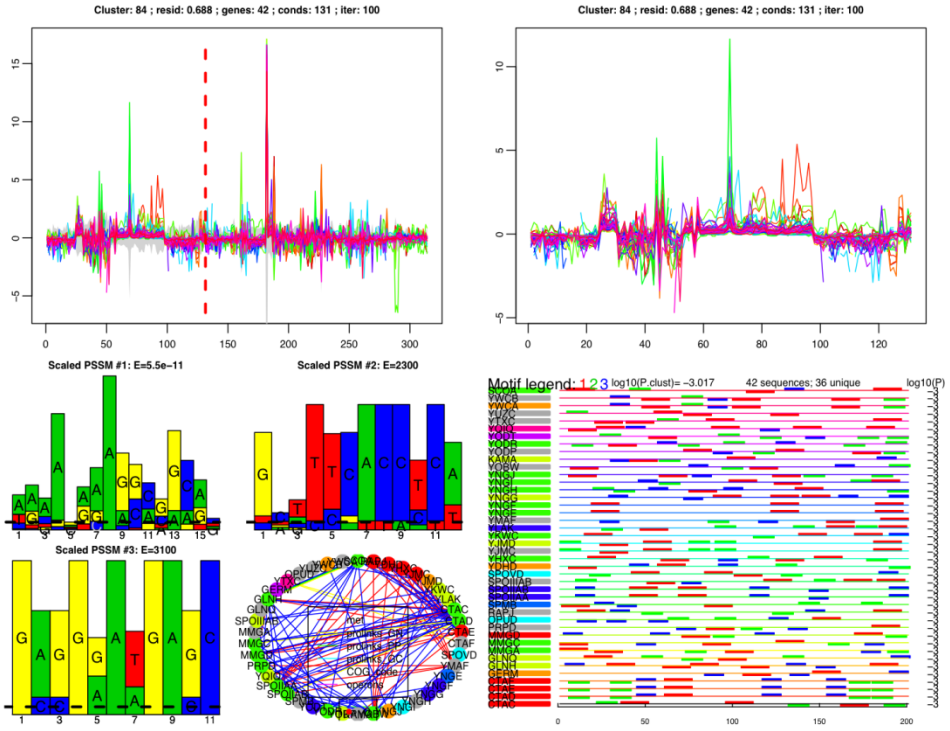


Table 7.9: *B. subtilis* cluster 84 core genes

Locus	Name	Function
BSU14890	*ctaC*	cytochrome caa3 oxidase (subunit II)
BSU14900	*ctaD*	cytochrome caa3 oxidase (subunit I)
BSU14910	*ctaE*	cytochrome caa3 oxidase (subunit III)
BSU14920	*ctaF*	cytochrome caa3 oxidase (subunit IV)
BSU19690	*kamA*	lysine 2,3-aminomutase

BSU24170	*mmgA*	acetyl-CoA acetyltransferase
BSU24150	*mmgC*	acyl-CoA dehydrogenase
BSU24140	*mmgD*	citrate synthase 3
BSU30070	*opuD*	glycine betaine transporter
BSU24130	*prpD*	2-methylcitrate dehydratase
BSU02820	*rapJ*	response regulator aspartate phosphatase
BSU23470	*spoIIAA*	anti-anti-sigma factor (antagonist of SpoIIAB)
BSU23460	*spoIIAB*	anti-sigma F factor
BSU18220	*yngF*	enoyl-CoA hydratase
BSU18230	*yngG*	hydroxymethylglutaryl-CoA lyase
BSU18240	*yngH*	acetyl-CoA carboxylase biotin carboxylase subunit
BSU13960	*ykwC*	hypothetical protein
BSU14810	*ylaK*	hypothetical protein
BSU18210	*yngE*	hypothetical protein
BSU18260	*yngJ*	hypothetical protein
BSU19700	*yodP*	hypothetical protein
BSU19720	*yodR*	hypothetical protein
BSU19740	*yodT*	hypothetical protein
BSU24120	*yqiQ*	hypothetical protein
BSU38240	*ywcA*	hypothetical protein
BSU38230	*ywcB*	hypothetical protein

Table 7.10: B. subtilis cluster 84 elaboration genes

Locus	Name	Function
BSU28380	*gerM*	germination (cortex hydrolysis) and sporulation (stage II, multiple polar septa)

BSU27440	*glnH*	glutamine ABC transporter (glutamine-binding protein)
BSU27430	*glnQ*	glutamine ABC transporter (ATP-binding protein)
BSU23170	*spmB*	spore maturation protein
BSU24420	*spoIIIAB*	stage III sporulation protein SpoAB
BSU15170	*spoVD*	penicillin-binding protein
BSU18250	*yngI*	acyl-CoA synthetase
BSU38990	*scoA*	succinyl CoA:3-oxoacid CoA-transferase (subunit A)
BSU05710	*ydhD*	hypothetical protein
BSU10400	*yhxC*	hypothetical protein
BSU12320	*yjmC*	hypothetical protein
BSU12330	*yjmD*	hypothetical protein
BSU17320	*ymaF*	hypothetical protein
BSU19110	*yobW*	hypothetical protein
BSU28960	*ytxC*	hypothetical protein
BSU31730	*yuzC*	hypothetical protein

7.1.1.1.3.2 B. anthracis cluster 84

Figure 7.6: *B. anthracis* cluster 84 image (post-elaboration)

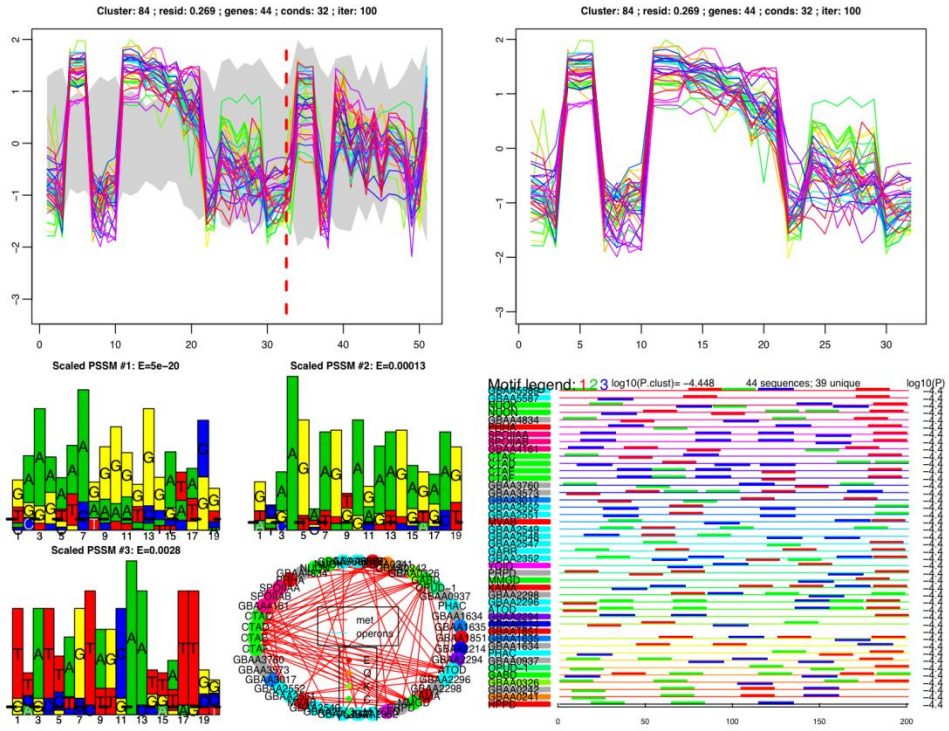


Table 7.11: *B. anthracis* cluster 84 core genes

Locus	Name	Function
GBAA4154	*ctaC*	cytochrome c oxidase, subunit ii
GBAA4153	*ctaD*	cytochrome c oxidase, subunit i
GBAA4152	*ctaE*	cytochrome c oxidase, subunit iii
GBAA4151	*ctaF*	cytochrome c oxidase, subunit ivb
GBAA2353	*garR*	2-hydroxy-3-oxopropionate reductase
GBAA2300	*kamA*	l-lysine 2,3-aminomutase
GBAA2348	*mmgD*	citrate synthase

GBAA2550	*mvaB*	hydroxymethylglutaryl-CoA lyase
GBAA0554	*opuD-1*	glycine betaine transporter
GBAA2349	*prpD*	2-methylcitrate dehydratase
GBAA4296	*spoIIAA*	anti-sigma f factor antagonist
GBAA4295	*spoIIAB*	anti-sigma F factor
GBAA2350	*yqiQ*	carboxyvinyl-carboxyphosphonate phosphorylmutase
GBAA5589	-	acetyl-CoA acetyltransferase
GBAA2548	-	acetyl-CoA carboxylase
GBAA2298	-	acetyltransferase, gnat family
GBAA2547	-	acyl-coa dehydrogenase
GBAA5587	-	acyl-coa dehydrogenase
GBAA2552	-	carboxyl transferase domain protein
GBAA2296	-	coa-transferase, beta subunit
GBAA2551	-	enoyl-CoA hydratase
GBAA4161	-	phoh family protein
GBAA3760	-	prophage lambdaba01, tpr domain protein, putative
GBAA1635	-	sodium/solute symporter family protein
GBAA1634	-	hypothetical protein
GBAA2294	-	hypothetical protein

Table 7.12: *B. anthracis* cluster 84 elaboration genes

Locus	Name	Function
GBAA2295	*atoD*	acetate coa-transferase, subunit a
GBAA0327	*gabD*	succinate-semialdehyde dehydrogenase (nadp+)
GBAA0240	*hppD*	4-hydroxyphenylpyruvate dioxygenase

GBAA1851	*ilvB-2*	acetolactate synthase III large subunit
GBAA5535	*nuoK*	NADH dehydrogenase kappa subunit
GBAA5532	*nuoN*	NADH dehydrogenase subunit N
GBAA1331	*phaC*	poly(r)-hydroxyalkanoic acid synthase, class iii, phac subunit
GBAA2549	-	acetyl-CoA carboxylase
GBAA2352	-	acyl-coa dehydrogenase
GBAA0241	-	fumarylacetoacetate hydrolase family protein
GBAA0242	-	homogentisate 1,2-dioxygenase, putative
GBAA4586	-	phenylalanine-4-hydroxylase, putative
GBAA0326	-	sensory box sigma-54 dependent dna-binding response regulator
GBAA0937	-	hypothetical protein
GBAA2214	-	hypothetical protein
GBAA3017	-	hypothetical protein
GBAA3573	-	hypothetical protein
GBAA4834	-	hypothetical protein

7.1.1.2 Gene lists for flagellar clusters

1. *B. subtilis - B. anthracis* cluster 58
2. *B. subtilis - L. monocytogenes* cluster 79
3. *B. anthracis - L. monocytogenes* cluster 102

7.1.1.2.1 B. subtilis - B. anthracis cluster 58:

7.1.1.2.1.1 *B. subtilis* cluster 58

Figure 7.7: *B. subtilis* cluster 58 image (post-elaboration)

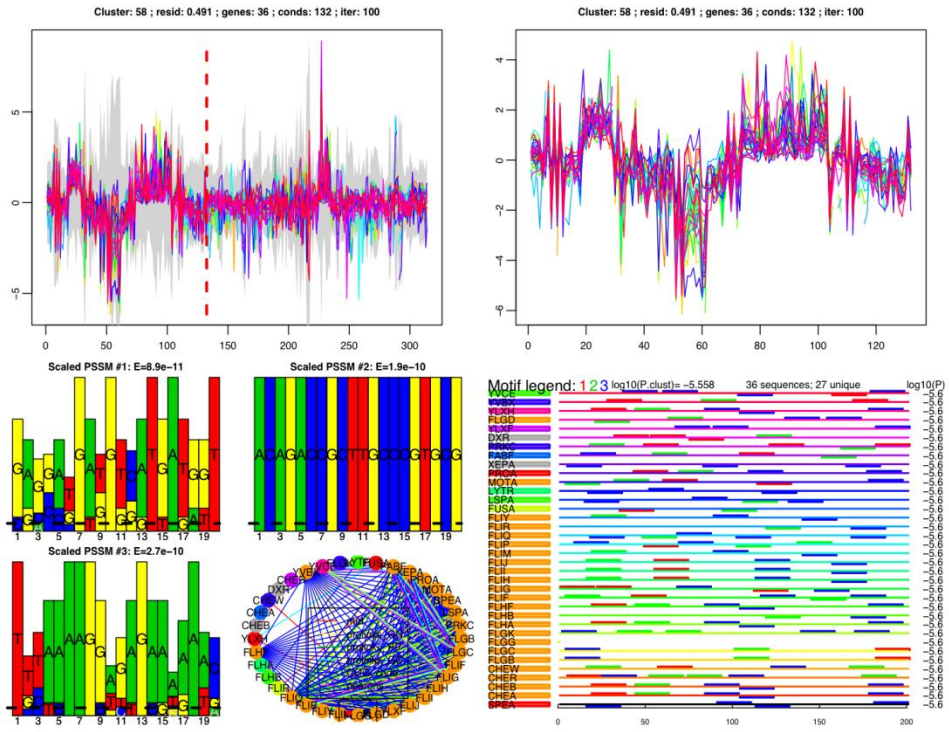


Table 7.13: *B. subtilis* cluster 58 shared

Locus	Name	Function
BSU13130	*proA*	gamma-glutamyl phosphate reductase
BSU13690	*motA*	flagellar motor protein MotA
BSU16180	*flgB*	flagellar basal body rod protein FlgB
BSU16190	*flgC*	flagellar basal body rod protein FlgC
BSU16220	*fliG*	flagellar motor switch protein G
BSU16240	*fliI*	flagellum-specific ATP synthase
BSU16290	*flgG*	flagellar basal body rod protein FlgG
BSU16320	*fliY*	flagellar motor switch protein
BSU16350	*fliP*	flagellar biosynthesis protein FliP

BSU16360	*fliQ*	flagellar biosynthesis protein FliQ
BSU16370	*fliR*	flagellar biosynthesis protein FliR
BSU16380	*flhB*	flagellar biosynthesis protein FlhB
BSU16390	*flhA*	flagellar biosynthesis protein A
BSU16400	*flhF*	flagellar biosynthesis regulator FlhF
BSU22720	*cheR*	methyl-accepting chemotaxis proteins (MCPs) methyltransferase
BSU34800	*yvcE*	hypothetical protein
BSU35410	*flgK*	flagellar hook-associated protein FlgK
BSU35650	*lytR*	membrane-bound transcriptional regulator LytR

Table 7.14: *B. subtilis* cluster 58 elaboration genes

Locus	Name	Function
BSU01120	*fusA*	elongation factor G
BSU11340	*fabF*	3-oxoacyl-(acyl carrier protein) synthase II
BSU12780	*xepA*	lytic exoenzyme associated with defective prophage PBSX
BSU14630	*speA*	arginine decarboxylase
BSU15450	*lspA*	lipoprotein signal peptidase
BSU15770	*prkC*	protein kinase
BSU16210	*fliF*	flagellar MS-ring protein
BSU16230	*fliH*	flagellar assembly protein H
BSU16250	*fliJ*	flagellar biosynthesis chaperone
BSU16260	*ylxF*	hypothetical protein
BSU16280	*flgD*	flagellar basal body rod modification protein
BSU16310	*fliM*	flagellar motor switch protein FliM
BSU16410	*ylxH*	hypothetical protein

BSU16420	*cheB*	chemotaxis-specific methylesterase
BSU16430	*cheA*	two-component sensor histidine kinase
BSU16440	*cheW*	modulation of CheA activity in response to attractants (chemotaxis)
BSU16550	*dxr*	1-deoxy-D-xylulose 5-phosphate reductoisomerase
BSU34020	*yvbX*	hypothetical protein

7.1.1.2.1.2 B. anthracis cluster 58

Figure 7.8: *B. anthracis* cluster 58 image (post-elaboration)

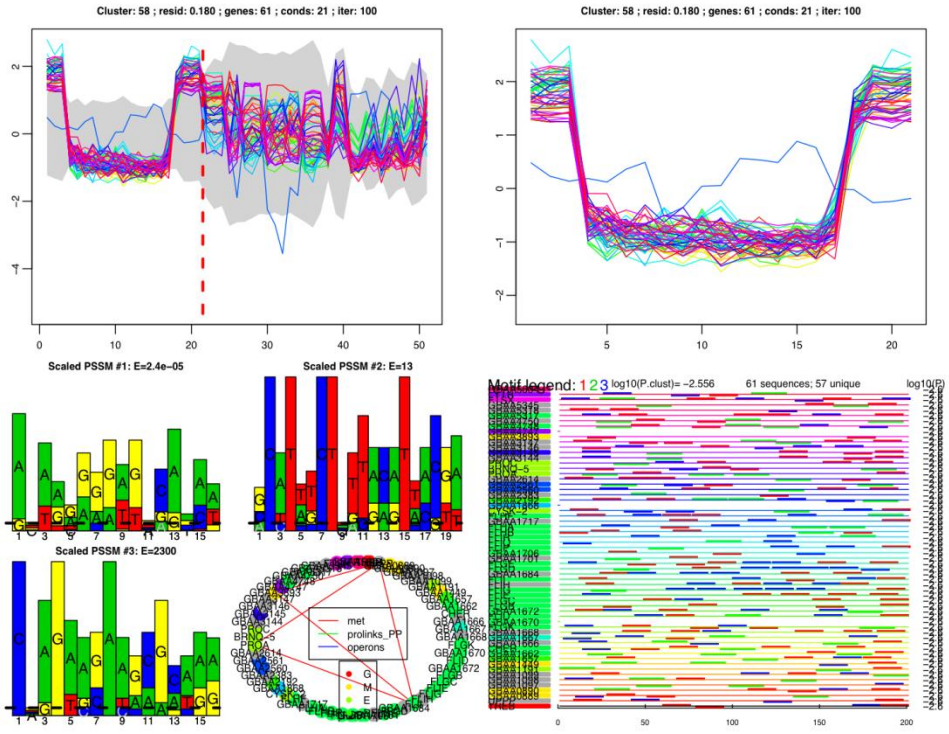


Table 7.15: *B. anthracis* cluster 58 core genes

Locus	Name	Function
GBAA1449	-	peptidase, m23/m37 family
GBAA1662	-	flagellar motor switch protein

GBAA1665	*cheR*	chemotaxis protein methyltransferase cher
GBAA1669	-	flagellar hook-associated protein
GBAA1674	*flgB*	flagellar basal body rod protein
GBAA1675	*flgC*	flagellar basal body rod protein
GBAA1679	*fliG*	flagellar motor protein
GBAA1681	-	flagellum-specific ATP synthase
GBAA1686	-	flagellar hook protein
GBAA1712	-	flagellar biosynthesis protein
GBAA1713	-	flagellar biosynthesis protein
GBAA1714	*fliR*	flagellar biosynthesis protein
GBAA1715	-	flagellar biosynthesis protein
GBAA1716	*flhA*	flagellar biosynthesis protein
GBAA1718	-	flagellar biosynthesis protein
GBAA2992	*proA*	gamma-glutamyl phosphate reductase
GBAA4748	-	flagellar motor protein
GBAA5506	*lytR*	membrane-bound transcriptional regulator LytR

Table 7.16: *B. anthracis* cluster 58 elaboration genes

Locus	Name	Function
GBAA0631	*treB*	pts system, trehalose-specific iibc component
GBAA0683	*uppP*	undecaprenyl pyrophosphate phosphatase
GBAA0889	-	alginate o-acetyltransferase, putative
GBAA0890	-	alginate o-acetyltransferase, putative
GBAA1097	-	hypothetical protein
GBAA1098	-	wall-associated domain protein

GBAA1099	-	hypothetical protein
GBAA1191	-	oligopeptide abc transporter, oligopeptide-binding protein
GBAA1657	-	hypothetical protein
GBAA1666	-	hypothetical protein
GBAA1667	-	hypothetical protein
GBAA1668	-	hypothetical protein
GBAA1670	-	flagellar hook-associated protein
GBAA1671	-	flagellar hook-associated protein
GBAA1672	-	flagellar protein flis, putative
GBAA1676	-	flagellar basal body protein
GBAA1680	-	hypothetical protein
GBAA1684	-	hypothetical protein
GBAA1685	-	flagellar hook assembly protein
GBAA1701	-	hypothetical protein
GBAA1706	-	flagellin
GBAA1717	-	hypothetical protein
GBAA1831	*cysK-2*	cysteine synthase a
GBAA1868	-	hydrolase, alpha/beta fold family
GBAA2192	-	hypothetical protein
GBAA2383	-	hypothetical protein
GBAA2560	-	sensor histidine kinase
GBAA2561	-	dna-binding response regulator
GBAA2614	-	hypothetical protein
GBAA3142	*brnQ-5*	branched-chain amino acid transport system ii carrier protein
GBAA3143	*proC*	pyrroline-5-carboxylate reductase

GBAA3144	-	hypothetical protein
GBAA3145	-	malate dehydrogenase, putative
GBAA3146	-	hypothetical protein
GBAA3147	-	hypothetical protein
GBAA3893	-	cell wall hydrolase, putative
GBAA4747	-	dna-binding protein
GBAA4750	-	d-alanyl-d-alanine carboxypeptidase family protein
GBAA5317	-	methyl-accepting chemotaxis protein
GBAA5318	-	endonuclease/exonuclease/phosphatase family
GBAA5345	-	hypothetical protein
GBAA5415	*ftsX*	cell division abc transporter, permease protein ftsx
GBAA5604	-	abc transporter, atp-binding protein

7.1.1.2.2 B. subtilis - L. monocytogenes cluster 79

7.1.1.2.2.1 *B. subtilis* cluster 79

Figure 7.9: *B. subtilis* cluster 79 image (post-elaboration)

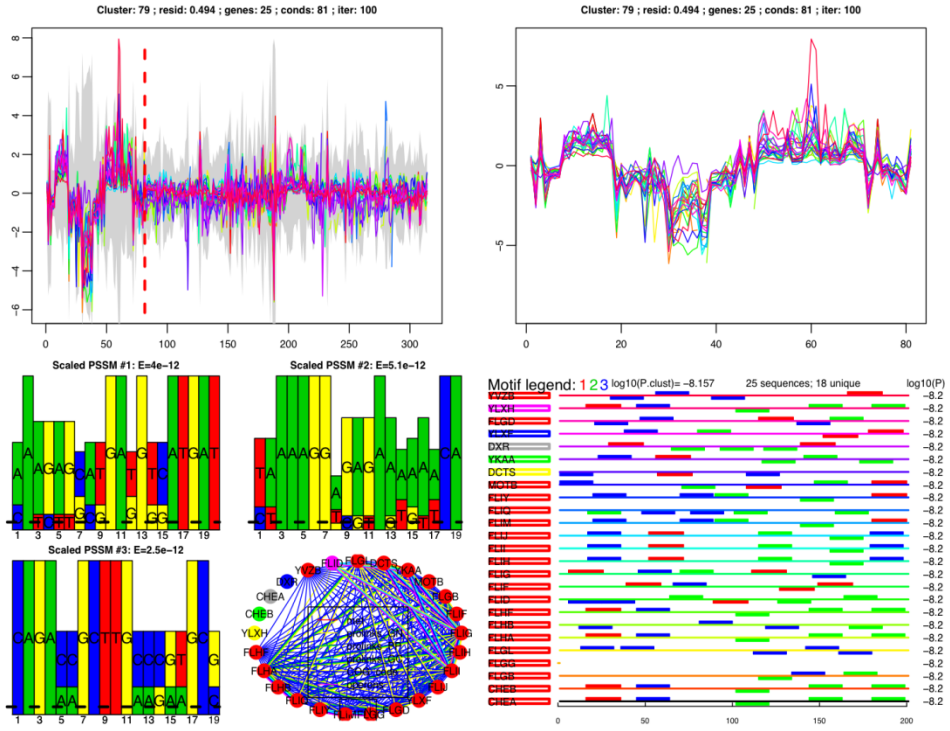


Table 7.17: *B. subtilis* cluster 79 shared genes

Locus	Name	Function
BSU12850	*ykaA*	hypothetical protein
BSU13680	*motB*	flagellar motor protein MotB
BSU16180	*flgB*	flagellar basal body rod protein FlgB
BSU16210	*fliF*	flagellar MS-ring protein
BSU16220	*fliG*	flagellar motor switch protein G

BSU16240	*fliI*	flagellum-specific ATP synthase
BSU16280	*flgD*	flagellar basal body rod modification protein
BSU16290	*flgG*	flagellar basal body rod protein FlgG
BSU16310	*fliM*	flagellar motor switch protein FliM
BSU16320	*fliY*	flagellar motor switch protein
BSU16380	*flhB*	flagellar biosynthesis protein FlhB
BSU16390	*flhA*	flagellar biosynthesis protein A
BSU16400	*flhF*	flagellar biosynthesis regulator FlhF
BSU16430	*cheA*	two-component sensor histidine kinase
BSU16550	*dxr*	1-deoxy-D-xylulose 5-phosphate reductoisomerase
BSU35150	*yvzB*	hypothetical protein
BSU35340	*fliD*	flagellar capping protein
BSU35400	*flgL*	flagellar hook-associated protein FlgL

Table 7.18: *B. subtilis* cluster 79 elaboration genes

Locus	Name	Function
BSU04450	*dctS*	two-component sensor histidine kinase
BSU16230	*fliH*	flagellar assembly protein H
BSU16250	*fliJ*	flagellar biosynthesis chaperone
BSU16260	*ylxF*	hypothetical protein
BSU16360	*fliQ*	flagellar biosynthesis protein FliQ
BSU16410	*ylxH*	hypothetical protein
BSU16420	*cheB*	chemotaxis-specific methylesterase

7.1.1.2.2.2 L. monocytogenes cluster 79

Figure 7.10: *L. monocytogenes* cluster 79 image (post-elaboration)

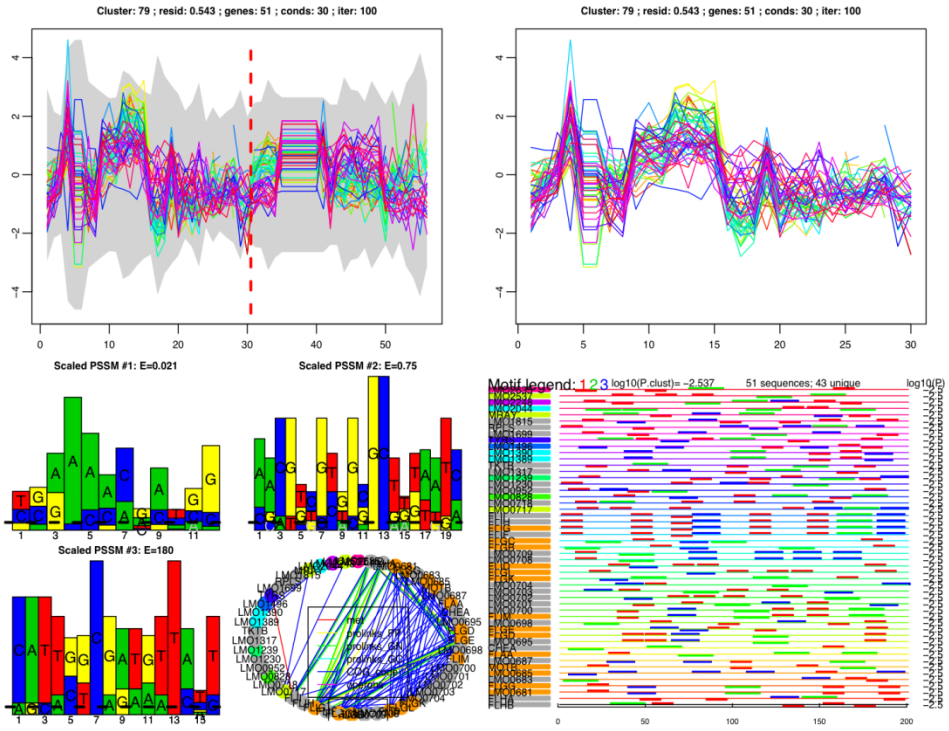


Table 7.19: L. monocytogenes cluster 79 shared

Locus	Name	Function
LMO0679	*flhB*	flagellar biosynthesis protein FlhB
LMO0680	*flhA*	flagellar biosynthesis protein A
LMO0681	-	flagellar biosynthesis regulator FlhF
LMO0686	*motB*	hypothetical protein
LMO0690	*flaA*	flagellin
LMO0692	*cheA*	two-component sensor histidine kinase CheA
LMO0696	*flgD*	flagellar basal body rod modification protein

LMO0697	*flgE*	flagellar hook protein FlgE
LMO0699	*fliM*	flagellar motor switch protein FliM
LMO0700	-	flagellar motor switch protein
LMO0706	*flgL*	flagellar hook-associated protein FlgL
LMO0707	*fliD*	flagellar capping protein
LMO0710	*flgB*	flagellar basal body rod protein FlgB
LMO0713	*fliF*	flagellar MS-ring protein
LMO0714	*fliG*	flagellar motor switch protein G
LMO0716	*fliI*	flagellum-specific ATP synthase
LMO1317	-	1-deoxy-D-xylulose 5-phosphate reductoisomerase
LMO2248	-	hypothetical protein

Table 7.20: *L. monocytogenes* cluster 79 elaboration genes

Locus	Name	Function
LMO0682	*flgG*	flagellar basal body rod protein FlgG
LMO0683	-	hypothetical protein
LMO0685	-	flagellar motor protein MotA
LMO0687	-	hypothetical protein
LMO0695	-	hypothetical protein
LMO0698	-	flagellar motor switch protein
LMO0701	-	hypothetical protein
LMO0702	-	hypothetical protein
LMO0703	-	hypothetical protein
LMO0704	-	hypothetical protein
LMO0705	*flgK*	flagellar hook-associated protein FlgK

LMO0708	-	hypothetical protein
LMO0709	-	hypothetical protein
LMO0711	*flgC*	flagellar basal body rod protein FlgC
LMO0715	*fliH*	flagellar assembly protein H
LMO0717	-	hypothetical protein
LMO0718	-	hypothetical protein
LMO0828	-	hypothetical protein
LMO0952	-	hypothetical protein
LMO1230	-	hypothetical protein
LMO1239	-	hypothetical protein
LMO1365	*tktB*	1-deoxy-D-xylulose-5-phosphate synthase
LMO1389	-	hypothetical protein
LMO1390	-	hypothetical protein
LMO1496	-	hypothetical protein
LMO1598	*tyrS*	tyrosyl-tRNA synthetase
LMO1699	-	hypothetical protein
LMO1787	*rplS*	50S ribosomal protein L19
LMO1815	-	hypothetical protein
LMO2037	*mraY*	hypothetical protein
LMO2044	-	hypothetical protein
LMO2537	-	hypothetical protein
LMO2635	-	1,4-dihydroxy-2-naphthoate octaprenyltransferase

7.1.1.2.3 B. anthracis - L. monocytogenes cluster 102

7.1.1.2.3.1 *B. anthracis* cluster 102

Figure 7.11: *B. anthracis* cluster 102 image (post-elaboration)

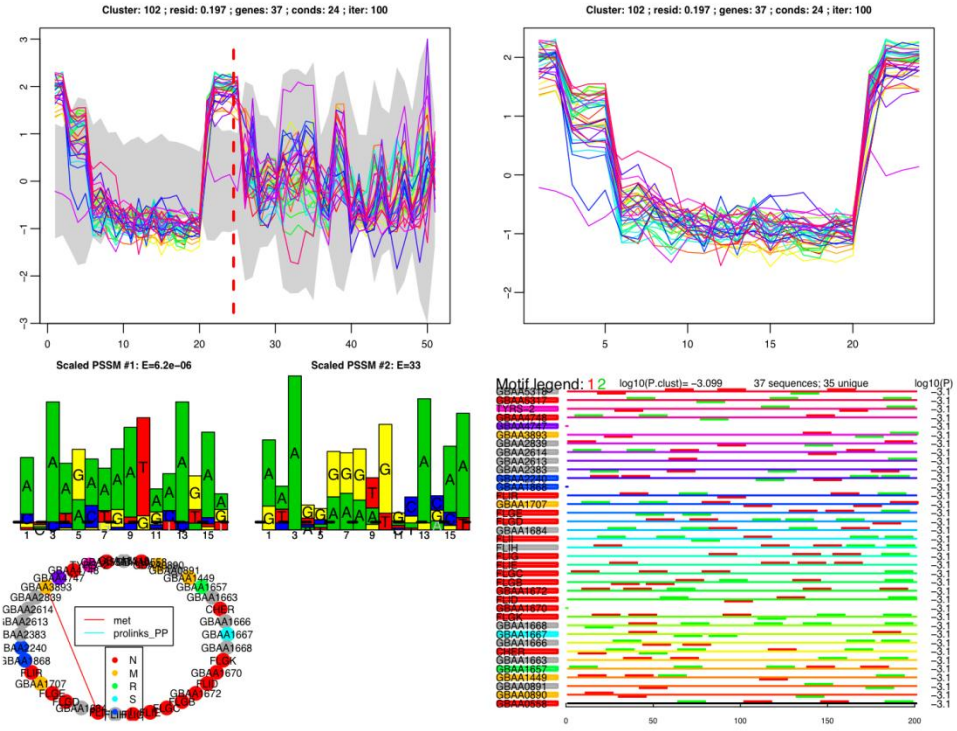


Table 7.21: *B. anthracis* cluster 102 core genes

Locus	Name	Function
GBAA1667	-	hypothetical protein
GBAA1669	-	flagellar hook-associated protein
GBAA1670	-	flagellar hook-associated protein
GBAA1672	-	flagellar protein flis, putative
GBAA1674	*flgB*	flagellar basal body rod protein

GBAA1675	*flgC*	flagellar basal body rod protein
GBAA1676	-	flagellar basal body protein
GBAA1679	*fliG*	flagellar motor protein
GBAA1680	-	hypothetical protein
GBAA1681	-	flagellum-specific ATP synthase
GBAA1685	-	flagellar hook assembly protein
GBAA1686	-	flagellar hook protein
GBAA1707	-	transglycosylase, slt family
GBAA1714	*fliR*	flagellar biosynthesis protein
GBAA5314	*tyrS-2*	tyrosyl-tRNA synthetase

Table 7.22: *B. anthracis* cluster 102 elaboration genes

Locus	Name	Function
GBAA0558	-	methyl-accepting chemotaxis protein
GBAA0890	-	alginate o-acetyltransferase, putative
GBAA0891	-	hypothetical protein
GBAA1449	-	peptidase, m23/m37 family
GBAA1657	-	hypothetical protein
GBAA1663	-	hypothetical protein
GBAA1665	*cheR*	chemotaxis protein methyltransferase cher
GBAA1666	-	hypothetical protein
GBAA1668	-	hypothetical protein
GBAA1671	-	flagellar hook-associated protein
GBAA1684	-	hypothetical protein
GBAA1868	-	hydrolase, alpha/beta fold family

GBAA2240	-	1-acyl-sn-glycerol-3-phosphate acyltransferase, putative
GBAA2383	-	hypothetical protein
GBAA2613	-	hypothetical protein
GBAA2614	-	hypothetical protein
GBAA2839	-	hypothetical protein
GBAA3893	-	cell wall hydrolase, putative
GBAA4747	-	dna-binding protein
GBAA4748	-	flagellar motor protein
GBAA5317	-	methyl-accepting chemotaxis protein
GBAA5318	-	endonuclease/exonuclease/phosphatase family

7.1.1.2.3.2 L. monocytogenes cluster 102

Figure 7.12: *L. monocytogenes* cluster 102 image (post-elaboration)

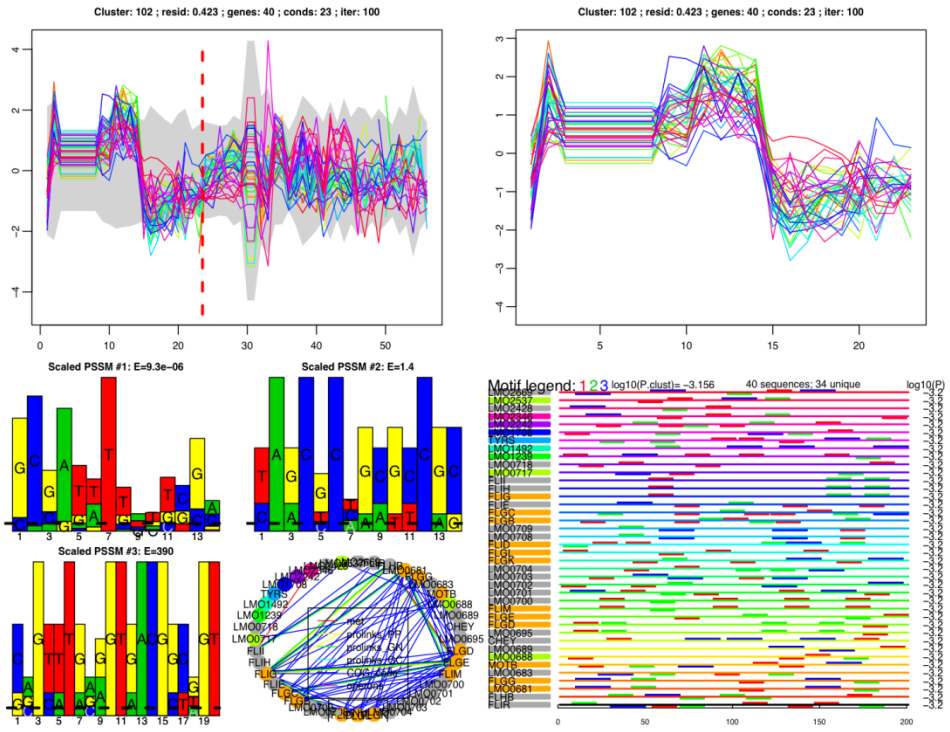


Table 7.23: *L. monocytogenes* cluster 102 core genes

Locus	Name	Function
LMO0678	*fliR*	flagellar biosynthesis protein FliR
LMO0696	*flgD*	flagellar basal body rod modification protein
LMO0697	*flgE*	flagellar hook protein FlgE
LMO0703	-	hypothetical protein
LMO0705	*flgK*	flagellar hook-associated protein FlgK
LMO0707	*fliD*	flagellar capping protein
LMO0708	-	hypothetical protein
LMO0710	*flgB*	flagellar basal body rod protein FlgB
LMO0711	*flgC*	flagellar basal body rod protein FlgC

LMO0712	*fliE*	flagellar hook-basal body protein FliE
LMO0714	*fliG*	flagellar motor switch protein G
LMO0715	*fliH*	flagellar assembly protein H
LMO0716	*fliI*	flagellum-specific ATP synthase
LMO0717	-	hypothetical protein
LMO1598	*tyrS*	tyrosyl-tRNA synthetase

Table 7.24: *L. monocytogenes* cluster 102 elaboration genes

Locus	Name	Function
LMO0679	*flhB*	flagellar biosynthesis protein FlhB
LMO0681	-	flagellar biosynthesis regulator FlhF
LMO0682	*flgG*	flagellar basal body rod protein FlgG
LMO0683	-	hypothetical protein
LMO0686	*motB*	hypothetical protein
LMO0688	-	hypothetical protein
LMO0689	-	hypothetical protein
LMO0691	*cheY*	Chemotaxis response regulator CheY
LMO0695	-	hypothetical protein
LMO0699	*fliM*	flagellar motor switch protein FliM
LMO0700	-	flagellar motor switch protein
LMO0701	-	hypothetical protein
LMO0702	-	hypothetical protein
LMO0704	-	hypothetical protein
LMO0706	*flgL*	flagellar hook-associated protein FlgL
LMO0709	-	hypothetical protein

LMO0718	-	hypothetical protein
LMO1239	-	hypothetical protein
LMO1492	-	hypothetical protein
LMO1708	-	hypothetical protein
LMO2242	-	hypothetical protein
LMO2346	-	hypothetical protein
LMO2428	-	hypothetical protein
LMO2537	-	hypothetical protein
LMO2669	-	hypothetical protein

7.2 Additional figures and tables from global validation

7.2.1 (bi)cluster coherence metric figures

7.2.1.1 Residuals

In each of the plots shown below are the distributions of the residuals from all methods considered by this study for a given pairing. Next to each distribution, in gray, are residuals from randomly shuffled (bi)clusters that match the size distribution for each method. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.1.1.1 Figures for the Gram-positive triplet

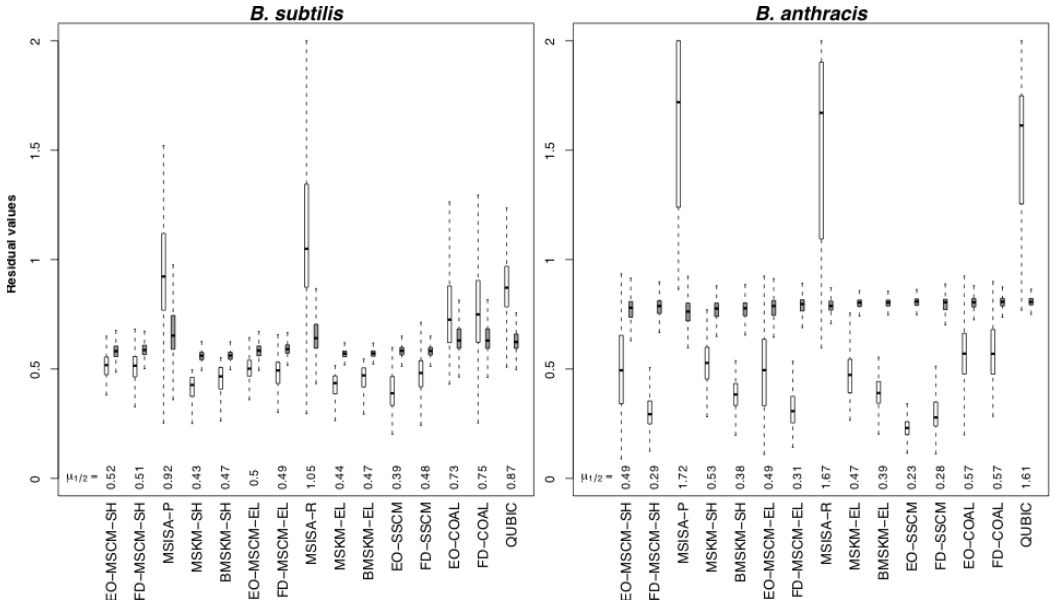


Figure 7.13: Residuals from the *B. subtilis – B. anthracis* pairing.

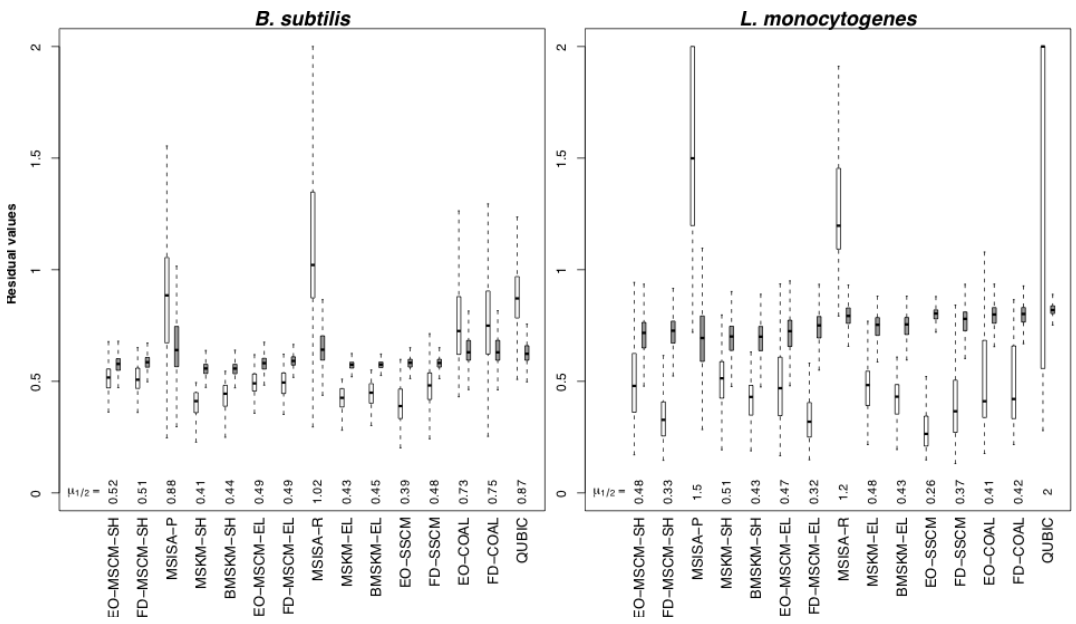


Figure 7.14: Residuals from the *B. subtilis – L. monocytogenes* pairing

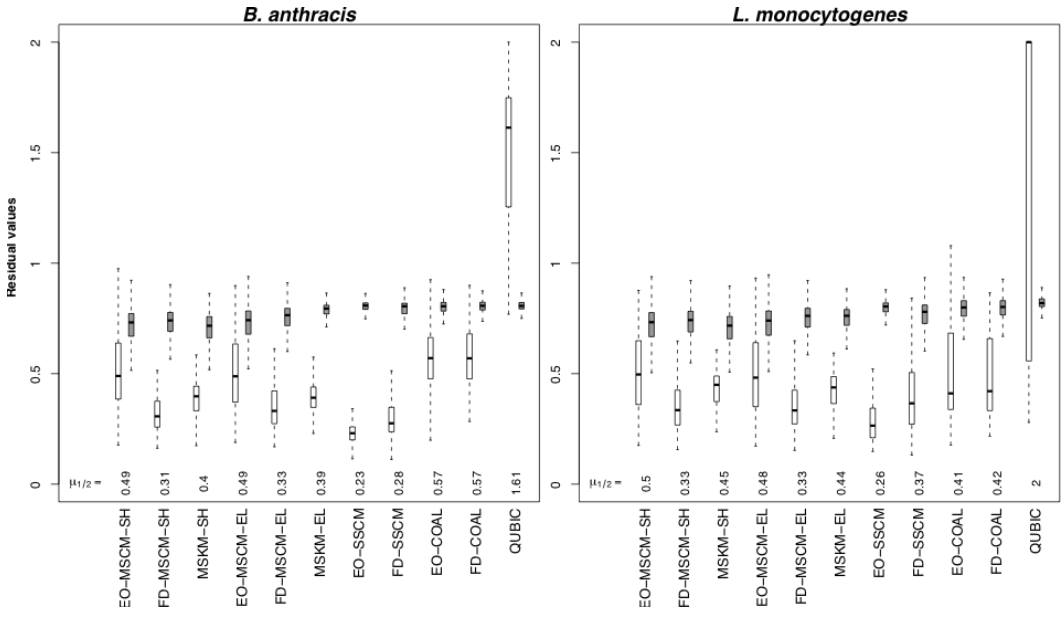


Figure 7.15: Residuals from the *B. anthracis – L. monocytogenes* pairing

7.2.1.1.2 Figures for the Gram-negative triplet

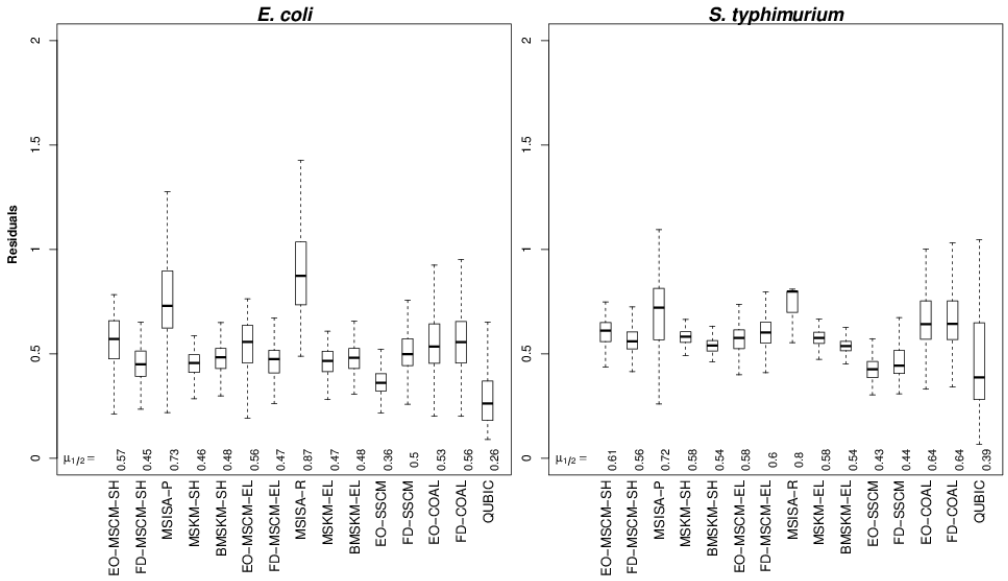


Figure 7.16: Residuals from the E. coli – S. typhimurium pairing.

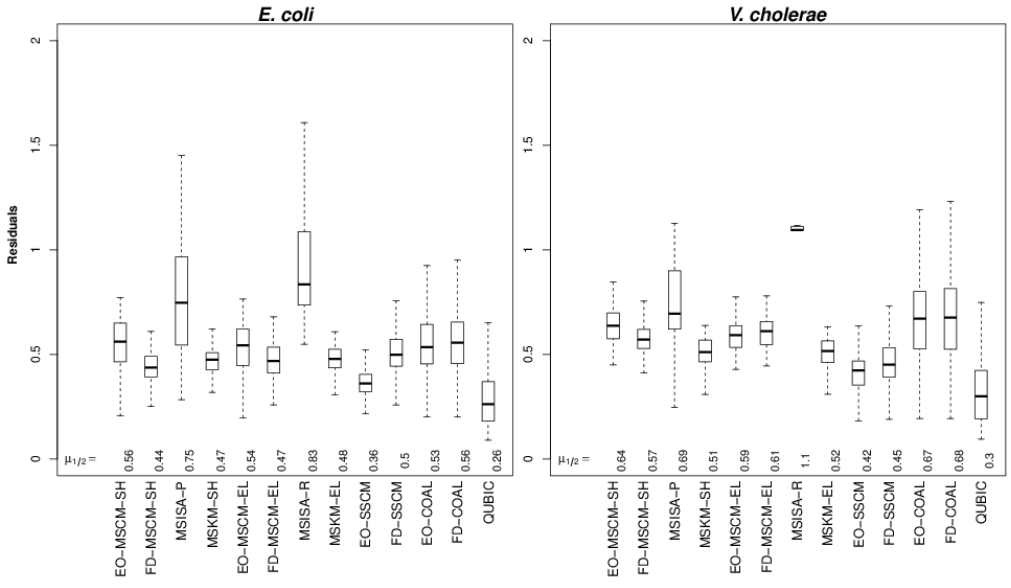


Figure 7.17: Residuals from the E. coli – V. Cholerae pairing.

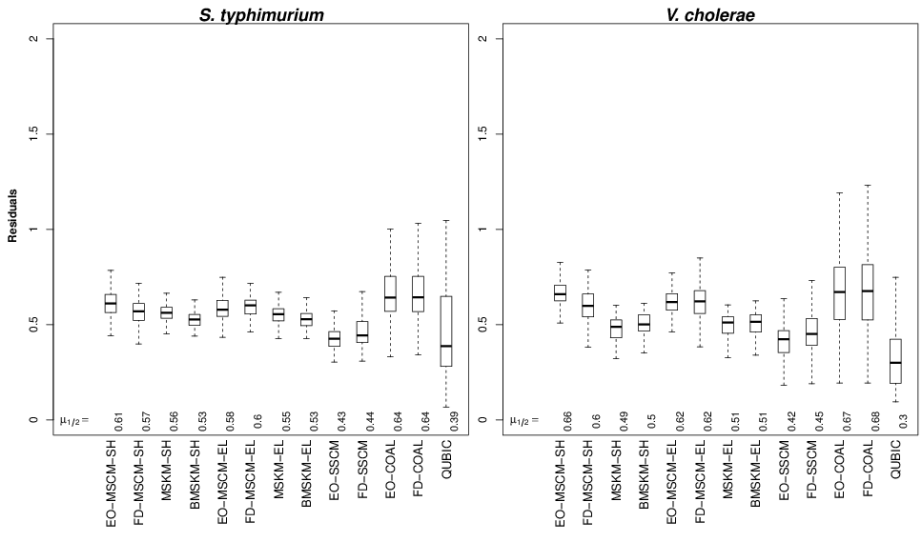


Figure 7.18: Residuals from the S. typhimurium – V. cholerae pairing

7.2.1.2 Average pairwise correlations

In each of the plots shown below are the distributions of the distributions of the mean correlations from all methods considered by this study for a given pairing. Next to each distribution, in gray, are residuals from randomly shuffled (bi)clusters that match the size distribution for each method. Explanations of the method name abbreviations can be found in Table 3.1**.**

7.2.1.2.1 Figures for the Gram-positive triplet

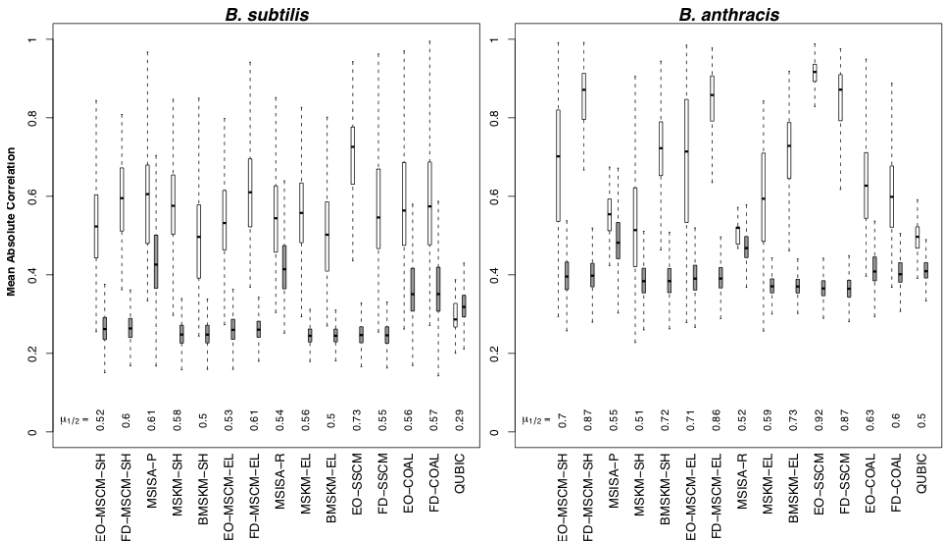


Figure 7.19: Mean correlations from the B. subtilis – B. anthracis pairing.

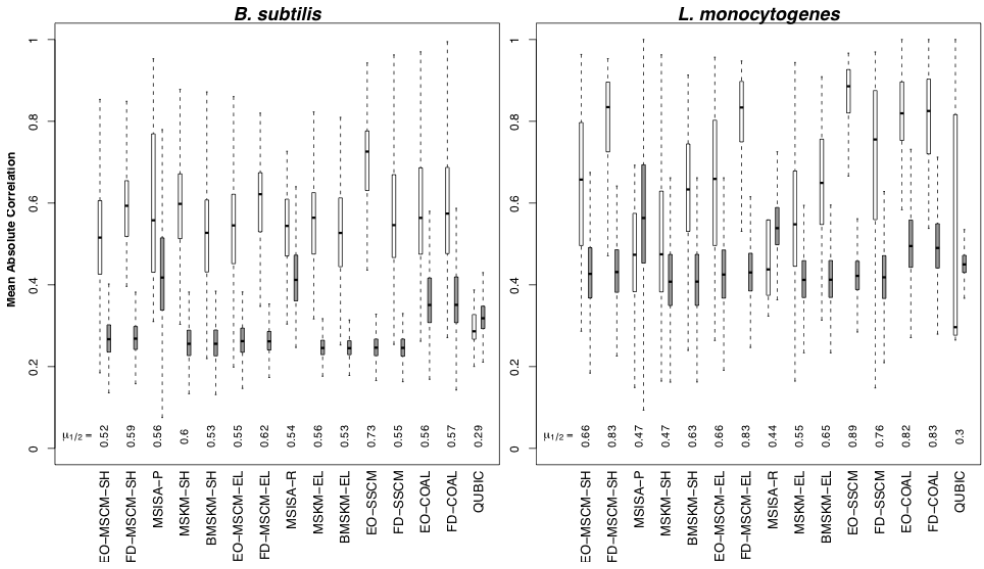


Figure 7.20: Mean correlations from the *B. subtilis – L. monocytogenes* pairing.

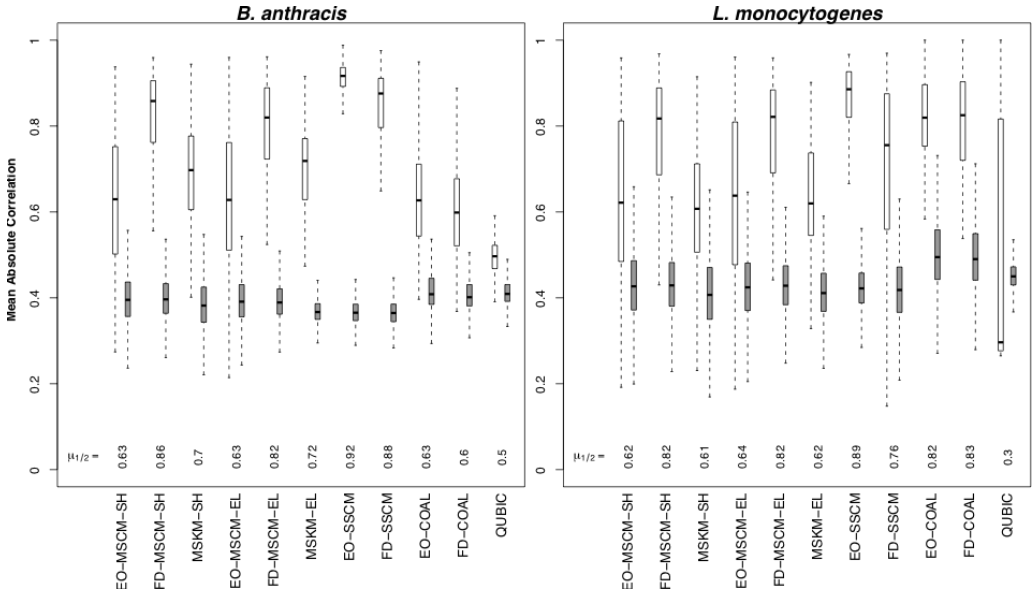


Figure 7.21: Mean correlations from the *B. anthracis – L. monocytogenes* pairing.

7.2.1.2.2 Figures for the Gram-negative triplet

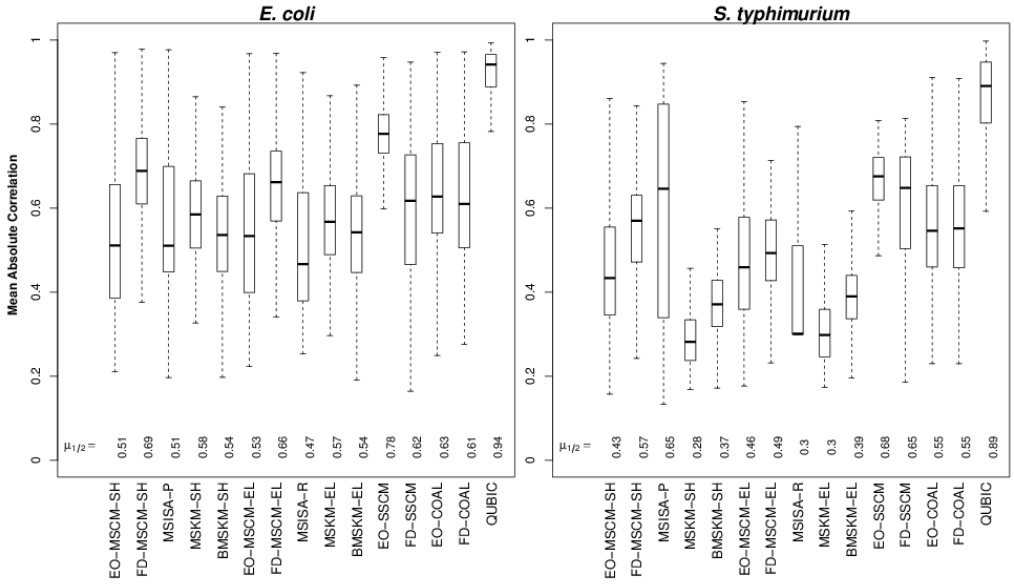


Figure 7.22: Mean correlations from the *E. coli – S. typhimurium* pairing.

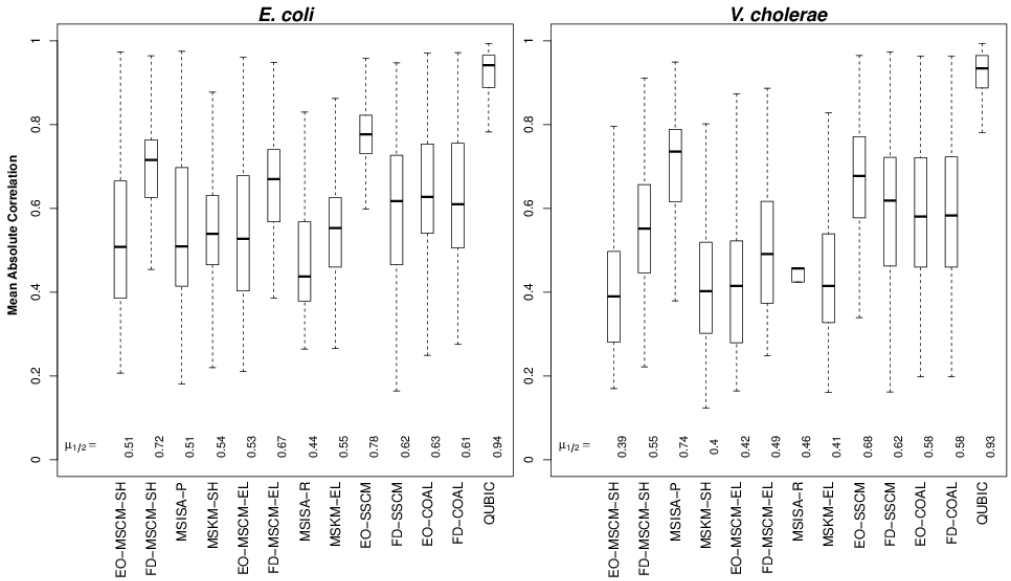


Figure 7.23: Mean correlations from the *E. coli – V. cholerae* pairing.

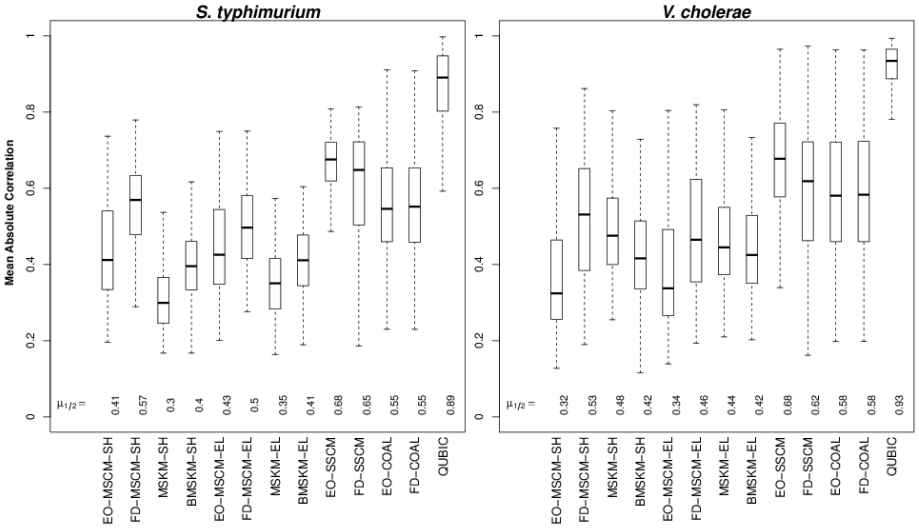


Figure 7.24: Mean correlations from the *S. typhimurium – V. cholerae* pairing.

7.2.1.3 Network Association p-values

In each of the plots shown below are the distributions of the distributions of the network association p-values ($-\log_{10}$) from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.1.3.1 Figures for the Gram-positive triplet

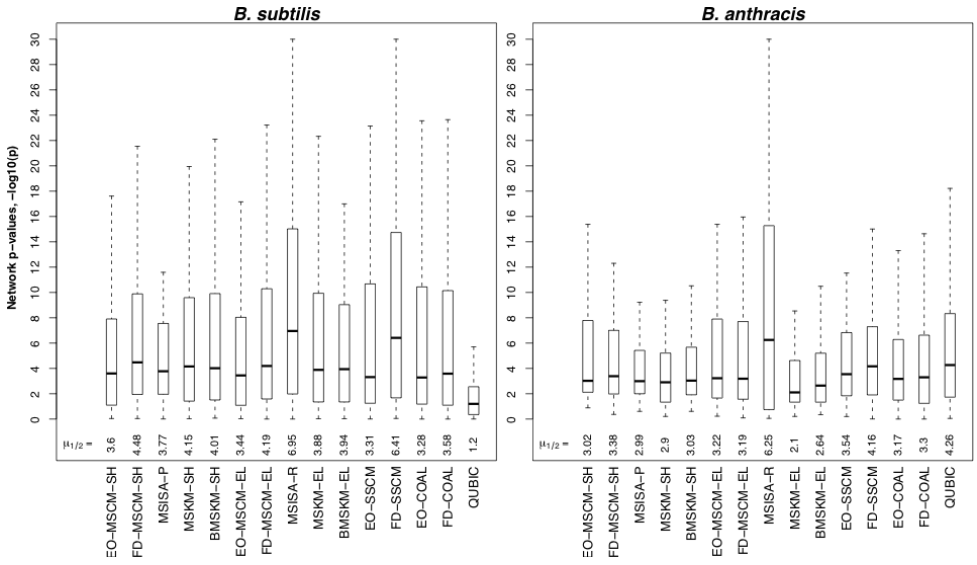


Figure 7.25: Network Association p-values from the *B. subtilis* – *B. anthracis* pairing.

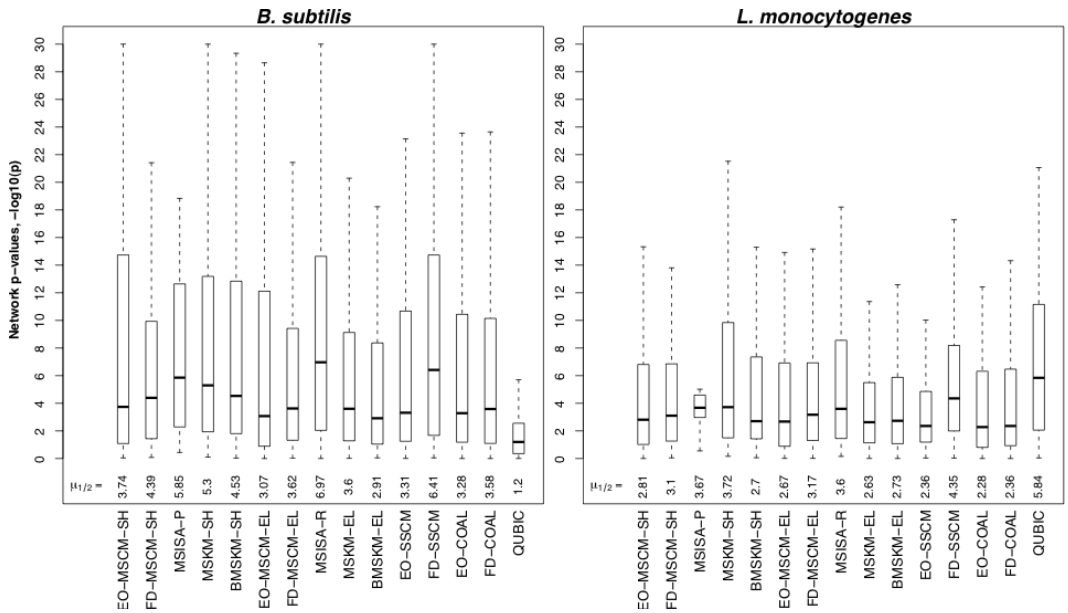


Figure 7.26: Network Association p-values from the *B. subtilis – L. monocytogenes* pairing

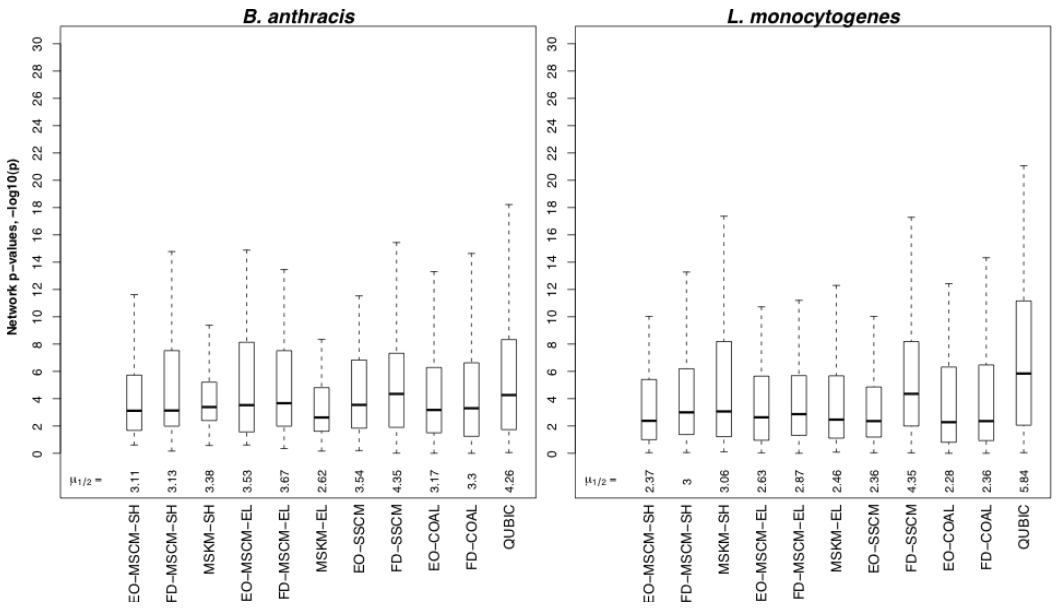


Figure 7.27: Network Association p-values from the *B. anthracis – L. monocytogenes* pairing.

7.2.1.3.2 Figures for the Gram-negative triplet

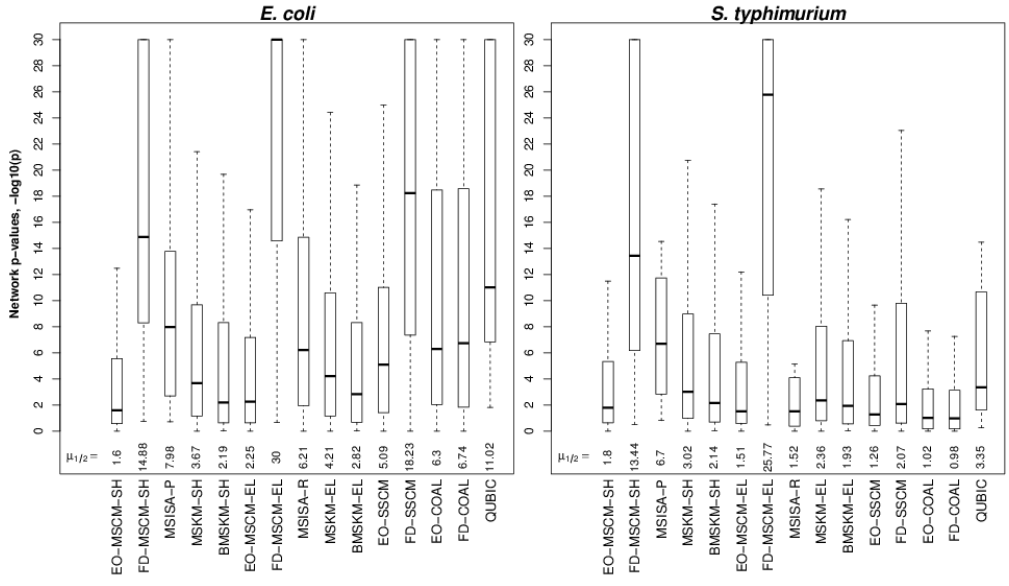


Figure 7.28: Network Association p-values from the *E. coli – S. typhimurium* pairing.

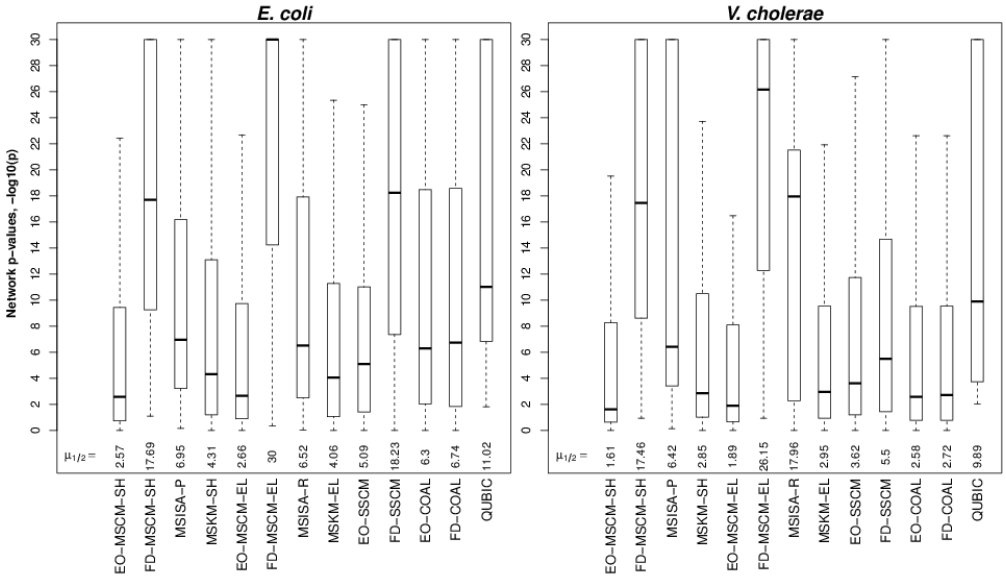


Figure 7.29: Network Association p-values from the *E. coli – V. cholerae* pairing

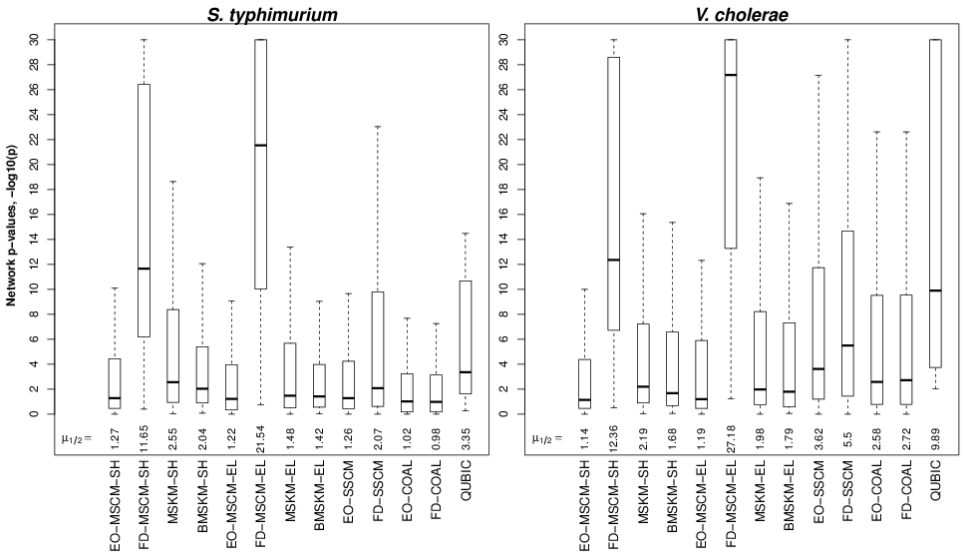


Figure 7.30: Network Association p-values from the *S. typhimurium* – *V. cholerae* pairing.

7.2.1.4 Motif E-values

In each of the plots shown below are the distributions of the motif E-values ($-\log_{10}$) from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.1.4.1 Figures for the Gram-positive triplet

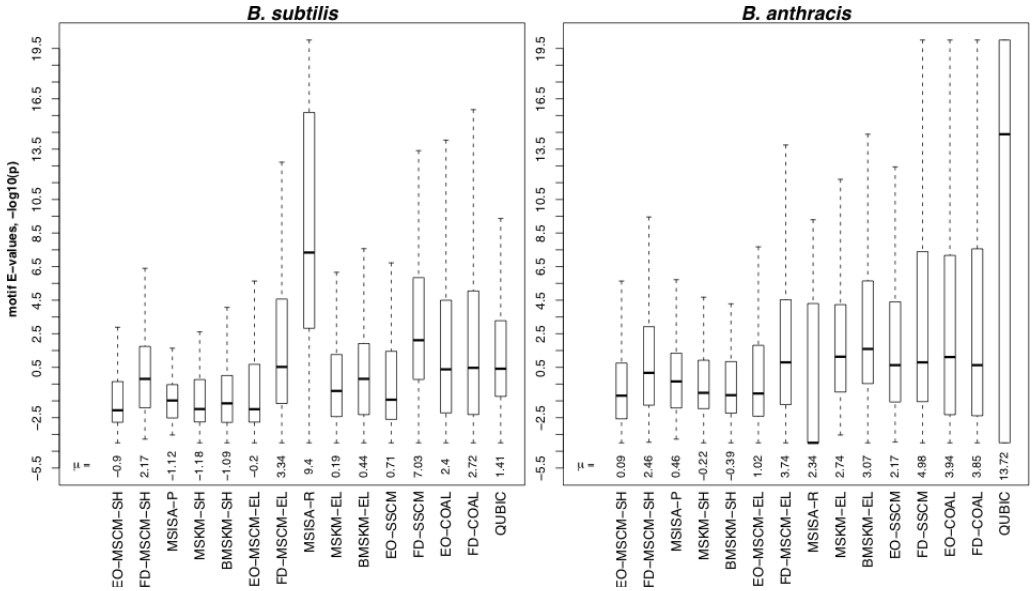


Figure 7.31: Motif E-values from the *B. subtilis*-*B. anthracis* pairing.

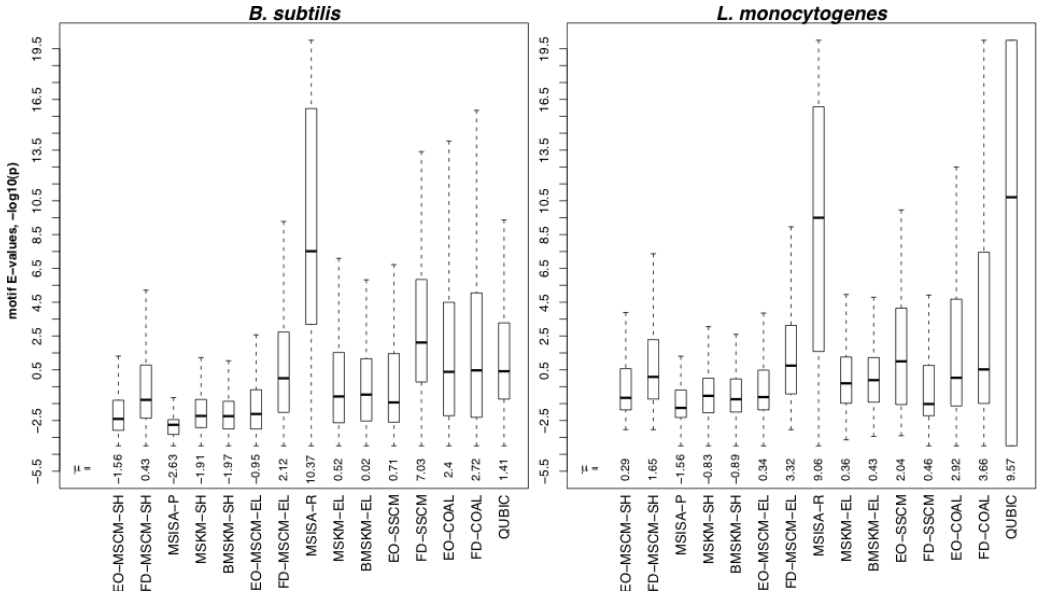


Figure 7.32: Motif E-values from the *B. subtilis*-*L. monocytogenes* pairing

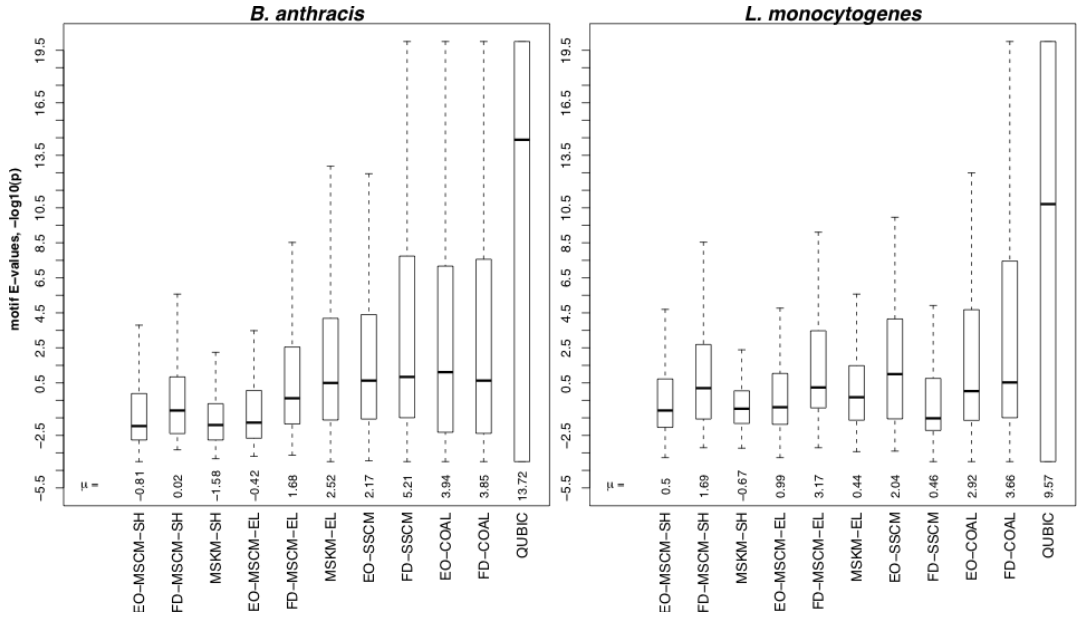


Figure 7.33: Motif E-values from the *B. anthracis*-*L. monocytogenes* pairing.

7.2.1.4.2 Figures for the Gram-negative triplet

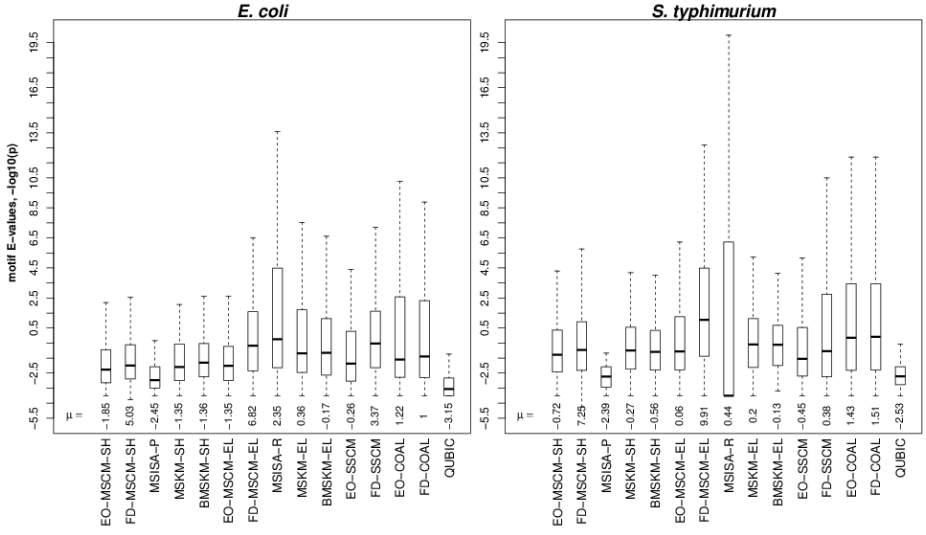


Figure 7.34: Motif E-values from the *E. coli – S. typhimurium* pairing

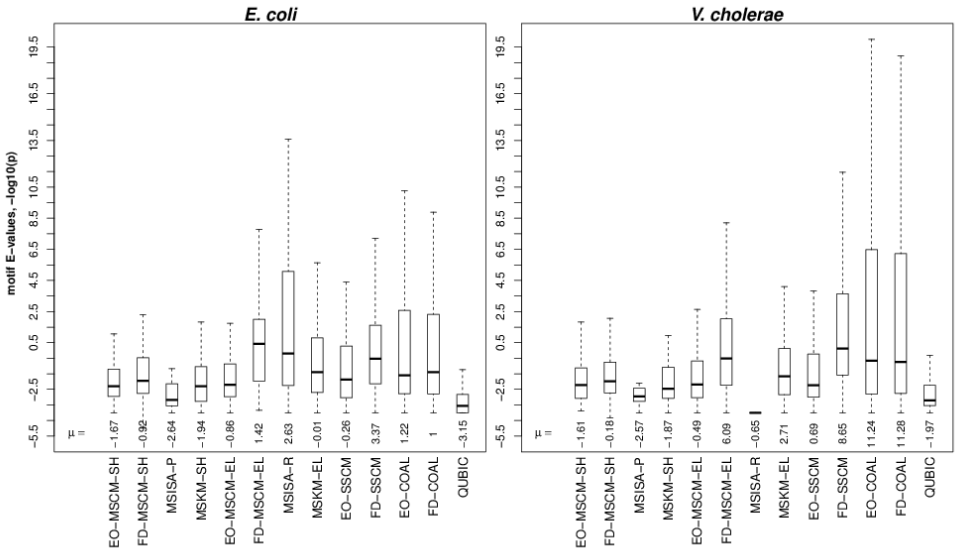


Figure 7.35: Motif E-values from the *E. coli – V. cholerae* pairing.

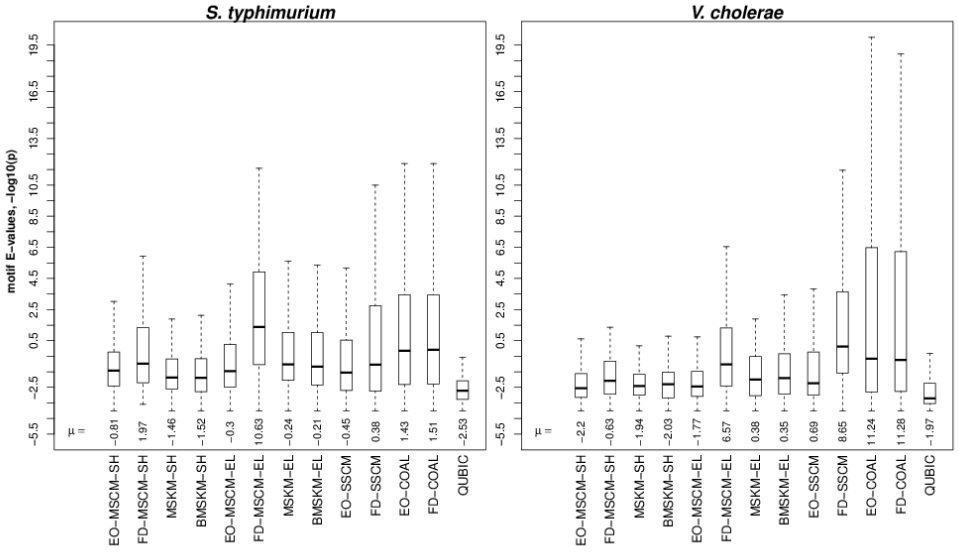


Figure 7.36: Motif E-values from the *S. typhimurium – V. cholerae* pairing.

7.2.1.5 Sequence p-values

In each of the plots shown below are the distributions of the sequence p-values ($-\log_{10}$) from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.1.5.1 Figures for the Gram-positive triplet

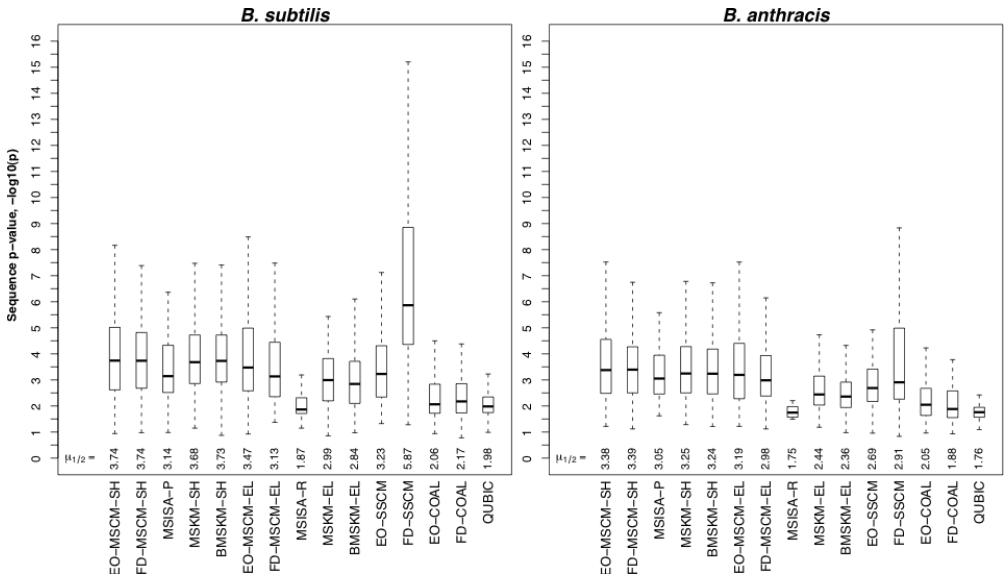


Figure 7.37: Sequence p-values from the *B. subtilis*-*B. anthracis* pairing.

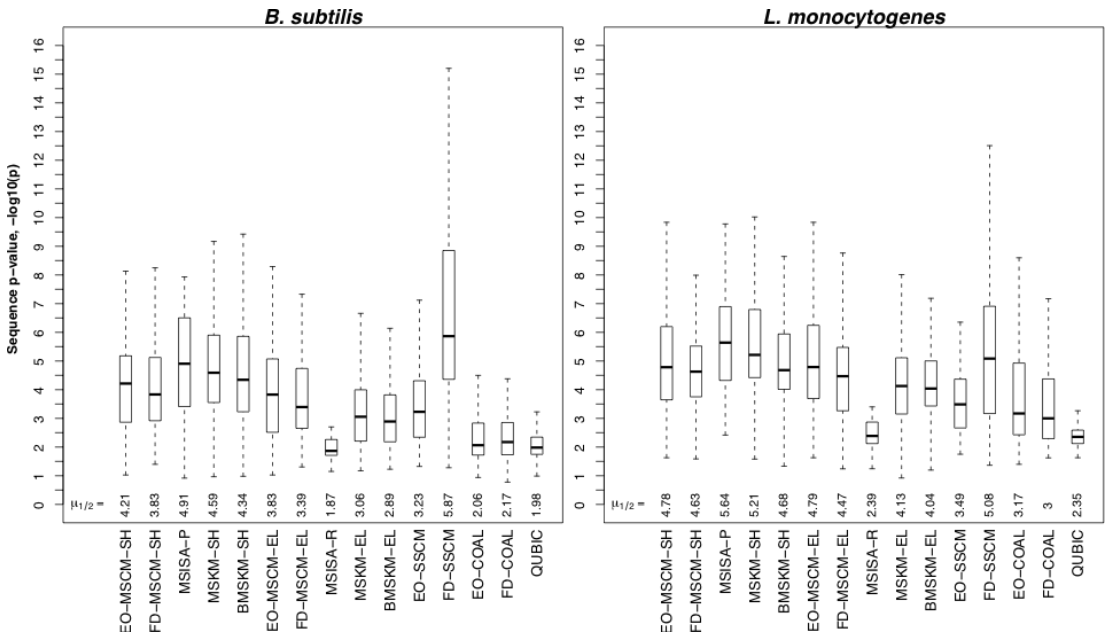


Figure 7.38: Sequence p-values from the *B. subtilis-L. monocytogenes* pairing.

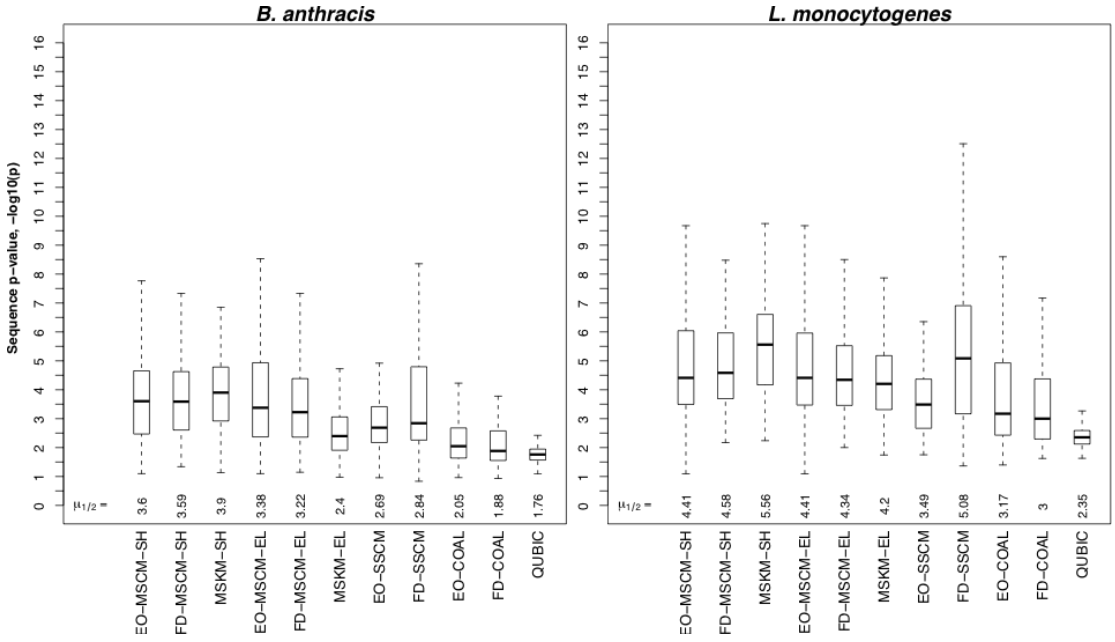


Figure 7.39: Sequence p-values from the *B. anthracis-L. monocytogenes* pairing.

7.2.1.5.2 Figures for the Gram-negative triplet

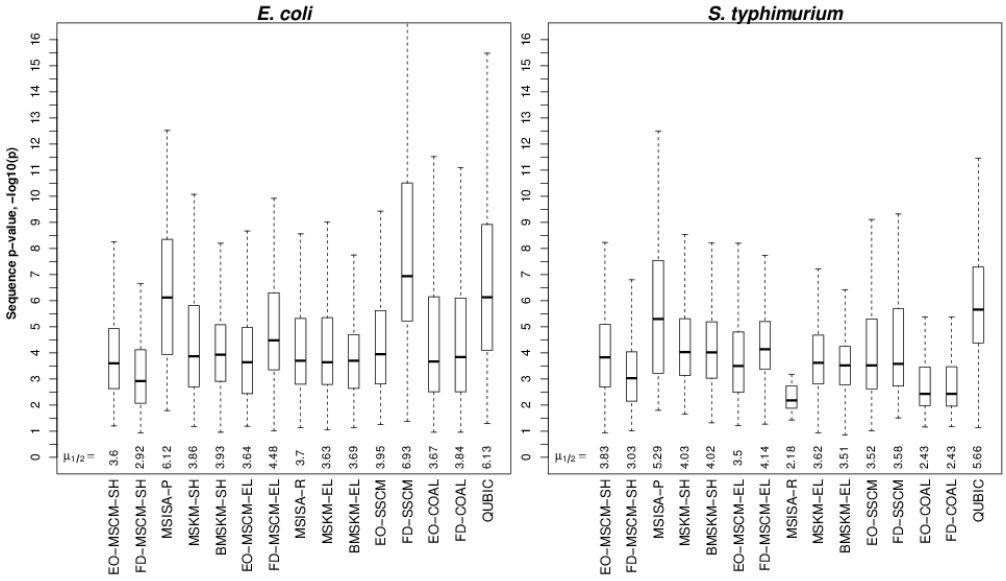


Figure 7.40: Sequence p-values from the *E. coli – S. typhimurium* pairing.

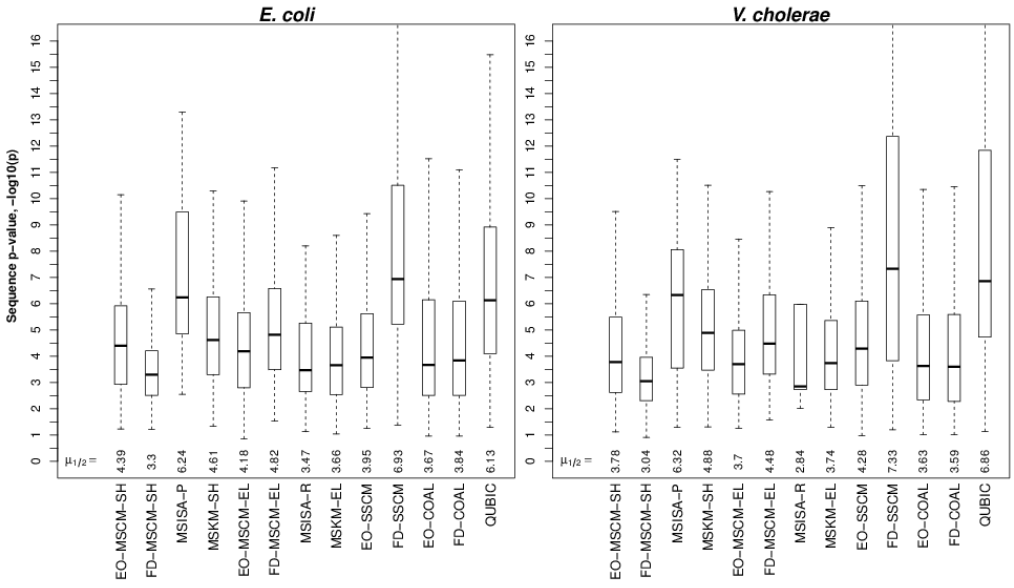


Figure 7.41: Sequence p-values from the *E. coli* – *V. cholerae* pairing.

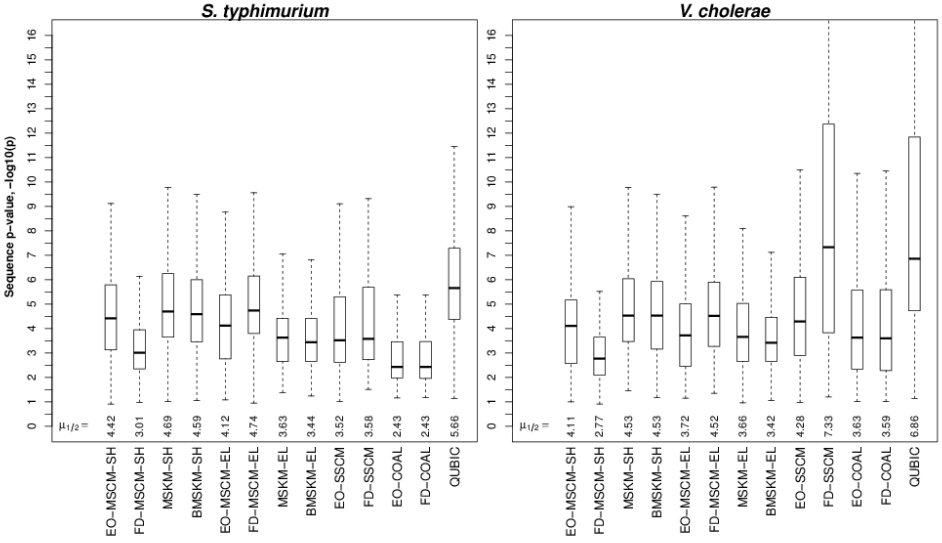


Figure 7.42: Sequence p-values from the *S. typhimurium* – *V. cholerae* pairing.

7.2.2 Additional size distribution, overlap and coverage figures

7.2.2.1 Number of genes

In each of the plots shown below are the distributions of the number of genes from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.2.1.1 Figures for the Gram-positive triplet

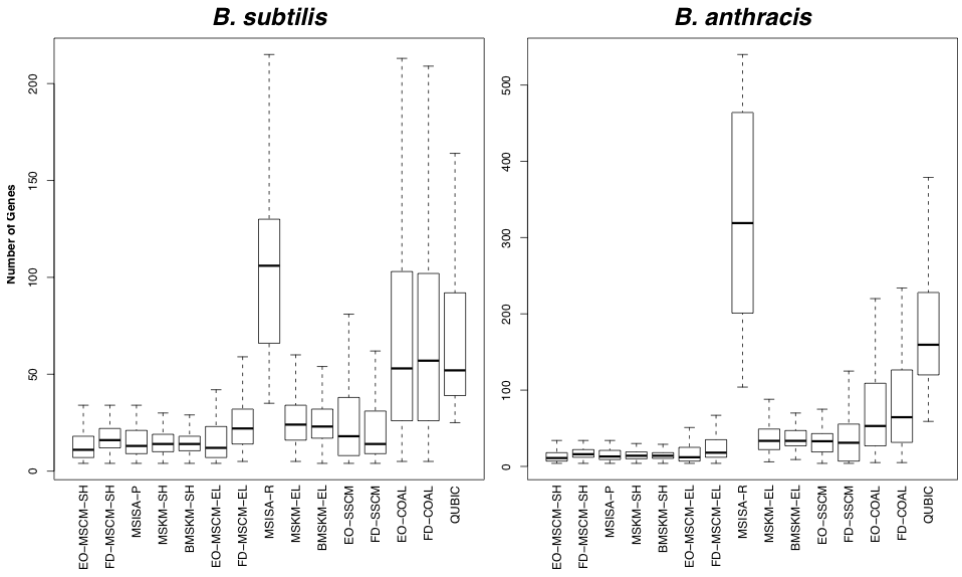


Figure 7.43: Number of genes from the *B. subtilis* – *B. anthracis* pairing.

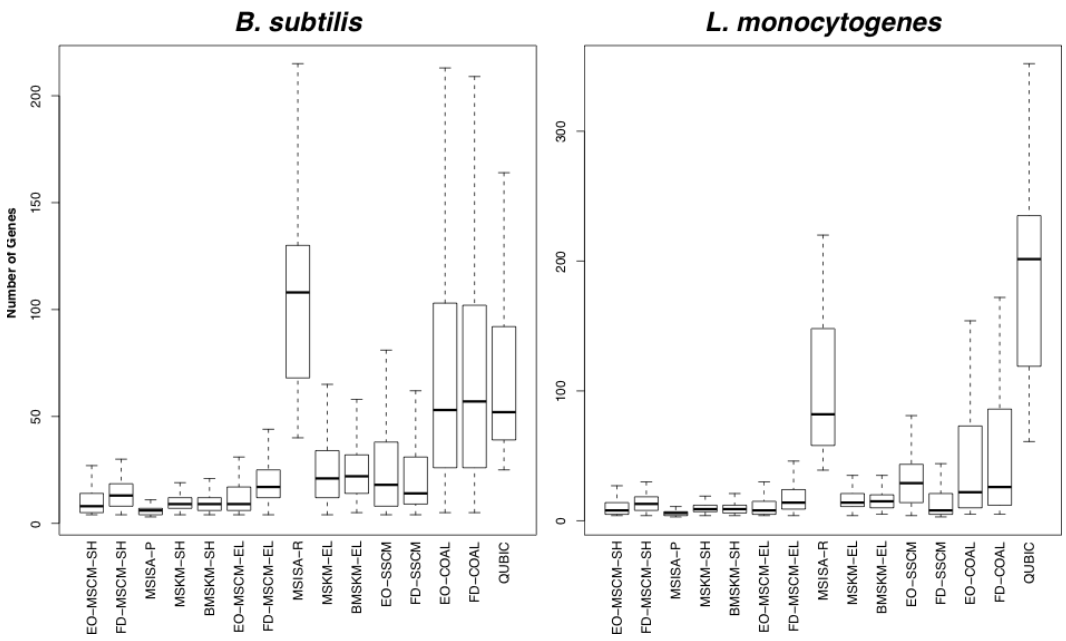


Figure 7.44: Number of genes from the *B. subtilis – L. monocytogenes* pairing

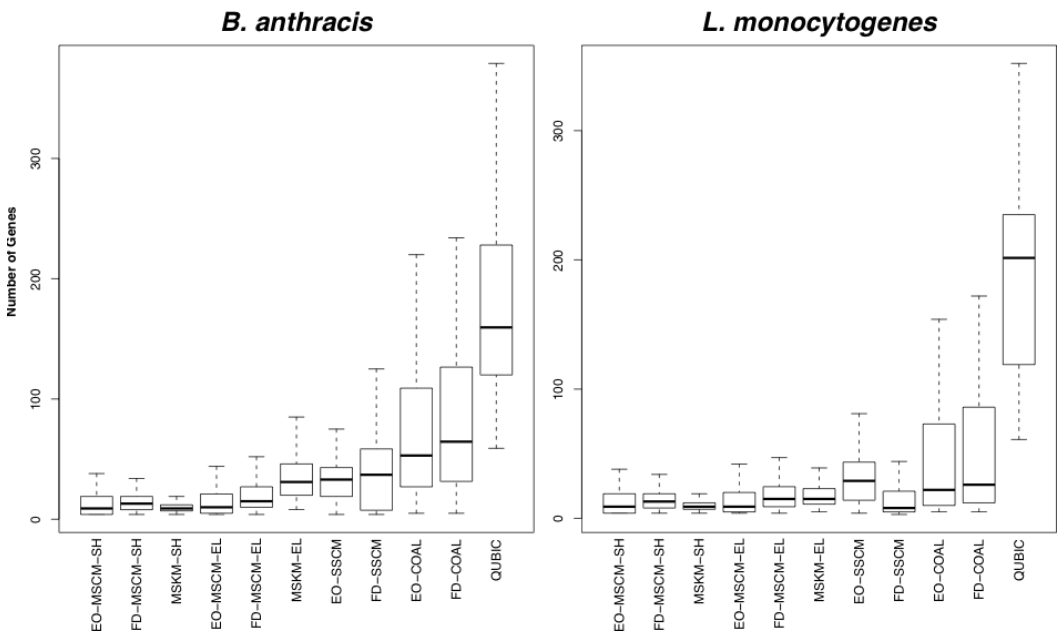


Figure 7.45: Number of genes from the *B. anthracis – L. monocytogenes* pairing

7.2.2.1.2 Figures for the Gram-negative triplet

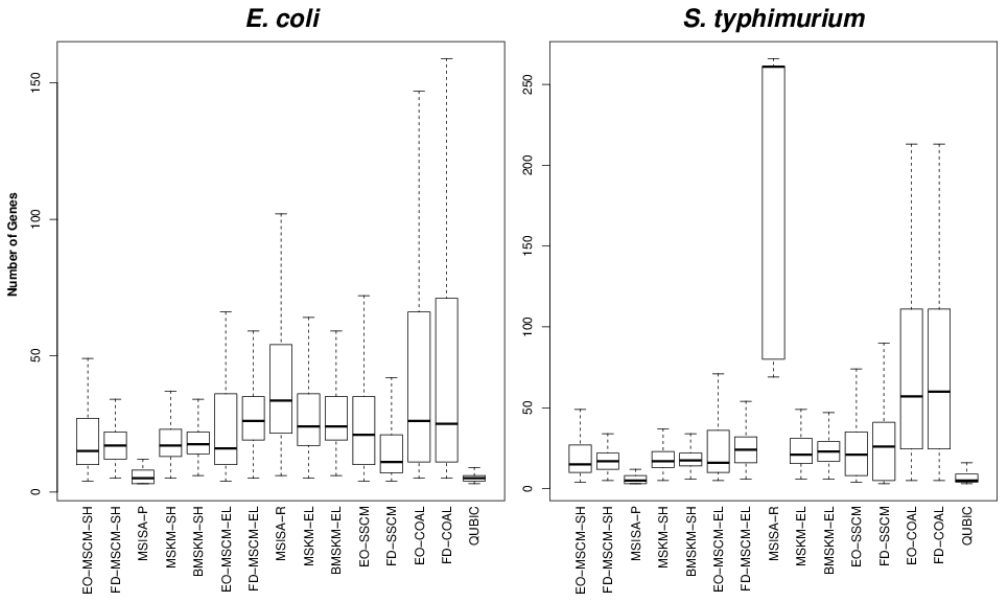


Figure 7.46: Number of genes from the *E. coli – S. typhimurium* pairing.

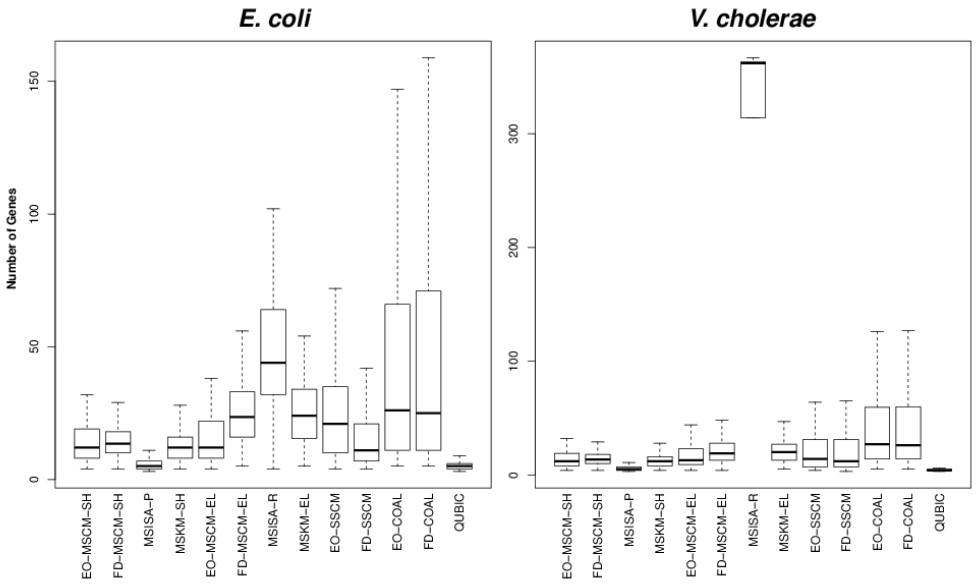


Figure 7.47: Number of genes from the *E. coli – V. cholerae* pairing

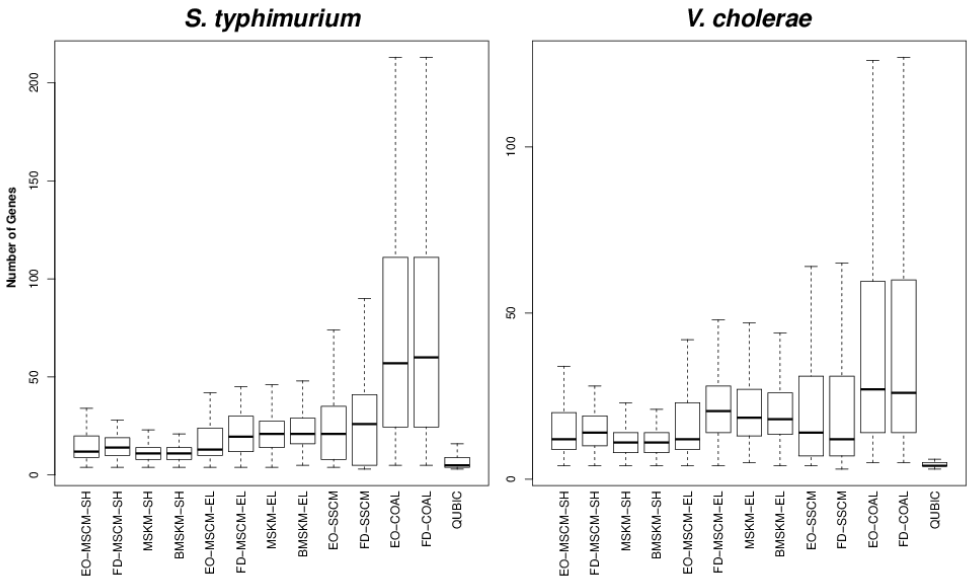


Figure 7.48: Number of genes from the *S. typhimurium* – *V. cholerae* pairing

7.2.2.2 Number of conditions

In each of the plots shown below are the distributions of the number of conditions from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.2.2.1 Figures for the Gram-positive triplet

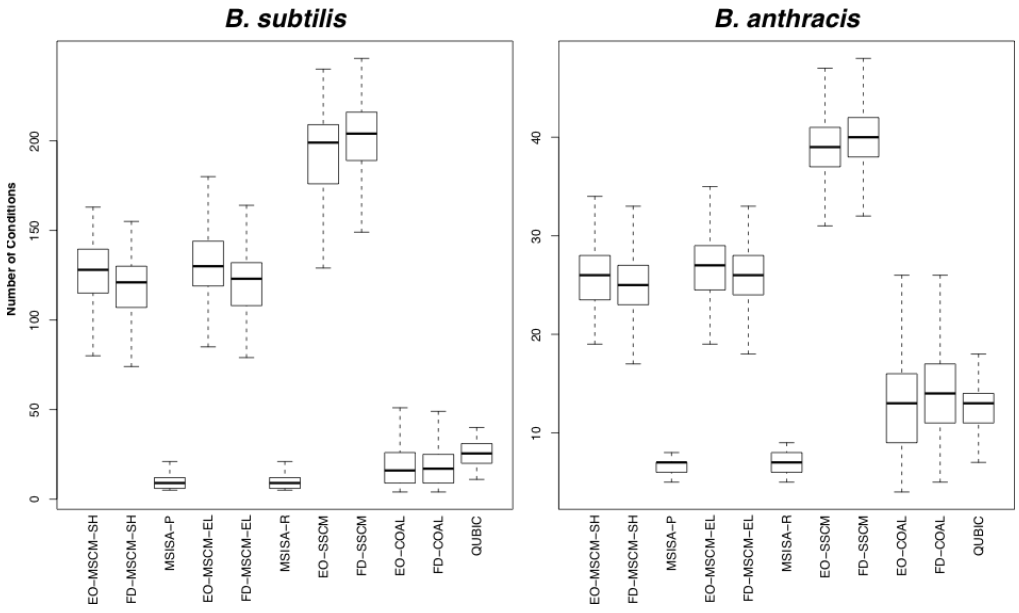


Figure 7.49: Number of conditions from the *B. subtilis – B. anthracis* pairing.

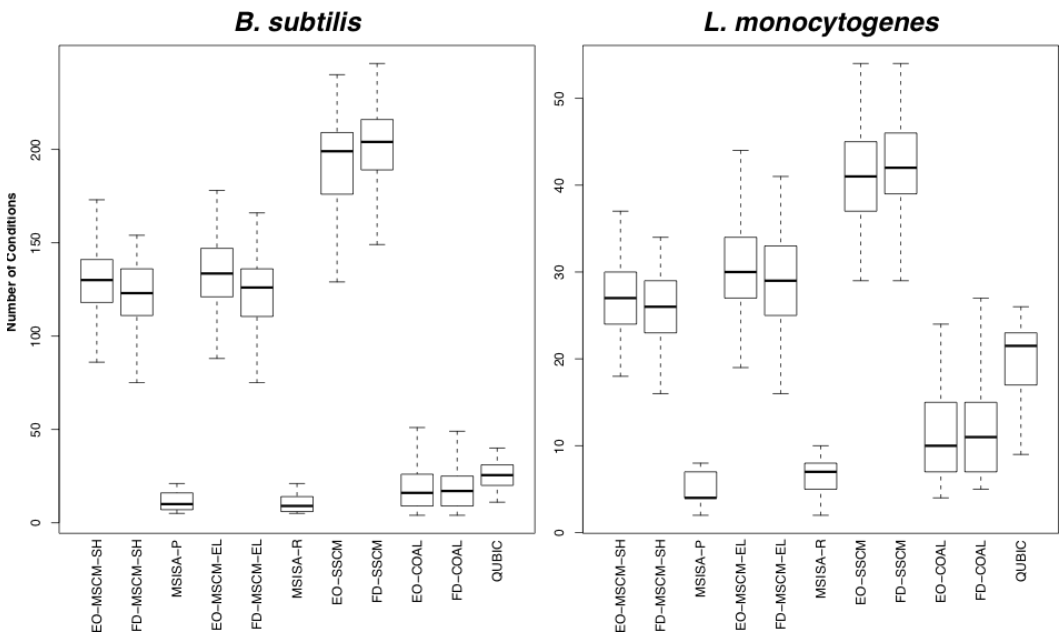


Figure 7.50: Number of conditions from the *B. subtilis – L. monocytogenes* pairing

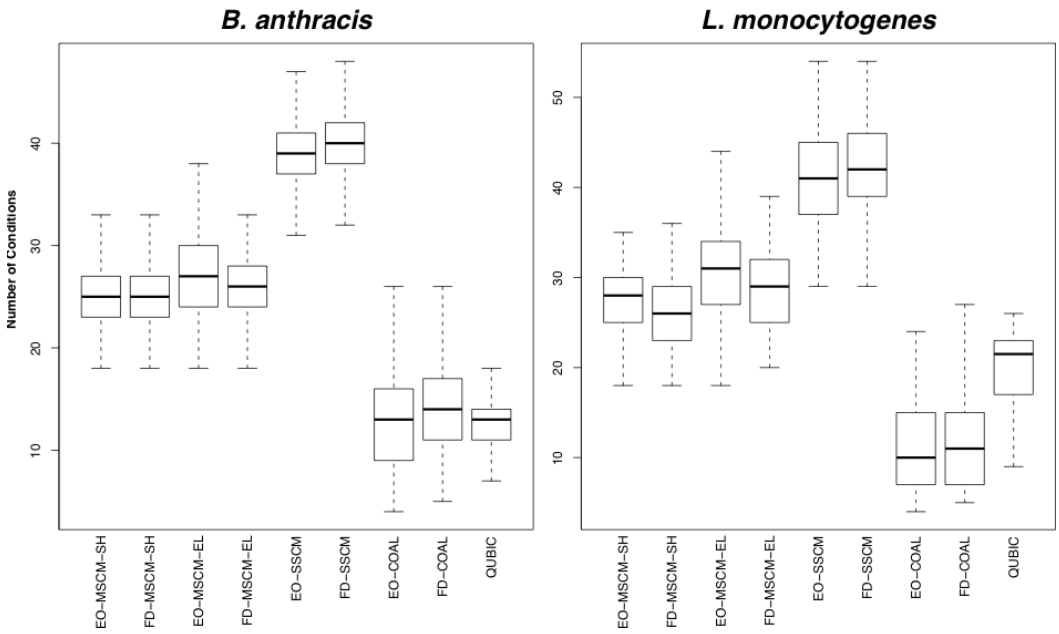


Figure 7.51: Number of conditions from the *B. anthracis – L. monocytogenes* pairing.

7.2.2.2.2 Figures for the Gram-negative triplet

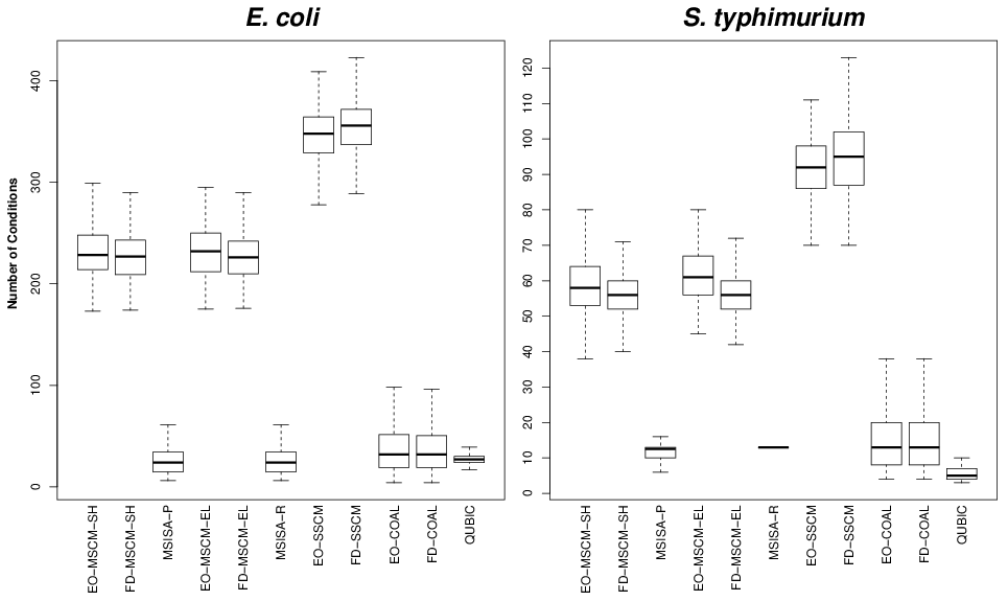


Figure 7.52: Number of conditions from the *E. coli – S. typhimurium* pairing.

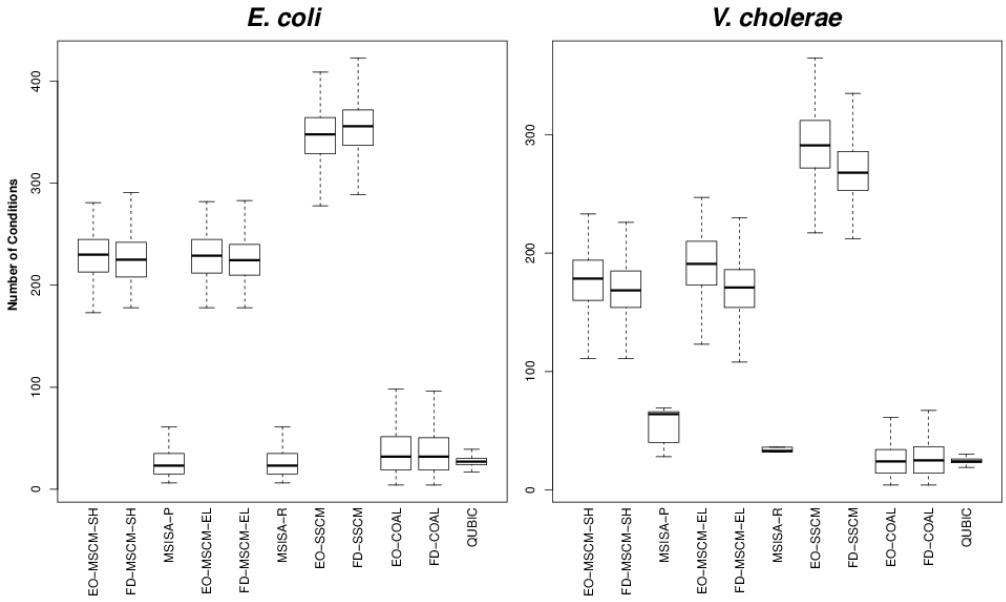


Figure 7.53: Number of conditions from the *E. coli – V. cholerae* pairing

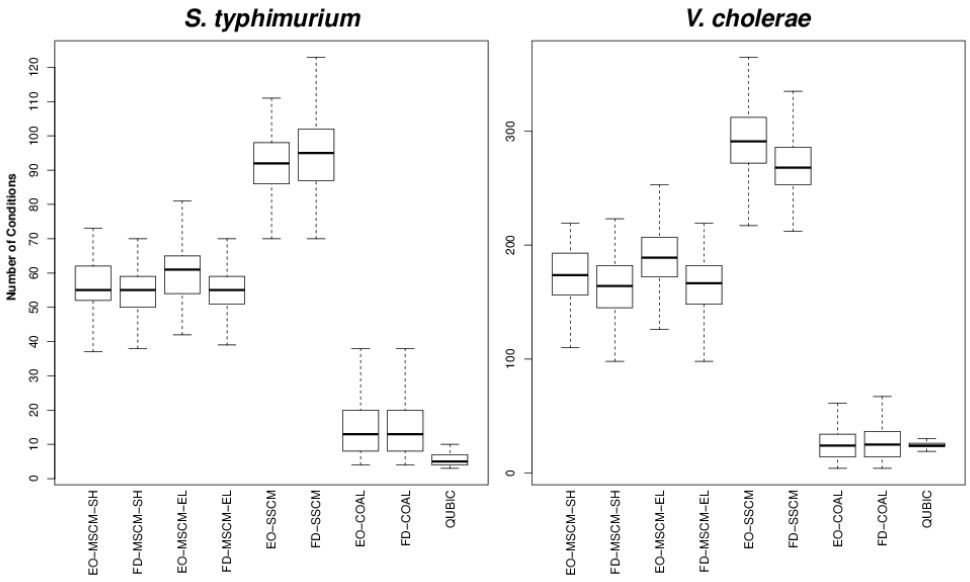


Figure 7.54: Number of conditions from the *S. typhimurium* – *V. cholerae* pairing.

7.2.2.3 Coverage (element-wise)

In each of the plots shown below are the distributions of the coverages (matrix element-wise) from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.2.3.1 Figures for the Gram-positive triplet

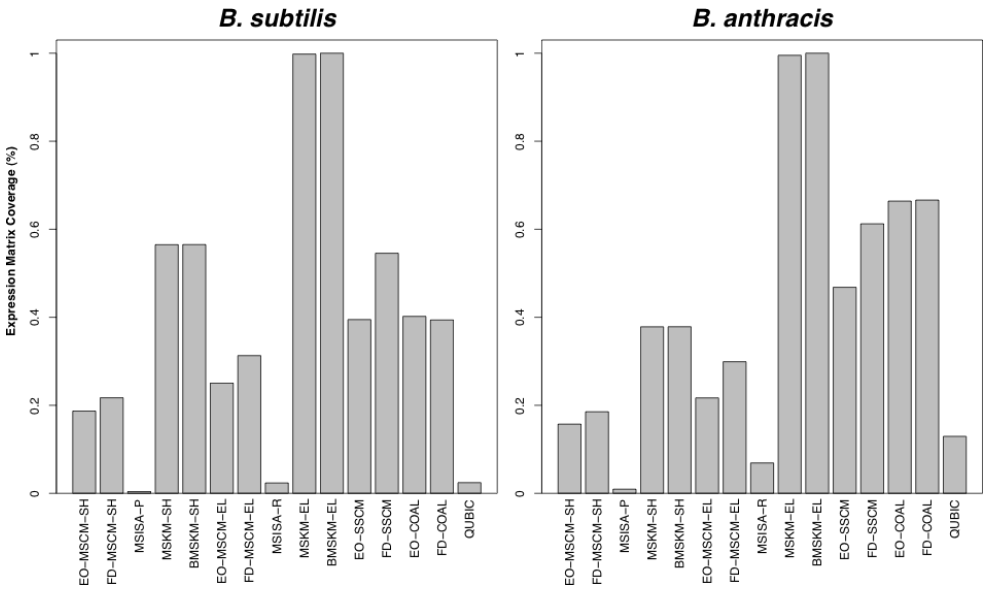


Figure 7.55: Coverages (matrix element-wise) from the *B. subtilis – B. anthracis* pairing.

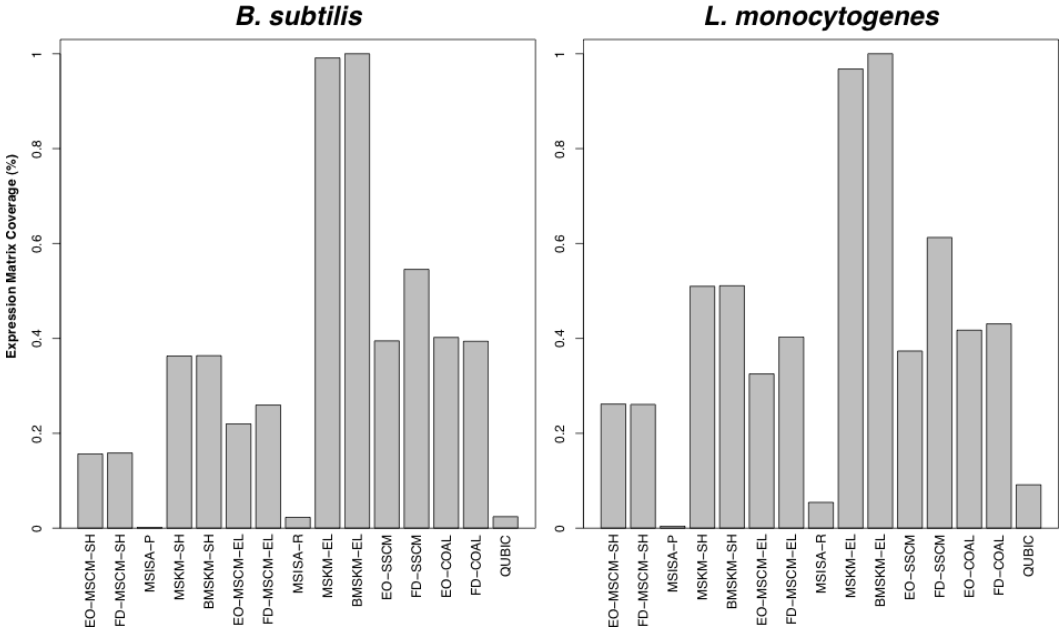


Figure 7.56: Coverages (matrix element-wise) from the *B. subtilis – L. monocytogenes* pairing.

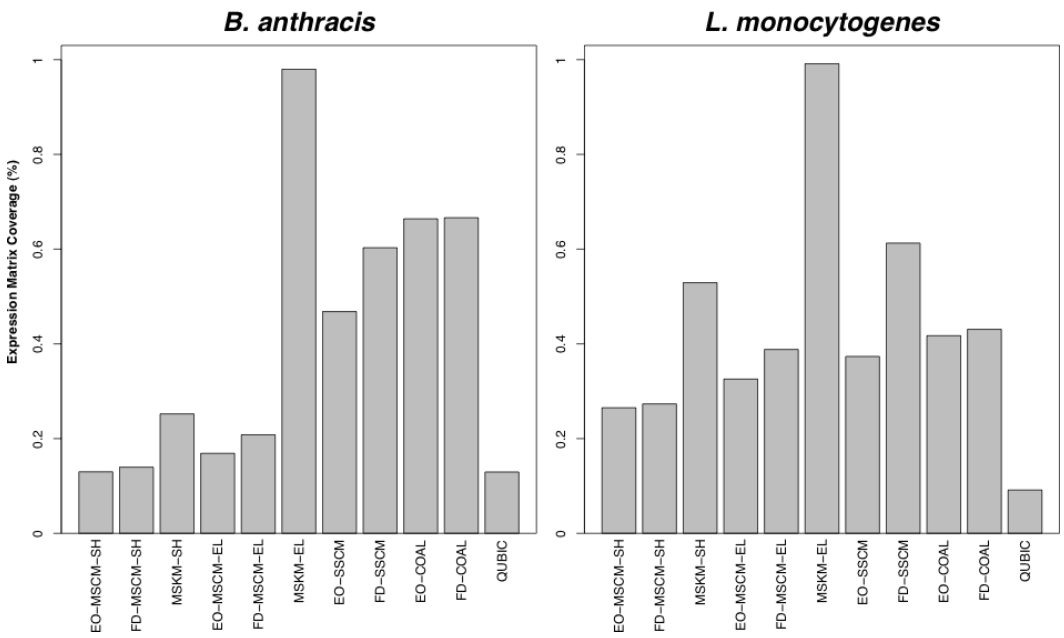


Figure 7.57: Coverages (matrix element-wise) from the *B. anthracis – L. monocytogenes* pairing.

7.2.2.3.2 Figures for the Gram-negative triplet

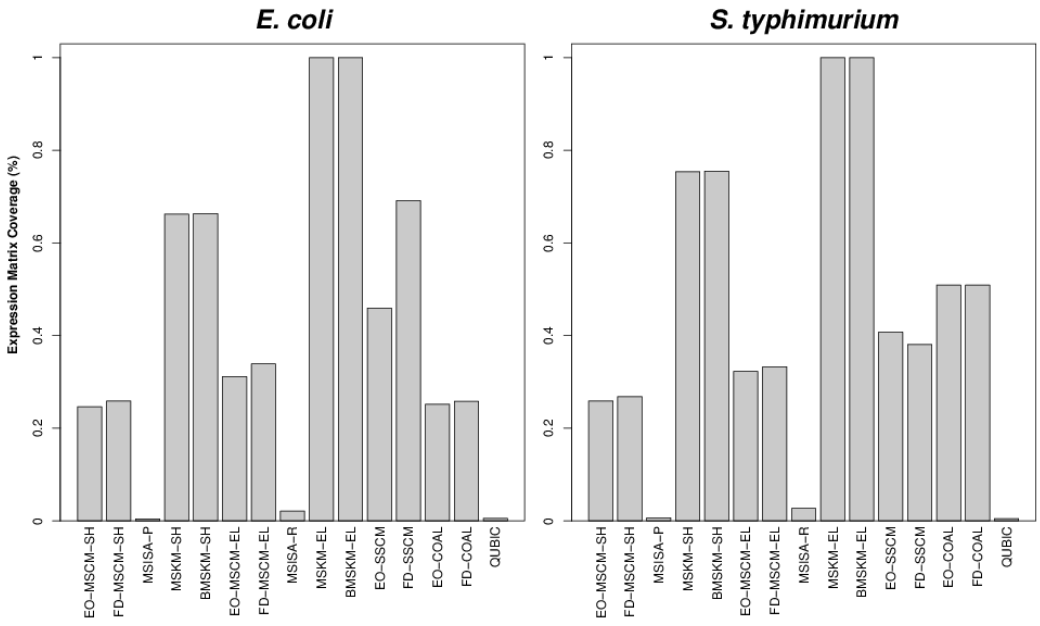


Figure 7.58: Coverages (matrix element-wise) from the *E. coli – S. typhimurium* pairing.

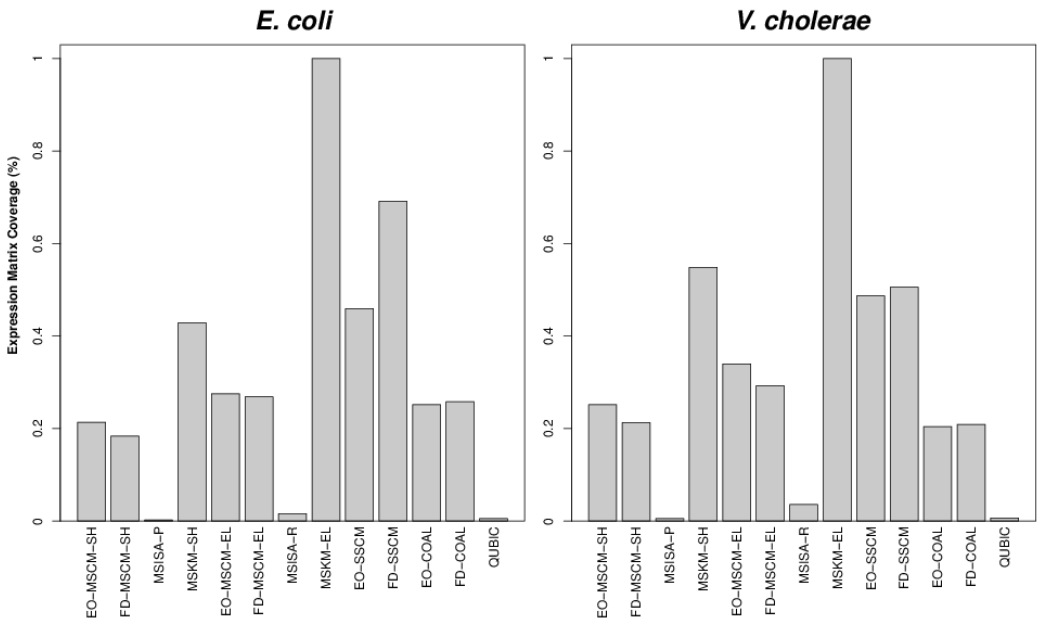


Figure 7.59: Coverages (matrix element-wise) from the *E. coli – V. cholerae* pairing.

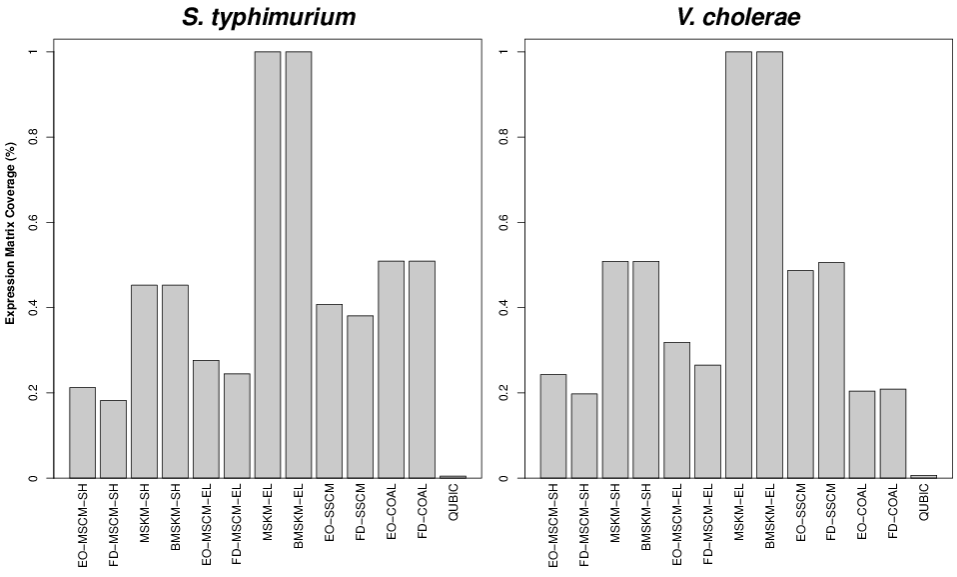


Figure 7.60: Coverages (matrix element-wise) from the *S. typhimurium – V. cholerae* pairing.

7.2.2.4 Coverage (gene-wise)

In each of the plots shown below are the distributions of the coverages (gene-wise) from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.2.4.1 Figures for the Gram-positive triplet

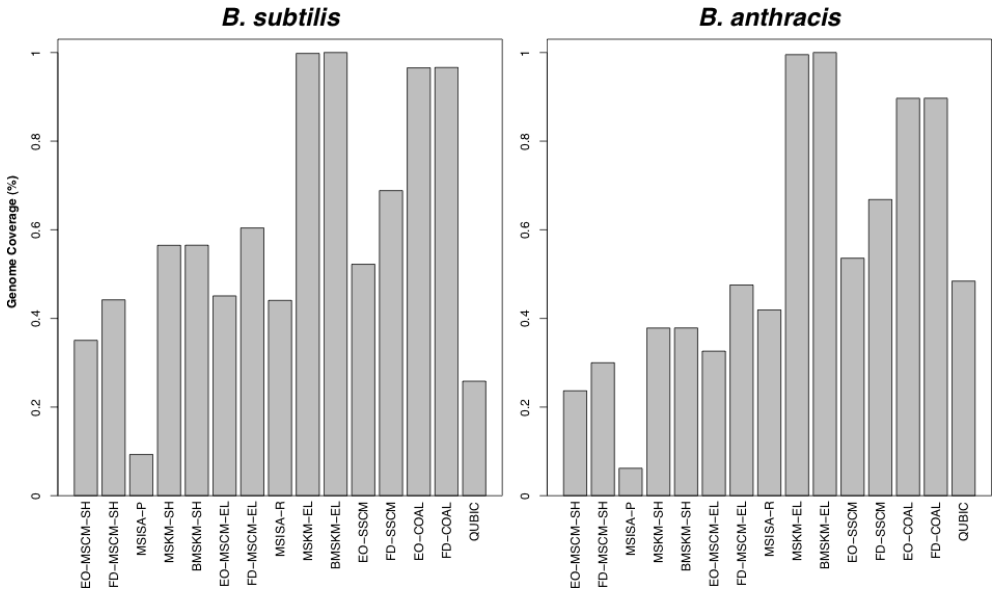


Figure 7.61: Coverages (gene-wise) from the *B. subtilis – B. anthracis* pairing.

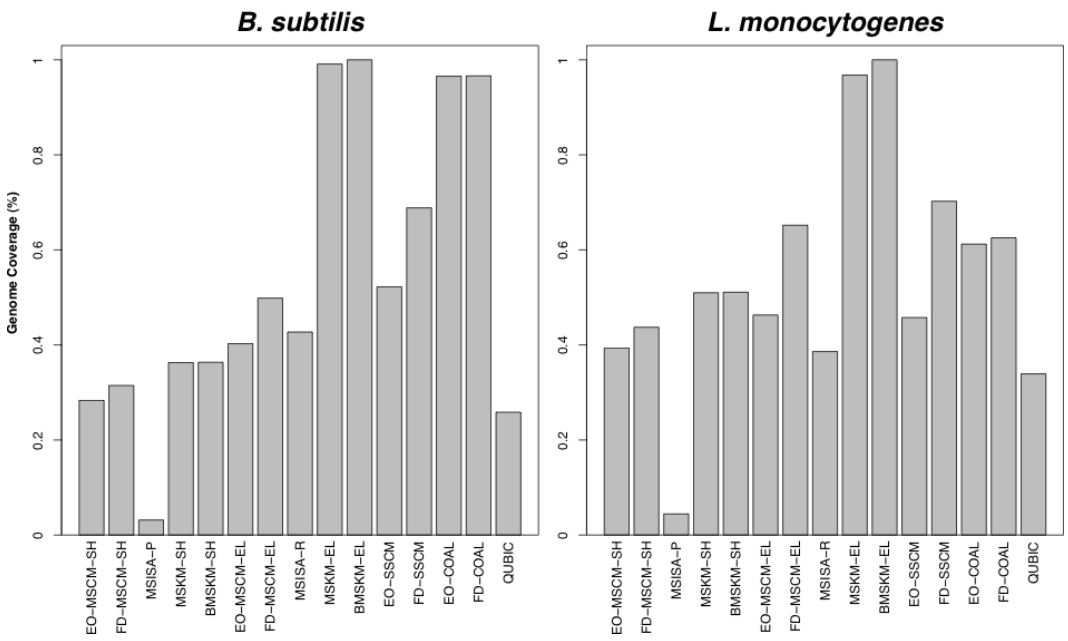


Figure 7.62: Coverages (gene-wise) from the *B. subtilis – L. monocytogenes* pairing.

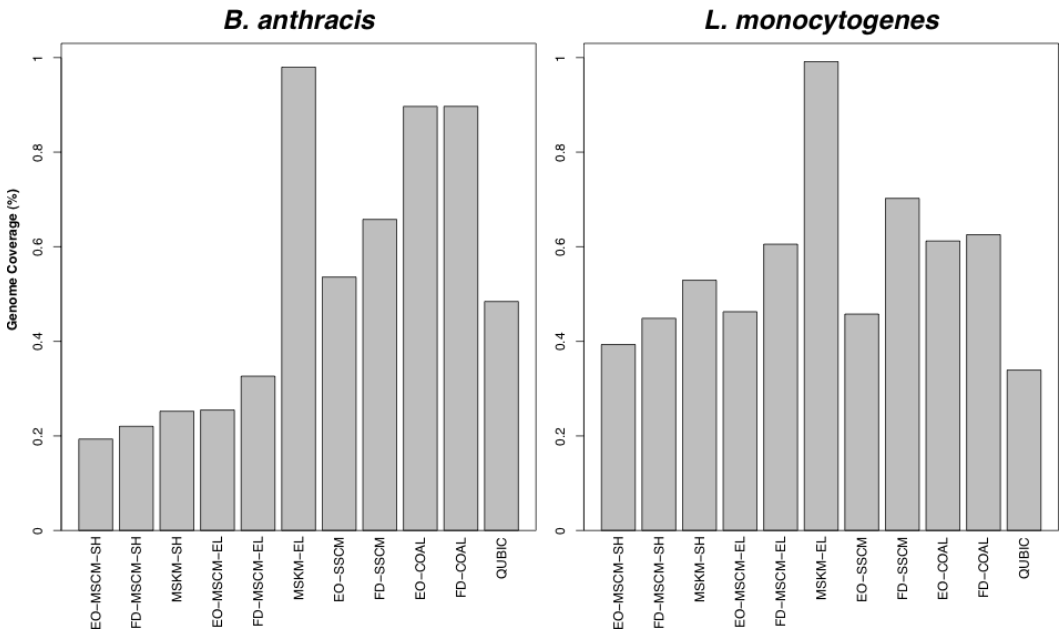


Figure 7.63: Coverages (gene-wise) from the *B. anthracis – L. monocytogenes* pairing.

7.2.2.4.2 Figures for the Gram-negative triplet

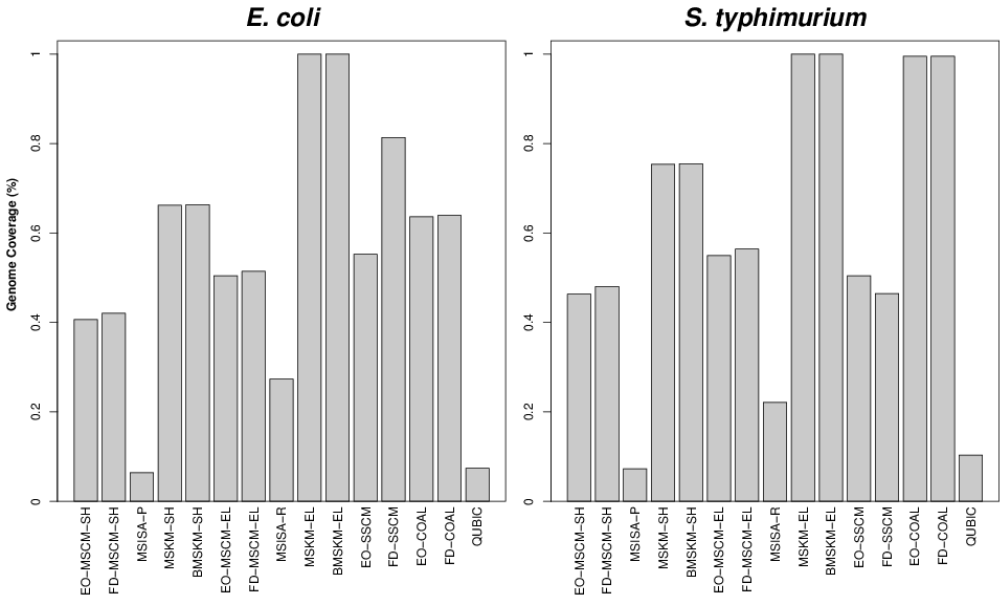


Figure 7.64: Coverages (gene-wise) from the *E. coli – S. typhimurium* pairing.

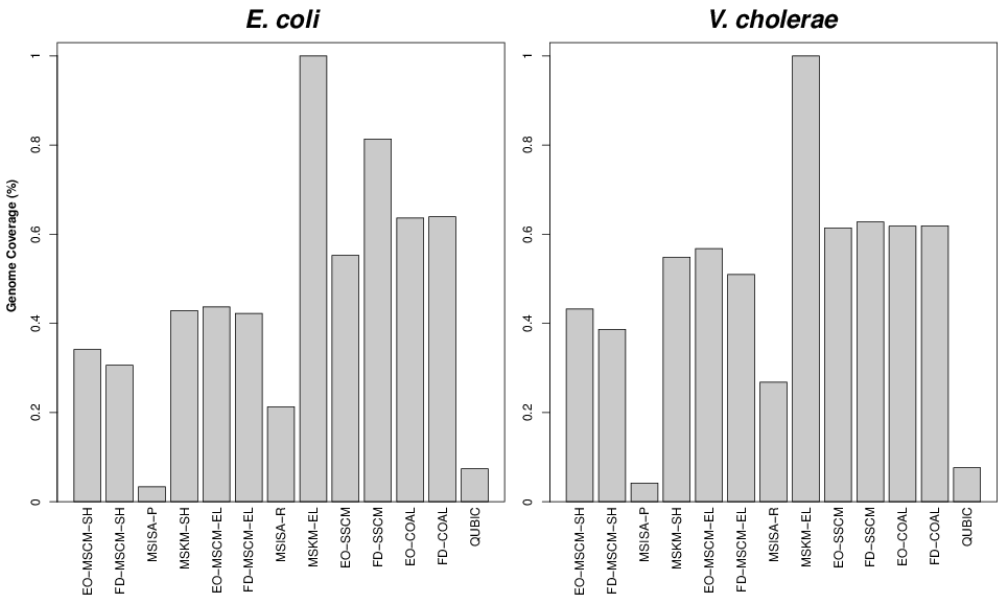


Figure 7.65: Coverages (gene-wise) from the *E. coli* – *V. cholerae* pairing.

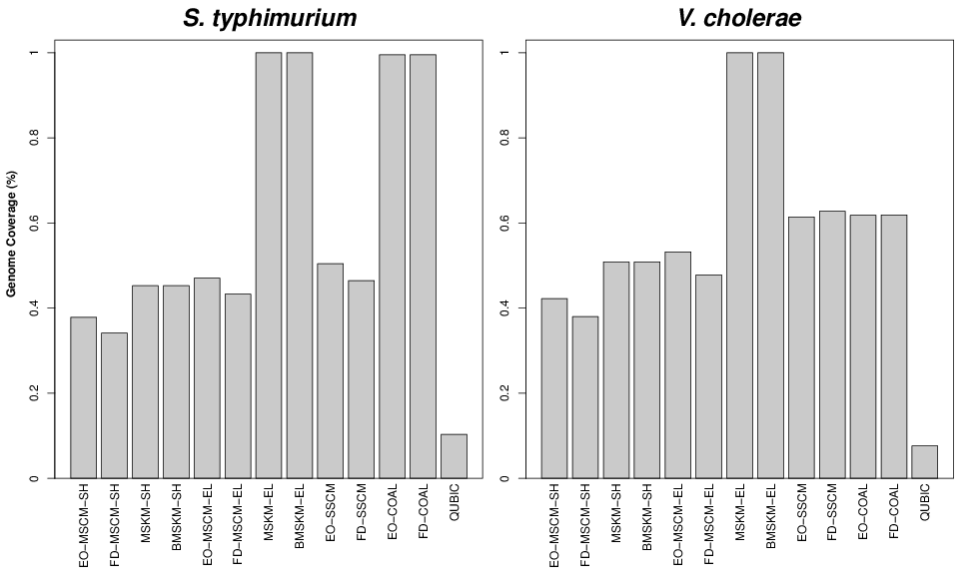


Figure 7.66: Coverages (gene-wise) from the *S. typhimurium* – *V. cholerae* pairing.

7.2.2.5 Overlap (element-wise)

In each of the plots shown below are the distributions of the overlaps (matrix element-wise) from all methods considered by this study for a given pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.2.5.1 Figures for the Gram-positive triplet

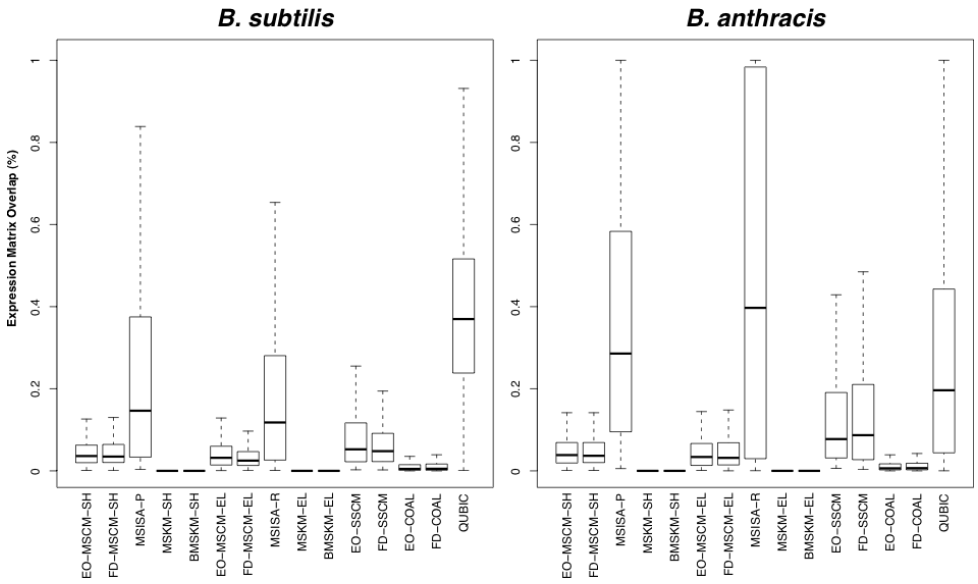


Figure 7.67: Overlaps (matrix element-wise) from the *B. subtilis – B. anthracis* pairing.

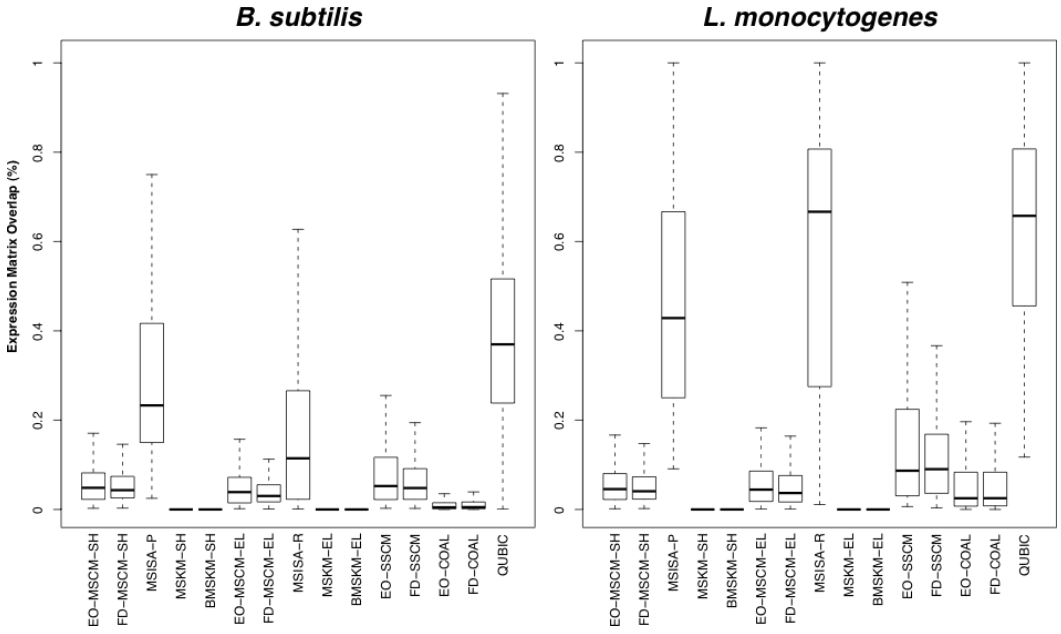


Figure 7.68: Overlaps (matrix element-wise) from the *B. subtilis – L. monocytogenes* pairing.

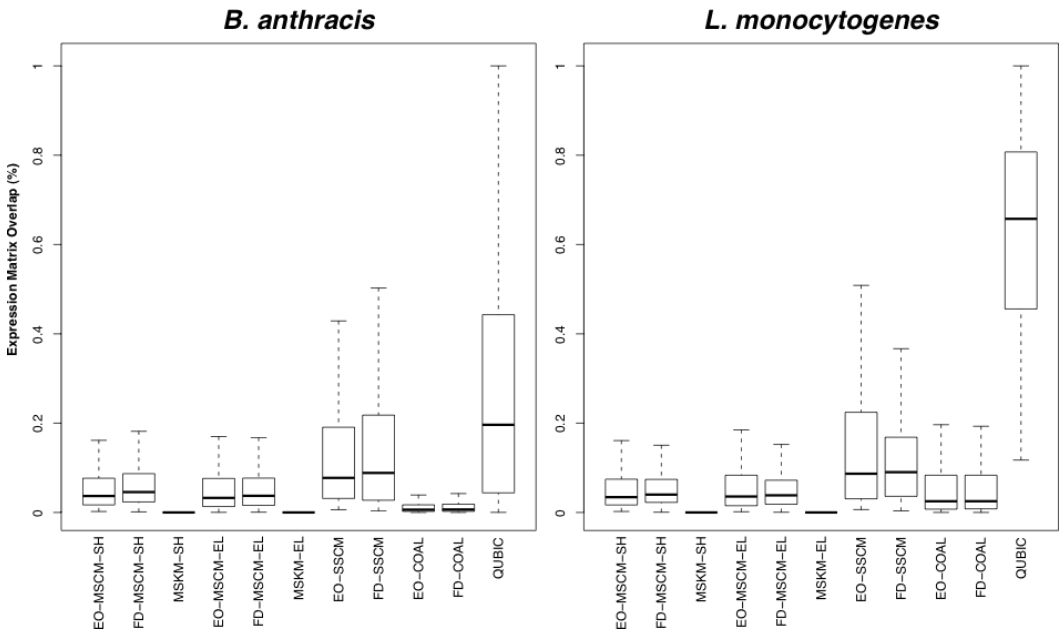


Figure 7.69: Overlaps (matrix element-wise) from the *B. anthracis – L. monocytogenes* pairing.

7.2.2.5.2 Figures for the Gram-negative triplet

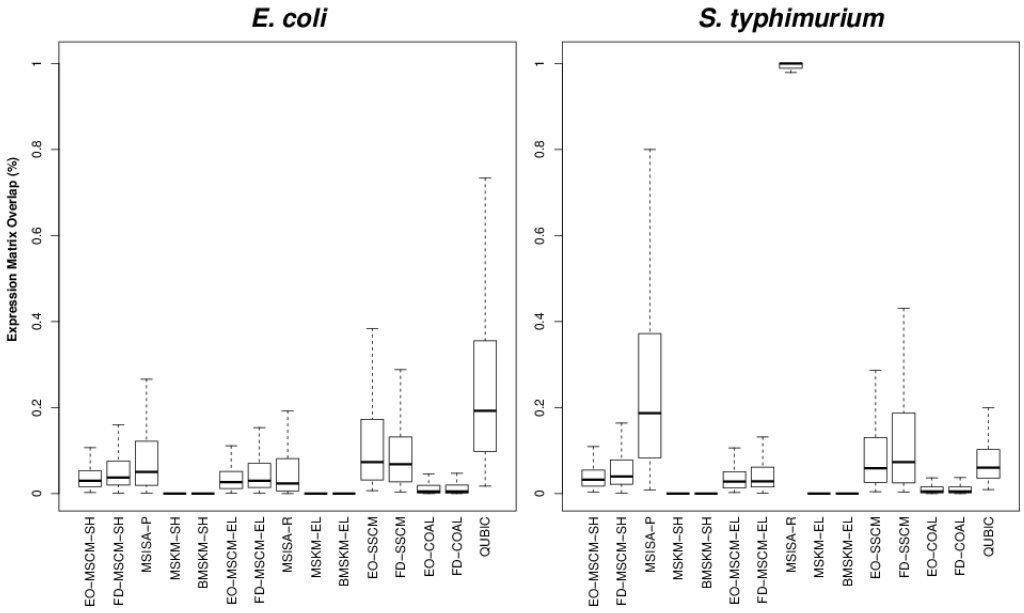


Figure 7.70: Overlaps (matrix element-wise) from the *E. coli – S. typhimurium* pairing.

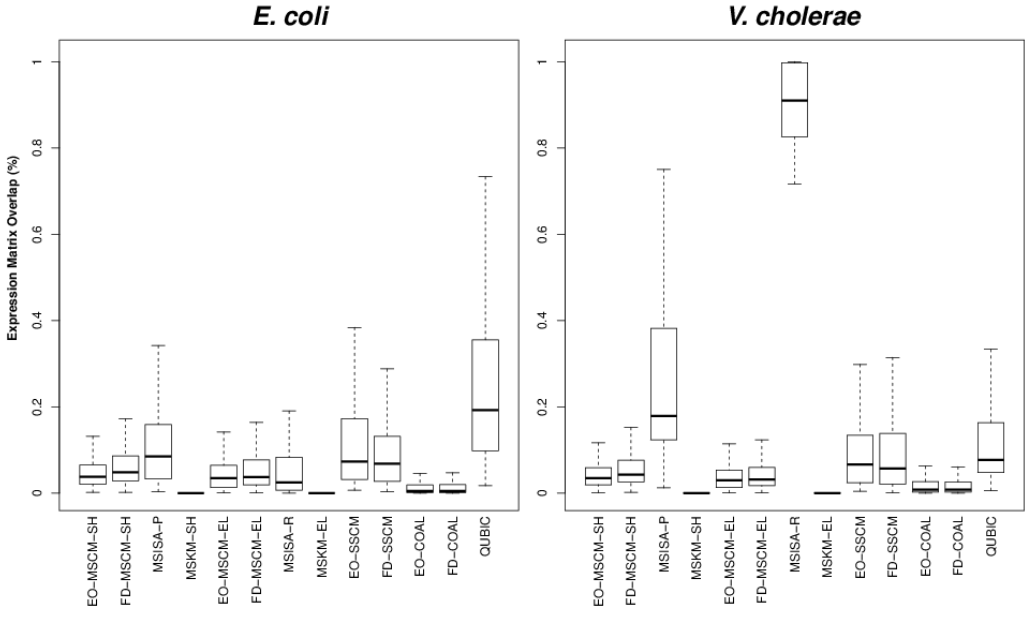


Figure 7.71: Overlaps (matrix element-wise) from the *E. coli – V. cholerae* pairing.

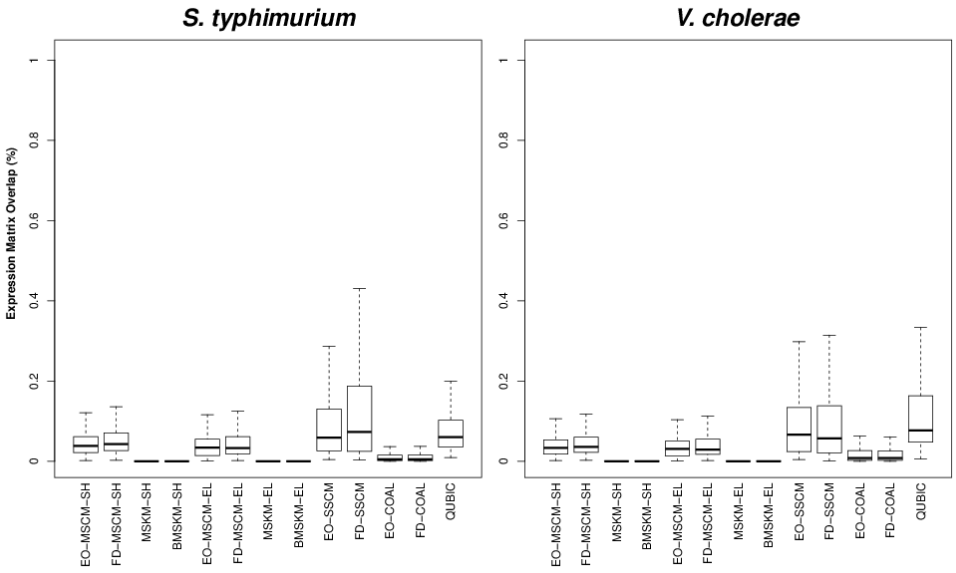


Figure 7.72: Overlaps (matrix element-wise) from the *S. typhimurium* – *V. cholerae* pairing.

7.2.2.6 Overlap (gene-wise)

In each of the plots shown below are the distributions of the Overlaps (gene-wise) from all methods considered by this study for the *B. subtilis*- *B. anthracis* pairing. Explanations of the method name abbreviations can be found in Table 3.1.

7.2.2.6.1 Figures for the Gram-positive triplet

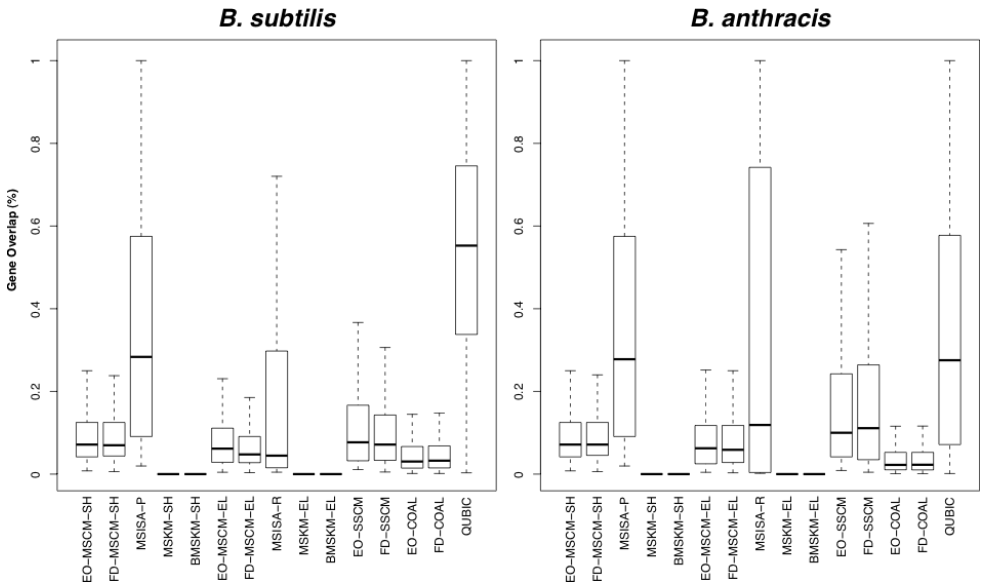


Figure 7.73: Overlaps (gene-wise) from the *B. subtilis – B. anthracis* pairing.

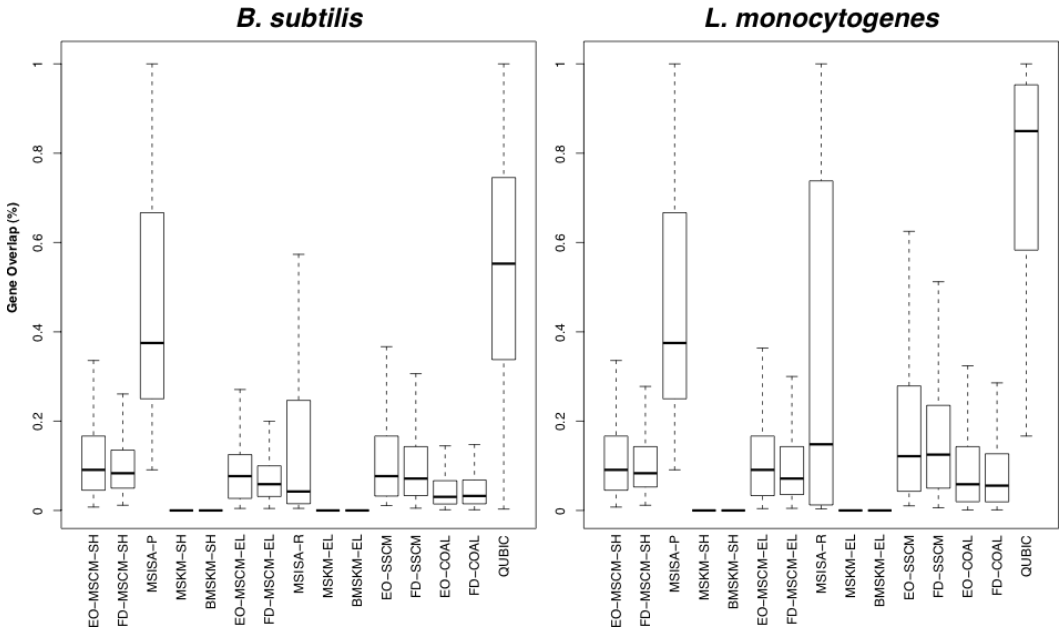


Figure 7.74: Overlaps (gene-wise) from the *B. subtilis – L. monocytogenes* pairing.

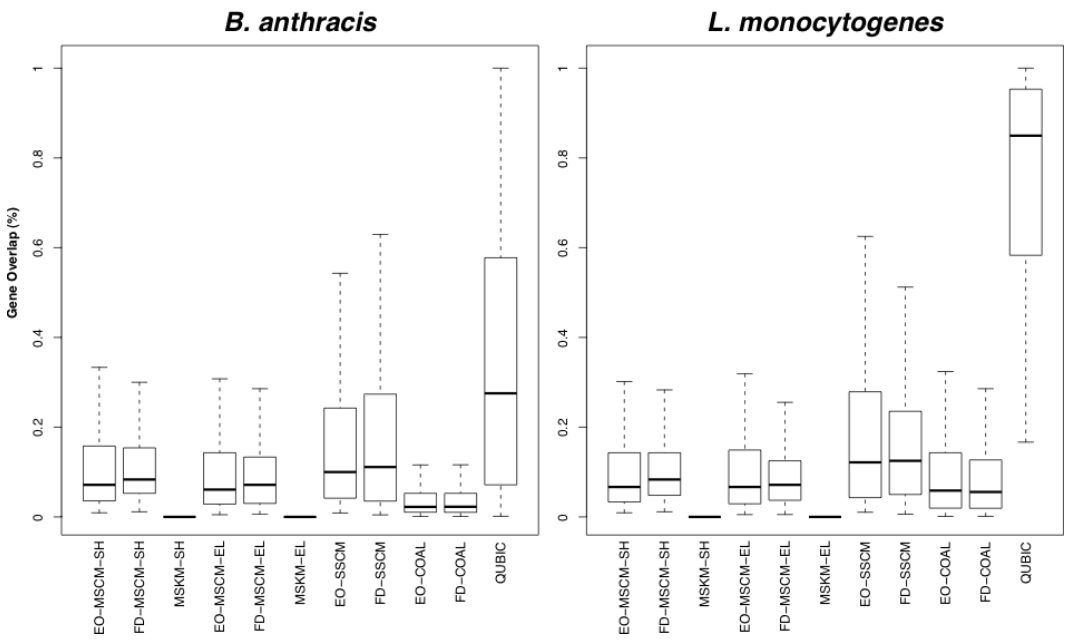


Figure 7.75: Overlaps (gene-wise) from the *B. anthracis – L. monocytogenes* pairing.

7.2.2.6.2 Figures for the Gram-negative triplet

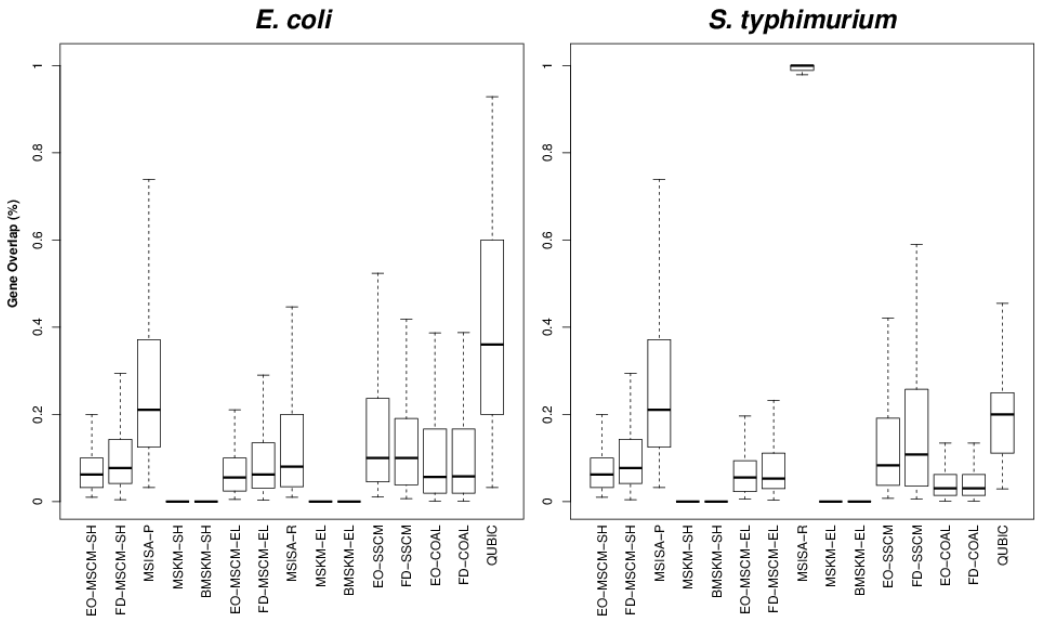


Figure 7.76: Overlaps (gene-wise) from the *E. coli – S. typhimurium* pairing.

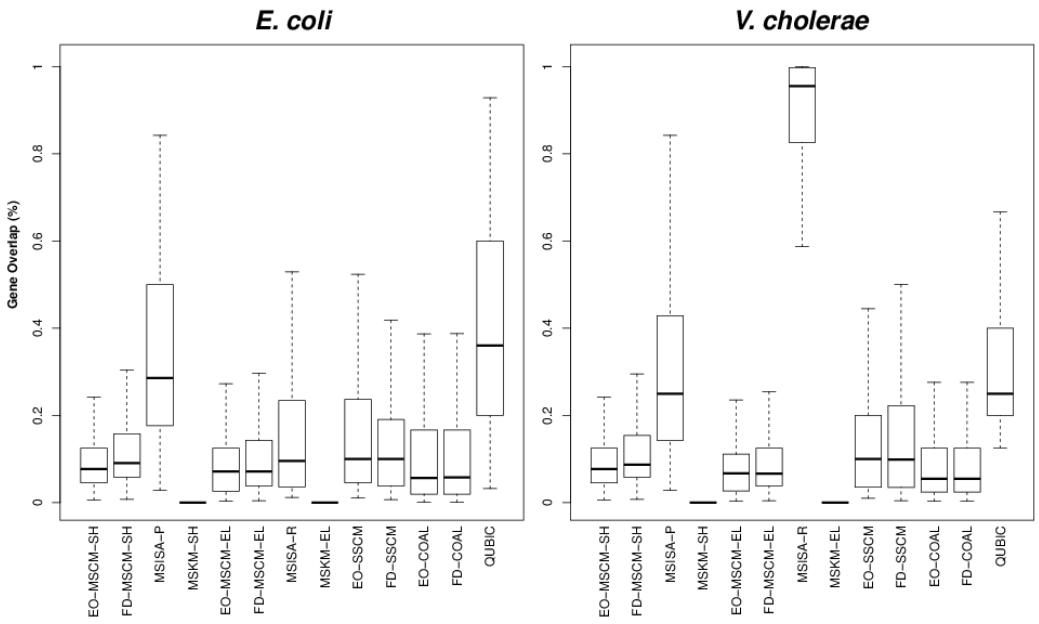


Figure 7.77: Overlaps (gene-wise) from the *E. coli – V. cholerae* pairing.

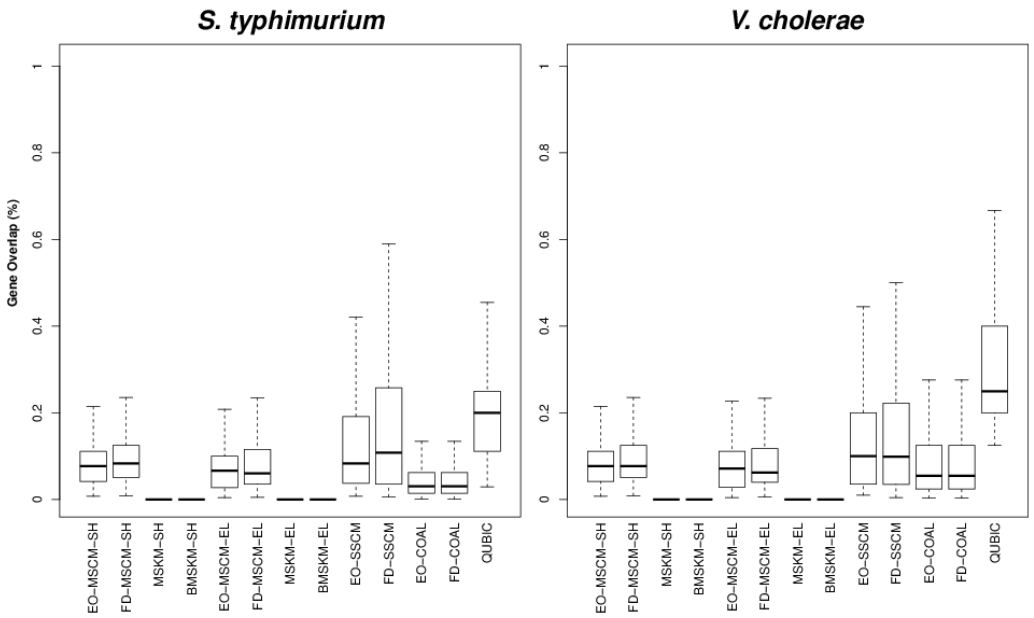


Figure 7.78: Overlaps (gene-wise) from the *S. typhimurium* – *V. cholerae* pairing.

7.2.3 Comparison of the (bi)cluster coherence metrics

7.2.3.1 Comparisons with FD-MScM

7.2.3.1.1 Residuals

In the tables below, we present a comparison of the residuals of the results from MScM (full data) with all other relevant methods for all 3 pairings of a given triplet that's examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as

well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We use 'dist.' as an abbreviation for distribution, and the "**dist1 vs. dist2**" column to represent both the distributions being compared and their order in the table. Therefore, for example, the **FD-MScM-SH vs. FD-SSCM** row displays the comparison of the distributions of residuals from the shared MS cMonkey results with those from the SS cMonkey, for the appropriate organism, with the **FD-MScM-SH** as **dist1** (and **FD-SSCM** as **dist2**). In addition, we color-code the **Wilcoxon's 2-sided** column for a given organism to indicate whether the test indicated the distributions were the same or different, and if different, the distribution with the better overall score, as determined by the metric (**Residuals**). In this scheme, we use **green** to indicate dist1 had a statistically better score, **red** if dist2, and **yellow** to indicate a tie. Therefore, as the MScM results are always the **dist1** in these comparisons, this color scheme allows one to quickly and easily determine the overall frequency with which MScM did as well or better than the other methods. In the case of the Gram-positive triplet, these results illustrate that in 71 of the 92 comparisons (77.2%) MScM step did as well or better than its competitors. Similarly, in 47 of the 92 comparisons (51%) for the Gram-negative triplet, MScM did as well or better than its competitors.

7.2.3.1.1.1 Gram-positive triplet

Table 7.25: Comparison of bicluster residuals from the full data methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*.

B. subtilis - B. anthracis pairing

dist1 vs. dist2	*B. subtilis*			*B. anthracis*		
	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
FD-MScM-SH vs. FD-SSCM	0.51 ± 0.08	**0.001**	0.49 ± 0.13	0.30 ± 0.09	**0.456**	0.31 ± 0.12
FD-MScM-SH vs. QUBIC	0.51 ± 0.08	**3.69E-37**	0.87 ± 0.21	0.30 ± 0.09	**1.08E-50**	1.51 ± 0.29
FD-MScM-SH vs. FD-COAL	0.51 ± 0.08	**9.94E-44**	0.80 ± 0.25	0.30 ± 0.09	**4.77E-38**	0.58 ± 0.17
FD-MScM-EL vs. FD-SSCM	0.49 ± 0.09	**0.273**	0.49 ± 0.13	0.32 ± 0.09	**0.036**	0.31 ± 0.12
FD-MScM-EL vs. QUBIC	0.49 ± 0.09	**5.60E-38**	0.87 ± 0.21	0.32 ± 0.09	**1.08E-50**	1.51 ± 0.29
FD-MScM-EL vs. FD-COAL	0.49 ± 0.09	**6.10E-47**	0.80 ± 0.25	0.32 ± 0.09	**8.43E-36**	0.58 ± 0.17
FD-MScM-SH vs. MSISA-P	0.51 ± 0.08	**2.62E-17**	0.98 ± 0.39	0.30 ± 0.09	**1.11E-22**	1.97 ± 0.94
FD-MScM-SH vs. MSISA-R	0.51 ± 0.08	**5.99E-20**	1.11 ± 0.41	0.30 ± 0.09	**1.11E-22**	1.58 ± 0.38
FD-MScM-SH vs. MSKM-SH	0.51 ± 0.08	**1.52E-25**	0.41 ± 0.07	0.30 ± 0.09	**7.05E-38**	0.53 ± 0.12
FD-MScM-SH vs. MSKM-EL	0.51 ± 0.08	**4.31E-23**	0.42 ± 0.06	0.30 ± 0.09	**2.63E-33**	0.48 ± 0.11
FD-MScM-SH vs. BMSKM-SH	0.51 ± 0.08	**2.68E-11**	0.45 ± 0.07	0.30 ± 0.09	**5.36E-16**	0.38 ± 0.07

FD-MScM-SH vs. BMSKM-EL	0.51 ± 0.08	**1.84E-11**	0.45 ± 0.06	0.30 ± 0.09	**4.79E-18**	0.39 ± 0.07
FD-MScM-EL vs. MSISA-P	0.49 ± 0.09	**9.21E-18**	0.98 ± 0.39	0.32 ± 0.09	**1.11E-22**	1.97 ± 0.94
FD-MScM-EL vs. MSISA-R	0.49 ± 0.09	**5.02E-20**	1.11 ± 0.41	0.32 ± 0.09	**1.14E-22**	1.58 ± 0.38
FD-MScM-EL vs. MSKM-SH	0.49 ± 0.09	**1.33E-17**	0.41 ± 0.07	0.32 ± 0.09	**2.40E-35**	0.53 ± 0.12
FD-MScM-EL vs. MSKM-EL	0.49 ± 0.09	**4.38E-15**	0.42 ± 0.06	0.32 ± 0.09	**2.41E-29**	0.48 ± 0.11
FD-MScM-EL vs. BMSKM-SH	0.49 ± 0.09	**6.91E-05**	0.45 ± 0.07	0.32 ± 0.09	**3.02E-11**	0.38 ± 0.07
FD-MScM-EL vs. BMSKM-EL	0.49 ± 0.09	**1.30E-04**	0.45 ± 0.06	0.32 ± 0.09	**3.34E-13**	0.39 ± 0.07

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.52 ± 0.08	**2.82E-04**	0.49 ± 0.13	0.34 ± 0.12	**0.006**	0.40 ± 0.18
FD-MScM-SH vs. QUBIC	0.52 ± 0.08	**1.31E-36**	0.87 ± 0.21	0.34 ± 0.12	**5.94E-37**	1.81 ± 0.85

FD-MScM-SH vs. FD-COAL	0.52 ± 0.08	**3.16E-42**	0.80 ± 0.25	0.34 ± 0.12	**1.91E-08**	1.70 ± 3.24
FD-MScM-EL vs. FD-SSCM	0.50 ± 0.10	**0.088**	0.49 ± 0.13	0.34 ± 0.12	**0.004**	0.40 ± 0.18
FD-MScM-EL vs. QUBIC	0.50 ± 0.10	**2.05E-36**	0.87 ± 0.21	0.34 ± 0.12	**3.99E-37**	1.81 ± 0.85
FD-MScM-EL vs. FD-COAL	0.50 ± 0.10	**1.35E-44**	0.80 ± 0.25	0.34 ± 0.12	**1.11E-08**	1.70 ± 3.24
FD-MScM-SH vs. MSISA-P	0.52 ± 0.08	**1.16E-09**	0.87 ± 0.34	0.34 ± 0.12	**3.21E-19**	1.59 ± 0.52
FD-MScM-SH vs. MSISA-R	0.52 ± 0.08	**1.90E-18**	1.11 ± 0.42	0.34 ± 0.12	**2.30E-21**	1.31 ± 0.34
FD-MScM-SH vs. MSKM-SH	0.52 ± 0.08	**1.07E-31**	0.40 ± 0.07	0.34 ± 0.12	**3.63E-22**	0.50 ± 0.12
FD-MScM-SH vs. MSKM-EL	0.52 ± 0.08	**1.07E-25**	0.42 ± 0.06	0.34 ± 0.12	**4.44E-19**	0.48 ± 0.11
FD-MScM-SH vs. BMSKM-SH	0.52 ± 0.08	**2.52E-19**	0.43 ± 0.07	0.34 ± 0.12	**1.37E-10**	0.42 ± 0.09
FD-MScM-SH vs. BMSKM-EL	0.52 ± 0.08	**3.47E-17**	0.44 ± 0.06	0.34 ± 0.12	**2.35E-11**	0.42 ± 0.09
FD-MScM-EL vs. MSISA-P	0.50 ± 0.10	**8.82E-10**	0.87 ± 0.34	0.34 ± 0.12	**3.21E-19**	1.59 ± 0.52
FD-MScM-EL vs. MSISA-R	0.50 ± 0.10	**2.99E-18**	1.11 ± 0.42	0.34 ± 0.12	**2.30E-21**	1.31 ± 0.34

FD-MScM-EL vs. MSKM-SH	0.50 ± 0.10	**2.23E-25**	0.40 ± 0.07	0.34 ± 0.12	**1.17E-22**	0.50 ± 0.12
FD-MScM-EL vs. MSKM-EL	0.50 ± 0.10	**1.13E-18**	0.42 ± 0.06	0.34 ± 0.12	**1.32E-19**	0.48 ± 0.11
FD-MScM-EL vs. BMSKM-SH	0.50 ± 0.10	**6.56E-13**	0.43 ± 0.07	0.34 ± 0.12	**7.95E-11**	0.42 ± 0.09
FD-MScM-EL vs. BMSKM-EL	0.50 ± 0.10	**8.10E-11**	0.44 ± 0.06	0.34 ± 0.12	**8.42E-12**	0.42 ± 0.09

B. anthracis - L. monocytogenes pairing

	B. anthracis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.33 ± 0.10	**0.004**	0.31 ± 0.12	0.36 ± 0.14	**0.097**	0.40 ± 0.18
FD-MScM-SH vs. QUBIC	0.33 ± 0.10	**2.30E-50**	1.51 ± 0.29	0.36 ± 0.14	**4.38E-35**	1.81 ± 0.85
FD-MScM-SH vs. FD-COAL	0.33 ± 0.10	**1.34E-33**	0.58 ± 0.17	0.36 ± 0.14	**1.83E-06**	1.70 ± 3.24
FD-MScM-EL vs. FD-SSCM	0.36 ± 0.11	**2.95E-07**	0.31 ± 0.12	0.36 ± 0.13	**0.049**	0.40 ± 0.18
FD-MScM-EL vs. QUBIC	0.36 ± 0.11	**2.30E-50**	1.51 ± 0.29	0.36 ± 0.13	**8.30E-36**	1.81 ± 0.85

FD-MScM-EL vs. FD-COAL	0.36 ± 0.11	**8.77E-28**	0.58 ± 0.17	0.36 ± 0.13	**4.48E-07**	1.70 ± 3.24
FD-MScM-SH vs. MSKM-SH	0.33 ± 0.10	**2.89E-11**	0.40 ± 0.08	0.36 ± 0.14	**2.64E-10**	0.43 ± 0.08
FD-MScM-SH vs. MSKM-EL	0.33 ± 0.10	**3.35E-11**	0.39 ± 0.07	0.36 ± 0.14	**3.47E-09**	0.43 ± 0.08
FD-MScM-EL vs. MSKM-SH	0.36 ± 0.11	**5.90E-05**	0.40 ± 0.08	0.36 ± 0.13	**8.73E-12**	0.43 ± 0.08
FD-MScM-EL vs. MSKM-EL	0.36 ± 0.11	**9.70E-05**	0.39 ± 0.07	0.36 ± 0.13	**1.27E-10**	0.43 ± 0.08

7.2.3.1.1.2 Gram-negative triplet

Table 7.26: Comparison of bicluster residuals from the full data methods considered by this study for all pairings of *E. coli, S. typhimurium* and *V. cholerae*.

E. coli – S. typhimurium pairing

	E. coli			***S. typhimurium***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.45 ± 0.09	**4.31E-08**	0.50 ± 0.10	0.57 ± 0.07	**1.34E-23**	0.46 ± 0.09
FD-MScM-SH vs. QUBIC	0.45 ± 0.09	**1.07E-23**	0.29 ± 0.13	0.57 ± 0.07	**6.48E-07**	0.58 ± 0.54
FD-MScM-SH vs. FD-COAL	0.45 ± 0.09	**1.87E-13**	0.61 ± 0.34	0.57 ± 0.07	**1.29E-12**	0.67 ± 0.15

FD-MScM-EL vs. FD-SSCM	0.47 ± 0.10	**2.69E-04**	0.50 ± 0.10	0.60 ± 0.08	**2.38E-30**	0.46 ± 0.09
FD-MScM-EL vs. QUBIC	0.47 ± 0.10	**2.73E-26**	0.29 ± 0.13	0.60 ± 0.08	**1.82E-08**	0.58 ± 0.54
FD-MScM-EL vs. FD-COAL	0.47 ± 0.10	**2.83E-10**	0.61 ± 0.34	0.60 ± 0.08	**6.21E-05**	0.67 ± 0.15
FD-MScM-SH vs. MSISA-P	0.45 ± 0.09	**9.12E-18**	0.78 ± 0.30	0.57 ± 0.07	**6.90E-09**	0.71 ± 0.20
FD-MScM-SH vs. MSISA-R	0.45 ± 0.09	**3.09E-28**	0.95 ± 0.34	0.57 ± 0.07	**8.53E-19**	0.83 ± 0.36
FD-MScM-SH vs. MSKM-SH	0.45 ± 0.09	**0.93**	0.45 ± 0.07	0.57 ± 0.07	**8.20E-03**	0.58 ± 0.05
FD-MScM-SH vs. MSKM-EL	0.45 ± 0.09	**0.27**	0.46 ± 0.07	0.57 ± 0.07	**0.10**	0.57 ± 0.05
FD-MScM-SH vs. BMSKM-SH	0.45 ± 0.09	**0.01**	0.47 ± 0.07	0.57 ± 0.07	**1.90E-04**	0.54 ± 0.04
FD-MScM-SH vs. BMSKM-EL	0.45 ± 0.09	**8.76E-03**	0.48 ± 0.07	0.57 ± 0.07	**6.50E-05**	0.54 ± 0.04
FD-MScM-EL vs. MSISA-P	0.47 ± 0.10	**9.45E-17**	0.78 ± 0.30	0.60 ± 0.08	**3.93E-06**	0.71 ± 0.20
FD-MScM-EL vs. MSISA-R	0.47 ± 0.10	**7.56E-28**	0.95 ± 0.34	0.60 ± 0.08	**2.07E-15**	0.83 ± 0.36
FD-MScM-EL vs. MSKM-SH	0.47 ± 0.10	**5.58E-02**	0.45 ± 0.07	0.60 ± 0.08	**6.56E-03**	0.58 ± 0.05

FD-MScM-EL vs. MSKM-EL	0.47 ± 0.10	**0.55**	0.46 ± 0.07	0.60 ± 0.08	**3.84E-04**	0.57 ± 0.05
FD-MScM-EL vs. BMSKM-SH	0.47 ± 0.10	**0.36**	0.47 ± 0.07	0.60 ± 0.08	**1.91E-15**	0.54 ± 0.04
FD-MScM-EL vs. BMSKM-EL	0.47 ± 0.10	**0.34**	0.48 ± 0.07	0.60 ± 0.08	**5.04E-16**	0.54 ± 0.04

E. coli - V. cholerae pairing

	E. coli			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.44 ± 0.08	**2.78E-12**	0.50 ± 0.10	0.58 ± 0.08	**1.97E-21**	0.50 ± 0.29
FD-MScM-SH vs. QUBIC	0.44 ± 0.08	**4.46E-23**	0.29 ± 0.13	0.58 ± 0.08	**7.06E-33**	0.33 ± 0.19
FD-MScM-SH vs. FD-COAL	0.44 ± 0.08	**1.12E-16**	0.61 ± 0.34	0.58 ± 0.08	**3.86E-08**	0.73 ± 0.35
FD-MScM-EL vs. FD-SSCM	0.47 ± 0.09	**4.65E-04**	0.50 ± 0.10	0.60 ± 0.09	**5.10E-27**	0.50 ± 0.29
FD-MScM-EL vs. QUBIC	0.47 ± 0.09	**2.16E-27**	0.29 ± 0.13	0.60 ± 0.09	**1.01E-34**	0.33 ± 0.19
FD-MScM-EL vs. FD-COAL	0.47 ± 0.09	**9.88E-10**	0.61 ± 0.34	0.60 ± 0.09	**5.78E-05**	0.73 ± 0.35
FD-MScM-SH vs. MSISA-P	0.44 ± 0.08	**9.44E-11**	0.76 ± 0.30	0.58 ± 0.08	**3.30E-07**	0.75 ± 0.28

FD-MScM-SH vs. MSISA-R	0.44 ± 0.08	**1.31E-20**	0.96 ± 0.35	0.58 ± 0.08	**1.51E-18**	1.06 ± 0.19
FD-MScM-SH vs. MSKM-SH	0.44 ± 0.08	**8.13E-03**	0.45 ± 0.09	0.58 ± 0.08	**9.43E-12**	0.50 ± 0.09
FD-MScM-SH vs. MSKM-EL	0.44 ± 0.08	**1.97E-04**	0.47 ± 0.08	0.58 ± 0.08	**9.89E-12**	0.50 ± 0.08
FD-MScM-EL vs. MSISA-P	0.47 ± 0.09	**3.16E-09**	0.76 ± 0.30	0.60 ± 0.09	**2.59E-05**	0.75 ± 0.28
FD-MScM-EL vs. MSISA-R	0.47 ± 0.09	**5.45E-20**	0.96 ± 0.35	0.60 ± 0.09	**3.60E-18**	1.06 ± 0.19
FD-MScM-EL vs. MSKM-SH	0.47 ± 0.09	**0.52**	0.45 ± 0.09	0.60 ± 0.09	**5.67E-20**	0.50 ± 0.09
FD-MScM-EL vs. MSKM-EL	0.47 ± 0.09	**0.70**	0.47 ± 0.08	0.60 ± 0.09	**1.07E-19**	0.50 ± 0.08

S. typhimurium - V. cholerae pairing

	S. typhimurium			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.57 ± 0.06	**5.97E-24**	0.46 ± 0.09	0.60 ± 0.09	**2.59E-26**	0.50 ± 0.29
FD-MScM-SH vs. QUBIC	0.57 ± 0.06	**6.44E-07**	0.58 ± 0.54	0.60 ± 0.09	**6.33E-35**	0.33 ± 0.19

FD-MScM-SH vs. FD-COAL	0.57 ± 0.06	**3.27E-12**	0.67 ± 0.15	0.60 ± 0.09	**3.22E-05**	0.73 ± 0.35
FD-MScM-EL vs. FD-SSCM	0.60 ± 0.07	**1.77E-31**	0.46 ± 0.09	0.62 ± 0.08	**8.65E-30**	0.50 ± 0.29
FD-MScM-EL vs. QUBIC	0.60 ± 0.07	**2.74E-08**	0.58 ± 0.54	0.62 ± 0.08	**4.67E-36**	0.33 ± 0.19
FD-MScM-EL vs. FD-COAL	0.60 ± 0.07	**1.12E-05**	0.67 ± 0.15	0.62 ± 0.08	**8.74E-04**	0.73 ± 0.35
FD-MScM-SH vs. MSKM-SH	0.57 ± 0.06	**0.20**	0.56 ± 0.05	0.60 ± 0.09	**6.76E-30**	0.47 ± 0.07
FD-MScM-SH vs. MSKM-EL	0.57 ± 0.06	**1.89E-02**	0.55 ± 0.05	0.60 ± 0.09	**1.97E-23**	0.49 ± 0.07
FD-MScM-SH vs. BMSKM-SH	0.57 ± 0.06	**5.35E-11**	0.52 ± 0.04	0.60 ± 0.09	**7.93E-22**	0.50 ± 0.07
FD-MScM-SH vs. BMSKM-EL	0.57 ± 0.06	**6.33E-09**	0.53 ± 0.04	0.60 ± 0.09	**1.66E-19**	0.51 ± 0.06
FD-MScM-EL vs. MSKM-SH	0.60 ± 0.07	**2.55E-08**	0.56 ± 0.05	0.62 ± 0.08	**2.60E-34**	0.47 ± 0.07
FD-MScM-EL vs. MSKM-EL	0.60 ± 0.07	**1.56E-10**	0.55 ± 0.05	0.62 ± 0.08	**5.31E-29**	0.49 ± 0.07
FD-MScM-EL vs. BMSKM-SH	0.60 ± 0.07	**4.93E-23**	0.52 ± 0.04	0.62 ± 0.08	**3.92E-27**	0.50 ± 0.07
FD-MScM-EL vs. BMSKM-EL	0.60 ± 0.07	**1.62E-20**	0.53 ± 0.04	0.62 ± 0.08	**5.33E-25**	0.51 ± 0.06

7.2.3.1.2 Mean correlations

A comparison of the mean correlations of the results from MScM (full data) with all other relevant methods for all 3 pairings of a given triplet that's examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to the description for section 7.2.3.1.1 for instructions on how to interpret the table. In this case, these results illustrate that in 92 of the 92 comparisons (100%) for the Gram-positive triplet MScM step did as well or better than its competitors. Similarly, in 65 of the 92 comparisons (70.7%) for the Gram-negative triplet, MScM did as well or better than its competitors.

7.2.3.1.2.1 Gram-positive triplet

Table 7.27: Comparison of bicluster mean correlations from the full data methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*.

***B. subtilis* - *B. anthracis* pairing**

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.59 ± 0.11	**0.013**	0.56 ± 0.14	0.85 ± 0.09	**0.351**	0.82 ± 0.15
FD-MScM-SH vs. QUBIC	0.59 ± 0.11	**7.33E-28**	0.36 ± 0.19	0.85 ± 0.09	**1.85E-50**	0.49 ± 0.05

FD-MScM-SH vs. FD-COAL	0.59 ± 0.11	**0.446**	0.59 ± 0.15	0.85 ± 0.09	**4.87E-35**	0.62 ± 0.13
FD-MScM-EL vs. FD-SSCM	0.61 ± 0.11	**3.11E-04**	0.56 ± 0.14	0.84 ± 0.09	**0.760**	0.82 ± 0.15
FD-MScM-EL vs. QUBIC	0.61 ± 0.11	**5.14E-29**	0.36 ± 0.19	0.84 ± 0.09	**1.37E-50**	0.49 ± 0.05
FD-MScM-EL vs. FD-COAL	0.61 ± 0.11	**0.067**	0.59 ± 0.15	0.84 ± 0.09	**1.89E-33**	0.62 ± 0.13
FD-MScM-SH vs. MSISA-P	0.59 ± 0.11	**0.963**	0.60 ± 0.14	0.85 ± 0.09	**9.59E-22**	0.56 ± 0.07
FD-MScM-SH vs. MSISA-R	0.59 ± 0.11	**0.041**	0.55 ± 0.13	0.85 ± 0.09	**1.38E-22**	0.51 ± 0.03
FD-MScM-SH vs. MSKM-SH	0.59 ± 0.11	**0.340**	0.58 ± 0.11	0.85 ± 0.09	**1.49E-43**	0.52 ± 0.14
FD-MScM-SH vs. MSKM-EL	0.59 ± 0.11	**0.020**	0.56 ± 0.11	0.85 ± 0.09	**1.33E-40**	0.58 ± 0.15
FD-MScM-SH vs. BMSKM-SH	0.59 ± 0.11	**1.11E-11**	0.49 ± 0.13	0.85 ± 0.09	**5.18E-25**	0.72 ± 0.10
FD-MScM-SH vs. BMSKM-EL	0.59 ± 0.11	**1.17E-10**	0.50 ± 0.12	0.85 ± 0.09	**9.12E-26**	0.71 ± 0.10
FD-MScM-EL vs. MSISA-P	0.61 ± 0.11	**0.500**	0.60 ± 0.14	0.84 ± 0.09	**7.49E-22**	0.56 ± 0.07
FD-MScM-EL vs. MSISA-R	0.61 ± 0.11	**0.009**	0.55 ± 0.13	0.84 ± 0.09	**1.18E-22**	0.51 ± 0.03

FD-MScM-EL vs. MSKM-SH	0.61 ± 0.11	**0.035**	0.58 ± 0.11	0.84 ± 0.09	**4.07E-43**	0.52 ± 0.14
FD-MScM-EL vs. MSKM-EL	0.61 ± 0.11	**0.001**	0.56 ± 0.11	0.84 ± 0.09	**1.01E-39**	0.58 ± 0.15
FD-MScM-EL vs. BMSKM-SH	0.61 ± 0.11	**2.10E-14**	0.49 ± 0.13	0.84 ± 0.09	**7.87E-23**	0.72 ± 0.10
FD-MScM-EL vs. BMSKM-EL	0.61 ± 0.11	**3.51E-13**	0.50 ± 0.12	0.84 ± 0.09	**1.31E-23**	0.71 ± 0.10

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.59 ± 0.11	**0.011**	0.56 ± 0.14	0.80 ± 0.13	**2.31E-04**	0.71 ± 0.20
FD-MScM-SH vs. QUBIC	0.59 ± 0.11	**5.45E-27**	0.36 ± 0.19	0.80 ± 0.13	**2.92E-18**	0.45 ± 0.27
FD-MScM-SH vs. FD-COAL	0.59 ± 0.11	**0.481**	0.59 ± 0.15	0.80 ± 0.13	**0.999**	0.80 ± 0.12
FD-MScM-EL vs. FD-SSCM	0.61 ± 0.10	**2.58E-04**	0.56 ± 0.14	0.81 ± 0.11	**3.00E-05**	0.71 ± 0.20
FD-MScM-EL vs. QUBIC	0.61 ± 0.10	**1.36E-28**	0.36 ± 0.19	0.81 ± 0.11	**3.29E-18**	0.45 ± 0.27

FD-MScM-EL vs. FD-COAL	0.61 ± 0.10	**0.064**	0.59 ± 0.15	0.81 ± 0.11	**0.674**	0.80 ± 0.12
FD-MScM-SH vs. MSISA-P	0.59 ± 0.11	**0.639**	0.60 ± 0.20	0.80 ± 0.13	**5.34E-13**	0.47 ± 0.23
FD-MScM-SH vs. MSISA-R	0.59 ± 0.11	**0.036**	0.55 ± 0.12	0.80 ± 0.13	**4.18E-08**	0.50 ± 0.27
FD-MScM-SH vs. MSKM-SH	0.59 ± 0.11	**0.716**	0.59 ± 0.11	0.80 ± 0.13	**8.05E-33**	0.51 ± 0.17
FD-MScM-SH vs. MSKM-EL	0.59 ± 0.11	**0.021**	0.56 ± 0.11	0.80 ± 0.13	**4.99E-29**	0.55 ± 0.16
FD-MScM-SH vs. BMSKM-SH	0.59 ± 0.11	**6.43E-06**	0.52 ± 0.14	0.80 ± 0.13	**5.58E-20**	0.63 ± 0.15
FD-MScM-SH vs. BMSKM-EL	0.59 ± 0.11	**1.28E-05**	0.53 ± 0.12	0.80 ± 0.13	**9.89E-19**	0.64 ± 0.14
FD-MScM-EL vs. MSISA-P	0.61 ± 0.10	**0.440**	0.60 ± 0.20	0.81 ± 0.11	**1.28E-13**	0.47 ± 0.23
FD-MScM-EL vs. MSISA-R	0.61 ± 0.10	**0.004**	0.55 ± 0.12	0.81 ± 0.11	**1.11E-08**	0.50 ± 0.27
FD-MScM-EL vs. MSKM-SH	0.61 ± 0.10	**0.347**	0.59 ± 0.11	0.81 ± 0.11	**1.47E-34**	0.51 ± 0.17
FD-MScM-EL vs. MSKM-EL	0.61 ± 0.10	**2.30E-04**	0.56 ± 0.11	0.81 ± 0.11	**3.83E-31**	0.55 ± 0.16
FD-MScM-EL vs. BMSKM-SH	0.61 ± 0.10	**3.53E-08**	0.52 ± 0.14	0.81 ± 0.11	**3.86E-22**	0.63 ± 0.15

FD-MScM-EL vs. BMSKM-EL	0.61 ± 0.10	**4.23E-08**	0.53 ± 0.12	0.81 ± 0.11	**9.29E-21**	0.64 ± 0.14	

B. anthracis - L. monocytogenes pairing

		B. anthracis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	
FD-MScM-SH vs. FD-SSCM	0.82 ± 0.11	**0.227**	0.82 ± 0.15	0.77 ± 0.14	**0.011**	0.71 ± 0.20	
FD-MScM-SH vs. QUBIC	0.82 ± 0.11	**5.78E-49**	0.49 ± 0.05	0.77 ± 0.14	**1.03E-17**	0.45 ± 0.27	
FD-MScM-SH vs. FD-COAL	0.82 ± 0.11	**4.30E-29**	0.62 ± 0.13	0.77 ± 0.14	**0.344**	0.80 ± 0.12	
FD-MScM-EL vs. FD-SSCM	0.80 ± 0.11	**3.92E-04**	0.82 ± 0.15	0.78 ± 0.13	**0.003**	0.71 ± 0.20	
FD-MScM-EL vs. QUBIC	0.80 ± 0.11	**2.76E-48**	0.49 ± 0.05	0.78 ± 0.13	**1.22E-17**	0.45 ± 0.27	
FD-MScM-EL vs. FD-COAL	0.80 ± 0.11	**1.34E-24**	0.62 ± 0.13	0.78 ± 0.13	**0.532**	0.80 ± 0.12	
FD-MScM-SH vs. MSKM-SH	0.82 ± 0.11	**7.41E-21**	0.69 ± 0.12	0.77 ± 0.14	**6.55E-19**	0.60 ± 0.14	
FD-MScM-SH vs. MSKM-EL	0.82 ± 0.11	**6.40E-20**	0.70 ± 0.10	0.77 ± 0.14	**1.47E-16**	0.63 ± 0.13	

FD-MScM-EL vs. MSKM-SH	0.80 ± 0.11	**1.09E-14**	0.69 ± 0.12	0.78 ± 0.13	**9.32E-22**	0.60 ± 0.14
FD-MScM-EL vs. MSKM-EL	0.80 ± 0.11	**8.90E-14**	0.70 ± 0.10	0.78 ± 0.13	**4.40E-19**	0.63 ± 0.13

7.2.3.1.2.2 Gram-negative triplet

Table 7.28: Comparison of bicluster mean correlations from the full data methods considered by this study for all pairings of *E. coli, S. typhimurium* and *V. cholerae*.

E. coli - S. typhimurium pairing

	E. coli			***S. typhimurium***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.68 ± 0.12	**1.40E-07**	0.59 ± 0.17	0.55 ± 0.11	**1.55E-04**	0.58 ± 0.18
FD-MScM-SH vs. QUBIC	0.68 ± 0.12	**1.27E-39**	0.91 ± 0.08	0.55 ± 0.11	**3.48E-36**	0.86 ± 0.12
FD-MScM-SH vs. FD-COAL	0.68 ± 0.12	**1.39E-03**	0.63 ± 0.16	0.55 ± 0.11	**0.95**	0.57 ± 0.15
FD-MScM-EL vs. FD-SSCM	0.66 ± 0.13	**4.23E-04**	0.59 ± 0.17	0.50 ± 0.11	**1.04E-09**	0.58 ± 0.18
FD-MScM-EL vs. QUBIC	0.66 ± 0.13	**8.01E-41**	0.91 ± 0.08	0.50 ± 0.11	**1.34E-39**	0.86 ± 0.12
FD-MScM-EL vs. FD-COAL	0.66 ± 0.13	**0.08**	0.63 ± 0.16	0.50 ± 0.11	**8.47E-05**	0.57 ± 0.15

FD-MScM-SH vs. MSISA-P	0.68 ± 0.12	**4.60E-06**	0.56 ± 0.20	0.55 ± 0.11	**0.14**	0.60 ± 0.26
FD-MScM-SH vs. MSISA-R	0.68 ± 0.12	**2.03E-09**	0.52 ± 0.18	0.55 ± 0.11	**3.91E-07**	0.46 ± 0.23
FD-MScM-SH vs. MSKM-SH	0.68 ± 0.12	**1.16E-10**	0.59 ± 0.12	0.55 ± 0.11	**1.03E-43**	0.29 ± 0.08
FD-MScM-SH vs. MSKM-EL	0.68 ± 0.12	**2.18E-13**	0.57 ± 0.12	0.55 ± 0.11	**1.74E-41**	0.31 ± 0.08
FD-MScM-SH vs. BMSKM-SH	0.68 ± 0.12	**8.03E-19**	0.54 ± 0.12	0.55 ± 0.11	**7.05E-33**	0.37 ± 0.09
FD-MScM-SH vs. BMSKM-EL	0.68 ± 0.12	**8.43E-19**	0.54 ± 0.12	0.55 ± 0.11	**1.33E-30**	0.38 ± 0.09
FD-MScM-EL vs. MSISA-P	0.66 ± 0.13	**1.09E-04**	0.56 ± 0.20	0.50 ± 0.11	**0.02**	0.60 ± 0.26
FD-MScM-EL vs. MSISA-R	0.66 ± 0.13	**5.42E-08**	0.52 ± 0.18	0.50 ± 0.11	**2.15E-05**	0.46 ± 0.23
FD-MScM-EL vs. MSKM-SH	0.66 ± 0.13	**1.17E-06**	0.59 ± 0.12	0.50 ± 0.11	**5.88E-40**	0.29 ± 0.08
FD-MScM-EL vs. MSKM-EL	0.66 ± 0.13	**5.95E-09**	0.57 ± 0.12	0.50 ± 0.11	**1.65E-35**	0.31 ± 0.08
FD-MScM-EL vs. BMSKM-SH	0.66 ± 0.13	**4.49E-14**	0.54 ± 0.12	0.50 ± 0.11	**7.54E-22**	0.37 ± 0.09
FD-MScM-EL vs. BMSKM-EL	0.66 ± 0.13	**8.55E-14**	0.54 ± 0.12	0.50 ± 0.11	**1.22E-18**	0.38 ± 0.09

E. coli - V. cholerae pairing

dist1 vs. dist2	*E. coli*			*V. cholerae*		
	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
FD-MScM-SH vs. FD-SSCM	0.70 ± 0.11	**2.92E-11**	0.59 ± 0.17	0.55 ± 0.15	**4.43E-03**	0.60 ± 0.19
FD-MScM-SH vs. QUBIC	0.70 ± 0.11	**3.32E-39**	0.91 ± 0.08	0.55 ± 0.15	**5.44E-49**	0.92 ± 0.06
FD-MScM-SH vs. FD-COAL	0.70 ± 0.11	**9.05E-06**	0.63 ± 0.16	0.55 ± 0.15	**0.03**	0.59 ± 0.17
FD-MScM-EL vs. FD-SSCM	0.66 ± 0.12	**2.68E-04**	0.59 ± 0.17	0.50 ± 0.15	**2.00E-07**	0.60 ± 0.19
FD-MScM-EL vs. QUBIC	0.66 ± 0.12	**7.46E-42**	0.91 ± 0.08	0.50 ± 0.15	**1.38E-49**	0.92 ± 0.06
FD-MScM-EL vs. FD-COAL	0.66 ± 0.12	**0.07**	0.63 ± 0.16	0.50 ± 0.15	**1.45E-06**	0.59 ± 0.17
FD-MScM-SH vs. MSISA-P	0.70 ± 0.11	**5.18E-05**	0.56 ± 0.21	0.55 ± 0.15	**3.09E-06**	0.69 ± 0.19
FD-MScM-SH vs. MSISA-R	0.70 ± 0.11	**3.73E-09**	0.51 ± 0.18	0.55 ± 0.15	**1.64E-03**	0.48 ± 0.12
FD-MScM-SH vs. MSKM-SH	0.70 ± 0.11	**1.30E-17**	0.56 ± 0.15	0.55 ± 0.15	**2.02E-11**	0.43 ± 0.16
FD-MScM-SH vs. MSKM-EL	0.70 ± 0.11	**1.40E-19**	0.55 ± 0.14	0.55 ± 0.15	**3.86E-10**	0.44 ± 0.15

FD-MScM-EL vs. MSISA-P	0.66 ± 0.12	**1.68E-03**	0.56 ± 0.21	0.50 ± 0.15	**3.09E-08**	0.69 ± 0.19
FD-MScM-EL vs. MSISA-R	0.66 ± 0.12	**1.81E-07**	0.51 ± 0.18	0.50 ± 0.15	**0.46**	0.48 ± 0.12
FD-MScM-EL vs. MSKM-SH	0.66 ± 0.12	**2.27E-10**	0.56 ± 0.15	0.50 ± 0.15	**1.41E-05**	0.43 ± 0.16

S. typhimurium - V. cholerae pairing

	S. typhimurium			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.55 ± 0.11	**6.81E-05**	0.58 ± 0.18	0.51 ± 0.17	**1.30E-05**	0.60 ± 0.19
FD-MScM-SH vs. QUBIC	0.55 ± 0.11	**1.33E-36**	0.86 ± 0.12	0.51 ± 0.17	**1.55E-49**	0.92 ± 0.06
FD-MScM-SH vs. FD-COAL	0.55 ± 0.11	**0.90**	0.57 ± 0.15	0.51 ± 0.17	**6.29E-05**	0.59 ± 0.17
FD-MScM-EL vs. FD-SSCM	0.50 ± 0.11	**6.42E-10**	0.58 ± 0.18	0.48 ± 0.17	**8.12E-09**	0.60 ± 0.19
FD-MScM-EL vs. QUBIC	0.50 ± 0.11	**6.93E-40**	0.86 ± 0.12	0.48 ± 0.17	**7.54E-50**	0.92 ± 0.06
FD-MScM-EL vs. FD-COAL	0.50 ± 0.11	**1.43E-04**	0.57 ± 0.15	0.48 ± 0.17	**3.54E-08**	0.59 ± 0.17
FD-MScM-SH vs. MSKM-SH	0.55 ± 0.11	**2.42E-41**	0.31 ± 0.09	0.51 ± 0.17	**0.16**	0.49 ± 0.13

FD-MScM-SH vs. MSKM-EL	0.55 ± 0.11	**6.12E-35**	0.35 ± 0.10	0.51 ± 0.17	**6.38E-03**	0.47 ± 0.13
FD-MScM-SH vs. BMSKM-SH	0.55 ± 0.11	**3.48E-28**	0.39 ± 0.10	0.51 ± 0.17	**1.03E-05**	0.43 ± 0.13
FD-MScM-SH vs. BMSKM-EL	0.55 ± 0.11	**6.95E-25**	0.41 ± 0.09	0.51 ± 0.17	**3.96E-05**	0.44 ± 0.12
FD-MScM-EL vs. MSKM-SH	0.50 ± 0.11	**3.17E-34**	0.31 ± 0.09	0.48 ± 0.17	**0.66**	0.49 ± 0.13
FD-MScM-EL vs. MSKM-EL	0.50 ± 0.11	**5.78E-24**	0.35 ± 0.10	0.48 ± 0.17	**0.39**	0.47 ± 0.13
FD-MScM-EL vs. BMSKM-SH	0.50 ± 0.11	**1.87E-15**	0.39 ± 0.10	0.48 ± 0.17	**5.48E-03**	0.43 ± 0.13
FD-MScM-EL vs. BMSKM-EL	0.50 ± 0.11	**5.57E-12**	0.41 ± 0.09	0.48 ± 0.17	**0.02**	0.44 ± 0.12

7.2.3.1.3 Network Association p-values

A comparison of the association p-values (-log10) from MScM (full data) with all other relevant methods for all 3 pairings of a given triplet that's examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to 7.2.3.1.1 for instructions on how to interpret the table. For example, these results indicate that in 77 of the 92 comparisons (83.7%) for the Gram-positive triplet MScM

does as well or better than its competitors. Similarly, in all of the 92 comparisons (100%) for the Gram-negative triplet, MScM did as well or better than its competitors.

7.2.3.1.3.1 Gram-positive triplet

Table 7.29: Comparison of bicluster network association p-values from the full data methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*

B. subtilis - B. anthracis pairing

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	7.82 ± 8.76	**0.009**	9.78 ± 9.76	6.16 ± 7.23	**0.935**	5.44 ± 5.38
FD-MScM-SH vs. QUBIC	7.82 ± 8.76	**5.42E-36**	2.52 ± 4.78	6.16 ± 7.23	**0.658**	6.73 ± 7.52
FD-MScM-SH vs. FD-COAL	7.82 ± 8.76	**0.042**	7.57 ± 9.16	6.16 ± 7.23	**0.132**	6.50 ± 8.74
FD-MScM-EL vs. FD-SSCM	7.79 ± 9.05	**0.001**	9.78 ± 9.76	6.38 ± 7.19	**0.753**	5.44 ± 5.38
FD-MScM-EL vs. QUBIC	7.79 ± 9.05	**1.09E-35**	2.52 ± 4.78	6.38 ± 7.19	**0.853**	6.73 ± 7.52
FD-MScM-EL vs. FD-COAL	7.79 ± 9.05	**0.114**	7.57 ± 9.16	6.38 ± 7.19	**0.098**	6.50 ± 8.74
FD-MScM-SH vs. MSISA-P	7.82 ± 8.76	**0.300**	5.56 ± 5.86	6.16 ± 7.23	**0.834**	5.61 ± 7.17

FD-MScM-SH vs. MSISA-R	7.82 ± 8.76	**0.041**	9.69 ± 9.37	6.16 ± 7.23	**0.186**	9.66 ± 9.95
FD-MScM-SH vs. MSKM-SH	7.82 ± 8.76	**0.416**	7.87 ± 9.35	6.16 ± 7.23	**0.084**	4.38 ± 5.10
FD-MScM-SH vs. MSKM-EL	7.82 ± 8.76	**0.505**	8.15 ± 9.65	6.16 ± 7.23	**0.003**	4.06 ± 5.32
FD-MScM-SH vs. BMSKM-SH	7.82 ± 8.76	**0.527**	7.27 ± 8.25	6.16 ± 7.23	**0.591**	5.54 ± 6.48
FD-MScM-SH vs. BMSKM-EL	7.82 ± 8.76	**0.128**	6.93 ± 8.19	6.16 ± 7.23	**0.021**	4.56 ± 5.86
FD-MScM-EL vs. MSISA-P	7.79 ± 9.05	**0.539**	5.56 ± 5.86	6.38 ± 7.19	**0.987**	5.61 ± 7.17
FD-MScM-EL vs. MSISA-R	7.79 ± 9.05	**0.017**	9.69 ± 9.37	6.38 ± 7.19	**0.241**	9.66 ± 9.95
FD-MScM-EL vs. MSKM-SH	7.79 ± 9.05	**0.742**	7.87 ± 9.35	6.38 ± 7.19	**0.093**	4.38 ± 5.10
FD-MScM-EL vs. MSKM-EL	7.79 ± 9.05	**0.867**	8.15 ± 9.65	6.38 ± 7.19	**0.005**	4.06 ± 5.32
FD-MScM-EL vs. BMSKM-SH	7.79 ± 9.05	**0.871**	7.27 ± 8.25	6.38 ± 7.19	**0.700**	5.54 ± 6.48
FD-MScM-EL vs. BMSKM-EL	7.79 ± 9.05	**0.296**	6.93 ± 8.19	6.38 ± 7.19	**0.018**	4.56 ± 5.86

B. subtilis - L. monocytogenes pairing

	B. subtilis			*L. monocytogenes*		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	7.70 ± 8.79	**0.004**	9.78 ± 9.76	5.84 ± 7.26	**0.012**	6.90 ± 7.75
FD-MScM-SH vs. QUBIC	7.70 ± 8.79	**8.94E-30**	2.52 ± 4.78	5.84 ± 7.26	**1.73E-05**	9.95 ± 10.82
FD-MScM-SH vs. FD-COAL	7.70 ± 8.79	**0.165**	7.57 ± 9.16	5.84 ± 7.26	**0.146**	5.93 ± 8.27
FD-MScM-EL vs. FD-SSCM	7.68 ± 9.27	**1.72E-04**	9.78 ± 9.76	5.69 ± 6.92	**0.007**	6.90 ± 7.75
FD-MScM-EL vs. QUBIC	7.68 ± 9.27	**3.99E-27**	2.52 ± 4.78	5.69 ± 6.92	**3.27E-06**	9.95 ± 10.82
FD-MScM-EL vs. FD-COAL	7.68 ± 9.27	**0.636**	7.57 ± 9.16	5.69 ± 6.92	**0.138**	5.93 ± 8.27
FD-MScM-SH vs. MSISA-P	7.70 ± 8.79	**0.185**	9.05 ± 8.89	5.84 ± 7.26	**0.948**	3.70 ± 1.79
FD-MScM-SH vs. MSISA-R	7.70 ± 8.79	**0.025**	9.61 ± 9.29	5.84 ± 7.26	**0.257**	6.20 ± 6.65
FD-MScM-SH vs. MSKM-SH	7.70 ± 8.79	**0.088**	9.76 ± 10.54	5.84 ± 7.26	**0.165**	7.88 ± 9.56
FD-MScM-SH vs. MSKM-EL	7.70 ± 8.79	**0.364**	7.68 ± 9.47	5.84 ± 7.26	**0.152**	4.91 ± 6.44
FD-MScM-SH vs. BMSKM-SH	7.70 ± 8.79	**0.392**	9.23 ± 10.39	5.84 ± 7.26	**0.874**	7.10 ± 9.45

FD-MScM-SH vs. BMSKM-EL	7.70 ± 8.79	**0.025**	6.79 ± 8.75	5.84 ± 7.26	**0.179**	4.86 ± 6.39
FD-MScM-EL vs. MSISA-P	7.68 ± 9.27	**0.137**	9.05 ± 8.89	5.69 ± 6.92	**0.941**	3.70 ± 1.79
FD-MScM-EL vs. MSISA-R	7.68 ± 9.27	**0.005**	9.61 ± 9.29	5.69 ± 6.92	**0.218**	6.20 ± 6.65
FD-MScM-EL vs. MSKM-SH	7.68 ± 9.27	**0.018**	9.76 ± 10.54	5.69 ± 6.92	**0.152**	7.88 ± 9.56
FD-MScM-EL vs. MSKM-EL	7.68 ± 9.27	**0.878**	7.68 ± 9.47	5.69 ± 6.92	**0.138**	4.91 ± 6.44
FD-MScM-EL vs. BMSKM-SH	7.68 ± 9.27	**0.127**	9.23 ± 10.39	5.69 ± 6.92	**0.854**	7.10 ± 9.45
FD-MScM-EL vs. BMSKM-EL	7.68 ± 9.27	**0.146**	6.79 ± 8.75	5.69 ± 6.92	**0.147**	4.86 ± 6.39

B. anthracis - L. monocytogenes pairing

	B. anthracis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	6.80 ± 8.19	**0.944**	5.47 ± 5.39	5.35 ± 6.97	**6.99E-04**	6.90 ± 7.75
FD-MScM-SH vs. QUBIC	6.80 ± 8.19	**0.679**	6.73 ± 7.52	5.35 ± 6.97	**6.82E-08**	9.95 ± 10.82

FD-MScM-SH vs. FD-COAL	6.80 ± 8.19	**0.226**	6.50 ± 8.74	5.35 ± 6.97	**0.281**	5.93 ± 8.27
FD-MScM-EL vs. FD-SSCM	6.70 ± 7.99	**0.869**	5.47 ± 5.39	5.03 ± 6.91	**8.08E-06**	6.90 ± 7.75
FD-MScM-EL vs. QUBIC	6.70 ± 7.99	**0.838**	6.73 ± 7.52	5.03 ± 6.91	**2.94E-11**	9.95 ± 10.82
FD-MScM-EL vs. FD-COAL	6.70 ± 7.99	**0.095**	6.50 ± 8.74	5.03 ± 6.91	**0.567**	5.93 ± 8.27
FD-MScM-SH vs. MSKM-SH	6.80 ± 8.19	**0.930**	5.67 ± 7.00	5.35 ± 6.97	**0.418**	6.83 ± 8.86
FD-MScM-SH vs. MSKM-EL	6.80 ± 8.19	**0.049**	3.86 ± 4.13	5.35 ± 6.97	**0.334**	4.94 ± 6.73
FD-MScM-EL vs. MSKM-SH	6.70 ± 7.99	**0.778**	5.67 ± 7.00	5.03 ± 6.91	**0.186**	6.83 ± 8.86
FD-MScM-EL vs. MSKM-EL	6.70 ± 7.99	**0.011**	3.86 ± 4.13	5.03 ± 6.91	**0.687**	4.94 ± 6.73

7.2.3.1.3.2 Gram-negative triplet

Table 7.30: Comparison of bicluster network association p-values from the full data methods considered by this study for all pairings of *E. coli, S. typhimurium* and *V. cholerae*.

E. coli - S. typhimurium pairing

	E. coli			***S. typhimurium***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**

FD-MScM-SH vs. FD-SSCM	17.22 ± 10.39	**0.45**	17.95 ± 10.94	16.55 ± 10.70	**1.21E-38**	6.57 ± 9.02
FD-MScM-SH vs. QUBIC	17.22 ± 10.39	**0.07**	15.63 ± 10.47	16.55 ± 10.70	**9.77E-05**	9.12 ± 11.64
FD-MScM-SH vs. FD-COAL	17.22 ± 10.39	**2.43E-27**	10.91 ± 10.78	16.55 ± 10.70	**1.83E-77**	3.85 ± 7.07
FD-MScM-EL vs. FD-SSCM	23.06 ± 9.27	**4.88E-22**	17.95 ± 10.94	20.69 ± 10.28	**2.29E-60**	6.57 ± 9.02
FD-MScM-EL vs. QUBIC	23.06 ± 9.27	**2.77E-19**	15.63 ± 10.47	20.69 ± 10.28	**8.25E-07**	9.12 ± 11.64
FD-MScM-EL vs. FD-COAL	23.06 ± 9.27	**4.34E-82**	10.91 ± 10.78	20.69 ± 10.28	**8.96E-105**	3.85 ± 7.07
FD-MScM-SH vs. MSISA-P	17.22 ± 10.39	**2.89E-06**	9.62 ± 7.69	16.55 ± 10.70	**7.25E-06**	8.54 ± 7.50
FD-MScM-SH vs. MSISA-R	17.22 ± 10.39	**5.82E-17**	10.12 ± 10.22	16.55 ± 10.70	**6.61E-56**	3.79 ± 5.04
FD-MScM-SH vs. MSKM-SH	17.22 ± 10.39	**6.55E-42**	7.37 ± 8.70	16.55 ± 10.70	**8.63E-36**	6.81 ± 8.49
FD-MScM-SH vs. MSKM-EL	17.22 ± 10.39	**5.26E-42**	7.78 ± 8.98	16.55 ± 10.70	**7.24E-45**	6.05 ± 8.05
FD-MScM-SH vs. BMSKM-SH	17.22 ± 10.39	**9.58E-50**	6.02 ± 8.14	16.55 ± 10.70	**5.71E-45**	5.34 ± 7.28
FD-MScM-SH vs. BMSKM-EL	17.22 ± 10.39	**6.98E-52**	6.37 ± 8.28	16.55 ± 10.70	**2.14E-52**	4.98 ± 7.05

FD-MScM-EL vs. MSISA-P	23.06 ± 9.27	**7.86E-17**	9.62 ± 7.69	20.69 ± 10.28	**1.04E-11**	8.54 ± 7.50
FD-MScM-EL vs. MSISA-R	23.06 ± 9.27	**1.13E-43**	10.12 ± 10.22	20.69 ± 10.28	**3.89E-78**	3.79 ± 5.04
FD-MScM-EL vs. MSKM-SH	23.06 ± 9.27	**8.25E-84**	7.37 ± 8.70	20.69 ± 10.28	**8.27E-59**	6.81 ± 8.49
FD-MScM-EL vs. MSKM-EL	23.06 ± 9.27	**9.64E-88**	7.78 ± 8.98	20.69 ± 10.28	**1.59E-70**	6.05 ± 8.05
FD-MScM-EL vs. BMSKM-SH	23.06 ± 9.27	**7.35E-87**	6.02 ± 8.14	20.69 ± 10.28	**2.04E-67**	5.34 ± 7.28
FD-MScM-EL vs. BMSKM-EL	23.06 ± 9.27	**2.32E-94**	6.37 ± 8.28	20.69 ± 10.28	**3.75E-77**	4.98 ± 7.05

E. coli - V. cholerae pairing

	E. coli			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	18.38 ± 10.17	**0.43**	17.95 ± 10.94	17.92 ± 10.12	**8.14E-35**	9.66 ± 10.28
FD-MScM-SH vs. QUBIC	18.38 ± 10.17	**2.07E-03**	15.63 ± 10.47	17.92 ± 10.12	**0.43**	16.69 ± 12.25
FD-MScM-SH vs. FD-COAL	18.38 ± 10.17	**1.72E-33**	10.91 ± 10.78	17.92 ± 10.12	**7.07E-64**	6.92 ± 9.00
FD-MScM-EL vs. FD-SSCM	22.53 ± 9.47	**1.97E-17**	17.95 ± 10.94	21.35 ± 9.62	**5.99E-61**	9.66 ± 10.28

FD-MScM-EL vs. QUBIC	22.53 ± 9.47	**2.30E-16**	15.63 ± 10.47	21.35 ± 9.62	**1.22E-02**	16.69 ± 12.25
FD-MScM-EL vs. FD-COAL	22.53 ± 9.47	**4.42E-76**	10.91 ± 10.78	21.35 ± 9.62	**9.71E-95**	6.92 ± 9.00
FD-MScM-SH vs. MSISA-P	18.38 ± 10.17	**3.25E-05**	11.12 ± 10.80	17.92 ± 10.12	**1.10E-03**	12.19 ± 12.04
FD-MScM-SH vs. MSISA-R	18.38 ± 10.17	**1.16E-12**	11.22 ± 10.86	17.92 ± 10.12	**4.15E-04**	14.88 ± 10.17
FD-MScM-SH vs. MSKM-SH	18.38 ± 10.17	**3.95E-31**	8.92 ± 9.93	17.92 ± 10.12	**9.51E-33**	8.04 ± 10.11
FD-MScM-SH vs. MSKM-EL	18.38 ± 10.17	**8.90E-47**	7.99 ± 9.13	17.92 ± 10.12	**3.94E-43**	7.33 ± 9.49
FD-MScM-EL vs. MSISA-P	22.53 ± 9.47	**1.09E-09**	11.12 ± 10.80	21.35 ± 9.62	**3.67E-06**	12.19 ± 12.04
FD-MScM-EL vs. MSISA-R	22.53 ± 9.47	**2.79E-26**	11.22 ± 10.86	21.35 ± 9.62	**3.56E-13**	14.88 ± 10.17
FD-MScM-EL vs. MSKM-SH	22.53 ± 9.47	**6.07E-56**	8.92 ± 9.93	21.35 ± 9.62	**9.87E-48**	8.04 ± 10.11
FD-MScM-EL vs. MSKM-EL	22.53 ± 9.47	**2.74E-82**	7.99 ± 9.13	21.35 ± 9.62	**4.41E-62**	7.33 ± 9.49

S. typhimurium - V. cholerae pairing

	S. typhimurium			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**

FD-MScM-SH vs. FD-SSCM	15.10 ± 10.15	**1.12E-33**	6.57 ± 9.02	15.61 ± 10.24	**2.47E-22**	9.66 ± 10.28
FD-MScM-SH vs. QUBIC	15.10 ± 10.15	**2.03E-04**	9.12 ± 11.64	15.61 ± 10.24	**0.75**	16.69 ± 12.25
FD-MScM-SH vs. FD-COAL	15.10 ± 10.15	**4.45E-71**	3.85 ± 7.07	15.61 ± 10.24	**4.51E-48**	6.92 ± 9.00
FD-MScM-EL vs. FD-SSCM	19.84 ± 10.08	**1.37E-55**	6.57 ± 9.02	21.63 ± 9.52	**6.04E-62**	9.66 ± 10.28
FD-MScM-EL vs. QUBIC	19.84 ± 10.08	**2.40E-06**	9.12 ± 11.64	21.63 ± 9.52	**1.10E-02**	16.69 ± 12.25
FD-MScM-EL vs. FD-COAL	19.84 ± 10.08	**1.10E-98**	3.85 ± 7.07	21.63 ± 9.52	**9.02E-96**	6.92 ± 9.00
FD-MScM-SH vs. MSKM-SH	15.10 ± 10.15	**1.40E-26**	6.75 ± 9.30	15.61 ± 10.24	**7.59E-33**	6.43 ± 9.17
FD-MScM-SH vs. MSKM-EL	15.10 ± 10.15	**5.31E-49**	4.57 ± 7.03	15.61 ± 10.24	**2.29E-36**	6.69 ± 9.35
FD-MScM-SH vs. BMSKM-SH	15.10 ± 10.15	**5.74E-30**	5.51 ± 8.19	15.61 ± 10.24	**2.82E-36**	5.57 ± 8.59
FD-MScM-SH vs. BMSKM-EL	15.10 ± 10.15	**4.18E-55**	3.76 ± 6.26	15.61 ± 10.24	**1.57E-41**	6.02 ± 8.86
FD-MScM-EL vs. MSKM-SH	19.84 ± 10.08	**1.69E-41**	6.75 ± 9.30	21.63 ± 9.52	**3.35E-56**	6.43 ± 9.17
FD-MScM-EL vs. MSKM-EL	19.84 ± 10.08	**2.69E-71**	4.57 ± 7.03	21.63 ± 9.52	**1.79E-64**	6.69 ± 9.35

FD-MScM-EL vs. BMSKM-SH	19.84 ± 10.08	**3.54E-43**	5.51 ± 8.19	21.63 ± 9.52	**1.22E-57**	5.57 ± 8.59
FD-MScM-EL vs. BMSKM-EL	19.84 ± 10.08	**5.07E-75**	3.76 ± 6.26	21.63 ± 9.52	**3.21E-70**	6.02 ± 8.86

7.2.3.1.4 Motif E-values

A comparison of the motif E-values (-log10) from MScM (full data) with all other relevant methods for all 3 pairings of a given triplet that's examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to 7.2.3.1.1 for instructions on how to interpret the table. For example, these results indicate that in 69 of the 92 of the comparisons for the Gram-positive triplet (75%) MScM does as well or better than its competitors. Similarly, in 85 of the 92 comparisons (92.4%) for the Gram-negative triplet, MScM did as well or better than its competitors.

7.2.3.1.4.1 Gram-positive triplet

Table 7.31: Comparison of bicluster motif E-values(-log10) from the full data methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*.

B. subtilis - B. anthracis pairing

	B. subtilis	***B. anthracis***

dist1 vs. dist2	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
FD-MScM-SH vs. FD-SSCM	2.17 ± 10.07	**2.91E-10**	7.03 ± 18.81	2.46 ± 8.08	**0.043**	4.98 ± 10.63
FD-MScM-SH vs. QUBIC	2.17 ± 10.07	**0.033**	1.41 ± 3.94	2.46 ± 8.08	**1.43E-07**	13.72 ± 14.49
FD-MScM-SH vs. FD-COAL	2.17 ± 10.07	**0.172**	2.72 ± 7.30	2.46 ± 8.08	**0.618**	3.85 ± 8.83
FD-MScM-EL vs. FD-SSCM	3.34 ± 8.22	**4.28E-04**	7.03 ± 18.81	3.74 ± 10.70	**0.427**	4.98 ± 10.63
FD-MScM-EL vs. QUBIC	3.34 ± 8.22	**0.802**	1.41 ± 3.94	3.74 ± 10.70	**5.09E-07**	13.72 ± 14.49
FD-MScM-EL vs. FD-COAL	3.34 ± 8.22	**0.464**	2.72 ± 7.30	3.74 ± 10.70	**0.685**	3.85 ± 8.83
FD-MScM-SH vs. MSISA-P	2.17 ± 10.07	**0.002**	-1.12 ± 2.03	2.46 ± 8.08	**0.226**	0.46 ± 3.43
FD-MScM-SH vs. MSISA-R	2.17 ± 10.07	**1.12E-08**	9.40 ± 9.19	2.46 ± 8.08	**4.99E-06**	2.34 ± 11.56
FD-MScM-SH vs. MSKM-SH	2.17 ± 10.07	**1.75E-07**	-1.18 ± 2.62	2.46 ± 8.08	**0.001**	-0.22 ± 2.96
FD-MScM-SH vs. MSKM-EL	2.17 ± 10.07	**0.045**	0.19 ± 4.26	2.46 ± 8.08	**0.037**	2.74 ± 5.66
FD-MScM-SH vs. BMSKM-SH	2.17 ± 10.07	**1.02E-06**	-1.09 ± 2.68	2.46 ± 8.08	**1.87E-04**	-0.39 ± 2.87

FD-MScM-SH vs. BMSKM-EL	2.17 ± 10.07	**0.378**	0.44 ± 4.06	2.46 ± 8.08	**0.002**	3.07 ± 5.44
FD-MScM-EL vs. MSISA-P	3.34 ± 8.22	**5.26E-05**	-1.12 ± 2.03	3.74 ± 10.70	**0.059**	0.46 ± 3.43
FD-MScM-EL vs. MSISA-R	3.34 ± 8.22	**7.22E-06**	9.40 ± 9.19	3.74 ± 10.70	**5.97E-06**	2.34 ± 11.56
FD-MScM-EL vs. MSKM-SH	3.34 ± 8.22	**2.91E-11**	-1.18 ± 2.62	3.74 ± 10.70	**2.12E-05**	-0.22 ± 2.96
FD-MScM-EL vs. MSKM-EL	3.34 ± 8.22	**1.88E-04**	0.19 ± 4.26	3.74 ± 10.70	**0.495**	2.74 ± 5.66
FD-MScM-EL vs. BMSKM-SH	3.34 ± 8.22	**2.18E-10**	-1.09 ± 2.68	3.74 ± 10.70	**3.64E-06**	-0.39 ± 2.87
FD-MScM-EL vs. BMSKM-EL	3.34 ± 8.22	**0.004**	0.44 ± 4.06	3.74 ± 10.70	**0.101**	3.07 ± 5.44

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.43 ± 5.96	**7.22E-18**	7.03 ± 18.81	1.65 ± 5.96	**3.46E-07**	0.46 ± 4.73
FD-MScM-SH vs. QUBIC	0.43 ± 5.96	**6.64E-06**	1.41 ± 3.94	1.65 ± 5.96	**0.352**	9.57 ± 13.54

FD-MScM-SH vs. FD-COAL	0.43 ± 5.96	**0.001**	2.72 ± 7.30	1.65 ± 5.96	**0.325**	3.66 ± 7.60
FD-MScM-EL vs. FD-SSCM	2.12 ± 7.86	**3.86E-08**	7.03 ± 18.81	3.32 ± 9.98	**1.17E-10**	0.46 ± 4.73
FD-MScM-EL vs. QUBIC	2.12 ± 7.86	**0.127**	1.41 ± 3.94	3.32 ± 9.98	**0.573**	9.57 ± 13.54
FD-MScM-EL vs. FD-COAL	2.12 ± 7.86	**0.370**	2.72 ± 7.30	3.32 ± 9.98	**0.917**	3.66 ± 7.60
FD-MScM-SH vs. MSISA-P	0.43 ± 5.96	**1.84E-07**	-2.63 ± 1.00	1.65 ± 5.96	**1.29E-06**	-1.56 ± 1.40
FD-MScM-SH vs. MSISA-R	0.43 ± 5.96	**5.52E-13**	10.37 ± 8.84	1.65 ± 5.96	**1.12E-07**	9.06 ± 7.74
FD-MScM-SH vs. MSKM-SH	0.43 ± 5.96	**5.11E-07**	-1.91 ± 1.56	1.65 ± 5.96	**4.50E-07**	-0.83 ± 1.64
FD-MScM-SH vs. MSKM-EL	0.43 ± 5.96	**0.985**	0.52 ± 5.33	1.65 ± 5.96	**0.169**	0.36 ± 2.68
FD-MScM-SH vs. BMSKM-SH	0.43 ± 5.96	**5.07E-08**	-1.97 ± 1.58	1.65 ± 5.96	**5.68E-08**	-0.89 ± 1.69
FD-MScM-SH vs. BMSKM-EL	0.43 ± 5.96	**0.992**	0.02 ± 3.81	1.65 ± 5.96	**0.177**	0.43 ± 3.17
FD-MScM-EL vs. MSISA-P	2.12 ± 7.86	**7.84E-10**	-2.63 ± 1.00	3.32 ± 9.98	**1.56E-08**	-1.56 ± 1.40
FD-MScM-EL vs. MSISA-R	2.12 ± 7.86	**6.40E-10**	10.37 ± 8.84	3.32 ± 9.98	**2.25E-06**	9.06 ± 7.74

FD-MScM-EL vs. MSKM-SH	2.12 ± 7.86	**2.41E-13**	-1.91 ± 1.56	3.32 ± 9.98	**1.69E-11**	-0.83 ± 1.64
FD-MScM-EL vs. MSKM-EL	2.12 ± 7.86	**0.013**	0.52 ± 5.33	3.32 ± 9.98	**0.002**	0.36 ± 2.68
FD-MScM-EL vs. BMSKM-SH	2.12 ± 7.86	**2.88E-14**	-1.97 ± 1.58	3.32 ± 9.98	**2.44E-12**	-0.89 ± 1.69
FD-MScM-EL vs. BMSKM-EL	2.12 ± 7.86	**0.007**	0.02 ± 3.81	3.32 ± 9.98	**0.002**	0.43 ± 3.17

B. anthracis - L. monocytogenes pairing

	B. anthracis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	0.02 ± 3.66	**3.12E-08**	5.21 ± 10.80	1.69 ± 5.21	**1.86E-06**	0.46 ± 4.73
FD-MScM-SH vs. QUBIC	0.02 ± 3.66	**5.33E-10**	13.72 ± 14.49	1.69 ± 5.21	**0.326**	9.57 ± 13.54
FD-MScM-SH vs. FD-COAL	0.02 ± 3.66	**0.004**	3.85 ± 8.83	1.69 ± 5.21	**0.337**	3.66 ± 7.60
FD-MScM-EL vs. FD-SSCM	1.68 ± 6.38	**0.001**	5.21 ± 10.80	3.17 ± 8.21	**3.92E-10**	0.46 ± 4.73
FD-MScM-EL vs. QUBIC	1.68 ± 6.38	**4.06E-08**	13.72 ± 14.49	3.17 ± 8.21	**0.686**	9.57 ± 13.54

FD-MScM-EL vs. FD-COAL	1.68 ± 6.38	**0.213**	3.85 ± 8.83	3.17 ± 8.21	**0.999**	3.66 ± 7.60
FD-MScM-SH vs. MSKM-SH	0.02 ± 3.66	**3.71E-05**	-1.58 ± 1.60	1.69 ± 5.21	**5.77E-06**	-0.67 ± 1.69
FD-MScM-SH vs. MSKM-EL	0.02 ± 3.66	**1.47E-04**	2.52 ± 6.60	1.69 ± 5.21	**0.118**	0.44 ± 3.00
FD-MScM-EL vs. MSKM-SH	1.68 ± 6.38	**2.02E-09**	-1.58 ± 1.60	3.17 ± 8.21	**3.65E-09**	-0.67 ± 1.69
FD-MScM-EL vs. MSKM-EL	1.68 ± 6.38	**0.099**	2.52 ± 6.60	3.17 ± 8.21	**0.004**	0.44 ± 3.00

7.2.3.1.4.2 Gram-negative triplet

Table 7.32: Comparison of bicluster motif E-values from the full data methods considered by this study for all pairings of *E. coli, S. typhimurium* and *V. cholerae*.

E. coli - S. typhimurium pairing

	E. coli			***S. typhimurium***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	5.03 ± 34.70	**5.06E-09**	3.37 ± 25.14	7.25 ± 38.52	**0.90**	0.38 ± 3.97
FD-MScM-SH vs. QUBIC	5.03 ± 34.70	**2.31E-19**	-3.15 ± 1.19	7.25 ± 38.52	**5.55E-15**	-2.53 ± 1.10
FD-MScM-SH vs. FD-COAL	5.03 ± 34.70	**0.03**	1.00 ± 6.43	7.25 ± 38.52	**0.12**	1.51 ± 5.76

FD-MScM-EL vs. FD-SSCM	6.82 ± 40.87	**0.51**	3.37 ± 25.14	9.91 ± 43.33	**5.71E-04**	0.38 ± 3.97
FD-MScM-EL vs. QUBIC	6.82 ± 40.87	**1.39E-29**	-3.15 ± 1.19	9.91 ± 43.33	**3.56E-21**	-2.53 ± 1.10
FD-MScM-EL vs. FD-COAL	6.82 ± 40.87	**0.12**	1.00 ± 6.43	9.91 ± 43.33	**0.04**	1.51 ± 5.76
FD-MScM-SH vs. MSISA-P	5.03 ± 34.70	**3.21E-05**	-2.45 ± 1.71	7.25 ± 38.52	**7.87E-10**	-2.39 ± 1.66
FD-MScM-SH vs. MSISA-R	5.03 ± 34.70	**6.58E-05**	2.35 ± 6.98	7.25 ± 38.52	**5.16E-04**	0.44 ± 7.23
FD-MScM-SH vs. MSKM-SH	5.03 ± 34.70	**0.58**	-1.35 ± 3.16	7.25 ± 38.52	**0.67**	-0.27 ± 3.92
FD-MScM-SH vs. MSKM-EL	5.03 ± 34.70	**2.24E-03**	0.36 ± 4.76	7.25 ± 38.52	**0.52**	0.20 ± 4.25
FD-MScM-SH vs. BMSKM-SH	5.03 ± 34.70	**0.77**	-1.36 ± 2.31	7.25 ± 38.52	**0.47**	-0.56 ± 2.60
FD-MScM-SH vs. BMSKM-EL	5.03 ± 34.70	**3.84E-03**	-0.17 ± 3.41	7.25 ± 38.52	**0.62**	-0.13 ± 2.88
FD-MScM-EL vs. MSISA-P	6.82 ± 40.87	**1.07E-11**	-2.45 ± 1.71	9.91 ± 43.33	**8.48E-14**	-2.39 ± 1.66
FD-MScM-EL vs. MSISA-R	6.82 ± 40.87	**0.21**	2.35 ± 6.98	9.91 ± 43.33	**8.03E-06**	0.44 ± 7.23
FD-MScM-EL vs. MSKM-SH	6.82 ± 40.87	**3.75E-07**	-1.35 ± 3.16	9.91 ± 43.33	**2.58E-06**	-0.27 ± 3.92

FD-MScM-EL vs. MSKM-EL	6.82 ± 40.87	**0.33**	0.36 ± 4.76	9.91 ± 43.33	**2.74E-04**	0.20 ± 4.25
FD-MScM-EL vs. BMSKM-SH	6.82 ± 40.87	**9.47E-06**	-1.36 ± 2.31	9.91 ± 43.33	**7.20E-07**	-0.56 ± 2.60
FD-MScM-EL vs. BMSKM-EL	6.82 ± 40.87	**0.19**	-0.17 ± 3.41	9.91 ± 43.33	**8.29E-05**	-0.13 ± 2.88

E. coli - V. cholerae pairing

	E. coli			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	-0.92 ± 3.47	**9.55E-08**	3.37 ± 25.14	-0.18 ± 8.74	**2.66E-12**	8.65 ± 31.60
FD-MScM-SH vs. QUBIC	-0.92 ± 3.47	**4.42E-23**	-3.15 ± 1.19	-0.18 ± 8.74	**3.63E-11**	-1.97 ± 3.33
FD-MScM-SH vs. FD-COAL	-0.92 ± 3.47	**0.07**	1.00 ± 6.43	-0.18 ± 8.74	**8.16E-05**	11.28 ± 35.97
FD-MScM-EL vs. FD-SSCM	1.42 ± 5.48	**0.15**	3.37 ± 25.14	6.09 ± 31.78	**0.04**	8.65 ± 31.60
FD-MScM-EL vs. QUBIC	1.42 ± 5.48	**7.77E-33**	-3.15 ± 1.19	6.09 ± 31.78	**1.34E-19**	-1.97 ± 3.33
FD-MScM-EL vs. FD-COAL	1.42 ± 5.48	**3.18E-03**	1.00 ± 6.43	6.09 ± 31.78	**0.87**	11.28 ± 35.97
FD-MScM-SH vs. MSISA-P	-0.92 ± 3.47	**2.04E-06**	-2.64 ± 1.48	-0.18 ± 8.74	**1.27E-05**	-2.57 ± 1.63

FD-MScM-SH vs. MSISA-R	-0.92 ± 3.47	**5.25E-03**	2.63 ± 7.50	-0.18 ± 8.74	**3.36E-13**	-0.65 ± 11.66
FD-MScM-SH vs. MSKM-SH	-0.92 ± 3.47	**5.98E-03**	-1.94 ± 1.74	-0.18 ± 8.74	**9.48E-03**	-1.87 ± 2.01
FD-MScM-SH vs. MSKM-EL	-0.92 ± 3.47	**0.18**	-0.01 ± 4.79	-0.18 ± 8.74	**0.43**	2.71 ± 29.53
FD-MScM-EL vs. MSISA-P	1.42 ± 5.48	**7.03E-11**	-2.64 ± 1.48	6.09 ± 31.78	**3.40E-10**	-2.57 ± 1.63
FD-MScM-EL vs. MSISA-R	1.42 ± 5.48	**0.92**	2.63 ± 7.50	6.09 ± 31.78	**5.80E-14**	-0.65 ± 11.66
FD-MScM-EL vs. MSKM-SH	1.42 ± 5.48	**4.37E-15**	-1.94 ± 1.74	6.09 ± 31.78	**7.35E-12**	-1.87 ± 2.01
FD-MScM-EL vs. MSKM-EL	1.42 ± 5.48	**3.82E-04**	-0.01 ± 4.79	6.09 ± 31.78	**2.42E-04**	2.71 ± 29.53

S. typhimurium - V. cholerae pairing

	S. typhimurium			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	1.97 ± 13.19	**0.35**	0.38 ± 3.97	-0.63 ± 9.18	**2.09E-15**	8.65 ± 31.60
FD-MScM-SH vs. QUBIC	1.97 ± 13.19	**1.44E-19**	-2.53 ± 1.10	-0.63 ± 9.18	**9.85E-09**	-1.97 ± 3.33
FD-MScM-SH vs. FD-COAL	1.97 ± 13.19	**0.28**	1.51 ± 5.76	-0.63 ± 9.18	**1.74E-06**	11.28 ± 35.97

FD-MScM-EL vs. FD-SSCM	10.63 ± 33.85	**1.97E-06**	0.38 ± 3.97	6.57 ± 32.51	**1.37E-03**	8.65 ± 31.60
FD-MScM-EL vs. QUBIC	10.63 ± 33.85	**1.39E-26**	-2.53 ± 1.10	6.57 ± 32.51	**1.15E-15**	-1.97 ± 3.33
FD-MScM-EL vs. FD-COAL	10.63 ± 33.85	**9.79E-04**	1.51 ± 5.76	6.57 ± 32.51	**0.27**	11.28 ± 35.97
FD-MScM-SH vs. MSKM-SH	1.97 ± 13.19	**4.57E-06**	-1.46 ± 1.83	-0.63 ± 9.18	**0.04**	-1.94 ± 2.27
FD-MScM-SH vs. MSKM-EL	1.97 ± 13.19	**0.53**	-0.24 ± 3.27	-0.63 ± 9.18	**0.70**	0.38 ± 12.80
FD-MScM-SH vs. BMSKM-SH	1.97 ± 13.19	**1.81E-06**	-1.52 ± 1.75	-0.63 ± 9.18	**0.06**	-2.03 ± 1.65
FD-MScM-SH vs. BMSKM-EL	1.97 ± 13.19	**0.28**	-0.21 ± 3.35	-0.63 ± 9.18	**0.27**	0.35 ± 15.26
FD-MScM-EL vs. MSKM-SH	10.63 ± 33.85	**1.88E-17**	-1.46 ± 1.83	6.57 ± 32.51	**3.05E-09**	-1.94 ± 2.27
FD-MScM-EL vs. MSKM-EL	10.63 ± 33.85	**1.23E-08**	-0.24 ± 3.27	6.57 ± 32.51	**4.65E-04**	0.38 ± 12.80
FD-MScM-EL vs. BMSKM-SH	10.63 ± 33.85	**3.78E-18**	-1.52 ± 1.75	6.57 ± 32.51	**7.13E-09**	-2.03 ± 1.65
FD-MScM-EL vs. BMSKM-EL	10.63 ± 33.85	**1.35E-08**	-0.21 ± 3.35	6.57 ± 32.51	**1.99E-03**	0.35 ± 15.26

7.2.3.1.5 Sequence p-values

A comparison of the sequence p-values (-log10) from MScM (full data) with all other relevant methods for all 3 pairings of the three organisms examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to the description for section 7.2.3.1.1 for instructions on how to interpret the table. As the table indicates, in 72 of the 92 of the comparisons (78.3%) MScM does as well or better than its competitors. In contrast, in only 38 of the 92 comparisons (41.3%) for the Gram-negative triplet, MScM did as well or better than its competitors.

7.2.3.1.5.1 Gram-positive triplet

Table 7.33: Comparison of bicluster sequence p-values (-log10) from the full data methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*.

***B. subtilis* - *B. anthracis* pairing**

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	3.86 ± 1.39	**2.84E-22**	6.73 ± 3.35	3.49 ± 1.31	**0.303**	3.90 ± 2.62
FD-MScM-SH vs. QUBIC	3.86 ± 1.39	**4.41E-33**	2.06 ± 0.50	3.49 ± 1.31	**1.78E-32**	1.77 ± 0.26
FD-MScM-SH vs. FD-COAL	3.86 ± 1.39	**7.40E-25**	2.47 ± 1.12	3.49 ± 1.31	**1.71E-18**	2.32 ± 1.57

FD-MScM-EL vs. FD-SSCM	3.47 ± 1.31	**3.39E-29**	6.73 ± 3.35	3.24 ± 1.22	**0.721**	3.90 ± 2.62
FD-MScM-EL vs. QUBIC	3.47 ± 1.31	**1.83E-25**	2.06 ± 0.50	3.24 ± 1.22	**5.27E-31**	1.77 ± 0.26
FD-MScM-EL vs. FD-COAL	3.47 ± 1.31	**1.64E-16**	2.47 ± 1.12	3.24 ± 1.22	**1.08E-15**	2.32 ± 1.57
FD-MScM-SH vs. MSISA-P	3.86 ± 1.39	**0.164**	3.65 ± 1.74	3.49 ± 1.31	**0.345**	3.34 ± 1.33
FD-MScM-SH vs. MSISA-R	3.86 ± 1.39	**5.63E-15**	2.02 ± 0.52	3.49 ± 1.31	**5.99E-07**	1.79 ± 0.27
FD-MScM-SH vs. MSKM-SH	3.86 ± 1.39	**0.879**	3.97 ± 1.81	3.49 ± 1.31	**0.985**	3.59 ± 1.53
FD-MScM-SH vs. MSKM-EL	3.86 ± 1.39	**8.16E-06**	3.24 ± 1.58	3.49 ± 1.31	**1.04E-09**	2.66 ± 1.03
FD-MScM-SH vs. BMSKM-SH	3.86 ± 1.39	**0.806**	4.05 ± 1.86	3.49 ± 1.31	**0.577**	3.42 ± 1.33
FD-MScM-SH vs. BMSKM-EL	3.86 ± 1.39	**1.02E-07**	3.06 ± 1.30	3.49 ± 1.31	**2.97E-11**	2.57 ± 0.88
FD-MScM-EL vs. MSISA-P	3.47 ± 1.31	**0.915**	3.65 ± 1.74	3.24 ± 1.22	**0.890**	3.34 ± 1.33
FD-MScM-EL vs. MSISA-R	3.47 ± 1.31	**1.63E-12**	2.02 ± 0.52	3.24 ± 1.22	**1.06E-06**	1.79 ± 0.27
FD-MScM-EL vs. MSKM-SH	3.47 ± 1.31	**0.029**	3.97 ± 1.81	3.24 ± 1.22	**0.077**	3.59 ± 1.53

FD-MScM-EL vs. MSKM-EL	3.47 ± 1.31	**0.046**	3.24 ± 1.58	3.24 ± 1.22	**2.57E-06**	2.66 ± 1.03
FD-MScM-EL vs. BMSKM-SH	3.47 ± 1.31	**0.008**	4.05 ± 1.86	3.24 ± 1.22	**0.238**	3.42 ± 1.33
FD-MScM-EL vs. BMSKM-EL	3.47 ± 1.31	**0.004**	3.06 ± 1.30	3.24 ± 1.22	**9.75E-08**	2.57 ± 0.88

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	4.31 ± 1.94	6.74E-15	6.73 ± 3.35	4.82 ± 1.63	**0.245**	5.24 ± 2.35
FD-MScM-SH vs. QUBIC	4.31 ± 1.94	**4.06E-33**	2.06 ± 0.50	4.82 ± 1.63	**2.19E-30**	2.36 ± 0.37
FD-MScM-SH vs. FD-COAL	4.31 ± 1.94	**9.99E-27**	2.47 ± 1.12	4.82 ± 1.63	**1.35E-08**	3.51 ± 1.51
FD-MScM-EL vs. FD-SSCM	3.85 ± 1.82	**1.47E-22**	6.73 ± 3.35	4.55 ± 1.60	**0.015**	5.24 ± 2.35
FD-MScM-EL vs. QUBIC	3.85 ± 1.82	**8.34E-29**	2.06 ± 0.50	4.55 ± 1.60	**1.38E-26**	2.36 ± 0.37
FD-MScM-EL vs. FD-COAL	3.85 ± 1.82	**3.61E-20**	2.47 ± 1.12	4.55 ± 1.60	**1.80E-06**	3.51 ± 1.51

FD-MScM-SH vs. MSISA-P	4.31 ± 1.94	**0.054**	5.06 ± 2.38	4.82 ± 1.63	**0.007**	5.77 ± 1.91
FD-MScM-SH vs. MSISA-R	4.31 ± 1.94	**4.83E-15**	1.99 ± 0.50	4.82 ± 1.63	**9.28E-16**	2.42 ± 0.56
FD-MScM-SH vs. MSKM-SH	4.31 ± 1.94	**0.007**	4.79 ± 1.75	4.82 ± 1.63	**4.44E-04**	5.49 ± 1.73
FD-MScM-SH vs. MSKM-EL	4.31 ± 1.94	**4.42E-06**	3.45 ± 1.88	4.82 ± 1.63	**0.004**	4.35 ± 1.67
FD-MScM-SH vs. BMSKM-SH	4.31 ± 1.94	**0.161**	4.61 ± 2.13	4.82 ± 1.63	**0.320**	5.02 ± 1.71
FD-MScM-SH vs. BMSKM-EL	4.31 ± 1.94	**3.12E-08**	3.19 ± 1.39	4.82 ± 1.63	**0.006**	4.43 ± 1.62
FD-MScM-EL vs. MSISA-P	3.85 ± 1.82	**0.002**	5.06 ± 2.38	4.55 ± 1.60	**0.001**	5.77 ± 1.91
FD-MScM-EL vs. MSISA-R	3.85 ± 1.82	**4.04E-13**	1.99 ± 0.50	4.55 ± 1.60	**1.78E-13**	2.42 ± 0.56
FD-MScM-EL vs. MSKM-SH	3.85 ± 1.82	**3.82E-07**	4.79 ± 1.75	4.55 ± 1.60	**5.17E-06**	5.49 ± 1.73
FD-MScM-EL vs. MSKM-EL	3.85 ± 1.82	**0.007**	3.45 ± 1.88	4.55 ± 1.60	**0.145**	4.35 ± 1.67
FD-MScM-EL vs. BMSKM-SH	3.85 ± 1.82	**2.38E-04**	4.61 ± 2.13	4.55 ± 1.60	**0.026**	5.02 ± 1.71
FD-MScM-EL vs. BMSKM-EL	3.85 ± 1.82	**4.73E-04**	3.19 ± 1.39	4.55 ± 1.60	**0.244**	4.43 ± 1.62

B. anthracis - L. monocytogenes pairing

	B. anthracis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	3.80 ± 1.49	**0.007**	3.83 ± 2.57	4.98 ± 1.73	**0.673**	5.24 ± 2.35
FD-MScM-SH vs. QUBIC	3.80 ± 1.49	**1.97E-33**	1.77 ± 0.26	4.98 ± 1.73	**1.03E-33**	2.36 ± 0.37
FD-MScM-SH vs. FD-COAL	3.80 ± 1.49	**1.47E-20**	2.32 ± 1.57	4.98 ± 1.73	**6.20E-10**	3.51 ± 1.51
FD-MScM-EL vs. FD-SSCM	3.38 ± 1.31	**0.662**	3.83 ± 2.57	4.66 ± 1.73	**0.044**	5.24 ± 2.35
FD-MScM-EL vs. QUBIC	3.38 ± 1.31	**2.49E-29**	1.77 ± 0.26	4.66 ± 1.73	**1.04E-30**	2.36 ± 0.37
FD-MScM-EL vs. FD-COAL	3.38 ± 1.31	**6.13E-16**	2.32 ± 1.57	4.66 ± 1.73	**3.57E-07**	3.51 ± 1.51
FD-MScM-SH vs. MSKM-SH	3.80 ± 1.49	**0.303**	3.94 ± 1.46	4.98 ± 1.73	**0.001**	5.61 ± 1.74
FD-MScM-SH vs. MSKM-EL	3.80 ± 1.49	**8.75E-14**	2.61 ± 1.04	4.98 ± 1.73	**0.001**	4.37 ± 1.51
FD-MScM-EL vs. MSKM-SH	3.38 ± 1.31	**4.908E-04**	3.94 ± 1.46	4.66 ± 1.73	**3.80E-07**	5.61 ± 1.74
FD-MScM-EL vs. MSKM-EL	3.38 ± 1.31	**3.44E-08**	2.61 ± 1.04	4.66 ± 1.73	**0.191**	4.37 ± 1.51

7.2.3.1.5.2 Gram-negative triplet

Table 7.34: Comparison of bicluster sequence p-values from the full data methods considered by this study for all pairings of *E. coli*, *S. typhimurium* and *V. cholerae*.

***E. coli* - *S. typhimurium* pairing**

	E. coli			***S. typhimurium***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	3.31 ± 2.19	**5.25E-49**	8.24 ± 4.43	3.29 ± 1.47	**2.86E-05**	4.67 ± 3.36
FD-MScM-SH vs. QUBIC	3.31 ± 2.19	**4.87E-17**	7.36 ± 4.69	3.29 ± 1.47	**1.81E-20**	5.98 ± 2.47
FD-MScM-SH vs. FD-COAL	3.31 ± 2.19	**6.97E-07**	4.75 ± 3.19	3.29 ± 1.47	**1.87E-03**	2.84 ± 1.31
FD-MScM-EL vs. FD-SSCM	5.24 ± 2.98	**9.21E-18**	8.24 ± 4.43	4.54 ± 1.94	**0.03**	4.67 ± 3.36
FD-MScM-EL vs. QUBIC	5.24 ± 2.98	**3.19E-04**	7.36 ± 4.69	4.54 ± 1.94	**2.92E-07**	5.98 ± 2.47
FD-MScM-EL vs. FD-COAL	5.24 ± 2.98	**0.01**	4.75 ± 3.19	4.54 ± 1.94	**7.68E-18**	2.84 ± 1.31
FD-MScM-SH vs. MSISA-P	3.31 ± 2.19	**3.90E-13**	6.59 ± 3.53	3.29 ± 1.47	**6.13E-10**	5.66 ± 2.66
FD-MScM-SH vs. MSISA-R	3.31 ± 2.19	**1.55E-04**	4.32 ± 2.19	3.29 ± 1.47	**6.10E-04**	2.28 ± 0.57
FD-MScM-SH vs. MSKM-SH	3.31 ± 2.19	**4.96E-07**	4.56 ± 2.71	3.29 ± 1.47	**1.58E-09**	4.51 ± 2.24

FD-MScM-SH vs. MSKM-EL	3.31 ± 2.19	**6.70E-06**	4.37 ± 2.67	3.29 ± 1.47	**1.37E-04**	4.07 ± 2.07
FD-MScM-SH vs. BMSKM-SH	3.31 ± 2.19	**6.26E-08**	4.39 ± 2.52	3.29 ± 1.47	**2.24E-08**	4.32 ± 1.90
FD-MScM-SH vs. BMSKM-EL	3.31 ± 2.19	**9.65E-05**	4.12 ± 2.59	3.29 ± 1.47	**2.05E-03**	3.83 ± 1.97
FD-MScM-EL vs. MSISA-P	5.24 ± 2.98	**5.95E-03**	6.59 ± 3.53	4.54 ± 1.94	**9.72E-03**	5.66 ± 2.66
FD-MScM-EL vs. MSISA-R	5.24 ± 2.98	**0.05**	4.32 ± 2.19	4.54 ± 1.94	**3.79E-11**	2.28 ± 0.57
FD-MScM-EL vs. MSKM-SH	5.24 ± 2.98	**0.02**	4.56 ± 2.71	4.54 ± 1.94	**0.52**	4.51 ± 2.24
FD-MScM-EL vs. MSKM-EL	5.24 ± 2.98	**1.50E-03**	4.37 ± 2.67	4.54 ± 1.94	**3.87E-03**	4.07 ± 2.07
FD-MScM-EL vs. BMSKM-SH	5.24 ± 2.98	**6.44E-03**	4.39 ± 2.52	4.54 ± 1.94	**0.36**	4.32 ± 1.90
FD-MScM-EL vs. BMSKM-EL	5.24 ± 2.98	**2.50E-05**	4.12 ± 2.59	4.54 ± 1.94	**3.93E-05**	3.83 ± 1.97

E. coli - V. cholerae pairing

	E. coli			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	3.50 ± 1.31	**1.69E-47**	8.24 ± 4.43	3.25 ± 1.24	**8.29E-28**	9.14 ± 6.57

FD-MScM-SH vs. QUBIC	3.50 ± 1.31	**6.52E-15**	7.36 ± 4.69	3.25 ± 1.24	**1.29E-26**	9.18 ± 6.74
FD-MScM-SH vs. FD-COAL	3.50 ± 1.31	**3.13E-03**	4.75 ± 3.19	3.25 ± 1.24	**2.70E-03**	4.77 ± 3.84
FD-MScM-EL vs. FD-SSCM	5.43 ± 2.66	**7.26E-15**	8.24 ± 4.43	5.02 ± 2.42	**2.64E-10**	9.14 ± 6.57
FD-MScM-EL vs. QUBIC	5.43 ± 2.66	**4.92E-03**	7.36 ± 4.69	5.02 ± 2.42	**1.24E-09**	9.18 ± 6.74
FD-MScM-EL vs. FD-COAL	5.43 ± 2.66	**2.39E-04**	4.75 ± 3.19	5.02 ± 2.42	**8.58E-04**	4.77 ± 3.84
FD-MScM-SH vs. MSISA-P	3.50 ± 1.31	**2.88E-10**	7.05 ± 3.51	3.25 ± 1.24	**7.91E-08**	5.94 ± 2.76
FD-MScM-SH vs. MSISA-R	3.50 ± 1.31	**0.25**	4.06 ± 2.05	3.25 ± 1.24	**0.64**	3.73 ± 1.77
FD-MScM-SH vs. MSKM-SH	3.50 ± 1.31	**4.45E-09**	5.28 ± 3.32	3.25 ± 1.24	**4.46E-16**	5.27 ± 2.56
FD-MScM-SH vs. MSKM-EL	3.50 ± 1.31	**0.10**	4.38 ± 3.11	3.25 ± 1.24	**9.73E-05**	4.42 ± 2.69
FD-MScM-EL vs. MSISA-P	5.43 ± 2.66	**7.92E-03**	7.05 ± 3.51	5.02 ± 2.42	**0.05**	5.94 ± 2.76
FD-MScM-EL vs. MSISA-R	5.43 ± 2.66	**1.34E-03**	4.06 ± 2.05	5.02 ± 2.42	**0.17**	3.73 ± 1.77
FD-MScM-EL vs. MSKM-SH	5.43 ± 2.66	**0.23**	5.28 ± 3.32	5.02 ± 2.42	**0.29**	5.27 ± 2.56

FD-MScM-EL vs. MSKM-EL	5.43 ± 2.66	**1.11E-06**	4.38 ± 3.11	5.02 ± 2.42	**4.43E-03**	4.42 ± 2.69

S. typhimurium - V. cholerae pairing

	S. typhimurium			***V. cholerae***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
FD-MScM-SH vs. FD-SSCM	3.26 ± 1.25	**7.83E-05**	4.67 ± 3.36	2.98 ± 1.28	**8.90E-32**	9.14 ± 6.57
FD-MScM-SH vs. QUBIC	3.26 ± 1.25	**1.18E-21**	5.98 ± 2.47	2.98 ± 1.28	**3.50E-29**	9.18 ± 6.74
FD-MScM-EL vs. FD-SSCM	5.10 ± 2.06	**2.59E-05**	4.67 ± 3.36	4.85 ± 2.25	**2.04E-11**	9.14 ± 6.57
FD-MScM-EL vs. QUBIC	5.10 ± 2.06	**2.92E-03**	5.98 ± 2.47	4.85 ± 2.25	**1.06E-10**	9.18 ± 6.74
FD-MScM-EL vs. FD-COAL	5.10 ± 2.06	**1.92E-23**	2.84 ± 1.31	4.85 ± 2.25	**3.03E-03**	4.77 ± 3.84
FD-MScM-SH vs. MSKM-SH	3.26 ± 1.25	**1.42E-16**	5.17 ± 2.34	2.98 ± 1.28	**8.44E-18**	5.06 ± 2.61
FD-MScM-SH vs. MSKM-EL	3.26 ± 1.25	**3.49E-03**	3.84 ± 1.93	2.98 ± 1.28	**1.71E-07**	4.32 ± 2.55
FD-MScM-SH vs. BMSKM-SH	3.26 ± 1.25	**7.36E-16**	5.03 ± 2.17	2.98 ± 1.28	**3.95E-14**	4.89 ± 2.53
FD-MScM-EL vs. MSKM-SH	5.10 ± 2.06	**0.86**	5.17 ± 2.34	4.85 ± 2.25	**0.62**	5.06 ± 2.61

FD-MScM-EL vs. BMSKM-SH	5.10 ± 2.06	**0.56**	5.03 ± 2.17	4.85 ± 2.25	**0.95**	4.89 ± 2.53
FD-MScM-EL vs. BMSKM-EL	5.10 ± 2.06	**4.46E-11**	3.78 ± 1.96	4.85 ± 2.25	**5.53E-06**	3.89 ± 2.22

7.2.3.2 Comparisons with EO-MScM

In the comparisons below, we only show results from the Gram-positive triplet, as the expression only results from the Gram-negative triplet were largely uninformative. We direct the reader to section 3.1.3.1 for a further explanation.

7.2.3.2.1 Residuals

Table 7.35: Comparison of bicluster residuals from the expression only methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*. A comparison of the residuals of the results from MScM (expression only) with all other relevant methods for all 3 pairings of the three organisms examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. In this case, these results illustrate that in 61 of the 116 comparisons (52.6%) MScM step did as well or better than its competitors.

***B. subtilis* - *B. anthracis* pairing**

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**

EO-MScM-SH vs. EO-SSCM	0.52 ± 0.09	**6.21E-20**	0.44 ± 0.20	0.50 ± 0.20	**1.89E-38**	0.23 ± 0.06
EO-MScM-SH vs. FD-SSCM	0.52 ± 0.09	**2.38E-05**	0.49 ± 0.13	0.50 ± 0.20	**1.79E-23**	0.31 ± 0.12
EO-MScM-SH vs. QUBIC	0.52 ± 0.09	**9.34E-36**	0.87 ± 0.21	0.50 ± 0.20	**3.12E-50**	1.51 ± 0.29
EO-MScM-SH vs. EO-COAL	0.52 ± 0.09	**7.01E-40**	0.78 ± 0.23	0.50 ± 0.20	**1.98E-04**	0.58 ± 0.17
EO-MScM-SH vs. FD-COAL	0.52 ± 0.09	**3.60E-40**	0.80 ± 0.25	0.50 ± 0.20	**3.53E-04**	0.58 ± 0.17
EO-MScM-EL vs. EO-SSCM	0.52 ± 0.10	**2.72E-18**	0.44 ± 0.20	0.49 ± 0.20	**3.63E-35**	0.23 ± 0.06
EO-MScM-EL vs. FD-SSCM	0.52 ± 0.10	**0.003**	0.49 ± 0.13	0.49 ± 0.20	**2.76E-20**	0.31 ± 0.12
EO-MScM-EL vs. QUBIC	0.52 ± 0.10	**1.35E-35**	0.87 ± 0.21	0.49 ± 0.20	**2.88E-50**	1.51 ± 0.29
EO-MScM-EL vs. EO-COAL	0.52 ± 0.10	**2.61E-42**	0.78 ± 0.23	0.49 ± 0.20	**3.81E-05**	0.58 ± 0.17
EO-MScM-EL vs. FD-COAL	0.52 ± 0.10	**1.63E-42**	0.80 ± 0.25	0.49 ± 0.20	**6.99E-05**	0.58 ± 0.17
EO-MScM-SH vs. MSISA-P	0.52 ± 0.09	**9.13E-17**	0.98 ± 0.39	0.50 ± 0.20	**1.71E-22**	1.97 ± 0.94
EO-MScM-SH vs. MSISA-R	0.52 ± 0.09	**2.34E-19**	1.11 ± 0.41	0.50 ± 0.20	**5.65E-22**	1.58 ± 0.38

EO-MScM-SH vs. MSKM-SH	0.52 ± 0.09	**4.07E-30**	0.41 ± 0.07	0.50 ± 0.20	**0.107**	0.53 ± 0.12
EO-MScM-SH vs. MSKM-EL	0.52 ± 0.09	**2.86E-27**	0.42 ± 0.06	0.50 ± 0.20	**0.455**	0.48 ± 0.11
EO-MScM-SH vs. BMSKM-SH	0.52 ± 0.09	**4.07E-14**	0.45 ± 0.07	0.50 ± 0.20	**8.32E-08**	0.38 ± 0.07
EO-MScM-SH vs. BMSKM-EL	0.52 ± 0.09	**1.08E-14**	0.45 ± 0.06	0.50 ± 0.20	**1.06E-06**	0.39 ± 0.07
EO-MScM-EL vs. MSISA-P	0.52 ± 0.10	**6.41E-17**	0.98 ± 0.39	0.49 ± 0.20	**1.51E-22**	1.97 ± 0.94
EO-MScM-EL vs. MSISA-R	0.52 ± 0.10	**2.20E-19**	1.11 ± 0.41	0.49 ± 0.20	**5.31E-22**	1.58 ± 0.38
EO-MScM-EL vs. MSKM-SH	0.52 ± 0.10	**5.39E-27**	0.41 ± 0.07	0.49 ± 0.20	**0.047**	0.53 ± 0.12
EO-MScM-EL vs. MSKM-EL	0.52 ± 0.10	**5.92E-24**	0.42 ± 0.06	0.49 ± 0.20	**0.703**	0.48 ± 0.11
EO-MScM-EL vs. BMSKM-SH	0.52 ± 0.10	**1.11E-09**	0.45 ± 0.07	0.49 ± 0.20	**1.04E-06**	0.38 ± 0.07
EO-MScM-EL vs. BMSKM-EL	0.52 ± 0.10	**1.24E-09**	0.45 ± 0.06	0.49 ± 0.20	**8.70E-06**	0.39 ± 0.07

B. subtilis - L. monocytogenes pairing

	B. subtilis	***L. monocytogenes***

dist1 vs. dist2	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
EO-MScM-SH vs. EO-SSCM	0.52 ± 0.08	**6.03E-20**	0.44 ± 0.20	0.49 ± 0.17	**1.42E-17**	0.29 ± 0.10
EO-MScM-SH vs. FD-SSCM	0.52 ± 0.08	**1.30E-04**	0.49 ± 0.13	0.49 ± 0.17	**5.74E-08**	0.40 ± 0.18
EO-MScM-SH vs. QUBIC	0.52 ± 0.08	**6.59E-37**	0.87 ± 0.21	0.49 ± 0.17	**1.06E-23**	1.81 ± 0.85
EO-MScM-SH vs. EO-COAL	0.52 ± 0.08	**1.33E-43**	0.78 ± 0.23	0.49 ± 0.17	**0.657**	1.63 ± 3.07
EO-MScM-SH vs. FD-COAL	0.52 ± 0.08	**2.21E-43**	0.80 ± 0.25	0.49 ± 0.17	**0.617**	1.70 ± 3.24
EO-MScM-EL vs. EO-SSCM	0.50 ± 0.09	**7.36E-17**	0.44 ± 0.20	0.48 ± 0.17	**1.85E-17**	0.29 ± 0.10
EO-MScM-EL vs. FD-SSCM	0.50 ± 0.09	**0.055**	0.49 ± 0.13	0.48 ± 0.17	**1.75E-07**	0.40 ± 0.18
EO-MScM-EL vs. QUBIC	0.50 ± 0.09	**1.10E-37**	0.87 ± 0.21	0.48 ± 0.17	**5.53E-24**	1.81 ± 0.85
EO-MScM-EL vs. EO-COAL	0.50 ± 0.09	**6.71E-47**	0.78 ± 0.23	0.48 ± 0.17	**0.755**	1.63 ± 3.07
EO-MScM-EL vs. FD-COAL	0.50 ± 0.09	**1.26E-46**	0.80 ± 0.25	0.48 ± 0.17	**0.734**	1.70 ± 3.24
EO-MScM-SH vs. MSISA-P	0.52 ± 0.08	**8.65E-10**	0.87 ± 0.34	0.49 ± 0.17	**4.97E-19**	1.59 ± 0.52

EO-MScM-SH vs. MSISA-R	0.52 ± 0.08	**1.64E-18**	1.11 ± 0.42	0.49 ± 0.17	**3.60E-21**	1.31 ± 0.34
EO-MScM-SH vs. MSKM-SH	0.52 ± 0.08	**1.11E-34**	0.40 ± 0.07	0.49 ± 0.17	**0.358**	0.50 ± 0.12
EO-MScM-SH vs. MSKM-EL	0.52 ± 0.08	**3.97E-28**	0.42 ± 0.06	0.49 ± 0.17	**0.683**	0.48 ± 0.11
EO-MScM-SH vs. BMSKM-SH	0.52 ± 0.08	**1.14E-20**	0.43 ± 0.07	0.49 ± 0.17	**3.32E-04**	0.42 ± 0.09
EO-MScM-SH vs. BMSKM-EL	0.52 ± 0.08	**1.16E-18**	0.44 ± 0.06	0.49 ± 0.17	**7.46E-04**	0.42 ± 0.09
EO-MScM-EL vs. MSISA-P	0.50 ± 0.09	**4.44E-10**	0.87 ± 0.34	0.48 ± 0.17	**4.97E-19**	1.59 ± 0.52
EO-MScM-EL vs. MSISA-R	0.50 ± 0.09	**1.38E-18**	1.11 ± 0.42	0.48 ± 0.17	**3.60E-21**	1.31 ± 0.34
EO-MScM-EL vs. MSKM-SH	0.50 ± 0.09	**4.41E-29**	0.40 ± 0.07	0.48 ± 0.17	**0.190**	0.50 ± 0.12
EO-MScM-EL vs. MSKM-EL	0.50 ± 0.09	**5.98E-22**	0.42 ± 0.06	0.48 ± 0.17	**0.972**	0.48 ± 0.11
EO-MScM-EL vs. BMSKM-SH	0.50 ± 0.09	**2.09E-14**	0.43 ± 0.07	0.48 ± 0.17	**0.001**	0.42 ± 0.09
EO-MScM-EL vs. BMSKM-EL	0.50 ± 0.09	**4.23E-12**	0.44 ± 0.06	0.48 ± 0.17	**0.002**	0.42 ± 0.09

B. anthracis - L. monocytogenes pairing

dist1 vs. dist2	*B. anthracis*			*L. monocytogenes*		
	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
EO-MScM-SH vs. EO-SSCM	0.52 ± 0.17	**1.52E-48**	0.23 ± 0.06	0.50 ± 0.18	**2.48E-17**	0.29 ± 0.10
EO-MScM-SH vs. FD-SSCM	0.52 ± 0.17	**9.86E-34**	0.31 ± 0.12	0.50 ± 0.18	**1.67E-08**	0.40 ± 0.18
EO-MScM-SH vs. QUBIC	0.52 ± 0.17	**4.65E-49**	1.51 ± 0.29	0.50 ± 0.18	**1.06E-22**	1.81 ± 0.85
EO-MScM-SH vs. EO-COAL	0.52 ± 0.17	**5.58E-04**	0.58 ± 0.17	0.50 ± 0.18	**0.568**	1.63 ± 3.07
EO-MScM-SH vs. FD-COAL	0.52 ± 0.17	**8.27E-04**	0.58 ± 0.17	0.50 ± 0.18	**0.532**	1.70 ± 3.24
EO-MScM-EL vs. EO-SSCM	0.50 ± 0.17	**2.27E-45**	0.23 ± 0.06	0.50 ± 0.19	**1.46E-16**	0.29 ± 0.10
EO-MScM-EL vs. FD-SSCM	0.50 ± 0.17	**1.11E-29**	0.31 ± 0.12	0.50 ± 0.19	**1.20E-07**	0.40 ± 0.18
EO-MScM-EL vs. QUBIC	0.50 ± 0.17	**4.46E-49**	1.51 ± 0.29	0.50 ± 0.19	**2.80E-23**	1.81 ± 0.85
EO-MScM-EL vs. EO-COAL	0.50 ± 0.17	**5.24E-05**	0.58 ± 0.17	0.50 ± 0.19	**0.810**	1.63 ± 3.07
EO-MScM-EL vs. FD-COAL	0.50 ± 0.17	**6.24E-05**	0.58 ± 0.17	0.50 ± 0.19	**0.753**	1.70 ± 3.24
EO-MScM-SH vs. MSKM-SH	0.52 ± 0.17	**1.15E-10**	0.40 ± 0.08	0.50 ± 0.18	**0.002**	0.43 ± 0.08

EO-MScM-SH vs. MSKM-EL	0.52 ± 0.17	**1.26E-11**	0.39 ± 0.07	0.50 ± 0.18	**4.66E-04**	0.43 ± 0.08
EO-MScM-EL vs. MSKM-SH	0.50 ± 0.17	**6.56E-09**	0.40 ± 0.08	0.50 ± 0.19	**0.010**	0.43 ± 0.08
EO-MScM-EL vs. MSKM-EL	0.50 ± 0.17	**1.83E-09**	0.39 ± 0.07	0.50 ± 0.19	**0.003**	0.43 ± 0.08

7.2.3.2.2 Mean correlations

Table 7.36: Comparison of bicluster mean correlations from the expression only methods considered by this study for all pairings of *B. subtilis, B. anthracis* and *L. monocytogenes*. A comparison of the mean correlations of the results from MScM (expression only) with all other relevant methods for all 3 pairings of the three organisms examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. In this case, these results illustrate that in 65 of the 116 comparisons (56%) MScM step did as well or better than its competitors.

B. subtilis - B. anthracis pairing

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
EO-MScM-SH vs. EO-SSCM	0.52 ± 0.12	**2.25E-29**	0.70 ± 0.11	0.69 ± 0.17	**1.02E-34**	0.91 ± 0.05

EO-MScM-SH vs. FD-SSCM	0.52 ± 0.12	**0.007**	0.56 ± 0.14	0.69 ± 0.17	**7.13E-15**	0.82 ± 0.15
EO-MScM-SH vs. QUBIC	0.52 ± 0.12	**1.23E-21**	0.36 ± 0.19	0.69 ± 0.17	**4.14E-23**	0.49 ± 0.05
EO-MScM-SH vs. EO-COAL	0.52 ± 0.12	**5.25E-05**	0.58 ± 0.14	0.69 ± 0.17	**0.002**	0.64 ± 0.13
EO-MScM-SH vs. FD-COAL	0.52 ± 0.12	**2.49E-05**	0.59 ± 0.15	0.69 ± 0.17	**1.05E-04**	0.62 ± 0.13
EO-MScM-EL vs. EO-SSCM	0.54 ± 0.12	**2.67E-26**	0.70 ± 0.11	0.69 ± 0.19	**2.90E-31**	0.91 ± 0.05
EO-MScM-EL vs. FD-SSCM	0.54 ± 0.12	**0.098**	0.56 ± 0.14	0.69 ± 0.19	**4.09E-13**	0.82 ± 0.15
EO-MScM-EL vs. QUBIC	0.54 ± 0.12	**3.57E-23**	0.36 ± 0.19	0.69 ± 0.19	**3.11E-19**	0.49 ± 0.05
EO-MScM-EL vs. EO-COAL	0.54 ± 0.12	**0.003**	0.58 ± 0.14	0.69 ± 0.19	**0.004**	0.64 ± 0.13
EO-MScM-EL vs. FD-COAL	0.54 ± 0.12	**0.001**	0.59 ± 0.15	0.69 ± 0.19	**2.44E-04**	0.62 ± 0.13
EO-MScM-SH vs. MSISA-P	0.52 ± 0.12	**0.003**	0.60 ± 0.14	0.69 ± 0.17	**5.35E-06**	0.56 ± 0.07
EO-MScM-SH vs. MSISA-R	0.52 ± 0.12	**0.304**	0.55 ± 0.13	0.69 ± 0.17	**7.54E-10**	0.51 ± 0.03
EO-MScM-SH vs. MSKM-SH	0.52 ± 0.12	**2.38E-05**	0.58 ± 0.11	0.69 ± 0.17	**1.58E-14**	0.52 ± 0.14

EO-MScM-SH vs. MSKM-EL	0.52 ± 0.12	**0.003**	0.56 ± 0.11	0.69 ± 0.17	**4.87E-07**	0.58 ± 0.15
EO-MScM-SH vs. BMSKM-SH	0.52 ± 0.12	**0.028**	0.49 ± 0.13	0.69 ± 0.17	**0.265**	0.72 ± 0.10
EO-MScM-SH vs. BMSKM-EL	0.52 ± 0.12	**0.107**	0.50 ± 0.12	0.69 ± 0.17	**0.341**	0.71 ± 0.10
EO-MScM-EL vs. MSISA-P	0.54 ± 0.12	**0.020**	0.60 ± 0.14	0.69 ± 0.19	**2.01E-05**	0.56 ± 0.07
EO-MScM-EL vs. MSISA-R	0.54 ± 0.12	**0.629**	0.55 ± 0.13	0.69 ± 0.19	**4.92E-08**	0.51 ± 0.03
EO-MScM-EL vs. MSKM-SH	0.54 ± 0.12	**0.002**	0.58 ± 0.11	0.69 ± 0.19	**2.01E-13**	0.52 ± 0.14
EO-MScM-EL vs. MSKM-EL	0.54 ± 0.12	**0.057**	0.56 ± 0.11	0.69 ± 0.19	**1.23E-06**	0.58 ± 0.15
EO-MScM-EL vs. BMSKM-SH	0.54 ± 0.12	**0.002**	0.49 ± 0.13	0.69 ± 0.19	**0.346**	0.72 ± 0.10
EO-MScM-EL vs. BMSKM-EL	0.54 ± 0.12	**0.012**	0.50 ± 0.12	0.69 ± 0.19	**0.413**	0.71 ± 0.10

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**

EO-MScM-SH vs. EO-SSCM	0.52 ± 0.13	**3.09E-28**	0.70 ± 0.11	0.64 ± 0.18	**1.43E-18**	0.86 ± 0.08
EO-MScM-SH vs. FD-SSCM	0.52 ± 0.13	**0.004**	0.56 ± 0.14	0.64 ± 0.18	**1.67E-04**	0.71 ± 0.20
EO-MScM-SH vs. QUBIC	0.52 ± 0.13	**3.89E-21**	0.36 ± 0.19	0.64 ± 0.18	**1.06E-13**	0.45 ± 0.27
EO-MScM-SH vs. EO-COAL	0.52 ± 0.13	**2.15E-05**	0.58 ± 0.14	0.64 ± 0.18	**1.37E-09**	0.81 ± 0.13
EO-MScM-SH vs. FD-COAL	0.52 ± 0.13	**9.27E-06**	0.59 ± 0.15	0.64 ± 0.18	**5.02E-09**	0.80 ± 0.12
EO-MScM-EL vs. EO-SSCM	0.54 ± 0.12	**1.67E-24**	0.70 ± 0.11	0.64 ± 0.18	**2.87E-19**	0.86 ± 0.08
EO-MScM-EL vs. FD-SSCM	0.54 ± 0.12	**0.187**	0.56 ± 0.14	0.64 ± 0.18	**8.67E-05**	0.71 ± 0.20
EO-MScM-EL vs. QUBIC	0.54 ± 0.12	**1.86E-23**	0.36 ± 0.19	0.64 ± 0.18	**2.98E-13**	0.45 ± 0.27
EO-MScM-EL vs. EO-COAL	0.54 ± 0.12	**0.007**	0.58 ± 0.14	0.64 ± 0.18	**8.12E-10**	0.81 ± 0.13
EO-MScM-EL vs. FD-COAL	0.54 ± 0.12	**0.003**	0.59 ± 0.15	0.64 ± 0.18	**1.67E-09**	0.80 ± 0.12
EO-MScM-SH vs. MSISA-P	0.52 ± 0.13	**0.099**	0.60 ± 0.20	0.64 ± 0.18	**8.39E-05**	0.47 ± 0.23
EO-MScM-SH vs. MSISA-R	0.52 ± 0.13	**0.268**	0.55 ± 0.12	0.64 ± 0.18	**0.001**	0.50 ± 0.27

EO-MScM-SH vs. MSKM-SH	0.52 ± 0.13	**1.40E-07**	0.59 ± 0.11	0.64 ± 0.18	**7.37E-10**	0.51 ± 0.17
EO-MScM-SH vs. MSKM-EL	0.52 ± 0.13	**0.002**	0.56 ± 0.11	0.64 ± 0.18	**1.87E-05**	0.55 ± 0.16
EO-MScM-SH vs. BMSKM-SH	0.52 ± 0.13	**0.737**	0.52 ± 0.14	0.64 ± 0.18	**0.403**	0.63 ± 0.15
EO-MScM-SH vs. BMSKM-EL	0.52 ± 0.13	**0.311**	0.53 ± 0.12	0.64 ± 0.18	**0.883**	0.64 ± 0.14
EO-MScM-EL vs. MSISA-P	0.54 ± 0.12	**0.324**	0.60 ± 0.20	0.64 ± 0.18	**6.38E-05**	0.47 ± 0.23
EO-MScM-EL vs. MSISA-R	0.54 ± 0.12	**0.874**	0.55 ± 0.12	0.64 ± 0.18	**0.002**	0.50 ± 0.27
EO-MScM-EL vs. MSKM-SH	0.54 ± 0.12	**8.84E-05**	0.59 ± 0.11	0.64 ± 0.18	**4.30E-10**	0.51 ± 0.17
EO-MScM-EL vs. MSKM-EL	0.54 ± 0.12	**0.121**	0.56 ± 0.11	0.64 ± 0.18	**1.46E-05**	0.55 ± 0.16
EO-MScM-EL vs. BMSKM-SH	0.54 ± 0.12	**0.298**	0.52 ± 0.14	0.64 ± 0.18	**0.406**	0.63 ± 0.15
EO-MScM-EL vs. BMSKM-EL	0.54 ± 0.12	**0.578**	0.53 ± 0.12	0.64 ± 0.18	**0.852**	0.64 ± 0.14

B. anthracis - L. monocytogenes pairing

	B. anthracis	***L. monocytogenes***

dist1 vs. dist2	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
EO-MScM-SH vs. EO-SSCM	0.63 ± 0.16	**2.22E-48**	0.91 ± 0.05	0.63 ± 0.19	**1.01E-18**	0.86 ± 0.08
EO-MScM-SH vs. FD-SSCM	0.63 ± 0.16	**1.47E-28**	0.82 ± 0.15	0.63 ± 0.19	**2.84E-05**	0.71 ± 0.20
EO-MScM-SH vs. QUBIC	0.63 ± 0.16	**3.47E-14**	0.49 ± 0.05	0.63 ± 0.19	**3.42E-12**	0.45 ± 0.27
EO-MScM-SH vs. EO-COAL	0.63 ± 0.16	**0.824**	0.64 ± 0.13	0.63 ± 0.19	**5.52E-10**	0.81 ± 0.13
EO-MScM-SH vs. FD-COAL	0.63 ± 0.16	**0.341**	0.62 ± 0.13	0.63 ± 0.19	**1.30E-09**	0.80 ± 0.12
EO-MScM-EL vs. EO-SSCM	0.63 ± 0.17	**6.29E-45**	0.91 ± 0.05	0.63 ± 0.19	**1.54E-18**	0.86 ± 0.08
EO-MScM-EL vs. FD-SSCM	0.63 ± 0.17	**1.56E-25**	0.82 ± 0.15	0.63 ± 0.19	**3.54E-05**	0.71 ± 0.20
EO-MScM-EL vs. QUBIC	0.63 ± 0.17	**3.37E-15**	0.49 ± 0.05	0.63 ± 0.19	**1.22E-11**	0.45 ± 0.27
EO-MScM-EL vs. EO-COAL	0.63 ± 0.17	**0.910**	0.64 ± 0.13	0.63 ± 0.19	**5.61E-10**	0.81 ± 0.13
EO-MScM-EL vs. FD-COAL	0.63 ± 0.17	**0.280**	0.62 ± 0.13	0.63 ± 0.19	**1.98E-09**	0.80 ± 0.12
EO-MScM-SH vs. MSKM-SH	0.63 ± 0.16	**0.002**	0.69 ± 0.12	0.63 ± 0.19	**0.250**	0.60 ± 0.14

EO-MScM-SH vs. MSKM-EL	0.63 ± 0.16	**4.39E-05**	0.70 ± 0.10	0.63 ± 0.19	**0.892**	0.63 ± 0.13
EO-MScM-EL vs. MSKM-SH	0.63 ± 0.17	**0.003**	0.69 ± 0.12	0.63 ± 0.19	**0.224**	0.60 ± 0.14
EO-MScM-EL vs. MSKM-EL	0.63 ± 0.17	**1.20E-04**	0.70 ± 0.10	0.63 ± 0.19	**0.817**	0.63 ± 0.13

7.2.3.2.3 Network Association p-values

Table 7.37: Comparison of bicluster network association p-values from the expression only methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*. A comparison of the residuals of the results from MScM (expression only) with all other relevant methods for all 3 pairings of the three organisms examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. In this case, these results illustrate that in 103 of the 116 comparisons (88.8%) MScM step did as well or better than its competitors.

B. subtilis - B. anthracis pairing

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
EO-MScM-SH vs. EO-SSCM	6.24 ± 7.63	**0.435**	7.63 ± 9.05	6.45 ± 7.29	**0.416**	5.39 ± 6.03

EO-MScM-SH vs. FD-SSCM	6.24 ± 7.63	**2.07E-05**	9.78 ± 9.76	6.45 ± 7.29	**0.721**	5.44 ± 5.38
EO-MScM-SH vs. QUBIC	6.24 ± 7.63	**6.42E-16**	2.52 ± 4.78	6.45 ± 7.29	**0.938**	6.73 ± 7.52
EO-MScM-SH vs. EO-COAL	6.24 ± 7.63	**0.568**	7.75 ± 9.50	6.45 ± 7.29	**0.198**	6.18 ± 7.98
EO-MScM-SH vs. FD-COAL	6.24 ± 7.63	**0.451**	7.57 ± 9.16	6.45 ± 7.29	**0.144**	6.50 ± 8.74
EO-MScM-EL vs. EO-SSCM	6.32 ± 7.80	**0.328**	7.63 ± 9.05	6.24 ± 7.31	**0.716**	5.39 ± 6.03
EO-MScM-EL vs. FD-SSCM	6.32 ± 7.80	**3.91E-06**	9.78 ± 9.76	6.24 ± 7.31	**0.950**	5.44 ± 5.38
EO-MScM-EL vs. QUBIC	6.32 ± 7.80	**6.13E-16**	2.52 ± 4.78	6.24 ± 7.31	**0.702**	6.73 ± 7.52
EO-MScM-EL vs. EO-COAL	6.32 ± 7.80	**0.402**	7.75 ± 9.50	6.24 ± 7.31	**0.316**	6.18 ± 7.98
EO-MScM-EL vs. FD-COAL	6.32 ± 7.80	**0.330**	7.57 ± 9.16	6.24 ± 7.31	**0.226**	6.50 ± 8.74
EO-MScM-SH vs. MSISA-P	6.24 ± 7.63	**0.490**	5.56 ± 5.86	6.45 ± 7.29	**0.814**	5.61 ± 7.17
EO-MScM-SH vs. MSISA-R	6.24 ± 7.63	**4.21E-04**	9.69 ± 9.37	6.45 ± 7.29	**0.464**	9.66 ± 9.95
EO-MScM-SH vs. MSKM-SH	6.24 ± 7.63	**0.190**	7.87 ± 9.35	6.45 ± 7.29	**0.060**	4.38 ± 5.10

EO-MScM-SH vs. MSKM-EL	6.24 ± 7.63	**0.137**	8.15 ± 9.65	6.45 ± 7.29	**0.005**	4.06 ± 5.32
EO-MScM-SH vs. BMSKM-SH	6.24 ± 7.63	**0.110**	7.27 ± 8.25	6.45 ± 7.29	**0.489**	5.54 ± 6.48
EO-MScM-SH vs. BMSKM-EL	6.24 ± 7.63	**0.346**	6.93 ± 8.19	6.45 ± 7.29	**0.024**	4.56 ± 5.86
EO-MScM-EL vs. MSISA-P	6.32 ± 7.80	**0.429**	5.56 ± 5.86	6.24 ± 7.31	**0.935**	5.61 ± 7.17
EO-MScM-EL vs. MSISA-R	6.32 ± 7.80	**0.000**	9.69 ± 9.37	6.24 ± 7.31	**0.282**	9.66 ± 9.95
EO-MScM-EL vs. MSKM-SH	6.32 ± 7.80	**0.130**	7.87 ± 9.35	6.24 ± 7.31	**0.113**	4.38 ± 5.10
EO-MScM-EL vs. MSKM-EL	6.32 ± 7.80	**0.090**	8.15 ± 9.65	6.24 ± 7.31	**0.012**	4.06 ± 5.32
EO-MScM-EL vs. BMSKM-SH	6.32 ± 7.80	**0.063**	7.27 ± 8.25	6.24 ± 7.31	**0.676**	5.54 ± 6.48
EO-MScM-EL vs. BMSKM-EL	6.32 ± 7.80	**0.247**	6.93 ± 8.19	6.24 ± 7.31	**0.054**	4.56 ± 5.86

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**

EO-MScM-SH vs. EO-SSCM	9.11 ± 10.61	**0.419**	7.63 ± 9.05	6.31 ± 8.71	**0.594**	4.42 ± 6.01
EO-MScM-SH vs. FD-SSCM	9.11 ± 10.61	**0.058**	9.78 ± 9.76	6.31 ± 8.71	**0.003**	6.90 ± 7.75
EO-MScM-SH vs. QUBIC	9.11 ± 10.61	**8.39E-16**	2.52 ± 4.78	6.31 ± 8.71	**2.23E-05**	9.95 ± 10.82
EO-MScM-SH vs. EO-COAL	9.11 ± 10.61	**0.279**	7.75 ± 9.50	6.31 ± 8.71	**0.422**	5.24 ± 7.41
EO-MScM-SH vs. FD-COAL	9.11 ± 10.61	**0.334**	7.57 ± 9.16	6.31 ± 8.71	**0.744**	5.93 ± 8.27
EO-MScM-EL vs. EO-SSCM	8.50 ± 10.57	**0.852**	7.63 ± 9.05	6.00 ± 8.15	**0.546**	4.42 ± 6.01
EO-MScM-EL vs. FD-SSCM	8.50 ± 10.57	**0.002**	9.78 ± 9.76	6.00 ± 8.15	**0.002**	6.90 ± 7.75
EO-MScM-EL vs. QUBIC	8.50 ± 10.57	**1.01E-13**	2.52 ± 4.78	6.00 ± 8.15	**1.34E-05**	9.95 ± 10.82
EO-MScM-EL vs. EO-COAL	8.50 ± 10.57	**0.999**	7.75 ± 9.50	6.00 ± 8.15	**0.440**	5.24 ± 7.41
EO-MScM-EL vs. FD-COAL	8.50 ± 10.57	**0.894**	7.57 ± 9.16	6.00 ± 8.15	**0.751**	5.93 ± 8.27
EO-MScM-SH vs. MSISA-P	9.11 ± 10.61	**0.317**	9.05 ± 8.89	6.31 ± 8.71	**0.533**	3.70 ± 1.79
EO-MScM-SH vs. MSISA-R	9.11 ± 10.61	**0.126**	9.61 ± 9.29	6.31 ± 8.71	**0.077**	6.20 ± 6.65

EO-MScM-SH vs. MSKM-SH	9.11 ± 10.61	**0.199**	9.76 ± 10.54	6.31 ± 8.71	**0.052**	7.88 ± 9.56
EO-MScM-SH vs. MSKM-EL	9.11 ± 10.61	**0.499**	7.68 ± 9.47	6.31 ± 8.71	**0.895**	4.91 ± 6.44
EO-MScM-SH vs. BMSKM-SH	9.11 ± 10.61	**0.504**	9.23 ± 10.39	6.31 ± 8.71	**0.327**	7.10 ± 9.45
EO-MScM-SH vs. BMSKM-EL	9.11 ± 10.61	**0.112**	6.79 ± 8.75	6.31 ± 8.71	**0.960**	4.86 ± 6.39
EO-MScM-EL vs. MSISA-P	8.50 ± 10.57	**0.139**	9.05 ± 8.89	6.00 ± 8.15	**0.603**	3.70 ± 1.79
EO-MScM-EL vs. MSISA-R	8.50 ± 10.57	**0.015**	9.61 ± 9.29	6.00 ± 8.15	**0.061**	6.20 ± 6.65
EO-MScM-EL vs. MSKM-SH	8.50 ± 10.57	**0.022**	9.76 ± 10.54	6.00 ± 8.15	**0.038**	7.88 ± 9.56
EO-MScM-EL vs. MSKM-EL	8.50 ± 10.57	**0.712**	7.68 ± 9.47	6.00 ± 8.15	**0.873**	4.91 ± 6.44
EO-MScM-EL vs. BMSKM-SH	8.50 ± 10.57	**0.110**	9.23 ± 10.39	6.00 ± 8.15	**0.304**	7.10 ± 9.45
EO-MScM-EL vs. BMSKM-EL	8.50 ± 10.57	**0.548**	6.79 ± 8.75	6.00 ± 8.15	**0.932**	4.86 ± 6.39

B. anthracis - L. monocytogenes pairing

	B. anthracis	***L. monocytogenes***

dist1 vs. dist2	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
EO-MScM-SH vs. EO-SSCM	6.75 ± 9.09	**0.621**	5.39 ± 6.03	5.63 ± 8.48	**0.982**	4.42 ± 6.01
EO-MScM-SH vs. FD-SSCM	6.75 ± 9.09	**0.440**	5.47 ± 5.39	5.63 ± 8.48	**6.20E-05**	6.90 ± 7.75
EO-MScM-SH vs. QUBIC	6.75 ± 9.09	**0.392**	6.73 ± 7.52	5.63 ± 8.48	**1.87E-07**	9.95 ± 10.82
EO-MScM-SH vs. EO-COAL	6.75 ± 9.09	**0.794**	6.18 ± 7.98	5.63 ± 8.48	**0.687**	5.24 ± 7.41
EO-MScM-SH vs. FD-COAL	6.75 ± 9.09	**0.623**	6.50 ± 8.74	5.63 ± 8.48	**0.929**	5.93 ± 8.27
EO-MScM-EL vs. EO-SSCM	6.82 ± 8.86	**0.872**	5.39 ± 6.03	5.93 ± 8.70	**0.859**	4.42 ± 6.01
EO-MScM-EL vs. FD-SSCM	6.82 ± 8.86	**0.653**	5.47 ± 5.39	5.93 ± 8.70	**1.29E-04**	6.90 ± 7.75
EO-MScM-EL vs. QUBIC	6.82 ± 8.86	**0.514**	6.73 ± 7.52	5.93 ± 8.70	**3.52E-07**	9.95 ± 10.82
EO-MScM-EL vs. EO-COAL	6.82 ± 8.86	**0.555**	6.18 ± 7.98	5.93 ± 8.70	**0.533**	5.24 ± 7.41
EO-MScM-EL vs. FD-COAL	6.82 ± 8.86	**0.451**	6.50 ± 8.74	5.93 ± 8.70	**0.959**	5.93 ± 8.27
EO-MScM-SH vs. MSKM-SH	6.75 ± 9.09	**0.623**	5.67 ± 7.00	5.63 ± 8.48	**0.087**	6.83 ± 8.86

EO-MScM-SH vs. MSKM-EL	6.75 ± 9.09	**0.401**	3.86 ± 4.13	5.63 ± 8.48	**0.693**	4.94 ± 6.73
EO-MScM-EL vs. MSKM-SH	6.82 ± 8.86	**0.749**	5.67 ± 7.00	5.93 ± 8.70	**0.118**	6.83 ± 8.86
EO-MScM-EL vs. MSKM-EL	6.82 ± 8.86	**0.190**	3.86 ± 4.13	5.93 ± 8.70	**0.847**	4.94 ± 6.73

7.2.3.2.4 Motif E-values

Table 7.38: Comparison of bicluster motif E-values (-log10) from the expression only methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*. A comparison of the motif E-values (-log10) from MScM (expression only) with all other relevant methods for all 3 pairings of the three organisms examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. As the table indicates, in 39 of the 116 of the comparisons (33.7%) MScM does as well or better than its competitors. This is by far the metric that EO-MScM does on.

B. subtilis - B. anthracis pairing

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
EO-MScM-SH vs. EO-SSCM	-0.90 ± 3.44	**0.026**	0.71 ± 5.69	0.09 ± 4.65	**6.24E-07**	2.17 ± 5.38

EO-MScM-SH vs. FD-SSCM	-0.90 ± 3.44	**2.01E-28**	7.03 ± 18.81	0.09 ± 4.65	**3.33E-09**	4.98 ± 10.63
EO-MScM-SH vs. QUBIC	-0.90 ± 3.44	**9.92E-13**	1.41 ± 3.94	0.09 ± 4.65	**1.28E-09**	13.72 ± 14.49
EO-MScM-SH vs. EO-COAL	-0.90 ± 3.44	**4.24E-08**	2.40 ± 7.33	0.09 ± 4.65	**1.56E-04**	3.94 ± 8.70
EO-MScM-SH vs. FD-COAL	-0.90 ± 3.44	**2.22E-08**	2.72 ± 7.30	0.09 ± 4.65	**0.001**	3.85 ± 8.83
EO-MScM-EL vs. EO-SSCM	-0.20 ± 4.56	**0.157**	0.71 ± 5.69	1.02 ± 6.11	**3.26E-04**	2.17 ± 5.38
EO-MScM-EL vs. FD-SSCM	-0.20 ± 4.56	**4.99E-22**	7.03 ± 18.81	1.02 ± 6.11	**2.59E-06**	4.98 ± 10.63
EO-MScM-EL vs. QUBIC	-0.20 ± 4.56	**4.60E-09**	1.41 ± 3.94	1.02 ± 6.11	**9.04E-09**	13.72 ± 14.49
EO-MScM-EL vs. EO-COAL	-0.20 ± 4.56	**1.25E-05**	2.40 ± 7.33	1.02 ± 6.11	**0.005**	3.94 ± 8.70
EO-MScM-EL vs. FD-COAL	-0.20 ± 4.56	**5.86E-06**	2.72 ± 7.30	1.02 ± 6.11	**0.022**	3.85 ± 8.83
EO-MScM-SH vs. MSISA-P	-0.90 ± 3.44	**0.295**	-1.12 ± 2.03	0.09 ± 4.65	**0.120**	0.46 ± 3.43
EO-MScM-SH vs. MSISA-R	-0.90 ± 3.44	**3.54E-13**	9.40 ± 9.19	0.09 ± 4.65	**2.25E-05**	2.34 ± 11.56
EO-MScM-SH vs. MSKM-SH	-0.90 ± 3.44	**0.915**	-1.18 ± 2.62	0.09 ± 4.65	**0.294**	-0.22 ± 2.96

EO-MScM-SH vs. MSKM-EL	-0.90 ± 3.44	**0.002**	0.19 ± 4.26	0.09 ± 4.65	**3.74E-09**	2.74 ± 5.66
EO-MScM-SH vs. BMSKM-SH	-0.90 ± 3.44	**0.620**	-1.09 ± 2.68	0.09 ± 4.65	**0.639**	-0.39 ± 2.87
EO-MScM-SH vs. BMSKM-EL	-0.90 ± 3.44	**8.72E-05**	0.44 ± 4.06	0.09 ± 4.65	**1.18E-11**	3.07 ± 5.44
EO-MScM-EL vs. MSISA-P	-0.20 ± 4.56	**0.701**	-1.12 ± 2.03	1.02 ± 6.11	**0.434**	0.46 ± 3.43
EO-MScM-EL vs. MSISA-R	-0.20 ± 4.56	**1.40E-11**	9.40 ± 9.19	1.02 ± 6.11	**1.57E-05**	2.34 ± 11.56
EO-MScM-EL vs. MSKM-SH	-0.20 ± 4.56	**0.353**	-1.18 ± 2.62	1.02 ± 6.11	**0.957**	-0.22 ± 2.96
EO-MScM-EL vs. MSKM-EL	-0.20 ± 4.56	**0.036**	0.19 ± 4.26	1.02 ± 6.11	**4.10E-06**	2.74 ± 5.66
EO-MScM-EL vs. BMSKM-SH	-0.20 ± 4.56	**0.663**	-1.09 ± 2.68	1.02 ± 6.11	**0.518**	-0.39 ± 2.87
EO-MScM-EL vs. BMSKM-EL	-0.20 ± 4.56	**0.006**	0.44 ± 4.06	1.02 ± 6.11	**6.89E-08**	3.07 ± 5.44

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**

EO-MScM-SH vs. EO-SSCM	-1.56 ± 3.27	**1.06E-05**	0.71 ± 5.69	0.29 ± 5.09	**4.46E-05**	2.04 ± 4.51
EO-MScM-SH vs. FD-SSCM	-1.56 ± 3.27	**3.94E-37**	7.03 ± 18.81	0.29 ± 5.09	**0.162**	0.46 ± 4.73
EO-MScM-SH vs. QUBIC	-1.56 ± 3.27	**4.26E-19**	1.41 ± 3.94	0.29 ± 5.09	**0.302**	9.57 ± 13.54
EO-MScM-SH vs. EO-COAL	-1.56 ± 3.27	**3.88E-13**	2.40 ± 7.33	0.29 ± 5.09	**0.001**	2.92 ± 6.80
EO-MScM-SH vs. FD-COAL	-1.56 ± 3.27	**3.68E-13**	2.72 ± 7.30	0.29 ± 5.09	**4.89E-04**	3.66 ± 7.60
EO-MScM-EL vs. EO-SSCM	-0.95 ± 4.01	**0.003**	0.71 ± 5.69	0.34 ± 5.08	**8.48E-05**	2.04 ± 4.51
EO-MScM-EL vs. FD-SSCM	-0.95 ± 4.01	**1.28E-30**	7.03 ± 18.81	0.34 ± 5.08	**0.121**	0.46 ± 4.73
EO-MScM-EL vs. QUBIC	-0.95 ± 4.01	**2.38E-14**	1.41 ± 3.94	0.34 ± 5.08	**0.269**	9.57 ± 13.54
EO-MScM-EL vs. EO-COAL	-0.95 ± 4.01	**2.27E-09**	2.40 ± 7.33	0.34 ± 5.08	**0.001**	2.92 ± 6.80
EO-MScM-EL vs. FD-COAL	-0.95 ± 4.01	**1.14E-09**	2.72 ± 7.30	0.34 ± 5.08	**0.001**	3.66 ± 7.60
EO-MScM-SH vs. MSISA-P	-1.56 ± 3.27	**0.023**	-2.63 ± 1.00	0.29 ± 5.09	**0.011**	-1.56 ± 1.40
EO-MScM-SH vs. MSISA-R	-1.56 ± 3.27	**6.06E-17**	10.37 ± 8.84	0.29 ± 5.09	**1.01E-09**	9.06 ± 7.74

EO-MScM-SH vs. MSKM-SH	-1.56 ± 3.27	**0.592**	-1.91 ± 1.56	0.29 ± 5.09	**0.617**	-0.83 ± 1.64
EO-MScM-SH vs. MSKM-EL	-1.56 ± 3.27	**5.02E-06**	0.52 ± 5.33	0.29 ± 5.09	**0.002**	0.36 ± 2.68
EO-MScM-SH vs. BMSKM-SH	-1.56 ± 3.27	**0.913**	-1.97 ± 1.58	0.29 ± 5.09	**0.284**	-0.89 ± 1.69
EO-MScM-SH vs. BMSKM-EL	-1.56 ± 3.27	**2.33E-06**	0.02 ± 3.81	0.29 ± 5.09	**0.002**	0.43 ± 3.17
EO-MScM-EL vs. MSISA-P	-0.95 ± 4.01	**0.002**	-2.63 ± 1.00	0.34 ± 5.08	**0.008**	-1.56 ± 1.40
EO-MScM-EL vs. MSISA-R	-0.95 ± 4.01	**1.89E-15**	10.37 ± 8.84	0.34 ± 5.08	**1.08E-09**	9.06 ± 7.74
EO-MScM-EL vs. MSKM-SH	-0.95 ± 4.01	**0.292**	-1.91 ± 1.56	0.34 ± 5.08	**0.459**	-0.83 ± 1.64
EO-MScM-EL vs. MSKM-EL	-0.95 ± 4.01	**0.001**	0.52 ± 5.33	0.34 ± 5.08	**0.004**	0.36 ± 2.68
EO-MScM-EL vs. BMSKM-SH	-0.95 ± 4.01	**0.143**	-1.97 ± 1.58	0.34 ± 5.08	**0.196**	-0.89 ± 1.69
EO-MScM-EL vs. BMSKM-EL	-0.95 ± 4.01	**0.001**	0.02 ± 3.81	0.34 ± 5.08	**0.004**	0.43 ± 3.17

B. anthracis - L. monocytogenes pairing

	B. anthracis	***L. monocytogenes***

dist1 vs. dist2	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)	dist1 mean (green)	Wilcoxon's 2-sided	dist2 mean (red)
EO-MScM-SH vs. EO-SSCM	-0.81 ± 3.42	**8.48E-12**	2.17 ± 5.38	0.50 ± 5.37	**1.24E-04**	2.04 ± 4.51
EO-MScM-SH vs. FD-SSCM	-0.81 ± 3.42	**1.14E-15**	5.21 ± 10.80	0.50 ± 5.37	**0.161**	0.46 ± 4.73
EO-MScM-SH vs. QUBIC	-0.81 ± 3.42	**3.82E-10**	13.72 ± 14.49	0.50 ± 5.37	**0.333**	9.57 ± 13.54
EO-MScM-SH vs. EO-COAL	-0.81 ± 3.42	**4.78E-07**	3.94 ± 8.70	0.50 ± 5.37	**0.002**	2.92 ± 6.80
EO-MScM-SH vs. FD-COAL	-0.81 ± 3.42	**8.93E-06**	3.85 ± 8.83	0.50 ± 5.37	**0.001**	3.66 ± 7.60
EO-MScM-EL vs. EO-SSCM	-0.42 ± 3.90	**1.18E-09**	2.17 ± 5.38	0.99 ± 6.39	**0.001**	2.04 ± 4.51
EO-MScM-EL vs. FD-SSCM	-0.42 ± 3.90	**2.52E-13**	5.21 ± 10.80	0.99 ± 6.39	**0.020**	0.46 ± 4.73
EO-MScM-EL vs. QUBIC	-0.42 ± 3.90	**1.15E-09**	13.72 ± 14.49	0.99 ± 6.39	**0.398**	9.57 ± 13.54
EO-MScM-EL vs. EO-COAL	-0.42 ± 3.90	**6.80E-06**	3.94 ± 8.70	0.99 ± 6.39	**0.012**	2.92 ± 6.80
EO-MScM-EL vs. FD-COAL	-0.42 ± 3.90	**8.02E-05**	3.85 ± 8.83	0.99 ± 6.39	**0.008**	3.66 ± 7.60
EO-MScM-SH vs. MSKM-SH	-0.81 ± 3.42	**0.595**	-1.58 ± 1.60	0.50 ± 5.37	**0.940**	-0.67 ± 1.69

EO-MScM-SH vs. MSKM-EL	-0.81 ± 3.42	**1.22E-09**	2.52 ± 6.60	0.50 ± 5.37	**0.009**	0.44 ± 3.00
EO-MScM-EL vs. MSKM-SH	-0.42 ± 3.90	**0.201**	-1.58 ± 1.60	0.99 ± 6.39	**0.271**	-0.67 ± 1.69
EO-MScM-EL vs. MSKM-EL	-0.42 ± 3.90	**5.61E-08**	2.52 ± 6.60	0.99 ± 6.39	**0.108**	0.44 ± 3.00

7.2.3.2.5 Sequence p-values

Table 7.39: Comparison of bicluster sequence p-values (-log10) from the expression only methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*. A comparison of the sequence p-values (-log10) from MScM (expression only) with all other relevant methods for all 3 pairings of the three organisms examined. In the comparisons, we compare both MScM steps to the other methods. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. As the table indicates, in 72 of the 92 of the comparisons (78.3%) MScM does as well or better than its competitors.

B. subtilis - B. anthracis pairing

	B. subtilis			***B. anthracis***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
EO-MScM-SH vs. EO-SSCM	4.11 ± 2.04	**0.026**	3.68 ± 1.88	3.77 ± 1.69	**4.87E-07**	2.92 ± 1.17
EO-MScM-SH vs. FD-SSCM	4.11 ± 2.04	**7.33E-18**	6.73 ± 3.35	3.77 ± 1.69	**0.097**	3.90 ± 2.62

EO-MScM-SH vs. QUBIC	4.11 ± 2.04	**7.49E-27**	2.06 ± 0.50	3.77 ± 1.69	**6.20E-31**	1.77 ± 0.26
EO-MScM-SH vs. EO-COAL	4.11 ± 2.04	**4.30E-21**	2.50 ± 1.30	3.77 ± 1.69	**1.25E-16**	2.36 ± 1.16
EO-MScM-SH vs. FD-COAL	4.11 ± 2.04	**1.07E-20**	2.47 ± 1.12	3.77 ± 1.69	**2.74E-18**	2.32 ± 1.57
EO-MScM-EL vs. EO-SSCM	3.86 ± 1.82	**0.235**	3.68 ± 1.88	3.55 ± 1.73	**7.07E-04**	2.92 ± 1.17
EO-MScM-EL vs. FD-SSCM	3.86 ± 1.82	**4.50E-21**	6.73 ± 3.35	3.55 ± 1.73	**0.831**	3.90 ± 2.62
EO-MScM-EL vs. QUBIC	3.86 ± 1.82	**1.00E-24**	2.06 ± 0.50	3.55 ± 1.73	**1.51E-24**	1.77 ± 0.26
EO-MScM-EL vs. EO-COAL	3.86 ± 1.82	**4.31E-18**	2.50 ± 1.30	3.55 ± 1.73	**5.34E-12**	2.36 ± 1.16
EO-MScM-EL vs. FD-COAL	3.86 ± 1.82	**1.79E-17**	2.47 ± 1.12	3.55 ± 1.73	**9.21E-14**	2.32 ± 1.57
EO-MScM-SH vs. MSISA-P	4.11 ± 2.04	**0.166**	3.65 ± 1.74	3.77 ± 1.69	**0.152**	3.34 ± 1.33
EO-MScM-SH vs. MSISA-R	4.11 ± 2.04	**1.18E-12**	2.02 ± 0.52	3.77 ± 1.69	**1.67E-06**	1.79 ± 0.27
EO-MScM-SH vs. MSKM-SH	4.11 ± 2.04	**0.742**	3.97 ± 1.81	3.77 ± 1.69	**0.442**	3.59 ± 1.53
EO-MScM-SH vs. MSKM-EL	4.11 ± 2.04	**4.81E-05**	3.24 ± 1.58	3.77 ± 1.69	**3.42E-10**	2.66 ± 1.03
EO-MScM-SH vs.	4.11 ± 2.04	**0.944**	4.05 ± 1.86	3.77 ± 1.69	**0.175**	3.42 ± 1.33

BMSKM-SH						
EO-MScM-SH vs. BMSKM-EL	4.11 ± 2.04	**9.38E-07**	3.06 ± 1.30	3.77 ± 1.69	**1.05E-11**	2.57 ± 0.88
EO-MScM-EL vs. MSISA-P	3.86 ± 1.82	**0.480**	3.65 ± 1.74	3.55 ± 1.73	**0.663**	3.34 ± 1.33
EO-MScM-EL vs. MSISA-R	3.86 ± 1.82	**1.66E-11**	2.02 ± 0.52	3.55 ± 1.73	**2.55E-05**	1.79 ± 0.27
EO-MScM-EL vs. MSKM-SH	3.86 ± 1.82	**0.421**	3.97 ± 1.81	3.55 ± 1.73	**0.502**	3.59 ± 1.53
EO-MScM-EL vs. MSKM-EL	3.86 ± 1.82	**0.002**	3.24 ± 1.58	3.55 ± 1.73	**2.00E-06**	2.66 ± 1.03
EO-MScM-EL vs. BMSKM-SH	3.86 ± 1.82	**0.240**	4.05 ± 1.86	3.55 ± 1.73	**0.891**	3.42 ± 1.33
EO-MScM-EL vs. BMSKM-EL	3.86 ± 1.82	**9.51E-05**	3.06 ± 1.30	3.55 ± 1.73	**1.27E-07**	2.57 ± 0.88

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
EO-MScM-SH vs. EO-SSCM	4.31 ± 1.90	**2.75E-04**	3.68 ± 1.88	5.07 ± 2.00	**2.69E-09**	3.62 ± 1.26
EO-MScM-SH vs. FD-SSCM	4.31 ± 1.90	**1.51E-14**	6.73 ± 3.35	5.07 ± 2.00	**0.692**	5.24 ± 2.35

EO-MScM-SH vs. QUBIC	4.31 ± 1.90	**5.23E-29**	2.06 ± 0.50	5.07 ± 2.00	**5.51E-28**	2.36 ± 0.37
EO-MScM-SH vs. EO-COAL	4.31 ± 1.90	**1.80E-24**	2.50 ± 1.30	5.07 ± 2.00	**6.16E-07**	3.74 ± 1.68
EO-MScM-SH vs. FD-COAL	4.31 ± 1.90	**1.28E-24**	2.47 ± 1.12	5.07 ± 2.00	**5.19E-09**	3.51 ± 1.51
EO-MScM-EL vs. EO-SSCM	3.99 ± 1.82	**0.050**	3.68 ± 1.88	5.08 ± 2.02	**3.38E-09**	3.62 ± 1.26
EO-MScM-EL vs. FD-SSCM	3.99 ± 1.82	**5.33E-19**	6.73 ± 3.35	5.08 ± 2.02	**0.722**	5.24 ± 2.35
EO-MScM-EL vs. QUBIC	3.99 ± 1.82	**9.61E-25**	2.06 ± 0.50	5.08 ± 2.02	**1.73E-27**	2.36 ± 0.37
EO-MScM-EL vs. EO-COAL	3.99 ± 1.82	**9.45E-20**	2.50 ± 1.30	5.08 ± 2.02	**7.06E-07**	3.74 ± 1.68
EO-MScM-EL vs. FD-COAL	3.99 ± 1.82	**2.81E-19**	2.47 ± 1.12	5.08 ± 2.02	**6.43E-09**	3.51 ± 1.51
EO-MScM-SH vs. MSISA-P	4.31 ± 1.90	**0.069**	5.06 ± 2.38	5.07 ± 2.00	**0.045**	5.77 ± 1.91
EO-MScM-SH vs. MSISA-R	4.31 ± 1.90	**6.55E-13**	1.99 ± 0.50	5.07 ± 2.00	**9.08E-15**	2.42 ± 0.56
EO-MScM-SH vs. MSKM-SH	4.31 ± 1.90	**0.021**	4.79 ± 1.75	5.07 ± 2.00	**0.015**	5.49 ± 1.73
EO-MScM-SH vs. MSKM-EL	4.31 ± 1.90	**3.99E-06**	3.45 ± 1.88	5.07 ± 2.00	**0.001**	4.35 ± 1.67
EO-MScM-SH vs.	4.31 ± 1.90	**0.263**	4.61 ± 2.13	5.07 ± 2.00	**0.965**	5.02 ± 1.71

BMSKM-SH						
EO-MScM-SH vs. BMSKM-EL	4.31 ± 1.90	**1.77E-08**	3.19 ± 1.39	5.07 ± 2.00	**0.001**	4.43 ± 1.62
EO-MScM-EL vs. MSISA-P	3.99 ± 1.82	**0.010**	5.06 ± 2.38	5.08 ± 2.02	**0.057**	5.77 ± 1.91
EO-MScM-EL vs. MSISA-R	3.99 ± 1.82	**9.88E-12**	1.99 ± 0.50	5.08 ± 2.02	**2.42E-14**	2.42 ± 0.56
EO-MScM-EL vs. MSKM-SH	3.99 ± 1.82	**1.04E-04**	4.79 ± 1.75	5.08 ± 2.02	**0.022**	5.49 ± 1.73
EO-MScM-EL vs. MSKM-EL	3.99 ± 1.82	**0.002**	3.45 ± 1.88	5.08 ± 2.02	**0.001**	4.35 ± 1.67
EO-MScM-EL vs. BMSKM-SH	3.99 ± 1.82	**0.009**	4.61 ± 2.13	5.08 ± 2.02	**0.979**	5.02 ± 1.71
EO-MScM-EL vs. BMSKM-EL	3.99 ± 1.82	**7.87E-05**	3.19 ± 1.39	5.08 ± 2.02	**0.001**	4.43 ± 1.62

B. anthracis - L. monocytogenes pairing

	B. anthracis			***L. monocytogenes***		
dist1 vs. dist2	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**	**dist1 mean (green)**	**Wilcoxon's 2-sided**	**dist2 mean (red)**
EO-MScM-SH vs. EO-SSCM	3.88 ± 1.80	**2.58E-07**	2.92 ± 1.17	5.02 ± 2.21	**5.62E-07**	3.62 ± 1.26
EO-MScM-SH vs. FD-SSCM	3.88 ± 1.80	**0.028**	3.83 ± 2.57	5.02 ± 2.21	**0.414**	5.24 ± 2.35

EO-MScM-SH vs. QUBIC	3.88 ± 1.80	**1.13E-27**	1.77 ± 0.26	5.02 ± 2.21	**2.09E-27**	2.36 ± 0.37
EO-MScM-SH vs. EO-COAL	3.88 ± 1.80	**6.71E-16**	2.36 ± 1.16	5.02 ± 2.21	**7.56E-06**	3.74 ± 1.68
EO-MScM-SH vs. FD-COAL	3.88 ± 1.80	**6.01E-17**	2.32 ± 1.57	5.02 ± 2.21	**2.04E-07**	3.51 ± 1.51
EO-MScM-EL vs. EO-SSCM	3.80 ± 1.80	**9.81E-06**	2.92 ± 1.17	4.90 ± 2.13	**1.37E-06**	3.62 ± 1.26
EO-MScM-EL vs. FD-SSCM	3.80 ± 1.80	**0.118**	3.83 ± 2.57	4.90 ± 2.13	**0.203**	5.24 ± 2.35
EO-MScM-EL vs. QUBIC	3.80 ± 1.80	**9.30E-27**	1.77 ± 0.26	4.90 ± 2.13	**2.78E-25**	2.36 ± 0.37
EO-MScM-EL vs. EO-COAL	3.80 ± 1.80	**2.42E-14**	2.36 ± 1.16	4.90 ± 2.13	**2.77E-05**	3.74 ± 1.68
EO-MScM-EL vs. FD-COAL	3.80 ± 1.80	**9.99E-16**	2.32 ± 1.57	4.90 ± 2.13	**9.20E-07**	3.51 ± 1.51
EO-MScM-SH vs. MSKM-SH	3.88 ± 1.80	**0.333**	3.94 ± 1.46	5.02 ± 2.21	**6.34E-04**	5.61 ± 1.74
EO-MScM-SH vs. MSKM-EL	3.88 ± 1.80	**7.27E-11**	2.61 ± 1.04	5.02 ± 2.21	**0.035**	4.37 ± 1.51
EO-MScM-EL vs. MSKM-SH	3.80 ± 1.80	**0.149**	3.94 ± 1.46	4.90 ± 2.13	**1.80E-04**	5.61 ± 1.74
EO-MScM-EL vs. MSKM-EL	3.80 ± 1.80	**3.18E-09**	2.61 ± 1.04	4.90 ± 2.13	**0.067**	4.37 ± 1.51

7.2.3.3 Comparisons with Randomized tests

In the comparisons below, we only show results from the Gram-positive triple. They were not generated for the Gram-negative triple as the initial results for the Gram-positive triple - displayed below - indicate that they are largely uninformative.

7.2.3.3.1 Residuals

Table 7.40: Comparison of bicluster residuals with randomized tests for all methods considered by this study for all pairings of *B. subtilis, B. anthracis* and *L. monocytogenes*. A comparison of the residuals of the results from all methods considered for all 3 pairings of the three organisms examined, where each method is compared with its equivalent randomized test. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. In this case, MSISA and Qubic always reported results worse than random, most likely due their identification of inversely correlated biclusters. Coalesce was worse than random for *B. subtilis* and *L. monocytogenes*, but better for *B. anthracis*.

B. subtilis - B. anthracis pairing

Method	*B. subtilis*			*B. anthracis*		
	derived mean (green)	Wilcoxon's 2-sided	shuffle mean (red)	derived mean (green)	Wilcoxon's 2-sided	shuffle mean (red)
FD-MScM-SH	0.51 ± 0.08	**3.61E-51**	0.59 ± 0.04	0.30 ± 0.09	**1.15E-99**	0.78 ± 0.06
FD-MScM-EL	0.49 ± 0.09	**7.10E-59**	0.59 ± 0.03	0.32 ± 0.09	**1.19E-99**	0.78 ± 0.05
FD-SSCM	0.49 ± 0.13	**6.97E-81**	0.58 ± 0.03	0.31 ± 0.12	**3.06E-202**	0.79 ± 0.05
MSKM-SH	0.41 ± 0.07	**7.34E-97**	0.56 ± 0.03	0.53 ± 0.12	**3.71E-86**	0.76 ± 0.05
MSKM-EL	0.42 ± 0.06	**1.51E-97**	0.57 ± 0.02	0.48 ± 0.11	**7.31E-98**	0.80 ± 0.03
MSISA-P	0.98 ± 0.39	**1.47E-10**	0.69 ± 0.15	1.97 ± 0.94	**3.74E-28**	0.76 ± 0.08

MSISA-R	1.11 ± 0.41	**3.92E-19**	0.66 ± 0.10	1.58 ± 0.38	**5.50E-26**	0.79 ± 0.03
EO-MScM-SH	0.52 ± 0.09	**8.04E-35**	0.58 ± 0.04	0.50 ± 0.20	**1.12E-57**	0.76 ± 0.07
EO-MScM-EL	0.52 ± 0.10	**1.25E-47**	0.58 ± 0.04	0.49 ± 0.20	**1.01E-66**	0.77 ± 0.06
EO-SSCM	0.44 ± 0.20	**9.78E-61**	0.58 ± 0.03	0.23 ± 0.06	**6.94E-139**	0.80 ± 0.04
QUBIC	0.87 ± 0.21	**7.37E-54**	0.63 ± 0.06	1.51 ± 0.29	**2.93E-97**	0.81 ± 0.02
EO-COAL	0.78 ± 0.23	**3.90E-27**	0.65 ± 0.11	0.58 ± 0.17	**3.14E-69**	0.80 ± 0.05
FD-COAL	0.80 ± 0.25	**2.63E-31**	0.65 ± 0.10	0.58 ± 0.17	**5.66E-66**	0.80 ± 0.04
BMSKM-SH	0.45 ± 0.07	**5.36E-88**	0.56 ± 0.03	0.38 ± 0.07	**2.41E-98**	0.77 ± 0.05
BMSKM-EL	0.45 ± 0.06	**5.82E-96**	0.57 ± 0.02	0.39 ± 0.07	**1.83E-98**	0.80 ± 0.02

B. subtilis - L. monocytogenes pairing

Method	***B. subtilis***			***L. monocytogenes***		
	derived mean (green)	**Wilcoxon's 2-sided**	**shuffle mean (red)**	**derived mean (green)**	**Wilcoxon's 2-sided**	**shuffle mean (red)**
FD-MScM-SH	0.52 ± 0.08	**1.56E-43**	0.58 ± 0.04	0.34 ± 0.12	**7.38E-93**	0.71 ± 0.08
FD-MScM-EL	0.50 ± 0.10	**7.54E-64**	0.59 ± 0.03	0.34 ± 0.12	**2.00E-93**	0.73 ± 0.08
FD-SSCM	0.49 ± 0.13	**6.97E-81**	0.58 ± 0.03	0.40 ± 0.18	**6.90E-158**	0.76 ± 0.08
MSKM-SH	0.40 ± 0.07	**3.81E-94**	0.55 ± 0.03	0.50 ± 0.12	**4.09E-62**	0.68 ± 0.09
MSKM-EL	0.42 ± 0.06	**2.01E-95**	0.57 ± 0.02	0.48 ± 0.11	**9.44E-88**	0.74 ± 0.07
MSISA-P	0.87 ± 0.34	**5.97E-05**	0.68 ± 0.20	1.59 ± 0.52	**4.72E-20**	0.71 ± 0.32
MSISA-R	1.11 ± 0.42	**2.40E-17**	0.66 ± 0.10	1.31 ± 0.34	**2.48E-21**	0.79 ± 0.09
EO-MScM-SH	0.52 ± 0.08	**4.19E-33**	0.57 ± 0.04	0.49 ± 0.17	**5.28E-50**	0.70 ± 0.10
EO-MScM-EL	0.50 ± 0.09	**4.34E-52**	0.58 ± 0.04	0.48 ± 0.17	**6.06E-54**	0.70 ± 0.10
EO-SSCM	0.44 ± 0.20	**9.78E-61**	0.58 ± 0.03	0.29 ± 0.10	**7.26E-56**	0.79 ± 0.05

QUBIC	0.87 ± 0.21	**7.37E-54**	0.63 ± 0.06	1.81 ± 0.85	**2.51E-24**	0.82 ± 0.03
EO-COAL	0.78 ± 0.23	**3.90E-27**	0.65 ± 0.11	1.63 ± 3.07	**7.52E-18**	0.79 ± 0.08
FD-COAL	0.80 ± 0.25	**2.63E-31**	0.65 ± 0.10	1.70 ± 3.24	**2.90E-19**	0.80 ± 0.07
BMSKM-SH	0.43 ± 0.07	**3.02E-85**	0.55 ± 0.03	0.42 ± 0.09	**1.73E-87**	0.68 ± 0.09
BMSKM-EL	0.44 ± 0.06	**9.03E-96**	0.57 ± 0.02	0.42 ± 0.09	**2.28E-95**	0.74 ± 0.06

B. anthracis - L. monocytogenes pairing

Method	***B. anthracis***			***L. monocytogenes***		
	derived mean (green)	**Wilcoxon's 2-sided**	**shuffle mean (red)**	**derived mean (green)**	**Wilcoxon's 2-sided**	**shuffle mean (red)**
FD-MScM-SH	0.33 ± 0.10	**2.75E-97**	0.73 ± 0.07	0.36 ± 0.14	**4.59E-88**	0.73 ± 0.08
FD-MScM-EL	0.36 ± 0.11	**7.87E-97**	0.75 ± 0.07	0.36 ± 0.13	**9.49E-91**	0.75 ± 0.08
FD-SSCM	0.31 ± 0.12	**1.02E-192**	0.79 ± 0.05	0.40 ± 0.18	**6.90E-158**	0.76 ± 0.08
MSKM-SH	0.40 ± 0.08	**2.42E-94**	0.70 ± 0.07	0.43 ± 0.08	**3.22E-89**	0.70 ± 0.08
MSKM-EL	0.39 ± 0.07	**1.68E-96**	0.78 ± 0.04	0.43 ± 0.08	**5.25E-95**	0.75 ± 0.06
EO-MScM-SH	0.52 ± 0.17	**1.08E-42**	0.71 ± 0.09	0.50 ± 0.18	**1.35E-41**	0.71 ± 0.10
EO-MScM-EL	0.50 ± 0.17	**8.12E-50**	0.72 ± 0.09	0.50 ± 0.19	**7.83E-45**	0.72 ± 0.10
EO-SSCM	0.23 ± 0.06	**6.94E-139**	0.80 ± 0.04	0.29 ± 0.10	**7.26E-56**	0.79 ± 0.05
QUBIC	1.51 ± 0.29	**2.93E-97**	0.81 ± 0.02	1.81 ± 0.85	**2.51E-24**	0.82 ± 0.03
EO-COAL	0.58 ± 0.17	**3.14E-69**	0.80 ± 0.05	1.63 ± 3.07	**7.52E-18**	0.79 ± 0.08
FD-COAL	0.58 ± 0.17	**5.66E-66**	0.80 ± 0.04	1.70 ± 3.24	**2.90E-19**	0.80 ± 0.07

7.2.3.3.2 Mean correlations

Table 7.41: Comparison of bicluster mean correlations with randomized tests for all methods considered by this study for all pairings of *B. subtilis*, *B. anthracis* and *L. monocytogenes*. A comparison of the mean correlations of the results from all methods considered for all 3 pairings of the three organisms examined, where each method is compared with its equivalent randomized test. Displayed are the means for each method and/or step compared, as well as the Wilcoxon's non-parametric rank test (2-sided) comparing their distributions. We direct the reader to section 7.2.3.1.1 for instructions on how to interpret the table. In nearly all comparisons, the method was significantly better than random. The sole exceptions were the biclusters produces by MSISA-R for *L. monocytogenes* from the pairing of *B. subtilis* and *L. monocytogenes*.

B. subtilis - B. anthracis pairing

Method	*B. subtilis*			*B. anthracis*		
	derived mean (green)	Wilcoxon's 2-sided	shuffle mean (red)	derived mean (green)	Wilcoxon's 2-sided	shuffle mean (red)
FD-MScM-SH	0.59 ± 0.11	**3.99E-99**	0.27 ± 0.04	0.85 ± 0.09	**1.28E-99**	0.40 ± 0.06
FD-MScM-EL	0.61 ± 0.11	**1.36E-99**	0.26 ± 0.04	0.84 ± 0.09	**1.10E-99**	0.39 ± 0.05
FD-SSCM	0.56 ± 0.14	**1.49E-190**	0.25 ± 0.04	0.82 ± 0.15	**1.36E-191**	0.37 ± 0.06
MSKM-SH	0.58 ± 0.11	**4.49E-98**	0.25 ± 0.04	0.52 ± 0.14	**1.18E-39**	0.39 ± 0.05
MSKM-EL	0.56 ± 0.11	**3.37E-98**	0.25 ± 0.03	0.58 ± 0.15	**9.14E-65**	0.37 ± 0.03
MSISA-P	0.60 ± 0.14	**2.43E-12**	0.44 ± 0.10	0.56 ± 0.07	**1.12E-07**	0.49 ± 0.08
MSISA-R	0.55 ± 0.13	**2.61E-10**	0.42 ± 0.08	0.51 ± 0.03	**6.32E-09**	0.47 ± 0.04
EO-MScM-SH	0.52 ± 0.12	**2.44E-90**	0.27 ± 0.05	0.69 ± 0.17	**1.29E-73**	0.40 ± 0.07
EO-MScM-EL	0.54 ± 0.12	**1.41E-93**	0.26 ± 0.05	0.69 ± 0.19	**1.80E-69**	0.40 ± 0.06
EO-SSCM	0.70 ± 0.11	**1.05E-106**	0.25 ± 0.04	0.91 ± 0.05	**6.91E-139**	0.37 ± 0.04
QUBIC	0.36 ± 0.19	**2.58E-10**	0.32 ± 0.04	0.49 ± 0.05	**8.79E-67**	0.41 ± 0.03

EO-COAL	0.58 ± 0.14	**1.28E-115**	0.37 ± 0.10	0.64 ± 0.13	**6.88E-84**	0.42 ± 0.06
FD-COAL	0.59 ± 0.15	**1.73E-99**	0.38 ± 0.10	0.62 ± 0.13	**3.19E-74**	0.41 ± 0.05
BMSKM-SH	0.49 ± 0.13	**3.89E-92**	0.25 ± 0.04	0.72 ± 0.10	**1.12E-97**	0.39 ± 0.05
BMSKM-EL	0.50 ± 0.12	**1.89E-97**	0.25 ± 0.03	0.71 ± 0.10	**2.07E-98**	0.37 ± 0.03

B. subtilis - L. monocytogenes pairing

	B. subtilis			***L. monocytogenes***		
Method	**derived mean (green)**	**Wilcoxon's 2-sided**	**shuffle mean (red)**	**derived mean (green)**	**Wilcoxon's 2-sided**	**shuffle mean (red)**
FD-MScM-SH	0.59 ± 0.11	**1.07E-95**	0.27 ± 0.05	0.80 ± 0.13	**1.06E-89**	0.44 ± 0.09
FD-MScM-EL	0.61 ± 0.10	**2.81E-97**	0.26 ± 0.04	0.81 ± 0.11	**1.05E-93**	0.43 ± 0.08
FD-SSCM	0.56 ± 0.14	**1.24E-190**	0.25 ± 0.04	0.71 ± 0.20	**2.05E-106**	0.42 ± 0.10
MSKM-SH	0.59 ± 0.11	**5.31E-95**	0.26 ± 0.05	0.51 ± 0.17	**4.57E-11**	0.42 ± 0.10
MSKM-EL	0.56 ± 0.11	**3.20E-96**	0.25 ± 0.03	0.55 ± 0.16	**4.76E-27**	0.42 ± 0.07
MSISA-P	0.60 ± 0.20	**4.42E-06**	0.44 ± 0.14	0.47 ± 0.23	**0.010**	0.55 ± 0.23
MSISA-R	0.55 ± 0.12	**3.47E-10**	0.42 ± 0.08	0.50 ± 0.27	**0.009**	0.51 ± 0.17
EO-MScM-SH	0.52 ± 0.13	**8.75E-85**	0.27 ± 0.06	0.64 ± 0.18	**1.24E-41**	0.44 ± 0.11
EO-MScM-EL	0.54 ± 0.12	**2.58E-91**	0.27 ± 0.06	0.64 ± 0.18	**1.52E-43**	0.43 ± 0.10
EO-SSCM	0.70 ± 0.11	**1.04E-106**	0.25 ± 0.04	0.86 ± 0.08	**6.49E-56**	0.43 ± 0.07
QUBIC	0.36 ± 0.19	**2.58E-10**	0.32 ± 0.04	0.45 ± 0.27	**4.59E-23**	0.45 ± 0.03
EO-COAL	0.58 ± 0.14	**1.28E-115**	0.37 ± 0.10	0.81 ± 0.13	**2.05E-39**	0.51 ± 0.10
FD-COAL	0.59 ± 0.15	**1.73E-99**	0.38 ± 0.10	0.80 ± 0.12	**3.70E-41**	0.50 ± 0.08
BMSKM-SH	0.52 ± 0.14	**2.18E-85**	0.26 ± 0.05	0.63 ± 0.15	**4.15E-54**	0.42 ± 0.10
BMSKM-EL	0.53 ± 0.12	**9.89E-96**	0.25 ± 0.03	0.64 ± 0.14	**3.92E-66**	0.42 ± 0.07

B. anthracis - L. monocytogenes pairing

Method	*B. anthracis* derived mean (green)	*B. anthracis* Wilcoxon's 2-sided	*B. anthracis* shuffle mean (red)	*L. monocytogenes* derived mean (green)	*L. monocytogenes* Wilcoxon's 2-sided	*L. monocytogenes* shuffle mean (red)
FD-MScM-SH	0.82 ± 0.11	**1.43E-97**	0.40 ± 0.07	0.77 ± 0.14	**2.14E-84**	0.44 ± 0.09
FD-MScM-EL	0.80 ± 0.11	**9.72E-98**	0.39 ± 0.06	0.78 ± 0.13	**5.76E-91**	0.43 ± 0.08
FD-SSCM	0.82 ± 0.15	**4.67E-182**	0.37 ± 0.05	0.71 ± 0.20	**2.23E-106**	0.42 ± 0.10
MSKM-SH	0.69 ± 0.12	**2.60E-88**	0.39 ± 0.07	0.60 ± 0.14	**2.13E-51**	0.42 ± 0.10
MSKM-EL	0.70 ± 0.10	**2.16E-95**	0.37 ± 0.03	0.63 ± 0.13	**2.53E-65**	0.42 ± 0.07
EO-MScM-SH	0.63 ± 0.16	**3.74E-56**	0.40 ± 0.08	0.63 ± 0.19	**7.13E-36**	0.43 ± 0.11
EO-MScM-EL	0.63 ± 0.17	**9.25E-57**	0.40 ± 0.08	0.63 ± 0.19	**7.82E-36**	0.43 ± 0.10
EO-SSCM	0.91 ± 0.05	**6.93E-139**	0.37 ± 0.04	0.86 ± 0.08	**6.28E-56**	0.43 ± 0.07
QUBIC	0.49 ± 0.05	**8.79E-67**	0.41 ± 0.03	0.45 ± 0.27	**4.59E-23**	0.45 ± 0.03
EO-COAL	0.64 ± 0.13	**6.88E-84**	0.42 ± 0.06	0.81 ± 0.13	**2.05E-39**	0.51 ± 0.10
FD-COAL	0.62 ± 0.13	**3.19E-74**	0.41 ± 0.05	0.80 ± 0.12	**3.70E-41**	0.50 ± 0.08

7.2.4 Additional GO term and KEGG pathway enrichment figures

GO term enrichments were initially introduced by Draghici et al (Draghici, Khatri et al. 2003) as a measure of the functional coherence of a set of genes. Effectively, GO term enrichments represent the probability, by chance, that a set of genes share the same functional annotation, which is approximated using the hypergeometric distribution:

$$P\left(b_k \mid G,T\right) = \frac{\binom{|T|}{|b_k \cap T|}\binom{|G|-|T|}{|b_k|-|b_k \cap T|}}{\binom{|G|}{|b_k|}}$$

where b_k is the set of genes in bicluster k; G is the set of genes in the genome ; and T is the set of genes having a particular GO term annotation. Similarly, KEGG pathway enrichments were approximated with a hypergeometric distribution, where T was instead the set of genes associated with a given KEGG pathway.

For all the pairings between *B. subtilis*, *B. anthracis* and *L. monocytogenes*, there is a consistent increase from the shared to elaboration steps of all the multi-species methods, with the percentage of FD-MScM-EL biclusters with significant GO term enrichments consistently greater than the SSCM results. Similar behavior is observed with the KEGG pathway enrichments. The higher percentages reported for the MSISA and Qubic methods are a reflection of the high redundancy of the biclusters identified by them.

Below, we show plots of the GO term and K pathway enrichments, where in panel**(A) GO Terms** are displayed the percentage of biclusters with enriched GO terms. In **(B) KEGG Pathways**, the percentage of biclusters with enriched KEGG pathways are displayed. Explanations of the method name abbreviations can be found in.

7.2.4.1 Figures for the Gram-positive triplet

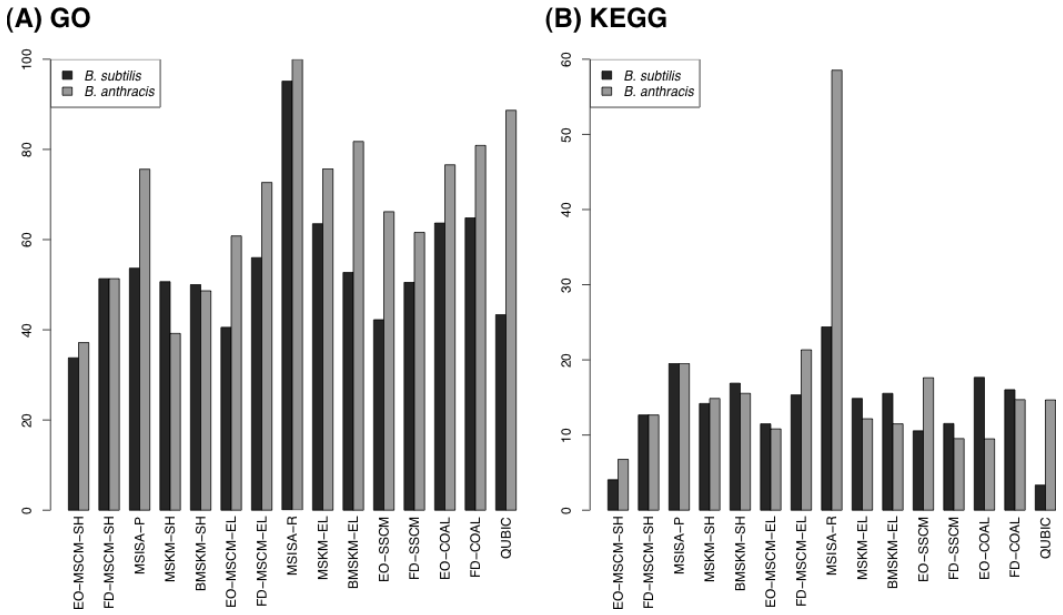


Figure 7.79: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments from all methods considered by this study for the *B. subtilis – B. anthracis* pairing.

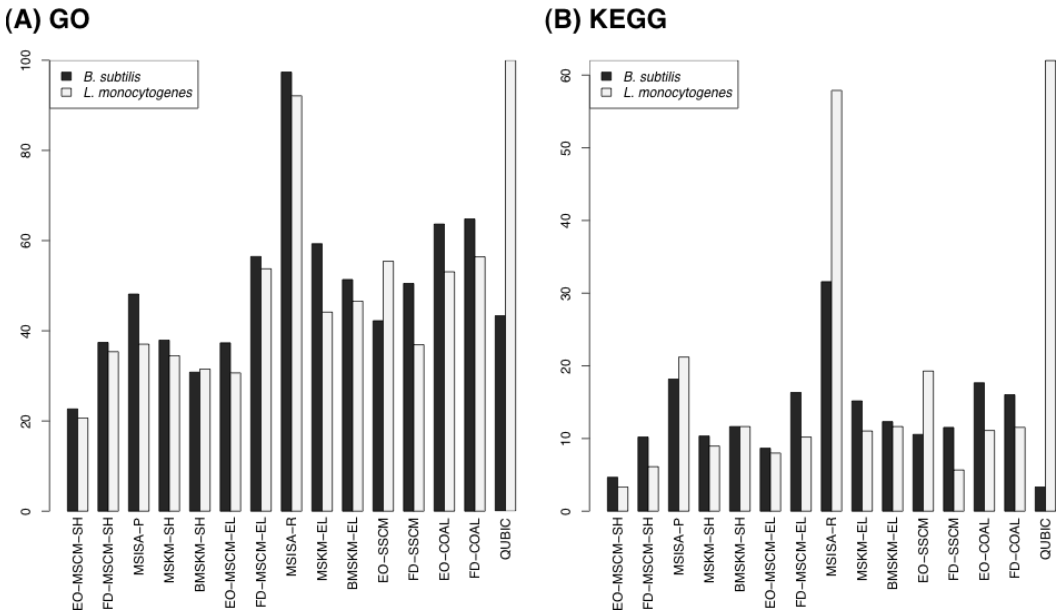


Figure 7.80: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments for the multi-species cMonkey, multi-species k-means and single-species cMonkey methods for the *B. subtilis – L. monocytogenes* pairing.

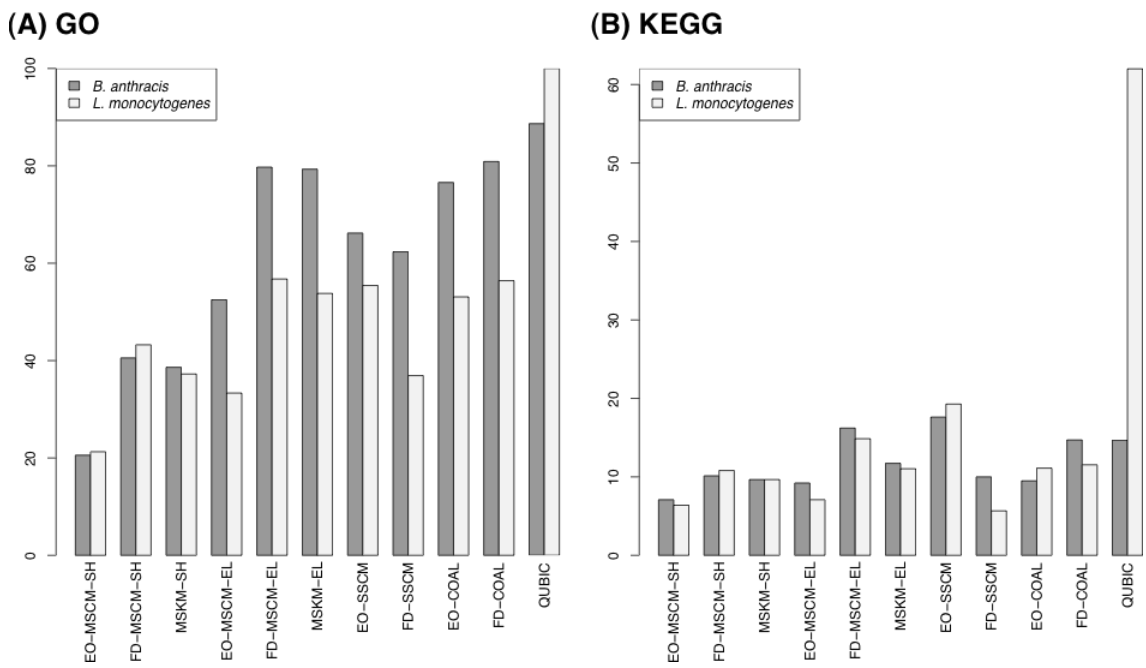


Figure 7.81: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments for the multi-species cMonkey, multi-species k-means and single-species cMonkey methods for the *B. anthracis – L. monocytogenes* pairing. (A) GO Terms.

7.2.4.2 Figures for the Gram-negative triplet

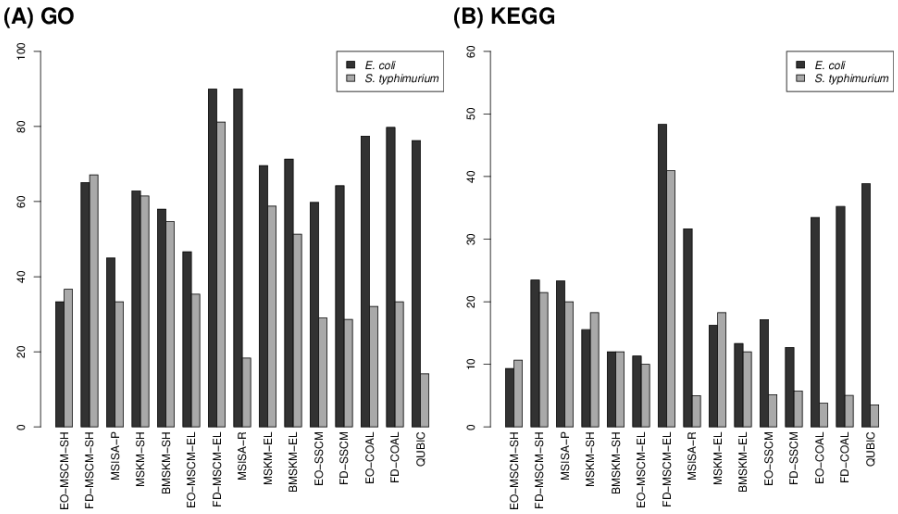


Figure 7.82: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments from all methods considered by this study for the *E. coli – S. typhimurium* pairing.

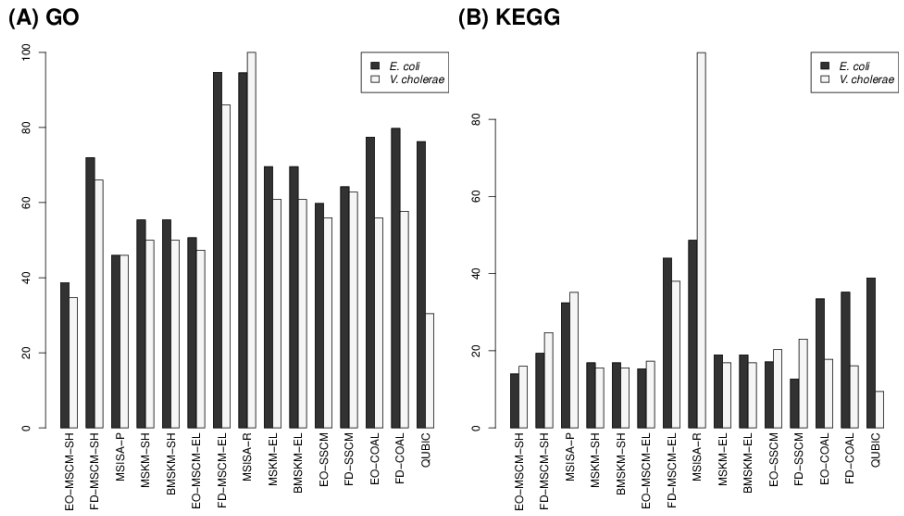


Figure 7.83: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments from all methods considered by this study for the *E. coli – V. cholerae* pairing.

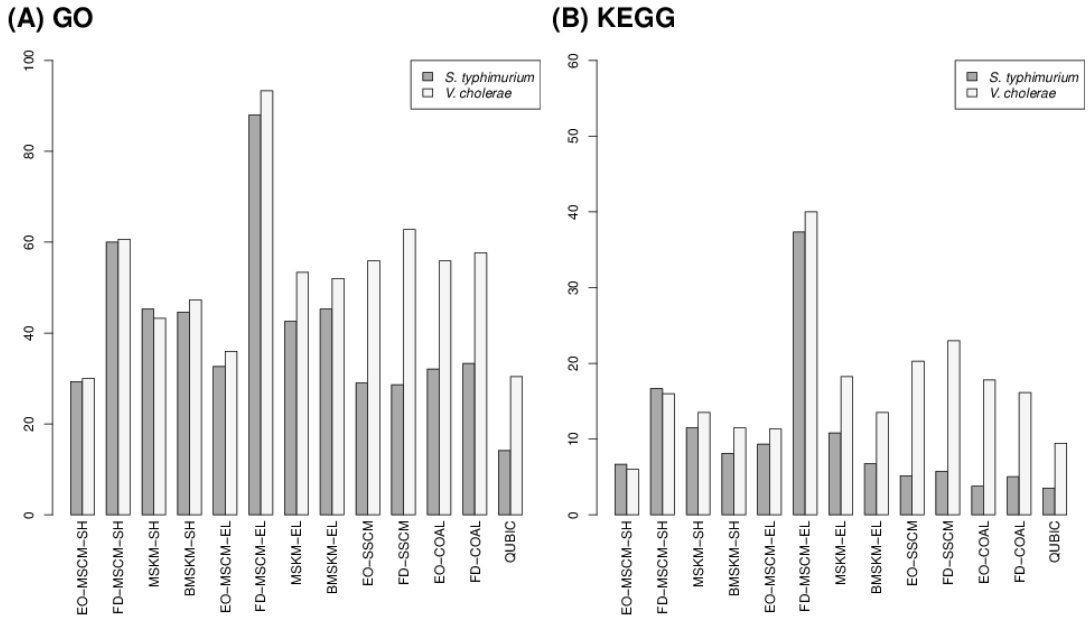


Figure 7.84: Comparison of the fraction of biclusters with significant GO and KEGG annotation enrichments from all methods considered by this study for the *S. typhimurium – V. cholerae* pairing.

7.3 Gene lists and bicluster images for biological highlights from the human and mouse immune system cell data analysis

7.3.1 Full descriptions of highlighted biclusters

7.3.1.1 Bicluster 31 (Human and Mouse listed together)

7.3.1.1.1 Shared Bicluster

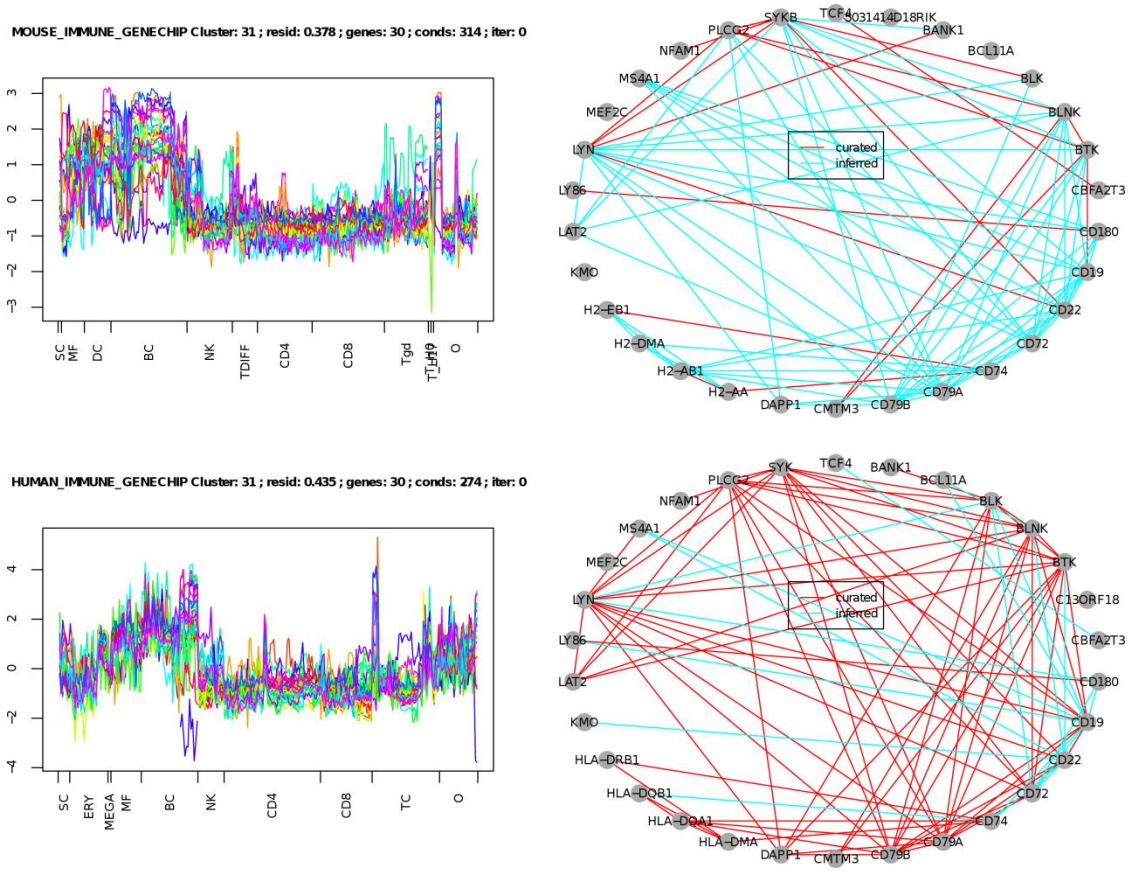


Figure 7.85: Shared Human-Mouse bicluster 32 image (pre-elaboration)

7.3.1.1.2 Gene list

Table 7.42: Human-Mouse Immune System bicluster 32 (Conserved Bicluster)

Mouse genes	Human genes
KMO	KMO
BCL11A	BCL11A
BANK1	BANK1
DAPP1	DAPP1
CD180	CD180
MEF2C	MEF2C
CD74	CD74
LY86	LY86
HLA-DMA	H2-DMA
HLA-DQA1	H2-AA
HLA-DQB1	H2-AB1
HLA-DRB1	H2-EB1
LAT2	LAT2
BLK	BLK
LYN	LYN
CD72	CD72
SYK	SYKB
BLNK	BLNK
MS4A1	MS4A1
C13ORF18	5031414D18RIK
CD19	CD19
CMTM3	CMTM3

CBFA2T3	CBFA2T3
PLCG2	PLCG2
CD79B	CD79B
TCF4	TCF4
CD22	CD22
CD79A	CD79A
NFAM1	NFAM1
BTK	BTK

7.3.1.2 Bicluster 87 (Human and Mouse listed together)

7.3.1.2.1 Shared Bicluster

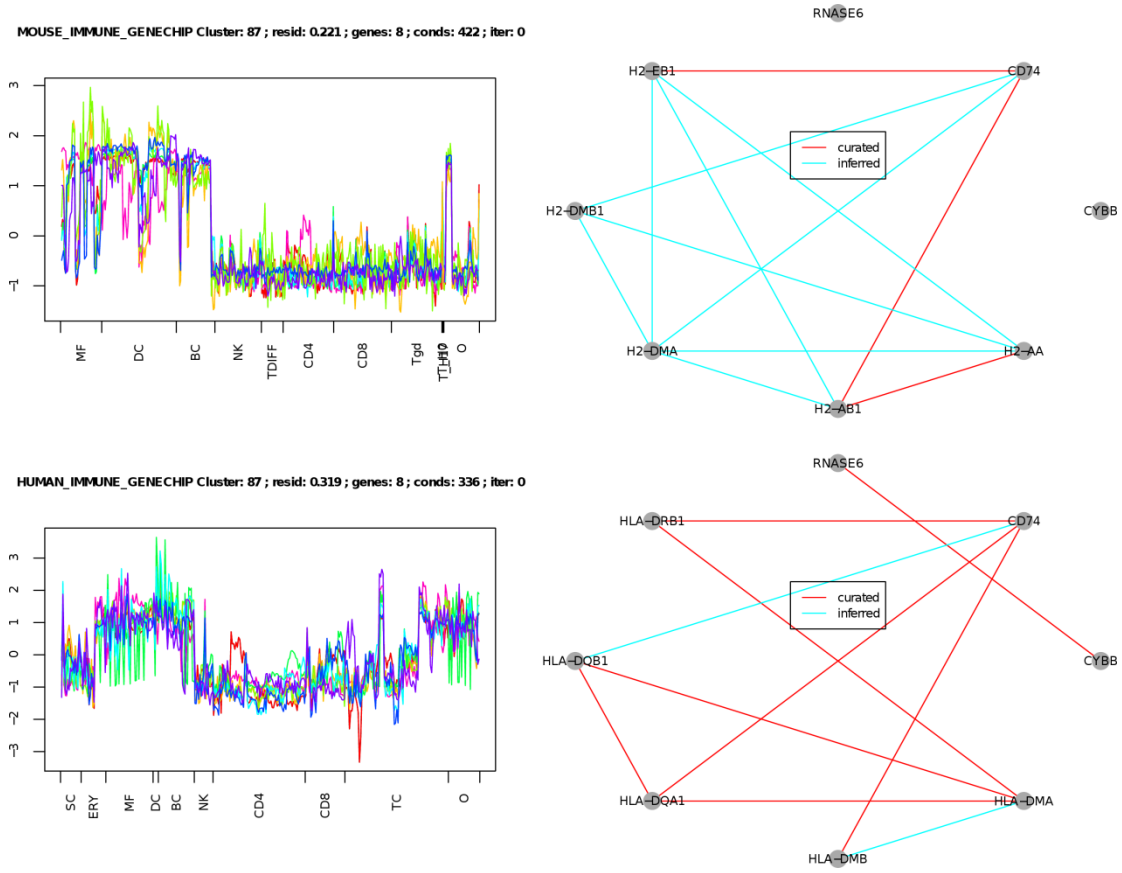


Figure 7.86: Shared Human-Mouse bicluster 87 image (pre-elaboration)

7.3.1.2.2 Gene list

Table 7.43: Human-Mouse Immune System bicluster 87 (Conserved Bicluster)

Mouse genes	Human genes
CD74	CD74
HLA-DMA	H2-DMA
HLA-DMB	H2-DMB1

HLA-DQA1	H2-AA
HLA-DQB1	H2-AB1
HLA-DRB1	H2-EB1
RNASE6	RNASE6
CYBB	CYBB

7.3.1.3 Bicluster 2 (Human and Mouse listed together)

7.3.1.3.1 Shared Bicluster

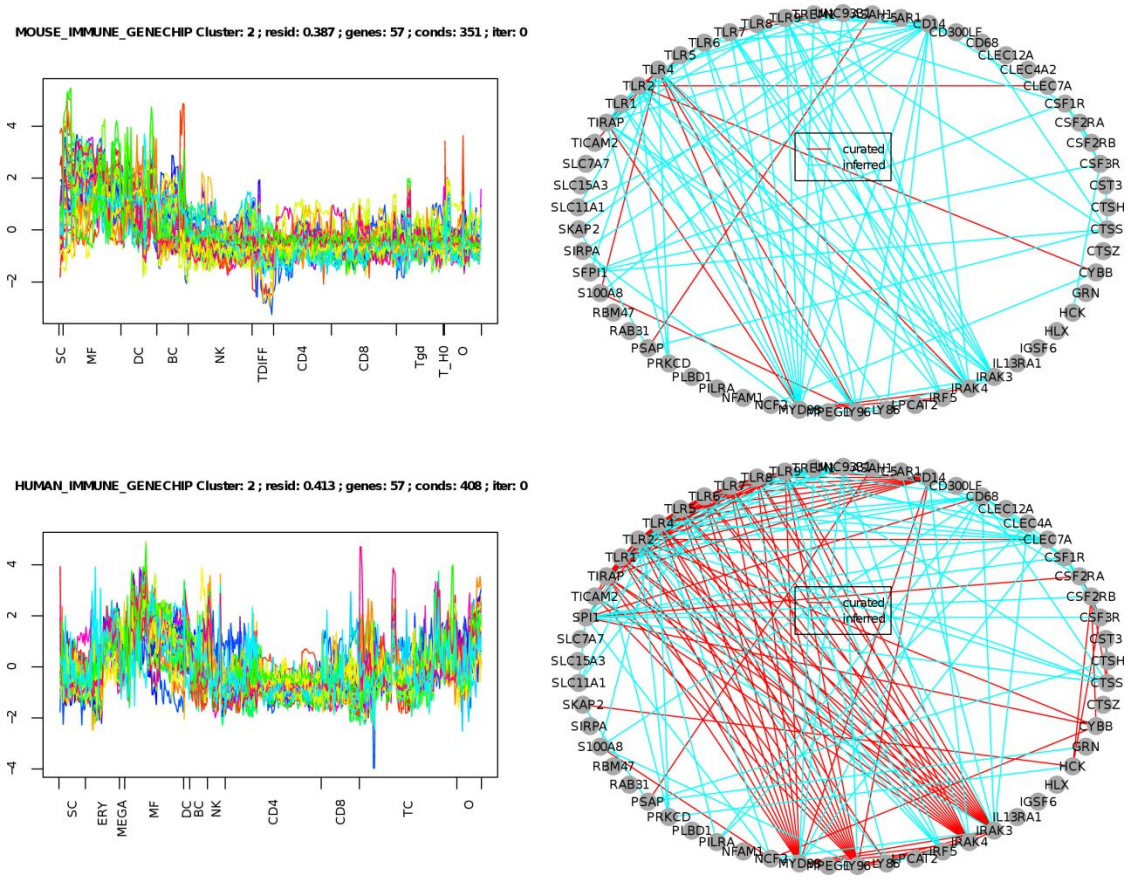


Figure 7.87: Shared Human-Mouse bicluster 2 image (pre-elaboration)

7.3.1.3.2 Gene list

Table 7.44: Human-Mouse Immune System bicluster 2 (Conserved Bicluster)

Mouse genes	Human genes
CSF3R	CSF3R
CTSS	CTSS
S100A8	S100A8
NCF2	NCF2
HLX	HLX
TLR5	TLR5
SLC11A1	SLC11A1
MYD88	MYD88
PRKCD	PRKCD
TLR9	TLR9
RBM47	RBM47
TLR1	TLR1
TLR6	TLR6
TLR2	TLR2
TICAM2	TICAM2
CD14	CD14
CSF1R	CSF1R
LY86	LY86
TREM1	TREM1
SKAP2	SKAP2

PILRA	PILRA
IRF5	IRF5
ASAH1	ASAH1
LY96	LY96
TLR4	TLR4
PSAP	PSAP
SPI1	SFPI1
MPEG1	MPEG1
SLC15A3	SLC15A3
UNC93B1	UNC93B1
TIRAP	TIRAP
CLEC12A	CLEC12A
CLEC7A	CLEC7A
PLBD1	PLBD1
CLEC4A	CLEC4A2
IRAK4	IRAK4
IRAK3	IRAK3
SLC7A7	SLC7A7
CTSH	CTSH
IGSF6	IGSF6
LPCAT2	LPCAT2
CD68	CD68
GRN	GRN
CD300LF	CD300LF
RAB31	RAB31

C5AR1	C5AR1
SIRPA	SIRPA
CST3	CST3
HCK	HCK
CTSZ	CTSZ
CSF2RB	CSF2RB
NFAM1	NFAM1
TLR7	TLR7
TLR8	TLR8
CYBB	CYBB
IL13RA1	IL13RA1
CSF2RA	CSF2RA

7.3.1.4 Bicluster 480 (Human and Mouse listed together)

7.3.1.4.1 Shared Bicluster

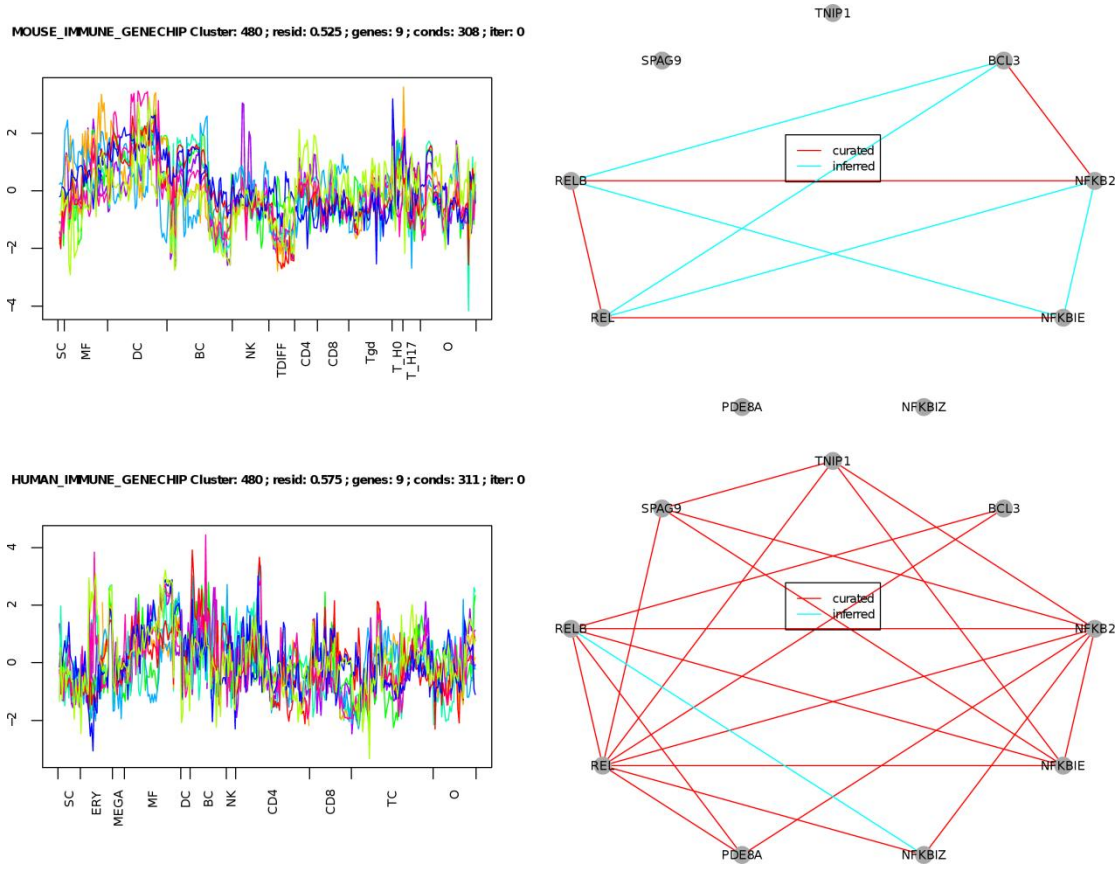


Figure 7.88: Shared Human-Mouse bicluster 482 image (pre-elaboration)

7.3.1.4.2 Gene list

Table 7.45: Human-Mouse Immune System bicluster 480 (Conserved Bicluster)

Mouse genes	Human genes
REL	REL
NFKBIZ	NFKBIZ
TNIP1	TNIP1

NFKBIE	NFKBIE
NFKB2	NFKB2
PDE8A	PDE8A
SPAG9	SPAG9
BCL3	BCL3
RELB	RELB

7.4 References

Draghici, S., P. Khatri, et al. (2003). "Global functional profiling of gene expression." Genomics **81**(2): 98-104.

Henriques, A. O. and C. P. Moran, Jr. (2007). "Structure, assembly, and function of the spore surface layers." Annual review of microbiology **61**: 555-588.

8. BIBLIOGRAPHY

Abbas, A. R., K. Wolslegel, et al. (2009). "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus." PLoS ONE **4**(7): e6098.

Aggarwal, K. and K. H. Lee (2003). "Functional genomics and proteomics as a foundation for systems biology." Brief Funct Genomic Proteomic **2**(3): 175-184.

Alexeyenko, A., I. Tamas, et al. (2006). "Automatic clustering of orthologs and inparalogs shared by multiple proteomes." Bioinformatics **22**(14): e9-15.

Allenby, N. E., N. O'Connor, et al. (2005). "Genome-wide transcriptional analysis of the phosphate starvation stimulon of Bacillus subtilis." J Bacteriol **187**(23): 8063-8080.

Alm, E. J., K. H. Huang, et al. (2005). "The MicrobesOnline Web site for comparative genomics." Genome Res **15**(7): 1015-1022.

Alon, U. (2007). An introduction to systems biology : design principles of biological circuits. Boca Raton, FL, Chapman & Hall/CRC.

Aloy, P., A. Stark, et al. (2003). "Predictions without templates: new folds, secondary structure, and contacts in CASP5." Proteins **53 Suppl 6**: 436-456.

Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proc Natl Acad Sci U S A **97**(18): 10101-10106.

Alter, O., P. O. Brown, et al. (2003). "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms." Proc Natl Acad Sci U S A **100**(6): 3351-3356.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Aranda, B., P. Achuthan, et al. (2010). "The IntAct molecular interaction database in 2010." Nucleic Acids Res **38**(Database issue): D525-531.

Arends, S. J. and D. S. Weiss (2004). "Inhibiting cell division in Escherichia coli has little if any effect on gene expression." J Bacteriol **186**(3): 880-884.

Arkin, A., J. Ross, et al. (1998). "Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells." Genetics **149**(4): 1633-1648.

Asai, K., H. Yamaguchi, et al. (2003). "DNA microarray analysis of Bacillus subtilis sigma factors of extracytoplasmic function family." FEMS Microbiol Lett **220**(1): 155-160.

Ausmees, N. and C. Jacobs-Wagner (2003). "Spatial and temporal control of differentiation and cell cycle progression in Caulobacter crescentus." Annu Rev Microbiol **57**: 225-247.

Avila-Campillo, I., K. Drew, et al. (2007). "BioNetBuilder: automatic integration of biological networks." Bioinformatics **23**(3): 392-393.

Babu, M., G. Musso, et al. (2009). "Systems-level approaches for identifying and analyzing genetic interaction networks in Escherichia coli and extensions to other prokaryotes." Mol Biosyst **5**(12): 1439-1455.

Bader, G. D., I. Donaldson, et al. (2001). "BIND--The Biomolecular Interaction Network Database." Nucleic acids research **29**(1): 242-245.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bailey, T. L. and M. Gribskov (1998). "Combining evidence using p-values: application to sequence homology searches." Bioinformatics **14**(1): 48-54.

Balasubramanian, R., T. LaFramboise, et al. (2004). "A graph-theoretic approach to testing associations between disparate sources of functional genomics data." Bioinformatics **20**(18): 3353-3362.

Baliga, N. S., S. J. Bjork, et al. (2004). "Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1." Genome Res **14**(6): 1025-1035.

Baliga, N. S., M. Pan, et al. (2002). "Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach." Proc Natl Acad Sci U S A **99**(23): 14913-14918.

Bar-Joseph, Z., G. K. Gerber, et al. (2003). "Computational discovery of gene modules and regulatory networks." Nat Biotechnol **21**(11): 1337-1342.

Barbosa, T. M. and S. B. Levy (2000). "Differential expression of over 60 chromosomal genes in Escherichia coli by constitutive expression of MarA." J Bacteriol **182**(12): 3467-3474.

Bare, J. C., P. T. Shannon, et al. (2007). "The Firegoose: two-way integration of diverse data from different bioinformatics web resources with desktop applications." BMC Bioinformatics **8**: 456.

Barrett, C. L., C. D. Herring, et al. (2005). "The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states." Proc Natl Acad Sci U S A **102**(52): 19103-19108.

Barrett, T. and R. Edgar (2006). "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis." Methods Enzymol **411**: 352-369.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: mining tens of millions of expression profiles--database and tools update." Nucleic Acids Res **35**(Database issue): D760-765.

Barrett, T., D. B. Troup, et al. (2007). "NCBI GEO: Mining tens of millions of expression profiles - Database and tools update." Nucleic Acids Research **35**(SUPPL. 1): D760-D765.

Battistuzzi, F. U., A. Feijao, et al. (2004). "A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land." BMC Evol Biol **4**: 44.

Beard, D. A., S. D. Liang, et al. (2002). "Energy balance for analysis of complex metabolic networks." Biophys J **83**(1): 79-86.

Ben-Dor, A., B. Chor, et al. (2003). "Discovering local structure in gene expression data: the order-preserving submatrix problem." J Comput Biol **10**(3-4): 373-384.

Ben-Yehuda, S., M. Fujita, et al. (2005). "Defining a centromere-like element in Bacillus subtilis by Identifying the binding sites for the chromosome-anchoring protein RacA." Mol Cell **17**(6): 773-782.

Berg, J. and M. Lassig (2006). "Cross-species analysis of biological networks by Bayesian alignment." Proc Natl Acad Sci U S A **103**(29): 10967-10972.

Bergman, N. H., E. C. Anderson, et al. (2006). "Transcriptional profiling of the Bacillus anthracis life cycle in vitro and an implied model for regulation of spore formation." J Bacteriol **188**(17): 6092-6100.

Bergmann, S., J. Ihmels, et al. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." Phys Rev E Stat Nonlin Soft Matter Phys **67**(3 Pt 1): 031902.

Bergmann, S., J. Ihmels, et al. (2004). "Similarities and differences in genome-wide expression data of six organisms." PLoS Biol **2**(1): E9.

Berka, R. M., J. Hahn, et al. (2002). "Microarray analysis of the Bacillus subtilis K-state: genome-wide expression changes dependent on ComK." Mol Microbiol **43**(5): 1331-1345.

Bigot, A., H. Pagniez, et al. (2005). "Role of FliF and FliI of Listeria monocytogenes in Flagellar Assembly and Pathogenicity." Infect. Immun. **73**(9): 5530-5539.

Biondi, E. G., S. J. Reisinger, et al. (2006). "Regulation of the bacterial cell cycle by an integrated genetic circuit." Nature **444**(7121): 899-904.

Biondi, E. G., J. M. Skerker, et al. (2006). "A phosphorelay system controls stalk biogenesis during cell cycle progression in Caulobacter crescentus." Mol Microbiol **59**(2): 386-401.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Blair, K. M., L. Turner, et al. (2008). "A molecular clutch disables flagella in the Bacillus subtilis biofilm." Science **320**(5883): 1636-1638.

Blake, J. A., C. J. Bult, et al. (2011). "The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics." Nucleic Acids Research **39**(Database issue): D842-848.

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-1474.

Blokesch, M. and G. K. Schoolnik (2007). "Serogroup conversion of Vibrio cholerae in aquatic reservoirs." PLoS Pathog **3**(6): e81.

Bogomolni, R. A. and J. L. Spudich (1982). "Identification of a third rhodopsin-like pigment in phototactic Halobacterium halobium." Proc Natl Acad Sci U S A **79**(20): 6250-6254.

Bon, E., S. Casaregola, et al. (2003). "Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns." Nucleic Acids Res **31**(4): 1121-1135.

Bonneau, R. (2006). The Inferelator Cytoscape Web Start.

Bonneau, R., N. S. Baliga, et al. (2004). "Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1." Genome Biol **5**(8): R52.

Bonneau, R., M. T. Facciotti, et al. (2007). "A predictive model for transcriptional control of physiology in a free living cell." Cell **131**(7): 1354-1365.

Bonneau, R., D. J. Reiss, et al. (2006). "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo." Genome Biol **7**(5): R36.

Bonneau, R., C. E. Strauss, et al. (2001). "Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation." Proteins **43**(1): 1-11.

Bonneau, R., J. Tsai, et al. (2001). "Rosetta in CASP4: progress in ab initio protein structure prediction." Proteins **Suppl 5**: 119-126.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome Biol **5**(5): R35.

Bowman, J. P., C. R. Bittencourt, et al. (2008). "Differential gene expression of Listeria monocytogenes during high hydrostatic pressure processing." Microbiology **154**(Pt 2): 462-475.

Bradley, P., D. Chivian, et al. (2003). "Rosetta predictions in CASP5: successes, failures, and prospects for complete automation." Proteins **53 Suppl 6**: 457-468.

Bray, D. (1995). "Protein molecules as computational elements in living cells." Nature **376**(6538): 307-312.

Brazma, A., H. Parkinson, et al. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res **31**(1): 68-71.

Breitkreutz, B. J., C. Stark, et al. (2008). "The BioGRID Interaction Database: 2008 update." Nucleic acids research **36**(Database issue): D637-640.

Britton, R. A., P. Eichenberger, et al. (2002). "Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of Bacillus subtilis." J Bacteriol **184**(17): 4881-4890.

Brocklehurst, K. R. and A. P. Morby (2000). "Metal-ion tolerance in Escherichia coli: analysis of transcriptional profiles by gene-array technology." Microbiology **146 (Pt 9)**: 2277-2282.

Brooun, A., J. Bell, et al. (1998). "An archaeal aerotaxis transducer combines subunit I core structures of eukaryotic cytochrome c oxidase and eubacterial methyl-accepting chemotaxis proteins." J Bacteriol **180**(7): 1642-1646.

Brown, K. R. and I. Jurisica (2005). "Online predicted human interaction database." Bioinformatics **21**(9): 2076-2082.

Bult, C. J., J. T. Eppig, et al. (2008). "The Mouse Genome Database (MGD): mouse biology and model systems." Nucleic Acids Res **36**(Database issue): D724-728.

Bunai, K., M. Ariga, et al. (2004). "Profiling and comprehensive expression analysis of ABC transporter solute-binding proteins of Bacillus subtilis membrane based on a proteomic approach." Electrophoresis **25**(1): 141-155.

Buttner, K., J. Bernhardt, et al. (2001). "A comprehensive two-dimensional map of cytosolic proteins of Bacillus subtilis." Electrophoresis **22**(14): 2908-2935.

Buzzeo, M. P., J. Yang, et al. (2007). "Hematopoietic stem cell mobilization with G-CSF induces innate inflammation yet suppresses adaptive immune gene expression as revealed by microarray analysis." Experimental hematology **35**(9): 1456-1465.

Celniker, S. E., L. A. Dillon, et al. (2009). "Unlocking the secrets of the genome." Nature **459**(7249): 927-930.

Ceol, A., A. Chatr Aryamontri, et al. (2010). "MINT, the molecular interaction database: 2009 update." Nucleic Acids Res **38**(Database issue): D532-539.

Cerami, E. G., B. E. Gross, et al. (2011). "Pathway Commons, a web resource for biological pathway data." Nucleic Acids Res **39**(Database issue): D685-690.

Chada, V. G., E. A. Sanstad, et al. (2003). "Morphogenesis of bacillus spore surfaces." J Bacteriol **185**(21): 6255-6261.

Chan, K., S. Baker, et al. (2003). "Genomic Comparison of Salmonella enterica Serovars and Salmonella bongori by Use of an S. enterica Serovar Typhimurium DNA Microarray." J. Bacteriol. **185**(2): 553-563.

Chan, K., C. C. Kim, et al. (2005). "Microarray-Based Detection of Salmonella enterica Serovar Typhimurium Transposon Mutants That Cannot Survive in Macrophages and Mice." Infect. Immun. **73**(9): 5438-5449.

Chandrasekaran, S. and N. D. Price (2010). "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis." Proc Natl Acad Sci U S A **107**(41): 17845-17850.

Chatr-aryamontri, A., A. Ceol, et al. (2007). "MINT: the Molecular INTeraction database." Nucleic Acids Res **35**(Database issue): D572-574.

Chautard, E., L. Ballut, et al. (2009). "MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions." Bioinformatics **25**(5): 690-691.

Chen, F., A. J. Mackey, et al. (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." Nucleic Acids Research **34**(suppl 1): D363-D368.

Chen, F., A. J. Mackey, et al. (2007). "Assessing performance of orthology detection strategies applied to eukaryotic genomes." PLoS ONE **2**(4): e383.

Cheng, Y. and G. M. Church (2000). "Biclustering of expression data." Proc Int Conf Intell Syst Mol Biol **8**: 93-103.

Chikina, M. D. and O. G. Troyanskaya (2011). "Accurate quantification of functional analogy among close homologs." PLoS Computational Biology **7**(2): e1001074.

Chivian, D., D. E. Kim, et al. (2003). "Automated prediction of CASP-5 structures using the Robetta server." Proteins **53 Suppl 6**: 524-533.

Cline, M. S., M. Smoot, et al. (2007). "Integration of biological networks and gene expression data using Cytoscape." Nat Protoc **2**(10): 2366-2382.

Courcelle, J., A. Khodursky, et al. (2001). "Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli." Genetics **158**(1): 41-64.

Covert, M. W., E. M. Knight, et al. (2004). "Integrating high-throughput and computational data elucidates bacterial networks." Nature **429**(6987): 92-96.

Covert, M. W. and B. O. Palsson (2002). "Transcriptional regulation in constraints-based metabolic models of Escherichia coli." J Biol Chem **277**(31): 28058-28064.

Covert, M. W., C. H. Schilling, et al. (2001). "Regulation of gene expression in flux balance models of metabolism." J Theor Biol **213**(1): 73-88.

Croft, D., G. O'Kelly, et al. (2011). "Reactome: a database of reactions, pathways and biological processes." Nucleic Acids Res **39**(Database issue): D691-697.

Csardi, G. (2010). "isa2: The Iterative Signature Algorithm. R package version 0.2.1." from http://cran.r-project.org/web/packages/isa2/index.html.

D'Haeseleer, P., S. Liang, et al. (2000). "Genetic network inference: from co-expression clustering to reverse engineering." Bioinformatics **16**(8): 707-726.

D'Haeseleer, P., X. Wen, et al. (1999). "Linear modeling of mRNA expression levels during CNS development and injury." Pac Symp Biocomput: 41-52.

Dai, M., P. Wang, et al. (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." Nucleic Acids Research **33**(20): e175.171-e175.179.

Das, D., N. Banerjee, et al. (2004). "Interacting models of cooperative gene regulation." Proc Natl Acad Sci U S A **101**(46): 16234-16239.

DasSarma, S. (1993). "Identification and analysis of the gas vesicle gene cluster on an unstable plasmid of Halobacterium halobium." Experientia **49**(6-7): 482-486.

Dassarma, S., B. R. Berquist, et al. (2006). "Post-genomics of the model haloarchaeon Halobacterium sp. NRC-1." Saline Systems **2**: 3.

de Hoon, M. J., P. Eichenberger, et al. (2010). "Hierarchical evolution of the bacterial sporulation network." Curr Biol **20**(17): R735-745.

Dehal, P. S., M. P. Joachimiak, et al. (2009). "MicrobesOnline: an integrated portal for comparative and functional genomics." Nucl. Acids Res.: gkp919.

Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res **27**(23): 4636-4641.

DeLuca, T. F., I.-H. Wu, et al. (2006). "Roundup: a multi-genome repository of orthologs and evolutionary distances." Bioinformatics **22**(16): 2044-2046.

DeRisi, J. L., V. R. Iyer, et al. (1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale." Science **278**(5338): 680-686.

Detweiler, C. S., D. M. Monack, et al. (2003). "virK, somA and rcsC are important for systemic Salmonella enterica serovar Typhimurium infection and cationic peptide resistance." Mol Microbiol **48**(2): 385-400.

DiMaggio, P. A., Jr., S. R. McAllister, et al. (2008). "Biclustering via optimal re-ordering of data matrices in systems biology: rigorous methods and comparative studies." BMC Bioinformatics **9**: 458.

Dittmar, K. A., E. M. Mobley, et al. (2004). "Exploring the regulation of tRNA distribution on the genomic scale." J Mol Biol **337**(1): 31-47.

Doan, T., P. Servant, et al. (2003). "The Bacillus subtilis ywkA gene encodes a malic enzyme and its transcription is activated by the YufL/YufM two-component system in response to malate." Microbiology **149**(Pt 9): 2331-2343.

Dower, K., D. K. Ellis, et al. (2008). "Innate immune responses to TREM-1 activation: overlap, divergence, and positive and negative cross-talk with bacterial lipopolysaccharide." Journal of immunology **180**(5): 3520-3534.

Draghici, S., P. Khatri, et al. (2003). "Global functional profiling of gene expression." Genomics **81**(2): 98-104.

Dutilh, B. E., M. A. Huynen, et al. (2006). "A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation." BMC Genomics **7**: 10.

Dybkaer, K., J. Iqbal, et al. (2007). "Genome wide transcriptional analysis of resting and IL2 activated human natural killer cells: gene expression signatures indicative of novel molecular signaling pathways." BMC Genomics **8**: 230.

Earl, A. M., R. Losick, et al. (2007). "Bacillus subtilis genome diversity." J Bacteriol **189**(3): 1163-1170.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Research **30**(1): 207-210.

Edgar, R., M. Domrachev, et al. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic Acids Res **30**(1): 207-210.

Edwards, J. S., R. U. Ibarra, et al. (2001). "In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data." Nat Biotechnol **19**(2): 125-130.

Edwards, J. S. and B. O. Palsson (2000). "The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities." Proc Natl Acad Sci U S A **97**(10): 5528-5533.

Edwards, J. S., R. Ramakrishna, et al. (2002). "Characterizing the metabolic phenotype: a phenotype phase plane analysis." Biotechnol Bioeng **77**(1): 27-36.

Efron, B. (2003). "Robbins, Empirical Bayes and Microarrays."

Eichenberger, P., M. Fujita, et al. (2004). "The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis." PLoS Biol **2**(10): e328.

Eichenberger, P., S. T. Jensen, et al. (2003). "The sigmaE regulon and the identification of additional sporulation genes in Bacillus subtilis." J Mol Biol **327**(5): 945-972.

Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-826.

Elemento, O., N. Slonim, et al. (2007). "A universal framework for regulatory element discovery across all genomes and data types." Mol Cell **28**(2): 337-350.

Elemento, O. and S. Tavazoie (2005). "Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach." Genome Biol **6**(2): R18.

Elo, L. L., H. Jarvenpaa, et al. (2010). "Genome-wide profiling of interleukin-4 and STAT6 transcription factor regulation of human Th2 cell programming." Immunity **32**(6): 852-862.

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.

Errington, J. (2003). "Regulation of endospore formation in Bacillus subtilis." Nat Rev Microbiol **1**(2): 117-126.

Eymann, C., A. Dreisbach, et al. (2004). "A comprehensive proteome map of growing Bacillus subtilis cells." Proteomics **4**(10): 2849-2876.

Fabret, C., V. A. Feher, et al. (1999). "Two-component signal transduction in Bacillus subtilis: how one organism sees its world." J Bacteriol **181**(7): 1975-1983.

Fadda, A., A. C. Fierro, et al. (2009). "Inferring the transcriptional network of Bacillus subtilis." Mol Biosyst **5**(12): 1840-1852.

Faith, J. J., M. E. Driscoll, et al. (2008). "Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata." Nucleic Acids Res **36**(Database issue): D866-870.

Faith, J. J., B. Hayete, et al. (2007). "Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles." PLoS Biology **5**(1): e8.

Fawcett, P., P. Eichenberger, et al. (2000). "The transcriptional profile of early to middle sporulation in Bacillus subtilis." Proc Natl Acad Sci U S A **97**(14): 8063-8068.

Ferreira, L. C., R. C. Ferreira, et al. (2005). "Bacillus subtilis as a tool for vaccine development: from antigen factories to delivery vectors." An Acad Bras Cienc **77**(1): 113-124.

Feucht, A., L. Evans, et al. (2003). "Identification of sporulation genes by genome-wide analysis of the sigmaE regulon of Bacillus subtilis." Microbiology **149**(Pt 10): 3023-3034.

Fiehn, O. (2002). "Metabolomics--the link between genotypes and phenotypes." Plant Mol Biol **48**(1-2): 155-171.

Fields, S. and O.-k. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-246.

Filen, S., E. Ylikoski, et al. (2010). "Activating transcription factor 3 is a positive regulator of human IFNG gene expression." Journal of immunology **184**(9): 4990-4999.

Fink, R. C., M. R. Evans, et al. (2007). "FNR is a global regulator of virulence and anaerobic metabolism in Salmonella enterica serovar Typhimurium (ATCC 14028s)." J Bacteriol **189**(6): 2262-2273.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-251.

Fischer, D., L. Rychlewski, et al. (2003). "CAFASP3: the third critical assessment of fully automated structure prediction methods." Proteins **53 Suppl 6**: 503-516.

Flannick, J., A. Novak, et al. (2006). "Graemlin: general and robust alignment of multiple large interaction networks." Genome Research **16**(9): 1169-1181.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Fontecave, M., R. Eliasson, et al. (1989). "Oxygen-sensitive ribonucleoside triphosphate reductase is present in anaerobic Escherichia coli." Proc Natl Acad Sci U S A **86**(7): 2147-2151.

Foster, J. W. (2004). "Escherichia coli acid resistance: tales of an amateur acidophile." Nat Rev Microbiol **2**(11): 898-907.

Friedman, N., M. Linial, et al. (2000). "Using Bayesian networks to analyze expression data." J Comput Biol **7**(3-4): 601-620.

Gama-Castro, S., H. Salgado, et al. (2011). "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)." Nucleic Acids Res **39**(Database issue): D98-105.

Gan, X., A. W. Liew, et al. (2008). "Discovering biclusters in gene expression data based on high-dimensional linear geometries." BMC Bioinformatics **9**: 209.

Gardy, J. L., D. J. Lynn, et al. (2009). "Enabling a systems biology approach to immunology: focus on innate immunity." Trends in Immunology **30**(6): 249-262.

Ge, H., A. J. Walhout, et al. (2003). "Integrating 'omic' information: a bridge between genomics and systems biology." Trends Genet **19**(10): 551-560.

Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Germain, R. N., M. Meier-Schellersheim, et al. (2011). "Systems biology in immunology: a computational modeling perspective." Annual review of immunology **29**: 527-585.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science **330**(6012): 1775-1787.

Gilad, Y., A. Oshlack, et al. (2006). "Expression profiling in primates reveals a rapid evolution of human transcription factors." Nature **440**(7081): 242-245.

Giotis, E. S., A. Muthaiyan, et al. (2008). "Genomic and proteomic analysis of the Alkali-Tolerance Response (AlTR) in Listeria monocytogenes 10403S." BMC Microbiol **8**: 102.

Glasner, J. D., P. Liss, et al. (2003). "ASAP, a systematic annotation package for community analysis of genomes." Nucleic Acids Res **31**(1): 147-151.

Goelzer, A., F. Bekkal Brikci, et al. (2008). "Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of Bacillus subtilis." BMC Systems Biology **2**(1): 20.

Golub, G. and W. Kahan (1965). "Calculating the Singular Values and Pseudo-Inverse of a Matrix." Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis **2**(2): 205-224.

Goto, S., S. Kawashima, et al. (2000). "KEGG/EXPRESSION: A Database for Browsing and Analysing Microarray Expression Data."

Grumann, D., S. S. Scharf, et al. (2008). "Immune cell activation by enterotoxin gene cluster (egc)-encoded and non-egc superantigens from Staphylococcus aureus." Journal of immunology **181**(7): 5054-5061.

Grundling, A., L. S. Burrack, et al. (2004). "Listeria monocytogenes regulates flagellar motility gene expression through MogR, a transcriptional repressor required for virulence." Proc Natl Acad Sci U S A **101**(33): 12318-12323.

Guardia, M. J., A. Gambhir, et al. (2000). "Cybernetic modeling and regulation of metabolic pathways in multiple steady states of hybridoma cells." Biotechnol Prog **16**(5): 847-853.

Guldener, U., M. Munsterkotter, et al. (2006). "MPact: the MIPS protein interaction resource on yeast." Nucleic Acids Res **34**(Database issue): D436-441.

Gygi, S. P., B. Rist, et al. (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**(10): 994-999.

Halbleib, J. M., A. M. Saaf, et al. (2007). "Transcriptional modulation of genes encoding structural characteristics of differentiating enterocytes during development of a polarized epithelium in vitro." Mol Biol Cell **18**(11): 4261-4278.

Hamon, M. A., N. R. Stanley, et al. (2004). "Identification of AbrB-regulated genes involved in biofilm formation by Bacillus subtilis." Mol Microbiol **52**(3): 847-860.

Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." Microbiol Mol Biol Rev **68**(4): 669-685.

Harbison, C. T., D. B. Gordon, et al. (2004). "Transcriptional regulatory code of a eukaryotic genome." Nature **431**(7004): 99-104.

Hartigan, J. A. (1972). "Direct Clustering of a Data Matrix." Journal of the American Statistical Association **67**(337): 123-129.

Hashimoto, R. F., S. Kim, et al. (2004). "Growing genetic regulatory networks from seed genes." Bioinformatics **20**(8): 1241-1247.

Hayashi, K., T. Kensuke, et al. (2006). "Bacillus subtilis RghR (YvaN) represses rapG and rapH, which encode inhibitors of expression of the srfA operon." Mol Microbiol **59**(6): 1714-1729.

Hayashi, K., T. Ohsawa, et al. (2005). "The H2O2 stress-responsive regulator PerR positively regulates srfA expression in Bacillus subtilis." J Bacteriol **187**(19): 6659-6667.

Helmann, J. D., M. F. Wu, et al. (2001). "Global transcriptional response of Bacillus subtilis to heat shock." J Bacteriol **183**(24): 7318-7328.

Heng, T. S. and M. W. Painter (2008). "The Immunological Genome Project: networks of gene expression in immune cells." Nature immunology **9**(10): 1091-1094.

Henriques, A. O. and C. P. Moran, Jr. (2007). "Structure, assembly, and function of the spore surface layers." Annual review of microbiology **61**: 555-588.

Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "IntAct: an open source molecular interaction database." Nucleic Acids Res **32**(Database issue): D452-455.

Herschkowitz, J. I., K. Simin, et al. (2007). "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors." Genome Biol **8**(5): R76.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7-8): 563-577.

Hilbert, D. W. and P. J. Piggot (2004). "Compartmentalization of gene expression during Bacillus subtilis spore formation." Microbiol Mol Biol Rev **68**(2): 234-262.

Holtzendorff, J., D. Hung, et al. (2004). "Oscillating global regulators control the genetic circuit driving a bacterial cell cycle." Science **304**(5673): 983-987.

Holtzendorff, J., J. Reinhardt, et al. (2006). "Cell cycle control by oscillating regulatory proteins in Caulobacter crescentus." Bioessays **28**(4): 355-361.

Hommais, F., E. Krin, et al. (2001). "Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS." Mol Microbiol **40**(1): 20-36.

Hong, R. W., M. Shchepetov, et al. (2003). "Transcriptional profile of the Escherichia coli response to the antimicrobial insect peptide cecropin A." Antimicrob Agents Chemother **47**(1): 1-6.

Hottes, A. K., L. Shapiro, et al. (2005). "DnaA coordinates replication initiation and cell cycle transcription in Caulobacter crescentus." Mol Microbiol **58**(5): 1340-1353.

Hsiao, T. L., O. Revelles, et al. "Automatic policing of biochemical annotations using genomic correlations." Nat Chem Biol **6**(1): 34-40.

Hu, Y., H. F. Oliver, et al. (2007). "Transcriptomic and phenotypic analyses suggest a network between the transcriptional regulators HrcA and sigmaB in Listeria monocytogenes." Appl Environ Microbiol **73**(24): 7981-7991.

Hu, Y., S. Raengpradub, et al. (2007). "Phenotypic and transcriptomic analyses demonstrate interactions between the transcriptional regulators CtsR and Sigma B in Listeria monocytogenes." Appl Environ Microbiol **73**(24): 7967-7980.

Hubble, J., J. Demeter, et al. (2009). "Implementation of GenePattern within the Stanford Microarray Database." Nucleic Acids Res **37**(Database issue): D898-901.

Hulsen, T., M. A. Huynen, et al. (2006). "Benchmarking ortholog identification methods using functional genomics data." Genome Biol **7**(4): R31.

Hung, D. Y. and L. Shapiro (2002). "A signal transduction protein cues proteolytic events critical to Caulobacter cell cycle progression." Proc Natl Acad Sci U S A **99**(20): 13160-13165.

Huttenhower, C., A. I. Flamholz, et al. (2007). "Nearest Neighbor Networks: clustering expression data based on gene neighborhoods." BMC Bioinformatics **8**: 250.

Huttenhower, C., K. T. Mutungu, et al. (2009). "Detailing regulatory networks through large scale data integration." Bioinformatics **25**(24): 3267-3274.

Huttenhower, C., K. T. Mutungu, et al. (2009). "Sleipnir Library." from http://function.princeton.edu/sleipnir.

Ibarra, R. U., J. S. Edwards, et al. (2002). "Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth." Nature **420**(6912): 186-189.

Ideker, T., O. Ozier, et al. (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics **18 Suppl 1**: S233-240.

Ihmels, J., S. Bergmann, et al. (2004). "Defining transcription modules using large-scale gene expression data." Bioinformatics **20**(13): 1993-2003.

Ihmels, J., S. Bergmann, et al. (2005). "Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program." PLoS Genet **1**(3): e39.

Iniesta, A. A., P. T. McGrath, et al. (2006). "A phospho-signaling pathway controls the localization and activity of a protease complex critical for bacterial cell cycle progression." Proc Natl Acad Sci U S A **103**(29): 10935-10940.

Ireton, K., S. Jin, et al. (1995). "Krebs cycle function is required for activation of the Spo0A transcription factor in Bacillus subtilis." Proc Natl Acad Sci U S A **92**(7): 2845-2849.

Iwasaki, H. and K. Akashi (2007). "Hematopoietic developmental pathways: on cellular basis." Oncogene **26**(47): 6687-6696.

Jacobs-Wagner, C. (2004). "Regulatory proteins with a sense of direction: cell cycle signalling network in Caulobacter." Mol Microbiol **51**(1): 7-13.

Jacobs, C., N. Ausmees, et al. (2003). "Functions of the CckA histidine kinase in Caulobacter cell cycle control." Mol Microbiol **47**(5): 1279-1290.

Jardine, O., J. Gough, et al. (2002). "Comparison of the small molecule metabolic enzymes of Escherichia coli and Saccharomyces cerevisiae." Genome Res **12**(6): 916-929.

Jensen, L. J., M. Kuhn, et al. (2009). "STRING 8--a global view on proteins and their functional interactions in 630 organisms." Nucleic Acids Res **37**(Database issue): D412-416.

Jensen, L. J., M. Kuhn, et al. (2009). "STRING 8--a global view on proteins and their functional interactions in 630 organisms." Nucl. Acids Res. **37**(suppl_1): D412-416.

Jin, S., P. A. Levin, et al. (1997). "Deletion of the Bacillus subtilis isocitrate dehydrogenase gene causes a block at stage I of sporulation." J Bacteriol **179**(15): 4725-4732.

Jürgen Richter-Gebert, B. S., Thorsten Theobald (2003). "First steps in tropical geometry." Proc. Conference on Idempotent Mathematics and Mathematical Physics.

Kacmarczyk, T. and R. Bonneau. (2010). "Comparative Microbial Module Resource." from http://meatwad.bio.nyu.edu/.

Kalaev, M., M. Smoot, et al. (2008). "NetworkBLAST: comparative analysis of protein networks." Bioinformatics **24**(4): 594-596.

Kaleta, C., A. Gohler, et al. (2010). "Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis." BMC Syst Biol **4**: 116.

Kanehisa, M. (2009). "Bacterial chemotaxis - Bacillus anthracis Sterne." Retrieved September 24, 2009, from http://www.genome.jp/dbget-bin/show_pathway?bat02030.

Kanehisa, M., M. Araki, et al. (2008). "KEGG for linking genomes to life and the environment." Nucleic Acids Res **36**(Database issue): D480-484.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Res **34**(Database issue): D354-357.

Kanehisa, M., S. Goto, et al. (2002). "The KEGG databases at GenomeNet." Nucleic Acids Res **30**(1): 42-46.

Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-280.

Karp, P. D., M. Riley, et al. (2000). "The EcoCyc and MetaCyc databases." Nucleic Acids Res **28**(1): 56-59.

Kaushansky, K. (2010). Hematopoietic Stem Cells, Progenitors, and Cytokines. Williams hematology. L. MA, K. TJ, S. U, Kaushansky K and P. JT. New York, McGraw-Hill Medical**:** xxiii, 2439 p.

Kazakov, A. E., M. J. Cipriano, et al. (2007). "RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes." Nucleic Acids Res **35**(Database issue): D407-412.

Kearns, D. B. and R. Losick (2003). "Swarming motility in undomesticated Bacillus subtilis." Mol Microbiol **49**(3): 581-590.

Keijser, B. J., A. T. Beek, et al. (2007). "Analysis of temporal gene expression during Bacillus subtilis spore germination and outgrowth." J Bacteriol **23**: 23.

Kerrien, S., Y. Alam-Faruque, et al. (2007). "IntAct--open source resource for molecular interaction data." Nucleic Acids Res **35**(Database issue): D561-565.

Keseler, I. M., J. Collado-Vides, et al. (2005). "EcoCyc: a comprehensive database resource for Escherichia coli." Nucleic Acids Res **33**(Database issue): D334-337.

Keymer, D. P., M. C. Miller, et al. (2007). "Genomic and phenotypic diversity of coastal Vibrio cholerae strains is linked to environmental factors." Appl Environ Microbiol **73**(11): 3705-3714.

Khaitovich, P., I. Hellmann, et al. (2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees." Science **309**(5742): 1850-1854.

Kim, C. C. and S. Falkow (2003). "Significance analysis of lexical bias in microarray data." BMC Bioinformatics **4**: 12.

Kim, C. C. and S. Falkow (2004). "Delineation of upstream signaling events in the salmonella pathogenicity island 2 transcriptional activation pathway." J Bacteriol **186**(14): 4694-4704.

Kim, J. W., I. Tchernyshyov, et al. (2006). "HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia." Cell metabolism **3**(3): 177-185.

Kinch, L. N., J. O. Wrabl, et al. (2003). "CASP5 assessment of fold recognition target predictions." Proteins **53 Suppl 6**: 395-409.

Kluger, Y., R. Basri, et al. (2003). "Spectral biclustering of microarray data: coclustering genes and conditions." Genome Res **13**(4): 703-716.

Kobayashi, K., M. Ogura, et al. (2001). "Comprehensive DNA microarray analysis of Bacillus subtilis two-component regulatory systems." J Bacteriol **183**(24): 7365-7370.

Kolbe, M., H. Besir, et al. (2000). "Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution." Science **288**(5470): 1390-1396.

Kolker, E., K. S. Makarova, et al. (2004). "Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae." Nucleic Acids Res **32**(8): 2353-2361.

Kunkel, B., L. Kroos, et al. (1989). "Temporal and spatial control of the mother-cell regulatory gene spoIIID of Bacillus subtilis." Genes Dev **3**(11): 1735-1744.

Kunst, F., N. Ogasawara, et al. (1997). "The complete genome sequence of the Gram-positive bacterium Bacillus subtilis." Nature **390**(6657): 249-256.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome biology **10**(3): R25.

Laub, M. T., S. L. Chen, et al. (2002). "Genes directly controlled by CtrA, a master regulator of the Caulobacter cell cycle." Proc Natl Acad Sci U S A **99**(7): 4632-4637.

Laub, M. T., H. H. McAdams, et al. (2000). "Global analysis of the genetic network controlling a bacterial cell cycle." Science **290**(5499): 2144-2148.

Lawley, T. D., K. Chan, et al. (2006). "Genome-wide screen for Salmonella genes required for long-term systemic infection of the mouse." PLoS Pathog **2**(2): e11.

Lazzeroni, L. and A. Owen (1999). Plaid models for gene expression data.

Lee, M. S., K. Hanspers, et al. (2004). "Gene expression profiles during human CD4+ T cell differentiation." International immunology **16**(8): 1109-1124.

Lee, T. I., N. J. Rinaldi, et al. (2002). "Transcriptional regulatory networks in Saccharomyces cerevisiae." Science **298**(5594): 799-804.

Lemmens, K., T. De Bie, et al. (2009). "DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli." Genome Biol **10**(3): R27.

Lemmens, K., T. De Bie, et al. (2009). "The condition-dependent transcriptional network in Escherichia coli." Ann N Y Acad Sci **1158**: 29-35.

Lewis, N. E., B. K. Cho, et al. (2009). "Gene expression profiling and the use of genome-scale in silico models of Escherichia coli for analysis: providing context for content." J Bacteriol **191**(11): 3437-3444.

Li, G., Q. Ma, et al. (2009). "CSBL Biclustering." from http://csbl.bmb.uga.edu/~maqin/bicluster/.

Li, G., Q. Ma, et al. (2009). "QUBIC: a qualitative biclustering algorithm for analyses of gene expression data." Nucleic Acids Res **37**(15): e101.

Li, J., D. M. Sze, et al. (2010). "Clonal expansions of cytotoxic T cells exist in the blood of patients with Waldenstrom macroglobulinemia but exhibit anergic properties and are eliminated by nucleoside analogue therapy." Blood **115**(17): 3580-3588.

Liao, C. S., K. Lu, et al. (2009). "IsoRankN: spectral methods for global alignment of multiple protein networks." Bioinformatics **25**(12): i253-258.

Lin, J., I. S. Lee, et al. (1995). "Comparative analysis of extreme acid survival in Salmonella typhimurium, Shigella flexneri, and Escherichia coli." J Bacteriol **177**(14): 4097-4104.

Lin, J. T., M. B. Connelly, et al. (2005). "Global transcriptional response of Bacillus subtilis to treatment with subinhibitory concentrations of antibiotics that inhibit protein synthesis." Antimicrob Agents Chemother **49**(5): 1915-1926.

Liu, R. and H. Ochman (2007). "Origins of flagellar gene operons and secondary flagellar systems." J Bacteriol **189**(19): 7098-7104.

Liu, R. and H. Ochman (2007). "Stepwise formation of the bacterial flagellar system." Proc Natl Acad Sci U S A **104**(17): 7116-7121.

Liu, X. and H. W. Taber (1998). "Catabolite regulation of the Bacillus subtilis ctaBCDEF gene cluster." J Bacteriol **180**(23): 6154-6163.

Loew, L. M. and J. C. Schaff (2001). "The Virtual Cell: a software environment for computational cell biology." Trends Biotechnol **19**(10): 401-406.

Longo, N. S., P. L. Lugar, et al. (2009). "Analysis of somatic hypermutation in X-linked hyper-IgM syndrome shows specific deficiencies in mutational targeting." Blood **113**(16): 3706-3715.

Lu, Y., P. Huggins, et al. (2009). "Cross species analysis of microarray expression data." Bioinformatics **25**(12): 1476-1483.

Luecke, H., B. Schobert, et al. (2000). "Coupling photoisomerization of retinal to directional transport in bacteriorhodopsin." J Mol Biol **300**(5): 1237-1255.

Lynn, D. J., C. Chan, et al. (2010). "Curating the innate immunity interactome." BMC Syst Biol **4**: 117.

Macnab, R. M. (2003). "How bacteria assemble flagella." Annu Rev Microbiol **57**: 77-100.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

Maglott, D., J. Ostell, et al. (2005). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **33**(Database issue): D54-58.

Malmstrom, L., M. Riffle, et al. (2007). "Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology." PLoS Biol **5**(4): e76.

Mangan, S. and U. Alon (2003). "Structure and function of the feed-forward loop network motif." Proc Natl Acad Sci U S A **100**(21): 11980-11985.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-753.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Marquez-Magana, L. M. and M. J. Chamberlin (1994). "Characterization of the sigD transcription unit of Bacillus subtilis." J Bacteriol **176**(8): 2427-2434.

Marr, A. K., B. Joseph, et al. (2006). "Overexpression of PrfA leads to growth inhibition of Listeria monocytogenes in glucose-containing culture media by interfering with glucose uptake." J Bacteriol **188**(11): 3887-3901.

Martinez-Llordella, M., I. Puig-Pey, et al. (2007). "Multiparameter immune profiling of operational tolerance in liver transplantation." American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons **7**(2): 309-319.

Mascher, T., J. D. Helmann, et al. (2006). "Stimulus perception in bacterial signal-transducing histidine kinases." Microbiol Mol Biol Rev **70**(4): 910-938.

Masuda, N. and G. M. Church (2002). "Escherichia coli gene expression responsive to levels of the response regulator EvgA." J Bacteriol **184**(22): 6225-6234.

Masuda, N. and G. M. Church (2003). "Regulatory network of acid resistance genes in Escherichia coli." Mol Microbiol **48**(3): 699-712.

Matsuno, K., T. Blais, et al. (1999). "Metabolic imbalance and sporulation in an isocitrate dehydrogenase mutant of Bacillus subtilis." J Bacteriol **181**(11): 3382-3391.

Matthews, L., G. Gopinath, et al. (2009). "Reactome knowledgebase of human biological pathways and processes." Nucleic Acids Res **37**(Database issue): D619-622.

Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Research **31**(1): 374-378.

McCarroll, S. A., C. T. Murphy, et al. (2004). "Comparing genomic expression patterns across species identifies shared transcriptional profile in aging." Nat Genet **36**(2): 197-204.

McGary, K. L., T. J. Park, et al. (2010). "Systematic discovery of nonobvious human disease models through orthologous phenotypes." Proc Natl Acad Sci U S A **107**(14): 6544-6549.

McQuitty, L. L. (1966). "Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data." Educational and Psychological Measurement **26**(4): 825-831.

Meibom, K. L., M. Blokesch, et al. (2005). "Chitin induces natural competence in Vibrio cholerae." Science **310**(5755): 1824-1827.

Meibom, K. L., X. B. Li, et al. (2004). "The Vibrio cholerae chitin utilization program." Proc Natl Acad Sci U S A **101**(8): 2524-2529.

Meile, J. C., L. J. Wu, et al. (2006). "Systematic localisation of proteins fused to the green fluorescent protein in Bacillus subtilis: identification of new proteins at the DNA replication factory." Proteomics **6**(7): 2135-2146.

Mellor, J. C., I. Yanai, et al. (2002). "Predictome: a database of putative functional links between proteins." Nucleic Acids Res **30**(1): 306-309.

Merrell, D. S., S. M. Butler, et al. (2002). "Host-induced epidemic spread of the cholera bacterium." Nature **417**(6889): 642-645.

Michaut, M., S. Kerrien, et al. (2008). "InteroPORC: automated inference of highly conserved protein interaction networks." Bioinformatics **24**(14): 1625-1631.

Mieczkowski, J., M. E. Tyburczy, et al. (2010). "Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements." BMC Bioinformatics **11**: 104.

Miller, M. B. and B. L. Bassler (2001). "Quorum sensing in bacteria." Annu Rev Microbiol **55**: 165-199.

Miller, M. C., D. P. Keymer, et al. (2007). "Detection and transformation of genome segments that differ within a coastal population of Vibrio cholerae strains." Appl Environ Microbiol **73**(11): 3695-3704.

Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science **298**(5594): 824-827.

Mirel, D. B., V. M. Lustre, et al. (1992). "An operon of Bacillus subtilis motility genes transcribed by the sigma D form of RNA polymerase." J Bacteriol **174**(13): 4197-4204.

Mirkin, B. G. (1996). Mathematical classification and clustering. Dordrecht ; Boston, Kluwer Academic Publishers.

Molle, V., M. Fujita, et al. (2003). "The Spo0A regulon of Bacillus subtilis." Mol Microbiol **50**(5): 1683-1701.

Molle, V., Y. Nakaura, et al. (2003). "Additional targets of the Bacillus subtilis global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis." J Bacteriol **185**(6): 1911-1922.

Moraru, II, J. C. Schaff, et al. (2002). "The virtual cell: an integrated modeling environment for experimental and computational cell biology." Ann N Y Acad Sci **971**: 595-596.

Moreno-Hagelsieb, G. and J. Collado-Vides (2002). "A powerful non-homology method for the prediction of operons in prokaryotes." Bioinformatics **18 Suppl 1**: S329-336.

Morgan, J. N. and J. A. Sonquist (1963). "Problems in the analysis of survey data, and a proposal." Journal of the American Statistical Association(58): 415-434.

Mosig, S., K. Rennert, et al. (2008). "Monocytes of patients with familial hypercholesterolemia show alterations in cholesterol metabolism." BMC medical genomics **1**: 60.

Moszer, I. (1998). "The complete genome of Bacillus subtilis: from sequence annotation to data management and analysis." FEBS Lett **430**(1-2): 28-36.

Moszer, I., L. M. Jones, et al. (2002). "SubtiList: the reference database for the Bacillus subtilis genome." Nucleic Acids Res **30**(1): 62-65.

Mukherjee, S. (2010). The emperor of all maladies : a biography of cancer. New York, Scribner.

Nakano, S., E. Kuster-Schock, et al. (2003). "Spx-dependent global transcriptional control is induced by thiol-specific oxidative stress in Bacillus subtilis." Proc Natl Acad Sci U S A **100**(23): 13603-13608.

Negre, N., C. D. Brown, et al. (2011). "A cis-regulatory map of the Drosophila genome." Nature **471**(7339): 527-531.

Ng, W. V., S. P. Kennedy, et al. (2000). "Genome sequence of Halobacterium species NRC-1." Proc Natl Acad Sci U S A **97**(22): 12176-12181.

Nielsen, A. T., N. A. Dolganov, et al. (2006). "RpoS controls the Vibrio cholerae mucosal escape response." PLoS Pathog **2**(10): e109.

Nielsen, A. T., N. A. Dolganov, et al. (2010). "A bistable switch and anatomical site control Vibrio cholerae virulence gene expression in the intestine." PLoS Pathog **6**(9).

Nierman, W. C., T. V. Feldblyum, et al. (2001). "Complete genome sequence of Caulobacter crescentus." Proc Natl Acad Sci U S A **98**(7): 4136-4141.

Nishino, K., Y. Inazumi, et al. (2003). "Global analysis of genes regulated by EvgA of the two-component regulatory system in Escherichia coli." J Bacteriol **185**(8): 2667-2672.

Nishino, K. and A. Yamaguchi (2001). "Analysis of a complete library of putative drug transporter genes in Escherichia coli." J Bacteriol **183**(20): 5803-5812.

Nobeli, I., H. Ponstingl, et al. (2003). "A structure-based anatomy of the E.coli metabolome." J Mol Biol **334**(4): 697-719.

Noirot-Gros, M. F., E. Dervyn, et al. (2002). "An expanded view of bacterial DNA replication." Proc Natl Acad Sci U S A **99**(12): 8342-8347.

Novershtern, N., A. Subramanian, et al. (2011). "Densely interconnected transcriptional circuits control cell states in human hematopoiesis." Cell **144**(2): 296-309.

Ogata, H., S. Goto, et al. (1998). "Computation with the KEGG pathway database." Biosystems **47**(1-2): 119-128.

Ogura, M. and Y. Fujita (2007). "Bacillus subtilis rapD, a direct target of transcription repression by RghR, negatively regulates srfA expression." FEMS Microbiol Lett **268**(1): 73-80.

Ogura, M., K. Tsukahara, et al. (2007). "The Bacillus subtilis NatK-NatR two-component system regulates expression of the natAB operon encoding an ABC transporter for sodium ion extrusion." Microbiology **153**(Pt 3): 667-675.

Ogura, M., H. Yamaguchi, et al. (2002). "Whole-genome analysis of genes regulated by the Bacillus subtilis competence transcription factor ComK." J Bacteriol **184**(9): 2344-2351.

Ogura, M., H. Yamaguchi, et al. (2001). "DNA microarray analysis of Bacillus subtilis DegU, ComA and PhoP regulons: an approach to comprehensive analysis of

B.subtilis two-component regulatory systems." Nucleic Acids Res **29**(18): 3804-3813.

Ohashi, Y., T. Inaoka, et al. (2003). "Expression profiling of translation-associated genes in sporulating Bacillus subtilis and consequence of sporulation by gene inactivation." Biosci Biotechnol Biochem **67**(10): 2245-2253.

Orkin, S. H. and L. I. Zon (2008). "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology." Cell **132**(4): 631-644.

Pagel, P., S. Kovac, et al. (2005). "The MIPS mammalian protein-protein interaction database." Bioinformatics **21**(6): 832-834.

Painter, M. W., S. Davis, et al. (2011). "Transcriptomes of the B and T lineages compared by multiplatform microarray profiling." Journal of immunology **186**(5): 3047-3057.

Palsson, B. (2006). Systems biology : properties of reconstructed networks. Cambridge ; New York, Cambridge University Press.

Park, D., R. Singh, et al. (2011). "IsoBase: a database of functionally related proteins across PPI networks." Nucleic Acids Research **39**(Database issue): D295-300.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature reviews. Genetics **10**(10): 669-680.

Parkinson, H., M. Kapushesky, et al. (2009). "ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression." Nucleic Acids Res **37**(Database issue): D868-872.

Paul, S., X. Zhang, et al. (2001). "Two ResD-controlled promoters regulate ctaA expression in Bacillus subtilis." J Bacteriol **183**(10): 3237-3246.

Peregrin-Alvarez, J. M., S. Tsoka, et al. (2003). "The phylogenetic extent of metabolic enzymes and pathways." Genome Res **13**(3): 422-427.

Piccaluga, P. P., C. Agostinelli, et al. (2007). "Gene expression analysis of peripheral T cell lymphoma, unspecified, reveals distinct profiles and new potential therapeutic targets." The Journal of clinical investigation **117**(3): 823-834.

Piggot, P. J. and J. G. Coote (1976). "Genetic aspects of bacterial endospore formation." Bacteriol Rev **40**(4): 908-962.

Piggot, P. J. and D. W. Hilbert (2004). "Sporulation of Bacillus subtilis." Curr Opin Microbiol **7**(6): 579-586.

Poetz, O., J. M. Schwenk, et al. (2005). "Protein microarrays: catching the proteome." Mech Ageing Dev **126**(1): 161-170.

Polen, T., D. Rittmann, et al. (2003). "DNA microarray analyses of the long-term adaptive response of Escherichia coli to acetate and propionate." Appl Environ Microbiol **69**(3): 1759-1774.

Pomposiello, P. J., M. H. Bennik, et al. (2001). "Genome-wide transcriptional profiling of the Escherichia coli responses to superoxide stress and sodium salicylate." J Bacteriol **183**(13): 3890-3902.

Poultney, C. S., R. A. Gutierrez, et al. (2007). "Sungear: interactive visualization and functional analysis of genomic datasets." Bioinformatics **23**(2): 259-261.

Prelic, A., S. Bleuler, et al. (2006). "A systematic comparison and evaluation of biclustering methods for gene expression data." Bioinformatics **22**(9): 1122-1129.

Price, M. N., K. H. Huang, et al. (2005). "MicrobesOnline Operon Predictions for Escherichia coli str. K-12 substr. MG1655." from http://www.microbesonline.org/operons/gnc511145.html.

Price, M. N., K. H. Huang, et al. (2005). "A novel method for accurate operon predictions in all sequenced prokaryotes." Nucleic Acids Res **33**(3): 880-892.

Price, N. D., J. A. Papin, et al. (2003). "Genome-scale microbial in silico models: the constraints-based approach." Trends Biotechnol **21**(4): 162-169.

Prots, I., A. Skapenko, et al. (2011). "Analysis of the transcriptional program of developing induced regulatory T cells." PLoS ONE **6**(2): e16913.

Prouty, A. M., I. E. Brodsky, et al. (2004). "Bile-salt-mediated induction of antimicrobial and bile resistance in Salmonella typhimurium." Microbiology **150**(Pt 4): 775-783.

Qian, J., Y. Kluger, et al. (2003). "Identification and correction of spurious spatial correlations in microarray data." Biotechniques **35**(1): 42-44, 46, 48.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing. Vienna, Austria.

R. A. Majewski and M. M. Domach (1990). "Simple constrained-optimization view of acetate overflow in *E. coli*." Biotechnology and Bioengineering **35**(7): 732-738.

Radom-Aizik, S., F. Zaldivar, Jr., et al. (2008). "Effects of 30 min of aerobic exercise on gene expression in human neutrophils." Journal of applied physiology **104**(1): 236-243.

Raengpradub, S., M. Wiedmann, et al. (2008). "Comparative analysis of the sigma B-dependent stress responses in Listeria monocytogenes and Listeria innocua strains exposed to selected stress conditions." Appl Environ Microbiol **74**(1): 158-171.

Rasko, D. A., J. Ravel, et al. (2004). "The genome sequence of Bacillus cereus ATCC 10987 reveals metabolic adaptations and a large plasmid related to Bacillus anthracis pXO1." Nucleic Acids Res **32**(3): 977-988.

Razick, S., G. Magklaras, et al. (2008). "iRefIndex: a consolidated protein interaction database with provenance." BMC Bioinformatics **9**: 405.

Reed, J. L. and B. O. Palsson (2003). "Thirteen years of building constraint-based in silico models of Escherichia coli." J Bacteriol **185**(9): 2692-2699.

Reed, J. L., T. D. Vo, et al. (2003). "An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)." Genome Biol **4**(9): R54.

Reiss, D. J., N. S. Baliga, et al. (2006). "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." BMC Bioinformatics **7**: 280.

Remm, M., C. E. Storm, et al. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons." J Mol Biol **314**(5): 1041-1052.

Ren, D., L. A. Bedzyk, et al. (2004). "Differential gene expression to investigate the effect of (5Z)-4-bromo- 5-(bromomethylene)-3-butyl-2(5H)-furanone on Bacillus subtilis." Appl Environ Microbiol **70**(8): 4941-4949.

Richmond, C. S., J. D. Glasner, et al. (1999). "Genome-wide expression profiling in Escherichia coli K-12." Nucleic Acids Res **27**(19): 3821-3835.

Riffle, M., L. Malmstrom, et al. (2005). "The Yeast Resource Center Public Data Repository." Nucleic Acids Res **33**(Database issue): D378-382.

Roberts, A., C. Trapnell, et al. (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome Biology **12**(3): R22.

Roth, F. P., J. D. Hughes, et al. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol **16**(10): 939-945.

Rudd, K. E. (2000). "EcoGene: a genome sequence database for Escherichia coli K-12." Nucleic Acids Res **28**(1): 60-64.

Ruepp, A., B. Brauner, et al. (2008). "CORUM: the comprehensive resource of mammalian protein complexes." Nucleic Acids Res **36**(Database issue): D646-650.

Ruepp, A. and J. Soppa (1996). "Fermentative arginine degradation in Halobacterium salinarium (formerly Halobacterium halobium): genes, gene products, and transcripts of the arcRACB gene cluster." J Bacteriol **178**(16): 4942-4947.

Rusch, D. B., A. L. Halpern, et al. (2007). "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific." PLoS Biol **5**(3): e77.

Salgado, H., S. Gama-Castro, et al. (2006). "RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions." Nucleic Acids Res **34**(Database issue): D394-397.

Salvetti, S., E. Ghelardi, et al. (2007). "FlhF, a signal recognition particle-like GTPase, is involved in the regulation of flagellar arrangement, motility behaviour and protein secretion in Bacillus cereus." Microbiology **153**(8): 2541-2552.

Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-451.

Salzberg, S. L., A. L. Delcher, et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-548.

Sandberg, R. and O. Larsson (2007). "Improved precision and accuracy for microarrays using updated probe set definitions." BMC Bioinformatics **8**: 48.

Sanger, F., S. Nicklen, et al. (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Schneider, A., C. Dessimoz, et al. (2007). "OMA Browser—Exploring orthologous relations across 352 complete genomes." Bioinformatics **23**(16): 2180-2182.

Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-176.

Segal, E., B. Taskar, et al. (2001). "Rich probabilistic models for gene expression." Bioinformatics **17 Suppl 1**: S243-252.

Selinger, D. W., K. J. Cheung, et al. (2000). "RNA expression analysis using a 30 base pair resolution Escherichia coli genome array." Nat Biotechnol **18**(12): 1262-1268.

Selkov, E., Jr., Y. Grechkin, et al. (1998). "MPW: the Metabolic Pathways Database." Nucleic Acids Res **26**(1): 43-45.

Serizawa, M., K. Kodama, et al. (2005). "Functional analysis of the YvrGHb two-component system of Bacillus subtilis: identification of the regulated genes by DNA microarray and northern blot analyses." Biosci Biotechnol Biochem **69**(11): 2155-2169.

Serizawa, M., H. Yamamoto, et al. (2004). "Systematic analysis of SigD-regulated genes in Bacillus subtilis by DNA microarray and Northern blotting analyses." Gene **329**: 125-136.

Serres, M. H., S. Gopal, et al. (2001). "A functional update of the Escherichia coli K-12 genome." Genome Biol **2**(9): RESEARCH0035.

Setlow, P. (2003). "Spore germination." Curr Opin Microbiol **6**(6): 550-556.

Severino, P., O. Dussurget, et al. (2007). "Comparative transcriptome analysis of Listeria monocytogenes strains of the two major lineages reveals differences in virulence, cell wall, and stress response." Appl Environ Microbiol **73**(19): 6078-6088.

Shamir, R., A. Maron-Katz, et al. (2005). "EXPANDER--an integrative program suite for microarray data analysis." BMC Bioinformatics **6**: 232.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Shannon, P. T., D. J. Reiss, et al. (2006). "The Gaggle: an open-source software system for integrating bioinformatics software and data sources." BMC Bioinformatics **7**: 176.

Shen-Orr, S. S., R. Milo, et al. (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." Nat Genet **31**(1): 64-68.

Shen, A. and D. E. Higgins (2006). "The MogR transcriptional repressor regulates nonhierarchal expression of flagellar motility genes and virulence in Listeria monocytogenes." PLoS Pathog **2**(4): e30.

Sheng, Q., Y. Moreau, et al. (2003). "Biclustering microarray data by Gibbs sampling." Bioinformatics **19 Suppl 2**: ii196-205.

Sherlock, G. (2009). "GO-TermFinder." from http://search.cpan.org/dist/GO-TermFinder/.

Sherlock, G., T. Hernandez-Boussard, et al. (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-155.

Shi, J., P. R. Romero, et al. (2006). "Evidence supporting predicted metabolic pathways for Vibrio cholerae: gene expression data and clinical tests." Nucleic Acids Res **34**(8): 2438-2444.

Shmulevich, I., H. Lahdesmaki, et al. (2003). "The role of certain Post classes in Boolean network models of genetic networks." Proc Natl Acad Sci U S A **100**(19): 10734-10739.

Sierro, N., Y. Makita, et al. (2008). "DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information." Nucleic Acids Res **36**(Database issue): D93-96.

Silvaggi, J. M., J. B. Perkins, et al. (2006). "Genes for small, noncoding RNAs under sporulation control in Bacillus subtilis." J Bacteriol **188**(2): 532-541.

Singh, R., J. Xu, et al. (2008). "Global alignment of multiple protein interaction networks with application to functional orthology detection." Proc Natl Acad Sci U S A **105**(35): 12763-12768.

Skerker, J. M. and M. T. Laub (2004). "Cell-cycle progression and the generation of asymmetry in Caulobacter crescentus." Nat Rev Microbiol **2**(4): 325-337.

Skerker, J. M., M. S. Prasol, et al. (2005). "Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis." PLoS Biol **3**(10): e334.

Slonim, N., N. Friedman, et al. (2006). "Multivariate information bottleneck." Neural Comput **18**(8): 1739-1789.

Snel, B., G. Lehmann, et al. (2000). "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." Nucl. Acids Res. **28**(18): 3442-3444.

Soga, T., Y. Ohashi, et al. (2003). "Quantitative metabolome analysis using capillary electrophoresis mass spectrometry." J Proteome Res **2**(5): 488-494.

Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proc Natl Acad Sci U S A **103**(32): 12115-12120.

Spudich, E. N., T. Takahashi, et al. (1989). "Sensory rhodopsins I and II modulate a methylation/demethylation system in Halobacterium halobium phototaxis." Proc Natl Acad Sci U S A **86**(20): 7746-7750.

Spudich, J. L. (1993). "Color sensing in the Archaea: a eukaryotic-like receptor coupled to a prokaryotic transducer." J Bacteriol **175**(24): 7755-7761.

Spudich, J. L. and R. A. Bogomolni (1984). "Mechanism of colour discrimination by a bacterial sensory rhodopsin." Nature **312**(5994): 509-513.

Stanley, J. T., R. P. Gunsalus, et al. (2007). Biosynthesis of Monomers, Nitrogen Assimilation. Microbial Life. J. T. Stanley. Sunderland, MA, Sinauer Associates Inc.**:** ???

Stark, C., B. J. Breitkreutz, et al. (2011). "The BioGRID Interaction Database: 2011 update." Nucleic acids research **39**(Database issue): D698-704.

Stark, C., B. J. Breitkreutz, et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic acids research **34**(Database issue): D535-539.

Steil, L., M. Serrano, et al. (2005). "Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of Bacillus subtilis." Microbiology **151**(Pt 2): 399-420.

Stein, B., S. M. Eissen, et al. (2003). On Cluster Validity and the Information Need of Users. Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03), Benalmádena, Spain. M. H. Hanza, ACTA Press**:** 216-221.

Sterne, M. and H. Proom (1957). "Induction of motility and capsulation in Bacillus anthracis." J Bacteriol **74**(4): 541-542.

Stevens, C. M. and J. Errington (1990). "Differential gene expression during sporulation in Bacillus subtilis: structure and regulation of the spoIIID gene." Mol Microbiol **4**(4): 543-551.

Stock, A. M., V. L. Robinson, et al. (2000). "Two-component signal transduction." Annu Rev Biochem **69**: 183-215.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Stragier, P. (2002). A Gene Odyssey: Exploring the Genomes of Endospore-Forming Bacteria. Bacillus subtilis and its closest relatives : from genes to cells. A. L. Sonenshein, J. A. Hoch and R. Losick. Washington, D.C., ASM Press**:** pp. 519-525.

Stragier, P. and R. Losick (1996). "Molecular genetics of sporulation in Bacillus subtilis." Annu Rev Genet **30**: 297-241.

Streips, U. N. and F. W. Polio (1985). "Heat shock proteins in bacilli." J Bacteriol **162**(1): 434-437.

Stuart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**(5643): 249-255.

Supper, J., M. Strauch, et al. (2007). "EDISA: extracting biclusters from multiple time-series of gene expression profiles." BMC Bioinformatics **8**: 334.

Tam le, T., H. Antelmann, et al. (2006). "Proteome signatures for stress and starvation in Bacillus subtilis as revealed by a 2-D gel image color coding approach." Proteomics **6**(16): 4565-4585.

Tanay, A., A. Regev, et al. (2005). "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast." Proc Natl Acad Sci U S A **102**(20): 7203-7208.

Tanay, A., R. Sharan, et al. (2004). "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proc Natl Acad Sci U S A **101**(9): 2981-2986.

Tanay, A., R. Sharan, et al. (2002). "Discovering statistically significant biclusters in gene expression data." Bioinformatics **18 Suppl 1**: S136-144.

Tao, H., C. Bausch, et al. (1999). "Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media." J Bacteriol **181**(20): 6425-6440.

Tatti, K. M., C. H. Jones, et al. (1991). "Genetic evidence for interaction of sigma E with the spoIIID promoter in Bacillus subtilis." J Bacteriol **173**(24): 7828-7833.

Tatusov, R. L., M. Y. Galperin, et al. (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Research **28**(1): 33-36.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.

Tatusov, R. L., D. A. Natale, et al. (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes." Nucleic Acids Res **29**(1): 22-28.

Thomas-Chollier, M., O. Sand, et al. (2008). "RSAT: regulatory sequence analysis tools." Nucl. Acids Res. **36**(suppl_2): W119-127.

Thomas-Chollier, M., O. Sand, et al. (2008). "RSAT: regulatory sequence analysis tools." Nucleic Acids Res **36**(Web Server issue): W119-127.

Thorsson, V., M. Hornquist, et al. (2005). "Reverse engineering galactose regulation in yeast through model selection." Stat Appl Genet Mol Biol **4**: Article28.

Tirosh, I. and N. Barkai (2007). "Comparative analysis indicates regulatory neofunctionalization of yeast duplicates." Genome Biol **8**(4): R50.

Tirosh, I., Y. Bilu, et al. (2007). "Comparative biology: beyond sequence analysis." Curr Opin Biotechnol **18**(4): 371-377.

Tirosh, I., A. Weinberger, et al. (2006). "A genetic signature of interspecies variations in gene expression." Nat Genet **38**(7): 830-834.

Todhanakasem, T. and G. M. Young (2008). "Loss of flagellum-based motility by Listeria monocytogenes results in formation of hyperbiofilms." J Bacteriol **190**(17): 6030-6034.

Tojo, S., M. Matsunaga, et al. (2003). "Organization and expression of the Bacillus subtilis sigY operon." J Biochem **134**(6): 935-946.

Tomasinsig, L., M. Scocchi, et al. (2004). "Genome-wide transcriptional profiling of the Escherichia coli response to a proline-rich antimicrobial peptide." Antimicrob Agents Chemother **48**(9): 3260-3267.

Tringe, S. G. and E. M. Rubin (2005). "Metagenomics: DNA sequencing of environmental samples." Nat Rev Genet **6**(11): 805-814.

Tucker, D. L., N. Tucker, et al. (2002). "Gene expression profiling of the pH response in Escherichia coli." J Bacteriol **184**(23): 6551-6558.

Tucker, D. L., N. Tucker, et al. (2003). "Genes of the GadX-GadW regulon in Escherichia coli." J Bacteriol **185**(10): 3190-3201.

Turnbaugh, P. J., M. Hamady, et al. (2009). "A core gut microbiome in obese and lean twins." Nature **457**(7228): 480-484.

Turnbaugh, P. J., R. E. Ley, et al. (2007). "The human microbiome project." Nature **449**(7164): 804-810.

Turnbaugh, P. J., R. E. Ley, et al. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**(7122): 1027-1031.

Tyson, G. W., J. Chapman, et al. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature **428**(6978): 37-43.

Ulitsky, I. and R. Shamir (2007). "Identification of functional modules using network topology and high-throughput data." BMC Syst Biol **1**: 8.

Valeru, S. P., P. K. Rompikuntal, et al. (2009). "Role of melanin pigment in expression of Vibrio cholerae virulence factors." Infect Immun **77**(3): 935-942.

van Helden, J. (2003). "Regulatory sequence analysis tools." Nucleic Acids Res **31**(13): 3593-3596.

van Someren, E. P., L. F. Wessels, et al. (2002). "Genetic network modeling." Pharmacogenomics **3**(4): 507-525.

van Someren, E. P., L. F. Wessels, et al. (2000). "Linear modeling of genetic networks from experimental data." Proc Int Conf Intell Syst Mol Biol **8**: 355-366.

Vanet, A., L. Marsan, et al. (2000). "Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals." J Mol Biol **297**(2): 335-353.

Varma, A. and B. O. Palsson (1994). "Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110." Appl Environ Microbiol **60**(10): 3724-3731.

Varner, J. and D. Ramkrishna (1998). "Application of cybernetic models to metabolic engineering: investigation of storage pathways." Biotechnol Bioeng **58**(2-3): 282-291.

Varner, J. and D. Ramkrishna (1999). "Metabolic engineering from a cybernetic perspective. 1. Theoretical preliminaries." Biotechnol Prog **15**(3): 407-425.

Vazquez, C. D., J. A. Freyre-Gonzalez, et al. (2009). "Identification of network topological units coordinating the global expression response to glucose in Bacillus subtilis and its comparison to Escherichia coli." BMC Microbiol **9**: 176.

Velculescu, V. E., L. Zhang, et al. (1995). "Serial analysis of gene expression." Science **270**(5235): 484-487.

Vemuri, G. N. and A. A. Aristidou (2005). "Metabolic engineering in the -omics era: elucidating and modulating regulatory networks." Microbiol Mol Biol Rev **69**(2): 197-216.

Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

von Dassow, G., E. Meir, et al. (2000). "The segment polarity network is a robust developmental module." Nature **406**(6792): 188-192.

Wahde, M. and J. Hertz (2001). "Modeling genetic regulatory dynamics in neural development." J Comput Biol **8**(4): 429-442.

Walhout, A. J. and M. Vidal (2001). "High-throughput yeast two-hybrid assays for large-scale protein interaction mapping." Methods **24**(3): 297-306.

Waltman, P., T. Kacmarczyk, et al. (2010). "Multi-species integrative biclustering." Genome Biololgy **11**(R96).

Waltman, P., T. Kacmarczyk, et al. (2009). Prokaryotic Systems Biology. Plant systems biology, Annual plant reviews. G. Coruzzi and R. A. Gutierrez. Ames, Iowa, Blackwell Pub.**:** 67-136.

Waltman, P., T. K. Kuppusamy, et al. (2010). "cMonkey2." from http://ms2.bio.nyu.edu/cMonkey2-trac/.

Wang, A. and D. E. Crowley (2005). "Global gene expression responses to cadmium toxicity in Escherichia coli." J Bacteriol **187**(9): 3259-3266.

Wang, Q. Z., C. Y. Wu, et al. (2006). "Integrating metabolomics into a systems biology framework to exploit metabolic complexity: strategies and applications in microorganisms." Appl Microbiol Biotechnol **70**(2): 151-161.

Wang, S. T., B. Setlow, et al. (2006). "The forespore line of gene expression in Bacillus subtilis." J Mol Biol **358**(1): 16-37.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews. Genetics **10**(1): 57-63.

Watanabe, S., M. Hamano, et al. (2003). "Mannitol-1-phosphate dehydrogenase (MtlD) is required for mannitol and glucitol assimilation in Bacillus subtilis: possible cooperation of mtl and gut operons." J Bacteriol **185**(16): 4816-4824.

Weaver, D. C., C. T. Workman, et al. (1999). "Modeling regulatory networks with weight matrices." Pac Symp Biocomput: 112-123.

Weber, A. and K. Jung (2002). "Profiling early osmostress-dependent gene expression in Escherichia coli using DNA macroarrays." J Bacteriol **184**(19): 5502-5507.

Wei, Y., J. M. Lee, et al. (2001). "High-density microarray-mediated gene expression profiling of Escherichia coli." J Bacteriol **183**(2): 545-556.

Wheeler, D. L., T. Barrett, et al. (2006). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **34**(Database issue).

Winteler, H. V. and D. Haas (1996). "The homologous regulators ANR of Pseudomonas aeruginosa and FNR of Escherichia coli have overlapping but distinct specificities for anaerobically inducible promoters." Microbiology **142 (Pt 3)**: 685-693.

Wodicka, L., H. Dong, et al. (1997). "Genome-wide expression monitoring in Saccharomyces cerevisiae." Nat Biotechnol **15**(13): 1359-1367.

Wolff, S., H. Antelmann, et al. (2007). "Towards the entire proteome of the model bacterium Bacillus subtilis by gel-based and gel-free approaches." J Chromatogr B Analyt Technol Biomed Life Sci **849**(1-2): 129-140.

Wolff, S., A. Otto, et al. (2006). "Gel-free and gel-based proteomics in Bacillus subtilis: a comparative study." Mol Cell Proteomics **5**(7): 1183-1192.

Woszczek, G., L. Y. Chen, et al. (2008). "Leukotriene D(4) induces gene expression in human monocytes through cysteinyl leukotriene type I receptor." The Journal of allergy and clinical immunology **121**(1): 215-221 e211.

Xenarios, I., E. Fernandez, et al. (2001). "DIP: The Database of Interacting Proteins: 2001 update." Nucleic Acids Res **29**(1): 239-241.

Xenarios, I., D. W. Rice, et al. (2000). "DIP: the database of interacting proteins." Nucleic Acids Res **28**(1): 289-291.

Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucleic Acids Res **30**(1): 303-305.

Yamane, K., K. Bunai, et al. (2004). "Protein traffic for secretion and related machinery of Bacillus subtilis." Biosci Biotechnol Biochem **68**(10): 2007-2023.

Yannakakis, M. (1981). "Node-Deletion Problems on Bipartite Graphs." SIAM J. Comput. **10**(2): 310-327.

Ye, R. W., W. Tao, et al. (2000). "Global gene expression profiles of Bacillus subtilis grown under anaerobic conditions." J Bacteriol **182**(16): 4458-4465.

Yooseph, S., G. Sutton, et al. (2007). "The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families." PLoS Biol **5**(3): e16.

Yoshida, K., K. Kobayashi, et al. (2001). "Combined transcriptome and proteome analysis as a powerful approach to study genes under glucose repression in Bacillus subtilis." Nucleic Acids Res **29**(3): 683-692.

Yoshida, K., Y. H. Ohki, et al. (2004). "Bacillus subtilis LmrA is a repressor of the lmrAB and yxaGH operons: identification of its binding site and functional analysis of lmrB and yxaGH." J Bacteriol **186**(17): 5640-5648.

Yoshida, K., H. Yamaguchi, et al. (2003). "Identification of additional TnrA-regulated genes of Bacillus subtilis associated with a TnrA box." Mol Microbiol **49**(1): 157-165.

Youngman, P., J. B. Perkins, et al. (1984). "Construction of a cloning site near one end of Tn917 into which foreign DNA may be inserted without affecting transposition in Bacillus subtilis or expression of the transposon-borne erm gene." Plasmid **12**(1): 1-9.

Yu, H., N. M. Luscombe, et al. (2004). "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs." Genome Research **14**(6): 1107-1118.

Yu, J. Y. a. H. W. a. W. W. a. P. (2003). "Enhanced biclustering on expression data." Yang,J., Wang,H., Wang,W., and Yu,P. 2003. Enhanced biclustering on expression data. In Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering (BIBE), pp. 321-327.

Yu, J. Y. a. W. W. a. H. W. a. P. S. (2002). delta-cluster: Capturing Subspace Correlation in a Large Data Set. {Icde}.

Yu, W. H., H. Hu, et al. (2010). "Bioinformatics analysis of macrophages exposed to Porphyromonas gingivalis: implications in acute vs. chronic infections." PLoS ONE **5**(12): e15613.

Yuh, C. H., H. Bolouri, et al. (1998). "Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene." Science **279**(5358): 1896-1902.

Zanzoni, A., L. Montecchi-Palazzi, et al. (2002). "MINT: a Molecular INTeraction database." FEBS Lett **513**(1): 135-140.

Zare, H., D. Sangurdekar, et al. (2009). "Reconstruction of Escherichia coli transcriptional regulatory networks via regulon-based associations." BMC Syst Biol **3**: 39.

Zhang, G., D. S. Spellman, et al. (2006). "Quantitative phosphotyrosine proteomics of EphB2 signaling by stable isotope labeling with amino acids in cell culture (SILAC)." J Proteome Res **5**(3): 581-588.

Zheng, L. B. and R. Losick (1990). "Cascade regulation of spore coat gene expression in Bacillus subtilis." J Mol Biol **212**(4): 645-660.

Zheng, M., X. Wang, et al. (2001). "DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide." J Bacteriol **183**(15): 4562-4570.

Zinman, G., S. Zhong, et al. (2011). "Biological interaction networks are conserved at the module level." BMC Systems Biology **5**(1): 134.

Zuberi, A. R., C. W. Ying, et al. (1990). "Transcriptional organization of a cloned chemotaxis locus of Bacillus subtilis." J. Bacteriol. **172**(4): 1870-1876.