Weekly Report

# 02/05/2022 & 09/05/2022

BOUDEMIA ALA EDDINE

TEAM CHEN

# Goal:

* Extend the analysis of mutational signatures further.

# Procedure:

1. Check if RNA-Seq data that I used to define active and inactive genes is from the same patients that provided samples for DNA-Seq data on which mutation calling was performed.
2. Analyze the mutations on TSS and TTS separately.
3. Plot expression profiles of the mutated genes.
4. Classify genes based on their orientation (Convergent, Divergent or Co-Directional):
   1. Read the file of non-overlapping genes previously generated.
   2. Sort genes by chromosome and starting position.
   3. Deduce the starting position of the gene (i+1) from the ending position of gene (i).
   4. If the distance is lower than 0, the region will be discarded (because the first time the non-overlapping was considered only on isoforms -same strand-).
   5. If the distance is superior to 10k bases the region will be noted ‘A’ for ‘Alone’.
   6. If the distance is between 0 and 10k bases, then:
      1. If the orientation of gene (i) is “+” and gene (i+1) is “-” then the region will be noted “C” for ‘Convergent’
      2. If it is the opposite, it will be noted “D” for ‘Divergent’
      3. If both genes are ‘+’ or ‘-’, it will be noted “T”.
   7. The lists of regions will be saved and used to classify the mutations. (-/+ 3kb)
5. Implement cosine similarity §taken directly from sigprofiler source code for reproducibility§ §Previous results that showed slight differences are quite similar with more than 85% similarity§

# Results:

**§ RNA-Seq data is concordant with mutational data:**

* BRCA
  + Number of unique patients in the MAF file is: 986
  + Number of unique patients in the RNA-Seq file is: 1093
  + Number of common patients between the files: 983
* BLCA
* Number of unique patients in the MAF file is: 412
* Number of unique patients in the RNA-Seq file is: 408
* Number of common patients between the files: 408

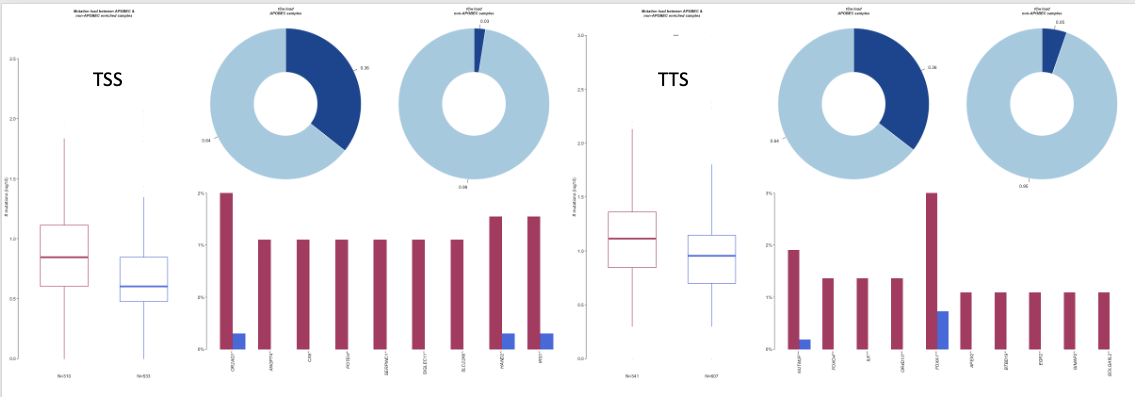
**§ Analysis of mutations at TSS and TTS:**

* Mostly not the same genes that are mutated at TSS and TTS

Tableau 1 Statistics about mutations at TSS and TTS

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cancer | TSS mutated | | | TTS mutated | | | TSS-TTS |
|  | Genes | Median | Mutations | Genes | Median | Mutations | Genes |
| BRCA | 2616 | 2 | 4272 | 4823 | 5 | 8679 | 730 |
| BRCA-Act | 2057 | 2 | 3275 | 3761 | 4 | 6542 | 515 |
| BRCA-Inact | 615 | 1 | 1089 | 1099 | 2 | 2181 | 233 |
| BLCA | 3334 | 10 | 6098 | 5367 | 18 | 10809 | 1124 |
| BLCA-Act | 2726 | 8 | 4916 | 4259 | 14 | 8393 | 876 |
| BLCA-Inact | 683 | 3 | 1315 | 1155 | 4 | 2509 | 274 |

* There are more genes that are mutated at TTS than genes mutated at TSS.
* The genes are more prone to be mutated at TTS than TSS (Median TTS > Median TSS).
* I still can’t figure why the numbers of active and inactive genes don’t add up to the non-separated data.
* Doesn’t seem to be any differences in APOBEC enrichment between TSS and TTS:

Figure APOBEC enrichement at TSS & TTS -BRCA-

* There seems to be some differences in the oncogenic pathways being impacted, in terms of fraction but not in pathways

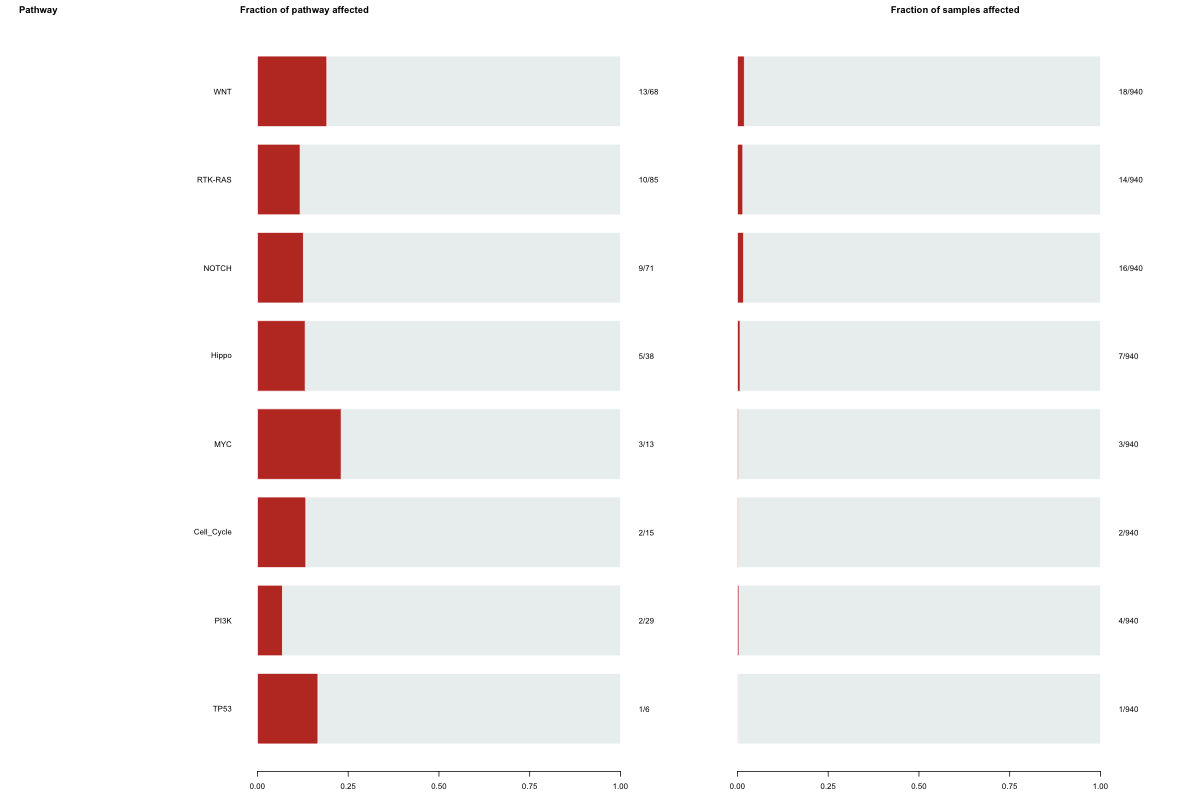


Figure 2 Oncogenic pathways impacted by mutations occurring at TSS -BRCA-

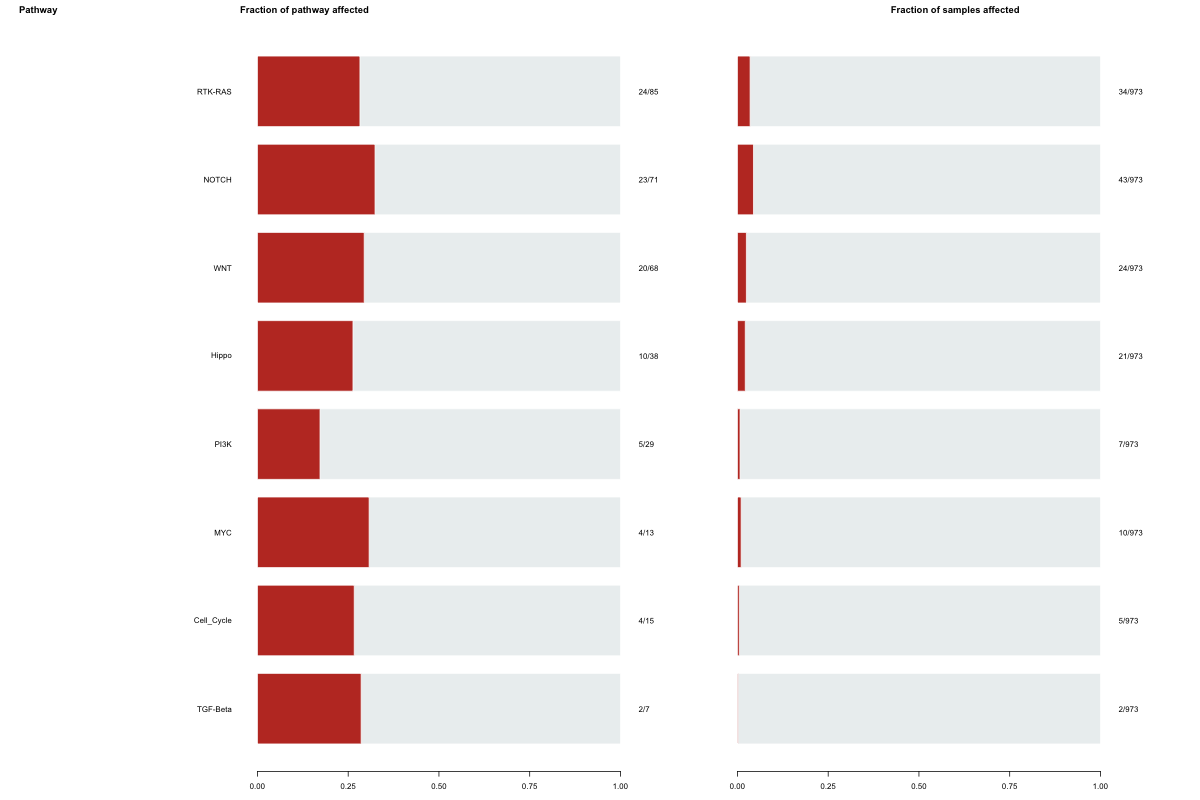


Figure 3 Oncogenic pathways impacted by mutations occurring at TTS -BRCA-

**§ Expression profiles**

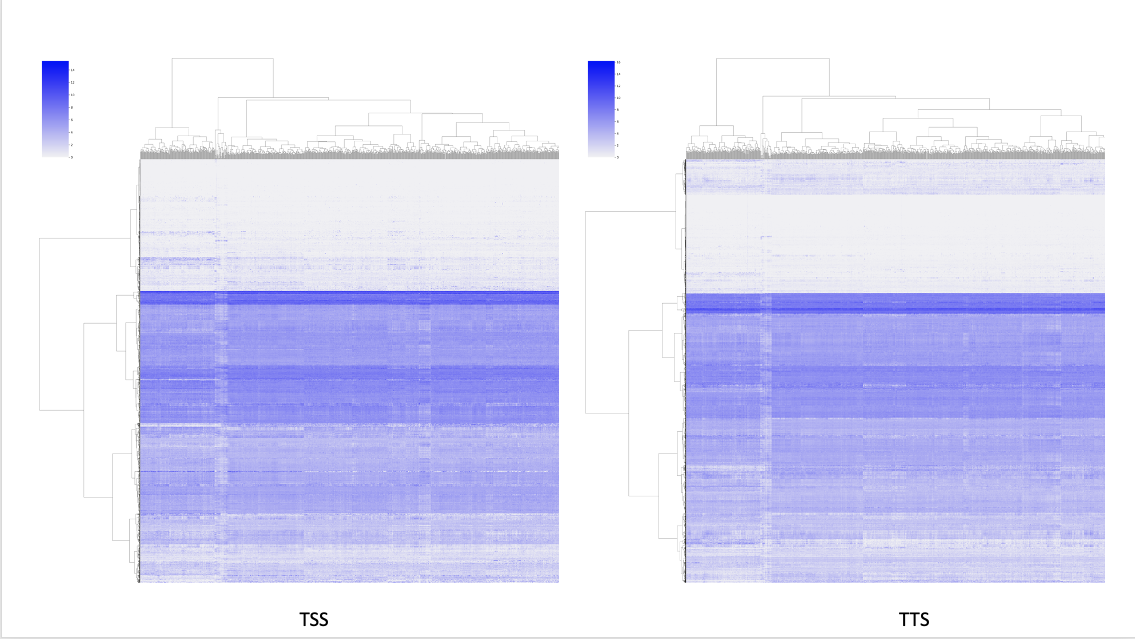


Figure 4 Expression profiles of genes mutated at TSS and TTS or both

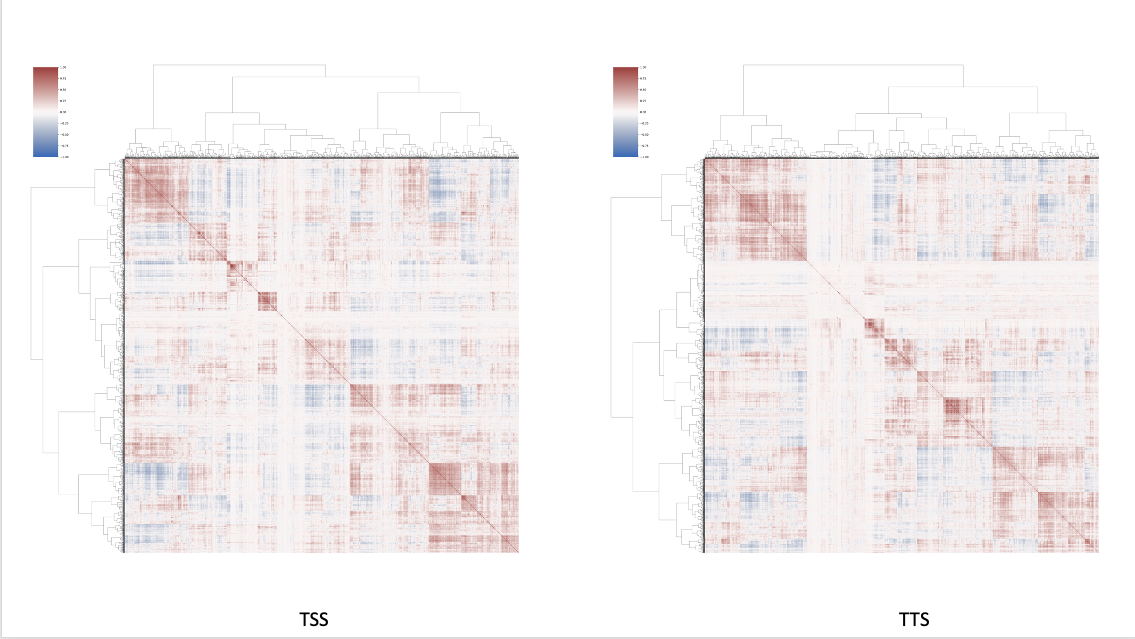


Figure 5 Co-expression profiles of the genes mutated at TSS and TTS or both

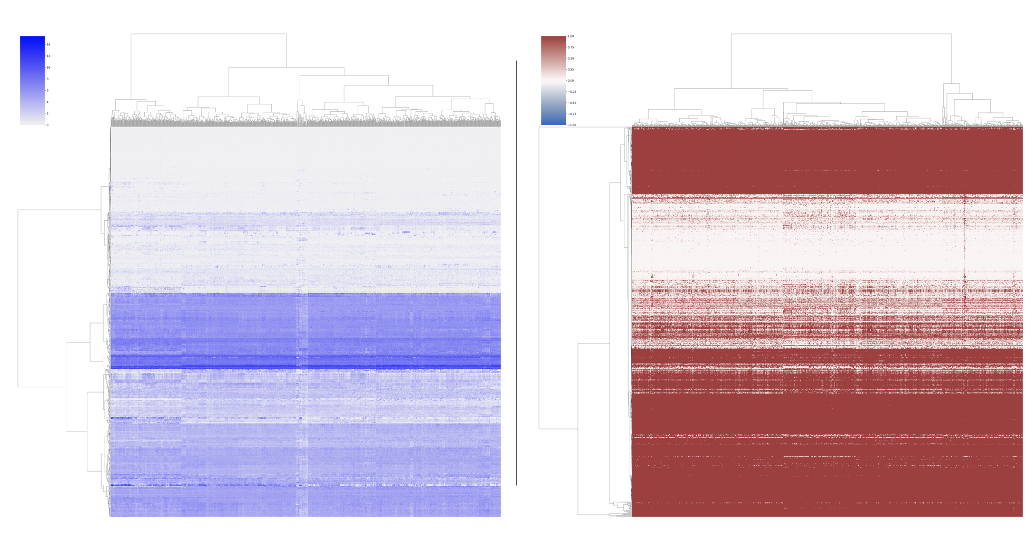


Figure 6 Expression and co-expression profiles of genes mutated at both TSS and TTS -BRCA-

* In all figures, the low levels of expression or where the correlation value is 0 are relative to inactivated genes (Clear in plots where I considered each set separately). Thus, it is remarkable that the active genes that are mutated at both TSS and TTS are highly correlated compared to genes that are mutated at TSS or TTS (with or without genes mutated at both regions). This may indicate that these genes share a lot in terms of transcriptional regulation. (May be gene network analysis would clarify more)

**§ Mutations based on gene orientation**

Tableau 2 Number of mutations at each region

|  |  |  |  |
| --- | --- | --- | --- |
| Genes Orientation | Convergent | Divergent | Co-Directional |
| Number of Genes | 1540 | 1182 | 2932 |
| BRCA Mutations | 111-4303 | 89-1882 | 174-6311 |
| BLCA Mutations | 153-5051 | 107-2706 | 234-8212 |

* WES does not capture intergenic regions which is probably the reason behind the reduced number of mutations