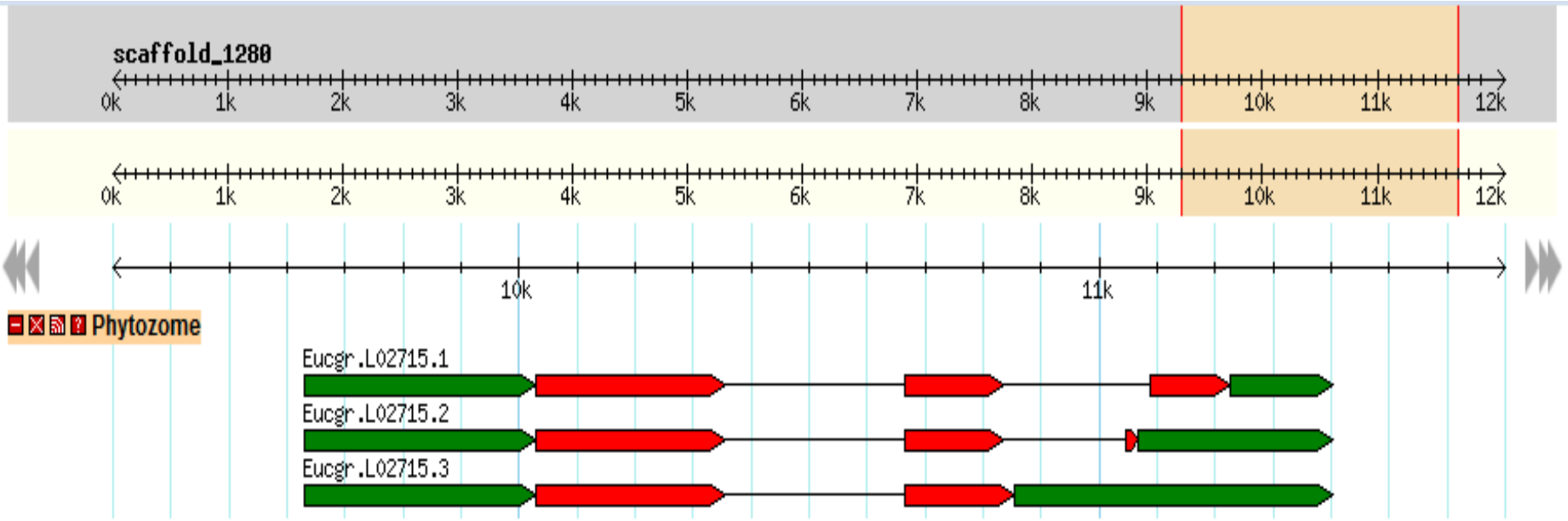
Montagem de RNA-Seq com genoma de referência

Leandro Costa do Nascimento

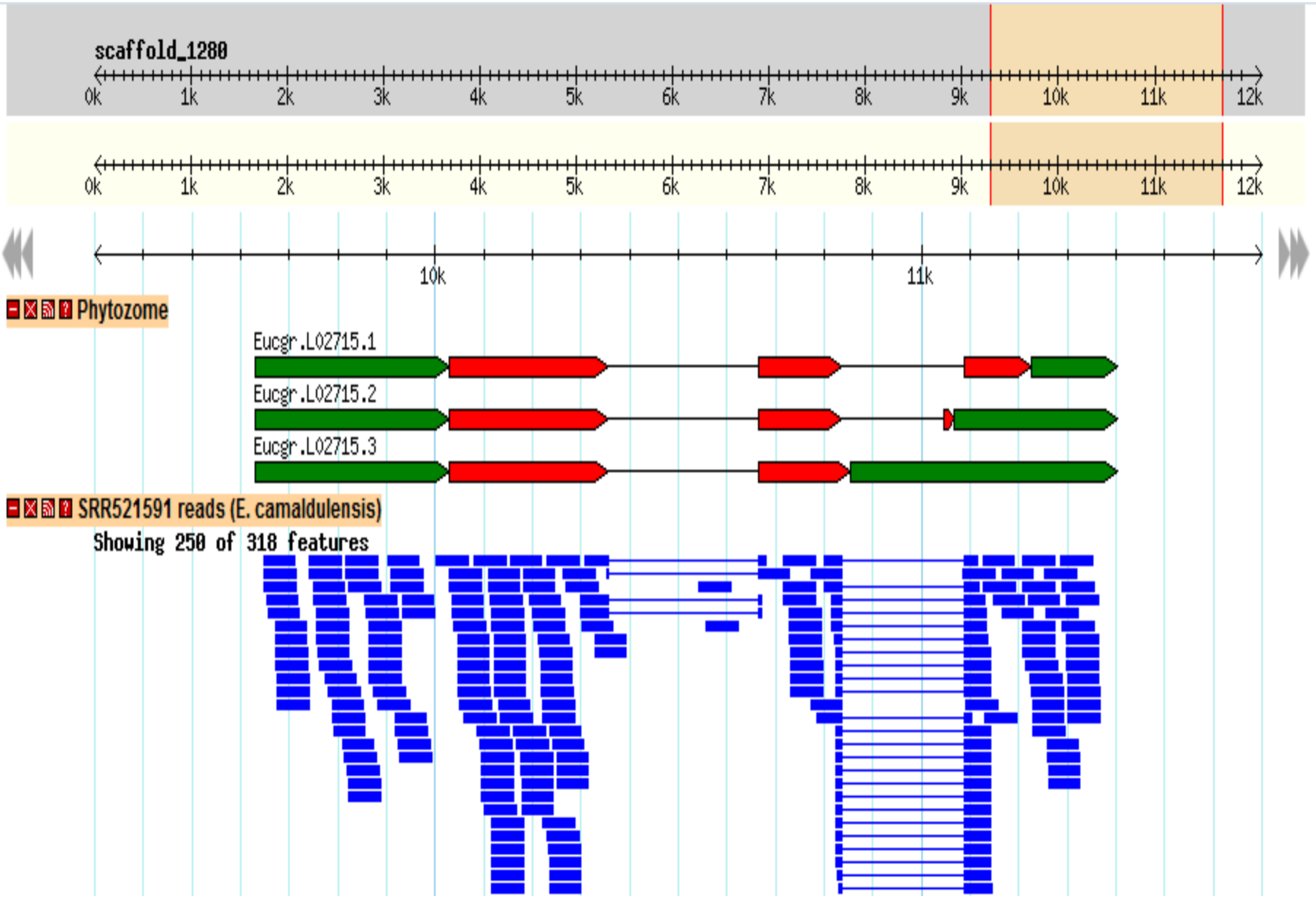
leandro@lge.ibi.unicamp.br

l.costa.nascimento@gmail.com

21/07/2015



- Scaffold1280: sequência genômica
- Eucgr.L02715: lócus gênico
- Eucgr.L02715.1, Eucgr.L02715.2 e Eucgr.L02715.3: transcritos do lócus


scaffold_1280
0k
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k
11k
12k
10k
11k
Phytozome
Eucgr.L02715.1
Eucgr.L02715.2
Eucgr.L02715.3
SRR521591 reads (E. camaldulensis)
Showing 250 of 318 features

GFF/GTF

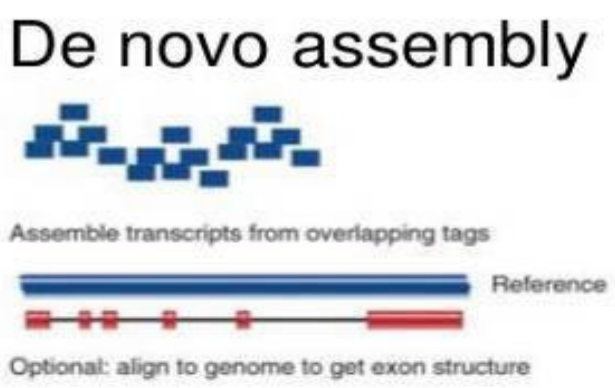
- Arquivos tabulares (separados por tabs).
- Contém 9 colunas, representando uma “feature” por linha.
- Cada “feature” pode representar um gene, um transcrito, exon etc.
- Em resumo, contém as posições dos genes já conhecidos no genoma de referência.

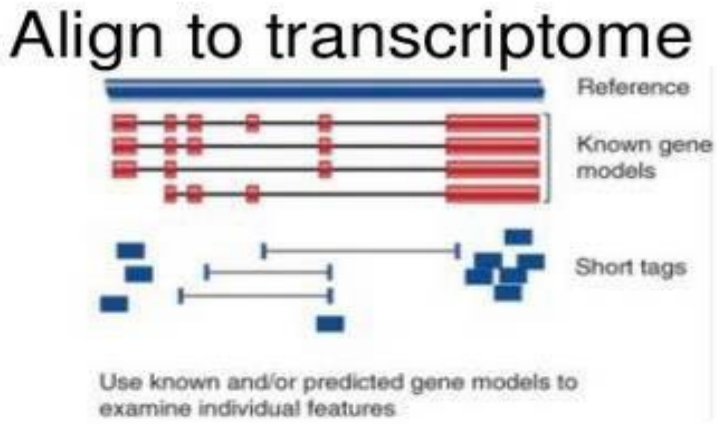
GFF/GTF

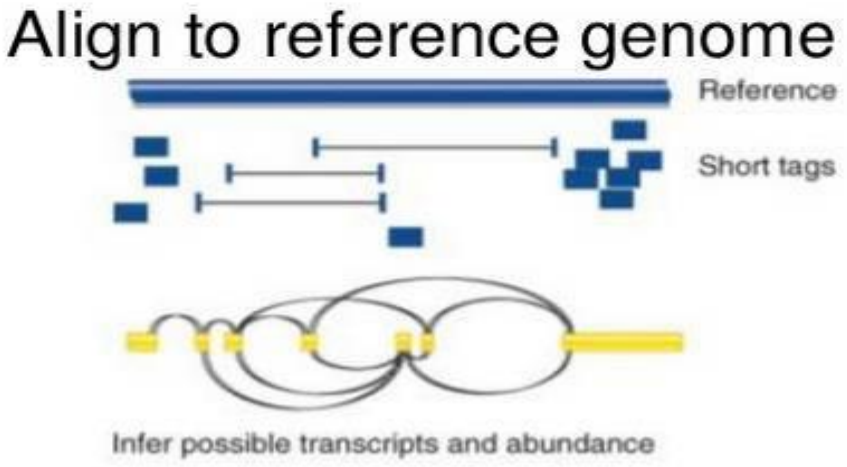
- Coluna 1: Sequência genômica.
- Coluna 2: Tipo de dado ou programa utilizado.
- Coluna 3: Feature (Gene, transcrito, CDS).
- Coluna 4: Posição inicial da feature na sequência genômica.
- Coluna 5: Posição final
- Coluna 6: Score.
- Coluna 7: Strand (foward ou reverse).
- Coluna 8: Frame.
- Coluna 9: Atributos (nome, id, etc).

```
scaffold_1	phytozome7	gene	12024	17350	.	+	.	ID=Eucgr.A00001;Name=Eucgr.A00001;
scaffold_1	phytozome7	mRNA	12024	17350	.	+	.	ID=PAC:18799928;Name=Eucgr.A00001.1;pacid=
18799928;Parent=Eucgr.A00001;
scaffold_1	phytozome7	five_prime_UTR	12024	12154	.	+	.	ID=PAC:18799928.five_prime_UTR.1;P
arent=PAC:18799928;pacid=18799928;
scaffold_1	phytozome7	five_prime_UTR	12668	12710	.	+	.	ID=PAC:18799928.five_prime_UTR.2;P
arent=PAC:18799928;pacid=18799928;
scaffold_1	phytozome7	CDS	12711	12896	.	+	0	ID=PAC:18799928.CDS.1;Parent=PAC:18799928;
pacid=18799928;
scaffold_1	phytozome7	CDS	12991	13086	.	+	0	ID=PAC:18799928.CDS.2;Parent=PAC:18799928;
pacid=18799928;
scaffold_1	phytozome7	CDS	13275	13395	.	+	0	ID=PAC:18799928.CDS.3;Parent=PAC:18799928;
pacid=18799928;
scaffold_1	phytozome7	CDS	13484	13590	.	+	2	ID=PAC:18799928.CDS.4;Parent=PAC:18799928;
pacid=18799928;
scaffold_1	phytozome7	CDS	13670	13887	.	+	0	ID=PAC:18799928.CDS.5;Parent=PAC:18799928;
pacid=18799928;
scaffold_1	phytozome7	CDS	14202	14379	.	+	1	ID=PAC:18799928.CDS.6;Parent=PAC:18799928;
pacid=18799928;
```

Estratégias para análise de RNA-Seq







Simplifying complexity (Cloonan and Grimmond, 2010)

Qual a melhor estratégia?

- Montagem *de novo*

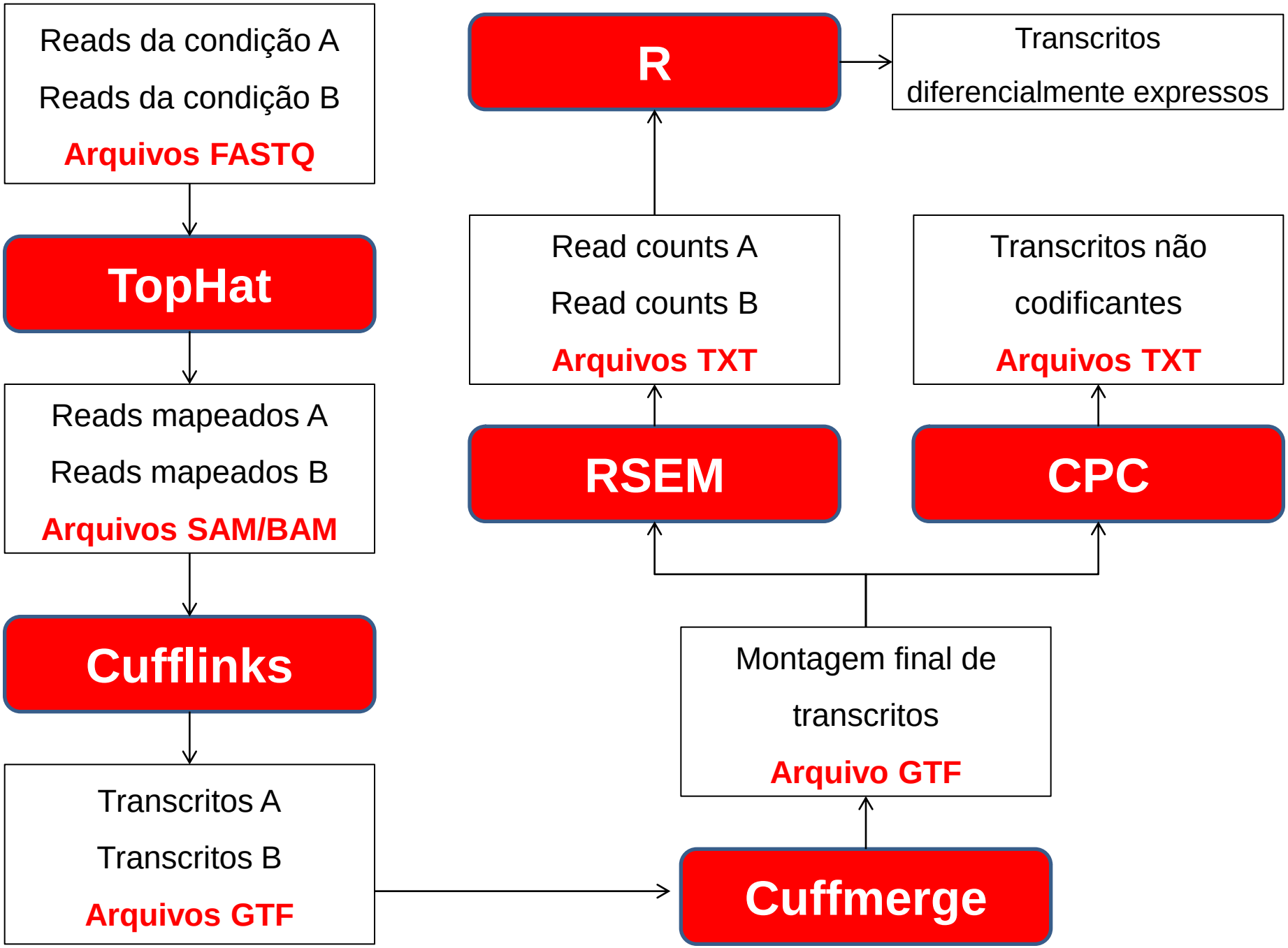
  - Não existe genoma/transcriptoma de referência.

- Alinhamento com transcritos

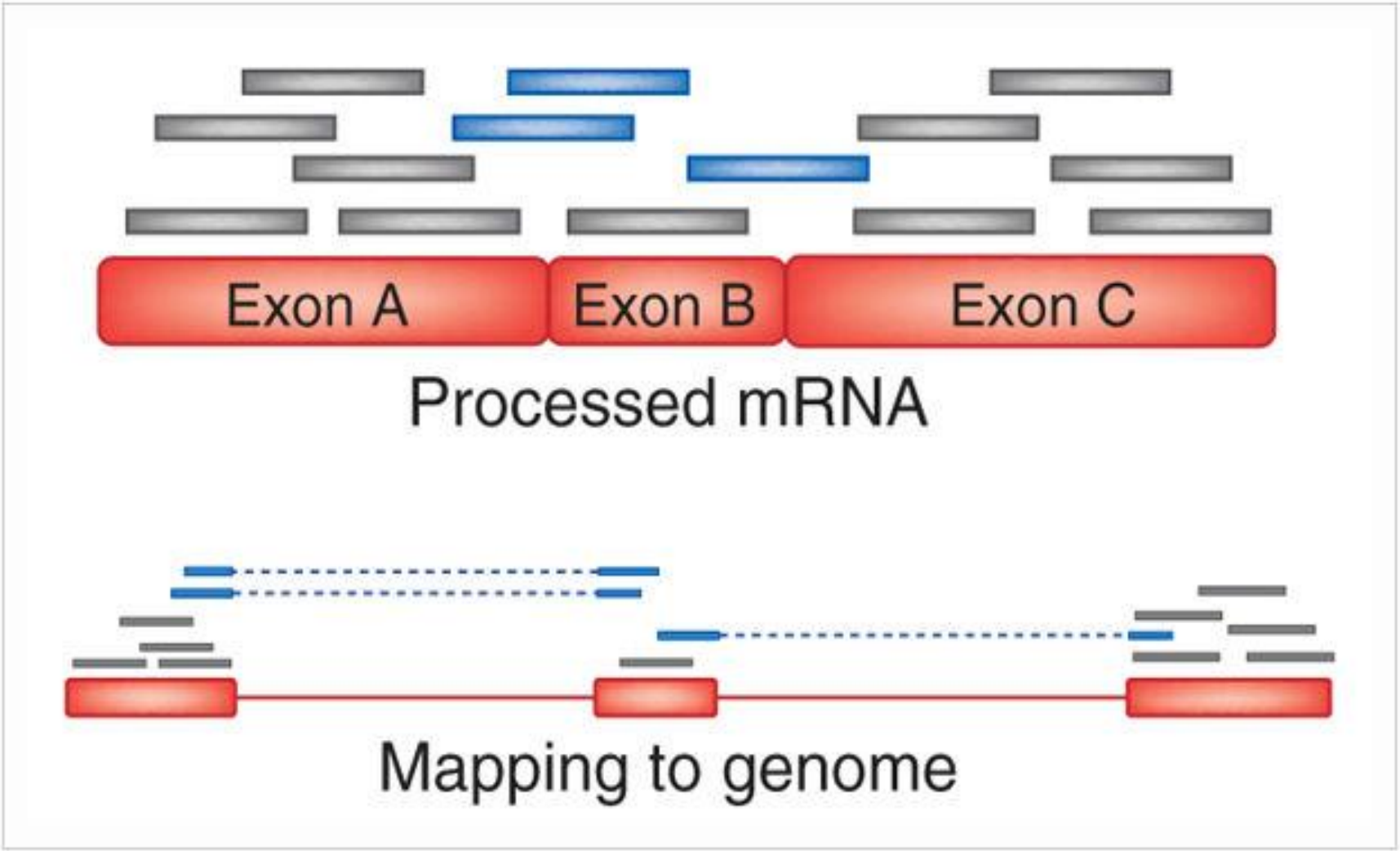
  - Visa identificar a expressão de genes/transcritos já conhecidos.

- Alinhamento com o genoma

  - Estudos voltados para a descoberta de novos genes (codificantes ou não) e novos transcritos para genes já conhecidos.


Reads da condição A
Reads da condição B
Arquivos FASTQ
TopHat
Reads mapeados A
Reads mapeados B
Arquivos SAM/BAM
Cufflinks
Transcritos A
Transcritos B
Arquivos GTF
Cuffmerge
Montagem final de transcritos
Arquivo GTF
RSEM
Read counts A
Read counts B
Arquivos TXT
CPC
Transcritos não codificantes
Arquivos TXT
R
Transcritos diferencialmente expressos

Como alinhar meus reads?



How to map billions of short reads onto genomes (Trapnell and Salzberg, 2009)

TopHat

- Faz mapeamento com “splicing”, permitindo a identificação de introns.
- Utiliza o Bowtie (tem que gerar índices do bowtie para o genoma).
- Permite a utilização de um catálogo de transcritos conhecidos.
- Arquivo de saída em formato BAM.
- http://ccb.jhu.edu/software/tophat/index.shtml

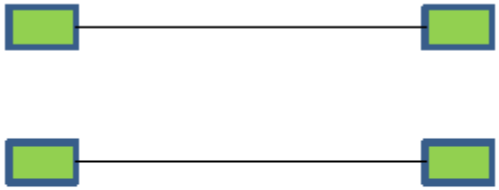
TopHat - pipeline

- Alinhamento dos reads nos transcritos (caso fornecido).
- Reads não alinhados nos transcritos são alinhados no genoma (sem splicing).
- Reads sem alinhamento com o genoma são divididos em seeds (sementes).
- As sementes são alinhadas no genoma: parâmetros de tamanho do intron limitam a distância entre sementes do mesmo read.

TopHat - pipeline

Mapped reads

Unmapped reads



Reads correspondentes a junções exon-exon

TopHat - parâmetros

- Tamanho mínimo do intron (-i).
- Tamanho máximo do intron (-I).
- Transcritos conhecidos (-G).
- Número máx. de alinhamentos por read (-g).
- Número máx. de mismatches (-N).
- Diretório de saída (-o).

TopHat - exemplo

- Supondo que eu tenho 2 amostras de RNAseq e quero alinhá-las no genoma permitindo introns de até 300.000 bp e permitir somente 1 mismatch nos alinhamentos, qual linha de comando vou usar?
- Importante, se você vai rodar o programa 2 vezes (uma por amostra) você deve modificar o diretório de saída!

TopHat - exemplo

```
[leandro@lactads04 Test]$
[leandro@lactads04 Test]$
[leandro@lactads04 Test]$ bowtie2-build meu_genoma.fasta meu_genoma.fasta  ^C
[leandro@lactads04 Test]$
[leandro@lactads04 Test]$ ls
meu_genoma.fasta        meu_genoma.fasta.4.bt2      SRR521590.fastq
meu_genoma.fasta.1.bt2  meu_genoma.fasta.rev.1.bt2  tophat-2.0.9.Linux_x86_64
meu_genoma.fasta.2.bt2  meu_genoma.fasta.rev.2.bt2
meu_genoma.fasta.3.bt2  SRR521589.fastq
[leandro@lactads04 Test]$
[leandro@lactads04 Test]$ ./tophat-2.0.9.Linux_x86_64/tophat -I 300000 -N 1 -o tophat_SRR521589 meu_gen
oma.fasta SRR521589.fastq
```

Cufflinks

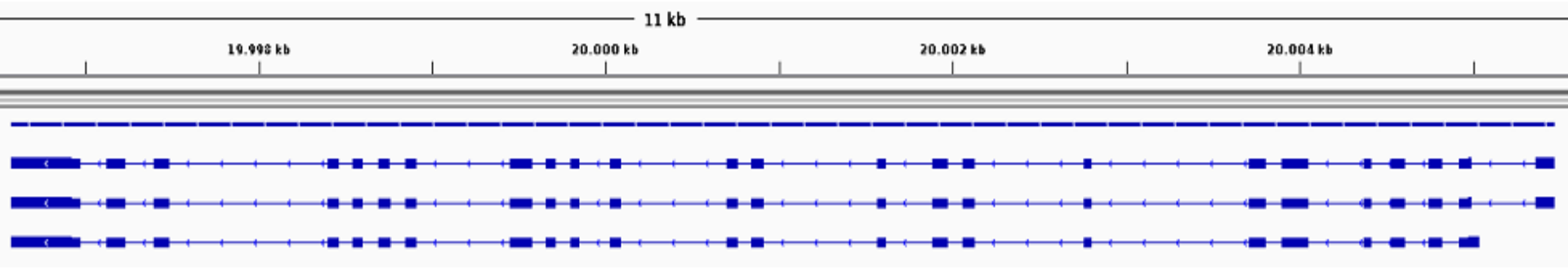
- Faz a montagem dos transcritos usando o alinhamento dos reads no genoma.
- 1 montagem por amostra.
- Permite a utilização de um catálogo de transcritos conhecidos.
- http://cole-trapnell-lab.github.io/cufflinks/

Cufflinks - parâmetros

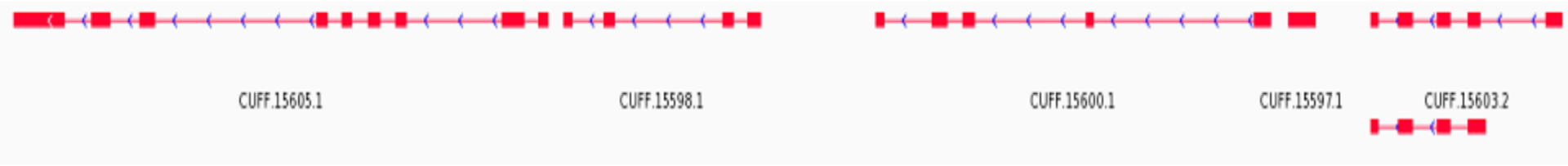
- Tamanho máximo do intron (-I).
- Descartar transcritos com abundância menor que a estabelecida (-F).
- Descartar transcritos intrônicos com abundância menor que a estabelecida (-j).
- Usar somente os transcritos conhecidos (-G).
- Usar os transcritos conhecidos para guiar a montagem (-g).

O que muda o –F?

- O parâmetro “–F” evita a montagem de transcritos falso positivos (transcrição de fundo).
- Corte muito baixo pode trazer transcritos que não existem.
- Corte muito alto pode descartar transcritos reais, principalmente não codificantes.

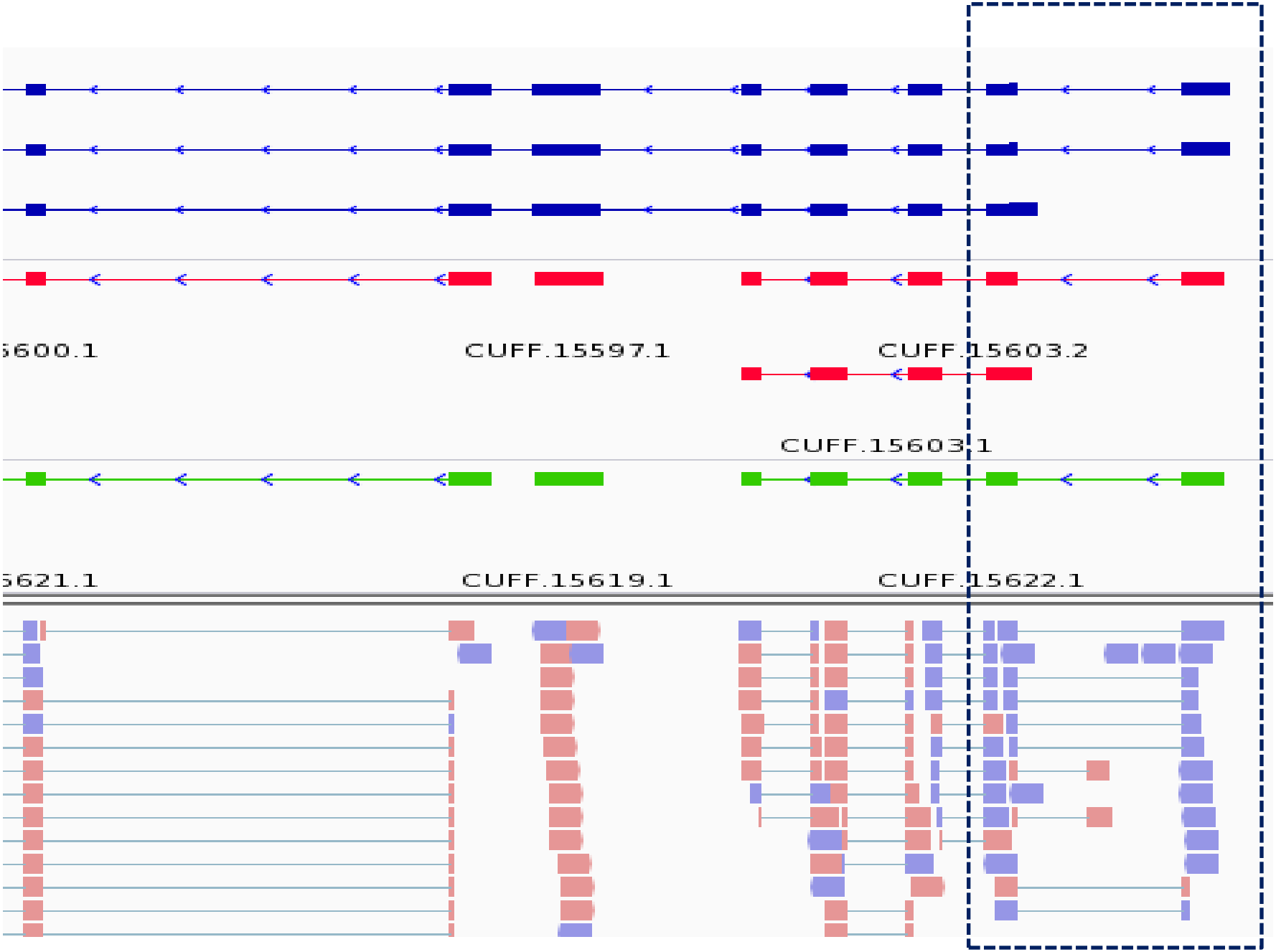


Transcritos conhecidos



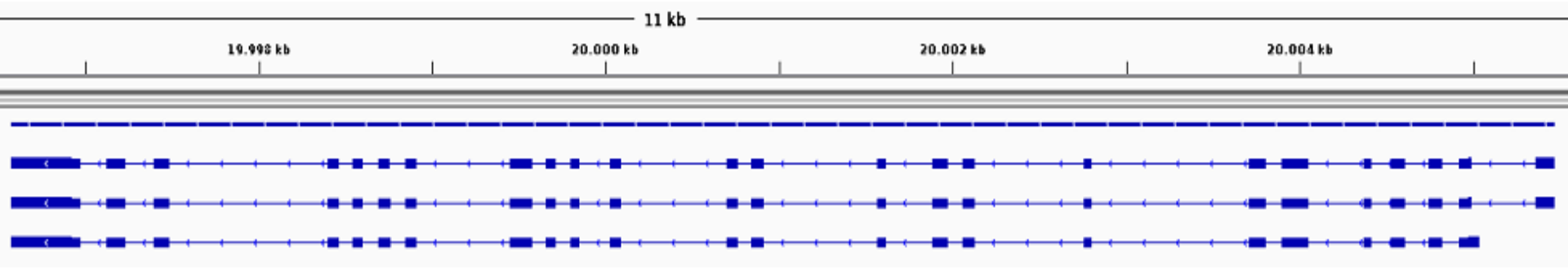
Montagem (cufflinks) sem alterar o –F (0,1)

Montagem (cufflinks) alterando o –F (0,4)

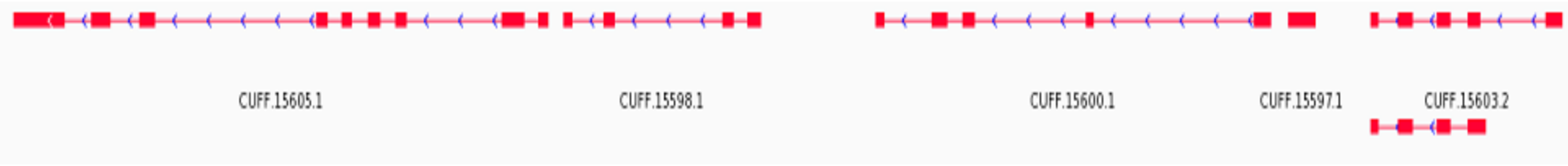

600.1
CUFF.15597.1
CUFF.15603.2
CUFF.15603.1
621.1
CUFF.15619.1
CUFF.15622.1

Porque usar o –g?

- O parâmetro “–g” permite fornecer ao cufflinks um catálogo de transcritos conhecidos para auxiliar a montagem.
- São aceitos os formatos GFF e GTF.
- Sem esse parâmetro, transcritos com pouca cobertura podem acabar “picados” em vários.

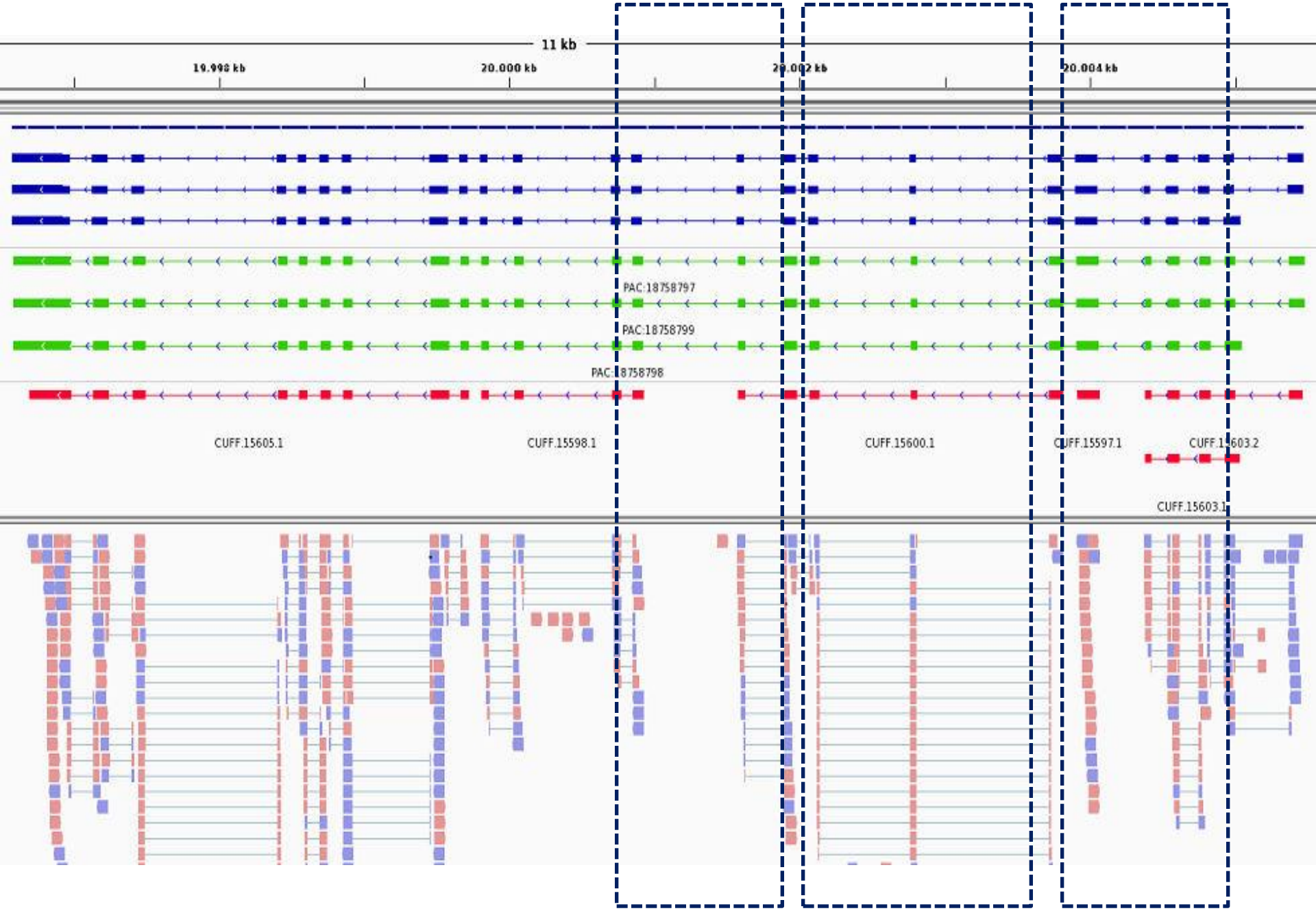


Transcritos conhecidos



Montagem (cufflinks) sem usar o –g

Montagem (cufflinks) utilizando o –g

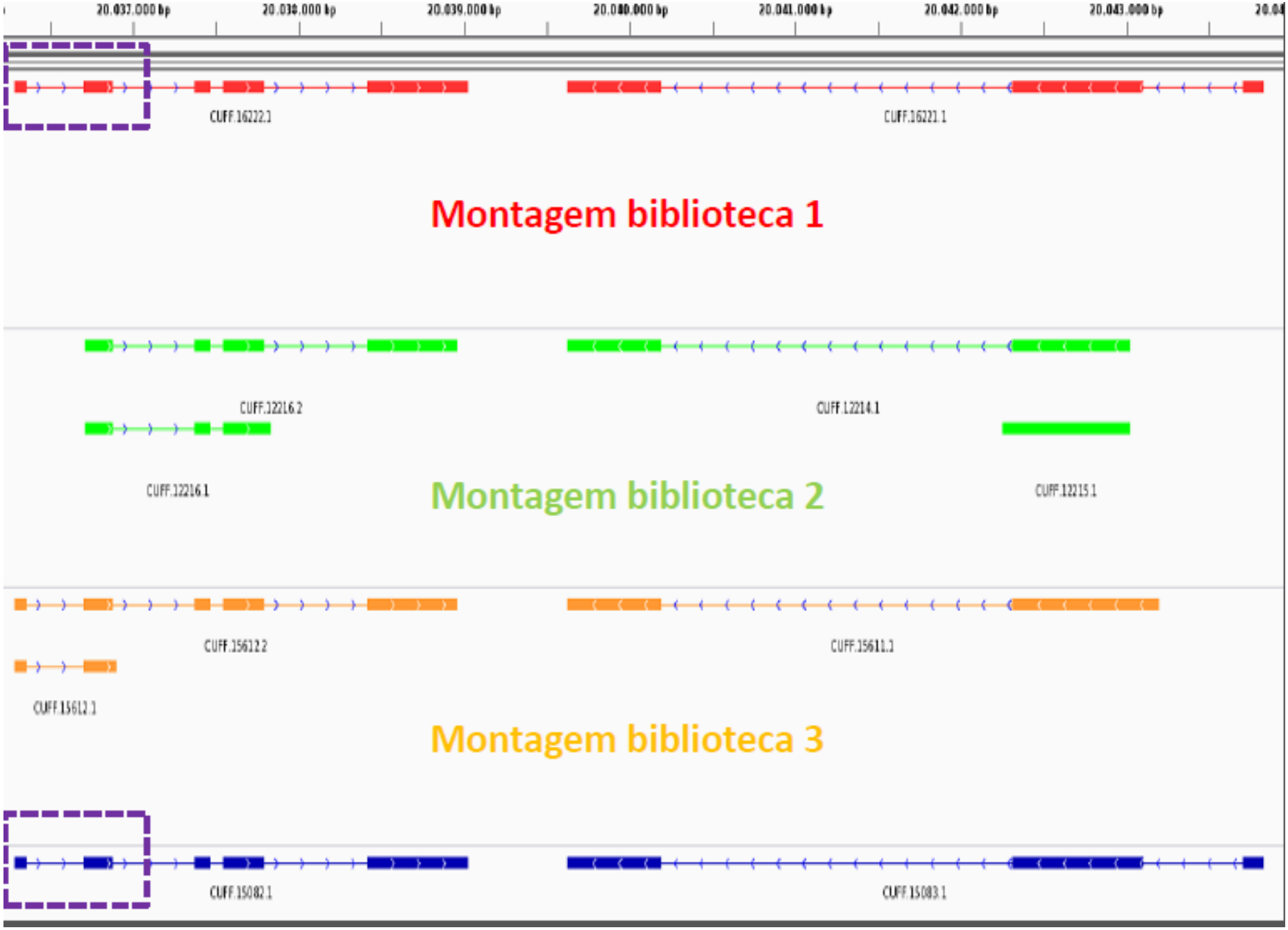

11 kb
19.998 kb
20.000 kb
20.002 kb
20.004 kb
PAC:18758797
PAC:18758799
PAC:18758798
CUFF.15605.1
CUFF.15598.1
CUFF.15600.1
CUFF.15597.1
CUFF.15603.2
CUFF.15603.1

E o –j?

- Parecido com o “–F”, mas diz respeito a transcritos localizados dentro de introns.
- Muito importante em estudos de ncRNA intrônicos.
- Corte muito baixo pode trazer transcritos que não existem (pré-mRNA).
- Corte muito alto pode descartar transcritos reais.

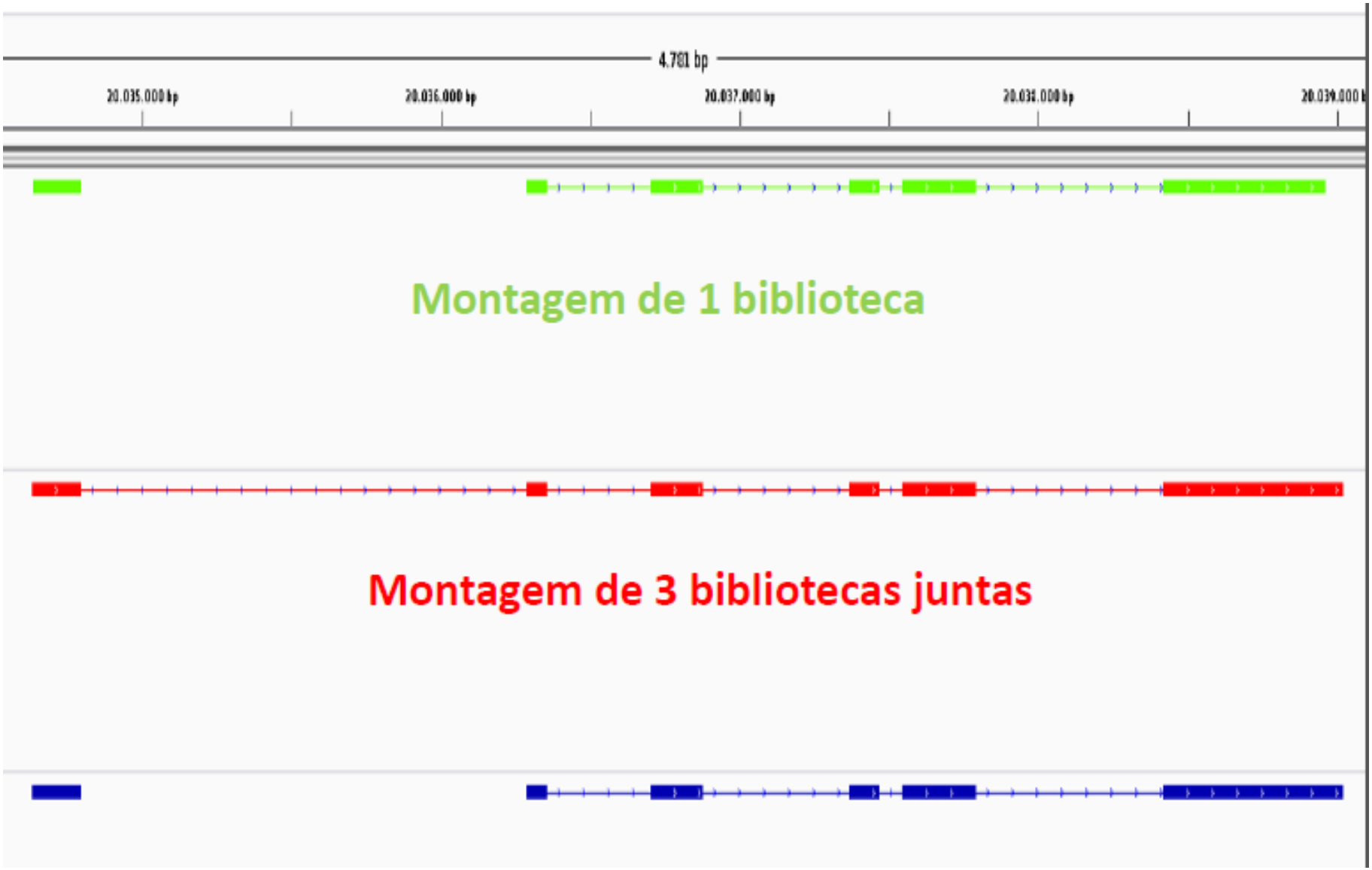
Cuffmerge

- Se eu fizer uma montagem por biblioteca, terei transcritos iguais com nomes diferentes.
- O cuffmerge faz um “merge” das montagens do cufflinks, gerando uma referência única.
- Usa como entrada os arquivos GTFs das montagens do cufflinks e gera um GTF único.


20.037.000 bp
20.038.000 bp
20.039.000 bp
20.040.000 bp
20.041.000 bp
20.042.000 bp
20.043.000 bp
CUFF.16222.1
CUFF.16221.1
Montagem biblioteca 1
CUFF.12216.2
CUFF.12214.1
CUFF.12216.1
CUFF.12215.1
Montagem biblioteca 2
CUFF.15612.2
CUFF.15611.1
CUFF.15612.1
Montagem biblioteca 3
CUFF.15082.1
CUFF.15083.1

Cuffmerge

- Os resultados são diferentes dos resultados de uma única montagem com várias bibliotecas.
- Em uma montagem com todas as bibliotecas, a soma de expressão de todas as bibliotecas vai causar um viés.



Montagem - Cuffmerge

Cuffcompare

- Compara os resultados de montagens do cufflinks/cuffmerge com o GTF/GFF inicial.
- Após a comparação, divide os transcritos em classes (class_codes) e grava os resultados no arquivo “cuffcmptracking”.

Class codes

- =

  - Transcrito igual a um conhecido.

- j

  - Nova variante de splicing.

- u

  - Transcrito em região intergênica (novo).

- i

  - Transcrito localizado em um intron.

- http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/

TCONS_00000001	XLOC_000001	SS11297|13100.1	o		q1:SS11297|TCONS_0
TCONS_00000002	XLOC_000001	SS11297|13100.1	=		q1:SS11297|TCONS_0
TCONS_00000003	XLOC_000002	SS11349|13155.1	=		q1:SS11349|TCONS_0
TCONS_00000004	XLOC_000003	-	u	q1:XLOC_000003	|TCONS_0000
TCONS_00000005	XLOC_000004	-	u	q1:XLOC_000004	|TCONS_0000
TCONS_00000006	XLOC_000005	-	u	q1:XLOC_000005	|TCONS_0000
TCONS_00000007	XLOC_000006	SS49599|53190.1	=		q1:SS49599|TCONS_0
TCONS_00000008	XLOC_000007	-	u	q1:XLOC_000007	|TCONS_0000
TCONS_00000009	XLOC_000008	-	u	q1:XLOC_000008	|TCONS_0000
TCONS_00000010	XLOC_000009	-	u	q1:XLOC_000009	|TCONS_0000
TCONS_00000011	XLOC_000010	SS49600|53191.1	=		q1:SS49600|TCONS_0
TCONS_00000012	XLOC_000011	SS16995|19325.1	=		q1:SS16995|TCONS_0
TCONS_00000013	XLOC_000012	SS49601|53192.1	=		q1:SS49601|TCONS_0
TCONS_00000014	XLOC_000013	SS49602|53193.1	=		q1:SS49602|TCONS_0
TCONS_00000015	XLOC_000014	SS49603|53194.1	=		q1:SS49603|TCONS_0
TCONS_00000016	XLOC_000015	-	u	q1:XLOC_000015	|TCONS_0000
TCONS_00000017	XLOC_000016	SS49604|53195.1	=		q1:SS49604|TCONS_0
TCONS_00000018	XLOC_000017	SS16996|19326.1	j		q1:SS16996|TCONS_0
TCONS_00000019	XLOC_000017	SS16996|19326.1	=		q1:SS16996|TCONS_0
TCONS_00000020	XLOC_000018	SS16997|19327.1	j		q1:SS16997|TCONS_0
TCONS_00000021	XLOC_000018	SS16997|19327.1	j		q1:SS16997|TCONS_0
TCONS_00000022	XLOC_000018	19328|19328.1	=		q1:19328|TCONS_00
TCONS_00000023	XLOC_000018	SS16997|19327.1	=		q1:SS16997|TCONS_0
TCONS_00000024	XLOC_000019	SS49605|53196.1	=		q1:SS49605|TCONS_0
TCONS_00000025	XLOC_000020	-	u	q1:XLOC_000020	|TCONS_0000
TCONS_00000026	XLOC_000020	-	u	q1:XLOC_000020	|TCONS_0000
TCONS_00000027	XLOC_000021	-	u	q1:XLOC_000021	|TCONS_0000
TCONS_00000028	XLOC_000022	SS49606|53197.1	=		q1:SS49606|TCONS_0
TCONS_00000029	XLOC_000023	SS16998|19329.1	=		q1:SS16998|TCONS_0
TCONS_00000030	XLOC_000024	-	u	q1:XLOC_000024	|TCONS_0000
TCONS_00000031	XLOC_000025	-	u	q1:XLOC_000025	|TCONS_0000
TCONS_00000032	XLOC_000026	SS49607|53198.1	=		q1:SS49607|TCONS_0
TCONS_00000033	XLOC_000027	SS16999|19330.1	x		q1:SS16999|TCONS_0
TCONS_00000034	XLOC_000028	SS17000|19331.1	=		q1:SS17000|TCONS_0
TCONS_00000035	XLOC_000029	SS16999|19330.1	j		q1:SS16999|TCONS_0
TCONS_00000036	XLOC_000029	SS16999|19330.1	j		q1:SS16999|TCONS_0
TCONS_00000037	XLOC_000029	SS16999|19330.1	j		q1:SS16999|TCONS_0
TCONS_00000038	XLOC_000029	SS16999|19330.1	j		q1:SS16999|TCONS_0
TCONS_00000039	XLOC_000029	SS16999|19330.1	=		q1:SS16999|TCONS_0
TCONS_00000040	XLOC_000030	-	u	q1:XLOC_000030	|TCONS_0000
TCONS_00000041	XLOC_000031	SS49608|53199.1	=		q1:SS49608|TCONS_0
TCONS_00000042	XLOC_000032	SS49609|53200.1	=		q1:SS49609|TCONS_0
TCONS_00000043	XLOC_000033	SS49610|53201.1	=		q1:SS49610|TCONS_0
TCONS_00000044	XLOC_000034	SS49611|53202.1	=		q1:SS49611|TCONS_0
TCONS_00000045	XLOC_000035	-	u	q1:XLOC_000035	|TCONS_0000
TCONS_00000046	XLOC_000035	-	u	q1:XLOC_000035	|TCONS_0000
TCONS_00000047	XLOC_000036	-	u	q1:XLOC_000036	|TCONS_0000
TCONS_00000048	XLOC_000037	SS49612|53203.1	=		q1:SS49612|TCONS_0
TCONS_00000049	XLOC_000038	SS49613|53204.1	=		q1:SS49613|TCONS_0
TCONS_00000050	XLOC_000039	SS17001|19332.1	j		q1:SS17001|TCONS_0
TCONS_00000051	XLOC_000039	SS17001|19332.1	j		q1:SS17001|TCONS_0

ncRNA

- Diversos estudos tem se voltado para descoberta de RNAs não codificadores, chamados de ncRNAs.
- Os ncRNAs estão relacionados a regulação gênica, pré ou pós transcrional.
- Não traduzem proteínas, ou seja, não possuem ORF.

CPC

- **C**oding **P**otential **C**alculator**:**
- http://cpc.cbi.pku.edu.cn/
- Programa que tenta calcular o potencial codificador de um transcrito.
- Usa diversas evidências:

  - Alinhamento com bancos de dados de proteínas.

  - Existência de ORFs.

  - Alinhamento com bancos de UTR.