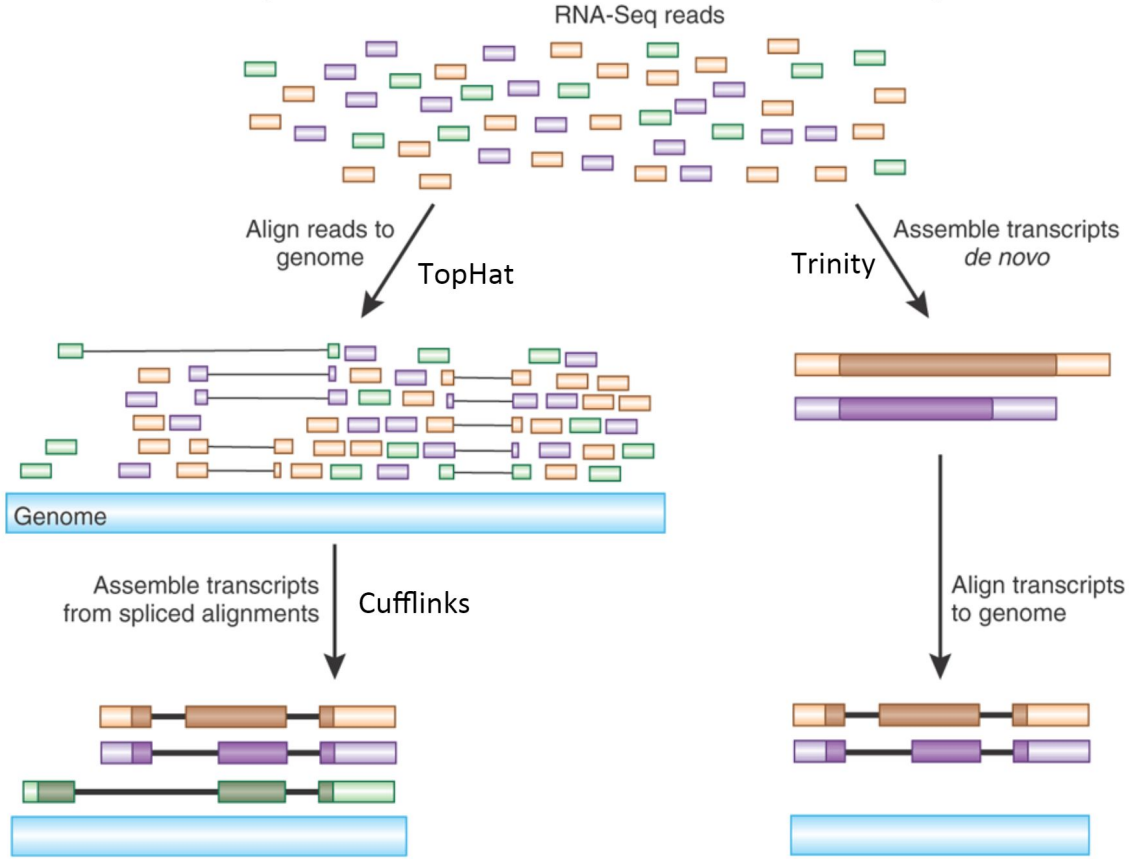
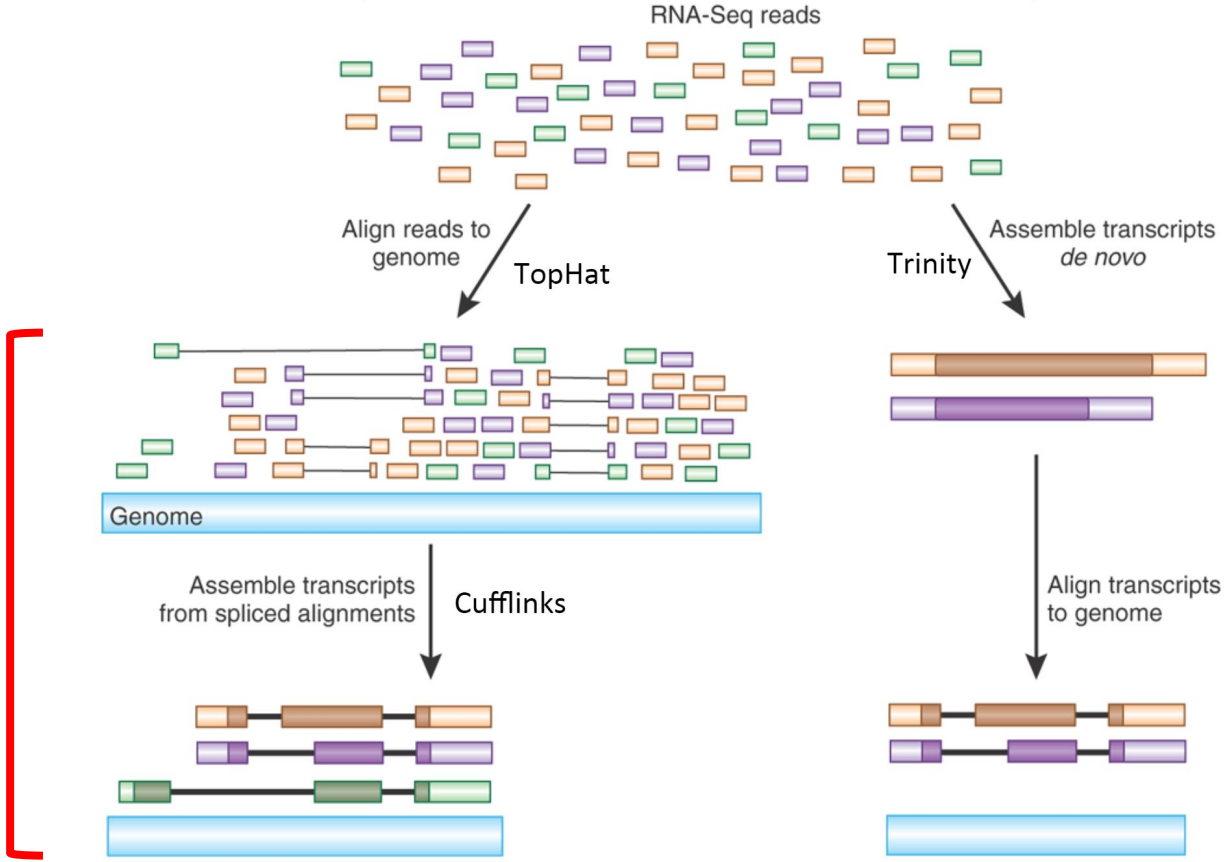
Montagem transcriptoma utilizando Trinity

Transcript Reconstruction from RNA-Seq Reads



https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

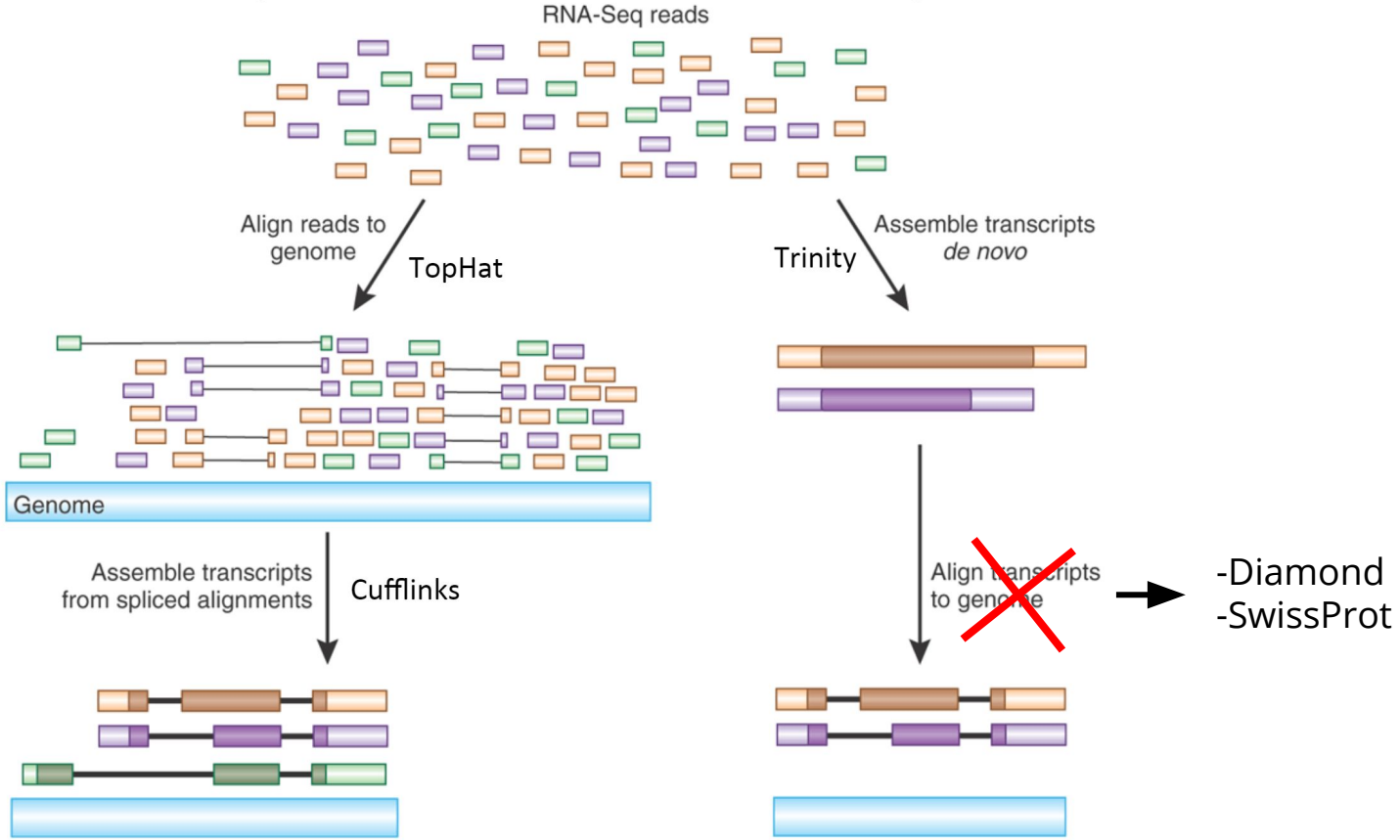
Transcript Reconstruction from RNA-Seq Reads



Pipeline Tuxedo

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Transcript Reconstruction from RNA-Seq Reads

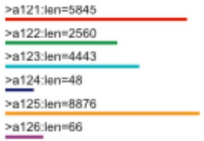


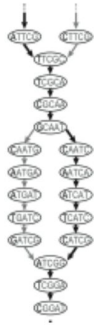
https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Trinity – How it works:

RNA-Seq reads → **Linear contigs** → **de-Bruijn graphs** →

>gene_01_isoform_01
>gene_02_isoform_01
>gene_02_isoform_02
>gene_03_isoform_01





...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

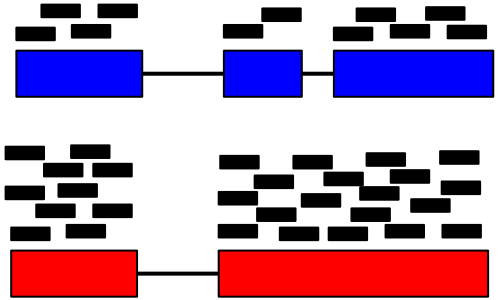
Thousands of disjoint graphs

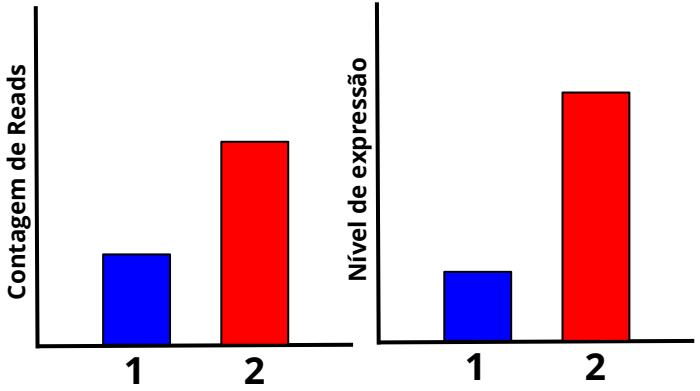
https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Montamos os transcritos, e agora?

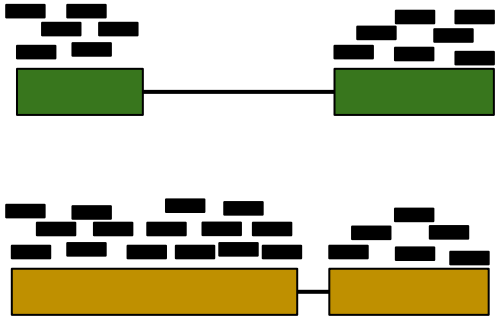
Contagem e estimação de abundância

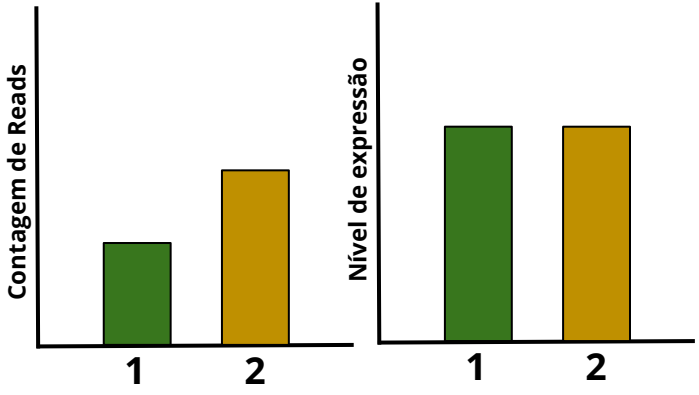
Mesmo Tamanho





Tamanhos Diferentes





https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Normalização dos dados

RPM = single-end

CPM = paired-end

RPM (**R**eads **P**er **M**illion mapped reads) = **CPM** (**C**ounts **P**er **M**illion mapped reads)

$$\text{RPM ou CPM} = \frac{\text{Número de reads mapeados nos genes} \times 10^6}{\text{Número total de reads mapeadas}}$$

Exemplo: Total de reads mapeadas 4 milhões e 5 mil reads pra um gene:

$$\text{RPM or CPM} = \frac{5000 \times 10^6}{4 \times 10^6} = 1250$$

https://www.reneshbedre.com/blog/expression_units.html

Normalização dos dados

RPKM = single-end

FPKM = paired-end

RPKM (**R**eads **P**er **K**ilo base per **M**illion mapped reads) = **FPKM** (**F**ragments **P**er **K**ilo base per **M**illion mapped reads)

$$\text{RPKM ou FPKM} = \frac{\text{Número de reads mapeados no gene} \times 10^3 \times 10^6}{\text{Número total de reads mapeadas} \times \text{Tamanho do gene em bp}}$$

Exemplo: Total de reads mapeadas 4 milhões e 5 mil reads pra um gene com tamanho de 2000 bp:

$$\text{RPKM} = \frac{5000 \times 10^3 \times 10^6}{4 \times 10^6 \times 2000} = 625$$

https://www.reneshbedre.com/blog/expression_units.html

Normalização dos dados

TPM (**T**ranscripts **P**er **M**illion)

- Permite comparar as bibliotecas (quantidade de reads totais) com tamanhos diferentes

$$TPM = \frac{RPKM \text{ ou } FPKM \times 10^6}{\Sigma \text{ RPKM ou FPKM}}$$

https://www.reneshbedre.com/blog/expression_units.html

Normalização dos dados

TMM (**T**rimmed **M**ean of **M**-values)

$$\log_2(TMM_k^{(r)}) = \frac{\sum_{g \in G*} w_{gk}^r M_{gk}^r}{\sum_{g \in G*} w_{gk}^r} \text{ where } M_{gk}^r = \frac{\log_2\left(Y_{gk}/N_k\right)}{\log_2\left(Y_{gr}/N_r\right)} \text{ and } w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}};$$

$Y_{gk}, Y_{gr} > 0.$

Robinson and Oshlack, 2010

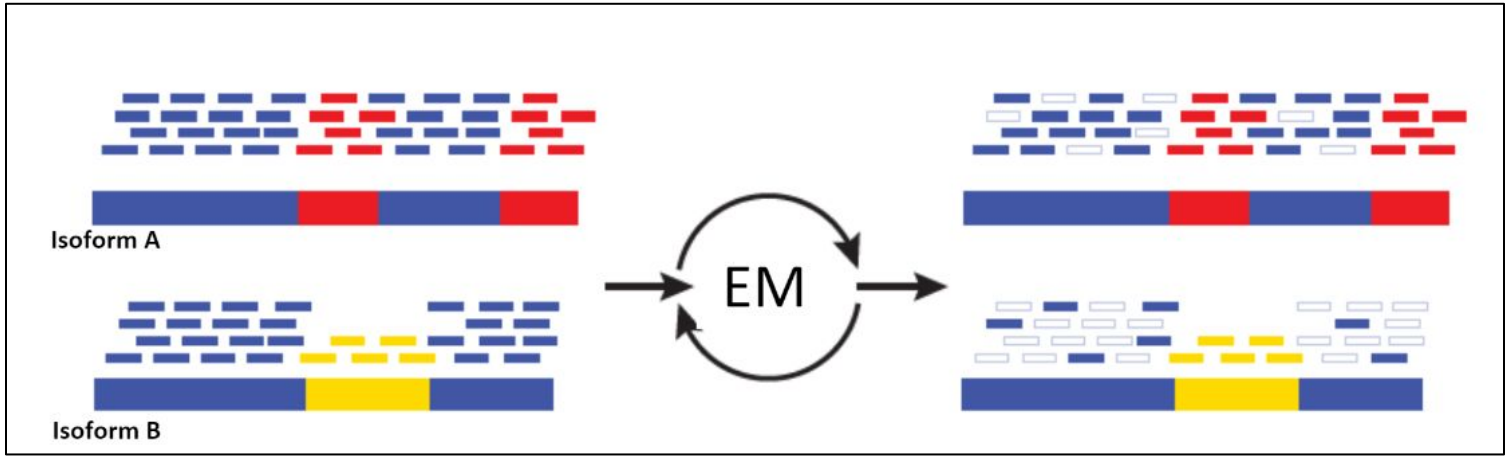
Normalização dos dados

TMM (**T**rimmed **M**ean of **M**-values)

- TMM é um método de normalização que permite comparações entre amostras diferentes;
- É uma boa escolha para remover os "*batch effects*" ao comparar as amostras de diferentes tecidos/genótipos ou nos casos em que a população de mRNA seria significativamente diferente entre as amostras.
- Batch effects = efeito do único sobre o todo;
- Alto valor de expressão de alguns genes = diminuição do valor de TPM dos outros;
- Retira-se os genes que são muito diferentes.

Robinson and Oshlack, 2010

Vamos utilizar o Salmon!



Azul = Reads multi-mapeadas

Vermelho e Amarelo = Reads únicas mapeadas

Usa-se o método *Expectation Maximization* (EM) para encontrar o melhor assinalamento das reads nos transcritos.

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Reads mapeadas, contadas e suas abundâncias estimadas: aplicar métodos estatísticos

Análises de expressão diferencial

Envolve:

- Contagem de reads;
- Testes de significância.

Vamos utilizar o edgeR:

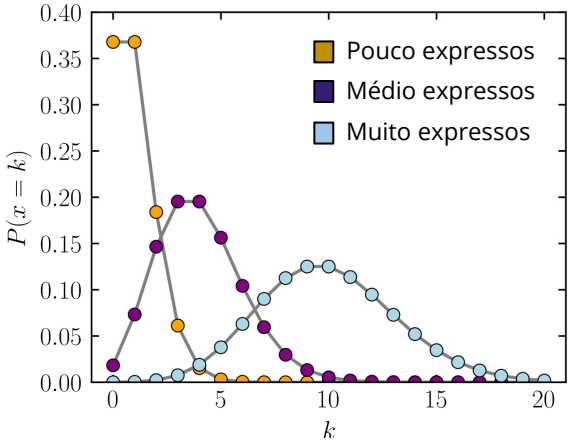
- As variações técnicas na contagem de reads em um experimento de RNA-Seq por *feature* é modelada pela Distribuição de Poisson:

$$P(X=k) = f(k;\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

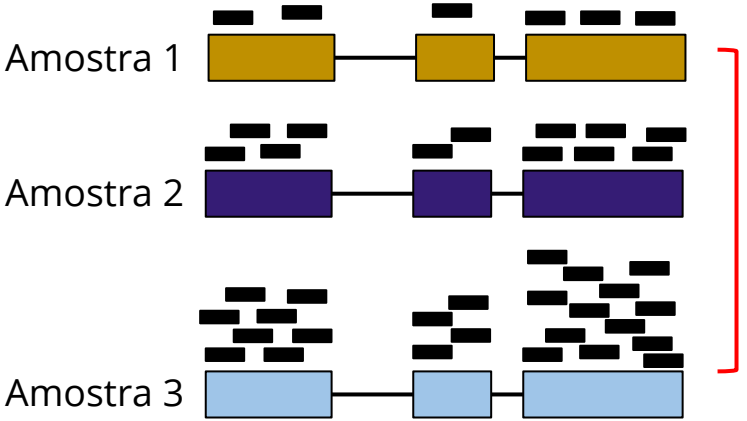
X = variável randômica;
k = número de ocorrências;
λ = representa o valor médio "esperado" de uma ocorrência se ela for repetida infinita vezes;
f(k;λ) = probabilidade de que ocorra k ocorrências dado λ.

edgeR

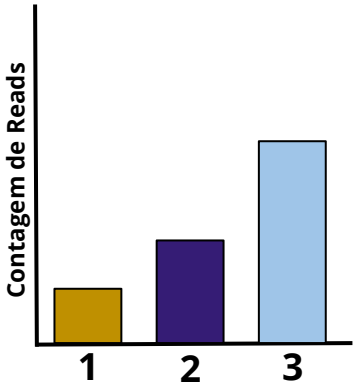
λ = média do número de reads



k = Número de reads



Avaliar a significância desta diferença na contagem de Reads

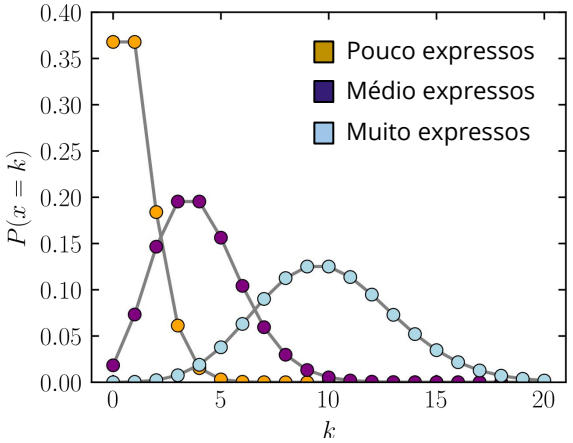


https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

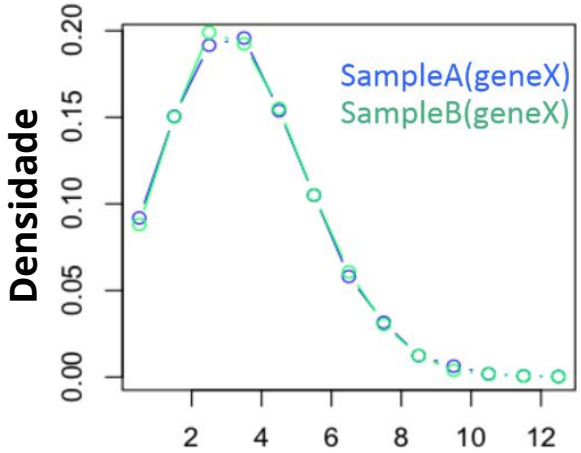
edgeR

amostra A (gene) = amostra B (gene) = 4 reads

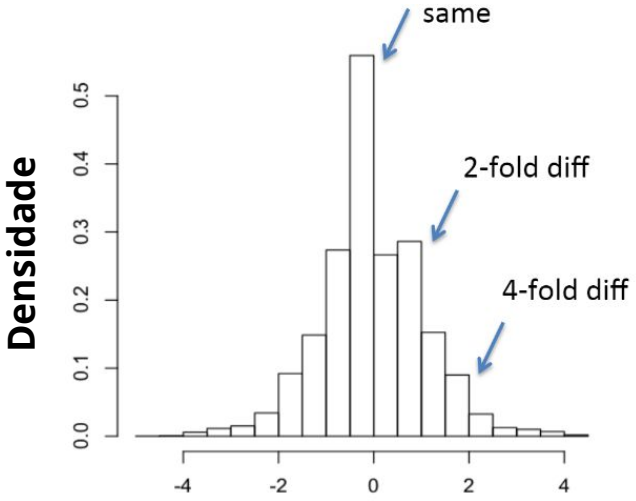
λ = média do número de reads



k = Número de reads



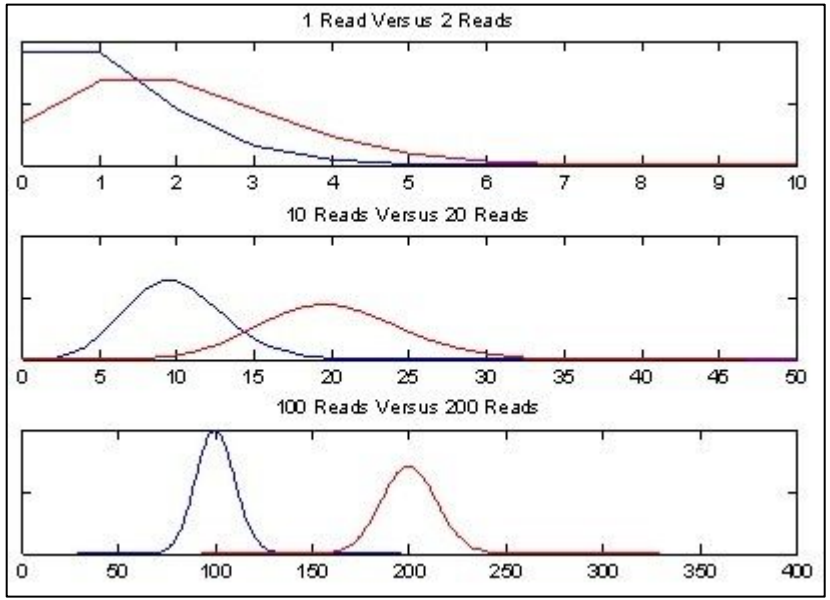
k = Número de reads



x = $\log_2$(amostraA/amostraB)

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

edgeR



Mais Reads = Maior o poder estatístico

Exemplo de 5000 reads por amostra:

	SampleA	Sample B	Fisher's Exact Test (P-value)
geneA	1	2	1.00
geneB	10	20	0.098
geneC	100	200	< 0.001

$$\frac{amostraA}{amostraB} = \frac{1}{2} = \frac{10}{20} = \frac{100}{200} = 0.5$$

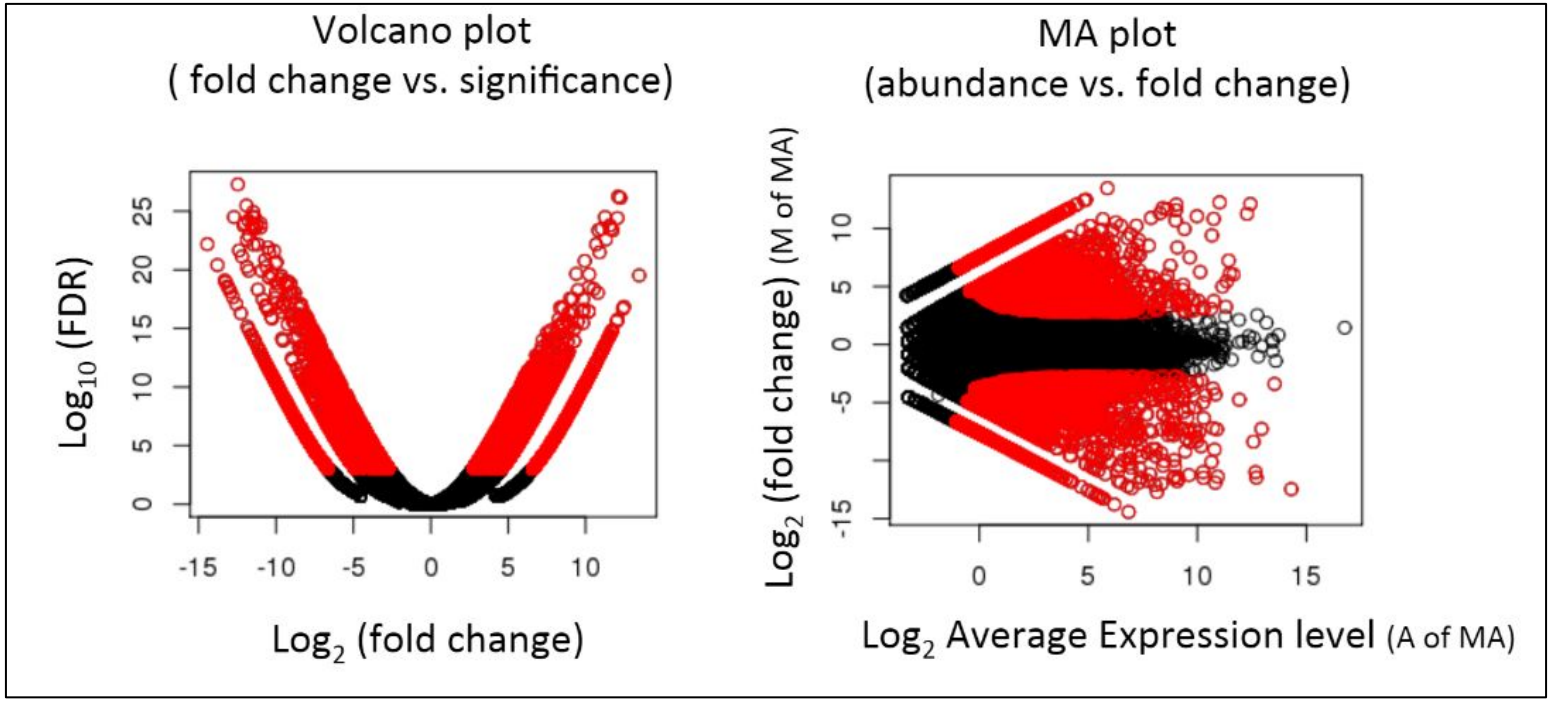
$$\frac{amostraB}{amostraA} = \frac{2}{1} = \frac{20}{10} = \frac{200}{100} = 2$$

$$\log_2 0.5 = -1$$
$$\log_2 2 = +1$$

https://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

edgeR

False Discovery Rate = probabilidade do gene não ser diferencialmente expresso. Representa a probabilidade do gene não ser diferencialmente expresso dado o seu p-valor e todos os outros testes estatísticos feitos. O p-valor não pode ser usado em condições com muito/pouco genes diferenciados.



https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Como descobrimos quem são os transcritos se não temos o genoma?

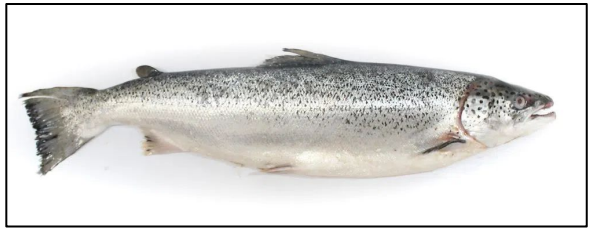
diamond

DIAMOND é um alinhador de sequências para proteínas e DNA traduzido, desenhado para ter alta performance em um banco de dados muito grande:

- Alinhamento pairwise com até 100x-10,000x mais velocidade que o BLAST.
- Baixa necessidade de recursos e adequado para execução em computadores pessoais.
- Diferentes formatos de saída, como por exemplo o formato BLAST, tabulado, e em formato XML.

March 26th	Dr Benjamin Buchfink, Max Planck Institute for Developmental Biology, Tübingen, Germany
Friday, 2PM	**Sensitive tree-of-life scale protein alignment using DIAMOND** Chair: Gabriel Fernandes, IRR Fiocruz Minas
	Link ID: 959 1978 7208 passwd: brazweb21

Montagens dos transcritos



Contagem e abundâncias dos genes

Testes Estatísticos

Anotação dos transcritos

Vias metabólicas

Rede de interações