Mapeamento das leituras no genoma de referência

Pablo Rodrigo Sanches

Departamento de Genética – FMRP/USP

psanches@usp.br

Roteiro de análise

1. RNA STAR (mapeamento das bibliotecas de leituras contra o genoma de referência)
2. Finalizar a tabela de resultados sobre as características gerais das bibliotecas sequenciadas

Recordando...

Controle (ex. não tratado)

Reads (leituras) do RNA-Seq

Desafio (ex. tratado com droga)

Genoma de referência

Arquivos necessários para o genoma de referência

Arquivo_de_supercontig.fasta – *conjunto ordenado e orientado de **contigs** que ainda possuem lacunas*

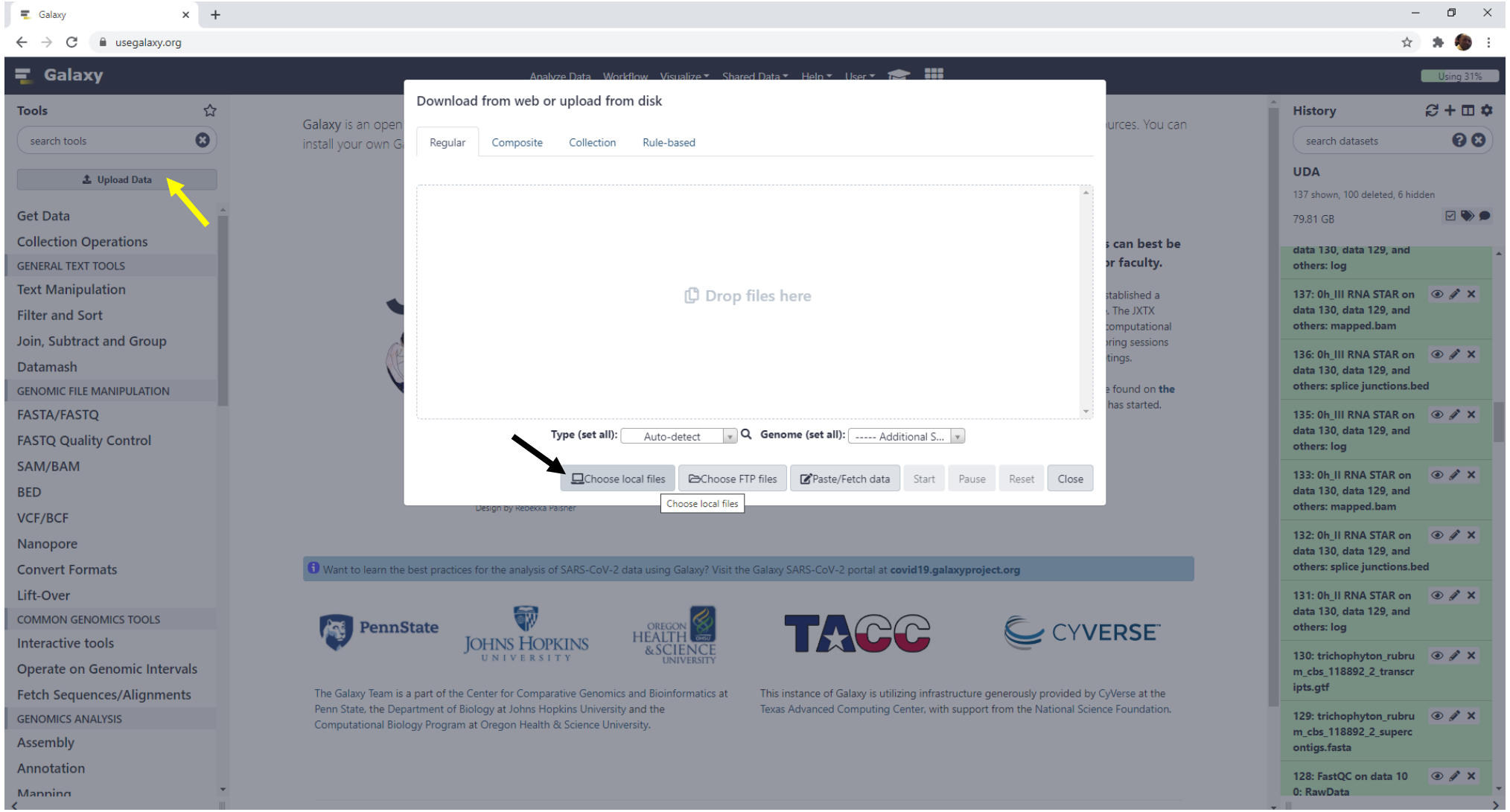
```
>Supercontig_2.1 of Trichophyton rubrum CBS 118892
ATAGTATTACTTACTACTACTATAAGTCCCTATTATAGAAGGTATGCTCTACTATTAGTA
TCTAATCTAGATAACTATAAAACTATATTAAAATATCTTTAAATACTTAGATATATAGCG
GTAGTTAAACTACTAGTATACCTCTATTTAGAAGGCGTAGATAGCTTATTTAACTCTCTA
GATAATACTAAAGTATAGAGATTTAATATTAAGTACTACCTATATTAGCCTTTTTATAAT
ATTAATAATAATCTTCTATAAAGTAGTAATACCTATAAAAAAGCTATTACTATTCCTATA
GAGTGTAATAATAATATTATATCTAATTCTAATAGTACTTTAGAGGATATTAATATAAAA
GATAACTCCACTACTATTTTAATAGCCCCTACTATATCTAGACCCCCTAAGTACCGTAAG
TATAAGGACGCGTCCTCTAATAGAGCTACTATAGCTAAGTAGCTAAAAGTAGATAGGCTA
ATATCTTATATTATATTAAGGACTAATATAGTACTAAATATAACTAATAAATCTAAAAAA
GATAGTAAAGGTTTTATATTTAGATTTTAGAATATATTAATATCTAGAGTATTATCTACT
AAATAACCTACTACTATATACTCTATAAACTCTATTATATTAATATCTAATAATAATATA
ATTAAGTTAGTTTATTTATAGCTACTAATACTTATTTACTACTATAATCTATCTAATATT
TTCTTTTTAGGCCCTATATTACTTAATATCGCTAAAACTCTTATTAAAGCCTTCTCTATA
TTTTTAGACCCTATAGATACTATATCTACTATTTAGACTATTAATTAGTATATAGATACT
ACCTATTACTTCTATAATAAGCTTAACTAATATATCTTATACTATATAGTATCTACTATA
GCTATTAATAAATATAATACTACTATTCTTACTTATTATATTATTTAAATATATTTATAG
--More--(0%)
```

Um contig é um conjunto de segmentos de DNA sobrepostos que juntos representam uma região de consenso do DNA

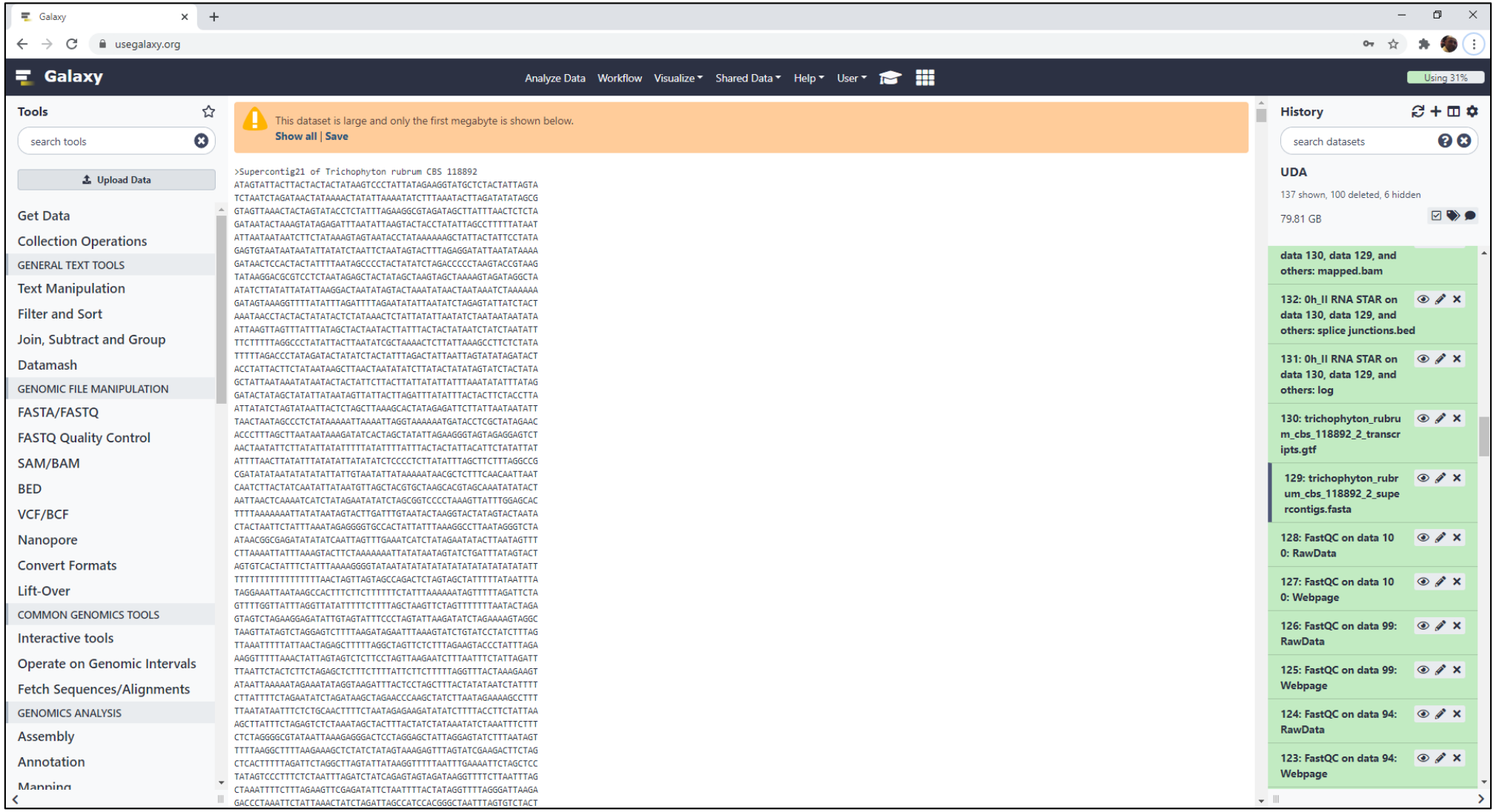
Arquivo.gtf - *usado para descrever genes e outras informações sobre sua estrutura*

```
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       start_codon     14925   14927   .       -       0       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       stop_codon      12841   12843   .       -       0       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    14609   14927   .       -       .       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       CDS     14609   14927   .       -       0       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    13933   14529   .       -       .       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       CDS     13933   14529   .       -       2       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    13702   13869   .       -       .       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       CDS     13702   13869   .       -       2       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    13470   13638   .       -       .       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       CDS     13470   13638   .       -       2       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    12841   13414   .       -       .       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       CDS     12844   13414   .       -       1       gene_id "TERG_00002"; transcript_id "TERG_00002T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       start_codon     16137   16139   .       +       0       gene_id "TERG_00003"; transcript_id "TERG_00003T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       stop_codon      17793   17795   .       +       0       gene_id "TERG_00003"; transcript_id "TERG_00003T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    16137   16610   .       +       .       gene_id "TERG_00003"; transcript_id "TERG_00003T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       CDS     16137   16610   .       +       0       gene_id "TERG_00003"; transcript_id "TERG_00003T0";
Supercontig_2.1 TR_CBS118892_V2_FINAL_CALLGENES_5       exon    16685   17514   .       +       .       gene_id "TERG_00003"; transcript_id "TERG_00003T0";
--More--(0%)
```

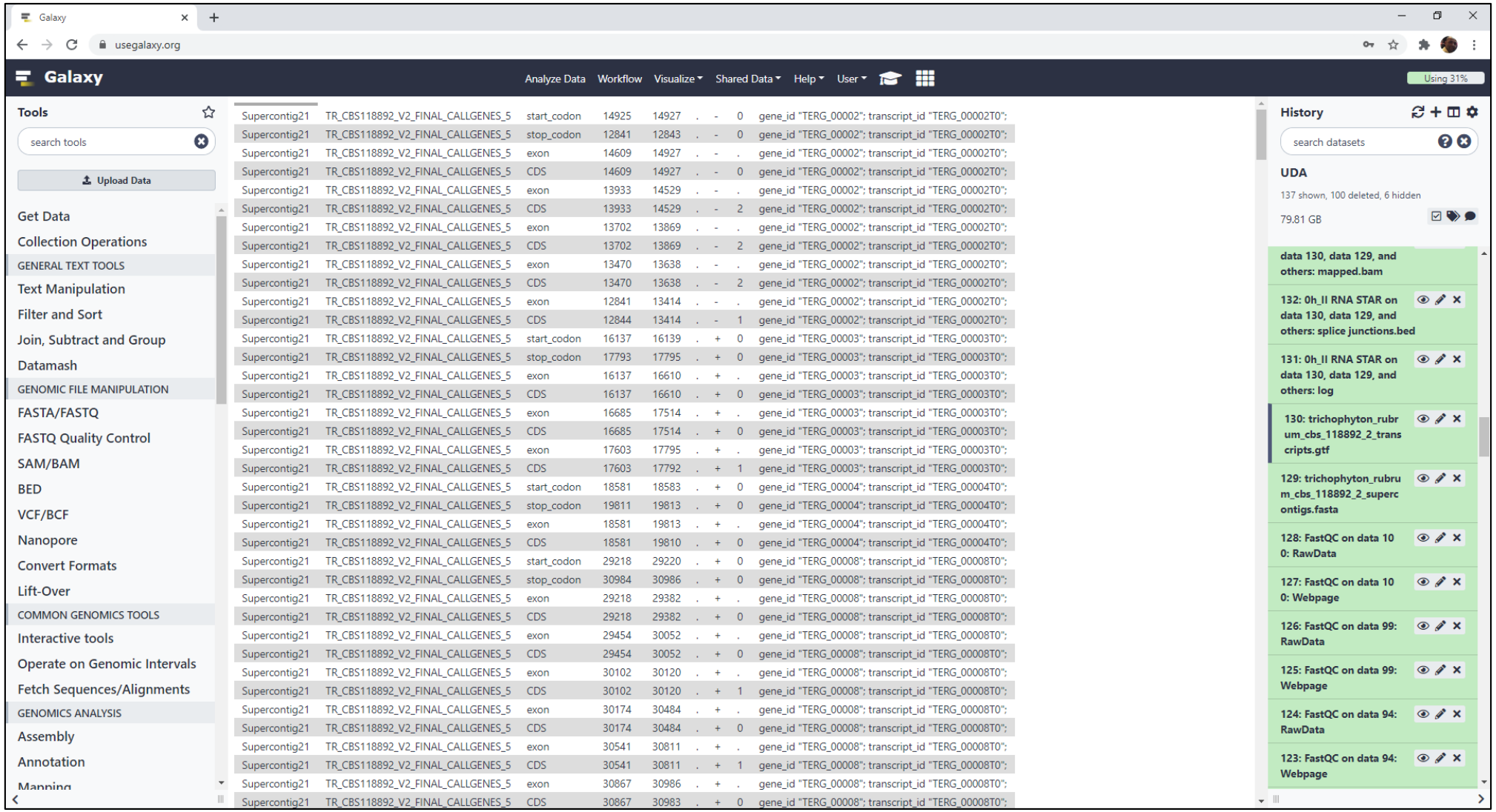
Enviar arquivos do genoma para o Galaxy



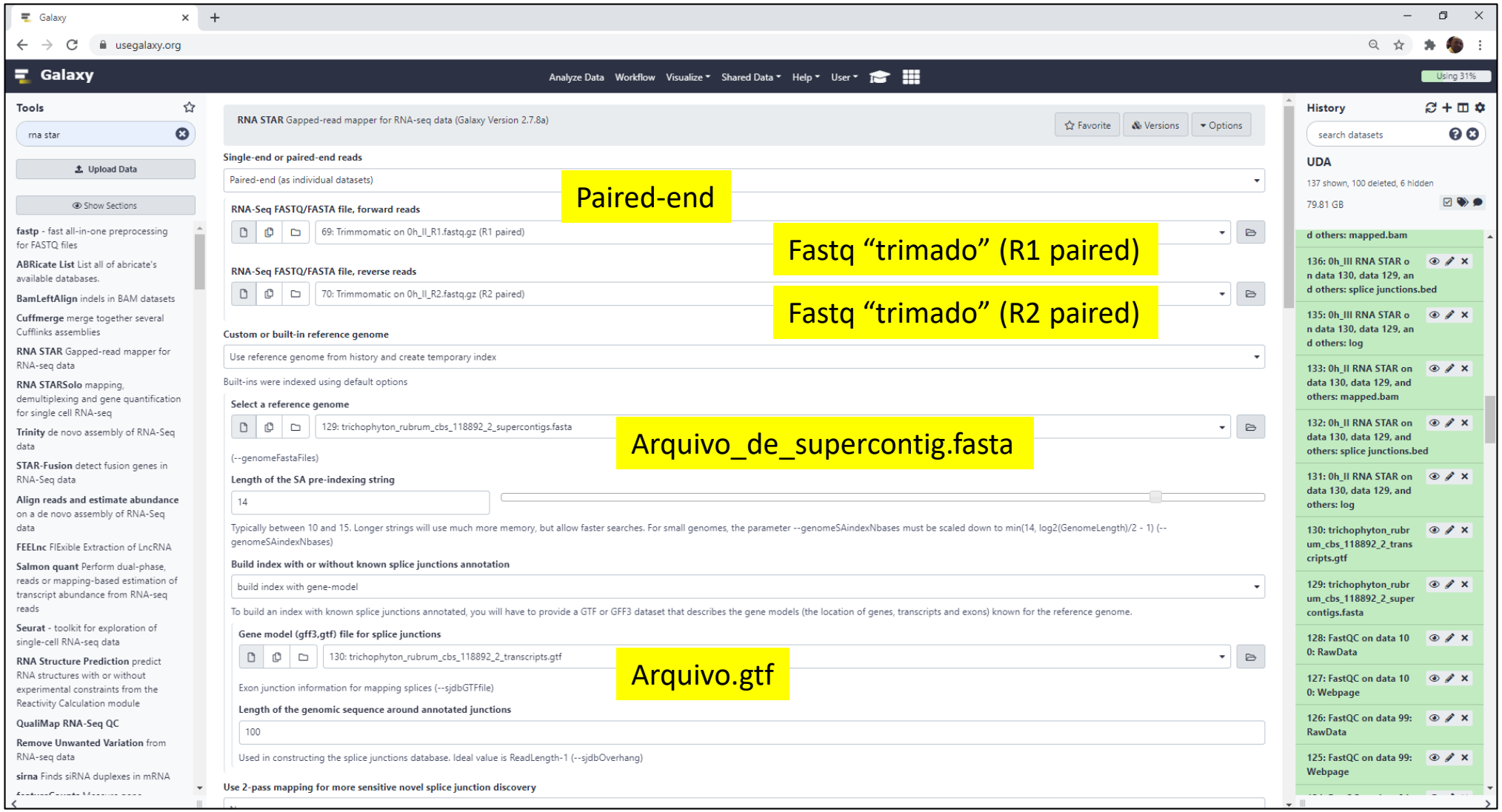
Arquivo_de_supercontig.fasta no Galaxy



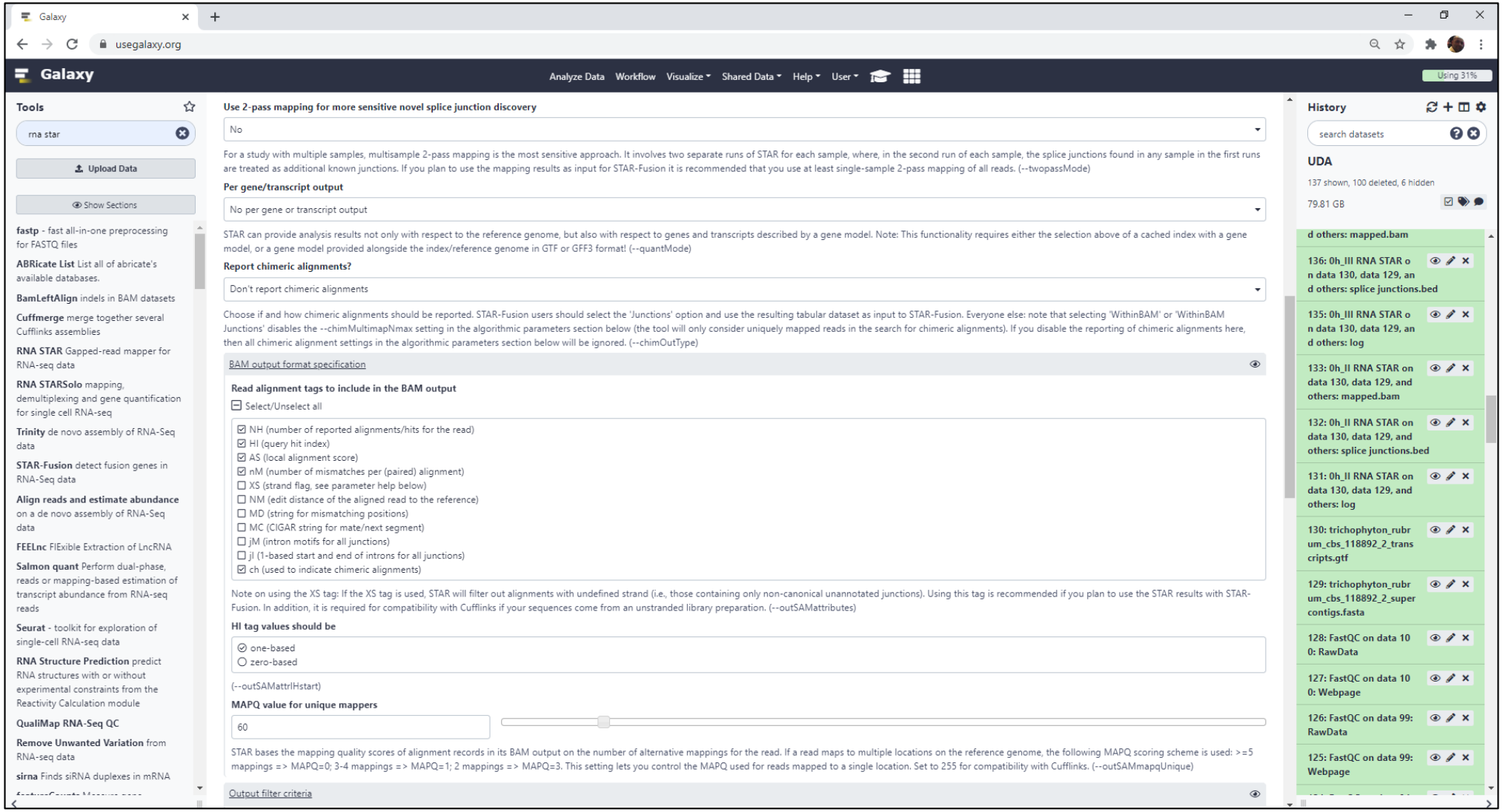
Arquivo.gtf no Galaxy



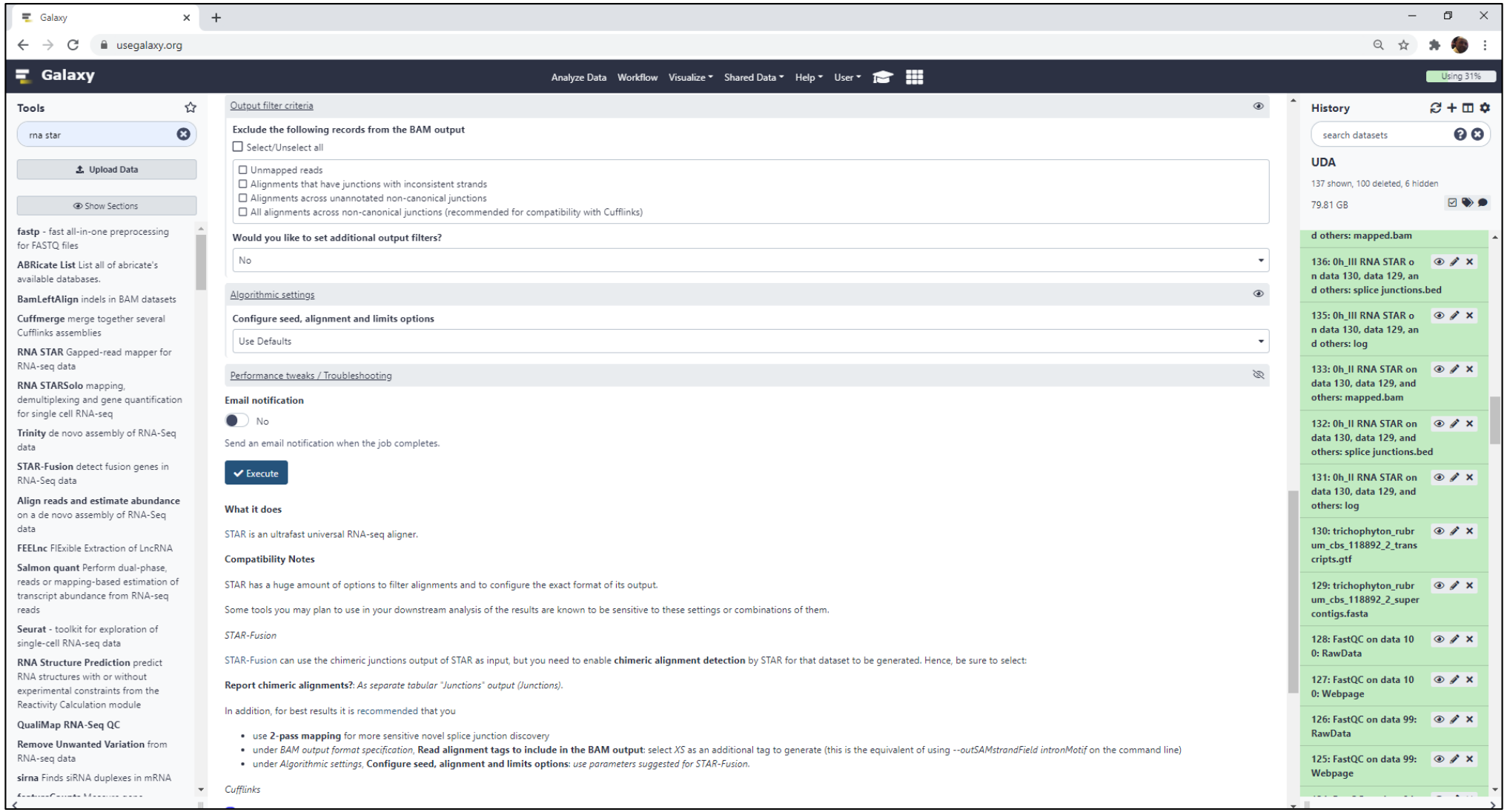
RNA STAR – Via Galaxy



RNA STAR – Via Galaxy (cont.)



RNA STAR – Via Galaxy (cont.)



Exemplo de Resultado RNA STAR (Log)

```
                                 Started job on |       Apr 09 17:17:25
                             Started mapping on |       Apr 09 17:17:56
                                    Finished on |       Apr 09 17:50:15
       Mapping speed, Million of reads per hour |       49.76

                           Number of input reads |       26800883
                       Average input read length |       91
                                    UNIQUE READS:
                      Uniquely mapped reads number |       23563054
                           Uniquely mapped reads % |       87.92%
                             Average mapped length |       91.49
                          Number of splices: Total |       3281588
              Number of splices: Annotated (sjdb) |       2879421
                          Number of splices: GT/AG |       3259914
                          Number of splices: GC/AG |       11595
                          Number of splices: AT/AC |       795
                  Number of splices: Non-canonical |       9284
                         Mismatch rate per base, % |       0.18%
                            Deletion rate per base |       0.00%
                           Deletion average length |       1.30
                           Insertion rate per base |       0.00%
                          Insertion average length |       1.07
                               MULTI-MAPPING READS:
            Number of reads mapped to multiple loci |       0
                 % of reads mapped to multiple loci |       0.00%
            Number of reads mapped to too many loci |       115902
                 % of reads mapped to too many loci |       0.43%
                                 UNMAPPED READS:
       Number of reads unmapped: too many mismatches |       0
            % of reads unmapped: too many mismatches |       0.00%
                 Number of reads unmapped: too short |       3049969
                      % of reads unmapped: too short |       11.38%
                     Number of reads unmapped: other |       71958
                          % of reads unmapped: other |       0.27%
                                       CHIMERIC READS:
                           Number of chimeric reads |       0
                                % of chimeric reads |       0.00%
```

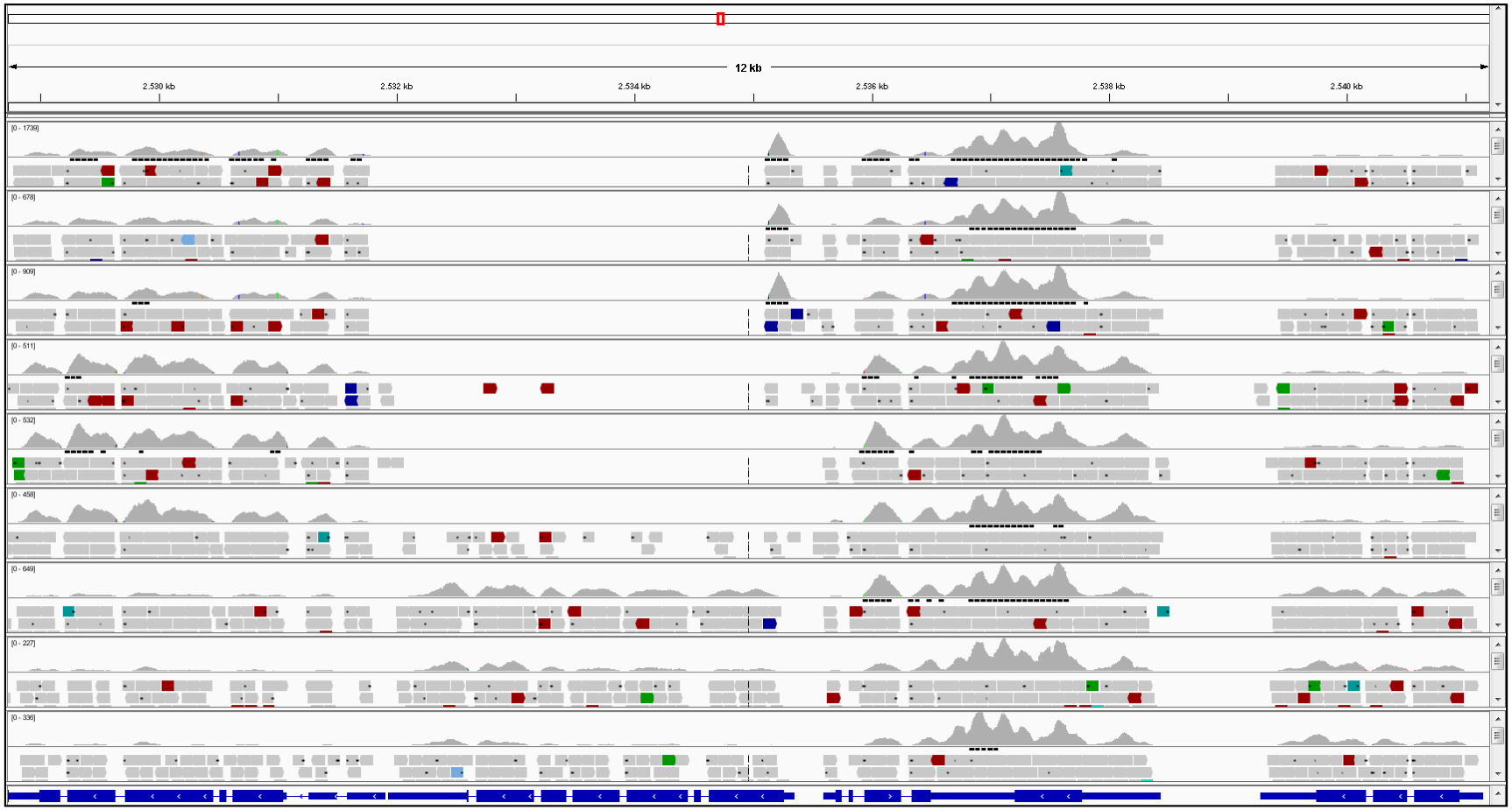
Nro. Reads de entrada (já trimados)

Nro. Reads mapeadas

Exemplo de Resultado RNA STAR (SAM/BAM)

```
NB501939:113:H3JNTBGXC:3:23609:19578:3727	163	Supercontig21	7375	255	12M26S	=	27626	20348	CTTAAAGAAGTACGCAGAAGAGGCTACTCCAGAGGTAA	/A//AE/E66EAE/A//E//E66/EEEA//E/E//E/E	NH:i:1	HI:i:1	AS:i:103	nM:i:1
NB501939:113:H3JNTBGXC:1:23305:15522:20187	99	Supercontig21	7801	255	150M	=	7892	241	GTCTATTTAGATATCTTAGCCTTCTAGAAGAAACTATATTTAGCTAAGCTTTCTAGAGGTAGCTAAGCTACTTTCTAGTAGTTCTTAGTAGCAGGGACCTAATAACTTAGCTAAACTATTAGATATTTTAGCTATAGCCGCGAATTTTTA
	AAAAAEEEEEEEEEEEEEEEEEEEEAEEEEAEEEEAEEEEEEEEEEEEEEEEEEEEEE/EEEEEEEEEEEEEAEEEEEEEEAEEEEAAEEEEEEEEEEEAEAEEAEEEEEEAAEEEAE/AAAAAEEAAEA<AAAEAE<A66A6/<AAEAAAA	NH:i:1	HI:i:1	AS:i:298	nM:i:0
NB501939:113:H3JNTBGXC:1:23305:15522:20187	147	Supercontig21	7892	255	150M	=	7801	-241	CAGGGACCTAATAACTTAGCTAAACTATTAGATATTTTAGCTATAGCCGCGAATTTTTATTAAGATAGATCAGGTAAGGAAAGATAAGGATTGATAAGTAGTAAGTAATAAGAGATAAAATAAGGAAAATCAAAATTATAGAAGGGAAG
	EEAA/EAAAA<EAAE<EA<AA66AEEEEAEEEEEE/<EAAEEEEEEEEAEAEEEEEAEEEAEE<EEEEEAEEEEEEEAEEEEEEEEE/EAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAAAA	NH:i:1	HI:i:1	AS:i:298	nM:i:0
NB501939:113:H3JNTBGXC:1:23112:1082:11971	163	Supercontig21	10726	255	134M	=	10726	134	CGCCTAGGCTCTCCTCAACGTCACTTCATCGACTCTTTCAACAACTAATGGATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATG	AAAAAEEEEEEEE
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEAEAEEEEEEEEAEEEAEEE<EEEEEEEEAEEEEA<EEEEEEAEEAAAEEEA	NH:i:1	HI:i:1	AS:i:266	nM:i:0
NB501939:113:H3JNTBGXC:1:23112:1082:11971	83	Supercontig21	10726	255	134M	=	10726	-134	CGCCTAGGCTCTCCTCAACGTCACTTCATCGACTCTTTCAACAACTAATGGATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATG	6<<A<<<AAA<A<E
EEEEEEEAEEEEEEEE<6EEEE/EEEEAEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAAAAA	NH:i:1	HI:i:1	AS:i:266	nM:i:0
NB501939:113:H3JNTBGXC:1:22308:16779:18879	163	Supercontig21	10767	255	151M	=	10782	166	CAACTAATGGATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCT
C	AAAAAEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEE/EEEEEEEEEEEEE/EEEEEAEEEEEEE<EEEEEEEEEAEEEEEEEEEAEAEEEEEEEEEEEEEEEEEEEEEEEAAEEEEEEAAEEEE/EAAA<AA<<<AEEAEEEEEA<<</EEEEA	NH:i:1	HI:i:1	AS:i:300	nM:i:0
NB501939:113:H3JNTBGXC:4:21602:17085:12720	163	Supercontig21	10767	255	118M	=	10767	118	CAACTAATGGATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATAC	AAAAAEEEEEEEEEE/EEEEEEEEEEEAE
EEEEEEEEEE<EEAEEEEEAEEEEEEEEEEEEEEEEEEEAEAE/EEEEE<EAEEEEAEEEAEEEEEEEEE<EAA<</EEEEEA<AEAA<EA	NH:i:1	HI:i:1	AS:i:234	nM:i:0
NB501939:113:H3JNTBGXC:4:21602:17085:12720	83	Supercontig21	10767	255	118M	=	10767	-118	CAACTAATGGATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATAC	AE/EEEEE6<AEEE<AEAEEEEEAEEEEEEE
AEAEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEAEEEEEEEEEEEEEEEEAEEEEEEE/<EEEEEEEEEEEEEEEEEEEEEEEEEAAAAA	NH:i:1	HI:i:1	AS:i:234	nM:i:0
NB501939:113:H3JNTBGXC:1:22105:7558:11669	163	Supercontig21	10776	255	150M	=	10896	388	GATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGCTGGCTCCAGTTTTG
	AAAAAE/EEEEE6/EAEEEAEE6EEEEA/</EEAEEE/EEE6EEEEEEA/EEEEE</EA/EA//AEEEEEEA/A<EEEEEE/EEEAAEEEEEEA</EAE/A<A<E/AE/E/EA/E//AA/A<<AEE<A<//AEA<</A//6AA</6<EA<<	NH:i:1	HI:i:1	AS:i:298	nM:i:1
NB501939:113:H3JNTBGXC:3:13502:13828:20383	163	Supercontig21	10776	255	114M	=	10821	194	GATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAG	AAAAA6EEEEEEEEAEE6EEAEEEE/EEAEE
/EEEEAAEEEEEAEEEEE6EEAEEEEEEEAE/AAAEAEEEA<EEEE/EAEE/EEE/EEEE/A</EEAEEE</6EE/EE<A/<A/<	NH:i:1	HI:i:1	AS:i:261	nM:i:0
NB501939:113:H3JNTBGXC:1:22107:10482:1322	163	Supercontig21	10777	255	46M	=	10782	156	ATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTA	AAAAAE6EEEEE/EE/AEEEEEEEEEEEAEEEEEEEA/EEEEEEEEE	NH:i:1	HI:i:1	AS:i:195	nM:i:0
NB501939:113:H3JNTBGXC:2:12210:6250:2862	163	Supercontig21	10777	255	151M	=	10803	177	ATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCTCCAGTTTTG
G	AAAAAEEEEEEEEEEEEEEEAEEEEEEE6EE6EEEEEAEEEEEEEEEEEEEEEEEEEEAE/A/EEEEEA/EAEEE/EEEEEEEEEE/AA6AEE/<E66EEEEEA//EEEE<EAE/EAEEEAE/<A<AA<EEA/6</AE/A<AEEAEEA/<A/A<	NH:i:1	HI:i:1	AS:i:300	nM:i:0
NB501939:113:H3JNTBGXC:2:13205:12362:10833	163	Supercontig21	10777	255	102M2D49M	=	10802	177	ATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCTCCAG
TTTTGCGCC	AAAAAEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEAEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AE/EEEEAEAEEEEEEEEEEEEEE6AAAEAAAEAEAAAAEAEA<A<	NH:i:1	HI:i:1	AS:i:287	nM:i:0
NB501939:113:H3JNTBGXC:3:12402:18084:9053	163	Supercontig21	10777	255	151M	=	10797	171	ATTTACACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCTCCAGTTTTG
G	AAAAAEEEEEEEEEEEEEAEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEAEEEEEEEEEEAEEEEEEEEEEEEEEEAE<EEEEEEAEEAEEEEEEEEE/<EEEEEEEEE<AAEE<EAEAAA<EAEEA<<AA<A//<AEA6<	NH:i:1	HI:i:1	AS:i:300	nM:i:0
NB501939:113:H3JNTBGXC:1:22107:10482:1322	83	Supercontig21	10782	255	151M	=	10777	-156	CACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCTCCAGTTTTGCGCCC
G	AA</<E/EA<EA<A6EEAEA/<EAE<E<AEEEE<EEA/EEEEE<EEEAAE/EEEEEEEAE<<EEA<E<E/EEEAAEEEEEEEEEEEEE/EEEAEEE/EEEEEEEEEEAAE/EEEEEA///EEEAAAEEEAE/E/EEEAEEEEE/EEEAAAAA	NH:i:1	HI:i:1	AS:i:195	nM:i:0
NB501939:113:H3JNTBGXC:1:22308:16779:18879	83	Supercontig21	10782	255	151M	=	10767	-166	CACAGCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCTCCAGTTTTGCGCCC
G	6AAAAAEEEAEEA<AEEEE/AEAEEEAEAA<AE/<AEAEEEAEEEEEAEEEEAEEEEEEEEEAEEEEEEEEEEE//EEEEE6EEEEEEEEEEEAEEEEEEEEAEEEEEEEEEEEE6EEEE/EEEEEEEEEEEEEEEEEEEEEEEEEAAAAA	NH:i:1	HI:i:1	AS:i:300	nM:i:0
NB501939:113:H3JNTBGXC:3:23506:19517:8309	163	Supercontig21	10784	255	150M	=	10989	474	CAGCTTAGACTAGATGATGATCACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGCCGAGATGGCTCCAGTTTTGCGCCCCGG
	A6AAAAEEEEEEEEEEAE/E/EEEAEE/E/E/EEEEEE6E/EEEEEEEEEEAEEEAEEEA/EEEEEEEEE6AEEE//EEA<6<<E</<EEAA/EEE</6/<AE/AAE/EAAAE<E</E///</</<</6A/A/A6<<</A<6A<A//A<<	NH:i:1	HI:i:1	AS:i:295	nM:i:3
NB501939:113:H3JNTBGXC:1:22203:5046:18199	163	Supercontig21	10786	255	151M	=	10942	425	GCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTAGTTAGTTAGACTCTAGGACGGTGATGAGTCCTCCGTTTTAACGAGATATACCCTAGCTTTATCGTTCAAAAGACGAGATGGCTCCAGTTTTGCGCCCCGGTG
C	AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/EEEAEE/EEEEEEEEEEEEEEA/EEEEEEE/EEEEEEEEEE<EEEEEEEEEEEEAEEAEAEEEEAEEEEEEEEEEEEEEAAE<AEEEEAAEE<<AAAAAAA/<<6A//<<<A	NH:i:1	HI:i:1	AS:i:302	nM:i:0
NB501939:113:H3JNTBGXC:4:12411:13676:10327	163	Supercontig21	10786	255	48M56N103M	=	10818	356	GCTTAGACTAGATGATGAACACTACTCTTTATAGCTAAACATATACTACTTTATCGTTCAAAAGACGAGATGGCTCCAGTTTTGCGCCCCGGTGCTATATCCTCATACCTCGTATTTATCATCATCACTGTGACCCTTGGTT
```

IGV – Software para visualização gráfica de um arquivo .bam



Finalizar a tabela de resultados

Table 1. General features of RNA-seq reads mapped to the *T. rubrum* reference genome

Sample	Raw reads	High-quality reads	Mapped reads STAR	Total mapped reads (%)
0h II				
0h III				
3h II				
3h III				
12h II				
12h III				