1 NGS Duplicates

1.1 Amplicon duplicates

Let N be the number of distinct segments (or seeds) before the amplification and M be the total number of amplicons in the library. For seed i ($i = 1, \ldots, N$), let k_i be the number of amplicons in the library and k_i is drawn from Poinsson distribution $\mathrm{Po}(\lambda)$. When N is sufficiently large, we have:

$$M = \sum_{i=1}^{N} k_i = N \sum_{k=0}^{\infty} k p_k = N\lambda$$

where $p_k = e^{-\lambda}\lambda^k/k!$.

At the sequencing step, we sample m amplicons from the library. On the condition that:

$$m \ll M \tag{1}$$

we can regard this procedure as sampling with replacement. For seed i, let:

$$X_i = \begin{cases} 1 & \text{seed } i \text{ has been sampled at least once} \\ 0 & \text{otherwise} \end{cases}$$

and then:

$$\mathrm{E}X_i = \Pr\{X_i = 1\} = 1 - \Big(1 - \frac{k_i}{M}\Big)^m \simeq 1 - e^{-k_i m/M}$$

Let:

$$Z = \sum_{i=1}^{N} X_i$$

be the number of seeds sampled from the library. The fraction of duplicates d is:

$$\begin{aligned} d &= 1 - \frac{\mathrm{E}(Z)}{m} \\ &\simeq 1 - \frac{N}{m} \sum_{k=0}^{\infty} \big(1 - e^{-km/M}\big) p_k \\ &= 1 - \frac{N}{m} + \frac{Ne^{-\lambda}}{m} \sum_k \frac{1}{k!} \big(\lambda e^{-m/M}\big)^k \\ &\simeq 1 - \frac{N}{m}\Big[1 - e^{-\lambda} \cdot e^{\lambda(1-m/M)}\Big] \end{aligned}$$

i.e.

$$d \simeq 1 - \frac{N}{m}\Big(1 - e^{-m/N}\Big) \tag{2}$$

irrelevant of λ. In addition, when m/N is sufficiently small:

$$d \approx \frac{m}{2N} \tag{3}$$

This deduction assumes that i) $k_i \ll M$ which should almost always stand; ii) $m \ll M$ which should largely stand because otherwise the fraction of duplicates will far more than half given $\lambda \sim 1000$ and iii) k_i is drawn from a Poisson distribution.

The basic message is that to reduce PCR duplicates, we should either increase the original pool of distinct molecules before amplification or reduce the number of reads sequenced from the library. Reducing PCR cycles, however, plays little role.

1.2 Random coincidence

For simplicity, we assume a read is as short as a single base pair. For m read pairs, define an indicator function:

$$Y_{ij} = \begin{cases} 1 & \text{if at least one read pair is mapped to } (i,j) \\ 0 & \text{otherwise} \end{cases}$$

Let $\{p_k\}$ be the distribution of insert size. Then:

$$\mathrm{E}Y_{ij} = \Pr\{Y_{ij}=1\} = 1 - \left[1 - \frac{p_{j-i}}{L-(j-i)}\right]^m \simeq 1 - e^{-p_{j-i}\cdot m/[L-(j-i)]}$$

where L is the length of the reference. The fraction of random coincidence is:

$$\begin{aligned} d' &= 1 - \frac{1}{m}\sum_{i=1}^{L}\sum_{j=i}^{L}\mathrm{E}Y_{ij} \\ &\simeq 1 - \frac{1}{m}\sum_{i=1}^{L}\sum_{j=i}^{L}\left(1 - e^{-p_{j-i}\cdot m/(L-(j-i))}\right) \\ &= 1 - \frac{1}{m}\sum_{k=0}^{L-1}(L-k)\left[1 - e^{-p_k m/(L-k)}\right] \end{aligned}$$

On the condition that L is sufficient large and:

$$m \ll L \tag{4}$$

$$d' \simeq \frac{m}{2}\sum_{k=0}^{L-1}\frac{p_k^2}{L-k} \tag{5}$$

We can calculate/approximate Equation 5 for two types of distributions. Firstly, if p_k is evenly distributed between $[k_0, k_0+k_1]$, $d' \simeq \frac{m}{2k_1L}$. Secondly, assume k is drawn from $N(\mu,\sigma)$ with $\sigma \gg 1$:

$$p_k = \frac{1}{\sqrt{2\pi}\sigma}\int_k^{k+1} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \simeq \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(k-\mu)^2}{2\sigma^2}}$$

If $p_0 \ll 1$, $\mu \ll L$ and $L \gg 1$:

$$\begin{aligned} d' &\simeq \frac{m}{4\pi\sigma^2}\int_0^1 \frac{1}{1-x}\cdot e^{-\frac{(Lx-\mu)^2}{\sigma^2}}\,dx \\ &\simeq \frac{m}{4\pi\sigma^2}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu/L)^2}{(\sigma/L)^2}}\,dx \\ &= \frac{m}{4\pi\sigma^2}\cdot\frac{\sqrt{2\pi}\cdot\sqrt{2}\sigma}{L} \\ &= \frac{m}{2\sqrt{\pi}\sigma L} \end{aligned}$$