Deciphering of Somatic Mutational Signatures of Cancer

Xiangwen Ji[1], Edwin Wang[2*], Qinghua Cui[1*]

1 Department of Biomedical Informatics, School of Basic Medical Science, Peking University Health Science Center, Beijing, China.

2 Department of Biochemistry and Molecular Biology, Medical Genetics, and Oncology, Cumming School of Medicine, University of Calgary, Calgary,Alberta, Canada.

*To whom correspondance should be addressed

Qinghua Cui, cuiqinghua@hsc.pku.edu.cn

Edwin Wang, edwin.wang@ucalgary.ca

Abstract

Somatic mutational signatures from genome sequencing play important roles in exploring the cause and development of cancer. However, it is still a challenge to infer their potential meanings (i.e., cancer causal or biological functions). Here we presented Gene Somatic Genome Pattern (GSGP), a computational framework which is able to decipher cancer casual factors and molecular mechanisms of the signatures to help in understanding causal biological processes leading to cancer.

Main

Understanding of causal and biological processes leading to cancer is important for cancer prevention and drug development. As more and more genome sequencing of tumors have been conducted, cancer genome information allows us to get critical insights from these data, among which the somatic mutational signatures (MSs) of cancer derived from these data greatly improved our understanding on the cause and development of cancer[1].

For example, single base substitutions (SBS) somatic MSs provided comprehension of various etiological factors of cancer, including genomic deficiencies and environmental exposures. It was reported that SBS2 and SBS13 are attributed to the activity of Apolipoprotein B mRNA editing enzyme catalytic (APOBEC) family[2], which is the major cause of hypermutation in several cancer types [3]. SBS6, together with SBS14, SBS15, SBS20, SBS21, SBS26, SBS44, are known to be associated with defective DNA mismatch repair (dMMR) [4, 5]. Recognized in cancer patients treated with temozolomide, SBS11 could generate large amounts of mutations[6].

However, signatures with validated evidence of biological function are still a minority, raising a new bottleneck for MS analysis. The key limitation is that the signatures revealed by the current studies are complex blends of single base mutation frequency, whose molecular mechanisms remain totally unknown. Here, we developed an algorithm, Gene Somatic Genome Pattern (GSGP) (Fig. 1a, Methods) to decipher the driver genes of signatures, which could yield new understandings of the tumorigenesis and molecular mechanisms of MSs.

Briefly, GSGP assigned mutation contexts to genes at or near the mutation sites by weights and solved signatures for each gene using a non-negative matrix factorization (NMF)-based algorithm (Fig. 1a, Methods). Specifically, for a sample, GSGP works in 3 steps.

Firstly, we curated the 96 (12 (base replacements) ×4 (5'-type) ×4 (3'-type) =96) different SBS mutation contexts of replacements of pyrimidine (C/T) bases for each mutation site. Same as traditional MS analysis [1], the replacements of purines (G/A) can be reverse-complemented to the other strand.

Secondly, the mutation sites are annotated to the exon, intron, upstream, and downstream of genes, so that their corresponding mutation contexts can be assigned to such genes. Particularly, for mutant loci that can be matched to multiple genes, its context is given to each gene with properly-calculated weights according to its position in the mapped genes. A 'NONE' gene is introduced when a mutation is far away from all genes.

Finally, given a matrix containing the weighted frequencies calculated previously ((number of genes plus NONE gene) ×96) and a known reference list, we use NMF to decompose and solve the GSGP. Benefiting from the introduction of the 'NONE' genes, we can simultaneously solve the sample-level signatures by adding a row vector to the matrix using

the sum of each column of it. As a result, we are able to obtain a new dimension of information, genes, for each signature of each sample, compared to the traditional MS analysis.

To test the GSGP algorithm, we applied it to The Cancer Genome Atlas (TCGA)[7] data. As a result, we obtained 78 SBS signatures covering 10,088 samples and 41,924 genes. To confirm the analysis result, we first focused on *TP53*, one of the most critical tumor suppressor genes. The result showed that GSGP ranks *TP53* in the top 10 out of the >40k genes in 11 signatures (Fig. 1b, Supplementary Table 1), which is indeed supported by established knowledge. For instance, the G>T mutation, the main component of SBS24 (aflatoxin exposure), in the third position of codon 249 of *TP53* is known to be indicative of aflatoxin exposure[8, 9]. Interestingly, a *TP53* mutation C380T, a typical base substitution in SBS11 (temozolomide treatment), has been reported in the recurrent tumor of a glioma patient treated with temozolomide[10]. Treatments with cisplatin[11] (SBS31/35) and thiopurine[12] (SBS87) have also been shown to induce new mutations in *TP53*. These findings corroborate the possibility of *TP53* mutations caused by these anticancer drugs, which result in the recurrence or progression of cancer.

From the aetiology perspective, SBS4 has been found to be involved in genomic damage caused by smoking[4, 13]. GSGP analysis of this signature indeed revealed a significant contribution of genes related to the sensory perception of chemical stimulus and the behavioral response to nicotine (Supplementary Fig. 1a, 1b). The latter is comprised mainly of cholinergic receptor genes, suggesting that smoking could lead to somatic mutations in these genes. Interestingly, several studies have revealed that germline polymorphisms in cholinergic receptor genes are indeed associated with tobacco addiction and are also associated with the

risk of smoking-related diseases including lung cancer [14, 15]. In our previous study, we introduced the cancer germline genome pattern (CGGP) and found a positive correlation between CGGP_E and SBS4[16]. These findings suggest an interaction between genomic variations and environmental exposure, whereby cholinergic receptor germline polymorphisms are associated with tobacco addiction, and tobacco smoking results in somatic mutations in such genes.

Furthermore, GSGP shows promising potential for exploring the aetiology of signatures. SBS7, including subtypes a/b/c/d, has been identified to be involved in DNA damage caused by ultraviolet (UV) exposure[13]. UV can be further classified as UVA, UVB, and UVC by wavelength, among which UVA is the main type that can pass through the atmosphere and cause exposure damage to human [17]. We found that the GSGP profiles of SBS7 subtypes showed significantly different clusters after t-distributed stochastic neighbor embedding (t-SNE)[18] (Fig. 2b), leading to the different genes and pathways affected by SBS7 subtypes. Through gene set enrichment analysis[19] (GSEA), we found that genes contributing to SBS7a were enriched in the 'response to UVC' pathway, while genes in SBS7c were identified as 'response to UVB' (Fig. 2c), revealing the potential wavelength-specificity of UV-related signatures.

However, none of the SBS7 subtypes showed significant affections to the genes responding to UVA, the main type causing exposure damage to human [17]. Here, we found that a signature presents only in melanomas, SBS55 was the only signature whose affected genes were enriched in the 'response to UVA' process, and the only one that was enriched in all the 3 processes of UVA/B/C simultaneously (Fig. 2c). SBS55 has been considered as a possible

signature due to sequencing artefacts. However, by de novo extraction of signatures, we successfully detected the presence of SBS55 (Supplementary Table 2, Methods). The results of clustering based on tissue distribution of signatures showed that SBS55 has a closer distance to SBS7 compared to SBS38, which is thought to be non-direct UV damage (Fig. 2a). Notably, patients with SBS55 have shorter overall survival days compared to those without (Fig. 2d). In conclusion, this analysis revealed that SBS55 represents a deficiency in the defense against UV exposure.

We observed that a cluster of signatures (Fig. 2a) contains 6 of the 7 known dMMR-related signatures. There are also 4 DNA polymerase-related signatures (SBS10a, SBS10b, SBS14, and SBS20) in this cluster. It is known that DNA polymerases participate in MMR[20] and base excision repair (BER)[21]. With the study of the genes contributing to the remaining 6 aetiology-unknown signatures by GSGP, we found that most (5/6) of them are related to MMR and/or BER processes (Supplementary Table 3), suggesting that they could also be involved in dMMR and/or dBER. SBS3 is proposed to be associated with mutations in BRCA1 and BRCA2 and the homologous recombination DNA repair (HRR) process in which they are involved[22, 23]. While SBS34 exhibits a similar tissue distribution (Fig. 2a). The GSGP results show that the genes relevant to SBS34 are enriched in the HR process (Supplementary Fig. 1c), suggesting that it could be a new dHRR-related signature.

Finally, we built a web tool (https://www.cuilab.cn/gsgp) to facilitate the GSGP analysis (Fig. 2e). The tool computes 78 SBS GSGPs for each gene of each given sample (3-dimensional, 3-D results). For the convenience of visualization, we provide filtering and aggregation functions for the results. Users can specify the axis and method to aggregate the 3-D results

into 2-D and display them as a heatmap. The results, as well as a python script that can have more personalized options (species, NMF parameters, multiprocessing, etc.), are available for download.

Overall, we presented an efficient algorithm, GSGP, for exploring the driver genes contributing to somatic signatures, a new challenge in cancer genomics. We confirmed the power of GSGP by investigating the *TP53*-contributed signatures and the signature SBS4-driver genes. More importantly, we showed that GSGP is able to explore aetiology for signatures. We revealed that SBS55, SBS17a, and SBS34 could be a new UV exposure-related signature, dMMR/dBER related signature, and dHRR related signature, respectively. In addition, an online tool for GSGP analysis was developed. This study provides new insights into deciphering the molecular mechanisms of MSs, which are critical for applying MSs to the development of new therapeutic strategies for cancer.

Methods

The gene somatic genome pattern

The workflow of the GSGP algorithm is illustrated in Fig. 1a. Specifically, the calculation of GSGP for a sample can be divided into 3 steps as follows.

Step 1: Like the SBS MS analysis[1], we obtained the bases of mutation sites and one base at each of their 5' and 3' sides, and identified the 96 different mutation contexts of replacements of pyrimidine (C/T) bases. The replacements of purines (G/A) can be reverse-complemented to the other strand. The reference genome FASTA files (hg38, used by TCGA and hg19, used by Australian melanoma biospecimen banks[24] (AMBB)) acquired by this step were obtained

from UCSC Genome Browser[25] (http://genome.ucsc.edu).

Step 2: The mutation contexts of each mutation site are assigned to the genes by weights. We used bedtools[26] (v2.30.0) and its python wrapper, pybedtools[27] (v0.8.2), to annotate the mutation sites to the exon, intron, upstream, and downstream of genes. For mutations that are at or near only one gene, its mutation context can be assigned to that gene entirely. Yet for mutation locus that can be matched to multiple genes, its context is assigned to each gene by weights according to its position in the gene it is matched to. Based on the research on expression quantitative trait locus (eQTL), we chose appropriate weights, which are exon: intron: up/downstream = 1.78: 1.57: 1, and distance thresholds, which are 121925bp upstream and 121388bp downstream, as described in the section 'Selection of weights and up/downstream distances'. Particularly, if a mutation locus is far from all genes, it will be assigned to a 'NONE' gene. The positions and biotype annotations of genes were collected from Ensembl[28] (https://www.ensembl.org), and pseudogenes were removed.

Step 3: NMF is used to solve GSGP. For the weighted counts of mutation contexts of genes counted in the previous step, we first convert them into frequencies by dividing them by the total number of mutations, denoted as matrix *X*. The rows of *X* represent genes (including 'NONE' gene) and the columns represent the 96 mutation contexts. We solve *W*, the SBS signature of genes by the following equation:

$$\begin{vmatrix} X \\ X' \end{vmatrix} \approx \begin{vmatrix} W \\ W' \end{vmatrix} \cdot H$$

where *H* (78*96) records the frequencies of 96 mutation contexts occurring in each of the 78 signatures, respectively, which can be downloaded from COSMIC (https://cancer.sanger.ac.uk/signatures/downloads) depending on the different reference

genomes. Keeping H unchanged, we solved the GSGP values W using the NMF algorithm from scikit-learn (v1.0.1), whose rows denote genes and columns denote 78 novel SBS signatures. Conveniently, benefiting from the introduction of the 'NONE' gene, we can solve for the sample-level signature W' simultaneously by adding a row vector X' to X using the sum of each column of X.

Selection of weights and up/downstream distances

We used a similar strategy to MAGENTA which is used to perform enrichment analysis of GWAS data[29], to select thresholds for upstream and downstream distances. Initially, all eQTL data were obtained from the Genotype-Tissue Expression (GTEx) project[30] v8 (https://gtexportal.org/home). Based on the gene position annotations of GENCODE[31] v26, the version used by GTEx, we annotated eQTL to the exon, intron, upstream, and downstream of the eQTL-correlated gene (eGene). Duplicate eQTL-eGene pairs are kept only once. We used the densities of eQTLs in each case as the weights of mutations. For mutations located on exons, the weights W_{exon} are:

$$\mathrm{W_{exon}} = \frac{\text{Total number of eQTLs on eGene exons}}{\text{Total length of eGene exons}}$$

For mutations located on introns, the weights W_{intron} are:

$$\mathrm{W_{intron}} = \frac{\text{Total number of eQTLs on eGene introns}}{\text{Total length of eGene introns}}$$

For the case of upstream/downstream, a distance from the gene start/end site was calculated at the same time. We used the median of the distances from all upstream/downstream eQTLs to their eGenes as the threshold for upstream/downstream distances used in GSGP, denoted as D_{up} and D_{down}. the weights for mutations within this threshold, $W_{up/downstream}$, are:

$$\mathrm{W_{up/downstream}} = \frac{\text{Total number of eQTLs on eGenes}' \text{ up/downstream} \times 0.5}{(\mathrm{D_{up}} + \mathrm{D_{down}}) \times \text{Total number of eGenes}}$$

When scaling $W_{up/downstream}$ to 1, the ratio of weights of a mutation located at exon, intron, and upstream/downstream is about 1.78: 1.57: 1 (in detail, 1.78074309122: 1.57446085424: 1).

Somatic mutation data collection

The somatic mutations of 10,114 samples of 33 different tumor types of TCGA project, as well as their corresponding clinical information, were downloaded from the GDC data portal (https://portal.gdc.cancer.gov). The mutations were called by VarScan2 and stored in Mutation Annotation Format (MAF). We used the 'maf2vcf' perl script (https://github.com/mskcc/vcf2maf) to convert the MAF to VCF files. In addition to the filtration implemented by TCGA, the mutations were further filtered using these criteria: (1) only SNVs were used in the SBS signatures we mainly focused on; (2) the read depths covering the mutated sites should be larger than 10; (3) the variant allele frequency (VAF) should be larger than 1%, consistent with our previous research[16]. As a result, 10,088 samples were retained for subsequent analysis.

Signature validation using SigProfiler

We used SigProfiler[1], the de novo MS extraction tool officially provided by COSMIC, to validate the existence of SBS55 in melanoma. Somatic mutation information for 183 melanoma samples was extracted from the AMBB cohort[24], and the same additional filtration was performed. Then, the python package SigProfilerExtractor (v1.1.4) was applied to the TCGA Skin Cutaneous Melanoma (SKCM) cohort and the AMBB cohort, respectively. Since the low proportion of patients with SBS55, SBS55 may be masked by major signatures such as SBS7a. We disabled the resample to improve the stability of the model to capture more signatures. We selected the optimal number of signatures by 100 replications of NMF solving

and assigned the signatures to COSMIC SBS signature v3.2. We also increased the requirement for read depth to >20x and repeated the same MS extraction to exclude the effect of sequencing artefacts.

Gene set enrichment analysis

For a given signature, we first picked all samples with the signature (sample level SBS signature>0) out. Next, we counted the number of samples with this signature (GSGP value>0) for each gene, which will serve as the scores of the genes (Supplementary Table 1). The genes that did not have the signature in all samples (i.e., those with score=0) were removed because too many equal values would affect the accuracy of gene set enrichment analysis (GSEA). Gene Ontology (GO) gene sets were curated from the Molecular Signatures Database (MSigDB)[32] v7.1 (http://www.gsea-msigdb.org/gsea/index.jsp). GSEA was implemented in positive score type mode using the R package ClusterProfiler[33] (v4.0.5).

Hierarchical clustering

We clustered signatures or tumor types using the number of samples with each signature in each tumor. Numbers of samples were obtained from the detail page of each signature from COSMIC (https://cancer.sanger.ac.uk/signatures/sbs). Min-max normalization was performed on such numbers by each signature. Using the Euclidean distance as the measure of similarity, the complete linkage method as the agglomeration strategy for branches (default parameters), we computed hierarchical clustering using the core function of R (v4.1.0). Leaves were properly ordered by dendsort[34] (v0.3.4).

t-SNE analysis of the SBS7 subtype GSGP profiles

For each subtype of SBS7, samples with corresponding signatures present at the sample level

were retained, and their GSGP profiles were then extracted and converted into a scanpy[35] annotated data matrix. To exclude outliers, we removed observations with less than 200 affected genes, and genes with less than 5 observations. A z-score normalization was performed prior to PCA. The first 10 principal components were used for further t-SNE dimension reduction. We used Multicore-tSNE to accelerate the t-SNE operation, which was considered to obtain a better convergence[36].

Statistical analysis

Survival analysis was performed using the R package survival (v3.2-13) and survminer (v0.4.9), and the log-rank test was used to evaluate the significance of the difference in survival days between the two groups.

Figure legends

Fig. 1. The GSGP workflow. (a) Taking a simulated observation (Obs.), which is aligned to a simulated reference (Ref.) genome and genes (GENE1 and GENE2), as an example, the figure shows how SBS mutation contexts for 5 different conditions are assigned to genes and used to compute MSs, i.e. GSGP. (b) The rank of GSGP of *TP53* among all genes for each signature with known aetiology.

Fig. 2. GSGP discover driver genes and aetiology of signatures. (a) Signatures and cancer types are clustered hierarchically according to the number of samples. The full names of the cancer type abbreviations can be found from https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations. (b) The distinguishability of the GSGP profile of SBS7 subtypes is shown by t-SNE. (c) The 'response to UV A/B/C' GSEA of GSGP scores of SBS7a/b/c/d, and SBS55. NES: normalized enrichment score. The waves representing the UV A/B/C (wavelength from long to short) are shown on the points, the waves in red mean significant ($p<0.05$) results. (d) The Kaplan-Meier curve showing the overall survival of SKCM patients with or without SBS55. (e) The main functions and workflow of the web tool for GSGP analysis.

Acknowledgements

This work was supported by the Natural Science Foundation of China (62025102).

Author contributions

Q.C. and E. W. conceived of the study. X.W. presented the GSGP algorithm and performed the analysis. X.W., Q.C., and E.W. wrote the paper.

Competing interests

The authors declare no competing interests.

Supplementary information

Supplementary Figures and Tables

PDF file containing Supplementary Fig. 1 and Supplementary Table 2-3. Detailed legends are provided below each figure/table.

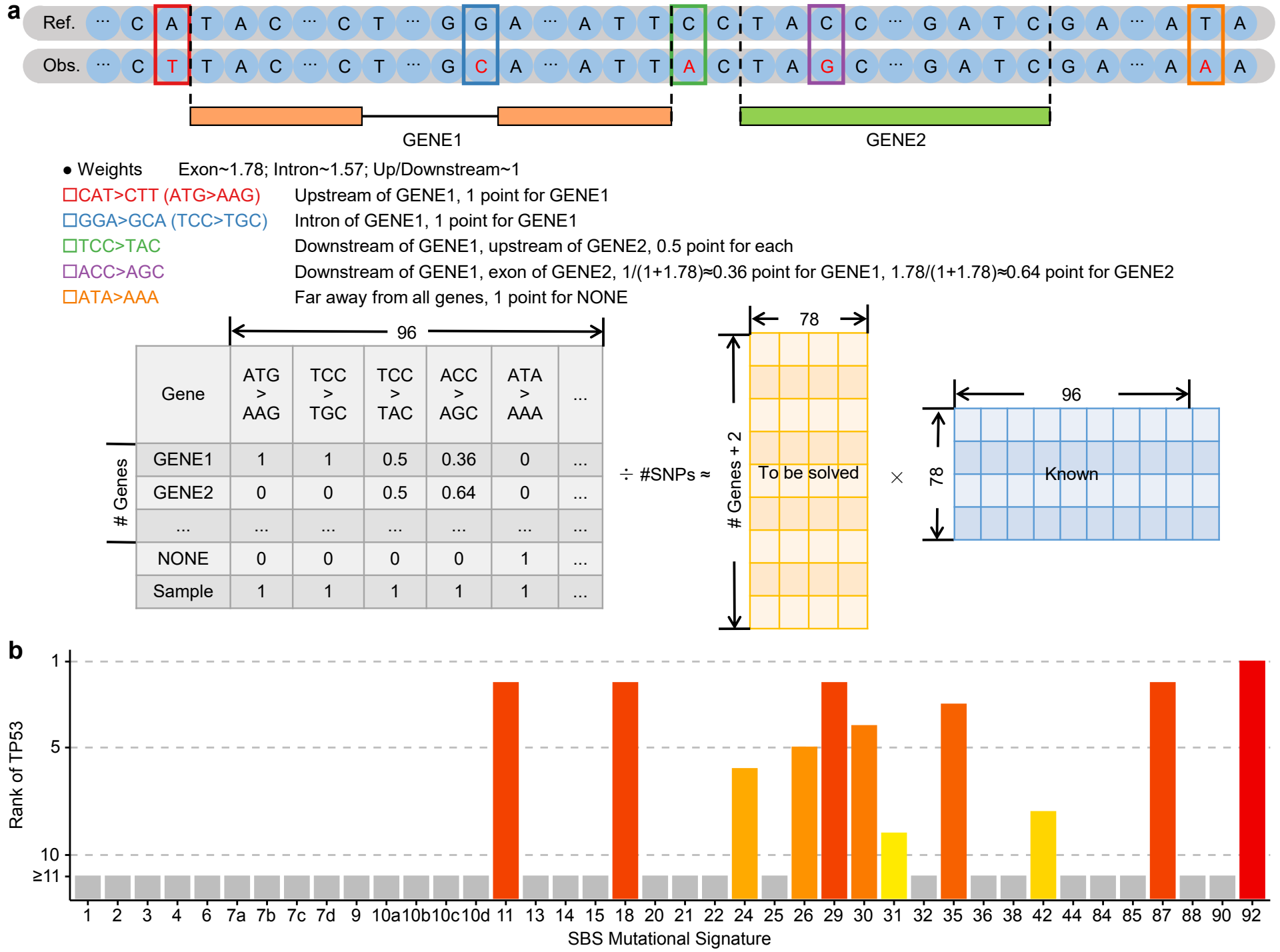
Supplementary Table 1

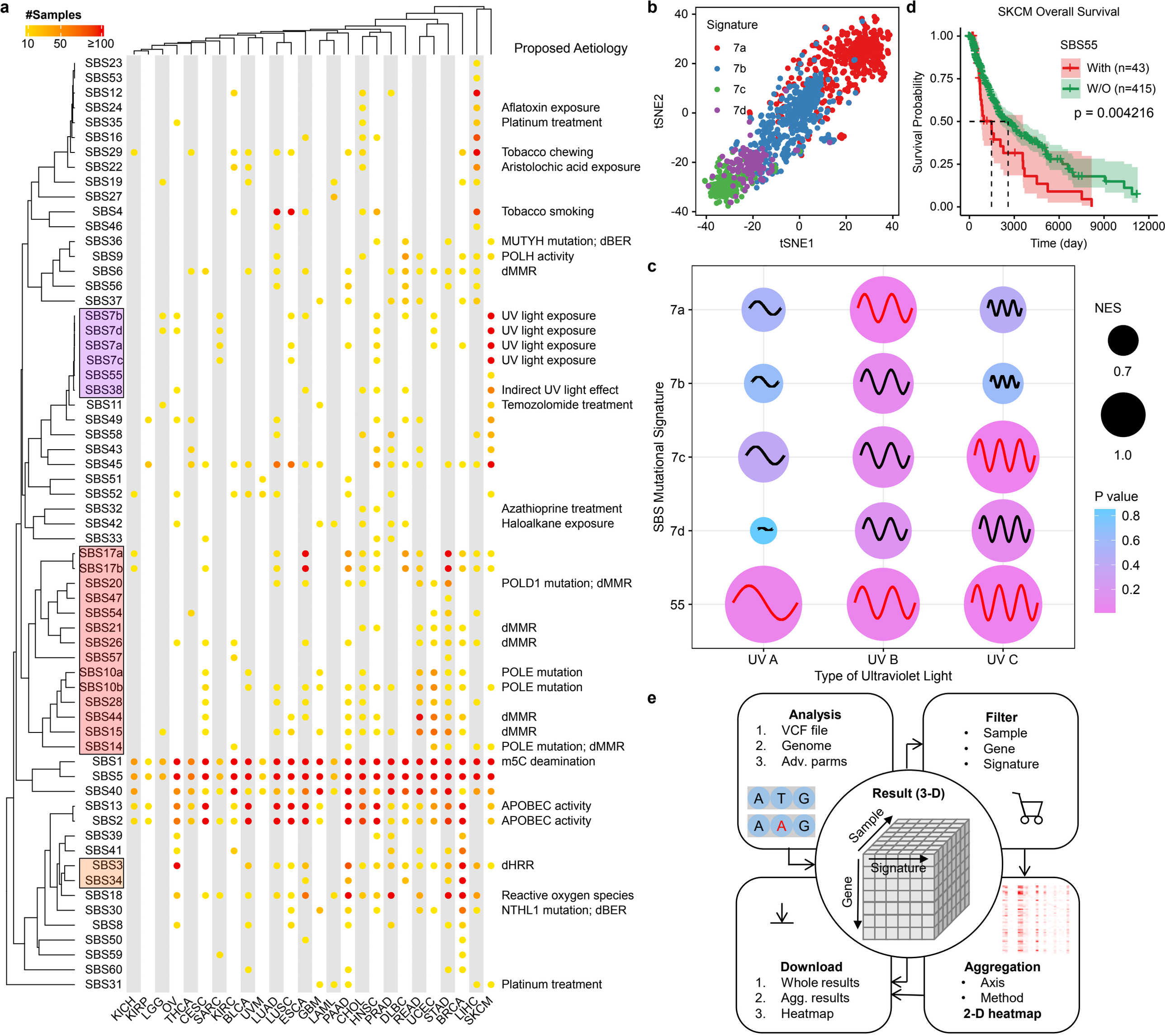
Excel file containing the scores (number of samples with GSGP>0) of genes in each SBS signatures, used for GSEA and the case study of *TP53*.

References

1. Alexandrov, L.B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
2. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* **47**, 1067-1072 (2015).
3. Roberts, S.A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**, 970-976 (2013).
4. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
5. Hodel, K.P. et al. POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. *Mol Cell* **78**, 1166-1177 e1166 (2020).
6. Kucab, J.E. et al. A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836 e816 (2019).
7. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
8. Lasky, T. & Magder, L. Hepatocellular carcinoma p53 G > T transversions at codon 249: the fingerprint of aflatoxin exposure? *Environ Health Perspect* **105**, 392-397 (1997).
9. Ozturk, M. p53 mutation in hepatocellular carcinoma after aflatoxin exposure. *Lancet* **338**, 1356-1359 (1991).
10. Johnson, B.E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189-193 (2014).
11. Murray, D. & Mirzayans, R. Cellular Responses to Platinum-Based Anticancer Drugs and UVC: Role of p53 and Implications for Cancer Therapy. *Int J Mol Sci* **21** (2020).
12. Li, B. et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. *Blood* **135**, 41-55 (2020).
13. Nik-Zainal, S. et al. The genome as a record of environmental exposure. *Mutagenesis* **30**, 763-770 (2015).
14. Hung, R.J. et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633-637 (2008).
15. Thorgeirsson, T.E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-642 (2008).

16. Xu, X. et al. Germline genomic patterns are associated with cancer risk, oncogenic pathways, and clinical outcomes. *Sci Adv* **6** (2020).
17. Sage, E., Girard, P.M. & Francesconi, S. Unravelling UVA-induced mutagenesis. *Photochem Photobiol Sci* **11**, 74-80 (2012).
18. Laurens, V.D.M. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).
19. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
20. Fishel, R. Mismatch repair. *J Biol Chem* **290**, 26395-26403 (2015).
21. Grundy, G.J. & Parsons, J.L. Base excision repair and its implications to cancer therapy. *Essays Biochem* **64**, 831-843 (2020).
22. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).
23. Zamborszky, J. et al. Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**, 5085-5086 (2017).
24. Hayward, N.K. et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175-180 (2017).
25. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
26. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
27. Dale, R.K., Pedersen, B.S. & Quinlan, A.R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423-3424 (2011).
28. Howe, K.L. et al. Ensembl 2021. *Nucleic Acids Res* **49**, D884-D891 (2021).
29. Segre, A.V. et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6** (2010).
30. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).
31. Frankish, A. et al. Gencode 2021. *Nucleic Acids Res* **49**, D916-D923 (2021).
32. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
33. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)* **2**, 100141 (2021).
34. Sakai, R., Winand, R., Verbeiren, T., Moere, A.V. & Aerts, J. dendsort: modular leaf ordering methods for dendrogram representations in R. *F1000Res* **3**, 177 (2014).
35. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
36. Laurens, V.D.M. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**, 3221-3245 (2014).


a
Ref.
Obs.
GENE1
GENE2
● Weights Exon~1.78; Intron~1.57; Up/Downstream~1
CAT>CTT (ATG>AAG) Upstream of GENE1, 1 point for GENE1
GGA>GCA (TCC>TGC) Intron of GENE1, 1 point for GENE1
TCC>TAC Downstream of GENE1, upstream of GENE2, 0.5 point for each
ACC>AGC Downstream of GENE1, exon of GENE2, 1/(1+1.78)≈0.36 point for GENE1, 1.78/(1+1.78)≈0.64 point for GENE2
ATA>AAA Far away from all genes, 1 point for NONE
96
Gene
ATG > AAG
TCC > TGC
TCC > TAC
ACC > AGC
ATA > AAA
...
Genes
GENE1 1 1 0.5 0.36 0 ...
GENE2 0 0 0.5 0.64 0 ...
NONE 0 0 0 0 1 ...
Sample 1 1 1 1 1 ...
÷ #SNPs ≈
78
Genes + 2
To be solved
×
96
78
Known
b
Rank of TP53
1
5
10
≥11
1 2 3 4 6 7a 7b 7c 7d 9 10a 10b 10c 10d 11 13 14 15 18 20 21 22 24 25 26 29 30 31 32 35 36 38 42 44 84 85 87 88 90 92
SBS Mutational Signature


a
#Samples
10
50
≥100
Proposed Aetiology
SBS23
SBS53
SBS12
SBS24
SBS35
SBS16
SBS29
SBS22
SBS19
SBS27
SBS4
SBS46
SBS36
SBS9
SBS6
SBS56
SBS37
SBS7b
SBS7d
SBS7a
SBS7c
SBS55
SBS38
SBS11
SBS49
SBS58
SBS43
SBS45
SBS51
SBS52
SBS32
SBS42
SBS33
SBS17a
SBS17b
SBS20
SBS47
SBS54
SBS21
SBS26
SBS57
SBS10a
SBS10b
SBS28
SBS44
SBS15
SBS14
SBS1
SBS5
SBS40
SBS13
SBS2
SBS39
SBS41
SBS3
SBS34
SBS18
SBS30
SBS8
SBS50
SBS59
SBS60
SBS31
Aflatoxin exposure
Platinum treatment
Tobacco chewing
Aristolochic acid exposure
Tobacco smoking
MUTYH mutation; dBER
POLH activity
dMMR
UV light exposure
UV light exposure
UV light exposure
UV light exposure
Indirect UV light effect
Temozolomide treatment
Azathioprine treatment
Haloalkane exposure
POLD1 mutation; dMMR
dMMR
dMMR
POLE mutation
POLE mutation
dMMR
dMMR
POLE mutation; dMMR
m5C deamination
APOBEC activity
APOBEC activity
dHRR
Reactive oxygen species
NTHL1 mutation; dBER
Platinum treatment
KICH
KIRP
LGG
OV
THCA
CESC
SARC
KIRC
BLCA
UVM
LUAD
LUSC
ESCA
GBM
LAML
PAAD
CHOL
HNSC
PRAD
DLBC
READ
UCEC
STAD
BRCA
LIHC
SKCM
b
Signature
7a
7b
7c
7d
tSNE2
tSNE1
d
SKCM Overall Survival
SBS55
With (n=43)
W/O (n=415)
p = 0.004216
Survival Probability
Time (day)
c
SBS Mutational Signature
7a
7b
7c
7d
55
UV A
UV B
UV C
Type of Ultraviolet Light
NES
0.7
1.0
P value
0.8
0.6
0.4
0.2
e
Analysis
1. VCF file
2. Genome
3. Adv. parms
A T G
A A G
Filter
• Sample
• Gene
• Signature
Result (3-D)
Sample
Signature
Gene
Download
1. Whole results
2. Agg. results
3. Heatmap
Aggregation
• Axis
• Method
2-D heatmap