Intro to RNA-Seq

- RNA-Seq replacing microarrys
- Issues with microarrys:
  - designing imaging
- RNA-Seq no designing probs just read over the transcriptome
- mRNA ⟼ Short Seq. Reads
  Biologist
- Genome Mapping bioinformatist

Advantages of mRNA
- not limited to known genes
- low background signal
- dynamic range of expression

Computational Issues

- Read mapping
  - ↳ Mapping to the genome
  - ↳ STAR best aligner
- Transcriptome reconstruction
  - ↳ Cufflinks
- Expression quantification
  - ↳ How much expression are we seeing?
  - ↳ Biologists: Cufflinks

Three Categories for regular RNA

(a) Genome Mapping

Reads $\xrightarrow[\text{w/Gaps}]{}$ genom. mapping $\longrightarrow$ trans. id.

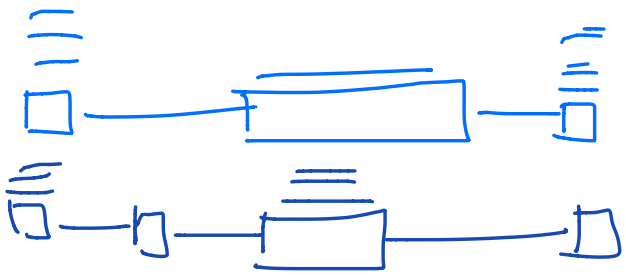
(b) Transcriptome Mapping

Reads $\longrightarrow$ Transcriptome $\longrightarrow$ trans. id

(c) Reference - free assembly

- No mapping exists
- reconstruct genom from scratch

- Need to normalize counts by gene length.

- 

What?

intron 1

intron 2

- Usually just combine into transcript union method.

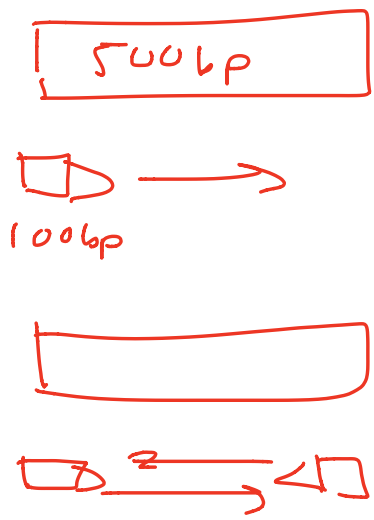
- The counts here are representing the number of the seq. fragments that they find in the mRNA

  ↳ Describes the gene expression.

- CPM = counts per million

$$= \frac{\#\ \text{reads}}{\text{total reads} \times 1{,}000{,}000}$$

- Pair versus unpaired read



Ex:

Gene	Samp1	Samp2	Samp3
A	10	12	30
B			60
C			15
D			1

$\Rightarrow$ / # reads RPKM

Tips:

- Ask for BAM
  - SamTools used to access these
  - (Mapped files in BigWig)
  - Map to count files

Modeling

Data:

	Cond A		Cond B
Gene	Samp 1	...	Samp N

Goal: Compare Group Means.

Issues: Count data, but would like to use normal models

Use VST for Pois, NBinomial.

↳ edgeR, DESeq2

Edge R

$Y \sim NB(\mu, \phi)$ $\quad \mathbb{E}(Y) = \mu$

$\mathrm{Var}(Y) = \mu(1 + \phi\mu)$

Let $Y_i \sim NB(\mu_i = m_i \lambda, \phi)$

library size

For equal library sizes

$$Z = \sum_{i=1}^{n} Y_i \sim NB(nm\lambda, \phi n^{-1})$$

sufficient stat. for λ.

$\ell_{Y|Z=z}(\phi) =$ (see slides)

If the library sizes are not equivalent, then we interpolated

to a common size, the de

the same CML.

- $m^* = \left(\prod m_i\right)^{1/n}$

- Interpolate counts as if the samp. were from $NB(m^* \lambda, \rho)$

  - Initialize φ

  - Estimate λ, given φ

  - See slides...

This is just one gene. Looking to use some Empirical Bayes ideas.

$$WL(\varphi_g) = l_g(\varphi_g) + \alpha\, l_c(\varphi_a)$$

Weighted likelihood with tuning α.

Estimate $\hat{\alpha}_{mom}$ using large

Sample to

Approx steps

- Pseudo Counts Interpolation
- Estimating ℓ

DESeq2:

$$Y_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_i \, q_{ij}$$

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}$$

$$s_j = \underset{i : K_i^R \neq 0}{\text{Median}} \frac{Y_{ij}}{Y_i^R} \qquad Y_i^R = \left(\prod Y_{ij}\right)^{1/m}$$

Interested in infering (α, β)

Using a Bayesian model

$$\log \alpha_i \sim N(\log \mathcal{L}_{tr}(\bar{\mu}_i), \sigma_\alpha^2)$$

— | ...

$$\mu_1 = \frac{1}{m} \frac{\cdots}{s_j}$$

parametric constraint.

$$\alpha_{tr}(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + a_0$$

Using the posterior max gives $\hat{\alpha}$

$\beta_{ir} \sim N(0, \sigma_r^2)$ σ_r ← matches percentiles $\hat{\beta}_r$

$$\hat{\beta}_i = \text{argmax}\left(\sum_j \log f_{NB}(Y_{ij} : \mu_j(\beta), \alpha_i)\right) + \Lambda(\beta_i)$$

Issues with DESeq2:

lots of approximations to accomodate NB.

<u>Voom + limma</u>

- LM with log CPM

- Really good.

Can we improve by aggregation?

- Still really unclear.
  - Read Cufflinks with supp. materials