Chapter 5

Contigs

> "*To the editor of* Science*: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists.* "
>
> —G.H. Hardy, *Mendelian Proportions in a Mixed Population*, SCIENCE, **28**:49–50, 1908.

5.1 Islands of Clones

Just as in the preceding chapter, consider a genome of length G that has been uniformly randomly sampled to collect N clones each one of length L. The parameters of interest are now extended to include the overlap threshold:

$G =$ Genome length (in bp).

$L =$ Length of a clone.

$N =$ Number of clones.

$\alpha = \left(\frac{N}{G}\right) =$ Expected # clones starting in a unit interval of G

$= $ Probability of a clone starting at a given site

$c = \left(\frac{LN}{G}\right) =$ *Coverage* $= L\alpha$

$T =$ *Overlap parameter*

$=$ # base pairs two clones must have in common to ensure their overlap.

$\theta = \left(\frac{T}{L}\right) =$ *Overlap threshold ratio*

$\sigma = 1 - \theta$

$$L - T = L(1-\theta) = L\sigma = \frac{c\sigma}{\alpha}.$$

In order to understand the finer structure of the overlapping clones, we need to consider few statistical properties of the "*islands*" and "*oceans*." An (apparent) *island* is defined to be a maximal set of clones that are closed under the reflexive and transitive closure of the relation induced by the overlap rule. Thus an apparent island will always cover a connected subinterval of the genome; however, note that it may still be possible that the union of several islands may also cover a connected subinterval. This happens because even when two islands overlap, if they overlap over a portion that is smaller than the overlap parameter T such an overlap may escape detection.

If an island contains exactly one clone, it is a *singleton island* or a trivial contig; otherwise, it is a non-trivial *contig* containing at least two clones. An *ocean* is a region of the genome between two neighboring islands.

In order to understand how the islands are distributed, let us introduce the indicator variable $S_{i,j}$ denoting the event that the clone i starts at the position j of the genome:

$$S_{i,j} = \begin{cases} 1 & \text{if clone } i \text{ starts at the base pair } j \\ 0 & \text{otherwise} \end{cases}$$

Let

$$V_a = \sum_{i=1}^{N} \sum_{j=a-L+T}^{a-1} S_{i,j}$$

be the random variable denoting the number of clones covering the position a from the left by the amount no larger than $L-T$. Thus every clone starting at position a will be detected to overlap with V_a clones "from the left," as they will each overlap in T base pairs or more with a clone starting at a.

Hence,

$$\begin{aligned} \mathbb{E}\left[\binom{V_a}{k}\right] &= \binom{(L-T)N}{k}\left(\frac{1}{G}\right)^k \\ &\approx \left(\frac{(L-T)N}{G}\right)^k / k! \\ &= \frac{(c-(T/L)c)^k}{k!} \\ &= \frac{(c\sigma)^k}{k!} \end{aligned}$$

Hence, by Brun's sieve, we have:

$$\Pr[V_a = k] = \mathbb{E}[\mathbb{I}_{V_a=k}] = e^{-c\sigma}\frac{(c\sigma)^k}{k!}.$$

In particular:

$$\Pr[V_a = 0] = e^{-c\sigma}, \quad \text{and} \quad \Pr[V_a \neq 0] = 1 - e^{-c\sigma}.$$

There are number of simple conclusions that one can make:

1. The probability that an island begins at a

$$I_a = \Pr[V_a = 0 \ \& \ \exists\, i \ \ S_{i,a} = 1] = \alpha e^{-c\sigma}.$$

2. The expected number of islands (= expected number of oceans =)
$$\sum_{a=1}^{G} I_a = G\alpha e^{-c\sigma} = Ne^{-c\sigma}.$$

3. Thus if we choose $c = \ln G/(1-\theta)$ (and thus make the effective total length of the clones $N(L-T) = NL(1-\theta) = Gc\sigma = G\ln G$) then the expected number of contigs is 1 with high probability, assuming that $\ln G < L\sigma$. Another way of saying the same would be that we must make
$$\theta \leq \max(1 - (\ln G/c), 0),$$
if we wish to get a genome wide complete map. For instance, if we have a 46 $\times$ coverage clone library for human (as claimed by Celera), then we need to use a $\theta \leq 0.474$.

4. The probability that the ith clone begins (or, symmetrically, ends) an island is
$$\Pr[\exists\, a \;\; S_{i,a} = 1 \;\; \& \;\; V_a = 0] = e^{-c\sigma}.$$

5. The probability that an island has exactly $j+1$ clones
$$Z_{I,j+1} = \left(1 - e^{-c\sigma}\right)^j e^{-c\sigma} \approx e^{-(c\sigma + je^{-c\sigma})}.$$
Thus the probability that an island is a singleton is $e^{-c\sigma}$ and the probability that it is a non-trivial contig is $1 - e^{-c\sigma}$.

6. The expected number of singleton islands
$$Ne^{-2c\sigma},$$
and the expected number of contigs is
$$Ne^{-c\sigma} - Ne^{-2c\sigma}.$$

7. The expected number of clones per island is then simply
$$\bar{j} = e^{c\sigma}.$$

8. Suppose an apparent island ends at position y. What is the probability that there is an ocean of length exactly x starting at y? This is simply

$$\begin{aligned}
&\Pr[\text{ No clone starts in the interval } [y-T, y+x] \\
&\qquad \text{and a clone starts at } x+1] \\
&= \alpha(1-\alpha)^{x+T} \\
&\approx e^{-\alpha T}\alpha e^{-\alpha x} \\
&= e^{-c\theta}\alpha e^{-\alpha x}.
\end{aligned}$$

Since the moment generating function in this case is

$$\Psi(t) = \frac{\alpha e^{-c\theta}}{\alpha - t},$$

the expected length of an ocean in base pairs is

$$\mathbb{E}[X] = \Psi'(0) = \frac{e^{-c\theta}}{\alpha} = \frac{L}{c}e^{-c\theta},$$

and the variance is

$$\begin{aligned}
\mathrm{Var}[X] &= \Psi''(0) - (\Psi'(0))^2 = \frac{e^{-c\theta}(2-e^{-c\theta})}{\alpha^2} \\
&= \frac{L^2 e^{-c\theta}(2-e^{-c\theta})}{c^2},
\end{aligned}$$

and

$$\text{Std. Dev.}[X] = \frac{L}{c}e^{-c\theta/2}\sqrt{(2-e^{-c\theta})}$$

9. Note also that the expected fraction of the genome in the oceans (i.e., not represented by the clones) is Ge^{-c} and the total number of oceans is $Ne^{-c\sigma}$. Thus the expected length of an ocean is

$$\frac{Ge^{-c}}{Ne^{-c\sigma}} = \frac{Ge^{-c(1-\sigma)}}{N} = \frac{Ge^{-c\theta}}{N} = \frac{L}{c}e^{-c\theta}.$$

10. Thus the probability that an ocean is of length greater than $N(2\ln N - c)/G$ is

$$\begin{aligned} & e^{-c\theta} \int_{\alpha(2\ln N - c)}^{\infty} e^{-\alpha x}\, \alpha\, dx \\ &= e^{-c\theta} e^{-(2\ln N - c)} \\ &= \frac{e^{c\sigma}}{N^2}. \end{aligned}$$

Since the expected number of oceans is $Ne^{-c\sigma}$, the probability that all the oceans are of length smaller than $N(2\ln N - c)/G$ is

$$\left(1 - \frac{e^{c\sigma}}{N^2}\right)^{Ne^{-c\sigma}} \approx e^{-(1/N)},$$

very close to 1, for large N.

In particular, if

$$\frac{2\ln N}{N} \leq \frac{L}{G},$$

then $2\ln N - c \leq 0$ and all oceans are of length 0 almost surely, and the contigs cover almost all of the genome.

5.2 Length of an Island

Let's try to estimate the expected length of an island in base pairs, with the following heuristic arguments. The expected length of all the oceans is $[(L/c)e^{-c\theta}][Ne^{-c\sigma}] = Ge^{-c}$. Thus the "total length" of all the islands (of course, without properly accounting for the undetected overlaps among the islands) is

$$G - Ge^{-c} = G(1 - e^{-c}),$$

and the expected "length" of an island in base pairs is

$$\begin{aligned} \frac{G(1 - e^{-c})}{Ne^{-c\sigma}} &= \frac{G}{N}\left(\frac{1 - e^{-c}}{e^{-c\sigma}}\right) \\ &= \frac{L}{c}(e^{c\sigma} - e^{-c\theta}) \end{aligned}$$

$$
\begin{aligned}
&\approx L\left(\frac{e^{c\sigma} - 1 + c\theta}{c}\right) \\
&= L\left(\frac{e^{c\sigma} - 1}{c} + \theta\right).
\end{aligned}
$$

For small θ, the above expression is correct, but may need to be modified appropriately, if we wish to account for significantly larger θ and hence the unaccounted for overlaps among the apparent islands. Interestingly enough, for $\theta = 1$ (thus, $\sigma = 0$), the above expressions yields for the expected length of an island a value of L, which is in fact the correct value! We will see shortly with a more detailed analysis that the above expression is correct for all values of θ.

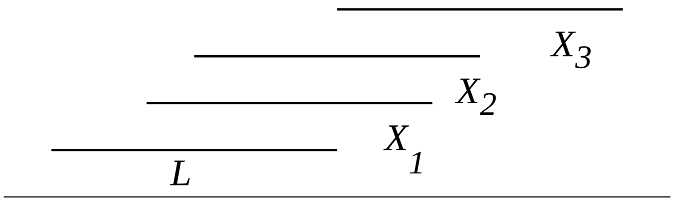
Let $0 = s_1$, s_2, $\ldots$, s_a, s_{a+1}, $\ldots$, s_{a+k}, s_{a+k+1}, $\ldots$, etc., be the left ends of the clones given in the increasing order. Let s_a, be the left end of the first clone of an island of interest. Let X_i be a random variable denoting $s_{a+i} - s_{a+i-1}$, the distance from the left end of the ith clone to the left end of the next clone in the same island, if there is one such clone. Of course $X_i = L$, if in fact it is the last clone in the island and the next clone starts a new island. More formally, let $x_i = s_{a+i} - s_{a+i-1}$

$$
X_i = \begin{cases} x_i, & \text{if } 0 < x_i < L\sigma; \\ L, & \text{otherwise.} \end{cases}
$$

Thus the length of the island is a random variable given as

$$
L = X_1 + X_2 + \cdots + X_k,
$$

such that $x_1 < L\sigma$, $\ldots$, $x_{k-1} < L\sigma$ and $x_k \geq L\sigma$. Thus $\mathbb{E}[L]$ is the expected length of an apparent island.



Definition 5.2.1 *A stochastic process* $\{Z_n, n \geq 1\}$ *is said to be a* martingale *process if*

$$
\mathbb{E}[|Z_n|] < \infty, \quad \textit{for all } n
$$

and

$$\mathbb{E}[Z_{n+1}|Z_1, Z_2, \ldots, Z_n] = Z_n.$$

Taking expectations of the last equation, we get

$$\mathbb{E}[Z_{n+1}] = \mathbb{E}[Z_n], \quad \Rightarrow \quad \forall\, n\ \mathbb{E}[Z_n] = \mathbb{E}[Z_1].$$

A classical example of a martingale is based on the fact that the partial sums of independent random variables having mean 0 is a martingale. Let X_i's be any random variables, for instance, the ones defined for the clones in an island. The random variables

$$X_i - \mathbb{E}[X_i|X_1, \ldots, X_{i-1}]$$

have mean zero and their partial sums

$$Z_n = \sum_{i=1}^{n} \{X_i - \mathbb{E}[X_i|X_1, \ldots, X_{i-1}]\}$$

form a martingale assuming that the expectations of its magnitudes are finite. It is a martingale of mean zero.

Note that

$$Z_{n+1} = Z_n + X_{n+1} - \mathbb{E}[X_{n+1}|X_1, \ldots, X_n]$$

and taking the expectations on both sides, we have

$$\begin{aligned} \mathbb{E}[Z_{n+1}|Z_1, \ldots, Z_n] &= Z_n + \mathbb{E}[X_{n+1}|X_1, \ldots, X_n] - \mathbb{E}[X_{n+1}|X_1, \ldots, X_n] \\ &= Z_n. \end{aligned}$$

Furthermore

$$\mathbb{E}[Z_{n+1}] = \mathbb{E}[Z_1] = \mathbb{E}[X_1] - \mathbb{E}[X_1] = 0.$$

Note, further, that if X_i's are independent (as in our case) then the definition above simplifies to

$$Z_n = \sum_{i=1}^{n} \{X_i - \mathbb{E}[X_i]\}$$

Definition 5.2.2 *For the process $\{X_n, n \geq 1\}$, the integer valued positive random variable N is said to be a* random time*, if the event $\{N = n\}$ is determined by the random variables X_1, X_2, …, X_n. If*

$$\Pr[N < \infty] = 1,$$

then the random time N is said to be a stopping time.

It follows from the *Martingale Stopping Theorem*, that if N is a stopping time for a martingale $\{Z_n, n \geq 1\}$ such that (1) $\mathbb{E}[N] < \infty$ and

$$(2) \exists\ M < \infty\ \ \mathbb{E}[|Z_{n+1} - Z_n| : Z_1, \ldots, Z_n] < M$$

then

$$\mathbb{E}[Z_N] = \mathbb{E}[Z_1].$$

Thus we have the following corollary:

Corollary 5.2.1 *Let X_i $(i \geq 1)$ be independent and identically distributed random variables with a stopping time N, $\mathbb{E}[N] < \infty$ and $\mathbb{E}[|X|] < \infty$, then*

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\mathbb{E}[X].$$

PROOF.

First define $Z_n = \sum_{i=1}^{n} X_i - \mathbb{E}[X_i]$ as before. Then

$$\begin{aligned}
\mathbb{E}[|Z_{n+1} - Z_n| : Z_1, \ldots, Z_n] &= \mathbb{E}[|X_{n+1} - \mathbb{E}[X]|] \\
&\leq \mathbb{E}[|X|] + |\mathbb{E}[X]| < 2\mathbb{E}[|X|] \\
&< \infty.
\end{aligned}$$

Thus $\mathbb{E}[Z_N] = \mathbb{E}[Z_1] = 0$, and

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} X_i - \mathbb{E}[X_i]\right] &= \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] - \mathbb{E}\left[\sum_{i=1}^{N} \mathbb{E}[X]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] - \mathbb{E}[N]\mathbb{E}[X] \\
&= 0,
\end{aligned}$$

as needed. ☐

Now, we are ready to estimate the expected length of an island. Recall that, by definition X_i's are i.i.d. and are defined as

$$X = \begin{cases} x, & \text{if } 0 < x < L\sigma; \\ L, & \text{otherwise.} \end{cases}$$

where x denotes the position where the next clone starts, and is distributed as $(1-\alpha)^x \alpha \approx \alpha e^{-\alpha x}$. Thus,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{L\sigma} x\alpha e^{-\alpha x}\, dx + \int_{L\sigma}^{\infty} L\alpha e^{-\alpha x}\, dx \\
&= \frac{1}{\alpha}\int_0^{c\sigma} y e^{-y}\, dy + L\int_{c\sigma}^{\infty} e^{-y}\, dy \\
&= \frac{L}{c}\left[-ye^{-y}\Big|_0^{c\sigma} - e^{-y}\Big|_0^{c\sigma}\right] + L\left[-e^{-y}\Big|_{c\sigma}^{\infty}\right] \\
&= -L\sigma e^{-c\sigma} + \frac{L}{c}\left(1 - e^{-c\sigma}\right) + Le^{-c\sigma} \\
&= L\left(\frac{1-e^{-c\sigma}}{c} + \theta e^{-c\sigma}\right)
\end{aligned}$$

Also, note that $0 < X \le L$ and $\mathbb{E}[|X|] < L$.

Let k be the stopping time denoting that the kth clone in the island terminates the island. It is easily seen that it is the stopping time as $X_1 < L\sigma$, …, $X_{k-1} < L\sigma$ and $X_k = L$.

Now the probability that a given $x_j < L\sigma$, and thus $j < k$, is given by

$$\begin{aligned}
\Pr[x < L\sigma] &= \int_0^{L\sigma} \alpha e^{-\alpha x}\, dx \\
&= \int_0^{c\sigma} e^{-y}\, dy = \left[-e^{-y}\Big|_0^{c\sigma}\right] = 1 - e^{-c\sigma} < 1.
\end{aligned}$$

Thus the $\mathbb{E}[k] = \sum_{j=1}^{\infty} j(1-e^{-c\sigma})^{j-1} e^{-c\sigma} = e^{c\sigma} < \infty$, and the expected length of an island is

$$\begin{aligned}
\mathbb{E}[k]\mathbb{E}[X] &= e^{c\sigma}\, L\left(\frac{1-e^{-c\sigma}}{c} + \theta e^{-c\sigma}\right) \\
&= L\left(\frac{e^{c\sigma}-1}{c} + \theta\right).
\end{aligned}$$

Rather familiar!

For large c, we can use the following approximation: Expected length of an island is $\approx Le^{c\sigma}/c$—recall that the expected length of an ocean is $Le^{-c\theta}/c$. In particular, their ratio is $\approx e^c$, giving an idea about the exponential growth in the sizes of the islands with the coverage.

5.2.1 Small Islands

We saw that an island is on the average of length $\approx Le^{c\sigma}/c$. Here, we will see that actually this estimate is quite sharp, in the sense almost all islands are of length larger than $Le^{3c\sigma/4}[1-o(1)]/c$.

First note that the probability that an island has at least k' clones is

$$(1-e^{-c\sigma})^{k'} \approx e^{-k'e^{-c\sigma}}.$$

Thus

$$\begin{aligned}
&\Pr[\text{ An island has more than } k' = e^{3c\sigma/4} \text{ clones }]\\
&\geq \quad e^{-e^{3c\sigma/4}e^{-c\sigma}}\\
&= \quad e^{-e^{-c\sigma/4}}.
\end{aligned}$$

Now, if an island has k' or more clones then, we know that its length is bigger than

$$\begin{aligned}
L' &= X_1 + X_2 + \cdots + X_{k'}\\
&= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_{k'}] + Z_{k'}\\
&= Le^{3c\sigma/4}\left(\frac{1-e^{-c\sigma}}{c} + \theta e^{-c\sigma}\right) + Z_{k'}.
\end{aligned}$$

Recall that $\{Z_n, n \geq 1\}$ is a martingale with mean zero. $Z_0 = 0$. $Z_n - Z_{n-1} = X_n - \mathbb{E}[X]$ and since $0 \leq X_n \leq L\sigma$, we have, for all n, $-L \leq -\alpha \leq Z_n - Z_{n-1} \leq \beta \leq L$ with $\alpha+\beta = L\sigma$.

By the *generalized Azuma inequality*, we have for any positive c (e.g., $c = L\sigma e^{-c\sigma/4}$)

$$\Pr[Z_{k'} \leq -k'c] \leq e^{-2k'c^2/(\alpha+\beta)^2}.$$

Thus the probability that $Z_{k'} \leq -e^{3c\sigma/4}(L\sigma e^{-c\sigma/4})$, is less than

$$e^{-2(e^{3c\sigma/4})(L^2\sigma^2 e^{-c\sigma/2})/(L^2\sigma^2)} = e^{-2e^{c\sigma/4}}.$$

Thus with probability close to 1, we have

$$Z_{k'} \geq -\frac{Le^{3c\sigma/4}}{c}(c\sigma e^{-c\sigma/4}).$$

Thus with probability close to 1, we have

$$\begin{aligned} L' &\geq \frac{Le^{3c\sigma/4}}{c}\left(1 - e^{-c\sigma} + c\theta e^{-c\sigma} - c\sigma e^{-c\sigma/4}.\right) \\ &= \frac{Le^{3c\sigma/4}}{c}(1 - o(1)). \end{aligned}$$

And, this also implies that almost all the islands are of length larger than this.

In summary, we have the following.

Theorem 5.2.2 *Consider N clones each of length L from a genome of length G constructed by sampling these uniformly over the genome. Let c denote their coverage $c = NL/G$ and θ denote an overlap threshold denoting the fractions of length that two clones must overlap in order for the overlap to be detected. The followings summarize our earlier results.*

- *The expected number of apparent islands and oceans is $Ne^{-c(1-\theta)}$. The expected number of singleton islands is $Ne^{-2c(1-\theta)}$.*
- *The expected number of clones in an island is $e^{c(1-\theta)}$.*
- *The expected length of an island is*

$$L\left(\frac{e^{c(1-\theta)} - 1}{c} + \theta\right).$$

- *The expected length of an ocean is*

$$\frac{Le^{-c\theta}}{c}.$$

- *An island is of length larger than*

$$(1+o(1))\frac{L}{c}e^{3c(1-\theta)/4},$$

almost surely.

- *An ocean is of length smaller than*

$$\max\left[0, \frac{L}{c}(2\ln(G/L) + 2\ln c - c)\right],$$

almost surely.

A particularly important parameter to consider is the effective coverage $\bar{c} = c(1-\theta)$ and then the number of clones in a contig is $e^{\bar{c}}$, which could be considered a good measure of the complexity of the islands.

Comparing the effective coverage under three different overlapping techniques (i.e., finger print, incomplete maps and complete maps), we see that

1. Finger Print:

$$\bar{c}_{fp} = c\left(1 - \frac{3\beta n}{2}\right) = \frac{NL}{G}\left(1 - \frac{3p_k\beta L}{2}\right).$$

2. Incomplete Maps:

$$\bar{c}_{IM} = c\left(1 - \frac{3\beta}{2p_c^3}\right) = \frac{NL}{G}\left(1 - \frac{3\beta}{2p_c^3}\right).$$

3. Complete Maps:

$$\bar{c}_{CM} = c\left(1 - \frac{K}{n}(1-\beta/2)\right). = \frac{NL}{G}\left(1 - \frac{K}{p_k L}(1-\beta/2)\right).$$