1 Cufflinks Transcript Abundance Model

1.1 Introduction

Now that we have established the gene-read matrix, we now look to an analyze the data using a popular software called Cufflinks. Cufflinks takes an aligned mRNA file and estimates the relative abundance of each transcript. Notice that this is not the gene-read matrix that we discussed in the previous lecture. Instead it uses a more fine grain matrix, the fragment-transcript matrix, $A_{R,T}$ where

$$a_{r,t} = \begin{cases} 1 & \text{fragment } r \text{ is entirely contained in transcript } t \\ 0 & \text{fragment } r \text{ is not contained in transcript } t \end{cases} \tag{1}$$

The task is that given this matrix, can we recover the "transcript abundances" $\{\rho_{t_1}, \rho_{t_2}, \ldots, \rho_{t_{|T|}}\}$ where T is the set of transcripts.

1.2 Maximum Likelihood Estimation

Eventually, we will need to incorporate the length of fragment in the transcript which we will denote $L_t(r)$. Now if we define f_i to be the the ith fragments alignment and let T_i be the true transcript from which the fragment f_i is stored, we can write out the likelihood it the transcript abundances given the number of fragments R as follows.

$$\begin{aligned} \mathcal{L}(\rho|R) &= \prod_{i=1}^{R} \mathbb{P}(f_i = r_i) \\ &= \prod_{i=1}^{R} \sum_{t \in T} \mathbb{P}(f_i = r_i | T_i = t) \mathbb{P}(T_i = t) \\ &= \prod_{i=1}^{R} \sum_{t \in T} \mathbb{P}(f_i = r_i | T_i = t) \frac{\rho_t \tilde{L}(t)}{\sum_{t \in T} \rho_t \tilde{L}(t)} \end{aligned}$$

Here we model $\mathbb{P}(T_i = t)$ based on the *adjusted length* $\tilde{L}(t) = \sum_{i=1}^{L(t)} F(i) \{L(t) - i + 1\}$ where $F(i)$ is the probability that any fragment has length i. In some sense, we are weighting the length amount that read t in transcript i by the length of that fragment. Moreover, the ratio in the last line of the conditional likelihood is the proportion of the transcript abundance t to the entire set. We expect this to be large for transcripts that are particularly abundant. Continuing the derivation we have

$$\mathcal{L}(\rho|R) = \prod_{i=1}^{R} \sum_{t \in T} \frac{F(L_t(r_i))}{L(t) - L_t(r_i) + 1} \alpha_t$$

Here we see that all functions are know except for α_t but notice that α_t must be estimated for far too many parameters. For this reason, we must introduce a new model that reduces the number of model parameters.

1.3 Genome Partitioning

Suppose we instead partition the genome $\mathcal{G} = \bigcup_{j=1}^{|G|} G_j$. Next, define L_i be a region that f_i belongs to. Lastly, define X_g be the number of fragments falling into region G_g. We wish to consider the probability that the region in which f_i is from is equal to some region $\{L_i = G\}$.

$$\begin{aligned}\beta_g \equiv \mathbb{P}(L_i = G) &= \sum_{t \in \mathcal{G}} \frac{\rho_t \tilde{L}(t)}{\sum_{t \in T} \rho_t \tilde{L}(t)} \\ &= \frac{\sum_{t \in \mathcal{G}} \rho_t \widetilde{L}(t)}{\sum_{u \in T} \rho_u \widetilde{L}(u)} = \frac{\sum_{t \in \mathcal{G}} \sigma_g \tau_t \widetilde{L}(t)}{\sum_{h=1}^{|G|} \sum_{u \in G_h} \sigma_h \tau_u \mathcal{L}(u)}\end{aligned}$$

where σ_g is the aggregated abundance over region G, $\sigma_g = \sum_{t \in G} \rho_t$ and τ_t is the proportional abundance in region G, $\tau_t = \frac{\rho_t}{\sum_{u \in G} \rho_u}$. From here, we can rederive the likelihood function as follows.

$$\begin{aligned}\mathcal{L}(\rho | R) &= \prod_{i=1}^{R} \mathbb{P}(f_i = r_i) \\ &= \prod_{i=1}^{R} \sum_{g \in G} \mathbb{P}(f_i = r_i | L_i = g) \mathbb{P}(L_i = g) \\ &= \prod_{i=1}^{r} \mathbb{P}(f_i = r_i | L_i = g) \beta_g \mathcal{I}(r_i \in G_g) \\ &= \left(\prod_{g=1}^{|G|} \prod_{r_i \in G_g} \sum_{t \in G_g} \mathbb{P}(f_i = r_i | L_i = g, T_i = t) \mathbb{P}(T_i = t | L_i = g) \right) \prod_{g=1}^{|\mathcal{G}|} \beta_g^{X_g} \\ &= \left(\prod_{g=1}^{|G|} \prod_{r_i \in G_g} \sum_{t \in G_g} \gamma_t \frac{F(L_t(r_i)}{L(t) - L_t(r_i) + 1} \right) \prod_{g=1}^{|\mathcal{G}|} \beta_g^{X_g}\end{aligned}$$

From here we can estimate γ_t by constrained optimization and $\hat{\beta}_g = \frac{X_g}{R}$.

We can approximate the variance by inverting the empirical Fisher information. However, as this technique is highly unstable in practice, an alternative method is importance sampling from the likelihood function Ψ from which we can estimate the mean and variance. A statistic normally used for abundance estimation in practice is the fragments per kilobase of transcript per million fragments mapped (FPKM).

$$FPKM = \frac{10^6 10^3 \beta_g \gamma_t}{\tilde{L}(t)} \tag{2}$$

From which the estimate is given by

$$\widehat{FPKM} = \frac{10^9 X_g \hat{\gamma}_t}{\tilde{L}(t) R} \tag{3}$$

where the variance is given by the following

$$\begin{aligned}\widehat{\text{Var}(FPKM)} &= \left(\frac{10^9}{\tilde{L}(t)R}\right)\left\{\text{Var}(X_g)\text{Var}(\hat{\gamma}_t) + \text{Var}(X)\mathbb{E}(\hat{\gamma}_t)^2 + \text{Var}(\hat{\gamma}_t)\mathbb{E}(X_g)^2\right\} \\ &= \left(\frac{10^9}{\tilde{L}(t)R}\right)\left\{\Psi^g_{t,t}X_g + \Psi^g_{t,t}X_g^2 + (\hat{\gamma}_t)^2 X_g\right\}\end{aligned}$$

From here, we can test for *differential expression* using the log ratio of RPKMs $\log\left(\frac{X_g^a\hat{\gamma}_t^a R^b}{X_g^b\hat{\gamma}_t^b R^a}\right)$

From here we can use the fact $\text{Var}(\log(X)) \sim \frac{\text{Var}(X)}{E(X)^2}$ we have the following variance estimate

$$\widehat{\text{Var}}\left[\log\left(\frac{X_g^a\hat{\gamma}_t^a R^b}{X_g^b\hat{\gamma}_t^b R^a}\right)\right] = \frac{\Psi^{g,a}_{t,t}(1+X_g^a) + (\hat{\gamma}_t{}^a)^2}{X_g^a(\hat{\gamma}_t{}^a)^2} + \frac{\Psi^{g,b}_{t,t}(1+X_g^a) + (\hat{\gamma}_t{}^b)^2}{X_g^b(\hat{\gamma}_t{}^b)^2} \tag{4}$$

From here, we can build the test statistic under the null hypothesis that $H_0 : RPKM_a = RPKM_b$

$$\frac{\log(\widehat{\text{ratio}}) - 0}{\sqrt{\widehat{\text{Var}}}} \dot{\sim} N(0,1) \tag{5}$$

We can extend this notation to the multiple testing framework by constructing an ANOVA like statistic.