Assignment – Margarida Pinheiro (up201805012)

In order to produce a full transcriptome assembly and annotation for two specific transcripts in the transcriptome of *S. pombe* using IGV we need to use the files obtained by completing the tasks during the class.

We also need the results of the quantification of the transcripts that are stored in the Sp_ds, Sp_hs, Sp_plat and Sp_log folders under the rnaseq folder. In each of these folders is contained the file **named quant.sf.genes**, that stores information about transcript expression measured by TPM (Transcripts per Million).

The first step in this work is to join the four **quant.sf.genes** files. In order to do this I used the `cat` command with the directory of the various files. This step resulted in the **quant.sf.genes_unsorted** file.

```
[up201805012@mbge assignment]$ cat /home/up201805012/rnaseq/Sp_ds/quant.sf.genes /home/up201805012/rnaseq/Sp_hs/quant.sf.genes /home/up201805012/rnaseq/Sp_log/quant.sf
.genes /home/up201805012/rnaseq/Sp_plat/quant.sf.genes > quant.sf.genes_unsorted
```

The most recent file is disordered is unordered. For this work we need to inspect the whole file, searching for the line which has the highest TPM. To do so, I used the command `sort`. This step resulted in the **quant.sf.genes_sorted** file.

Next we need to remove the isoforms. For that, I used the `awk` command and the `cat` command to add an order to the lines of the file. This step resulted in the **quant.sf.genes_clean**.

```
[up201805012@mbge assignment]$ sort -nrk4 quant.sf.genes_unsorted > quant.sf.genes_sorted
[up201805012@mbge assignment]$ awk '{print}' quant.sf.genes_sorted | awk '!x[substr($1, 1, length($1)-3)]++' | cat -n | head -n 25 | column -t > quant.sf.genes_clean
```

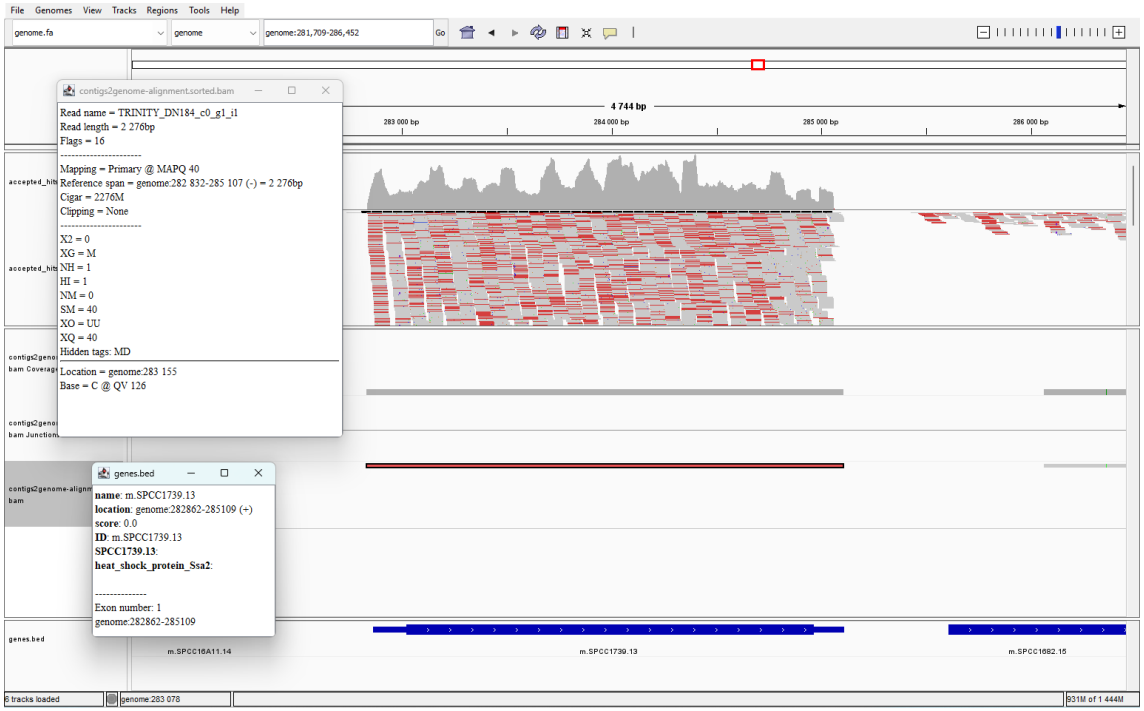
Finally, I added the titles to the file using the `awk` command. This step resulted in the **quant.sf.genes_final**.

```
[up201805012@mbge assignment]$ awk -F, 'NR==1 {print "Order","Name","Length","EffectiveLength","TPM","NumReads"} {gsub(/"/,""); print}' quant.sf.genes_clean | column -
t > quant.sf.genes_final
```

Below is the final file needed to find the actual transcripts given their order.

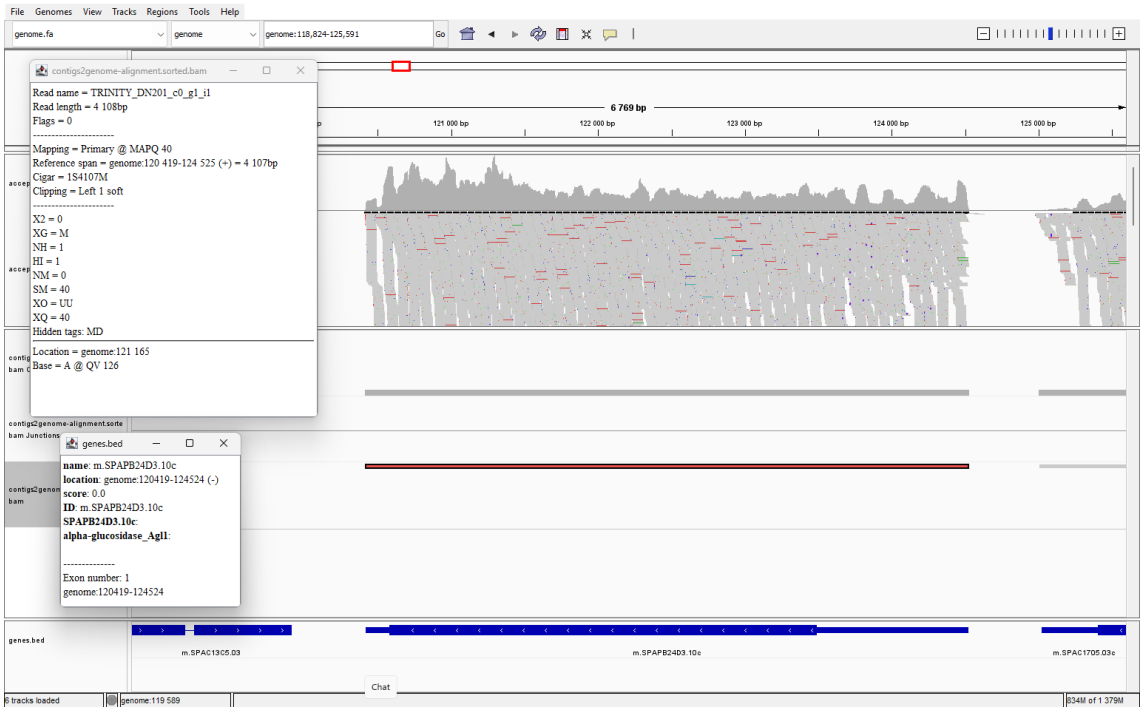
```
[up201805012@mbge assignment]$ cat quant.sf.genes_final
Order  Name                 Length   EffectiveLength  TPM        NumReads
1      TRINITY_DN200_c0_g1  870.00   604.88           213214.09  20920.00
2      TRINITY_DN15_c0_g1   488.00   222.89           124573.49  4504.00
3      TRINITY_DN184_c0_g1  2276.00  1995.65          85248.89   21278.06
4      TRINITY_DN1_c0_g1    1303.00  1039.69          84232.04   10462.00
5      TRINITY_DN197_c0_g1  309.00   71.44            77864.89   433.00
6      TRINITY_DN220_c0_g1  640.00   374.89           63392.87   3855.00
7      TRINITY_DN190_c0_g1  2573.00  2292.65          54079.30   15507.00
8      TRINITY_DN186_c0_g1  546.00   279.78           50832.86   1107.00
9      TRINITY_DN185_c0_g1  1701.00  1437.69          42462.67   7293.00
10     TRINITY_DN162_c0_g5  211.00   18.73            38850.80   91.00
11     TRINITY_DN139_c0_g1  234.00   22.17            34265.49   95.00
12     TRINITY_DN168_c0_g1  2771.00  2507.69          34221.66   10252.00
13     TRINITY_DN167_c0_g1  2947.00  2666.65          31428.17   10482.00
14     TRINITY_DN203_c0_g1  703.00   436.78           26795.45   911.00
15     TRINITY_DN191_c0_g1  2094.00  1830.69          25038.88   5476.00
16     TRINITY_DN32_c0_g1   1456.00  1190.88          24283.93   4691.00
17     TRINITY_DN201_c0_g1  4108.00  3844.69          23966.98   11008.00
18     TRINITY_DN48_c0_g1   1141.00  875.88           22874.99   3258.00
19     TRINITY_DN210_c0_g1  1935.00  1671.69          19368.55   3868.00
20     TRINITY_DN44_c0_g1   1264.00  983.65           19361.67   2382.00
21     TRINITY_DN74_c0_g1   217.00   21.41            18604.31   31.00
22     TRINITY_DN57_c0_g1   222.00   23.45            18078.84   33.00
23     TRINITY_DN161_c0_g1  244.00   35.81            16011.34   93.00
24     TRINITY_DN188_c0_g1  1984.00  1717.78          15750.51   2106.00
25     TRINITY_DN158_c0_g1  269.00   47.44            15723.32   121.00
```

The two transcripts assigned to me are the 3º (Gene 1) and 17º (Gene 2).
After placing the files obtained by completing the tasks into the IGV, I started by looking for the transcript in the line 3 (**TRINITY_DN184_c0_g1**). The result obtained in the IGV was the following:



The transcript **TRINITY_DN184_c0_g1_i1** with 2276bp length is located between positions 282,832 and 285,107 in the genome of *S. pombe*. It corresponds to a heat_shock_protein_Ssa2 protein.

After following the same process for the transcript of the line 17 (**TRINITY_DN201_c0_g1**), the result was the following:



The transcript **TRINITY_DN201_c0_g1_i1** with 4108bp length is located between positions 120,419 and 124,525 in the genome of *S. pombe*. It corresponds to a alpha-glucosidase_Agl1 protein.