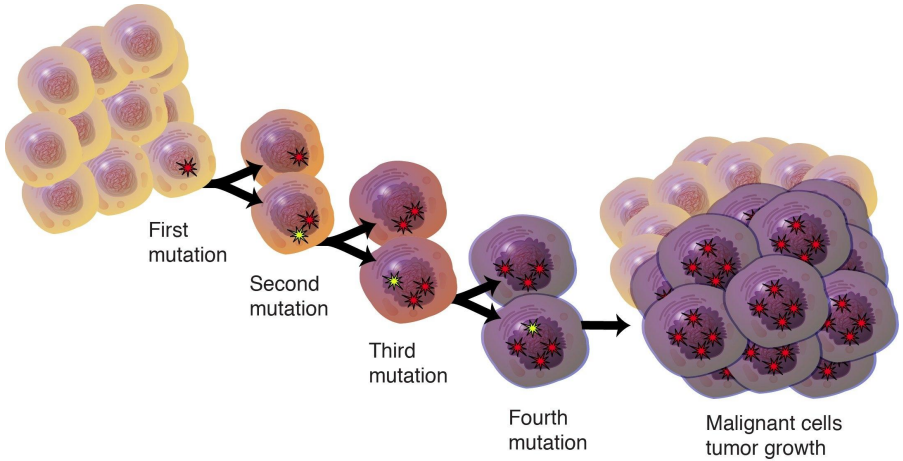
Computational Analysis of TCGA Mutation Signatures



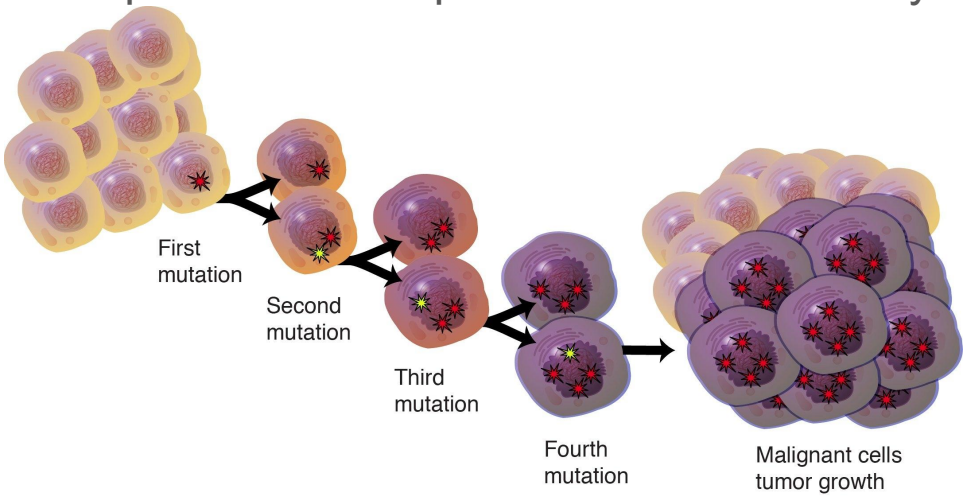
Konur Bayrak - Waleed Iqbal - Blas Rodriguez

Outline

- Background
- Project Overview
- Mutation Data
- Mutation Signatures
- Unfinished Project Parts
- Conclusion

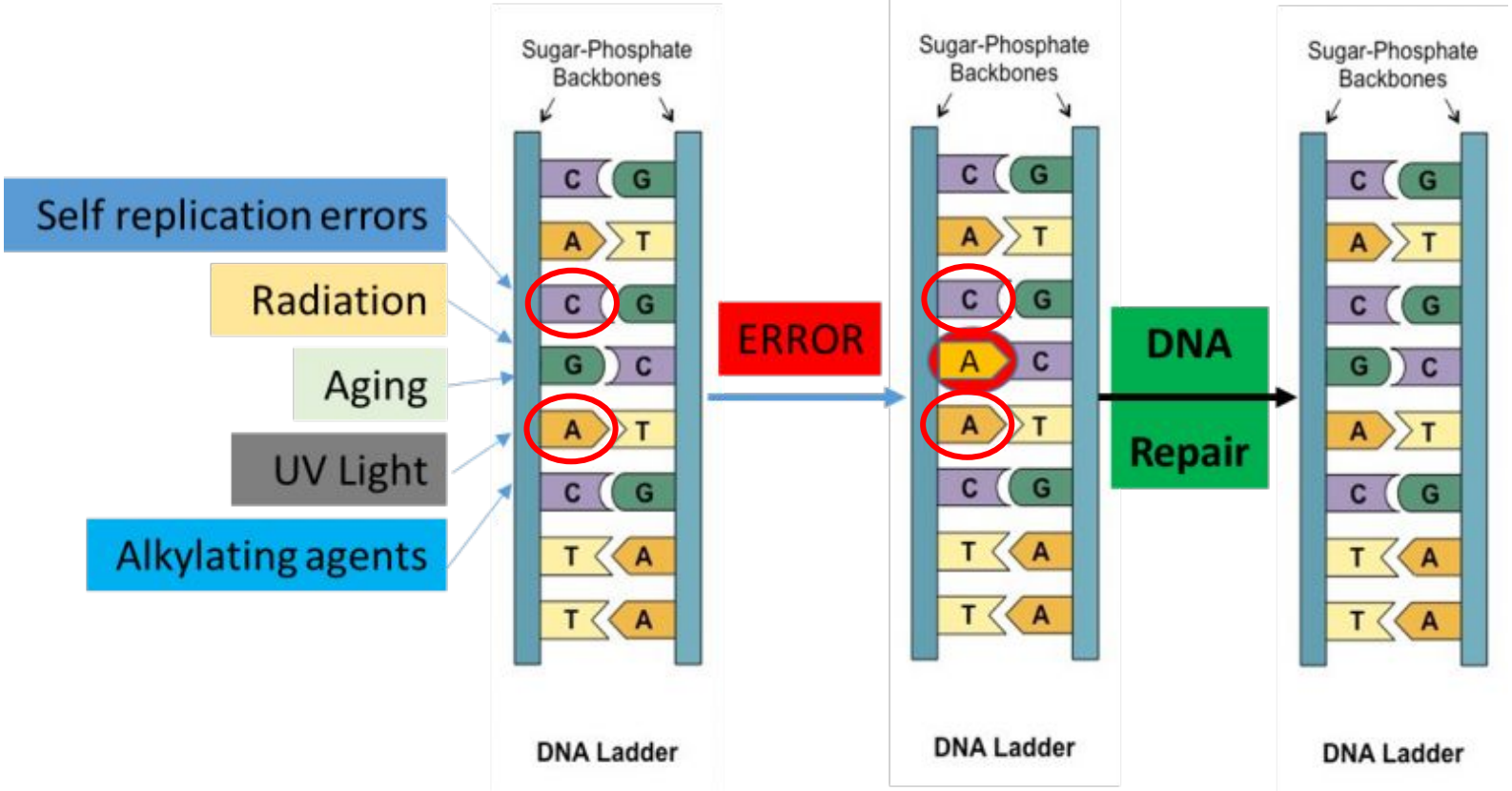
Background

- There are mechanisms in place to prevent and repair mutations that may occur through:
  - Replication errors
  - Radiation
  - Aging
  - UV exposure
  - Carcinogenic chemicals



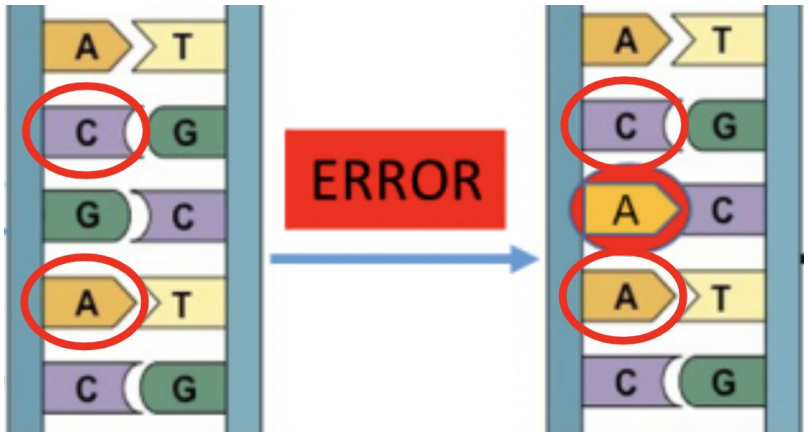
- However, mutations can become too damaging to healthy cell maintenance and lead to tumorigenesis

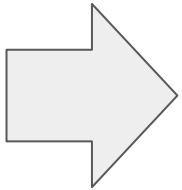
DNA damage and Mutation Signatures

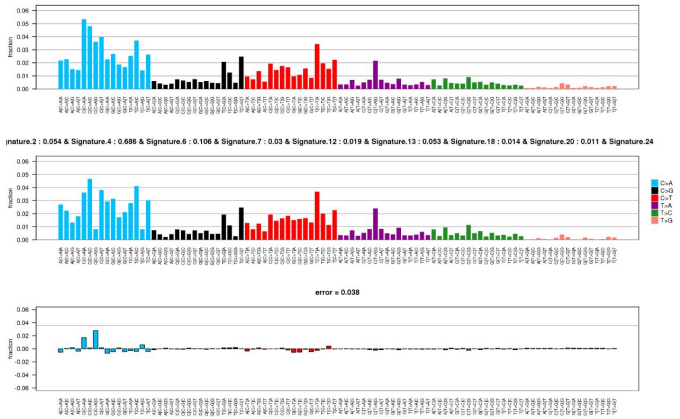


Background

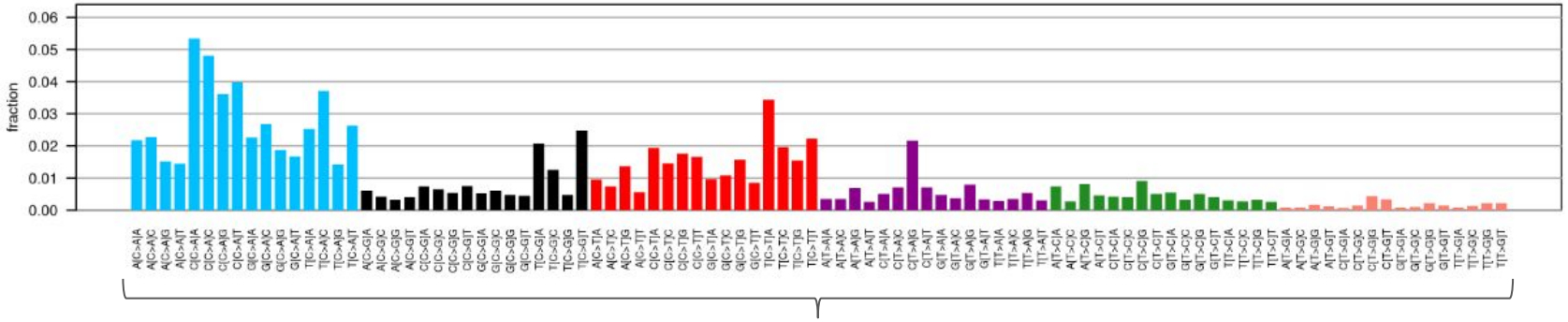
- DNA is the code that makes proteins in our body to perform cellular processes
- This code can be compared to reference healthy DNA to check for alterations
- These alterations can then be clustered into different subtypes based on the nucleotide that is before and after the mutation of interest





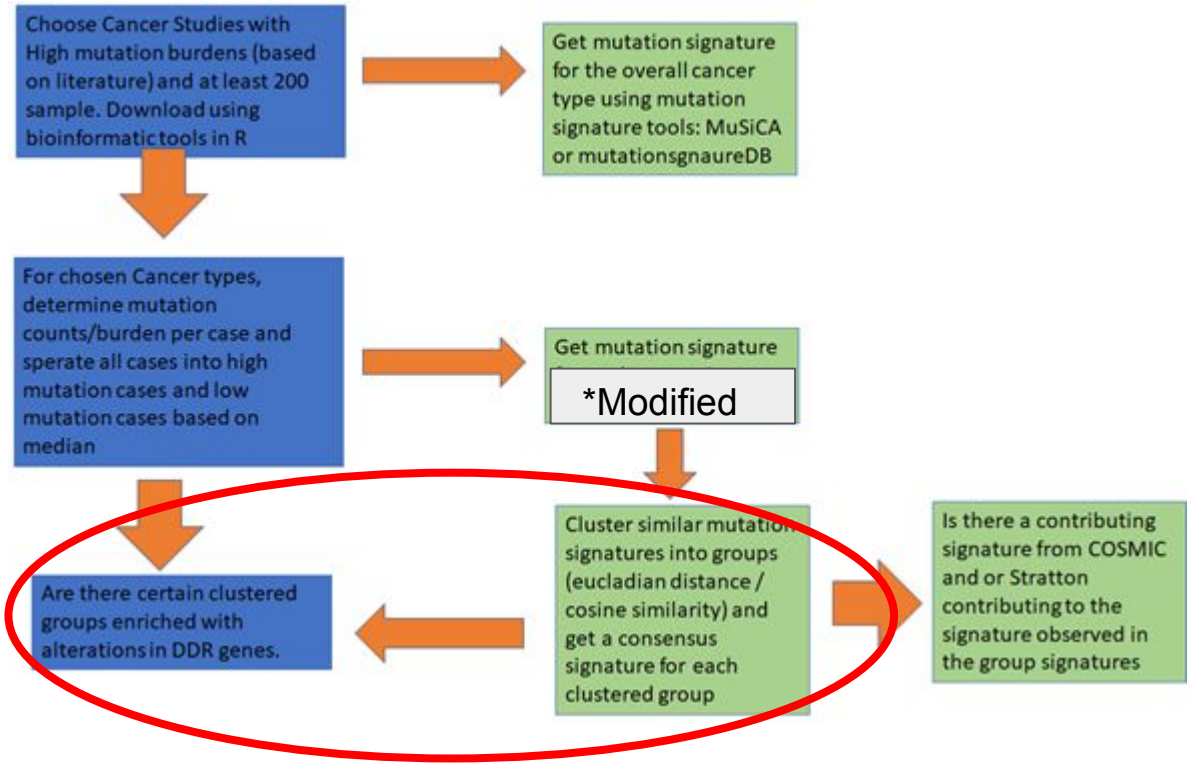


COSMIC Signatures



- Mutations can be classified into 96 possible types based on the nucleotide before and after the mutation
- This classification is based the six substitution subtypes: C>A, C>G, C>T, T>A, T>C, and T>G AND...
- … each of the substitutions is examined by incorporating information on the bases immediately 5’ and 3’ to each mutated base generating 96 possible mutation types

Project Overview



Circled in red are parts we are attempting to complete

Tools used

-TCGAbiolinks : Used to download individual Tumor data sets from The Cancer genome Atlas database (TCGA)

https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html

-MSignatureDB: Database containing TCGA data. Can be used to get overall cancer signatures

http://tardis.cgu.edu.tw/msignaturedb/

-NMF R package: Provides a framework to perform Non-negative Matrix Factorization (NMF)

https://cran.r-project.org/web/packages/NMF/index.html

Tools used

-BSgenome : Specifically the BSgenome.Hsapiens.UCSC.hg38 reference genome

http://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg38.html

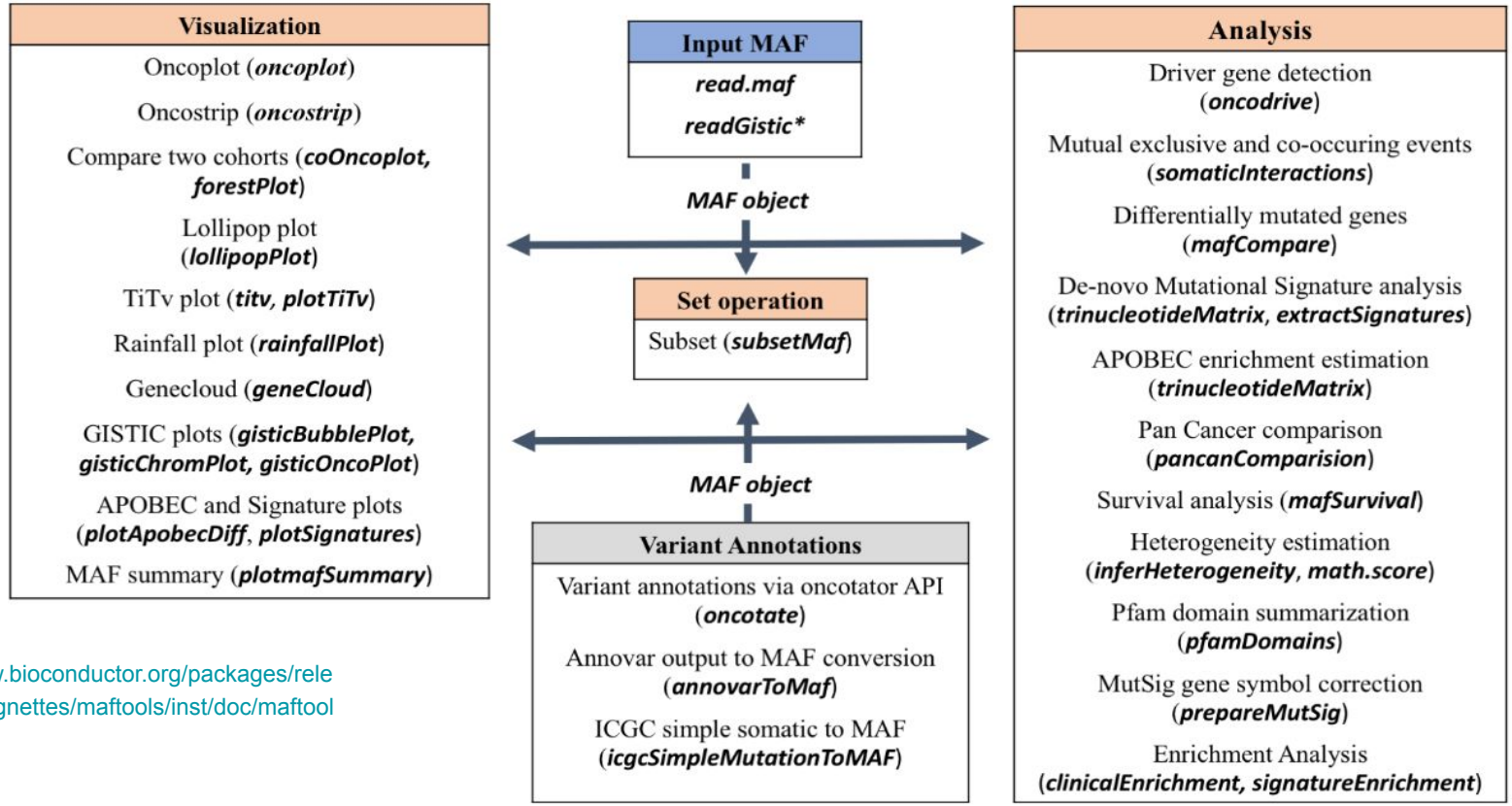
-Maftools: Analyze and visualize Mutation Annotation Format (MAF) files from TCGA

https://bioconductor.org/packages/release/bioc/html/maftools.html

Functions used:

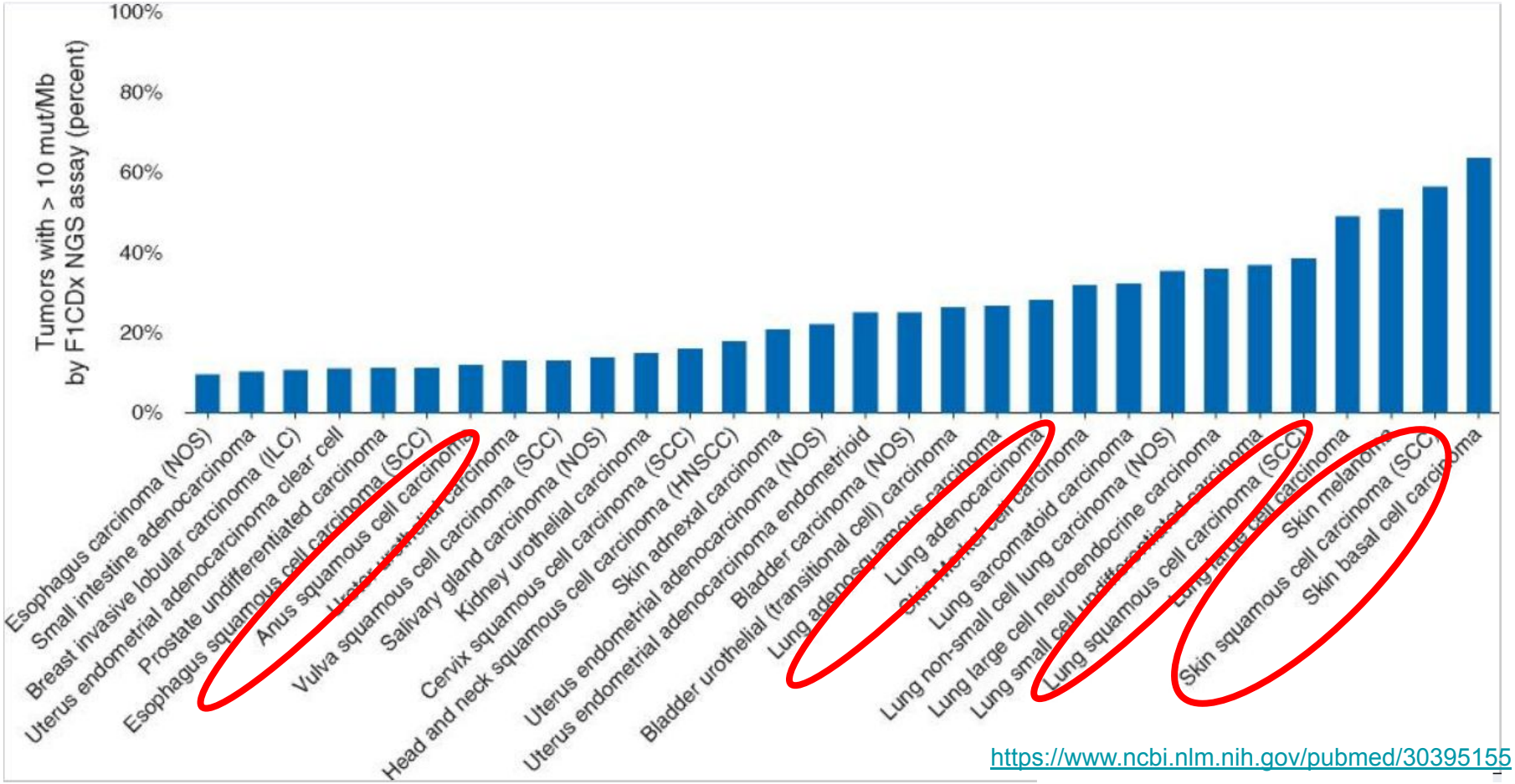
-read.maf, -subset.maf, -mafsummarry, -Oncogenic Pathways, -mutCountMatrix, -trinucleotideMatrix, -extractSignatures

MAFtools is very versatile and contains many other useful functions



https://www.bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html

Cancer Types used

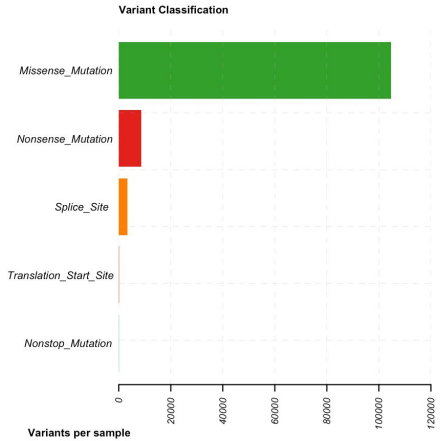


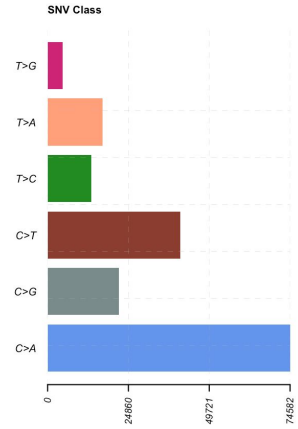
https://www.ncbi.nlm.nih.gov/pubmed/30395155

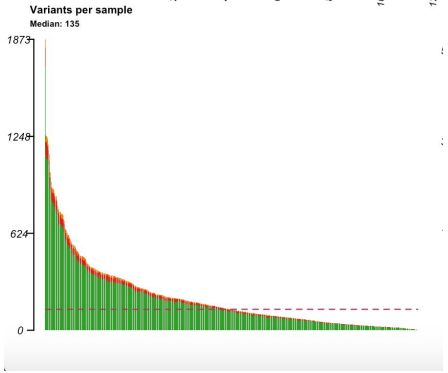
Cancer Types Chosen

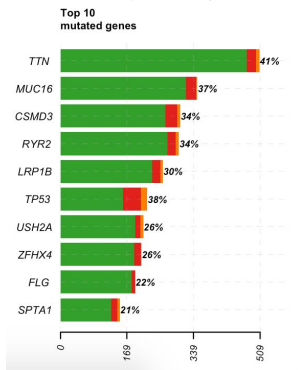
Type of cancer	Amount of Samples	Amount of Mutations	Minimum	Maximum	Median	Mean	Standard Deviation
LUAD	563	172086	1	2823	197	305.65	342.16
LUSC	490	158757	1	3945	257.5	323.99	299.22
SKCM	465	374099	1	26230	416	804.51	1578.4
COAD	399	264786	59	12393	175	663.62	1360.4

Summary of LUAD mutation data



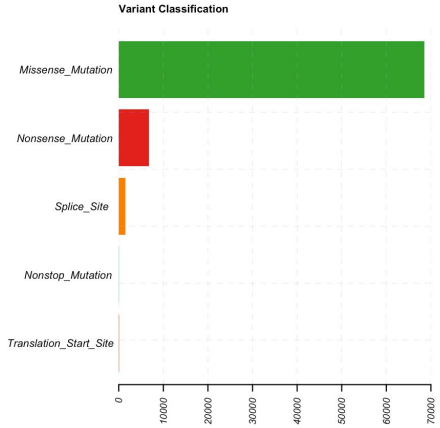


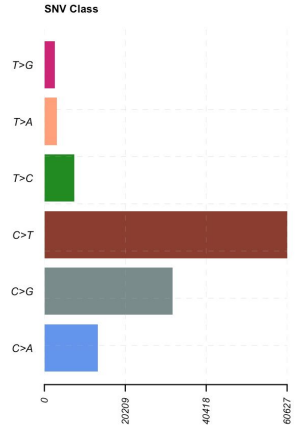


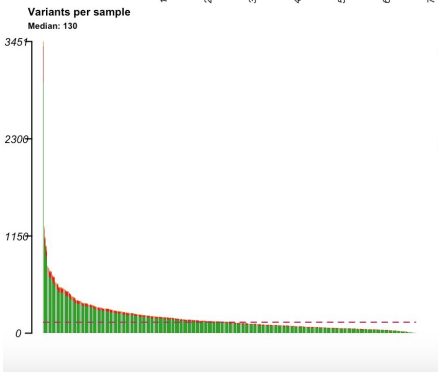


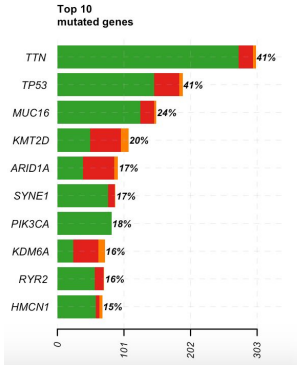
***Summary of LUAD (Lung adenocarcinoma)*.** In LUAD the median variants for per sample is 135 but it can be observed that there are some samples with very high number of variants (uptil ~1873) which is lower than the max of COAD. The top 3 mutated genes are TTN, MUC16 and CSMO3; most of the mutations in all three genes are mostly missense mutations (green). Most single nucleotide variants (CNVs) are C>A changes.

Summary of LUSC mutation data



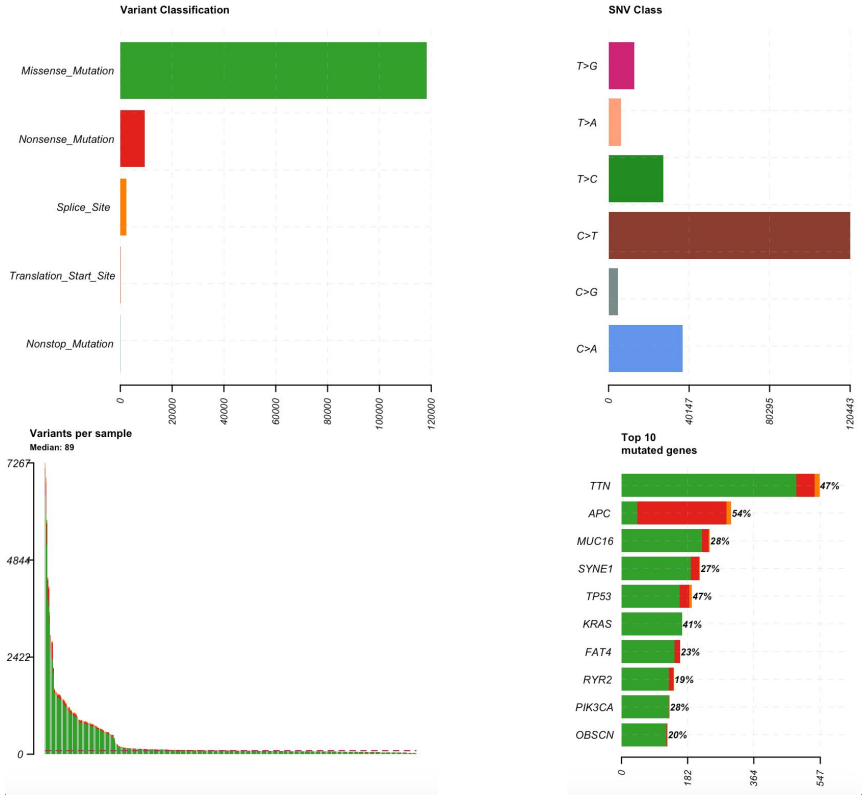






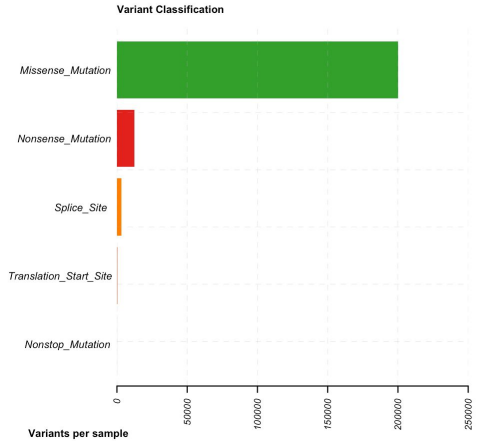
Summary of LUSC (Lung squamous cell carcinoma). In LUSC the median variants for per sample is 130, it can be observed that there are some samples with a high number of variants (uptil ~3451) which is lower than the max of COAD. The top 3 mutated genes are TTN, TP53 and MUC16; most of the mutations in all three genes are mostly missense mutations (green). Most single nucleotide variants (CNVs) are C>T changes.

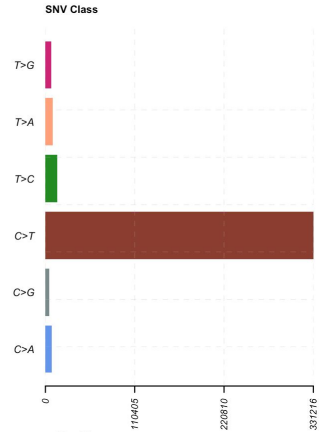
Summary of COAD mutation data

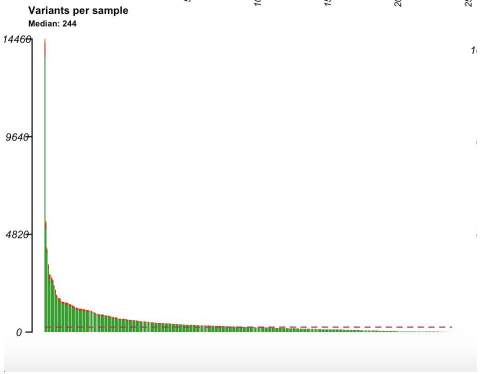


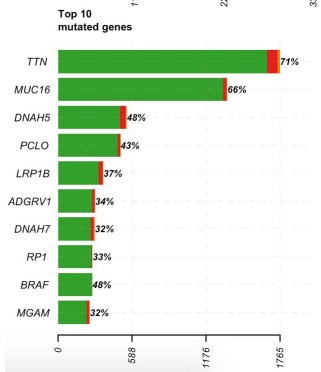
***Summary of COAD (Colon adenocarcinoma)*.** In COAD the median variants for per sample is 89 but it can be observed that there are some samples with very high number of variants (uptil ~7200). The top 3 mutated genes are TTN, APC and MUC16; APC is characterized with mostly loss of function mutations (red). Most single nucleotide variants (CNVs) are C>T changes.

Summary of SKCM mutation data



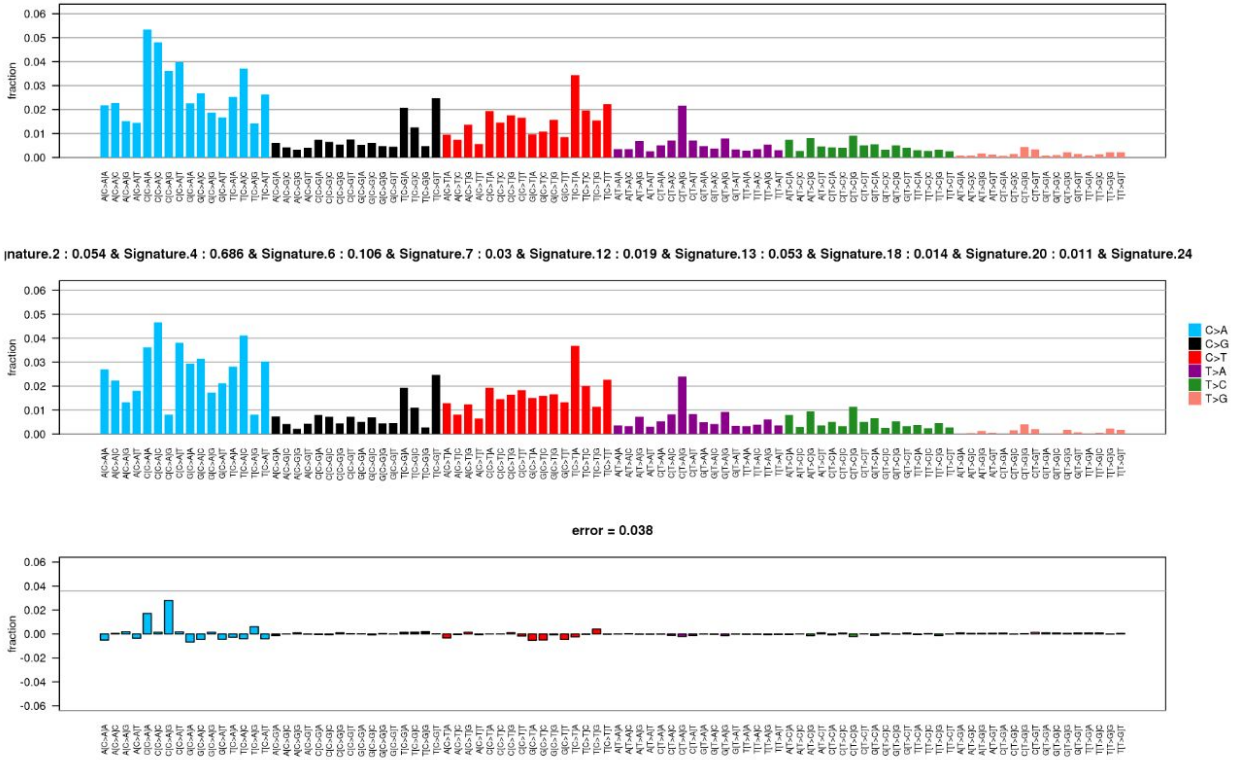




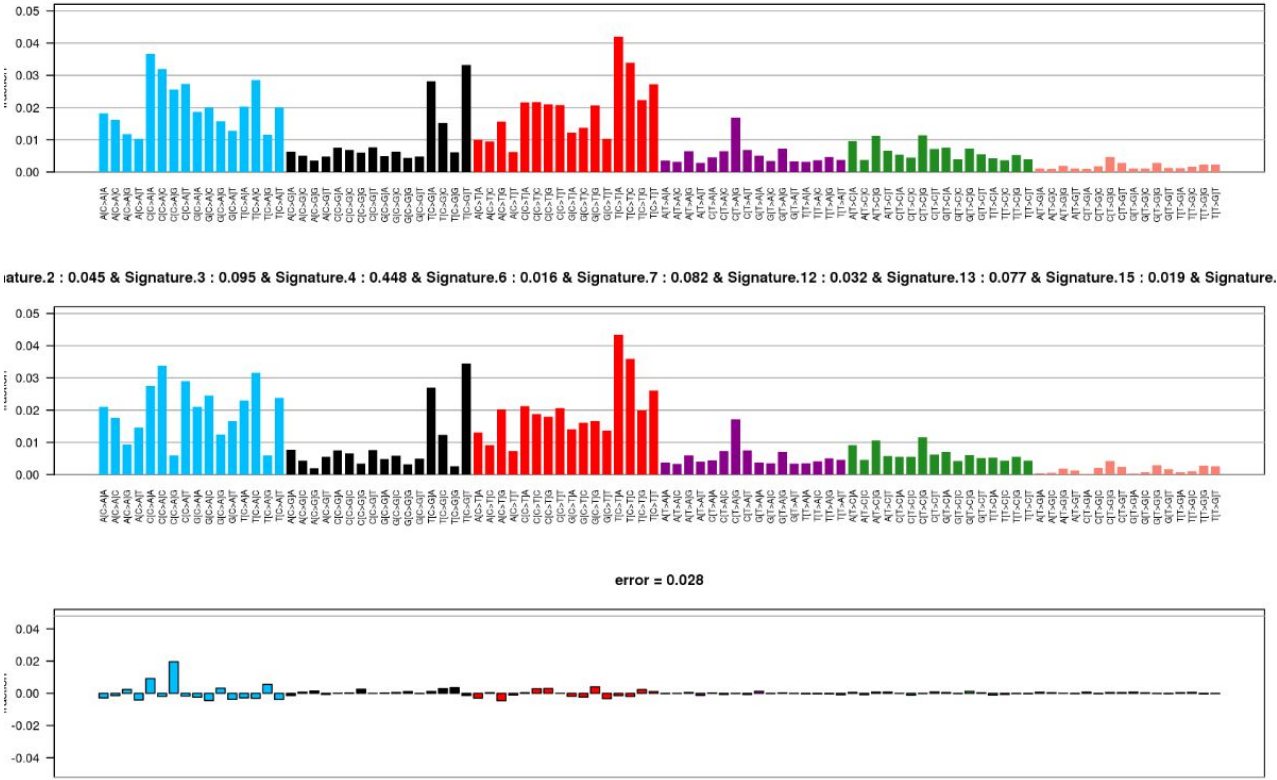


Summary of SKCM (Skin Cutaneous Melanoma). In SKCM the median variants for per sample is 244 but it can be observed that there are some samples with very high number of variants (uptil ~14466) which is lower than the max of COAD. The top 3 mutated genes are TTN, MUC16 and DNAH5; most of the mutations in all three genes are mostly missense mutations (green). Most single nucleotide variants (CNVs) are C>T changes.

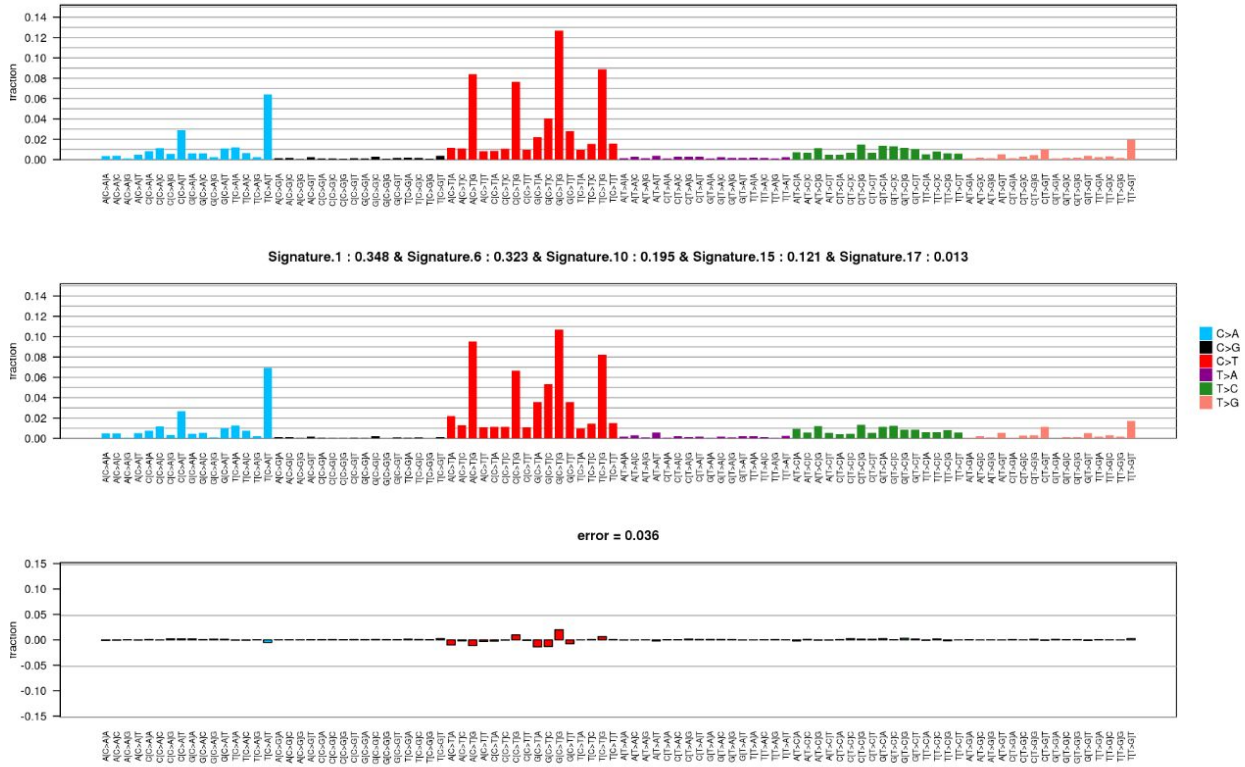
Overall Mutation Signatures: LUAD has Significant C>A and some C>T as well



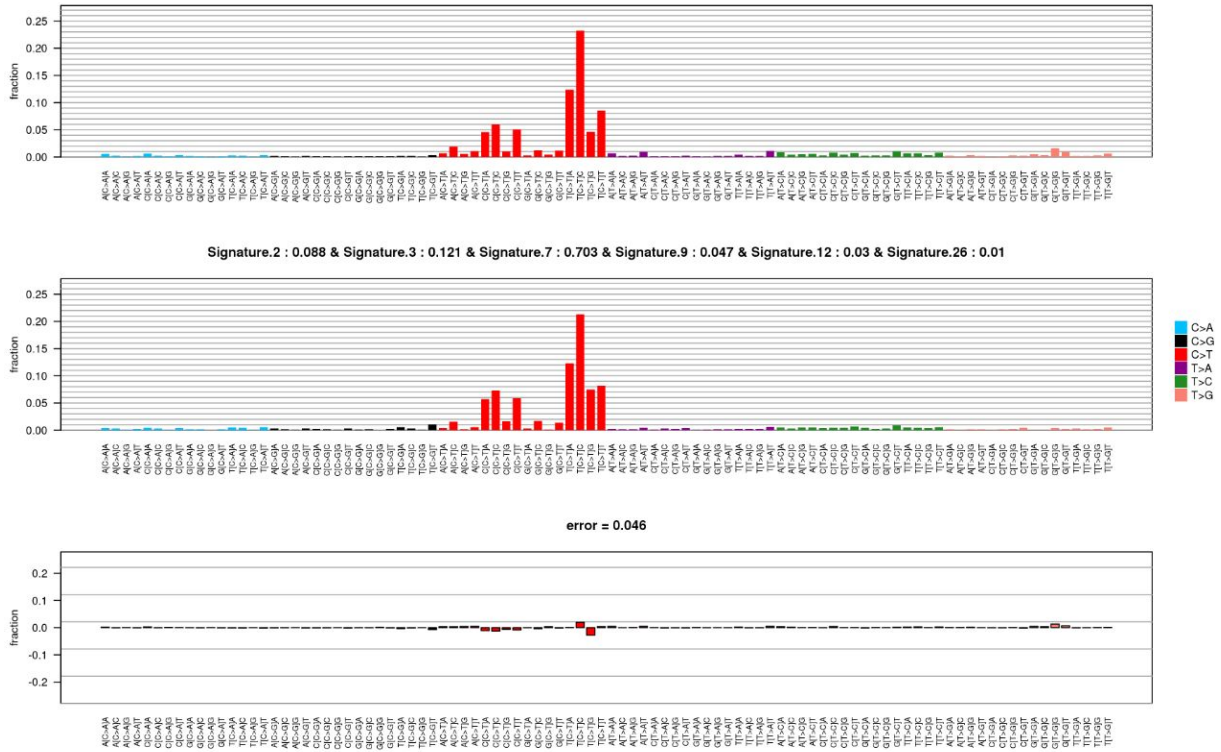
Overall Mutation Signatures: LUSC has a similar mutation Signature to LUAD



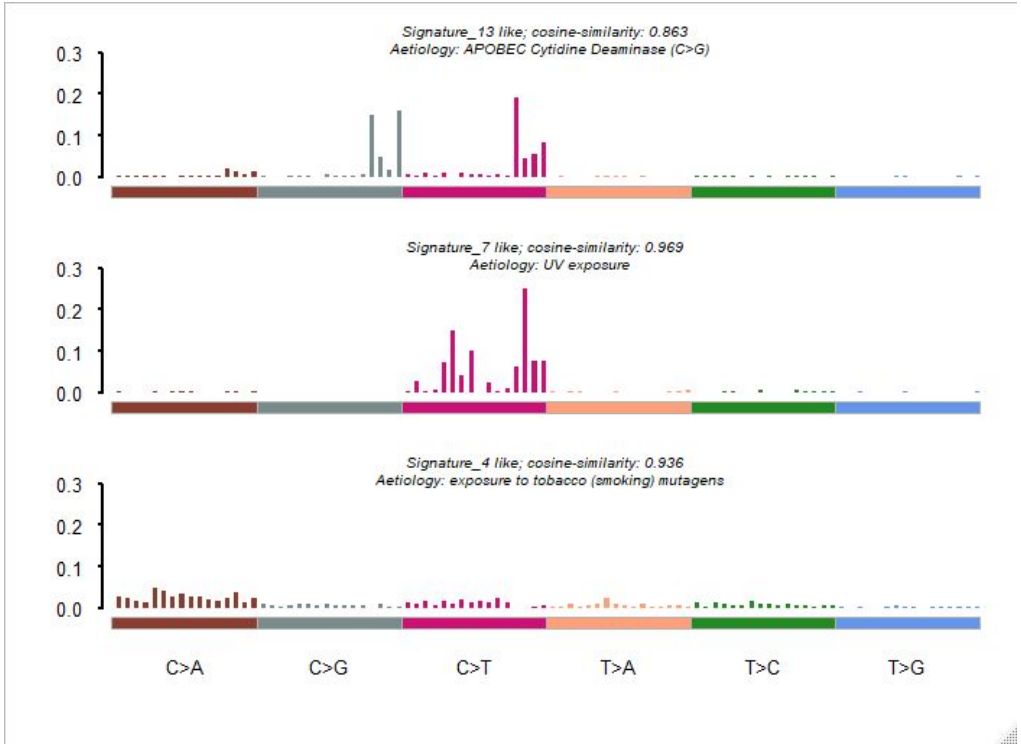
Overall Mutation Signatures: COAD has significant C>T mutations and some C>A mutation



Overall Mutation Signatures: SKCM has mainly C>T mutations



Multiple Decomposed Signatures and comparison to COSMIC signatures: LUAD



-Found Signature_1 most similar COSMIC Signature_13.

Aetiology: APOBEC Cyidine deaminase [cosine-similarity: 0.863]

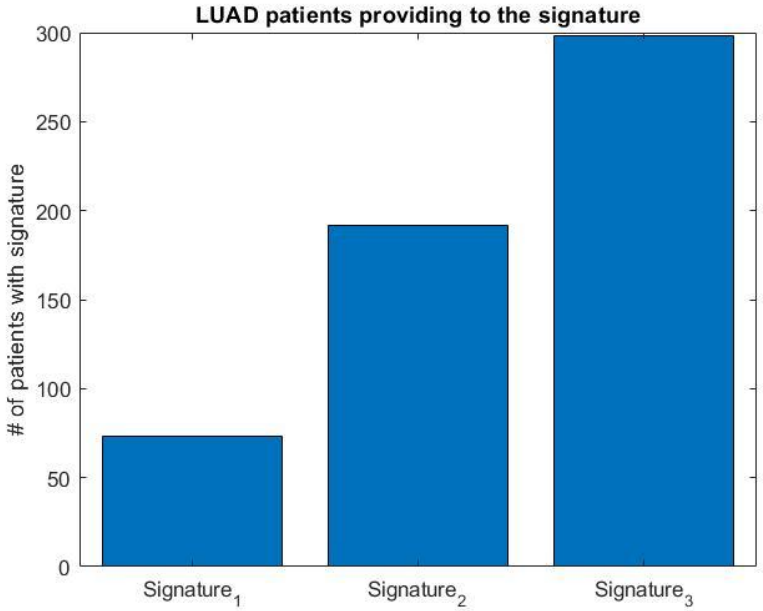
-Found Signature_2 most similar COSMIC Signature_1.

Aetiology: UV exposure [cosine-similarity: 0.969]

-Found Signature_3 most similar COSMIC Signature_4.

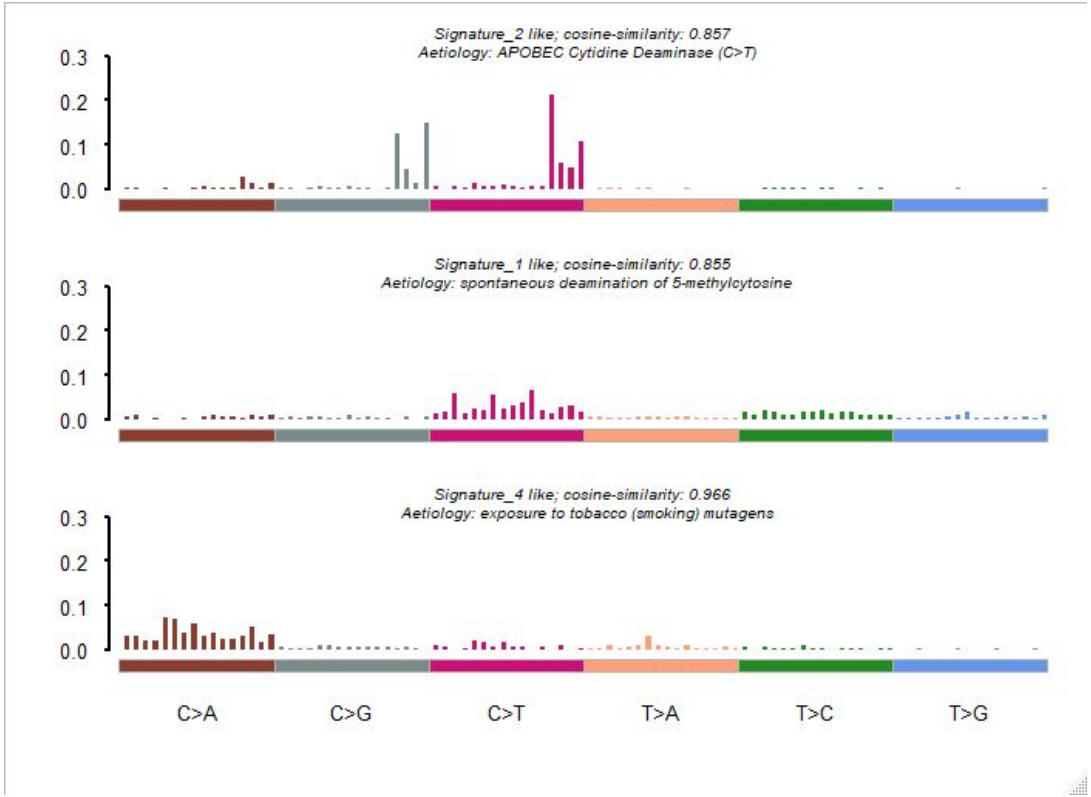
Aetiology: Exposure to Tobacco [cosine-similarity: 0.936]

Demographics of patients LUAD



Total patients = 563
Total mutations = 172086
Min mutation per patient = 1
Max mutation per patient = 2823
Median = 1412
Average = 305.65
Standard Deviation = 342.16

Multiple Decomposed Signatures and comparison to COSMIC signatures: LUSC



-Found Signature_1 most similar COSMIC Signature_2.

Aetiology: APOBEC Cyidine deaminase [cosine-similarity: 0.857]

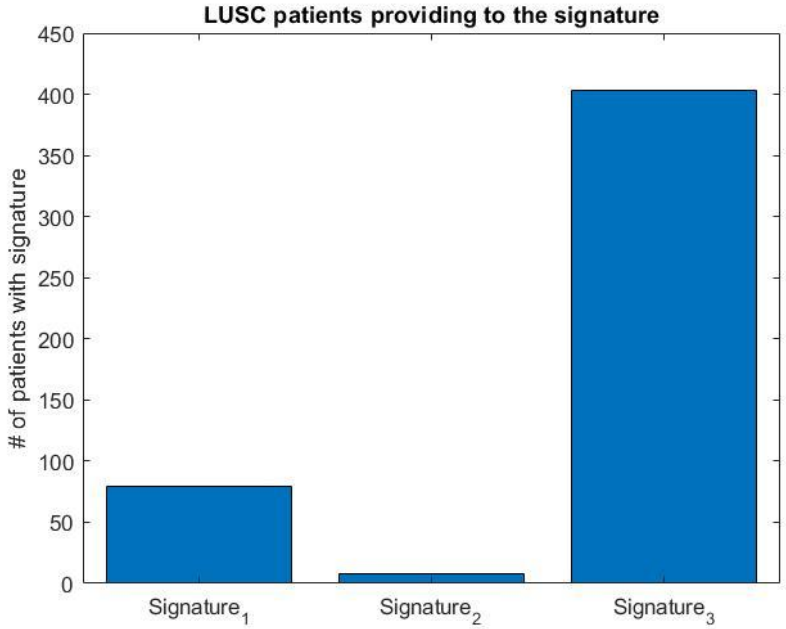
-Found Signature_2 most similar COSMIC Signature_1.

Aetiology: Spontaneous deamination of 5-methylcytosine deamination [cosine-similarity: 0.855]

-Found Signature_3 most similar COSMIC Signature_4.

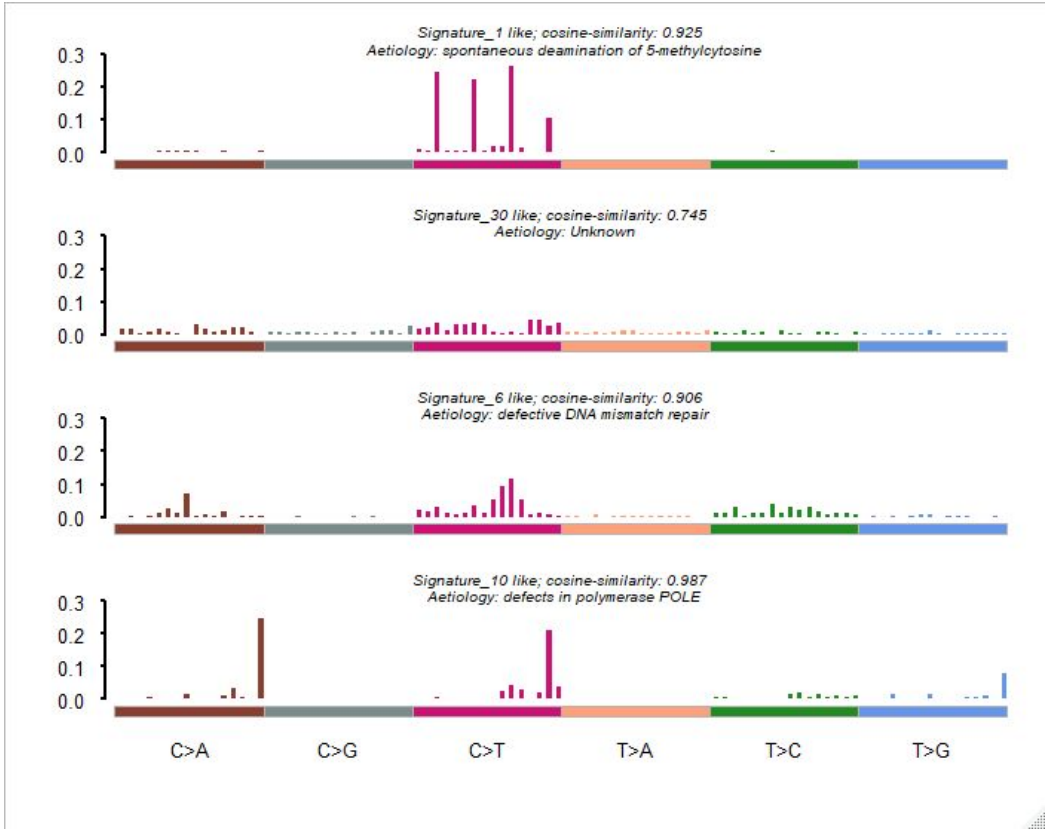
Aetiology: Exposure to Tobacco [cosine-similarity: 0.966]

Demographics of patients LUSC



Total patients = 490
Total mutations = 158757
Min mutation per patient = 1
Max mutation per patient = 1977.5
Median = 257.5
Average = 323.99
Standard Deviation = 299.22

Multiple Decomposed Signatures and comparison to COSMIC signatures: COAD



-Found Signature_1 most similar COSMIC Signature_1.

Aetiology: Spontaneous deamination of 5-methylcytosine [cosine-similarity: 0.925]

-Found Signature_2 most similar COSMIC Signature_30.

Aetiology: Unknown [cosine-similarity: 0.745]

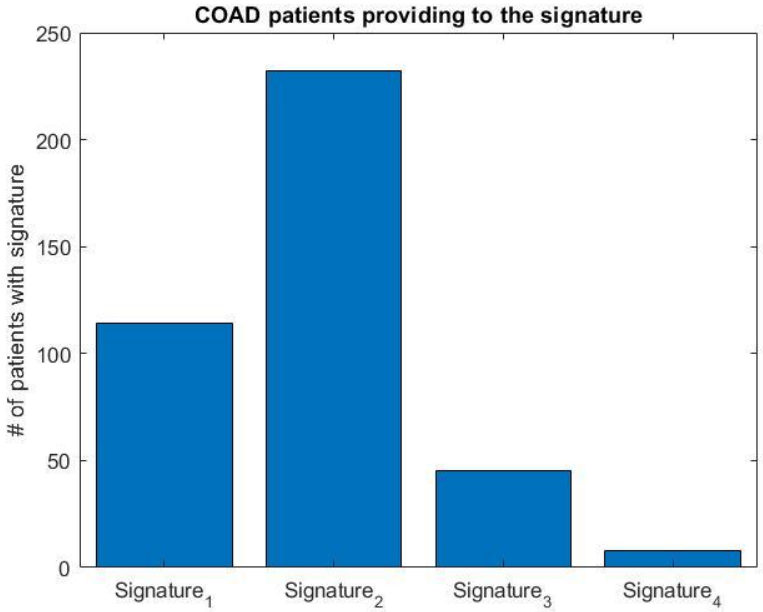
-Found Signature_3 most similar COSMIC Signature_6.

Aetiology: defective DNA mismatch repair [cosine-similarity: 0.906]

-Found Signature_4 most similar COSMIC Signature_10.

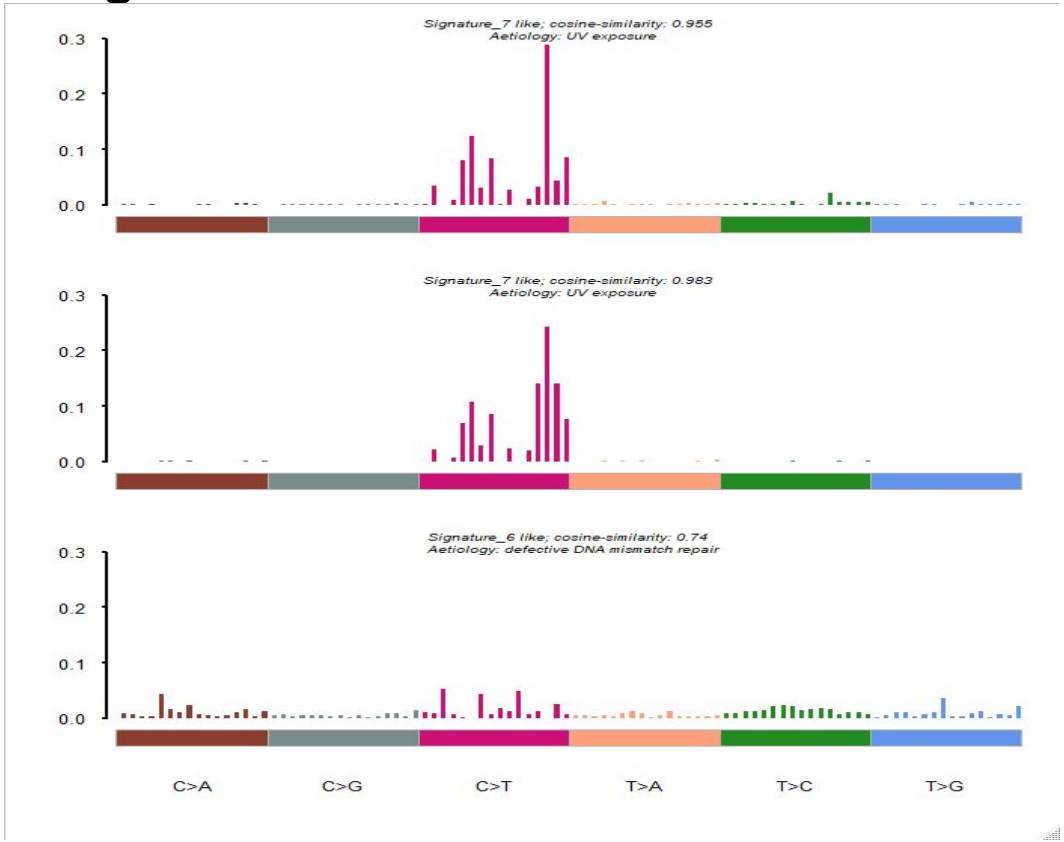
Aetiology: defects in POLE gene repair [cosine-similarity: 0.987]

Demographics of patients COAD



Total patients = 399
Total mutations = 264786
Min mutation per patient = 59
Max mutation per patient = 12393
Median = 6226
Average = 663.62
Standard Deviation = 1360.4

Multiple Decomposed Signatures and comparison to COSMIC signatures: SKCM



-Found Signature_1 most similar COSMIC Signature_7.

Aetiology: UV exposure [cosine-similarity: 0.955]

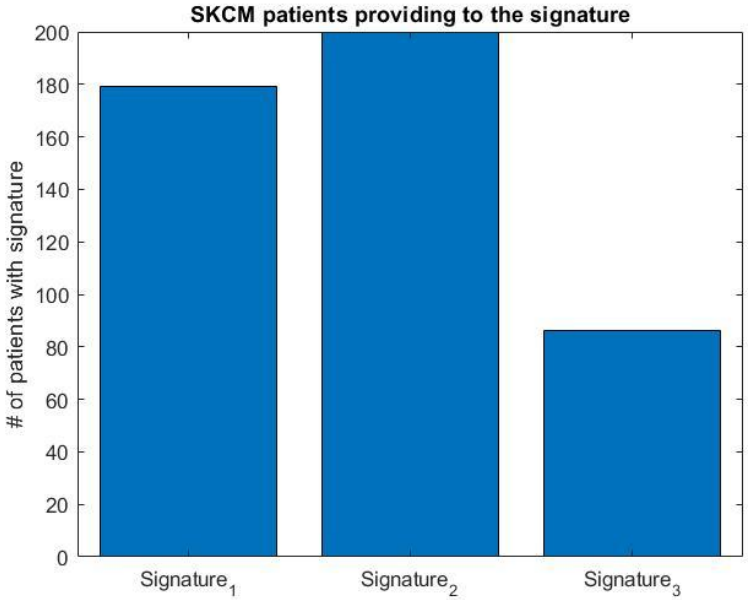
-Found Signature_2 most similar COSMIC Signature_7.

Aetiology: UV exposure [cosine-similarity: 0.983]

-Found Signature_3 most similar COSMIC Signature_6.

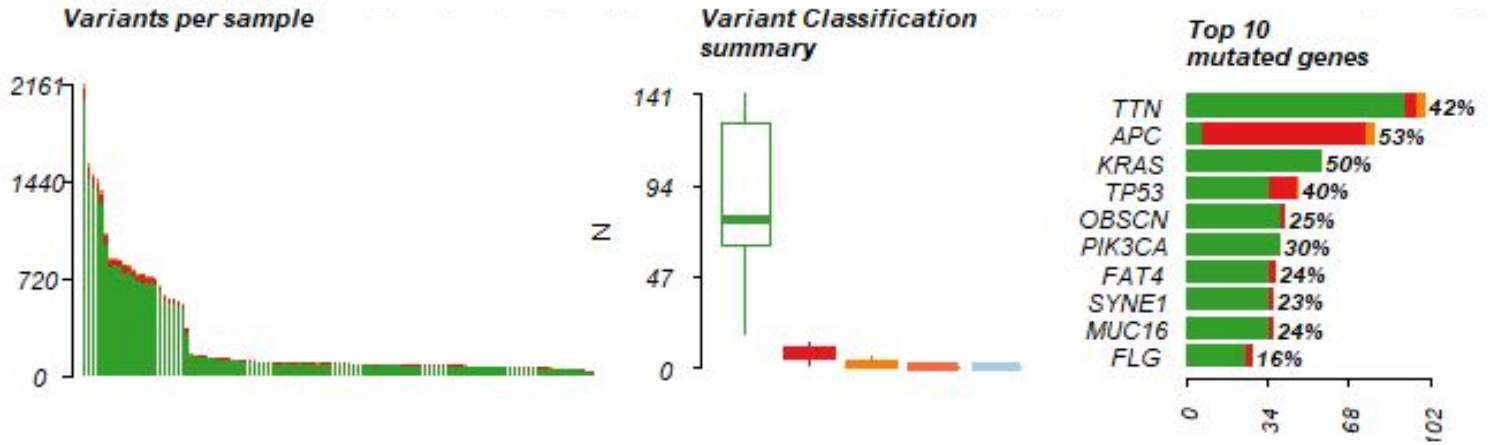
Aetiology: defective DNA mismatch repair [cosine-similarity: 0.74]

Demographics of patients SKCM

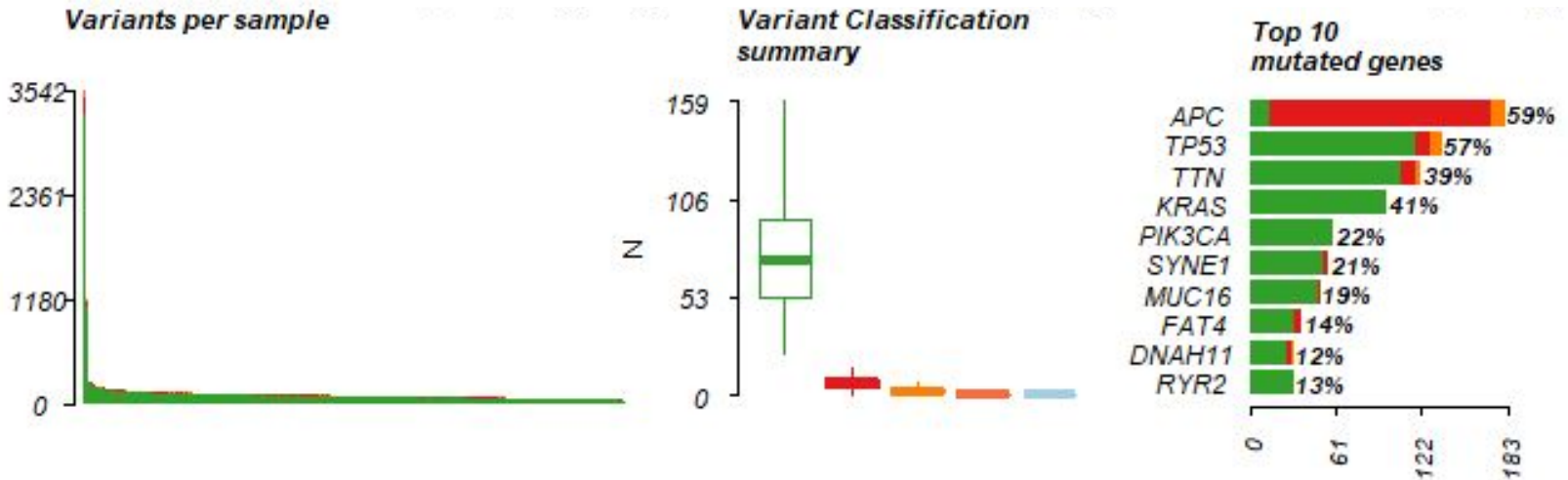


Total patients = 465
Total mutations = 374099
Min mutation per patient = 1
Max mutation per patient = 26230
Median = 13116
Average = 804.51
Standard Deviation = 1578.4

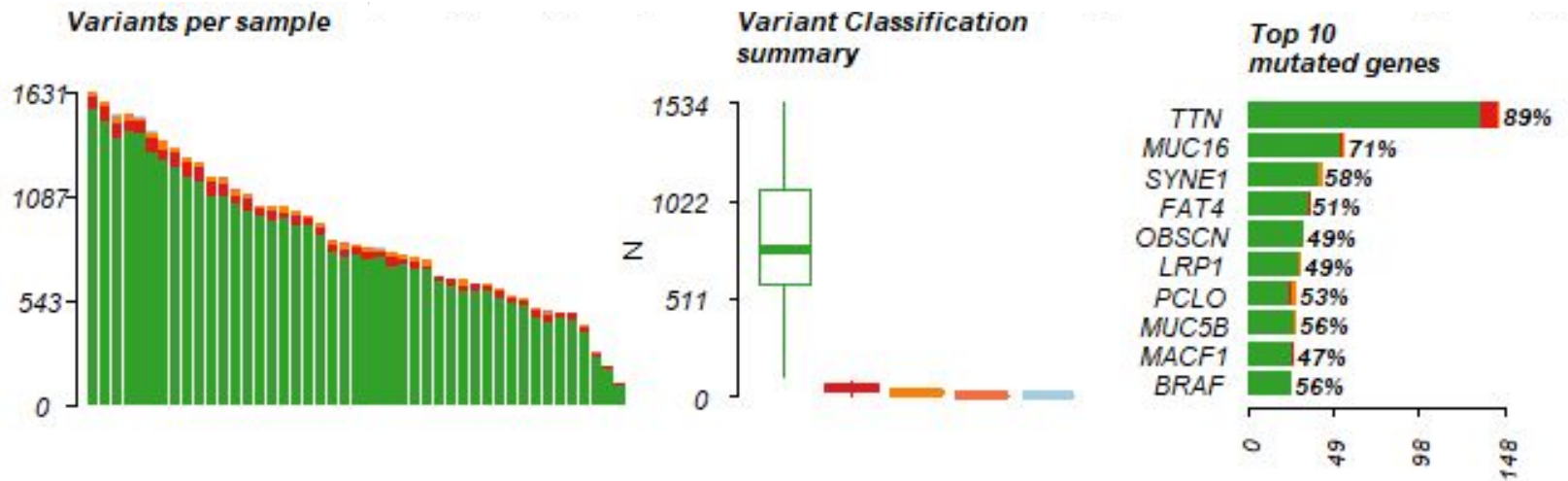
Analysis of COAD Signature 1 Patients



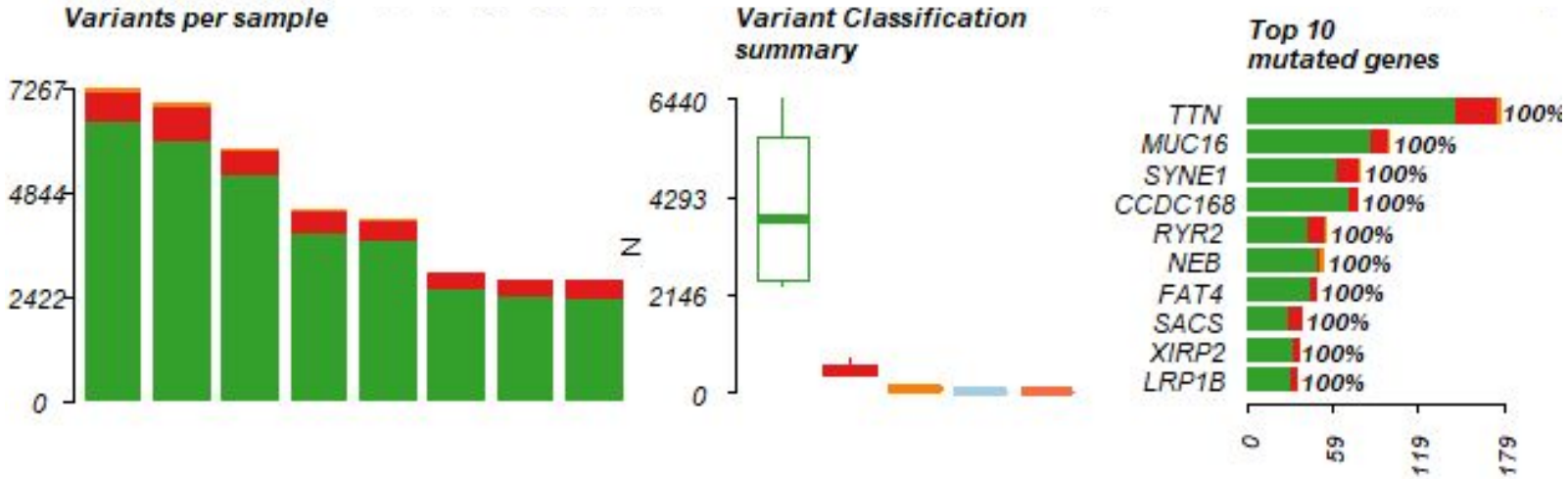
Analysis of COAD Signature 2 Patients



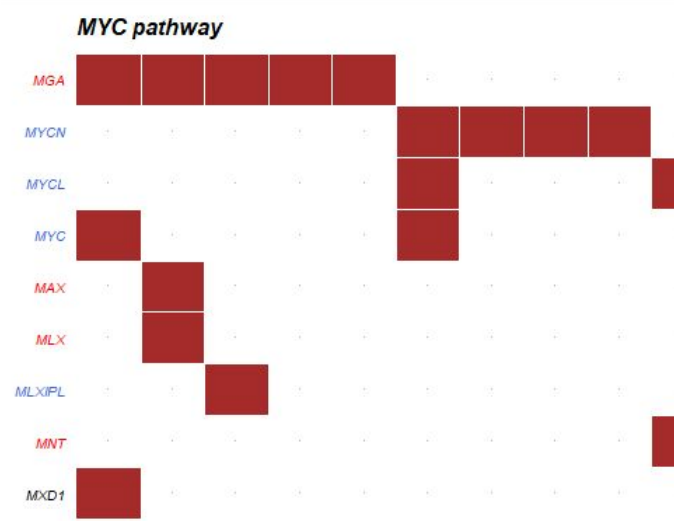
Analysis of COAD Signature 3 Patients

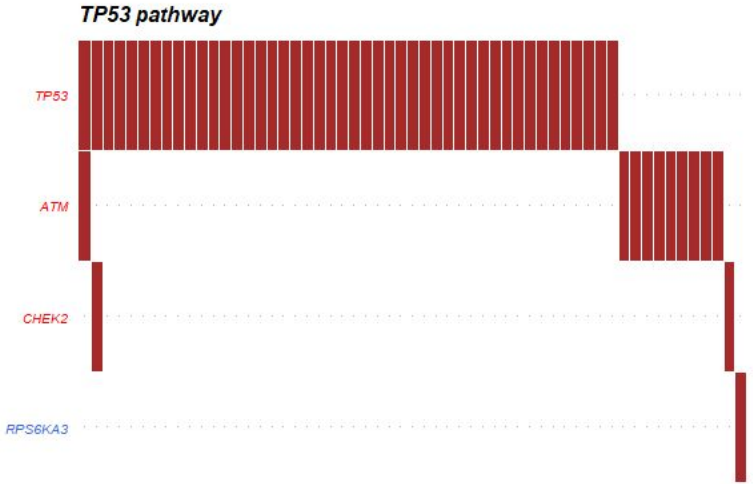


Analysis of COAD Signature 3 Patients



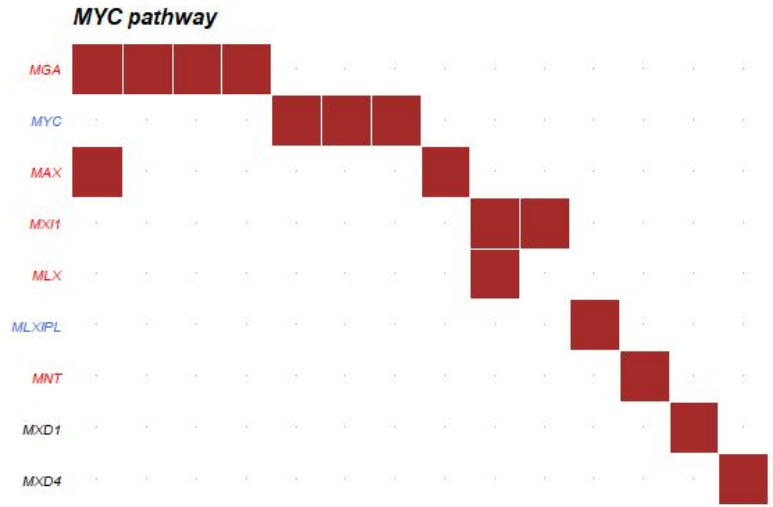
Analysis of COAD Signature 1 Patients: Oncogenic Pathways

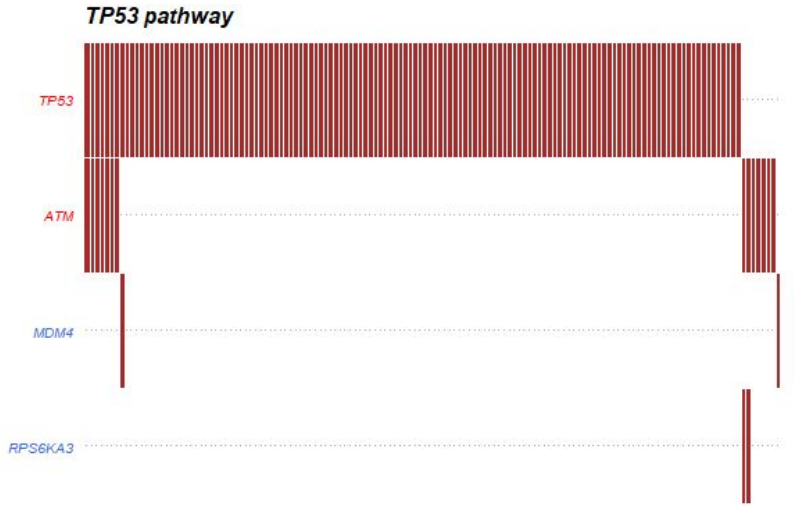




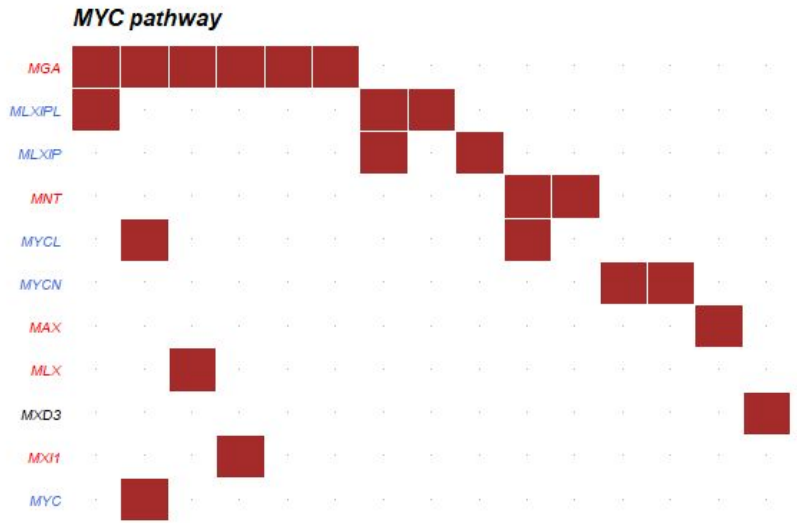
Columns represent Individual Samples. Blue names are Oncogenes and Red are Tumor suppressor genes in pathway. Brown boxes represent mutations

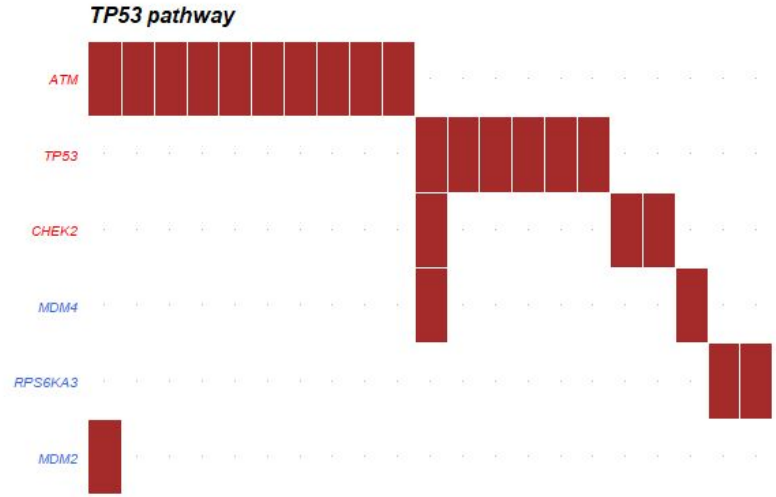
Analysis of COAD Signature 2 Patients: Oncogenic Pathways



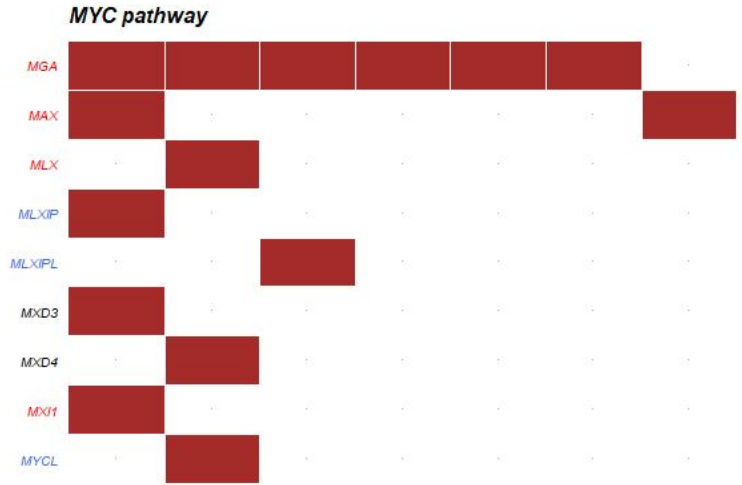


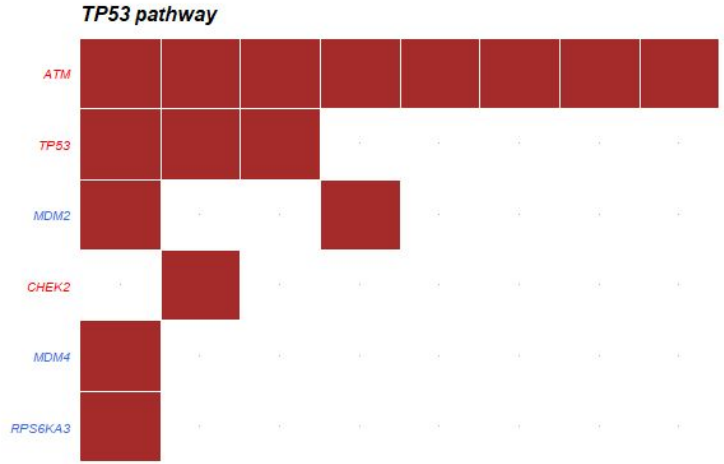
Analysis of COAD Signature 3 Patients: Oncogenic Pathways





Analysis of COAD Signature 4 Patients: Oncogenic Pathways





Relevance and Significance

- These methods can be applied to patient sequenced DNA
- Can check for specific mutation signatures that may or may not put the patient at risk for developing tumors
- We can use this to develop a patient risk analysis and pre-diagnosis

Future Work - Clustering (Methodology)

Using the high mutation burden cases with the 96 trinucleotide mutation types

- Use Euclidean distance to group most similar vectors (96 mutation types per sample)
- Determine the mutation burden of each clustered group
- Compare signatures study to find differences to signatures of the whole patient pool
- Look for differences by excluding non-synonymous mutations and removing common germline variants

References

1. Ludmil Alexandrov, Jaegil Kim, Nicholas Haradhvala, Mi Ni Huang and Anoud Boot. “The Repertoire of Mutational Signatures in Human Cancer”. *BioRxiv Magazine*. (March 12, 2018). https://www.biorxiv.org/content/biorxiv/early/2018/05/15/322859.full.pdf
2. Jennifer Ma, Jeremy Setton, Nancy Lee, Nadeem Riaz and Simon Powell. “The therapeutic significance of mutational signatures from DNA repair deficiency in cancer”. *Nature Communications.* Volume 9, Article number: 3292 (2018). https://www.nature.com/articles/s41467-018-05228-y
3. “Cosmic V88”. Catalog on Somatic Mutations In Cancer, last modified March 19, 2019. https://cancer.sanger.ac.uk/cosmic
4. “Signatures of Mutational Processes in Human Cancer”. Catalog on Somatic Mutations In Cancer, last modified March 19, 2019. https://cancer.sanger.ac.uk/cosmic
5. Young Kwang Chae, Jonathan F. Anker, Michael S. Oh, Preeti Bais, Sandeep Namburi and Sarita Agte. “Mutations in DNA repair genes are associated with increased neoantigen burden and a distinct immunophenotype in lung squamous cell carcinoma”. *Scientific Reports 9*, Article number: 3235 (2019): https://www.nature.com/articles/s41598-019-39594-4
6. Rodrigo Drummond, Renan Valieris, Rafael Rosales and Israel Tojal da Silva. “signeR”. BioConductor.org. https://bioconductor.org/packages/release/bioc/vignettes/signeR/inst/doc/signeR-vignette.html
7. Julian Gehring. “Somatic Signatures”. BioConductor.org, last modified 2015. http://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html
8. N. J. Haradhvala, J. Kim, Y. E. Maruvka, P. Polak, D. Rosebrock, D. Livitz, J. M. Hess and I. Leshchiner, A. Kamburov. “Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair”. *Nature Communications.* Volume 9, Article number: 1746 (2018). https://www.nature.com/articles/s41467-018-04002-4
9. Díaz-Gay, M., Vila-Casadesús, M., Franch-Expósito, S., Hernández-Illán, E., Lozano, J. J., & Castellví-Bel, S. (2018). Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics, 19*(1), 224. doi:10.1186/s12859-018-2234-y
10. Huang, P. J., Chiu, L. Y., Lee, C. C., Yeh, Y. M., Huang, K. Y., Chiu, C. H., & Tang, P. (2018). mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Res, 46*(D1), D964-D970. doi:10.1093/nar/gkx1133