Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

Human Population Genomics

Outline

But the tale of history forms a very strong bulwark against the stream of time, and to some extent checks its irresistible flow, and, of all things done in it, as many as history has taken over, it secures and binds together, and does not allow them to slip away into the abyss of oblivion.
–Anna Comnena, *The Alexiad*

Summary of the lecture (Mon Jan 26, 2009) / discussion points

1. Graphical Models for structuring population. Wright-Fisher model is well suited for changes in a more or less ÒstaticÓ population. Is it different if we know how the population is structured? Can we build a graph corresponding to a population structure and observe how this graph evolves?
2. What is a rationale for using a reference genome? Can we use several genomes as reference genomes?
3. Sequence genome validation (part of the new sequence and assembly technology topic). How do we know when to trust a reference genome?
4. How a shift in technology can change problem from being NP to P?
5. Is it possible to predict new technology? What does Biotechnology need? Moore's Law in biotechnology?
6. What about population genetics of viruses? It is possible to apply this to viruses, but humans are more interesting.
7. Do we want any standard topics from a standard Bioinformatics course? No, we want to study population genetics.

Possible Sets of Lectures

- **Lecture 1:** Introduction to Biology (Genomics)
- **Lectures 2, 3 & 4:** Reading Human Genomes Haplotypically: New generation sequencing technologies. The challenges. Resequencing algorithms. Sequence assembly algorithms
- **Lecture 5, 6, 7 & 8:** Population Genomics (Wright-Fisher Model, Moran Model, Coalescent Model, Testing the Neutral Theory, Population and Species Comparison)
- **Lecture 9:** Genome Evolution: (Point Mutations; Rearrangements; Evolution by Duplication)

Possible Sets of Lectures (Contd.)

- **Lecture 10:** Genome Structure: (Retro-Elements and their distributions; Physical Properties of a genome; Large Segmental Duplications; Models of Segmental Duplications); Polymorphisms: (SNPS & CNPS; Haplotyping and Haplotype phasing);
- **Lecture 11 & 12:** Genetics: (Linkage Analysis; Association Studies)
- **Lecture 13:** Whole Genome Comparison
- **Lecture 14:** Personalized Medicine

Outline

Information Processing inside a Cell

- Biology — A study of certain special kinds of information processing systems. Or, multi-agent repeated game working under some *replicator dynamics*.
- This view disregards most of what biologists (still) study: *biochemistry*, *molecular biology*, *cell biology*, etc. as these can *only* lead up to an understanding of the structural machinery underlying the biological systems.
- This view would be analogous to saying that one can understand a computer by simply looking at how p-doped and n-doped areas tile a silicon surface.

Games Biomolecules Play

- **Who are the agents in this game?**
- Depending on one's viewpoint one could say these are the *genes* (are they selfish?), or the *cells* or even the *species*, or *phyla*.
- At each level, the molecular substrates for encoding the information and chemical reactions for transforming the informations differ.
  1. **Macroscopic/Population level**: Hereditary and evolutionary roles of the information.
  2. **Microscopic/Cell level**: Cell biological roles of the information.

We start with the *Genome*

Genomes

- Hereditary information of an organism is encoded in its DNA and enclosed in a cell (unless it is a virus).
- All the information contained in the DNA of a single organism is its *genome*.
- Understanding information encoding in DNA: Envision a DNA molecule to be just a *very* long sequence of *nucleotides* or *bases*:

$$\Sigma = \{\mathbf{A}, \mathbf{T}, \mathbf{C}, \mathbf{G}\}.$$

Genomes (Contd.)

- DNA is a double-stranded polymer ... Think of it as a pair of sequences over Σ.
- There is a relation of complementarity between the two sequences:
  - That is if there is an **A** (respectively, **T**, **C**, **G**) on one sequence at a particular position then the other sequence must have a **T** (respectively **A**, **G**, **C**) at the same position.
  - **A** and **T** form one complementary pair and **C** and **G** another.
- It suffices to simply describe one sequence, as the other one is completely determined by the first.

Genomes (Contd.)

- We will measure the sequence length (or the DNA length) in terms of *base pairs* (bp):
  - for instance, human (*H. sapiens*) DNA is 3.3×10^9 bp measuring about 6 ft of DNA polymer completely streched out!
- The genomes vary widely in size: measuring from few thousand base pairs for viruses to $2 \sim 3 \times 10^{11}$ bp for certain amphibian and flowering plants.

Genomes (Contd.)

- Coliphage MS2 (a virus) has the smallest genome: only 3.5×10^3 bp.
- *Mycoplasmas genitallium* (a unicellular organism) has the smallest cellular genome: 5×10^5 bp.
- *Mycoplasma laboratorium* — An artificially constructed organism with a genome size 5.8×10^5 bp (381 genes).
- *C. elegans* (nematode worm, a primitive multicellular organism) has a genome of size $\sim 10^8$ bp.

Goals of a Genome Study

- For example, the *Human Genome Project*, the *HapMap Project* or the *ENCODE (ENCyclopedia Of DNA Elements) Project* or *TCGA (The Cancer Genome Atlas) Project*...

Goals of a Genome Study (Contd.)

1. *Genetic Maps*
2. *Physical Maps*
3. *DNA Sequencing*
4. *Gene Identification*: Identify parts of the DNA involved in controlling the metabolic processes through proteins they encode.
5. *Informatics*
    1. **Diagnostic and Therapeutic Tools:**
    2. **Phylogenetic Tools:**
6. *Polymorphism Analysis*
7. *Population Studies*

Outline

DNA—Structure and Components

- The usual configuration of DNA is in terms of a *double helix*
- DNA consists of two *chains* or *strands* coiling around each other with two alternating grooves of slighltly different spacing.
- The "backbone" in each strand is made of alternating big sugar molecules (Deoxyribose residues: $C_5O_4H_{10}$) and small phosphate ($(PO_4)^{-3}$) molecules.
- One of the four bases (the letters in our alphabet Σ), each one an almost planar nitrogenic organic compound, is connected to the sugar molecule.

DNA—Structure and Components (Contd.)

- The bases are: *adenine* (**A**), *thymine* (**T**), *cytosine* (**C**) and *guanine* (**G**).
- If one reads the sequence of bases then that defines the information encoded by the DNA.
- Complementary base pairs (**A**-**T**, and **C**-**G**) are connected by hydrogen bonds and the base-pair forms an essentially coplanar “rung” connecting the two strands.
- Note: *cytosine* and *thymine* are smaller (lighter) molecules, called *pyrimidines*, where as *guanine* and *adenine* are bigger (bulkier) molecules, called *purines*.

DNA—Structure and Components (Contd.)

- Also, note: *adenine* and *thymine* allow only for double hydrogen bonding, while *cytosine* and *guanine* allow for triple hydrogen bonding.
  Thus the chemical (through hydrogen bonding) and the mechanical (purine to pyrimidine) constraints on the pairing lead to the complementarity and makes the double stranded DNA both chemically inert and mechanically quite rigid and stable.

DNA—Structure and Components (Contd.)

- From a chemist's point of view, the building blocks of the DNA molecule are four kinds of deoxyribonucleotides...
- Each deoxyribonucleotide is made up of a sugar residue, a phosphate group and a base. Out of such building blocks (or related, dNTPs deoxyribonucleoside triphosphates), one can synthesize a strand of DNA.

DNA—Structure and Components (Contd.)

- The sugar molecule in the strand is in the shape of a pentagon (4 carbons and 1 oxygen) in a plane parallel to the helix axis and with the 5th carbon (5' C) sticking out.
- The phosphodiester bond (-O-P-O-) between the sugars connects this 5' C to a carbon in the pentagon (3' C) and provides a directionality to each strand.
- The strands in a double-stranded DNA molecule has opposite directions—the strands are *antiparallel*.
- When DNA molecule breaks (say by interacting with a restriction enzyme) it breaks at one of these -O-P-O-bonds.

DNA—Structure and Components (Contd.)

- Note: Most of the enzymes moving along the backbone moves in the 5'-3' direction.
- When we represent a DNA sequence, say by writing **GATTACA**, what we mean is the following:

$$5' - \mathbf{GATTACA} - \mathbf{3}'$$
$$3' - \mathbf{CTAATGT} - \mathbf{5}'$$

which is also (the unpronounciable) **TGTAATC**.

Outline

The Cell

- The next more complicated player in the game of life
  - Cell is a small coalition of a set of genes held together in a set of chromosomes and unrelated extrachromosomal elements. It also has a set of machinery made of proteins, enzymes, lipids and organelles taking part in a dynamic process of information processing.
  - In *eukaryotic* cells the genetic materials are enclosed in the *cell nucleus* separated from the other organelles in the *cytoplasm* by a membrane.
  - In *prokaryotic* cells the genetic materials are distributed homogeneously as it does not have a nucleus. Example of prokaryotic cells are bacteria with a considerably simple genome.

The Cell

The organelles common to eukaryotic plant and animal cells include

- *Mitochondria* in animal cells and *chloroplasts* in plant cells (responsible for energy production);
- A Golgi apparatus (responsible for modifying, sorting and packaging various macromolecules for distribution within and outside the cell);
- Endpolastic reticulum (responsible for synthesizing protein); and
- Nucleus (responsible for holding the DNA as chromosomes and replication and transcription).

Chromosmes

- The entire cell is contained in a sack made of plasma membrane. In plant cells, they are further surrounded by a cellulose cell wall.
- The nucleus of the eukaryotic cells contain its genome in several chromosomes, where each chromosome is simply a single molecule of DNA as well as some proteins (primarily histones).
- The chromosomes can be a circular molecule or linear, in which case the ends are capped with special sequence of *telomeres*.
- The protein in the nucleus binds to the DNA and effects the compaction of the very long DNA molecules.

Chromosomes: Ploidy

- In somatic cells (as opposed to gametes: egg and sperm cells) of most eukaryotic organisms, the chromosomes occur in homologous pairs, with the only exceptions being the X and Y chromosomes—*sex chromosomes*.
- Gemetes contain only unpaired chromosomes; the egg cell contains only X chromosome and the sperm cell either an X or an Y chromosome. The male has X and Y chromosomes; the female, 2 X's.
- Cells with single unpaired chromosomes are called *haploid*; the cells with homologous pairs, *diploid*; the cells with homologous triplet, quadruplet, etc., chromosomes are called *polyploid*—many plant cells are polyploid.

Cell Dynamics

- The dynamics of cell is manifested in several manners:
  - The cell cycle (the set of events that occur within a cell between its birth by mitosis and its division into daughter cells again by mitosis) made up of an *interphase* period when DNA is synthesized and a *mitotic phase*; the cell division by *mitosis* (into 2 daughter cells) and *meiosis* (into 4 gametes from germ-line cells); and working of the machinery within the cell—mainly the ones involving replication of DNA, transcription of DNA into RNA and translation of RNA into protein.

Outline

Central Dogma

- The intermediate molecule carrying the information out of the nucleus of an eukaryotic cell is RNA, a single stranded polymer with the same bases as DNA except the base *thymine* is replaced by *uracil*, **U**.
- RNA also controls the translation process in which amino acids are created making up the proteins.
- The central dogma (due to Francis Crick in 1958) states that these information flows are all unidirectional...

Central Dogma

" The central dogma states that once 'information' has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein ."

RNA

- The polymer RNA (*ribonucleic acid*) is similar to DNA but differ in several ways:
  - It's single stranded
  - Its nucleotide has a ribose sugar (instead of deoxyribose) and
  - It has the pyrimidine base uracil, **U**, substituting *thymine* **T—U** is complementary to **A** just as thymine is.

RNA: Secondary Structure

- One consequence of an RNA molecule's single-strandedness is that it tends to fold back on itself to make helical twisted and rigid segments.
- For instance, if a segment of an RNA is

$$5' - \mathbf{GGGGAAAACCCC} - \mathbf{3}',$$

  then the **C**'s fold back on the **G**'s to make a hairpin structure (with a 4 bp *stem* and a 5 bp *loop*).
- The secondary RNA structure can even be more complicated, for instance, in case of *E. coli Ala* tRNA (transfer RNA) forming a *cloverleaf* .
- Prediction of RNA structure is an interesting computational problem.

Genes

- A specific region of DNA that ultimately determines the synthesis of proteins (through the transcription and translation) is called a *gene*
- NOTE: Originally, a gene meant something more abstract—a unit of hereditary inheritance. Understanding of molecular biological basis of heredity has led to an understanding of a gene with a physical molecular existence.
- Transcription of a gene to a *messenger RNA* is keyed by an RNA polymerase enzyme, which attaches to a *core promoter* (a specific sequence adjacent to the relevant structural gene).

Gene Regulation

- Regulatory sequences such as *silencers* and *enhancers* are responsible in controlling the rate of transcription by their influence on the RNA polymerase
- Regulation involves a feedback control loop involving many large families of *activator* and *repressor* proteins that bind with DNA
- These in turn, transpond the RNA polymerase by *coactivator proteins* and *basal factors*.

Gene Regulation

- The entire structure of transcriptional regulation of gene expression is rather dispersed and fairly complicated:
  - The enhancer and silencer sequences occur over a wide region spanning many Kb's from the core promoter on either directions
  - A gene may have many silencers and enhancers and can be shared among the genes
  - They are not unique—different genes may have different combinations

Gene Regulation

- The proteins involved in control of the RNA polymerase number around fifty and different cliques of transcriptional factors operate in different cliques.
- Any disorder in their proper operation can lead to cancer, immune disorder, heart disease, etc.

Gene Transcription

- The transcription of DNA in to m-RNA is performed with a single strand of DNA (the sense strand) around the region corresponding to a gene.
- The double helix untwists momentarily to create a transcriptional bubble which moves along the DNA in the 3' - 5' direction (of the sense strand) as the complementary m-RNA synthesis progresses adding one RNA nucleotide at a time at the 3' end of the RNA, attaching an **U** (respectively, **A**, **G** and **C**) for the corresponding DNA base of **A** (respectively, **T**, **C** and **G**).

Gene Transcription

- The transcription process ends when a special sequence called the *termination signal* is encountered.
- This newly synthesized m-RNA are capped by attaching special nucleotide sequences to the 5' and 3' ends. This molecule is called a *pre-m-RNA*.
- In eukaryotic cells, the region of DNA that is transcribed into a pre-m-RNA involves more than just the information needed to synthesize the proteins.

Gene Transcription

- The DNA subsequences that contain the information or *code* for protein (somewhat indirectly) are the so-called *exons* which are interrupted by regions of *introns*, the non-coding regions.
- Note that pre-m-RNA contains both exons and introns and needs to be altered to excise all the intronic subsequences in preparation for the translation process—this is done by the *spliceosome*.
- The location of splice sites, separating the introns and exons, is dictated by short sequences and simple rules (which are frequently violated) such as “introns begin with the dinucleotide **GT** and end with the dinucleotide **AG**” (the **GT-AG** rule).

Translation of a Gene

- The translation process begins at a particular location of the m-RNA called the translation start sequence (usuall **AUG**) and is mediated by the *transfer RNA* (t-RNA), made up of a group of small RNA molecules, each with specifity for a particular amino acid.
- The t-RNA's carry the amino acids to the *ribosomes*, the site of protein synthesis, where they are attached to a growing polypeptide. The translation stops when one of the three trinucleotides **UAA**, **UAG**, **UGA** is encountered.

Codons

- Each 3 consecutive (nonoverlapping) bases of m-RNA (corresponding to a *codon*) codes for a specific amino acid. There are $4^3 = 64$ possible trinucleotide *codons* belonging to the set

$$\{\mathbf{U}, \mathbf{A}, \mathbf{G}, \mathbf{C}\} \times \{\mathbf{U}, \mathbf{A}, \mathbf{G}, \mathbf{C}\} \times \{\mathbf{U}, \mathbf{A}, \mathbf{G}, \mathbf{C}\}.$$

- The codon **AUG** is the *start codon* and the codons **UAA**, **UAG**, **UGA** are the *stop codons*.

Codons

- The line of nucleotides between and including the start and stop codons is called an *open reading frame* (ORF) and one can assume that all the information of interest to us resides in the ORF's.
- The mapping from the codons to amino acid (and naturally extended to a mapping from ORF's polypeptides by a homomorphism) given by

$$\begin{aligned} F_P : \{\mathbf{U},\mathbf{A},\mathbf{G},\mathbf{C}\}^{\mathbf{3}} &\rightarrow \{A,R,D,N,C,E,Q,G,H, \\ &\qquad I,L,K,M,F,P,S,T,W,Y,V\} \\ \mathbf{UUU} &\mapsto F(= \text{Phe} = \text{phenylamine}) \end{aligned}$$

Coding of the Amino Acids

	RF_1				
RF_0	**G**	**A**	**C**	**U**	RF_1
G	Gly	Glu	Ala	Val	**G**
	Gly	GLu	Ala	Val	**A**
	Gly	Asp	Ala	Val	**C**
	Gly	Asp	Ala	Val	**U**
A	Arg	Lys	Thr	Met	**G**
	Arg	Lys	Thr	Ile	**A**
	Ser	Asn	Thr	Ile	**C**
	Ser	Asn	Thr	Ile	**U**
C	Arg	Gln	Pro	Leu	**G**
	Arg	Gln	Pro	Leu	**A**
	Arg	His	Pro	Leu	**C**
	Arg	His	Pro	Leu	**U**
U	Trp	Stop	Ser	Leu	**G**
	Stop	Stop	Ser	Leu	**A**
	Cys	Tyr	Ser	Phe	**C**
	Cys	Tyr	Ser	Phe	**U**

Genetic Codes

- The genetic code for each triplet can be read of by looking at the entry given by the first letter (RF_0, base in the reading frame 0) along the left column, the second letter (RF_1, base in the reading frame 1) along the row and the third letter (RF_2, base in the reading frame 2).
- In an ORF, a given occurrence of a base is said to be in *reading frame* 0, 1, or 2, if it is the first, second or third letter in a codon, respectively.
- A codon is said to be *in-frame* if its first base is in reading frame 0.

Reading Frames

- The ribosome is simply a transducer that reads the open reading frame one codon at a time to create the amino acids and subsequently a protein.
- The translation process is carried out by two non-protein coding RNA molecules, r-RNA (ribosomal RNA) and t-RNA (transfer RNA).

Genetic Codes

1 ltr code	3 ltr code	amino acid	inverse homomorphism
A	Ala	alanine	$\mathbf{GC}(\mathbf{U}+\mathbf{A}+\mathbf{C}+\mathbf{G})$
C	Cys	cysteine	$\mathbf{UG}(\mathbf{U}+\mathbf{C})$
D	Asp	aspertic acid	$\mathbf{GA}(\mathbf{U}+\mathbf{C})$
E	Glu	glutamic acid	$\mathbf{GA}(\mathbf{G}+\mathbf{A})$
F	Phe	phenylanine	$\mathbf{UU}(\mathbf{U}+\mathbf{C})$
G	Gly	glycine	$\mathbf{GG}(\mathbf{U}+\mathbf{A}+\mathbf{C}+\mathbf{G})$
H	His	histine	$\mathbf{CA}(\mathbf{U}+\mathbf{C})$
I	Ile	isoleucine	$\mathbf{AU}(\mathbf{U}+\mathbf{A}+\mathbf{C})$
K	Lys	lysine	$\mathbf{AA}(\mathbf{A}+\mathbf{G})$
L	Leu	leucine	$(\mathbf{C}+\mathbf{U})\mathbf{U}(\mathbf{A}+\mathbf{G})+\mathbf{CU}(\mathbf{U}+\mathbf{C})$
M	Met	methionine	$\mathbf{AUG}$
N	Asn	asparginine	$\mathbf{AA}(\mathbf{U}+\mathbf{C})$
P	Pro	proline	$\mathbf{CC}(\mathbf{U}+\mathbf{A}+\mathbf{C}+\mathbf{G})$
Q	Gln	glutamine	$\mathbf{CA}(\mathbf{A}+\mathbf{G})$
R	Arg	arginine	$(\mathbf{A}+\mathbf{C})\mathbf{G}(\mathbf{A}+\mathbf{G})+\mathbf{CG}(\mathbf{U}+\mathbf{C})$
S	Ser	serine	$(\mathbf{AG}+\mathbf{UC})(\mathbf{U}+\mathbf{C})+\mathbf{UC}(\mathbf{A}+\mathbf{G})$
T	Thr	threonine	$\mathbf{AC}(\mathbf{U}+\mathbf{A}+\mathbf{C}+\mathbf{G})$
V	Val	valine	$\mathbf{GU}(\mathbf{U}+\mathbf{A}+\mathbf{C}+\mathbf{G})$
W	Trp	tryptophan	$\mathbf{UGG}$
Y	Tyr	tyrosine	$\mathbf{UA}(\mathbf{U}+\mathbf{C})$

Outline

Population Genetics

- Study of Genetic Basis of Evolution.
- Frequencies and fitness of genotypes in natural population.
- Evolution is the change in the frequencies of genotypes through time, perhaps due to their differences in fitness.
- Fitness of genotypes are determined by the phenotypes.
- Associating genotypes to phenotypes (either in an individual or a population).

Forces in Genotypic Evolution

- **Mutations/Polymorphisms**
- **Genetic Drift**
- **Selection**
- **Migration**
- What dominates? Survival of the Fittest vs. Survival of the First-Mover or Survival of the Luckiest?

Standard Approach: “Bean Bag Genetics”

- Strategy:
  - Ignore the complexities of real populations and focus on the evolution of just one (or a few loci)
  - Treat the population as mating at random or, if subdivided, cross-migrating in a simple pattern
- Though successful, this strategy was mocked as ”Bean Bag Genetics:” Ernst Mayr.
- In the absence of large number of genomic sequence data and powerful statistical algorithms, this strategy appeared to be the only viable one...

Example: DNA Variation in Drosophila

- Marty Kreitman: “Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*.”
- Sequence variation in a sample of natural (wild-type) alleles: 11 alleles from Florida (F1), Washington (Wa), Africa (Af), Japan (Ja) and France (Fr)
- In the region of 11 ADH alleles: No two alleles matched in their DNA sequences

Polymorphisms

- In the coding regions, some alleles did have the same sequence; 14 sites have two alternative nucleotides (**biallelic**); A site with different nucleotides in independently samples: **a sgregating site**, or **a polymorphic site**.

allele	39	226	387	393	441	513	519	531	540	578	606	615	645	68
Reference	T	C	C	C	C	C	T	C	C	A	C	T	A	G
Wa-S	.	T	T	.	A	A	C	.	.	.	.	.	.	.
F1-1S	.	T	T	.	A	A	C	.	.	.	.	.	.	.
Af-S	.	.	.	.	.	.	.	.	.	.	.	.	.	A
Fr-S	.	.	.	.	.	.	.	.	.	.	.	.	.	A
F1-2S	G	.	.	.	.	.	.	.	.	.	.	.	.	.
Ja-S	G	.	.	.	.	.	.	.	T	.	T	.	C	A
F1-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Fr-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Wa-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Af-F	G	.	.	.	.	.	.	G	T	C	T	C	C	.
Ja-F	G	.	.	A	.	.	.	G	T	C	T	C	C	.

Polymorphisms

- About 1.8 of every 100 sites are segregating in the ADH sample (typical for *D. melanogaster*
- The variation at 13 of the 14 segregating sites is **silent** (i.e., they represent synonymous mutations, that changes the codon but not the encoded amino acid)
- The variation at the 578th nucleotide position results in a change of the amino acid at position 192 in protein, where a lysine (**AAG**) or a threonine (**ACG**) is found. This is a *replacement or non-synonymous polymorphism* as this nucleotide polymorphism causes an amino acid polymorphism.

Questions

- What causes these diversities within the same species?
- Why are there so many synonymous polymorphisms? Random mutations are mostly lethal...
- Note: Alcohol Dehydrogenase is an important enzyme as flies and their larvae are often found in fermenting fruits with high alcohol concentration
- Alcohol Dehydrogenase is used in the detoxification of ingested alcohol... A small change in the protein could have a serious consequence.

Variation Across Species

- Comparison of the coding region of the ADH loci in *D. melanogaster* vs. *D. erecta*:
- 36 out of 768 nucleotides differ between the two species
- Of the 36 differences, only 10 (26%) are non-synonymous

Loci and Alleles

- **Locus**: A chromosomal location referring to a segment of DNA (which may or may not have a phenotypic effect). A locus is a template for an allele.
- **Allele**: A segment of DNA sequence at a locus. An allele is an instantiation of a locus.
- The genome consists of a sequence of loci: one for each haploid chromosome. A diploid human has two alleles at a particular autosomal locus (one from father and the other from the mother). If they differ in nucleotide sequences, the human is heterozygote; otherwise, homozygote. They can also vary in copy-numbers... (things get a bit complicated)...

Different Alleles

- **By Origin**: They come from the same locus on different chromosomes (perhaps belonging to different individuals)
- **By State**: They have different nucleotide sequences
- **By Descent**: They do not share a common ancestor allele (i.e., during a relatively short time period in the recent past)
- **Identity by origin**, **Identity by state** or **Identity by descent**...

[End of Lecture #2]