Mutational Signatures

Massimo Maiolo

Institute of Tissue Medicine and Pathology – University of Bern | 2.10.2025

Introduction

- Understanding Mutational Signatures
- Why mutational signatures matter
- Common Causes of DNA Damage
- Mutagenic factors

Understanding Mutational Signatures

What is a Mutational Signature?

- A specific pattern of mutation types in DNA
- Caused by a particular biological process, like a mutagen or a cellular defect

- **Not a 1-to-1 Match:**
  - A signature is usually not unique to a single tumor type or a specific mutagen
  - Many different causes can lead to the same signature

- **The Big Picture:**
  - Deciphering signatures in a patient's tumor provides insight into the biological mechanisms that led to the cancer
  - This helps us understand the cause and can guide treatment strategies

Terminology

- CUP (Cancer of Unknown Primary):
  - A cancer whose original location in the body cannot be determined
  - Mutational signatures can be used to infer the tissue of origin, helping doctors to better treat the disease

Why mutational signatures matter

Mutational signatures matter because they act as a forensic record of a cell's history, **revealing the processes that have damaged its DNA**. By analyzing these "fingerprints" left on the genome, scientists and doctors can gain crucial insights into the causes of a disease, particularly cancer.

Why mutational signatures matter

- **Uncovering the Causes of Cancer**

  Mutational signatures can help scientists understand what causes cancer. Each signature is linked to a specific process. For example, a signature caused by ultraviolet (UV) radiation is found in melanoma, while another is linked to cigarette smoke and is found in lung cancer. This provides direct evidence of the role of these factors in tumor development.

- **Improving Cancer Diagnosis**

  Signatures can provide more information than a standard diagnosis. For instance, two tumors might look identical under a microscope, but their mutational signatures could be very different. This difference can reveal a lot about the tumor's origin, which can help in choosing the right treatment.

- **Guiding Treatment Decisions**

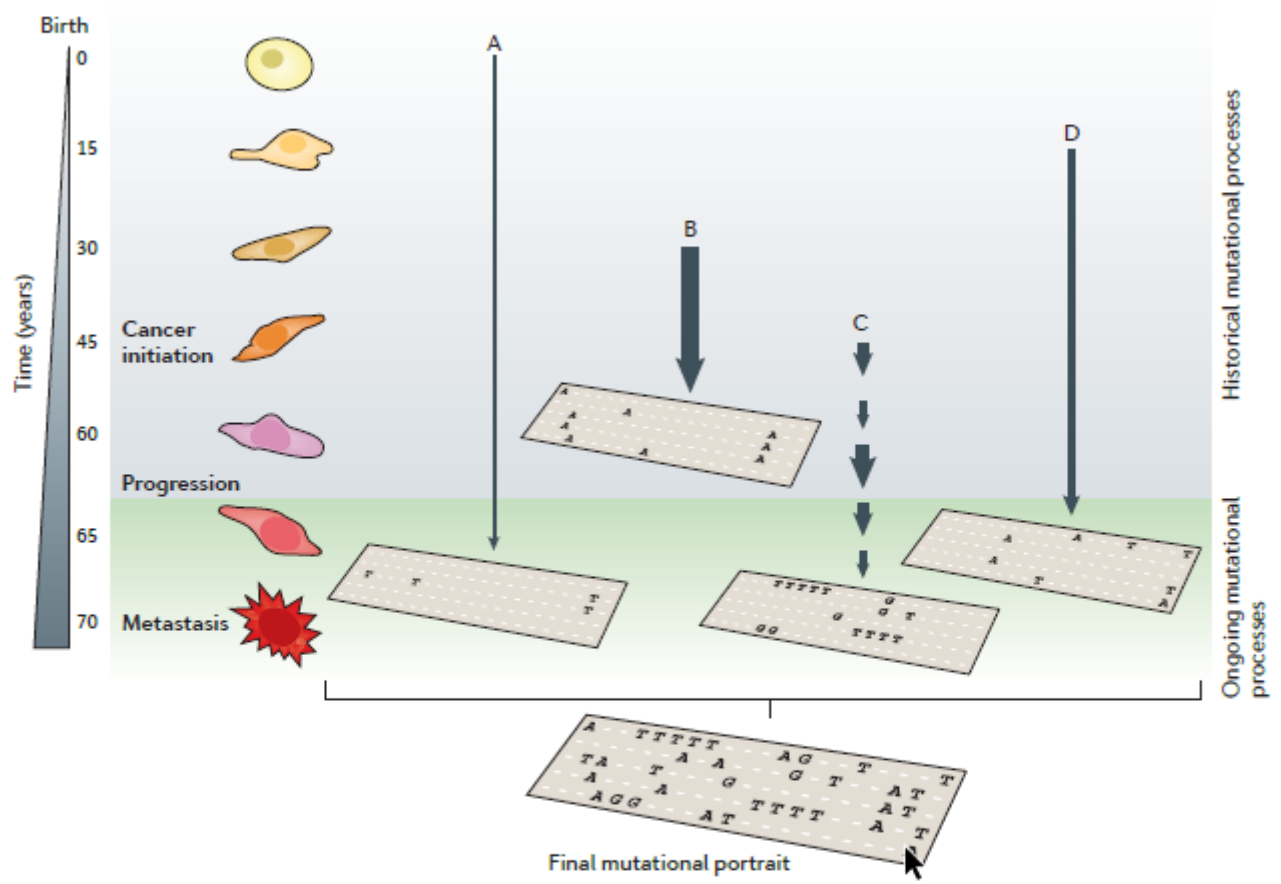
  Mutational signatures can help predict how a patient will respond to certain therapies. Some signatures indicate that a tumor will be sensitive to specific drugs. For example, a signature caused by a defect in DNA repair (e.g., in BRCA-mutated breast cancer) suggests that the tumor will likely respond well to targeted therapies like PARP inhibitors.

- **Monitoring Treatment Response**

  In some cases, changes in mutational signatures over time can be used to track a tumor's evolution and its response to treatment. If a new signature appears, it might indicate that the tumor is developing resistance to a drug, allowing doctors to adjust the treatment plan.

Mutational signatures transform our understanding of genetic damage from a simple list of changes into a rich, informative story about how and why a disease developed. They are a powerful tool for precision medicine, helping to make diagnosis more accurate and treatment more effective.

Active mutational processes



Mutational signature
The pattern of mutations produced by a mutational process.

Mutational portrait
The total genetic changes observed in a cancer genome; that is, the sum of all mutational signatures occurring in a lifetime.

Base substitutions
A type of mutation in which one base is replaced by another in DNA.

Insertions and deletions
(Indels). A type of mutation that arises from the insertion or deletion of one or more nucleotides within a DNA sequence.

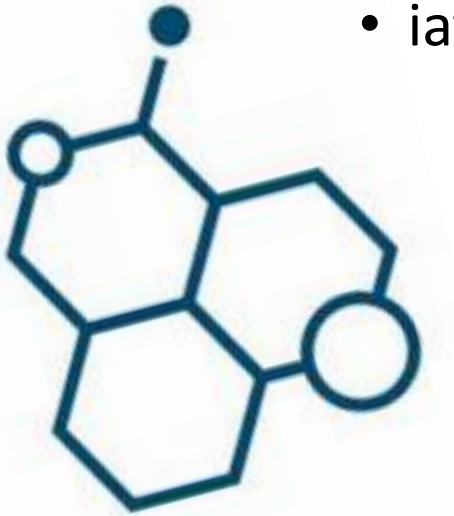
Figure 1 | **Active mutational processes over the course of cancer development.** Each mutational process leaves a characteristic imprint — a mutational signature — in the cancer genome and comprises both a DNA damage component and a DNA repair component. In this hypothetical cancer genome, arrows indicate the duration and intensity of exposure to a mutational process. The final mutational portrait is the sum of all of the different mutational processes (A–D) that have been active in the entire lifetime. Ongoing mutational processes reflect active biological processes in the cancer that could be exploited either as biomarkers to monitor treatment response or as therapeutic anticancer targets. By contrast, historical mutational processes are no longer active. Signature A represents deamination of methylated cytosines, which is ongoing through life. Signature B can be matched up with the signatures of tobacco smoking. Signature C can represent bursts of APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide)-induced deamination, and Signature D represents a DNA repair pathway that is awry.

Helleday et al, Nat Rev Genet 2014

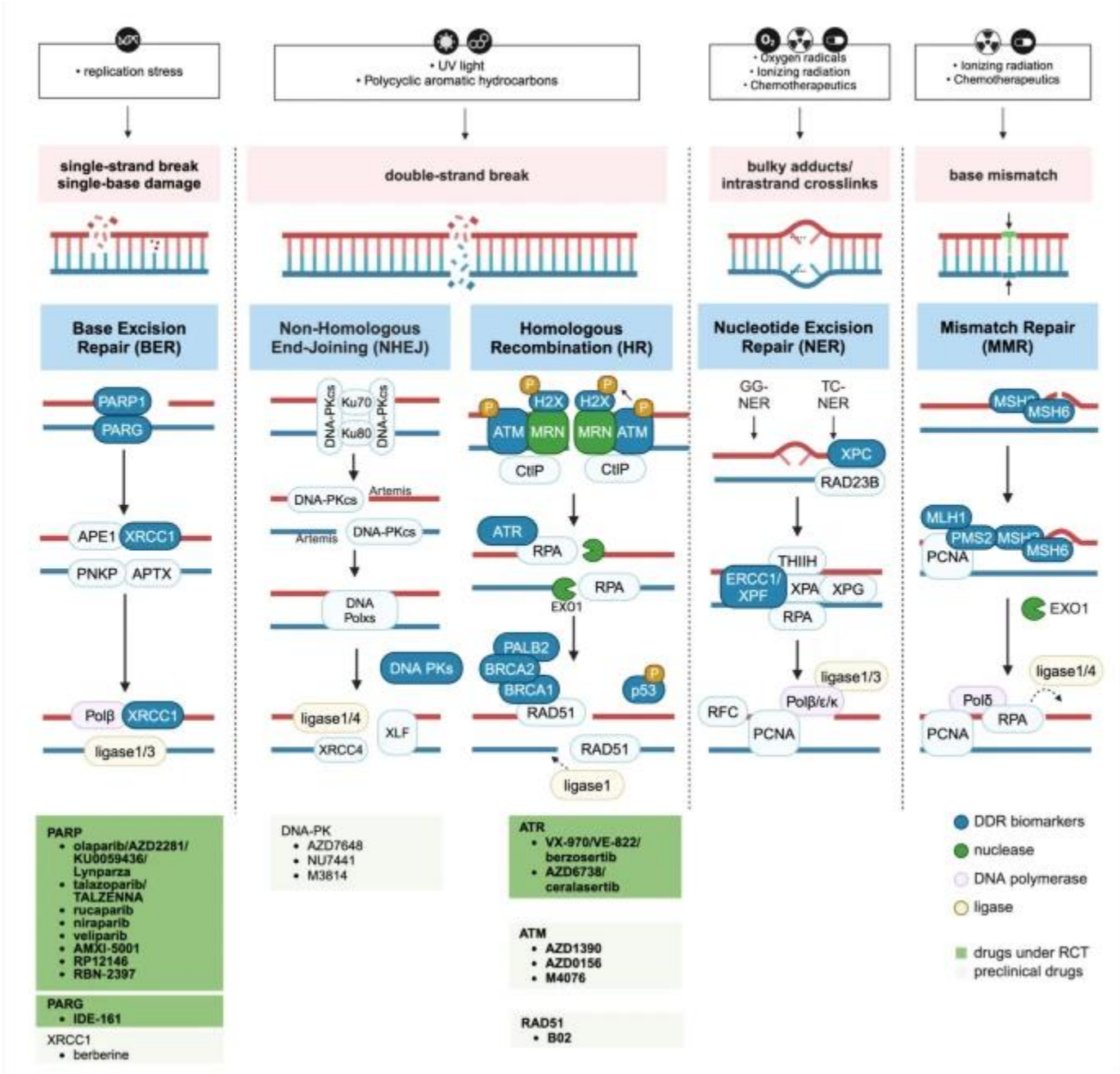
Mutagenic factors

Mutagenic factors are agents that can cause mutations by damaging DNA. These factors can be broadly classified into three categories:

- environmental
- endogenous (internal to the body)
- iatrogenic (resulting from medical treatment)



Common Causes of DNA Damage



Li et al., Biomark Res 12, 111 (2024)

Environmental Mutagens

- **Tobacco Smoke**

  Contains a complex mixture of chemicals, including polycyclic aromatic hydrocarbons (PAHs), which are potent mutagens that can cause specific mutational signatures in lung, bladder, and other cancers

- **Ultraviolet (UV) Light**

  The UV radiation from sunlight or tanning beds is a primary cause of melanoma and other skin cancers. It creates characteristic "UV damage" signatures on the DNA

- **Aflatoxin**

  A toxic compound produced by certain molds found on food products like corn, peanuts, and tree nuts. It is a major cause of liver cancer, particularly in regions with poor food storage practices

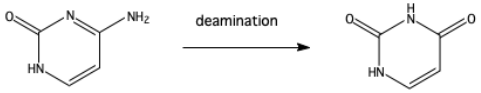
- **Aristolochic Acid**

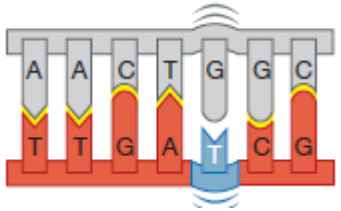
  Found in plants of the Aristolochia genus, often used in traditional herbal medicines. It is a powerful carcinogen linked to kidney and liver cancers

- **Air Pollution**

  Particulate matter and other pollutants in the air contain various carcinogens that can contribute to lung cancer

Endogenous Mutagens





- **Reactive Oxygen Species (ROS)**

  Produced during normal metabolic processes, ROS can damage DNA, leading to mutations

- **Spontaneous Deamination**

  This is a natural chemical reaction where DNA bases can be altered, such as cytosine changing to uracil, which can lead to mutations during DNA replication

- **DNA Replication Errors**

  While DNA polymerase has proofreading capabilities, mistakes can still occur during the replication process, leading to a natural background rate of mutations

Iatrogenic Mutagens

- **Chemotherapy**

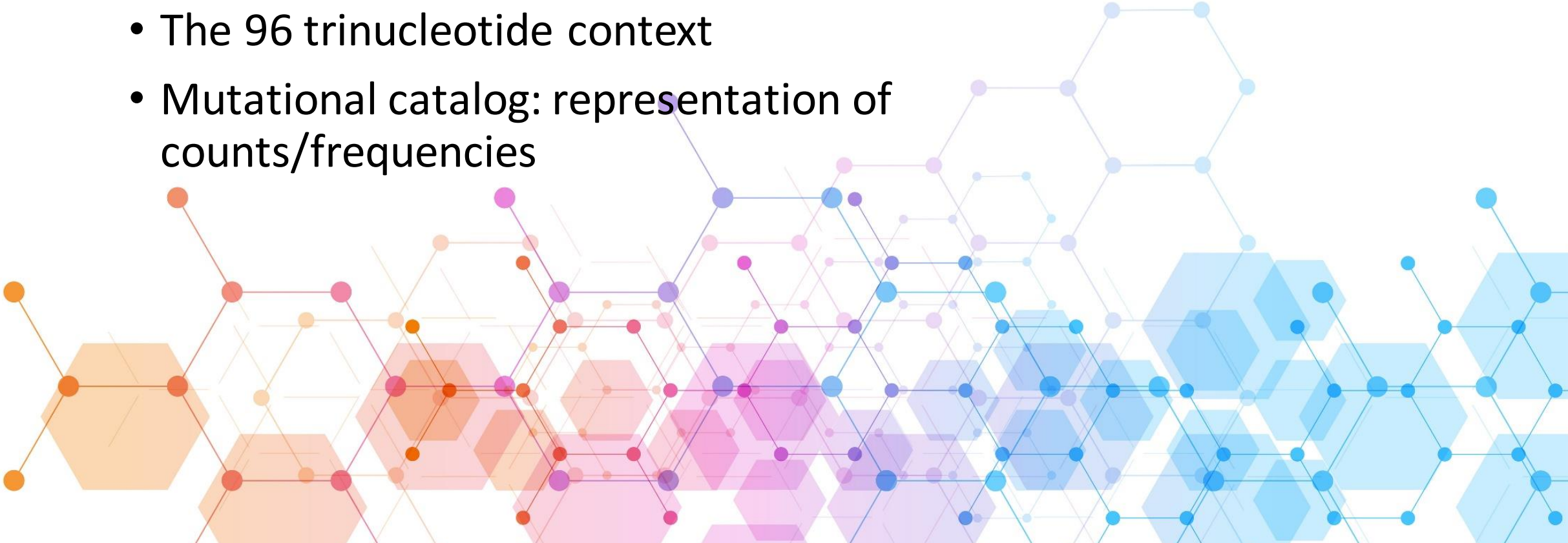
  Many chemotherapy drugs, such as platinum compounds, are designed to damage DNA and kill cancer cells. As a side effect, they can also induce new mutational signatures in the surviving cells, sometimes leading to therapy-related cancers.

- **Radiation Therapy**

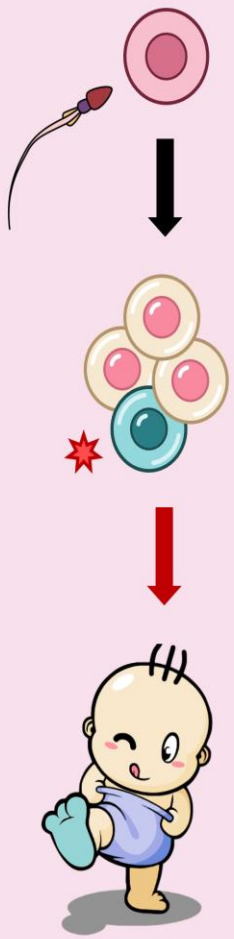
  Similar to UV light, high-energy radiation used in cancer treatment can cause extensive DNA damage and introduce unique mutational patterns in the exposed tissue.

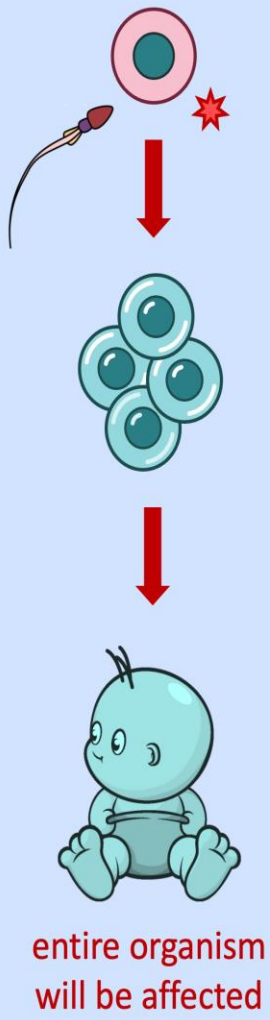
Concepts and Definitions

- Somatic mutations vs germline mutations
- Types of mutations: SNVs, indels, structural variants
- The 96 trinucleotide context
- Mutational catalog: representation of counts/frequencies



Somatic mutations vs germline mutations





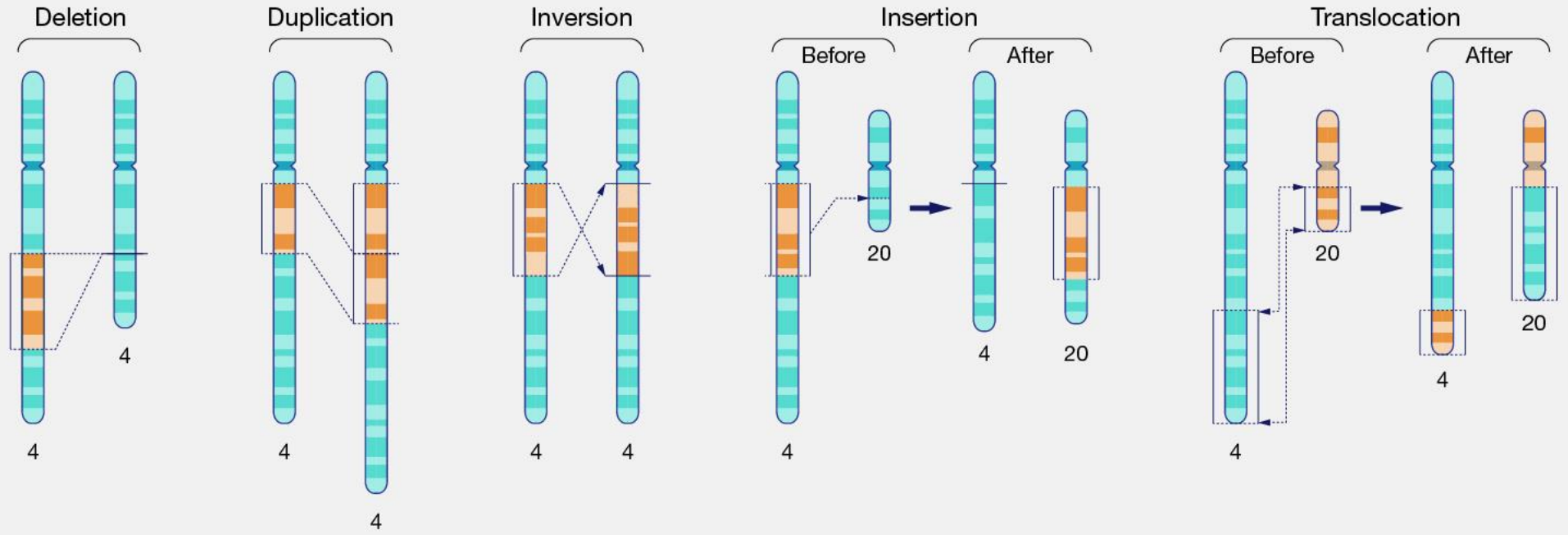
Somatic mutation	Germline mutation
It is defined as any alteration in the genetic sequence of genes of the somatic cells.	It is also called “acquired mutation” as it is acquired during an individual’s life.
It is defined as any alteration in the genetic sequence of genes of the germinal cells.	It is also called “hereditary mutation” as it is passed to offspring.
It occurs in somatic or body cells.	It occurs in germ cells.
It is not inheritable.	It is inheritable.
It affects only mutated cells and their progeny.	It affects all the cells of the organism.
It can occur at any stage of the life cycle.	It occurs only during gametogenesis.
It plays no role in evolution.	It is the basis of evolution.
It does not cause genetic disorders but may cause cancer.	It is responsible for genetic disorders and also germline cancers.
It is finished along with the death of the individual.	It continues till the offspring of the individual breed and may result in separate sub-species. Hence it is genetically more important.
In most cases, they show observable effects.	In most cases, they are silent and don’t show detectable effects.
They can be treated or cured.	They can’t be treated or cured.
Mosaicism occurs.	Usually, mosaicism doesn’t occur.
Examples include cancers, tumor development, a neurodegenerative disorder, etc.	Examples include Hemophilia, Down’s Syndrome, 18-Trisomy, etc.

Germline mutation filtering

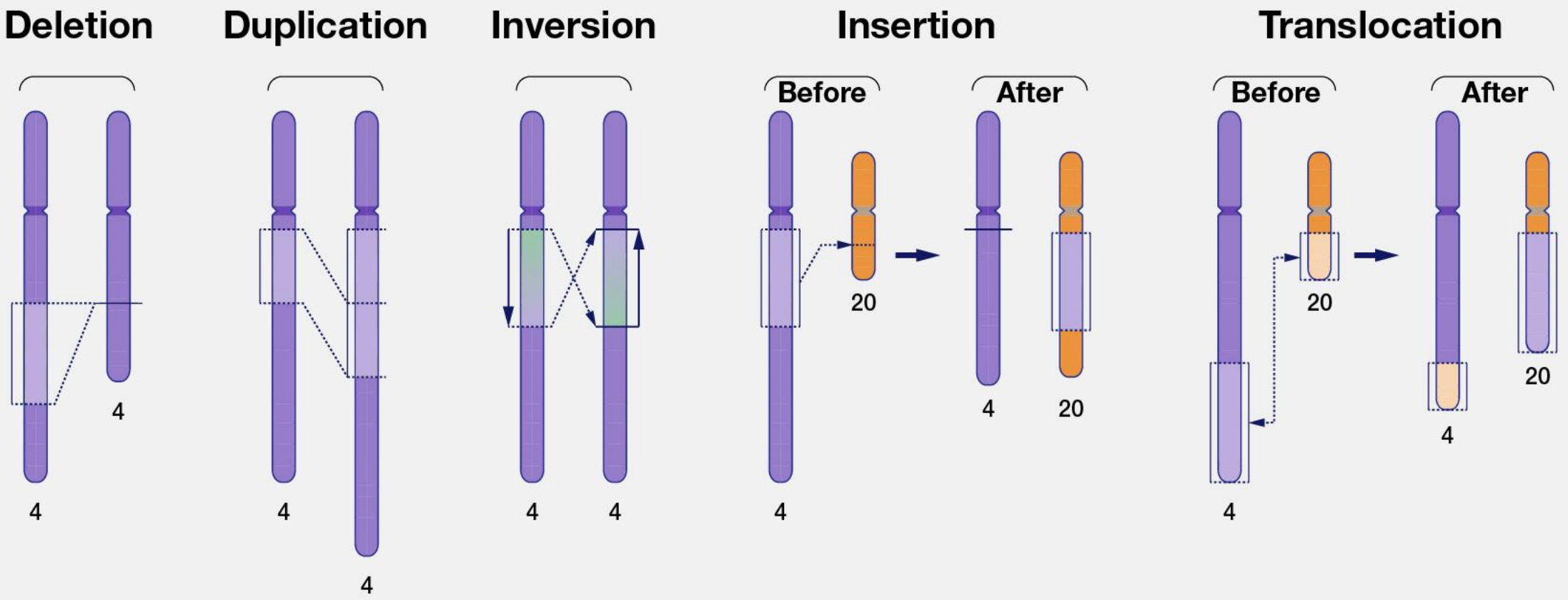
- For mutational signature analysis, the fundamental principle is to filter out all germline variants and focus exclusively on somatic mutations.
- **Germline Mutations**
    - *These are genetic changes present in every cell of an organism. They are inherited from parents and are not associated with a specific disease process, such as cancer.*
- **Somatic Mutations**
    - *These are genetic changes that occur after conception in a single cell and are subsequently passed on to its daughter cells. They are not inherited and are typically caused by environmental factors, endogenous processes (like DNA replication errors), or specific mutational processes that lead to diseases like cancer.*
- If you were to include germline variants in your analysis, you would be mixing inherited, non-cancer-related mutations with the acquired, disease-causing somatic mutations. This would completely obscure the subtle patterns that define mutational signatures, making it impossible to accurately identify and attribute them to the correct underlying causes.

Analyzing mutational signatures requires a pristine dataset of only the mutations that have arisen in the tumor, which is why a robust filtering pipeline to identify and remove all germline variants is a mandatory first step.

Cancer : a disease driven by accumulated genetic alterations



Types of mutations



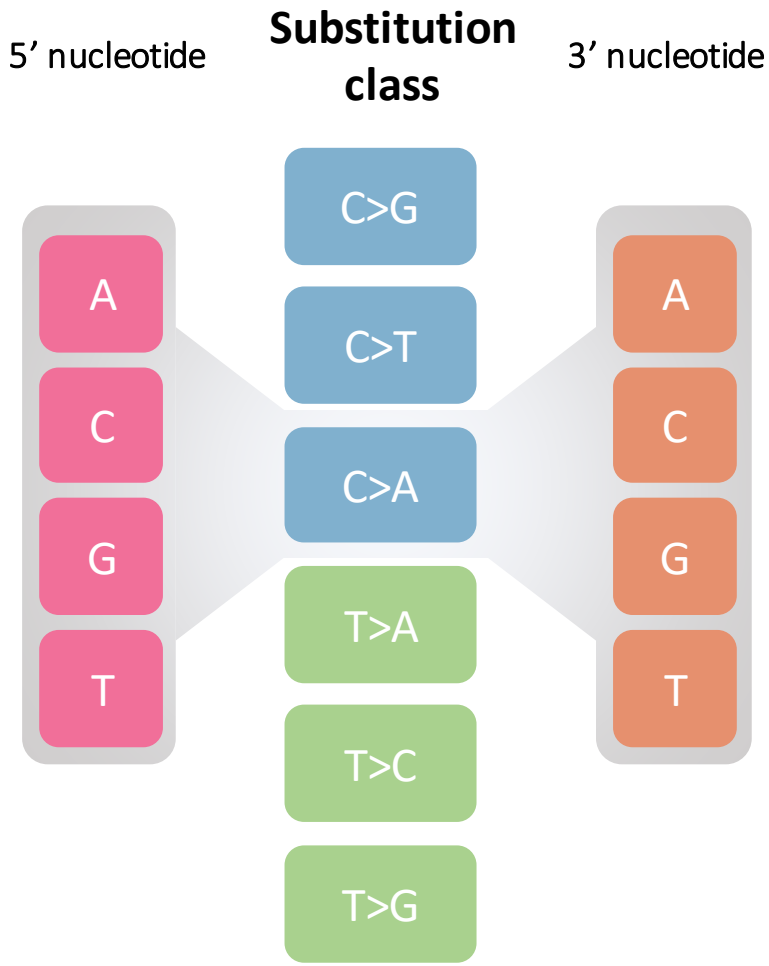
Single base substitutions (SBS)

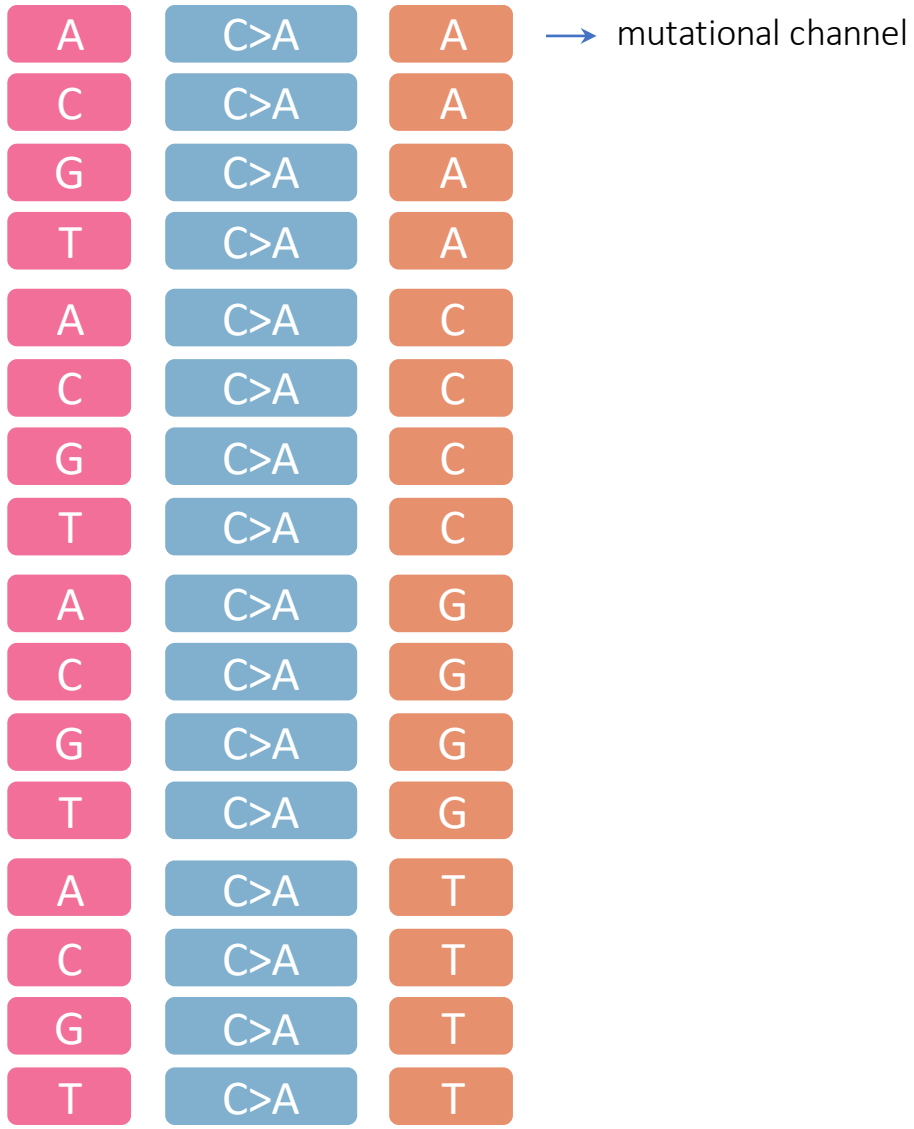
SBS-6 classification (base substitution)

- single base substitutions can be described using only the mutated base-pair
- 6 possible mutational channels known as **SBS-6 classification**

SBS - 96

trinucleotide context (± 1 bp context)

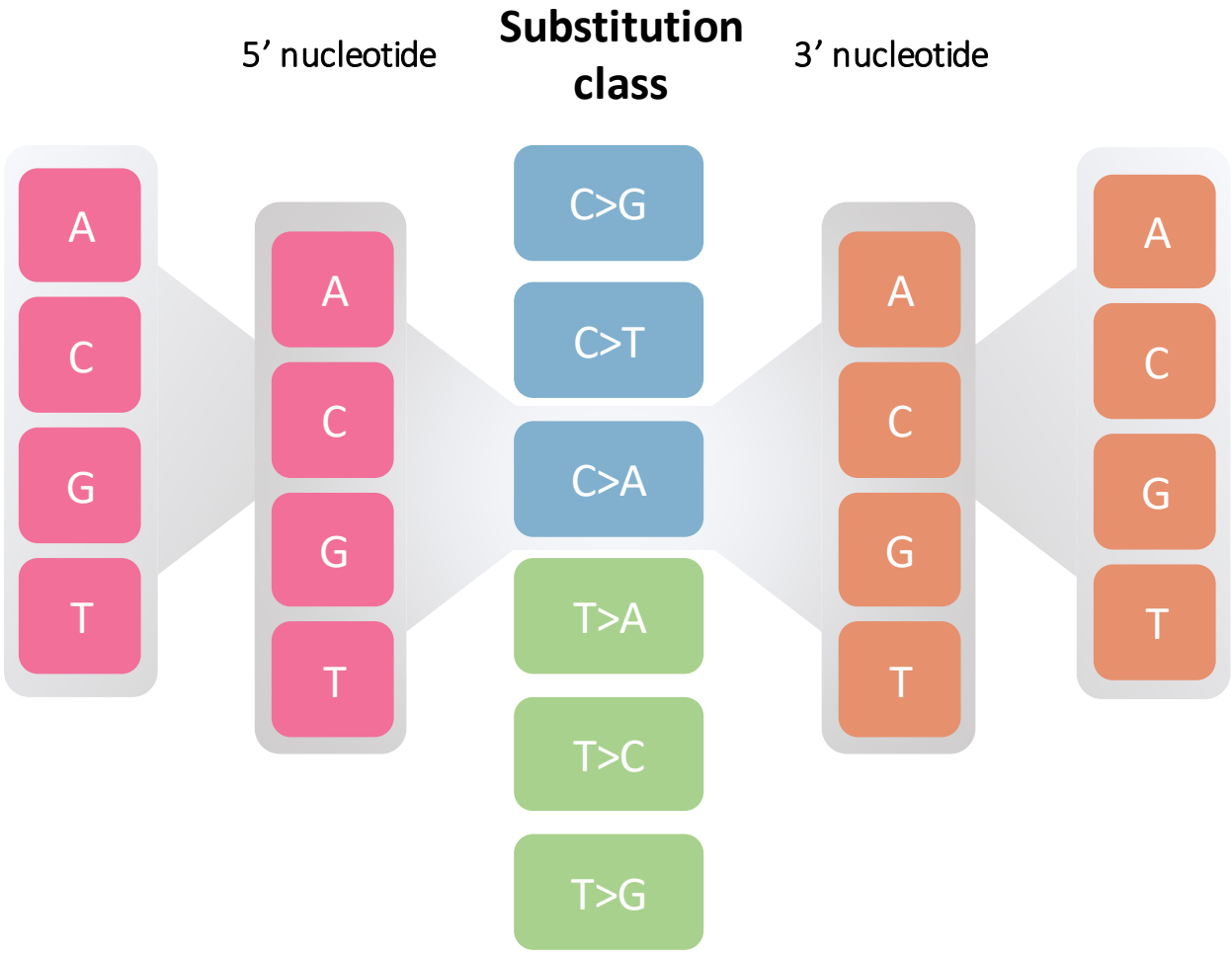




Example C>A 16 mutational channels

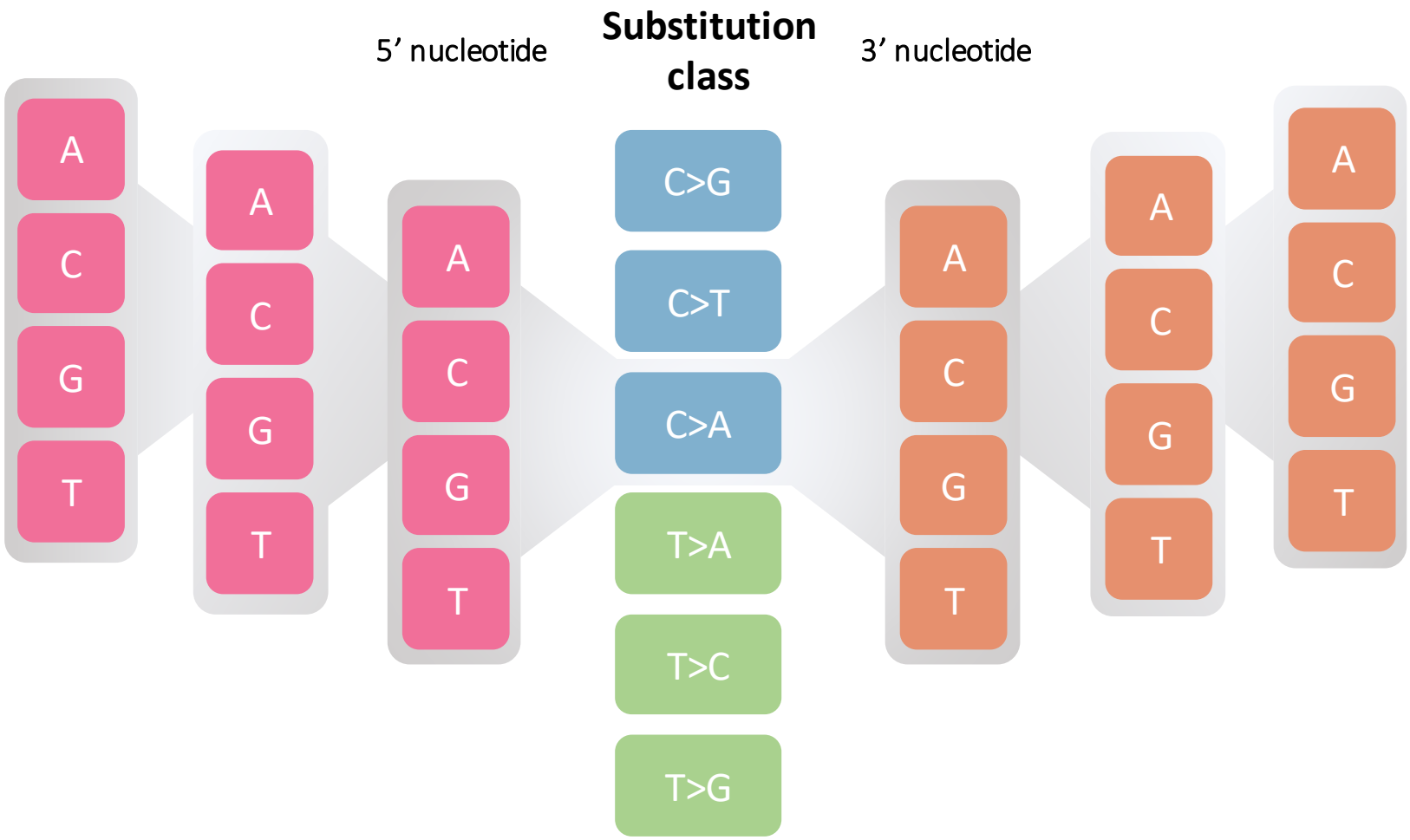
SBS - 1536

5-nucleotide context (± 2 bp context)



SBS - 24576

7-nucleotide context (± 3 bp context)

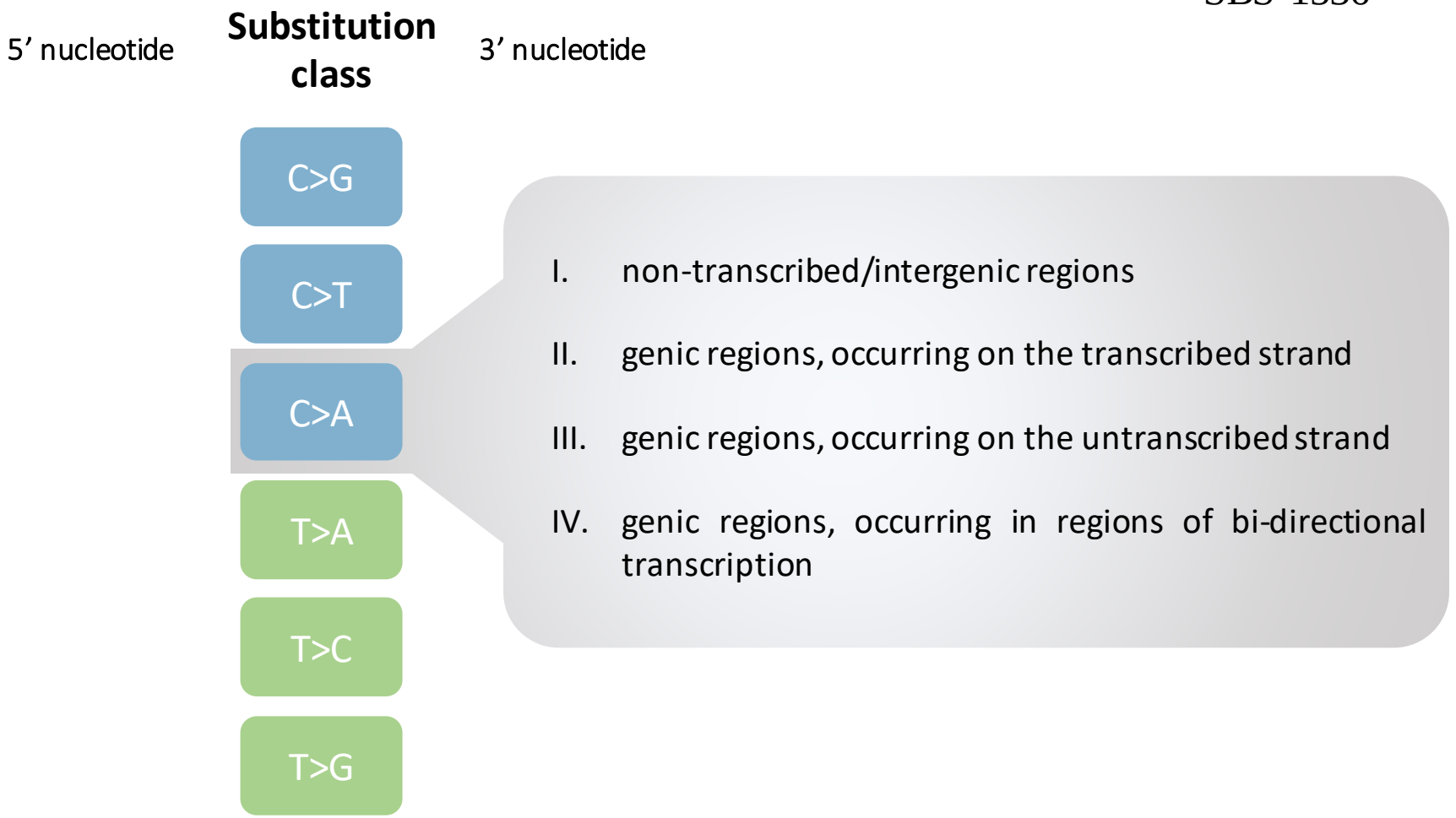


SBS - 24

extends SBS-6

SBS-96 → SBS-384

SBS-1536 → SBS-6144



SBS – 24 : transcriptional strand bias

1. Non-transcribed/Intergenic Regions

This category refers to mutations that occur in the vast stretches of the genome that are not genes or are on the non-transcribed strand of a gene. Most of our DNA is non-coding, and these regions have very little or no transcription activity.

- ***Significance**: Mutations in these regions are a good baseline for a mutational signature because they are less influenced by the complex processes of transcription and DNA repair that happen when a gene is being actively read.*

2. Genic Regions, Transcribed Strand

This category captures mutations that happen within the boundaries of a gene on the transcribed strand. This strand is actively being "read" by RNA polymerase to create messenger RNA (mRNA).

- ***Significance**: Because this strand is constantly being opened and processed, it's more exposed to certain mutagens. It also has a special DNA repair mechanism called transcription-coupled nucleotide excision repair (TC-NER). If a mutational process leaves a different pattern of mutations on the transcribed strand versus the non-transcribed strand, it suggests that the mutagen is interacting with the transcription machinery, or that TC-NER is either working or failing in a specific way.*

3. Genic Regions, Untranscribed Strand

These are mutations that occur within the boundaries of a gene on the untranscribed strand. This is the complementary DNA strand that isn't being directly read by the RNA polymerase.

- ***Significance**: The untranscribed strand is less exposed and is not a direct substrate for transcription-coupled repair. Therefore, comparing the mutation patterns between the transcribed and untranscribed strands can give you powerful clues about the specific DNA repair deficiencies at play. A difference in the number of mutations between these two strands is called transcriptional strand bias.*

4. Genic Regions, Bidirectional Transcription

This is a more specialized category for mutations that occur in a genomic region that is transcribed on both strands. This happens in certain parts of the genome where two genes are very close together and are transcribed in opposite directions, with their start sites near each other.

- ***Significance**: In these areas, both strands are subject to the processes and repair mechanisms of transcription. Mutations in these regions provide a unique insight into mutational processes that are influenced by transcription on both strands.*

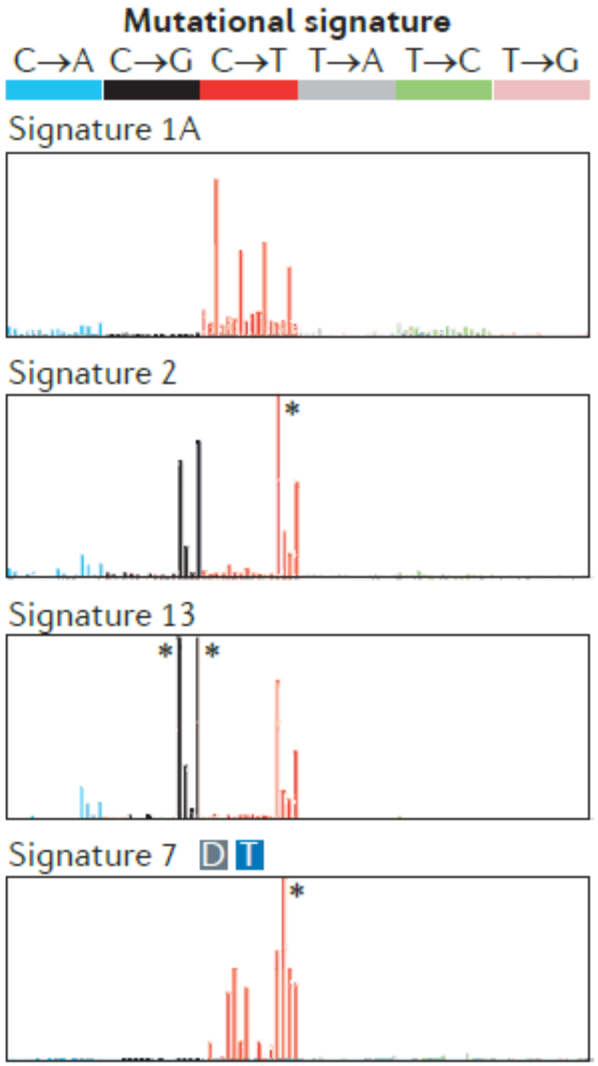
Mutational Processes and Signatures

- Examples of mutational processes (UV, tobacco, aging, APOBEC, mismatch repair deficiency, homologous recombination deficiency)
- COSMIC database (SBS, DBS, ID signatures)
- Clinical and biological interpretations

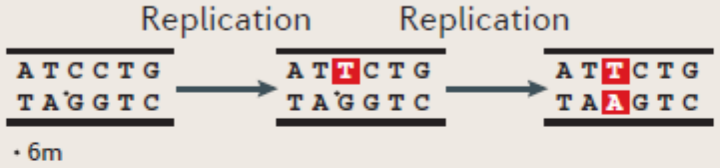
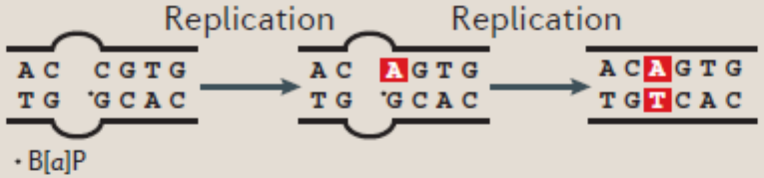
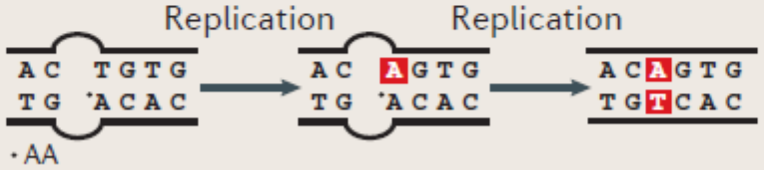
Examples of mutational processes

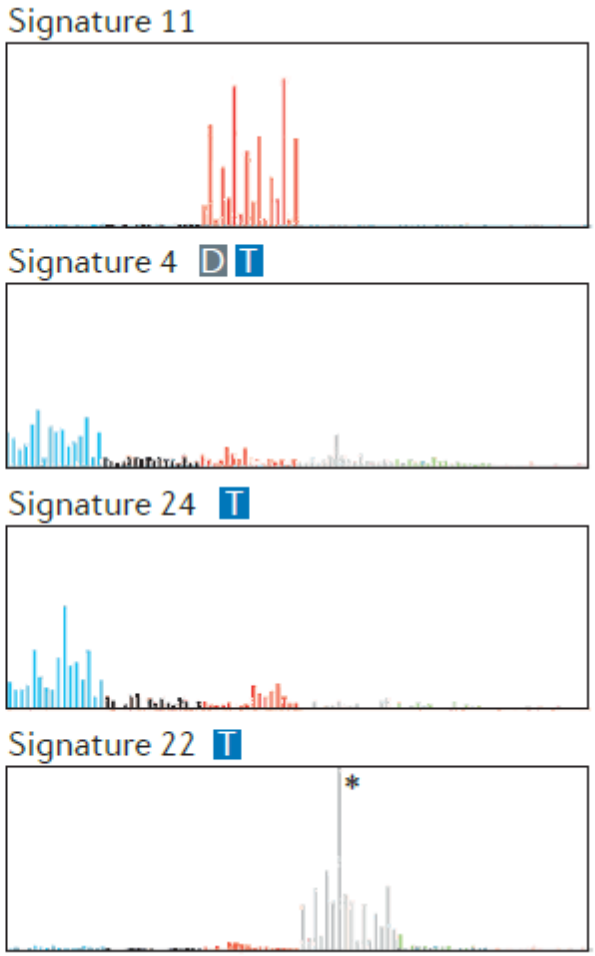
Mutational process

DNA damage	DNA repair pathway	Replicative process	Mutational signature
C:G→T:A at methylated CpGs correlates with age	Replicative polymerases	Deamination → Replication	Signature 1A
APOBEC editing	Base excision repair with 'A-rule' or with excess of REV1 activity?	Deamination → Replication (A-rule / REV1?)	Signature 2
			Signature 13
UV radiation on pyridimines and dipyrimidines	Transcription-coupled repair	Replication → Replication	Signature 7 D T



Helleday et al, Nat Rev Genet 2014

Temozolomide-induced O⁶-methyl-guanine lesions	Direct repair using methylguanine DNA methyltransferase	
Benzo[*a*]pyrene (B[*a*]P) adducts on guanine	Transcription-coupled repair	
Aflatoxin adducts on guanine		
Aristolochic acid (AA) adducts on adenine	Transcription-coupled repair	



Helleday et al, Nat Rev Genet 2014

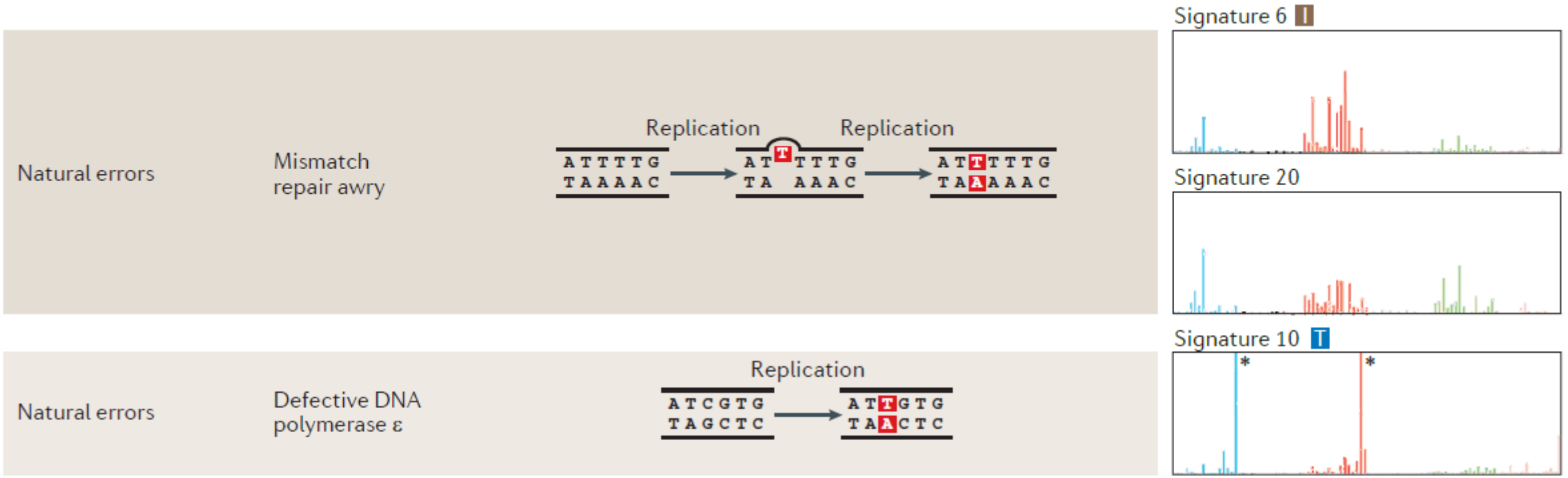


Figure 2 | **Summary of known mutational signatures, and the components of DNA damage and repair that constitute the mutational processes.** There are marked differences among the 96-element mutational signatures, which are dominated by specific elements, including enrichment of various base substitutions (shown in the graphs on the right), transcriptional strand bias (T), excess of dinucleotide mutations (D), and association with insertions and deletions (I). The asterisks mark instances at which the limits of the *y* axes, which represent the likelihood of specific mutations being present in a signature, are exceeded. 5m, 5′ methyl group; 6m, O^6 methyl group; APOBEC, apolipoprotein B mRNA editing enzyme, catalytic polypeptide; REV1, DNA repair protein REV1; UV, ultraviolet.

Helleday et al, Nat Rev Genet 2014

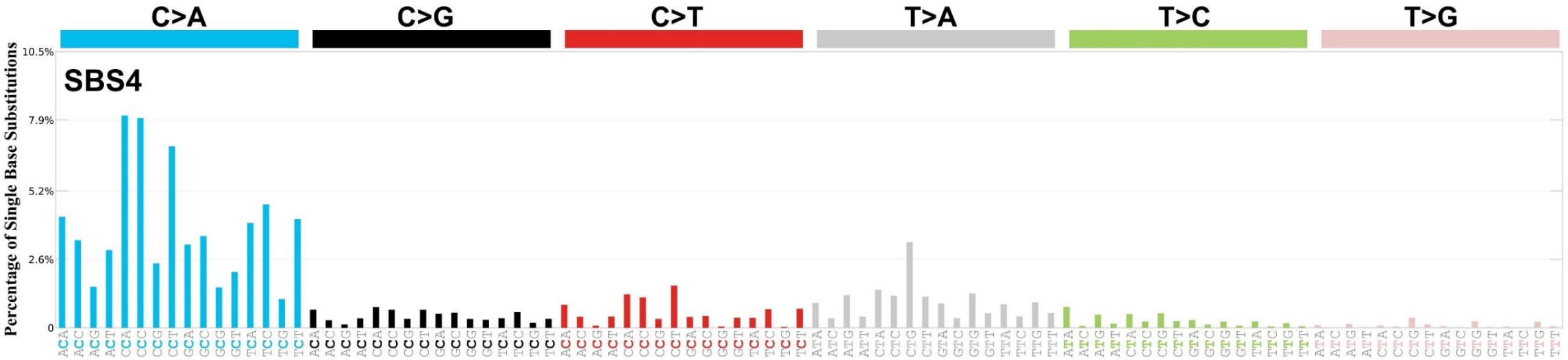
COSMIC Mutational Catalog (The Reference Library)

- The COSMIC (Catalogue Of Somatic Mutations In Cancer) database is the most widely used public repository of mutational signatures. It's a gold-standard reference, meticulously curated by a team of experts based on the analysis of thousands of cancer
- It as the encyclopedia of known mutational causes. It contains signatures with well-established aetiologies (causes), such as SBS7 from UV light and SBS4 from tobacco smoke. When you fit your tumor data to the COSMIC catalog, you're asking, "Does my tumor's mutation pattern match any of these known causes?
- It as a pre-defined, non-negative matrix factorization (NMF) basis matrix. The goal is to solve the equation $V=W\times H$, where V is your mutation matrix, W is the fixed COSMIC signature matrix, and you're solving for H (the contribution of each signature to your samples). This is a supervised learning approach.

Key advantage: High confidence. You're attributing your mutations to a known, validated cause.

Single Base Substitution (SBS) Signatures

SBS 4 : tobacco smoking signature

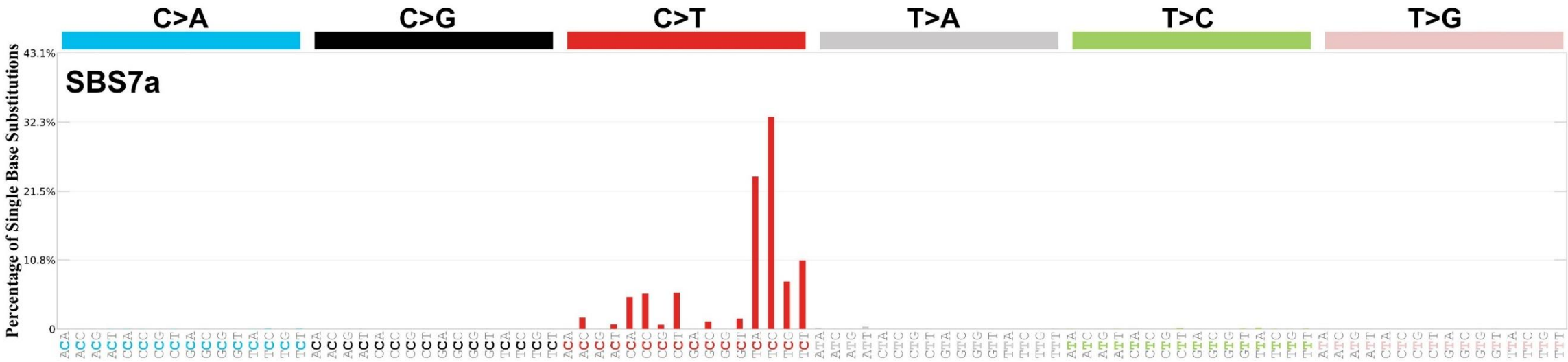


Proposed aetiology

Associated with **tobacco smoking**. Its profile is similar to the mutational spectrum observed in experimental systems exposed to tobacco carcinogens such as benzo[a]pyrene. SBS4 is, therefore, likely due to direct DNA damage by tobacco smoke mutagens.

https://cancer.sanger.ac.uk/cosmic

SBS 7a : UV light exposure signature

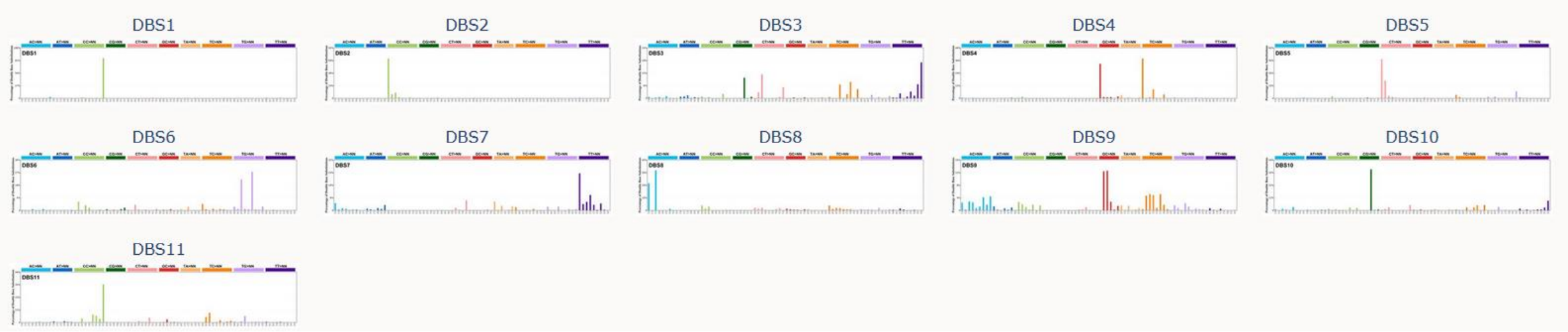


Proposed aetiology

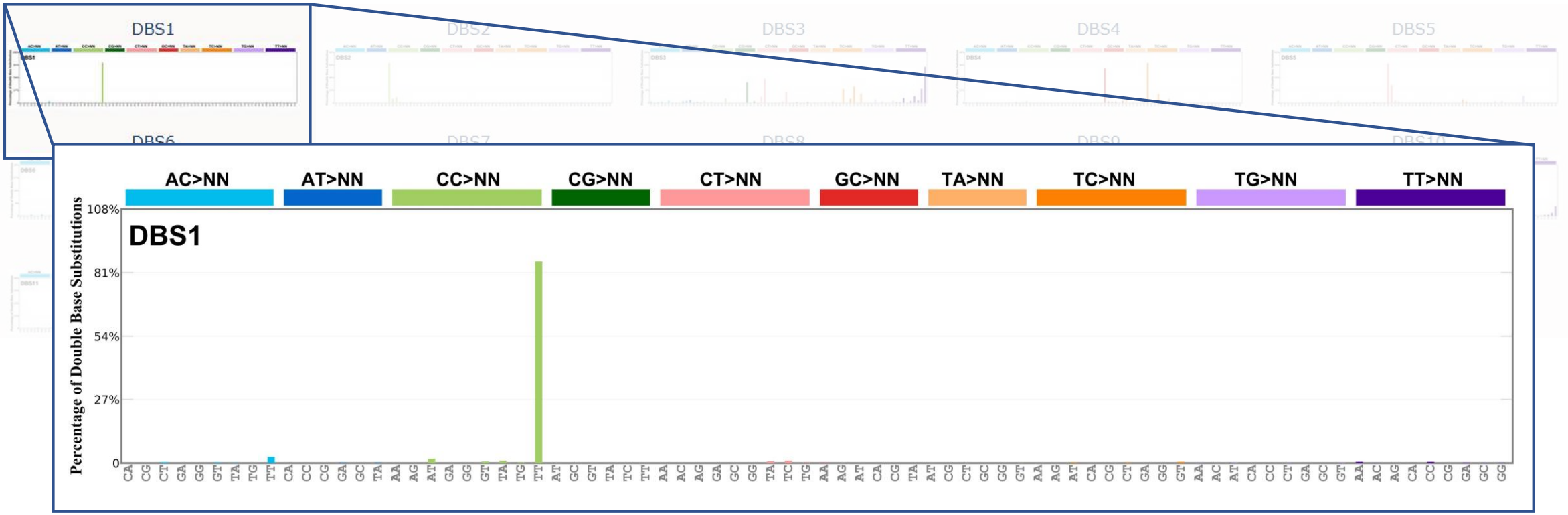
SBS7a/SBS7b/SBS7c/SBS7d are found in cancers of the skin from sun exposed areas and are thus likely to be due to exposure to **ultraviolet light**. SBS7a may possibly be the consequence of just one of the two major known UV photoproducts, cyclobutane pyrimidine dimers or 6-4 photoproducts. However, there is currently no evidence for this hypothesis and it is unclear which of these photoproducts may be responsible for SBS7a.

https://cancer.sanger.ac.uk/cosmic

Doublet Base Substitution (DBS) Signatures



Doublet Base Substitution (DBS) Signatures



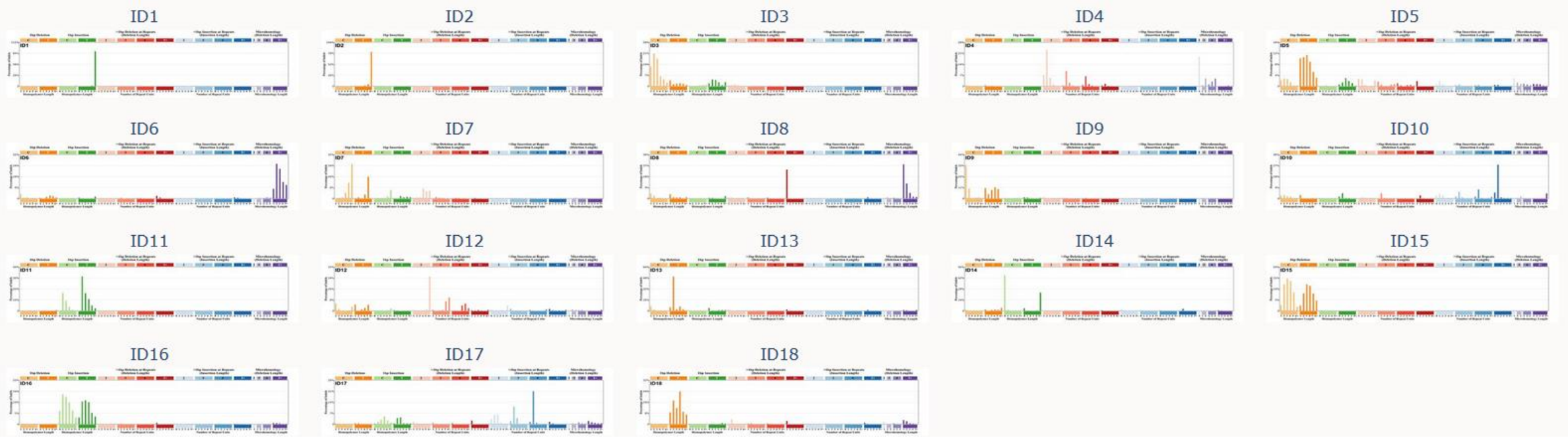
Proposed aetiology

Exposure to ultraviolet light.

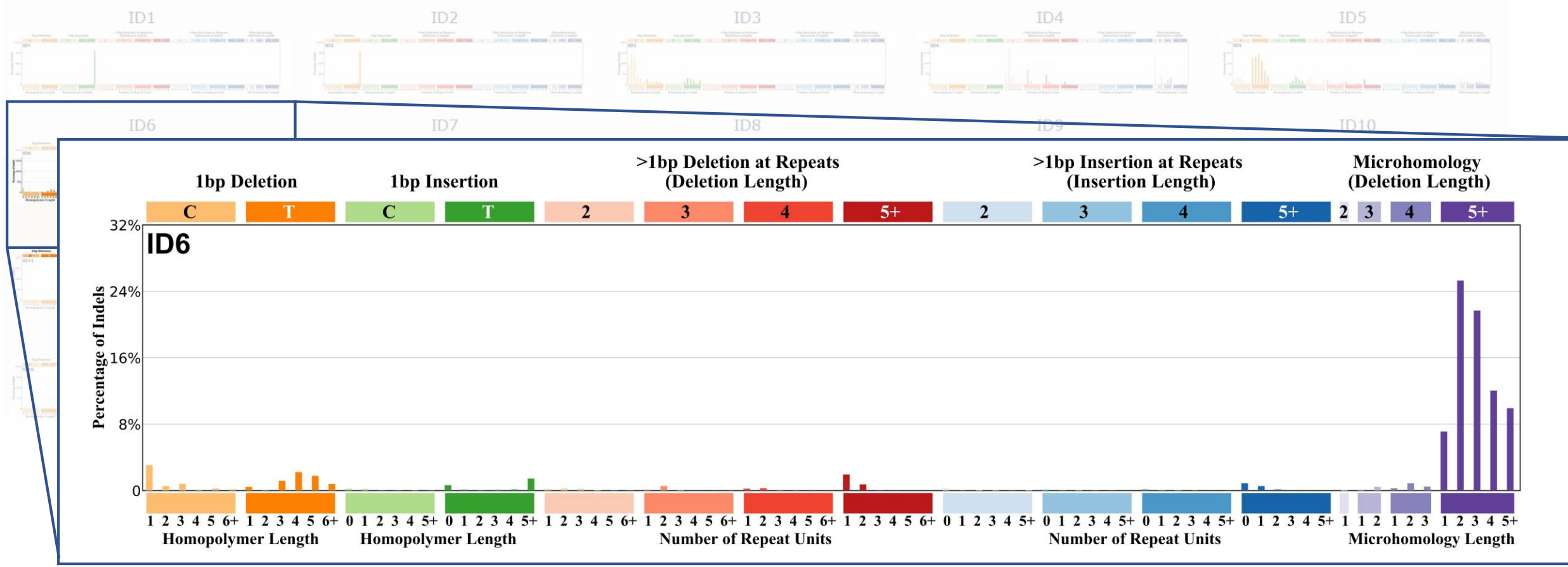
Comments

DBS1 exhibits transcriptional strand bias with more CC>TT mutations than GG>AA on the untranscribed strands of genes indicative of damage to cytosine and repair by transcription coupled nucleotide excision repair.

Small insertions and deletions (ID) Signatures



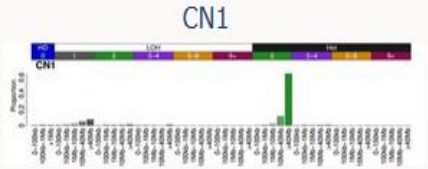
Small insertions and deletions (ID) Signatures

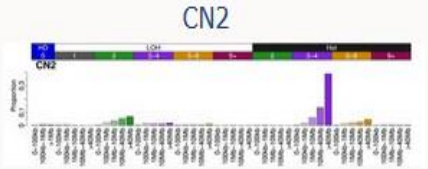


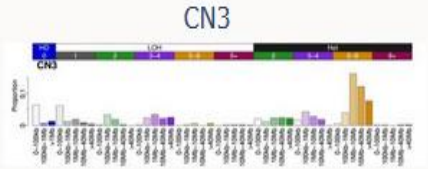
Proposed aetiology

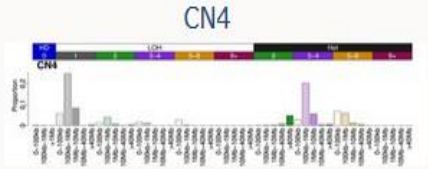
Defective homologous recombination-based DNA damage repair, often due to inactivating BRCA1 or BRCA2 mutations, leading to non-homologous DNA end-joining activity.

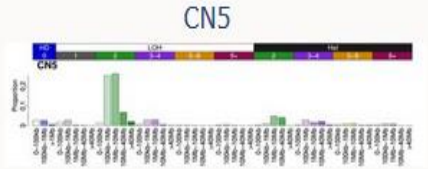
Copy Number Variations (CN) Signatures

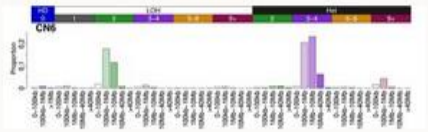


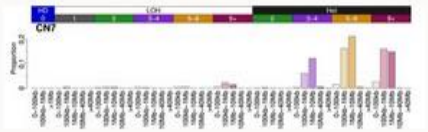


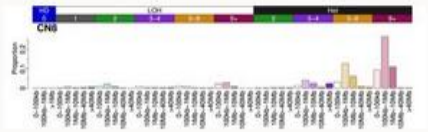


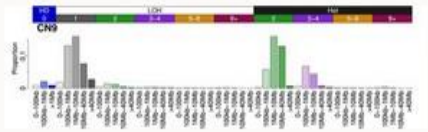


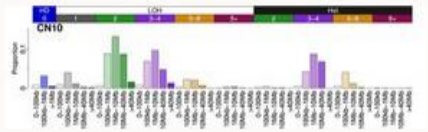


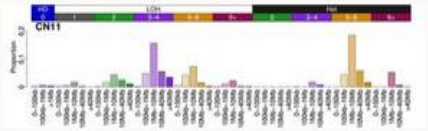


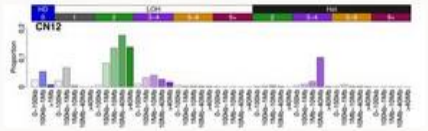


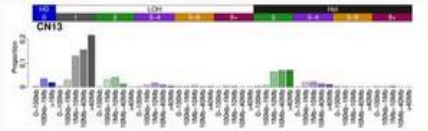


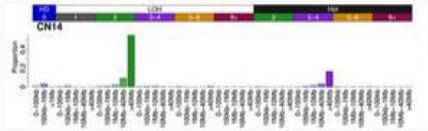


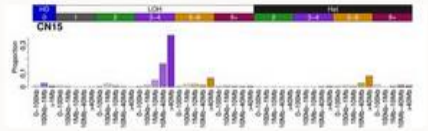


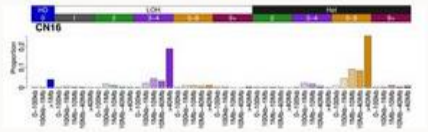


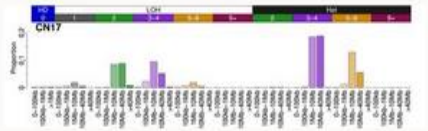


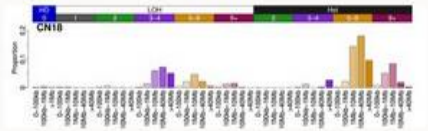


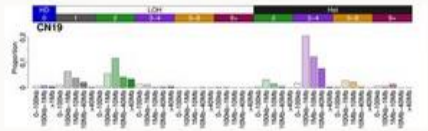


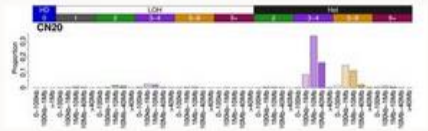


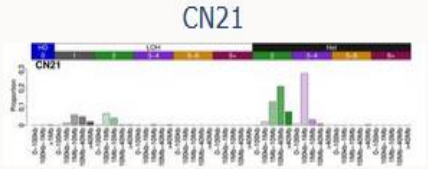




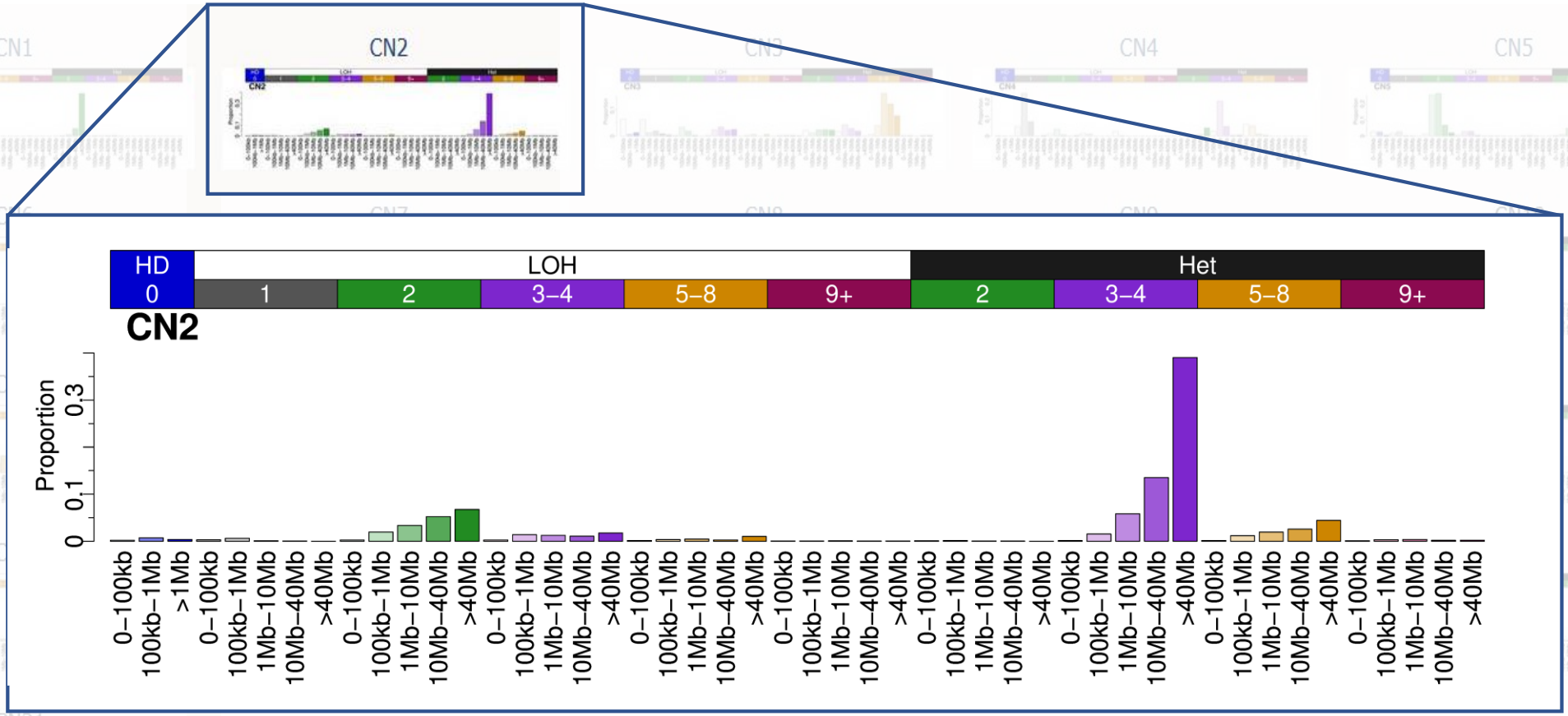








Copy Number Variations (CN) Signatures



CN2 is primarily a signature of a tetraploid genome; this has been verified through associations with once-genome-doubled samples, simulations of genome doubling, and artificial genome doubling of signature CN1. Note that due to the copy number class that most strongly defines this signature (TCN 3-4), a triploid cancer may be assigned this signature.
However, if CN2 is observed at <<100% proportion of segments, it may instead indicate aneuploidy, with gained segments (on a diploid background) being assigned CN2, and unaltered segments being assigned CN1 or lost segments being assigned CN9 or CN13, depending on size and genomic background.

Notes

CN2 is associated with a poor disease specific survival in thymomas.

De Novo Catalog (Building Your Own Library)

- A de novo catalog is a set of mutational signatures that you extract directly from your own dataset without any prior knowledge. You're creating new "fingerprints" based on the patterns present in your specific cohort of samples.
- Pure discovery. You're not looking for a match to a known pattern; you're letting the data speak for itself to find new, previously uncharacterized mutational processes that might be unique to your study population or cancer type.
- This as an unsupervised learning approach. You're performing NMF on your mutation matrix (V) to simultaneously solve for both the signature matrix (W) and the contribution matrix (H). The algorithm finds the optimal number of signatures and their patterns based on the data provided.

Key advantage: Discovery. This method can identify novel signatures that might not yet be in the COSMIC database.

When to Use Each Approach

- Start with COSMIC: Always begin your analysis by fitting your data to the COSMIC catalog. It's the standard first step and will tell you if your tumors are driven by known, common mutational processes.
- Use De Novo for Discovery: If your data doesn't fit the COSMIC signatures well, or if you're working with a new or rare cancer type, a de novo extraction is appropriate. The signatures you discover can then be compared to the COSMIC catalog to see if they are similar to any known signatures or if they are truly novel.

Biological meaning of signatures

Each signature reflects a mutagenic process or a DNA repair defect

- Exogenous exposures
  - **SBS4**: tobacco smoking → lung, head & neck cancers.
  - **SBS7**: UV light → skin cancers (C>T at dipyrimidines).
  - **SBS11**: alkylating agents → chemotherapy-induced mutations.
- Endogenous processes
  - **SBS1**: spontaneous deamination of 5-methylcytosine → age-related.
  - **SBS2/SBS13**: APOBEC enzyme activity → kataegis patterns in breast, bladder.
  - **SBS3**: defective homologous recombination (HR), linked to BRCA1/2 mutations.
  - **SBS6, 15, 20, 26**: mismatch repair (MMR) deficiency, microsatellite instability (MSI+).

Clinical applications

- **Cancer etiology**
  - Link cancer type to environmental exposure (e.g. SBS4 in lung → smoking history).
  - Distinguish sporadic vs hereditary predisposition (e.g. SBS3 → hereditary BRCA deficiency).
- **Diagnosis & classification**
  - Signatures help define molecular subtypes (e.g. MSI-high colorectal cancers have SBS6/15).
  - Can confirm tumor type when histology is ambiguous.
- **Prognosis**
  - Certain signatures correlate with better/worse outcomes.
  - Example: HR-deficiency (SBS3) predicts sensitivity to platinum therapy.
- **Therapy guidance (Precision Oncology)**
  - BRCA/HRD signatures (SBS3): PARP inhibitors (Olaparib, Niraparib).
  - MMR-deficiency (SBS6/15): immune checkpoint inhibitors (anti-PD1).
  - Chemotherapy exposure (SBS11, SBS31): signatures of past treatment can be identified.

Methods for Extraction and Attribution

- **Brief intro to tools:**
  - SigProfiler
- **Matrix decomposition methods:**
  - Non-negative matrix factorization (NMF).
- **Attribution methods:**
  - Fitting known COSMIC signatures to new samples.
- **Practical considerations:**
  - Sequencing depth, tumor purity, mutational burden.
  - Noise and overfitting.

Brief intro to tools

Table 1. Overview of bioinformatics tools for *de novo* extraction of mutational signatures

Tool name	Input	Platform	Factorization method	Factorization engine	GPU	Manual selection	Automatic selection	Automatic algorithm	Mutational catalog support	Plotting support	COSMIC comparison
EMu[20]	matrix	C++	EM	original implementation[20]	no	yes	yes[a]	BIC[21]	SBS-96	no	no
Maftools[22]	matrix, MAF	R-Bioconductor	NMF	NMF R package[23]	no	yes	no	–	SBS-96	SBS-96	1 to 1
Mutational Patterns[24]	matrix, VCF	R-Bioconductor	NMF	NMF R package[23]	no	yes	no	–	SBS-96, SBS-192	SBS-96, SBS-192	1 to 1
MutSignatures[25]	matrix, VCF, MAF	R	NMF	Brunet et al.[26]	no	no	no	–	SBS-96	SBS-96	1 to 1
MutSpec[27]	matrix, VCF, custom	Galaxy, Perl, R	NMF	NMF R package[23]	no	yes	no	–	SBS-96, SBS-192	SBS-96, SBS-192	1 to 1
SigFit[28]	matrix	R	Bayesian inference	Stan R package[29]	no	yes	yes[a]	Elbow method[30]	SBS-96	SBS-96, SBS-192	1 to 1

Islam et al., 2022, Cell Genomics 2, 100179

Brief intro to tools

Table 1. Overview of bioinformatics tools for *de novo* extraction of mutational signatures

Tool name	Input	Platform	Factorization method	Factorization engine	GPU	Manual selection	Automatic selection	Automatic algorithm	Mutational catalog support	Plotting support	COSMIC comparison
SigMiner[31]	matrix, MAF	R	(automatic) Bayesian NMF, (manual) NMF	(automatic) Signature Analyzer implementation,[32] (manual) NMF R package[23]	no	yes[a]	yes	ARD[33]	SBS-96, DBS-78, ID-83	generic	1 to 1
Signature Analyzer[32,34]	matrix, MAF	R (CPU),[18] Python (GPU)[19]	Bayesian NMF	original implementation[32,34]	yes	no	yes	ARD[33]	SBS-96, DBS-78, ID-83	SBS-96, DBS-78, ID-83	1 to 1
Signature ToolsLib[35]	matrix, VCF, custom	R	NMF	NMF R package[23]	no	yes	no	–	SBS-96, DBS-78, ID-83, SV-32	SBS-96, SV-32, generic	1 to 2
SigneR[36]	matrix, VCF	R-Bioconductor, C++	Bayesian NMF	original implementation[36]	no	yes	yes[a]	BIC[21]	SBS-96	SBS-96	no
SigProfiler Extractor	matrix, VCF, MAF, custom	Python, R wrapper	NMF	(current work) original implementation	yes	yes	yes[a]	NMFk[37]	SBS-96, DBS-78, ID-83, CN-48, others,[15] any	SBS-96, DBS-78, ID-83, CN-48, SV-32, others,[15] generic	1 to many

Islam et al., 2022, Cell Genomics 2, 100179

Brief intro to tools

Table 1. *Continued*

Tool name	Input	Platform	Factorization method	Factorization engine	GPU	Manual selection	Automatic selection	Automatic algorithm	Mutational catalog support	Plotting support	COSMIC comparison
SigProfiler_ PCAWG[12]	matrix, VCF, MAF, custom	Python, MATLAB	NMF	Brunet et al.[26]	no	yes	no	–	SBS-96, DBS-78, ID-83, others,[15] any	SBS-96, DBS-78, ID-83	no
Somatic Signatures[38]	matrix, VCF	R-Bioconductor	NMF, PCA	NMF R package[23] pcaMethods R package[39]	no	yes	no	–	SBS-96	SBS-96	no
Tensor Signatures[40]	VCF	Python	NTF	TensorFlow[41]	yes	yes	yes[a]	BIC[21]	tensor	SBS-96 with strand bias	no

Tools are ordered alphabetically. 1 to 1 refers to one *de novo* signature being matched with exactly one COSMIC signature; 1 to 2 refers to one *de novo* signature being matched with a combination of up to two COSMIC signatures; 1 to many refers to one *de novo* signature being matched with a combination of one or more COSMIC signatures. MAF, mutation annotation format; VCF, variant call format; EM, expectation maximization algorithm; NMF, nonnegative matrix factorization; PCA, principal component analysis; NTF, nonnegative tensor factorization; ARD, automatic relevance determination; BIC, Bayesian information criterion; COSMIC, catalog of somatic mutations in cancer; SBS, single base substitutions; DBS, doublet base substitutions; ID, small insertions and deletions; CN, copy number; SV, structural variants.

[a]The default approach for selecting the total number of signatures when a tool supports both manual and automatic selection.

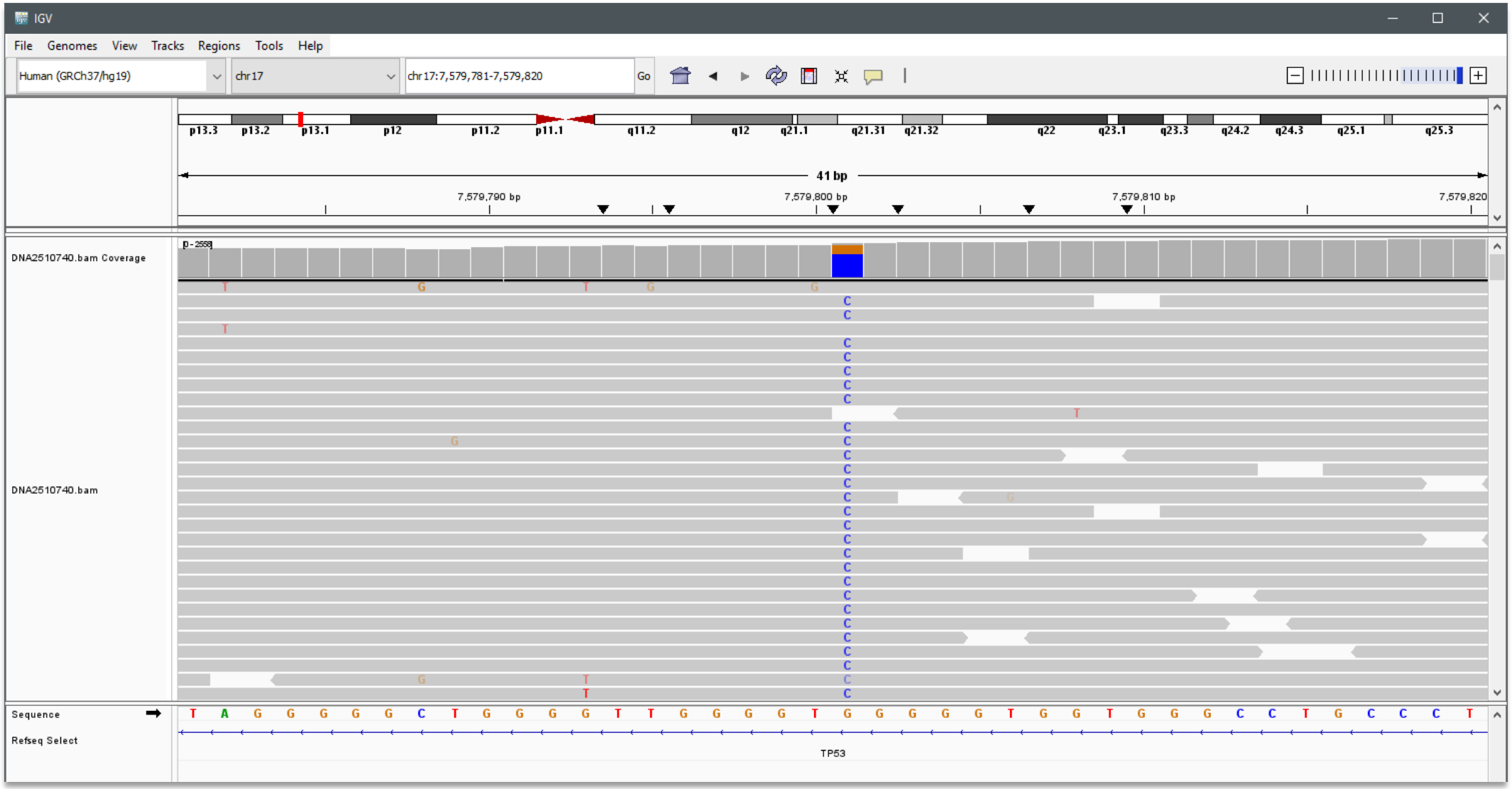
Islam et al., 2022, Cell Genomics 2, 100179

Somatic genomic mutations

Current gold standard workflow for analysis of whole genome sequencing data.

Supernat, A., Vidar
three variant caller
Rep 8, 17851 (2018
36177-7

Somatic genomic mutations

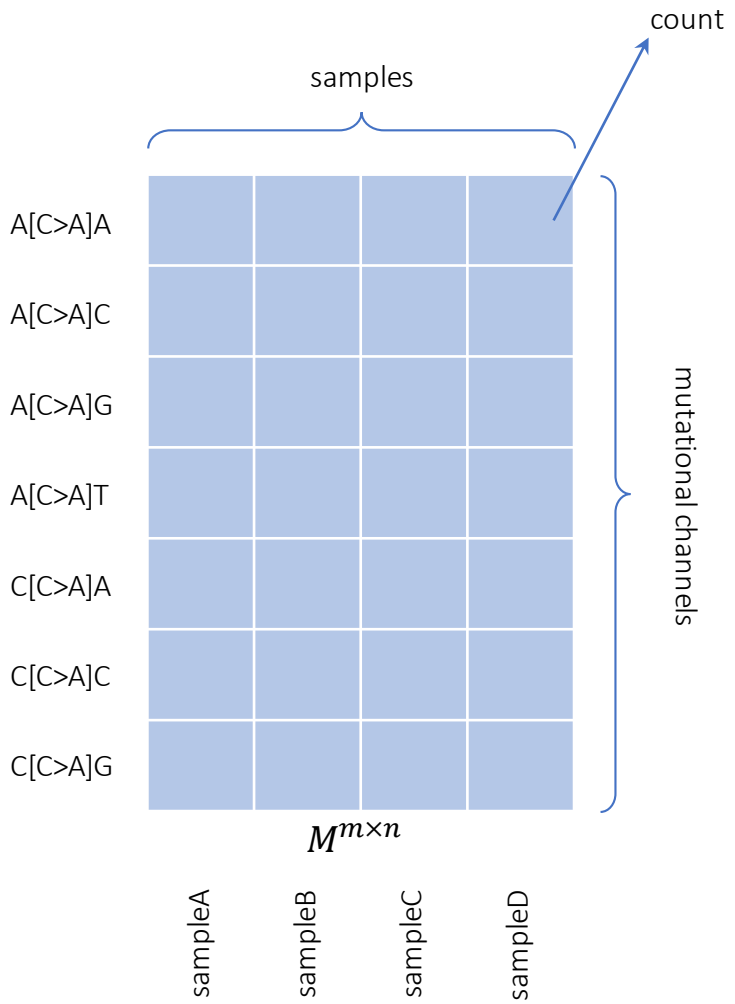


Transformation of somatic mutations into a matrix

somatic mutations in a cancer sample

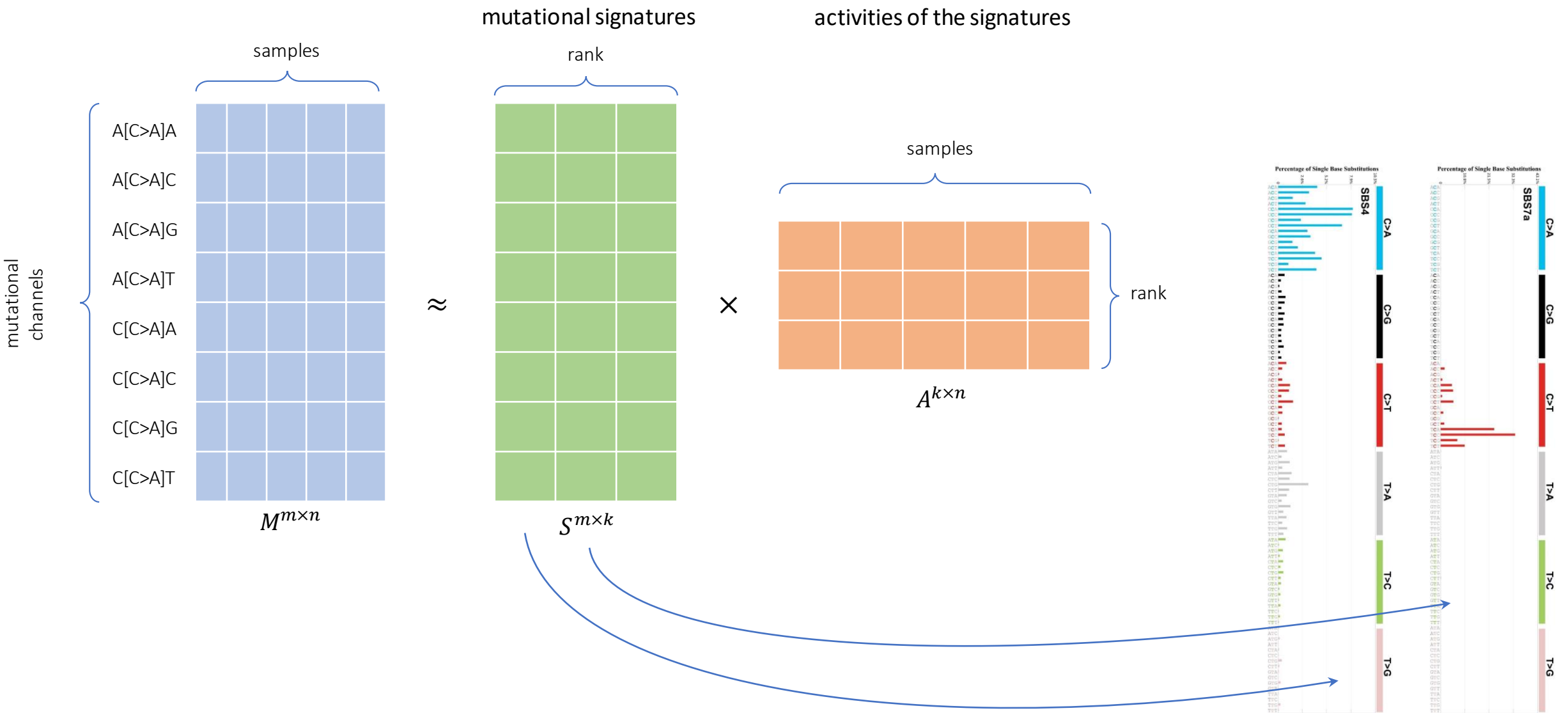
Chromosome	Position	RefCall	AltCall	VAF	Depth	CytoBand	GeneName				
chr1	2488153	A	G	0.4406	202	1p36.32	TNFRSF14	SNV	COSM4999217;CO		
chr1	4367323	G	A	0.9966	290	1p36.32		SNV	NA	0	NA	NA	0	14
chr1	8074334	C	T	0.5011	459	1p36.23	ERRFI1	SNV	COSM3689848	1	NA			
chr1	11169789	A	G	0.4684	316	1p36.22	MTOR	SNV	NA	0	NA	NA		
chr1	11181327	C	T	0.5641	234	1p36.22	MTOR	SNV	COSM3996743	3				
chr1	11190646	G	A	0.5659	334	1p36.22	MTOR	SNV	COSM3996746	2				
chr1	11205058	C	T	0.5804	367	1p36.22	MTOR	SNV	COSM4142146	3				
chr1	16256007	T	C	1	386	1p36.13	SPEN	SNV	COSM4142934	4	NA			
chr1	16259813	A	G	1	294	1p36.13	SPEN	SNV	COSM4142938	3	NA			
chr1	16260803	G	A	0.5018	277	1p36.13	SPEN	SNV		0	NA	NA		
chr1	18957546	G	A	0.543	221	1p36.13	PAX7	SNV	NA	0	NA	NA		
chr1	19018405	G	A	0.5121	207	1p36.13	PAX7	SNV	COSM3750707;CO					
chr1	19018432	A	C	0.4781	228	1p36.13	PAX7	SNV	COSM6351865;CO					
chr1	19027239	A	G	0.3538	130	1p36.13	PAX7	SNV	COSM424864;COS					
chr1	19071752	T	C	1	251	1p36.13	PAX7	SNV	NA	0	NA	NA	0	
chr1	19072926	A	G	1	347	1p36.13	PAX7	SNV		0	NA	NA	0	
chr1	19075225	A	G	1	343	1p36.13	PAX7	SNV	NA	0	NA	NA	0	
chr1	23885498	T	C	0.9973	371	1p36.12	ID3	SNV	NA	0	NA	NA	20	
chr1	40356155	A	G	0.4735	359	1p34.2		SNV	NA	0	NA	NA	0	
chr1	40363054	G	C	1	238	1p34.2	MYCL	SNV	COSM3927628	1	NA			
chr1	40364803	C	A	0.4777	224	1p34.2	MYCL	SNV	NA	0	NA	NA		
chr1	40366222	C	A	0.5227	308	1p34.2	MYCL	SNV	NA	0	NA	NA		
chr1	45797505	C	G	0.4784	255	1p34.1	MUTYH	SNV	COSM3751252	21				
chr1	45805566	G	C	1	94	1p34.1	MUTYH;TOE1	SNV	NA	0	NA	NA		

sampleA



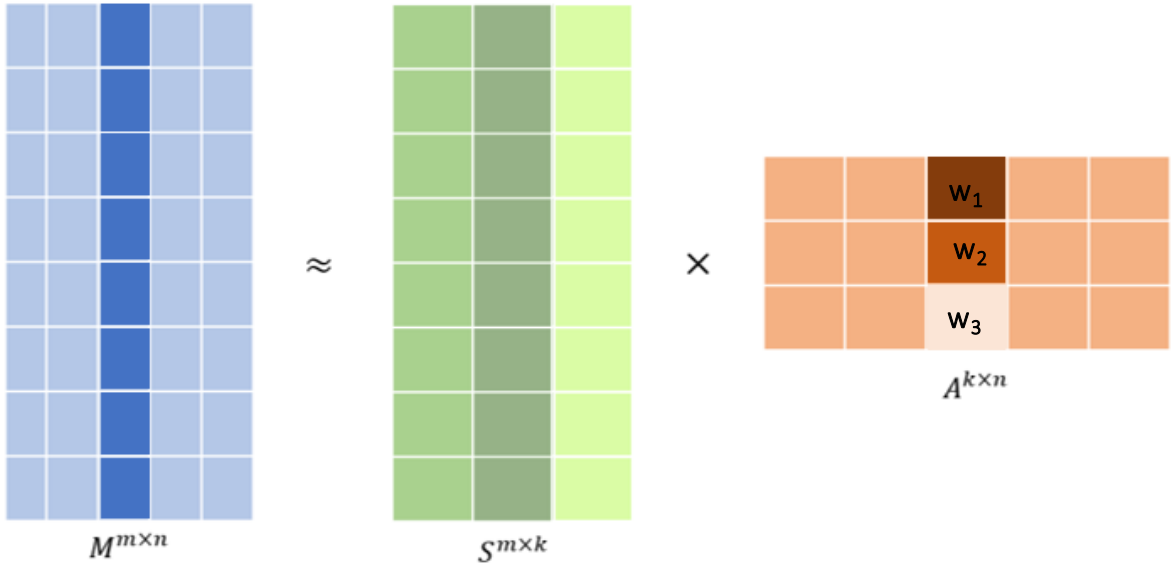
The value of each cell in the matrix, M, corresponds to the number of somatic mutations from a particular mutational channel in each sample

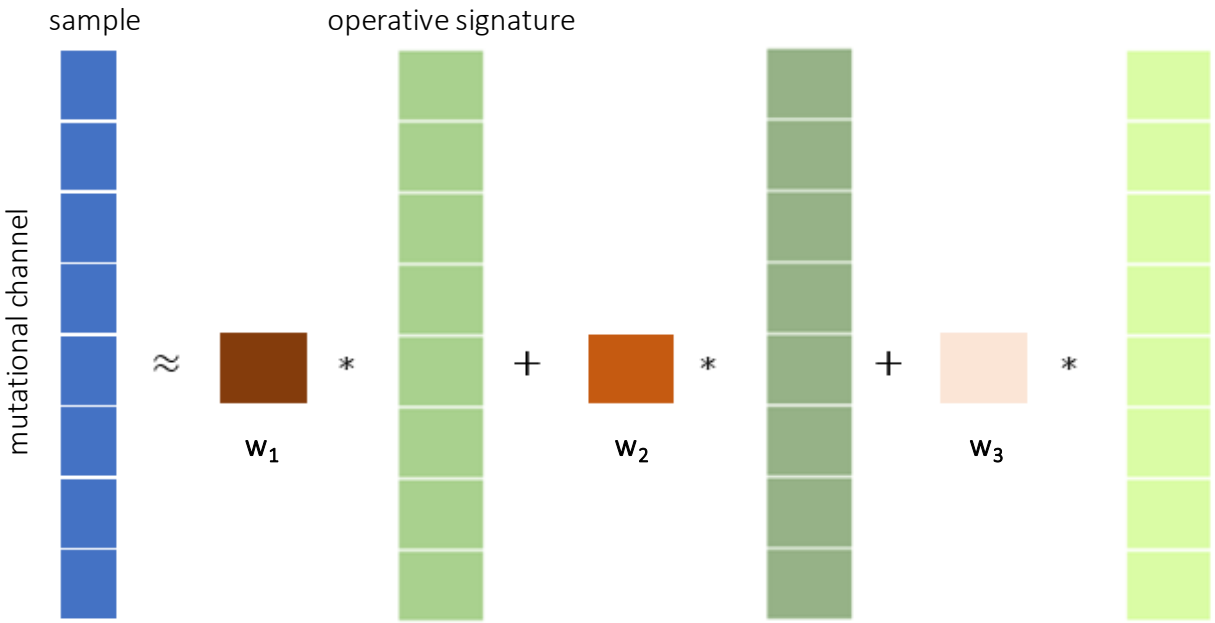
Nonnegative Matrix Factorization (NMF)



Note: the *S* and *A* matrices are not unique!

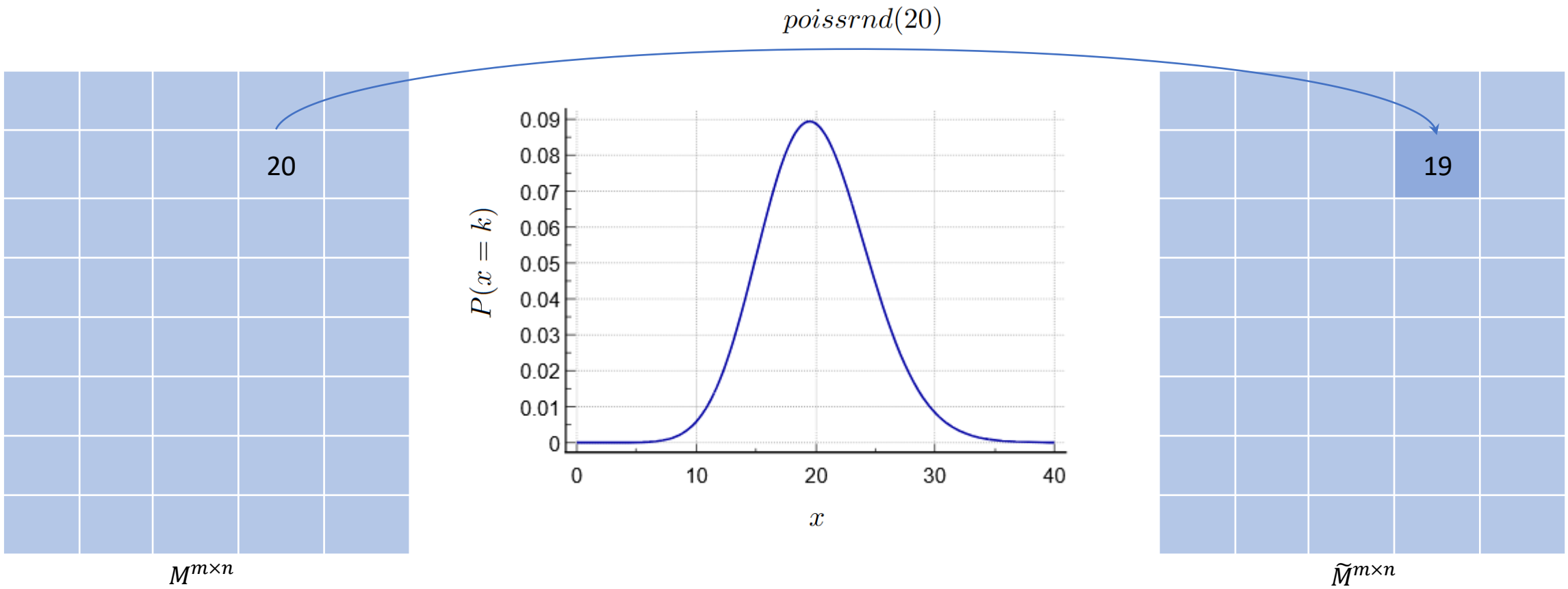
Nonnegative Matrix Factorization (NMF)





Resampling of the input mutational matrix

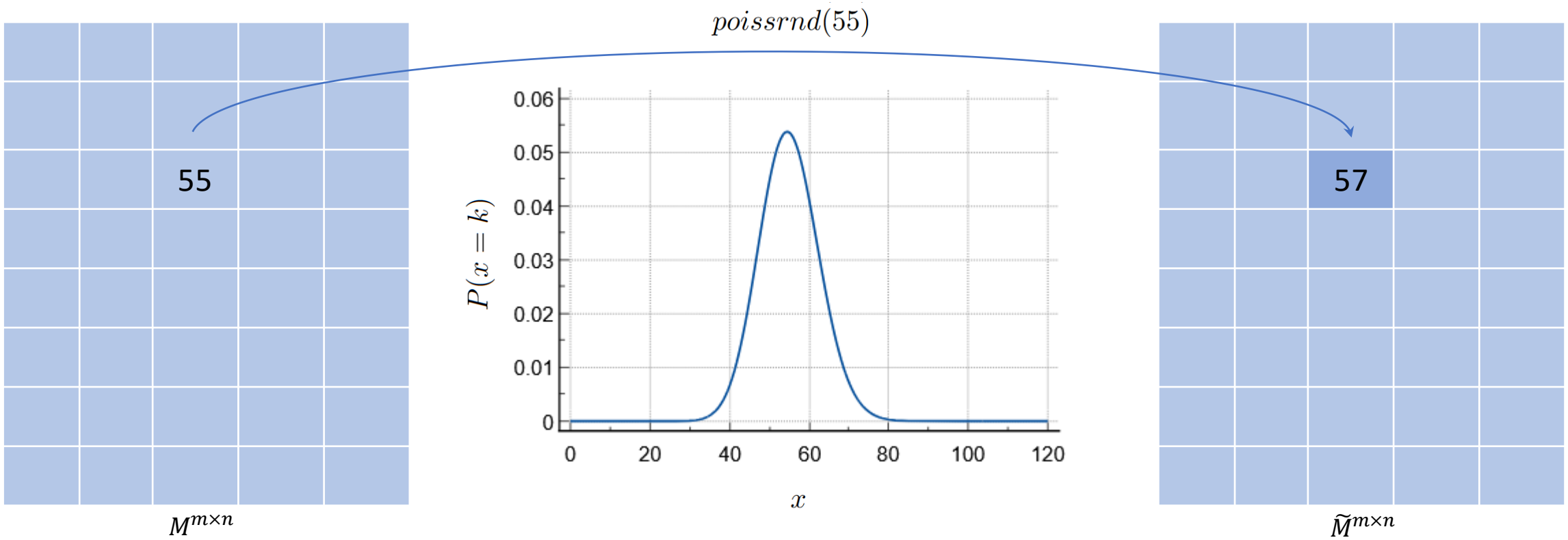
independent Poisson resampling of the original matrix for each replicate



The resampling is performed to ensure that Poisson fluctuations of the matrix do not impact the stability of the factorization results

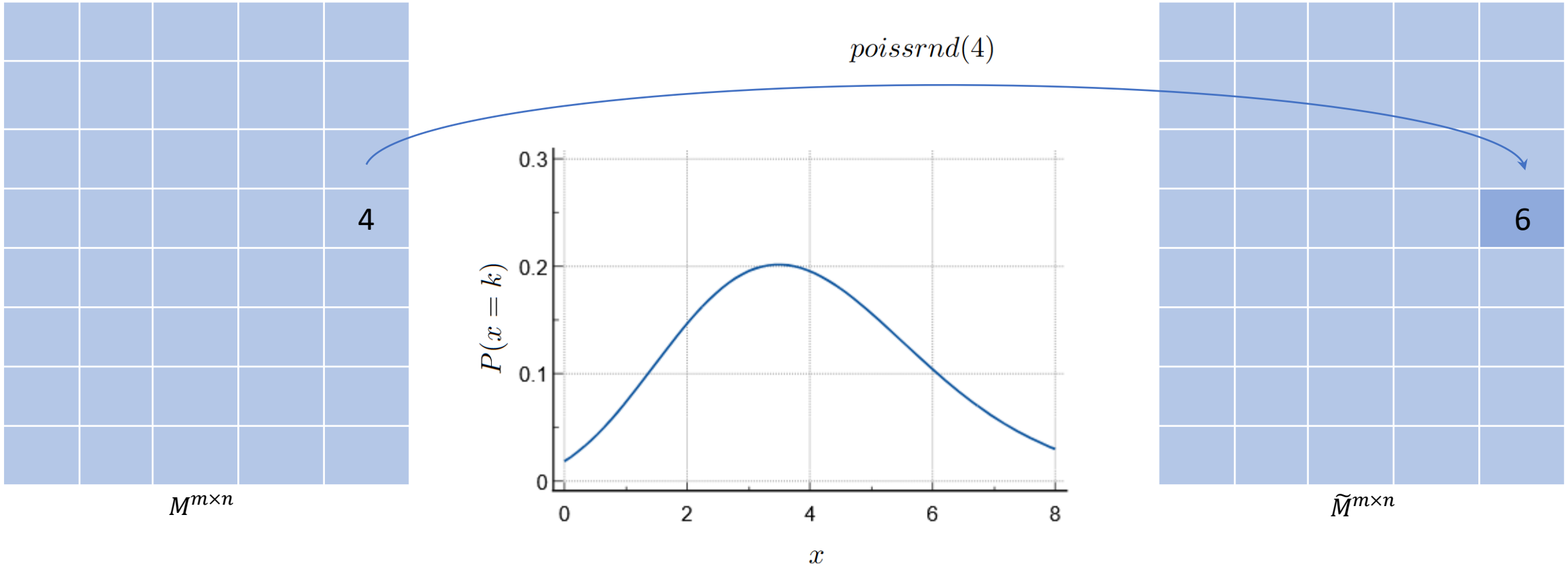
Resampling of the input mutational matrix

independent Poisson resampling of the original matrix for each replicate



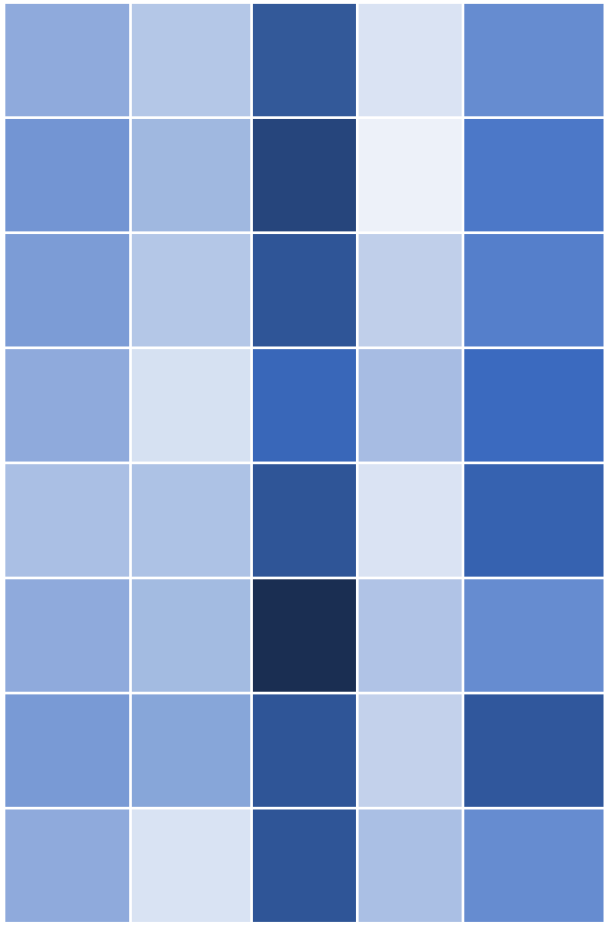
Resampling of the input mutational matrix

independent Poisson resampling of the original matrix for each replicate



Normalizing the resampled matrix

normalization to overcome potential skewing



$\widetilde{M}^{m \times n}$

- **Log2 normalization**

$$M(i,j) = M(i,j) \times \frac{\log_2 \left(\sum_i M(i,j)\right)}{\sum_i M(i,j)}$$

- **100x normalization**

$$\text{if } \sum_i M(i,j) > 100 \times m$$

$$M(i,j) = \frac{M(i,j) \times 100 \times m}{\sum_i M(i,j)}$$

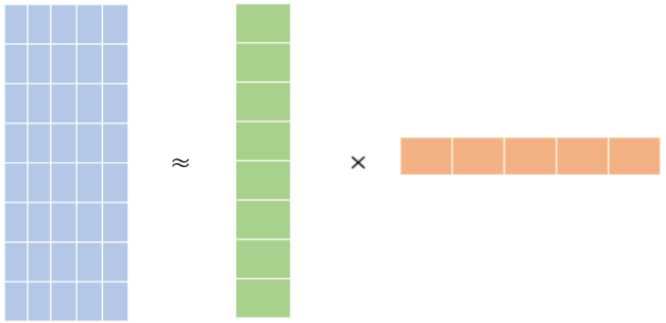
- **Gaussian mixture model (GMM) normalization**

- **No normalization**

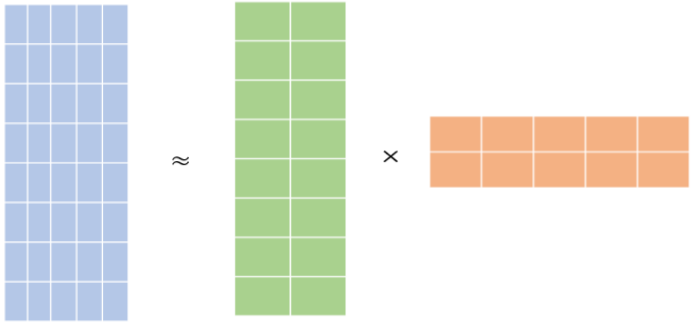
Module 3: Non-Negative Matrix factorization with replicates

rank

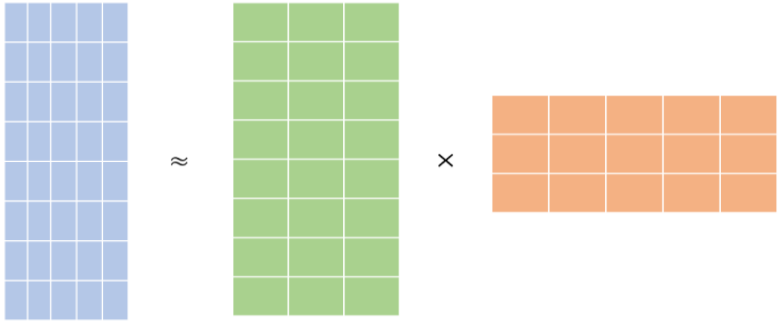
k = 1



k = 2



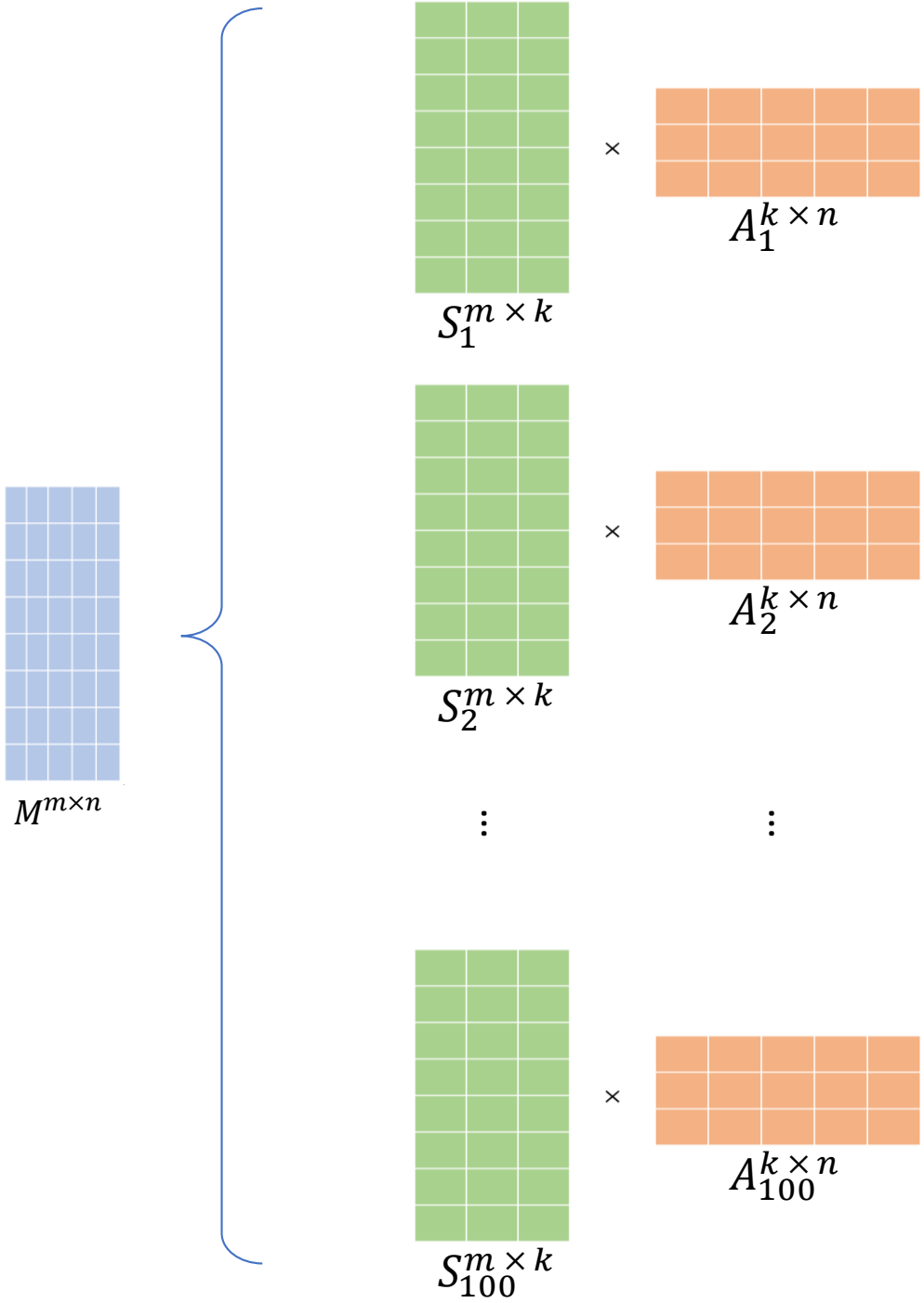
k = 3



...

k = 25 ...

factorizes the matrix *M* with different ranks searching for an optimal solution between k = 1 and k = 25 mutational signatures (number of operative signatures)

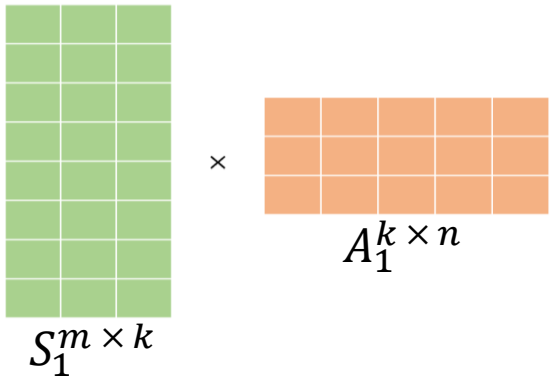


NMF

Multiplicative update algorithm

- Kullback-Leibler divergence
- Euclidean distance
- Itakura-Saito divergence

Itakura-Saito divergence



The Itakura-Saito cost function:

$$D_{\mathrm{IS}}(\mathbf{X} \mid \mathbf{WH}) = \sum_{i,j} \left(\frac{\mathbf{X}_{ij}}{(\mathbf{WH})_{ij}} - \log \frac{\mathbf{X}_{ij}}{(\mathbf{WH})_{ij}} - 1 \right)$$

$M^{m \times n}$ $S^{m \times k}$ $A^{k \times n}$

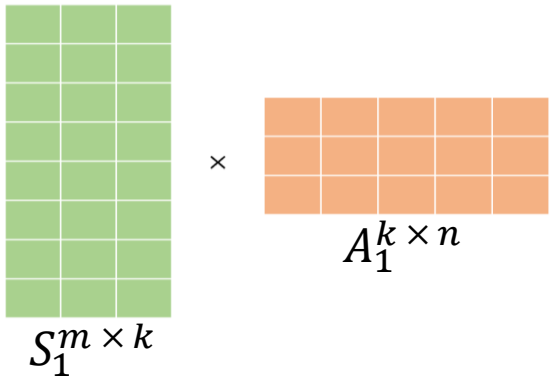
$$\frac{\partial f(x)}{\partial x} = \frac{c}{x} - \log \frac{c}{x} - 1 = \frac{\partial}{\partial x} \left(\frac{c}{x} - \log c + \log x - 1 \right) = -\frac{c}{x^2} + \frac{1}{x}$$

Deriving with respect to w_{ij}

$$\frac{\partial D_{\mathrm{IS}}(\mathbf{X}, \mathbf{WH})}{\partial w_{ij}} = \sum_{m,n} x_{mn} \frac{\partial}{\partial w_{ij}} \frac{1}{(\mathbf{WH})_{mn}} + \sum_{m,n} \frac{\partial}{\partial w_{ij}} \log (\mathbf{WH})_{mn}$$

$$\nabla_{\mathbf{W}} D_{\mathrm{IS}} (\mathbf{X}, \mathbf{WH}) = -\frac{\mathbf{X}}{(\mathbf{WH})^{\circ 2}} \mathbf{H}^{\intercal} + \frac{1}{\mathbf{WH}} \mathbf{H}^{\intercal}$$

Itakura-Saito divergence



The Itakura-Saito cost function:

$$D_{\mathrm{IS}}(\mathbf{X} \mid \mathbf{WH}) = \sum_{i,j} \left(\frac{\mathbf{X}_{ij}}{(\mathbf{WH})_{ij}} - \log \frac{\mathbf{X}_{ij}}{(\mathbf{WH})_{ij}} - 1 \right)$$

$M^{m \times n}$ $S^{m \times k}$ $A^{k \times n}$

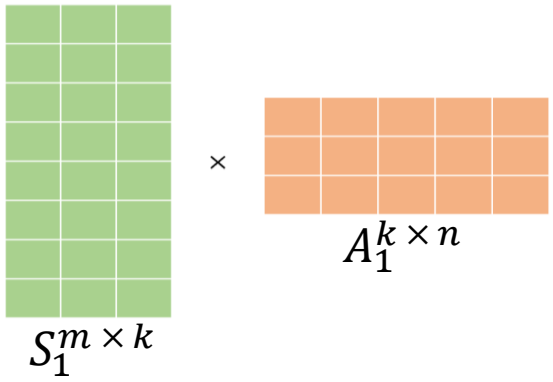
Minimization problem

$$\frac{\partial f(x)}{\partial x} = \frac{c}{x} - \log \frac{c}{x} - 1 = \frac{\partial}{\partial x} \left(\frac{c}{x} - \log c + \log x - 1 \right) = -\frac{c}{x^2} + \frac{1}{x}$$

Deriving with respect to h_{ij}

$$\nabla_{\mathbf{H}} D_{\mathrm{IS}} \left(\mathbf{X}, \mathbf{WH} \right) = -\mathbf{W}^{\intercal} \frac{\mathbf{X}}{(\mathbf{WH})^{\circ 2}} + \mathbf{W}^{\intercal} \frac{1}{\mathbf{WH}}$$

Itakura-Saito divergence



$$M^{m \times n} \approx S^{m \times k} \times A^{k \times n}$$

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^\intercal \frac{\mathbf{X}}{(\mathbf{WH})^{\circ 2}}}{\mathbf{W}^\intercal \frac{1}{\mathbf{WH}}}$$

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\frac{\mathbf{X}}{(\mathbf{WH})^{\circ 2}} \mathbf{H}^\intercal}{\frac{1}{\mathbf{WH}} \mathbf{H}^\intercal}$$

$S^{m \times k}$

$A^{k \times n}$

10’000 – 1’000’000 NMF multiplicative update steps

Slide 1 — Concept

Title: Nonnegative Matrix Factorization (NMF)

What it is:

- A **dimensionality reduction** technique for nonnegative data (all entries ≥ 0).
- Factorizes a data matrix into two smaller nonnegative matrices:

$$V \approx WH$$

Interpretation:

- V: original data ($m \times n$) — e.g., mutational spectra (channels × samples)
- W: basis/signatures ($m \times k$) — patterns (mutational signatures)
- H: coefficients/exposures ($k \times n$) — how much each signature contributes to each sample
- k: reduced rank (number of signatures, chosen << min(m,n))

Visual Idea:

Data ≈ **(Signatures)** × **(Exposures)**

→ Each sample is explained as a weighted combination of latent signatures.

Slide 2 — Main Algebra

Optimization problem:

$$\min_{W,H \geq 0} \|V - WH\|_F^2$$

where $\|\cdot\|_F$ = Frobenius norm (sum of squared differences).

Constraints:

- $W_{ij} \geq 0,\ H_{ij} \geq 0$
- Nonnegativity makes the decomposition **additive and interpretable** (no cancellation).

Update rules (multiplicative, Lee & Seung 1999):

$$H \leftarrow H \odot \frac{W^T V}{W^T W H}, \quad W \leftarrow W \odot \frac{V H^T}{W H H^T}$$

($\odot$ = elementwise multiplication; fractions = elementwise division).

Outcome:

- W: basis vectors (mutational signatures)
- H: weights (per-sample exposures)

The "-log(a/x)-1" formula for the Itakura-Saito (IS) divergence in NMF leads to multiplicative update rules through a careful manipulation of the gradient. Here's how the gradient becomes the widely-used iterative formula:

IS Divergence and NMF

For input matrix X and reconstruction WH, the IS divergence is:

$$D_{IS}(X|WH) = \sum_{i,j} \left(\frac{X_{ij}}{(WH)_{ij}} - \log \frac{X_{ij}}{(WH)_{ij}} - 1 \right)$$

The goal is to minimize D_{IS} with respect to W and H under non-negativity.

Gradient and Update Rule

Compute the gradient of D_{IS} with respect to H (similarly for W):

$$\frac{\partial D_{IS}}{\partial H_{kn}} = \sum_f W_{fk} \left(-\frac{X_{fn}}{(WH)_{fn}^2} + \frac{1}{(WH)_{fn}} \right)$$

If you naively use gradient descent, you'd get additive updates. But NMF multiplicative updates maintain non-negativity, so we manipulate the gradient into a positive-negative ratio (often using a majorization-minimization framework). jjburred +3

Multiplicative Update Derivation

- Split the gradient into positive/negative terms. For IS divergence (see Févotte 2009, JJ Burred's notes): hil.u-tokyo +1
  - The negative part (for H): $W^T\left(\frac{X}{(WH)^2}\right)$
  - The positive part: $W^T\left(\frac{1}{WH}\right)$
- The update formula:

$$H \leftarrow H \cdot \frac{\text{Negative part}}{\text{Positive part}} = H \cdot \frac{W^T\left(\frac{X}{(WH)^2}\right)}{W^T\left(\frac{1}{WH}\right)}$$

Where all multiplications and divisions are element-wise.

Why Multiplicative?

The element-wise multiplication/division maintains non-negativity—at every step, H remains non-negative. This is the advantage over simple gradient descent. perso.telecom-paristech +1

If you naively use gradient descent, you'd get additive updates. But NMF multiplicative updates maintain non-negativity, so we manipulate the gradient into a positive-negative ratio (often using a majorization-minimization framework). jjburred +3

Multiplicative Update Derivation

- Split the gradient into positive/negative terms. For IS divergence (see Févotte 2009, JJ Burred's notes): hil.u-tokyo +1
  - The negative part (for H): $W^T\left(\frac{X}{(WH)^2}\right)$
  - The positive part: $W^T\left(\frac{1}{WH}\right)$
- The update formula:

$$H \leftarrow H \cdot \frac{\text{Negative part}}{\text{Positive part}} = H \cdot \frac{W^T\left(\frac{X}{(WH)^2}\right)}{W^T\left(\frac{1}{WH}\right)}$$

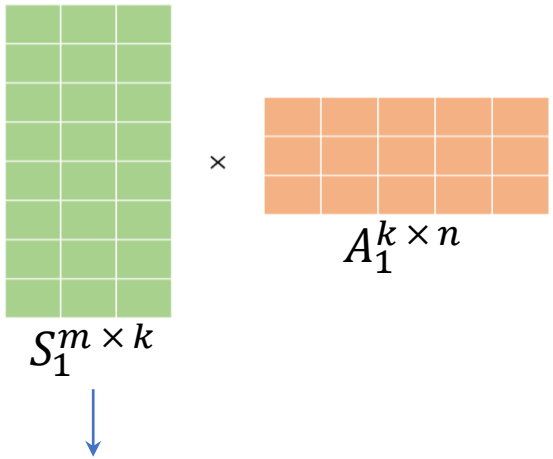
Where all multiplications and divisions are element-wise.

Why Multiplicative?

The element-wise multiplication/division maintains non-negativity—at every step, H remains non-negative. This is the advantage over simple gradient descent. perso.telecom-paristech +1

Partition clustering

k = 3 (number of operative signatures)

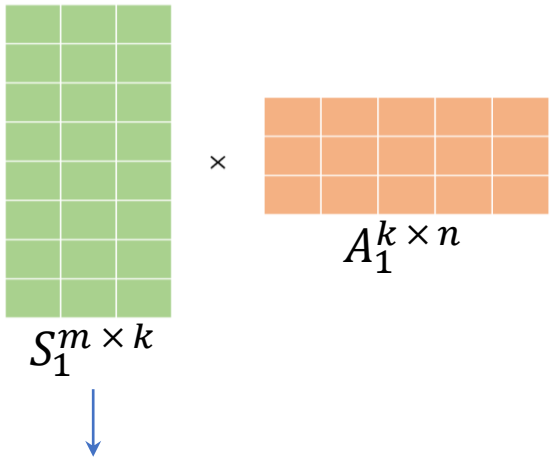


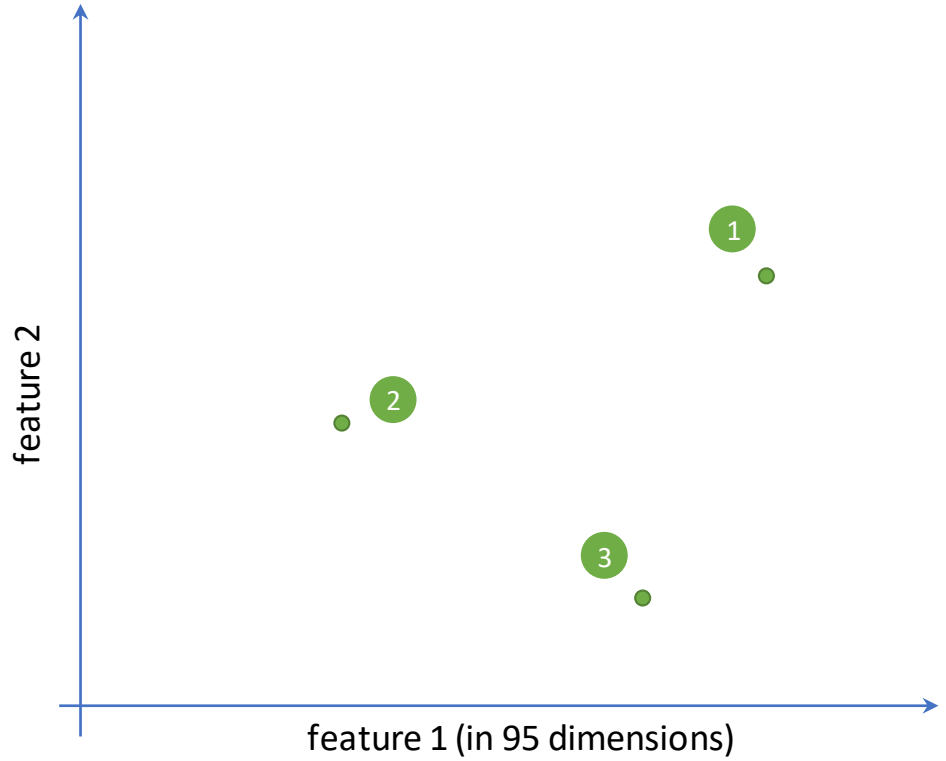
Clustering algorithm

1. **k clusters initialized randomly**
2. Each column of Si assigned to a different centroid (in 96 dimensions)
3. Repeat for all 100 Si matrices
4. After assigning all columns to a cluster, the centroid of each cluster is recalculated

Partition clustering

k = 3 (number of operative signatures)





Clustering algorithm

1. k clusters initialized randomly
2. **Each column of Si assigned to a different centroid (in 96 dimensions)**
3. Repeat for all 100 Si matrices
4. After assigning all columns to a cluster, the centroid of each cluster is recalculated

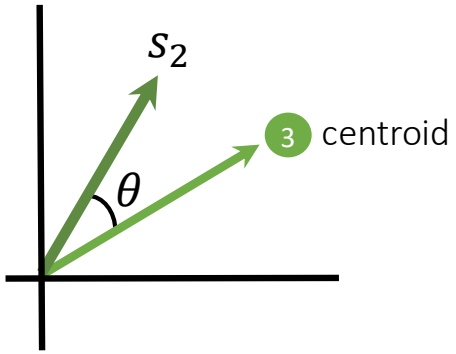
Partition clustering

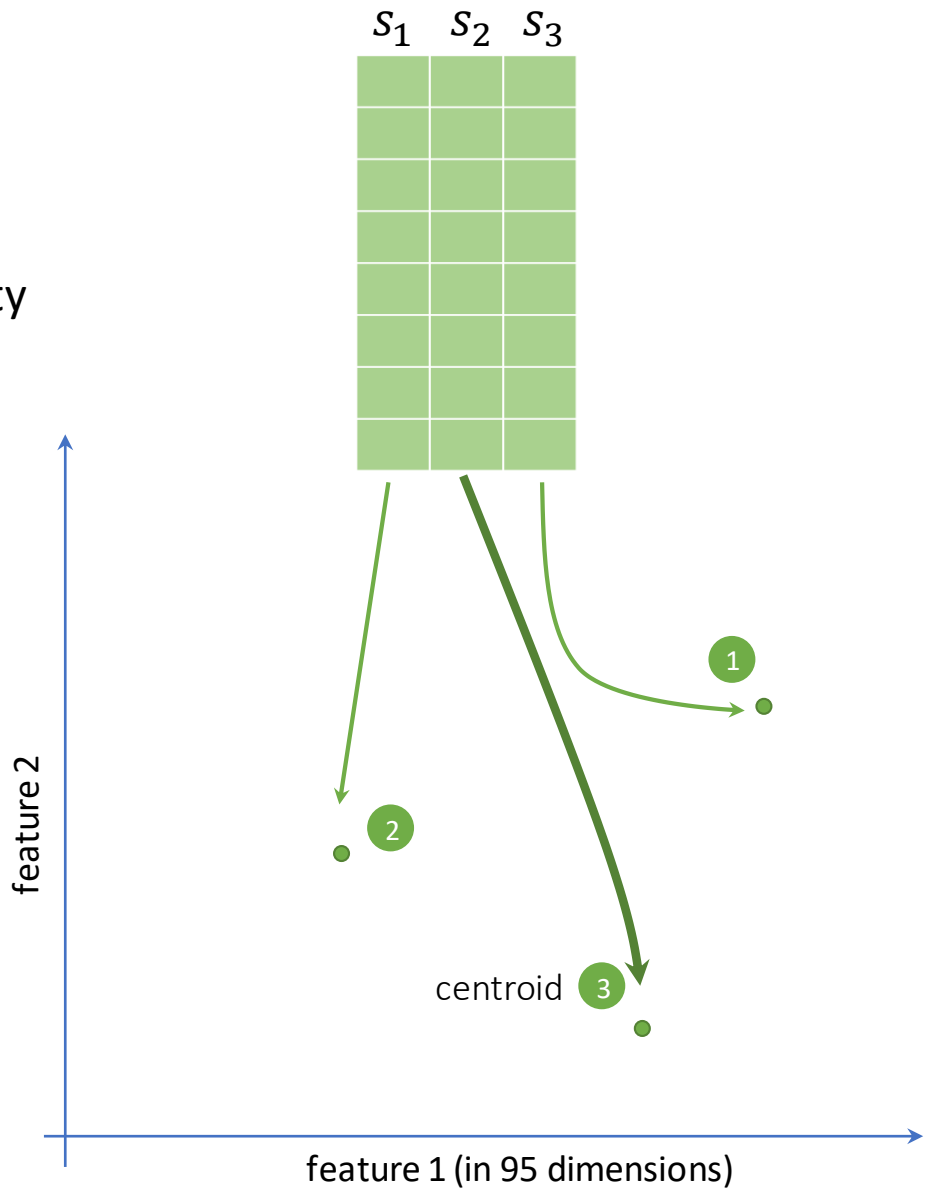
Each column of Si assigned to a different centroid

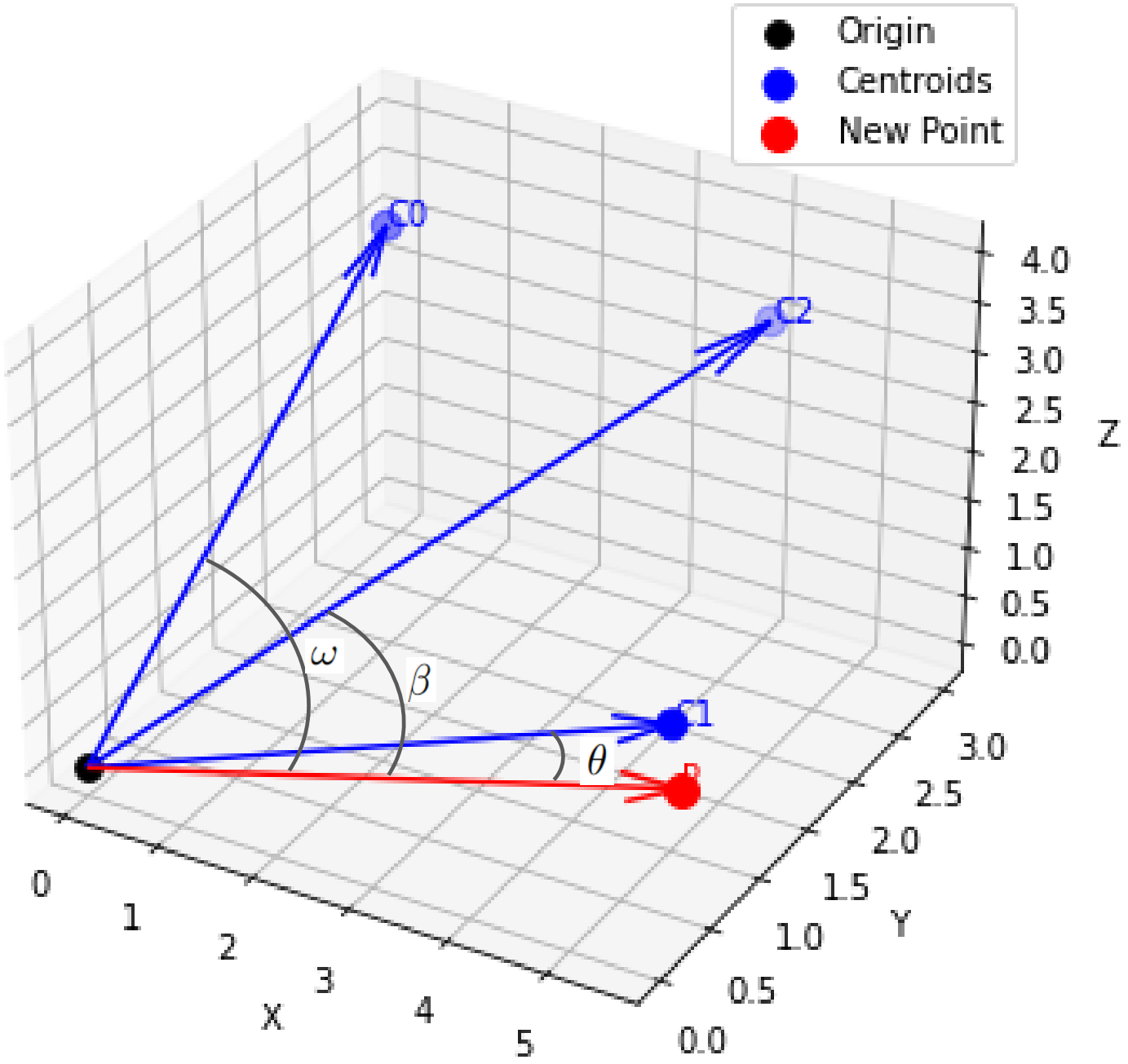
the Hungarian algorithm pair consensus vectors (cluster centroids and mutational signature from a matrix) by maximizing the cosine similarity

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

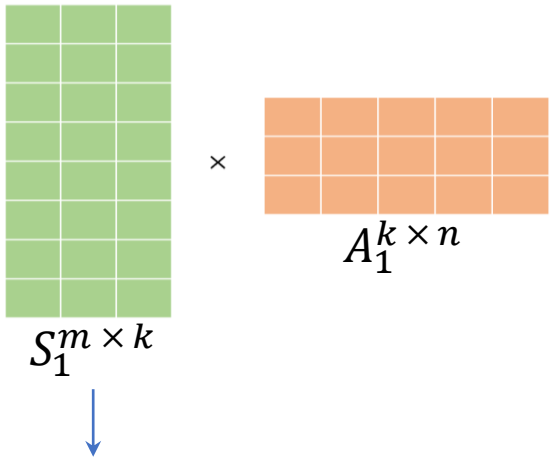


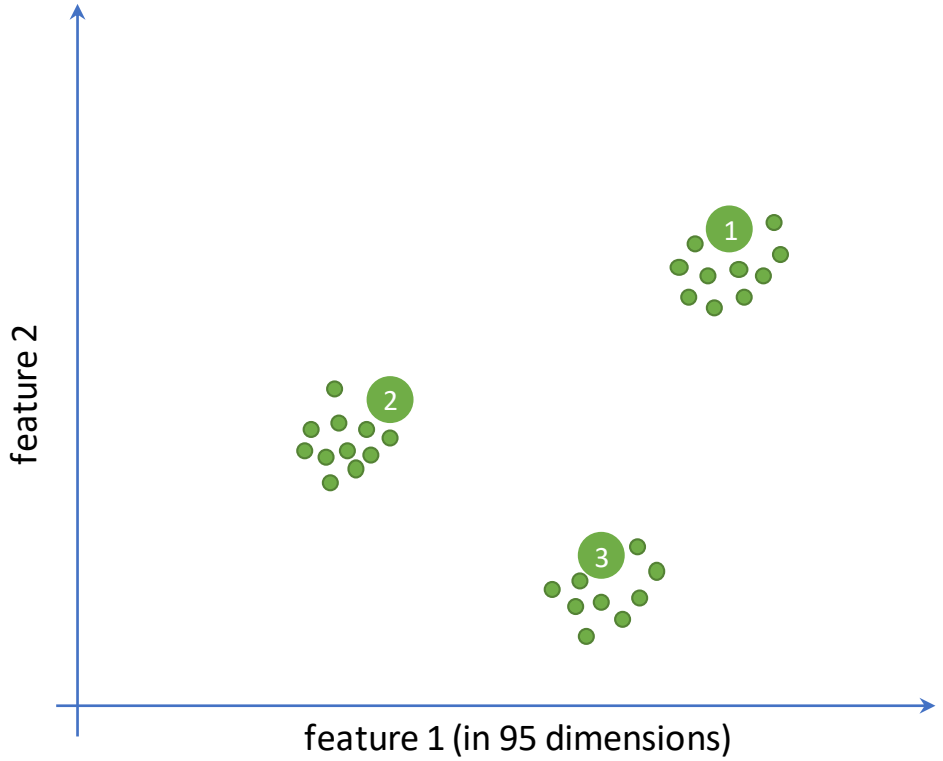



Origin
Centroids
New Point
C0
C1
C2
P
ω
β
θ
X
Y
Z
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Partition clustering

k = 3 (number of operative signatures)

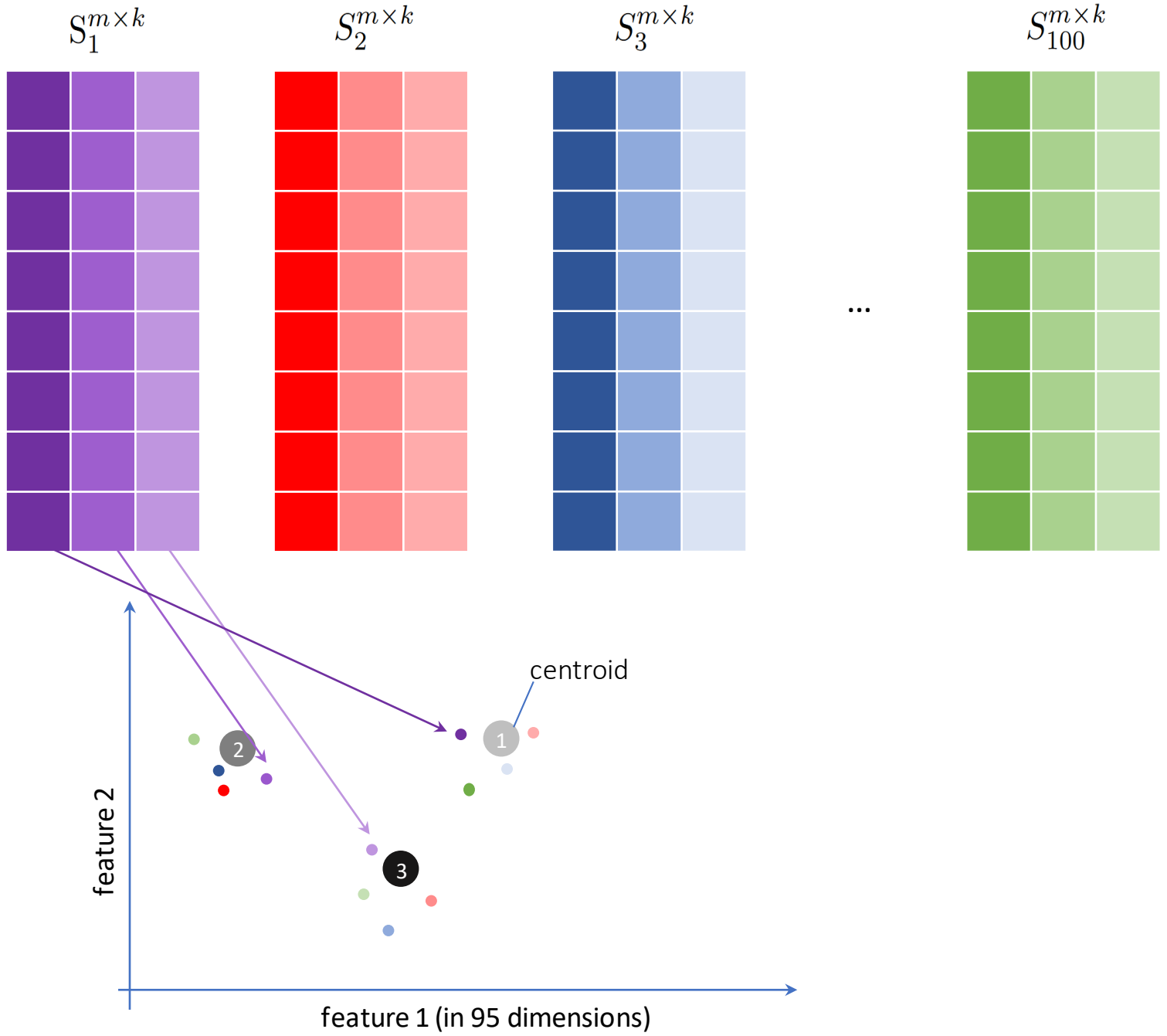




Clustering algorithm

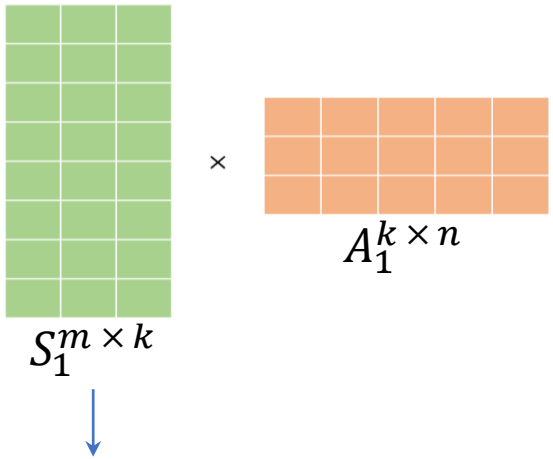
1. k clusters initialized randomly
2. Each column of Si assigned to a different centroid (in 96 dimensions)
3. **Repeat for all 100 Si matrices**
4. After assigning all columns to a cluster, the centroid of each cluster is recalculated

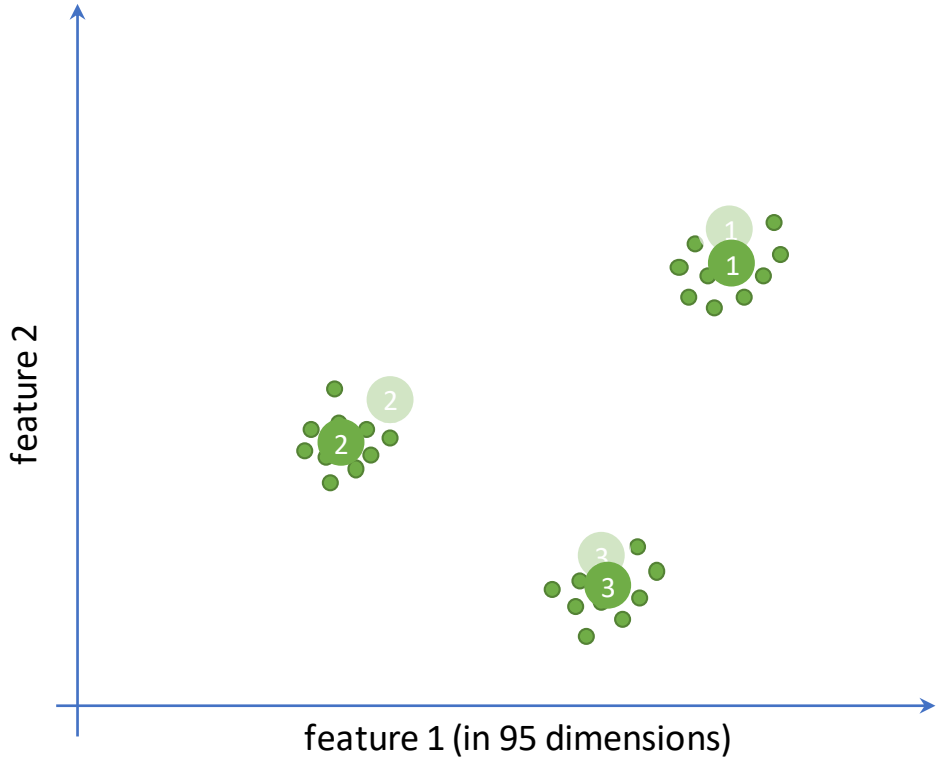
Partition clustering



Partition clustering

k = 3 (number of operative signatures)





Clustering algorithm

1. k clusters initialized randomly
2. Each column of Si assigned to a different centroid (in 96 dimensions)
3. Repeat for all 100 Si matrices
4. **After assigning all columns to a cluster, the centroid of each cluster is recalculated until convergence**

Partition clustering

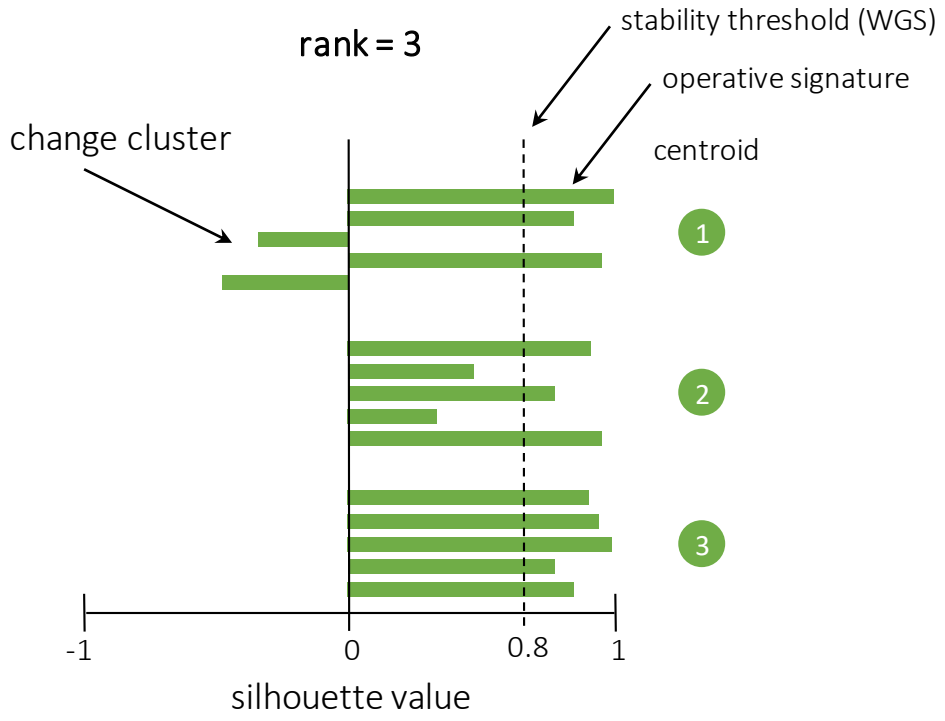
This process continues iteratively until the average silhouette coefficient converges

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

silhouette

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$



After convergence for a given value of k, the centroids of the clusters are reported as **consensus mutational signatures**

- $|C_I|$ represents the **cardinality** of the cluster, i.e., the **number of points** in cluster I.
- Each C_I is the set of points assigned to cluster I.
- The centroid of cluster I is a single point (often denoted c_I or similar) which is the mean position of all points in C_I, but C_I itself refers to the whole cluster membership (the set of points), not the centroid.

So your sums over $j \in C_I$ mean iterating over each point in the cluster I, and the denominator counts how many members are in that cluster (or one less, when excluding i itself).

Summary:

- C_I: cluster membership (set of points)
- $|C_I|$: number of points in cluster I
- centroid c_I: usually the average coordinates of points in C_I

This is standard notation in clustering literature.

Stability thresholds

WGS stable signatures:

- average stability > 0.80
- no individual signature having stability below 0.20

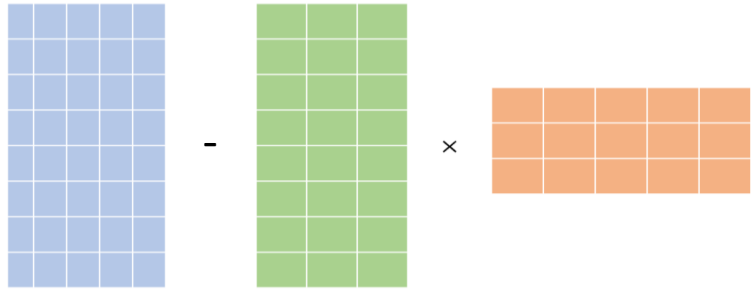
WES stable signatures:

- average stability > 0.70
- no individual signature having stability below 0.10

Model selection

Overfitting reduction(i.e. minimize the rank)

reconstructed error (residual) across all samples



rank = 24 R_{24} → Wilcoxon rank-sum tests (R_{25} , R_{24})

rank = 23 R_{23} → Wilcoxon rank-sum tests (R_{25} , R_{23})

⋮ ⋮ ⋮

rank = 1 R_1 → Wilcoxon rank-sum tests (R_{25} , R_1)

stop when p-value < 0.05

The stable solution with the lowest number of signatures and a Wilcoxon rank-sum test p-value above 0.05 is selected as the optimal solution

1. What data we have

For each tested solution (say **10 signatures, 11 signatures, 12 signatures**), the tool reconstructs each sample's mutational profile using those signatures.

That gives you, per sample j:

$$\text{Error}_j^{(k)} = \|M_j - \hat{M}_j^{(k)}\|$$

- M_j = observed mutational spectrum for sample j
- $\hat{M}_j^{(k)}$ = reconstructed spectrum using k signatures
- Error metric = usually **cosine distance** or **L1/L2 norm difference**

So for a dataset with N samples you get an **array of N errors per solution**.

2. What the Wilcoxon test compares

Take two solutions, e.g.:

- Array A = reconstruction errors across all samples with 12 signatures → $[e_1^{(12)}, e_2^{(12)}, \ldots, e_N^{(12)}]$
- Array B = reconstruction errors across all samples with 11 signatures → $[e_1^{(11)}, e_2^{(11)}, \ldots, e_N^{(11)}]$

Then apply a **Wilcoxon rank-sum test (a.k.a. Mann–Whitney U test)**:

$$H_0 : \text{The distribution of errors with 11 signatures is the same as with 12 signatures}$$

$$H_A : \text{Errors with 11 signatures are higher (worse fit) than with 12 signatures}$$

3. Interpretation

- If **p ≥ 0.05**: errors are not significantly worse → fewer signatures are acceptable → prefer 11 over 12.
- If **p < 0.05**: errors are significantly worse → keep 12.

Then the comparison continues downward (11 vs 10, etc.) until a significant difference is found.

---

4. Example with numbers

Imagine 5 samples, errors computed like this:

- 12 signatures: `[0.02, 0.03, 0.01, 0.05, 0.04]`
- 11 signatures: `[0.03, 0.04, 0.02, 0.06, 0.05]`

The 11-signature errors are consistently higher.

Wilcoxon test will give p < 0.05 → reject 11, keep 12.

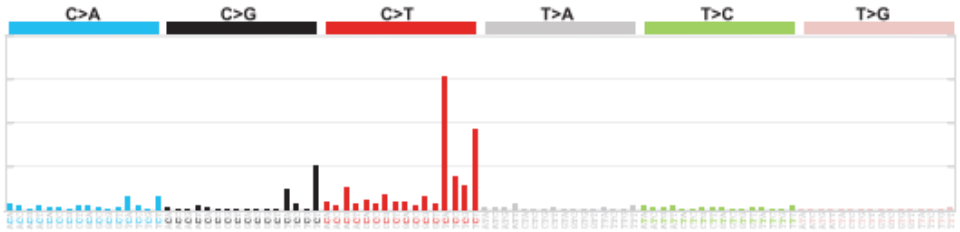
If instead the arrays were very close:

- 12 signatures: `[0.02, 0.03, 0.01, 0.05, 0.04]`
- 11 signatures: `[0.021, 0.031, 0.011, 0.051, 0.041]`

The distributions are nearly identical → p > 0.05 → keep 11 (simpler).

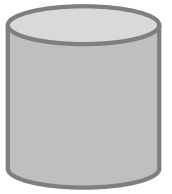
Decomposing de novo extracted signatures to known COSMIC signatures

De novo extracted signature

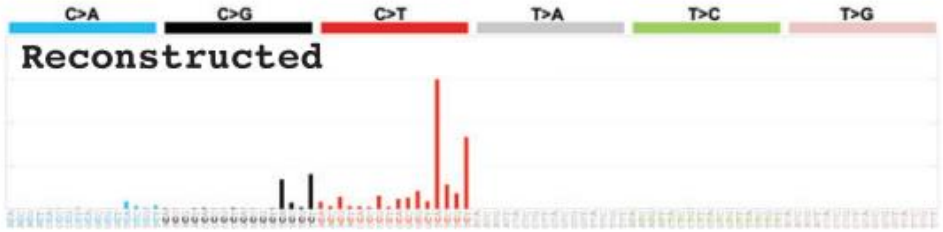


$\rightarrow y$

COSMIC
(SBS,DBS,ID)

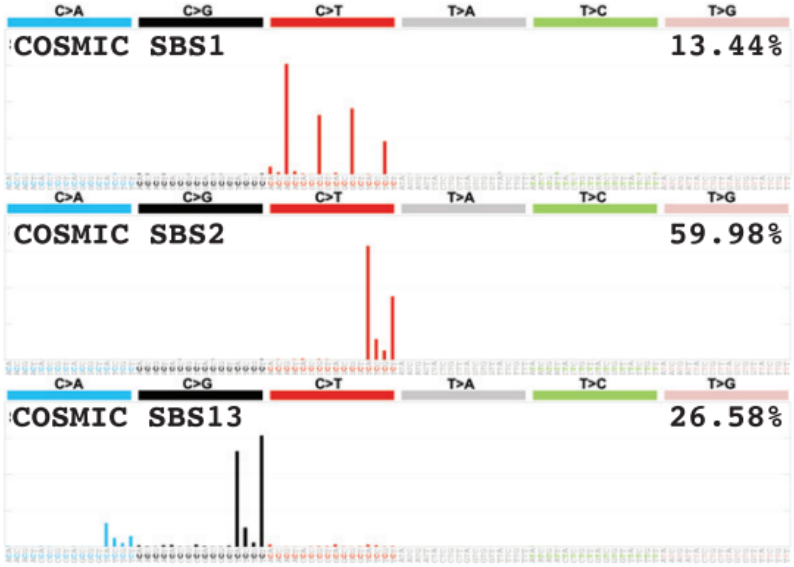


Reconstruction of a *de novo* signature using COSMIC signatures



$z = Ax$

Decomposition of a *de novo* signature using COSMIC signatures



The decomposition functionality leverages the nonnegative least squares (NNLS) algorithm

Minimizes the difference between the extracted signature and the reconstructed signature (using COSMIC signatures)

$$\arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \quad \text{subject to } \mathbf{x} \geq 0$$

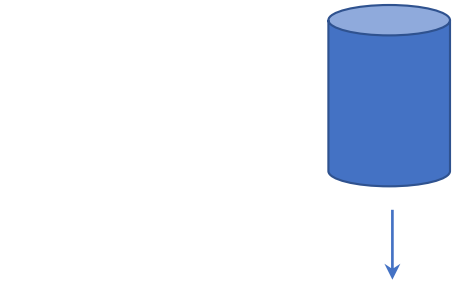
reconstructed signature

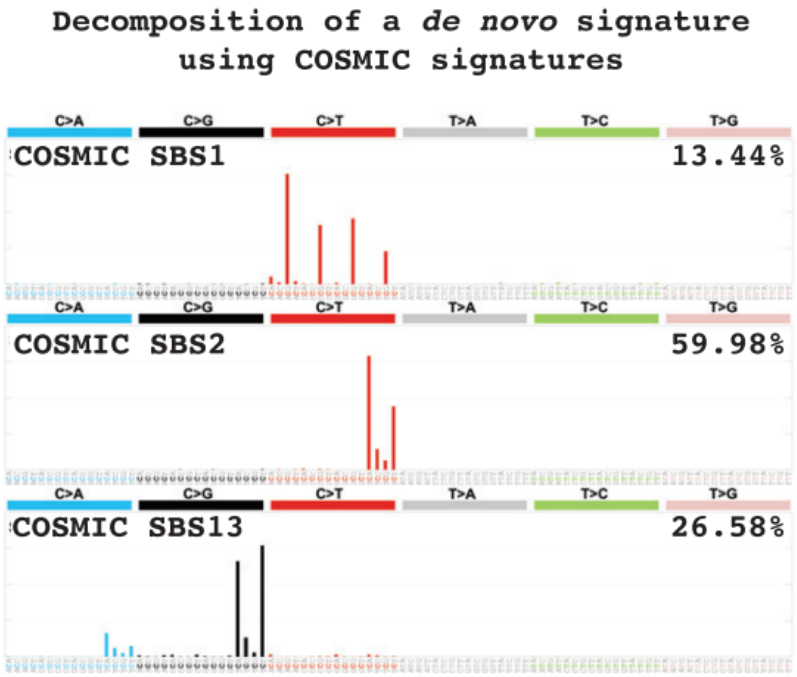
De novo extracted signature

L2-norm

$$\|\mathbf{e}\|_2 = \sqrt{\sum_{k=1}^{n} |e_k|^2}$$

COSMIC
(SBS,DBS,ID)

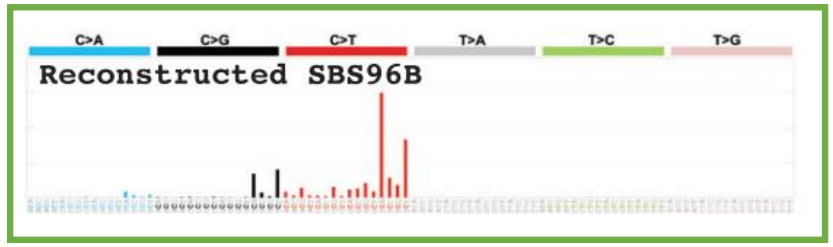




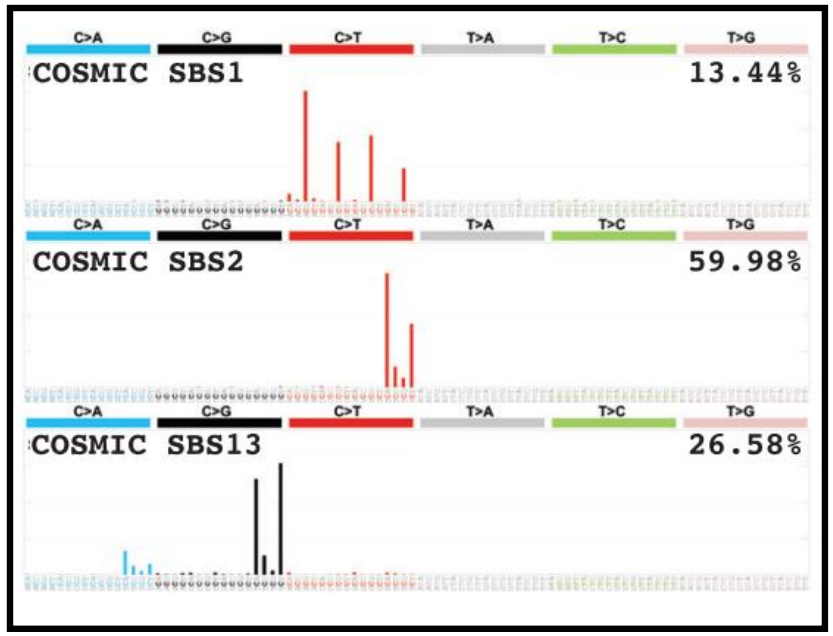
Evaluating activities of mutational signatures in individual samples

SigProfilerExtractor

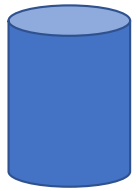
De novo extracted signature



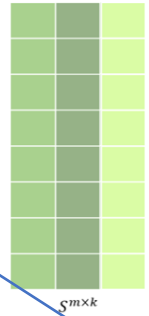
COSMIC derived signatures

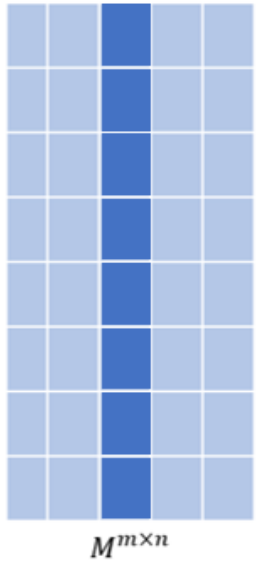


COSMIC
(SBS,DBS,ID)



first reconstructed

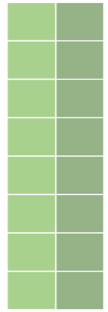


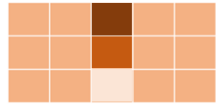


sample by sample

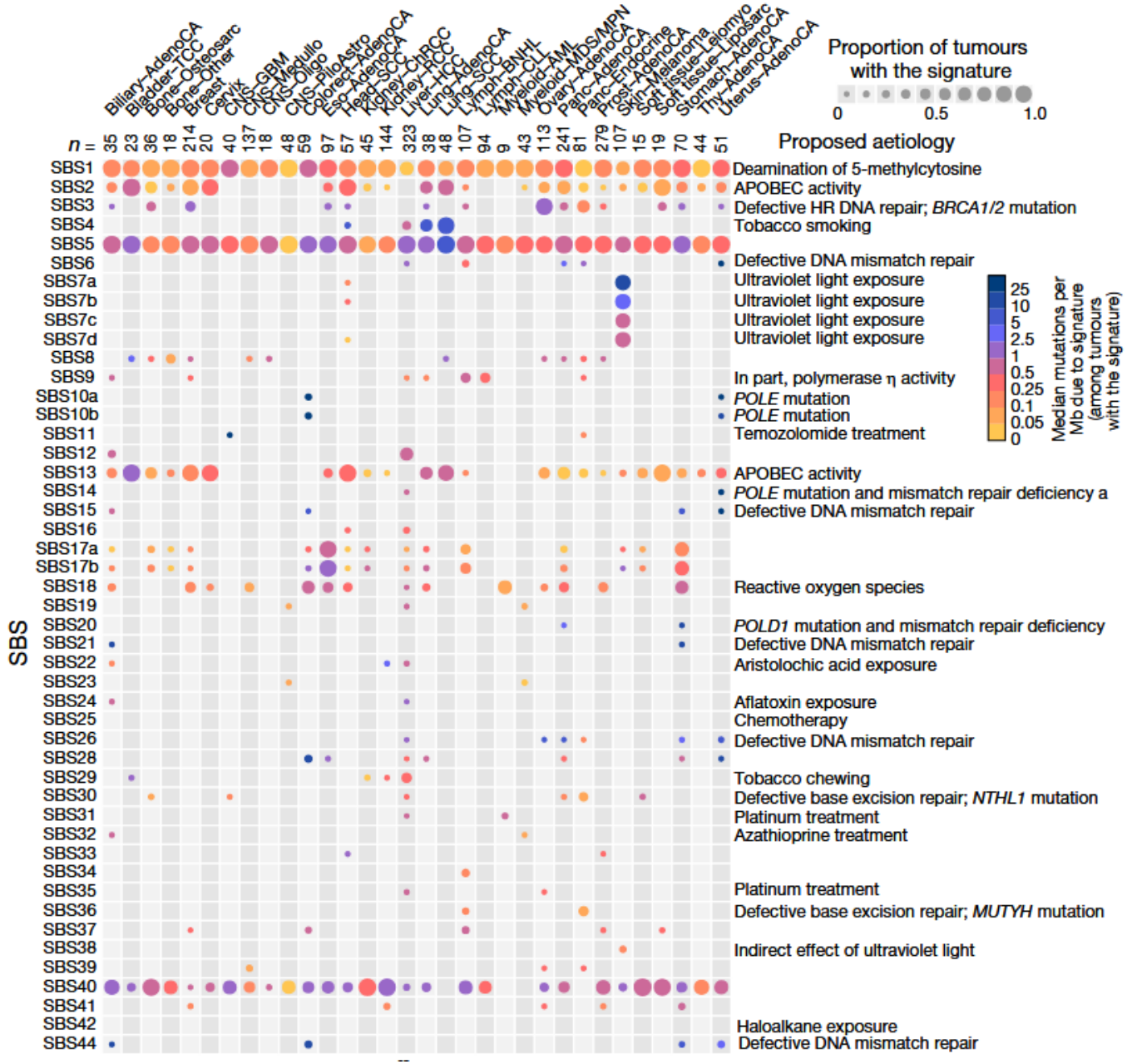
nonnegative least squares (NNLS)

all de novo signatures are initially added to the list, removal step with 2% cutoff





Activity of signatures in the individual sample



https://www.nature.com/articles/s41586-020-1943-3 -

Limitations & caveats

- Signatures are not always unique or specific (multiple processes can look similar)
- Need enough mutations for reliable extraction.
- Clinical translation still evolving (not all signatures are actionable)
- Noise and overfitting

Exercise 1: Exploring Mutational Contexts

- Dataset: VCF file (somatic mutations from a tumor sample)
- Tasks:
  - Count SNVs by type (C>A, C>G, etc.)
  - Build trinucleotide context (with reference genome, e.g., hg19 FASTA)
  - Visualize as a barplot (96-channel profile)
- Tools: R, Python

Exercise 2: Signature Extraction with NMF

- Provide a pre-computed mutation count matrix (samples × 96 contexts)
- run NMF (R: MutationalPatterns::extract_signatures) to identify 2–3 signatures.
- Compare extracted signatures against COSMIC reference.
- How many signatures are “real”? what about overfitting?