Emerging Approaches for Tumor Analyses
in Epidemiological Studies

Session 5: Mutational Signatures

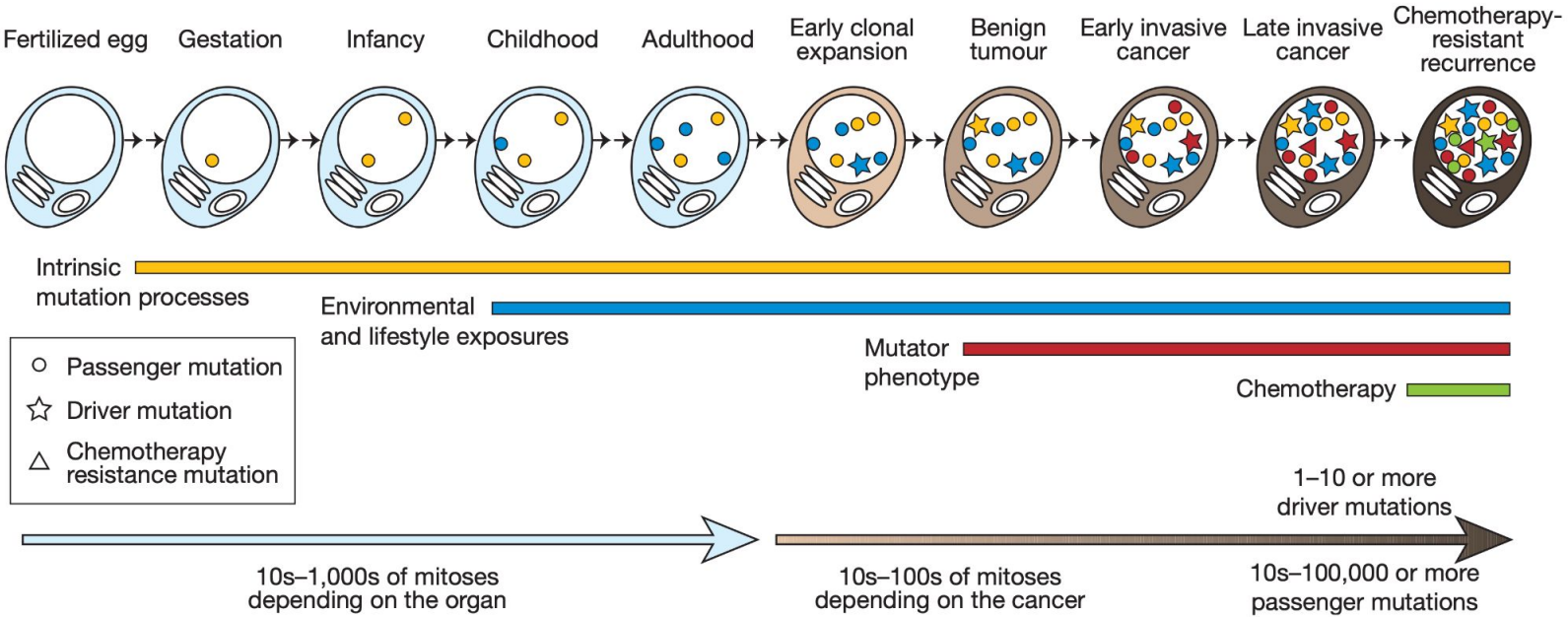
January 18, 2023
9:30 AM- 12:00 PM

Overview

- ➔ Mutational signature background
- ➔ **De novo identification of mutational signatures**
- ➔ **Decomposition mutational signatures based on known reference signatures**
- ➔ Emerging mutational signatures in cancer genomic studies
- ➔ Downstream analysis for mutational signature data
- ➔ Practical session (Mutational Signature analysis and explore data portals)

Mutational signature background

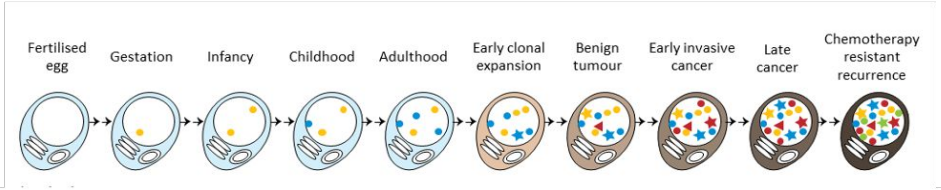
Acquisition of somatic mutations in cancer genomes



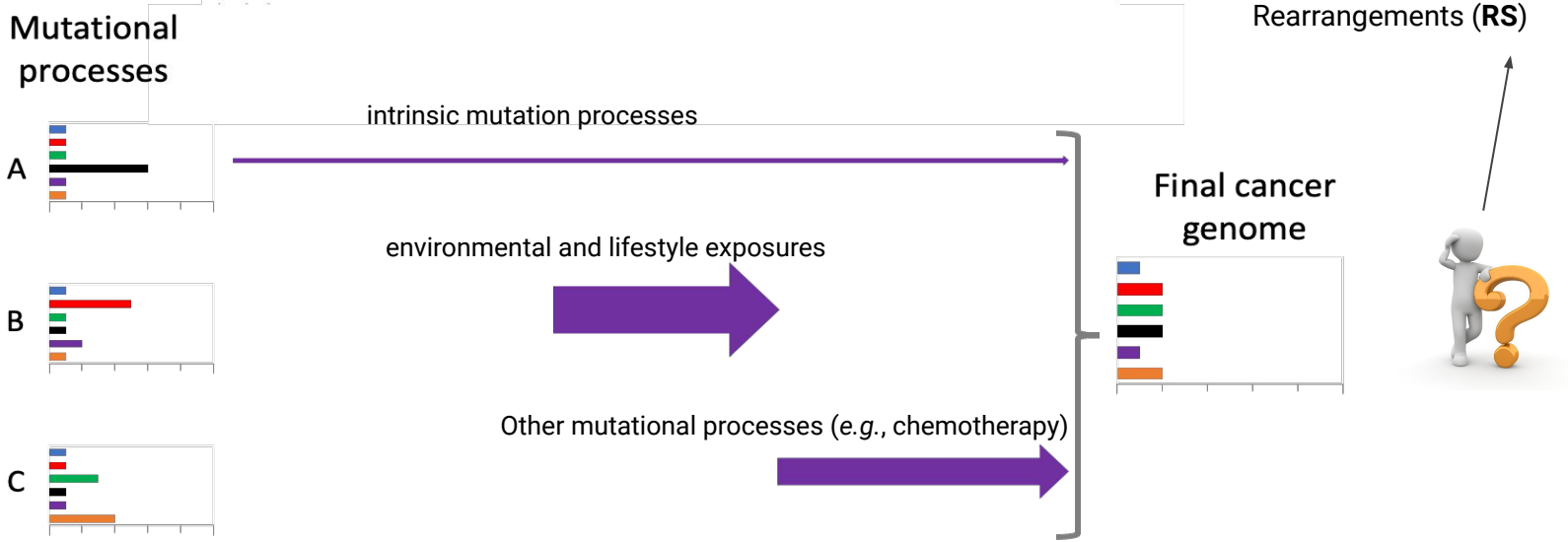
Stratton et al., *Nature*, 2009

The catalogue of somatic mutations in a cancer genome

The final cancer genome represents an archaeological record of the effect of the different mutagenic and DNA repair processes

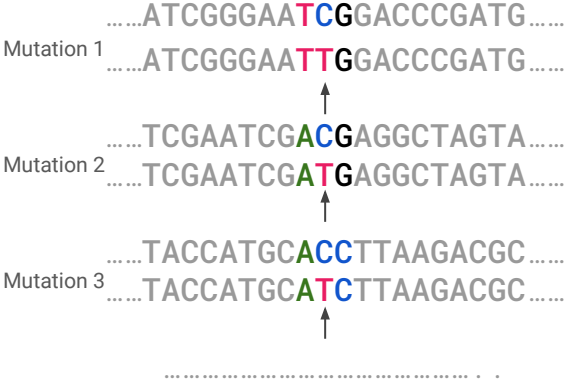


Single Base Substitutions (**SBS**)
Doublet Base Substitutions (**DBS**)
Small insertions and deletions (**ID**)
Copy number (**CN**)
Rearrangements (**RS**)

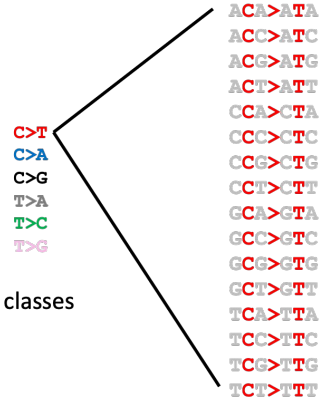


Mutations → mutational profiles/spectra (*e.g.*, SBS96)

Mutation Calling from one sample

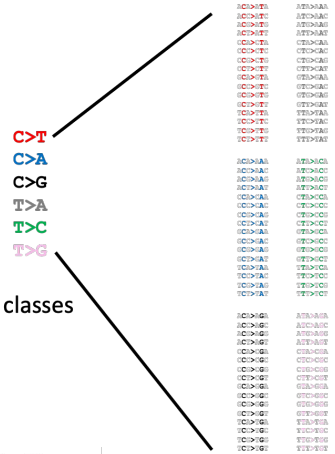


Six classes of single-base mutations (Reported by pyrimidine)



6 mutation classes

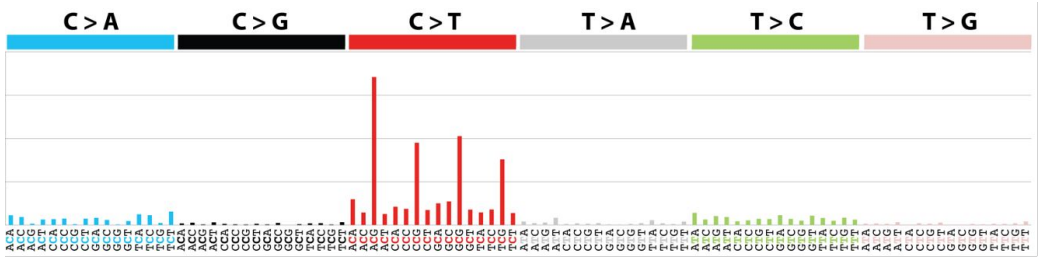
96 possibilities considering context (adding 5’ and 3’ adjacent bases)



6 mutation classes

96 mutation subtypes

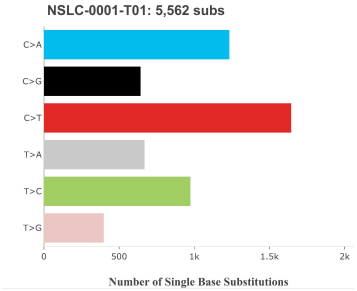
SBS96



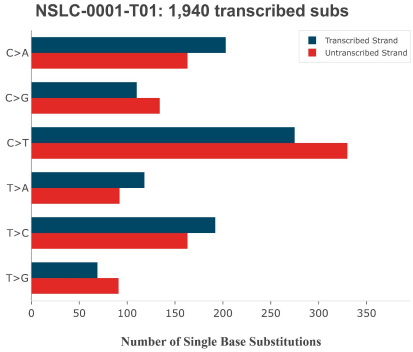
Mutational signatures can be determined based on the mutational profiles across a set of individuals

SBS Mutational profiles

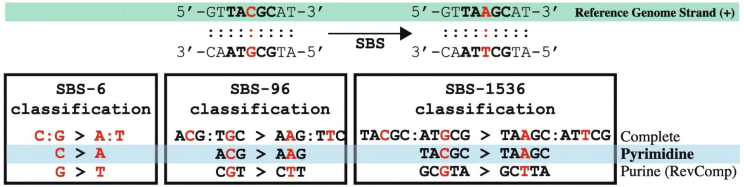
SBS6



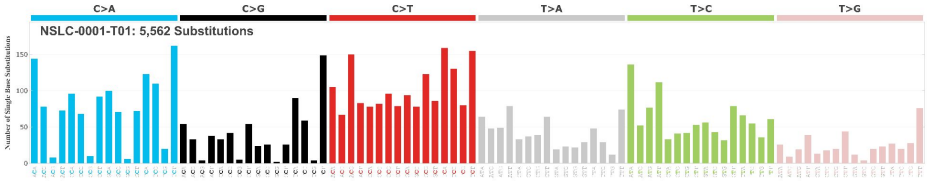
SBS12



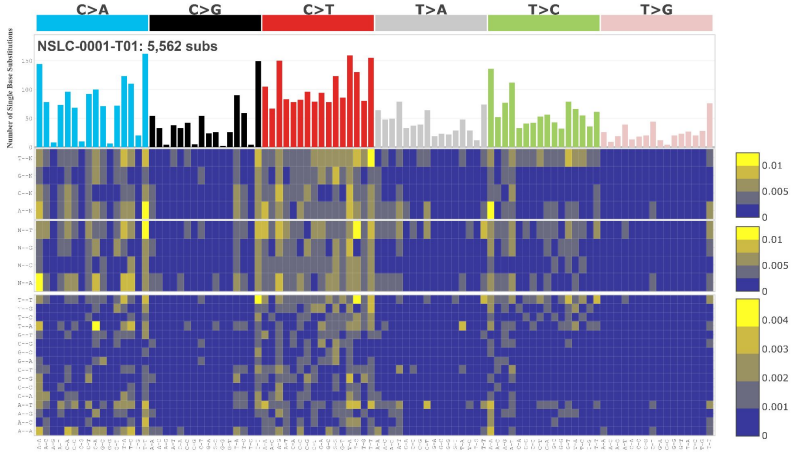
Example



SBS96

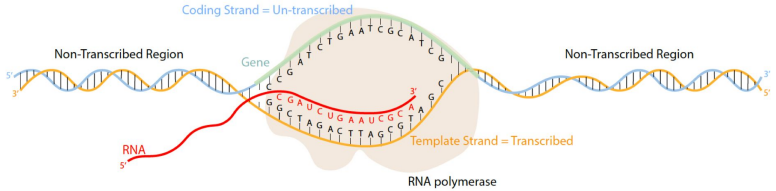


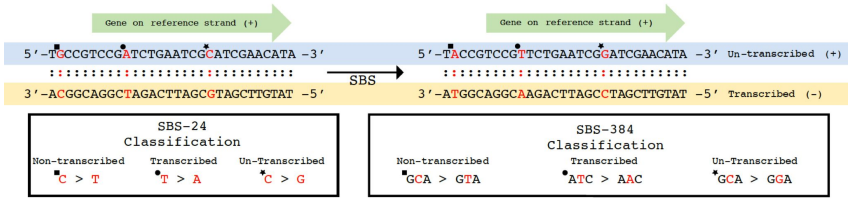
SBS1536



SBS Mutational profiles

Transcribed strand information

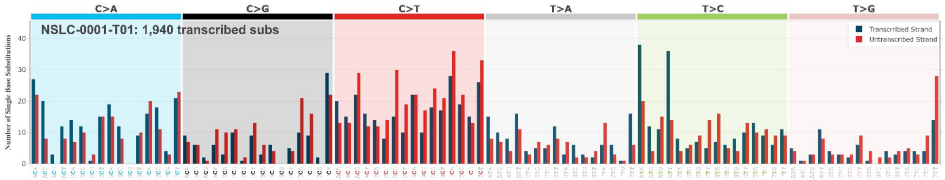




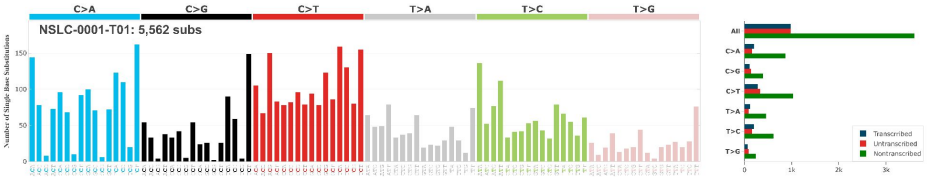
Transcribed Region: (**Transcribed** or **Un-transcribed** strand)

Non-Transcribed Region: **Non-transcribed**

SBS192

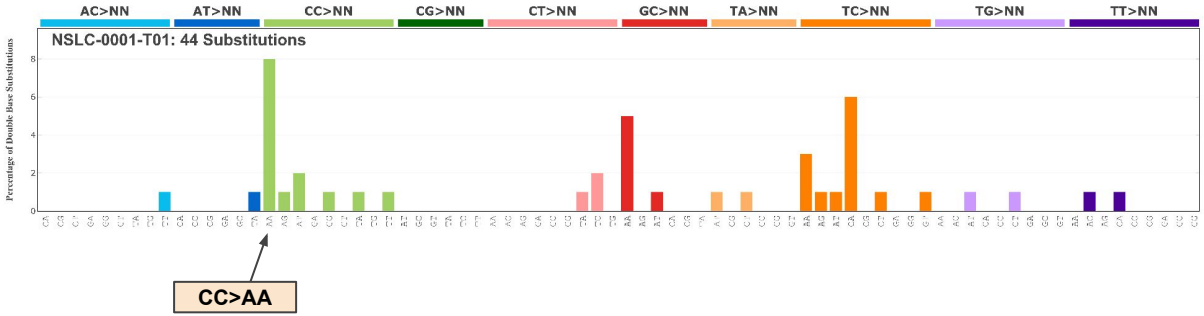


SBS288



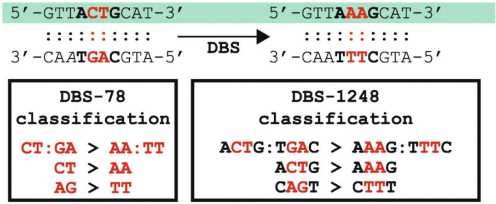
DBS Mutational profiles

- DBS are generated after the concurrent modification of two consecutive nucleotide bases.
- There are 78 strand-agnostic DBS mutation types
- More specifically, there are 16 possible source doublet bases (4 x 4)
- Of these, AT, TA, CG, and GC are their own reverse complement
- The remaining 12 can be represented as 6 possible strand-agnostic doublets
- Thus, there are 4+6=10 source doublet bases
- Because they are their own reverse complements, AT, TA, CG, and GC can each be substituted by only 6 doublets
- For the remaining doublets, there are 9 possible DBS mutation types (3 x 3)
- Therefore, in total there are 4 x 6 + 6 x 9 = **78** strand-agnostic DBS mutation types.



DBS78

Example



Other uncommon DBS profiles: DBS150/DBS186/DBS1248/DBS2400/DBS2976.

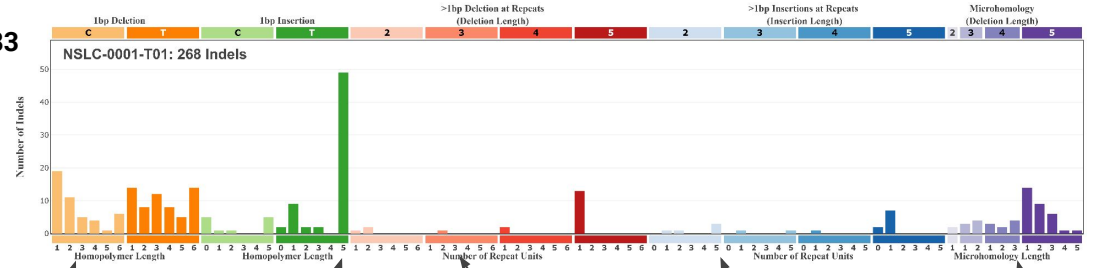
Check the SigProfilerMatrixGenerator for details;

Bergstrom et al., *BMC Genomics*, 2019

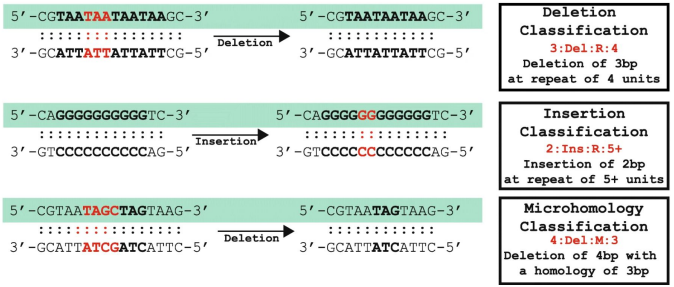
ID Mutational profiles

- Also known as indels, ID are defined as the incorporation or loss of small fragments of DNA (usually between 1 and 50 base pairs) in a specific genomic location
- Although there is no single intuitive and naturally constrained set of ID mutation types (as there arguably are for single base substitutions and doublet base substitutions), a compilation of **83** different types considering size, nucleotides affected and presence on repetitive and/or microhomology regions was used to extract mutational signatures.
- Other uncommon ID profiles: ID28/ID29/ID96/ID166/ID332/ID415/ID8628.
- More details can be found here: https://cancer.sanger.ac.uk/signatures/documents/4/PCAWG7_indel_classification_2021_08_31.xlsx

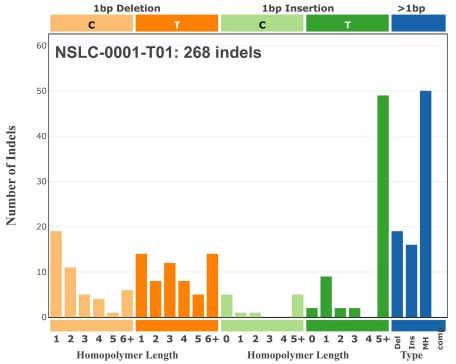
ID83



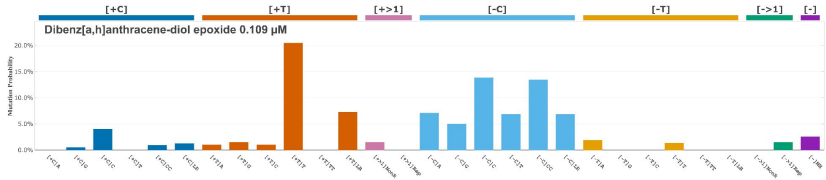
Example



ID28



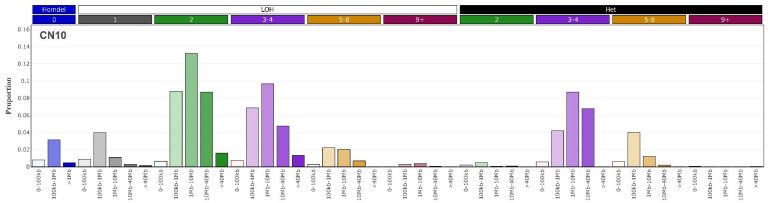
ID29



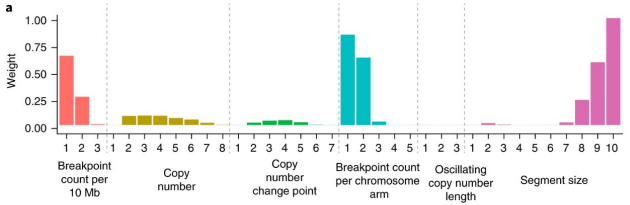
CN or RS Mutational profiles

CN48

- Copy number variants are characterized using a **48**-channel copy number classification scheme
- To categorise segments from allele-specific copy number profiles (as major copy number and minor copy number respectively i.e. non-phased profiles) the scheme incorporates: **loss-of-heterozygosity status**, **total copy number state**, **segment length.**

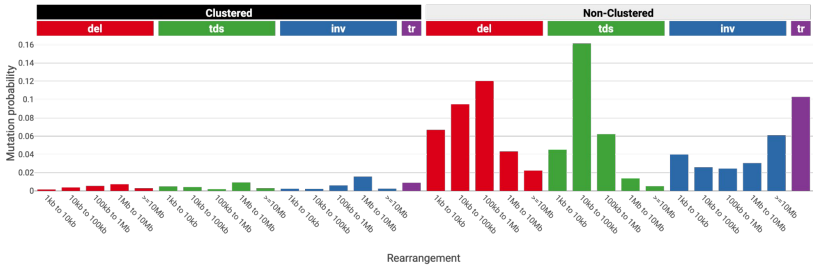


Steele, et al., *Nature*, 2022



Macintyre, et al., *Nature Genetics*, 2018

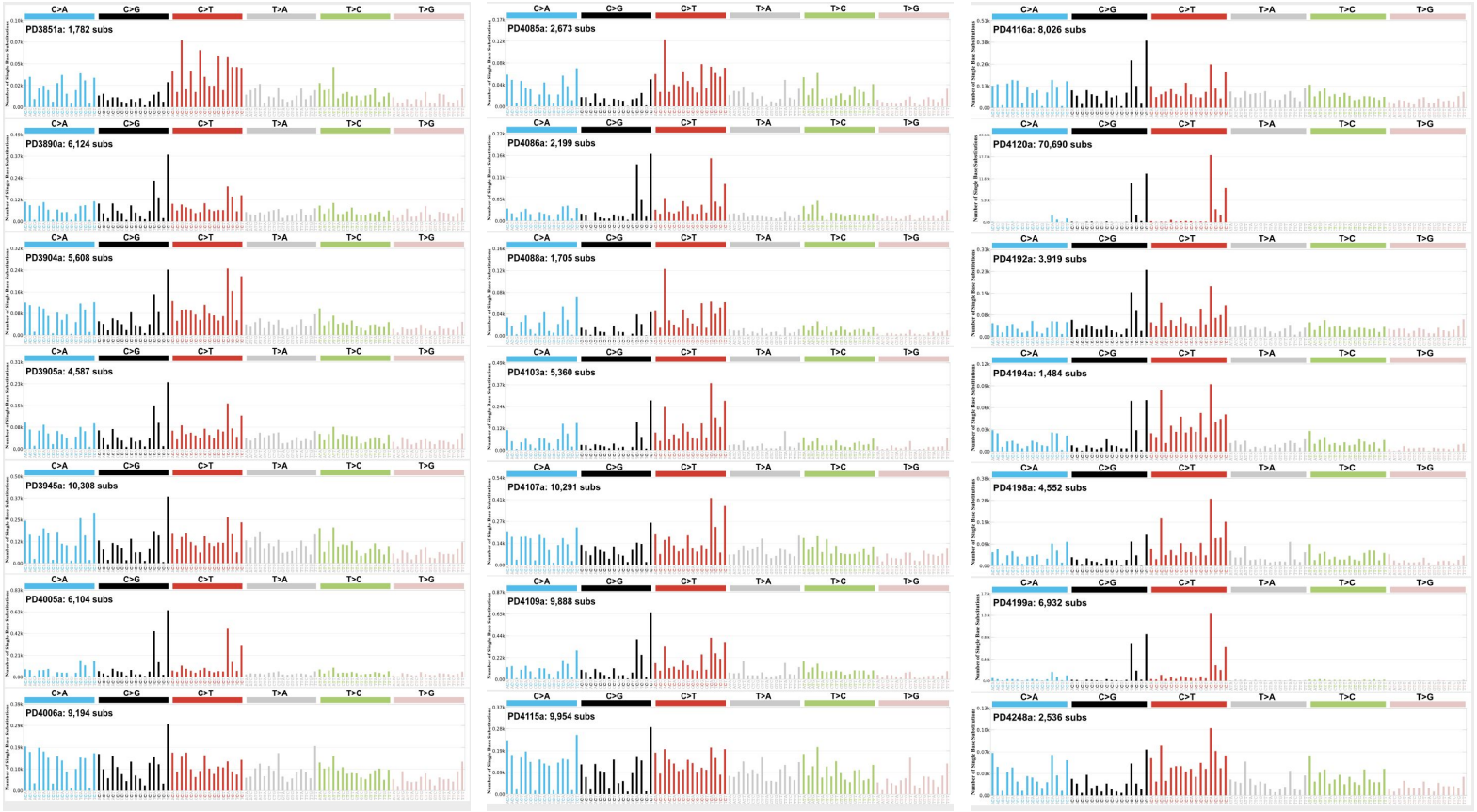
RS32



Degasperi, et al., *Nature Cancer*, 2020

- Rearrangement fragments are broadly categorised based on four types of rearrangements, namely tandem duplications (**tds**), deletions (**del**), inversions (**inv**), and translocations (**tr**), with further consideration of **sizes** and **clustering** of the rearranged fragments.

The life History of 21 Breast Cancers

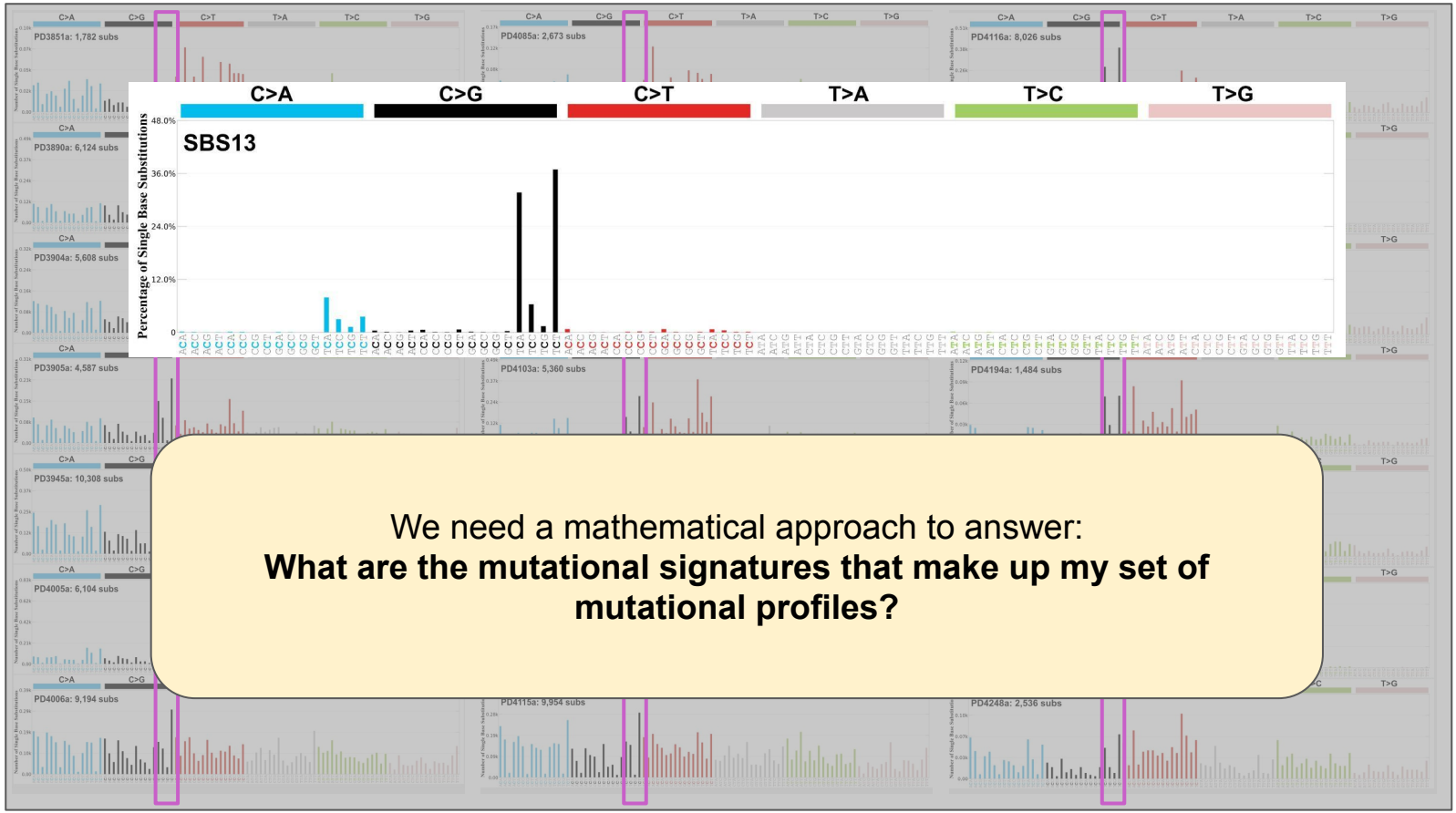


Nik-Zainal et al., *Cell*, 2012

The life History of 21 Breast Cancers

Nik-Zainal et al., *Cell*, 2012

The life History of 21 Breast Cancers



We need a mathematical approach to answer:
What are the mutational signatures that make up my set of mutational profiles?

Nik-Zainal et al., *Cell*, 2012

De novo identification of mutational signatures

Computational identification of mutational signatures

- Mutational signatures can be determined based on mutational profiles across a set of individuals

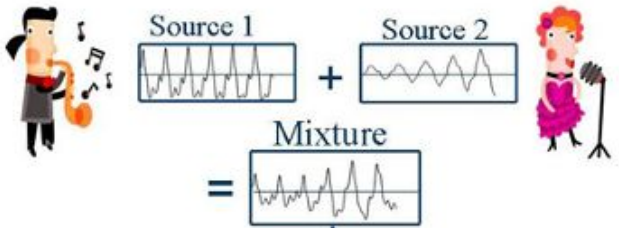
Computational identification of mutational signatures

– – –

- Mutational signatures can be determined based on mutational profiles across a set of individuals
- Mathematical models allows the *un-mixing* and extraction of mutational signatures by solving a **blind source separation problem**

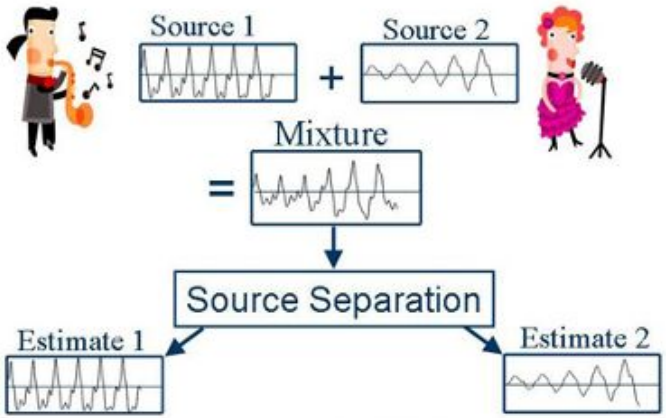
Computational identification of mutational signatures

- Mutational signatures can be determined based on mutational profiles across a set of individuals
- Mathematical models allows the *un-mixing* and extraction of mutational signatures by solving a **blind source separation problem**



Computational identification of mutational signatures

- Mutational signatures can be determined based on mutational profiles across a set of individuals
- Mathematical models allows the *un-mixing* and extraction of mutational signatures by solving a **blind source separation problem**



Computational identification of mutational signatures

– – –

- **Non-negative matrix factorization (NMF)** for solving the blind source separation (BSS) problem

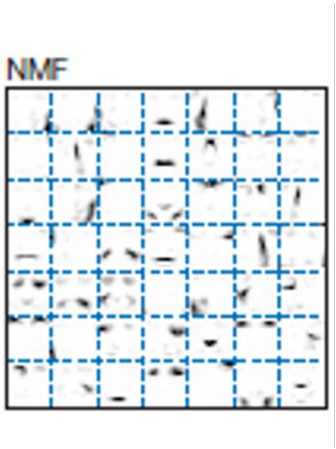
Computational identification of mutational signatures

- **Non-negative matrix factorization (NMF)** for solving the blind source separation (BSS) problem

Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung*†

* *Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA*
† *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*



Lee & Seung 1999 Nature

Computational identification of mutational signatures

– – –

- **Non-negative matrix factorization (NMF)** for solving the blind source separation (BSS) problem
  - Infinite solutions as a matrix can be approximately decomposed into two matrices in an infinite number of ways
  - BSS problem is usually solved by constraining the solutions
  - Intrinsic nonnegative constraints from our theoretical model
  - One main hyperparameter, the rank *k* of the latent matrices S and A, which corresponds to the number of mutational signatures present in the input data (matrix M)

Computational identification of mutational signatures

- **Non-negative matrix factorization (NMF)** for solving the blind source separation (BSS) problem
  - Infinite solutions as a matrix can be approximately decomposed into two matrices in an infinite number of ways
  - BSS problem is usually solved by constraining the solutions
  - Intrinsic nonnegative constraints from our theoretical model
  - One main hyperparameter, the rank *k* of the latent matrices S and A, which corresponds to the number of mutational signatures present in the input data (matrix M)

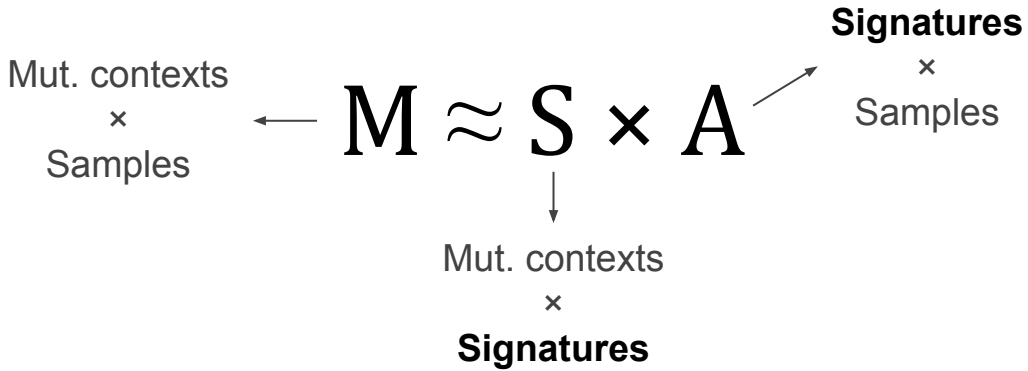
$$M \approx S \times A$$

Computational identification of mutational signatures

- **Non-negative matrix factorization (NMF)** for solving the blind source separation (BSS) problem
  - Infinite solutions as a matrix can be approximately decomposed into two matrices in an infinite number of ways
  - BSS problem is usually solved by constraining the solutions
  - Intrinsic nonnegative constraints from our theoretical model
  - One main hyperparameter, the rank *k* of the latent matrices S and A, which corresponds to the number of mutational signatures present in the input data (matrix M)

Computational identification of mutational signatures

- **Non-negative matrix factorization (NMF)** for solving the blind source separation (BSS) problem
  - Infinite solutions as a matrix can be approximately decomposed into two matrices in an infinite number of ways
  - BSS problem is usually solved by constraining the solutions
  - Intrinsic nonnegative constraints from our theoretical model
  - One main hyperparameter, the rank *k* of the latent matrices S and A, which corresponds to the number of mutational signatures present in the input data (matrix M)



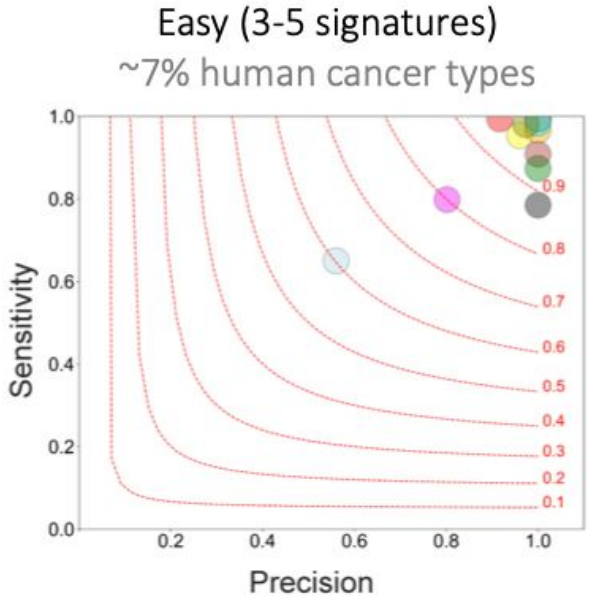
Benchmark of tools for *de novo* signature extraction

Tool	Platform	Factorization Approach		Selection Approach		Reference
		Method	Computational Engine	Type	Algorithm	
EMu	C++	EM	Original implementation	M/A	BIC	Fischer *et al.* 2013
Maftools	R-Bioconductor	NMF	NMF R package	M	-	Mayakonda *et al.* 2018
MutationalPatterns	R-Bioconductor	NMF	NMF R package	M	-	Blokzijl *et al.* 2018
MutSignatures	R	NMF	Brunet *et al.* 2004	-	-	Fantini *et al.* 2020
MutSpec	R/Galaxy	NMF	NMF R package	M	-	Ardin *et al.* 2016
SigFit	R	Bayesian inference	Stan R package	M/A	Elbow method	Gori *et al.* 2020
SigMiner	R	NMF/Bay. NMF	NMF R package/SA	M/A	ARD	Wang *et al.* 2021
SignatureAnalyzer	R/Python	Bayesian NMF	Original implementation	A	ARD	Kasar *et al.* 2015
SignatureToolsLib	R	NMF	NMF R package	M	-	Degasperi *et al.* 2020
SigneR	C++/R-Bioconductor	Bayesian NMF	Original implementation	M/A	BIC	Rosales *et al.* 2017
SigProfilerExtractor	Python/R	NMF	Original implementation	M/A	NMFk	Islam *et al.* 2021
SigProfiler_PCAWG	Python/MATLAB	NMF	Brunet *et al.* 2004	M	-	Alexandrov *et al.* 2013
SomaticSignatures	R-Bioconductor	NMF	NMF R package	M	-	Gehring *et al.* 2015
TensorSignatures	Python	NTF	TensorFlow	M/A	BIC	Vöhringer *et al.* 2021

Islam *et al.* 2022
Cell Genomics

Benchmark of tools for *de novo* signature extraction

– – – **No noise (WGS synthetic data)**

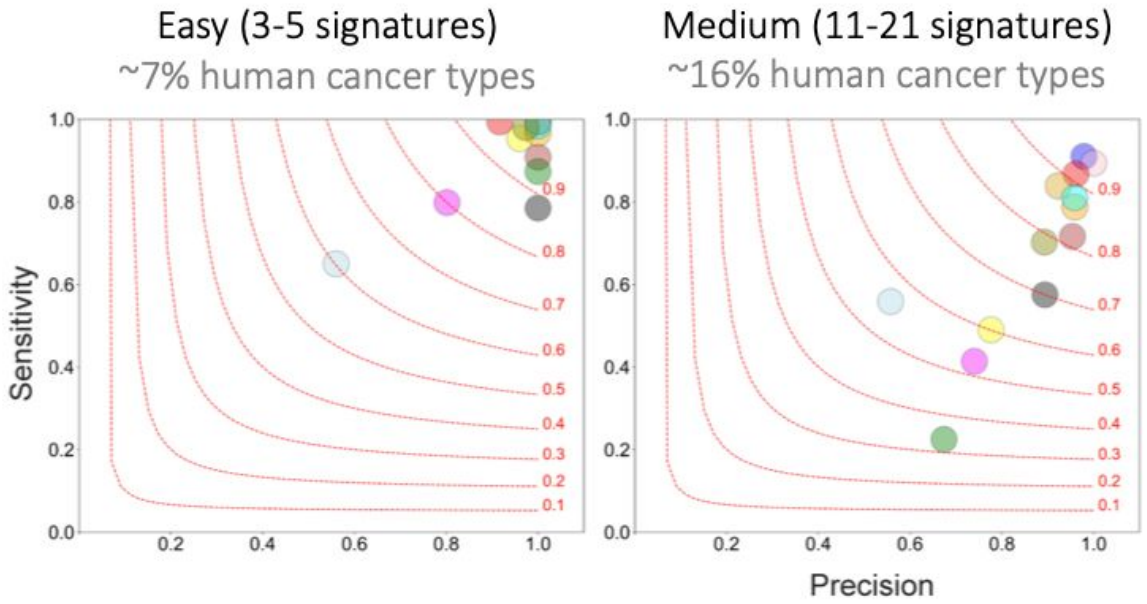


SigProfilerExtractor
SignatureAnalyzer
SigProfiler_PCAWG
SigneR
MutationalPatterns
SomaticSignatures
MutSpec
SignatureToolsLib
SigMiner
MutSignatures
Maftools
EMu
SigFit
TensorSignatures

Islam *et al.* 2022 Cell Genomics

Benchmark of tools for *de novo* signature extraction

– – – **No noise (WGS synthetic data)**

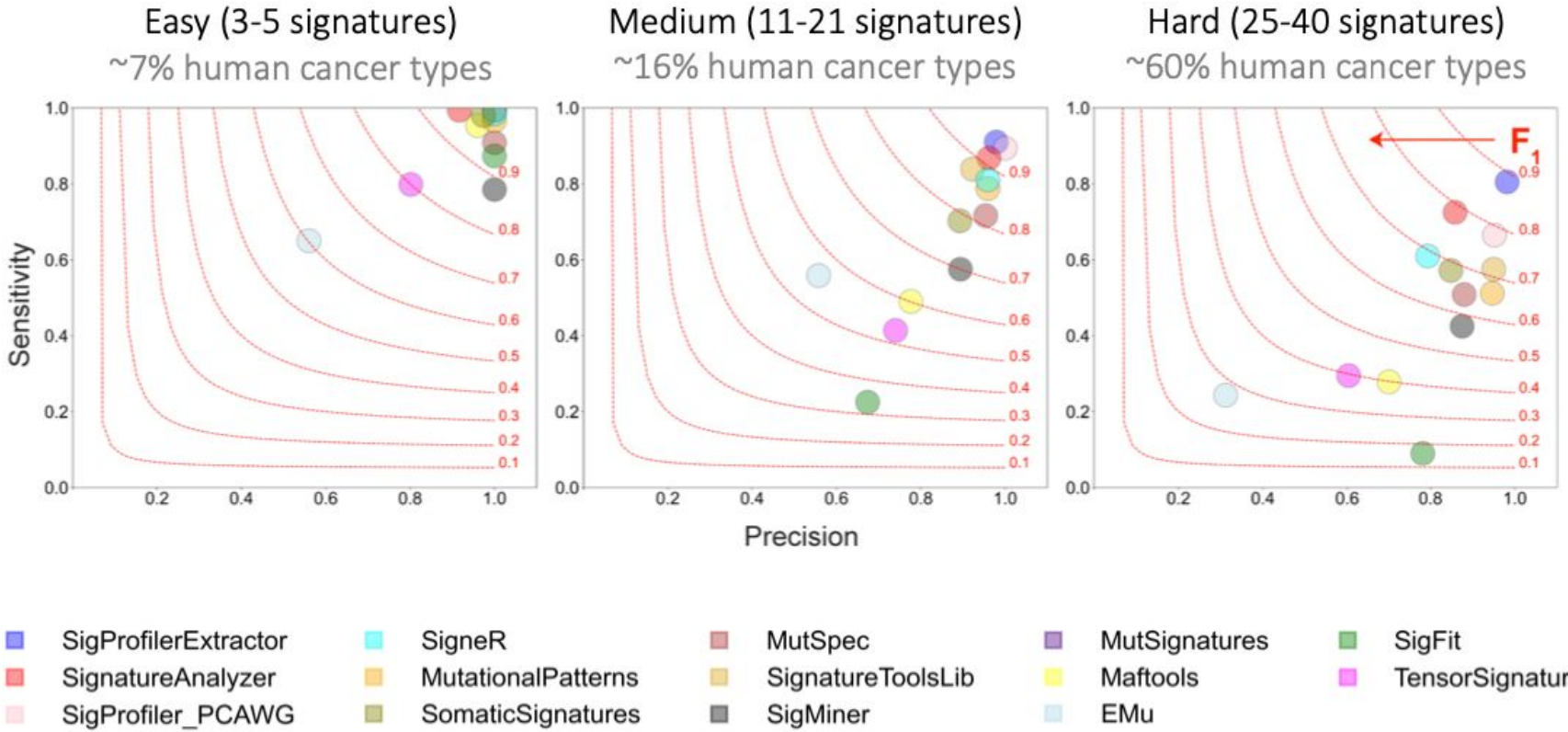


SigProfilerExtractor
SignatureAnalyzer
SigProfiler_PCAWG
SigneR
MutationalPatterns
SomaticSignatures
MutSpec
SignatureToolsLib
SigMiner
MutSignatures
Maftools
EMu
SigFit
TensorSignatures

Islam *et al.* 2022 Cell Genomics

Benchmark of tools for *de novo* signature extraction

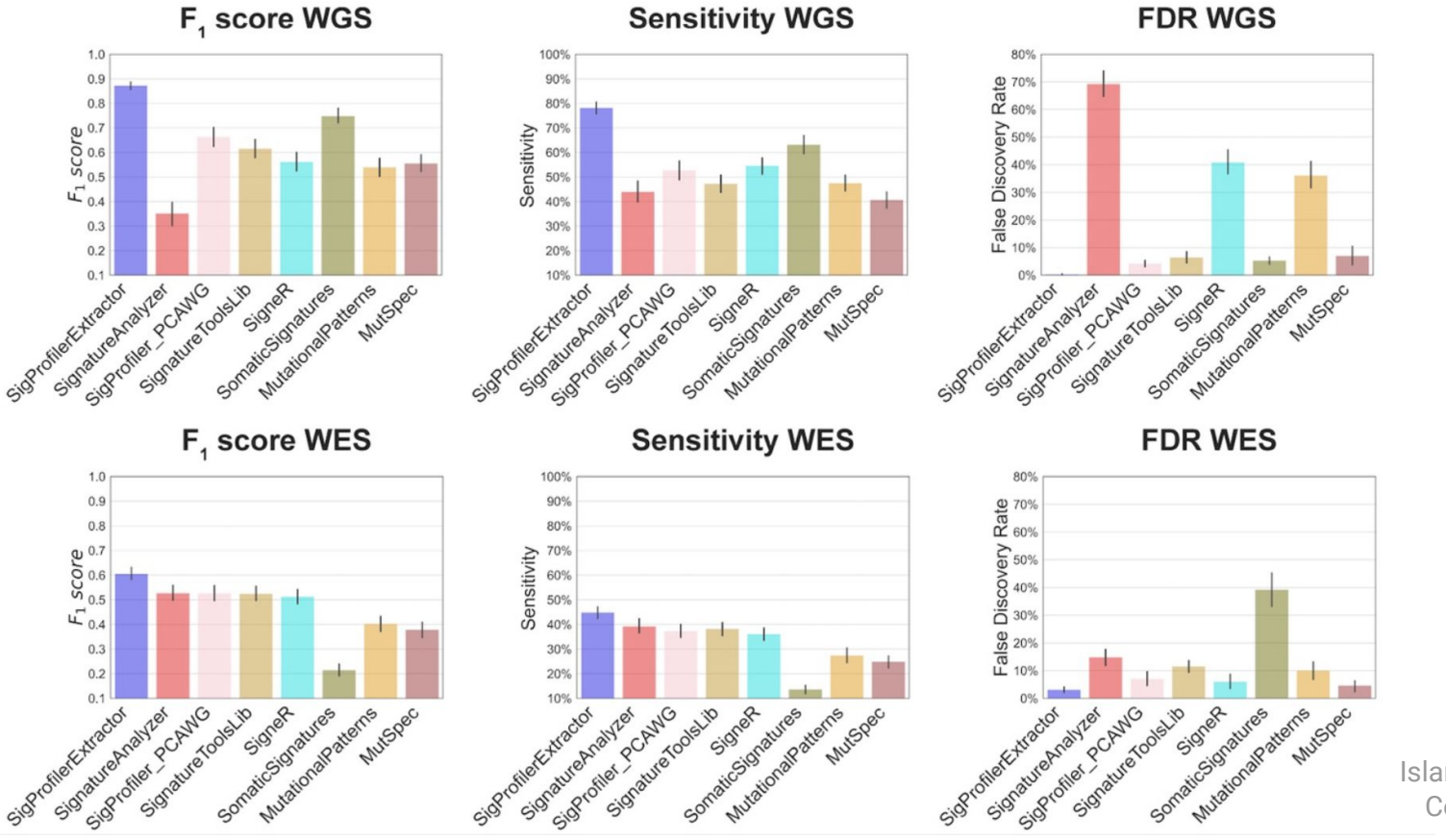
– – – **No noise (WGS synthetic data)**



Islam *et al.* 2022 Cell Genomics

Benchmark of tools for *de novo* signature extraction

– – – **5% noise**



Islam *et al.* 2022
Cell Genomics

Benchmark of tools for *de novo* signature extraction

- Although most tools achieved high performance on easy scenarios using noiseless synthetic data, this is not the case for medium or hard scenarios
- When the number of signatures increases, different tools experience drops in both sensitivity and precision
- This reduced performance is more noticeable in hard scenarios, based in over 25 signatures, and representing >60% of human cancer types
- As real sequencing data contains different levels of noise, it is important to consider it in the benchmarking
- When noise is introduced in the synthetic dataset, some of the top performing tools without noise for WGS data suffer a reduced precision, giving rise to false positive signatures
- Benchmarking with WES synthetic data did not achieve 50% sensitivity for any tool, indicating the lack of statistical power to identify all signatures present in these data

Reference mutational signatures

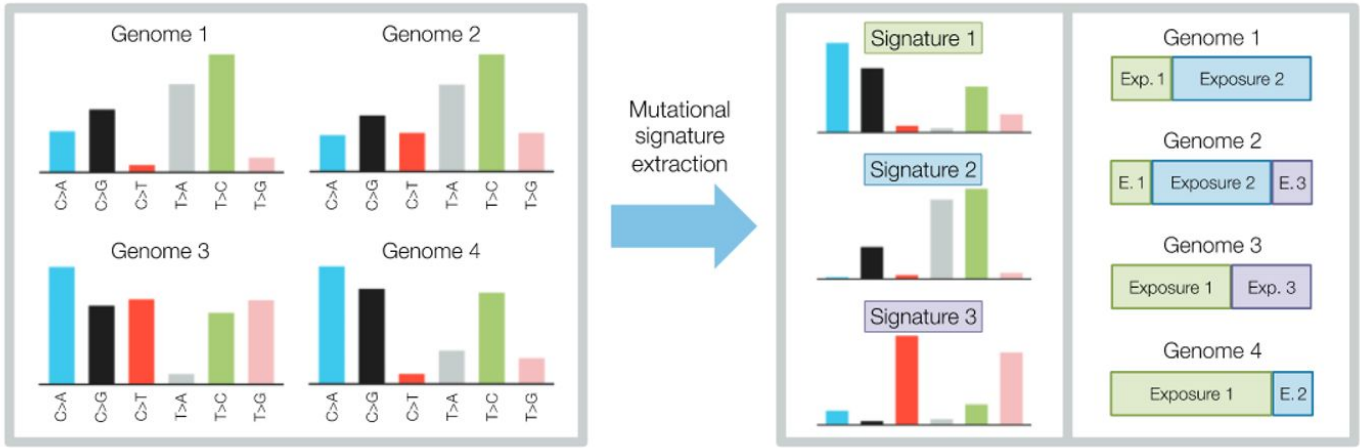
- Mutational signature extraction relies on a large number of samples (and mutations) to get accurate results

Reference mutational signatures

- Mutational signature extraction relies on a large number of samples (and mutations) to get accurate results
- Large international consortia have analyzed thousands of whole genome and whole exome sequenced samples to generate a consensus set of reference mutational signatures (deposited in the COSMIC database)

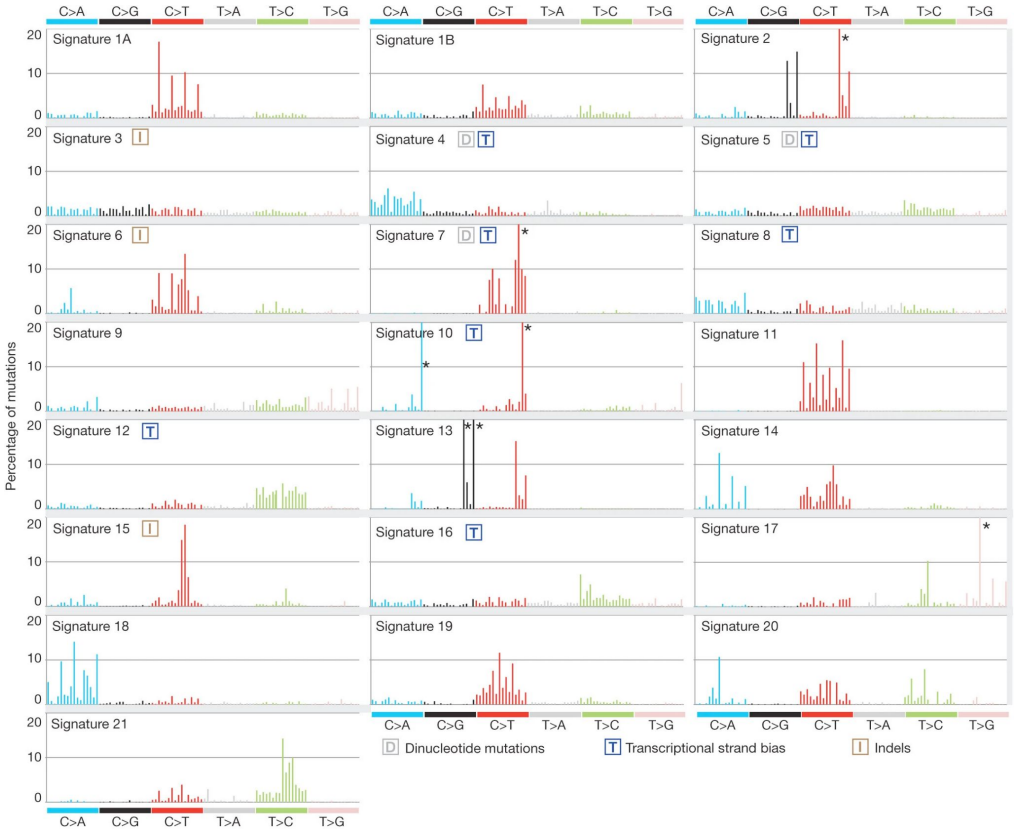
Reference mutational signatures

- Mutational signature extraction relies on a large number of samples (and mutations) to get accurate results
- Large international consortia have analyzed thousands of whole genome and whole exome sequenced samples to generate a consensus set of reference mutational signatures (deposited in the COSMIC database)



Baez-Ortega & Gori 2019 Briefings in Bioinformatics

Reference mutational signatures



The COSMIC database has been growing over the years with the addition of novel samples and different variant classes

v1 (August 2013)

- 22 SBS signatures

Reference mutational signatures

The COSMIC database has been growing over the years with the addition of novel samples and different variant classes

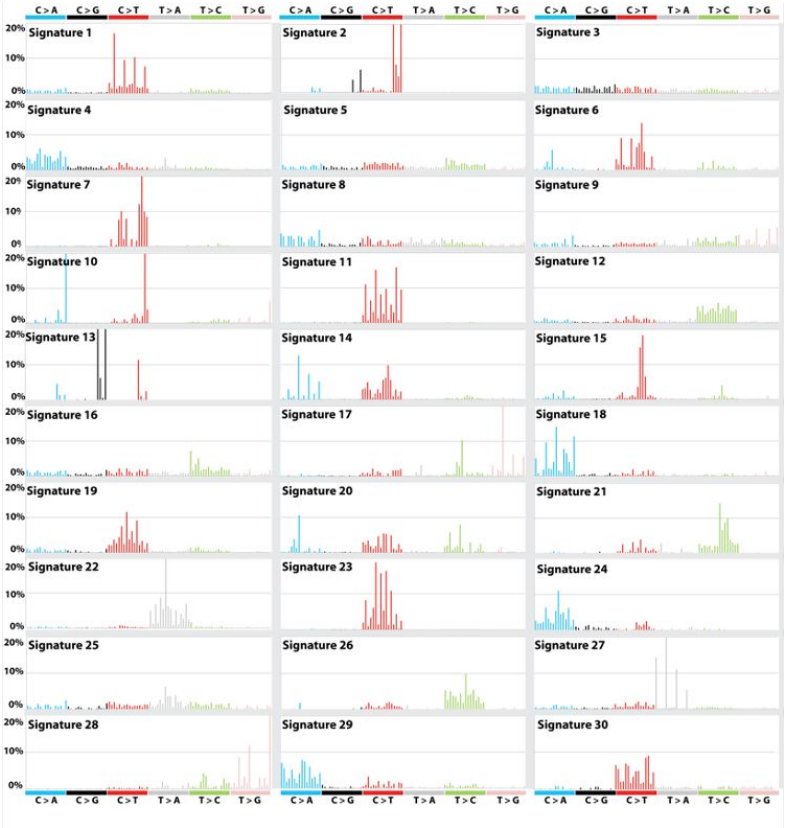
v1 (August 2013)

- 22 SBS signatures

v2 (March 2015)

- 30 SBS signatures

Reference mutational signatures



The COSMIC database has been growing over the years with the addition of novel samples and different variant classes

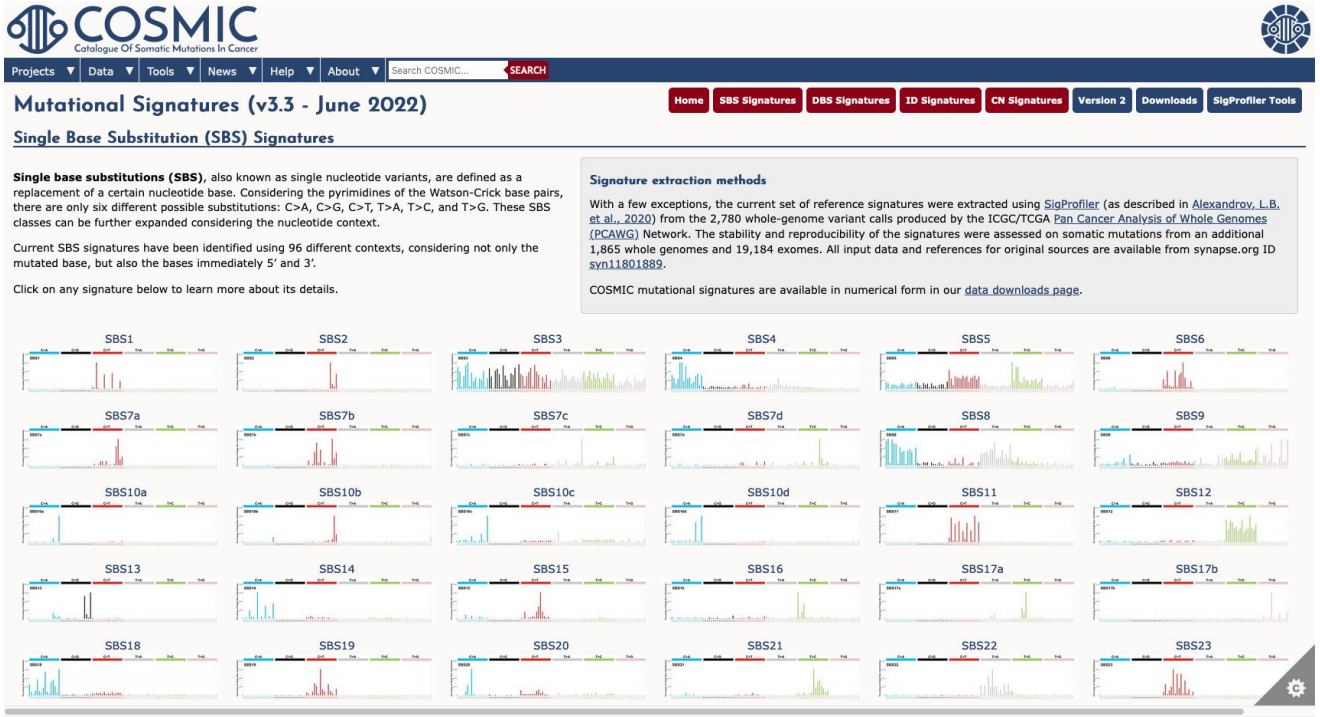
v3 (May 2019)

- 67 SBS signatures
- 11 DBS signatures
- 17 ID signatures

Reference mutational signatures

The current set of COSMIC reference signatures (v3.3 - June 2022) is available at https://cancer.sanger.ac.uk/signatures/, and encompasses:

- 79 SBS signatures
- 11 DBS signatures
- 18 ID signatures
- 24 CN signatures

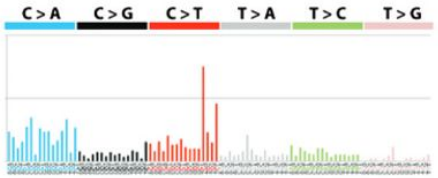


Signature decomposition based on known reference signatures

Refitting mutational signatures

- For mutational signature refitting, the set of mutational signatures is given (matrix S) apart from the input mutational matrix (matrix M), and the goal is to infer the activities or exposures of each signature in each sample (matrix A)
- Most methods are based on the non-negative least squares algorithm
- The signature matrix can consist of either the full set of COSMIC signatures, a subset thereof, or signatures extracted from a specific cancer cohort using a *de novo* method
- The refitting methods are especially useful when the analyzed set of mutations is too small for *de novo* signature extraction, for example, in the case of small sample size, targeted sequencing panels, or samples with few mutations such as in healthy tissues or in slowly growing tumors
- Also, refitting allows extending the applicability of validated mutational signatures in small targeted studies and even in clinical settings for individual patients

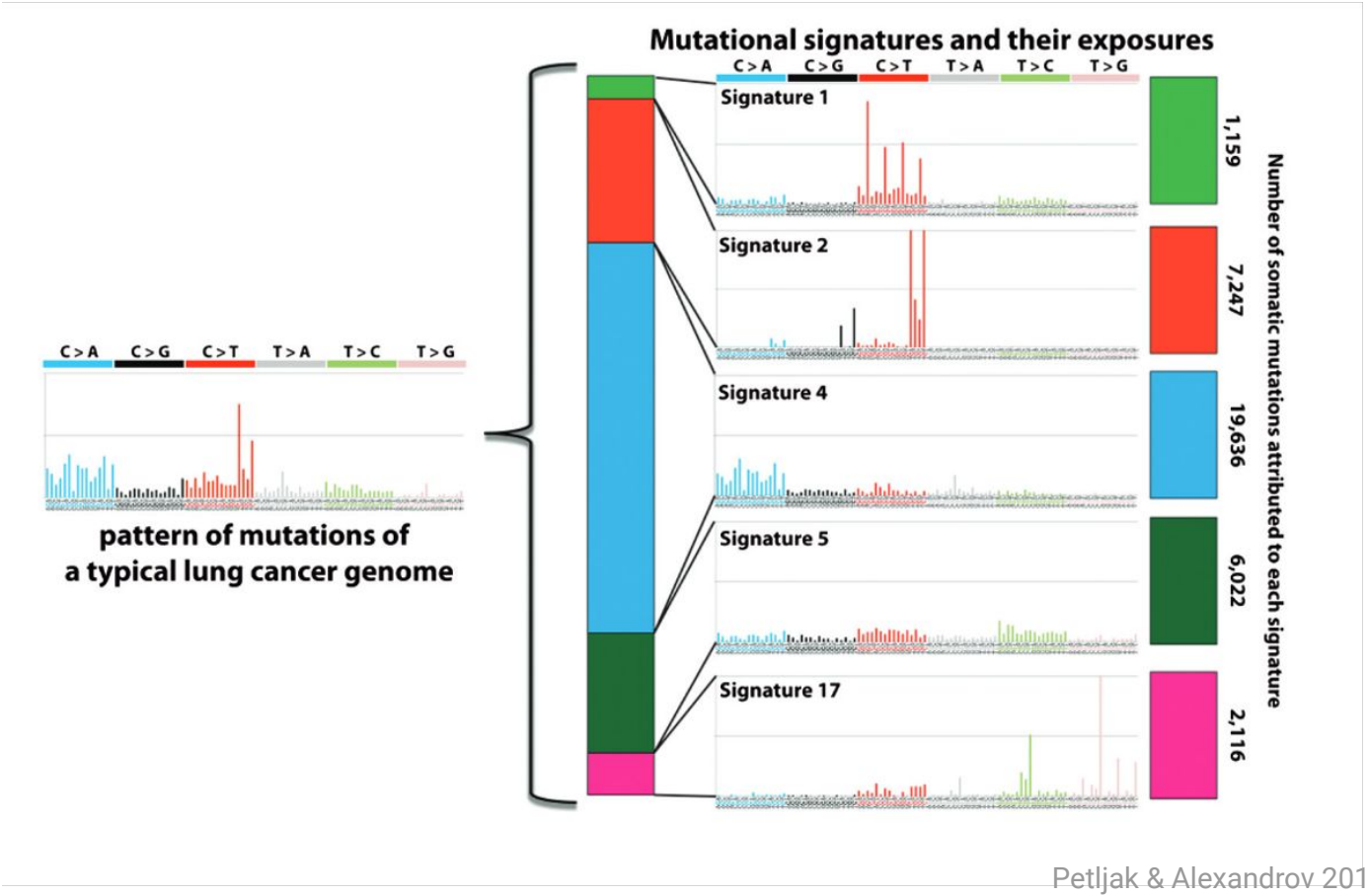
Refitting mutational signatures



pattern of mutations of
a typical lung cancer genome

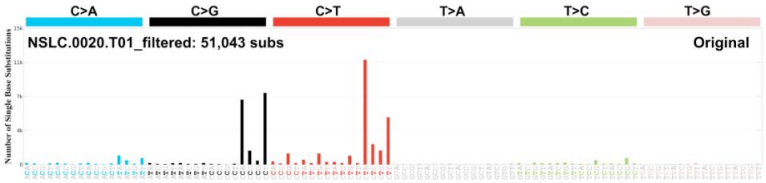
Petljak & Alexandrov 2016 Carcinogenesis

Refitting mutational signatures

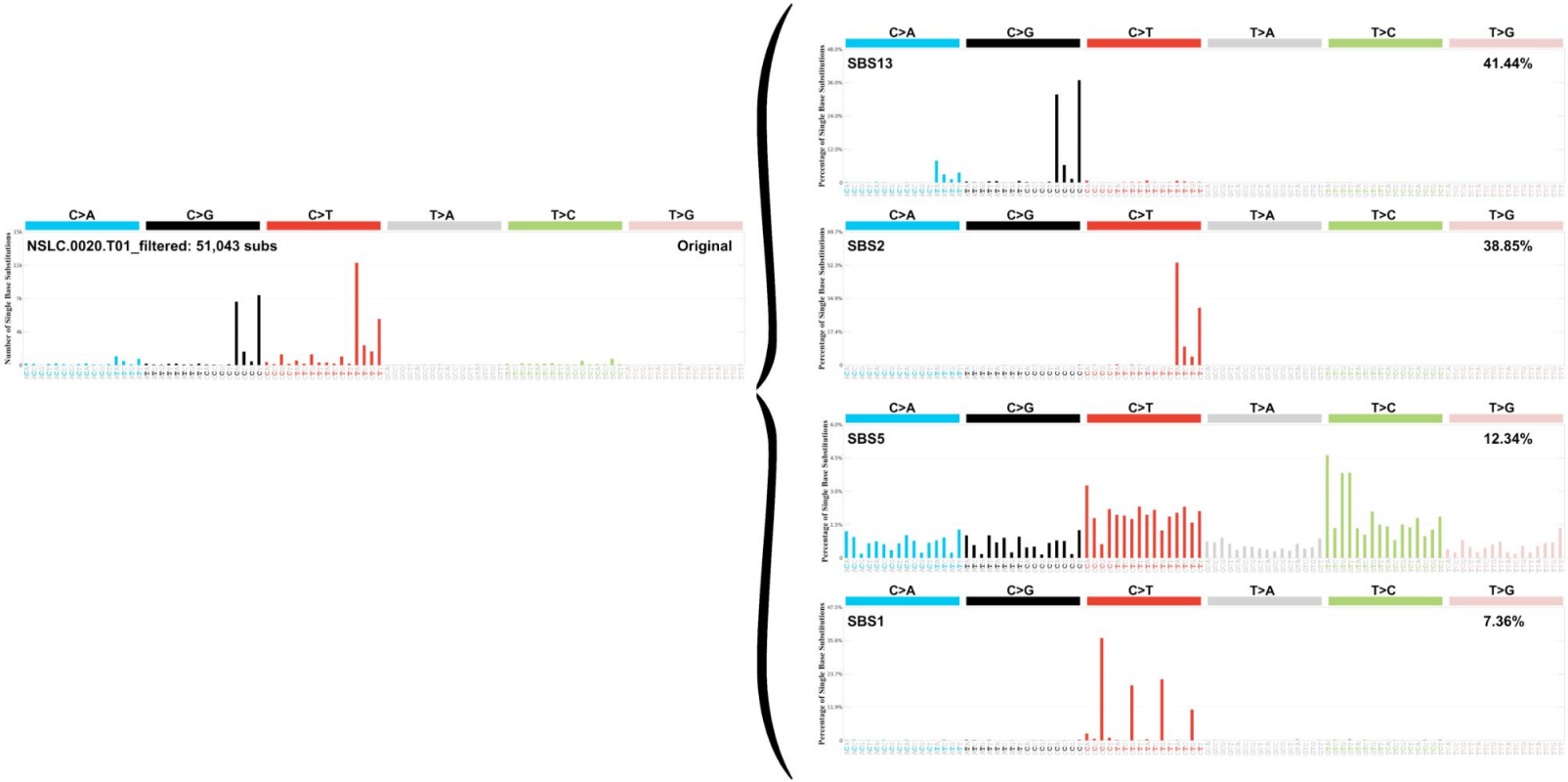


Petljak & Alexandrov 2016 Carcinogenesis

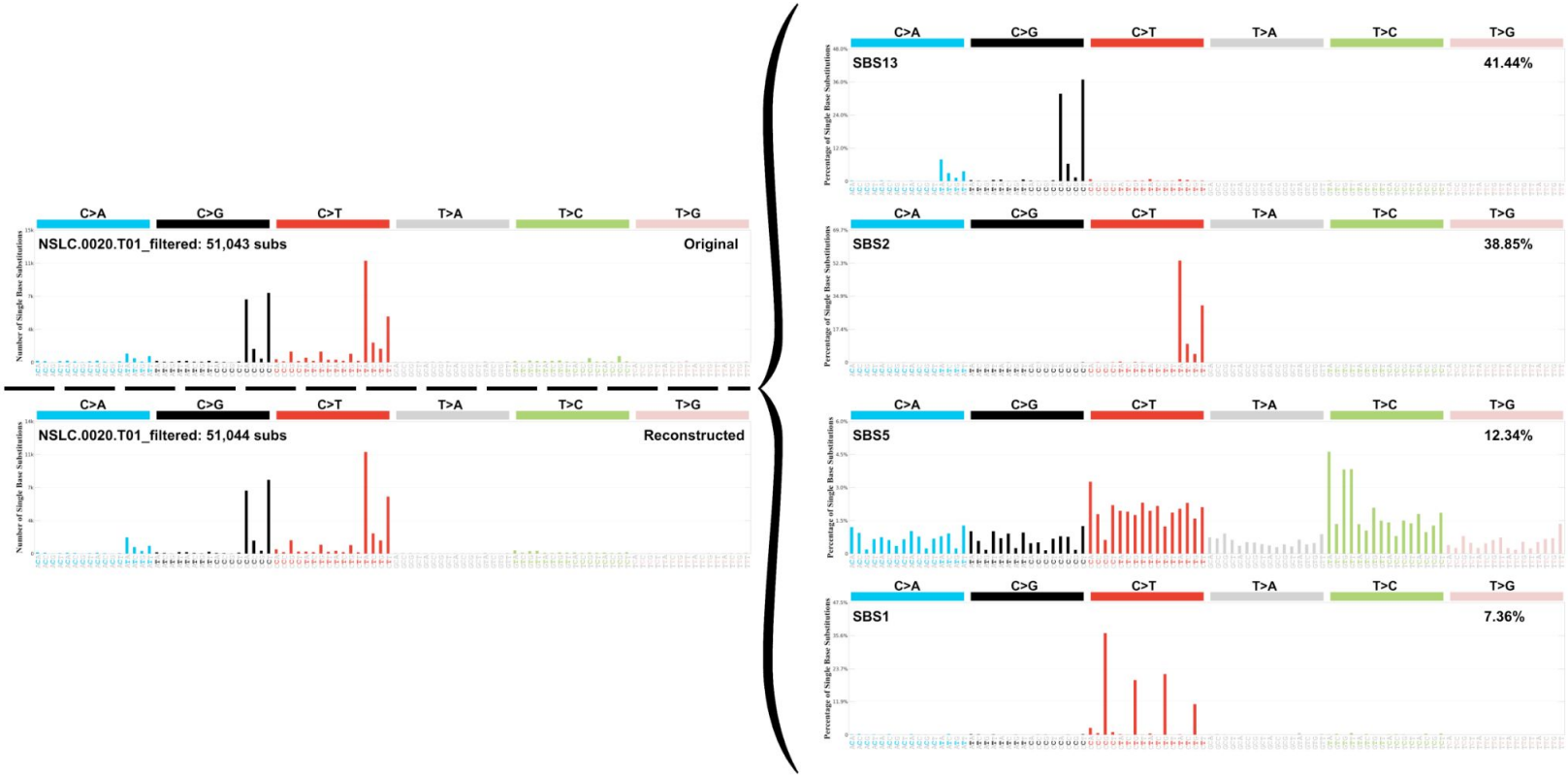
Refitting mutational signatures



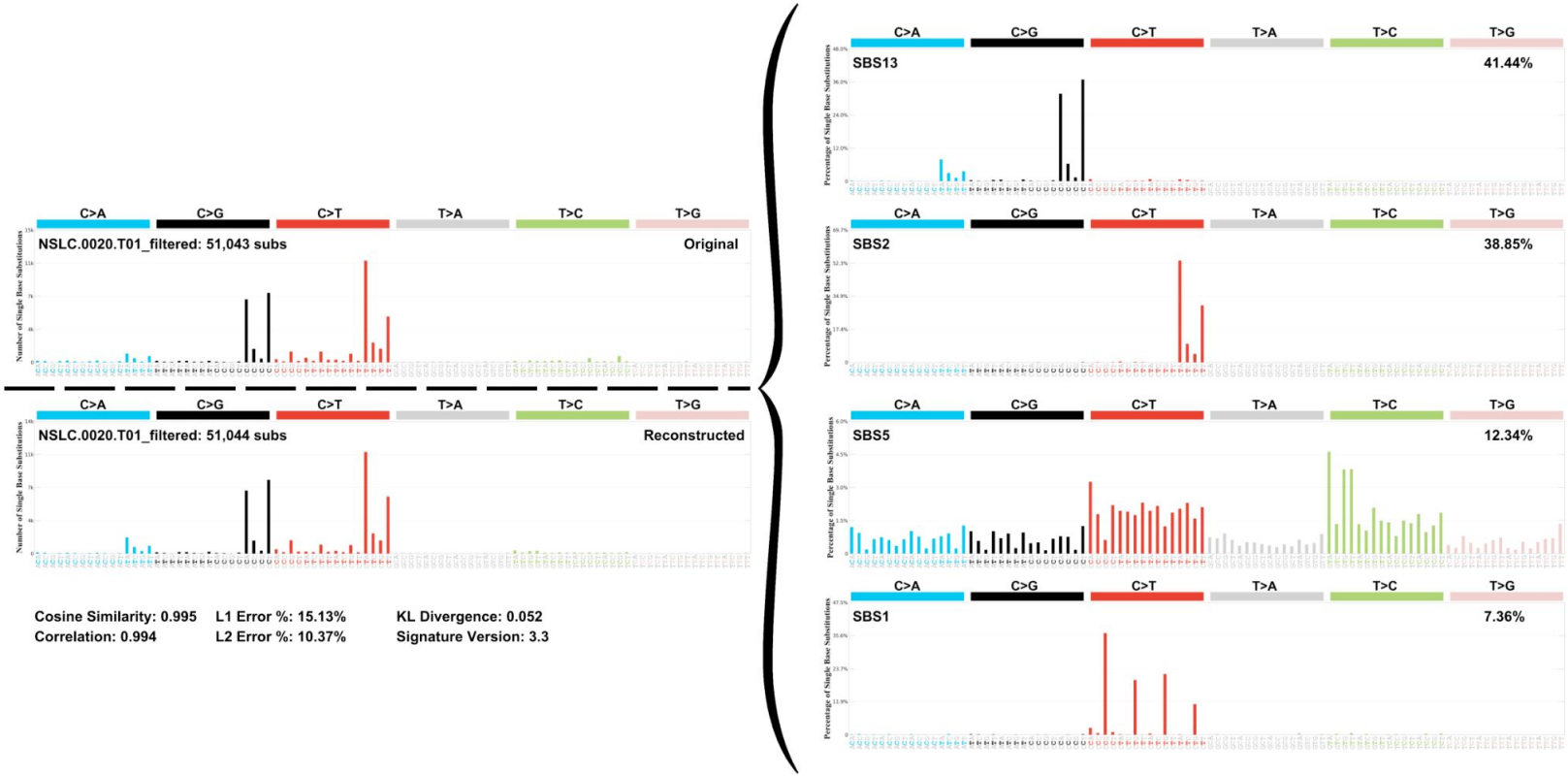
Refitting mutational signatures



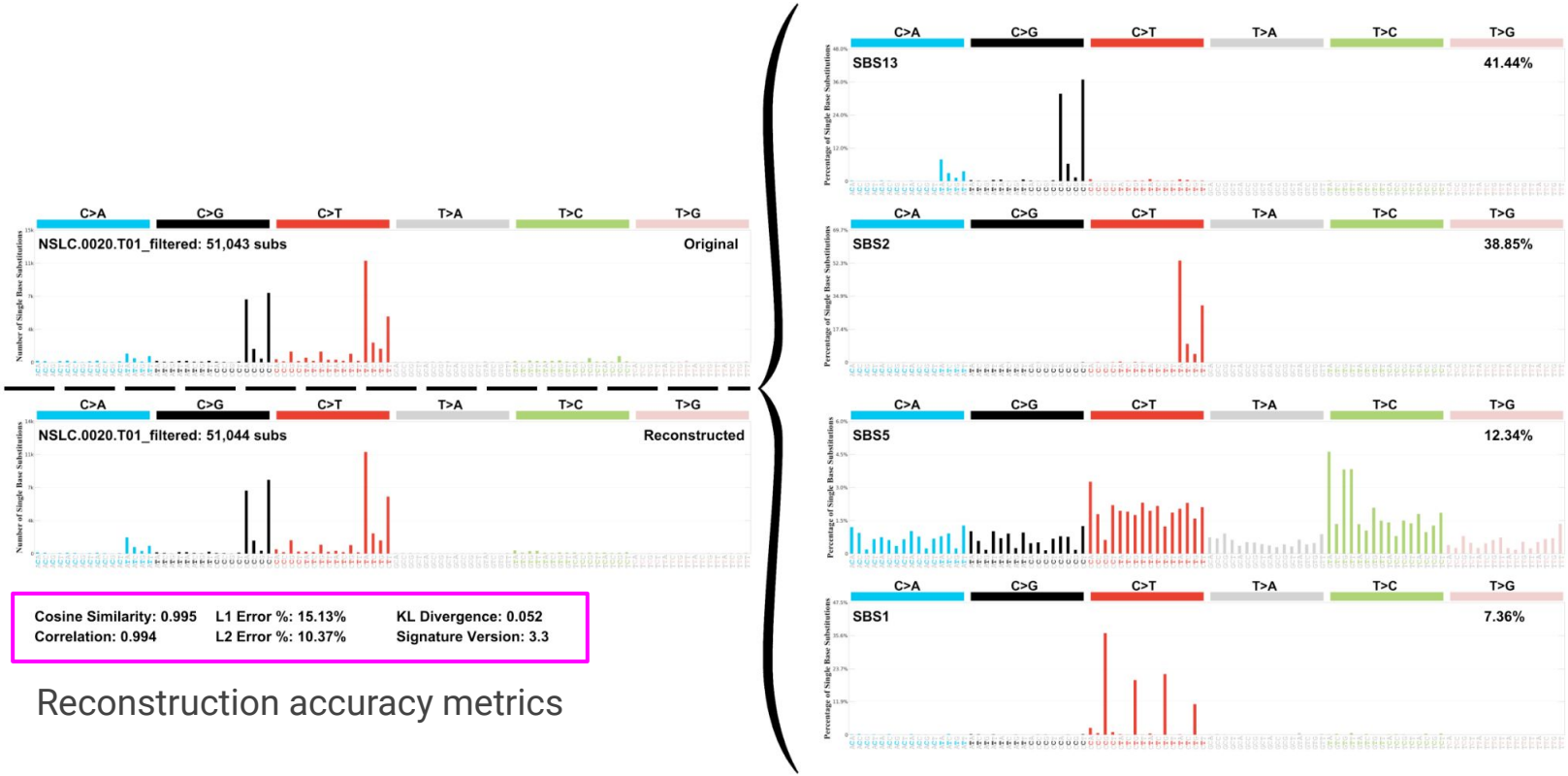
Refitting mutational signatures



Refitting mutational signatures



Refitting mutational signatures



Reconstruction accuracy metrics

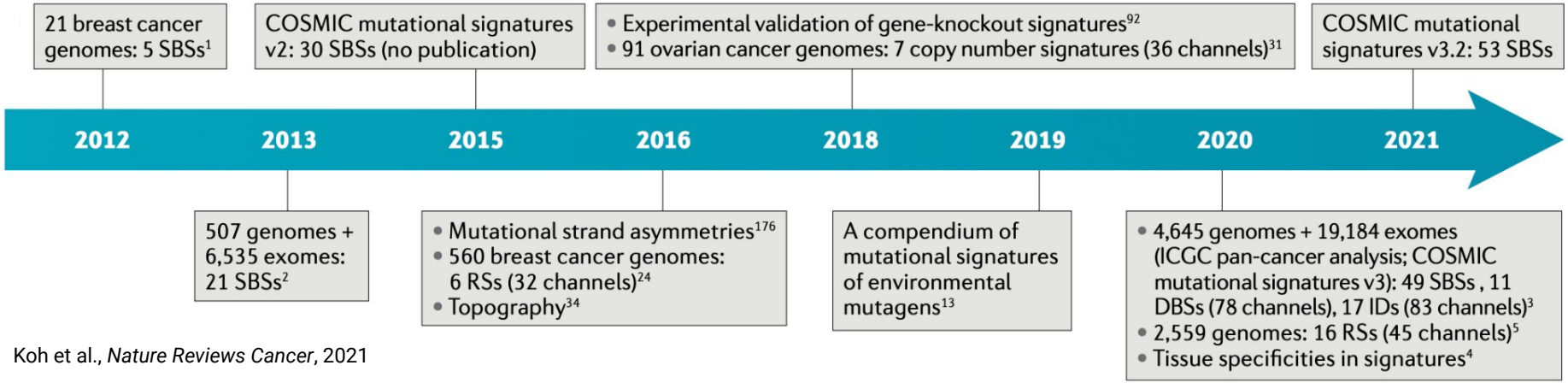
Tools for signature refitting analysis

Tool	Platform	Refitting Approach		Reference
		Method	Computational Engine	
deconstructSigs	R	Non-negative linear regression	Original implementation	Rosenthal *et al.* 2016 Genome Biology
MSA	Python / Nextflow	NNLS	Original implementation / Scipy python package	Senkin 2021 BMC Bioinformatics
MutationalPatterns (standard)	R	NNLS	Pracma R package	Blokzijl *et al.* 2018 Genome Medicine
MutationalPatterns (strict)	R	NNLS	Original implementation / Pracma R package	Manders *et al.* 2022 BMC Genomics
sigLASSO	R	Lasso regression	Original implementation / glmnet R package	Li *et al.* 2020 Nature Communications
SignatureToolsLib	R / Web app	Non-negative linear regression	NNLM R package	Degasperi *et al.* 2022 Science
SigProfilerAssignment	Python / R / Web app	NNLS	Original implementation / Scipy python package	https://github.com/AlexandrovLab/SigProfilerAssignment/

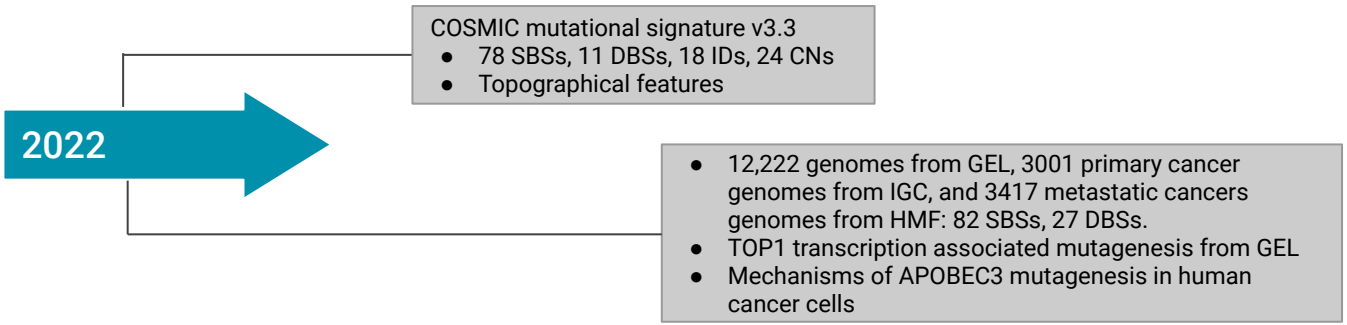
NNLS: non-negative least squares

Emerging mutational signatures in cancer genomics studies

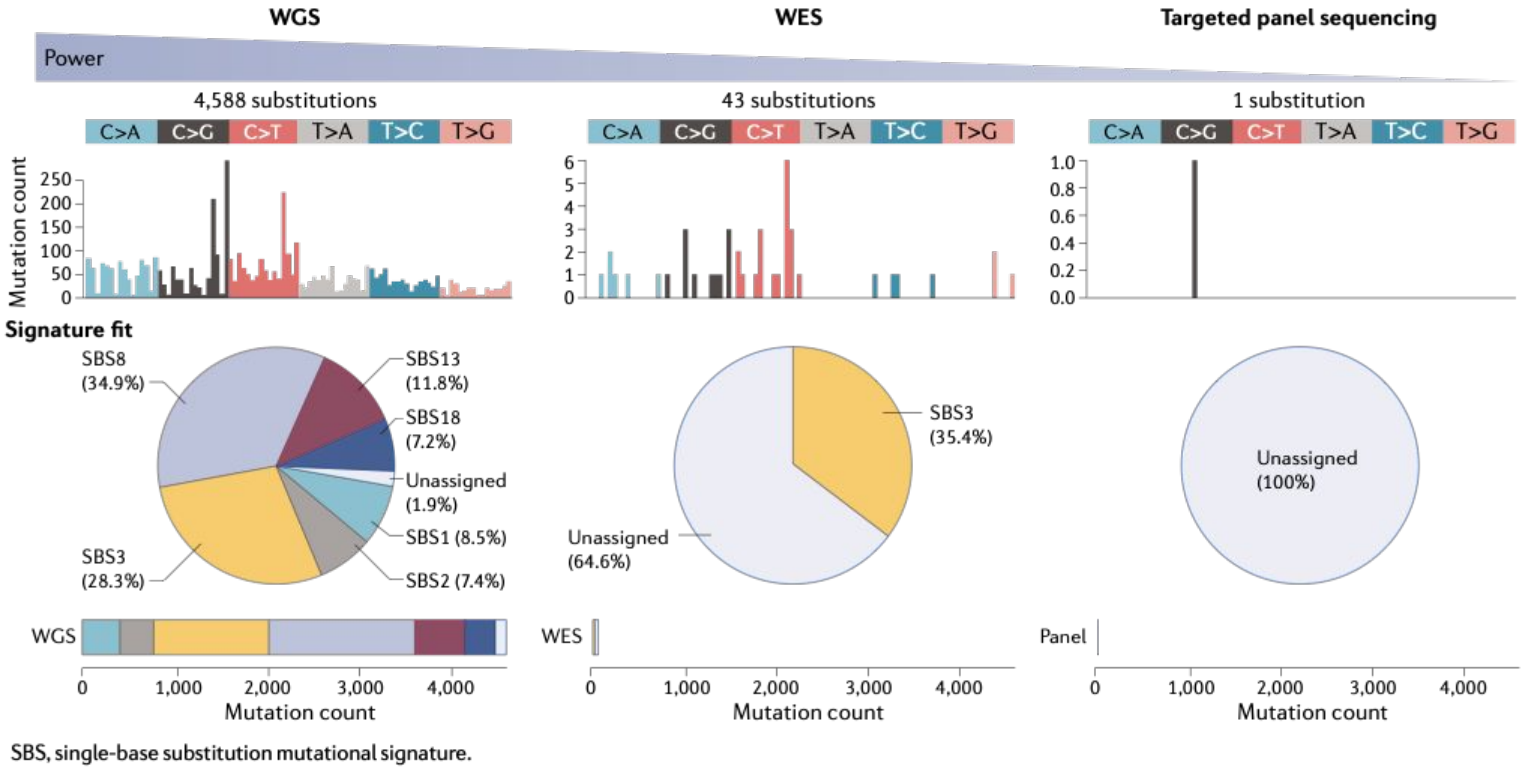
History of mutational signature analysis



Koh et al., *Nature Reviews Cancer*, 2021



Power for signature detection with different sequencing approaches

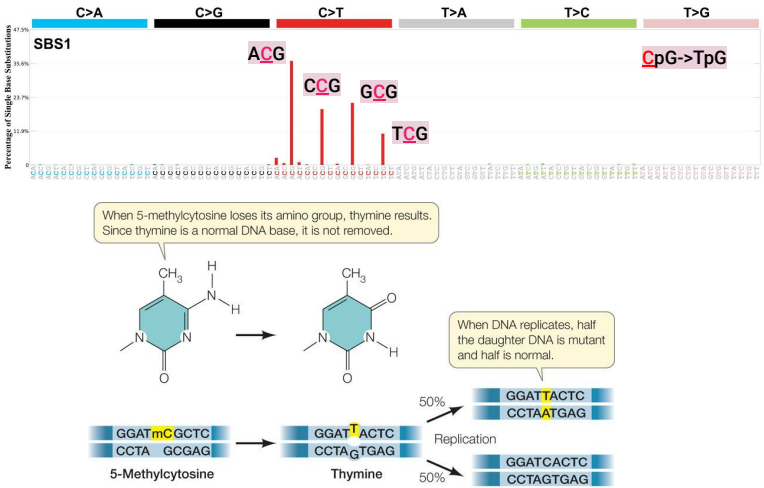


SBS, single-base substitution mutational signature.

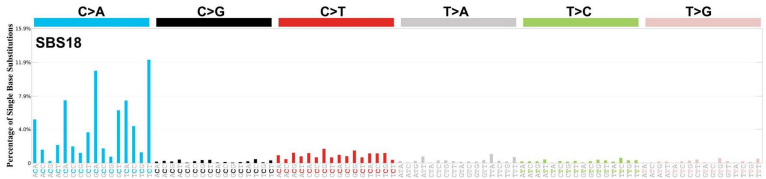
Koh et al., *Nature Reviews Cancer*, 2021

Etiologies of SBS mutational signatures - *Endogenous*

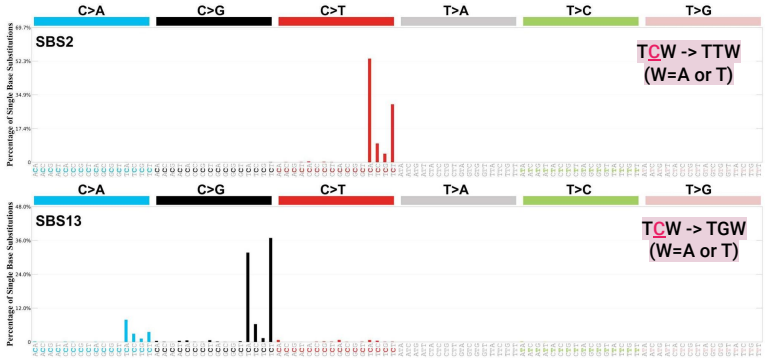
Spontaneous deamination of 5-methylcytosine (clock-like signature)



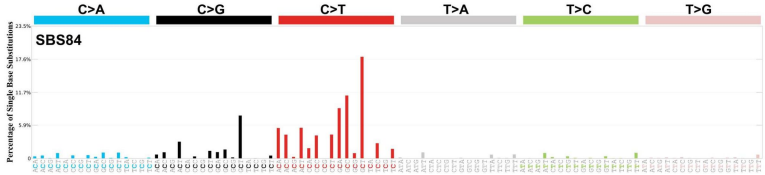
Damage by reactive oxygen species



Activity of APOBEC family of cytidine deaminases (*e.g.*, APOBEC3A, APOBEC3B)



Activity of activation-induced cytidine deaminase (AID)



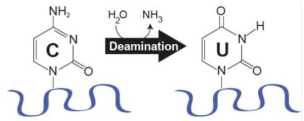
Gene family: (12p13.1)

adaptive immunity

RNA editing

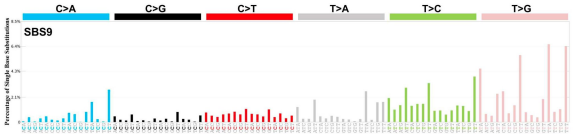
★innate immunity: restriction of retroviruses/retrotransposons

★hypermutation by accidental access to chromosomal DNA

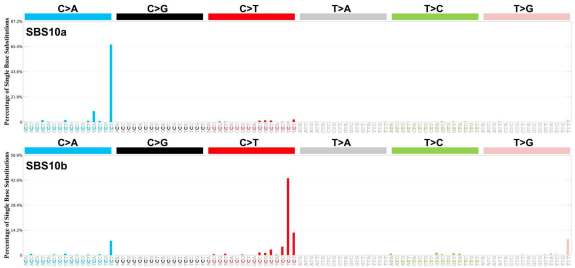


Etiologies of SBS mutational signatures - *Endogenous*
(DNA replication or repair deficiency)

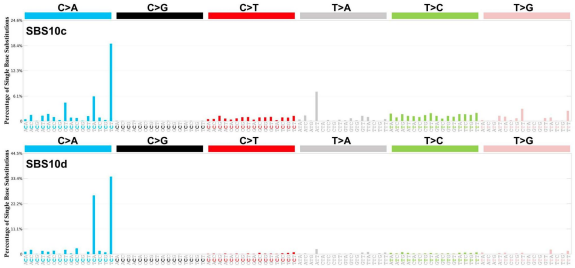
Polymerase eta somatic hypermutation activity



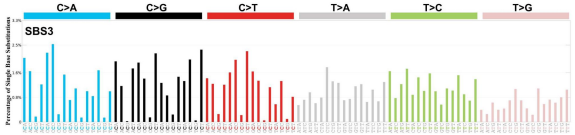
Polymerase epsilon (POLE) exonuclease domain mutations



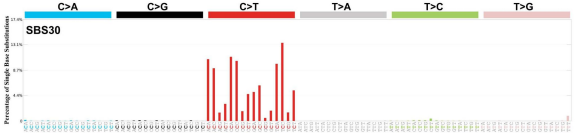
Defective *POLD1* proofreading



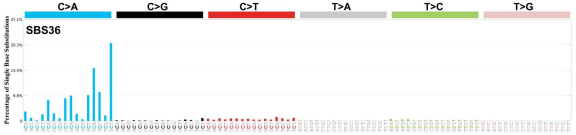
Defective homologous recombination DNA damage repair



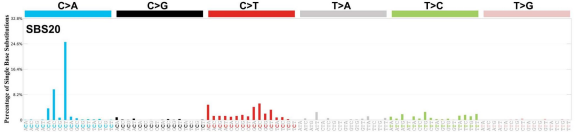
Defective DNA base excision repair due to *NTHL1* mutations



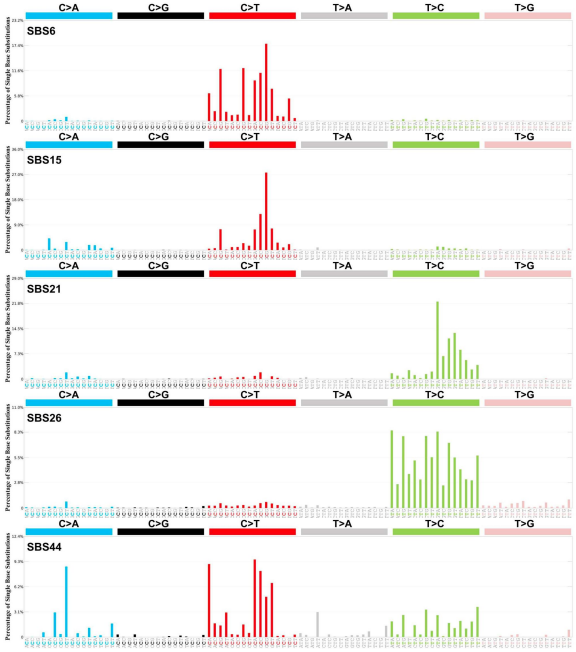
Defective DNA base excision repair due to *MUTYH* mutations (or reactive oxygen species)



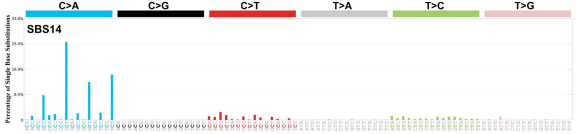
Concurrent *POLD1* mutation and defective DNA mismatch repair



Defective DNA mismatch repair

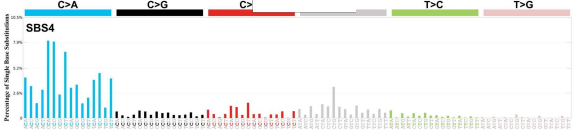


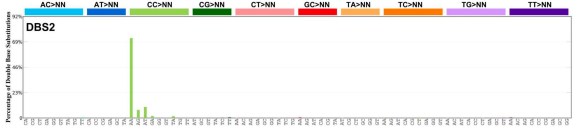
Concurrent polymerase epsilon mutation and defective DNA mismatch repair

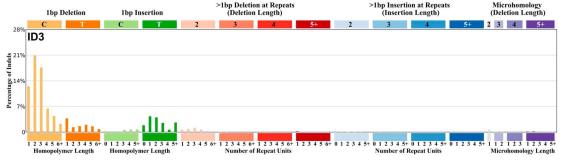


Etiologies of SBS mutational signatures - *Exogenous*

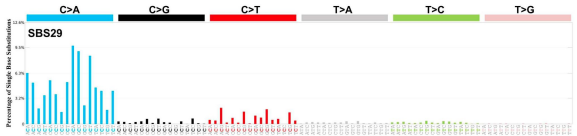
Tobacco smoking





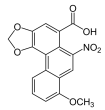


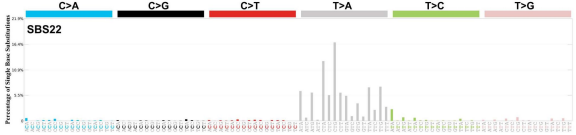
Tobacco chewing



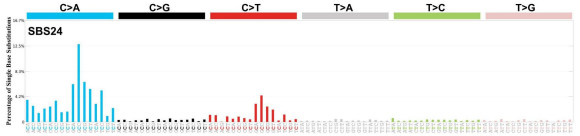
Ultraviolet light exposure

Aristolochic acid exposure

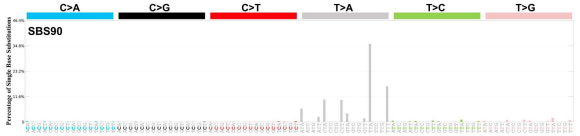




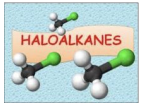
Aflatoxin exposure

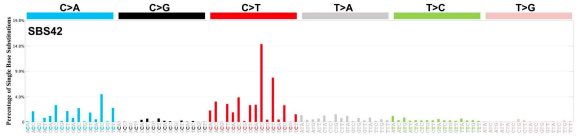


Duocarmycin exposure (DNA-alkylating agents)

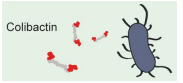


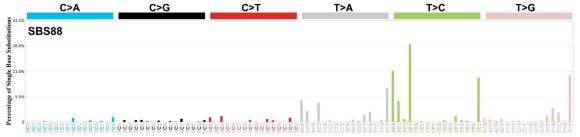
Occupational exposure to haloalkanes





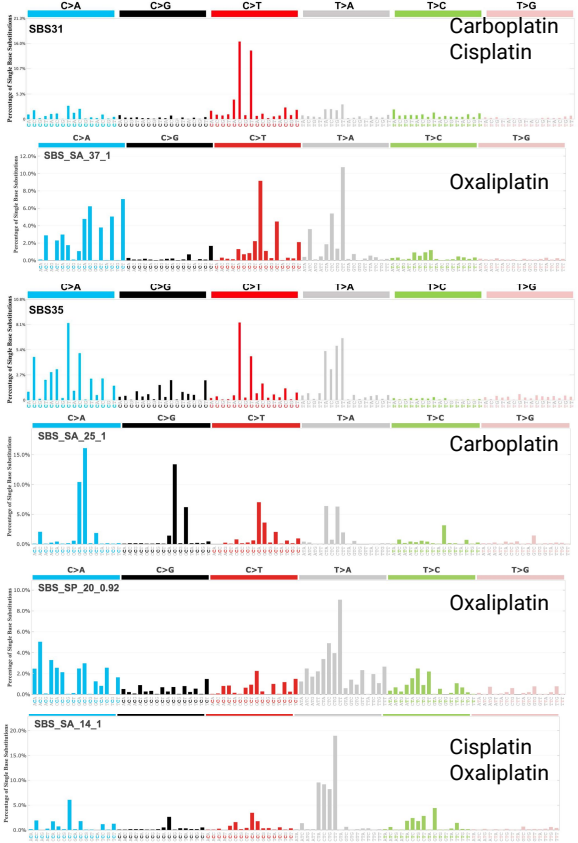
Colibactin exposure (E.coli bacteria carrying pks pathogenicity island)



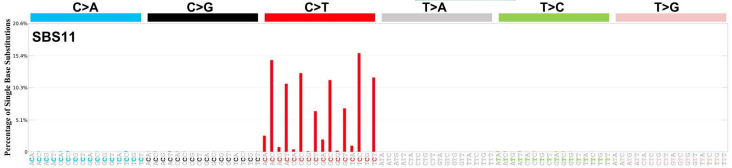


Etiologies of SBS mutational signatures - *Exogenous* (Cancer therapies)

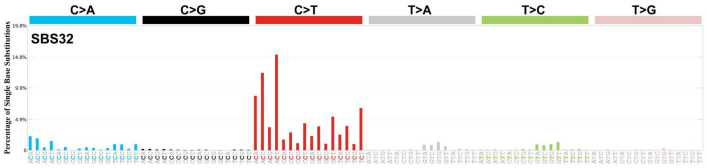
Platinum chemotherapy treatment

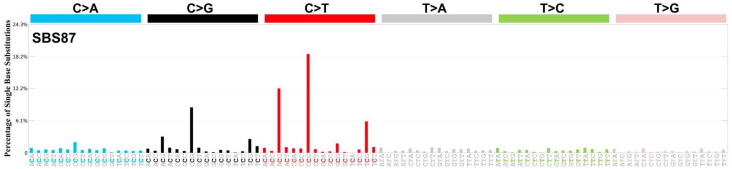
Temozolomide treatment



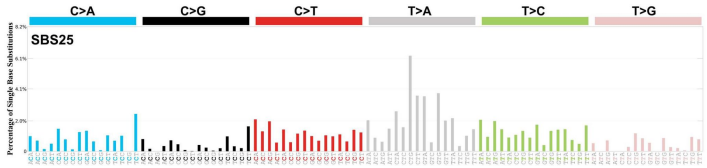
Azathioprine treatment



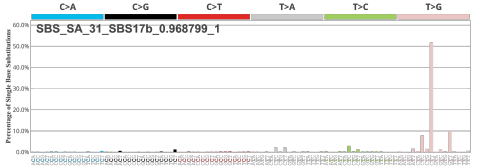
Thiopurine chemotherapy treatment



Chemotherapy treatment



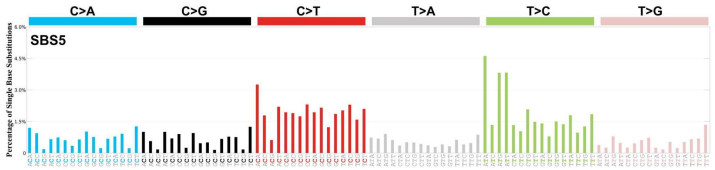
Nucleoside Metabolic inhibitor (Capecitabine)



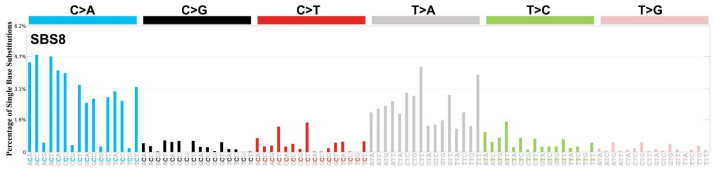
- More than 3,500 metastatic tumors originating from different organs (WGS) (Pich et al., *Nature Genetics*, 2019)
- Signatures extracted using SignatureAnalyzer, SigProfiler, and a third non-NMF method.
- Identified SBS and/or DBS signatures in several anticancer therapies
- Platinum-based: Carboplatin, Cisplatin, Oxaliplatin.
- Nucleoside Metabolic Inhibitor (Capecitabine)

Mutational signatures with *Unknown* etiology

Clock-like mutations associated with age



HR/NER deficiency?

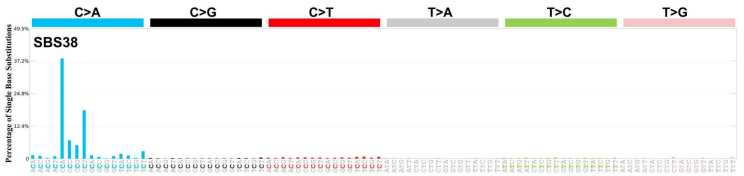


Possible sequencing artefacts:

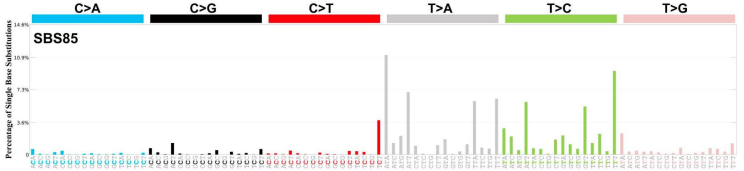
SBS27, SBS43, SBS45, SBS46, SBS47, SBS48, SBS49, SBS50, SBS51, SBS52, SBS53, SBS54, SBS55, SBS56, SBS57, SBS58, SBS59, SBS60, SBS95....

[Mutational Signatures (v3.3 - June 2022)]

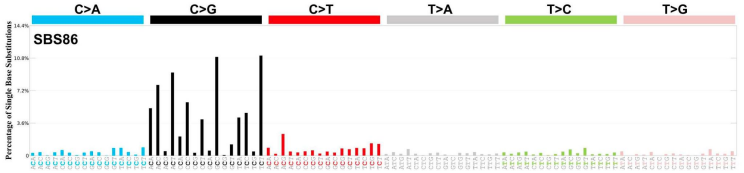
Indirect effect from UV-light exposure



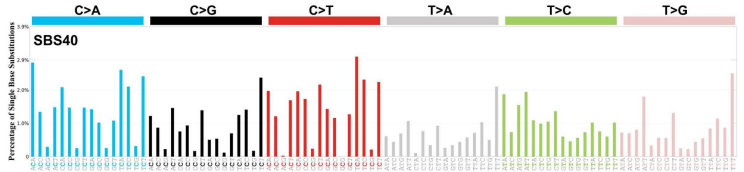
Indirect effect from of activation-induced cytidine deaminase (AID)



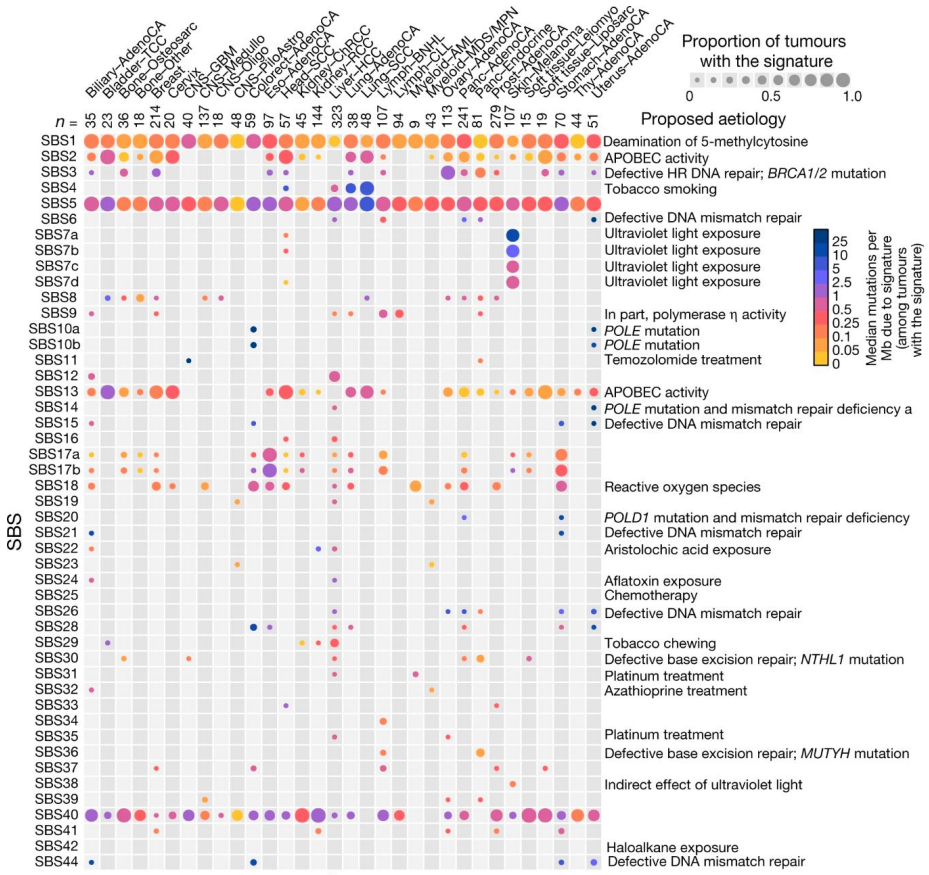
Unknown chemotherapy treatment

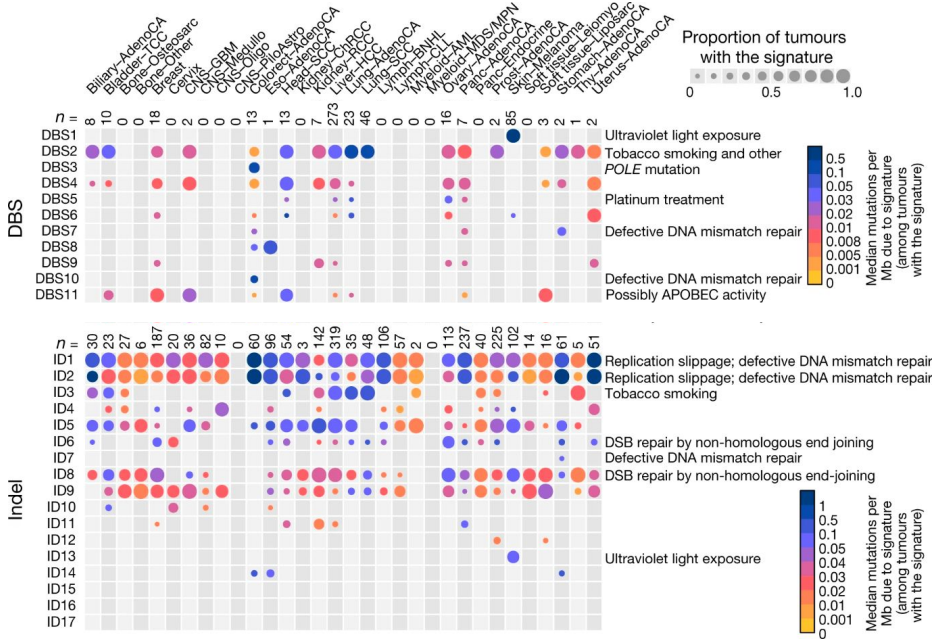


Unknown but very common



The repertoire of mutational signatures in human cancer





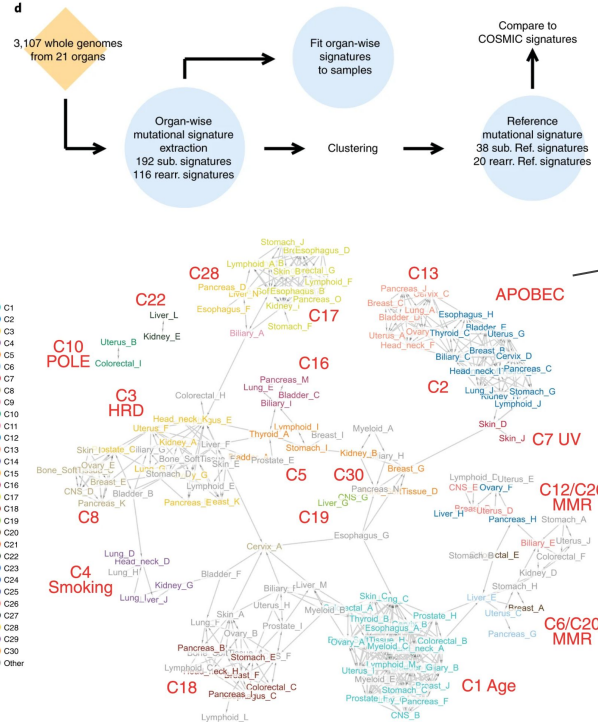
PCAWG study [Alexandrov et al., *Nature*, 2020]

Repertoire form Signal website:
https://signal.mutationalsignatures.com/explore/cancer
(PCAWG, GEL, Hartwig)

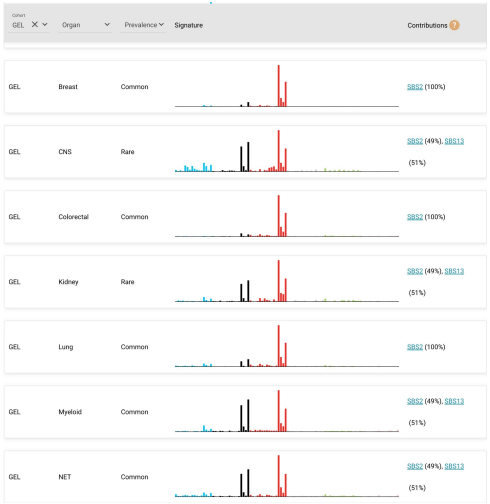
[Degasperi et al., Science, 2022]

Organ-specific signatures (Cancer Specific Signatures)

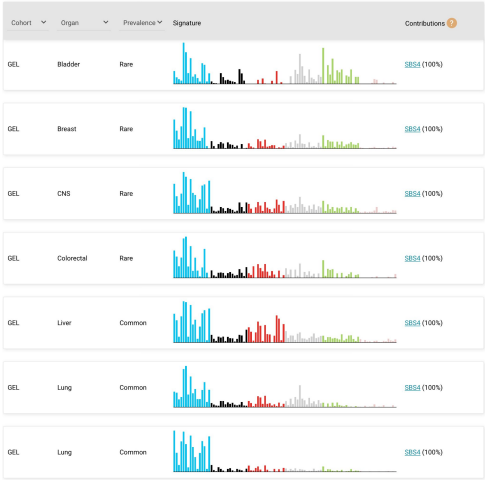
- 3,107 WGS primary cancers across 21 organs (Degasperi *et al.*, *Nature Cancer*, 2020)
- GEL (12,222), ICGC (3,001), Hartwig (3,417) WGS studies across 21 organs (Degasperi et al., *Science*, 2022).
- Signatures are initially extracted from subsets of samples from each cohort and organ. The organ-specific signatures were then clustered and the averages of these clusters are the reference signatures (SBS, DBS). While we encourage the use of reference signatures primarily, organ-specific signatures can highlight signature variability across organs and cohorts.



APOBEC



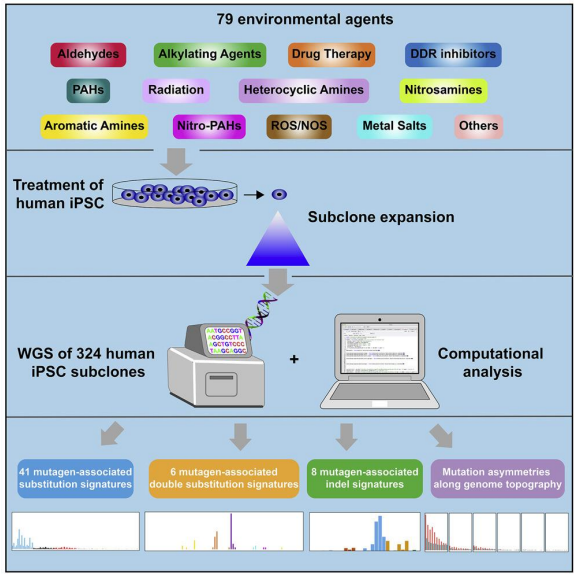
Tobacco Smoking



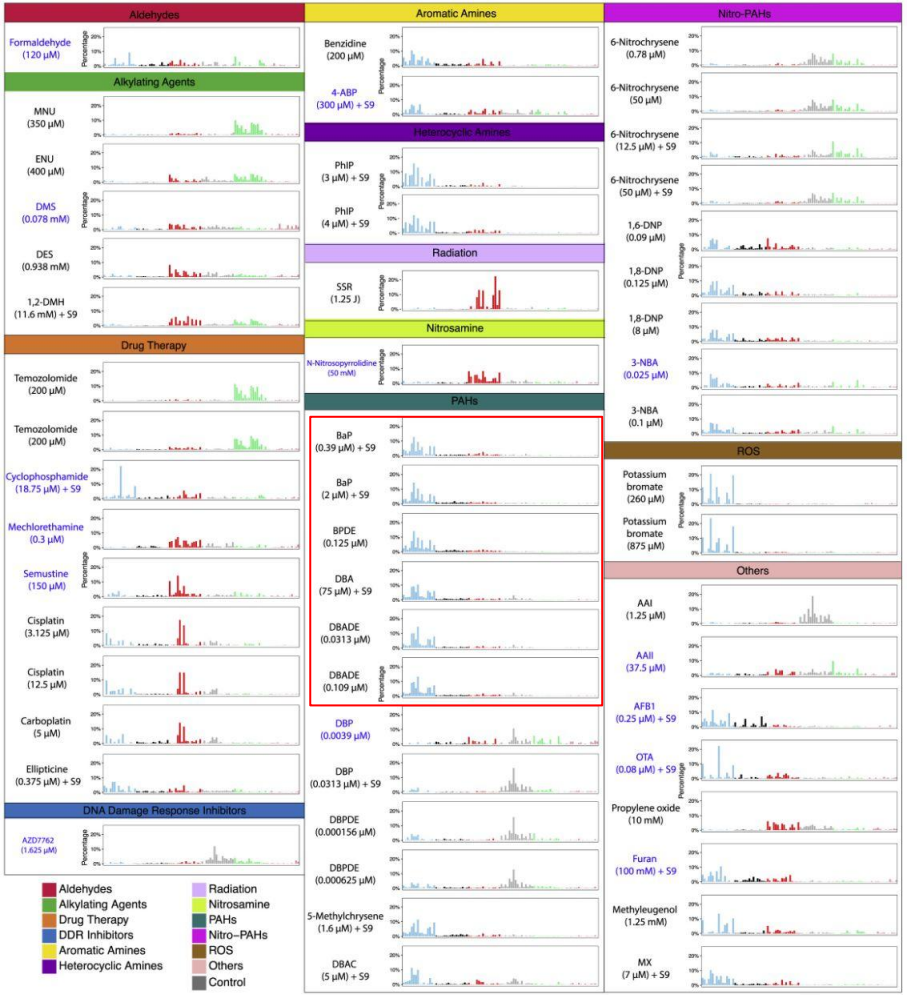
Environmental Mutagenesis

- 324 WGS human-induced pluripotent stem cells
- 79 known or suspected environmental carcinogens
  - 41 yielded SBS signatures
  - 6 yielded DBS signatures
  - 8 yielded ID signatures

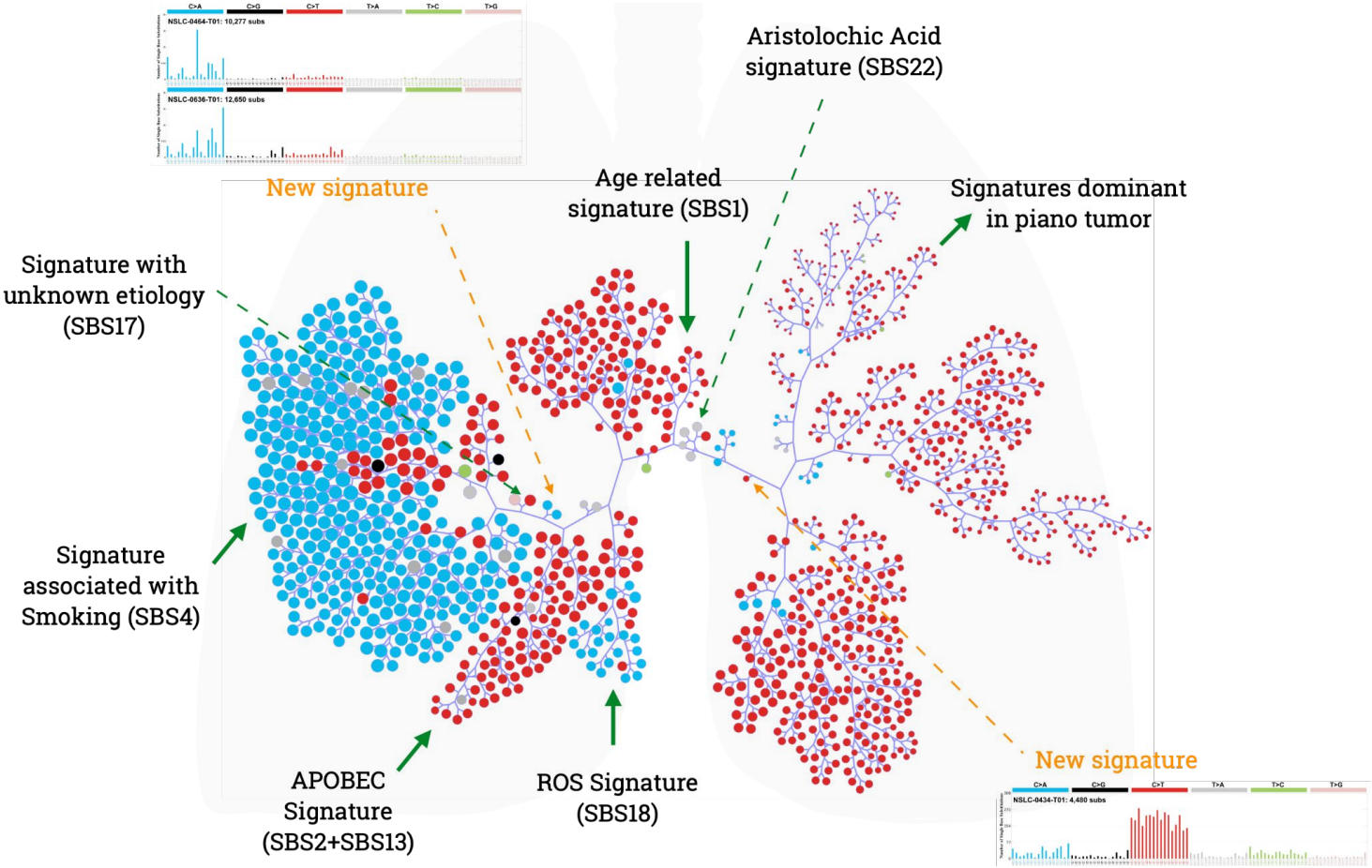
This compendium of experimentally induced mutational signatures permits further exploration of roles of environmental agents in cancer etiology and underscores how human stem cell DNA is directly vulnerable to environmental agents.



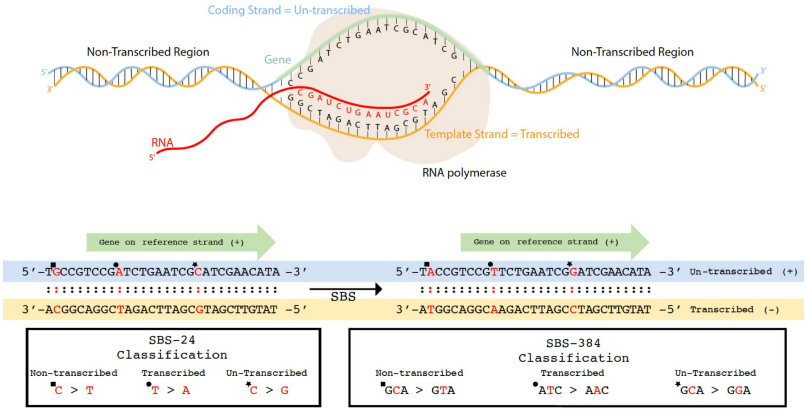
Kucab *et al.*, *Cell*, 2019



Discovery of new signatures in Sherlock-Lung by mSigPortal

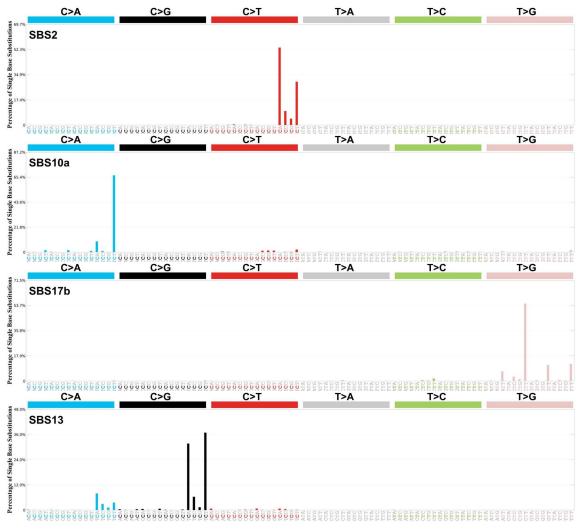


Strand bias of mutational signatures



Mutational signatures exhibit asymmetric number of mutations due to either one of the DNA strands being preferentially repaired or one of the DNA strands having a higher propensity for being damaged. One common example of strand asymmetry is transcription-strand asymmetry which can be due to the activity of transcription-coupled nucleotide excision repair (TC-NER) or transcription-coupled damage amongst others.

Shape of signatures - Shannon equitability index

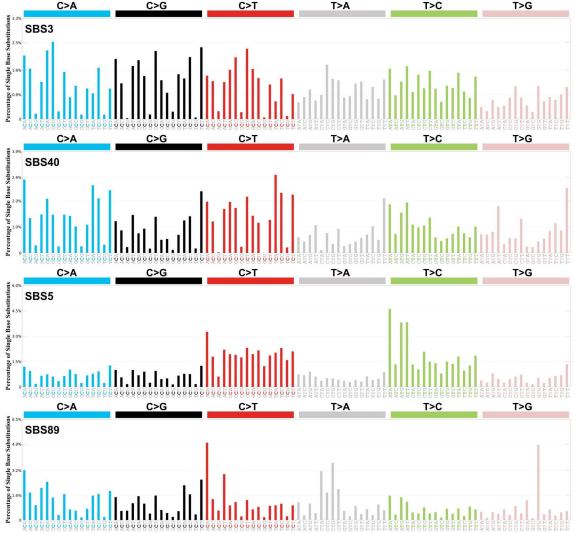


Equitability=0.27

Equitability=0.36

Equitability=0.38

Equitability=0.42

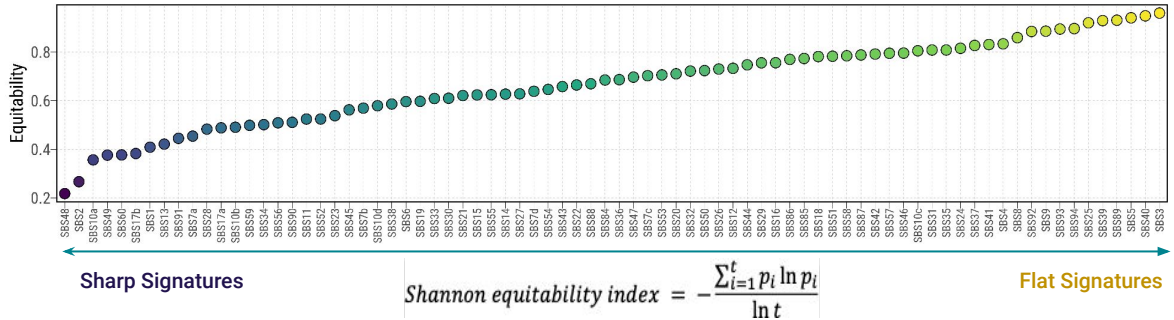


Equitability=0.96

Equitability=0.95

Equitability=0.94

Equitability=0.93



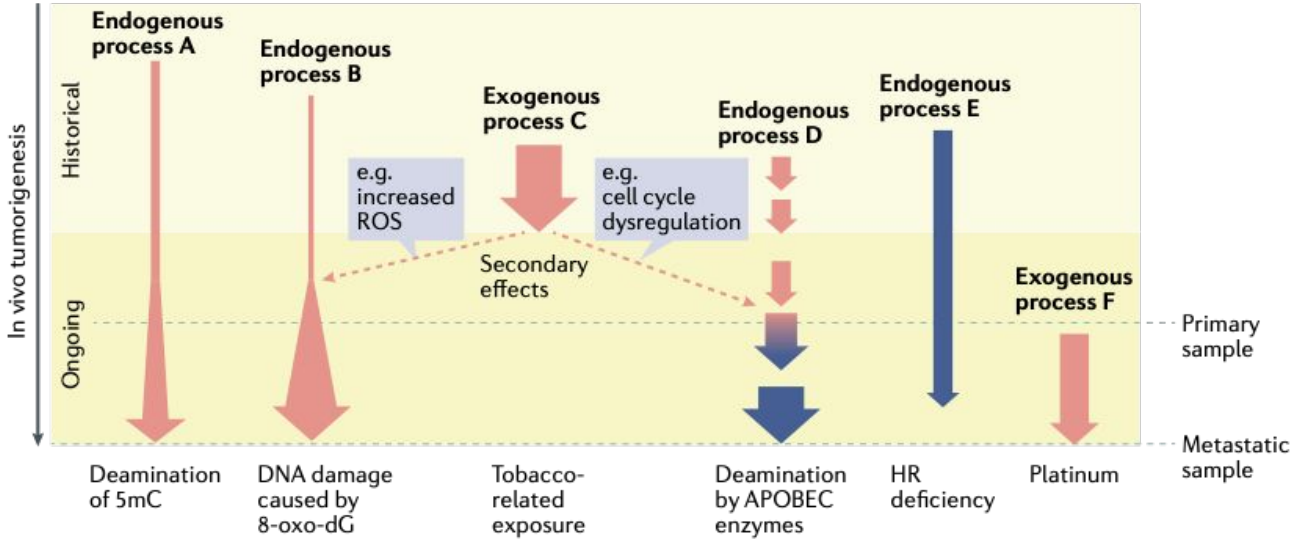
Sharp Signatures

$$Shannon\ equitability\ index = -\frac{\sum_{i=1}^{t} p_i \ln p_i}{\ln t}$$

Flat Signatures

Islam *et al. Cell Genomics*, 2022

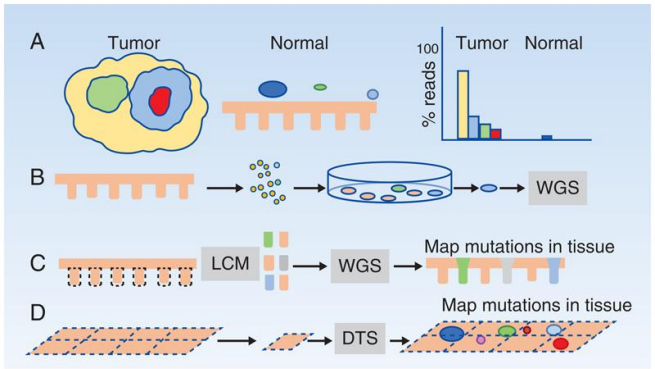
Mutational Signature Interaction



Mutational signatures are the imprints of various endogenous and exogenous mutational processes (labelled 'A' to 'F'). Some processes are historical, while others are ongoing and even intermittent (process D). Mutational processes that cause signatures in a direct manner can be considered primary signatures. There may also be augmentation of certain signatures secondary to cellular abnormalities that arise due to primary exogenous mutagen exposure (red dashed arrows). Some mutational processes may be clinically informative (highlighted in dark blue); for example, process D, which when amplified may signal dysregulation of the cell cycle, or process E, which may indicate a deficiency of a DNA repair pathway that has synthetically lethal interactions with particular therapeutic agents. Process F is an example of a late-onset iatrogenic exposure due to treatment. The horizontal turquoise dashed lines indicate different sampling times. APOBEC, apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like; HR, homologous recombination; 5mC, 5-methylcytosine; 8-oxo-dG, 8-oxo-2'-deoxyguanosine; ROS, reactive oxygen species.

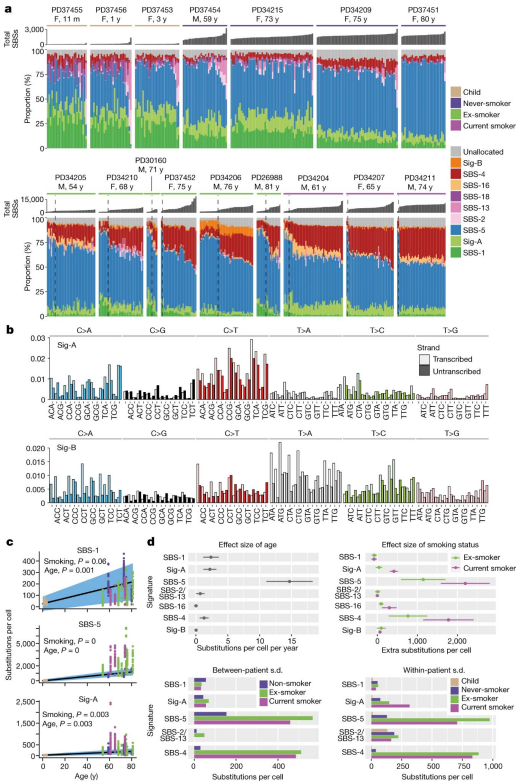
Koh *et al.*, *Nature Reviews Cancer*, 2021

Mutational signatures observed in normal tissue studies



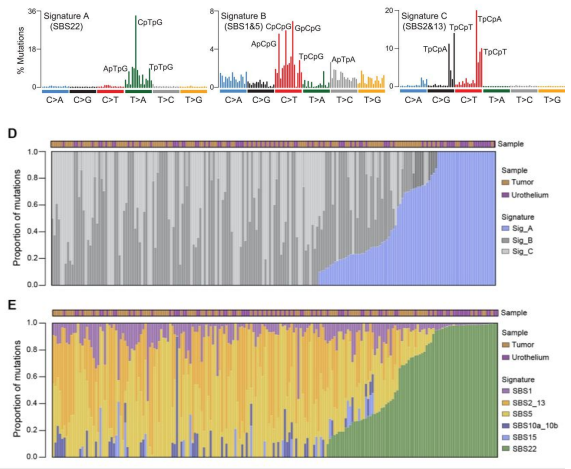
Detecting somatic mutations in normal epithelia
LCM: Laser capture microdissection
DTS: deep targeted sequencing

Fowler *et al.*, *Cancer Discovery*, 2022



Tobacco smoking and somatic mutations in human bronchial epithelium

Yoshida *et al.*, *Nature*, 2020

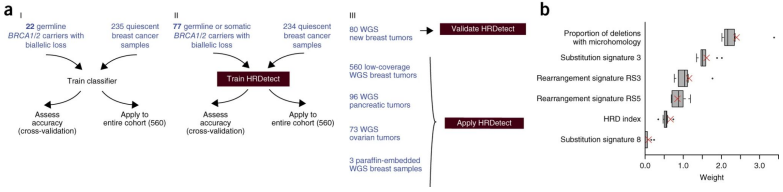


Macroscopic somatic clonal expansion in morphologically normal human urothelium (MNU; epithelium lining the bladder and ureter)

Li *et al.*, *Science*, 2020

Clinical applications for mutational signatures

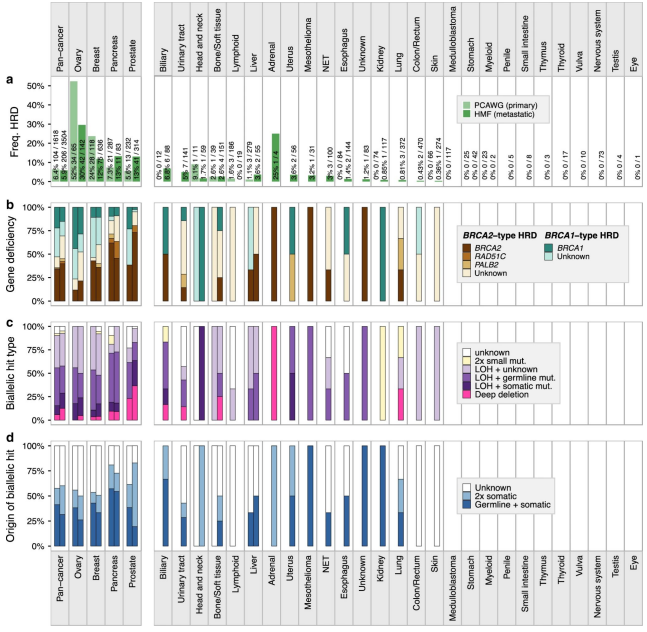
- Provide evidences for unexpected carcinogens. (e.g., Azathioprine caused SBS32 mutations, which was used as one of the most effective and safe immunosuppressive medicines according to WHO).
- Identify the origin of unknown primary cancer or known carcinogen in unexpected cancer types (*e.g.*, UV-light signatures SBS7 observed in non-skin cancer)
- Establish the link between known carcinogens and suspected cancer types (*e.g.*, exposure to aristolochic acids SBS22 and hepatocellular carcinomas)
- HRD prediction based on SBS3 and other signatures provides an opportunity to identify larger populations of cancer patients who may benefit from treatment with PARP inhibitors (PARPi).
- MMR-deficiency prediction using MMRDetect with implications for responsiveness to immunotherapies.



HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures

Davies et al., *Nature Medicine*, 2017

Machine learning-based approach for estimating HRD status from target sequencing



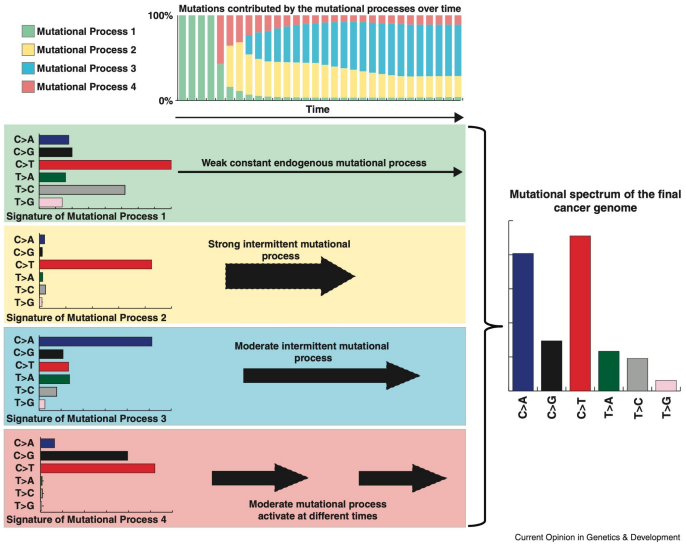
Percentage breakdown of the incidence and genetic causes of HRD in CHORD-HRD patients pan-cancer and by cancer type.

Nguyen et al., *Nature Communications*, 2020

Downstream analysis for mutational signature data

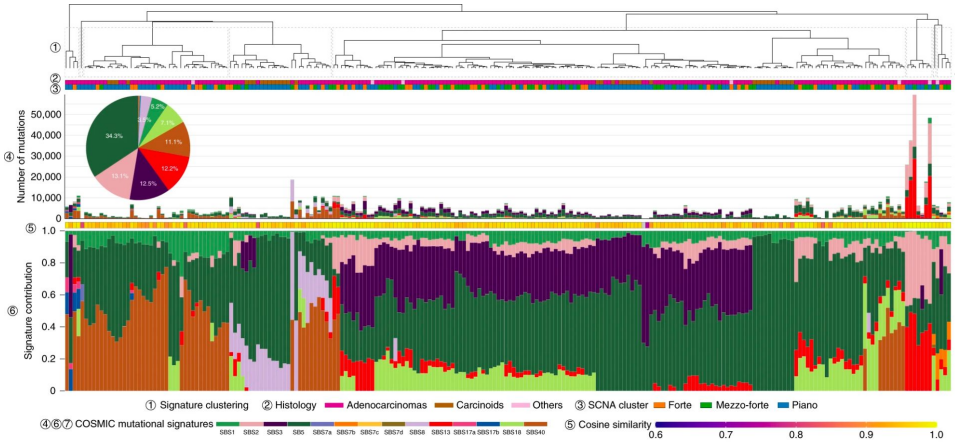
Mutational signature activities in cancer genome

Mutational processes operative in a cancer

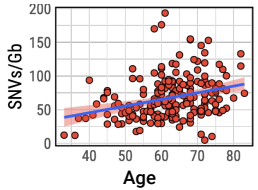


Alexandrov and Stratton, *Current Opinion in Genetics* & Development. 2014

Landscape of mutational processes in Sherlock-*Lung*

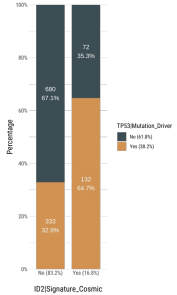


Zhang, *et al.*, Nature Genetics, 20201

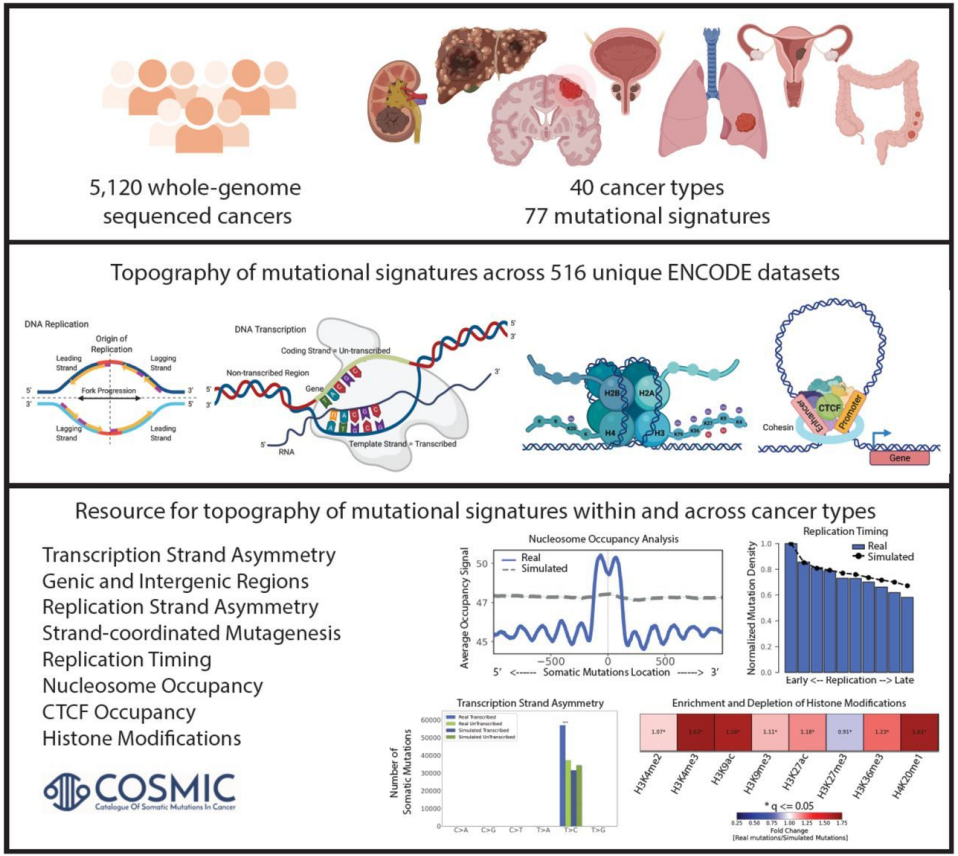


Association analysis

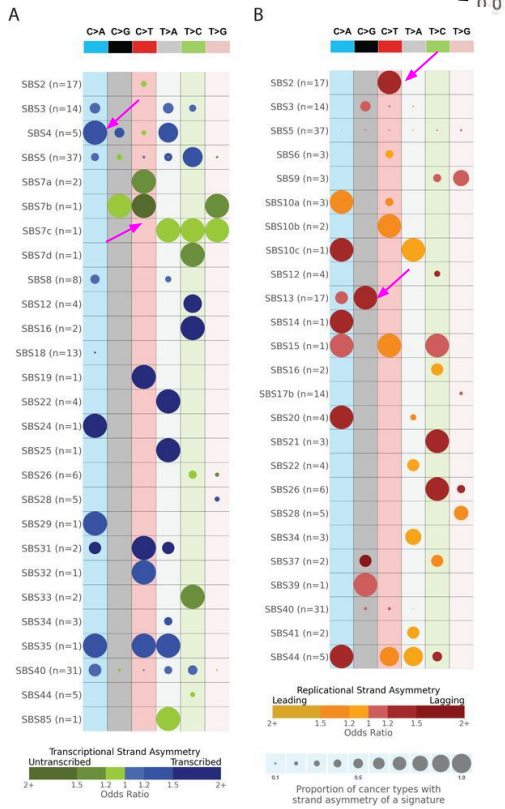
Enrichment analysis



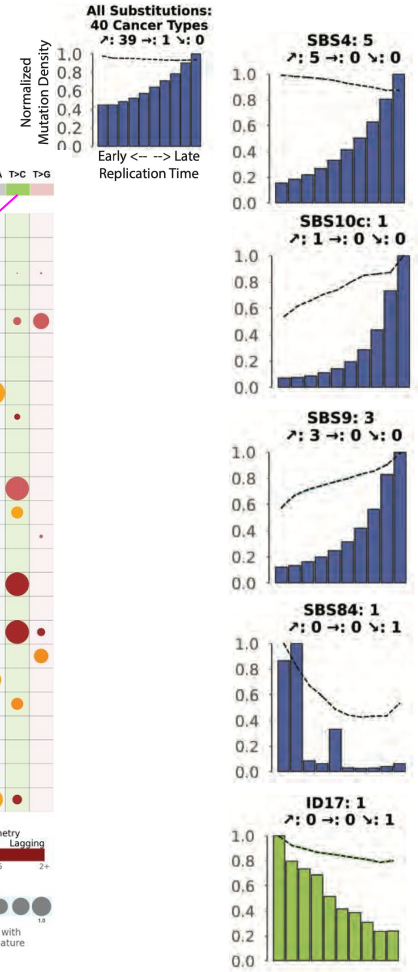
Topography analysis of mutational signatures



Otlu et al., *BioRxiv*, 2022



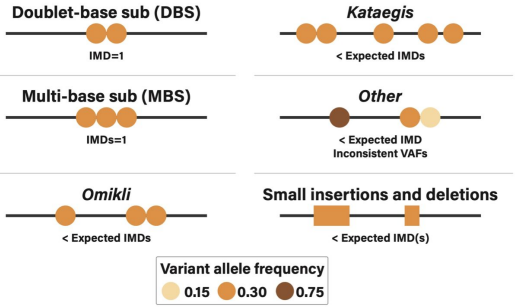
Strand asymmetries and strand-coordinated mutagenesis of mutational signatures



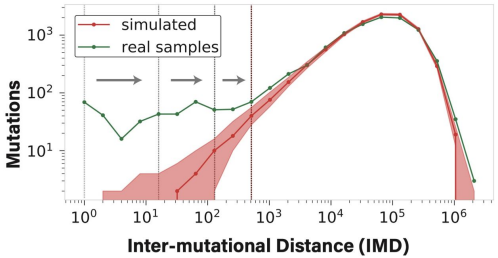
Interplay between replication timing and mutational signatures

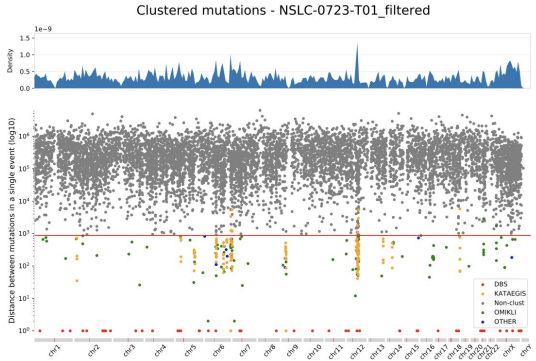
Analysis of clustered mutations

Examining clustered somatic mutations with SigProfilerClusters

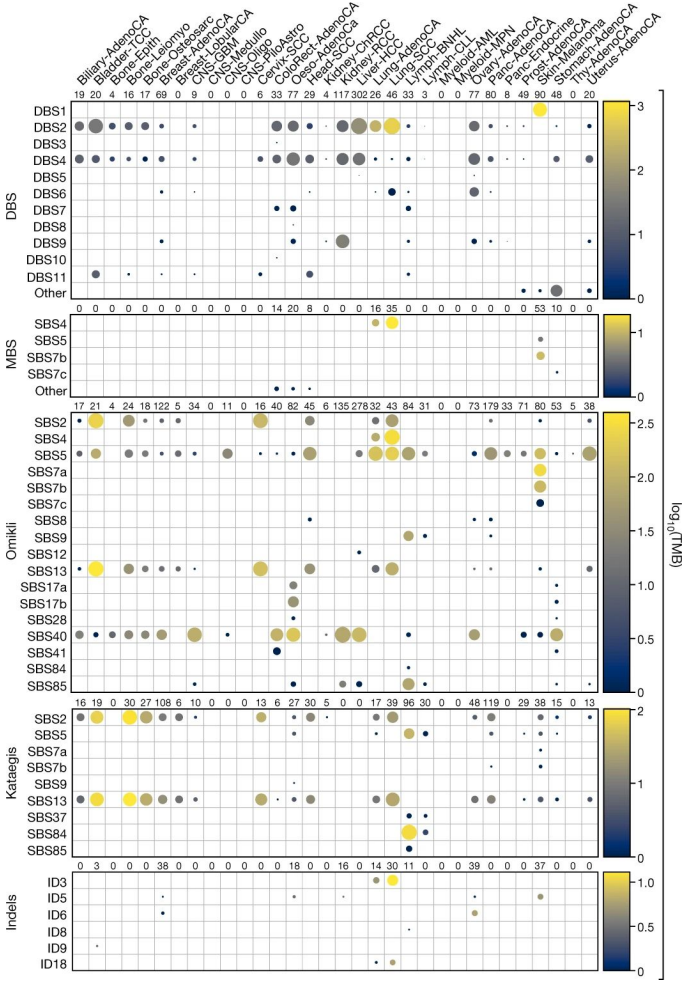


Detection of global IMD threshold





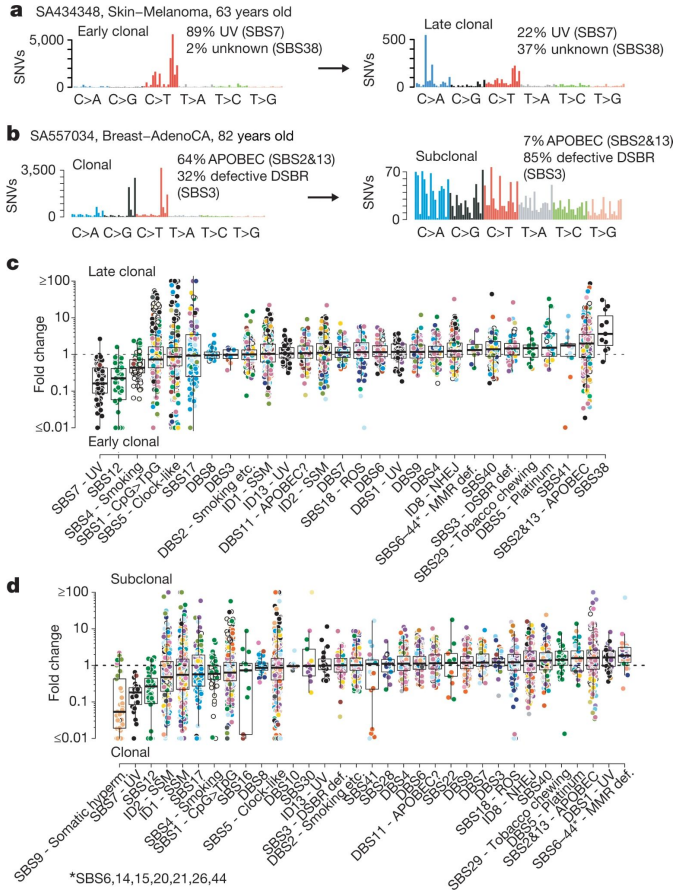
Bergstrom, *et al.*, *Bioinformatics*, 2022



Mutational process that underlie clustered events

Bergstrom et al., *Nature*, 2022

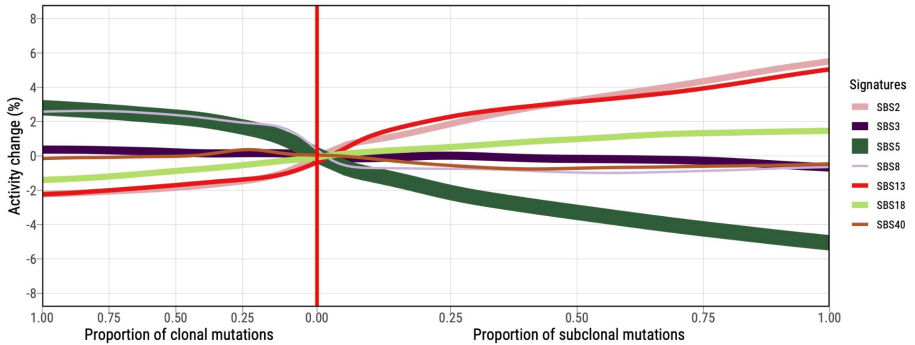
Dynamics of mutational signatures over cancer evolutionary time



Timing of mutational signatures

Gerstung, M. et al., *Nature*, 2020

Signature activity trajectories for Sherlock-*Lung* samples



Zhang, *et al.*, Nature Genetics, 20201

TrackSig Rubanova *et al.*, *Nature communications*, 2020

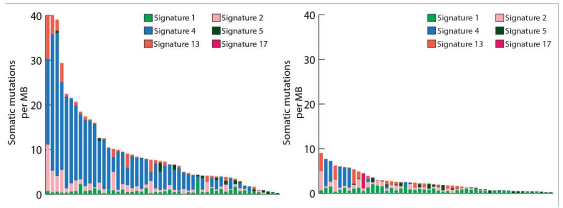
Validation of mutational signatures

(Supporting evidence for mutational signature validity)

- Mutational signature can be replicated in multiple studies or validated in orthogonal techniques (NGS techniques, variant callers, sequencing centers *etc*.)
- Proposed etiology associated with mutational signature.

For example, signature SBS4 is likely related to tobacco smoking, How this is validated?

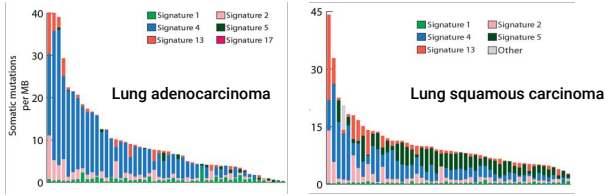
❏ Contributions of signatures to lung adenocarcinoma in smokers vs. non-smokers



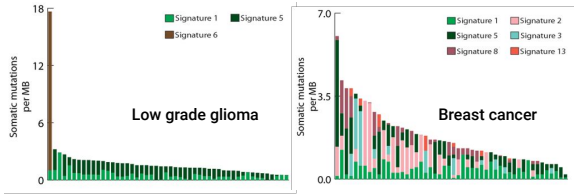
Tobacco smokers | Lifelong non-smokers

❏ Contributions of signatures to smoking induced and non-smoking induced cancer types

Smoking induced cancer types



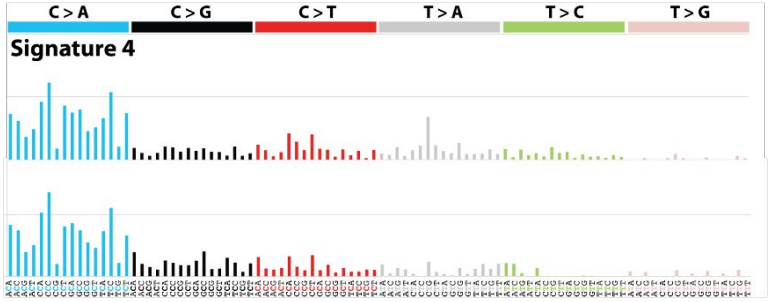
Non-smoking induced cancer types



❏ Similarity with experimental benzo[a]pyrene (known tobacco carcinogen) signature

Signature SBS4 extracted from human cancers

Signature of benzo[a]pyrene exposure *in vitro*

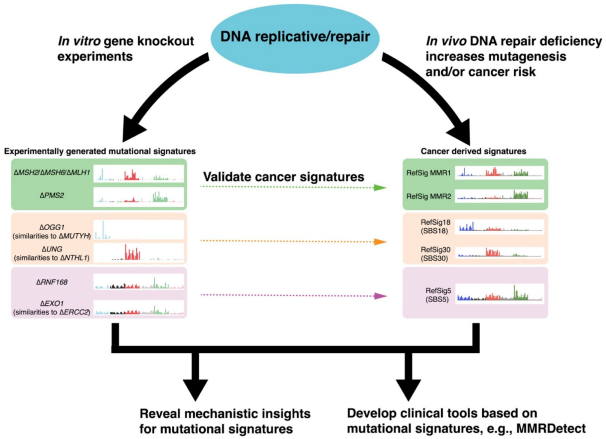


Validation of mutational signatures

(Experimental validation)

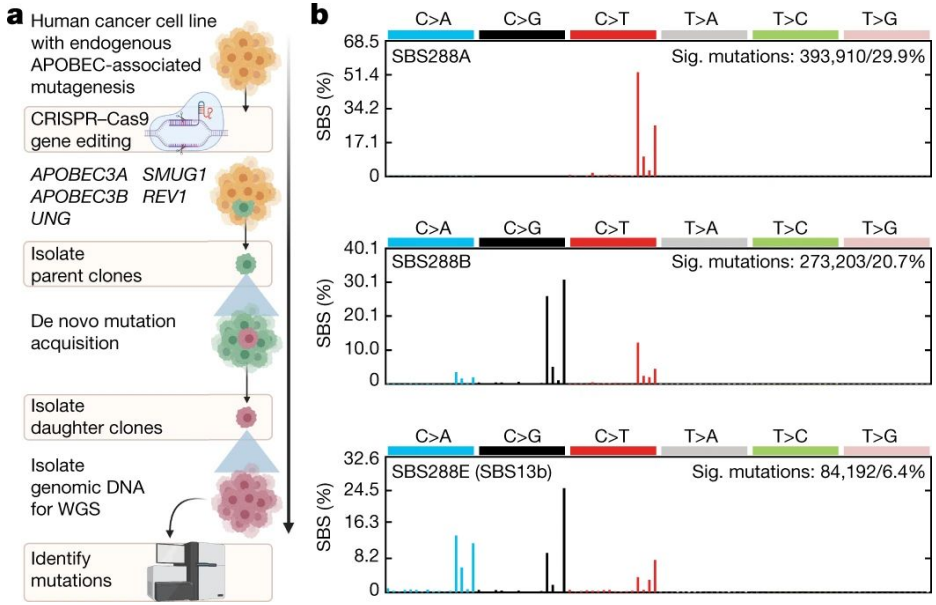
Mutational signature can be validated in experimental study:

- Cellular model systems, including *C. elegans*, yeast, human cancer cell lines, organoids, and human induced pluripotent stem cells.
- Experimental design including genetic manipulation (*e.g.*, CRISPR KO) and treatments (exposure to environmental carcinogens).



Impact of experimental validation of cancer-derived mutational signatures on biological understanding and development of clinical applications

Zou *et al.*, *Nature Cancer*, 2021



Using human cancer cell lines to investigate the origins of APOBEC3-associated mutagenesis.

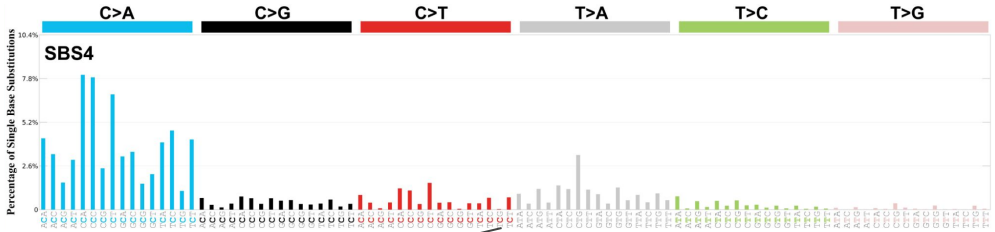
Petljak *et al.*, *Nature*, 2022

Mutational Signature webtools and data portals

COSMIC Mutational Signatures

- Most recent version: v3.3 (June 2022)
- Signatures extracted using SigProfiler from 2,780 whole-genome variants calls produced by ICGC/TCGA PCAWG Network

Signature Type	Number of Signatures
SBS	60 real, 18 possible artefacts
DBS	11 real
ID	18 real
CN	21 real, 3 artifacts



SBS4 signature features

- **Mutational profile**
- Proposed aetiology
- Acceptance criteria
- Tissue distribution
- Associated signatures
- Differences

SBS4 topographical features

- Replication timing
- Nucleosome occupancy
- CTCF occupancy
- Histone modifications
- Transcriptional strand asymmetry
- Genic and intergenic regions
- Replicational strand asymmetry
- Strand-coordinated mutagenesis

Supporting evidence for mutational signature validity

Validated evidence for *real signature*
Unclear evidence for *real signature*
Evidence for *artefact signature*

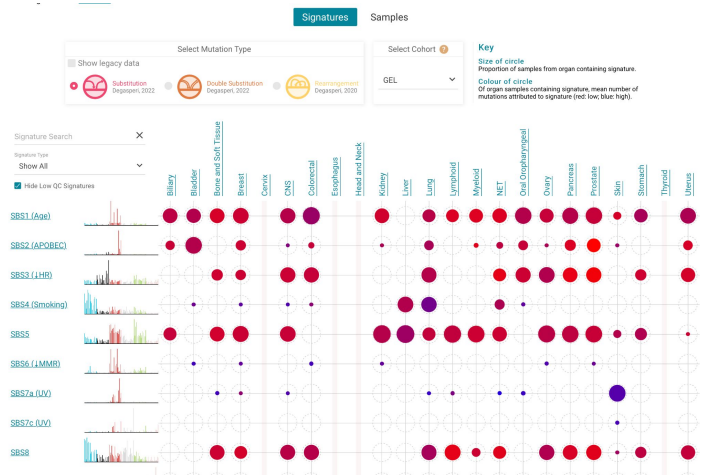
Background	**Identification study**		**First included in COSMIC**
	Alexandrov *et al.* 2013 Nature		v1
Identification	**NGS technique**	**Different variant callers**	**Multiple sequencing centres**
	WES & WGS	Yes	Yes
Technical validation	**Validated in orthogonal techniques**	**Replicated in additional studies**	**Extended context enrichment**
	Yes	Yes	-
Proposed aetiology	**Mutational process**		**Support**
	Tobacco smoking		Experimental confirmation
Experimental validation	**Experimental study**		**Species**
	Nik-Zainal *et al.* 2015 Mutagenesis		Mouse

Summary of the technical and experimental evidence available in the scientific literature regarding the validation of the mutational signature.

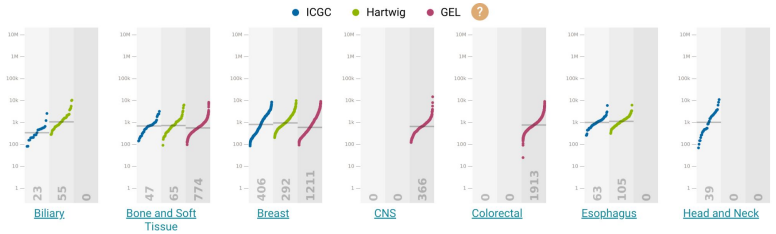
Signal

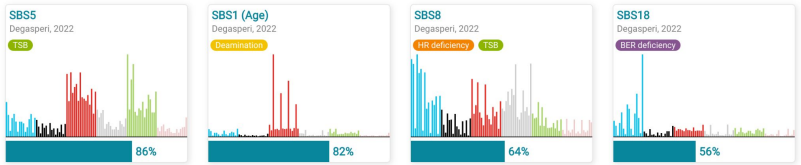
- Explore different source of signatures (cancer, gene edits, environmental mutagenesis)
- Support different mutational profiles (SBS/ID/DBS/RS)
- Support signature data analyses using signature.tools.lib
- Explore largest cancer genomic studies, allow to discovery the rare mutational signatures in cancer

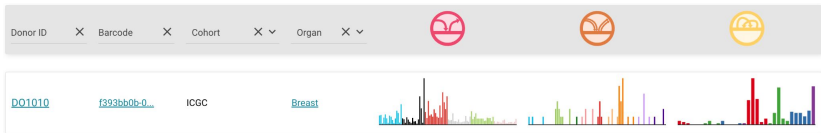
Enhanced Interactive visualization for:
Study, Cancer Type, Signature, Sample

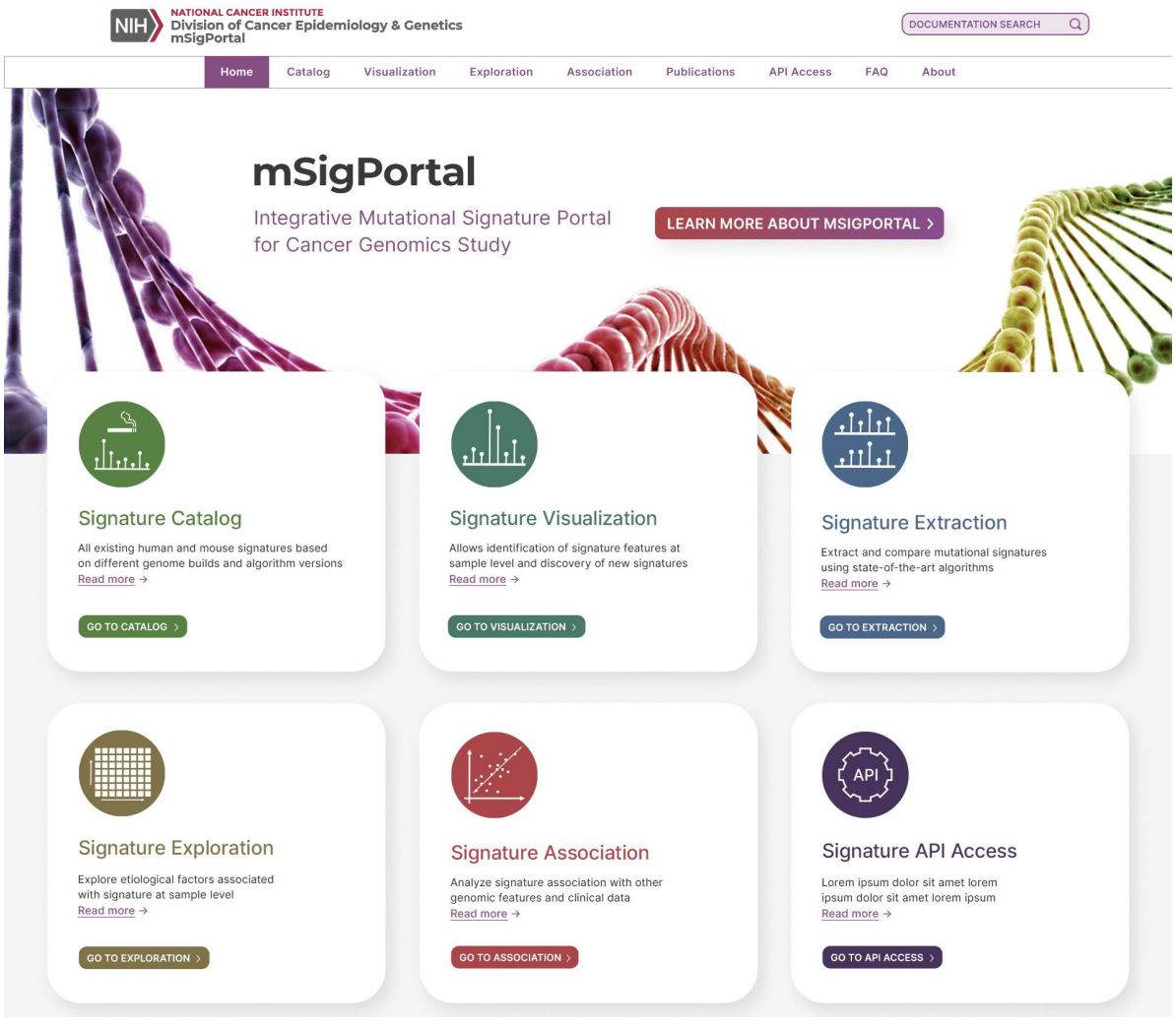


Degasperi *et al.*, *Nature Cancer*, 2020









mSigPortal production site:

https://analysistools.cancer.gov/mutational-signatures

mSigPortal development site:

https://analysistools-qa.cancer.gov/mutational-signatures

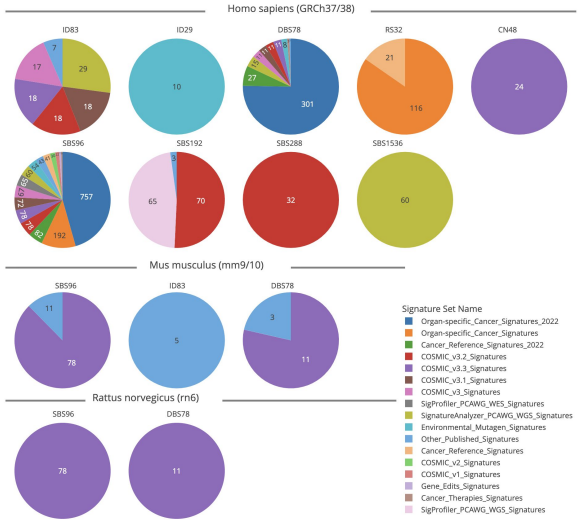
- **Broad topics related to mutational signatures (knowledges, findings, studies, data, analyses)**
- **Comprehensive signature analyses for user (VCF -> profiler generation -> signature extraction -> signature evaluation)**
- **Interactive data visualization across all modules.**
- **API access following FAIR principles.**

Studies includes: TCGA, PCAWG, ICGC, Breast Cancer 560, Chernobyl Thyroid, Mutographs ESCC, GEL, Hartwig, Sherlock-Lung, LCM normal tissues.

mSigPortal: Signature Catalog

Explore all existing multi-species and different sources of mutational signatures (e.g., genome builds and algorithm versions) with detailed annotations.

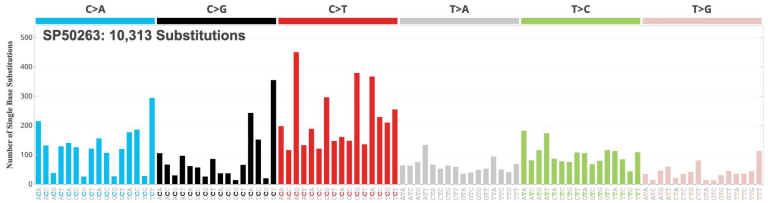
- Search current reference mutational signatures in the literature across major categories, including:
  - COSMIC Mutational Signatures
  - Environmental Mutagenesis
  - DNA Repair Gene Edits
  - Cancer Specific Signatures
  - Cancer Therapies
  - Others from literature
- Include 3,247 mutational signatures collected form literature across different profiles (SBS/DBS/ID/CN/RS).
- Provide signature comparison functions



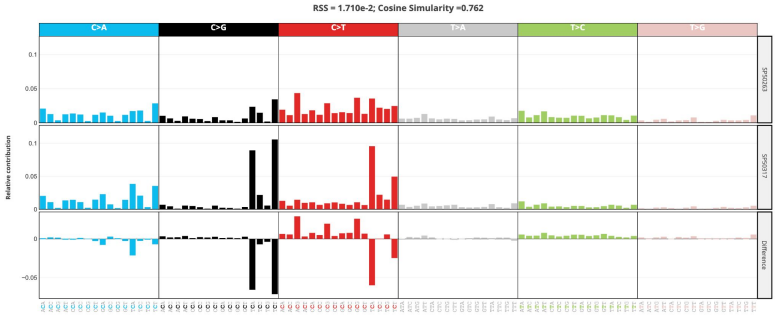
mSigPortal: Signature Visualization

Allows identification of signature features at sample level and discovery of new signatures

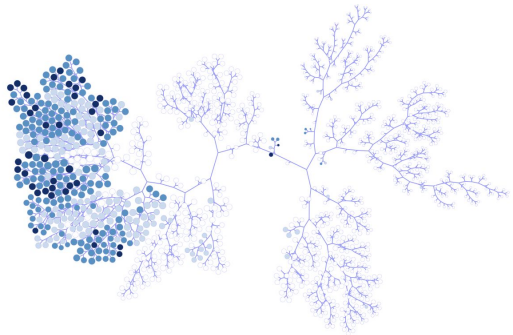
- Visualize and analyze mutational profiles from user input and published studies in the literature
- Perform wide range of analyses, including:
  - Mutation profile extractions (SigProfilerMatrixGenerator)
  - TreeAndLeaf visualizaiton
  - Cosine similarity calculation
  - Mutational pattern enrichment analysis
  - Profile comparison
  - Principal components analysis
  - Clustered mutations identification (SigProfilerClusters)



Mutational Profiles



Profile comparison

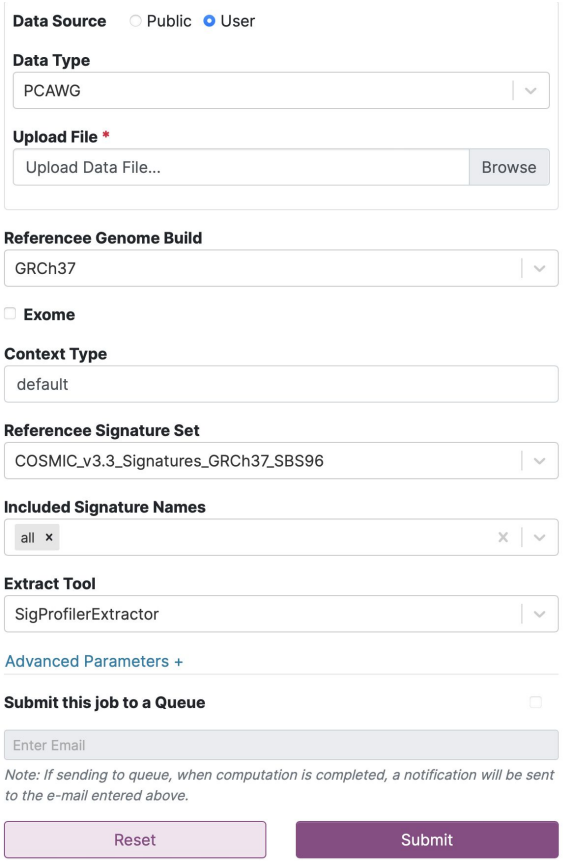


TreeAndLeaf

mSigPortal: Signature Exaction

Extract and compare mutational signatures using state-of-the-art algorithms

- Supported CPU/GPU use for signature de novo extraction and decomposition based on the most popular algorithms: **SigProfiler**; Signal, MuSiCal and SignatureAnalyzar will be added soon.
- Analysis result will be automatically imported to ‘Signature Exploration’ module for visualization.
- Flexible selection of number of reference signatures.

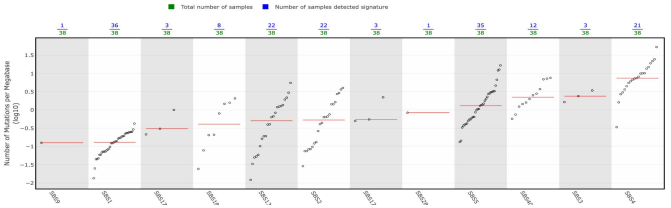


Signature Extraction input parameters

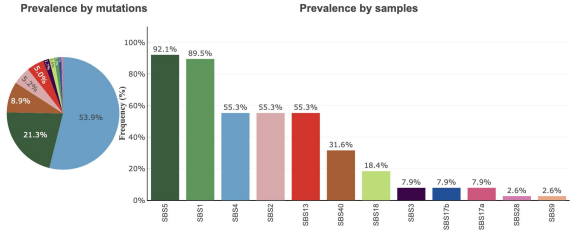
mSigPortal: Signature Exploration

Explore etiological factors association with signature at sample level

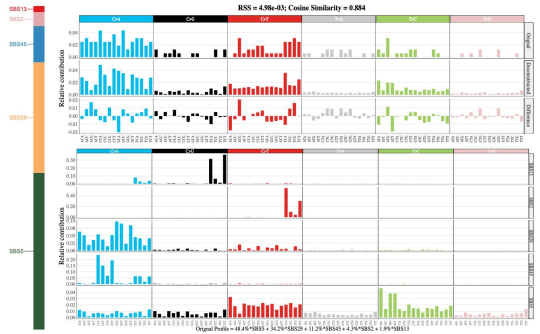
- Perform analyses for mutational signatures activities.
- Perform wide range of visualization, including:
  - Tumor mutational burden (overall or by signature)
  - Evaluation of mutational signature decomposition
  - Mutational signature association
  - Mutational signature landscape
  - Mutational signature prevalence
  - Mutational signature decomposition in single sample



Mutational Burden by Signatures



Mutational signature Prevalence



Visualization of mutational signature in single sample

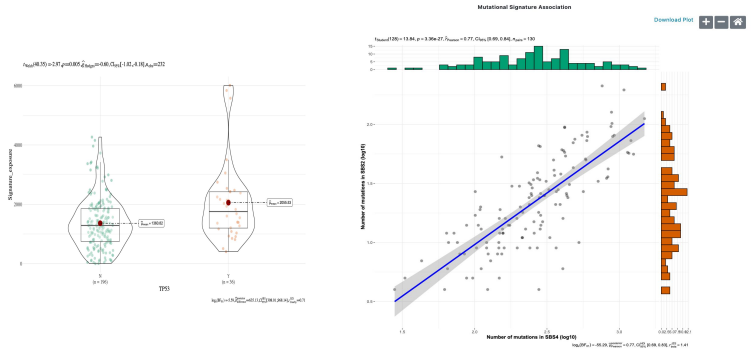
mSigPortal: Signature Association

Analyze signature association with other genomic features and clinical data

- Analyze and visualize associations between mutational signature activities and sample level variables (*e.g.*, other genomic features, epigenomic features, mutational status, copy number alterations and clinical data).
- Allows of selection of different statistical approaches for both univariable and multivariable association analyses.
- Current version supported the following studies: TCGA, PCAWG, and Sherlock-*Lung*.

Data Source	Data Type	Number of Variables
quality control	sequencing metrics	13
clinical data	clinical variables	22
genomic data	complex sv rearrangement	2
genomic data	consensus cnv features	3
genomic data	evolution_and_heterogeneity	32
genomic data	GISTIC_actual_copy_change	131
genomic data	GISTIC_lesions_level	131
genomic data	msi	1
genomic data	mutational burden	15
genomic data	panorama driver mutations	573
genomic data	sv count	6
genomic data	telomere maintenance	23
genomic data	transcriptome fusion	4515
genomic data	viral genome	18
germline data	ancestry	7
Gene-based data	mutation status	-
Gene-based data	expression level	-
Gene-based data	Copy nunber status	-

Table1. Example of variables included in Signature Association for the PCAWG study



THANKS FOR YOUR ATTENTION!
Questions?

Next: Practical session 5 (10:45am)

- Deciphering mutational signatures using SigProfiler tools (including profile extraction, de-novo and decomposition signature analyses)
- Explore the mutational signature data portals (*e.g.*, mSigPortal)

Invited Speakers:

Title: Anthology of unusual patterns of somatic mutations in cancer genomes

Ludmil Alexandrov, M.Phil., Ph.D.

University of California San Diego

January 19th, 2023

Ludmil Alexandrov is an Associate Professor at the University of California, San Diego (UCSD). Dr. Alexandrov received his Ph.D. in 2014 from the University of Cambridge researching mutational processes and signatures in human cancers at the Wellcome Sanger Institute. Dr. Alexandrov then went on to research as an Oppenheimer Fellow at the Los Alamos National Laboratory from 2014 to 2017 before becoming an Assistant Professor of Bioengineering and of Cellular and Molecular Medicine at UCSD in 2018. He was appointed as an Associate Professor at UCSD in 2021.

His research on mutational signatures and algorithms for mutational signature decomposition in human cancers has received numerous awards and recognition: recognition from the American Society of Clinical Oncology (2014), the Fred Hutchinson Cancer Center's Harold M. Weintraub Award (2015), Science magazine's Prize for Young Scientists in Genomics and Proteomics (2015), Oxford University Press' Carcinogenesis Young Investigator Award (2016), Alfred P. Sloan Research Fellowship in Computational & Evolutionary Molecular Biology (2018), the Balfour Prize Lecture of the Genetics Society (2018), The International Academy for Medical and Biological Engineering's Early Career Award (2018), the Packard Foundation's Packard Fellowship for Science and Engineering (2019), and the Outstanding New Environmental Scientist Award from the National Institute of Environmental Health Sciences (2020).

Dr. Alexandrov's many publications have been cited 36,858 times as of August 2022. His lab develops and maintains the highly popular SigProfiler software suite for mutational signature analysis, and collaborates with Wellcome Sanger Institute to maintain the COSMIC catalogue of mutational signatures.