Cufflinks

Notation:

$F(101)$ $\sum_{i=1}^{\infty} F(i) = 1$

Our data

Task: Given this genome we try and figure out the attribution to some transcriptome assignments

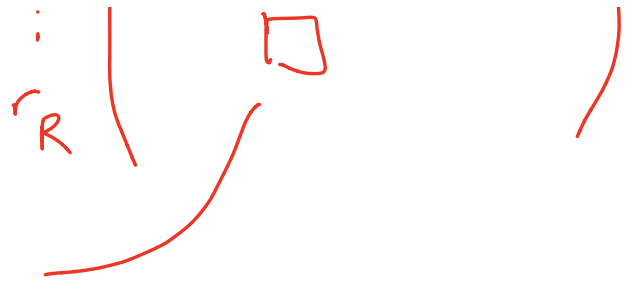
$\rho = \{\rho_{t_1}, \ldots, \rho_{t_{|T|}}\}$

"abundances of transcripts"

$t_1 \quad \cdots \quad t_{|T|}$

r_1

n

H_{RT} = 

1 if fragment is completely contained in t

0 o.w.

- $I_t(r)$ implied length in fragment t

- f_i = ith frag. alignment.

$$L(f|R) = \prod_{i=1}^{R} P(f_i = r_i)$$

$$= \prod_{i=1}^{R} \sum_{t \in T} P(f_i = r_i | T_i = t) P(T_i = t)$$

$$= \prod_{i=1}^{R} \sum_{t} P(f_i = r_i | T_i = t) \rho_t \tilde{I}(t)$$

$$= \prod_{i=1}^{R} \sum_{t \in T} \frac{F(l_t(r_i))}{\sum_{u \in T} \rho_u \hat{T}(u)}$$

with $\hat{T}(t_i) = \sum_{i=1}^{\infty} F(i)\left(\ell(t_i) - i + 1\right)$

Called the "adjusted length"

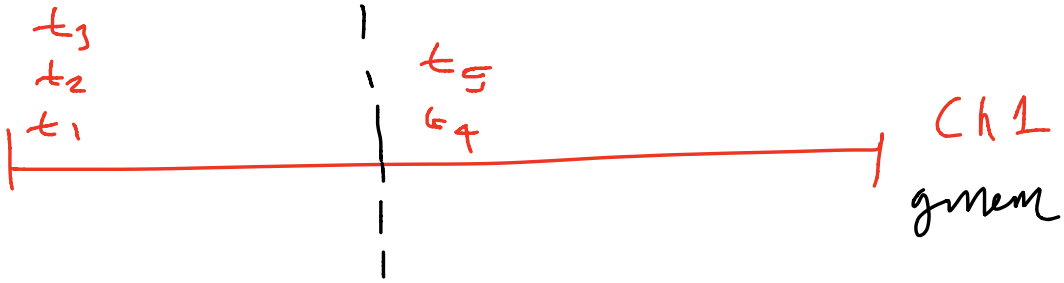
$$= \prod_{i=1}^{R} \sum_{t \in T} \frac{F(l_t(r_i))}{\ell(t) - l_t(r_i) + 1} \alpha_t$$

need to estimate

We could optimize but $|\alpha_t| = 60{,}000$

Too big to optimize.

Idea partition the genome.



Introduce L_i a region where f_i corresponds to and $X_G = \#$ of fragments falling into region G_g.

$$\beta_g \equiv \mathbb{P}(L_i = g) = \sum_{t \in G_g} \alpha_t$$

$$= \frac{\sum_{t \in G_g} \sigma_t T_t \vec{l}(t)}{\sum_{h=1}^{|\mathcal{G}|} \sum_{u \in G_h} \sigma_u T_u \vec{l}(u)}$$

$$= \frac{\sigma_g \vec{l}(G_g)}{\sum_{h=1}^{|\mathcal{G}|} \sigma_h \vec{l}(G_h)}$$

$$\sigma_g = \sum_{t \in G_g} \rho_t \qquad T_t = \frac{\rho_t}{\sum \rho_u}$$

$u \in G_g$

So the full likelihood

$$L(\rho/R) = \prod_{i=1}^{v} P(f_i = r_i / L_i = g)\, \beta_g\, I(r_i \in G_g)$$

$$= \left\{ \prod_{g=1}^{|G|} \prod_{r_i \in G_g} \sum_{t \in G_g} \gamma_t \frac{F(l_t(r_i))}{\ell(t) - I_t(r_i) + 1} \right\}$$

$$\times \prod_{g=1}^{|G|} \beta_g^{x_G}$$

So $\vec{\beta} = \frac{x_G}{R}$ γ_t can be

found via constrained optimization.

- We can have variance estimate
- Unstable
  - identifiability

- Use importance sampling from likelihood. for mean & variance estimate. $\psi = Cov(\gamma)$.

- Using this we can define the FPKM statistic with its corresp. variance.

Testing Diff. Expression

$$\log\left(\frac{x_g^a \, r_\epsilon^a \, R^b}{x_g^b \, \hat{r}_\epsilon^b \, R^a}\right)$$

and we can estimate the variance as well.

Then the test stat. is

$$\sqrt{n}\,\frac{\log(\hat{ratio}) - 0}{\quad} \;\dot{\sim}\; N(0,1)$$

Next up

- HW: how can we test differences between multiple groups.
  - Differential gene transcript expression analysis of ...
- Question HW: how do we do multiple testing in Cufflinks.