CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes

Masroor Bayati[1], Hamid R. Rabiee[1], Mehrdad Mehrbod[1], Fatemeh Vafaee[2], Diako Ebrahimi[3], Alistair R.R. Forrest[4], Hamid Alinejad-Rokny[4]

[1]ICT Innovation Center, Department of Computer Engineering, Sharif University of Technology, Tehran, 11365, IR

[2]School of Biotechnology and Biomolecular Sciences, University of New South Wales, UNSW Sydney, 2033, NSW, AU

[3]Department of Biochemistry, Molecular Biology and Biophysics; Masonic Cancer Center, Institute for Molecular Virology, University of Minnesota, MN, US

[4]Harry Perkins Institute of Medical Research, QEII Medical Centre and Centre for Medical Research, The University of Western Australia, Nedlands 6009, AU

Corresponding author's email: *h.alinejad@ieee.org*

ABSTRACT. Analyses of large somatic mutation datasets, using advanced computational algorithms, have revealed at least 30 independent mutational signatures in tumor samples. These studies have been instrumental in identification and quantification of responsible endogenous and exogenous molecular processes against cancer. The quantitative approach used to deconvolute mutational signatures is becoming an integral part of cancer research. Therefore, development of a stand-alone tool with a user-friendly graphical interface for analysis of cancer mutational signatures is necessary. In this manuscript, we introduce CANCERSIGN as an open access[1] bioinformatics tool that uses raw mutation data (BED files) as input, and generates 3-mer and 5-mer mutational signatures. Additionally, this tool enables users to perform clustering on tumor samples based on the raw mutation counts as well as using the proportion of mutational signatures in each sample. Using this tool, we analysed all the whole genome somatic mutation datasets of International Cancer Genome Consortium (ICGC) samples and identified a number of novel signatures.

INTRODUCTION

Aberrant somatic changes in DNA resulting from endogenous sources (e.g. APOBEC-induced mutagenesis and DNA repair defects) and exogenous factors (e.g. tobacco smoking and UV radiation) are the hallmark of cancer. These alternations in DNA may have different forms, ranging from gross chromosomal rearrangements to single base substitutions [1]. The whole genome sequencing of tumor cells has shown that the number of mutations varies from less than one hundred per genome to hundreds of thousands depending on the cancer type and patient. Moreover, the type of mutation and sequence context of many cancer mutations are not random. For instance, C-to-T mutation within the CG (a.k.a. CpG) dinucleotide is a prevalent mutation in cancer and as its abundance is proportional to the age of patient it is referred to as an "aging" signature [2]. Many cancers also have a large number of C-to-T and C-to-G mutations within TCA and TCT trinucleotides [3]. These mutations are attributed to the aberrant changes in the level and activity of APOBEC enzymes. The mutational landscape of

[1] HAR is corresponding author of the paper. HRR is second corresponding author of this paper.

each cancer genome is thus a cumulative result of multiple mutational signatures, each caused by a unique process such as methylation, APOBEC mediated changes, etc. [1].

Typically, signatures of mutational processes are determined by considering the trinucleotide context of single base substitutions. If all mutations are presented based on changes in the same DNA strand, there are 96 possible different types of mutations within trinucleotide motifs [4]. In 2013, Alexandrov *et al*. proposed a mathematical framework for analysing mutational signatures [4] based on these 96 types of mutations. Using a matrix factorization algorithm, the authors uncovered 30 independent mutational signatures. Details of these signatures including their prevalence in each cancer type and potential etiology are available at the COSMIC database (http://cancer.sanger.ac.uk/cosmic/signatures [5]).

The discovery of mutational signatures was a breakthrough in the field of cancer research. Therefore, the mathematical framework developed by Alexandrov *et al*. [4] is now routinely used to identify novel mutational signatures and to study the processes involved in different cancers and in different patients. To help the progress of this field, we have developed a computational tool, CANCERSIGN, which enables the users to easily apply a matrix factorization analysis to cancer mutation datasets and receive a complete set of mutational signatures. Compared to the previously developed packages in R [6-8], CANCERSIGN is unique in that it is a stand-alone package (i.e. it does not require additional software programming). Therefore, to use this tool, no programming skills are required. Additionally, it enables the users to perform *de novo* mutational signature analyses.

Application of CANSERSIGN is not limited to extracting mutational signatures based on nucleotides immediately flanking the mutated site (i.e. tri-nucleotide motifs). It allows the users to extend the analysis to two bases on each side of the mutated base (i.e. penta-nucleotides motifs). According to a recent study [9], taking larger sequence contexts into consideration provides a greater power to explain variability in genomic substitution probabilities. In addition, CANCERSIGN allows the user to select trinucleotides of interest, and determine their penta-nucleotide mutational signatures. Furthermore, it has a built-in clustering option to study the groupings of cancer samples based on the raw mutation counts and/or composition of mutational signatures. In this manuscript, we introduce CANCERSIGN and show the new mutational signatures obtained from a *de novo* analysis of whole genome ICGC dataset. This analysis was performed for each cancer type separately and resulted in 78 mutational signatures. Each of the obtained signatures were shown to be highly similar to at least one of the 30 signatures discovered previously [2], except 5 signatures that can be considered as novel.

DATA and METHODS

To develop CANCERSIGN and demonstrate its application, we used all the whole genome mutation datasets available at the International Cancer Genome Consortium (ICGC) data portal. A summary of the datasets used is shown in **Supplementary table 1.**

A schematic of CANCERSIGN features is given in **Supplementary figure 1**. This tool deciphers mutational signatures in somatic mutation datasets using a previously reported non-negative matrix factorization (NMF) model [4]. It only requires a dataset of mutations in a simple format described in the tool manual (https://github.com/masroor-bayati/cancer_analysis_tool and https://github.com/ictic-bioinformatics/CANCERSIGN).

CANCERSIGN then performs the requested analyses and outputs mutational signatures (3-mer and/or 5-mer formats) and clustering figures. The numerical outputs of the analyses are saved in a folder named "output" (refer to the CANCERSIGN manual). This data can be used to gain further insight into mutational processes in cancer. For example, the exposure matrix obtained from the NMF analysis, can be used to determine the prevalence of each mutational signature in each sample.

Estimation of mutational signatures

Each mutational signature is defined by a total of 96 types of mutations within tri-nucleotide motifs [4]. Assuming that m_g^i represents the number of mutations of type i recorded in the mutational catalogue of sample g, then the mutational catalogue of multiple samples is represented by a matrix *M* (*K* × *G*):

$$M = \begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix}$$

Where, *K* represents the number of mutation types (*K*=96 for tri-nucleotide signature analysis) and *G* represents the number of cancer samples. Therefore, the ith column of *M* has the number of each of 96 mutation types in the ith sample. Here, it is assumed that the mutational catalogue of each sample (i.e. each column of *M*) is the result of linear superposition of several mutational signatures [4], each of which corresponds to a particular mutational process. Assuming that the number of mutational signatures is *N*, we need to factorize *M* into two matrices *P* and *E* with sizes *K* × *N* and *N* × *G*, respectively:

$$M \approx P \times E \qquad \textbf{(Equation 1)}$$

The expanded representation of the above equation is given by:

$$\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix} \approx \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix} \times \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$$

Here, each column of *P* is interpreted as one mutational signature and each row of *E* represents the exposures of each mutational signature in the corresponding sample genome (i.e. the prevalence of each mutational signature in that sample). The process of estimating independent mutational signatures from a mutational catalogue matrix is done using a nonnegative matrix factorization (NMF) method. Note that the elements of the input catalogue matrix are nonnegative because they represent the number of mutations.

The algorithm used in this framework consists of several iterative steps. The overall procedure is as follows. Each iteration starts with sampling from matrix *M* and creating a new bootstrapped matrix $\breve{M}$. To do this, each column of *M* (mutation counts of one input cancer genome) is considered a discrete

probability distribution and is resampled to create the corresponding column of $\breve{M}$. Next, the bootstrapped matrix $\breve{M}$ is factorized into matrices *P* and *E* by applying the NMF algorithm. These steps will be repeated for a certain number of iterations (typically 600 iterations are enough to obtain a stable result). The resulting matrices obtained after each iteration are stored and grouped into two sets: *P* matrices and *E* matrices named S_P and S_E, respectively. The set of columns of matrices in S_P are then clustered into *N* groups using a variation of the $k-means$ clustering algorithm [10]. A similarity measure between two columns is calculated using cosine similarity and the centroids of the clusters are obtained by averaging over the members of each cluster. The cosine similarity between the vectors A and B is calculated as $A \cdot B / \sqrt{(A \cdot A) \times (B \cdot B)}$, where $A \cdot B$ represents the dot product of vectors A and B. Finally, the *N* centroids are combined to form the columns of a single matrix $\overline{P}$, which is the matrix of mutational signatures. Since each column of matrix *P* corresponds to one specific row of matrix *E*, the procedure of clustering columns of matrices in S_P naturally leads to clustering of rows of matrices in S_E. Therefore, the final exposure matrix $\overline{E}$ (prevalence of mutational signatures) is constructed in a manner similar to $\overline{P}$ after the clustering step. Finally, the results are evaluated by calculating the reproducibility of mutational signatures and the Frobenius reconstruction error of the NMF solution. The reproducibility is measured by calculating the average silhouette width of the result of clustering step, while the reconstruction error of the solution is quantified by the Frobenius norm of difference between matrix *M*, and its estimation obtained by the factorization algorithm, i.e., $\|M - \overline{P} \times \overline{E}\|_F^2$ [4]. The overall procedure described above is carried out for different values of *N*; the number of deciphered signatures. The range of values for *N* can be specified by the user and the optimum value of *N* can be selected based on the aforementioned evaluation measures [4]. An optimum *N* leads to high reproducibility and low reconstruction error. CANCERSIGN produces an evaluation plot similar to the one proposed by Alexandrov *et al*. [4].

CANCERSIGN is designed to carry out each iteration of the aforementioned algorithm, which consists of one bootstrap and a complete NMF decomposition, independent and in parallel with other iterations. This enables the user to distribute the whole task across CPU cores and speed up the procedure of extracting mutational signatures. This tool also allows the user to specify the convergence criteria, the number of NMF iterations, the number of bootstrapping iterations and the number of CPU cores for parallelization.

Penta-nucleotide mutational signatures

In addition to extracting mutational signatures based on the immediately flanking nucleotides around the mutated site (tri-nucleotides), CANCERSIGN enables the user to further investigate the underlying mutational process, by expanding the mutational context to two bases upstream and downstream of the mutated base (penta-nucleotides). For this purpose, the user can choose an arbitrary set of mutations in tri-nucleotide motifs (up to 10 combinations). For each of these selected motifs, our tool quantifies the number of mutations in the corresponding 16 penta-nucleotide motifs (e.g. when the motif C[G>T]A is selected, 16 penta-nucleotide motifs of the form NC[G>T]AN are considered where N is each of four nucleotides). The mutation counts of penta-nucleotides in all samples are then used as

the input to the NMF analysis, which produces a set of penta-nucleotide (or 5-mer) mutational signatures.

Clustering of samples

One important application of CANCERSIGN is clustering of tumors based on mutations within selected motifs (Matrix *M* in eq. 1) or based on the prevalence of mutational signatures (Exposure Matrix *E* in eq. 1). This feature allows the user to investigate between sample heterogeneity and to identify potential outlier samples. Additionally, this information can be used to investigate correlation with clinical data including treatment history and survival outcomes.

The clustering of samples is done using the k-means algorithm with Euclidian distance measure and random initial seeds via k-means++ method [11]. The algorithm repeats the clustering procedure several times to find a stable solution. The optimum value for the number of clusters is heuristically determined by measuring the *within cluster sum of squares* (WSS) against the number of clusters and finding the point after which this measure does not decrease dramatically (i.e. approximately stabilizes). To visualize the result of clustering, CANCERSIGN produces two types of plots. The first plot visualizes the clustering centroids, which are obtained by averaging over features of the members of each cluster. The second plot visualizes a projection of the samples in a 2D plane to show how samples within a cluster are similar to each other and different from samples in other clusters. To do this, the tool performs a principal component analysis (PCA) and generates a PC1-PC2 plot. The user can find the sample IDs associated with each cluster by using tabular files produced by the tool (see CANCERSIGN manual).

A snapshot of CANCERSIGN is provided in **Figure 1**.

We have applied our tool to the mutational data of various cancers (18 types- **Supplementary table 1**) and presented the results. The ICGC somatic mutations dataset (last updated June 2017) is used for this analysis. General information and statistics about the data can be found in **Supplementary table1**. Details about our analyses, including the tool parameters, are provided in supplementary information.

RESULTS AND DISCUSSION

The results of separate analyses of whole genome mutation data from each of 18 tumor types reported in the ICGC database are presented in **Figures 2A-2R**. The evaluation diagrams for each tumor type are presented in **Supplementary figures 2-19**. These diagrams were used to determine the total number of mutational signatures presented in Figure 2. For each tumor type, the numerical values of the deciphered 3-mer signatures and the prevalence of each signature in each tumor sample are obtained and provided in **Supplementary tables 2-19**. Totally, The CANCERSIGN program identified 78 signatures in 18 tumor-types (**Figures 2A-2R**). The majority of these signatures were highly similar (Cosine Similarity > 50) to one or more of the previously reported 30 signatures [2]. But five of these signatures had a Cosine similarity of <50, suggesting that they are possibly novel signatures that have not been identified previously (**Figure 3**). For example, **Figure 2A** indicates 10

mutational signatures deciphered from the mutational data of breast cancer samples by using CANCERSIGN. For each of these 10 signatures there is at least one previously reported signature which can be considered as a close match (**Figure 3**). **Figure 2B** shows the mutational signatures that CANCERSIGN has identified in blood cancer samples. For this cancer type, one signature (Blood Signature No. 1) does not resemble any of the previously reported signatures, i.e. it does not have >50% similarity to any of the previously reported signatures (**Figure 2B**). Details of the correlations between all 78 signatures discovered by CANCERSIGN, and the signatures reported by Alexandrov *et al*. [2] are given in **Supplementary table 20**.

The 30 mutational signatures reported previously, is based on the analysis of all tumor types together. These signatures are like a consensus and represent trends across all tumor types. However, this global mutational signature analysis of multiple cancer types might produce results that are biased toward signatures of the tumors with more samples in the dataset. Furthermore, pooling somatic mutations obtained from whole exome sequencing with that of whole genome data can be another source of bias that is not considered in the previous studies. These issues have been discussed in a recent review by Nik-Zainal and Morganella [12]. Nevertheless, our tool can be used to extract mutational signatures for all tumors together, or for each tumor separately. Usually, researchers used mixed dataset (whole genome + whole exome) to extract mutational signatures. CANCERSIGN is able to provide an option for user to extract mutational signatures from the whole genome, whole exome or pooled data. For example, **Figures 4 and 5** show the analysis of whole exome alone or whole exome + whole genome for breast and skin cancers, respectively. **Figures 2A** and **2N** also show the analyses on whole genome data for breast and skin cancer, respectively. The evaluation diagrams for these analyses are presented in **Supplementary figures 20-23.**

Identification of differences between tumor samples using clustering feature of CANCERSIGN can be informative, because this output can be used along with clinical data such as subtype, response to therapies, and overall survival to identify potential associations. To demonstrate the clustering feature of CANCERSIGN, we used breast cancer samples and performed both of the clustering options available in this tool. One option is clustering based on the proportion of mutational signatures. The results of this clustering is shown in **Figure 6**. Based on the estimated proportion of the 10 signatures of breast cancer, CANCERSIGN classified the samples into three distinct groups (**Figure 6**). The second option is clustering based on the proportions of 3-mer motifs. The selected motifs were T(C->G)A, T(C->G)T, T(C->T)A and T(C->T)T. The result of this clustering is shown in **Figure 7.**

Another feature of CANCERSIGN is to decipher mutational signatures based on mutations within specific 5-mer motifs. To demonstrate this feature, we selected two 3-mer motifs (TCA and TCT) and two mutation types (C>T and C>G) as an input (i.e. T(C>G)A, T(C>G)T, T(C>T)A and T(C>T)T). These mutations are attributed to APOBEC enzymes and are the most abundant mutations in many cancer types [13] [14]. The CANSERSIGN program quantified the number of mutations within all possible 5-mer motifs that have been obtained from these 3-mer motifs. As an example, for C>G mutation within the 3-mer TCA motif, the program correctly counts the number of mutations for all 16 possible 5-mers (i.e. NT(C>G)AN where N:A, C, G, or T). The analysis by CANCERSIGN, over all the

selected mutations in breast cancer samples, revealed four 5-mer mutational signatures, as shown in **Figure 8**. The evaluation diagram for this analysis is given in **Supplementary figure 24**. The numerical values of the resultant 5-mer signatures and their prevalence for each breast cancer sample are provided in **Supplementary table 21**.

Comparison with other tools

We have compared our tool with its well-known counterparts which have been developed for similar purposes: **SomaticSignatures** [6], **SigneR** [7] and **deconstructSigs** [8]. The deconstructSigs package is different from other tools as it is designed to determine the optimal linear combination of pre-defined mutational signatures that most accurately reconstructs the mutational profile of a single tumor sample [8], whereas others are designed to extract *de novo* mutational signatures from a cohort. **Table 1** compares the tools based on a set of features. Note that this comparison only considers relevant features of these tools rather than all capabilities. Please refer to the corresponding papers for more information [6-8].

The tools were applied to a simulated dataset of mutations. We constructed the simulated dataset as follows: 6 signatures were selected from the list of mutational signatures presented at the COSMIC database [5] (signatures 1, 2, 3, 17, 21 and 29) as the underlying mutational processes, producing our simulated mutational catalogue. The mutational load (total mutations) for each sample was determined by taking a random number from a Rayleigh distribution with parameter $\sigma = 8000$ (this distribution was used since we observed that the histogram of mutational loads for samples of a cancer is often unimodal and positively skewed, and the average of mutational loads across all samples and all cancer types is near 10'000 which is approximately equal to the mean of a Rayleigh distribution with scale parameter of 8000). In the next step, the vector of mutations for each sample was generated by a linear combination of the 6 signatures where the 6 coefficients were determined randomly (for each set of coefficients, we selected 6 random numbers between 0 to 1, then normalized them such that they sum up to 1, and finally multiply them with the amount of mutational load of the sample). The aforementioned algorithm were used to generate 400 samples to form the simulated mutational catalogue.

We tried to apply CANCERSIGN, SomaticSignatures and SigneR to the simulated dataset in an equal condition. However, due to its Bayesian framework, SigneR is not scalable to large mutational catalogues containing several hundreds of samples. Consequently, it is not feasible to test SigneR on the simulation dataset, and instead performed the comparison between CANCERSIGN and SomaticSignatures.

The parameters of the analysis were set as follows. The range of values of ***N*** (number of signatures to decipher) was set from 2 to 12. The maximum number of bootstraps for each ***N*** was set to 100 for CANCERSIGN, and the number of replicates (nReplicates) for SomaticSignatures was set to 20. With these settings, the tools consumed approximately the same amount of time (~45 minutes) to decipher mutational signatures from our simulated dataset (using a typical computer with four 1.7 GHz CPU cores and 8GB memory). According to **Figure 9**, both tools have correctly found ***N* = 6** as the optimal number of underlying mutational signatures (the knee point in the diagram of summary statistics of

SomaticSignatures [6], and the point with a high reproducibility and the lowest reconstruction error in the evaluation diagram of CANCERSIGN). The obtained mutational signatures are shown in **Figure 10**. By a simple visual comparison, we can conclude that both tools have deciphered almost identical signatures which are also identical to the original selected signatures (signatures 1, 2, 3, 17, 21 and 29 from the COSMIC database). This test shows that both tools can produce valid results when analysing catalogues of somatic signatures.

We also compared the tools in terms of time efficiency. Note that the number of replicates for SomaticSignatures is equivalent to the number of bootstraps for CANCERSIGN. It is observed that in the same amount of time (~45 minutes), the number of bootstraps performed by CANCERSIGN was more than the number of replicates in the process of SomaticSignatures (100 vs. 20). Thus, we can conclude that CANCERSIGN performs the analyses faster than SomaticSignatures.

CONCLUSION

Systematic analyses of mutations in tumor biopsies from a large number of cancer types have identified at least 30 mutational signatures, each pointing to distinct molecular mechanisms acting on cellular DNA. The computational method used to extract mutational signatures is becoming an integral part of cancer research. Therefore, in recent years, a number of studies have focused specifically on the development of bioinformatics tools for analysis of mutational signatures. In the present study, we reported the development of a stand-alone tool called CANCERSIGN, which does not require any programming skills to be used. This tool has several unique features. Firstly, it has been optimized to run parallel computational analysis in order to speed up *de novo* mutational signature extraction. Secondly, it enables the extraction of 5-mer mutational signature profiles in addition to the commonly used 3-mer signature patterns. Thirdly, using the clustering features of this tool, differences between patients and/or tumor types can be investigated using all mutation profiles and/or subsets of selected mutations. In addition, using this tool, we identified a number of novel signatures. Overall, CANCERSIGN is a multi-functional and user-friendly computational tool for accurate and quick analysis of mutation signatures and clustering of tumor samples. This tool is a stand-alone package, is freely available, and does not require any specific computational skills to run.

AUTHOR CONTRIBUTION

HAR designed the study. MB, HAR, HRR, DE, ARRF wrote and revised the manuscript. HAR, FV provided the data and revised the manuscript. MB, HRR and HAR carried out tool implementation. MB generated all figures and tables. MB, HAR and HRR and MM performed statistical analyses. All authors have read and approved the final version of the paper.

SUPPLEMENTARY DATA

The CANCERSIGN is accessible at https://github.com/ictic-bioinformatics/CANCERSIGN and https://github.com/masroor-bayati/cancer_analysis_tool.

ACKNOWLEDGMENT

We would like to kindly thank Ludmil Alexandrov (Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California, San Diego) for useful discussion on the signatures and his insightful comments on the paper and CANCERSIGN. H.A.R. is a research associate in the Harry Perkins Institute of Medical Research. A.R.R.F. is supported by a Senior Cancer Research Fellowship from the Cancer Research Trust. This work was supported by funds raised by the MACA Ride to Conquer Cancer. This work has also been supported by Iran National Science Foundation (INSF) Grant No. 96006077.

DATA AVAILIBILITY

All supplementary files are available on the journal website.

CONFLICT OF INTEREST

The authors declare no competing financial interests.

REFERENCES

1. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-724.
2. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-421.
3. Roberts, S.A., et al., *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.* Nat Genet, 2013. **45**(9): p. 970-976.
4. Alexandrov, Ludmil B., et al., *Deciphering Signatures of Mutational Processes Operative in Human Cancer.* Cell Reports. **3**(1): p. 246-259.
5. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Research, 2015. **43**(D1): p. D805-D811.
6. Gehring, J.S., et al., *SomaticSignatures: inferring mutational signatures from single-nucleotide variants.* Bioinformatics, 2015. **31**(22): p. 3673-3675.
7. Rosales, R.A., et al., *signeR: an empirical Bayesian approach to mutational signature discovery.* Bioinformatics, 2017. **33**(1): p. 8-16.
8. Rosenthal, R., et al., *DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution.* Genome Biol, 2016. **17**: p. 31.

9. Aggarwala, V. and B.F. Voight, *An expanded sequence context model broadly explains variability in polymorphism levels across the human genome.* Nat Genet, 2016. **48**(4): p. 349-355.
10. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. **31**(8): p. 651-666.
11. Arthur, D. and S. Vassilvitskii, *k-means++: the advantages of careful seeding*, in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007, Society for Industrial and Applied Mathematics: New Orleans, Louisiana. p. 1027-1035.
12. Nik-Zainal, S. and S. Morganella, *Mutational Signatures in Breast Cancer: The Problem at the DNA Level.* Clinical Cancer Research, 2017. **23**(11): p. 2617-2629.
13. Seplyarskiy, V.B., M.A. Andrianova, and G.A. Bazykin, *APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context.* Genome Res, 2017. **27**(2): p. 175-184.
14. Burns, M.B., et al., *APOBEC3B is an enzymatic source of mutation in breast cancer.* Nature, 2013. **494**(7437): p. 366-370.
15. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**: p. 47.

FIGURES AND TABLES LEGENDS

Table 1 | A comparison between packages for analysis of mutational catalogues.

	Usage	**Method of selecting optimum number of extracted signatures**	**Option of extracting 5-mer mutational signatures**	**Option of clustering the samples based on motifs or signatures**	**Graphical User Interface**	**Implementation**
CANSERSIGN	Deciphering *de novo* mutational signatures	Specified by user by inspecting the diagrams of reproducibility and reconstruction error	Yes	Yes	Yes	R
SomaticSignature	Deciphering *de novo* mutational signatures	Specified by user by inspecting the diagrams of residual sum of squares and explained variance	No	No	No	R

SigneR	Deciphering *de novo* mutational signatures	Automatically specified by empirical Bayesian treatment to the Poissonian NMF model	No	No	No	R
deconstructSigs	Obtain the optimal linear combination of pre-defined mutational signatures	–	No	No	No	R

Figure 1 | Tool functionality map.

This diagram abstracts the functional features of our tool. The user can either choose to extract mutational signatures in the conventional format which are the spectrums over 96 mutation types (3-mer motifs) or extract 5-mer mutational signatures. The box indicating "NMF algorithm" consists of several steps. Its main steps are bootstrapping from mutational catalogue matrix, NMF iteration and clustering the results obtained from the bootstrapped data. In the clustering part, the user can choose different sets of features to cluster the samples. These feature sets are based on either mutational motifs or mutational signatures. In the former case, counts or proportions of 3-mer or 5-mer mutational motifs can be selected, and in the latter case, contributions of 3-mer or 5-mer mutational signatures can be chosen for clustering features. Finally, in the simulation part, there are two options to simulate the mutational catalogues and in either case, it produces simulated signatures to examine the statistical significance of deciphered signatures from the input data.

Figure 2 | Deciphering whole genome 3-mer mutational signatures for various cancer types.

The 3-mer mutational signatures extracted from whole-genome samples of 18 cancer types. Each cancer type is analysed separately. **(A)** Breast cancer 3-mer signatures. **(B)** Brain cancer 3-mer signatures. **(C)** Esophagus cancer 3-mer signatures … **(R)** Nervous system cancer 3-mer signatures.

Figure 3 | Correlation of 78 signatures identified by CANCERSIGN with 30 previously reported signatures by Nik-Zainal et al. [15].

Here, each cell indicates percentage of similarity between signatures identified by our tool and previously reported signatures. Those signatures with less than 50% similarity with previously reported signatures have been highlighted.

Figure 4 | Mutational signatures deciphered from the whole exome and pooled (whole genome and whole exome) data of breast cancer.

(A) For the pooled data. **(B)** For the whole exome data.

Figure 5 | Mutational signatures deciphered from the whole exome and pooled (whole genome and whole exome) data of skin cancer.

(A) For the pooled data. **(B)** For the whole exome data.

Figure 6 | Clustering the samples of breast cancer based on contributions of 3-mer mutational signatures.

This clustering is applied to whole-genome samples of breast cancer. The features of samples for this clustering are the contributions of 3-mer mutational signatures to each sample. **(A)** Average contributions of signatures within each cluster. **(B)** The result of principal component analysis of all samples in the space of clustering features. The axes of the plot are the first two principal components.

Figure 7 | Clustering the samples of breast cancer based on proportions of 3-mer mutational motifs.

This clustering was applied to whole-genome samples of breast cancer. The features of samples for this clustering are proportions of the 3-mer mutational motifs recorded in each sample. **(A)** Average counts of motifs within each cluster. **(B)** The result of principal component analysis of all samples in the space of clustering features. The axes of the plot are the first two principal components.

Figure 8 | Deciphering whole genome 5-mer mutational signatures for breast cancer.

The 5-mer mutational signatures extracted from whole-genome samples of breast cancer. Here, we have chosen four 3-mer motifs, namely T(C>G)A, T(C>G)T, T(C > T)A and T(C > T)T, to expand to 5-mer motifs and extracted the corresponding 5-mer signatures.

Figure 9 | Evaluation diagrams produced by SomaticSignatures and CANCERSIGN packages when applied to the sample dataset (for tool comparison).

(A) For SomaticSignatures package. **(B)** For CANCERSIGN package.

Figure 10 | Mutational signatures deciphered by SomaticSignatures and CANCERSIGN packages when applied to the sample dataset (for tool comparison).

(A) For the SomaticSignatures package. **(B)** For the CANCERSIGN package.

Figure 1

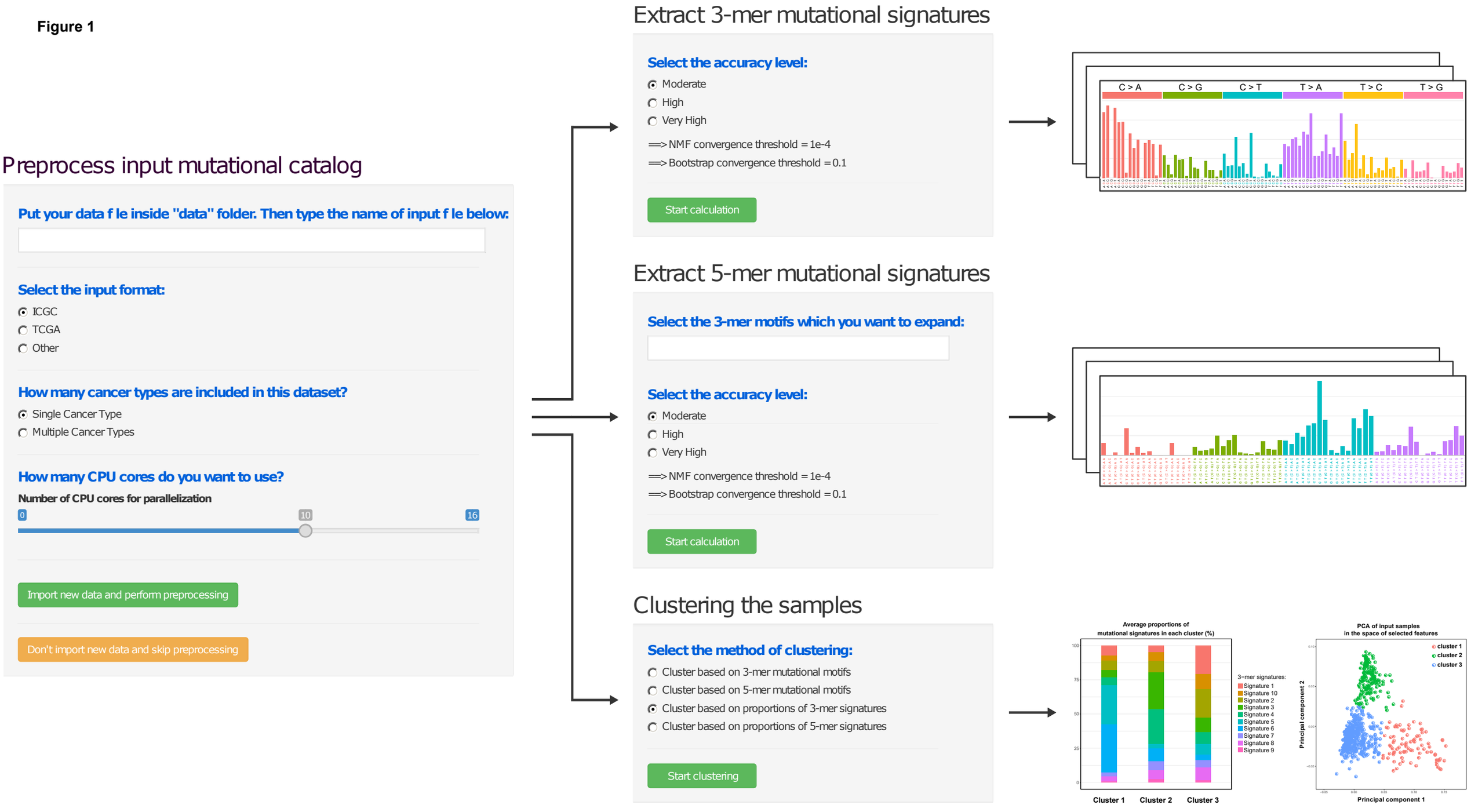

Preprocess input mutational catalog
Put your data f le inside "data" folder. Then type the name of input f le below:
Select the input format:
ICGC
TCGA
Other
How many cancer types are included in this dataset?
Single Cancer Type
Multiple Cancer Types
How many CPU cores do you want to use?
Number of CPU cores for parallelization
0
10
16
Import new data and perform preprocessing
Don't import new data and skip preprocessing
Extract 3-mer mutational signatures
Select the accuracy level:
Moderate
High
Very High
==> NMF convergence threshold = 1e-4
==> Bootstrap convergence threshold = 0.1
Start calculation
C > A
C > G
C > T
T > A
T > C
T > G
Extract 5-mer mutational signatures
Select the 3-mer motifs which you want to expand:
Select the accuracy level:
Moderate
High
Very High
==> NMF convergence threshold = 1e-4
==> Bootstrap convergence threshold = 0.1
Start calculation
Clustering the samples
Select the method of clustering:
Cluster based on 3-mer mutational motifs
Cluster based on 5-mer mutational motifs
Cluster based on proportions of 3-mer signatures
Cluster based on proportions of 5-mer signatures
Start clustering
Average proportions of mutational signatures in each cluster (%)
3−mer signatures:
Signature 1
Signature 10
Signature 2
Signature 3
Signature 4
Signature 5
Signature 6
Signature 7
Signature 8
Signature 9
Cluster 1
Cluster 2
Cluster 3
PCA of input samples in the space of selected features
cluster 1
cluster 2
cluster 3
Principal component 1
Principal component 2

Figure 2-1

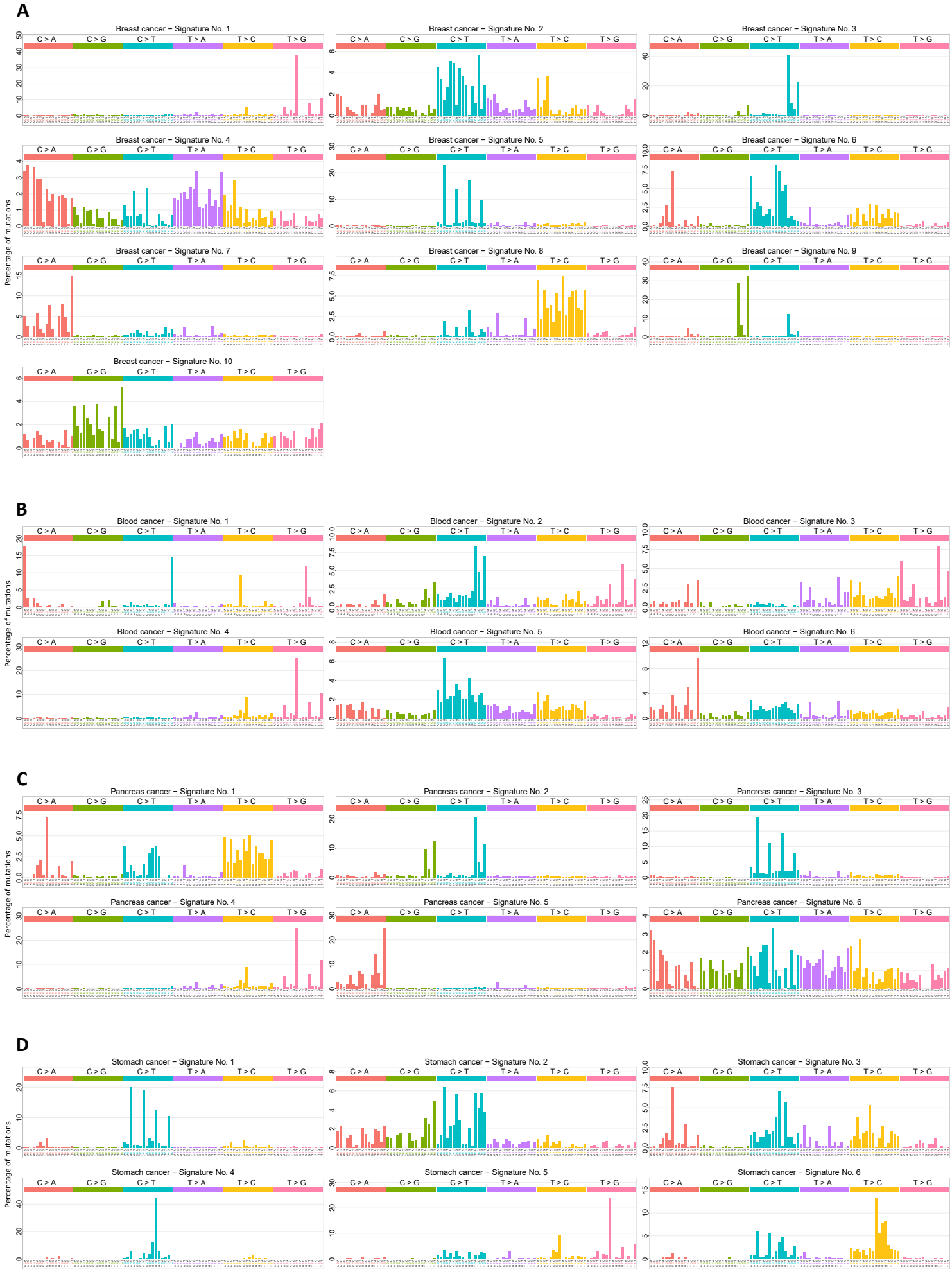
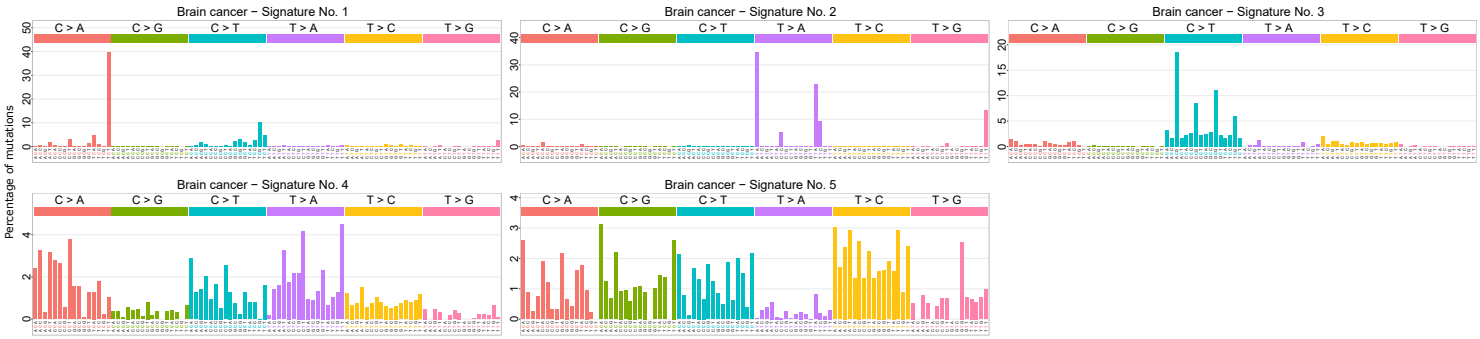
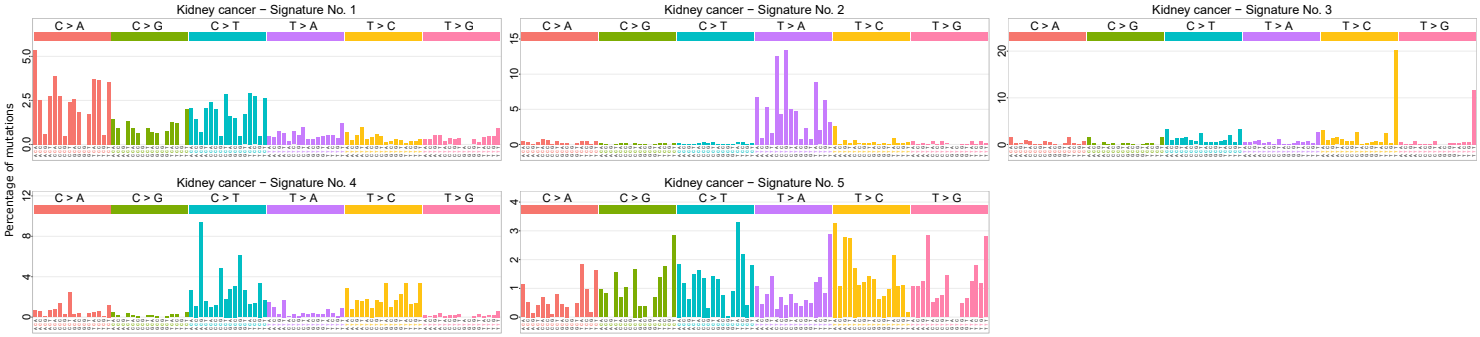

A
Breast cancer − Signature No. 1
Breast cancer − Signature No. 2
Breast cancer − Signature No. 3
Breast cancer − Signature No. 4
Breast cancer − Signature No. 5
Breast cancer − Signature No. 6
Breast cancer − Signature No. 7
Breast cancer − Signature No. 8
Breast cancer − Signature No. 9
Breast cancer − Signature No. 10
B
Blood cancer − Signature No. 1
Blood cancer − Signature No. 2
Blood cancer − Signature No. 3
Blood cancer − Signature No. 4
Blood cancer − Signature No. 5
Blood cancer − Signature No. 6
C
Pancreas cancer − Signature No. 1
Pancreas cancer − Signature No. 2
Pancreas cancer − Signature No. 3
Pancreas cancer − Signature No. 4
Pancreas cancer − Signature No. 5
Pancreas cancer − Signature No. 6
D
Stomach cancer − Signature No. 1
Stomach cancer − Signature No. 2
Stomach cancer − Signature No. 3
Stomach cancer − Signature No. 4
Stomach cancer − Signature No. 5
Stomach cancer − Signature No. 6
C > A
C > G
C > T
T > A
T > C
T > G
Percentage of mutations

Figure 2-2

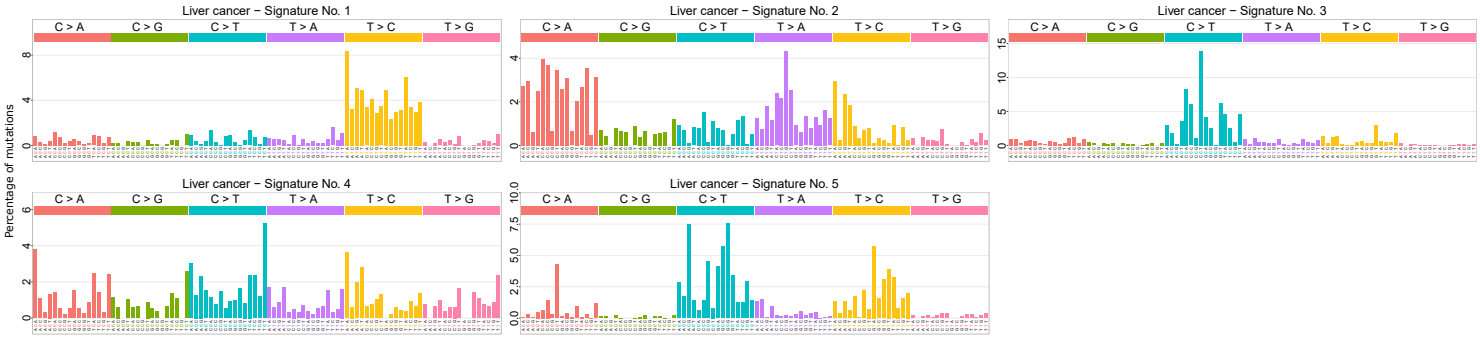
E



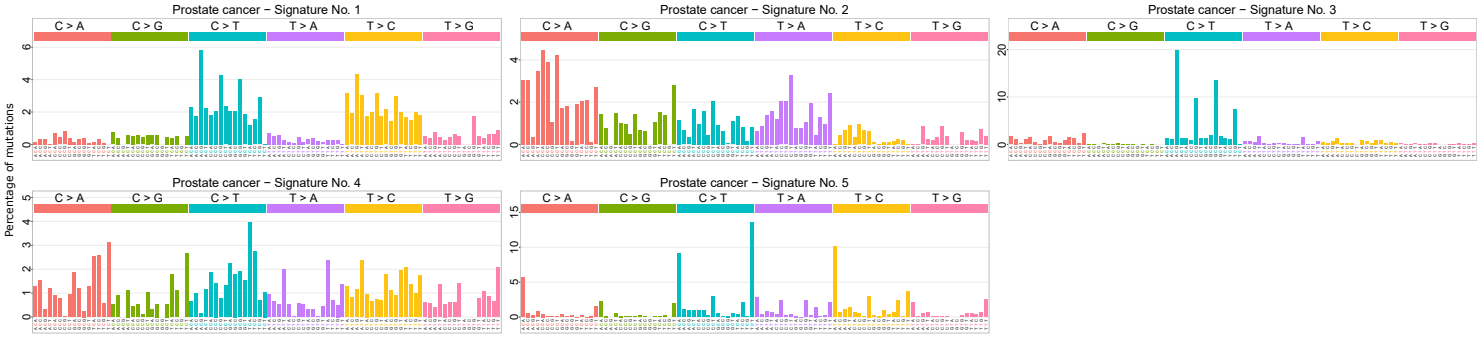
F



G



H



I

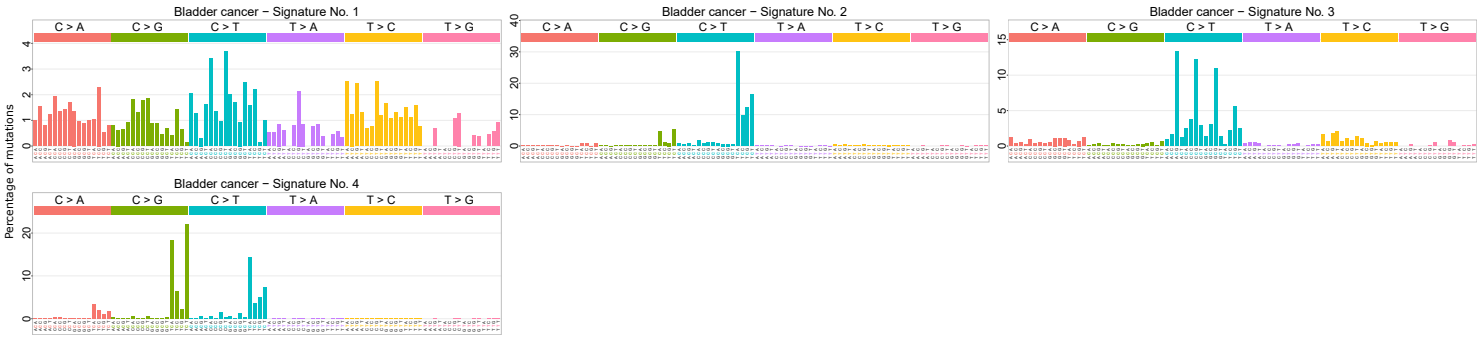
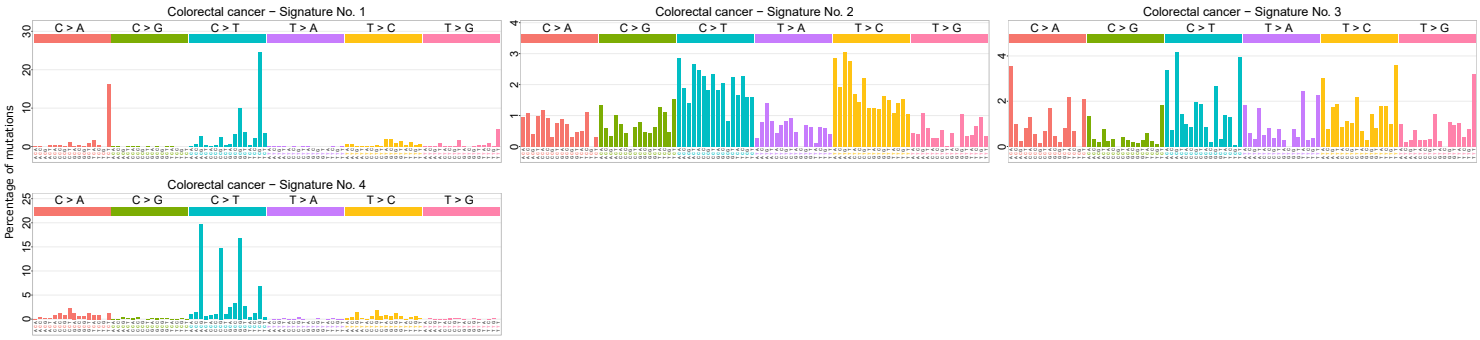
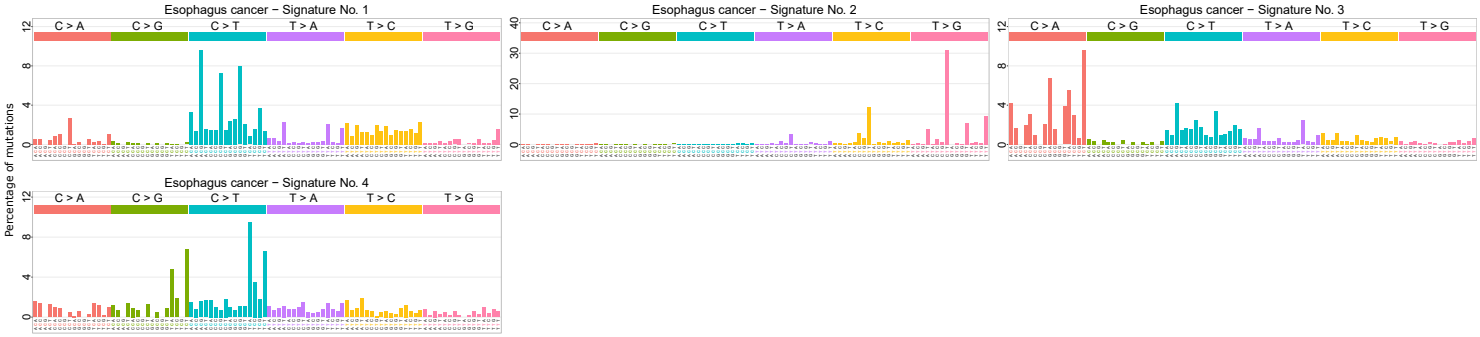


Figure 2-3

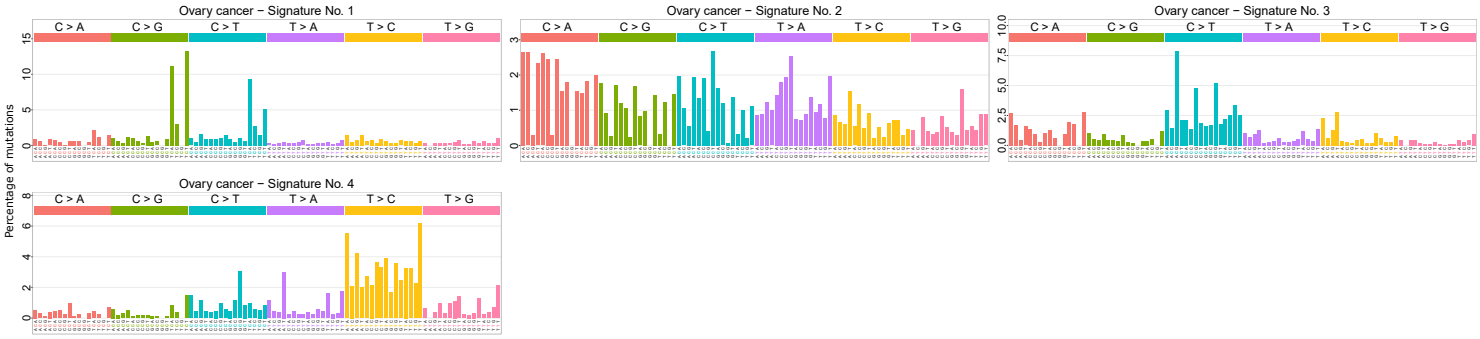
J



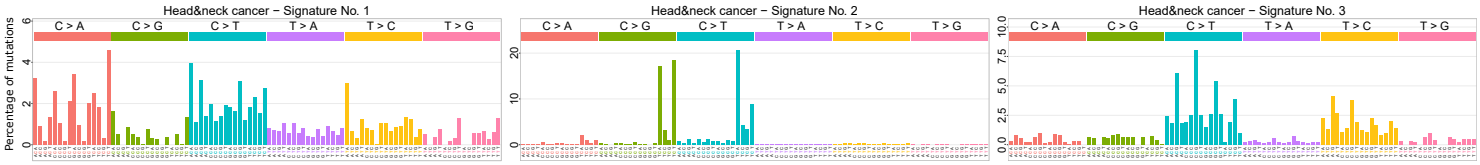
K



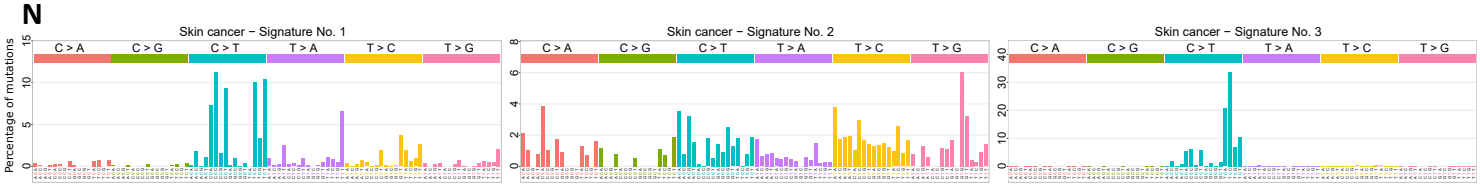
L



M



N



O

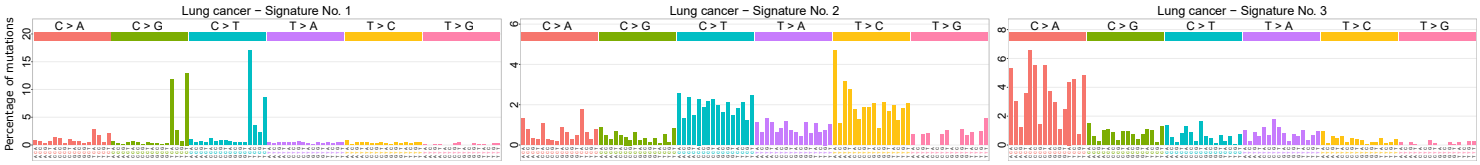
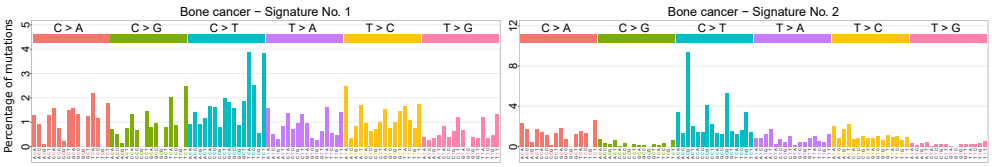
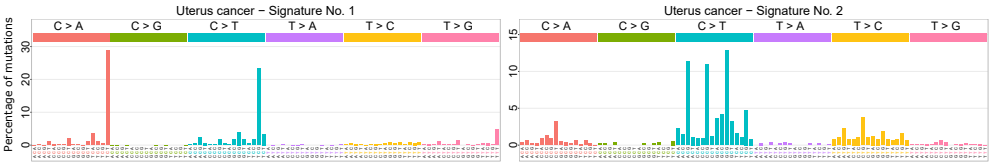


Figure 2-4

P



Q



R

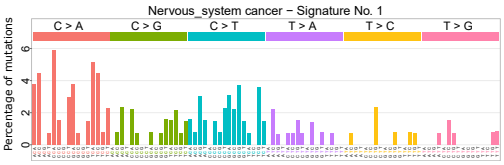


Figure 3

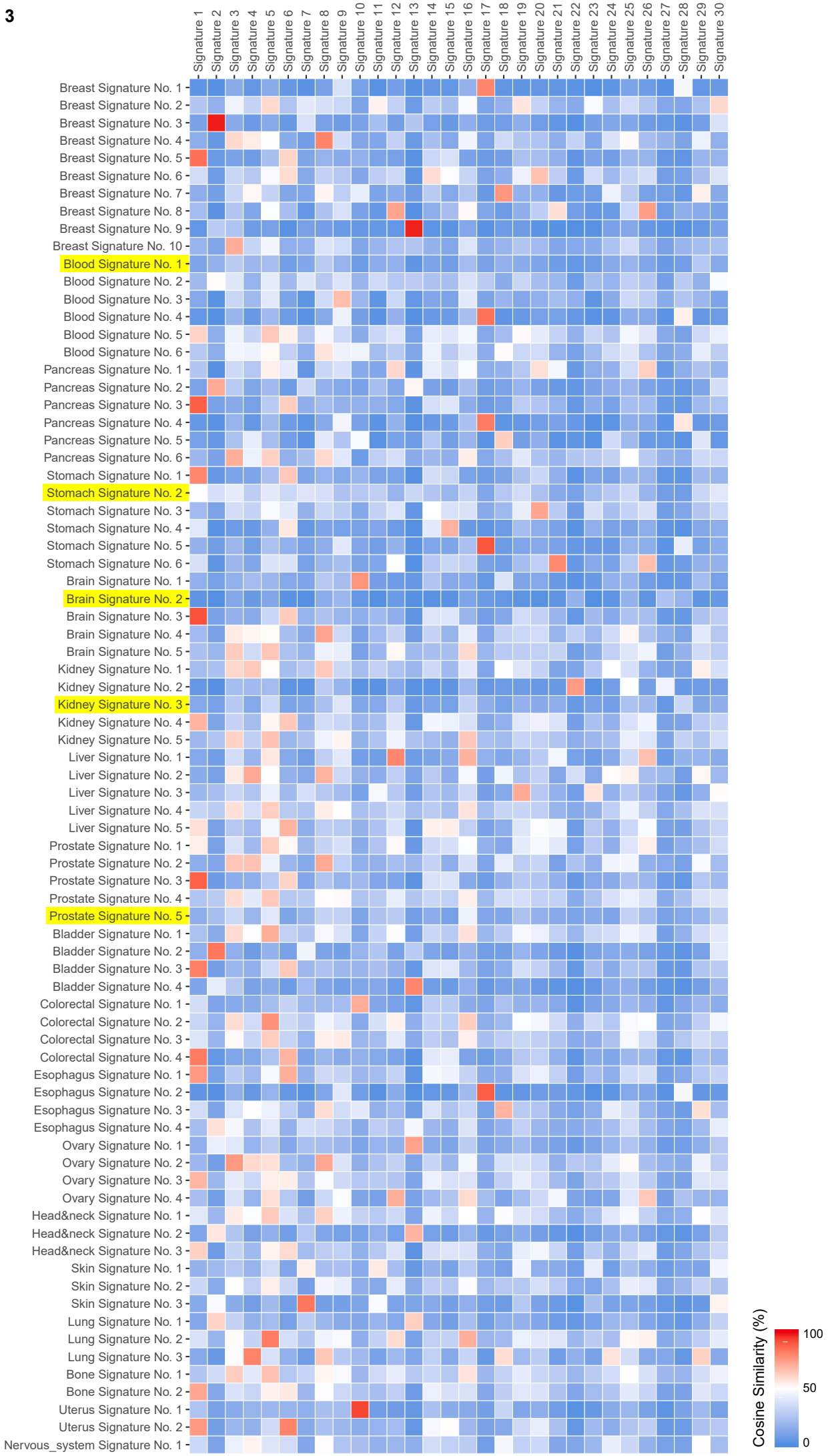

Signature 1
Signature 2
Signature 3
Signature 4
Signature 5
Signature 6
Signature 7
Signature 8
Signature 9
Signature 10
Signature 11
Signature 12
Signature 13
Signature 14
Signature 15
Signature 16
Signature 17
Signature 18
Signature 19
Signature 20
Signature 21
Signature 22
Signature 23
Signature 24
Signature 25
Signature 26
Signature 27
Signature 28
Signature 29
Signature 30
Breast Signature No. 1
Breast Signature No. 2
Breast Signature No. 3
Breast Signature No. 4
Breast Signature No. 5
Breast Signature No. 6
Breast Signature No. 7
Breast Signature No. 8
Breast Signature No. 9
Breast Signature No. 10
Blood Signature No. 1
Blood Signature No. 2
Blood Signature No. 3
Blood Signature No. 4
Blood Signature No. 5
Blood Signature No. 6
Pancreas Signature No. 1
Pancreas Signature No. 2
Pancreas Signature No. 3
Pancreas Signature No. 4
Pancreas Signature No. 5
Pancreas Signature No. 6
Stomach Signature No. 1
Stomach Signature No. 2
Stomach Signature No. 3
Stomach Signature No. 4
Stomach Signature No. 5
Stomach Signature No. 6
Brain Signature No. 1
Brain Signature No. 2
Brain Signature No. 3
Brain Signature No. 4
Brain Signature No. 5
Kidney Signature No. 1
Kidney Signature No. 2
Kidney Signature No. 3
Kidney Signature No. 4
Kidney Signature No. 5
Liver Signature No. 1
Liver Signature No. 2
Liver Signature No. 3
Liver Signature No. 4
Liver Signature No. 5
Prostate Signature No. 1
Prostate Signature No. 2
Prostate Signature No. 3
Prostate Signature No. 4
Prostate Signature No. 5
Bladder Signature No. 1
Bladder Signature No. 2
Bladder Signature No. 3
Bladder Signature No. 4
Colorectal Signature No. 1
Colorectal Signature No. 2
Colorectal Signature No. 3
Colorectal Signature No. 4
Esophagus Signature No. 1
Esophagus Signature No. 2
Esophagus Signature No. 3
Esophagus Signature No. 4
Ovary Signature No. 1
Ovary Signature No. 2
Ovary Signature No. 3
Ovary Signature No. 4
Head&neck Signature No. 1
Head&neck Signature No. 2
Head&neck Signature No. 3
Skin Signature No. 1
Skin Signature No. 2
Skin Signature No. 3
Lung Signature No. 1
Lung Signature No. 2
Lung Signature No. 3
Bone Signature No. 1
Bone Signature No. 2
Uterus Signature No. 1
Uterus Signature No. 2
Nervous_system Signature No. 1
Cosine Similarity (%)
100
50
0

Figure 4

A

Breast cancer (MIXED) − Signature No. 1
Breast cancer (MIXED) − Signature No. 2
Breast cancer (MIXED) − Signature No. 3
Breast cancer (MIXED) − Signature No. 4
Breast cancer (MIXED) − Signature No. 5
Breast cancer (MIXED) − Signature No. 6
Breast cancer (MIXED) − Signature No. 7
Breast cancer (MIXED) − Signature No. 8
Breast cancer (MIXED) − Signature No. 9
Breast cancer (MIXED) − Signature No. 10
C > A
C > G
C > T
T > A
T > C
T > G
Percentage of mutations

B

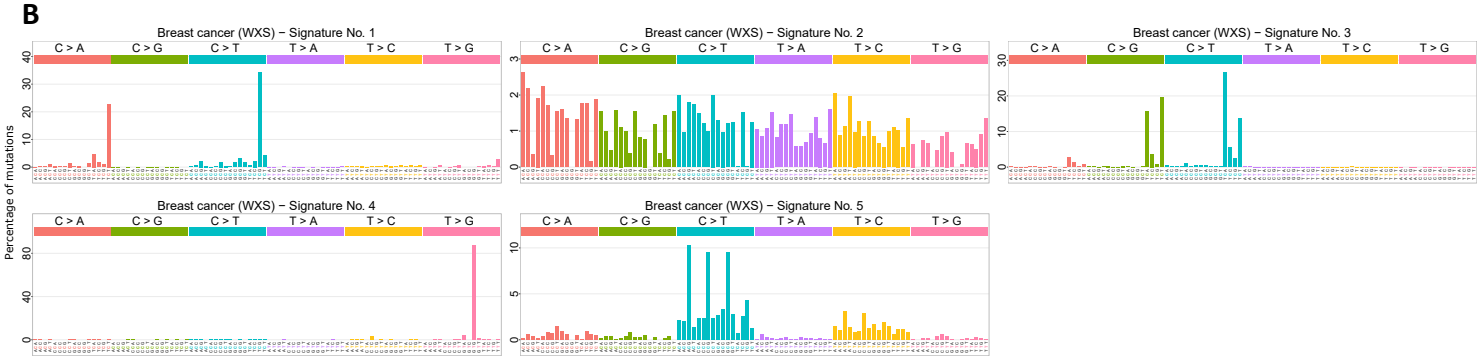
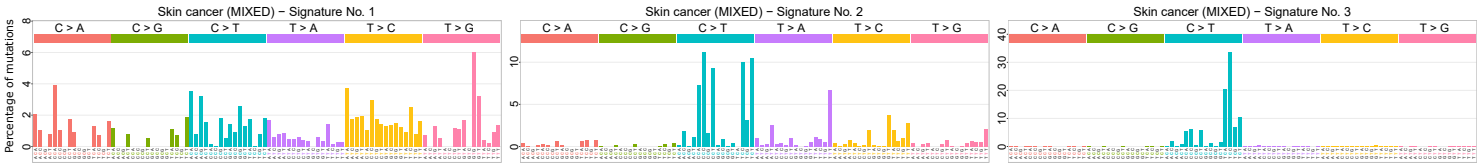

Breast cancer (WXS) − Signature No. 1
Breast cancer (WXS) − Signature No. 2
Breast cancer (WXS) − Signature No. 3
Breast cancer (WXS) − Signature No. 4
Breast cancer (WXS) − Signature No. 5
C > A
C > G
C > T
T > A
T > C
T > G
Percentage of mutations

Figure 5

A


Skin cancer (MIXED) − Signature No. 1
Skin cancer (MIXED) − Signature No. 2
Skin cancer (MIXED) − Signature No. 3
C > A
C > G
C > T
T > A
T > C
T > G
Percentage of mutations

B

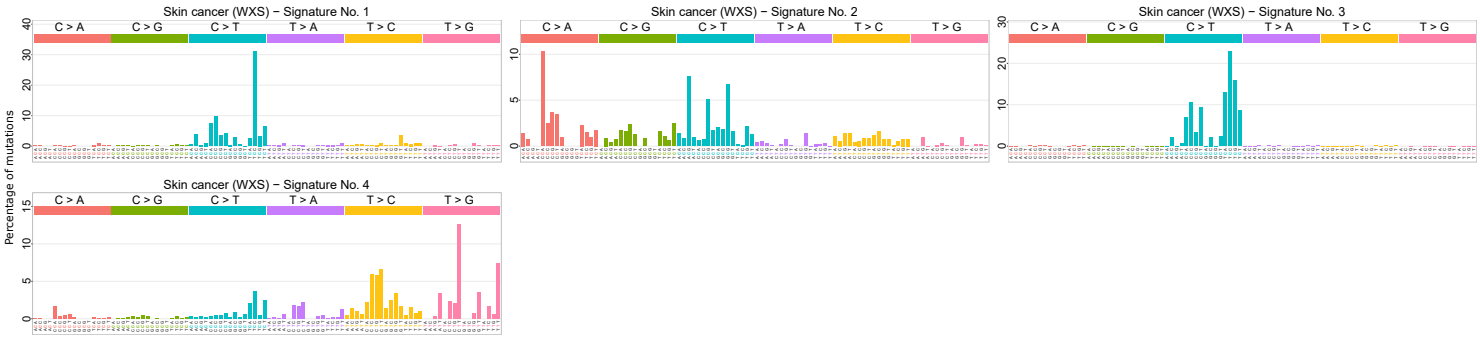

Skin cancer (WXS) − Signature No. 1
Skin cancer (WXS) − Signature No. 2
Skin cancer (WXS) − Signature No. 3
Skin cancer (WXS) − Signature No. 4
C > A
C > G
C > T
T > A
T > C
T > G
Percentage of mutations

Figure 6

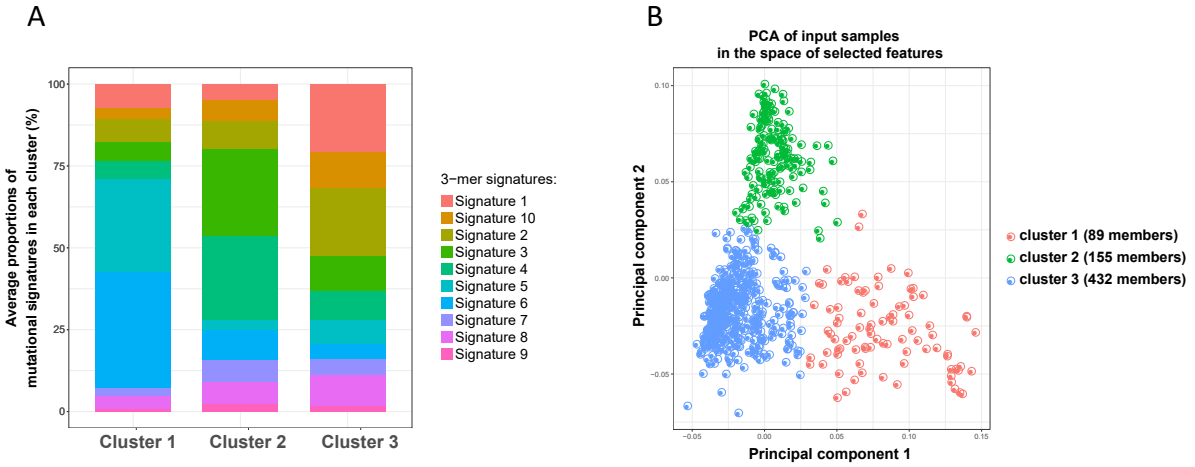

A
Average proportions of mutational signatures in each cluster (%)
100
75
50
25
0
Cluster 1
Cluster 2
Cluster 3
3−mer signatures:
Signature 1
Signature 10
Signature 2
Signature 3
Signature 4
Signature 5
Signature 6
Signature 7
Signature 8
Signature 9
B
PCA of input samples
in the space of selected features
Principal component 2
Principal component 1
cluster 1 (89 members)
cluster 2 (155 members)
cluster 3 (432 members)

Figure 7

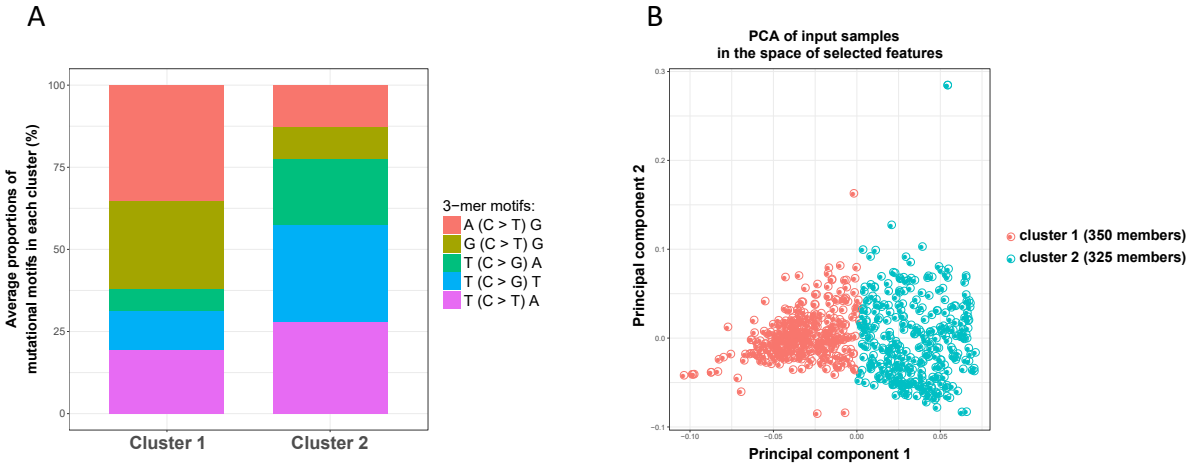

A
Average proportions of
mutational motifs in each cluster (%)
100
75
50
25
0
Cluster 1
Cluster 2
3−mer motifs:
A (C > T) G
G (C > T) G
T (C > G) A
T (C > G) T
T (C > T) A
B
PCA of input samples
in the space of selected features
Principal component 2
Principal component 1
cluster 1 (350 members)
cluster 2 (325 members)

Figure 8

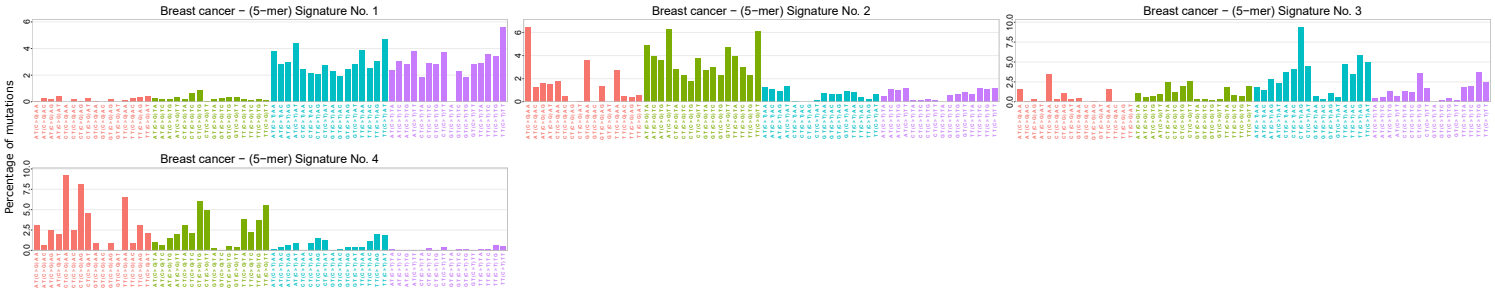
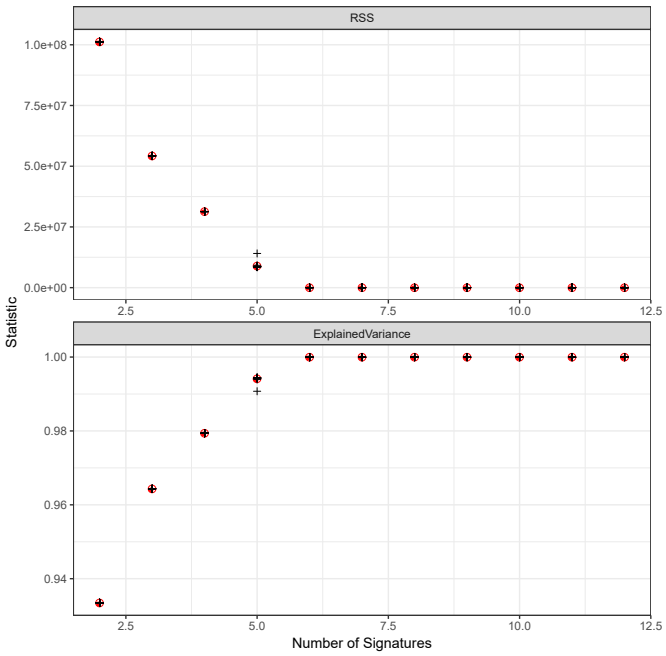

Breast cancer − (5−mer) Signature No. 1
Breast cancer − (5−mer) Signature No. 2
Breast cancer − (5−mer) Signature No. 3
Breast cancer − (5−mer) Signature No. 4
Percentage of mutations

Figure 9

A


RSS
1.0e+08
7.5e+07
5.0e+07
2.5e+07
0.0e+00
Statistic
ExplainedVariance
1.00
0.98
0.96
0.94
2.5
5.0
7.5
10.0
12.5
Number of Signatures

B

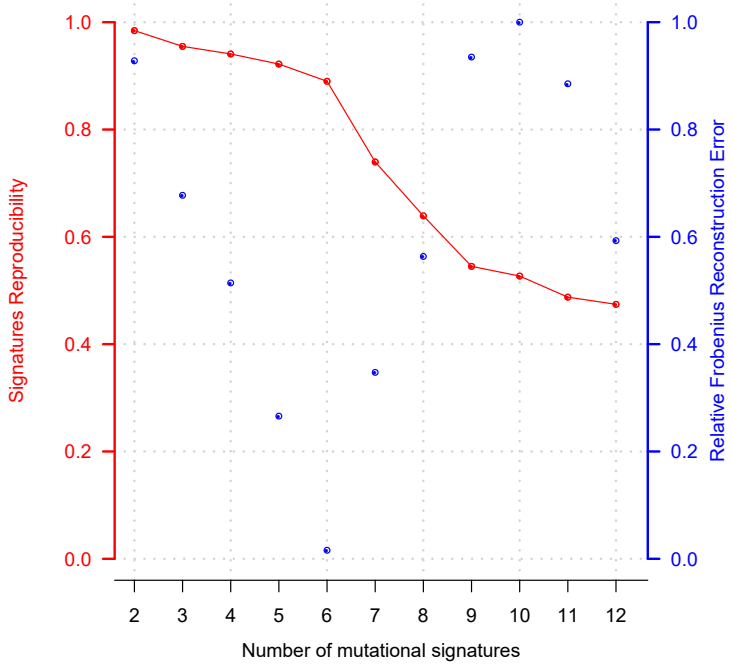
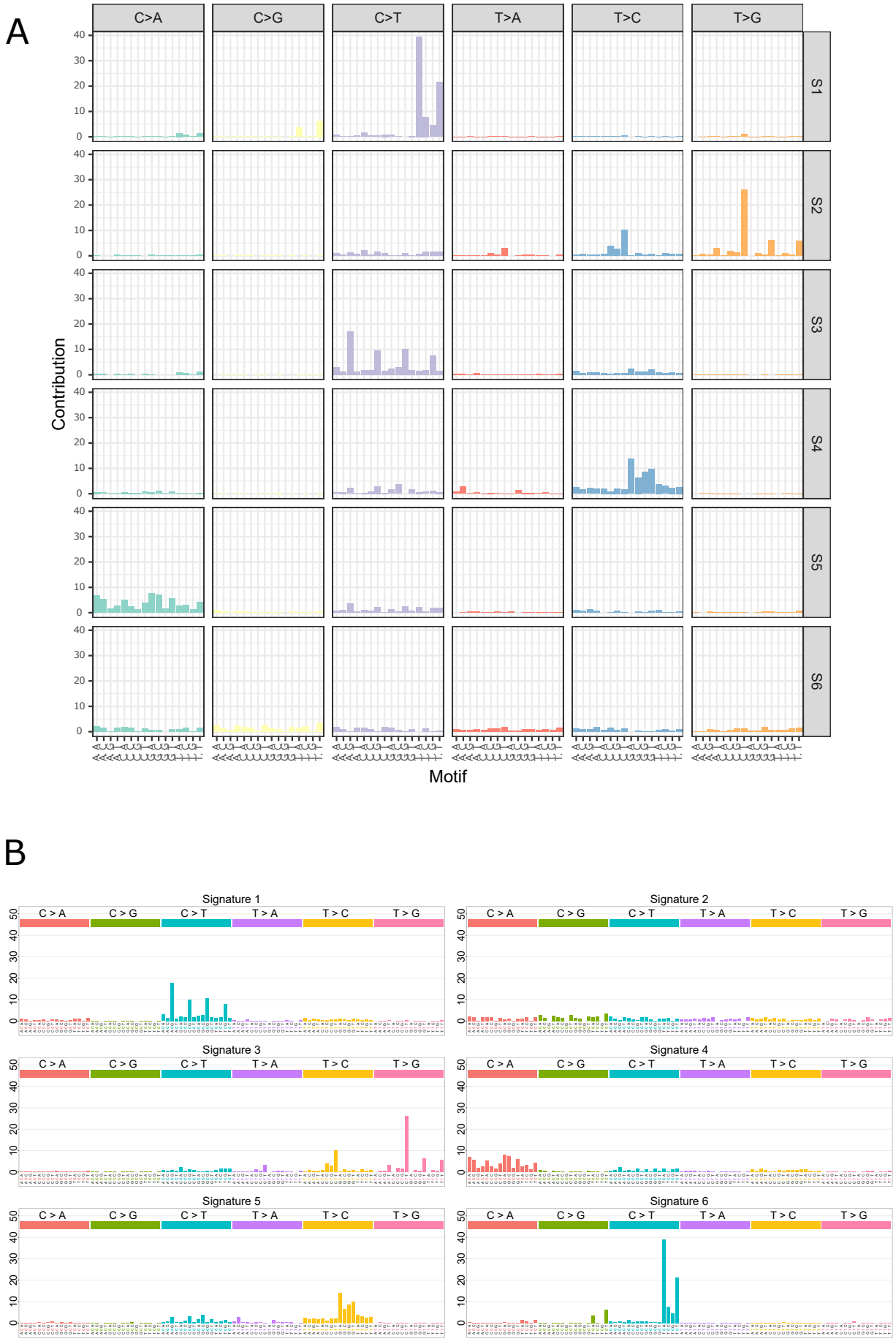

Signatures Reproducibility
Relative Frobenius Reconstruction Error
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
8
9
10
11
12
Number of mutational signatures

Figure 10


A
C>A
C>G
C>T
T>A
T>C
T>G
S1
S2
S3
S4
S5
S6
Contribution
Motif
B
Signature 1
Signature 2
Signature 3
Signature 4
Signature 5
Signature 6
C > A
C > G
C > T
T > A
T > C
T > G
Percentage of mutations