wellcome connecting science

connecting people with science

Dr Jia-Wern Pan
Cancer Research Malaysia

Introduction to mutational signatures
(adapted from Dr Marcos Díaz Gay, WCS CGA 2023)

April 2025

Recap: Variant Calling

- What does a VCF file contain?
- How does the data in a VCF file help a patient?
- How do you aggregate VCF files from multiple patients?

Why does X patient have X variants?

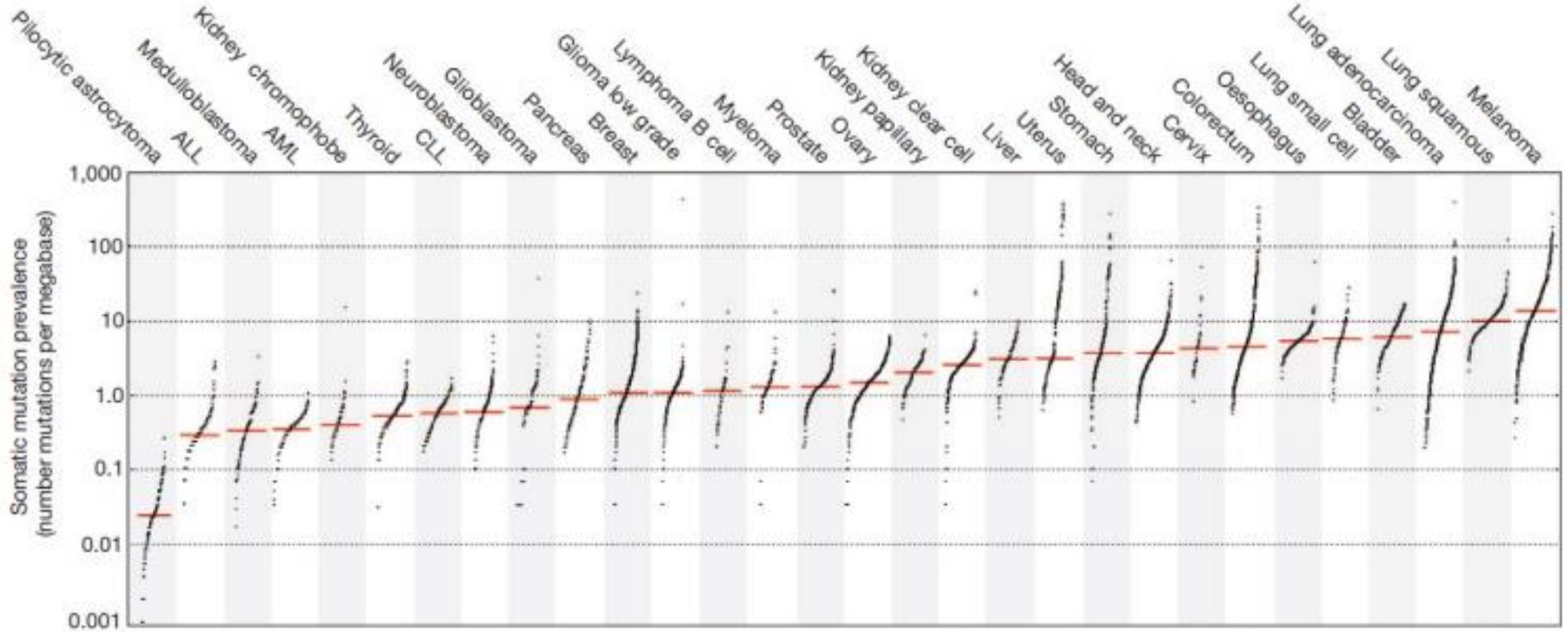
Mutational signature analysis is an attempt to determine the causal sources (aetiology) of somatic variants and mutations in an individual or set of samples

Discuss: What are common (or rare) causes of somatic mutations? How might these differ across cancer types?

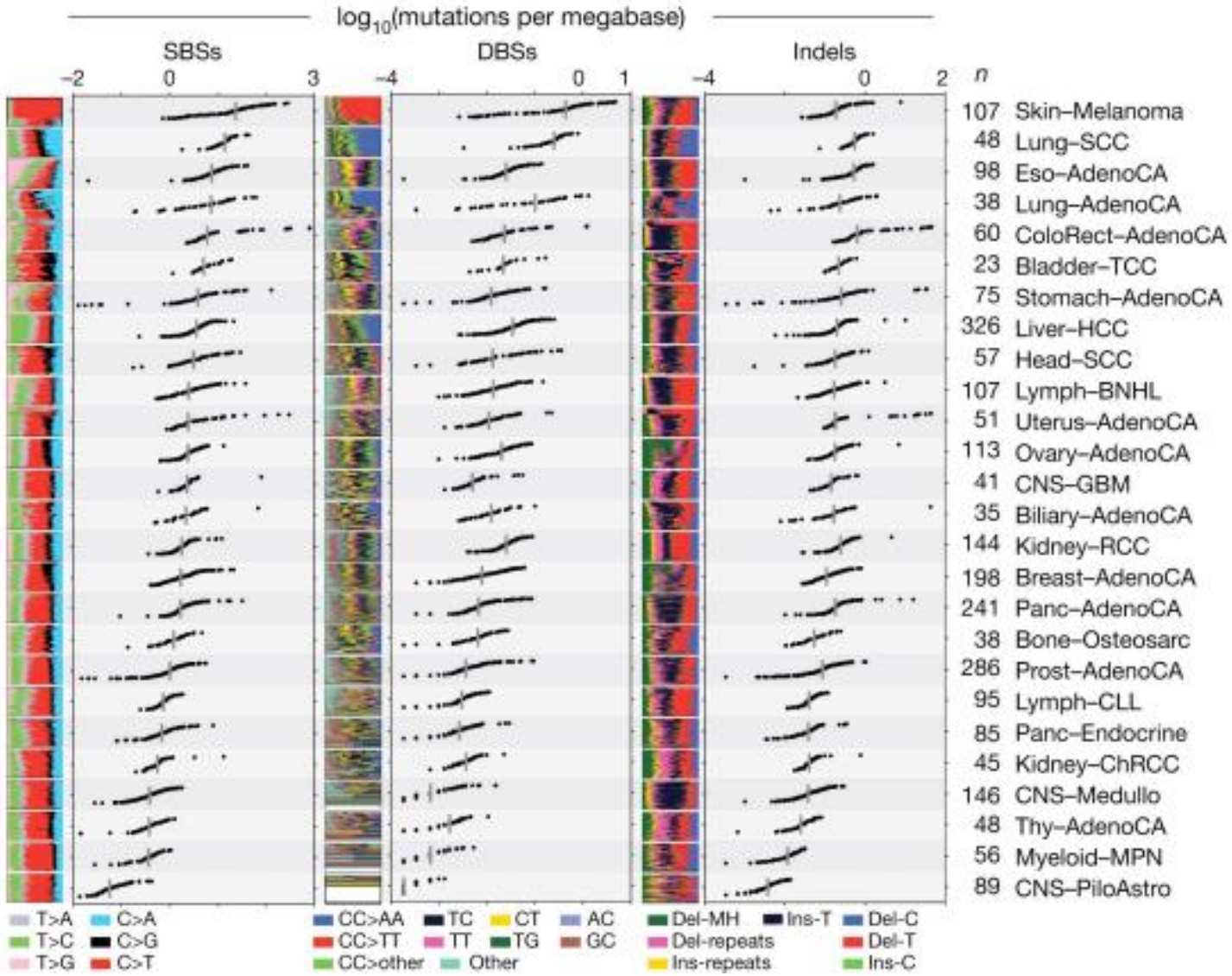
Well-known carcinogens and carcinogenic processes

- Aging
- UV light
- DNA repair deficiencies
- Tobacco smoke
- Alcohol
- Asbestos
- Radiation
- Bacteria i.e. *Heliobacter pylori*
- Viruses i.e. HPV, EBV

The burden of somatic mutations is highly variable across different types of tumours



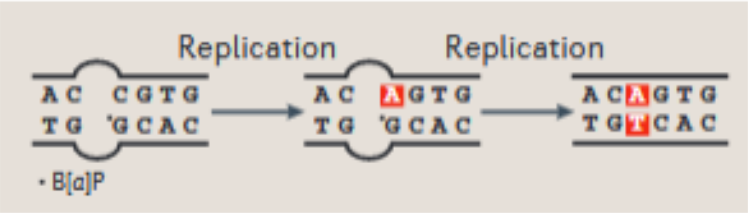
The burden of somatic mutations is highly variable across different types of tumours

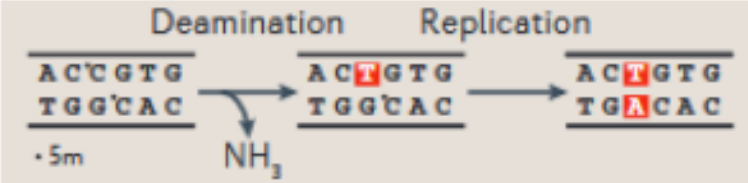


Different sources of DNA damage have different patterns of mutations

Environmental exposures

Tobacco smoking or chewing



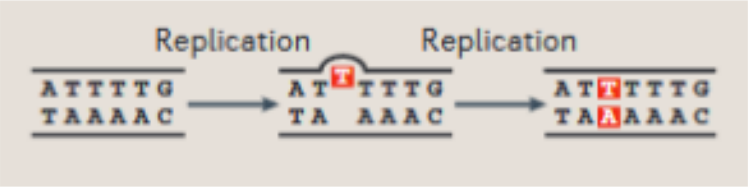


Normal cellular activities

Spontaneous deamination of methylated cytosines

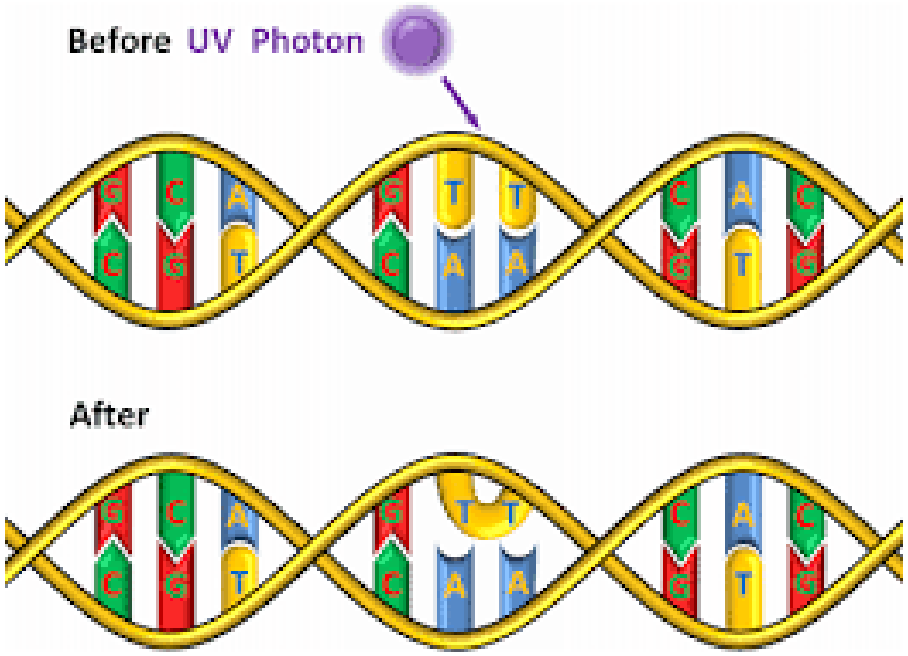
Failure in DNA replication or repair

Aberrant mismatch repair pathway

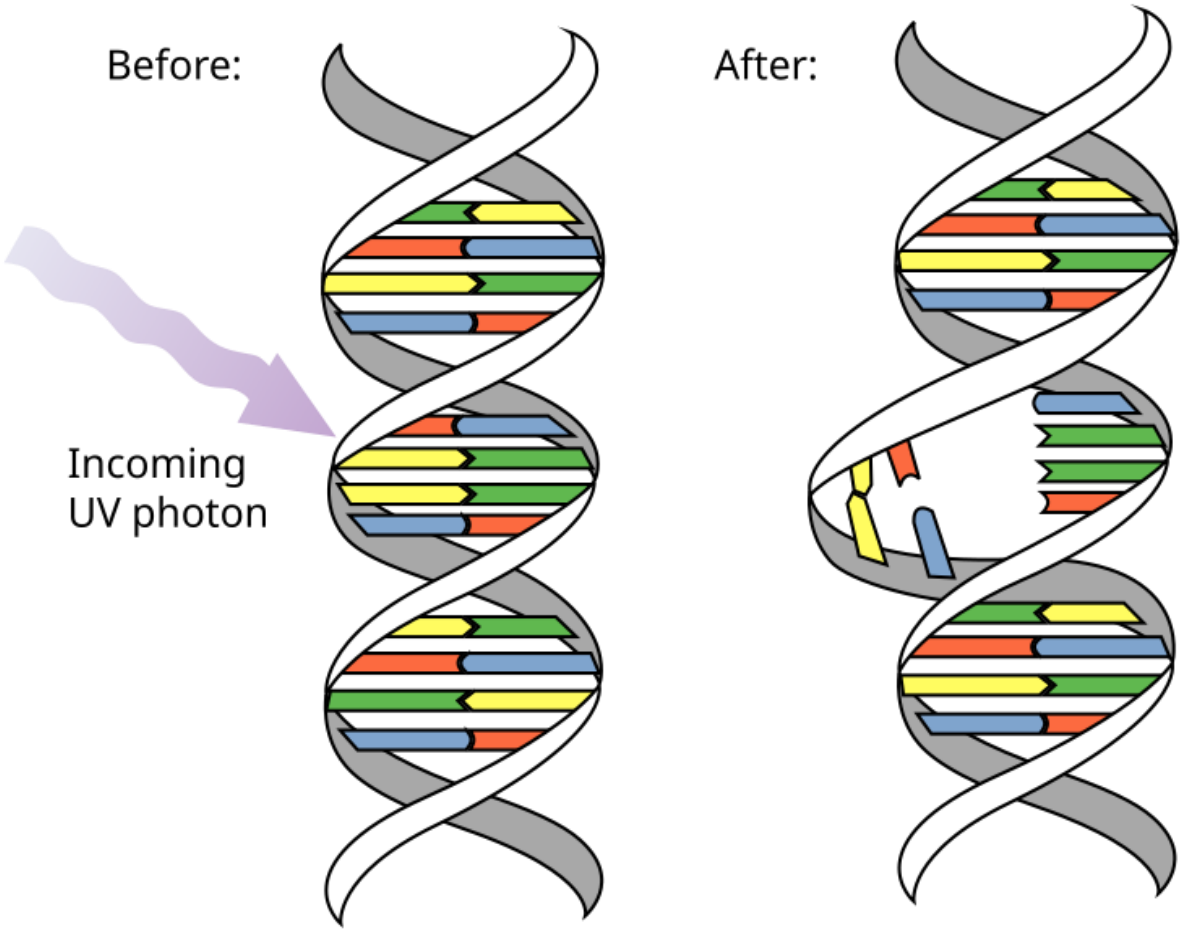


UV radiation

Thymine dimers

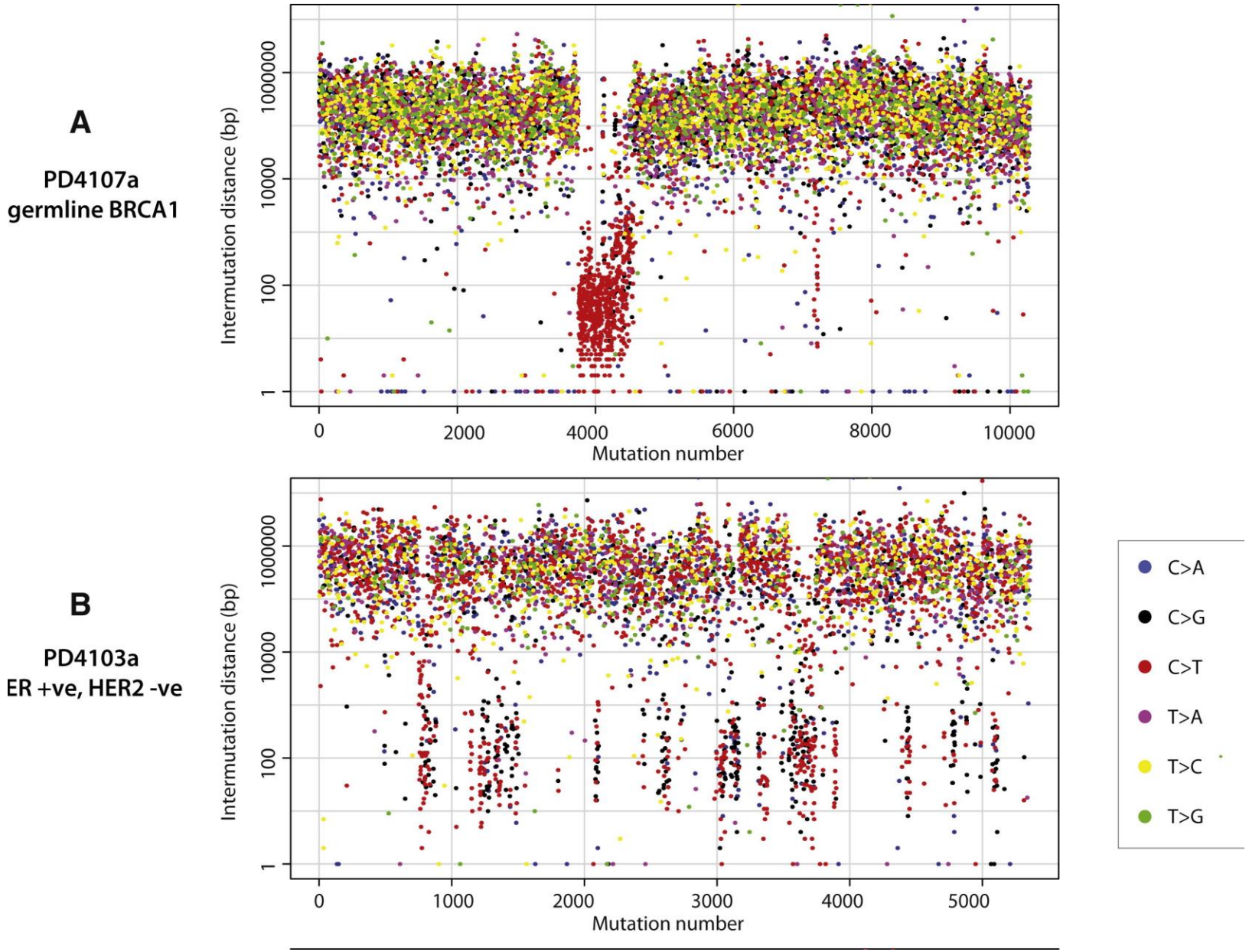


Schmid et al. 2017



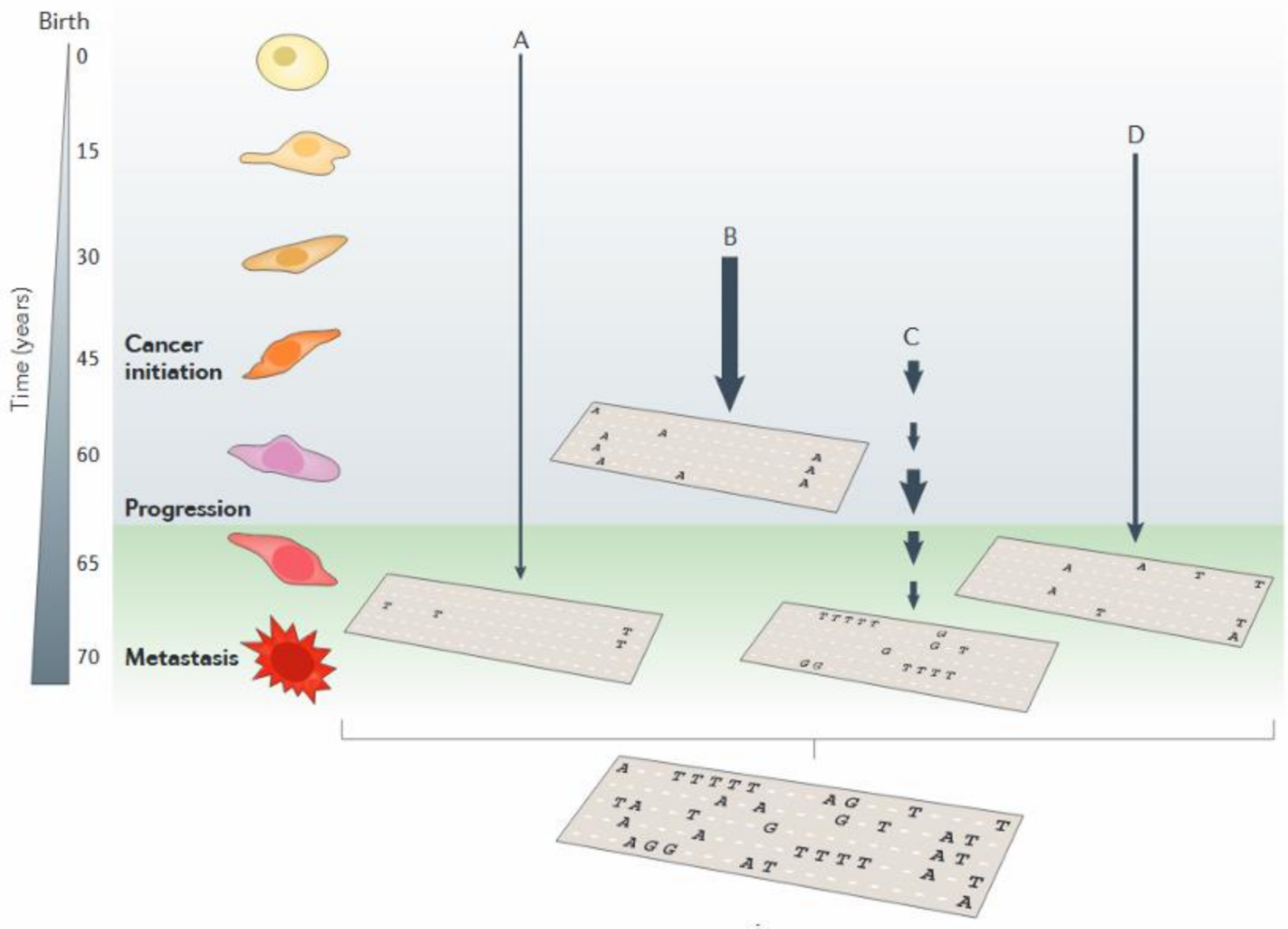
Kataegis

Localized somatic hypermutation associated with APOBEC enzymes and TLS DNA polymerases

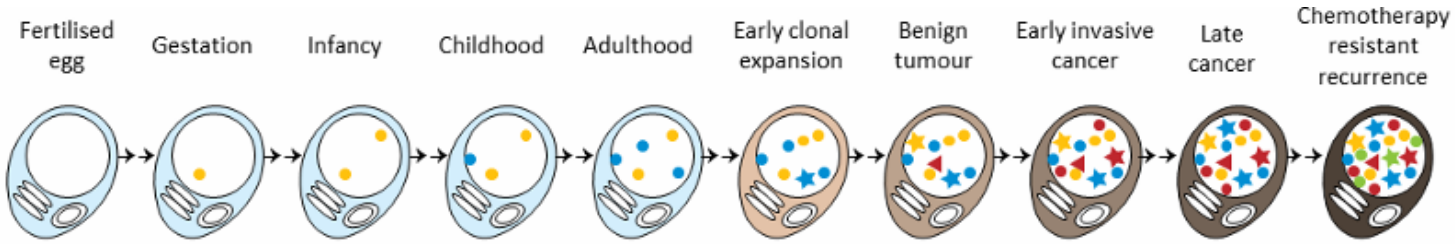


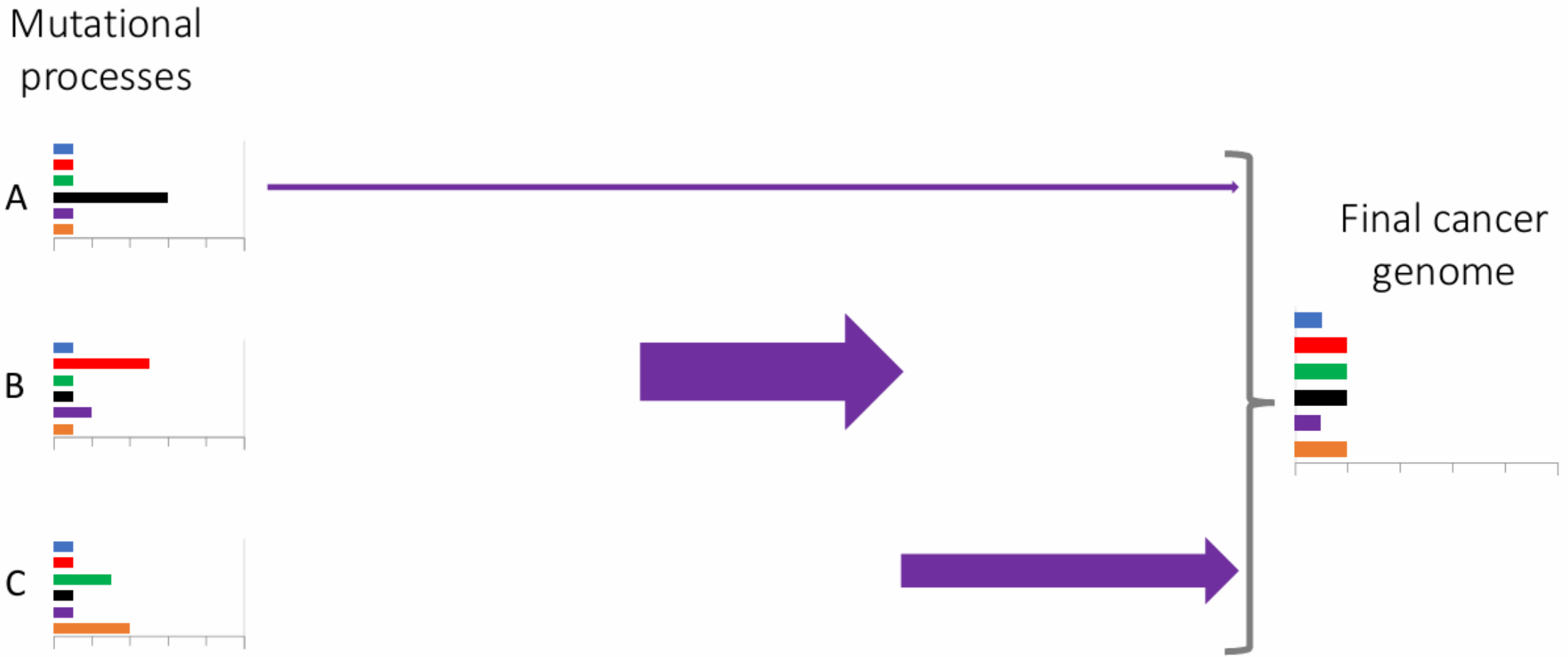
April 2025

Multiple mutational processes can occur simultaneously



Mutational processes accumulate over the life history of each cell





Looking for patterns of somatic mutations in the cancer genome

How do we actually find and discover mutational signatures from genomic data?

Looking for patterns of somatic mutations in the cancer genome

The first step is to classify or group different somatic mutations into categories

Discuss: What are different ways that you can group somatic mutations occurring at different positions throughout the genome?

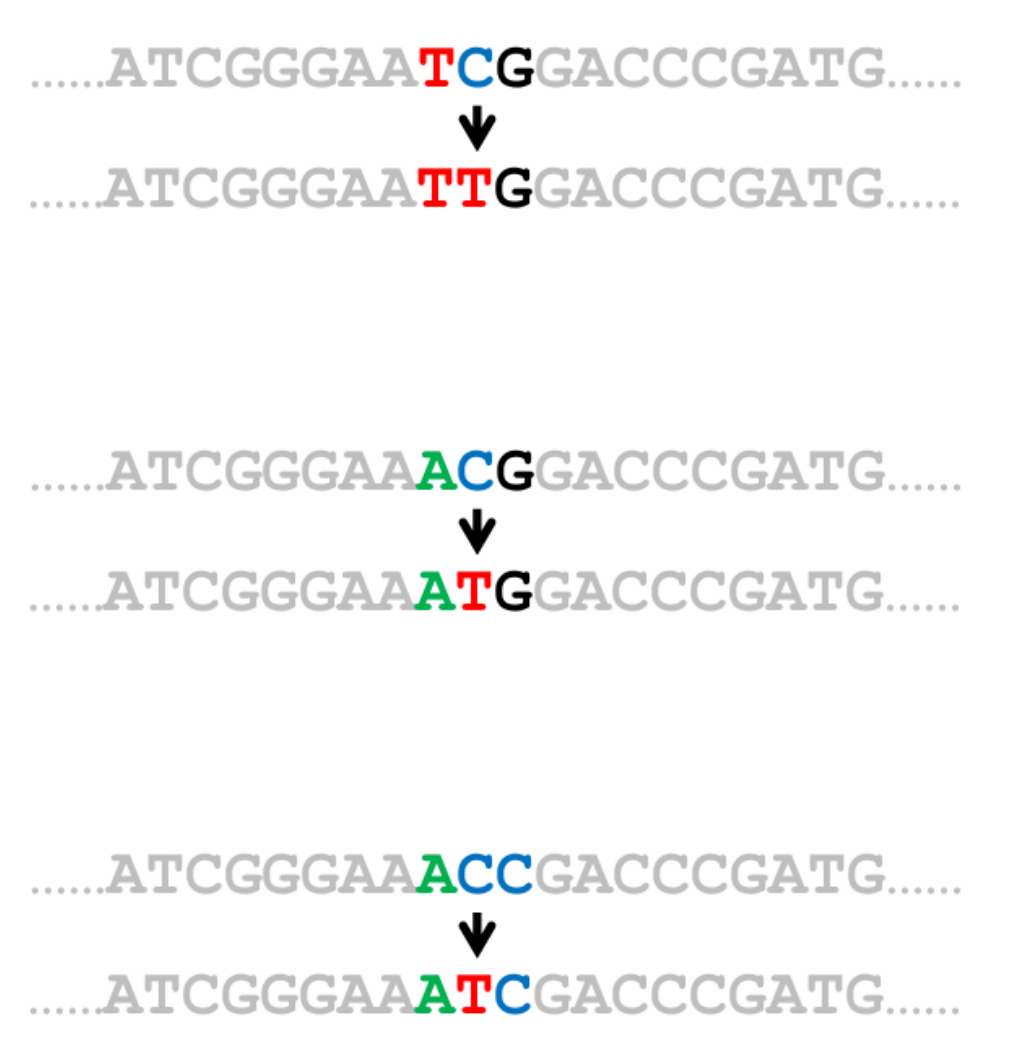
Classification of SNVs

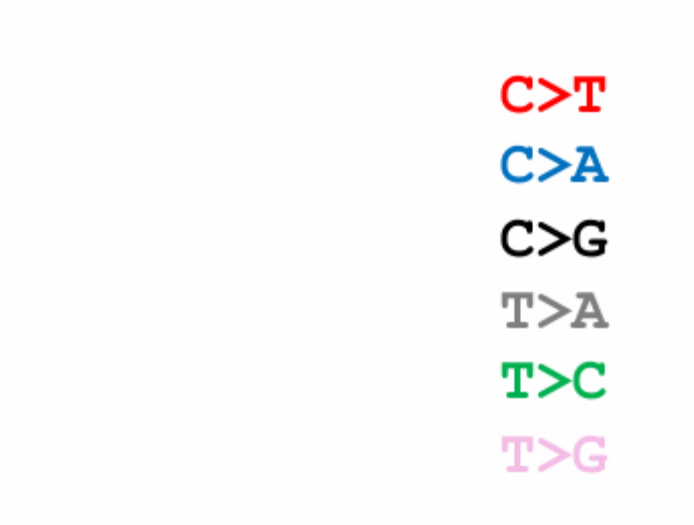
- Transitions versus transversions
- Synonymous versus non-synonymous
- Location/position
- Base substitutions
- **Mutational context**

Single base substitution

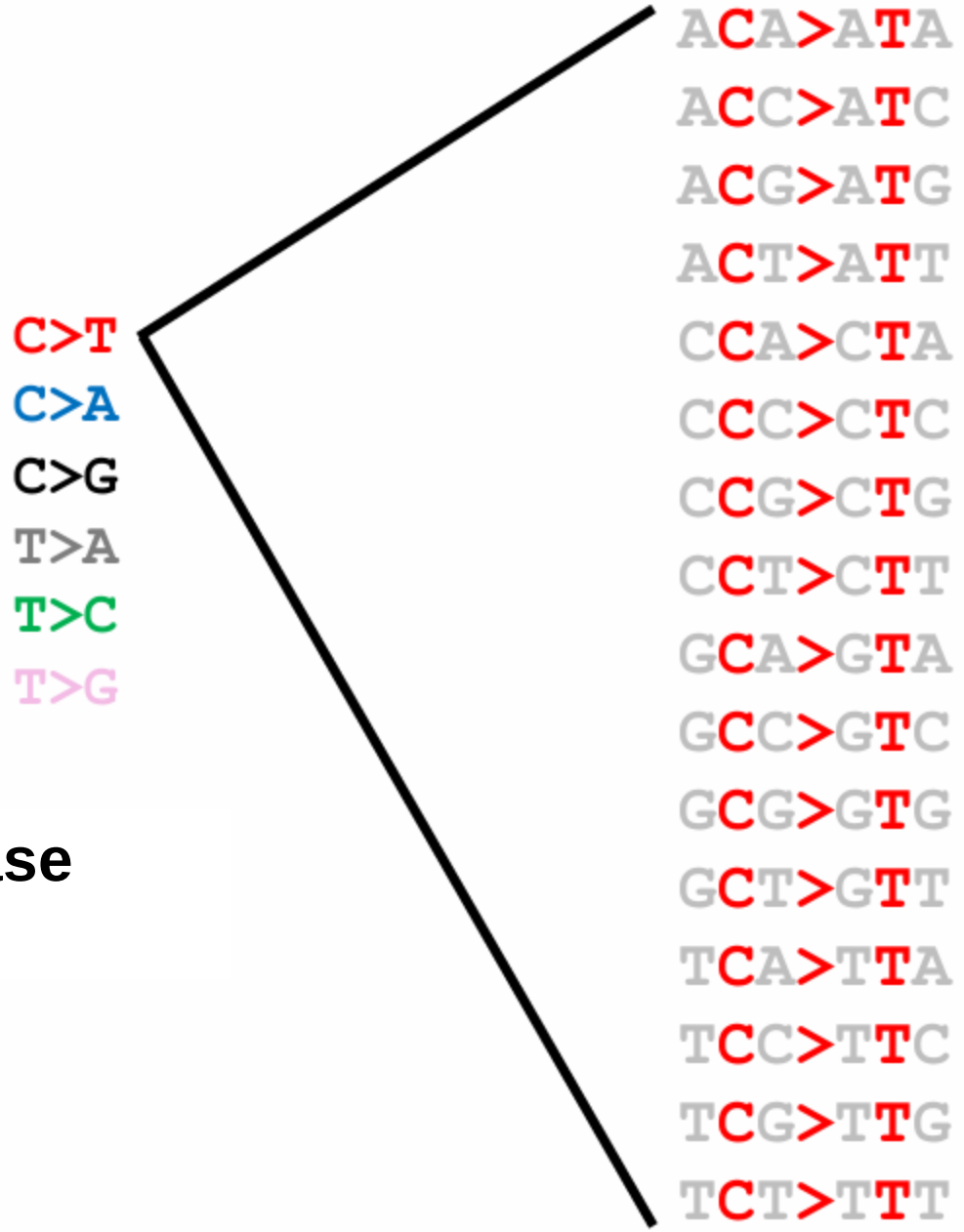
Trinucleotide context of the single base substitution

Trinucleotide context of the single base substitution



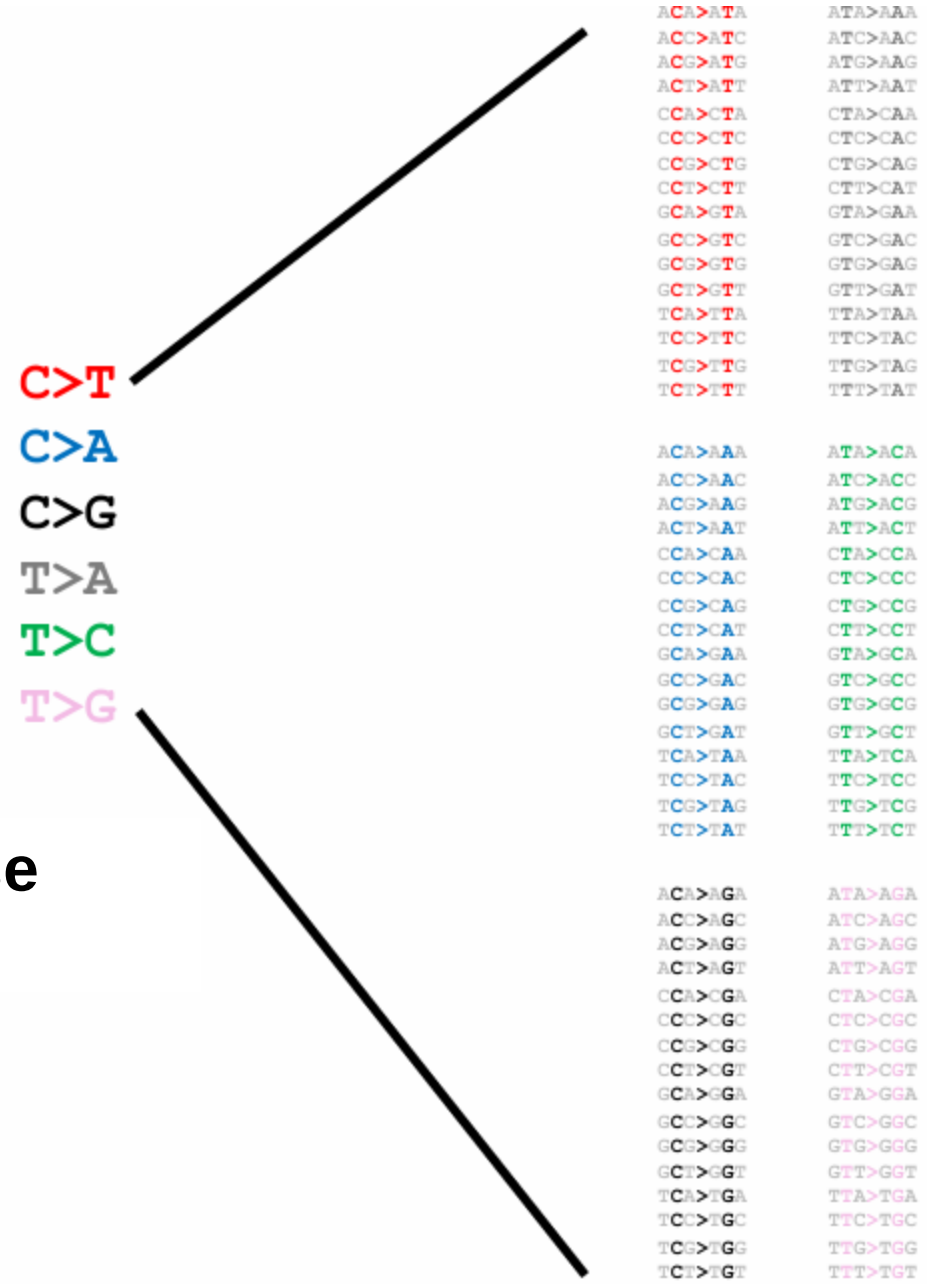


6 classes of base substitutions



6 classes of base substitutions

16 different trinucleotide contexts per class



6 classes of base substitutions

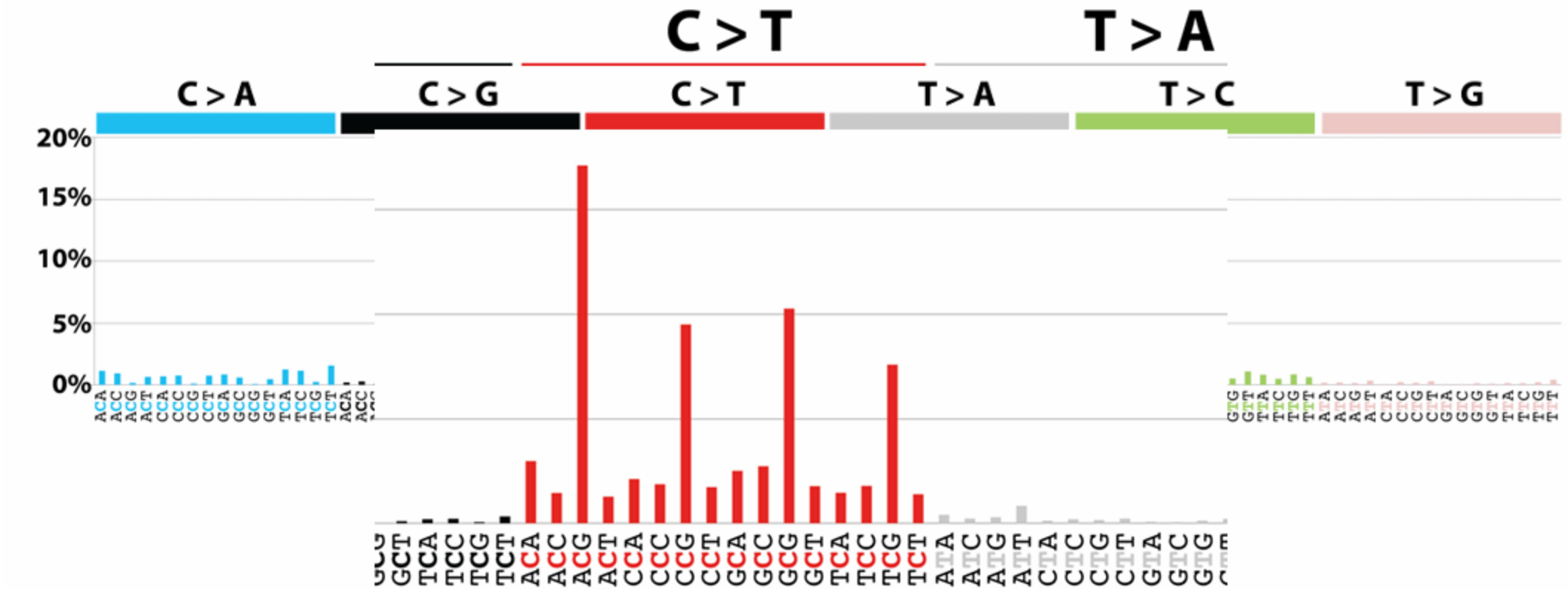
96 classes of SNV mutations

Six classes of single-base mutations

Reported by pyrimidine

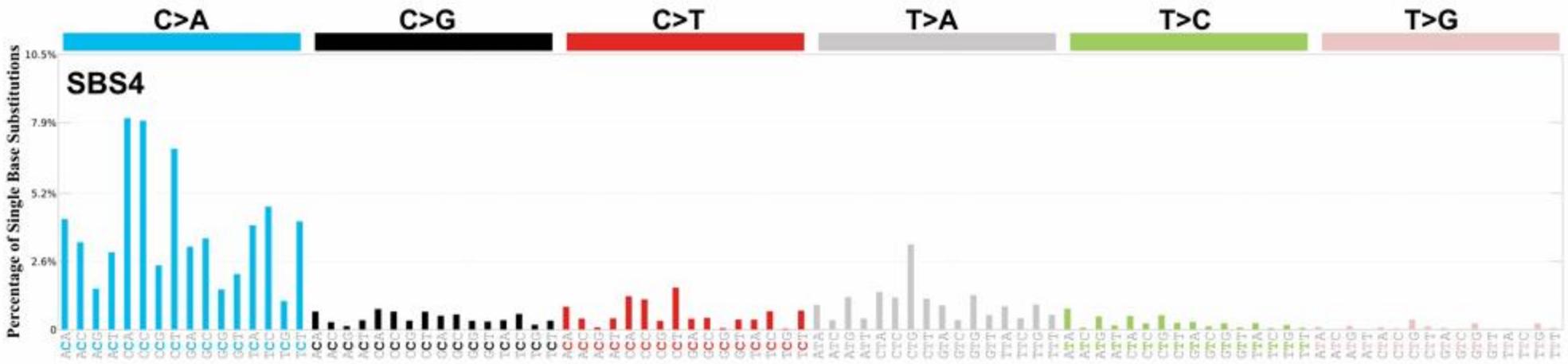
Adding 5’ and 3’ adjacent bases

96 possibilities considering context

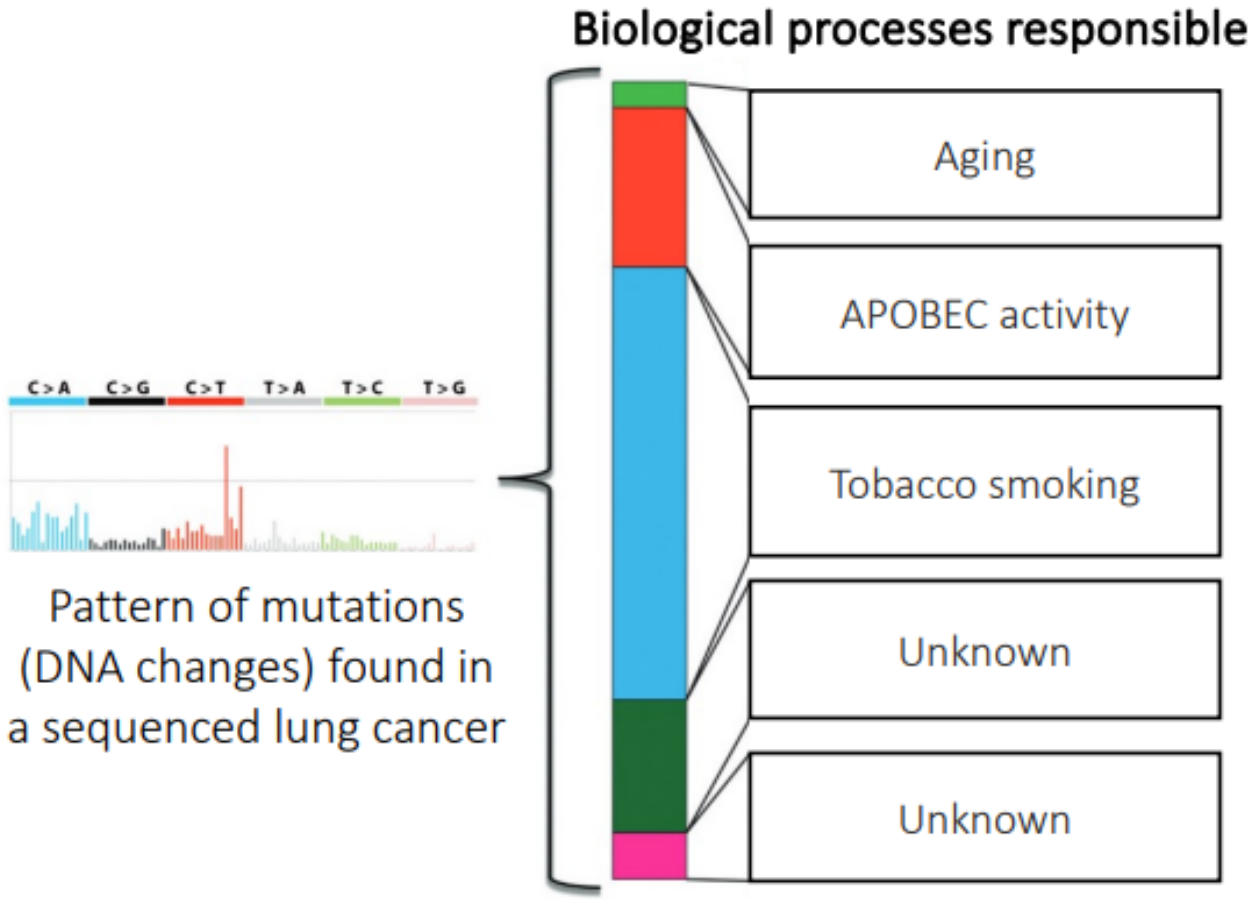


Tobacco smoking

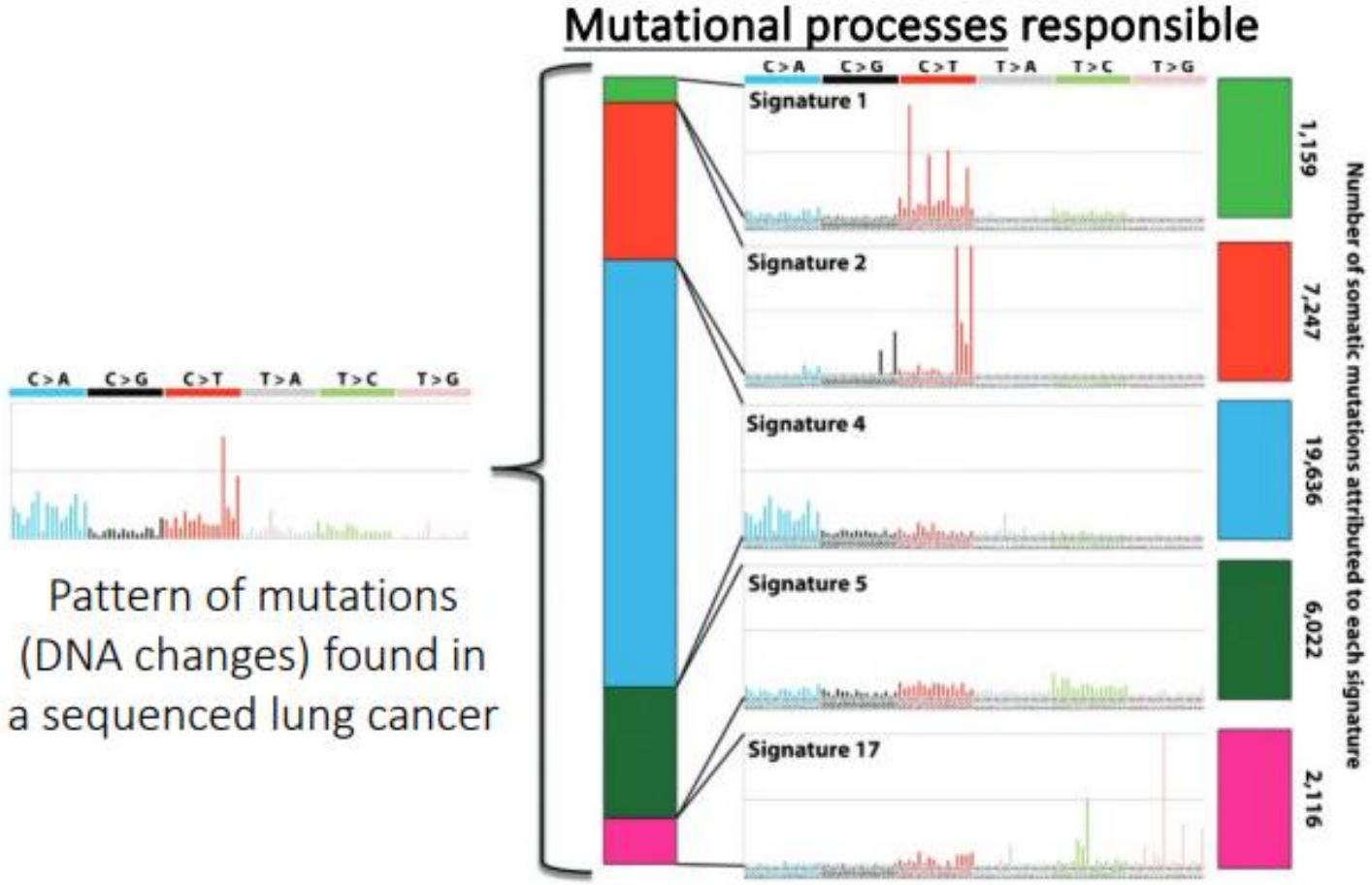
Cause of ~80% of lung cancers



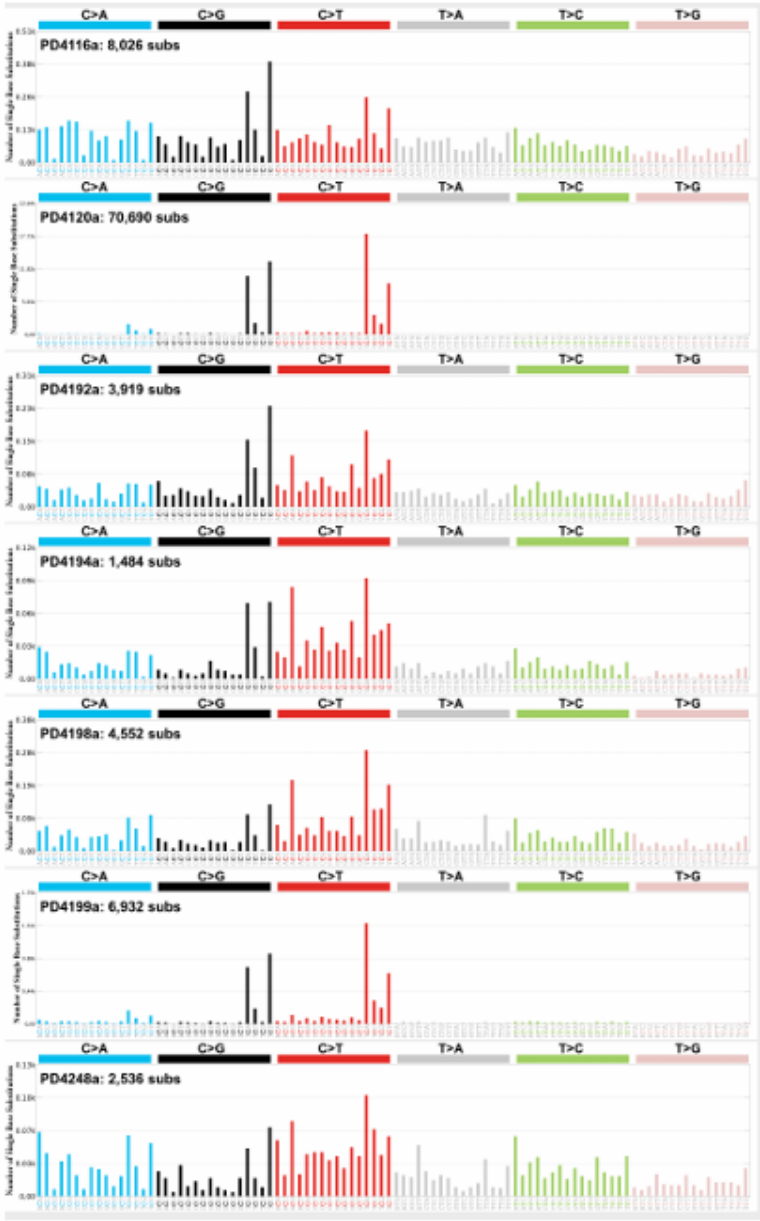
The mutational profile of a tumour is the sum of all the biological processes acting on the cells in the tumour



The mutational profile of a tumour is the sum of all the mutational processes acting on the cells in the tumour

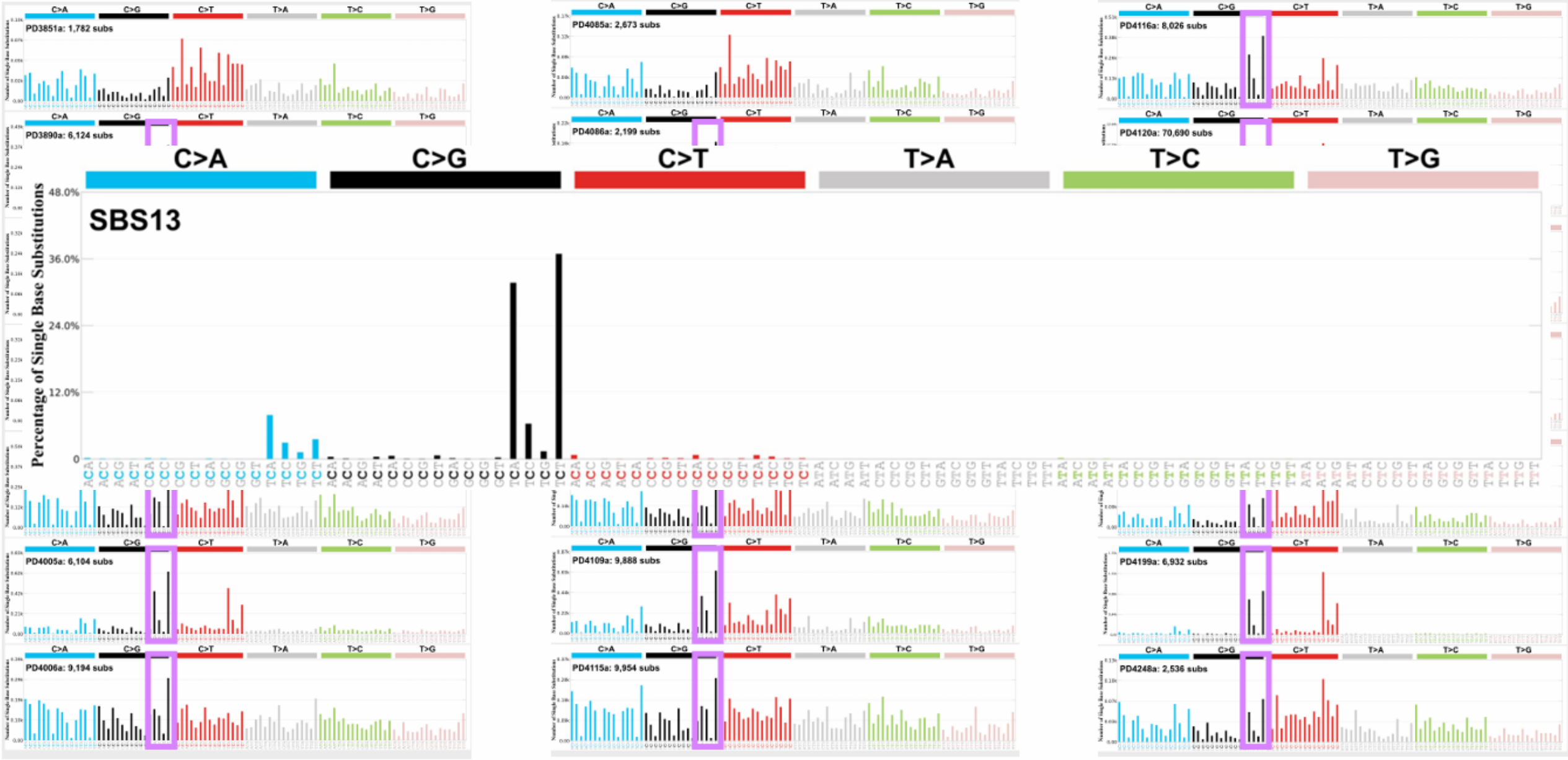


Patterns of the 96 mutation classes across individual breast tumours

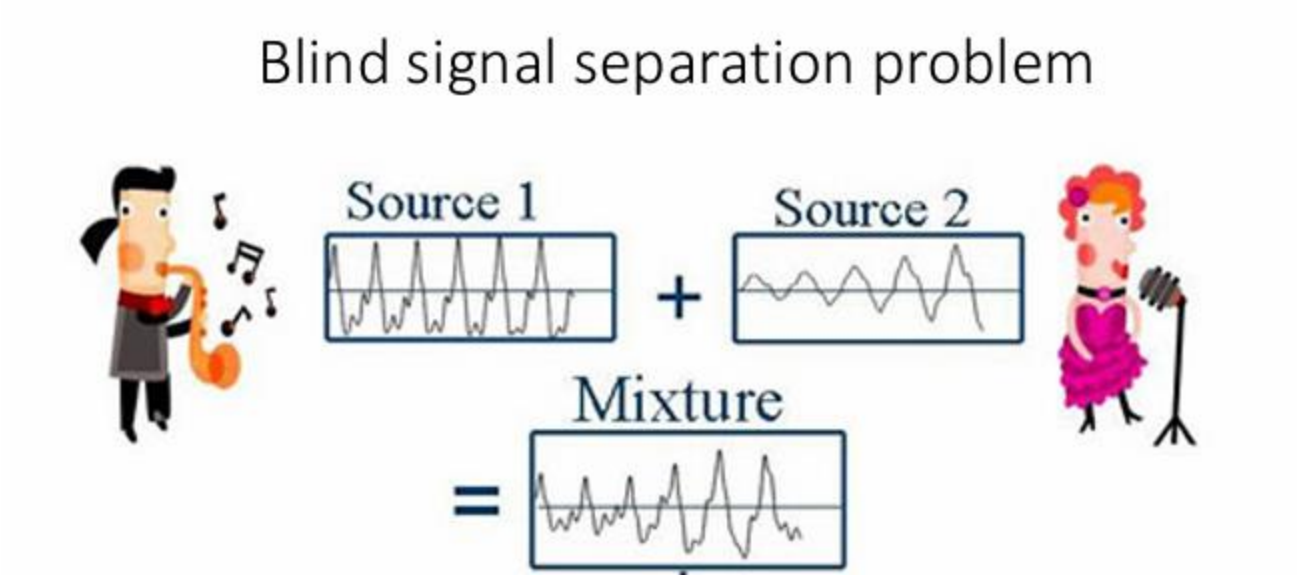


Patterns of the 96 mutation classes across individual breast tumours

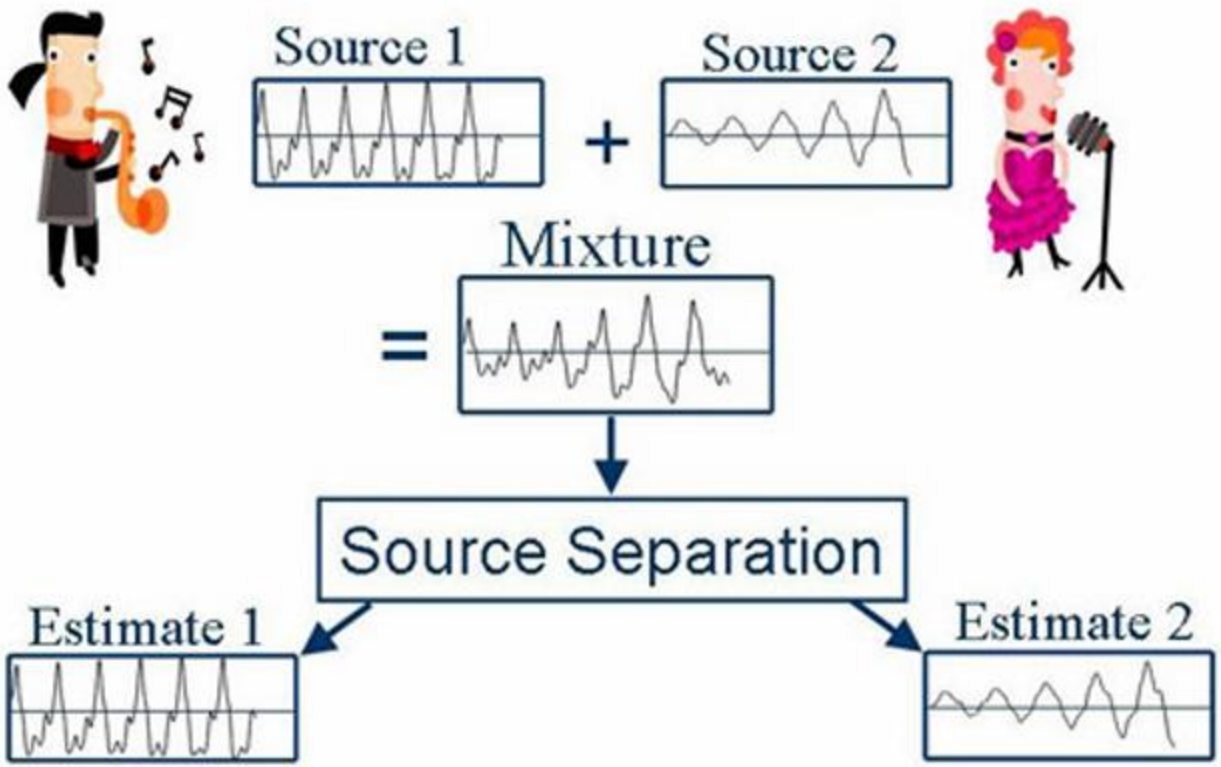
Patterns of the 96 mutation classes across individual breast tumours

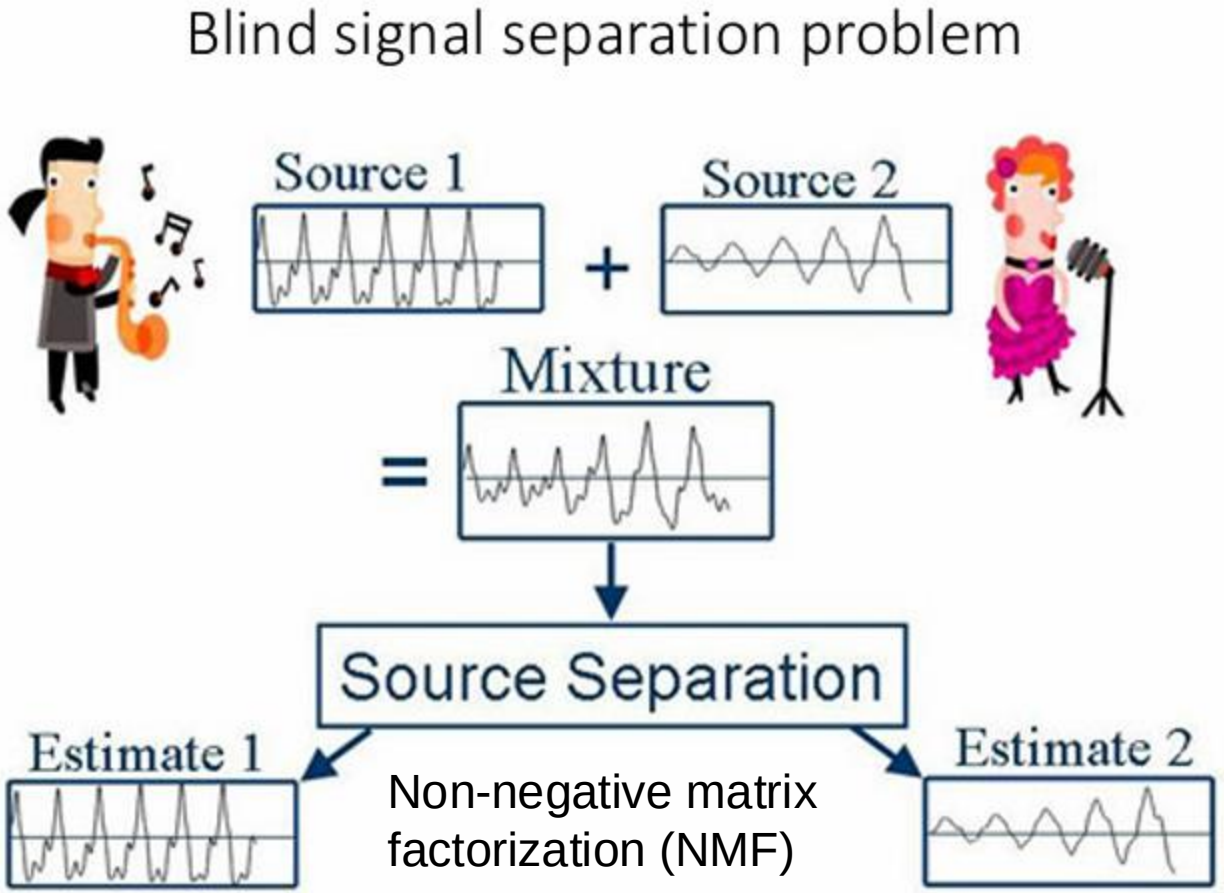


Isolating the contribution of individual signatures to an overall pattern = blind signal separation problem

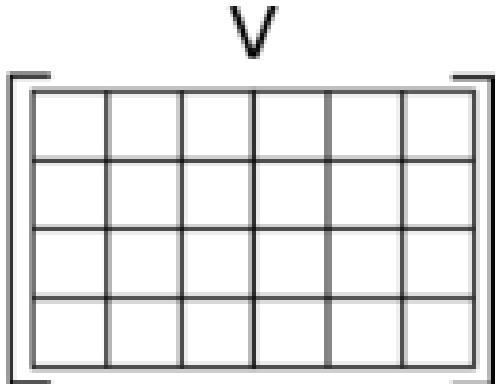


Blind signal separation problem



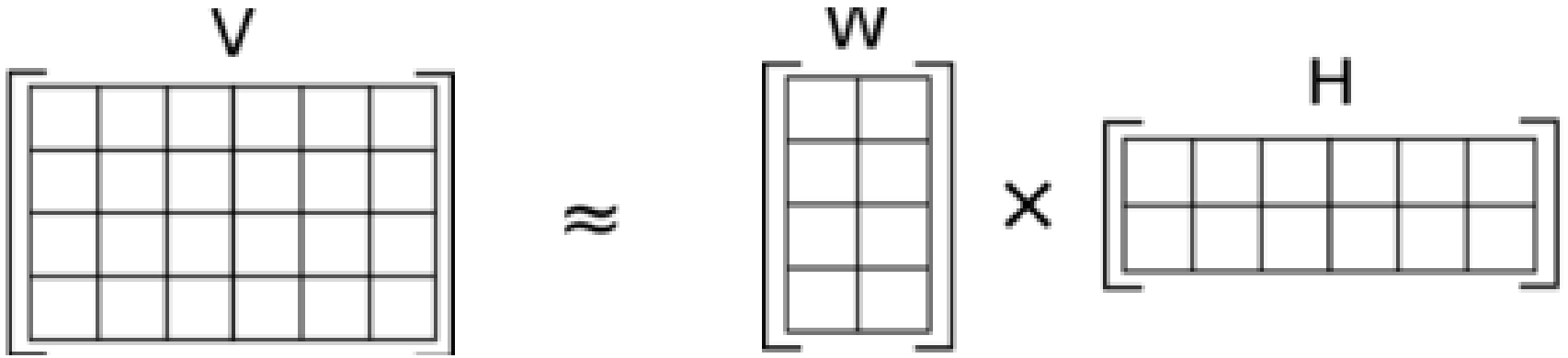

Blind signal separation problem
Source 1
Source 2
+
Mixture
=
Source Separation
Estimate 1
Estimate 2
Non-negative matrix factorization (NMF)

Non-negative matrix factorization (NMF)



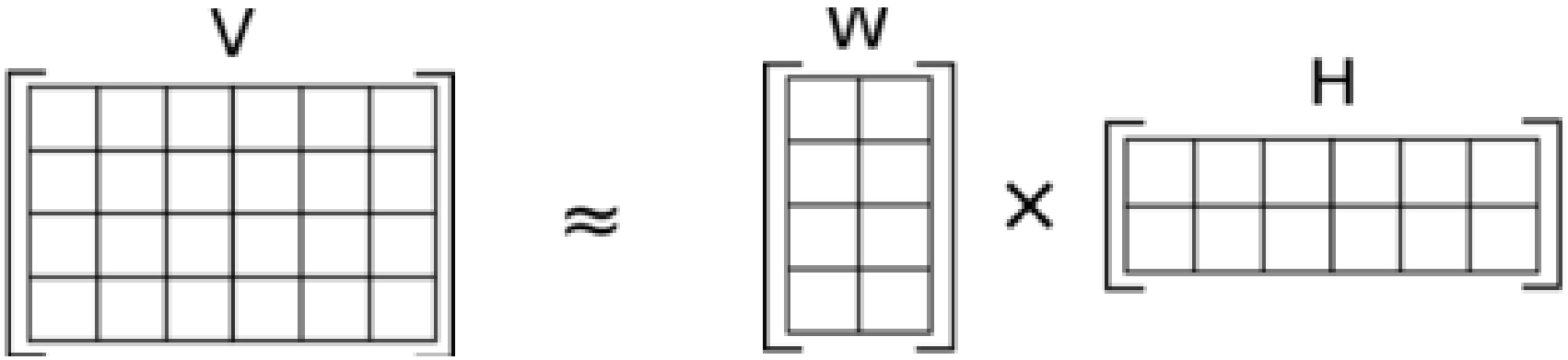
Non-negative matrix factorization (NMF)

Factorization/Decomposition



Constrains: W and H must contain positive values

Non-negative matrix factorization (NMF)



NMF is often implemented as a multi-step optimization/selection process:

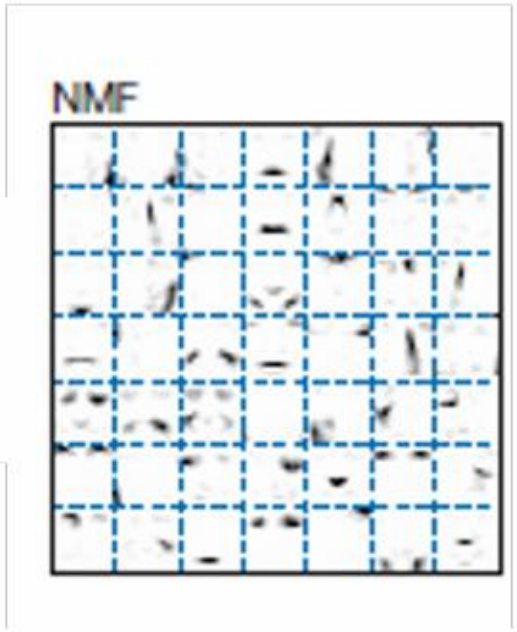
1. Find an initial solution for W and H (options include complete randomization, SVD, k-means clustering)
2. Compare the solution to the actual data to quantify error
3. Iterate on the solution to minimize the error value

NMF to learn the components of human faces (facial recognition)

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*



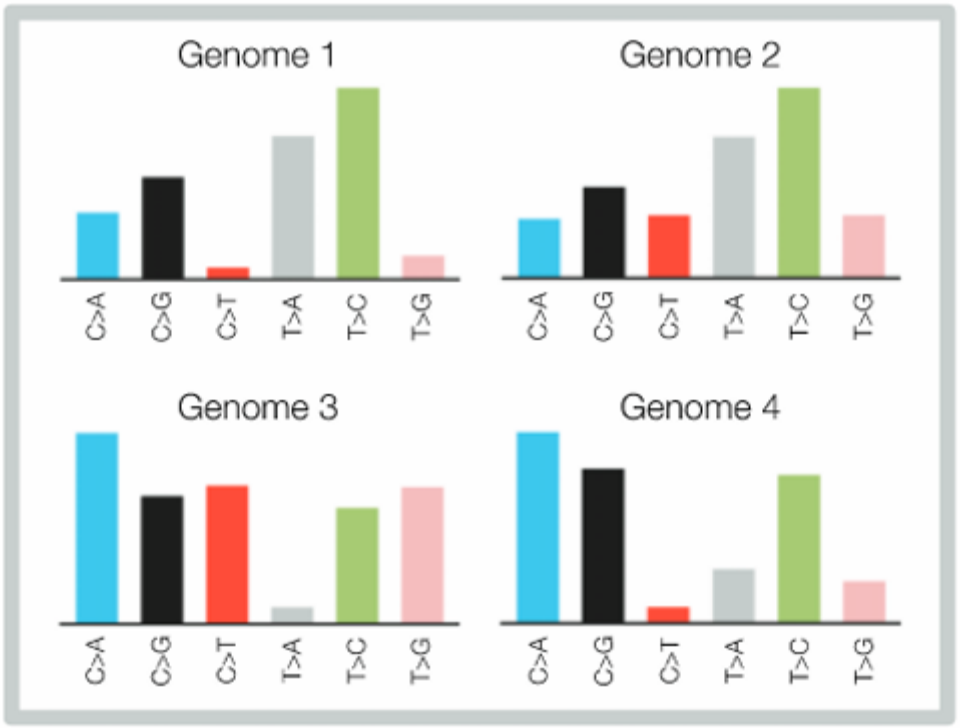
NMF to isolate different instruments from an audio file

Audio examples:

1. d-kitamura.net/demo-defNMF_en.html [Daichi Kitamura]
2. Audio Source Separation. Unleashing the Power of Non-Negative Matrix Factorization: A Python Implementation.

NMF to isolate SNV mutational signatures from a collection of genomes

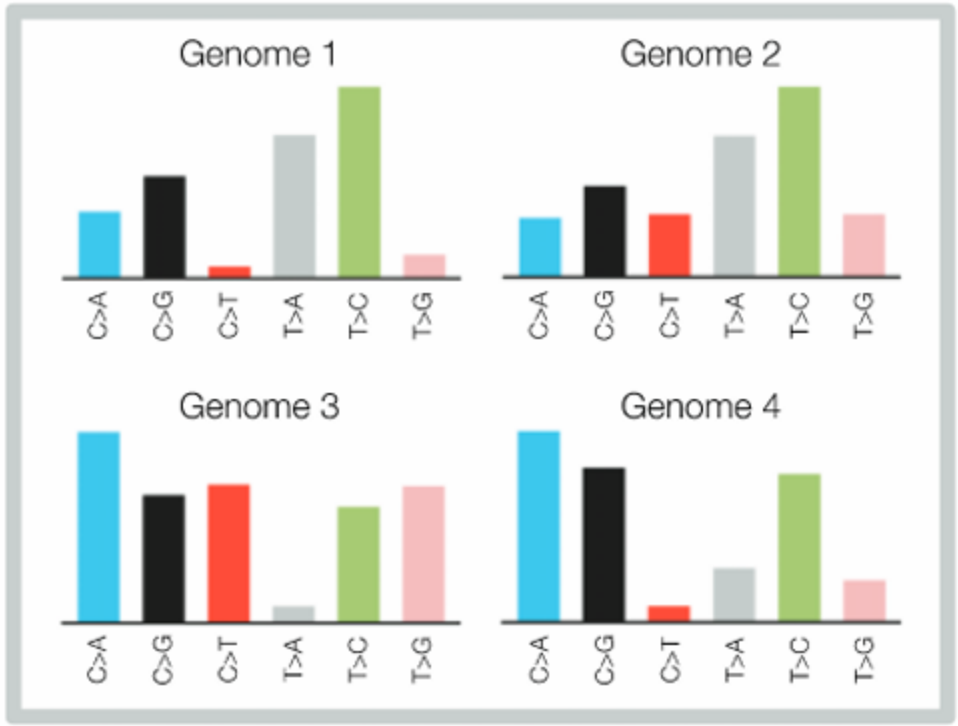
V



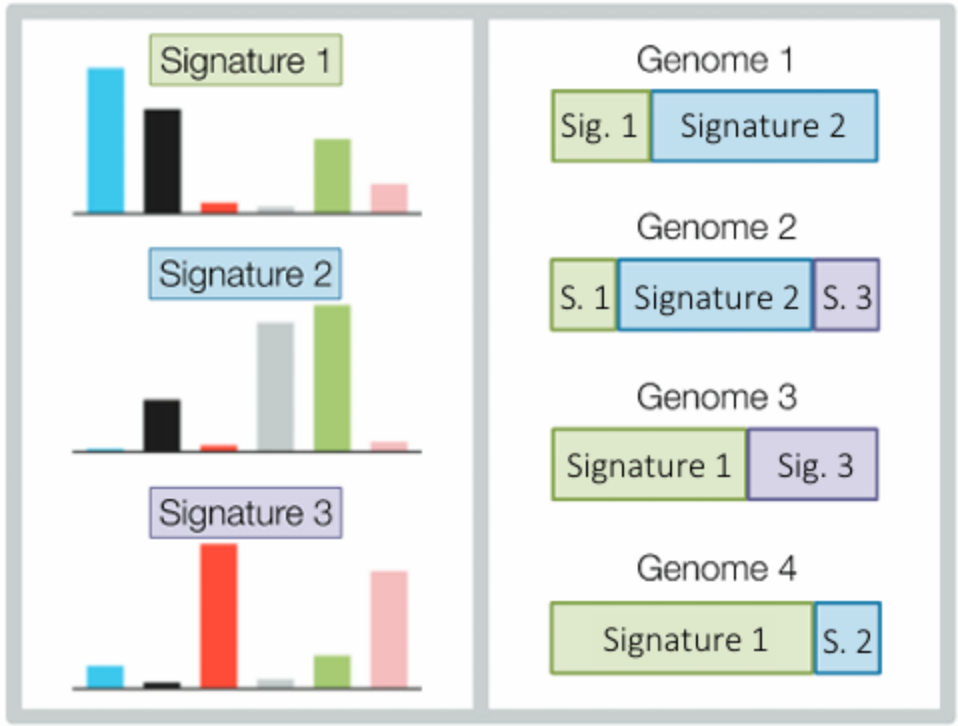
Large number of samples

NMF to isolate SNV mutational signatures from a collection of genomes

V = W x H



Large number of samples



Set of consensus signatures

Considerations for mutational signature experiments

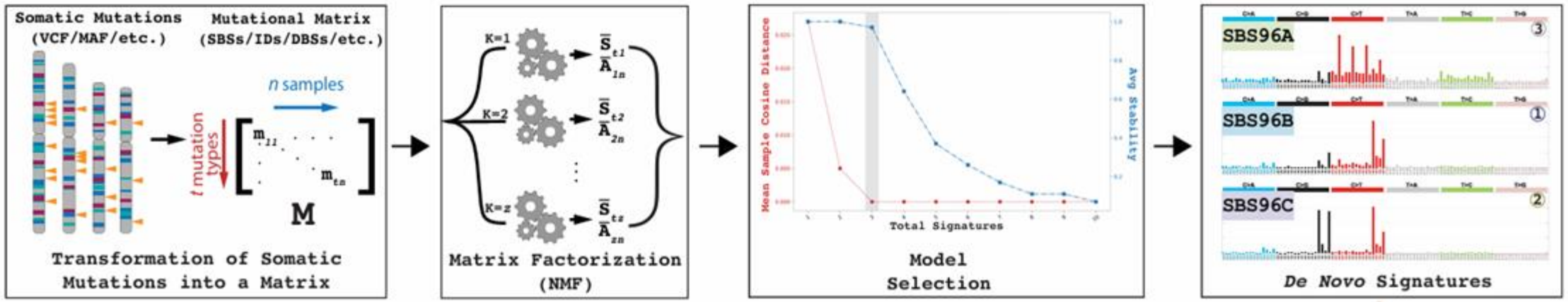
Discuss:

- What kind of sequencing strategy will give you the best data for mutational signatures, and why?
- What other things should you consider before doing a de novo mutational signature analysis? [Sample/cohort size, sequencing depth, cancer type?]

SigProfilerExtractor is a commonly used implementation of NMF to extract mutational signatures

Step 1: Transform VCFs/MAFs into mutational matrices (classification/categorization)

Step 2: Do NMF



Step 3: Select best NMF solution by minimizing k number of clusters that are both stable and have small intracluster distance

The best NMF solution is your set of signatures

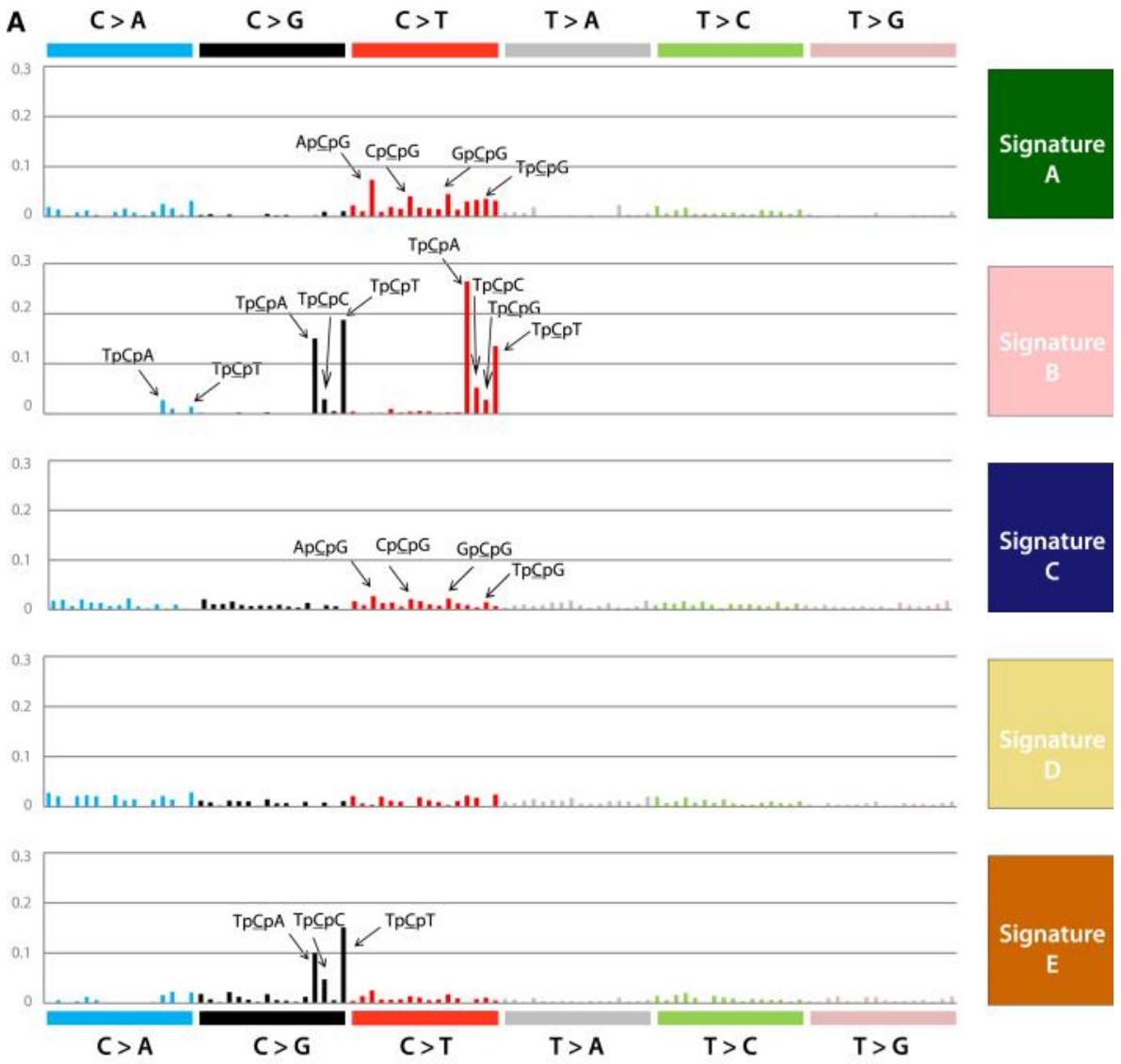
Many other tools and algorithms exist to extract mutational signatures

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

Early work on mutational signatures: Breast Cancer (Nik-Zainal 2012)

21 breast cancer whole genomes

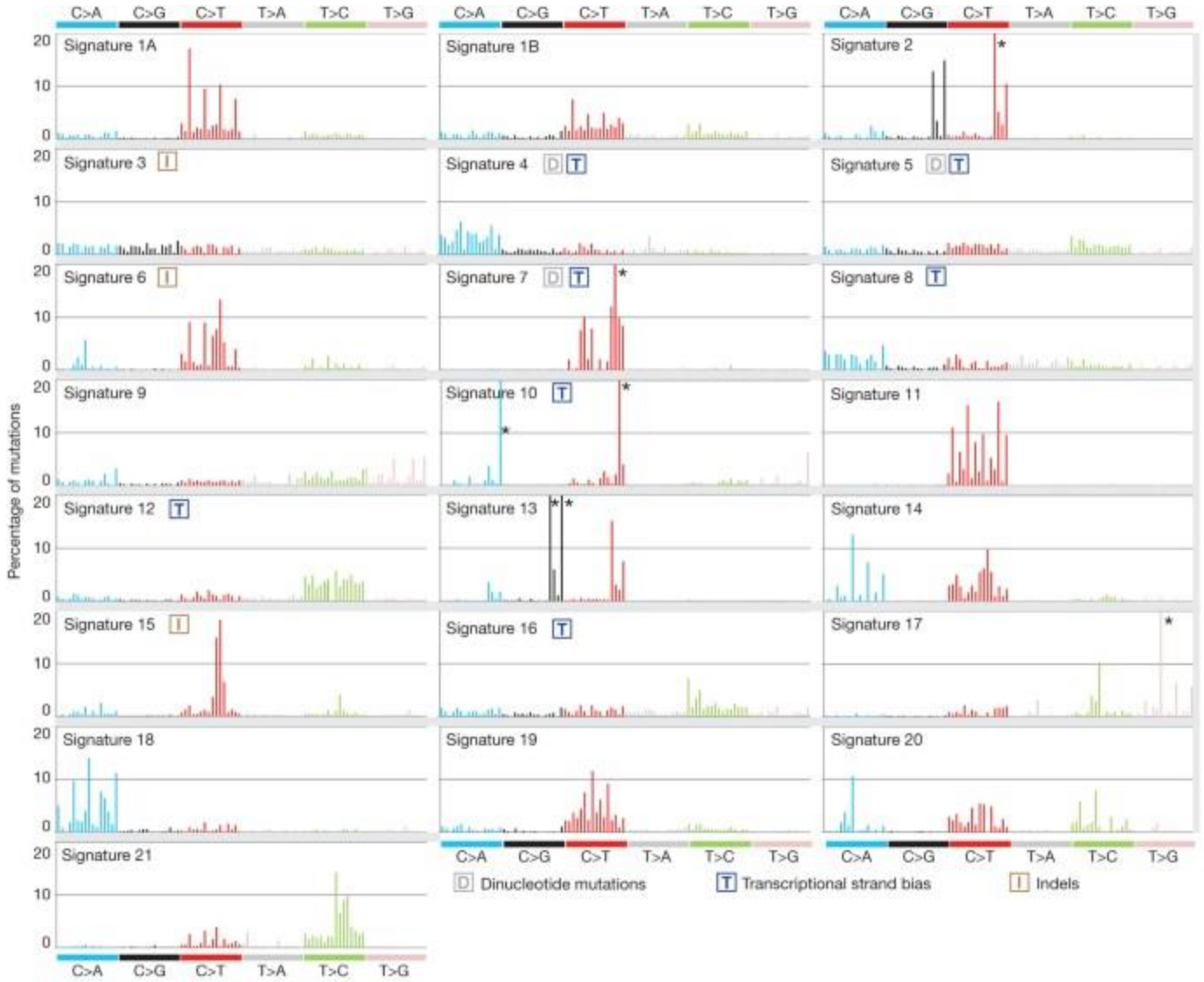
5 mutational signatures, A-E



Early work on mutational signatures: Pan-Cancer (Alexandrov 2013)

7,042 cancers (exomes + genomes from TCGA and ICGC)

21 single base substitution (SBS) mutational signatures



COSMIC Mutational Signatures Database

TCGA

7,042 cancers (exomes + genomes)

21 single base substitution (SBS) mutational signatures

COSMIC v1 (August 2013)

COSMIC Mutation Signature Database

TCGA

COSMIC v2 (March 2015)

- 30 SBS signatures

COSMIC Mutation Signature Database

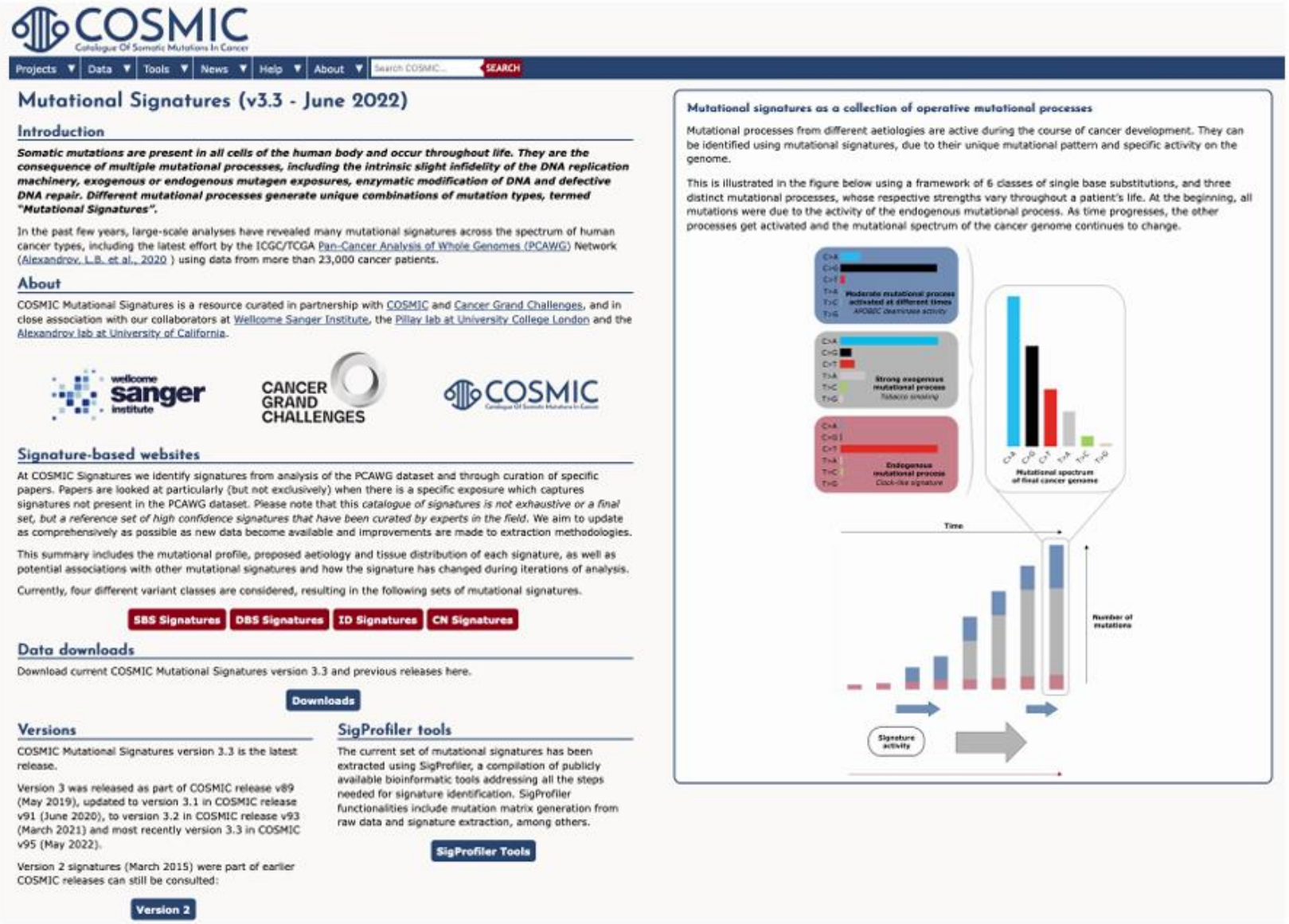
PCAWG
PanCancer Analysis of WHOLE GENOMES

v3 (May 2019)

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

Whole genome sequencing data from 2658 cancers across 38 tumor types

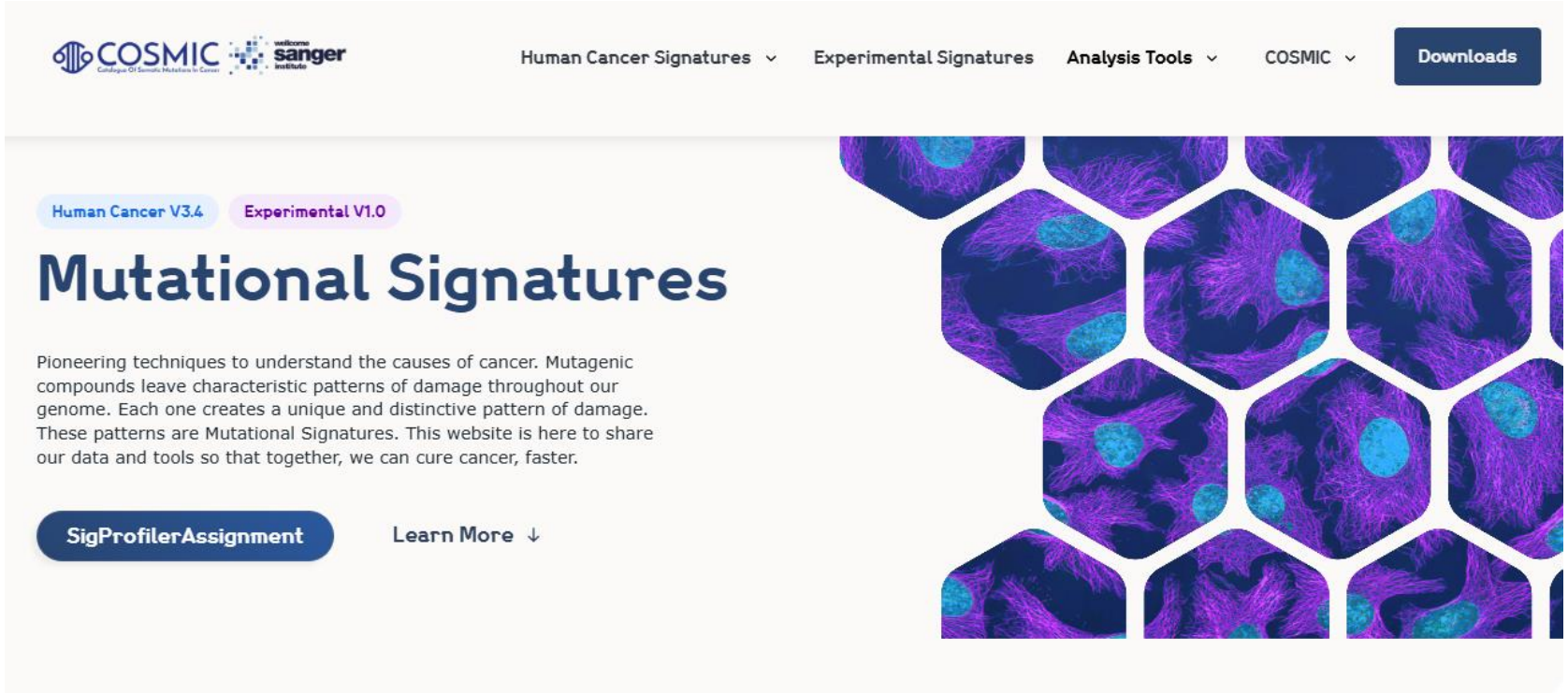
COSMIC Mutational Signatures Database



COSMIC v3.3 (June 2022)

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

COSMIC Mutational Signatures Database



October v3.4 (June 2023)

Genomics
- 99 SBS signatures
- 20 DBS signatures
- 23 ID signatures
- 25 CN signatures
- 10 SV signatures
- 5 RNA-SBS signatures

Experimental
- 140 signatures

Other Mutational Signatures Databases

MUTAGENE

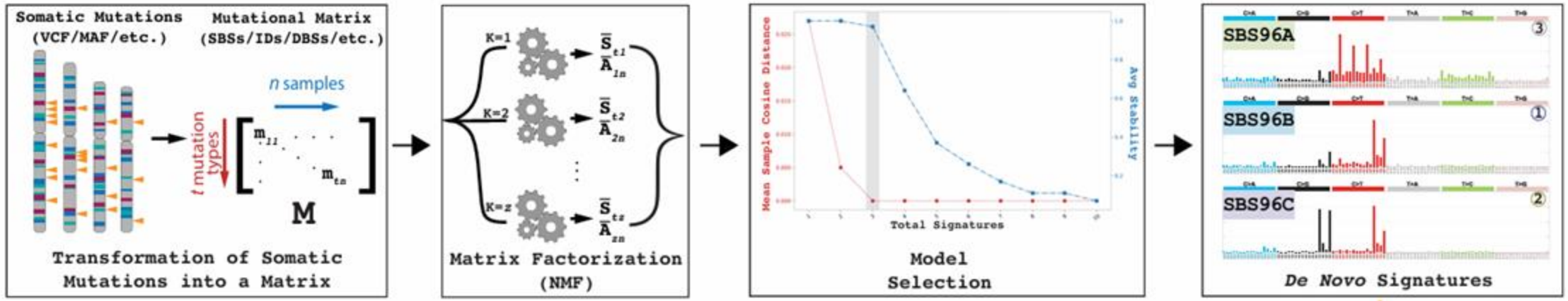
https://www.ncbi.nlm.nih.gov/research/mutagene/

signal

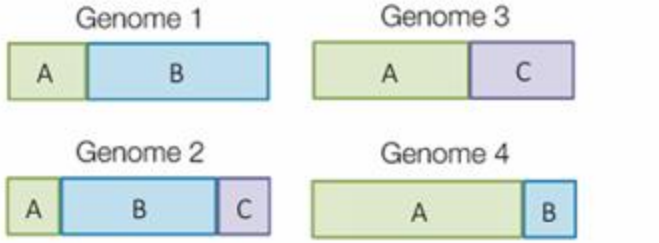
https://signal.mutationalsignatures.com/

SigProfiler is an open-source toolkit for mutational signatures analyses

SIGPROFILER Extractor

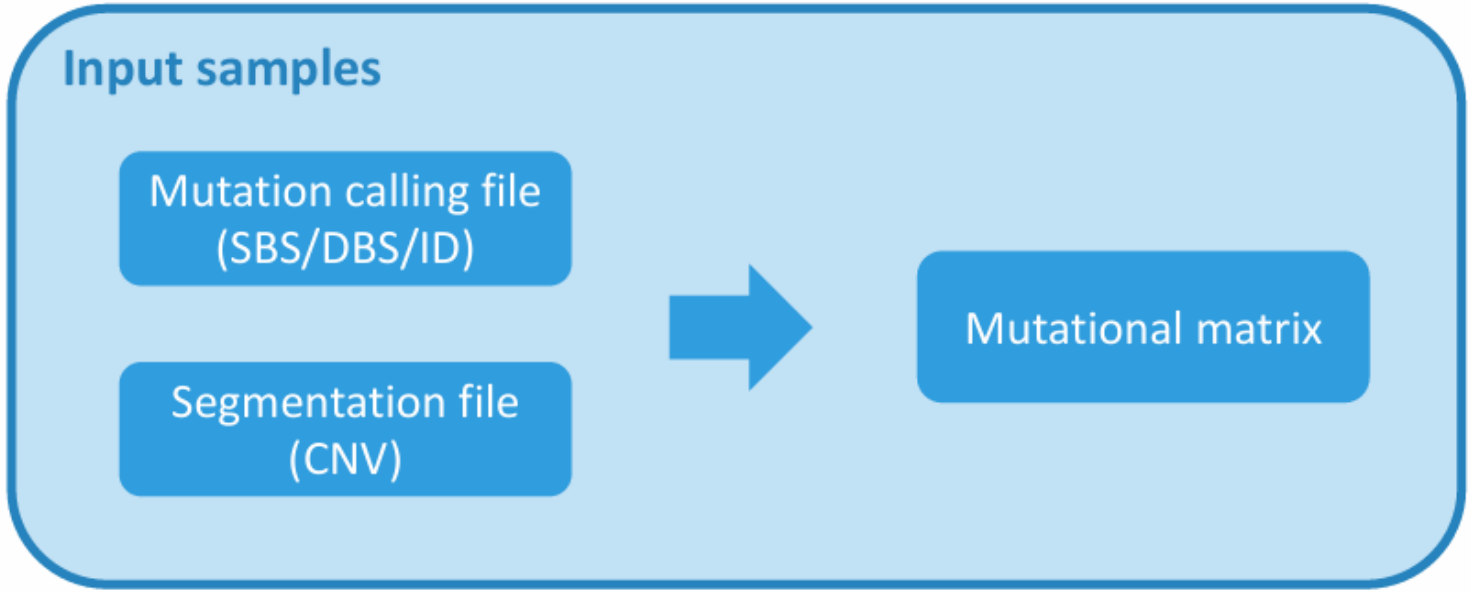


$$\underset{t\,x\,n}{\mathbf{M}} = \underset{t\,x\,k}{\mathbf{S}} \times \underset{k\,x\,n}{\mathbf{A}}$$



SIGPROFILER Assignment

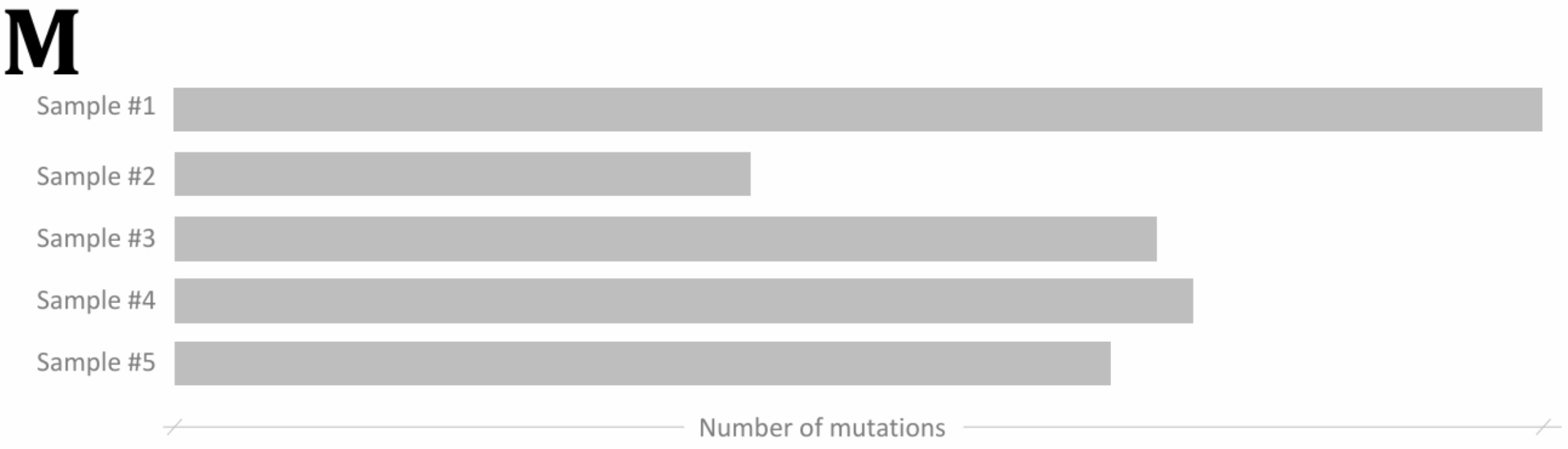
SigProfilerAssignment assigns mutational signature weights to individual samples

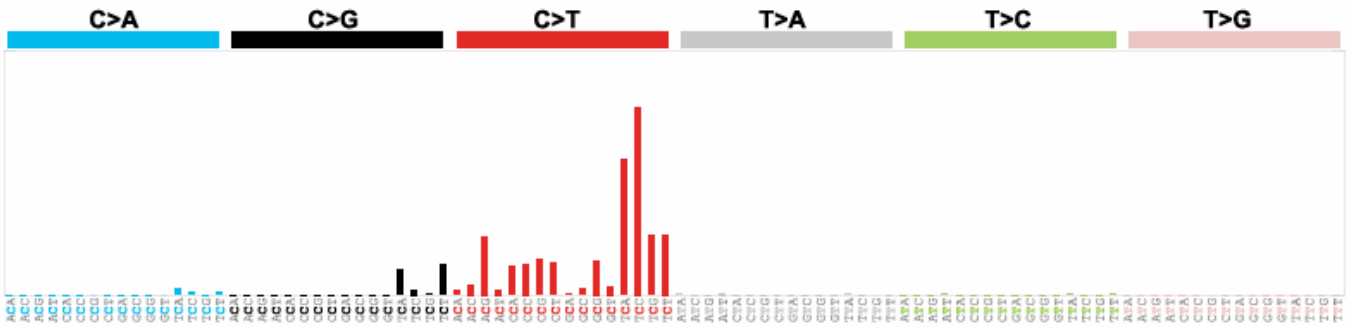


$$\underset{t\,x\,n}{M} = \underset{t\,x\,k}{S} \times \underset{k\,x\,n}{A}$$

t mutational contexts
n samples
k signatures

Assigning reference signatures to individual samples





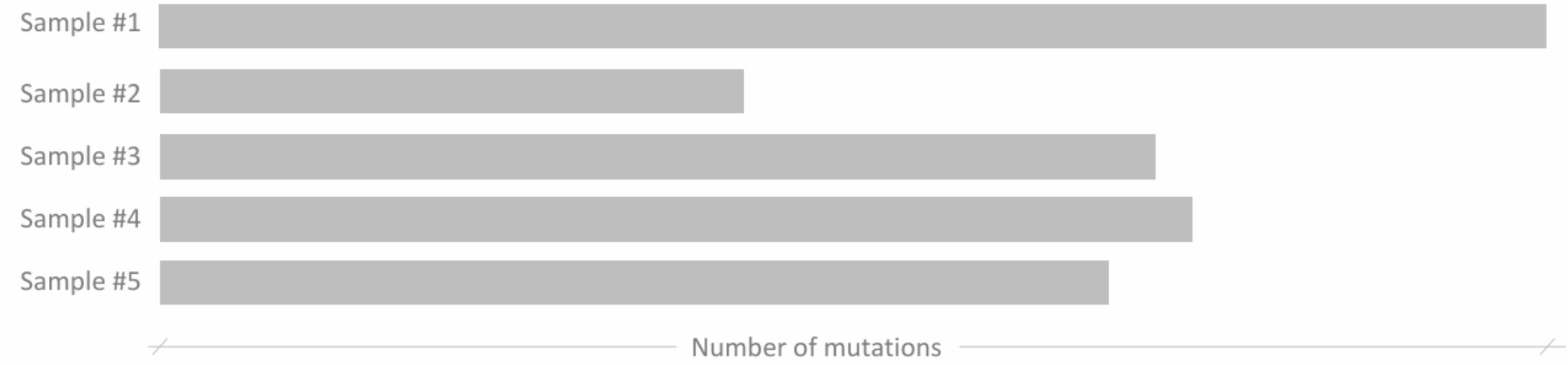
Mutational profile

#CHROM	POS	FILTER	REF	ALT
1	809687	PASS	G	C
1	819245	PASS	G	T
1	1911011	PASS	C	G
1	2112413	PASS	T	C
1	2927666	PASS	A	G
1	3359791	PASS	C	T
1	4347912	PASS	G	A
1	4961889	PASS	G	C
1	5949138	PASS	C	T
1	7806339	PASS	A	C
1	9648435	PASS	G	A
1	9705025	PASS	C	T

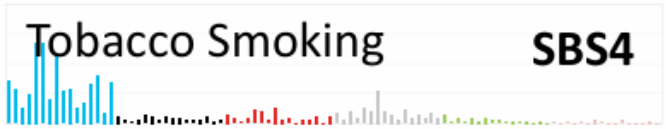
Mutation calling file

Assigning reference signatures to individual samples

M

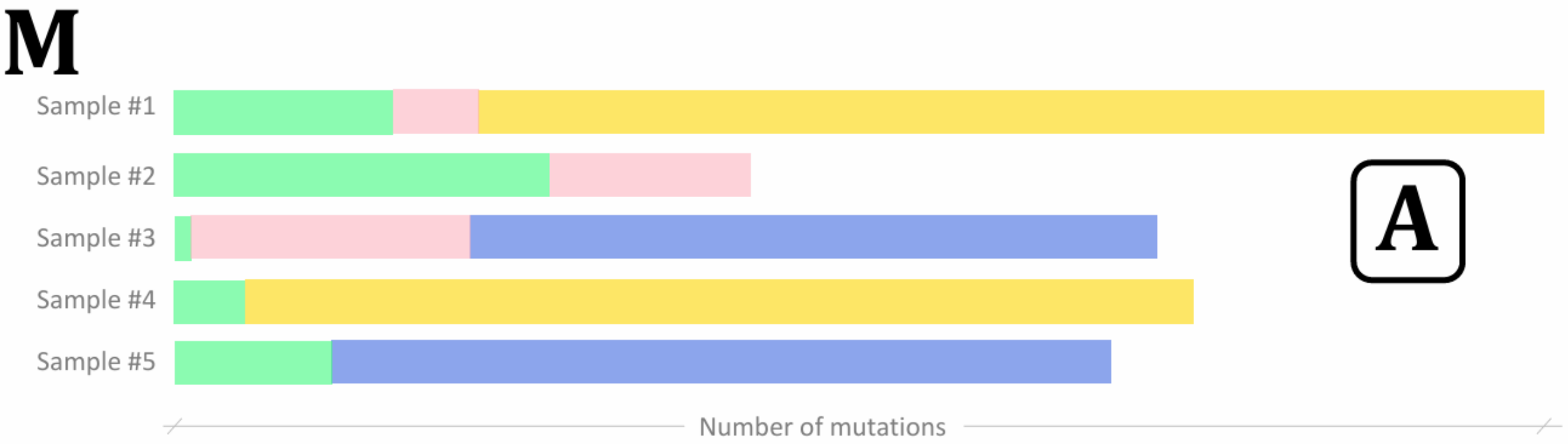


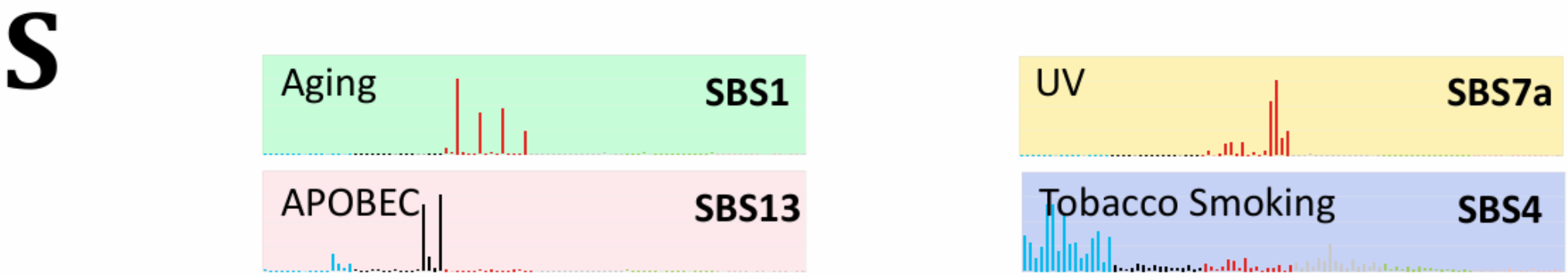
S



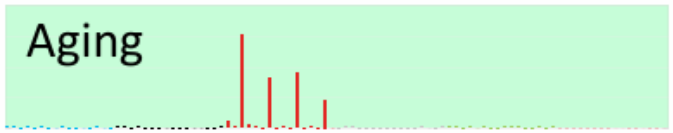
COSMIC
Catalogue Of Somatic Mutations In Cancer

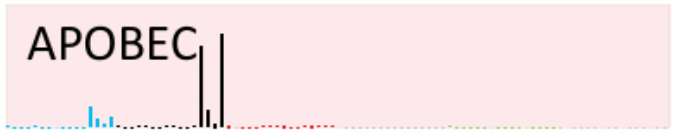
Assigning reference signatures to individual samples

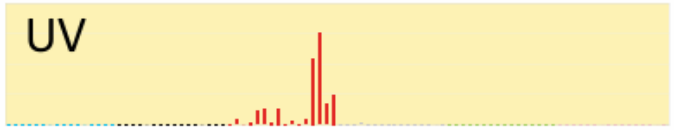


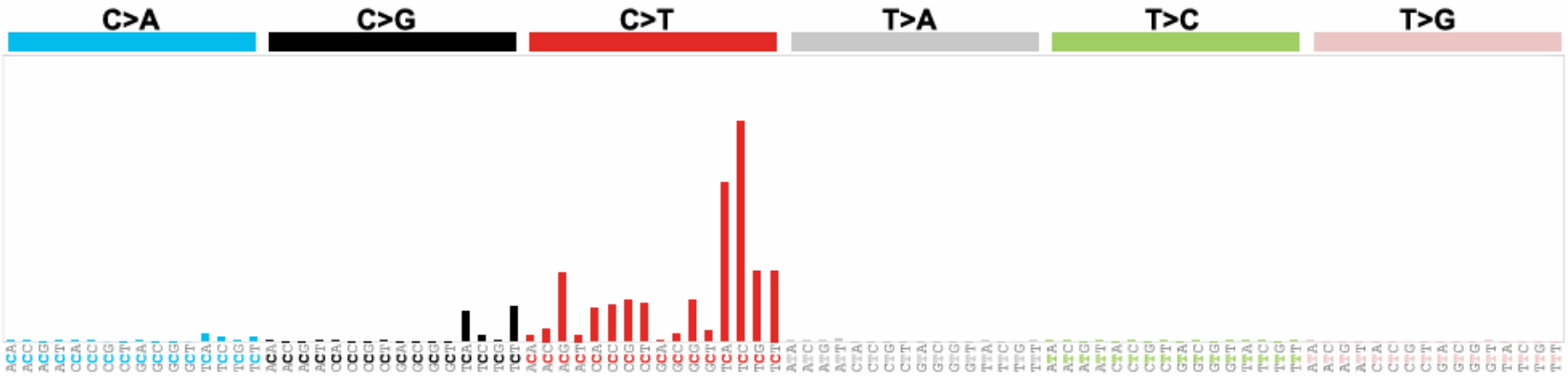


Assigning reference signatures to individual mutations







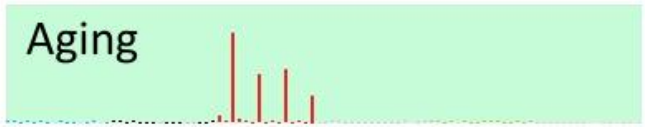


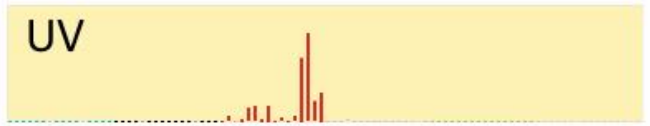
CHROM	POS	REF	ALT	Mut. Context
1	221436661	G	A	A[C>T]G
3	178936091	G	A	T[C>T]A
4	119785075	C	G	T[C>G]T
6	162294115	C	T	T[C>T]G
...	...	...	...	...

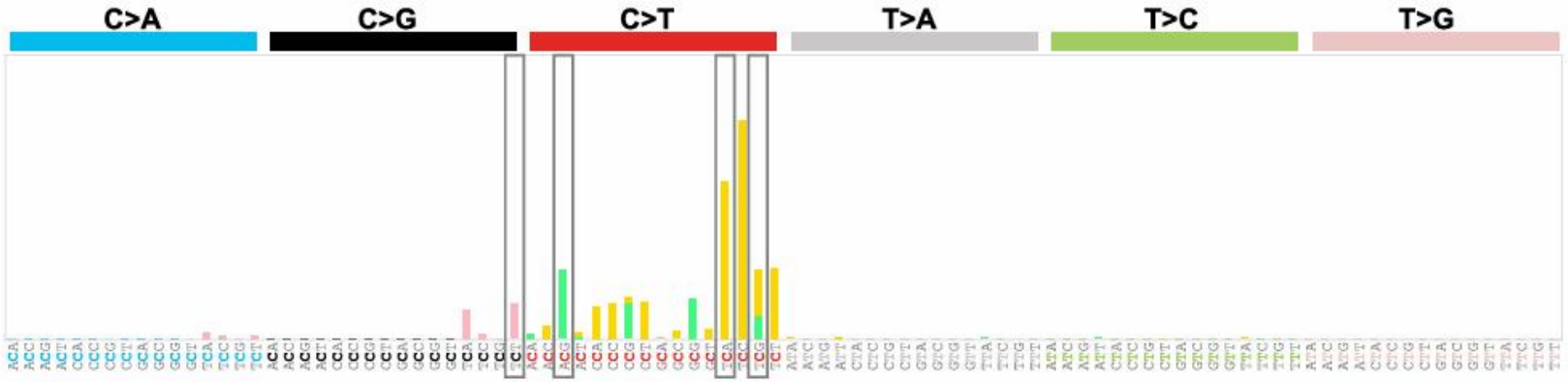
Assigning reference signatures to individual mutations

$$P(Signature \mid Mut.context) = \frac{\overbrace{P(Mut.context \mid Signature)}^{\mathbf{S}} \times \overbrace{P(Signature)}^{\mathbf{A}}}{\underbrace{P(Mut.context)}_{\mathbf{M}}}$$

Assigning reference signatures to individual mutations

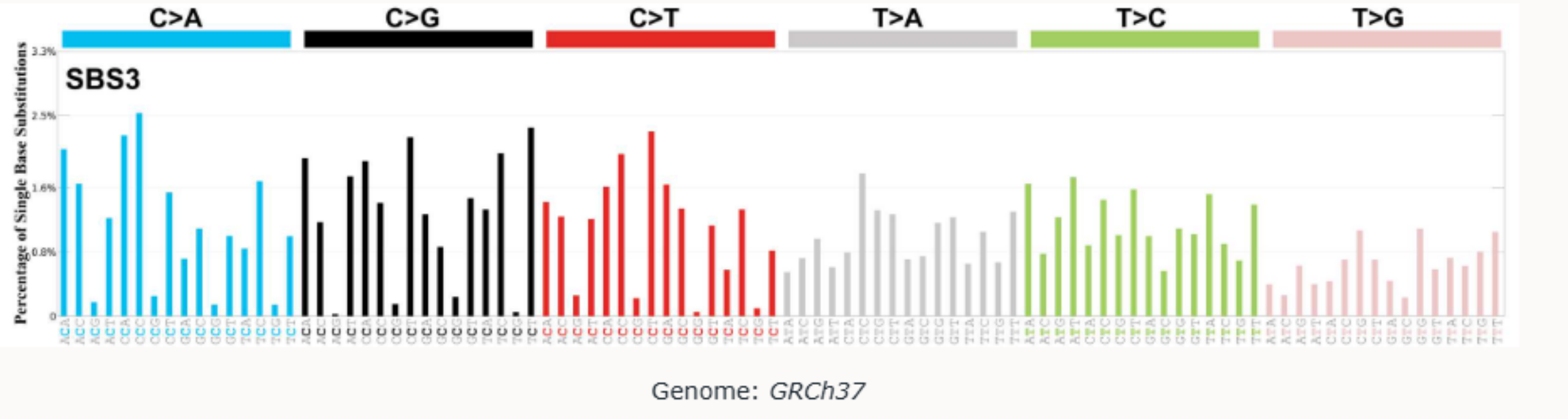






CHROM	POS	REF	ALT	Most Prob. Signature	Prob.
1	221436661	G	A	SBS1	99.9%
3	178936091	G	A	SBS7a	99.5%
4	119785075	C	G	SBS13	98.6%
6	162294115	C	T	SBS7a	70.1%
...	...	...	...	...	...

Warning: watch out for flat signatures!



When you have limited data, tools to assign signature can often give too much weight to flat signatures like SBS3 to match the weird shape of your data

Many other tools and algorithms exist to assign mutational signatures and their weights to individual samples

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	Díaz-Gay *et al.* 2023 bioRxiv

NNLS: non-negative least squares

Summary and takeaways

- Mutational signatures are patterns of mutations in the genome associated with a distinct biological source or process

- SBS mutational signatures were discovered by applying NMF to 96-class mutation profiles of large genomics cohorts

- The SigProfiler toolkit can conduct both de novo extraction of mutational signatures as well as assignment of known signatures to individual samples

wellcome connecting science

questions?

Please contact wellcomeconnectingscience.org for more information.