Cálculo de expressão de transcritos

Leandro Costa do Nascimento

25/07/2015

l.costa.nascimento@gmail.com

Conceitos

- Read:
- Fragmento:
- Read count: número de reads/fragmentos que alinharam com um transcrito/gene em uma amostra.

Transcriptômica

- Identificação e anotação de novos transcritos (Montagem).
- Comparar a expressão de transcritos diferentes em uma amostra.
- Comparar a expressão de um mesmo transcrito entre amostras diferentes.
- Mas, após montar os transcritos, como calculo a expressão deles em cada uma das minhas amostras?

Expressão com RNA-Seq

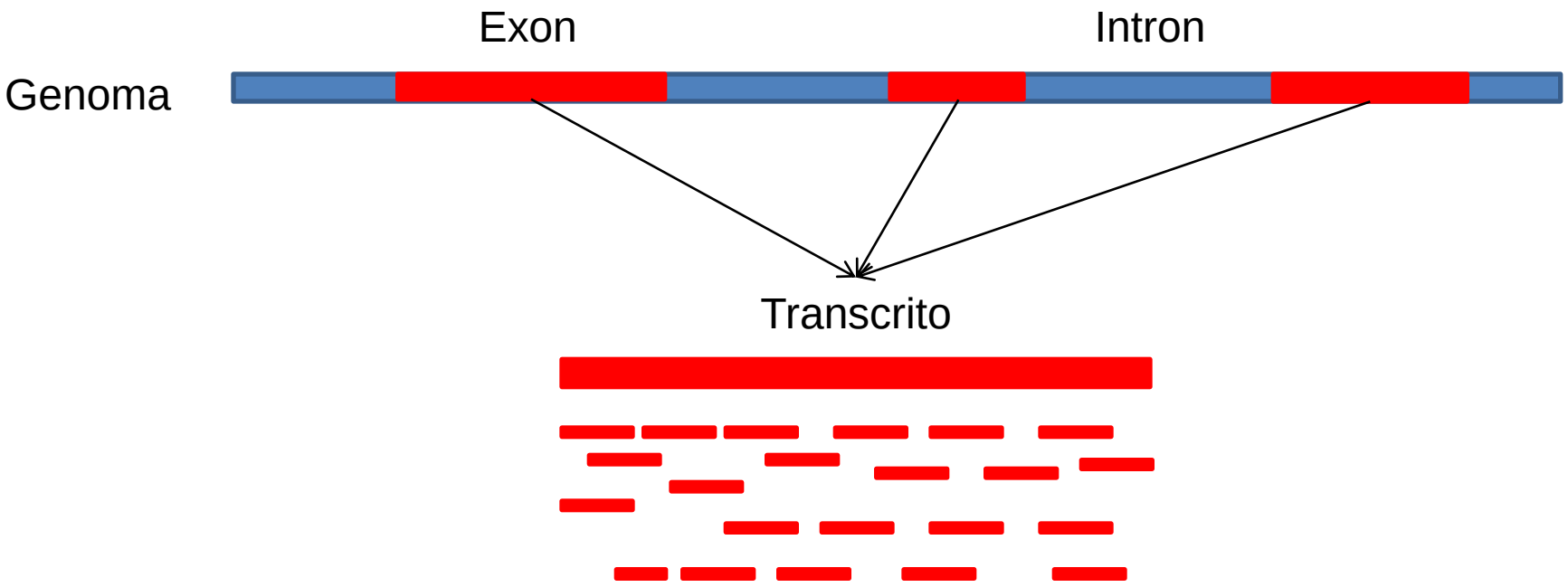
- Os valores de expressão em RNA-Seq são baseados no número de reads que mapeiam/alinham em um gene/transcrito.
- Dois tipos de alinhamento:

  - Alinhamento padrão

  - Alinhamento com splicing

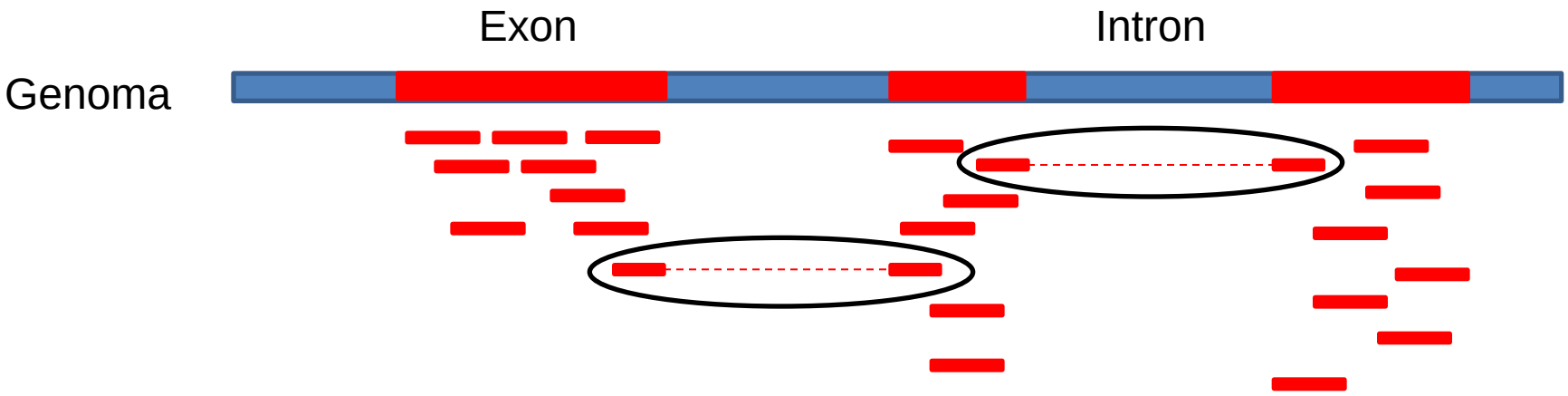
Alinhamento padrão

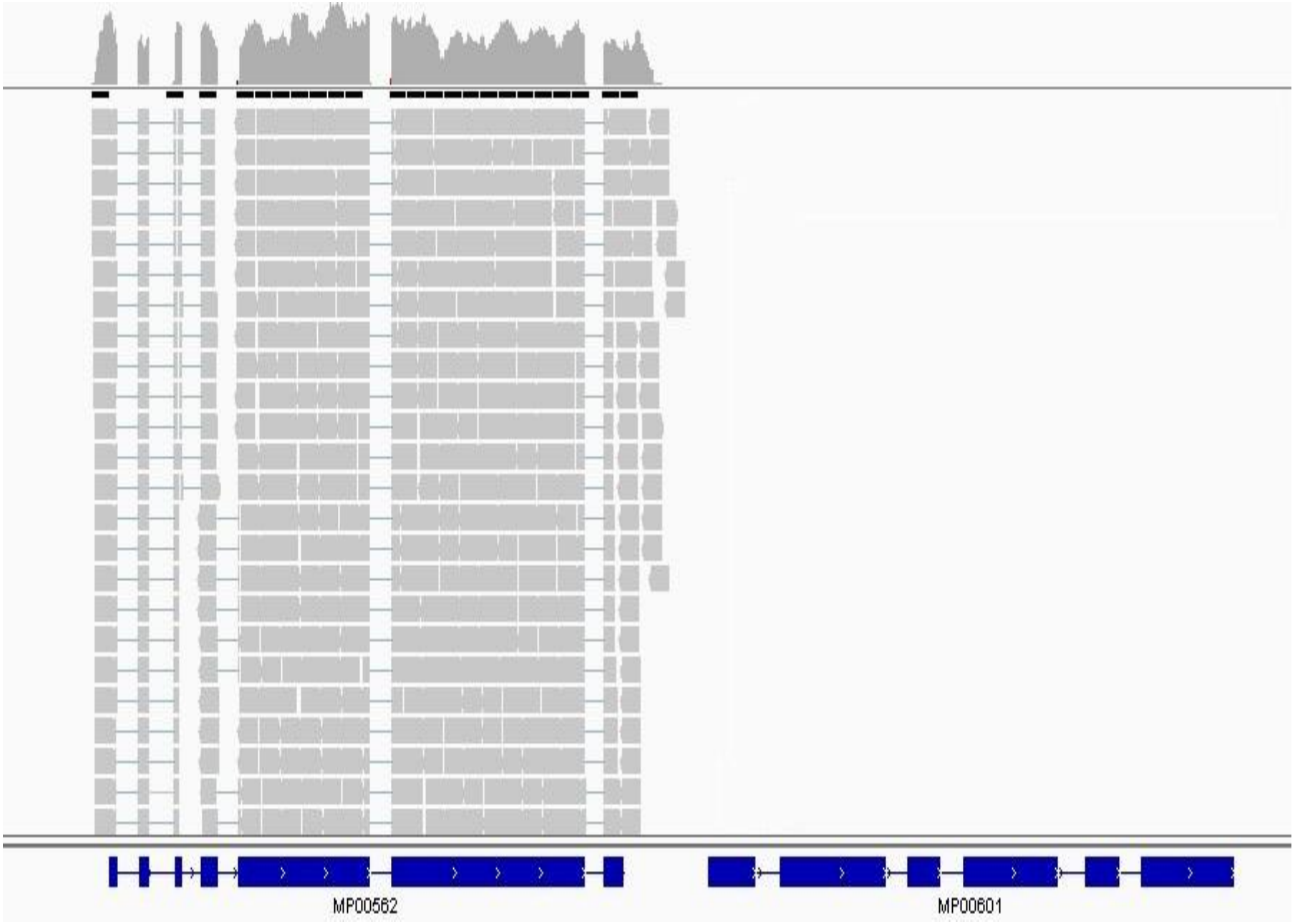
- Utilizado para alinhamento de reads contra o transcrito montados.
- Exemplos: Bowtie, SOAP, RSEM.



Alinhamento com splicing

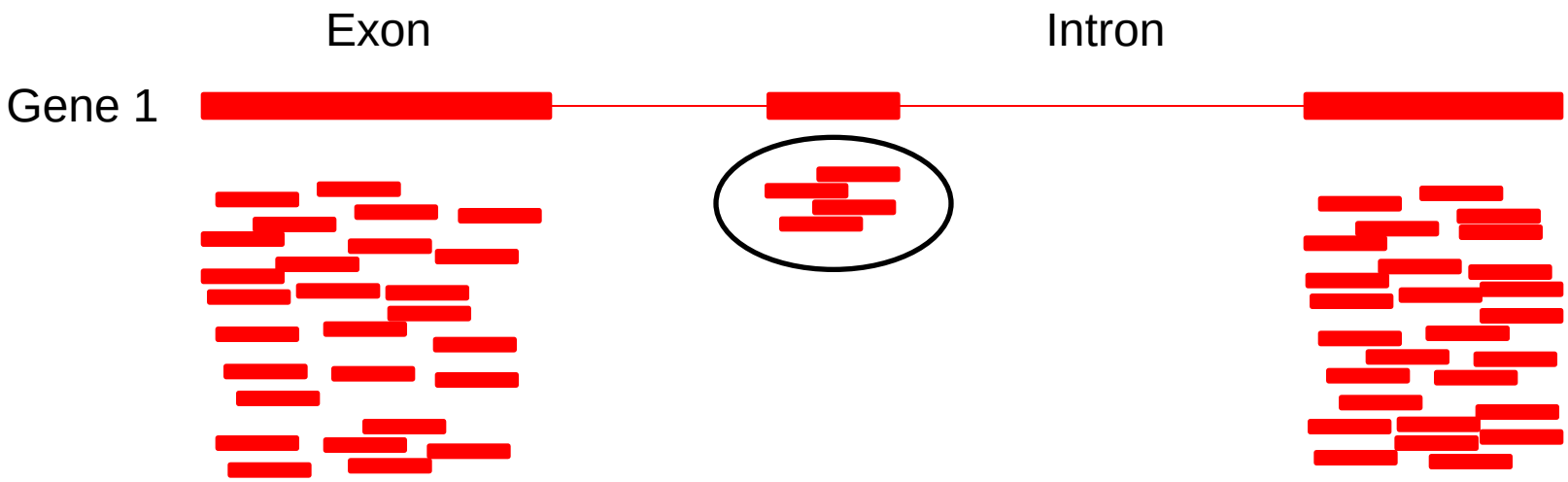
- Utilizado para alinhamento de reads contra o genoma.
- Exemplo: TopHat




MP00562
MP00601

Expressão – nível de gene

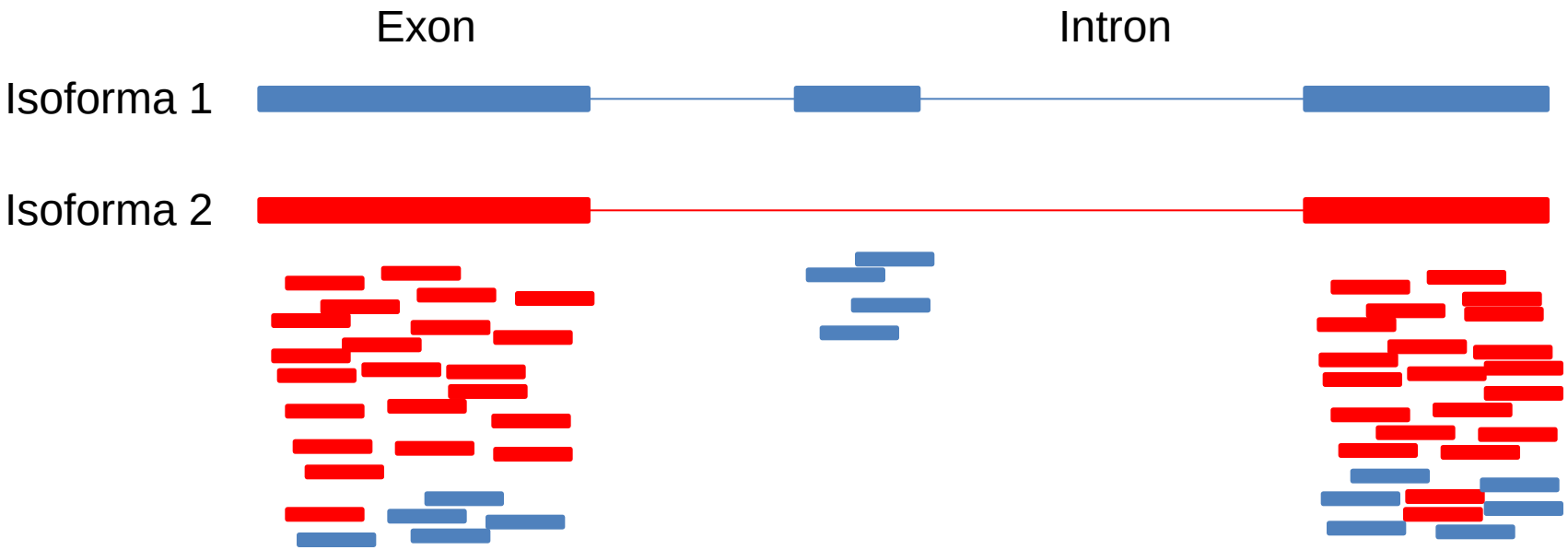
- Considera todos os reads alinhados no lócus "Gene 1" para calcular a expressão.



Gene	Read count
Gene 1	85

Expressão – nível de isoforma

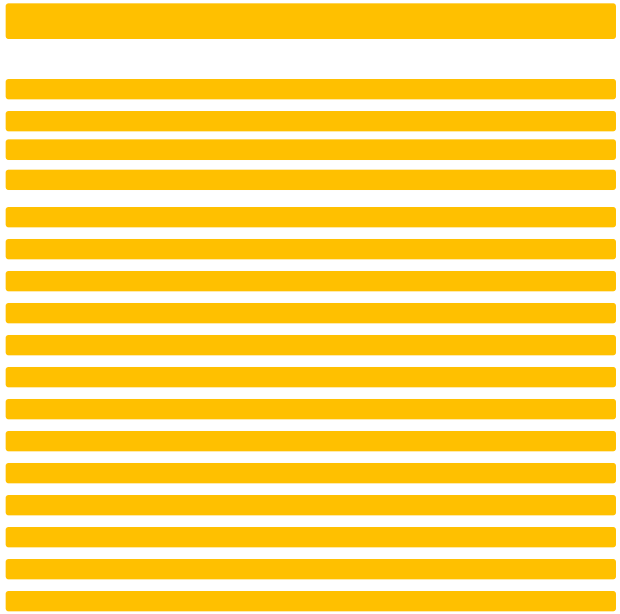
- Dividir os reads entre todas as isoformas do lócus “Gene 1” e calcular a expressão de cada.



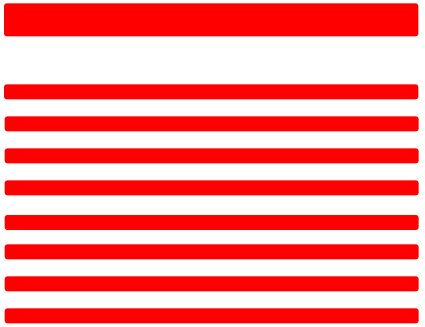
Gene	Isoforma	Read count
Gene X	Isoforma A	24
	Isoforma B	61

Expressão com RNA-Seq

- Comparar a expressão de transcritos diferentes



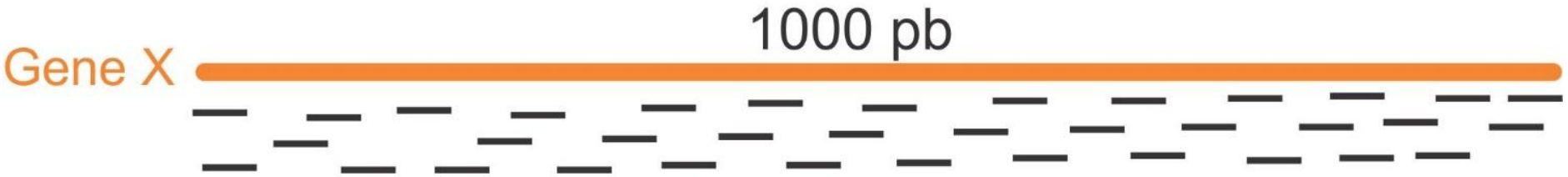
Transcrito 1 - Controle

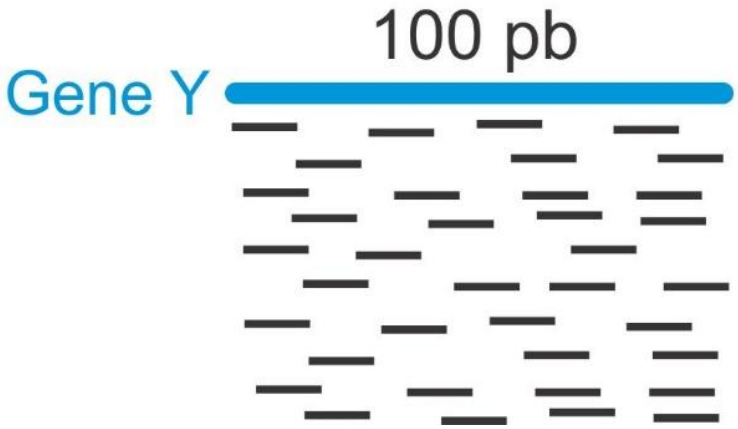


Transcrito 2 - Controle

Expressão com RNA-Seq

- Importante: genes ou transcritos tem tamanhos diferentes

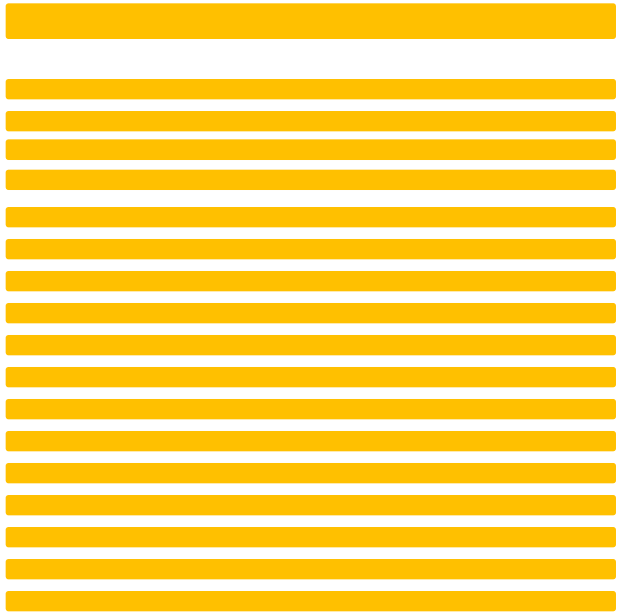




Gene	Amostra A
Gene X	36
Gene Y	36

Expressão com RNA-Seq

- Comparar a expressão do mesmo transcrito em amostras diferentes.



Transcrito 1 - Controle

Transcrito 1 - Tratado

Expressão com RNA-Seq

- Importante: número de reads sequenciados em cada biblioteca é diferente.

Amostra	Total reads
Amostra A	15 milhões
Amostra B	20 milhões

Gene	Amostra A	Amostra B
Gene Y	36	36

RPKM/FPKM

- Reads/Fragmentos por Kilobase de exon por Milhão de reads mapeados.

$$RPKM = [R/(T * N)] * 10^9$$

R – nº de reads/fragmentos que alinharam com o transcrito/gene.

T – tamanho efetivo do transcrito/gene (em bp).

N – total de reads mapeado na amostra.

Média de expressão relativa

- Em duas amostras, a média de expressão relativa dos transcritos é a mesma.
- Exemplo: amostra X com 100 transcritos e amostra Y com 500 transcritos no total (5 diferentes).

Amostra	#Total transcritos	A	B	C	D	E
X	100	80	10	6	3	1
Y	500	20	20	10	50	400

Média de expressão relativa

- Se calcularmos a expressão relativa de cada transcrito em cada amostra, teremos:

  - Média da expressão relativa amostra X: (0,8 + 0,1 + 0,06 + 0,03 + 0,01)/5 = 0,2

  - Média da expressão relativa amostra Y: (0,04 + 0,04 + 0,02 + 0,1 + 0,8)/5 = 0,2

Amostra	A	B	C	D	E
X	0,8	0,1	0,06	0,03	0,01
Y	0,04	0,04	0,02	0,1	0,8

Incosistência do FPKM

- Se calcularmos o FPKM de cada transcrito em cada amostra, teremos:

  - Média do FPKM amostra X: (8.000 + 2.000 + 2.400 + 6.000 + 10.000)/5 = 5.680

  - Média do FPKM amostra Y: (400 + 800 + 800 + 20.000 + 800.000)/5 = 164.400

Amostra	#Total reads	FPKM A	FPKM B	FPKM C	FPKM D	FPKM E
X	100	8.000	2.000	2.400	6.000	10.000
Y	500	400	800	800	20.000	800.000

TPM

- Transcripts per million.

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6$$

Amostra	#Total reads	TPM A	TPM B	TPM C	TPM D	TPM E
X	100	281690,14	70422,53	84507,04	211267,60	352112,67
Y	500	486,61	973,23	973,23	24330,90	973236,00

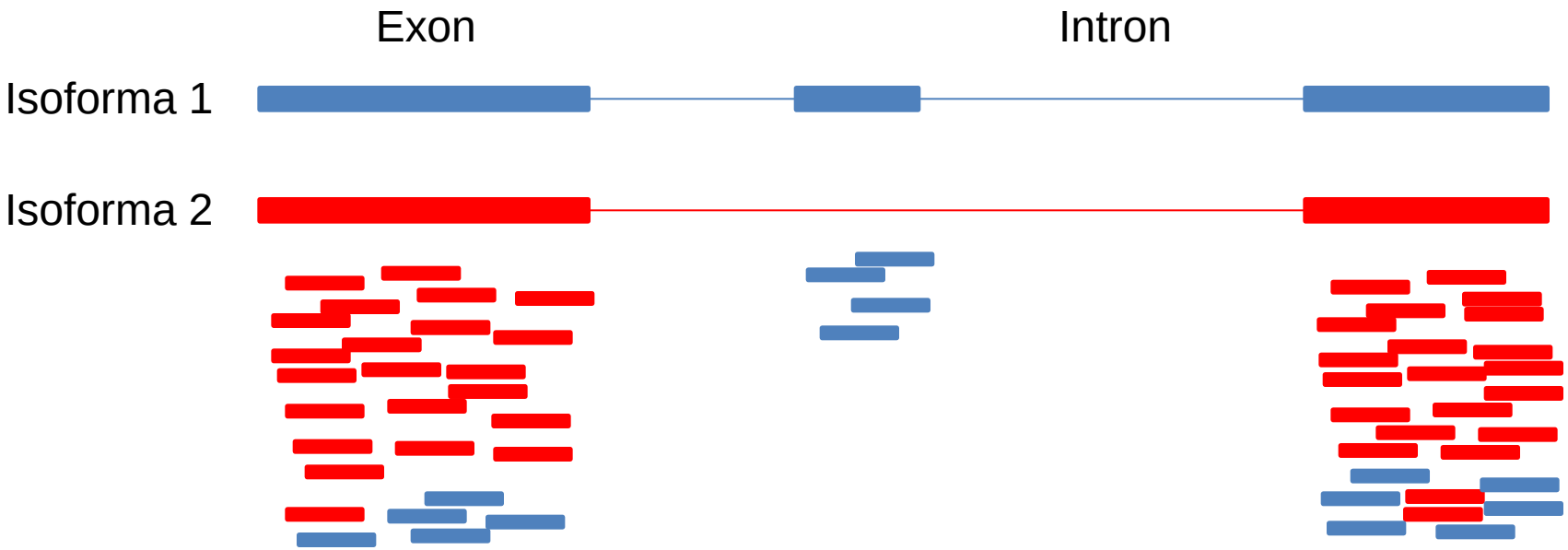
A média do TPM para as duas amostras é ~200.000

RSEM

- Utiliza o bowtie.
- Calcula a expressão a nível de gene e de transcrito.
- Retorna valores de read count, FPKM, TPM.
- Utiliza reads com alinhamento único para tentar separar os com alinhamento múltiplo.

Expressão – nível de isoforma

- Dividir os reads entre todas as isoformas do lócus “Gene 1” e calcular a expressão de cada.



Gene	Isoforma	Read count
Gene X	Isoforma A	24
	Isoforma B	61

RSEM – Preparando a referência

```
NAME
    rsem-prepare-reference

SYNOPSIS
    rsem-prepare-reference [options] reference_fasta_file(s) reference_name
```

- reference_fasta_file: arquivo FASTA contendo os transcritos de referência.
- reference_name: nome que você quer dar para o seu banco de dados.
- --transcript-to-gene-map: arquivo texto que informa de qual lócus é cada transcrito.

Ligando os transcritos aos genes

TR1	TR1|c0_g1_i1
TR2	TR2|c0_g1_i1
TR2	TR2|c0_g1_i2
TR3	TR3|c0_g1_i1
TR3	TR3|c0_g2_i1
TR4	TR4|c0_g1_i1
TR4	TR4|c0_g2_i1
TR4	TR4|c0_g2_i2
TR4	TR4|c0_g2_i3
TR4	TR4|c0_g2_i4
TR5	TR5|c0_g1_i1
TR5	TR5|c0_g1_i2
TR5	TR5|c0_g1_i3
TR5	TR5|c0_g1_i4
TR6	TR6|c0_g1_i1
TR6	TR6|c0_g1_i2
TR7	TR7|c0_g1_i1
TR7	TR7|c0_g2_i1
TR7	TR7|c1_g1_i1
TR8	TR8|c0_g1_i1
TR9	TR9|c0_g1_i1
TR9	TR9|c0_g2_i1
TR10	TR10|c0_g1_i1
TR10	TR10|c0_g2_i1
TR11	TR11|c0_g1_i1

- Arquivo tabular com duas colunas:
  - Nome do lócus
  - Nome do transcrito
- Um transcrito por linha.

RSEM – Quantificando a expressão

rsem–calculate–expression [opções] FASTQ–files [index] [output]

- paired–end: trabalha com reads paired–end.
- p: número de processadores.
- Index: nome do índice.
- Output: prefixo dos arquivos de saída.

Kallisto – Preparando a referência

Kallisto index [opções] FASTA–file

- –i: Nome do indíce
- –k: Tamanho de k mer (Default: 31; Máximo: 31)

Kallisto – Quantificando a expressão

Kallisto quant [opções] [FASTQ–files]

- i: Nome do indíce (obrigatório)
- o: Nome do diretório de saída (obrigatório)
- single: single –end reads
- bias: parâmetro correção

DESeq - Normalização

Genes	Amostra A Rep 1	Amostra A Rep 2	Amostra B Rep 1	Amostra B Rep 2	Média geométrica
Gene A	15	13	17	14	14.68
Gene B	452	430	420	444	436.32
Gene C	1024	1053	1002	987	1016.20
Gene D	120	132	523	582	263.51
Gene E	74	62	55	63	63.14

Calcula a média geométrica dos read counts de cada gene/transcrito

DESeq - Normalização

Genes	Amostra A Rep 1	Amostra A Rep 2	Amostra B Rep 1	Amostra B Rep 2
Gene A	1.02	0.89	1.16	0.95
Gene B	1.04	0.99	0.96	1.02
Gene C	1.01	1.04	0.99	0.97
Gene D	0.46	0.50	1.98	2.21
Gene E	1.17	0.98	0.87	1.00

Divide cada read count pela média geométrica do transcrito.

DESeq - Normalização

Genes	Amostra A Rep 1	Amostra A Rep 2	Amostra B Rep 1	Amostra B Rep 2
Gene A	1.02	0.89	1.16	0.95
Gene B	1.04	0.99	0.96	1.02
Gene C	1.01	1.04	0.99	0.97
Gene D	0.46	0.50	1.98	2.21
Gene E	1.17	0.98	0.87	1.00
Mediana	1.02	0.98	0.99	1.00

Calcula a mediana dos novos valores para cada amostra.

DESeq - Normalização

Genes	Amostra A Rep 1	Amostra A Rep 2	Amostra B Rep 1	Amostra B Rep 2
Gene A	15.33	12.76	16.76	13.97
Gene B	461.93	422.21	414.13	442.99
Gene C	1046.50	1033.92	988.00	984.75
Gene D	122.64	129.61	515.69	580.67
Gene E	75.63	60.88	54.23	62.86

Multiplica os reads counts originais pela mediana de cada amostra.