Are mutational signatures a valid tool to detect tobacco consumption among patients with NSCLC cancer?

The detection of mutational signatures connected to tobacco use in TRACERx database

PhDr. Ján Sliacky
*Institute of Informatics, Information Systems and Software Engineering**
Faculty of Informatics and Information Technologies, STU in Bratislava
Bratislava, Slovakia
jan.sliacky@stuba.sk

***Abstract*— This study evaluates whether mutational signatures can reliably detect tobacco exposure in whole-exome sequencing (WES) samples from non-small cell lung cancer (NSCLC) patients. Using five refitting tools (SigProfiler, SigFit, SigMiner, MusiCaTK, MutationalPatterns) on TRACERx 100 data (426 WES samples), we found no significant differences in single base substitution (SBS) signatures between ex-smokers (n=62 patients) and never-smokers (n=38). Canonical tobacco-associated signatures (SBS4, SBS29, SBS92) were undetectable, with only ubiquitous clock-like signatures (SBS1, SBS5) dominating. Substantial inter-tool discordance persisted despite parameter harmonization. These results question the utility of SBS analysis in WES for smoking inference and highlight limitations in sensitivity, coverage, and standardization**

Keywords— TRACERx, WES, mutational signatures, SBS, smoking, NSCLC, NMF, COSMIC

I. INTRODUCTION

Mutational signatures capture endogenous and exogenous mutational processes imprinted in cancer genomes. Tobacco smoking, a primary NSCLC risk factor, generates characteristic SBS profiles (e.g., SBS4: C>A transversions from polycyclic aromatic hydrocarbons). Whole-genome studies consistently detect these in smokers, but WES utility (~1–2% coverage, protein-coding focus) remains underexplored.

We hypothesized that refitting tools applied to TRACERx NSCLC WES would: (i) consistently identify tobacco signatures and (ii) differentiate ex-smokers from never-smokers. Unexpected null results prompted evaluation of tool variability, WES power, and methodological pitfalls.

II. METHODOLOGY

To calculate the mutational signatures, five tools were used. The extraction of the signatures was using refitting methods. The respective tools are ***Sigminer****: Mutational Signature Analysis and Visualization in R [1]*, ***sigfit****: flexible Bayesian inference of mutational signatures[2]*, ***Mutational Signatures [3]***, ***musicatk****: Mutational Signature Comprehensive Analysis Toolkit[4]*. and ***SigProfilerAssignment*** [5]. The first four tools are written in R language. *SigProfilerAssigment* is written in Python by creators of the mutational signature concept. As a reference database for refitting was used the COSMIC Mutational Signatures which is a catalogue of curated reference mutational signatures. In our dataset, we worked with single base substitutions. This kind of mutation was the first to be used in Non-Negative Matrix Factorization method. For refitting, the abovementioned tools are using Non-Negative Least Squares apart from *sigfit* that uses Markov Chain Monte Carlo Algorithm.

These tools could calculate the mutational signatures using **de novo** methods or **refitting**. De novo extraction estimates both signatures (*W*) and exposures (*H*) from scratch, requiring substantial mutation counts (typically >10,000 SBS/sample). Refitting fixes *W* to COSMIC references, solving only *H* via NNLS/MCMC—more robust for sparse data. For our experiment, **refitting** was optimal since WES datasets (~1-2% genomic coverage, ~100-500 SBS/sample vs. WGS thousands) lack power for de novo discovery, prioritizing known etiology assignment over novel signature detection.

The idea behind the mutational signatures

The human genome consists of approximately 3.2 billion nucleotides (bases). The concept of mutational signatures simplifies analysis of the genome by identifying patterns relevant to specific tumors. Counting single nucleotide mutations is not a new idea; scientists in genomics have been measuring single-point mutations for years.

Rather than tracking individual mutations, signatures aggregate them into reproducible patterns reflecting underlying mutational processes (e.g., APOBEC, tobacco, aging). Mutational signatures span SBS (96 trinucleotide contexts), DBS, Indel, SV, and CNA. This study focuses on SBS96, derived from 6 substitution types (C>A, C>G, C>T,

Doctoral study program in field: Applied Informatics,
Supervisor: Assoc. Professor Mgr. Monika Kováčová, PhD., Institute of Informatics,
Information Systems and Software Engineering, Faculty of Informatics and Information Technologies STU in Bratislava

T>A, T>C, T>G) contextualized by flanking bases (trinucleotide expansion from 6 to 96 classes). Each of the 96 channels (e.g., C[A>T]G → C[A>T]C) provides sufficient resolution to distinguish etiologically distinct processes while remaining computationally tractable for NMF decomposition.Types of Mutations

Mutational signatures are classified based on the types of mutations they capture and the underlying biological processes. The main types are :

Single Base Substitution (SBS) signatures: These describe patterns of single nucleotide changes in the genome, specifically the probability of a particular nucleotide substitution at any position. Examples include C>A mutations at CpG dinucleotides (associated with aging and smoking) and T>C mutations at non-CpG sites (often linked to defects in DNA replication and repair).

Doublet Base Substitution (DBS) signatures: These capture patterns of two nucleotide changes occurring in close proximity. Less common than SBS signatures, they provide insights into mutational processes leading to certain diseases.

Insertion-Deletion (Indel) signatures: These describe patterns of nucleotide insertions or deletions. Often associated with DNA replication/repair defects or exposure to chemicals/radiation, they are particularly dangerous due to frameshifts that disrupt the normal codon sequence.

Structural Variant (SV) signatures: These characterize large-scale genomic changes such as translocations, inversions, and deletions. Commonly linked to DNA repair defects, they contribute to diseases like cancer.

Copy Number Alteration (CNA) signatures: These describe changes in the copy number of genes or genomic regions. Associated with DNA replication/repair defects and exposures to chemicals or radiation.

In summary, mutational signatures include SBS, DBS, Indel, SV, and CNA types, each capturing distinct aspects of genomic alterations that contribute to diseases such as cancer.

In this paper we worked with the SBS signatures and Whole Exome data.

Non-Negative Matrix Factorization: the mathematical foundation behind mutational signatures

In refitting, the connection between Single Base Substitution (SBS) signatures and the Non-negative Matrix Factorization (NMF) algorithm is derived by referencing the COSMIC signature catalogue.

The basic formula defining the mutual relationship between mutational signatures, exposure and overall mutational catalogue is:

$$M \approx S * W$$

M – mutational catalogue (the overall number of mutational classes for the sample)

S – mutational signatures

E - exposure (the weight of mutational signature in the mutational catalogue)

In the previous paragraphs, we defined six main mutation types in the introduction. However, using only these six base substitutions provides insufficient resolution to distinguish between illnesses caused by different mutations. Consequently, different tumors exhibited identical mutational signatures. To address this, researchers incorporated the preceding and following bases (trinucleotide context).

This expanded the number of mutational classes to 96—sufficiently detailed to capture mutational processes in individual genome samples, yet computationally feasible with current technology. Applying this method to whole-genome sequencing (WGS) data produces a mutational catalogue: the final observed signature, which reflects contributions from various underlying mutational processes over the cell's lifetime.

Fig. 1. Example of final mutation spectrum formation from contributing mutational processes. Each process contributes to varying degrees, as shown.

Shortcomings of NMF

The number of iterations required for estimating mutational signatures varies by method. Signature extraction typically involves iterative factorization or optimization to identify the optimal set of signatures explaining observed mutation patterns.

For instance, Non-negative Matrix Factorization (NMF often requires hundreds or thousands of iterations for convergence. Factors influencing this include dataset size, number of signatures, and convergence criteria.

Similar iterative processes apply to other methods, such as Bayesian NMF (BNMF) and Non-negative Tensor Factorization (NTF).

In summary, estimating mutational signatures is computationally intensive, with iteration counts depending on the method and dataset complexity. Careful optimization and convergence criteria are essential for robust, accurate results.

DATASET

Original dataset

For signatures extraction we have used the dataset named TRACERx 100: whole exome data of the first 100 TRACERx tumours, EGAS00001002247. This dataset consists of 426 WES samples for NSCLC cancer, 100 of them are control samples. TRACERx (TRAcking Cancer Evolution through therapy (Rx)) is a prospective cohort study designed to

investigate intratumor heterogeneity (ITH) in relation to clinical outcome, and to determine the clonal nature of driver events and evolutionary processes in early stage non-small cell lung cancer (NSCLC).

Multiregion high-depth whole-exome sequencing (M-seq) was performed on 100 early stage NSCLC tumors resected prior to systemic therapy. A total of 327 tumor regions were sequenced and analyzed to define evolutionary histories, obtain a census of clonal and subclonal events, and assess the relationship between ITH and recurrence-free survival (RFS).

The overall ethnicity of sampled patients is white (mainly White-British). Only three patients are of Caribbean origin. The gender is slightly skewed to males – 62 vs 38. The average age of the patient is 68.69 years of age.

Data cleaning and preparation

The dataset was originally sequenced with the reference genome Hg 19 so we have to overlift it to Hg 38 reference genome. The data were filtered using VQSR tool from gatk [6] with the threshold of 90.0, artifacts and decoys were filtered. The more strict filtering would result in very small number of usable mutations.

EXPERIMENT

As a first step in the experiment, matrices with their respective signatures were extracted using the selected refitting tools against COSMIC v3.2, which catalogs 49 curated SBS96 signatures with established etiologies. For reference, version 3.2 of the COSMIC database was used. The results were normalized using the min-max method to ensure comparability across tools and samples despite varying mutation burdens:

$$\frac{(Value - Min_{Value})}{Max_{Value} - Min_{Value}} = Normalized_{Value}$$

Parameter harmonization included fixed signature rank (top-5 by exposure >1% total mutations), convergence tolerance (10^{-6}), and maximum iterations (10,000) where configurable. Five resulting matrices (SigProfiler, SigFit, SigMiner, MusiCaTK, MutationalPatterns) were merged with metadata from the TRACERx 100 database (EGAS00001002247). This metadata included granular smoking status (e.g., Ex-Smoker, Recent Ex-Smoker, Current Smoker, Never Smoked), pack-years, and cessation timing. The first three categories were combined into an "Ex-Smoker" group (n=62 patients), while "Never Smoked" (<100 cigarettes lifetime) formed the second group (n=38). This created two patient groups for comparing signatures between those exposed to tobacco smoke and those who never smoked.

Since the number of samples per patient varied (median 4, range 1–8), we calculated patient-level average exposures to mitigate intratumor heterogeneity.

$$\bar{H}_p = \left(\frac{1}{n_p}\right) \times \Sigma_{i=1}^{n_p} H_i$$

We then filtered to retain only the five highest-ranking signatures per sample (exposure >1%, FDR q<0.05).

The resulting signatures were systematically compared with known tobacco-associated profiles (SBS4, SBS29, SBS92).

The mutational signatures connected to tobacco

Several mutational signatures are directly causally linked to tobacco consumption or strongly correlated with tobacco use, according to published studies.

The most prominent signature directly associated with smoking is **SBS4 [7], [8], [9]**. It is characterized primarily by C>A transversions (also reported in the literature as G>T due to damaged guanine), which are typical of carcinogens in tobacco smoke, such as benzo(a)pyrene (Hollstein et al., 2017). This signature predominates in lung cancer but also appears in head and neck cancers.

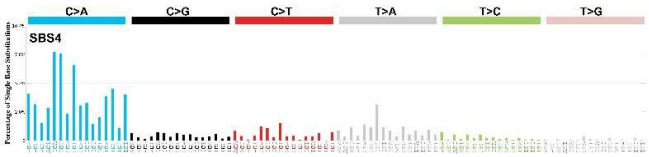

Fig. 2. Mutational profile of SBS4 signature. Mutational profile using the conventional 96 mutation type classification. This classification is based on the six substitution subtypes: C>A, C>G, C>T, T>A, T>C, and T>G, as well as the nucleotides immediately 5' and 3' to the mutation.[10]

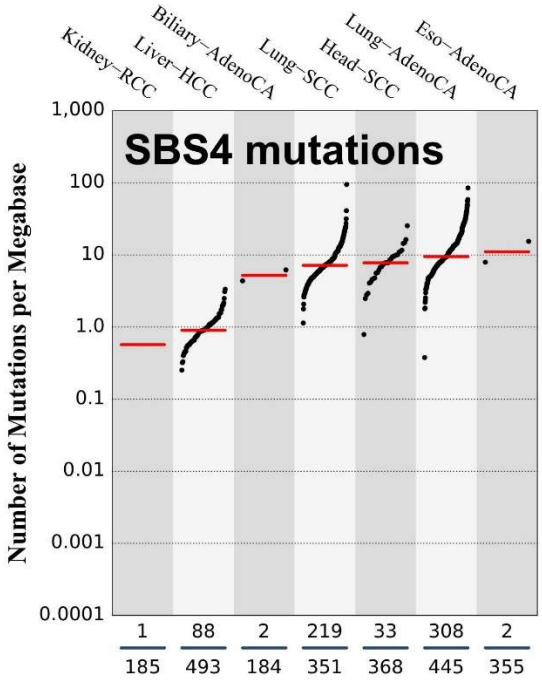

Fig. 3. Numbers of mutations per megabase attributed to the mutational signature across the cancer types in which the signature was found. Each dot represents an individual sample and only samples where the signature is found are shown. The number of mutations per megabase was calculated by assuming that an average whole-exome has 30 Mb with sufficient coverage, whereas an average whole-genome has 2,800 Mb with sufficient coverage.[10]

SBS92 is linked to tobacco consumption and consists mainly of T>C transitions [11]. However, SBS92 activity was found increased in bladder tumours, as well as in normal urothelium of tobacco ever smokers compared to never smokers.[10]. There is no such predominance in NSCLC cancer and is not regarded as the canonical smoking siganture in lung cancer.

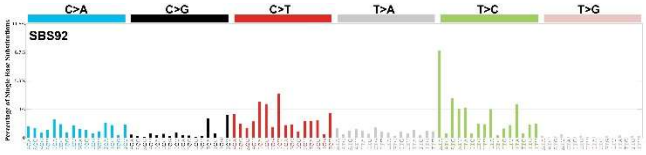

Fig. 4. Mutational profile of SBS4 signature. [10]

SBS29 is strongly associated with tobacco chewing and is prevalent in populations where this habit is common [8]

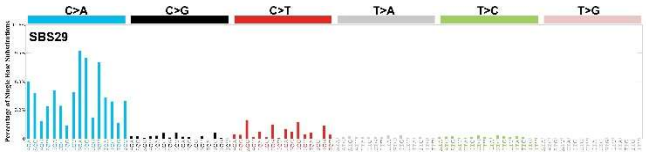

Fig. 5. Mutational profile of SBS29 signature. [10]. This mutational signature strongly resembles the SBS4 mutational signature strongly associated with the lung cancer. Both signatures have elevated C>A point mutations.

Since these signatures were not among the top five detected by the aforementioned tools, we examined signatures correlated with tobacco use. The most common was **SBS5**, a universal clock-like signature detected across nearly all cancer types. Its mutational burden is elevated 1.3–5.1-fold in lifelong smokers compared to never-smokers [12]. The mutational burden of this signature is elevated in bladder cancer samples with ERCC2 mutations as well.[10]

SBS5 is clock-like in that the number of mutations in most cancers and normal cells correlates with the age of the individual. Rates of acquisition of SBS5 mutations over time differ between different cancer types and different normal cell types. These differences do not clearly correlate with estimated rates of stem cell division in different tissues nor with differences in SBS1 mutation rates. SBS5 may be contaminated by SBS16.[10]

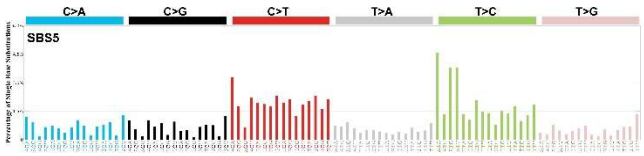

Fig. 6. Mutational profile of SBS29 signature[10].

The second most common signature, **SBS1**, is not correlated with smoking. It is a clock-like signature associated with patient age rather than tobacco exposure [12]

This mutational signature is caused by endogenous mutational process initiated by spontaneous or enzymatic deamination of 5-methylcytosine to thymine which generates G:T mismatches in double stranded DNA. Failure to detect and remove these mismatches prior to DNA replication results in fixation of the T substitution for C.

Rates of acquisition of Signature SBS1 mutations over time differ markedly between different cancer types and different normal cell types. These differences correlate with estimated rates of stem cell division in different tissues and Signature SBS1 may therefore be a cell division/mitotic clock.[10]

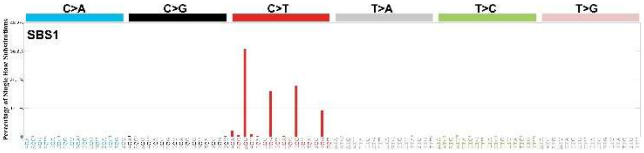

Fig. 7. Mutational profile of SBS1 signature[10].

III. Results and Observations

After applying the selected tools (SigProfiler, SigFit, SigMiner, MusiCaTK, and MutationalPatterns) to extract mutational signatures from exome samples, the results were unexpected. Contrary to our hypothesis that the majority of signatures would be consistent across tools with clear differentiation between ex-smokers and never-smokers, we observed substantial inter-tool variability.

Tool	SBS Mutational Signatures	
	Ex-Smoker	*Never Smoked*
Sigminer	SBS1, SBS5, SBS32, SBS37, SBS39, SBS46, SBS54	SBS1, SBS5, SBS32, SBS37, SBS39, SBS46, SBS54
Musicatk	SBS1, SBS5, SBS32, SBS37, SBS39, SBS46, SBS54	SBS1, SBS5, SBS32, SBS37, SBS39, SBS46, SBS54
sigfit	SBS24, SBS25, SBS37, SBS39, SBS43	SBS24, SBS25, SBS37, SBS39, SBS43
SigProfiler	SBS1, SBS2, SBS5, SBS12, SBS16, SBS32, SBS37, SBS39, SBS54, SBS89	SBS1, SBS2, SBS5, SBS12, SBS16, SBS32, SBS37, SBS54, SBS89
MutPatterns	SBS1, SBS5, SBS6, SBS26, SBS32, SBS33, SBS39, SBS87	SBS1, SBS5, SBS26, SBS32, SBS39, SBS87

Fig. 8. Final mutational signatures for respective tools. Red mutational signatures arose among ex-smokers.

This heterogeneity can be attributed to methodological differences in signature extraction algorithms and inherent parameter settings. Despite harmonizing user-configurable parameters (e.g., signature rank, iteration limits, and convergence criteria), tool-specific defaults (e.g., NMF initialization in SigProfiler vs. Bayesian approaches in SigFit) persisted. Notably, SigMiner and MusiCaTK yielded identical top signatures for most samples, while SigFit consistently produced divergent profiles. SigProfiler and MutationalPatterns exhibited patient-specific variability, resulting in >5 signatures passing our filtering threshold ($q < 0.05$, exposure >1% of total mutations) in the final consensus matrix.

Absence of Tobacco-Associated Signatures

None of the tools reliably detected canonical tobacco-associated signatures (SBS4, SBS29, SBS92) despite their established enrichment in smoking-related cancers [1–3]. The only signature showing modest smoking correlation was SBS5—a ubiquitous clock-like signature present across tumor types and even normal tissues. SBS5 exposures were uniformly elevated (median 25–35% of total SBS burden) across all samples, precluding its use as a smoking discriminator (Wilcoxon rank-sum test, $p > 0.1$ between groups).

Lack of Differentiation Between Smoking Groups

TRACERx 100 metadata classified patients by smoking status (Current Smoker, Recent Ex-Smoker, Ex-Smoker, Never Smoker). We aggregated these into two metagroups: Ex-Smokers ($n = 62$) and Never Smokers ($n = 38$). Cosine similarity of results revealed near-identical profiles (mean cosine similarity = 0.96 ± 0.08).

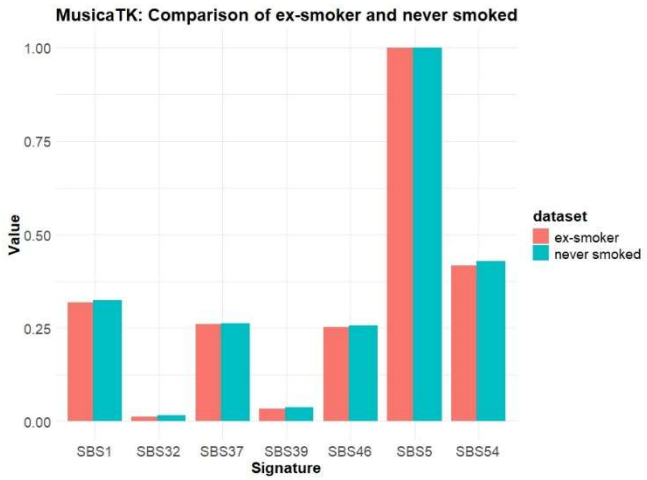

Fig. 9. Comparison of the volume of respective signatures among ex-smokers and never smoked groups in MusicaTK.

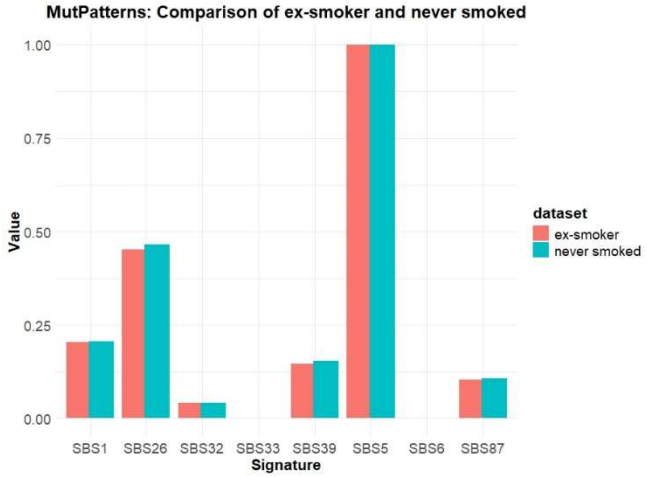

Fig. 10. Comparison of the volume of respective signatures among ex-smokers and never smoked groups in Mutational Patters.

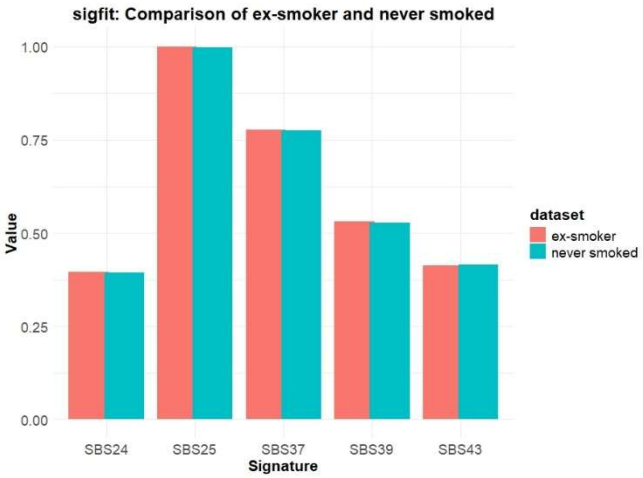

Fig. 11. Comparison of the volume of respective signatures among ex-smokers and never smoked groups in sigfit.

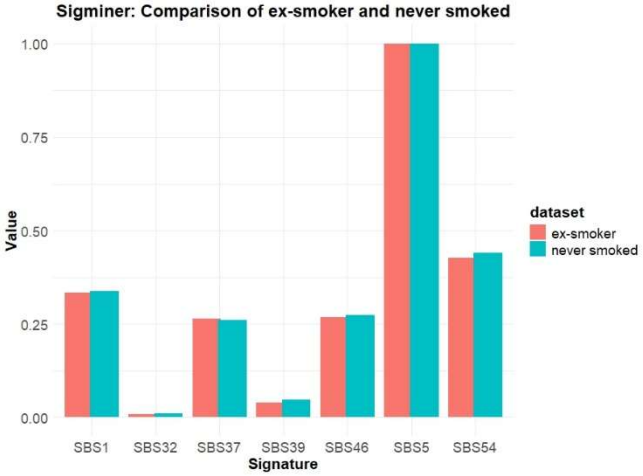

Fig. 12. Comparison of the volume of respective signatures among ex-smokers and never smoked groups in Sigminer.

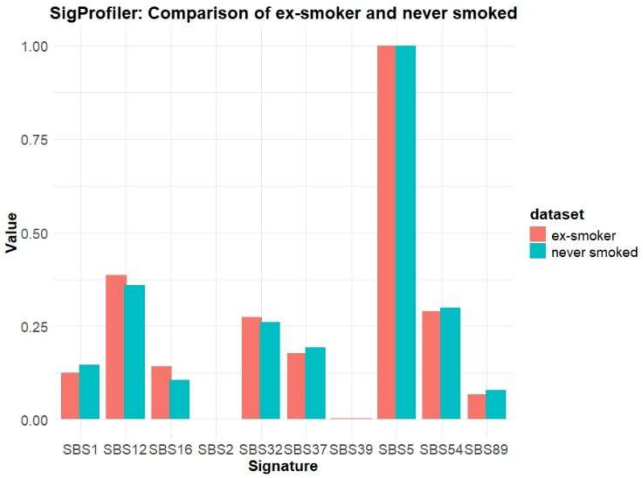

Fig. 13. Comparison of the volume of respective signatures among ex-smokers and never smoked groups in Sigprofiler.

Dataset Limitations

Our analysis was constrained to whole-exome sequencing (WES) data (~1–2% genomic coverage), focusing on protein-coding regions. While exomes capture functionally critical mutations, they undersample non-coding regions where tobacco signatures may accumulate preferentially. We anticipated at least directional trends (e.g., SBS4 enrichment in ex-smokers), but power calculations suggest >500 samples per group are needed for 80% power to detect 5% absolute exposure differences ($\alpha = 0.05$).

IV. CONCLUSION

This study demonstrates that mutational signature analysis in whole-exome sequencing (WES) data cannot reliably distinguish between ex-smokers and never-smokers with NSCLC, which questions its current utility for clinical smoking inference. Despite utilizing five different refitting tools and harmonizing parameters, canonical tobacco signatures (SBS4, SBS29, SBS92) remained undetectable. Instead, mutational profiles were dominated by ubiquitous "clock-like" signatures SBS1 and SBS5, resulting in a near-identical mean cosine similarity of 0.96 between groups.

These null results are largely attributed to the limitations of WES, which covers only 1–2% of the genome, thereby "diluting" the detectability of tobacco carcinogens like benzo[a]pyrene that induce signatures genome-wide.

Furthermore, these findings highlight significant inter-tool discordance and sensitivity disparities among extraction tools. Even with parameter harmonization, algorithmic differences—such as NMF rank selection in SigProfiler, sparsity penalties in SigFit, or ensemble approaches in MutationalPatterns—yield divergent outputs. In the absence of a "COSMIC-endorsed gold standard," tool selection remains a subjective decision that complicates reproducibility and heavily influences experimental outcomes.

Ultimately, these discrepancies underscore an urgent need for the transition to whole-genome sequencing (WGS), the use of larger cohorts (>1000 samples) to achieve sufficient statistical power, and the development of standardized pipelines. Future work should also prioritize tool benchmarking via simulated datasets to quantify false negative rates and the implementation of exome-to-genome correction factors to improve the accuracy of smoking signature studies.

ACKNOWLEDGMENT

This research project received financial support from VEGA under grant number 1/0603/24, as well as from the Slovak Research and Development Agency through grants VV-MVP-24-0290 and APVV-21-0448.

Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project 07I04-04-V05-00114 (AI-ONKO).

REFERENCES

[1] S. Wang *et al.*, "Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes," *PLoS Genet.*, vol. 17, no. 5, p. e1009557, 2021, [Online]. Available: https://doi.org/10.1371/journal.pgen.1009557

[2] K. Gori and A. Baez-Ortega, "sigfit: flexible Bayesian inference of mutational signatures," *bioRxiv*, p. 372896, 2018, doi: 10.1101/372896.

[3] Y. A. Kim, M. D. M. Leiserson, P. Moorjani, R. Sharan, D. Wojtowicz, and T. M. Przytycka, "Mutational Signatures: From Methods to Mechanisms," *Annu. Rev. Biomed. Data Sci.*, vol. 4, pp. 189–206, 2021, doi: 10.1146/annurev-biodatasci-122320-120920.

[4] A. Chevalier, N. Gurevich, T. Guo, and J. D. Campbell, "musicatk: Mutational Signature Comprehensive Analysis Toolkit," 2025. doi: 10.18129/B9.bioc.musicatk.

[5] M. Díaz-Gay, R. Vangara, M. Barnes, and L. B. Alexandrov, "Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment," *Bioinformatics*, 2023, doi: 10.1093/bioinformatics/btad756.

[6] Broad Institute, "Genome Analysis Toolkit (GATK)," 2025.

[7] M. Hollstein, L. B. Alexandrov, C. P. Wild, M. Ardin, and J. Zavadil, "Base changes in tumour DNA have the power to reveal the causes and evolution of cancer," *Oncogene*, vol. 36, pp. 158–167, 2017.

[8] A. R. Huber, A. Van Hoeck, and R. Van Boxtel, "The Mutagenic Impact of Environmental Exposures in Human Cells and Cancer: Imprints Through Time," *Front. Genet.*, vol. 12, p. 760039, Oct. 2021.

[9] A. L. Moreira and J. K. Sabari, "Seeing Beyond the Smoke: Reclassifying Lung Cancer by Smoking-Related Mutational Signatures," *Journal of Thoracic Oncology*, vol. 18, no. 4, pp. 396–398, Apr. 2023.

[10] J. G. Tate *et al.*, "COSMIC: the Catalogue Of Somatic Mutations In Cancer," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D941–D947, Jan. 2019, doi: 10.1093/nar/gky1015.

[11] L. Torrens, S. Moody, A. C. de Carvalho, and others, "The complexity of tobacco smoke-induced mutagenesis in head and neck cancer," *Nat. Genet.*, vol. 57, pp. 884–896, 2025, doi: 10.1038/s41588-025-02134-0.

[12] L. B. Alexandrov *et al.*, "Mutational signatures associated with tobacco smoking in human cancer," *Science (1979).*, vol. 354, no. 6312, pp. 618–622, Nov. 2016, doi: 10.1126/science.aag0299.