

**wellcome**  
**connecting**  
**science**

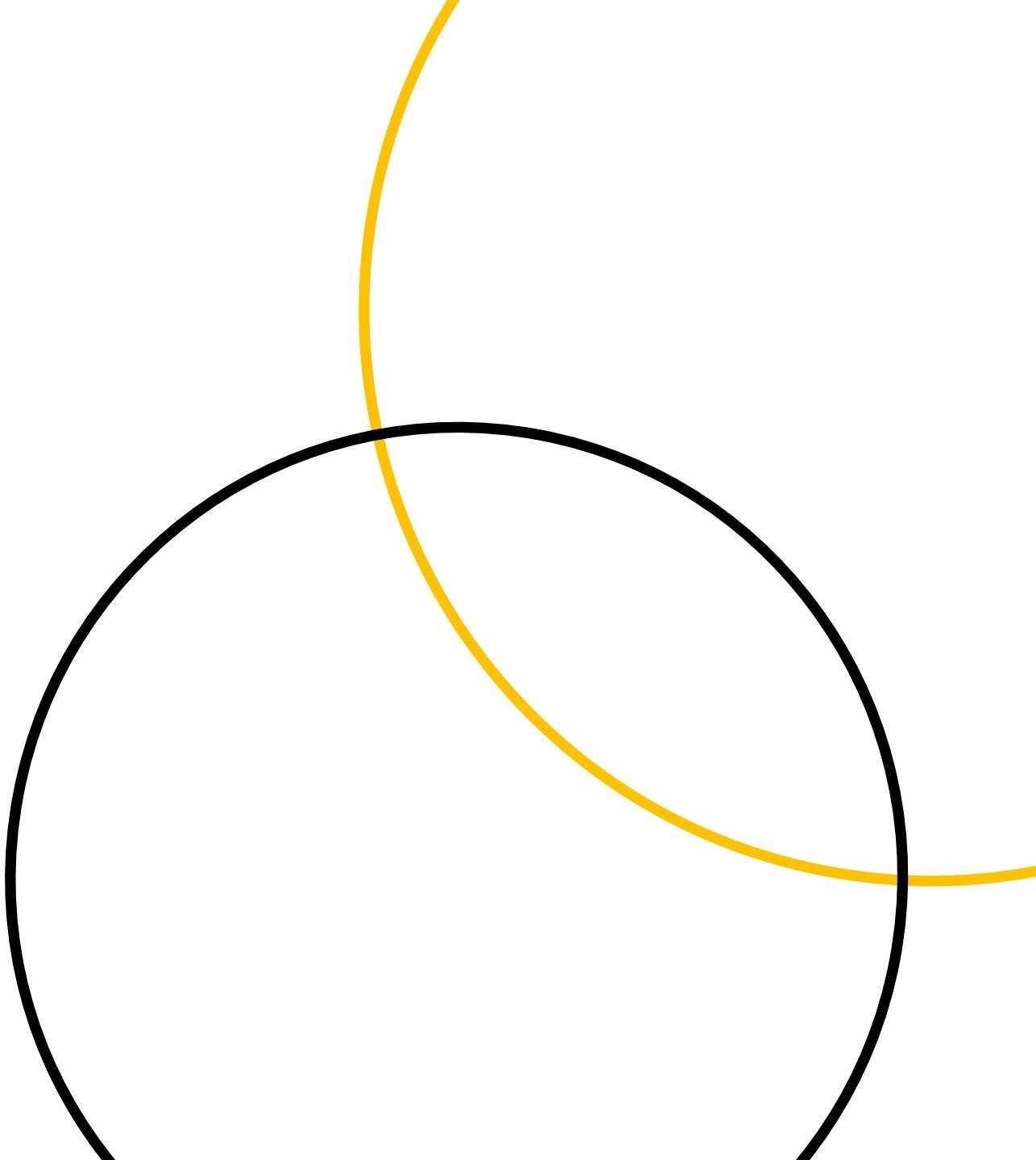
# connecting people with science

**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?

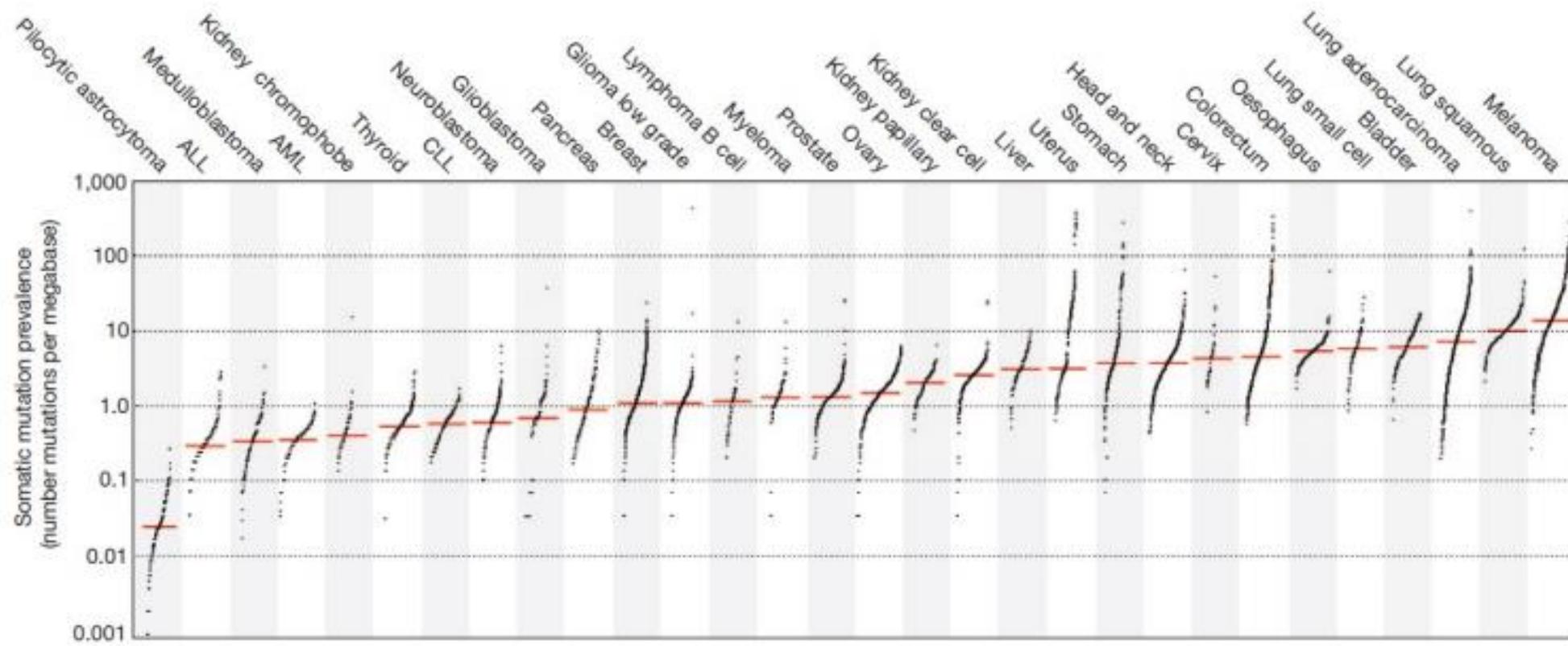
Mutational signature analysis is an attempt to determine the causal sources (aetiology) of somatic variants and mutations in an individual or set of samples

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

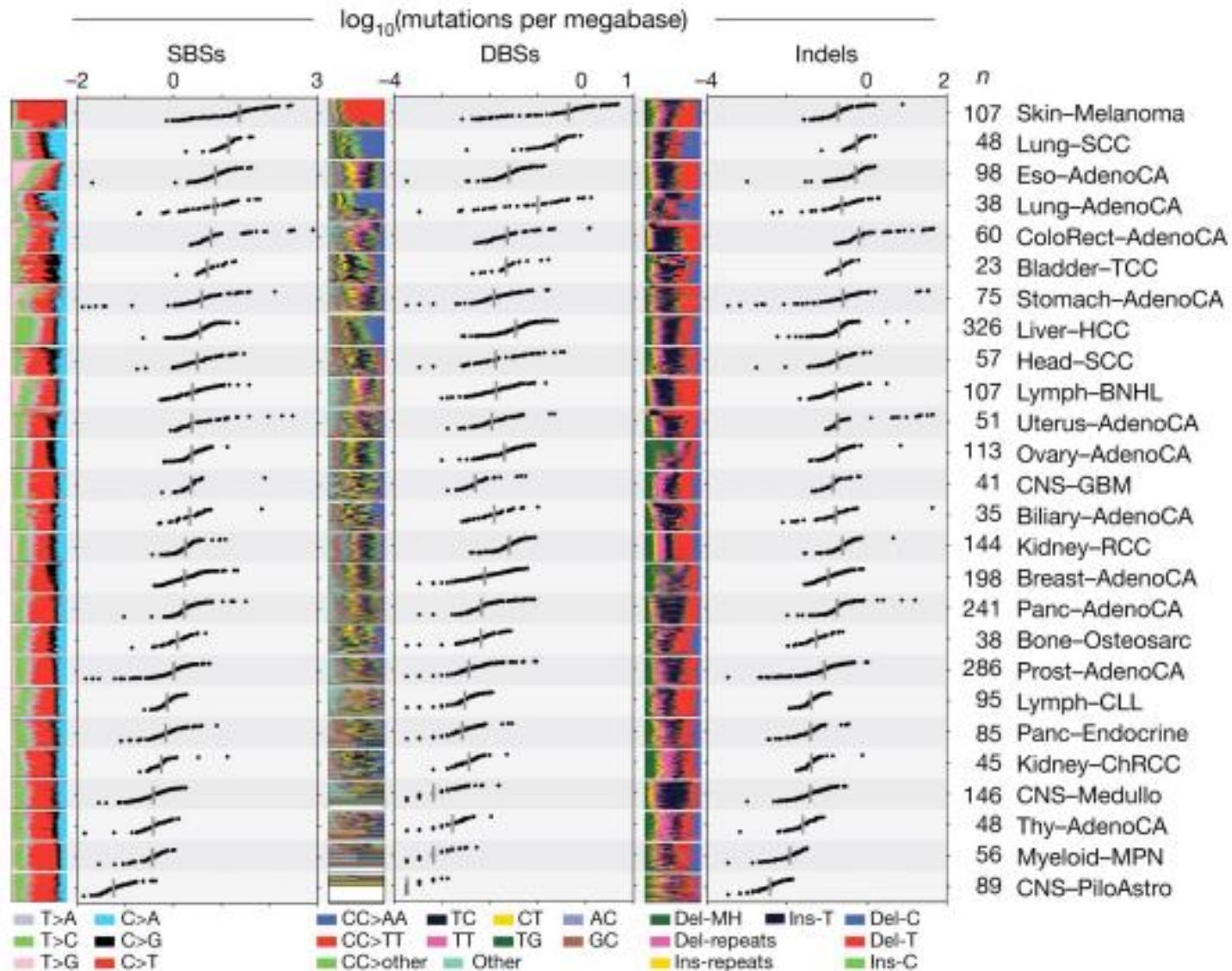
# Well-known carcinogens and carcinogenic processes

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

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

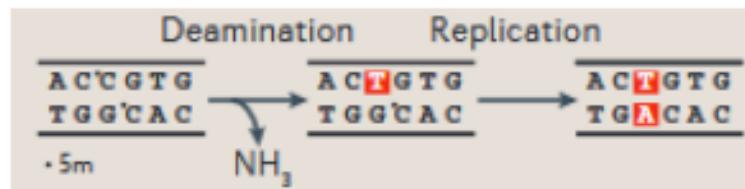
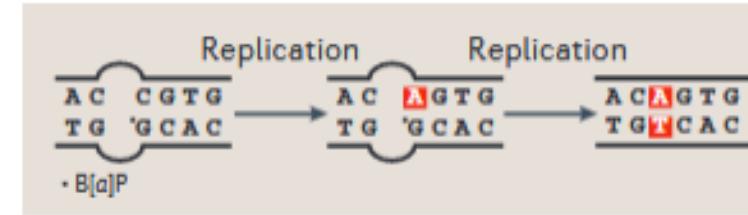


# The burden of somatic mutations is highly variable across different types of tumours



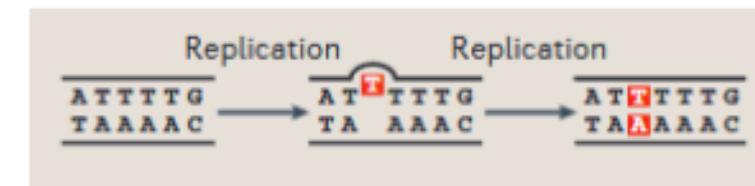
# Different sources of DNA damage have different patterns of mutations

Environmental exposures  
Tobacco smoking or chewing



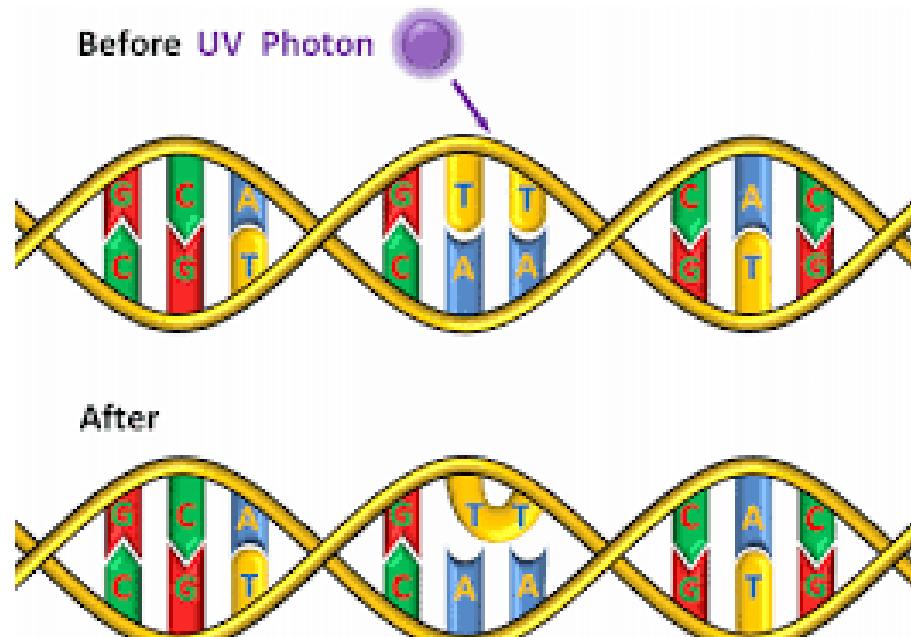
Normal cellular activities  
Spontaneous deamination of methylated cytosines

Failure in DNA replication or repair  
Aberrant mismatch repair pathway

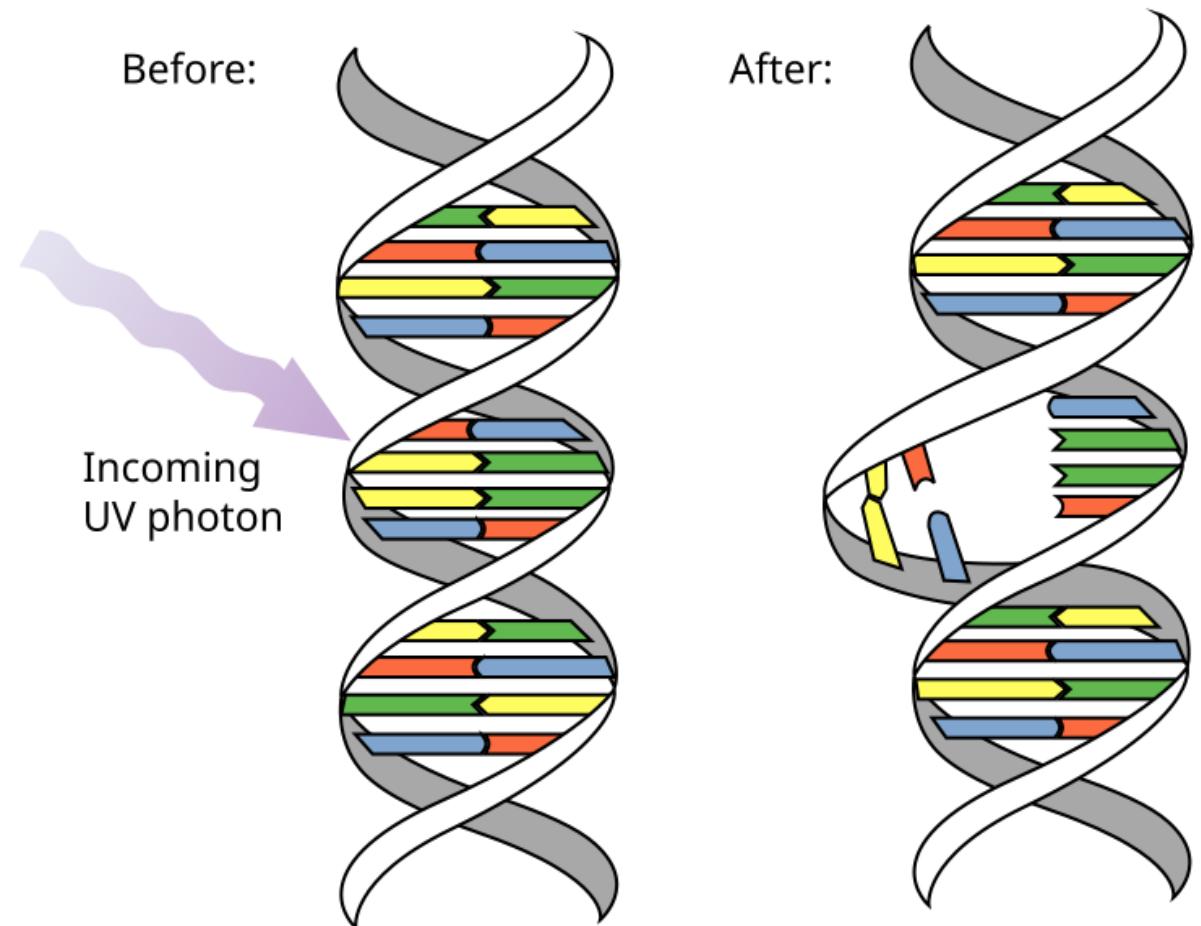


# UV radiation

Thymine dimers



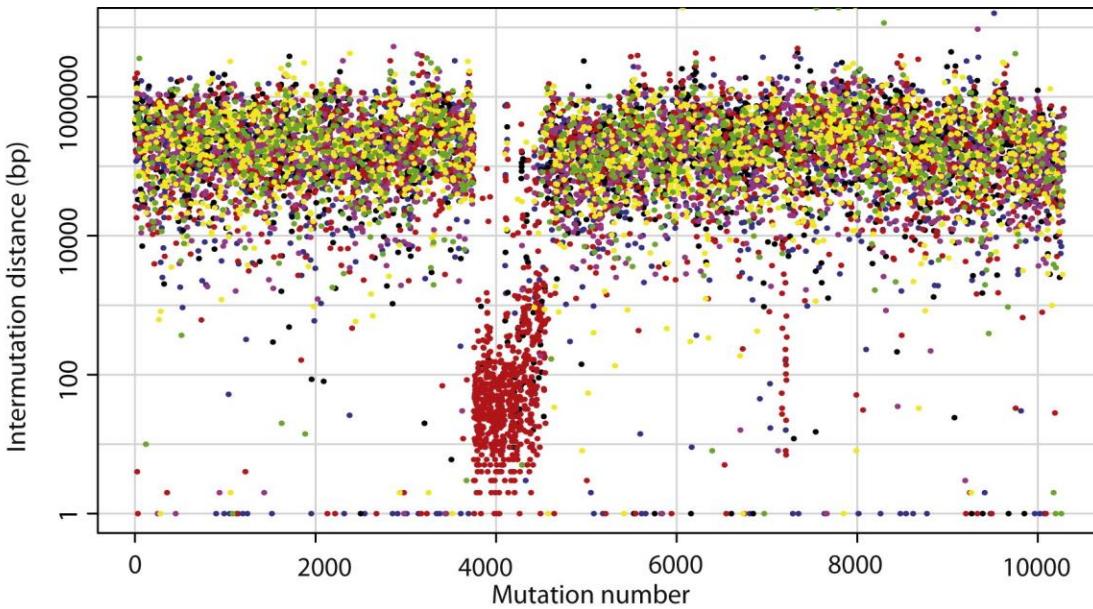
Schmid et al. 2017



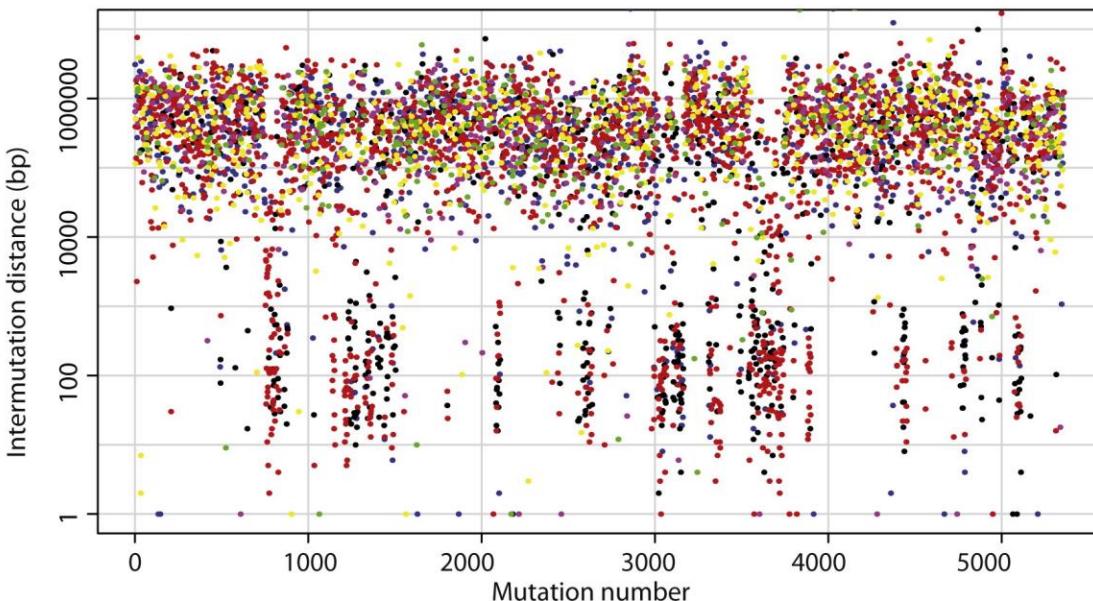
# Kataegis

Localized somatic  
hypermutation  
associated with  
APOBEC enzymes and  
TLS DNA polymerases

**A**  
PD4107a  
germline BRCA1

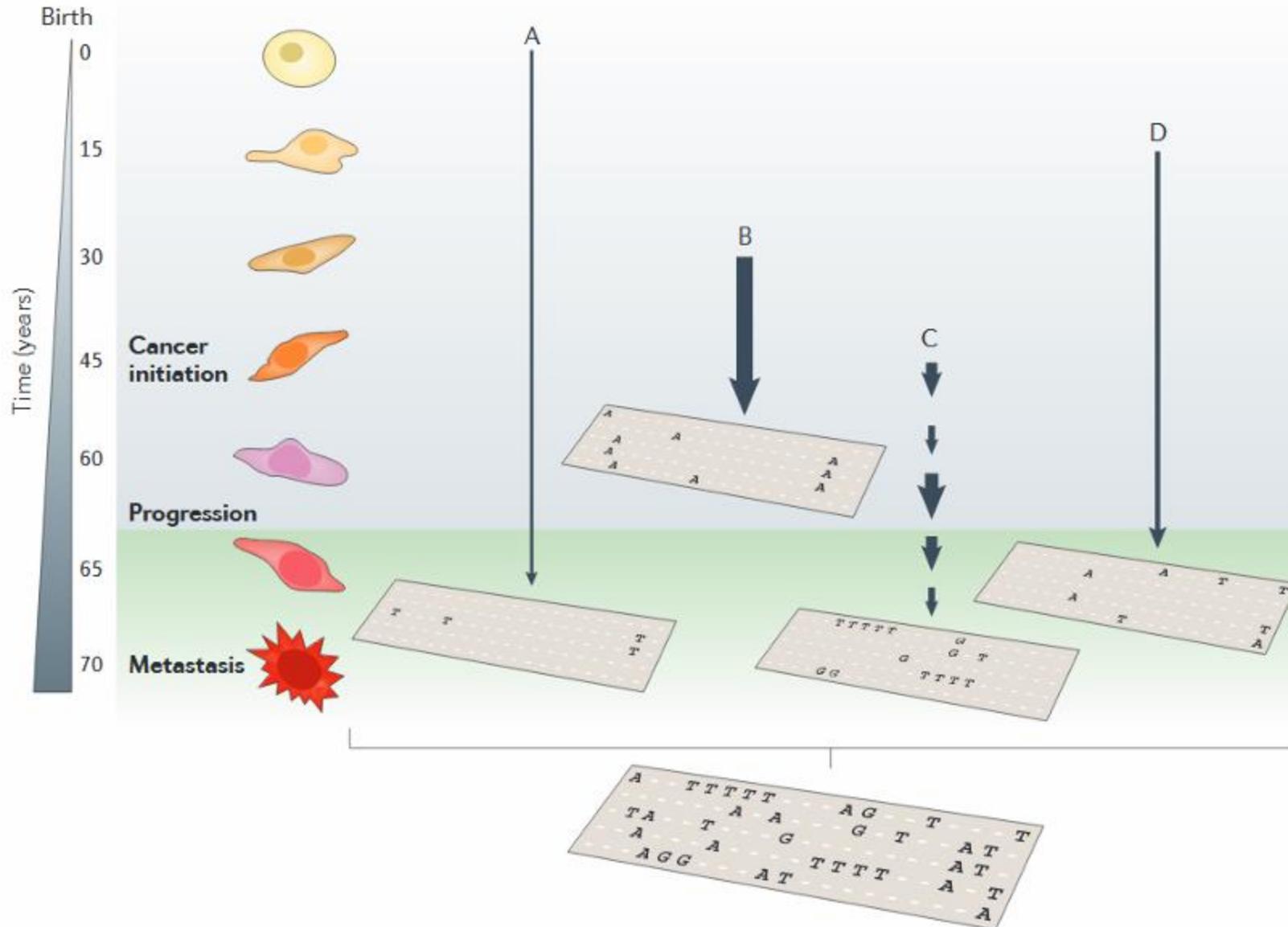


**B**  
PD4103a  
ER +ve, HER2 -ve



- C>A
- C>G
- C>T
- T>A
- T>C
- T>G

# 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

.....ATCGGGAAAT**C**GGACCCGATG.....  
                        ↓  
.....ATCGGGAAAT**T**GGACCCGATG.....

# Trinucleotide context of the single base substitution

.....ATCGGGAA**TCG**GACCCGATG.....  
                  ↓  
.....ATCGGGAA**TTG**GACCCGATG.....

# Trinucleotide context of the single base substitution

.....ATCGGGAA**TCG**GACCCGATG.....  
                  ↓  
.....ATCGGGAA**TTG**GACCCGATG.....

.....ATCGGGAA**ACG**GACCCGATG.....  
                  ↓  
.....ATCGGGAA**ATG**GACCCGATG.....

.....ATCGGGAA**ACC**GACCCGATG.....  
                  ↓  
.....ATCGGGAA**ATC**GACCCGATG.....

**C>T**

**C>A**

**C>G**

**T>A**

**T>C**

**T>G**

## **6 classes of base substitutions**

## 6 classes of base substitutions

C>T  
C>A  
C>G  
T>A  
T>C  
T>G

ACA>ATA  
ACC>ATC  
ACG>ATG  
ACT>ATT  
CCA>CTA  
CCC>CTC  
CCG>CTG  
CCT>CTT  
GCA>GTA  
GCC>GTC  
GCG>GTG  
GCT>GTT  
TCA>TTA  
TCC>TTC  
TCG>TTG  
TCT>TTT

16 different  
trinucleotide  
contexts per class

**6 classes of base substitutions**

C>T

C>A

C>G

T>A

T>C

T>G

A CA > ATA      ATA > AAA  
A CC > ATC      ATC > AAC  
A CG > ATG      ATG > AAG  
A CT > ATT      ATT > AAT  
C CA > CTA      CTA > CAA  
C CC > CTC      CTC > CAC  
C CG > CTG      CTG > CAG  
C CT > CTT      CTT > CAT  
G CA > GTA      GTA > GAA  
G CC > GTC      GTC > GAC  
G CG > GTG      GTG > GAG  
G CT > GTT      GTT > GAT  
T CA > TTA      TTA > TAA  
T CC > TTC      TTC > TAC  
T CG > TTG      TTG > TAG  
T CT > TTT      TTT > TAT

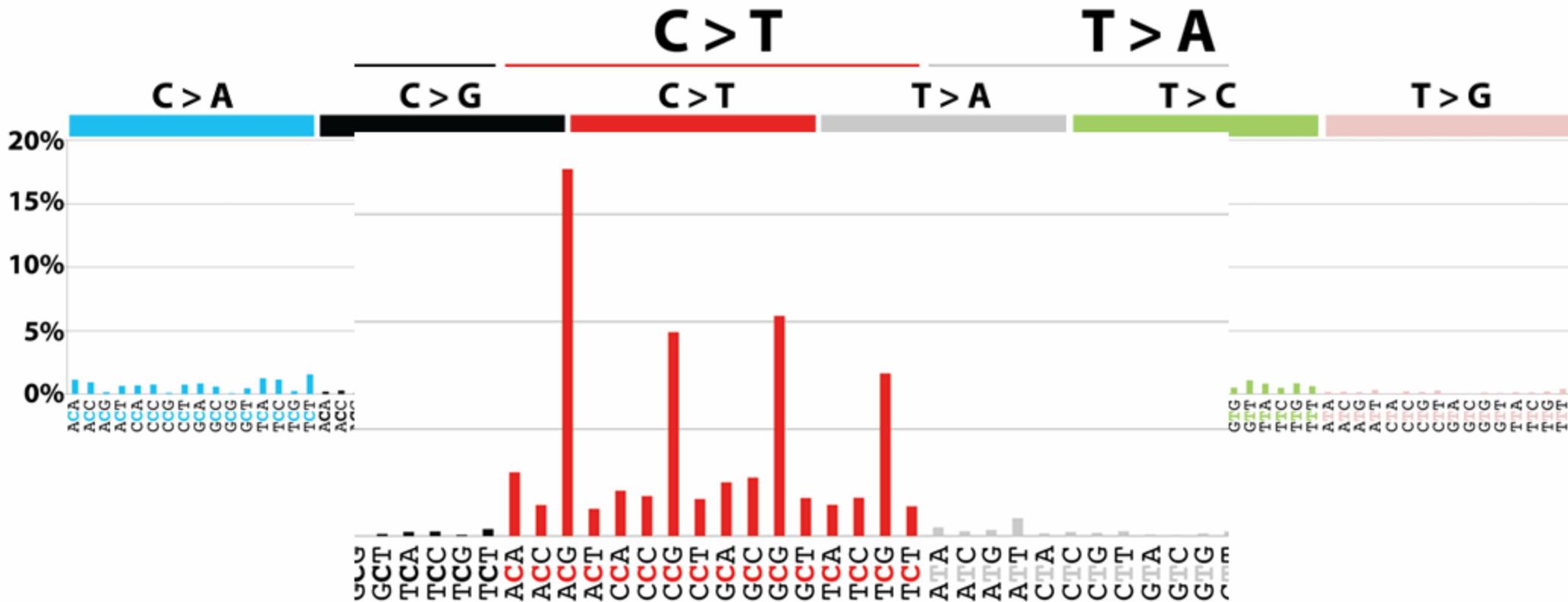
A CA > AAA      ATA > ACA  
A CC > AAC      ATC > ACC  
A CG > AAG      ATG > ACG  
A CT > AAT      ATT > ACT  
C CA > CAA      CTA > CCA  
C CC > CAC      CTC > CCC  
C CG > CAG      CTG > CCG  
C CT > CAT      CTT > CCT  
G CA > GAA      GTA > GCA  
G CC > GAC      GTC > GCC  
G CG > GAG      GTG > GCG  
G CT > GAT      GTT > GCT  
T CA > TAA      TTA > TCA  
T CC > TAC      TTC > TCC  
T CG > TAG      TTG > TCG  
T CT > TAT      TTT > TCT

A CA > AGA      ATA > AGA  
A CC > AGC      ATC > AGC  
A CG > AGG      ATG > AGG  
A CT > AGT      ATT > AGT  
C CA > CGA      CTA > CGA  
C CC > CGC      CTC > CGC  
C CG > CGG      CTG > CGG  
C CT > CGT      CTT > CGT  
G CA > GGA      GTA > GGA  
G CC > GGC      GTC > GGC  
G CG > GGG      GTG > GGG  
G CT > GGT      GTT > GGT  
T CA > TGA      TTA > TGA  
T CC > TGC      TTC > TGC  
T CG > TGG      TTG > TGG  
T CT > TGT      TTT > TGT

**96 classes of SNV mutations**

# Six classes of single-base mutations Reported by pyrimidine

Adding 5' and 3' adjacent bases  
96 possibilities considering context

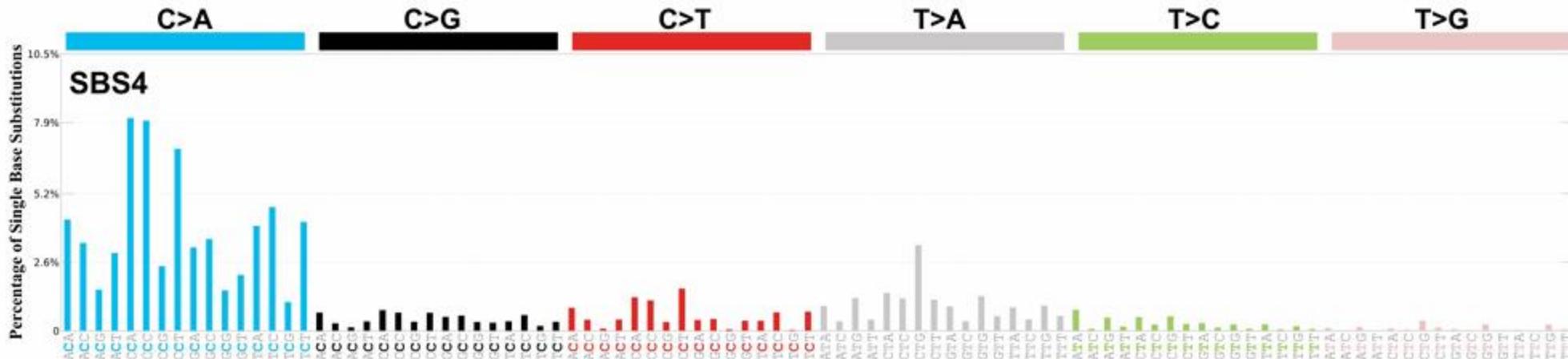




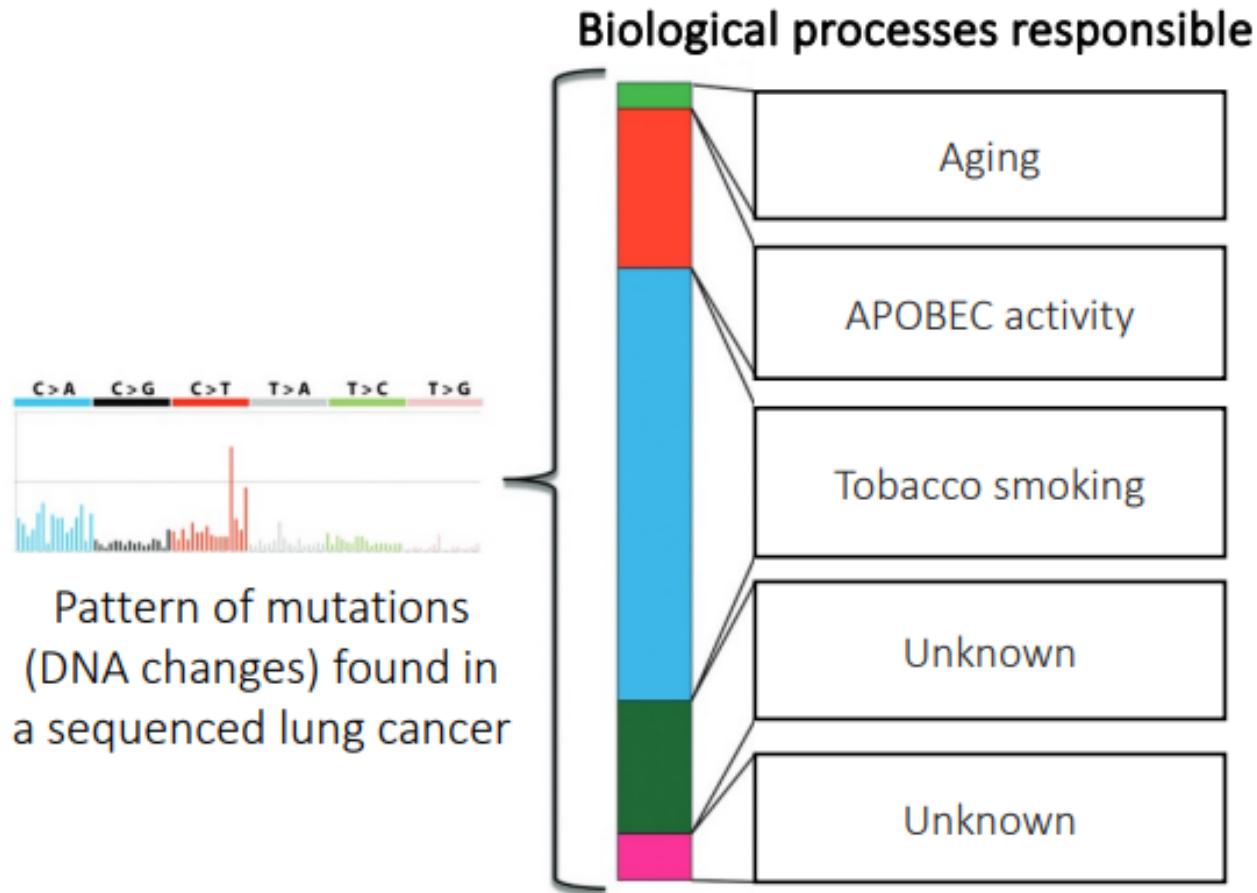
Tobacco smoking



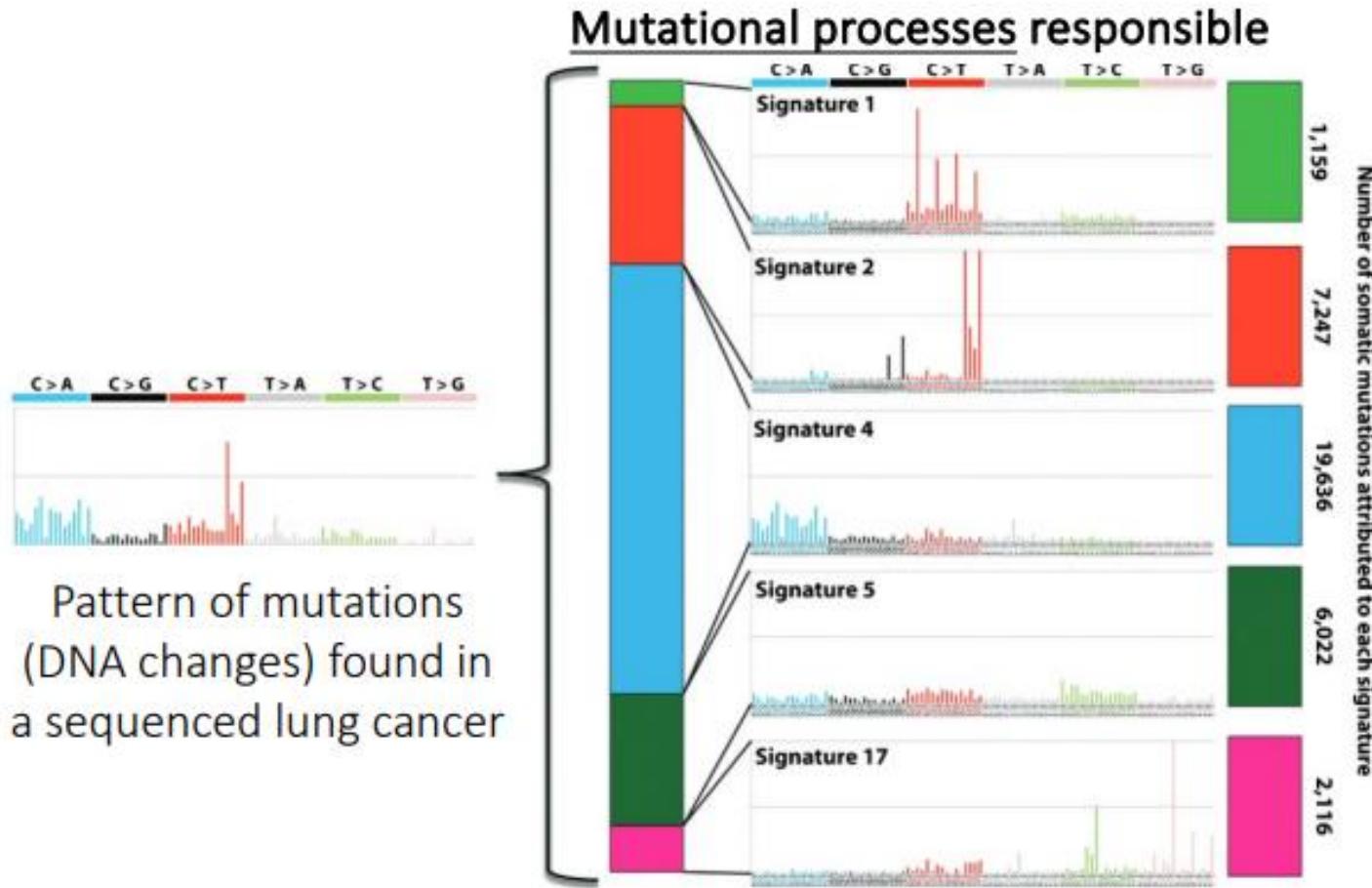
Cause of ~80% of lung cancers



# 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



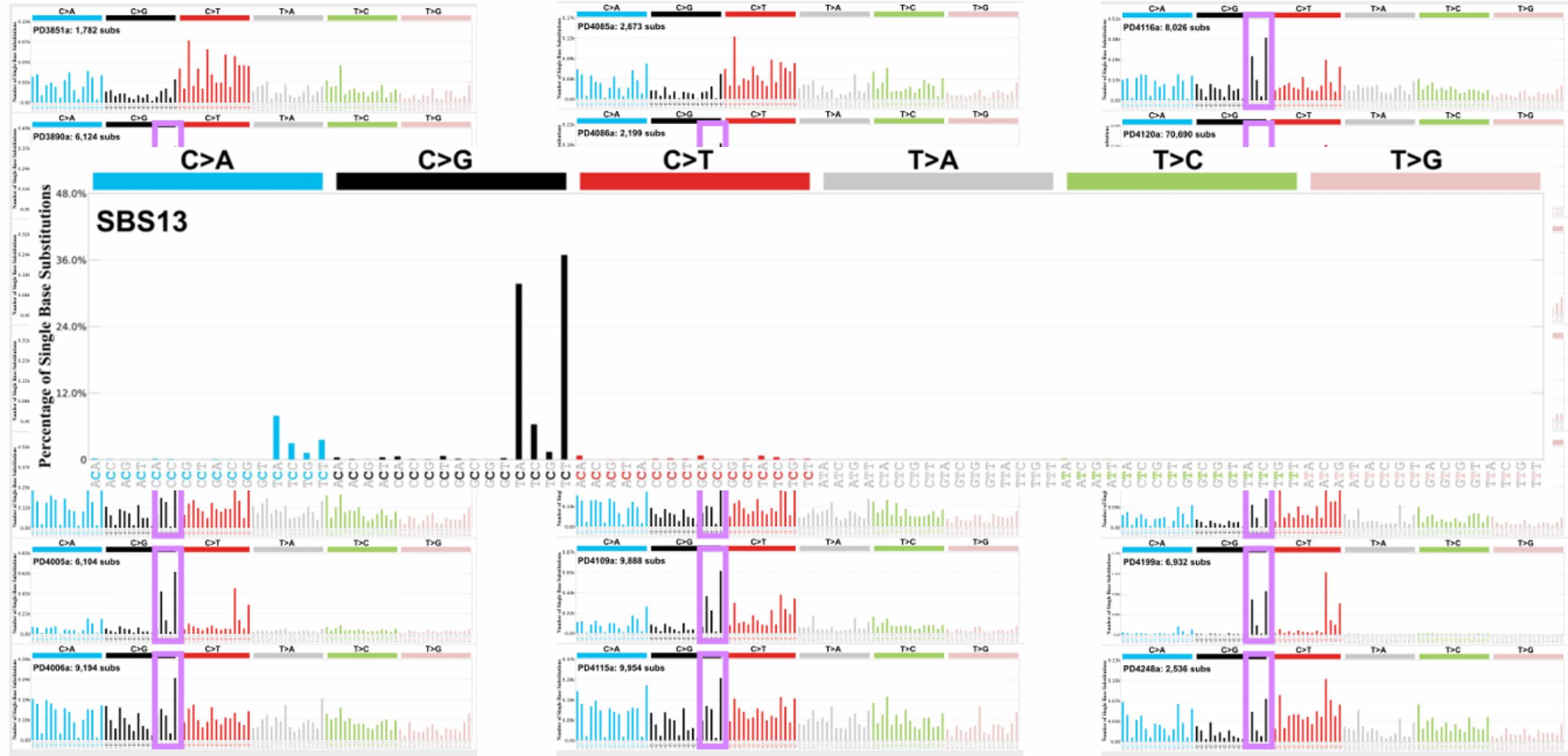
# Patterns of the 96 mutation classes across individual breast tumours



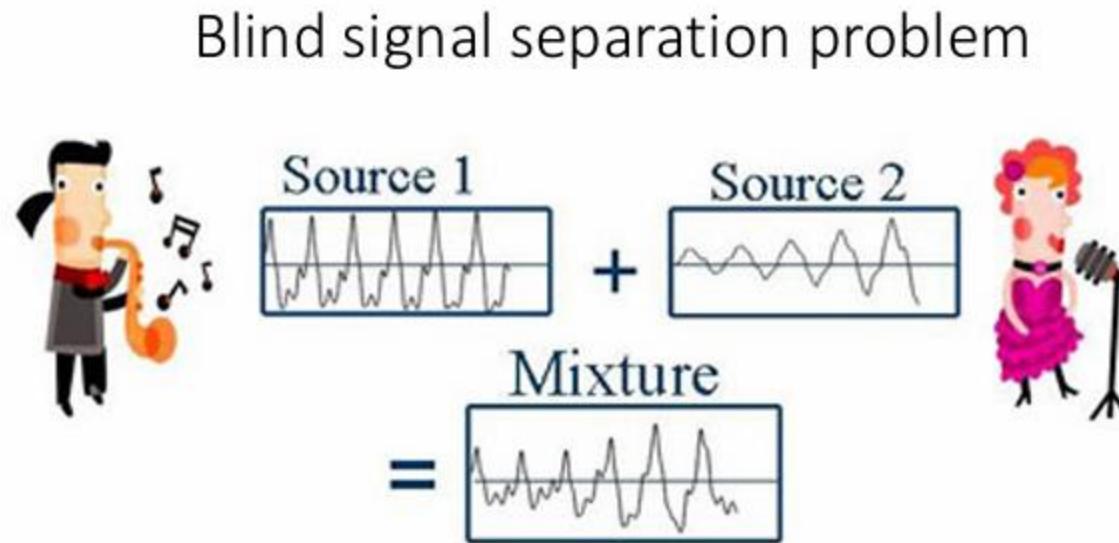
# Patterns of the 96 mutation classes across individual breast tumours



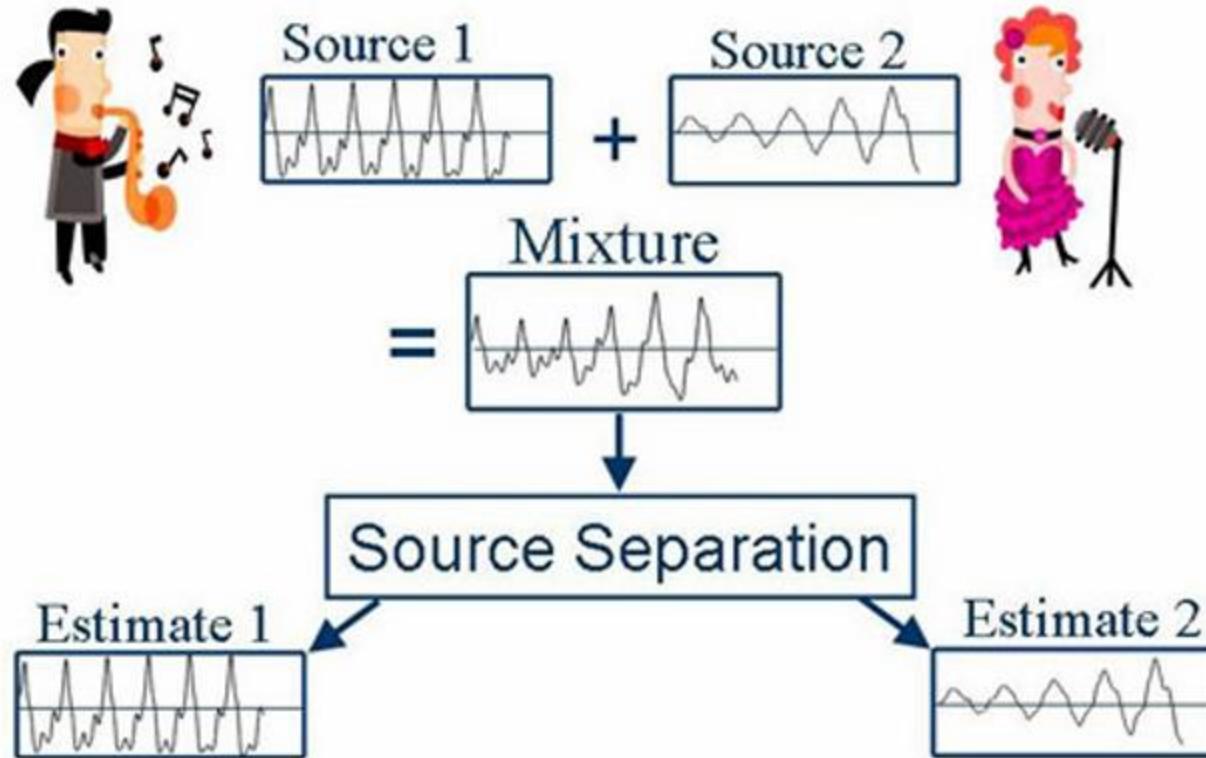
# Patterns of the 96 mutation classes across individual breast tumours



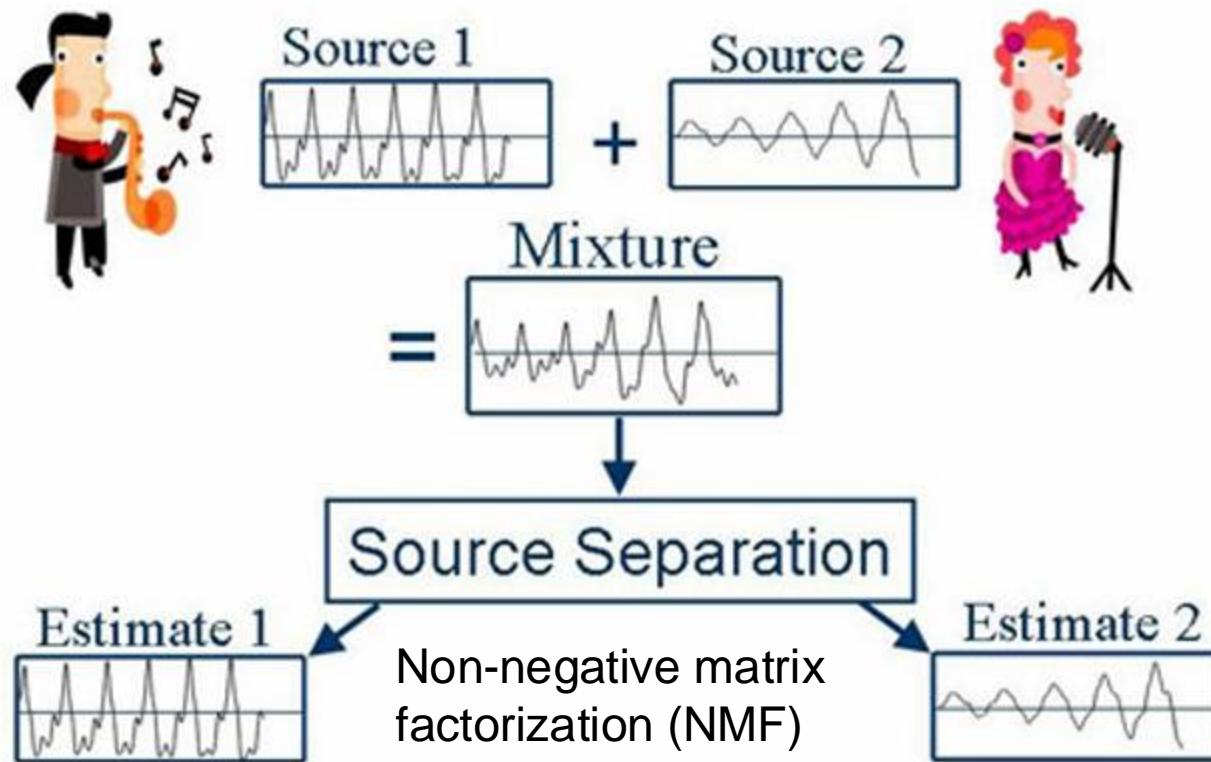
# Isolating the contribution of individual signatures to an overall pattern = blind signal separation problem



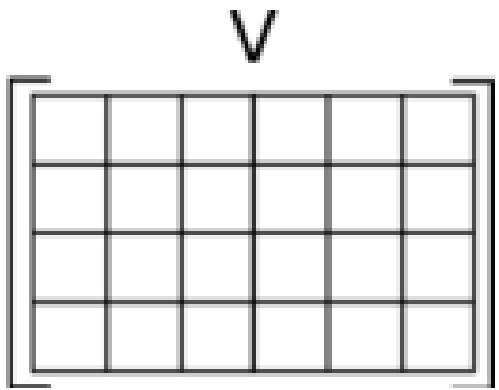
## Blind signal separation problem



## Blind signal separation problem

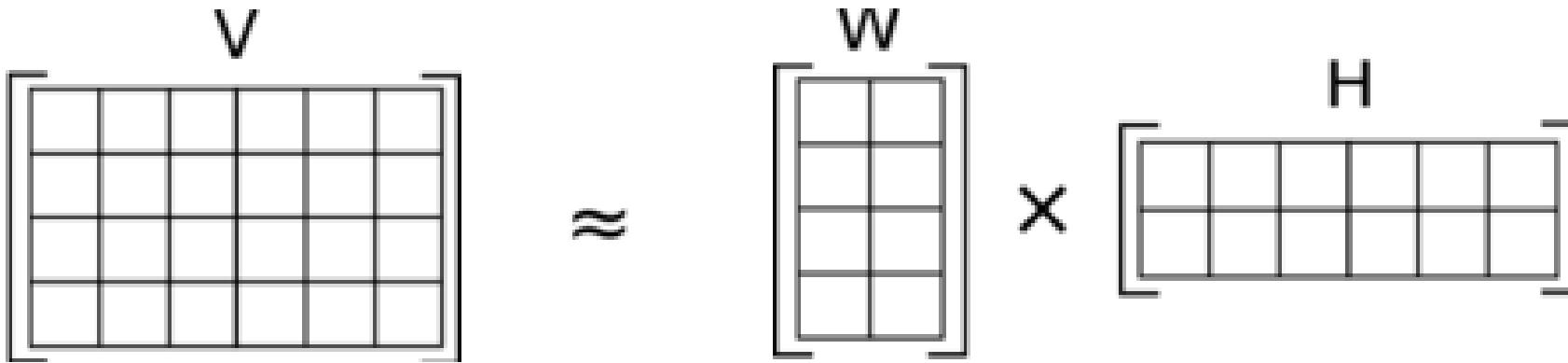


# Non-negative matrix factorization (NMF)



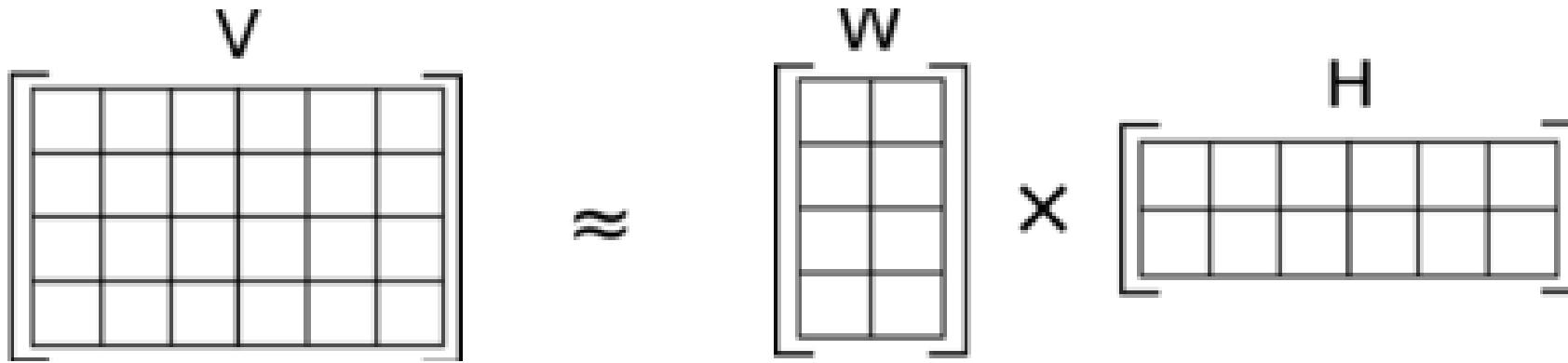
# Non-negative matrix factorization (NMF)

Factorization/Decomposition



Constraints:  $W$  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

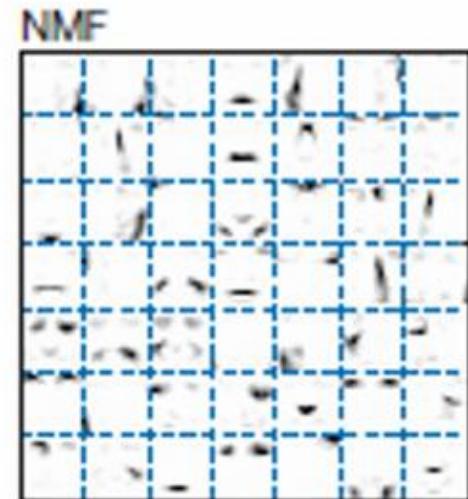
# NMF to learn the components of human faces (facial recognition)

## 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



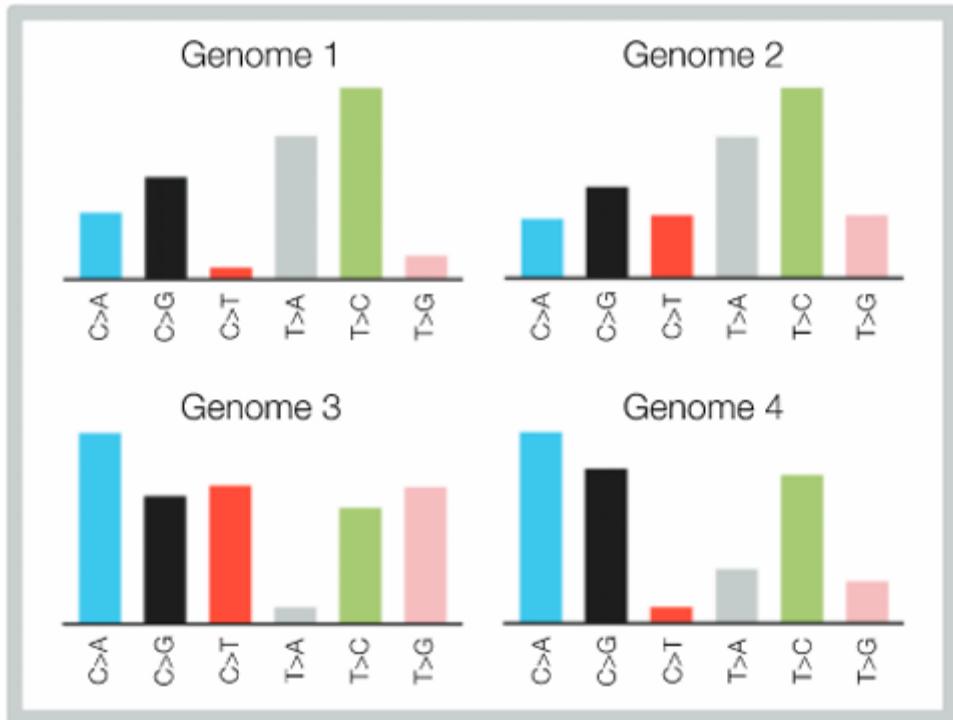
# NMF to isolate different instruments from an audio file

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.](#)

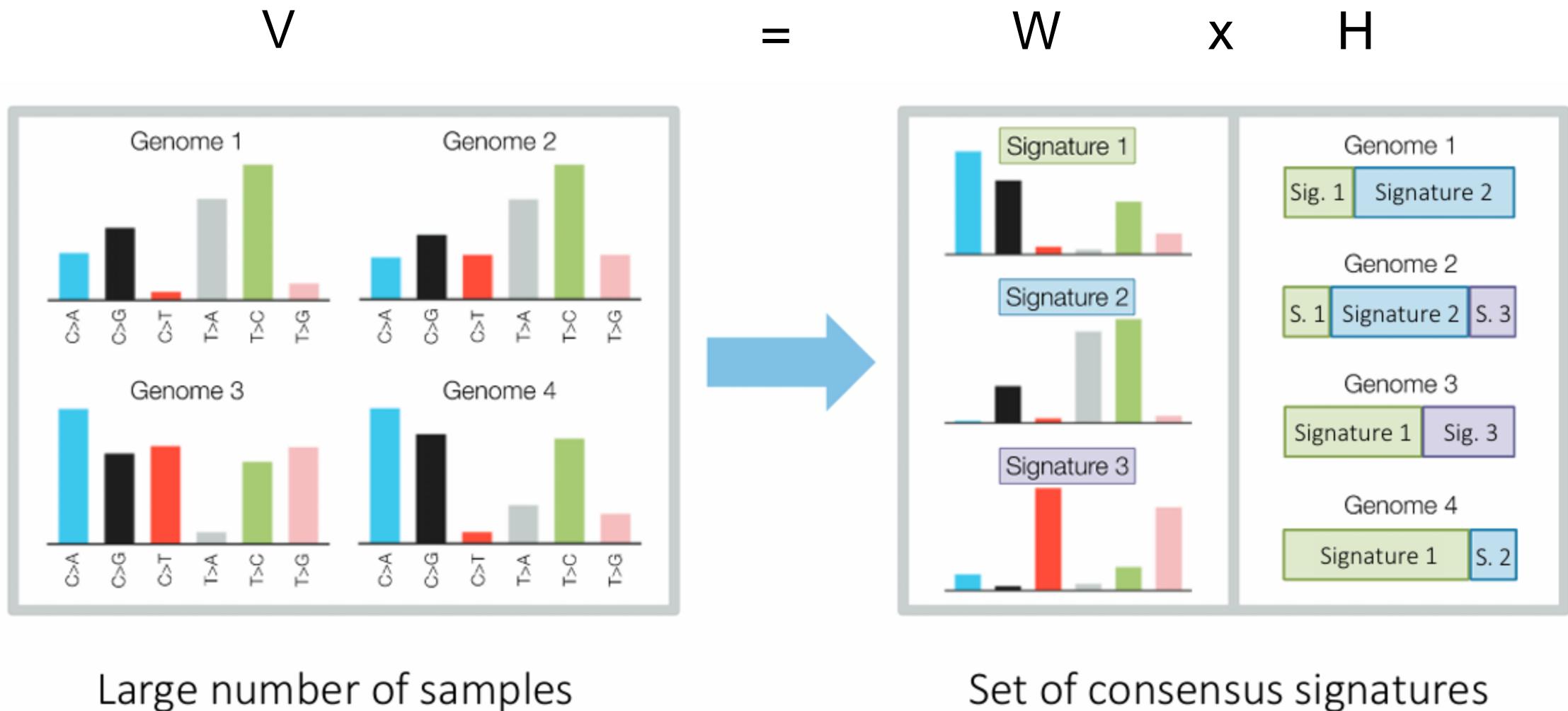
# NMF to isolate SNV mutational signatures from a collection of genomes

V



Large number of samples

# NMF to isolate SNV mutational signatures from a collection of genomes



# Considerations for mutational signature experiments

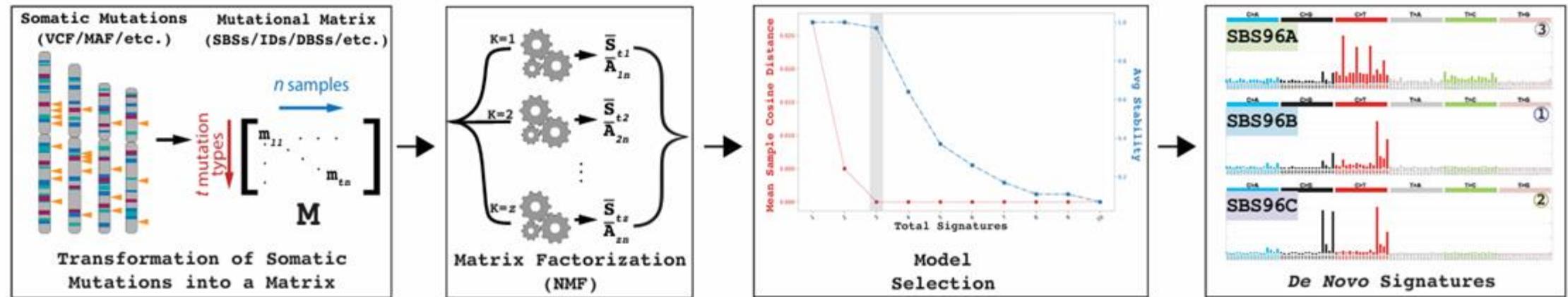
## Discuss:

- 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

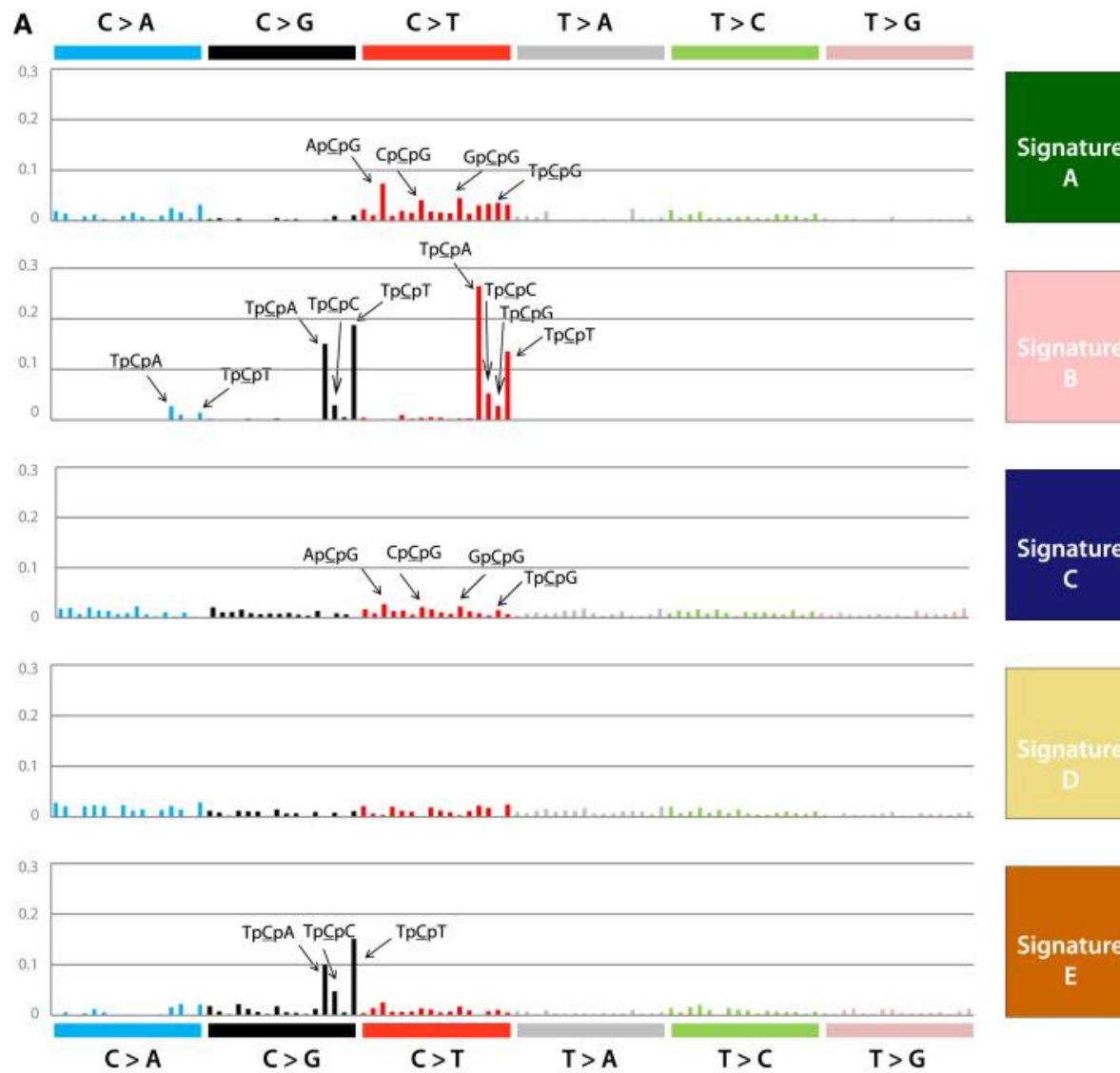
# Many other tools and algorithms exist to extract mutational signatures

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

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

21 breast cancer whole genomes

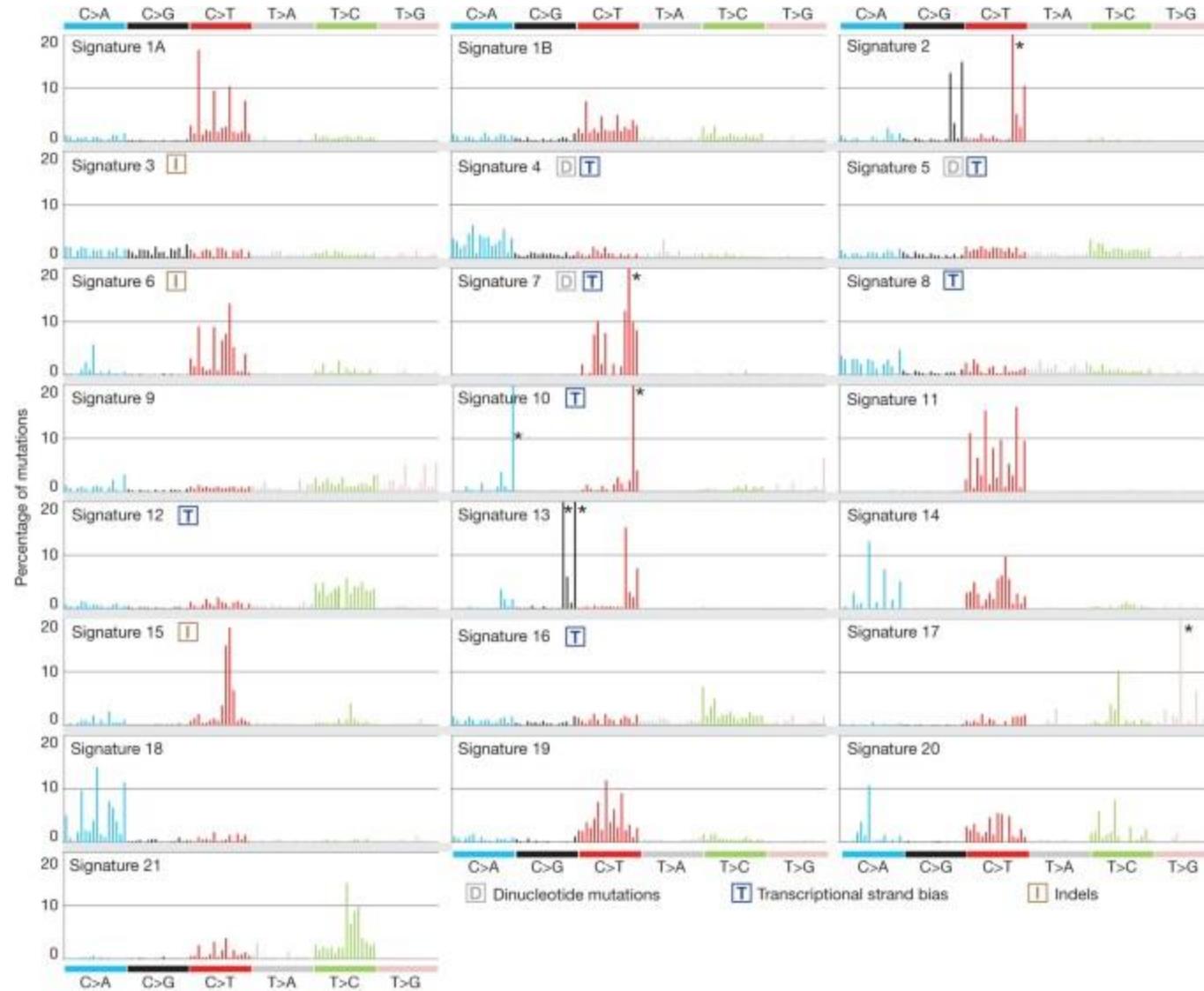
5 mutational signatures, A-E



# 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



7,042 cancers (exomes + genomes)

21 single base substitution (SBS) mutational signatures

COSMIC v1 (August 2013)

# COSMIC Mutation Signature Database



COSMIC v2 (March 2015)

- 30 SBS signatures

# COSMIC Mutation Signature Database



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



Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Search COSMIC... SEARCH

## Mutational Signatures (v3.3 - June 2022)

### Introduction

Somatic mutations are present in all cells of the human body and occur throughout life. They are the consequence of multiple mutational processes, including the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair. Different mutational processes generate unique combinations of mutation types, termed "Mutational Signatures".

In the past few years, large-scale analyses have revealed many mutational signatures across the spectrum of human cancer types, including the latest effort by the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Network (Alexandrov, L.B. et al., 2020) using data from more than 23,000 cancer patients.

### About

COSMIC Mutational Signatures is a resource curated in partnership with COSMIC and Cancer Grand Challenges, and in close association with our collaborators at Wellcome Sanger Institute, the Pillay lab at University College London and the Alexandrov lab at University of California.



### Signature-based websites

At COSMIC Signatures we identify signatures from analysis of the PCAWG dataset and through curation of specific papers. Papers are looked at particularly (but not exclusively) when there is a specific exposure which captures signatures not present in the PCAWG dataset. Please note that this catalogue of signatures is not exhaustive or a final set, but a reference set of high confidence signatures that have been curated by experts in the field. We aim to update as comprehensively as possible as new data become available and improvements are made to extraction methodologies.

This summary includes the mutational profile, proposed aetiology and tissue distribution of each signature, as well as potential associations with other mutational signatures and how the signature has changed during iterations of analysis.

Currently, four different variant classes are considered, resulting in the following sets of mutational signatures.

[SBS Signatures](#) [DBS Signatures](#) [ID Signatures](#) [CN Signatures](#)

### Data downloads

Download current COSMIC Mutational Signatures version 3.3 and previous releases here.

[Downloads](#)

### Versions

COSMIC Mutational Signatures version 3.3 is the latest release.

Version 3 was released as part of COSMIC release v89 (May 2019), updated to version 3.1 in COSMIC release v91 (June 2020), to version 3.2 in COSMIC release v93 (March 2021) and most recently version 3.3 in COSMIC v95 (May 2022).

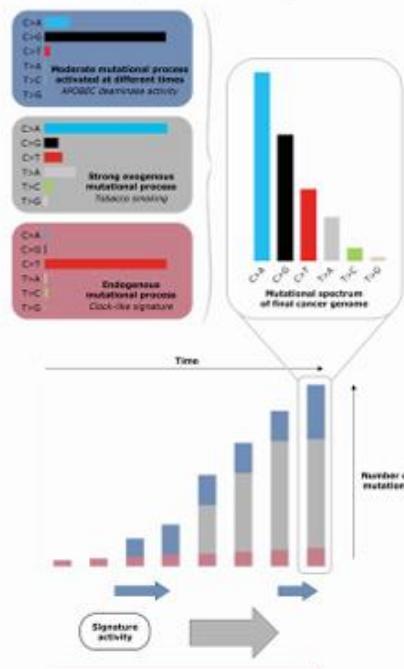
Version 2 signatures (March 2015) were part of earlier COSMIC releases can still be consulted:

[Version 2](#)

### Mutational signatures as a collection of operative mutational processes

Mutational processes from different aetiologies are active during the course of cancer development. They can be identified using mutational signatures, due to their unique mutational pattern and specific activity on the genome.

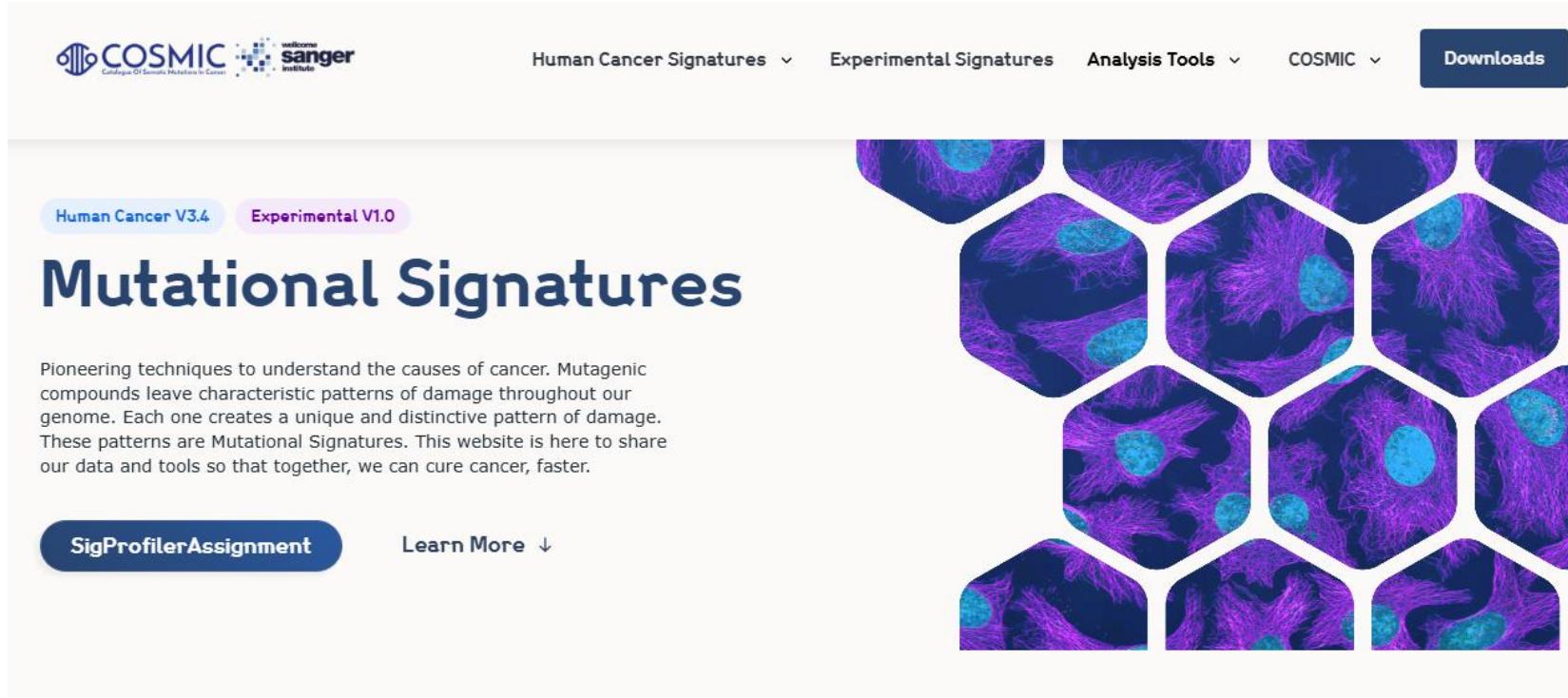
This is illustrated in the figure below using a framework of 6 classes of single base substitutions, and three distinct mutational processes, whose respective strengths vary throughout a patient's life. At the beginning, all mutations were due to the activity of the endogenous mutational process. As time progresses, the other processes get activated and the mutational spectrum of the cancer genome continues to change.



## COSMIC v3.3 (June 2022)

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

# COSMIC Mutational Signatures Database



Human Cancer Signatures ▾ Experimental Signatures Analysis Tools ▾ COSMIC ▾ Downloads

Human Cancer V3.4 Experimental V1.0

## Mutational Signatures

Pioneering techniques to understand the causes of cancer. Mutagenic compounds leave characteristic patterns of damage throughout our genome. Each one creates a unique and distinctive pattern of damage. These patterns are Mutational Signatures. This website is here to share our data and tools so that together, we can cure cancer, faster.

SigProfilerAssignment Learn More ↓

October v3.4 (June 2023)

### Genomics

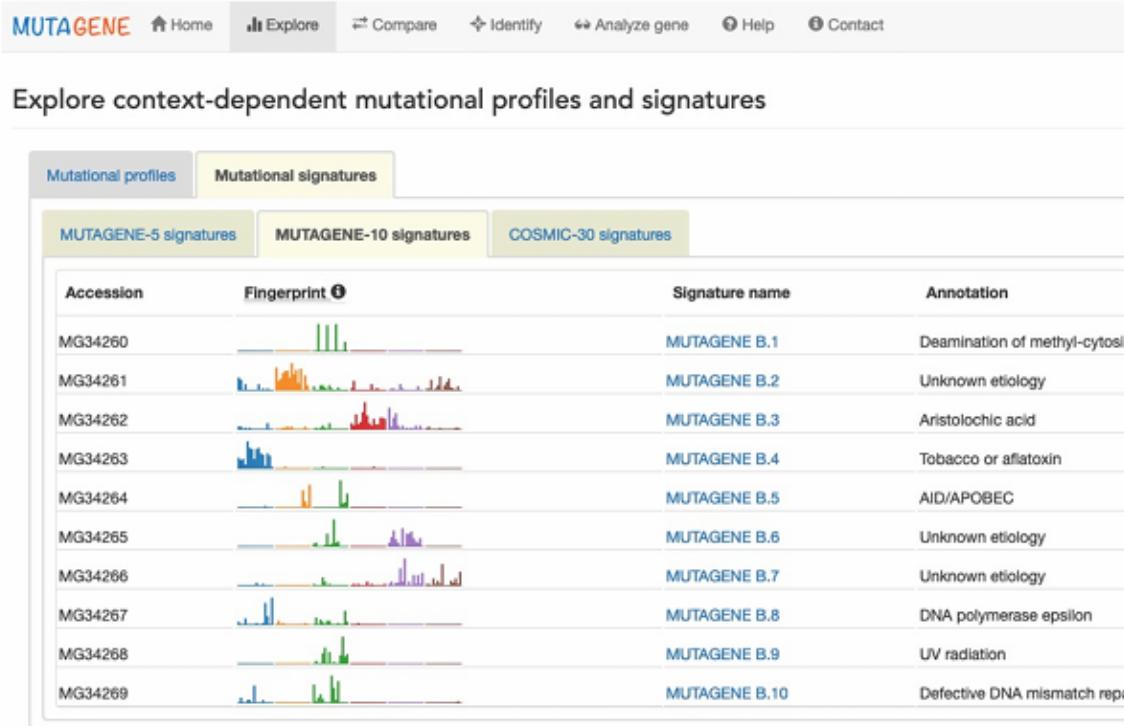
- 99 SBS signatures
- 20 DBS signatures
- 23 ID signatures
- 25 CN signatures
- 10 SV signatures
- 5 RNA-SBS signatures

### Experimental

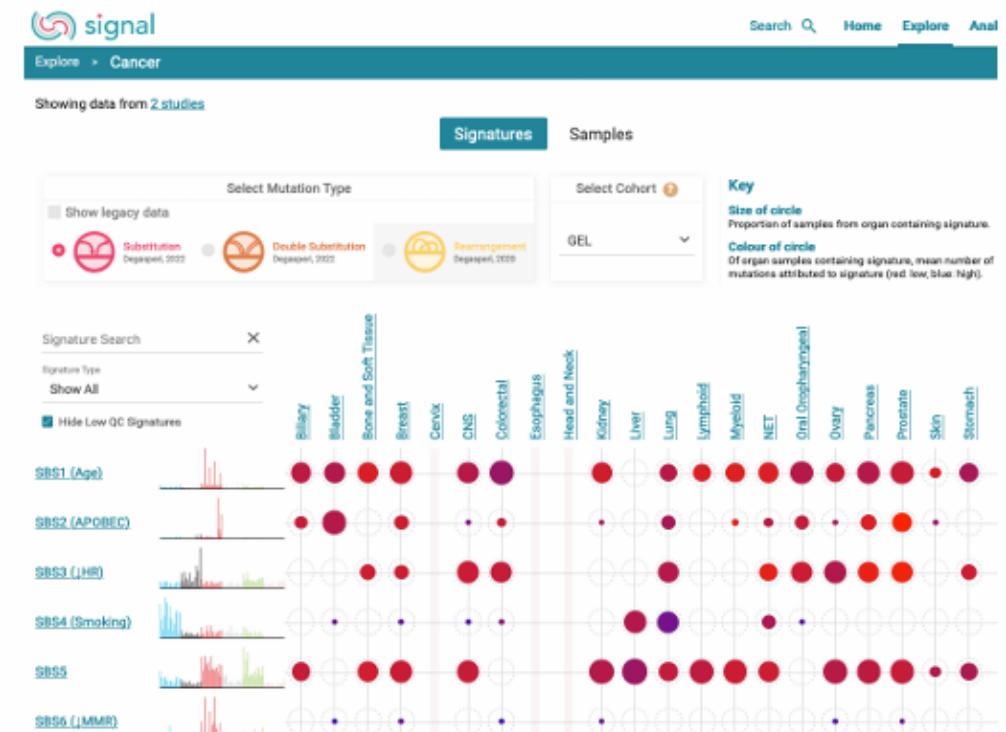
- 140 signatures

# Other Mutational Signatures Databases

# MUTAGENE



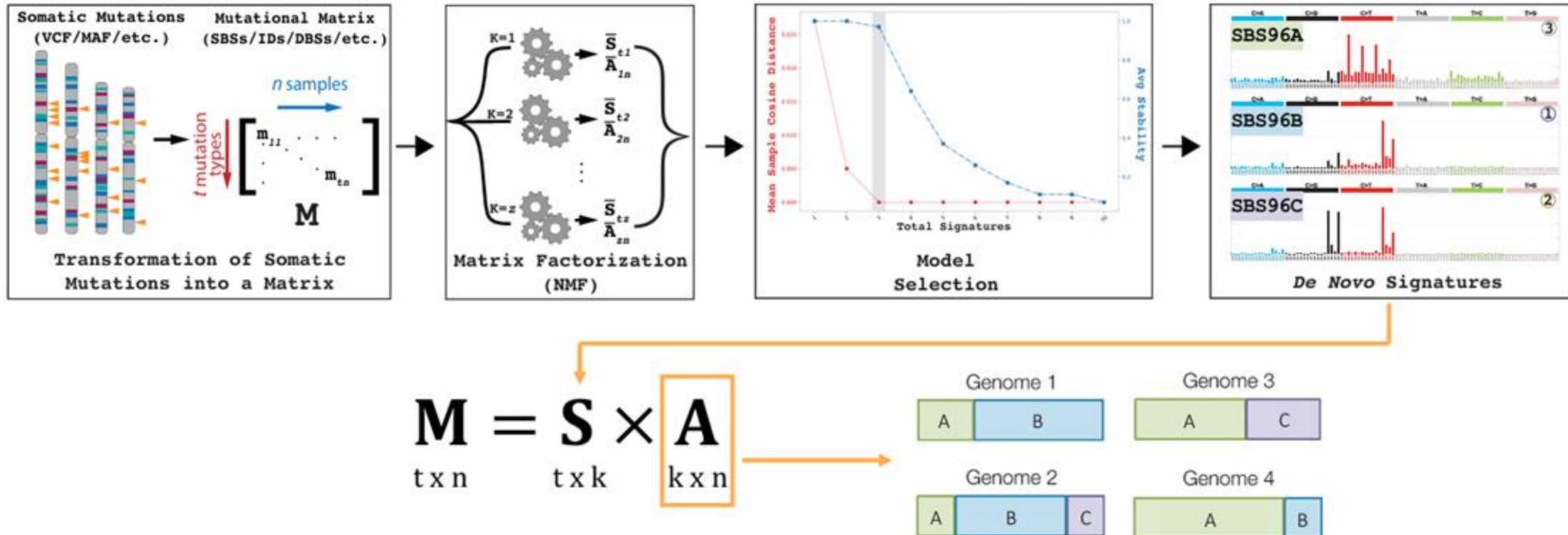
<https://www.ncbi.nlm.nih.gov/research/mutagene/>



<https://signal.mutationalsignatures.com/>

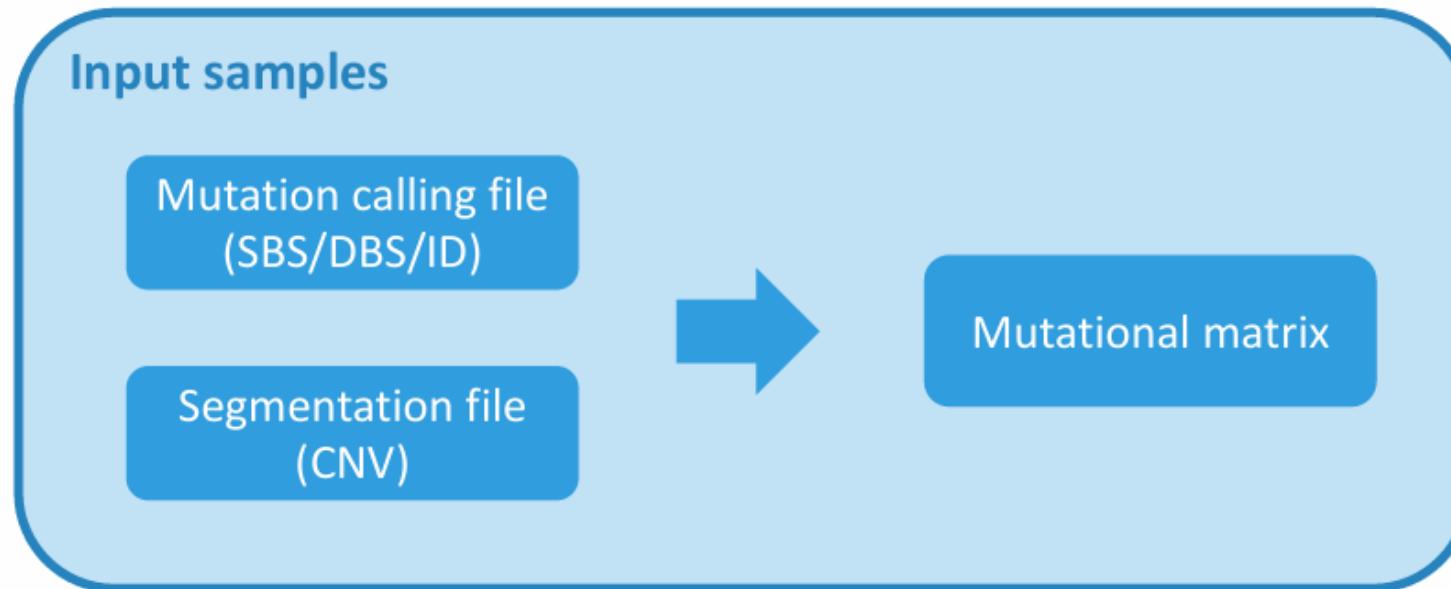
# SigProfiler is an open-source toolkit for mutational signatures analyses

**SIGPROFILER**  
Extractor



**SIGPROFILER**  
Assignment

# SigProfilerAssignment assigns mutational signature weights to individual samples



$$M = S \times A$$

$t \times n$        $t \times k$        $k \times n$

The equation  $M = S \times A$  is shown in blue and red text. To its right, a matrix multiplication diagram is displayed:  $t \times n$  is written in blue above a red square labeled  $S$ ;  $t \times k$  is written in red below a black square labeled  $A$ ; and  $k \times n$  is written in black below the red square  $S$ .



**t** mutational contexts  
**n** samples  
**k** signatures

# Assigning reference signatures to individual samples

M

Sample #1



Sample #2



Sample #3



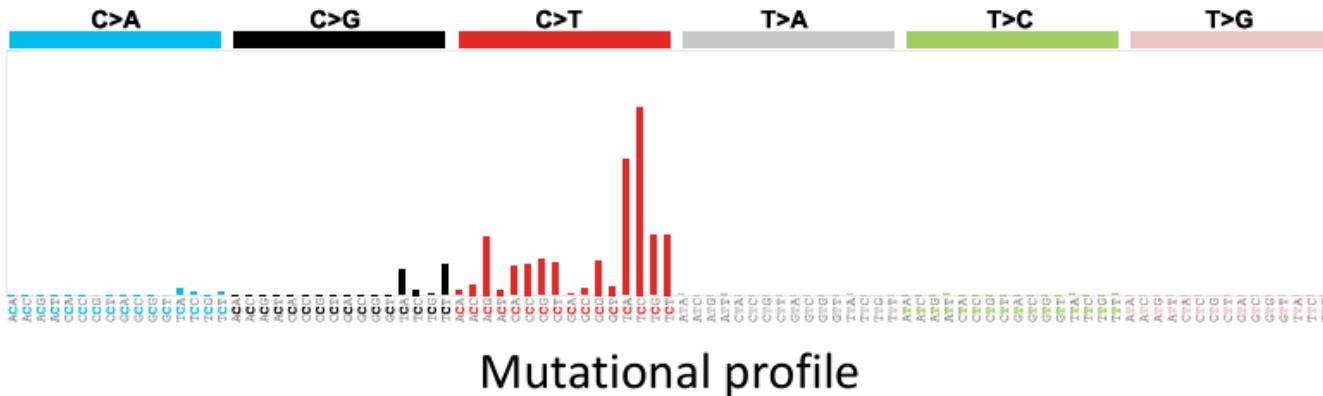
Sample #4



Sample #5



Number of mutations

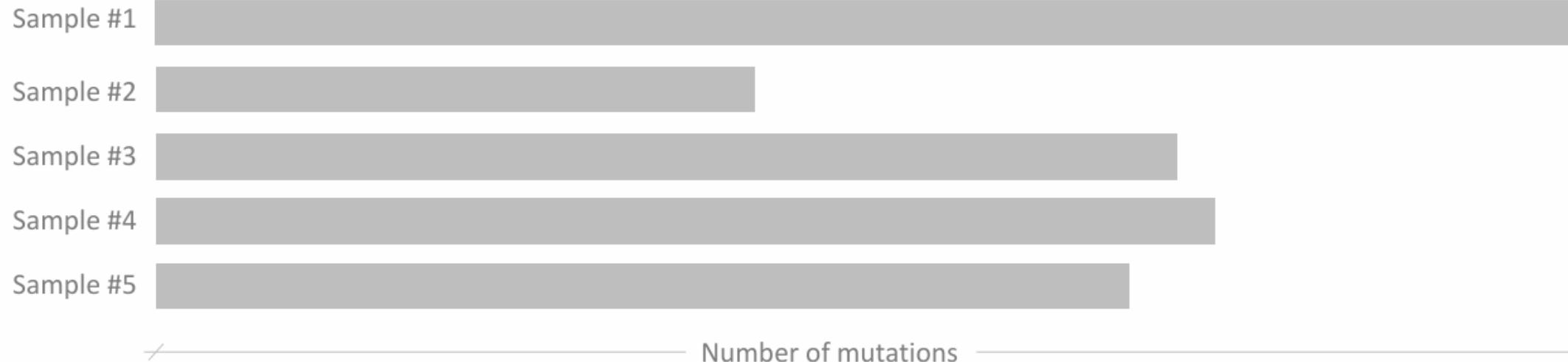


#CHROM	POS	FILTER	REF	ALT
1	809687	PASS	G	C
1	819245	PASS	G	T
1	1911011	PASS	C	G
1	2112413	PASS	T	C
1	2927666	PASS	A	G
1	3359791	PASS	C	T
1	4347912	PASS	G	A
1	4961889	PASS	G	C
1	5949138	PASS	C	T
1	7806339	PASS	A	C
1	9648435	PASS	G	A
1	9705025	PASS	C	T

Mutation calling file

# Assigning reference signatures to individual samples

M

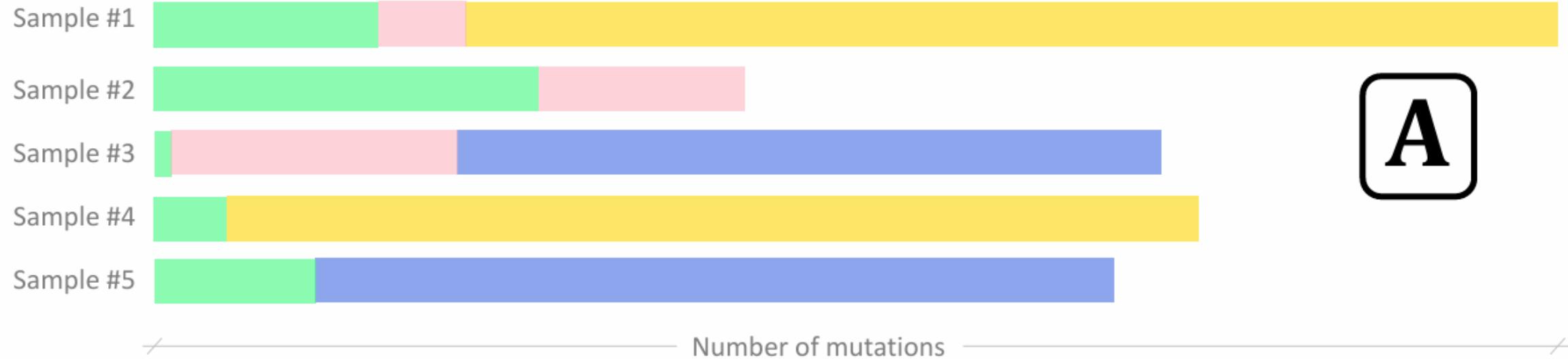


S

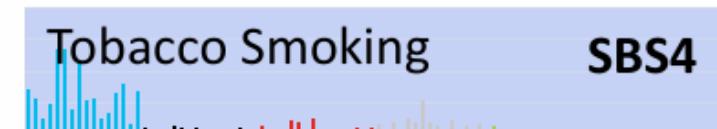
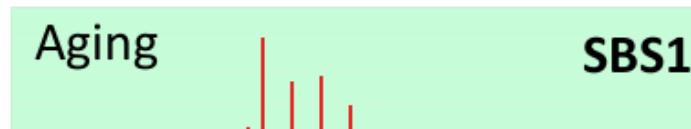


# Assigning reference signatures to individual samples

M



S



# Assigning reference signatures to individual mutations

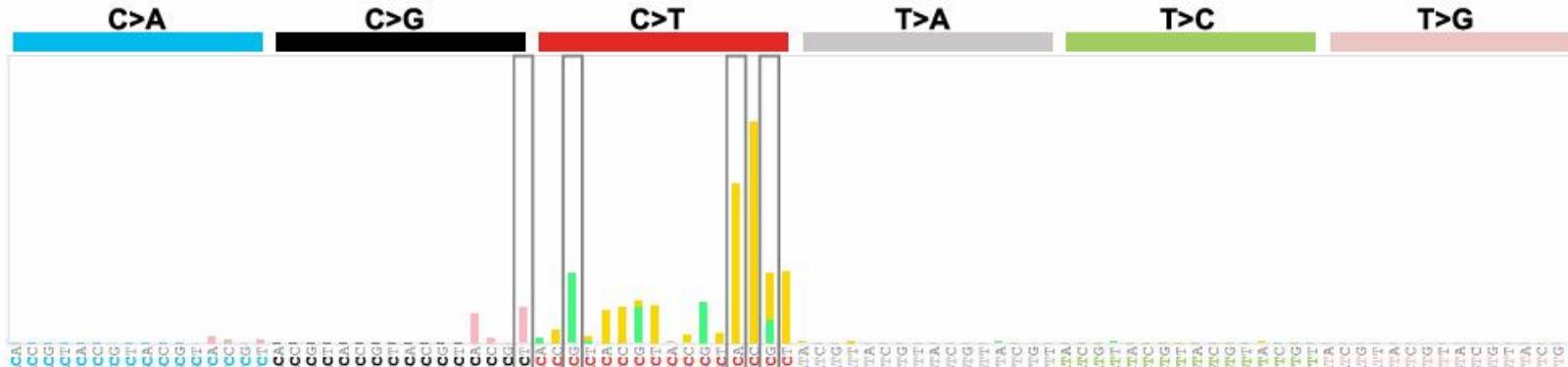


# Assigning reference signatures to individual mutations



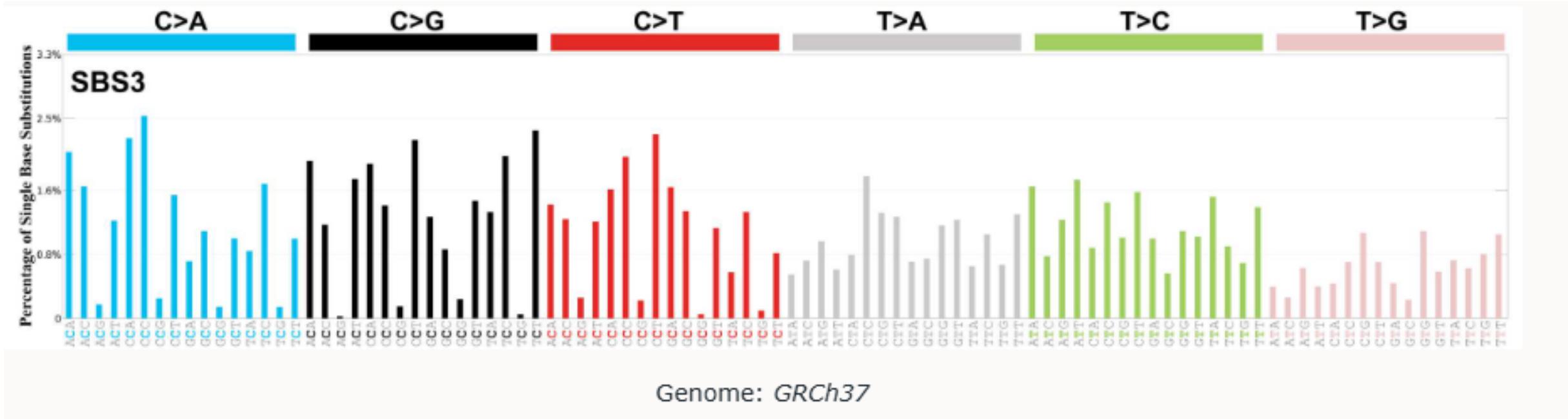
$$P(\text{Signature} \mid \text{Mut. context}) = \frac{S}{M} \cdot A$$
$$S = \frac{P(\text{Mut. context} \mid \text{Signature}) \times P(\text{Signature})}{P(\text{Mut. context})}$$

# 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

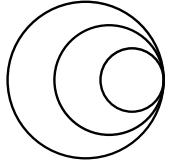
# Many other tools and algorithms exist to assign mutational signatures and their weights to individual samples

Tool	Platform	Refitting Approach		Reference
		Method	Computational Engine	
<b>deconstructSigs</b>	R	Non-negative linear regression	Original implementation	Rosenthal <i>et al.</i> 2016 Genome Biology
<b>MSA</b>	Python / Nextflow	NNLS	Original implementation / Scipy python package	Senkin 2021 BMC Bioinformatics
<b>MutationalPatterns (standard)</b>	R	NNLS	Pracma R package	Blokzijl <i>et al.</i> 2018 Genome Medicine
<b>MutationalPatterns (strict)</b>	R	NNLS	Original implementation / Pracma R package	Manders <i>et al.</i> 2022 BMC Genomics
<b>sigLASSO</b>	R	Lasso regression	Original implementation / glmnet R package	Li <i>et al.</i> 2020 Nature Communications
<b>SignatureToolsLib</b>	R / Web app	Non-negative linear regression	NNLM R package	Degasperi <i>et al.</i> 2022 Science
<b>SigProfilerAssignment</b>	Python / R / Web app	NNLS	Original implementation / Scipy python package	Díaz-Gay <i>et al.</i> 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](http://wellcomeconnectingscience.org)  
for more information.

