Análise de Expressão Diferencial

Normalização dos dados

Dr. Pablo Rodrigo Sanches

Departamento de Genética – FMRP/USP

psanches@usp.br

Roteiro de análise

1. DESeq2 (normalização e calculo da expressão diferencial)
   a) tratamento vs. controle
2. Obter os resultados:
   a) PCA
   b) Distância entre amostras
   c) Histograma de p valor
   d) Genes significativamente DE
   e) Tabelas de genes significativamente DE

Normalização

- Descobrir como os genes se expressam diferencialmente em diferentes condições ou tecidos.
- Número de reads é correlacionado com o nível de expressão do gene
- RNA-Seq oferece uma aproximação quantitativa da abundância dos transcritos na forma de contagens.
- Contagens precisam ser normalizadas
  - Comprimento das diferentes moléculas de RNAs
  - Profundidade do sequenciamento de diferentes bibliotecas

Normalização

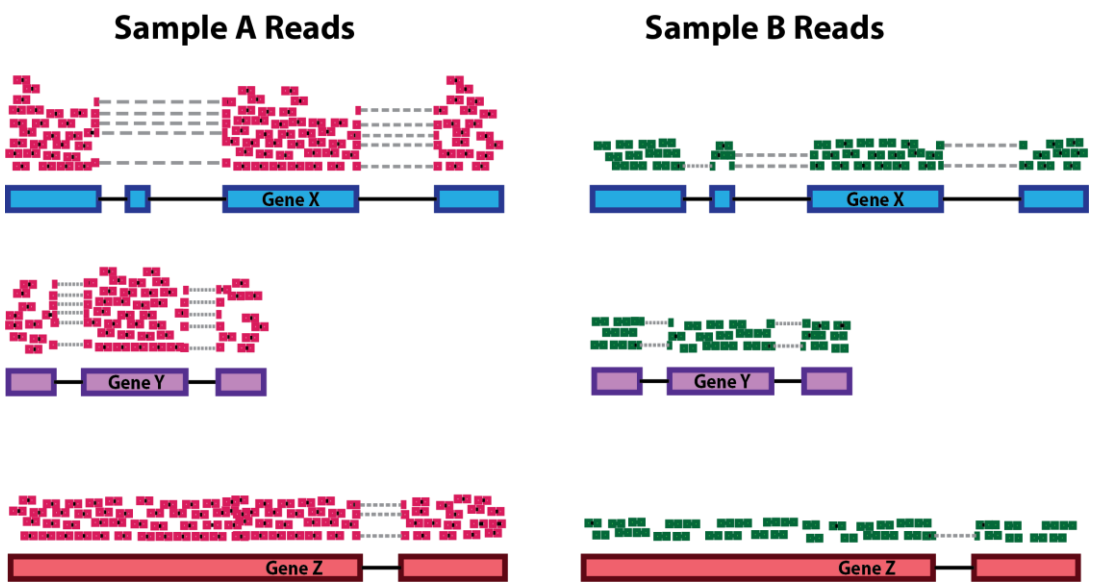
- Exemplo:
  - Biblioteca 1 → 12 milhões de reads mapeados
  - Biblioteca 2 → 16 milhões de reads mapeados

Loco	Tamanho loco (pb)	Nro. Reads Biblioteca 1	Nro. Reads Biblioteca 2
GeneA	800	24	38

É possível afirmar que temos maior expressão do GeneA na Biblioteca 2 quando comparada à Biblioteca 1 ???

Normalização

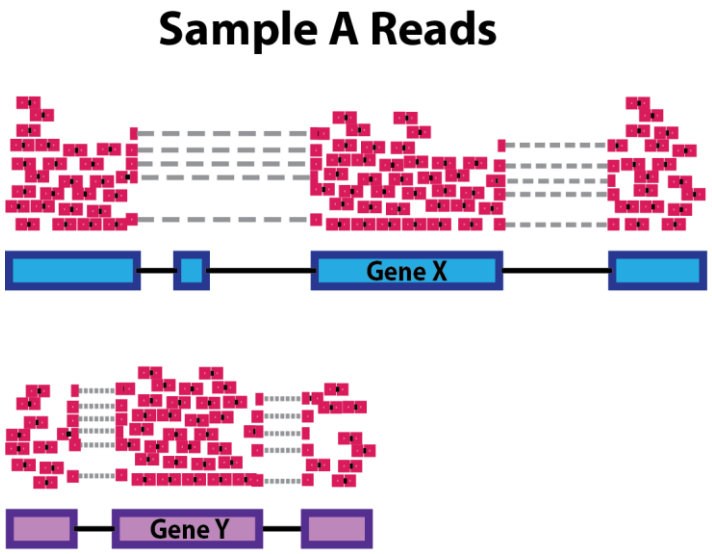
Profundidade do sequenciamento:



- No exemplo, cada gene parece ter dobrado a expressão na Amostra A em relação à Amostra B, no entanto, isso é uma consequência da Amostra A ter o dobro da profundidade de sequenciamento.

Normalização

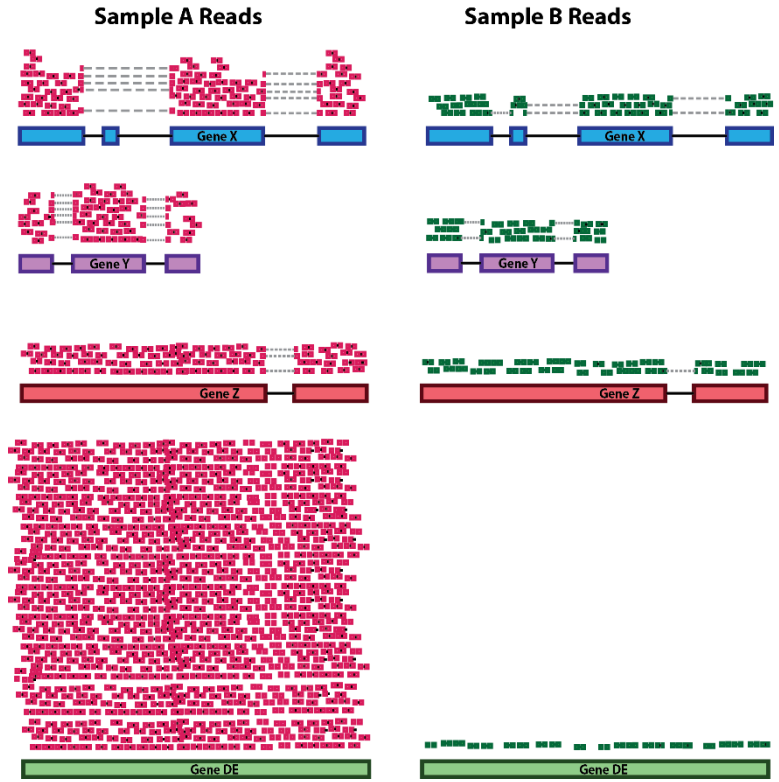
Comprimento do gene:



- No exemplo, *Gene X* e *Gene Y* têm níveis semelhantes de expressão, mas o número de leituras mapeadas para *Gene X* seria muito maior do que o número mapeado para *Gene Y* porque *Gene X* é mais longo.

Normalização

Composição de RNA:



- No exemplo, se fosse para dividir cada uma das amostras pelo número total de contagens para normalização, as contagens seriam muito enviesadas pelo Gene DE (ocupa a maior parte das contagens para uma das amostras).

DESeq2 – Mediana das razões

- O DESeq2 executa uma normalização onde a média geométrica é calculada para cada gene em todas as amostras. A contagem de um gene em cada amostra é então dividida por essa média. A mediana dessas proporções em uma amostra é o fator de tamanho dessa amostra. Este procedimento corrige o tamanho da biblioteca e o viés da composição do RNA, que pode surgir, por exemplo, quando apenas um pequeno número de genes são altamente expressos em uma condição de experimento, mas não na outra.

Exemplo de aplicação

	amostraA	amostraB	referência	amostraA/referência	amostraB/referência	normalizadoA	normalizadoB	FC	log2FC
geneA	2800	1000	1673,32	1,67	0,60	3429,29	816,50	0,24	-2,07
geneB	30	15	21,21	1,41	0,71	36,74	12,25	0,33	-1,58
geneC	500	750	612,37	0,82	1,22	612,37	612,37	1,00	0,00
geneD	40	80	56,57	0,71	1,41	48,99	65,32	1,33	0,42
geneE	500	1200	774,60	0,65	1,55	612,37	979,80	1,60	0,68
			Mediana	0,82	1,22				