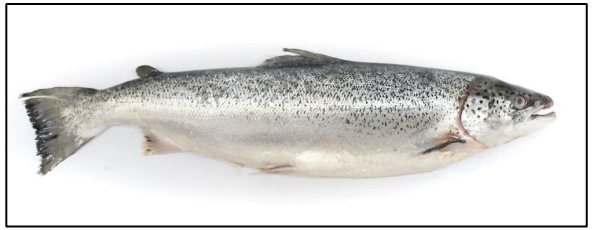
Montagem transcriptoma utilizando Trinity

Montagens dos transcritos



Contagem e abundâncias dos genes

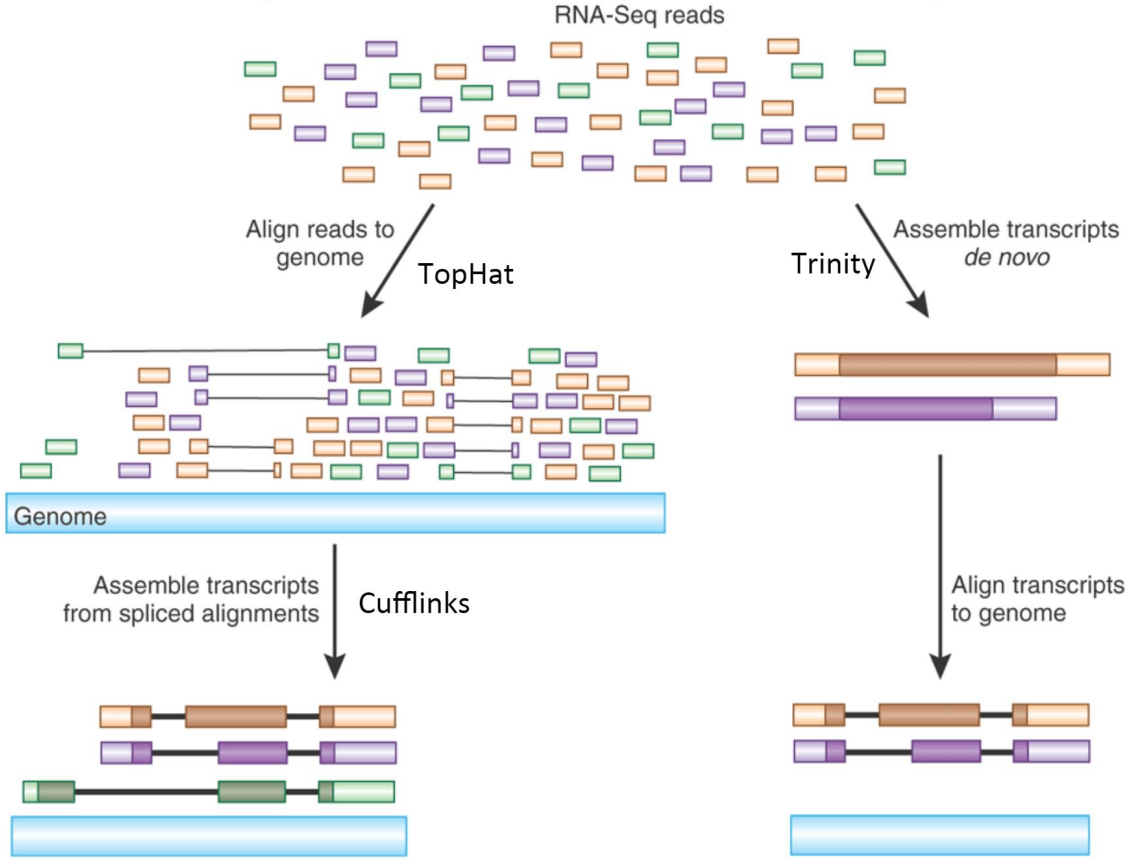
Testes Estatísticos

Anotação dos transcritos

Vias metabólicas

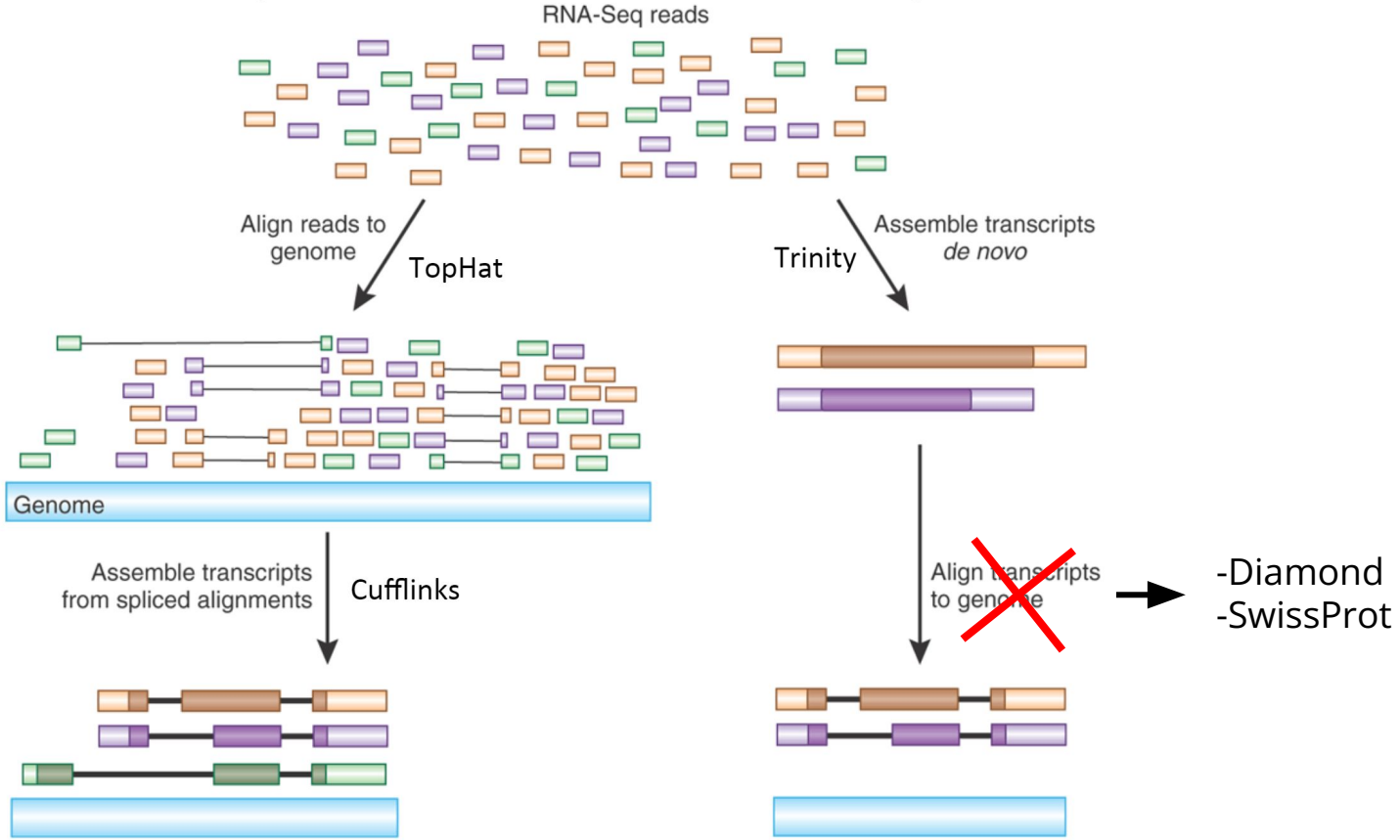
Rede de interações

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

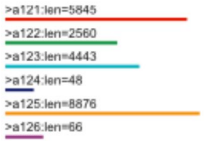


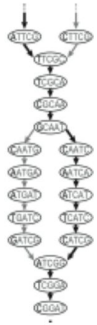
https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

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...

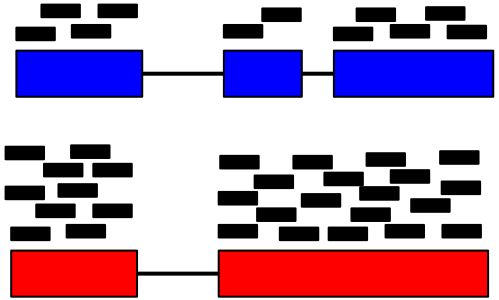
Thousands of disjoint graphs

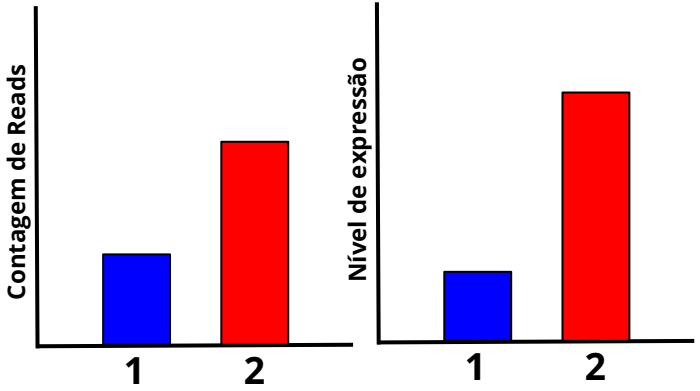
https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/blob/master/docs/rnaseq_workshop_slides.pdf

Montamos os transcritos, e agora?

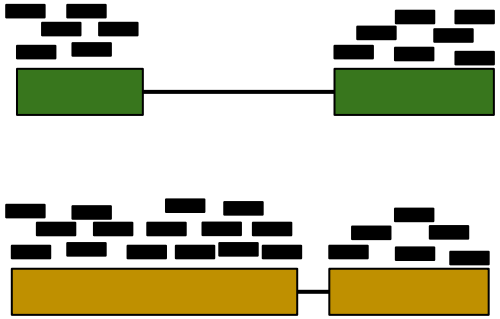
Contagem e estimação de abundância

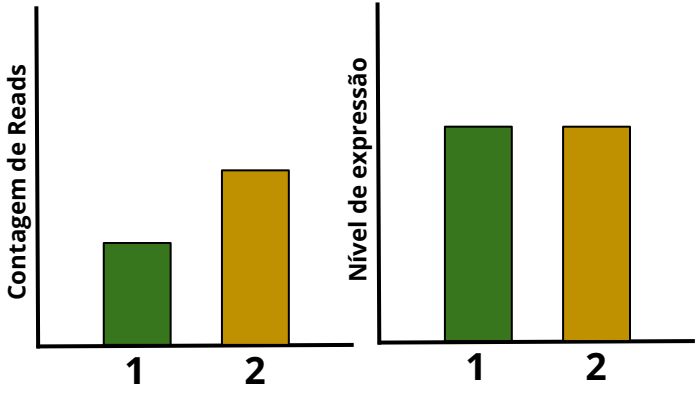
Mesmo Tamanho





Tamanhos Diferentes





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

Normalização dos dados

RPM = single-end	**RPKM** = single-end
CPM = paired-end	**FPKM** = paired-end

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

$$\text{RPM ou CPM} = \frac{\text{Número de reads mapeadas do gene} \times 10^6}{\text{Número total de reads mapeadas}}$$

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)

$$\text{RPKM ou FPKM} = \frac{\text{Número de reads mapeadas do gene} \times 10^3 \times 10^6}{\text{Número total de reads mapeadas} \times \text{Tamanho do gene em bp}}$$

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

Normalização dos dados

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

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

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

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

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

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

Normalização dos dados

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

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

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;
- Retira-se os genes que são muito diferentes.

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

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

Análises de expressão diferencial

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:

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

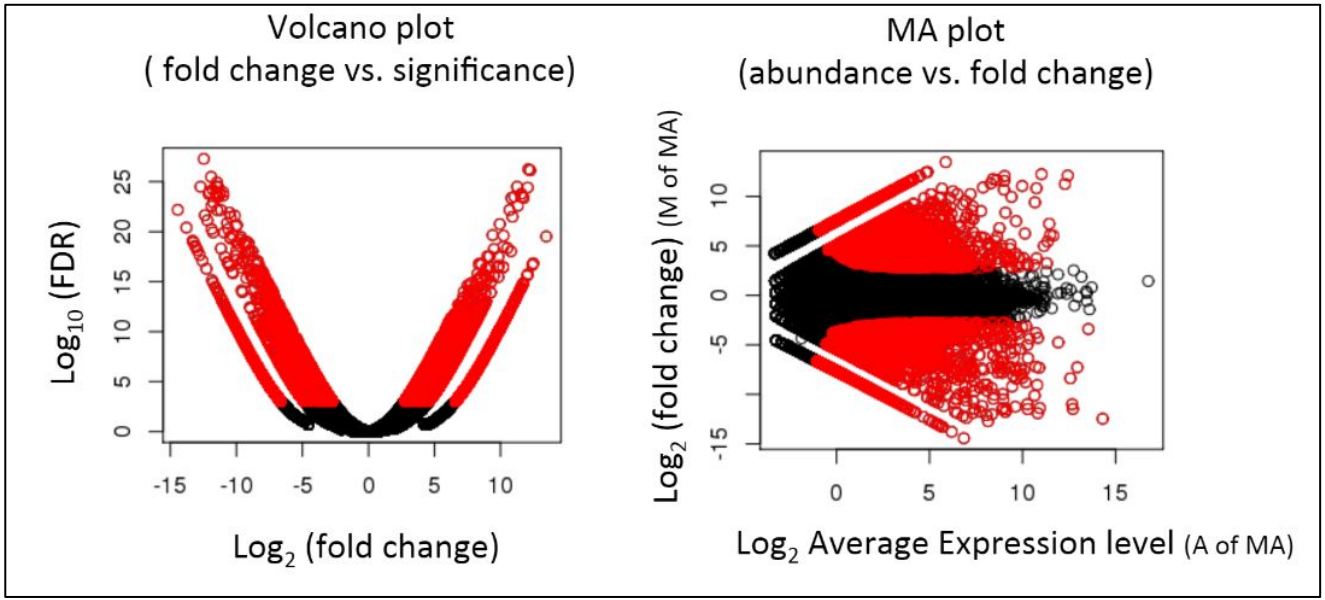
X = variável randômica;
k = número de ocorrências;
λ = representa o valor médio "esperado" de uma ocorrência se ela for repetida infinita vezes;
f(k;λ) = probabilidade de que ocorra k ocorrências dado λ.

Robinson et. al. (2010): edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

edgeR

FDR = FP / (FP + TP)

Método de Benjamini-Hochberg sobre o p-valor



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

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

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

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

- Até 100x-10,000x mais rapido que o BLAST+.