

**ECES 650 – Statistical Analysis of Genomics**

**Spring 2018-2019**

**Dr. Gail Rosen**

**TITLE**: Mutation Signature Final Project Report

**AUTHOR**: Waleed Iqbal

Konur Bayrak  
 Blas Rodriguez

**Table of Contents**

|  |  |
| --- | --- |
|  | **Page #** |
| **Abstract** | **3** |
| **Literature Review** | **4** |
| **Materials and Methods** | **13** |
| **Results** | **18** |
| **Discussion** | **68** |
| **References** | **70** |

**Abstract**

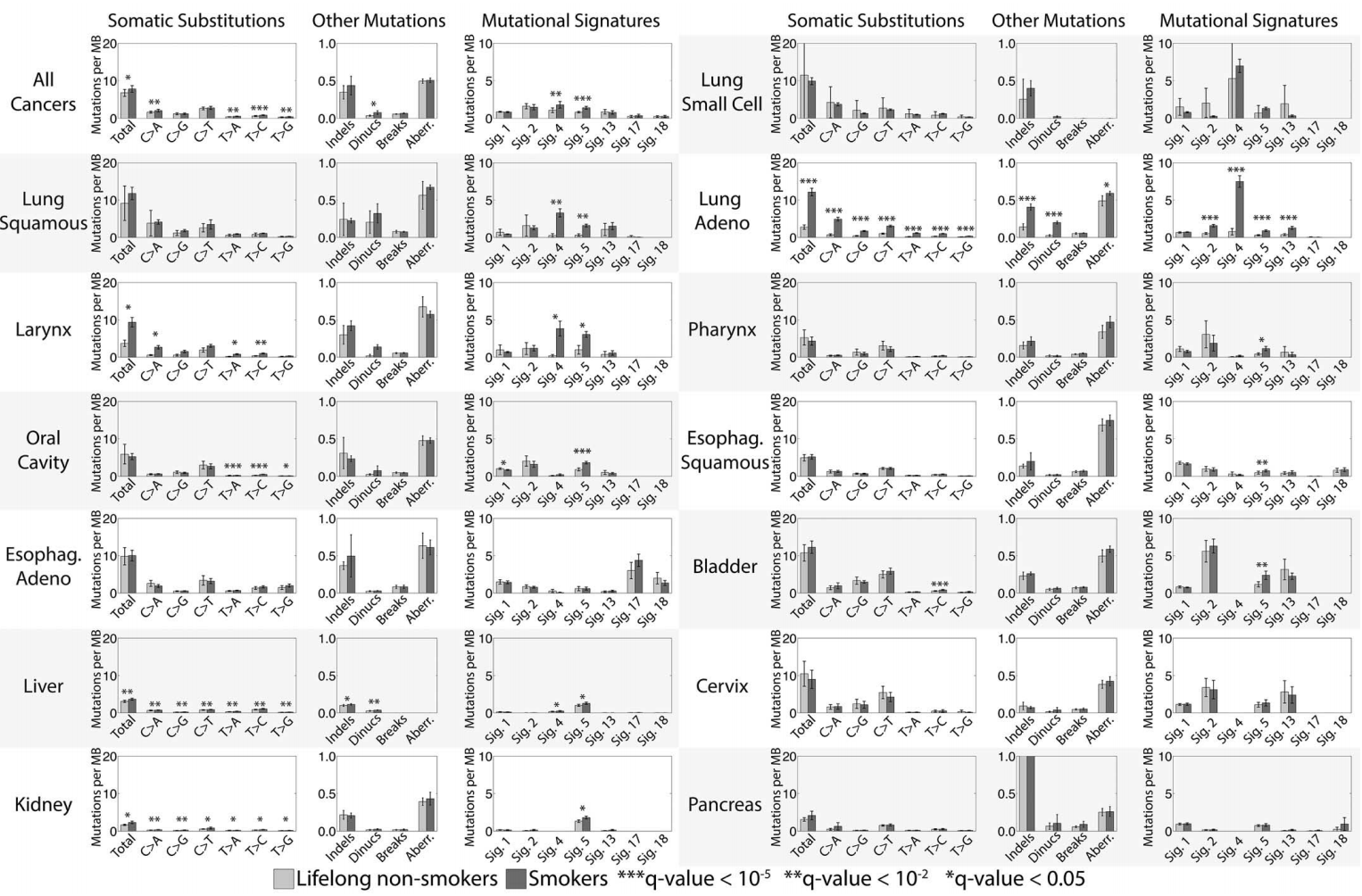
Genetic alterations are one of the main underlying mechanisms driving cancer development and progression. Differences among cancer subtypes and phenotypes is in a large part due to the differences in their genetic composition. The underlying mutations that are correlated with such differences is being systematically researched in cancer biology. An emerging idea has been that a mutational signature of a cancer would take into account the various different underlying processes responsible for it current phenotype or characteristic. Different subtypes could have different signatures that could in theory be correlated with a subtype differences such as treatment response, survival or cancer progression. Defects in certain genes such as those belonging to DNA replication and repair (*POLE*, Mis-match repair genes) have been correlated to distinct signatures and despite the discovery of new signatures and development of better analytical tools many of the established signatures are of unknown aetiology. We utilized two different mutation signature tools to construct the overall signatures present in 4 different cancer types lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM) and colorectal adenocarcinoma and observed the contribution of different established signatures mostly agreed upon by both methods. There were some clear differences between patients among the same cancers evident in differing contribution scores to decomposed signatures of different etiologies. Next we clustered our patients into groups based on similarity of trinucleotide mutations, and found novel signatures not evident in decomposed signatures from the combined cancer set in all 4 cancer types; surprisingly most of these were distinct signatures matching established signatures of unknown aetiology. This may be due to underlying differences in mutations not yet correlated with specific signatures or mechanisms more complex and a combination of various mechanisms not yet fully understood.

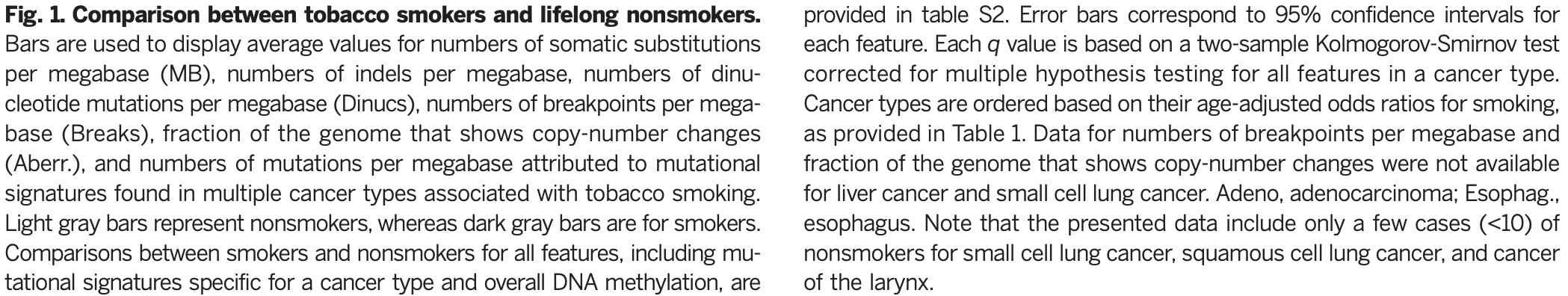
**Literature Review:**

At its core, cancer arises from the accumulation of somatic mutations in the DNA genetic code. These mutations lead to incorrect RNA translation and subsequently incorrect protein assembly. These faulty proteins can oftentimes have drastic effects on an extensive range of pathways involved in cellular homeostasis and metabolism. Until recently, our knowledge of the mutational processes that give rise to these somatic mutations had been very limited. However, computational analysis of pan-cancer patient data has allowed researchers to identify signatures of mutational processes thought to be responsible for the pattern of mutations in any given cancer type. Advances in sequencing technologies and the development of novel mathematical approaches have allowed deciphering the patterns of somatic mutations caused by different mutational processes. This allows researchers to analyze the mutations for patterns and signatures thereby allowing further classification of molecular subtypes within cancers. The identification of certain cancer pathways from patient genomic sequencing would also allow better targeted therapeutics and have a substantial impact on healthcare.

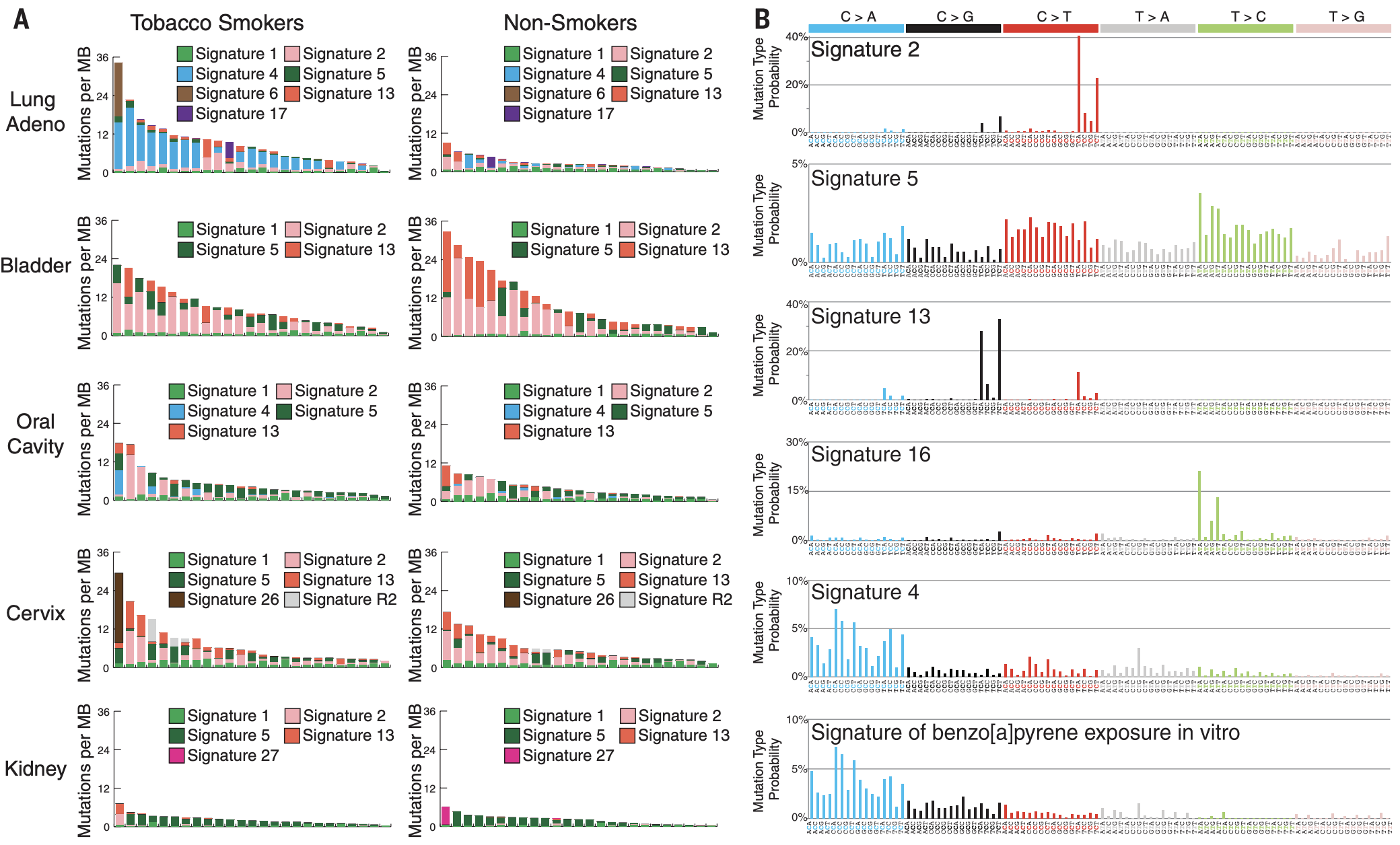
The first and seminal mutation signatures observed in cancer were compiled in COSMIC (3) and are based on six single nucleotide substitutions, based on the alteration of the pyrimidine in watson-crick pairing: C>A, C>G, C>T, T>A, T>C, and T>G. In addition, the adjacent context of bases present at the 5' and the 3’ position is also accounted, which leads to 96 trinucleotide changes being possible; For example for an T>G change, you would look at neighboring bases and the trinucleotide might be A**T**G>A**G**G or it can be a C**T**C>C**G**C. Current established mutation signatures have been compiled by combining large scale tumor studies. The COSMIC mutation signatures are based on 10,952 exomes and 1,048 whole-genomes across 40 distinct types of human cancer from TCGA and ICGC data as well peer reviewed somatic study data (4). These serve as reference signatures to compare experimental results with.

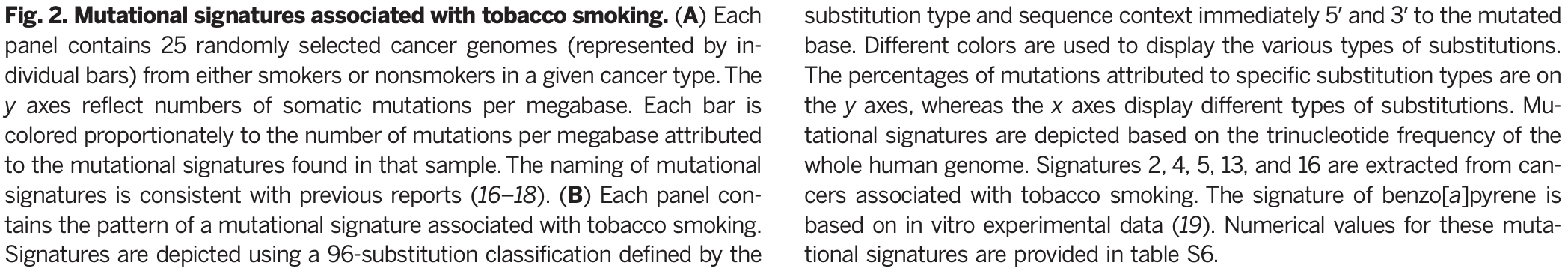
Several others have already paved the way for this new area of computational genomic research. One example of mutation signature discovery is the early finding that cellular exposure to UV irradiation has a tendency to cause C->T and CC->TT DNA mutations at dipyrimidine sites. This was subsequently the first defined mutation signature, as it was the first characterized pattern in DNA changes occurring due to external carcinogenic stimulus. Another notable study with far-reaching implications was done on a much more man-made external carcinogenic stimulus: tobacco smoking. With the foreknowledge that tobacco smoking increases the risk of 17 classes of cancer, *Alexandrov et al* [11] analyzed the somatic mutations and DNA methylation in 5243 cancer genome sequences (4633 exomes and 610 whole genomes) of types associated with smoking tobacco. The data was curated from three sources: (i) The Cancer Genome Atlas (TCGA) data portal, (ii) the International Cancer Genome Consortium (ICGC) data portal, and (iii) data previously generated for 17 articles published in peer-reviewed journals. Of the samples studied, 2490 were from tobacco smokers and 1063 from people who have never smoked. This enabled them to look at differences in mutational signatures of smokers vs non-smokers in lung, larynx, pharynx, oral cavity, esophageal, bladder, liver, cervical, kidney, and pancreatic cancers. They chose to look at regions that would have the most impact from smoking and also those that would have less of an impact to compare the two. First they compared total numbers of single base substitutions (C>A, C>G, C>T, T>A, T>C, & T>G), small insertions and deletions (indels), and genomic rearrangements. The total number of base substitutions was higher in smokers compared with nonsmokers for all cancer types together and, for individual cancer types, in lung adenocarcinoma, larynx, liver, and kidney cancers (Alexandrov figure 1). Total numbers of indels were higher in smokers compared with nonsmokers in lung adenocarcinoma and liver cancer (Alexandrov figure1). They then extracted mutation signatures, estimated the contributions of each signature to each cancer, and compared the numbers of mutations attributable to each signature in smokers and non-smokers. Mutation signatures were found using the conventional method of accounting for the adjacent bases present at the 5' and the 3’ position, which leads to 96 trinucleotide changes being possible. From these, signatures can be determined by comparing to the defined cancer signatures on the Catalogue of Somatic Mutations in Cancer (COSMIC) site. Increases in smokers compared with nonsmokers were seen for signatures 2, 4, 5, 13, and 16 (Alexandrov figure 1 & 2). They found that signature 4 was most enhanced in samples from cancer areas that have direct contact with cigarette smoke. Signature 4 is very similar to the mutational signature induced in vitro by exposing cells to benzo[a]pyrene (cosine similarity = 0.94). Which is why the signature induced was included in Alexandrov figure 2. Signatures 2 and 13 are characterized by C>T and C>G mutations, respectively, at TpC dinucleotides and have been attributed to overactive DNA editing by APOBEC deaminases. Signature 5 is characterized by mutations distributed across all 96 subtypes of base substitution, with a predominance of T>C and C>T mutations. Signature 16 is predominantly characterized by T>C mutations at ApT dinucleotides. It was found that not only did smokers and non-smokers have deviations in the signatures observed in their respective cancer types, but also that the mutation signatures differed across areas closest to direct smoke contact to those with the least direct smoke contact [evident in the data shown in figures 1 & 2]. This study was able to shed light on the genomic implications tobacco smoking may have on an individual. We also see how far reaching these implications can be depending on the organ or tissue of interest.





Alexandrov Figure 1 (ALexandrov, 2016) [11]

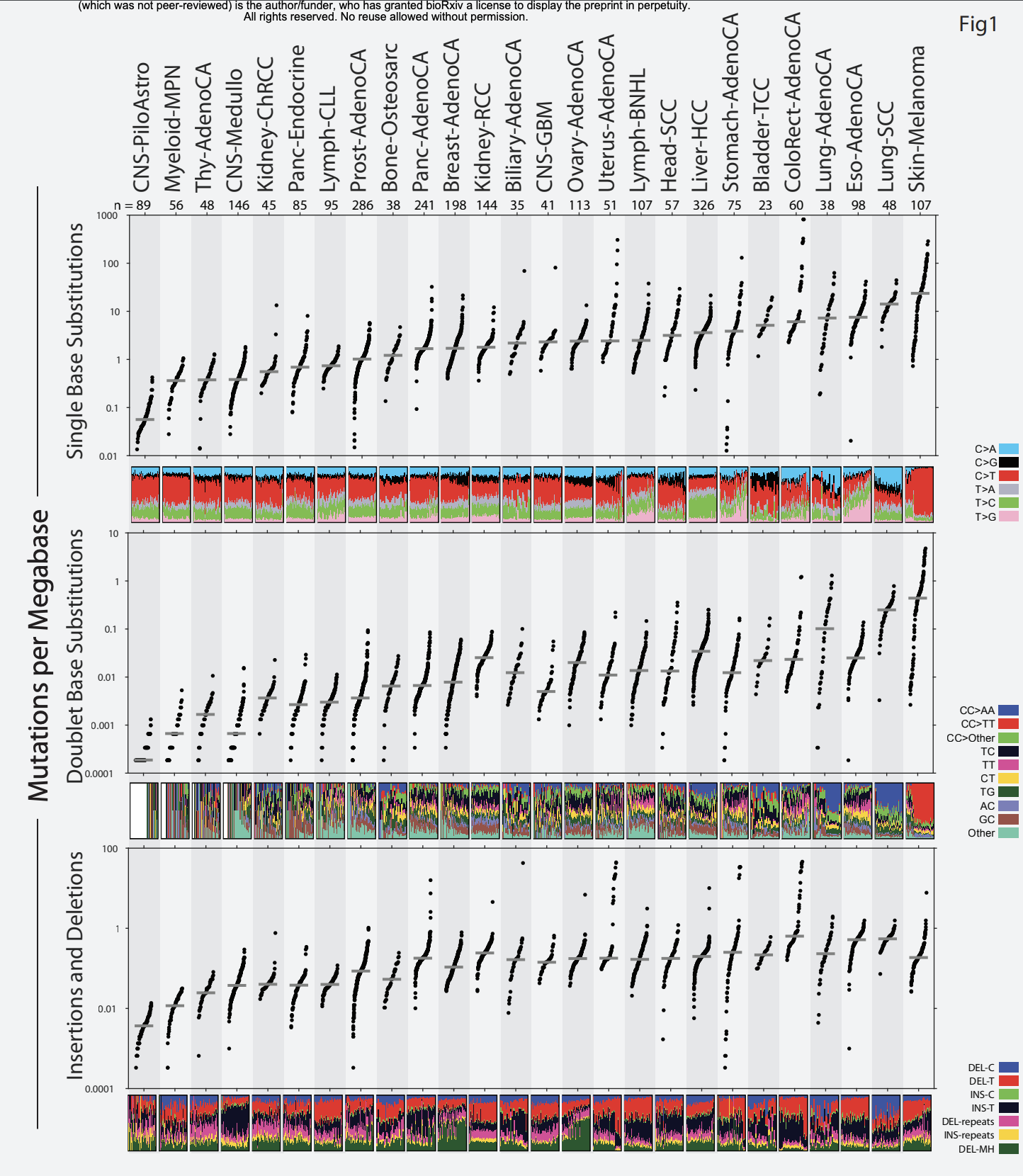




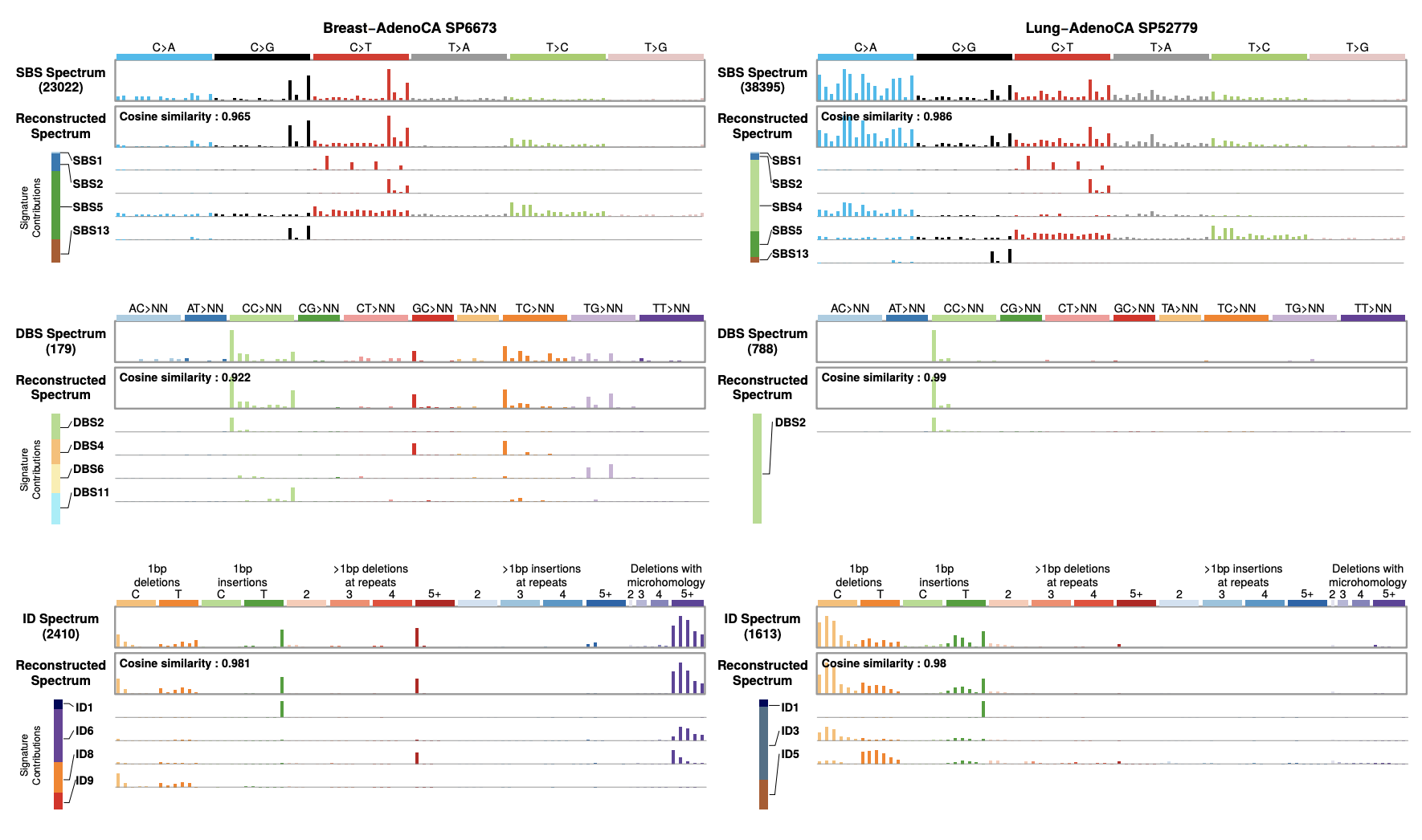
Alexandrov Figure 2 (Alexandrov, 2016)[11]

Additionally, a recent comprehensive study was done on a much larger and broader set of patient data by *Alexandrov et al* [1]. Instead of looking at two phenotypes (smoker vs. non-smoker), the aim of this study was to do a pan-cancer mutation signature analysis in human cancer in general. They identified 49 single base substitution, 11 doublet base substitution, four cluster base substitution, and 17 small insertion and deletion mutational signatures using 84,729,690 somatic mutations from 4,645 whole cancer genome and 19,184 exome sequences comprising most cancer types (Alexandrov figure 3 and 4). The substantial dataset size compared to previous analyses enabled discovery of new signatures, separation of overlapping signatures and decomposition of signatures into components that may represent associated, but distinct, DNA damage, repair and/or replication mechanisms. In order to achieve this, somatic mutational catalogues from 23,829 samples of most cancer types were studied and from these, 79,793,266 somatic single base substitutions, 814,191 doublet base substitutions and 4,122,233 small insertions and deletions (indels) were analysed for mutational signatures. That is 10 fold more mutations analyzed than any previous study. For single base substitutions, the primary classification comprised 96 classes for the 6 possible nucleotide changes C>A, C>G, C>T, T>A, T>C, & T>G and the flanking 5’ and 3’ bases. In some analyses, two flanking bases 5’ and 3’ to the mutated base were considered (generating 1,536 classes) or mutations within transcribed genome regions were selected and classified according to whether the pyrimidine of the mutated base pair fell on the transcribed or non-transcribed strand (192 classes). A classification was also derived for doublet base substitutions (78 classes).

Signature extraction was done using methods based on non-negative matrix factorization (NMF) on each cancer type individually and also on all cancer types together. Individual analyses were carried out for single base subs (SBS signatures), doublet base subs (DBS signatures), and indels (ID signatures) Signatures were extracted using two independently developed NMF-based methods: (i) SigProfiler, a more elaborate version of the framework used to generate the signatures displayed in COSMIC, and (ii) SignatureAnalyzer, based on a Bayesian variant of NMF used in several previous publications. NMF determines both the signature profiles and the contributions of each signature to each cancer genome as part of its factorization of the input matrix of mutation spectra. The two methods had many similarities but also showed some distinct differences. The number of SBS signatures found in low mutation burden tumors were similar across methods (31 by SigProfiler vs 35 by Signature Analyzer), however, the number of additional SBS signatures extracted from hyper-mutated samples displayed marked differences across methods (13 by SigProfiler vs 25 by SignatureAnalyzer). Conversely, the DBS and IS signatures were generally similar between methods.



Alexandrov Figure 3 (Alexandrov 2018)[1]



Alexandrov Figure 4 (Alexandrov 2018)[1]

Taking into consideration that both of the aforementioned research articles were published by the same lab, the techniques in both do not deviate that much. In terms of differences, the second *Alexandrov et al.* article essentially follows the same general school of thought as the initial, but expands the range or tumor samples analyzed and also implements further analyses. One such expansion of their approach was the inclusion of the doublet base substitution signatures and the indels signature. This is a great improvement because it allows for the identification of more signatures that would have otherwise been ignored completely. In addition to this improvement, the second article goes even further by using two distinct NMF based signature extraction methods. These methods proved to have differences as previously discussed (Alexandrov Figure 5, below), but they also provided a useful perspective on both the consistency and variability of signature extraction depending on the methodology used.

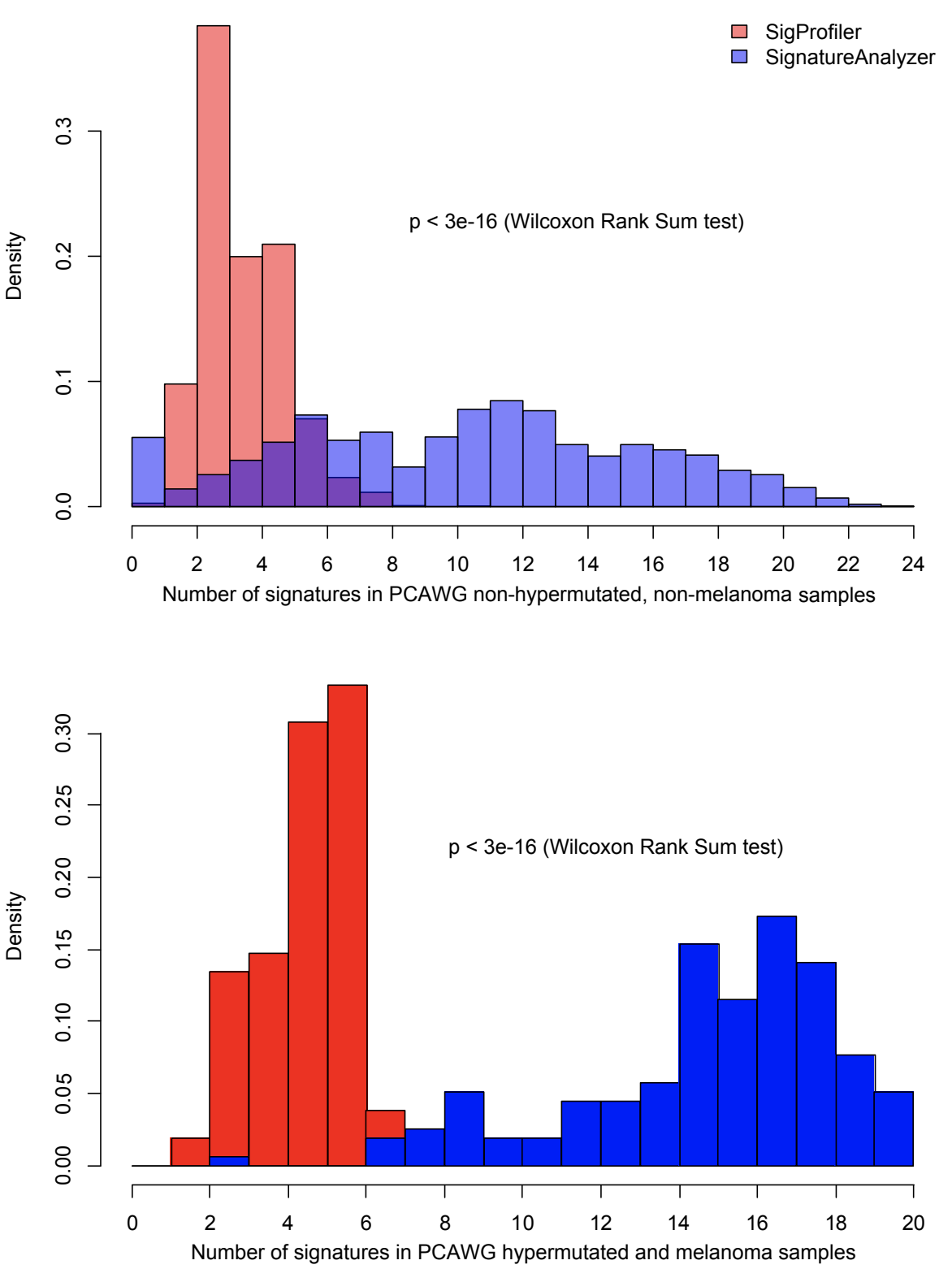


Figure 5 (Alexandrov 2018)[1]

**Materials and Methods:**

***Cancer Data selection and download***

Somatic mutation data based on human genome hg38 whole exome sequencing (WES) was downloaded from the cancer genome atlas database (<https://portal.gdc.cancer.gov/>) in MAF format on May 06 2019 using the TCGAbiolinks package (<https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html>). The cancer studies chosen were skin cutaneous melanoma (SKCM), lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD) and COAD (COAD). We chose SKCM because it has been reported to have some of the highest mutation burden across cancer types. The two lung cancers were chosen due to our interest in examining differences across the same organ with two cancer types arising in different tissues with squamous cell tissue having a greater exposure to the outside (exposure to air and the environment and hence also having a higher mutation burden. The MAF file data was read and analyzed using the maftools R package (<https://bioconductor.org/packages/release/bioc/html/maftools.html>)

***Analysis of mutation burden and summary characteristics***

Determining the mutation burden per each sample ID and sorting of patients based on mutation rate was done using MATLAB (<https://www.mathworks.com/help/matlab/>). Data obtained from R, this data was converted into .csv format (see R code) for import into MATLAB using the excel toolbox in MATLAB. From all the columns, the only data taken into consideration were the patients IDs as well and the mutation name on each row. A for loop was created to count how much mutation each patient had. The result was a matrix with each different patient and the number of mutations, using the second column of this matrix (count of mutations per patient) some analytical characteristics were obtained (total number of patients, total number of mutations, mean number of mutations, maximum, minimum, standard deviation and the median). Using the median, the patient IDS with greater number of mutations than the median were stored in different vectors. Then this vector was exported to a .csv file for further analysis using the packages in R (see subsetMaf function under the clustering section).

***Obtaining Signatures***

To obtain the overall signatures present in cancer and contribution of COSMIC signatures we used two separate and distinct methods. The first method was using the mSignatureDB (Huang et al 2019) web app. It uses the R mutSignatures package which performs non negative matrix factorization (NMF) decomposition to estimate and generate the most probable number of decomposed signatures within a cancer study and uses bootstrapped cosine similarity function that is part of R supraHex package (<https://www.bioconductor.org/packages/release/bioc/html/supraHex.html>) was used to calculate statistical significance of the similarity between observed mutation signatures in cohort being analyzed to published and established signatures such as in COSMIC and. The mSignatureDB allows the observation of signatures based on patient IDS but the web app only allowed us to look at one patient at a time. We used a similar pipeline and tools described in their methodology. We generated trinucleotide mutation counts using the maftools function trinucleotideMatrix (<https://rdrr.io/bioc/maftools/man/trinucleotideMatrix.html>), uses a reference genome to find5' and 3' bases flanking the mutation position listed in the MAF file. We used the hg38 human genome (bsgenome.hsapiens.ucsc.hg38) obtained from the bsgenome package (<https://www.bioconductor.org/packages/release/bioc/html/BSgenome.html>)It also allows filtering of synonymous variants and also variants arising from mitochondria termed chrM or other specific chromosome). The signatures were extracted for the overall cancer types using maftools function extractSignatures

(<https://rdrr.io/bioc/maftools/man/extractSignatures.html>). This function also utilizes NMF and thus requires the NMF package in R (<https://cran.r-project.org/web/packages/NMF/index.html>). This too; similar to the mSignatureDB can try to create a given number of signatures using NMF and choose the most probable ones (we used the number 10) using a cophenetic correlation coefficient. Also similar to the mSignatureDB it compartes the signatures constructed to validated signatures using cosine similarity. Lastly because NMF relies on positive values, due to a problem in your dataset or certain cells in your matrix you can get an error stating ‘non-conformable arrays’ error; To bypass this, a small pConstant value can be given when calling the function.

***Determining contribution of each patient to decomposed signatures***

Once we obtained the overall signatures present in the cancer type, a matrix with scores of contribution of each patient to each signature (this was part of the file generated from extractSignatures tool) was exported for analysis in MATLAB as a.csv file to determine what was the dominant signature being contributed to by each patient.. In the code, the matrix from R was split into sub matrices (one for each signature) in which the first was a vector with all the patient IDs, the second was a vector of the signatures and finally, the rest of the matrix had the contribution of each patient to each signature. Then, a for loop was created to go through each patient and compare the different values in their contribution to each signature. An empty matrix created to store the patient IDs corresponding to their dominant signatures. Using this information, a histogram was created as well as the data was exported a .csv file to be used for analysis in R (see subsetMaf function under the clustering section)

***Clustering of patients based on similarity of Trinucleotide mutation counts***

This code requires the trinucleotide mutation counts per patient ID obtained using R package trinucleotideMatrix. Clustering was done using Euclidean distance to cluster the patients using their mutation counts, and as a default we seperated into 4 clusters (but this can be changed), and generated a matrix, with the IDs of the patients belonging in each cluster, separated by a column (column 1 is for patients in cluster 1), as well as a dendrogram with the cluster numbers. The code first splits the patient ids into vector separate from the matrix with trinucleotide mutation counts. Using the matrix with all trinucleotide mutation counts and pdist() function, the Euclidean distance between each patient is determined. Then, using linkage() function the data obtained from pdist() is encoded into a tree containing hierarchical clusters of the rows. Which then this data is parsed analyzed using the clusters() function to define clusters from the agglomerative hierarchical cluster tree based on the linkage provided: in this code we defined the numbers of clusters we wanted as 4. In the final steps a dendrogram is created for the clustering and data of patient’s IDs in the different clusters is imported into R as a .csv file.

A new subset maf file can be created for different groups or clusters of patients of interest from the main maf file containing all patient ids and mutation information using the subsetMaf function (<https://www.rdocumentation.org/packages/maftools/versions/0.99.30/topics/subsetMaf>)

This function can create smaller subsetmaf file based on list of specific tumor barcodes (in this case lists of tumor IDs belonging to specific clusters, or even genes of interest (only patients with mutations in those genes of interest would be included). Additionally this function also allows for filtering of synonymous variants or querying based on specific variant type (for example missense mutations)

***Analysis of pathway enrichments per cluster***

Maftools allows for the generation of summary tables depicting variant classification counts, variation count variation across samples and top mutated genes present in a given maf file using the plotmafSummary function (<https://rdrr.io/bioc/maftools/man/plotmafSummary.html>)

Comparison of mutation loads in maf samples of different clusters compared to other TCGA cancer type studies was done using the tcgaCompare function (h[ttps://rdrr.io/bioc/maftools/man/tcgaCompare.html](https://rdrr.io/bioc/maftools/man/tcgaCompare.html)**).** For enrichment of biological pathways we used the PlotOncogenicPathways function (<https://rdrr.io/bioc/maftools/man/PlotOncogenicPathways.html>). Only a few pathways are available such as *WNT*, *NOTCH*, *TNF-kappaB*, *TP53*, *MYC and K-RAS.* We chose to focus on *TP53* and *MYC* pathways because they can have opposing roles in cancer; *MYC* is one of the most frequently amplified genes across many cancer types while *TP53* is one of the most highly mutated genes across most cancer types)

**Code Available onGithub**

**Link:** <https://github.com/mwi468/Mutation-Signature-Project>

**Results:**

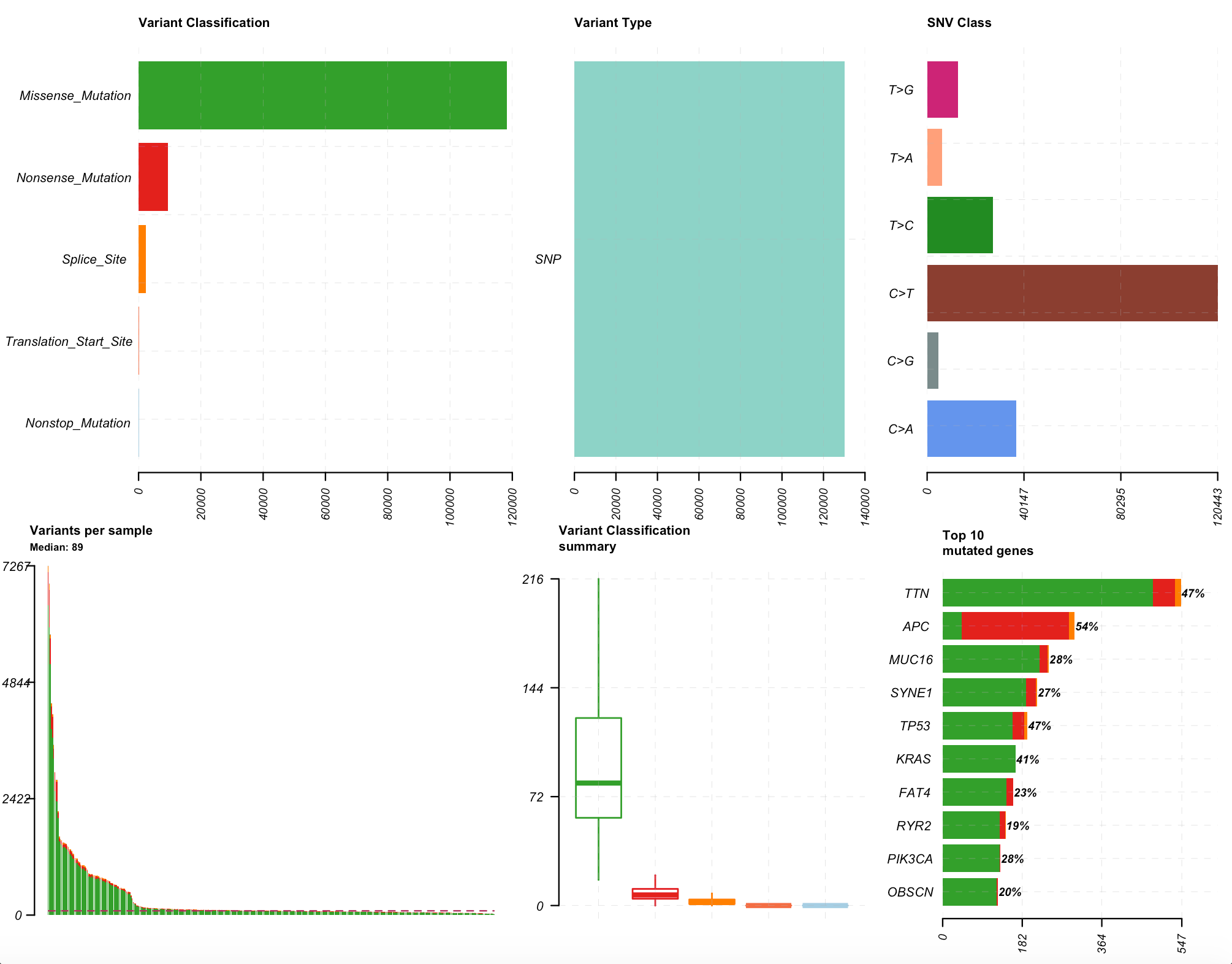
Of the 4 cancer studies. SKCM had the highest mutation burden as well as having the highest amount of variation across 465 samples. COAD had the second highest variation, while also having the lost median mutation count from 399 samples. Both lung cancer studies had much smaller variation across samples compared to COAD and SKCM while also having a median mutation count between the COAD and SKCM studies; However the median mutation counts were much higher in the LUSC study.

***Table 1. Distribution of all mutation counts across cancer types by Patient ID***

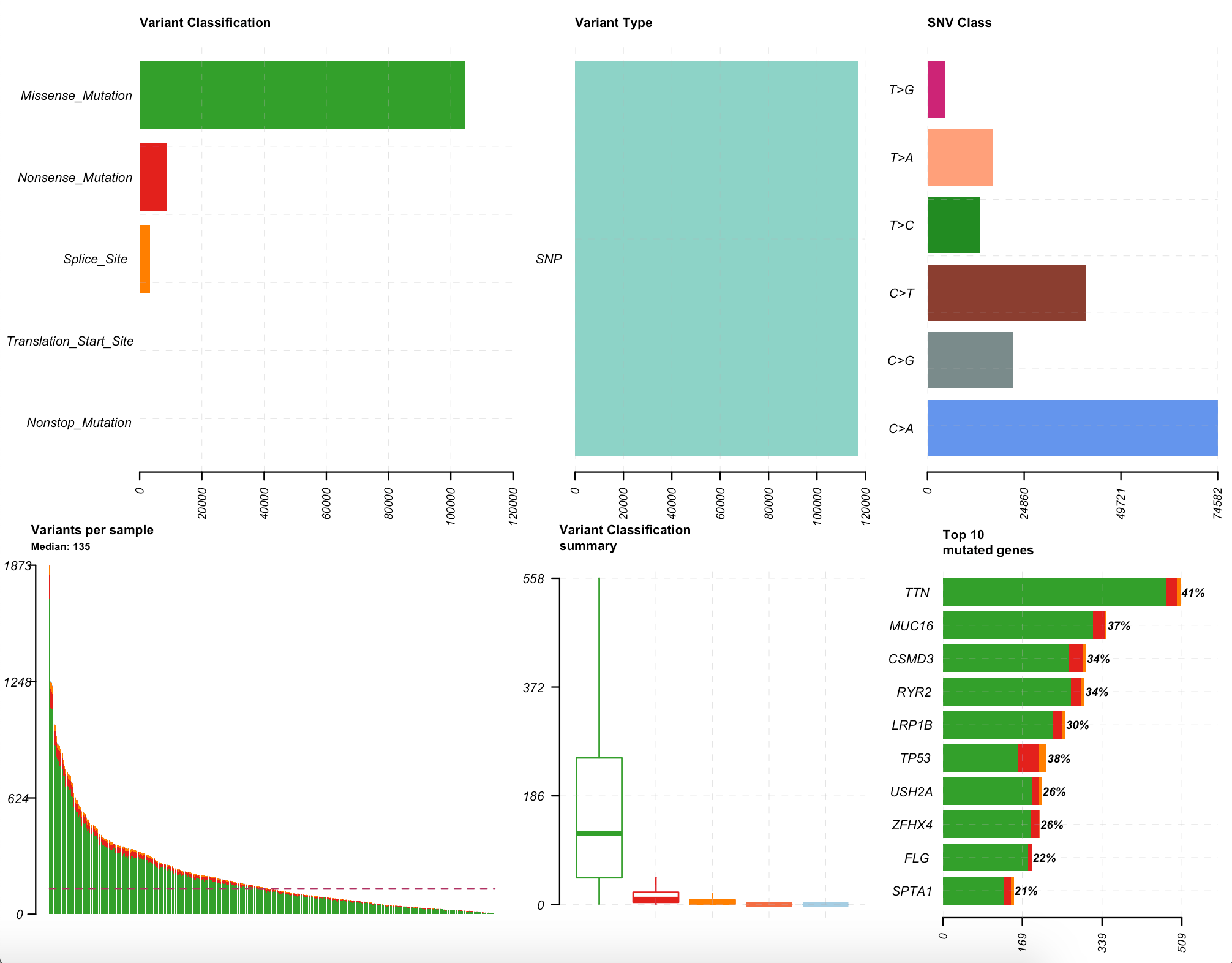
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Type of cancer** | **Amount of Samples** | **Amount of Mutations** | **Minimum** | **Maximum** | **Median** | **Mean** | **Standard Deviation** |
| **LUAD** | 563 | 172086 | 1 | 2823 | 197 | 305.65 | 342.16 |
| **LUSC** | 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 |

From the results, it is clear different tumor types have different amounts of mutations. However, in each of the 4 sampled tumors, there is a huge standard deviation. This implies high amounts of outliers throughout the data, which tell us that the average is not as reliable as we expected. As a result, the median to determine our cutoff value might be more valid than using the median for separating cases based on high or low mutation. Additionally this table has values for all mutations (synonymous mutations were not filtered) but they were, before the generation of the summary plots in the next figures. As expected we observed distinct differences in the type of mutations across all cancer types ( even between the two lung cancer studies) based on the overall SNV change as well as in top mutated genes.

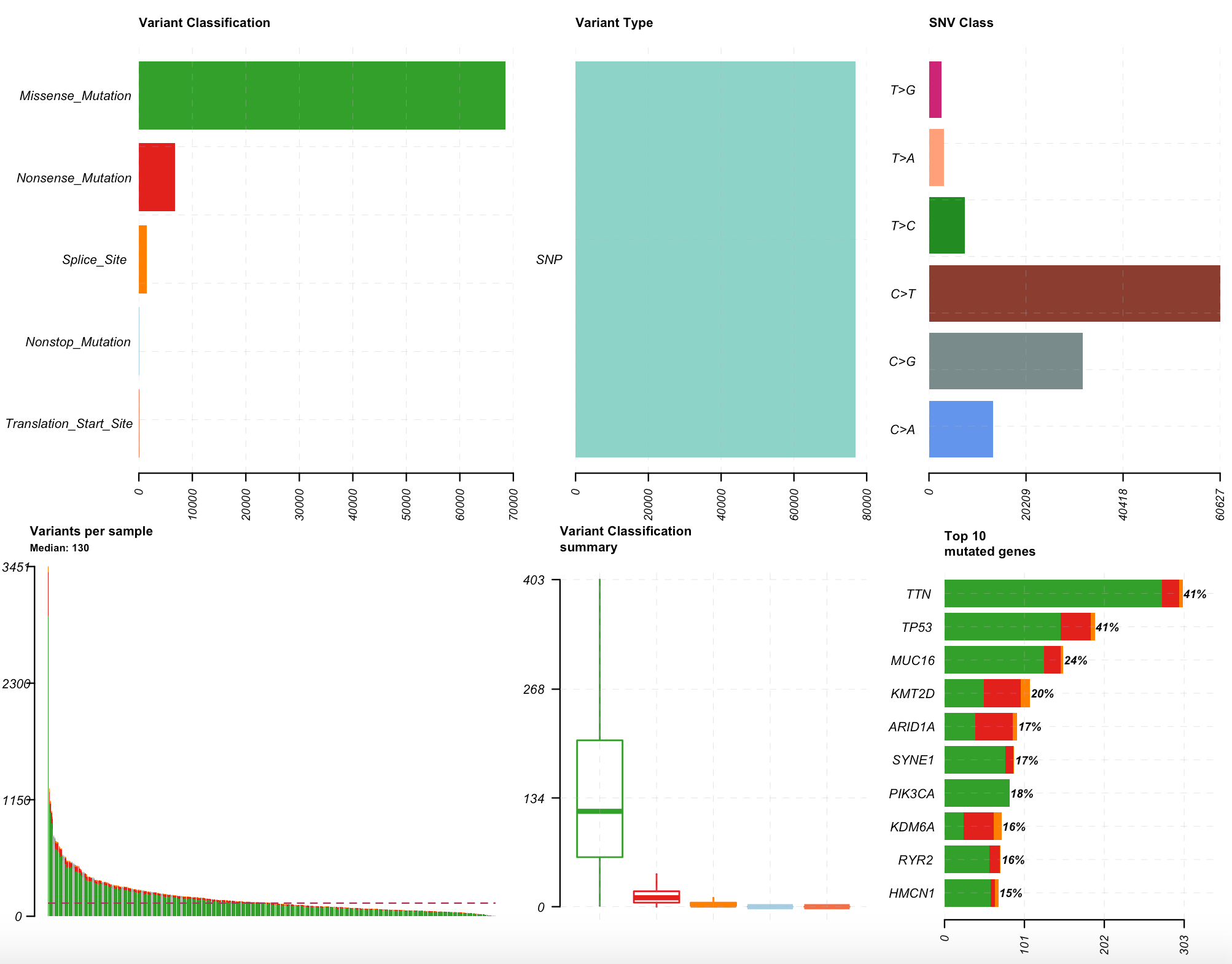
**Summary of Downloaded MAF files**

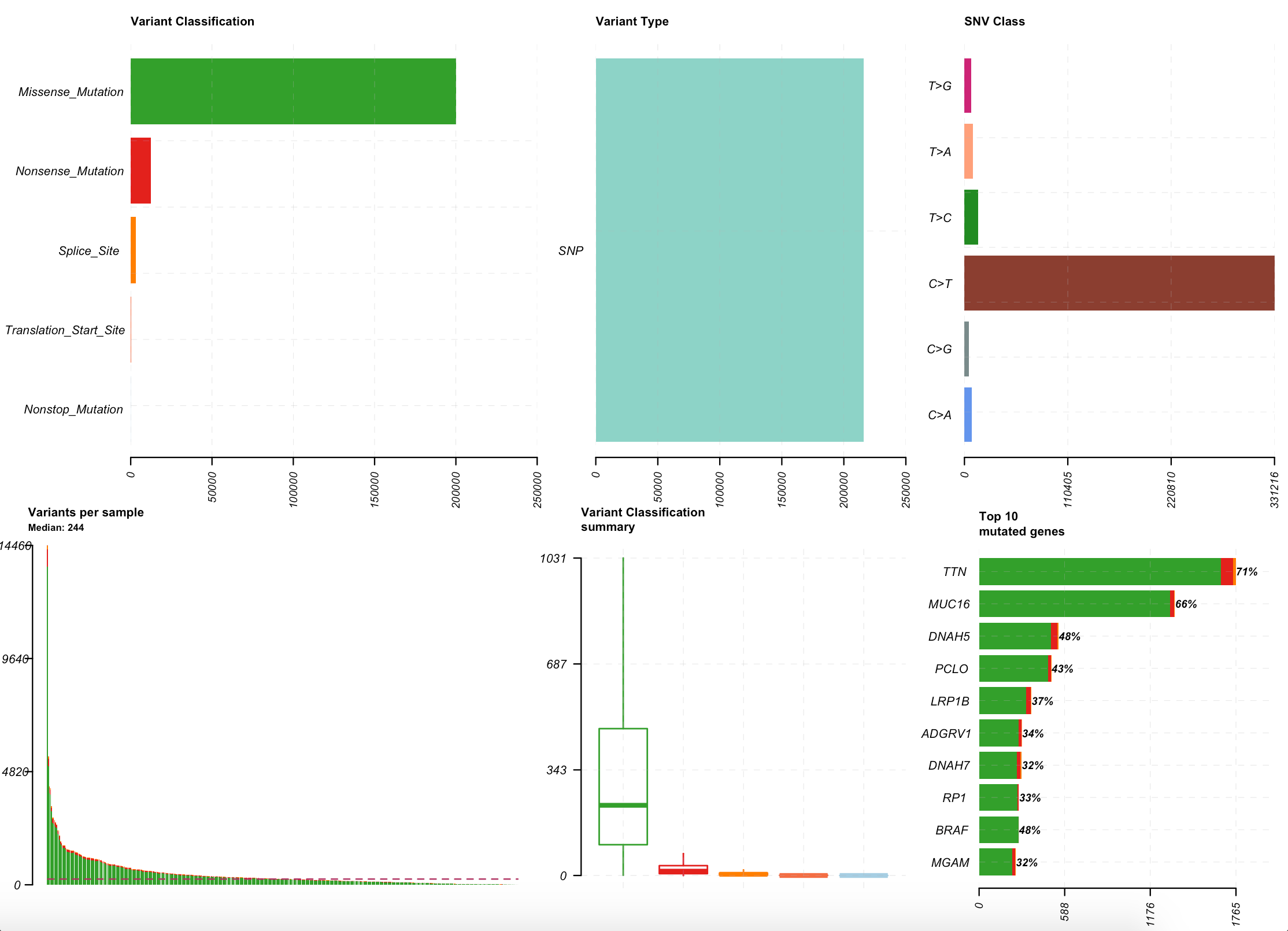
**

***Figure 2. Summary of COAD (Colon adenocarcinoma)*.** In COAD the median variants 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.



***Figure 3. Summary of LUAD (Lung adenocarcinoma)*.** In LUAD the median variants 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.

***Figure 4. Summary of LUSC (Lung squamous cell carcinoma)*.** In LUSC the median variants 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.

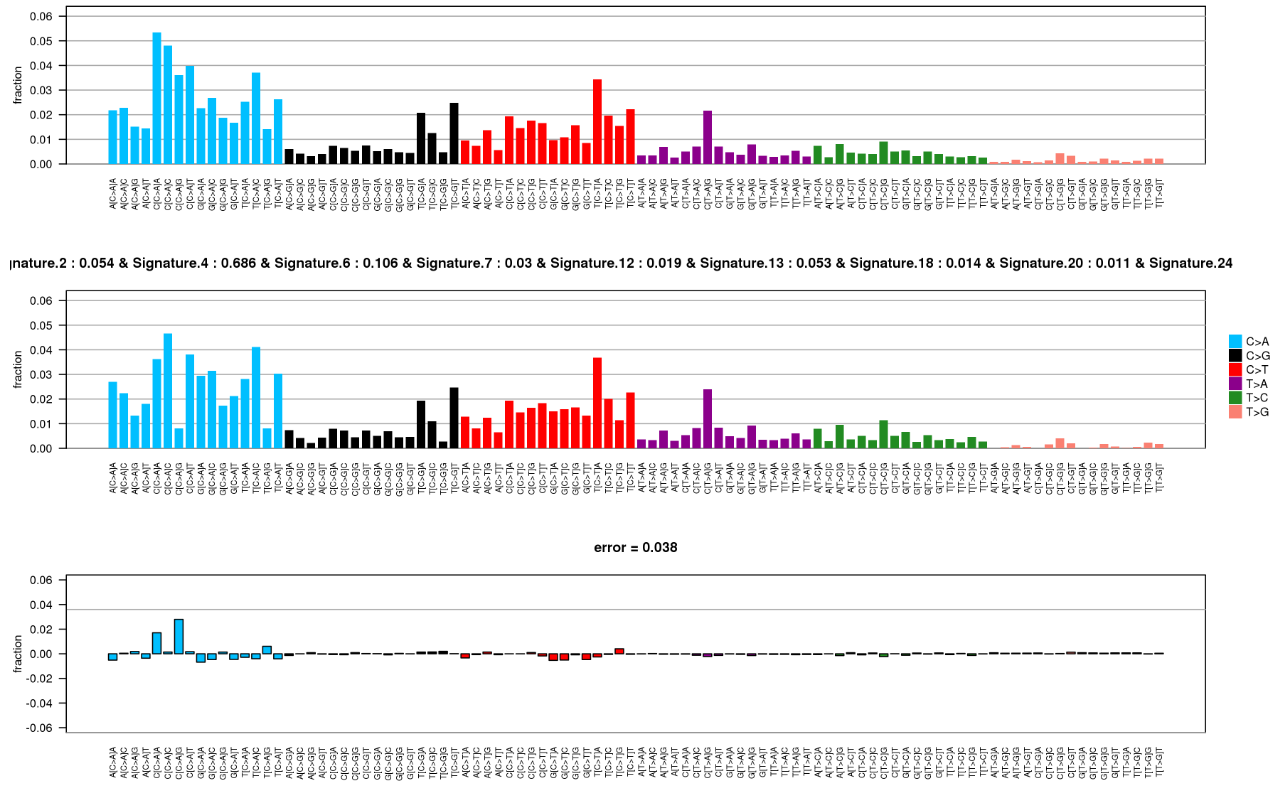


***Figure 5. Summary of SKCM (Skin Cutaneous Melanoma)*.** In SKCM the median variants per sample is 244 but it can be observed that there are some samples with very high number of variants ( uptil ~14466) which is lower than the max of COAD. The top 3 mutated genes are TTN, MUC16 and DNAH5; most of the mutations in all three genes are mostly missense mutations (green). Most single nucleotide variants (CNVs) are C>T changes.

***\***

**Overall Mutation Signatures by Cancer type**

Based on observation of the overall cancer signatures from mSignatureDB, as expected we found differences across all cancer types with some shared similarity among the lung cancer signatures. For LUAD the mutation signature was characterized by significant C>A changes followed by some significant C>T changes and it was reported that Signature 4 from COSMIC predominantly contributes the most to the overall signature (Figure 6). Similarly for LUSC we observed mutation signature characterized by significant C>A changes but there were also some more significant C>T changes and C>G changes (Figure 7). Additionally it was reported that Signature 4 from COSMIC contributes close to 50% (much less than in LUAD) to the overall signature and many other smaller signatures signatures contribute the other 50%. Signature 4 in COSMIC is associated with mutations in guanine due problems in transcription-coupled nucleotide excision repair as a result of tobacco smoking. For COAD, the mutation signature is characterized by significant C>T changes followed by some significant C>A changes and it was reported that Signature 1 and signature 6 from COSMIC contribute the most to the overall signature (Figure 8). Signature 1 is associated with enzymatic deamination of 5-methyl-cytosine to thymine causing G:T mismatches in double stranded DNA and signature 6 is associated with defective DNA mismatch repair due to microsatellite instability. In SKCM The mutation signature is characterized by significant C>T changes with few other mutations observed (figure 9). Additionally it was reported that Signature 7 from COSMIC contributes the most to the overall signature



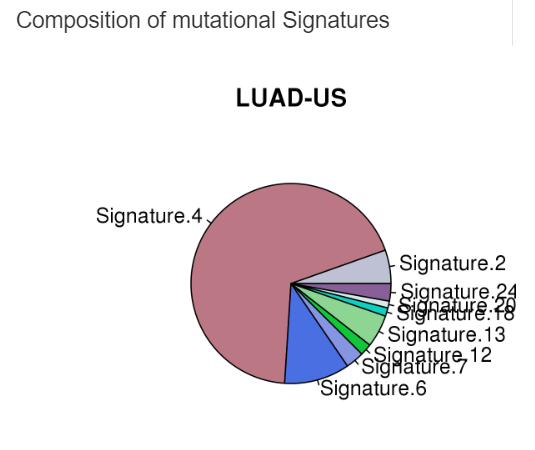
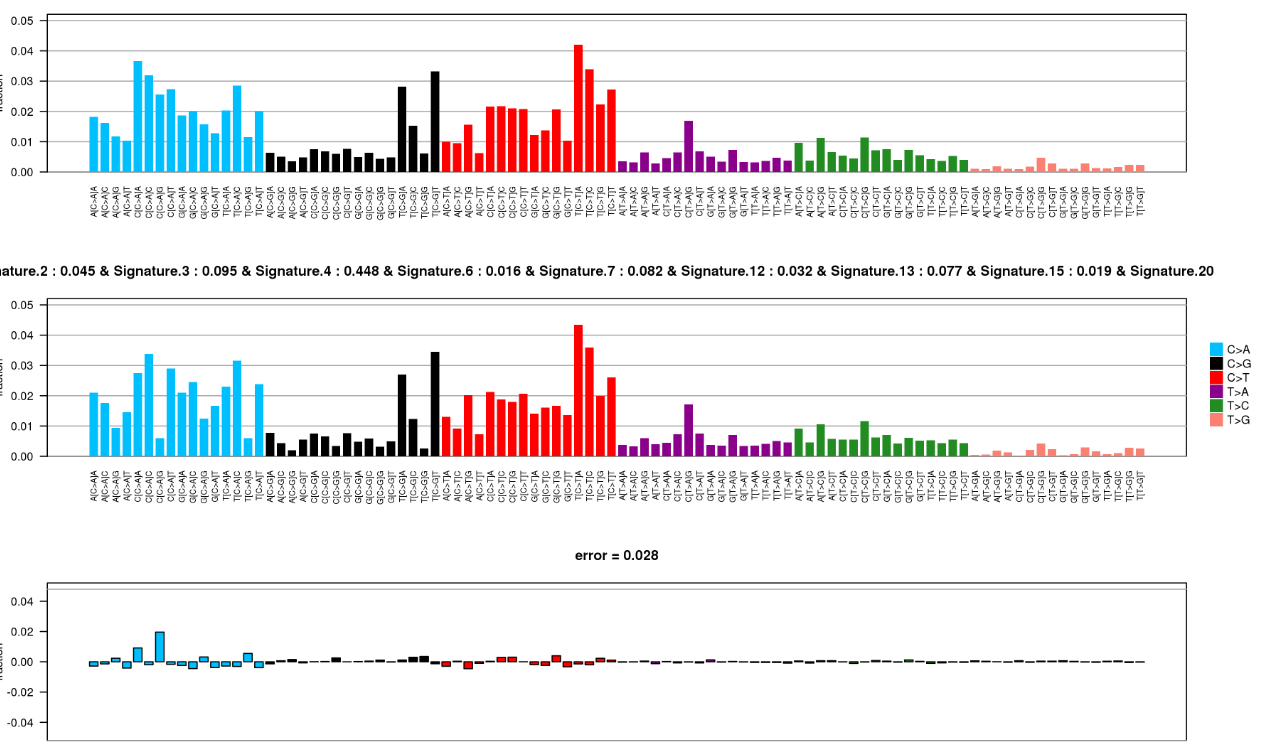


Figure 6 Mutation Signature of LUAD. Top: Overall mutation signature observed across the whole cancer. Bottom: the contribution of different COSMIC signatures to the overall signature observed



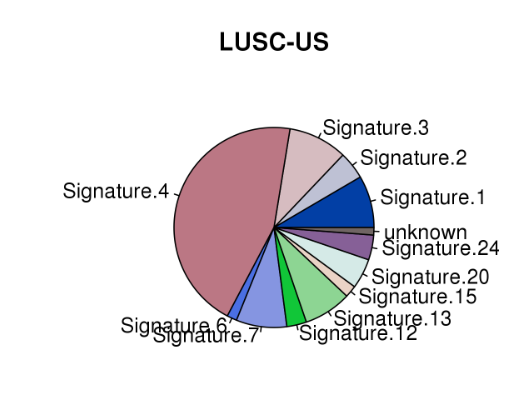
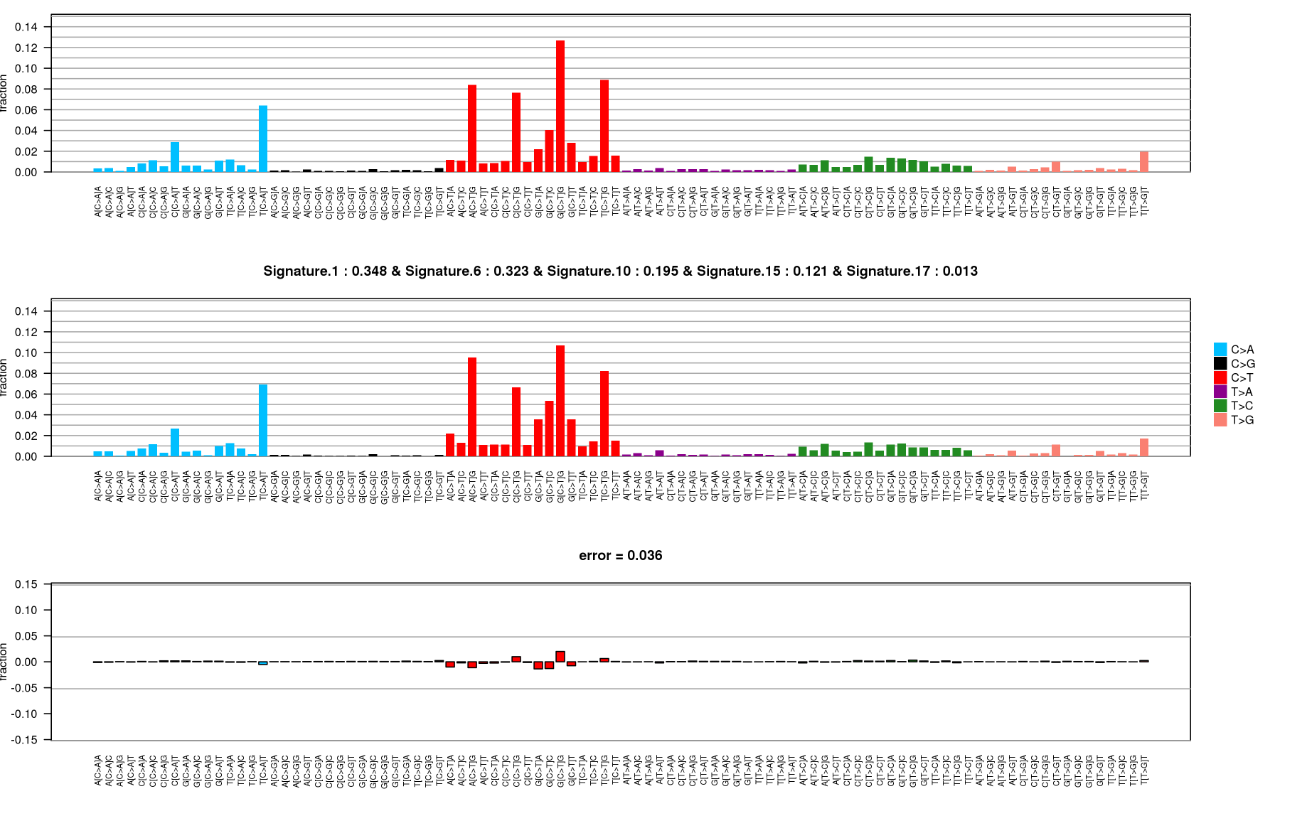


Figure 7 Mutation Signature of LUSC Top: Overall mutation signature observed across the whole cancer. Bottom: the contribution of different COSMIC signatures to the overall signature observed

****

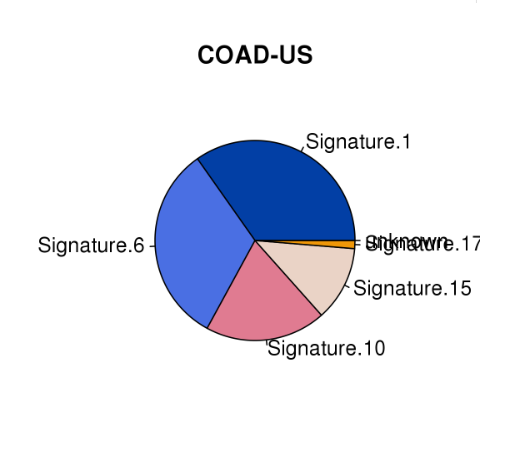
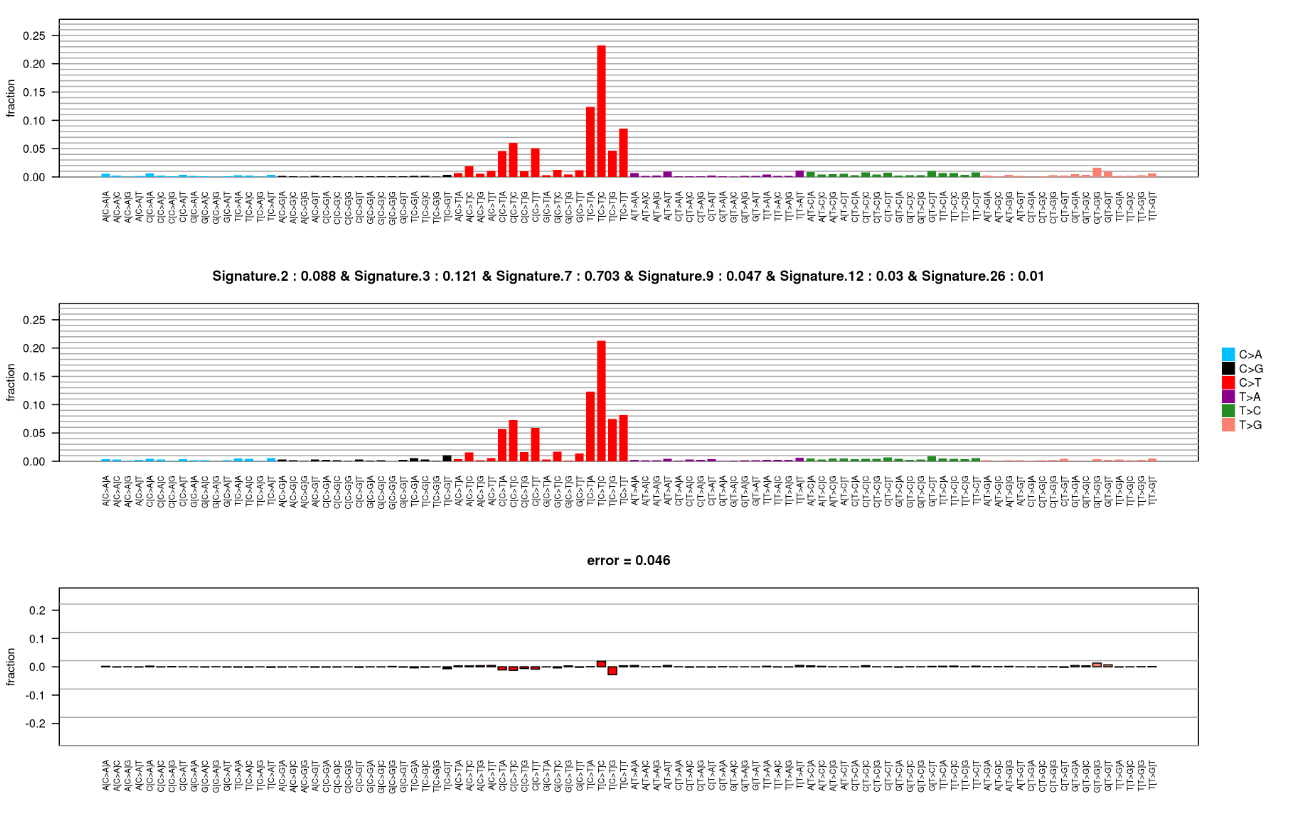


Figure 8 Mutation Signature of COAD

Top: Overall mutation signature observed across the whole cancer. Bottom: the contribution of different COSMIC signatures to the overall signature observed

****

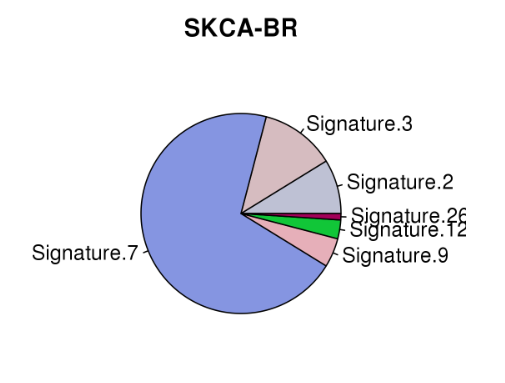


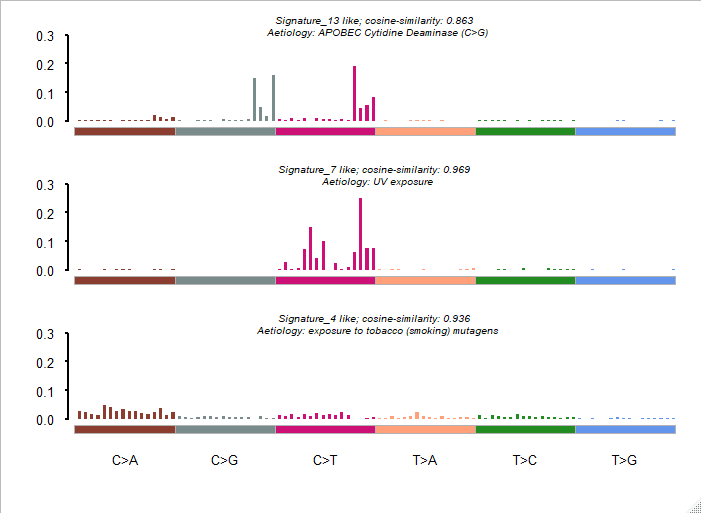
Figure 9 Mutation Signature of SKCM

Top: Overall mutation signature observed across the whole cancer. Bottom: the contribution of different COSMIC signatures to the overall signature observed

**Decomposed Mutation Signatures using MAF tools**

By using the extract signatures tool we were able to better visualize the decomposed signatures that may be contributing to the overall signature as well as contribution scores of patients to all the signatures found. For LUSC (Figure 10) similar to the previous overall signature found using the mSignatureDB method, the predominant signature in ~300 patients was signature 4 associated with tobacco smoking followed by signature 13 (U.V exposure) in ~200 patients and in close to 70 patients it was signature 7 (DNA damage due to APOBEC cytidine deaminase). For LUAD (figure 11) results were similar as well with signature 4 remaining the predominant signature. Interestingly close to 70 had signature 2 (also APOBEC cytidine deaminase) as the predominant type which is a different signature for the same aetiology observed in LUSC and interestingly a much smaller subset had signature 1 (5-methylcytosine deamination) not observed in LUAD. For COAD the most common predominant signature was signature 2 (aetiology unknown) for approximately 240 cases followed by signature 1 (5-methyl-cytosine deamination) in approximately 110 patients. The other signatures, signatures 6 (mismatch repair defect) and 10 (POLE) were observed predominant in less than 50 patients. This is also differs from mSignatureDB method which did not mention signature 30 at all. Interestingly for SKCM two different signatures were found both having 180-200 patients and both matching signature 7 (U.V exposure). This was the predominant signature shown using the mSignatureDB method but signature 6 predominant in close to 80 patients was missed by the previous method.

*LUSC*

****

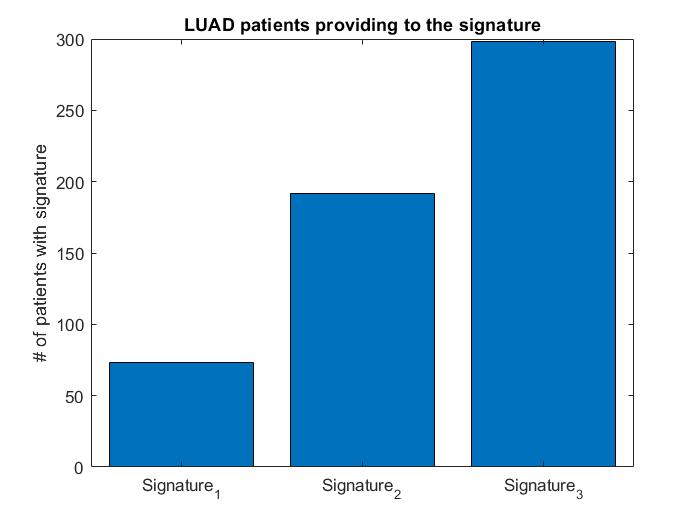
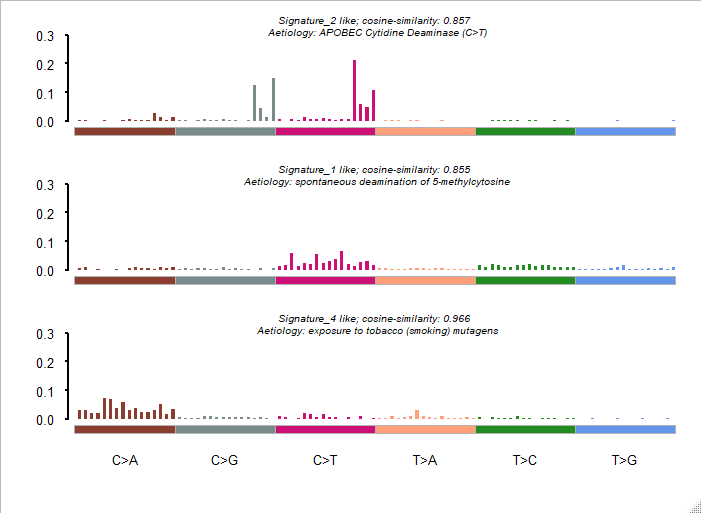


Figure 10 Decomposed LUSC signatures Top: Overall mutation signatures observed across the whole cancer. Bottom: Number cases based on their dominant signature

*LUAD*

****

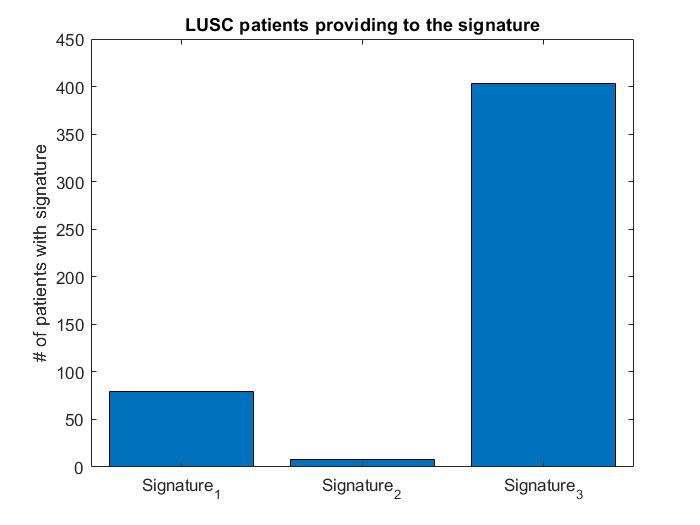
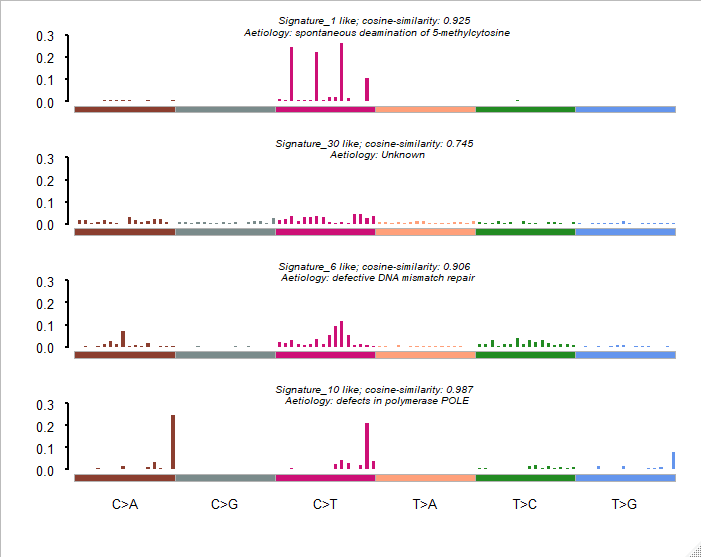


Figure 11 Decomposed LUAD signatures Top: Overall mutation signatures observed across the whole cancer. Bottom: Number cases based on their dominant signature

*COAD*

****

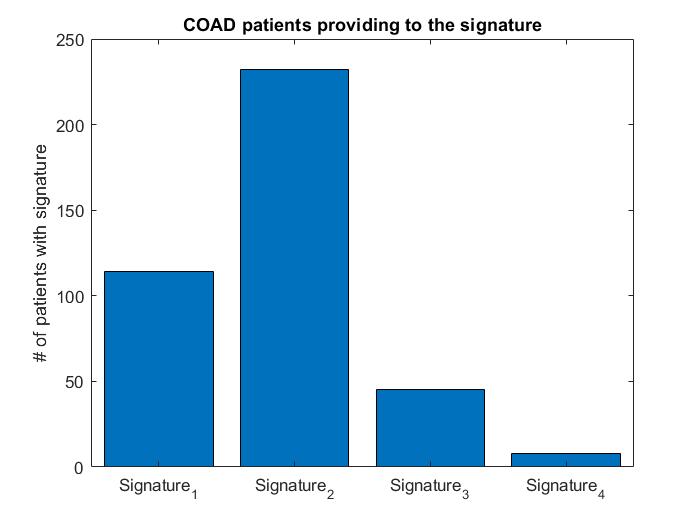
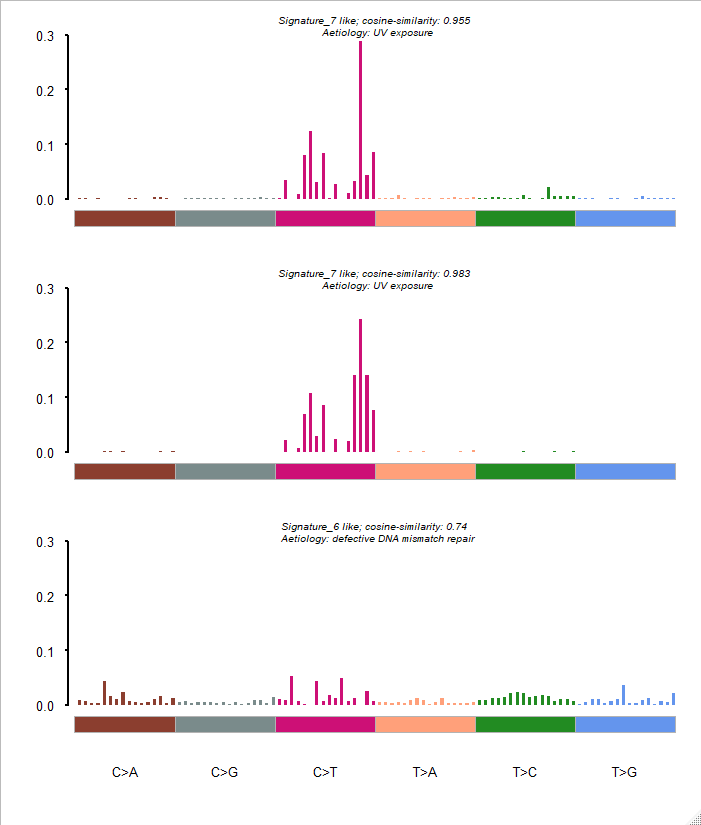


Figure 12 Decomposed COAD signatures Top: Overall mutation signatures observed across the whole cancer. Bottom: Number cases based on their dominant signature

*SKCM*

****

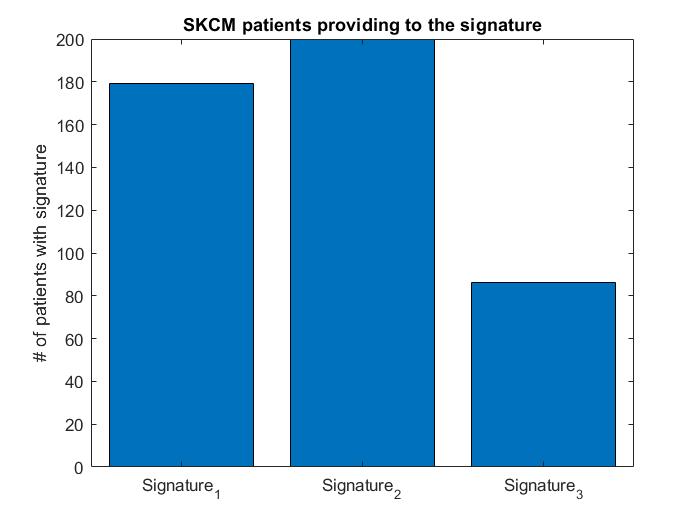
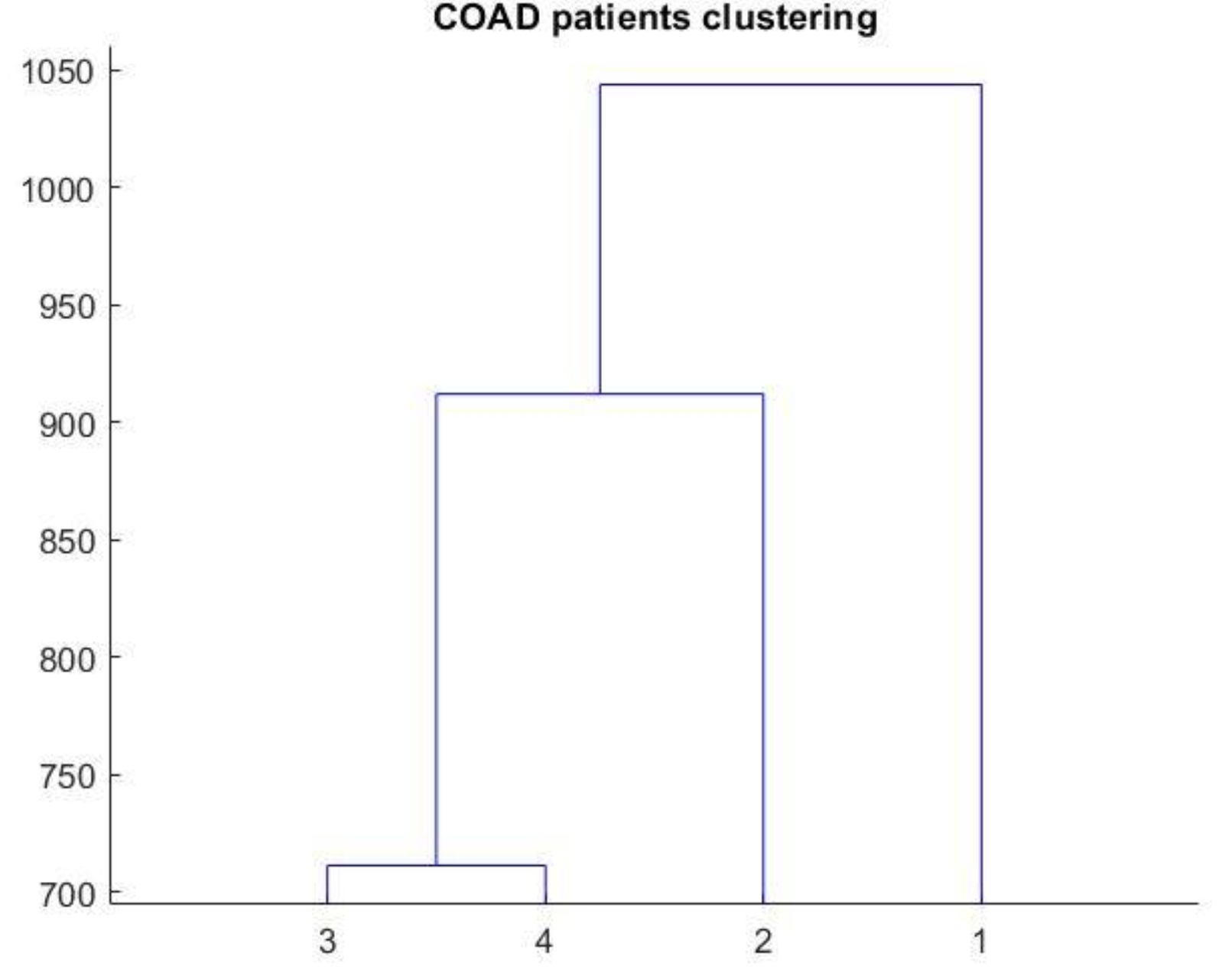
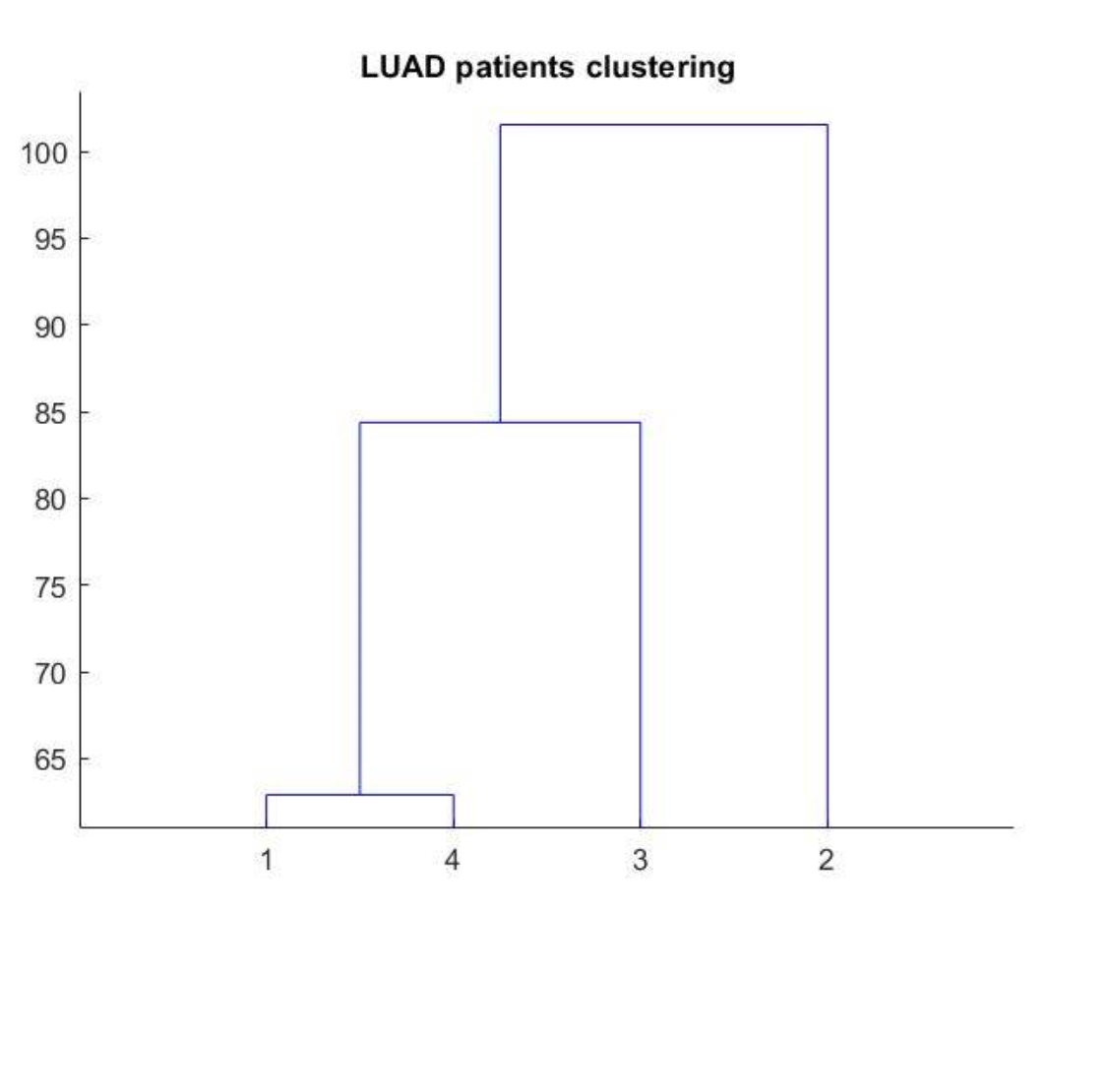
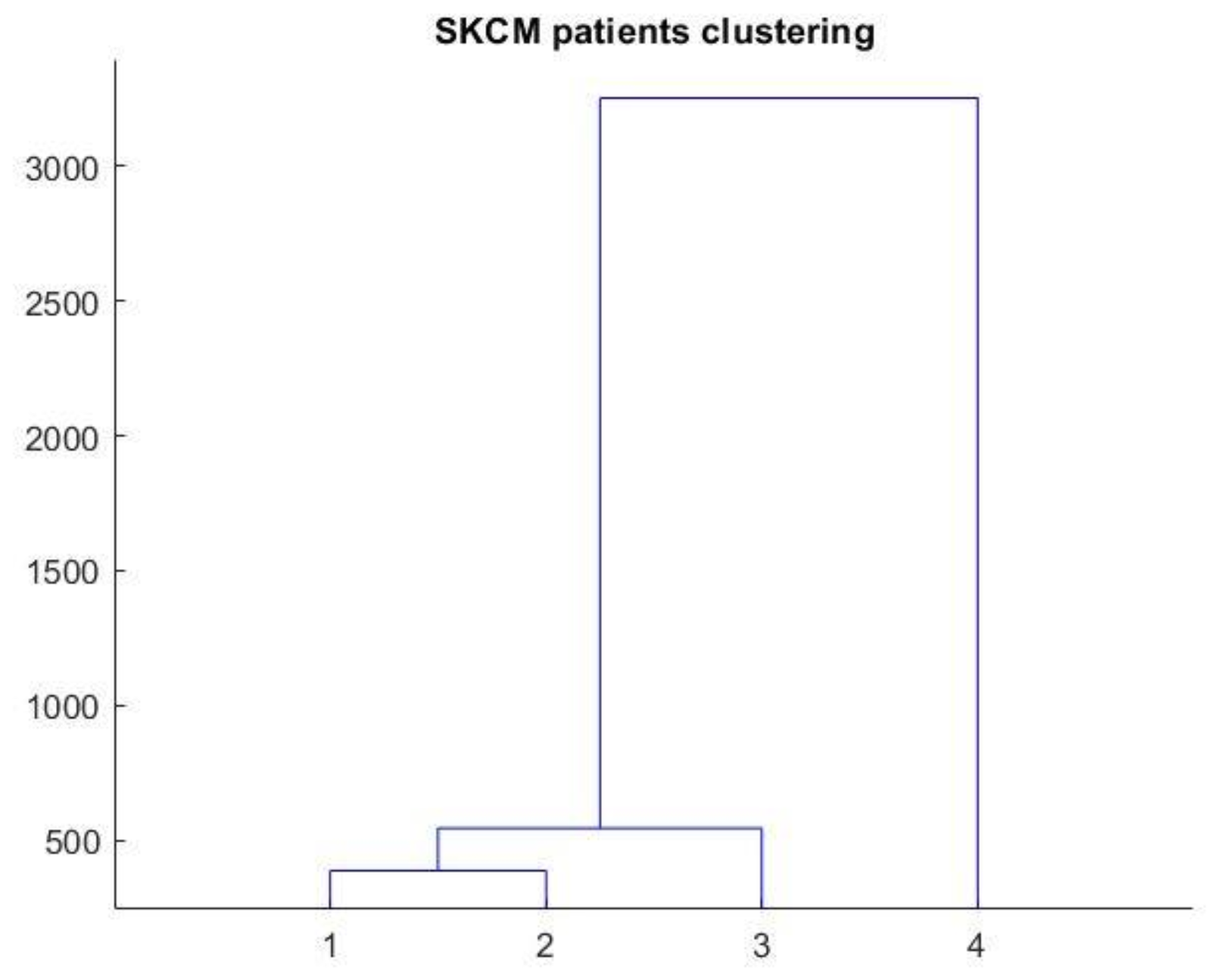
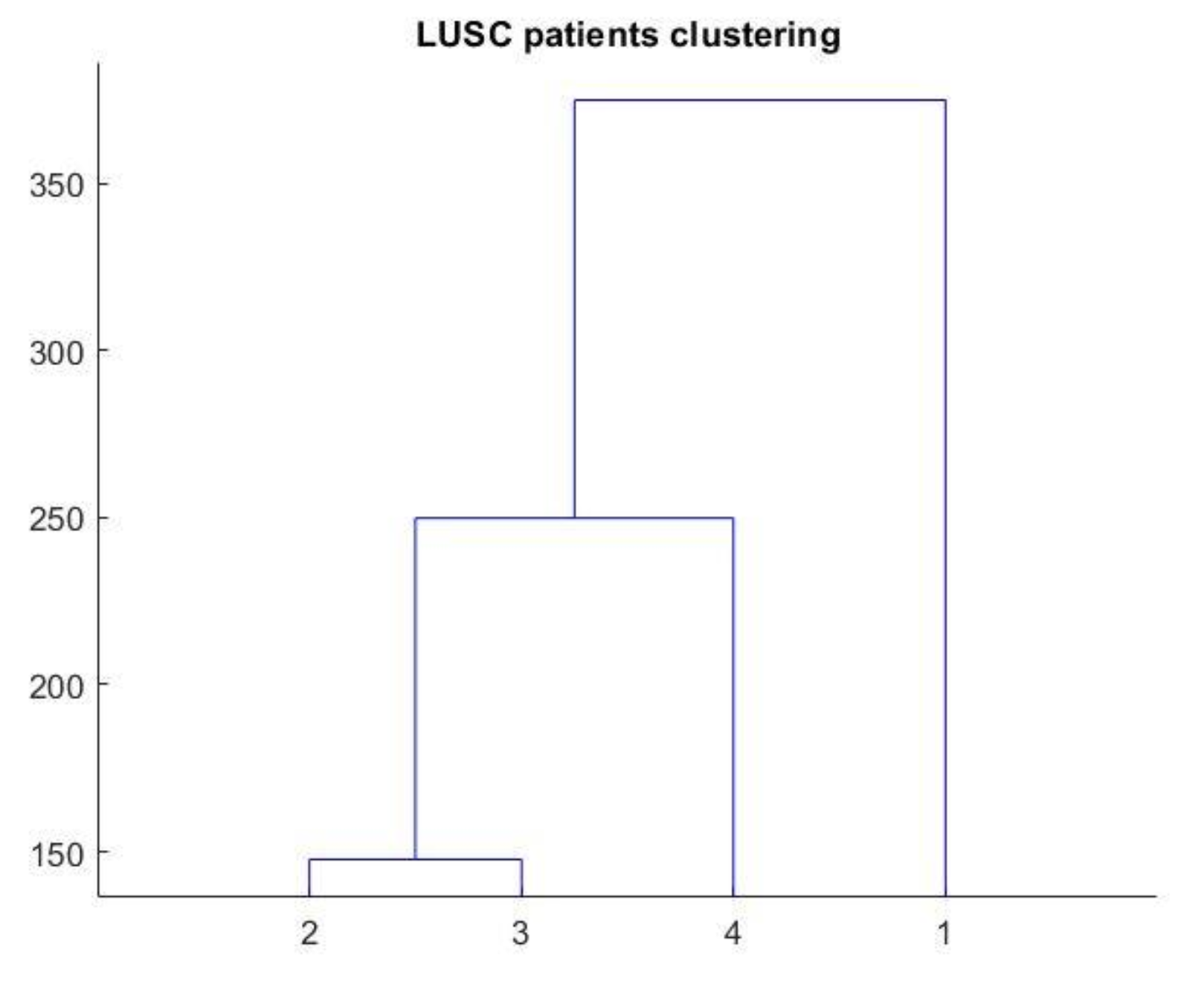


Figure 13 Decomposed SKCM signatures Top: Overall mutation signatures observed across the whole cancer. Bottom: Number cases based on their dominant signature

**Clustering**

Patients in the overall cancer studies were grouped into 4 clusters based on similarity of trinucleotide mutation counts between the different patients (see methods). In LUAD cluster1 subset is most similar to cluster4 subset, with cluster 3 and 2 being the closest branches, respectively. In COAD cluster subset3 and 4 are the closest with clusters 2 and 1 being the next closest branches, respectively. In LUSC the closest clusters are subset 2 and 3 with subset 4 and 1 being the furthest. In SKCM subset 1, 2 and 3 are relatively close with 4 being the farthest away from the other three.



\

Figure 13 Clustering Dendrograms comparing distances between clustered groups Top Left: LUAD, Bottom Left: LUSC, Top right: COAD Bottom Right: SKCM

**Mutation loads in clusters**

Below are the quantified breakdown of the cohort subsets per cancer type. The subsets tend to be in similar size apart from subset 3 in the SKCM cohort, subsets 2 and 4 in the COAD cohort, subset 3 in the LUSC cohort, and subset 3 in the LUAD cohort. Some issues we faced were with the subsets with extremely low numbers of samples, namely subsets 2 and 4 in the COAD cohort was generation of signatures and pathway analysis. Due to their small sample sizes, 3 and 5 respectively, we weren't able to generate figures for these subsets due to an error. On the other hand, median mutations per sample did not vary as drastically. At most, the maximum is three fold the minimum value. We also found that when we generated figures comparing where the subsets stand with the rest of tumour types, the subsets generally tended to stay local to their respective parent cohort (shown in the figures below). The scatter plot below are based on the log of the median mutations per sample.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cancer Cohort** | **Number of Samples** | **Median Mutations per sample** | **Log of Median Mutations** |
| **SKCM overall** | **465** | **244** | **2.3873898** |
| subset1 | 152 | 249.0 | 2.3961993 |
| subset2 | 130 | 235.5 | 2.3719909 |
| subset3 | 62 | 267.0 | 2.4265113 |
| subset4 | 120 | 240 | 2.3802112 |
| **COAD Overall** | **399** | **89.0** | **1.9493900** |
| subset1 | 235 | 89.0 | 1.9493900 |
| subset2 | 3 | 145.0 | 2.1613680 |
| subset3 | 155 | 86.0 | 1.9344985 |
| subset4 | 5 | 101.0 | 2.0043214 |
| **LUSC Overall** | **490** | **173.0** | **2.2380461** |
| subset1 | 195 | 185.0 | 2.2671717 |
| subset2 | 106 | 164.5 | 2.2161659 |
| subset3 | 31 | 159.0 | 2.2013971 |
| subset4 | 157 | 168.0 | 2.2253093 |
| **LUAD Overall** | **559** | **135.0** | **2.1303338** |
| subset1 | 158 | 160.5 | 2.2054750 |
| subset2 | 163 | 128.0 | 2.1072100 |
| subset3 | 36 | 110.5 | 2.0433623 |
| subset4 | 201 | 122.0 | 2.0863598 |

**Table 2. Clustered Cohorts Summary Table**

*SKCM*

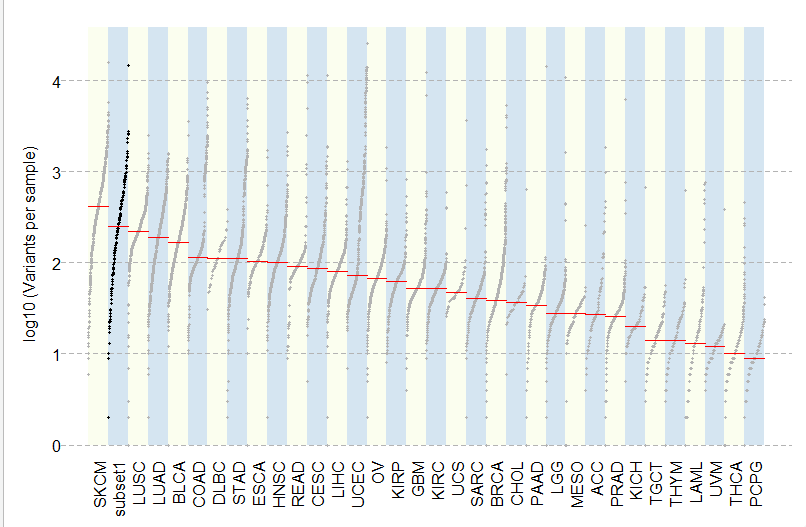


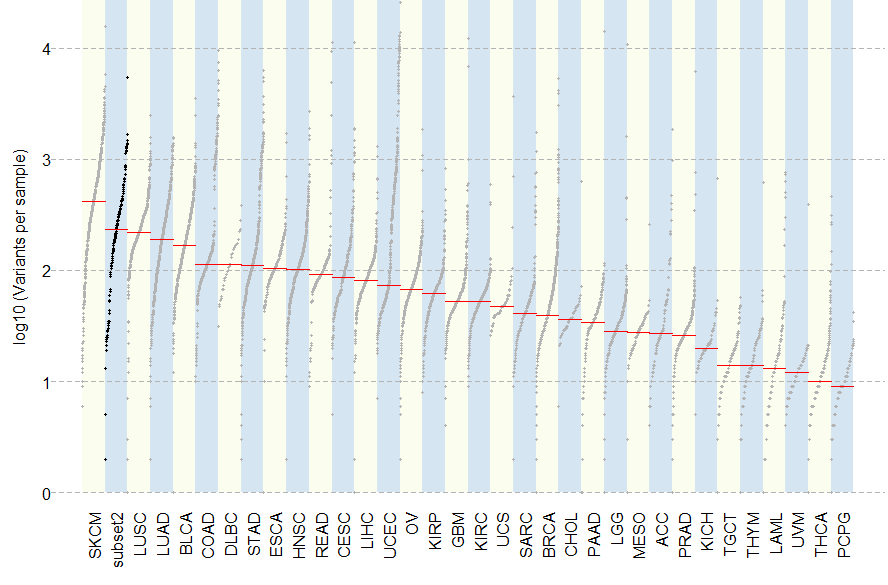
Figure 14a Mutation load comparison of SKCM subset1 to TCGA cancer types

Figure 14b Mutation load comparison of SKCM subset2 to TCGA cancer types

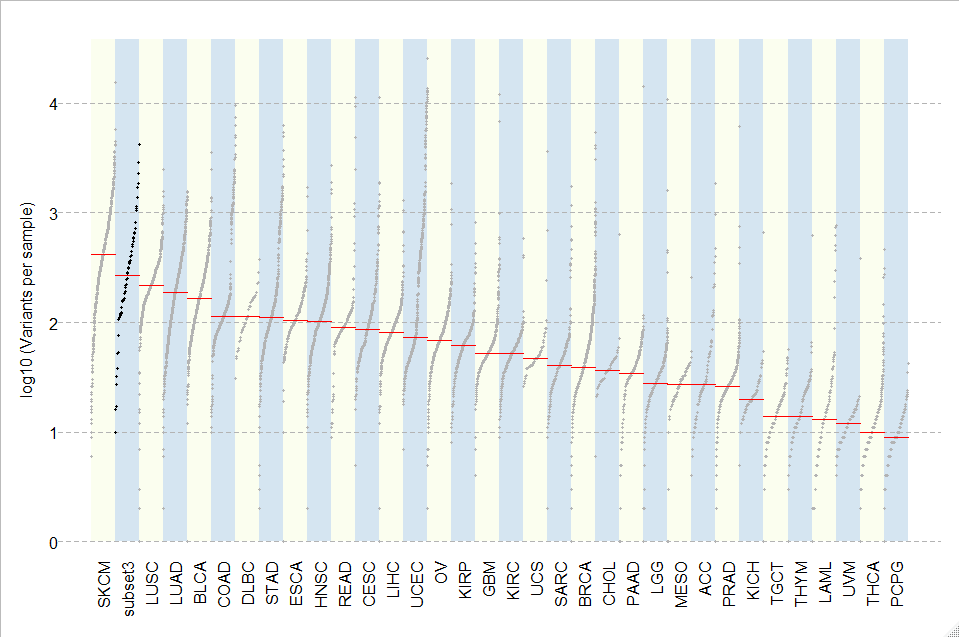


Figure 14c Mutation load comparison of SKCM subset3 to TCGA cancer types

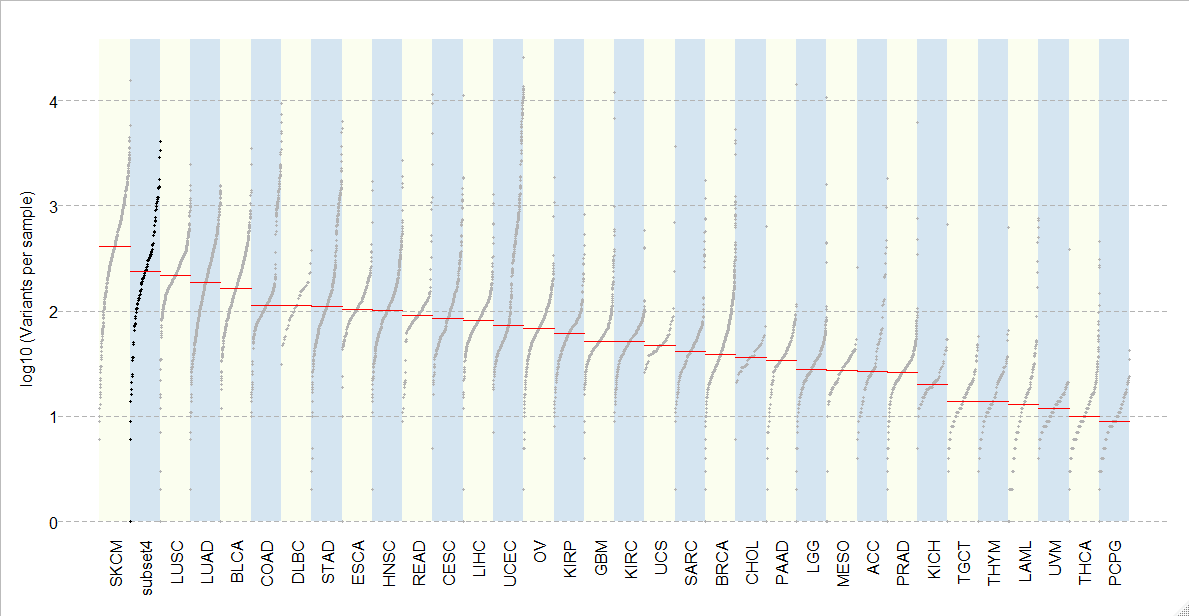


Figure 14d Mutation load comparison of SKCM subset4 to TCGA cancer types

*COAD*

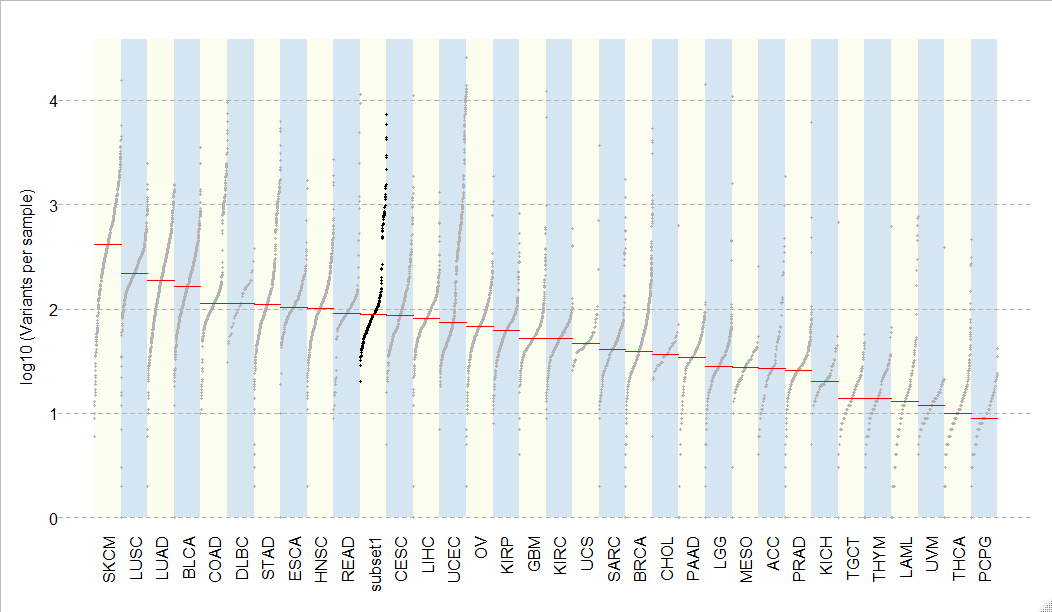


Figure 15a Mutation load comparison of COAD subset1 to TCGA cancer types

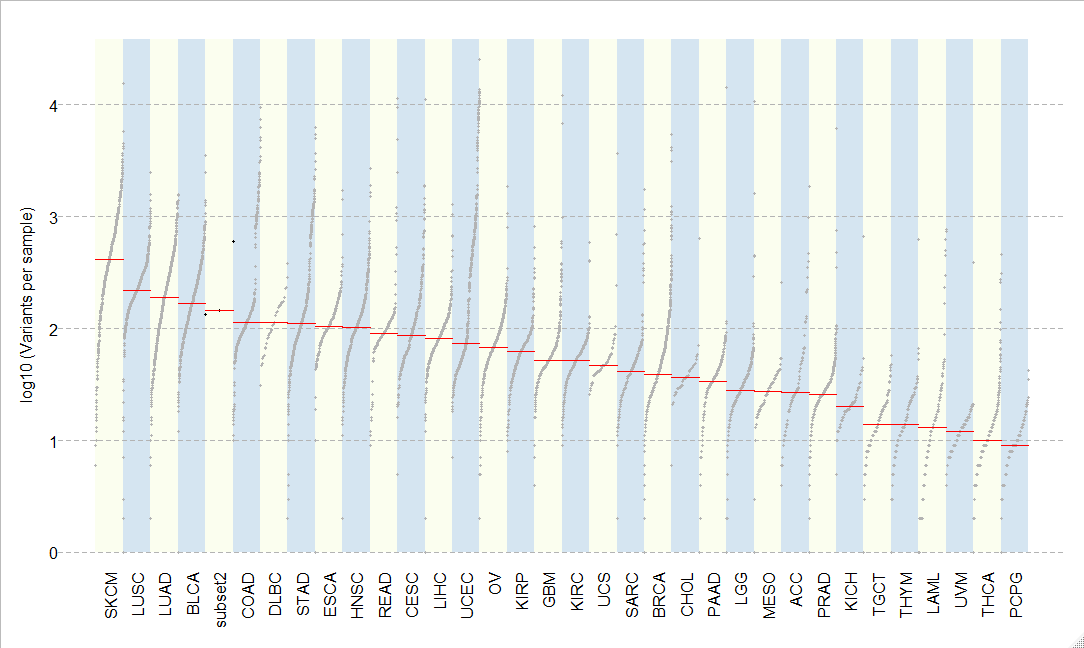


Figure 15b Mutation load comparison of COAD subset2 to TCGA cancer types

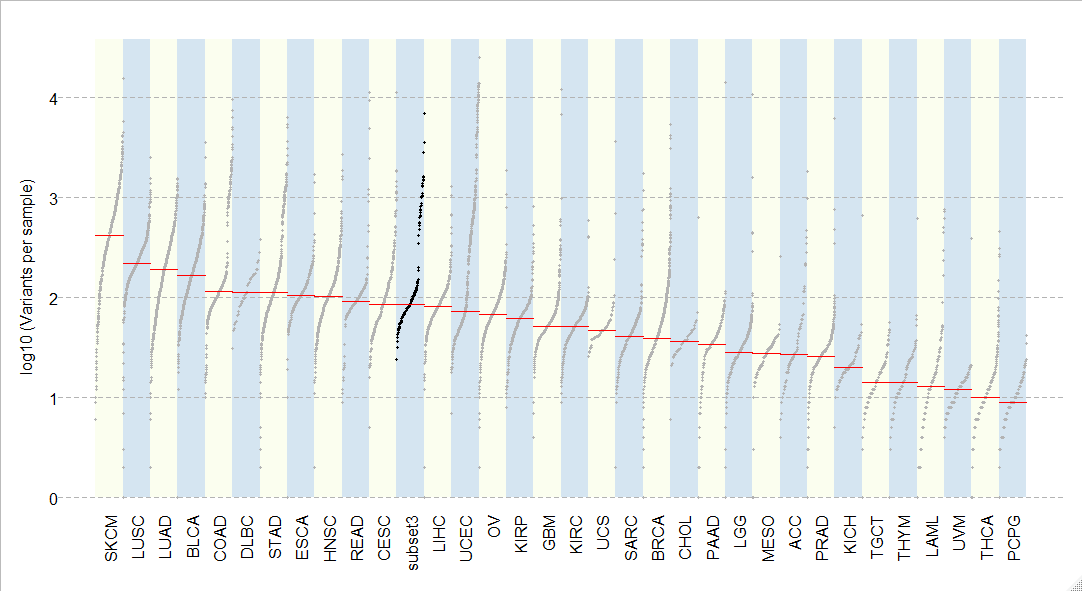


Figure 15c Mutation load comparison of COAD subset3 to TCGA cancer types

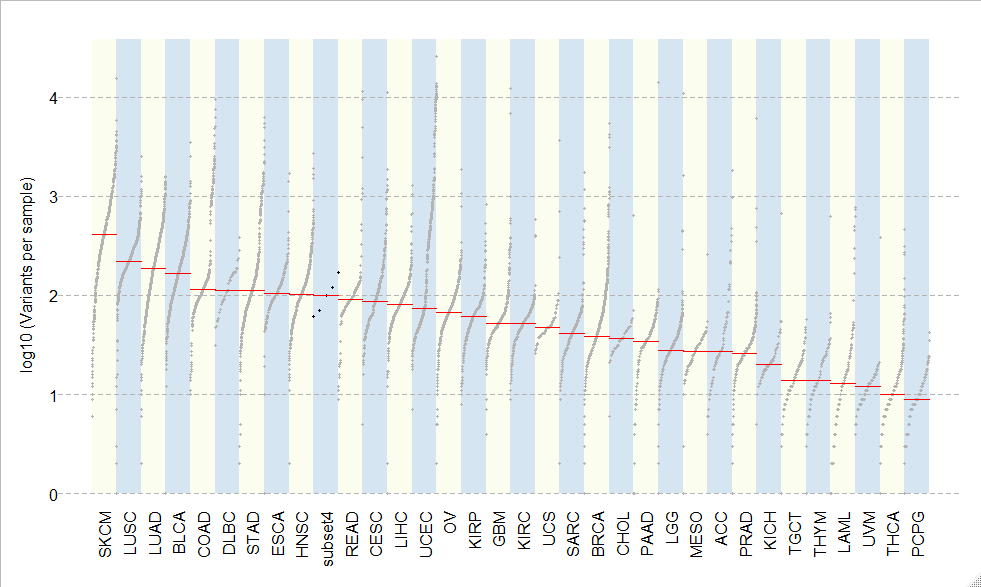


Figure 15d Mutation load comparison of COAD subset4 to TCGA cancer types

*LUSC*

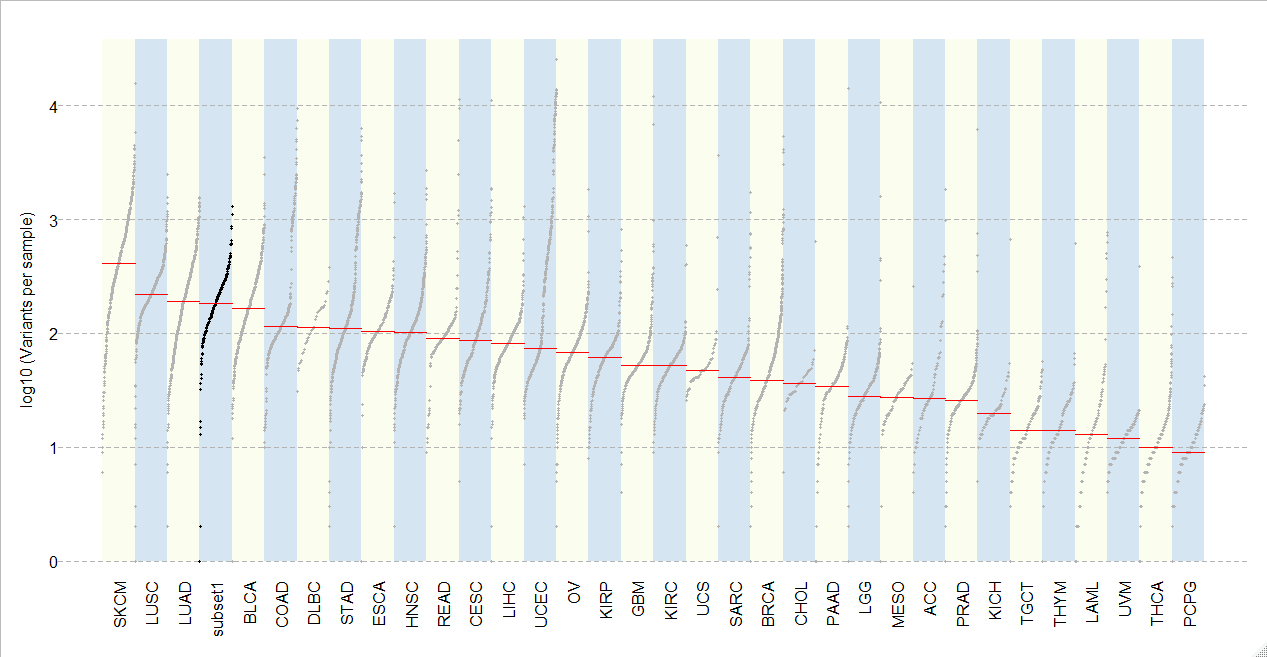


Figure 16a Mutation load comparison of LUSC subset1 to TCGA cancer types

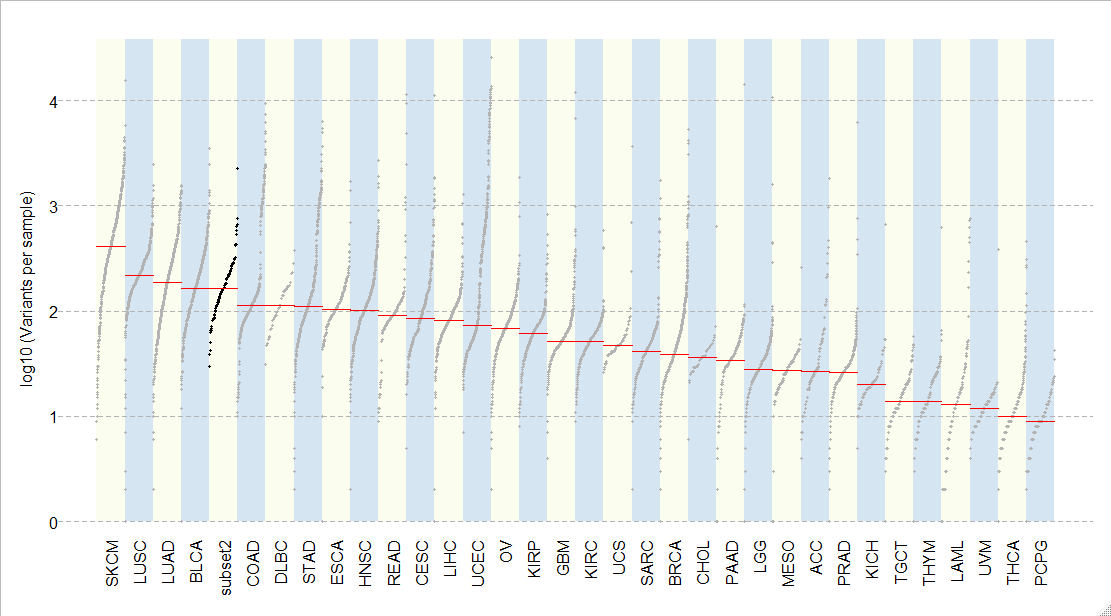


Figure 16b Mutation load comparison of LUSC subset2 to TCGA cancer types



Figure 16c Mutation load comparison of LUSC subset3 to TCGA cancer types

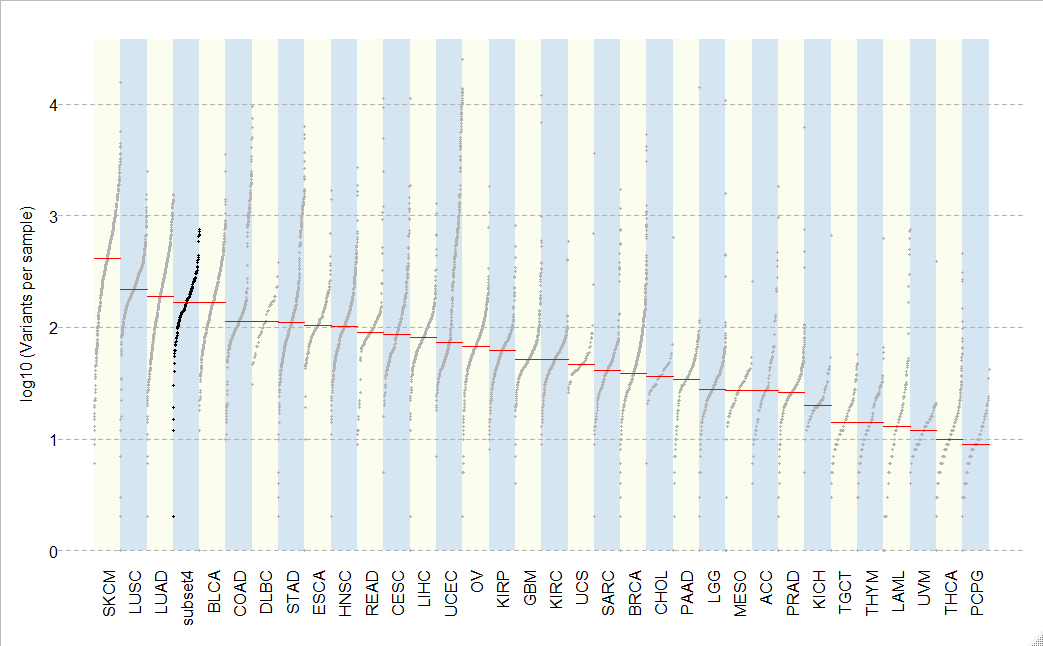
4

Figure 16d Mutation load comparison of LUSC subset3 to TCGA cancer types

*LUAD*

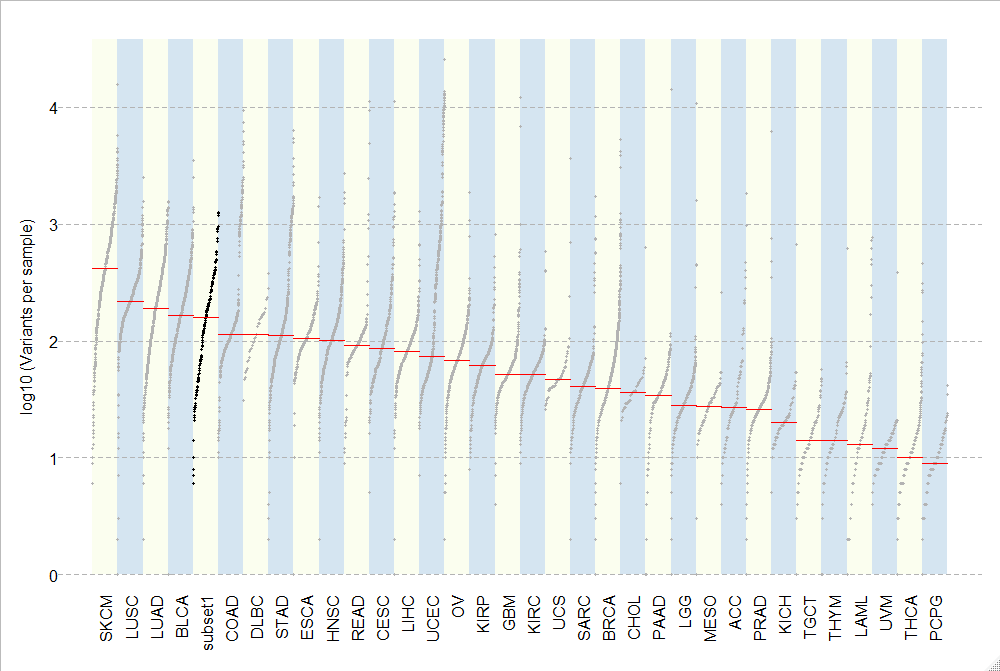


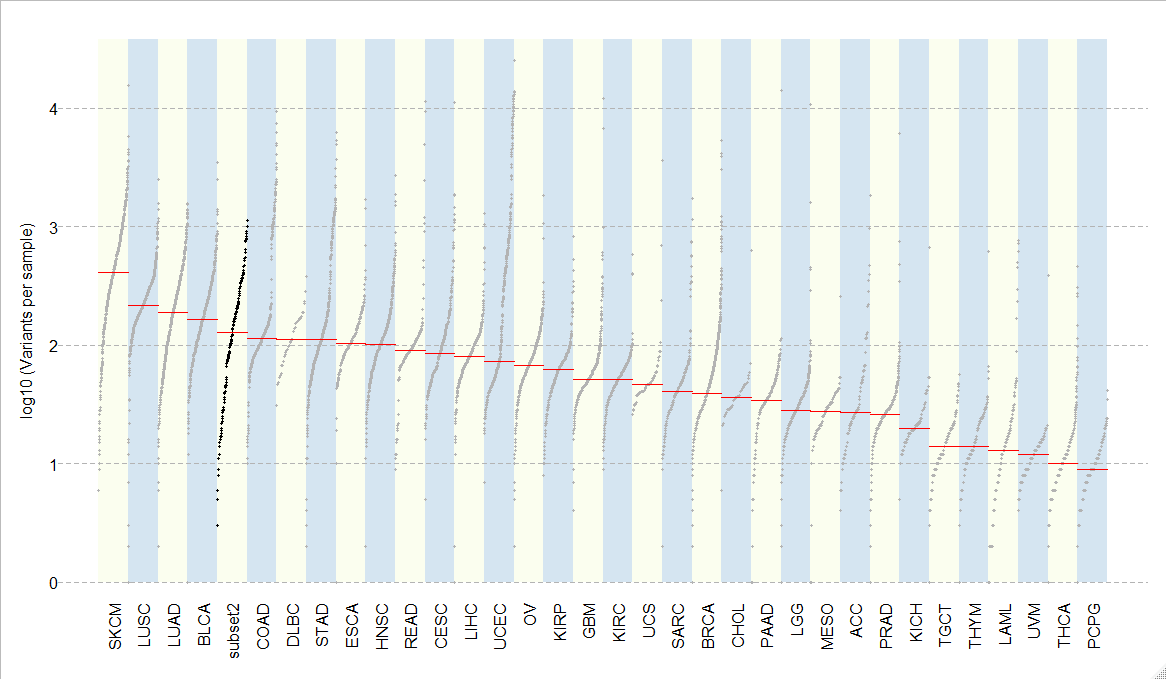
Figure 17a Mutation load comparison of LUAD subset1 to TCGA cancer types

Figure 17b Mutation load comparison of LUAD subset2 to TCGA cancer types

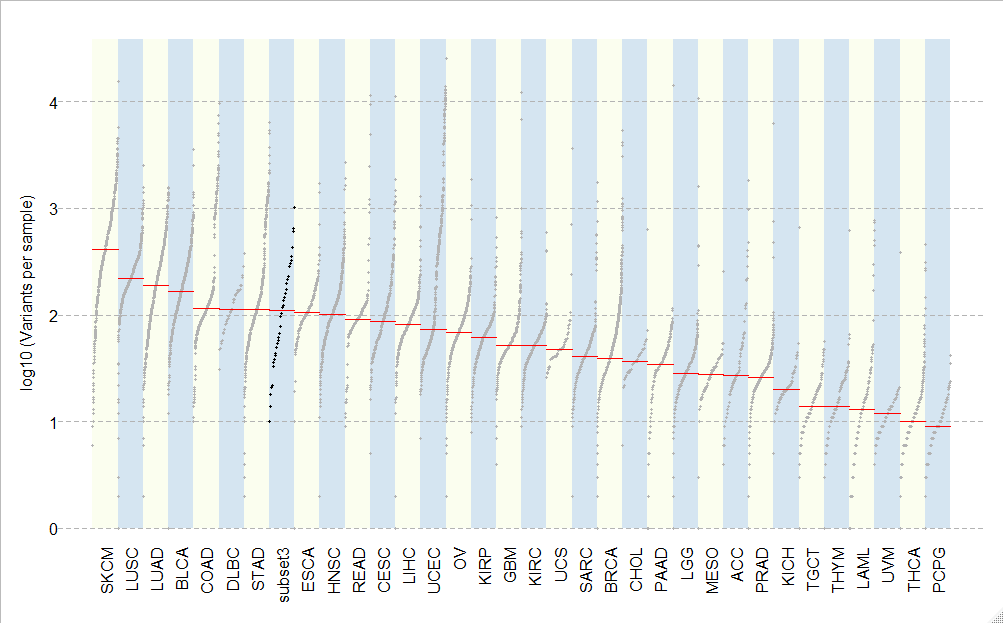


Figure 17c Mutation load comparison of LUAD subset3 to TCGA cancer types

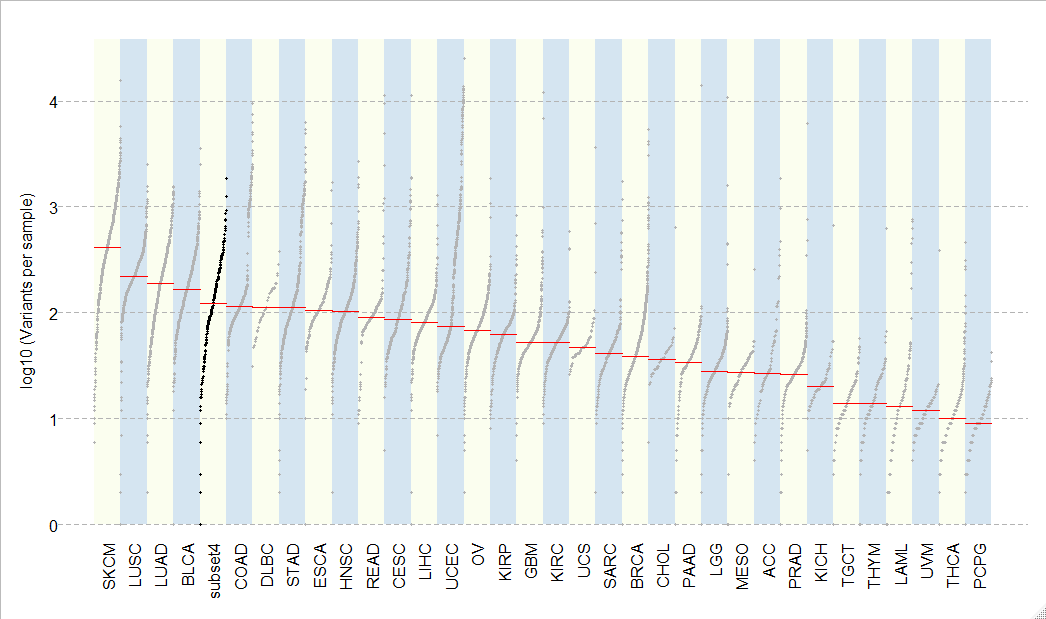
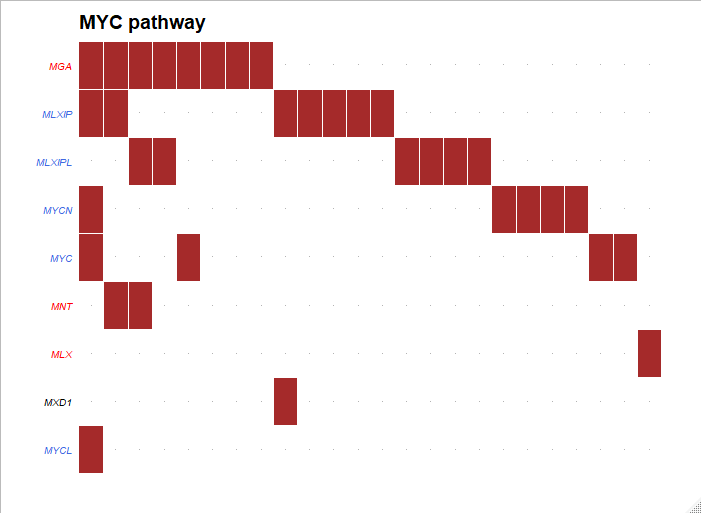


Figure 17d Mutation load comparison of LUAD subset4 to TCGA cancer types

**Pathway Analysis in clustered subsets**

In the pathway analysis below, we get to visualize which pathways are involved and enriched within the clustered patient subsets per cancer type. Columns represent Individual Samples and each individual brown box represents a distinct mutation. On the left-hand side are gene pathways affected. Those labeled in red are genes involved in tumor suppressor pathways (T*p53*). Those labeled in blue are canonical oncogenes (MYC). some samples/patients had several mutations per sample, others had only one that was associated with a single gene pathway. Most interesting observations were when some subsets had much higher enrichment in one pathway type over another (*Tp53* vs *MYC*), this was seen in COAD, LUAD, LUSC. It was also surprising to see an enrichment of mutations in DNA repair pathways in COAD. Despite only choosing these two pathways and using a basic tool for analysis we were able to see differences among the subset even only using these small set of genes and differences may be even more pronounced using and expanded gene set and a more comprehensive tool.

*SKCM*

****

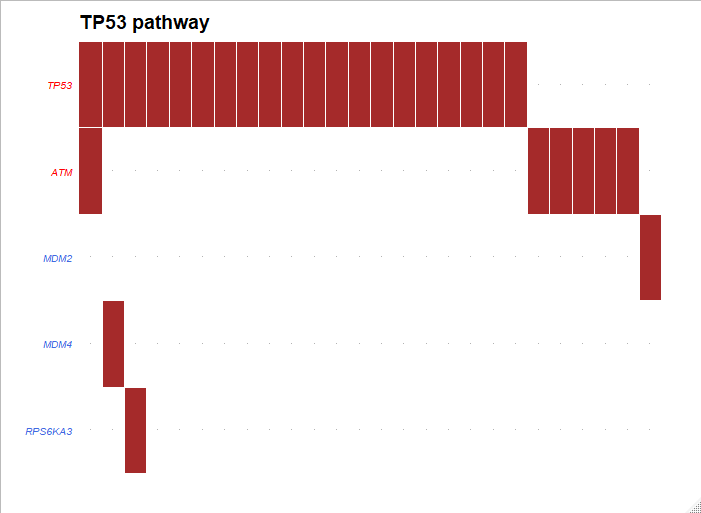
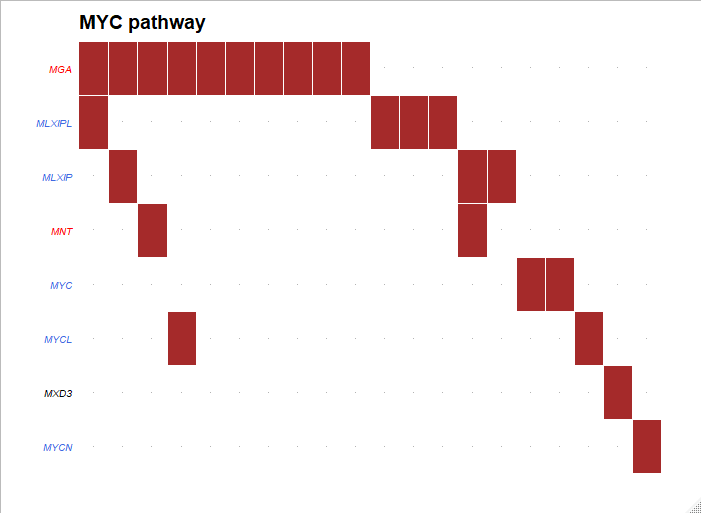


Figure 18a Pathway enrichment Analysis for SKCM subset1 Top (*MYC* pathway). Bottom (*TP53* pathway)



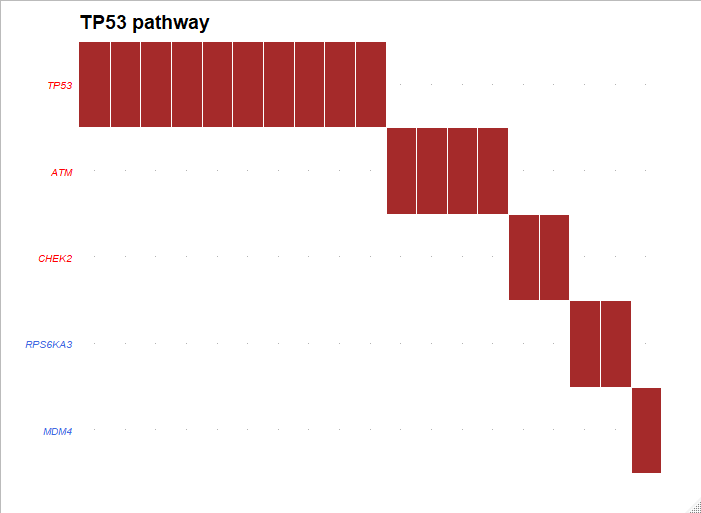
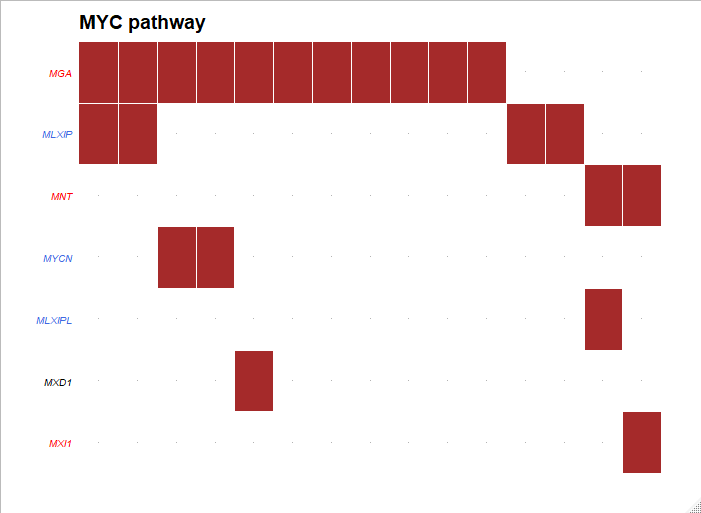


Figure 18b Pathway enrichment Analysis for SKCM subset2 Top (MYC pathway). Bottom (TP53 pathway)



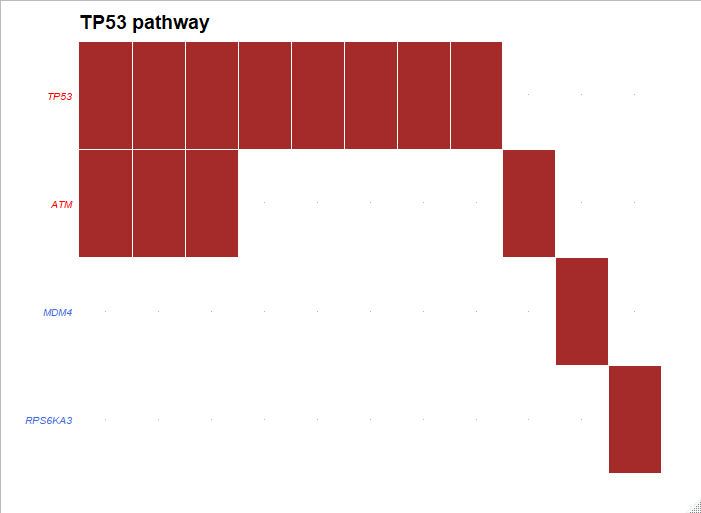
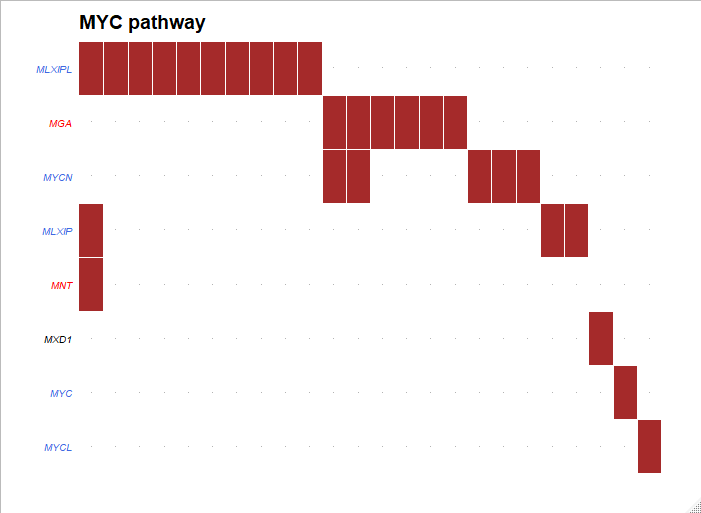


Figure 18c Pathway enrichment Analysis for SKCM subset3 Top (MYC pathway). Bottom (TP53 pathway)



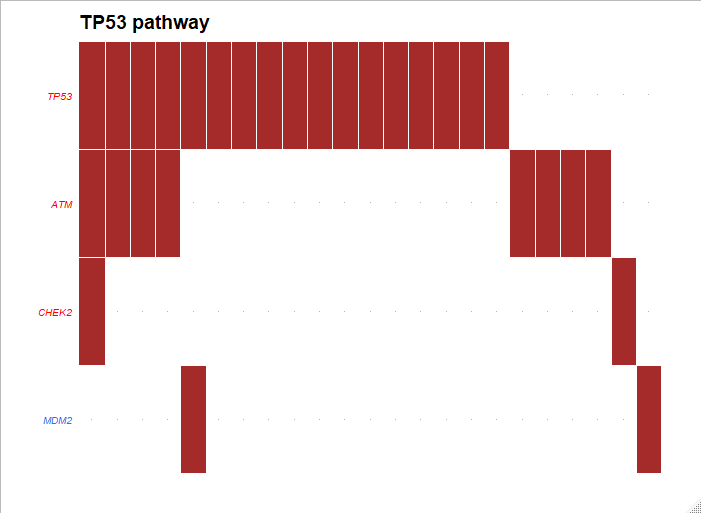
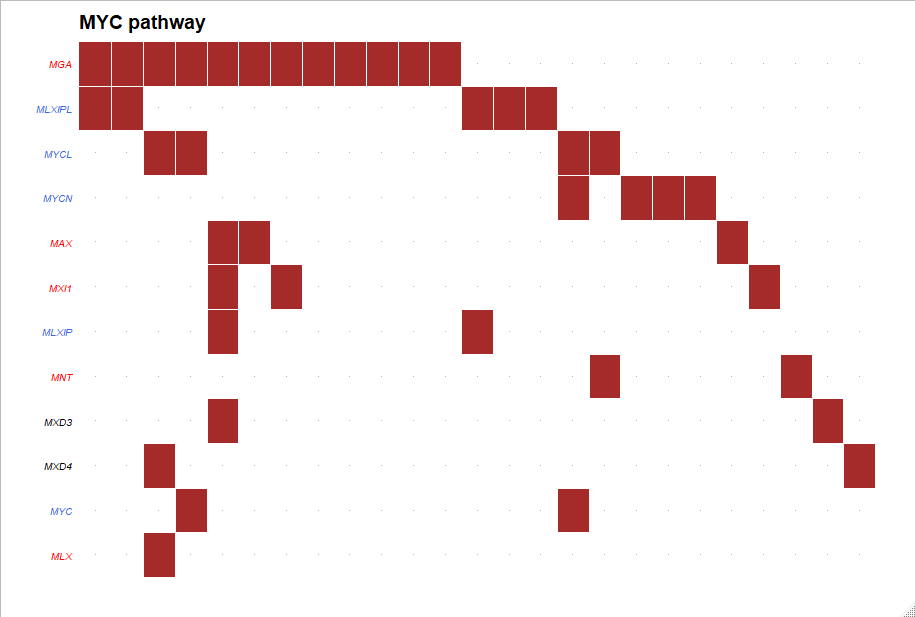


Figure 18d Pathway enrichment Analysis for SKCM subset4 Top (MYC pathway). Bottom (TP53 pathway)

*COAD*



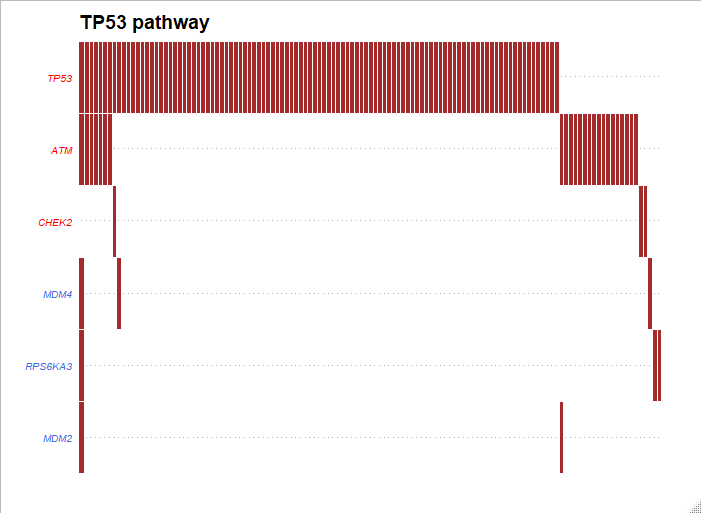
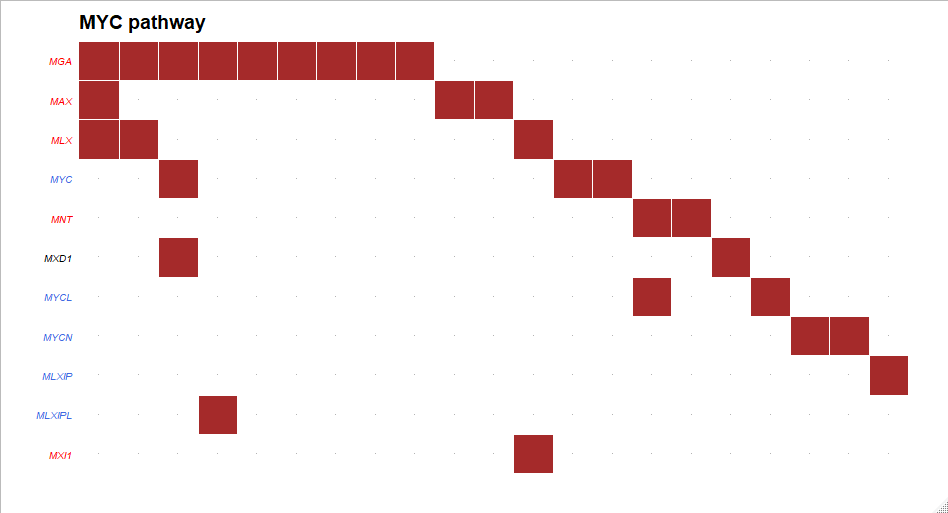


Figure 19a Pathway enrichment Analysis for COAD subset1 Top (MYC pathway). Bottom (TP53 pathway)



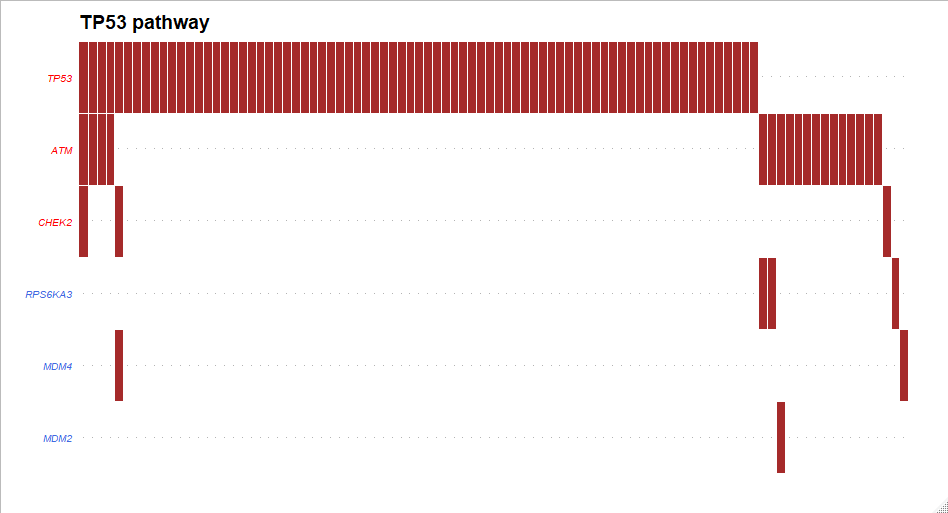
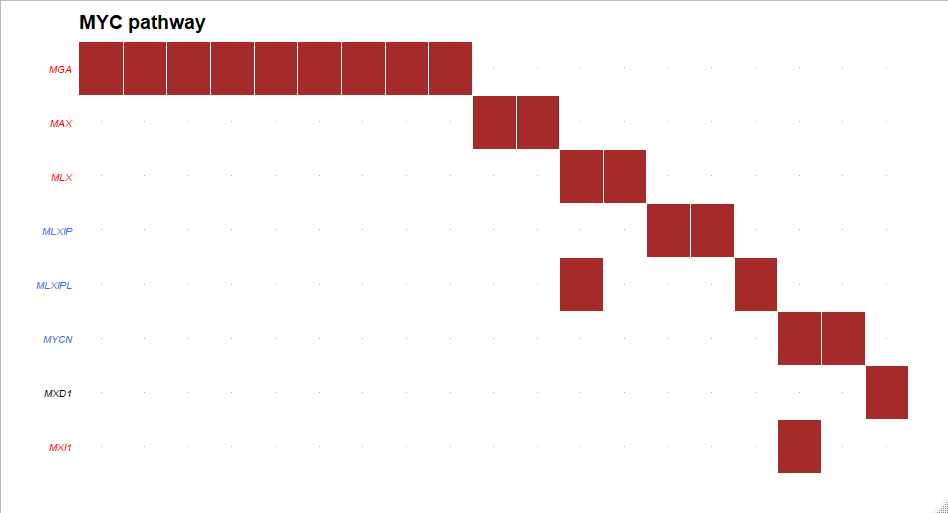


Figure 19b Pathway enrichment Analysis for COAD subset3 Top (MYC pathway). Bottom (TP53 pathway)

*LUAD*



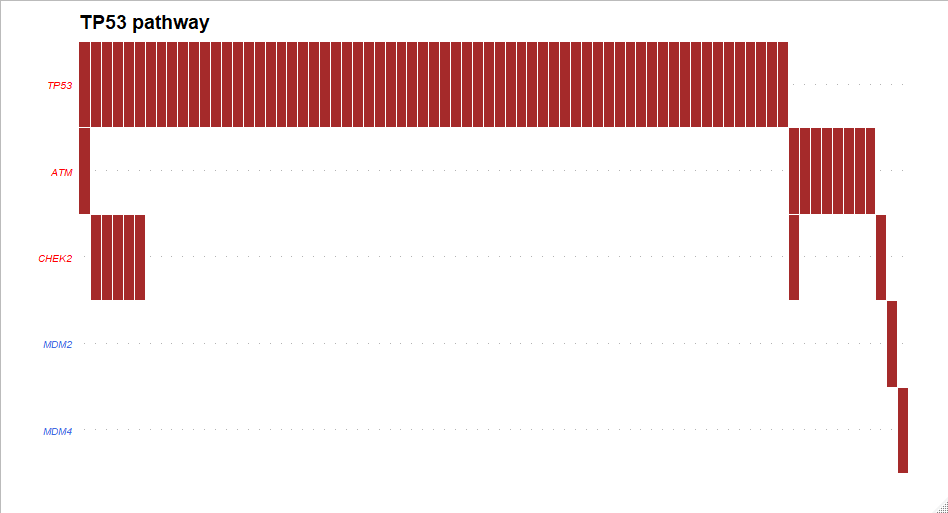
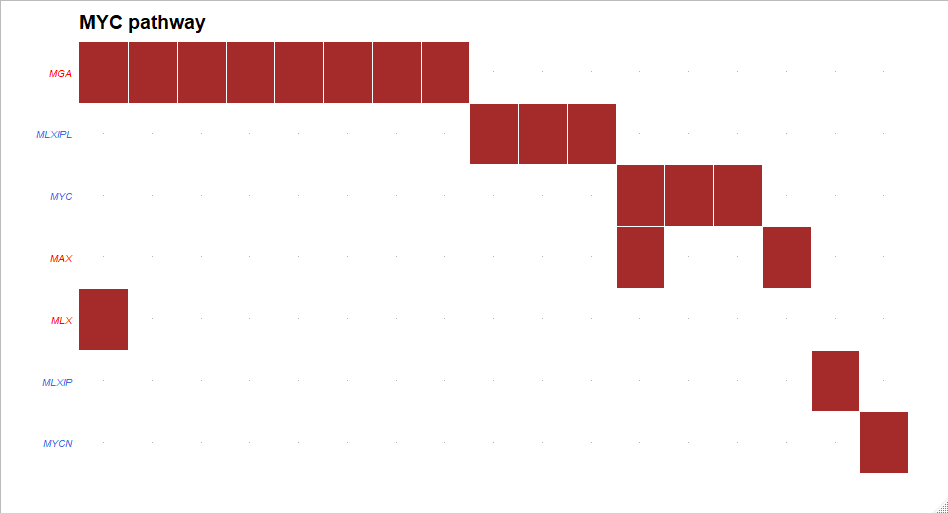


Figure 20a Pathway enrichment Analysis for LUAD subset1 Top (MYC pathway). Bottom (TP53 pathway)



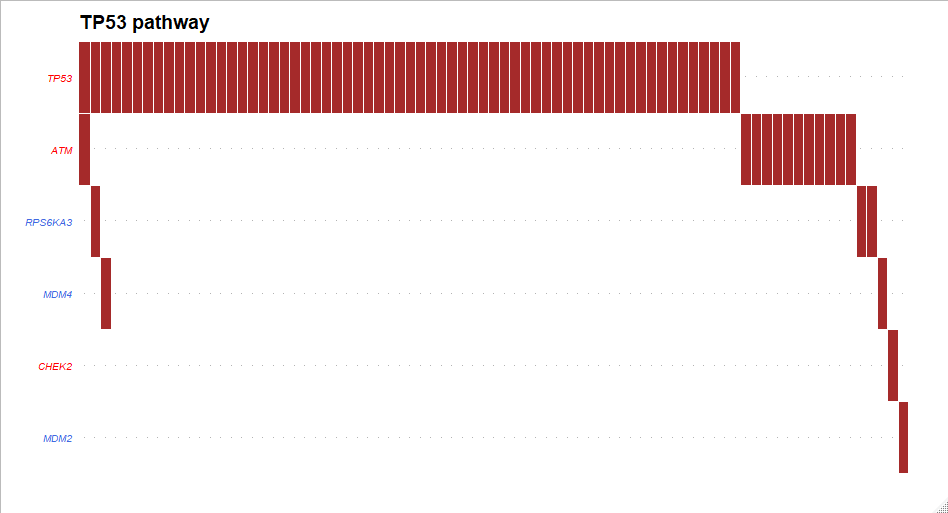
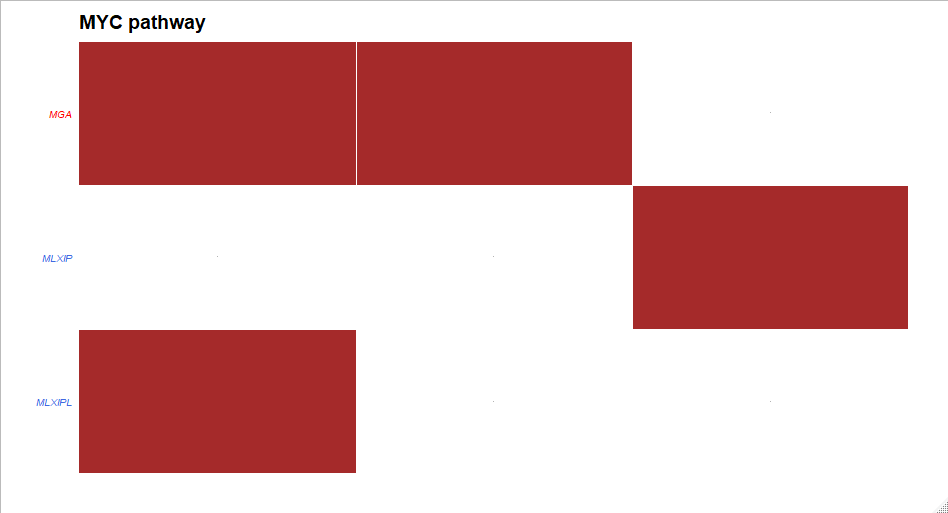


Figure 20b Pathway enrichment Analysis for LUAD subset2 Top (MYC pathway). Bottom (TP53 pathway)



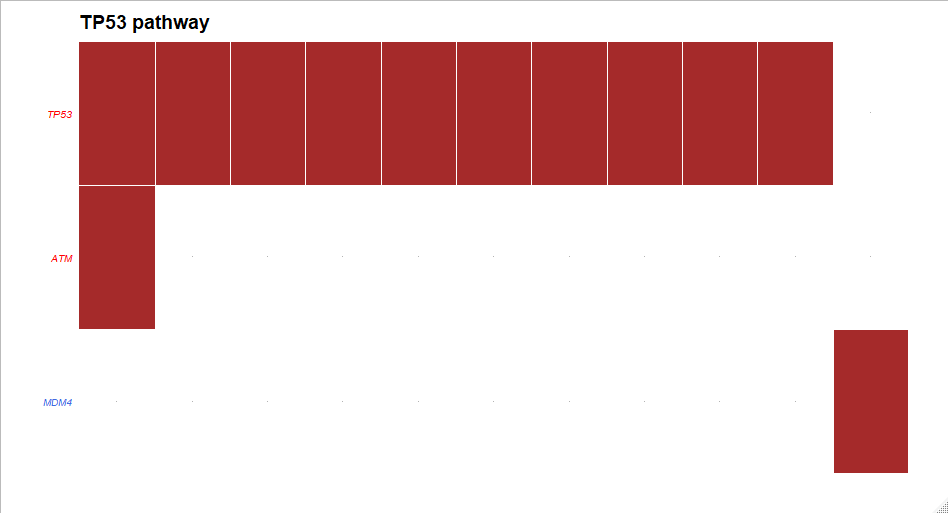
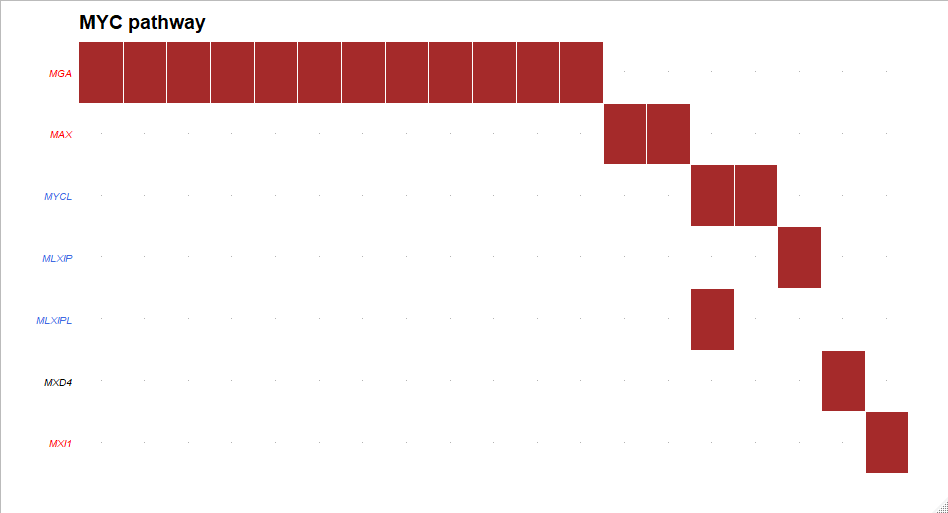


Figure 20c Pathway enrichment Analysis for LUAD subset3 Top (MYC pathway). Bottom (TP53 pathway)



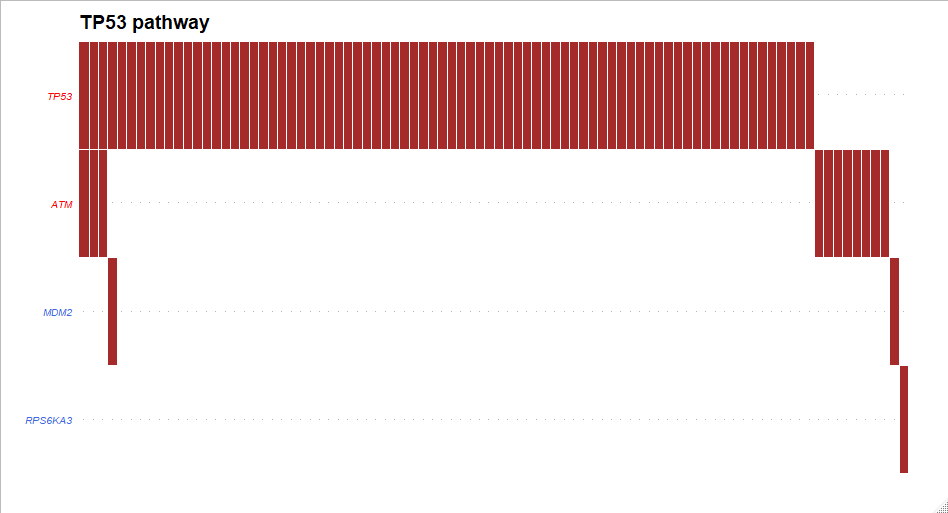
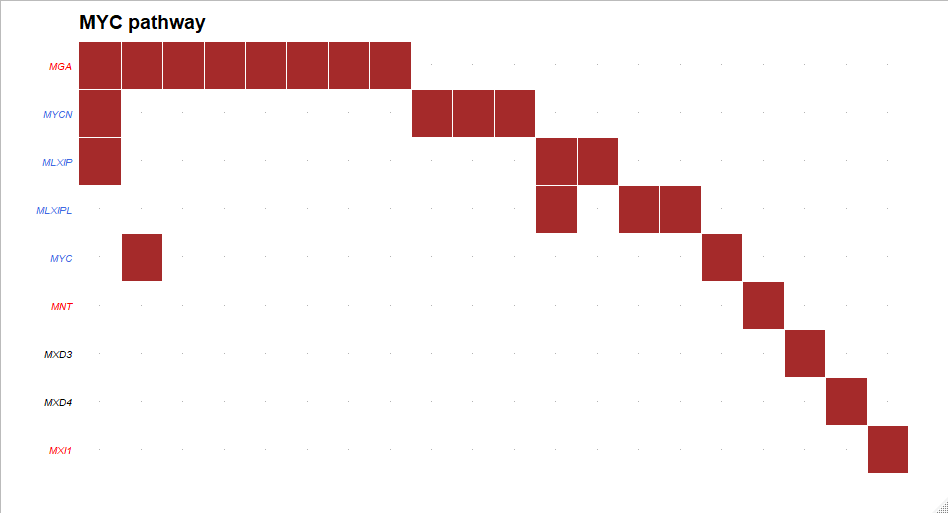


Figure 20d Pathway enrichment Analysis for LUAD subset4 Top (MYC pathway). Bottom (TP53 pathway)

*LUSC*



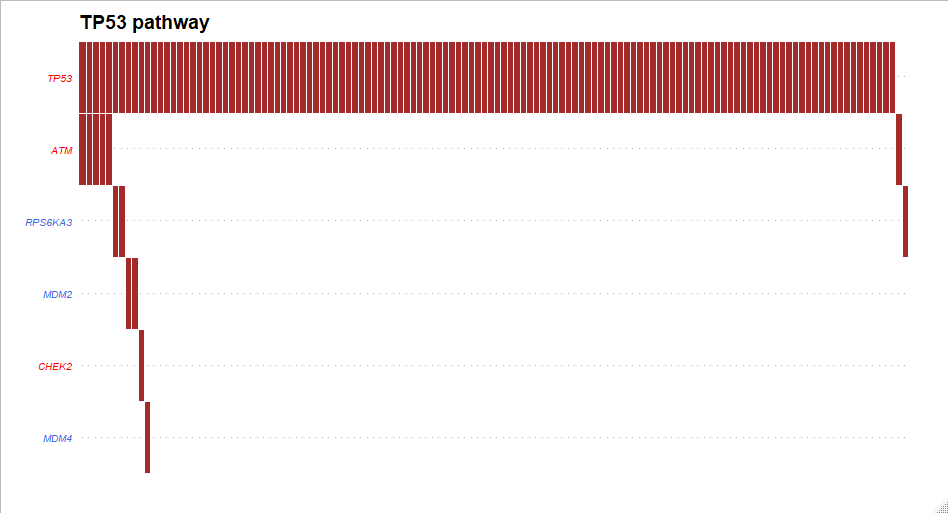
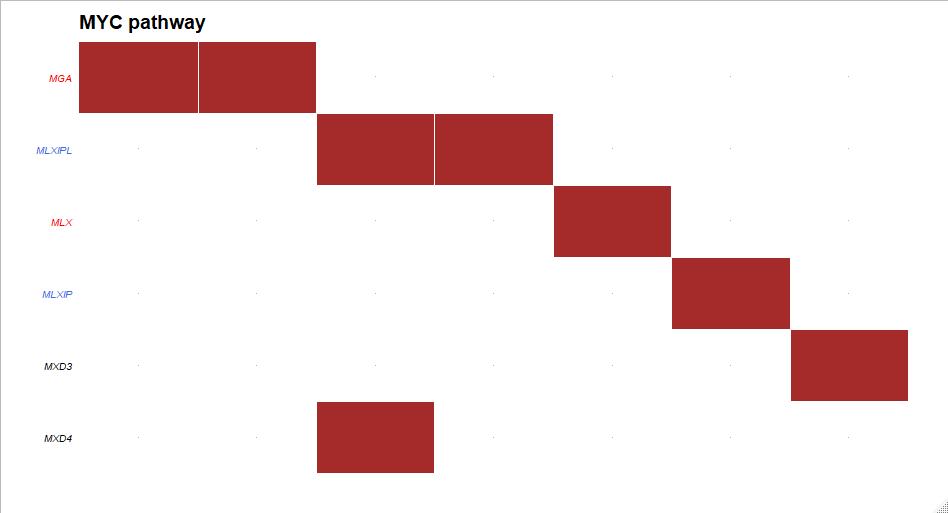


Figure 21a Pathway enrichment Analysis for LUSC subset1 Top (MYC pathway). Bottom (TP53 pathway)



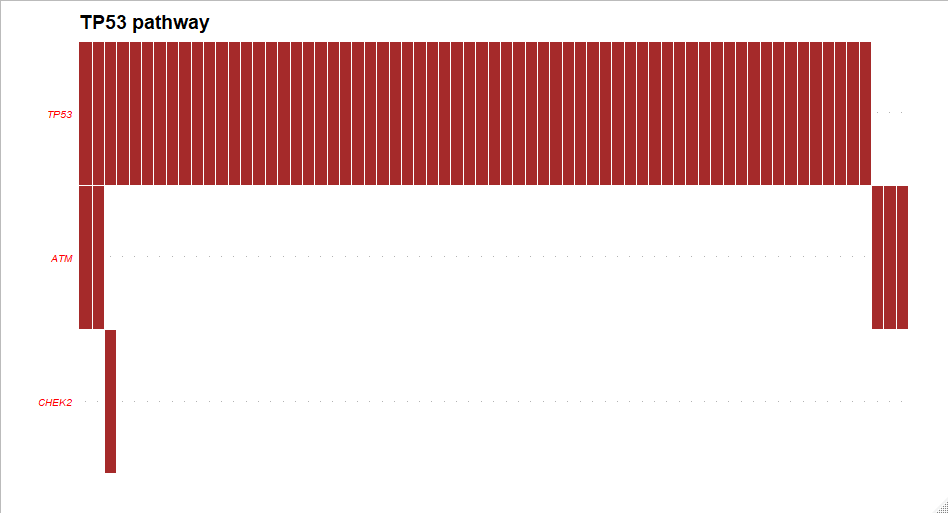
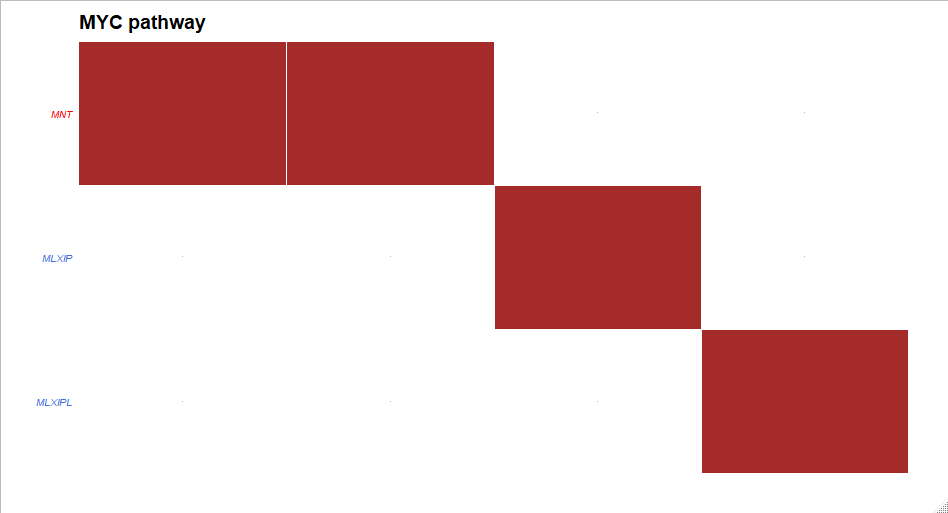


Figure 21b Pathway enrichment Analysis for LUSC subset2 Top (MYC pathway). Bottom (TP53 pathway)



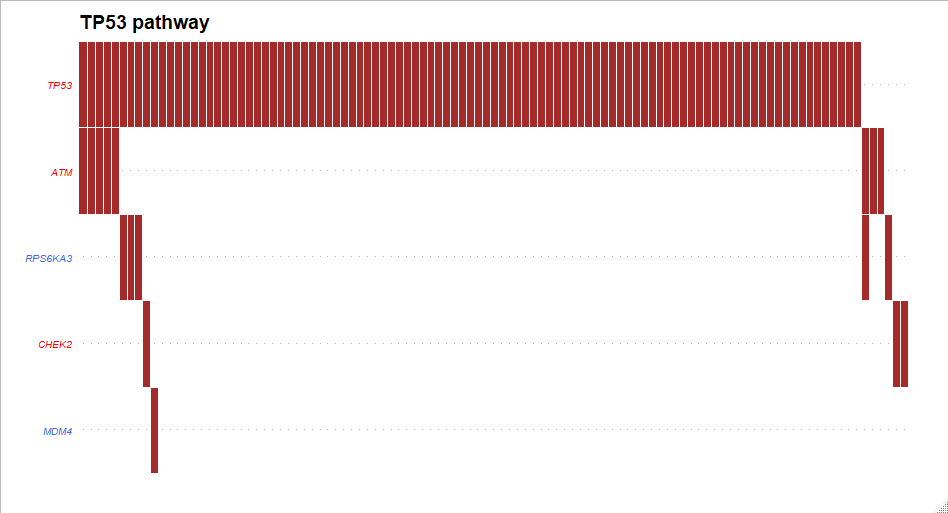
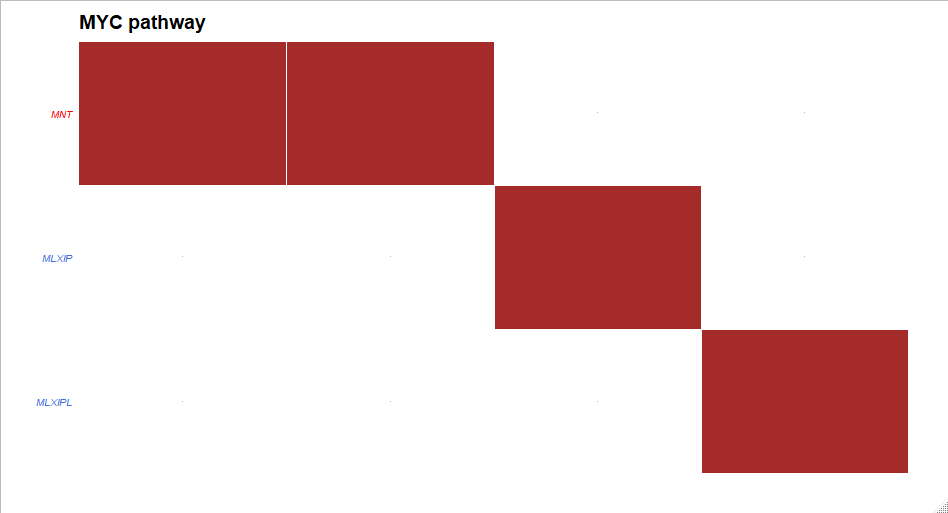


Figure 21c Pathway enrichment Analysis for LUSC subset3 Top (MYC pathway). Bottom (TP53 pathway)



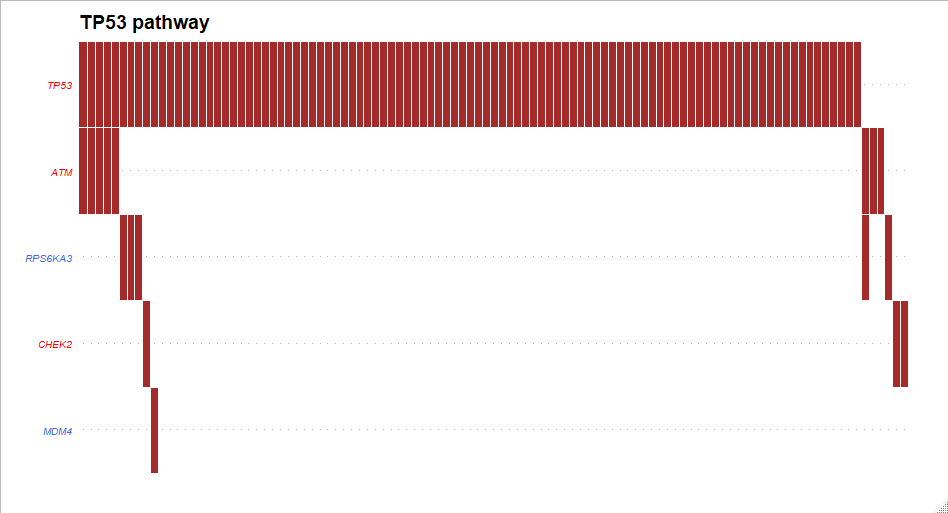


Figure 21d Pathway enrichment Analysis for LUSC subset4 Top (MYC pathway). Bottom (TP53 pathway)

**Signatures found in subsets**

For COAD subset1 (n= 235) two different signatures matching to COSMIC signature 10 (POLE defects) as well as a signature similar to signature 6 (defective DNA mismatch repair)For Subset 3 (n=155 ) we found two different signatures matching COSMIC signature 1 (5-methylcytosine deamination) as well as two different signatures matching signatures 10 and signature 15, which interestingly is a different signature than 6 but also for defective DNA mismatch repair. Interestingly Signature 2 of unknown aetiology was not observed either. We were not able to get results for clusters 3 (n=3) and 4 (n=5) due to low sample sizes in both and function output was an error.

For SKCM all of the subsets even after clustering had very similar decomposed signatures. In subset1 (n=152) matched COSMIC signatures were two different signatures matching to COSMIC signature 7 (U.V exposure) and signature 6. In subset2 (n= 130) matching signatures for the three types were signature 7, signature 7 and signature 1 (spontaneous deamination of 5-methylcytosine). For subset 3 (n=62) same exact signatures were found as in subset2. For subset 4 (n=120) two of the signatures also matched to signature 7 but signature matching to COSMIC signature 5 of unknown aetiology was also observed. Finally, in subset4 we observed 3 signatures matching to signature 7 and one signature similar to signature 6.

After clustering we found differing results in lung cancer as well. In LUAD study subset1(n=158) we observed matches to signatures 13 (APOBEC Cytidine Deaminase), 1 (deamination of 5-methylcytosine) and 4 (tobacco smoking). For subset 2 (n=163) we observed three different signatures matching signatures 13, 6 and 2 signatures matching 4. In subset 3 (n=36) two of the signatures were similar to signature 4 and one matched to signature 13. In subset 4 (n=201) the matching signatures were 13, 6 and 4. The only signatures matched before clustering other than signature 4 had been signatures 1 and 2.

In the LUSC study for subset1 (n=195) we found signatures matching COSMIC signatures 13,4,7 and two matching 5 (unknown aetiology). For subset 2 we found matches to signatures 13,7 and 4. For subset 4 we found matches to signatures 13,7,4 and a different signature of unknown aetiology, signature 8. Signatures 13, 7 and 4 had been observed in the overall signature before, however signatures 5 and 8 had not.

*COAD*



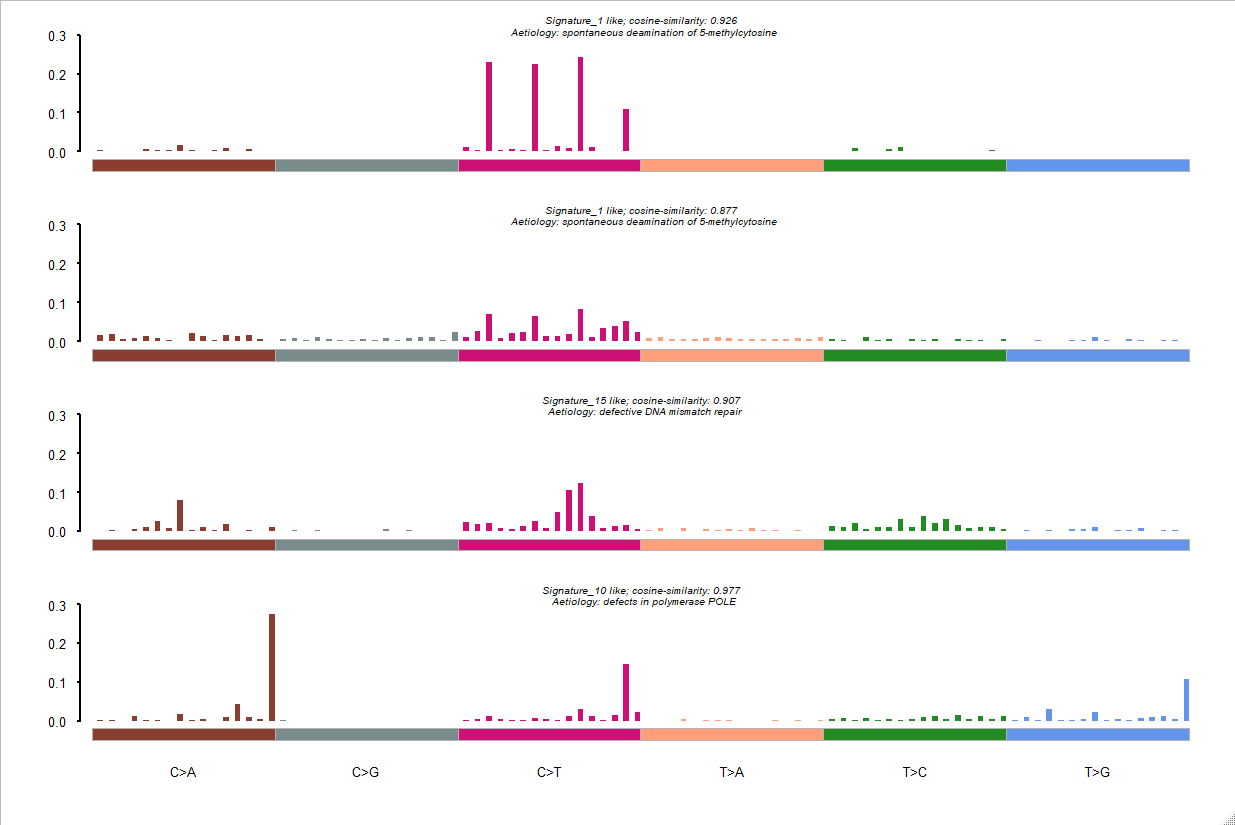
Figure 22a Mutation Signature for COAD subset1 

Figure 22b Mutation Signature for COAD subset2

*SKCM*



Figure 23a Mutation Signature for SKCM subset1

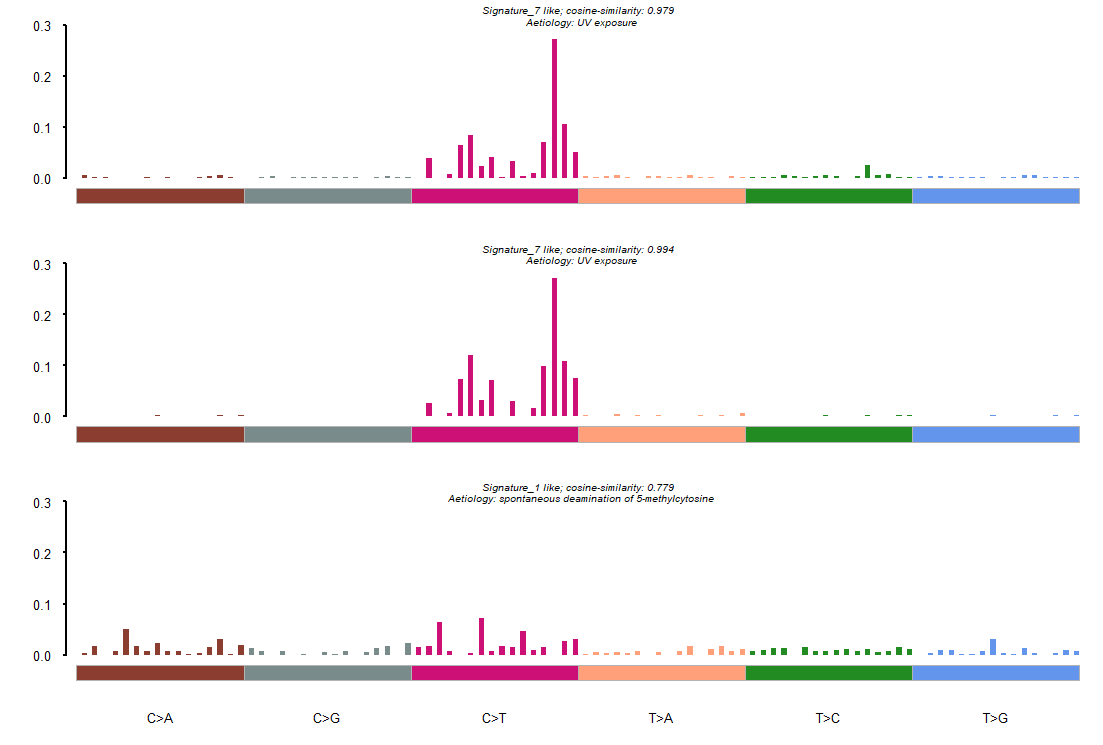


Figure 23b Mutation Signature for SKCM subset2

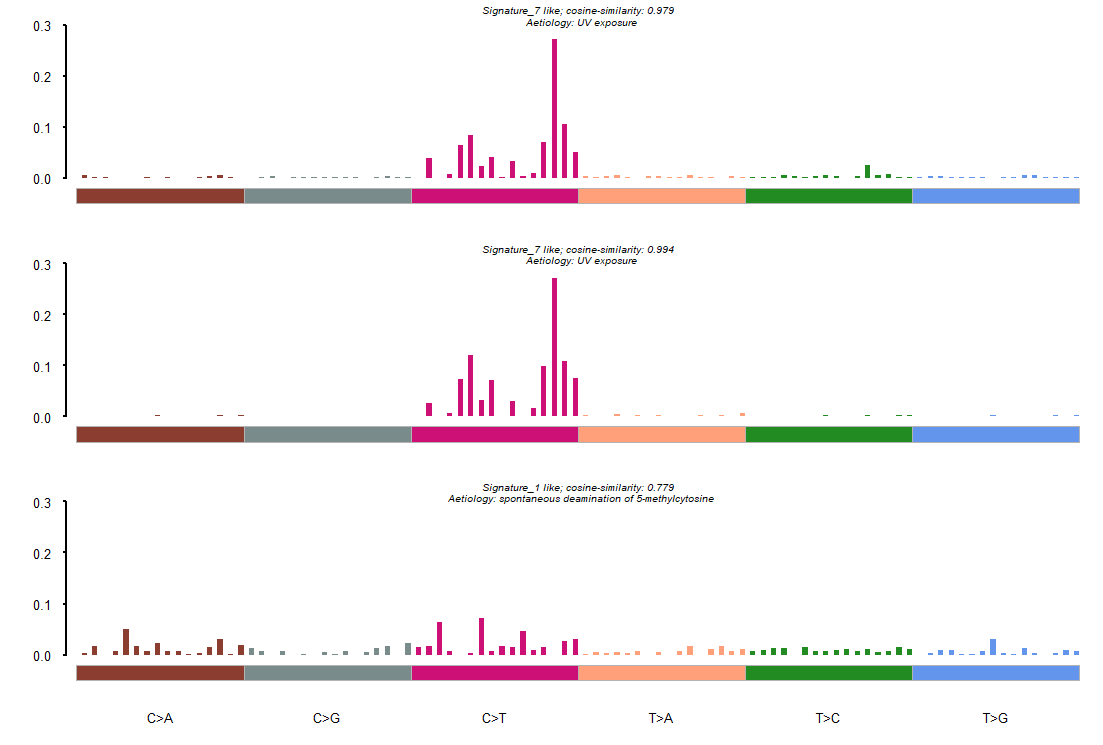


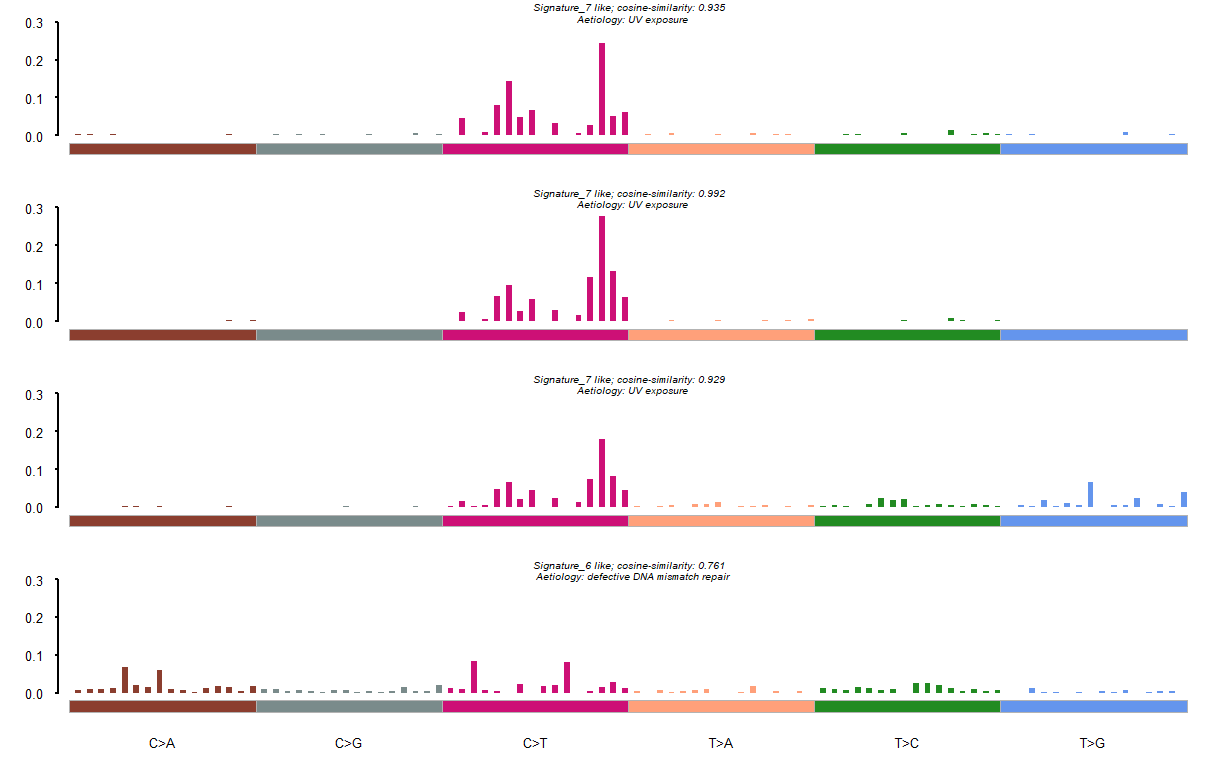
Figure 23c Mutation Signature for SKCM subset3

Figure 23d Mutation Signature for SKCM subset4

*LUAD*

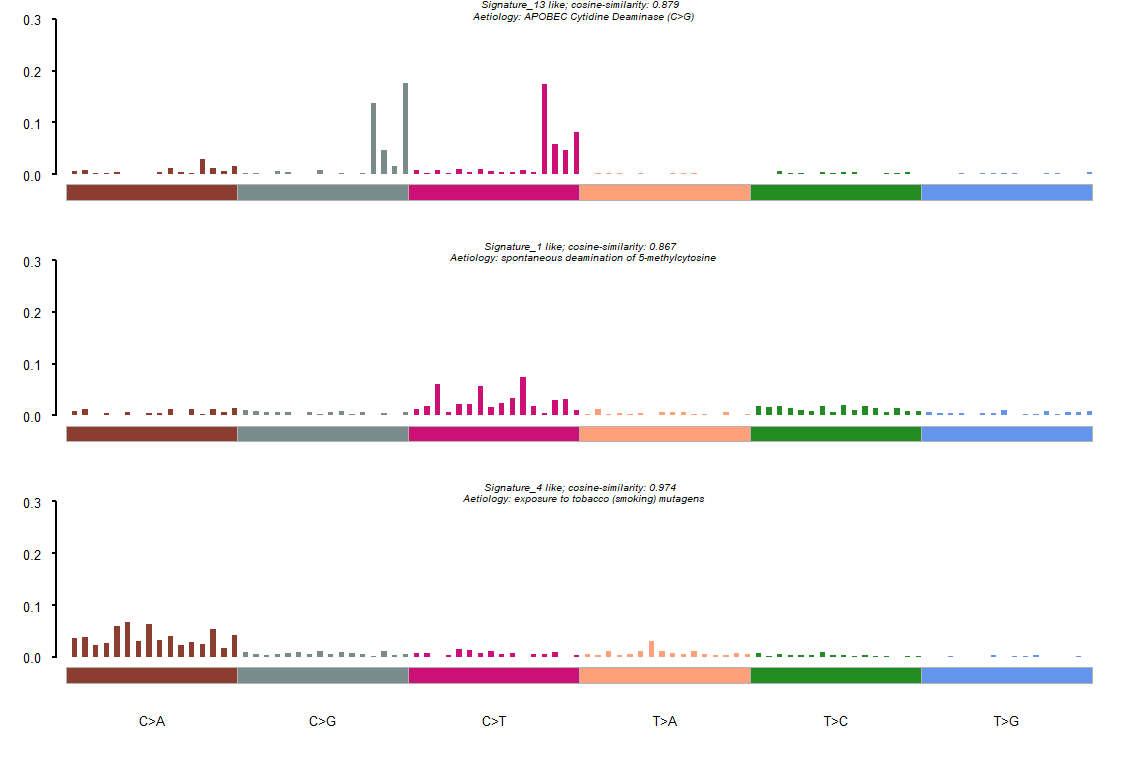


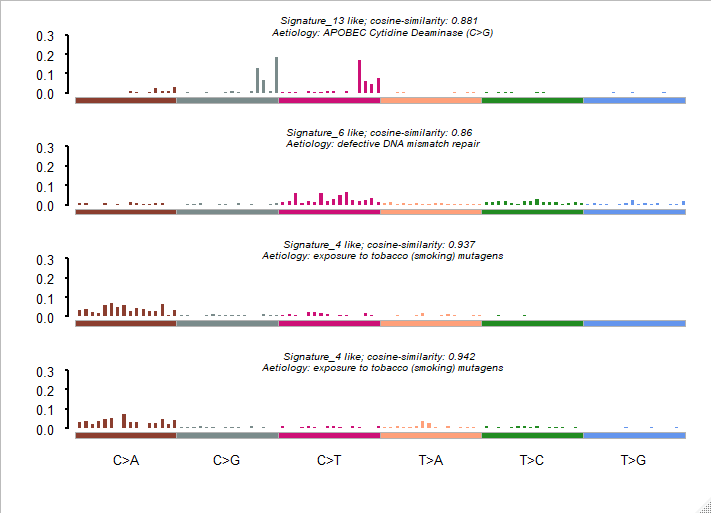
Figure 24a Mutation Signature for LUAD subset1

Figure 24b Mutation Signature for LUAD subset2



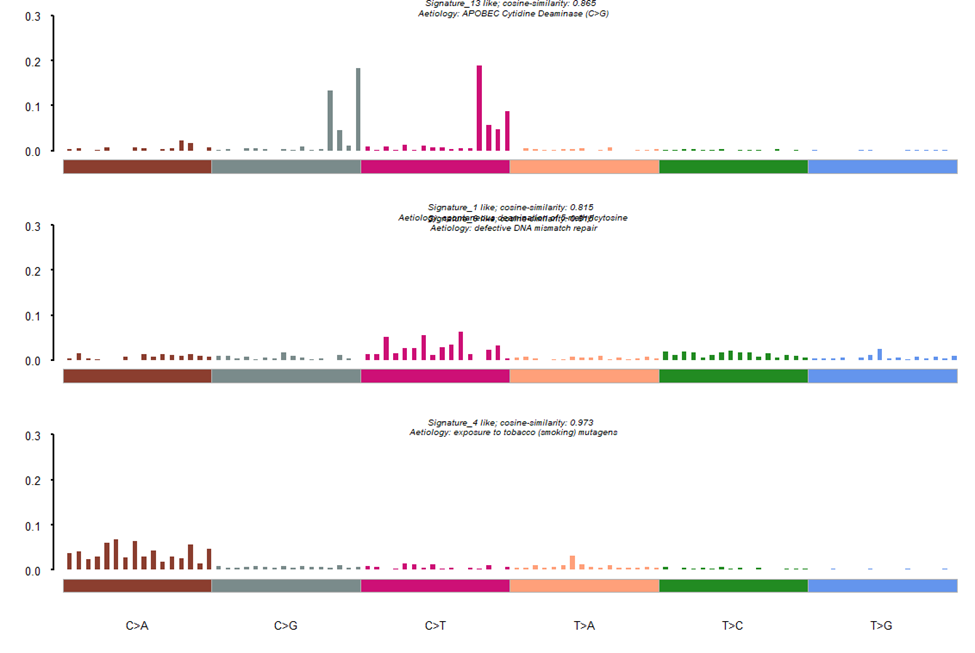
Figure 24c Mutation Signature for LUAD subset3

Figure 24d Mutation Signature for LUAD subset4

*LUSC*

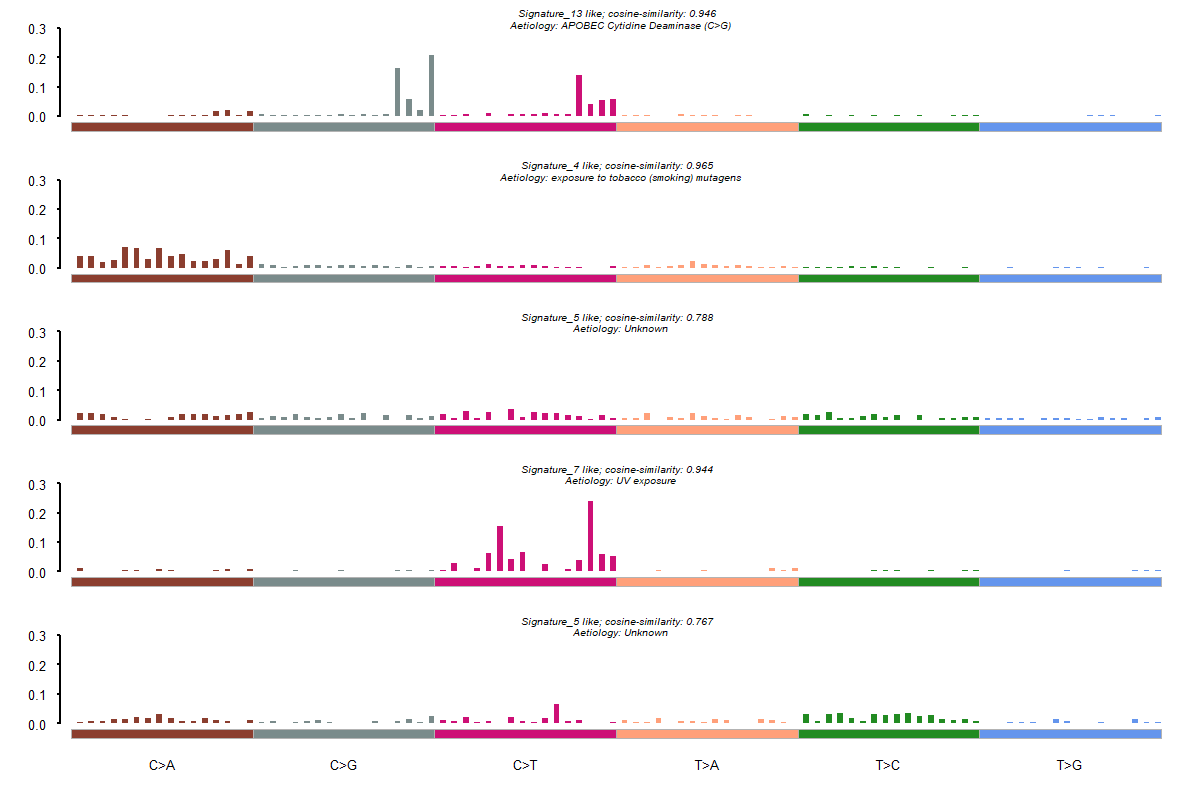


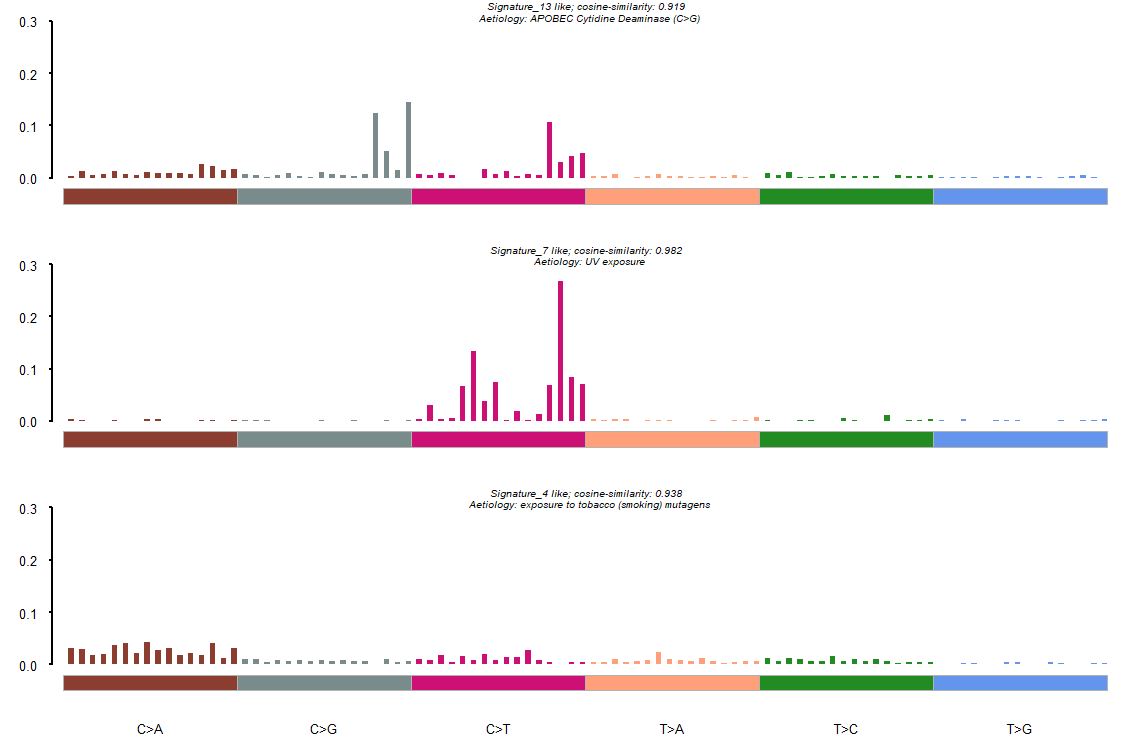
Figure 25a Mutation Signature for LUSC subset1

Figure 25b Mutation Signature for LUSC subset2



Figure 25c Mutation Signature for LUSC subset3

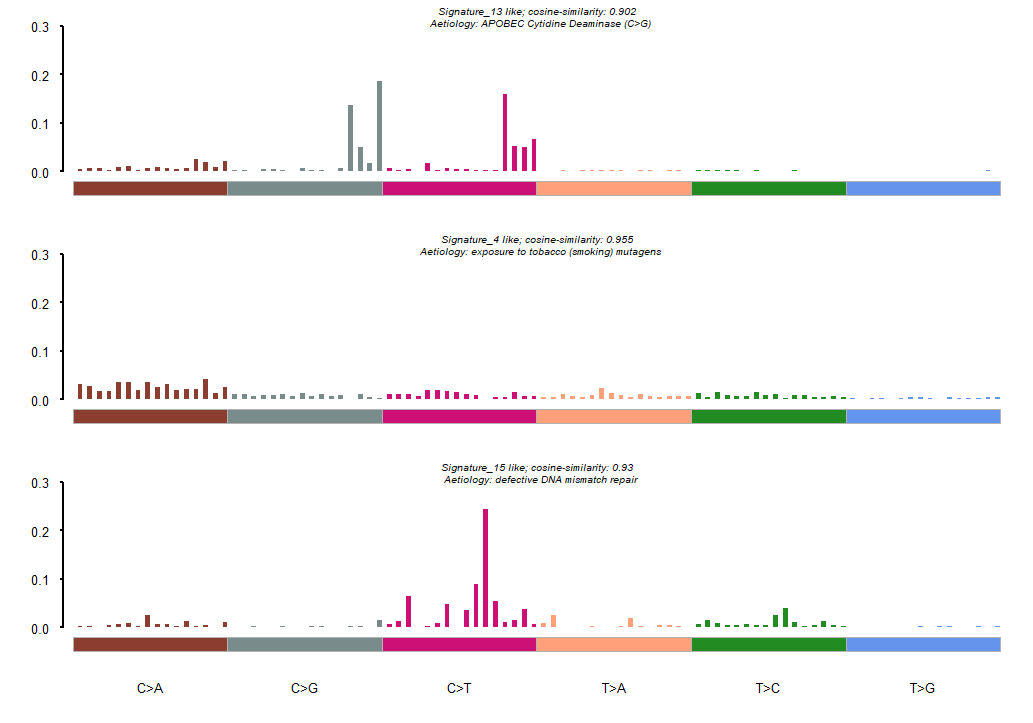


Figure 25d Mutation Signature for LUSC subset4

**Discussion**

As we had expected as has been established regarding mutation signatures, different underlying mechanisms drive differences in mutation signatures. We analyzed completely different cancer types that had different mutation burdens as well as differences in underlying mutation types. The overall signatures that we obtained using two similar methods, mSignatureDB and extractSignatures, resulted in some differences that we pointed out. However we cannot say which method is more accurate or precise based on a similar conclusion made by Alexandrov et al. who found differences after utilizing two different methods based on Sigprofiler and SigAnalyzer and provided the results for both and stated they did not know which is relevant or that it would be right to combine them (Alexandrov et al, 2018). However, there is no doubt that for our pipeline extract signature provides much greater utility since it allows us to base comparisons by dividing into subsets as well as providing multiple decomposed signatures rather than to combine it as one. This way we can judge and analyze the relevant contribution score of each patient to the different signatures. However, there is one shortcoming. This is that it requires more samples to extract signatures which is the reason we could not obtain signatures for COAD subsets 2 and 4 (which had a sample size of 3 and 5 respectively). The mSignatureDB however allows construction of signatures based on a single TCGA patient id. This could perhaps be overcome by modifying our cluster algorithm and our grouping criteria, for example in the clustering dendrogram it can be seen that COAD subset4 and 1 are very close in distance. So perhaps they could be combined, and the result could be compared to what we had for subset 1 by itself. It was also surprising and very interesting that we after clustering we observed signatures not present in the multiple decomposed signatures for example observations of signatures of multiple different signatures of unknown aetiologies found in certain subsets of all cancer types in the study including SKCM which was an extremely homogeneous cancer type with signature 7 (U.V exposure) being the most predominant type. Additionally, certain known signatures not found in the overall signature such as signature 13 and 6 were found in the subsets of LUAD.

This finding is of significance because it shows that there are mutation patterns being driven due to underlying mechanisms not yet fully understood. To attempt to answer these questions several steps can be taken building upon our current pipeline. As mentioned earlier can refine our clustering algorithms more carefully now that we have evidence that clustering based on a trinucleotide context can cause differences in signatures observed. Additionally since we had focused on somatic data we can try to correlate germline variations in DNA repair mechanisms that could contribute to differences at the somatic level we did see evidence of more *POLE* or MMR defects in certain subsets but what about other DDR genes; Perhaps such correlations could explain signatures of unknown aetiology. In addition to utilizing germline data we can also refine our filtering, such as a common practice removing germline variants that might be present TCGA samples (Mia et al 2019) (Sherry et al 2001). Additionally, Maftools documentation lists support for Annovar formats so it is also possible to analyze pathogenicity based on clinical designations from ClinVar (Landrum et al, 2014) and frequencies from the EXAC and GnomAD databases (lek et 2016) in addition to pathogenicity prediction tools being designed using machine learning using Annovar (Wang, 2010). To establish correlation and interactions of mutations occurring in subsets of interest we can use more comprehensive biological pathway analysis tools such as Cytoscape than what is available in maftools since it is limited to a handful of pathways, and genes in those pathways is limited also, with no way of looking at interactions among different pathways. Finally, this pipeline could simply be much improved and be efficient if the certain steps for which MATLAB code is being used, would be integrated into the main R script.

**Works cited**

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
11. Alexandrov, L. B., Ju, Y. S., Haase, K., Loo, P. V., Martincorena, I., Nik-Zainal, S., . . . Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. doi:10.1101/051417
12. Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–311
13. Petljak, Mia et al. “Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis.” *Cell* vol. 176,6 (2019): 1282-1294.e20. doi:10.1016/j.cell.2019.02.012
14. Landrum, Melissa J et al. “ClinVar: public archive of relationships among sequence variation and human phenotype.” *Nucleic acids research* vol. 42,Database issue (2014): D980-5. doi:10.1093/nar/gkt1113
15. Lek, Monkol et al. “Analysis of protein-coding genetic variation in 60,706 humans.” *Nature* vol. 536,7616 (2016): 285-91. doi:10.1038/nature19057
16. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing dataNucleic Acids Research, 38:e164, 2010