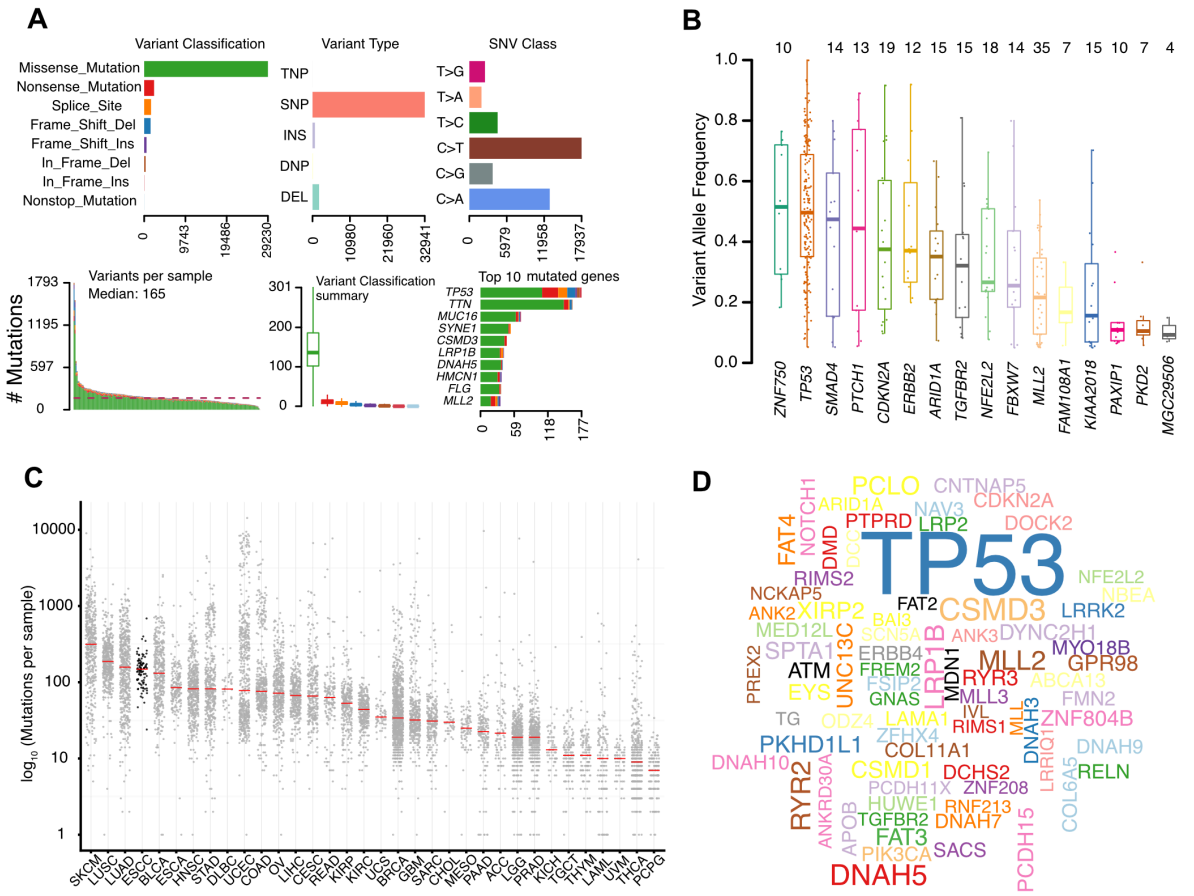
Supplemental Material

Maftools: Efficient and comprehensive analysis of somatic variants in cancer

Anand Mayakonda[1,2,6], De-Chen Lin[3,6], Yassen Assenov[2,5], Christoph Plass[2,5], H. Phillip Koeffler[1,3,4]

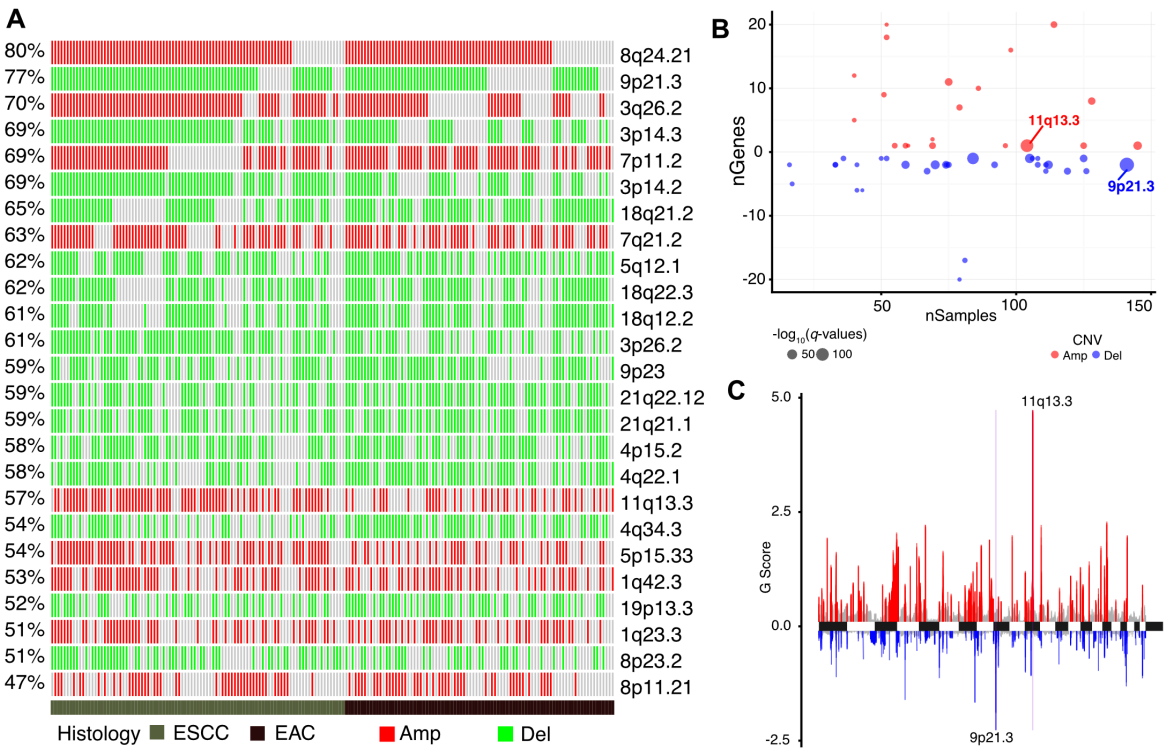
[1]Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore
[2]Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[3]Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA.
[4]National University Cancer Institute, National University Hospital, 119074, Singapore
[5]German Centre for Cardiovascular Research (DZHK), Partner Site Heidelberg/Mannheim, 69120 Heidelberg, Germany
[6]To whom correspondence should be addressed (dchlin11@gmail.com, a.mayakonda@dkfz-heidelberg.de)

Table of Contents:



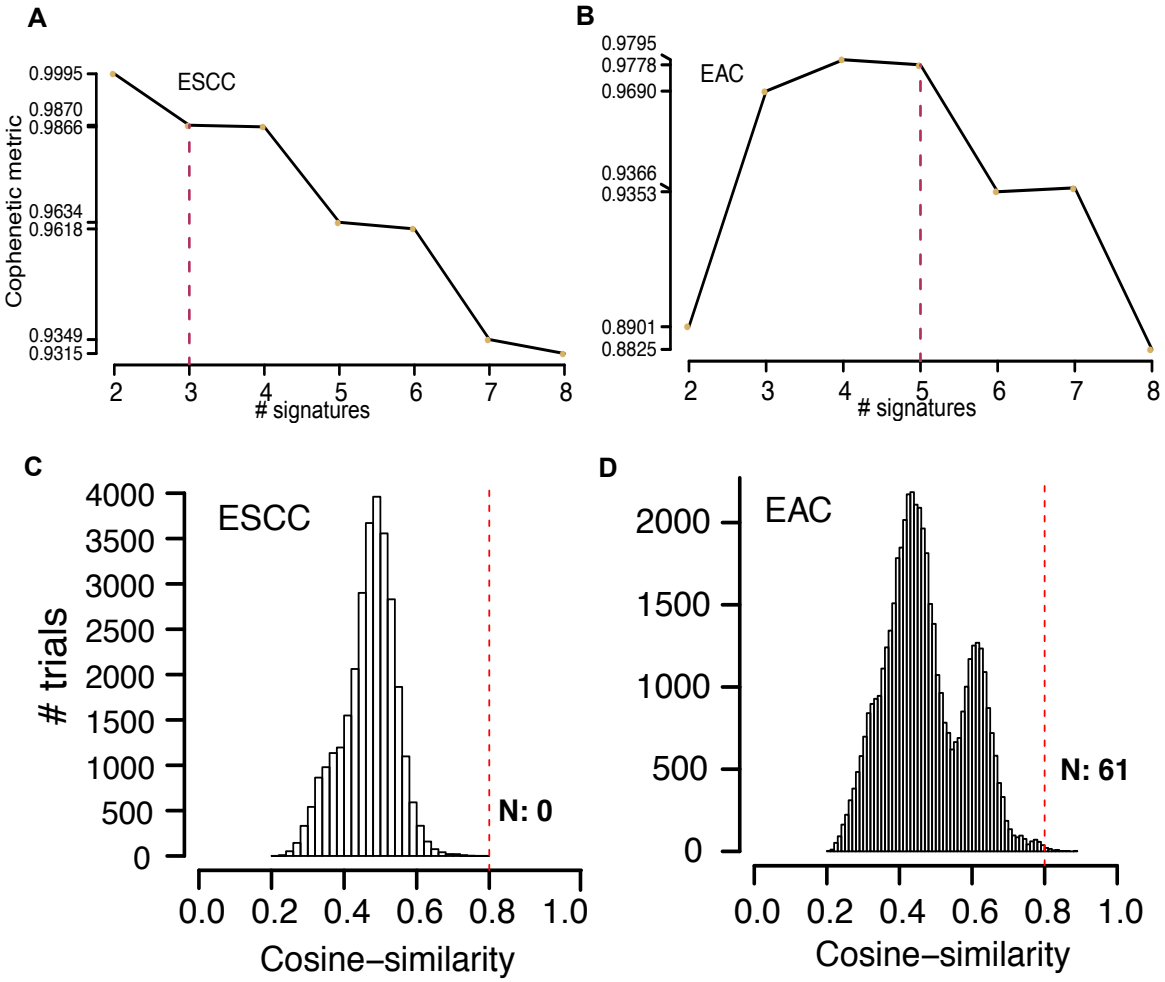
Supplemental Figure S1: Visualization of somatic variants from ESCA

A. Cohort summary plot displaying distribution of variants according to variant classification, type and SNV class. Bottom part (from left to right) indicates mutation load for each sample, variant classification type. A stacked barplot shows top ten mutated genes. **B.** Boxplot of variant allele frequencies for significant disease associated genes identified by MutSigCV ($q < 0.01$). **C.** *TCGAcompare* plot for comparing mutation load of a given MAF against all 33 TCGA cohorts. Mutation load from ESCC is compared against all TCGA cohorts. Red line indicates median; each dot is a sample derived from TCGA cohort. **D.** A word cloud generated based on frequency of mutated genes in ESCA.



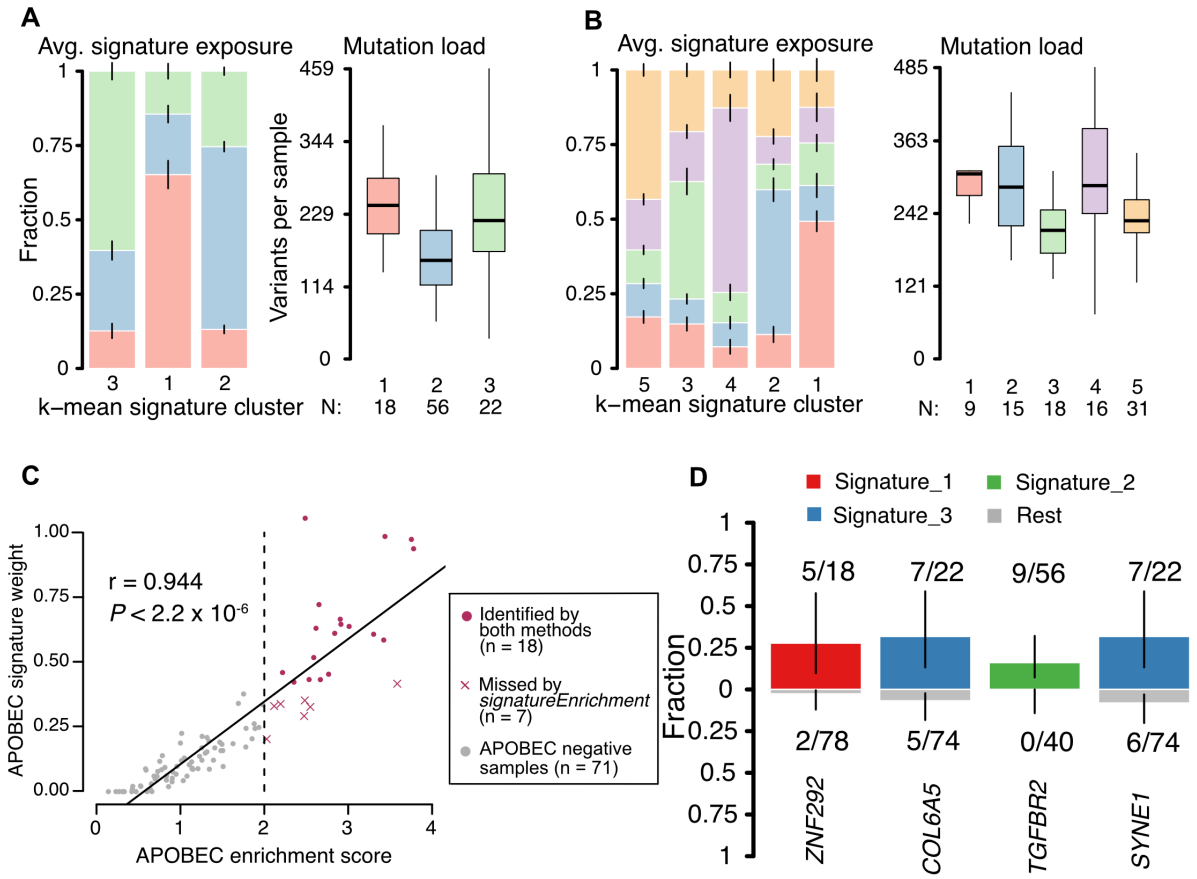
Supplemental Figure S2: Visualizing GISTIC results from ESCA.

A. Oncoplot displays most frequently altered (amplifications or deletions) copy number events ordered according to the frequency. Each columns represents a sample and each row represent a CNV segment. Samples are further sorted according to histology. **B.** GISTIC results plotted as function of altered cytobands, mutated samples, and genes involved within the cytoband. Two of the most significantly altered cytobands 11q13.3 (amplified) and 9q21.3 (deleted) are highlighted. **C.** G-scores assigned by GISTIC for every cytoband plotted along the chromosome.



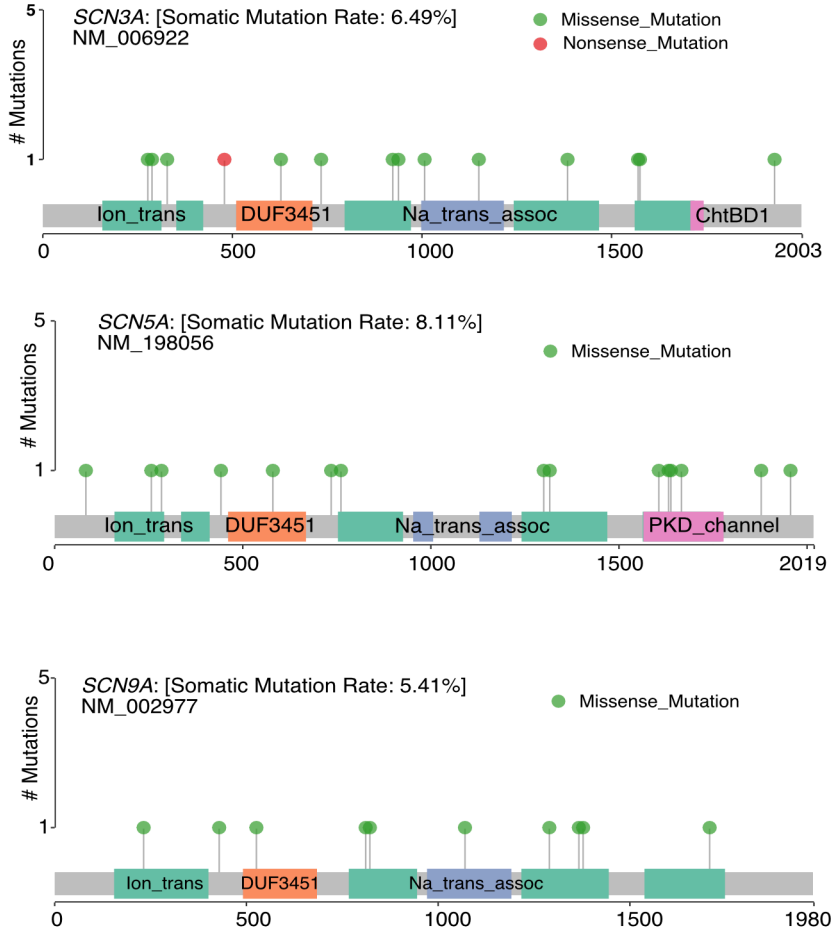
Supplemental Figure S3: Parameter estimation and False Positive Rate for mutational signature analysis.

A and B Cophenetic correlation coefficient (y-axis) measured for a range of values (*r* = 2 .. 8, x-axis) and an optimal decomposition value was chosen at which the coefficient starts decreasing (red line; *r* = 3 and 5 for ESCC and EAC respectively). **C and D.** False positive rate for signature analysis. Weights of each signature identified in ESCC and EAC were shuffled 10,000 times and cosine-similarity was estimated against 30 known COSMIC signatures. Highest similarity value from each trial was noted and plotted as a histogram. Of the 10,000 simulations, zero trials in ESCC had high similarity (> 0.8) whereas 61 trials crossed the threshold of 0.8 in EAC suggesting an approximate low false positive rate of <0.01% and <0.06% respectively.



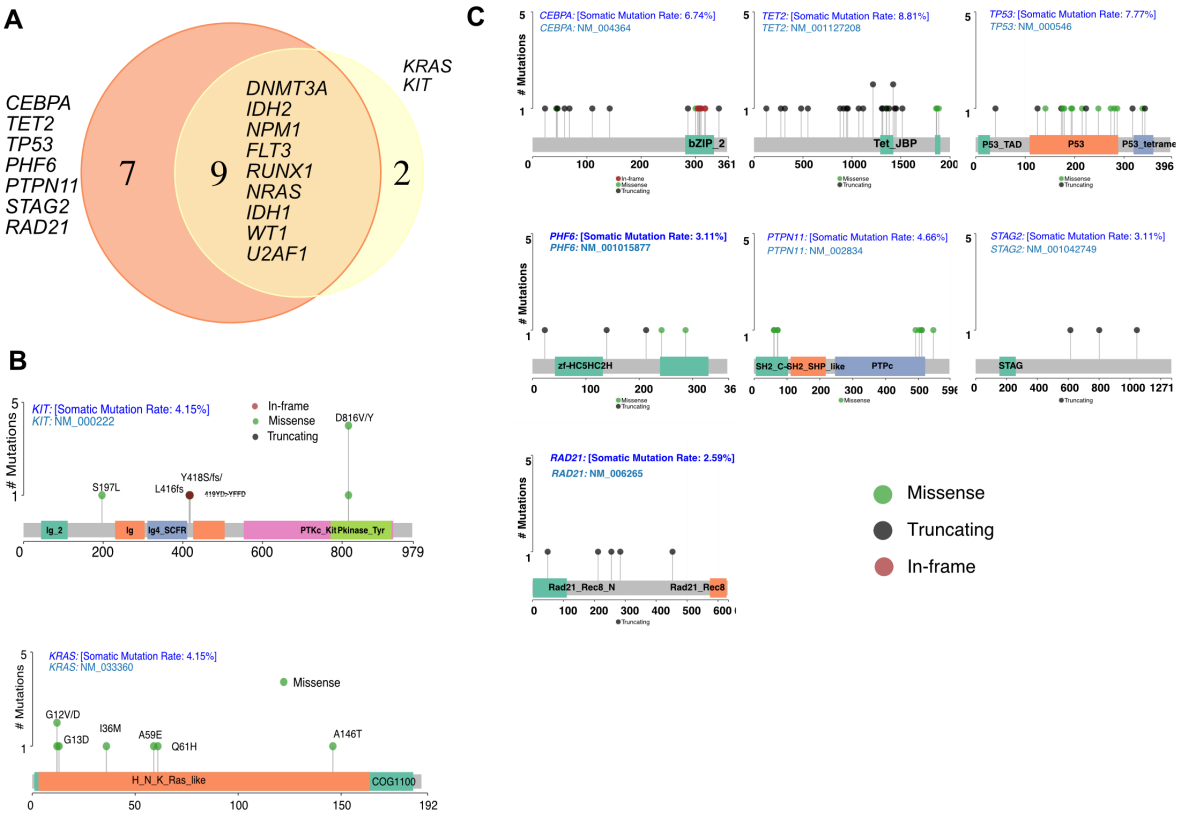
Supplemental Figure S4: Signature enrichment analysis.

A and B. Samples were clustered into *r* groups (*k*-mean clustering approach) based on signature exposures in ESCC and EAC, respectively. Stacked bar plots show average weights of signatures to the cluster. Error bars show standard error of mean. Boxplots show differences in mutation load between signature clusters. (N: number of samples within a cluster) **C.** Scatter plot showing correlation between APOBEC enrichment scores estimated by *trinucleotideMatrix* function and weights of Signature_1 (similar to COSMIC signature 13: APOBEC related, cosine similarity 0.838) as measured by matrix factorization. Both independent analysis showed a high correlation, while *trinucleotideMatrix* function classified significantly large number of samples as APOBEC enriched. **D.** Differentially enriched genes among signatures in ESCC ($P < 0.01$; Fishers exact Test). *TGFBR2* was exclusively mutated among samples belonging to Signature_2 (similar to COSMIC Signature 6: defects in DNA mismatch repair). Bars are annotated with the ratio of mutated samples to total samples. Error bars show 95% confidence interval for binomial ratios.



Supplemental Figure S5: Mutations in genes associated with Voltage Gated Sodium Channels (VGSCs) in EAC.

Activating mutations are randomly distributed across the protein with no recurring hotspots.



Supplemental Figure S6: Comparison of MutSigCV and *oncodrive* results

A. Venn diagram showing overlap between significant genes identified by MutSigCV and oncodriveCLUST algorithm. **B.** *KRAS* and *KIT* identified by oncodrive with the highlighted oncogenic hotspots (G12, G13 in *KRAS*, and D816 in *KIT*). **C.** Significant genes identified by MutSigCV without any recurrent hotspots.

Supplemental Tables (See Excel files)

Supplemental Table S1: Genomic change points identified in 96 Whole Genome Sequenced Breast cancer samples from TCGA.
Supplemental Table S2: Mutation category and APOBEC enrichment scores for 96 ESCC samples.
Supplemental Table S3: Gene enrichment analysis for identified signatures in ESCC.
Supplemental Table S4: Differentially mutated genes between ESCC and EAC.
Supplemental Table S5: Top ten mutated protein domains in ESCC and EAC.
Supplemental Table S6: Genes mutated in mutually exclusive and co-occurring manner in AML.
Supplemental Table S7: Gene sets (*N* = 3) mutated in mutually exclusive manner in AML.
Supplemental Table S8: Time taken by major computationally expensive functions.

Supplemental Data

Supplemental Data S1: Reproducible R code and datasets used for analysis.
Supplemental Data S2: Maftools R package source code.