Weekly Report

# 18/04/2022

BOUDEMIA ALA EDDINE

TEAM CHEN

# Goal:

* Remove overlapping genes and validate RFD profiles
* Re-extract mutational signatures after removal of overlapping genes

# Procedure:

* Extract protein coding genes from gff3 annotation file -previously done-
* Sort the genes by strand, then by chromosome, then by the starting position and finally by the ending position. If two or more genes start at the same position, this will sort them from shortest to longest.
* For each gene, it will be compared to the following gene and if there is an overlap the longest gene will be kept.
* Depending on which gene is removed, the index will be corrected, this is useful if more than two genes are overlapping so none of them will be skipped.

|  |  |  |
| --- | --- | --- |
| Gene | Start | End |
| 1 | 1 | 130 |
| 2 | 20 | 120 |
| 3 | 50 | 170 |
| 4 | 140 | 250 |

* In this example, genes 2 and 3 will be removed. The intersection is calculated as the Start of gene (i+1) – the End of gene (i) and if it is inferior to 0 than it is considered as an overlap.
* As a result, the number of genes was reduced from 19969 to 18340 genes.
* Following, I plotted the average profiles using the two modes in deeptools (reference-point and scale-region) for a window size of 10 and 3 kb.
* I extracted the TSS and TTS regions from the non-overlapping genes file instead of gff3 annotation.
* Then, I redistributed the mutations from the original MAF file on these regions and removed any common mutations. The number of mutations is as follows:

|  |  |  |
| --- | --- | --- |
| Region/Cancer | BRCA | BLCA |
| TSS | 14682 | 19777 |
| TTS | 23781 | 26742 |
| Common | 6431 | 8346 |

* It seems that there is a significant difference between the number of mutations occurring on the TSS and TTS with TTS being more prone to accumulate mutations.
* I will need to check the number of mutations on each of these regions after removing the mutations in common. Globally, it is something that worth investigating because maybe the difference is not in the mutational signatures but in the way they are distributed.
* I re-ran the signature extraction on the cluster, it should be ready by next week.

# Results:

* The profiles seem better than last time after removal of overlapping genes but the TSS of the genes on the – strand (lefward plots) seems shifted a little bit from the peak

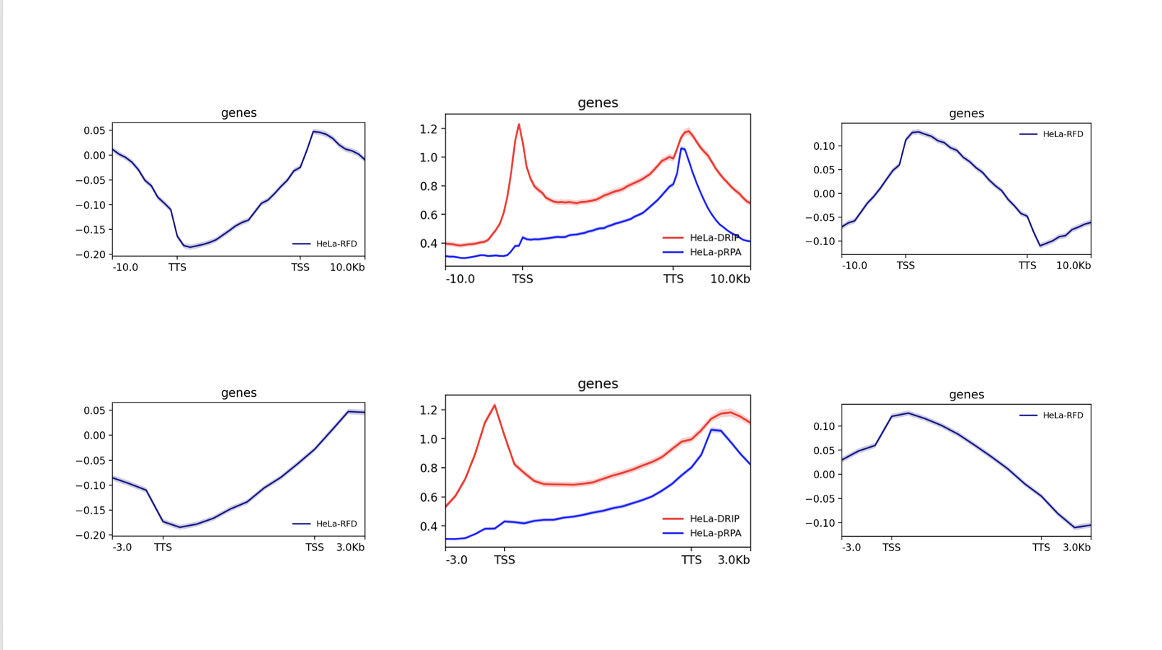


Figure Average profiles (scale-region)

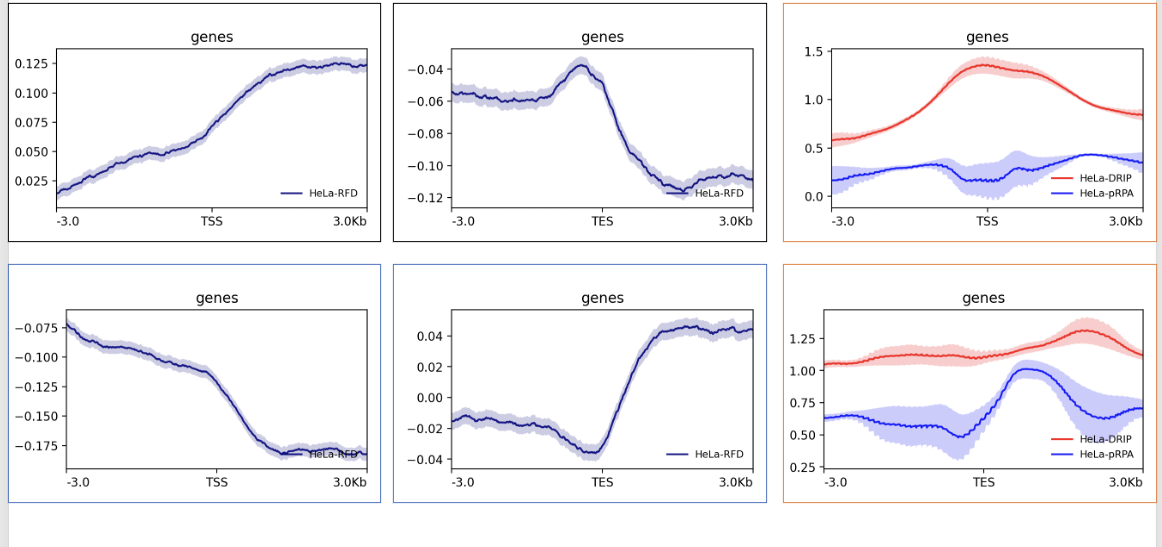


Figure Average profiles (reference-point)

# KDI projects:

Here, I list some KDI projects that could be helpful to us alongside the PM and the amount of data:

|  |  |  |  |
| --- | --- | --- | --- |
| Project link | Type of data | Amount | PM |
| https://bioinfo-portal.curie.fr/#!/project/1821/ | WGS | 4 | Kamal Maud  Le Tourneau Christophe |
| https://bioinfo-portal.curie.fr/#!/project/606/ | Target | 114 | Golmard Lisa |
| https://bioinfo-portal.curie.fr/#!/project/904/ | Target | 219 | Bieche Ivan |
| https://bioinfo-portal.curie.fr/#!/project/624/ | WGS | 2 | Chaligne Ronan  Heard Edith |
| https://bioinfo-portal.curie.fr/#!/project/884/ | WGS | 1 | Andrieu Nadine  Lesueur Fabienne |
| https://bioinfo-portal.curie.fr/#!/project/1274/ | WGS | 10 | Carreira Aura |
| https://bioinfo-portal.curie.fr/#!/project/1184/ | WGS | 9 | Vallot Celine |

The Target data seems interesting regarding its size, while the rest of WGS is not enough to perform the extraction. We can use them with the fitting approach to check if there will be any differences in the signature’s composition. However, using the fitting approach could impact the results and if there will be a difference it might be due to the method. One way possible to counter this is by fitting the current data as well, but still the fitting approaches are a bit complicated because on the current data a bigger number of signatures could be fitted compared to few samples were not all processes are operating.

# Next week planning:

* Perform the analysis using on active and inactive genes (I am worried that there will not be enough mutations for the inactive genes)