Day 4. Cancer Genome Analysis - Latin America and the Caribbean

Extraction of *de novo* mutational signatures

November 30th, 2023

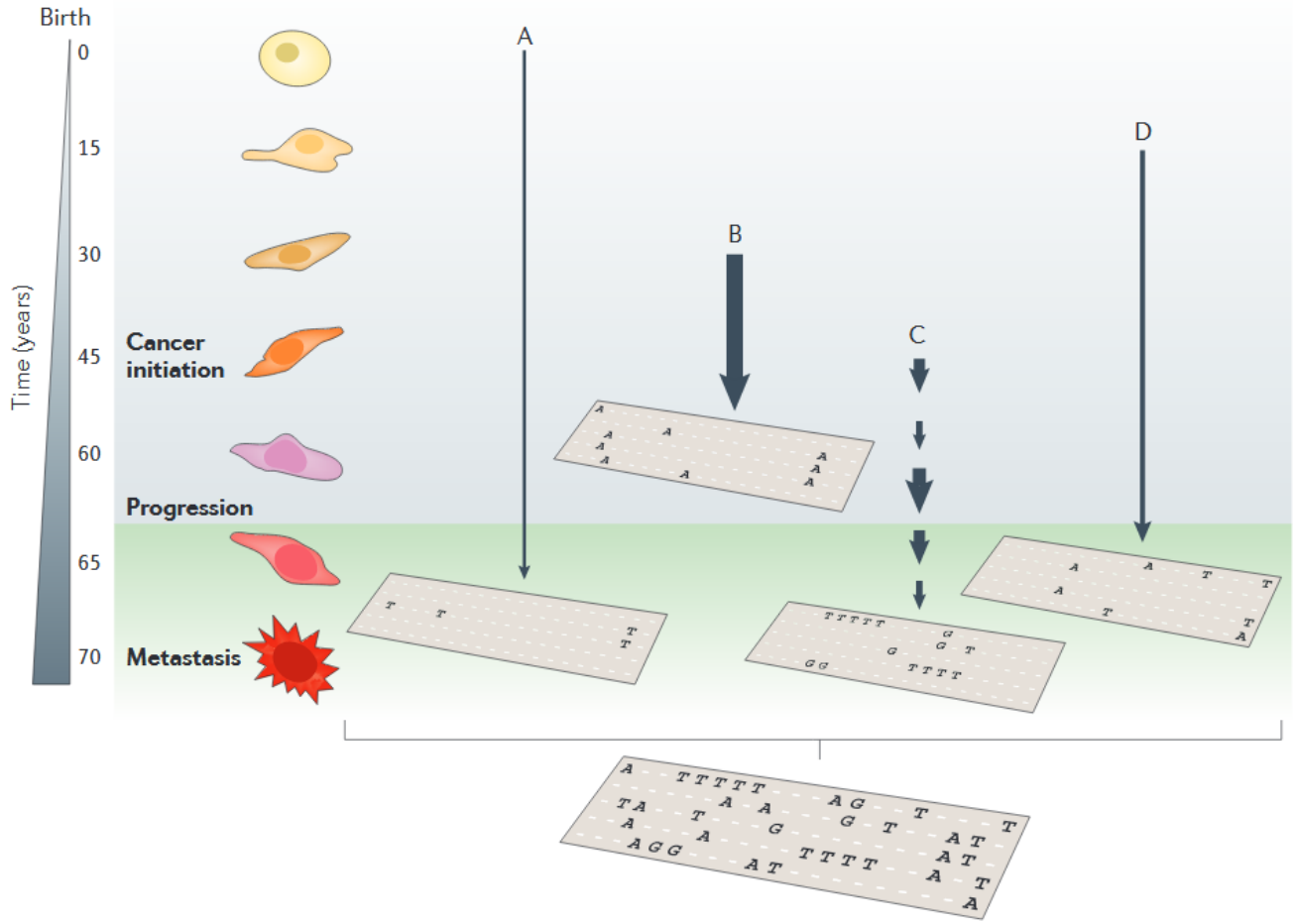
Marcos Díaz Gay

Alexandrov lab, University of California San Diego

UC San Diego

wellcome connecting science

The mutational profile of a cancer patient is a mix of different processes characterized by specific mutational signatures

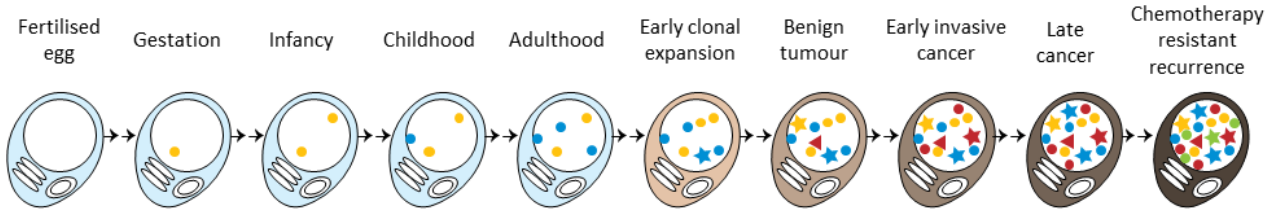


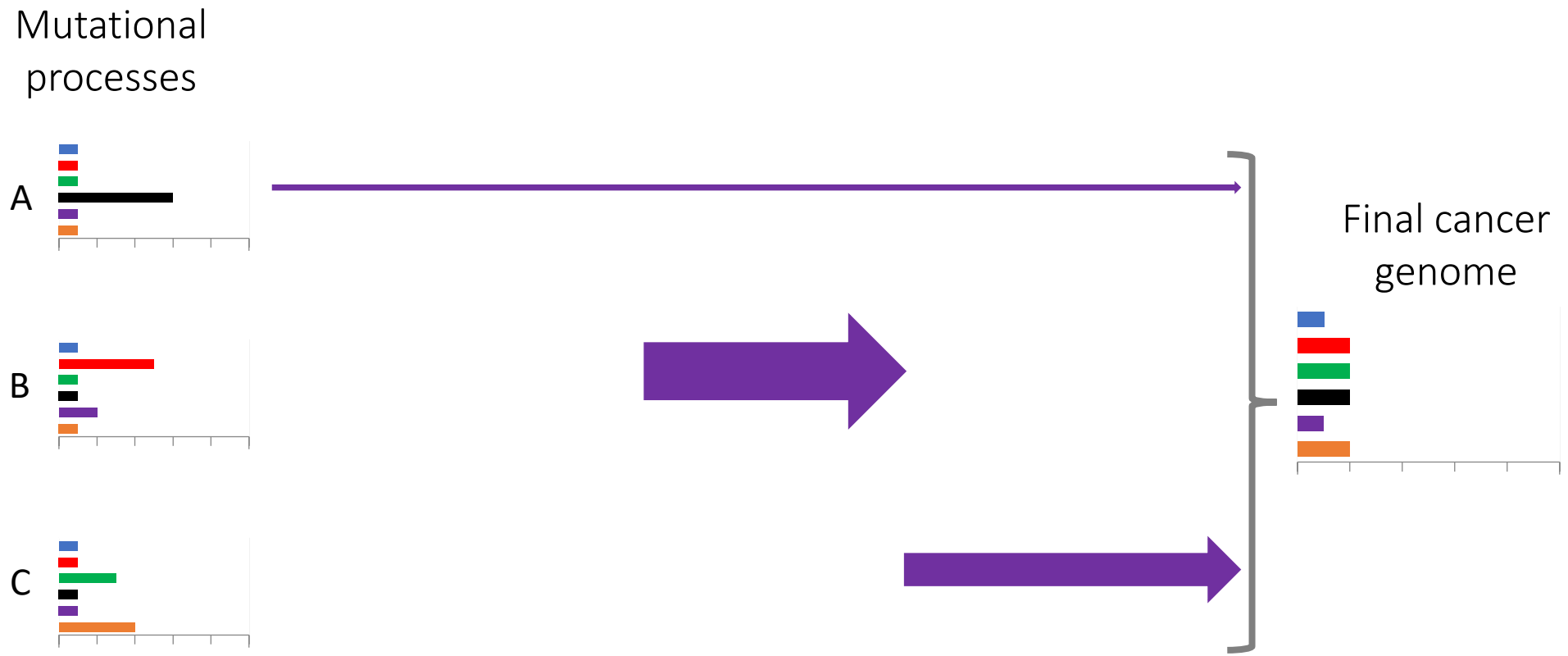
Helleday *et al.* 2014 Nature Reviews Genetics

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

Stratton *et al.* 2009 Nature

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





Stratton *et al.* 2009 Nature

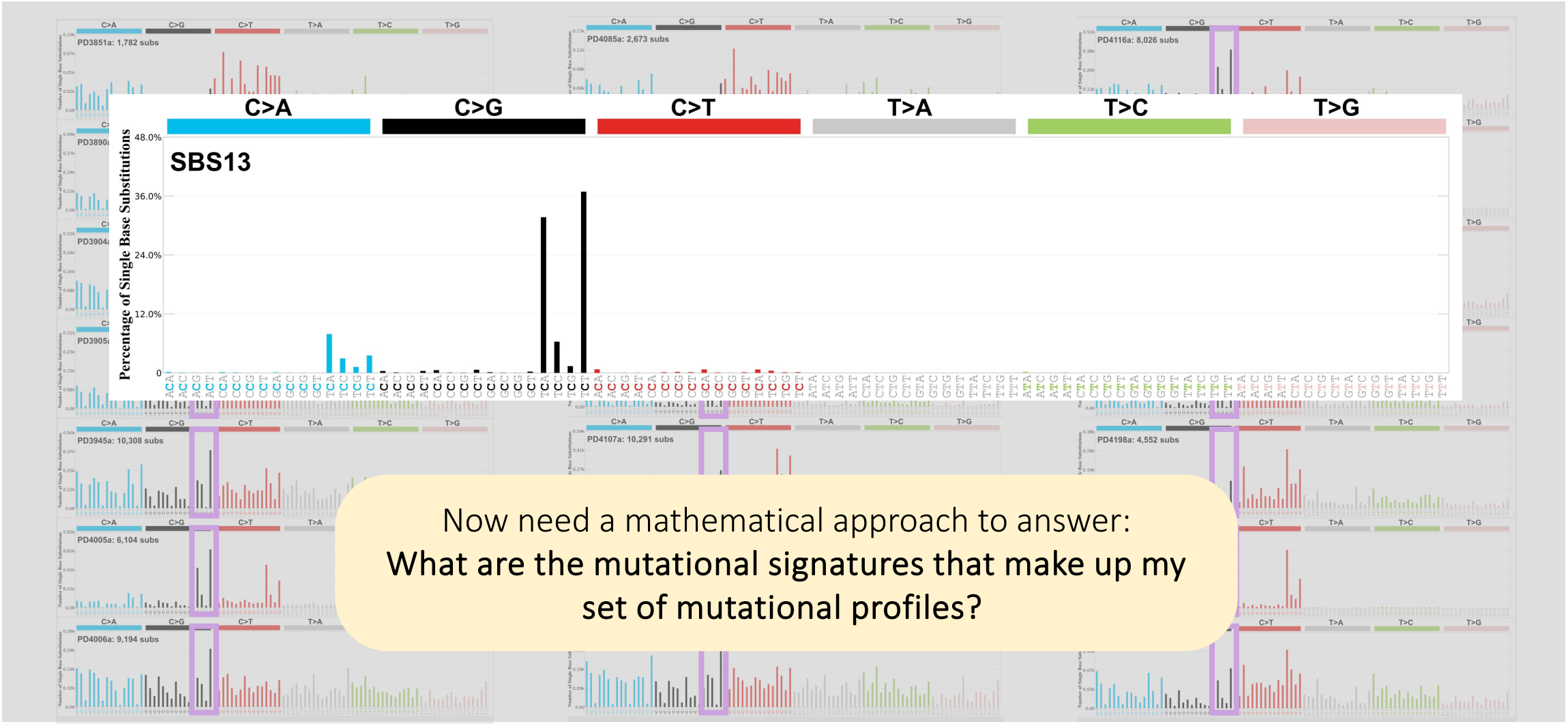
Mutational signatures can be determined based on mutational profiles across a set of individuals

Nik-Zainal *et al.* 2012 Cell

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

Nik-Zainal *et al.* 2012 Cell

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

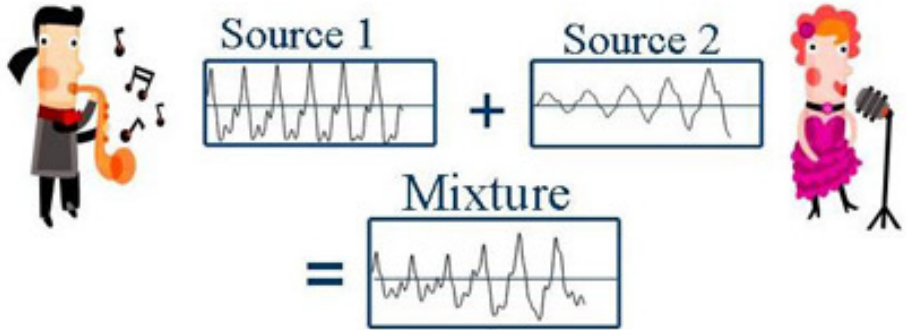


Now need a mathematical approach to answer:
What are the mutational signatures that make up my set of mutational profiles?

Nik-Zainal *et al.* 2012 Cell

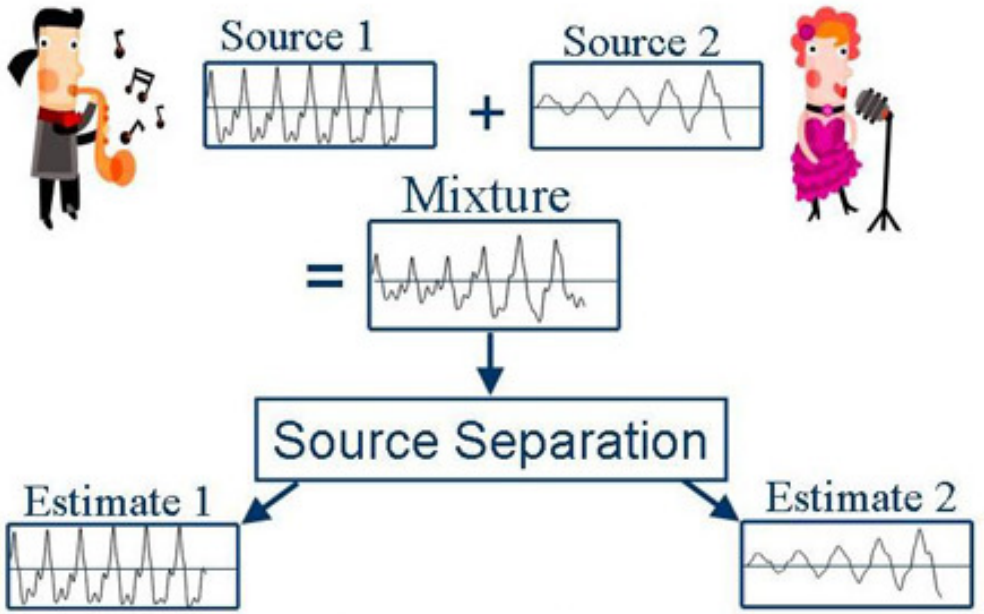
Mathematical models allows the *un-mixing* and the extraction of mutational signatures

Blind signal separation problem



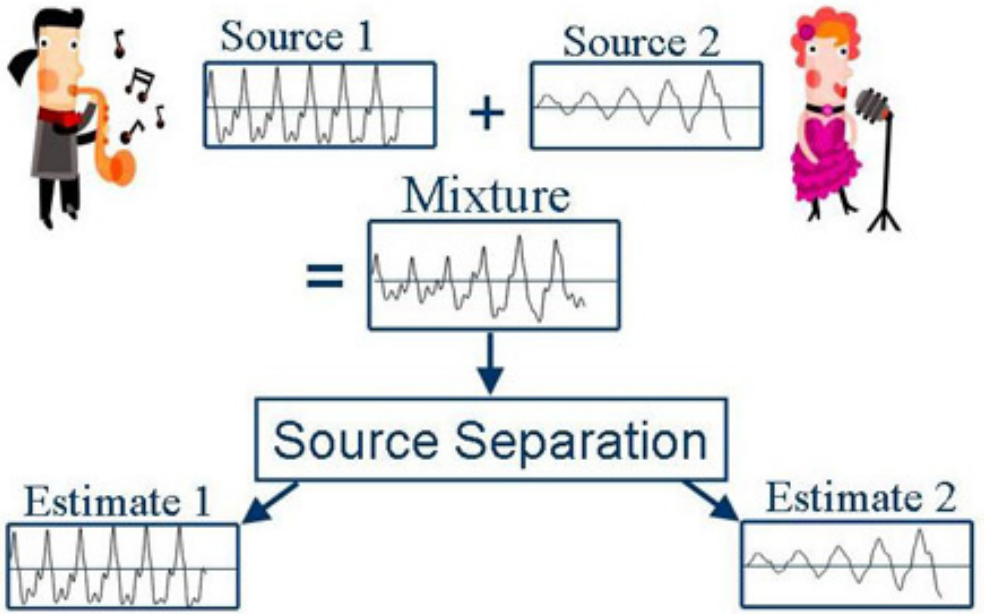
Mathematical models allows the *un-mixing* and the extraction of mutational signatures

Blind signal separation problem



Mathematical models allows the *un-mixing* and the extraction of mutational signatures

Blind signal separation problem



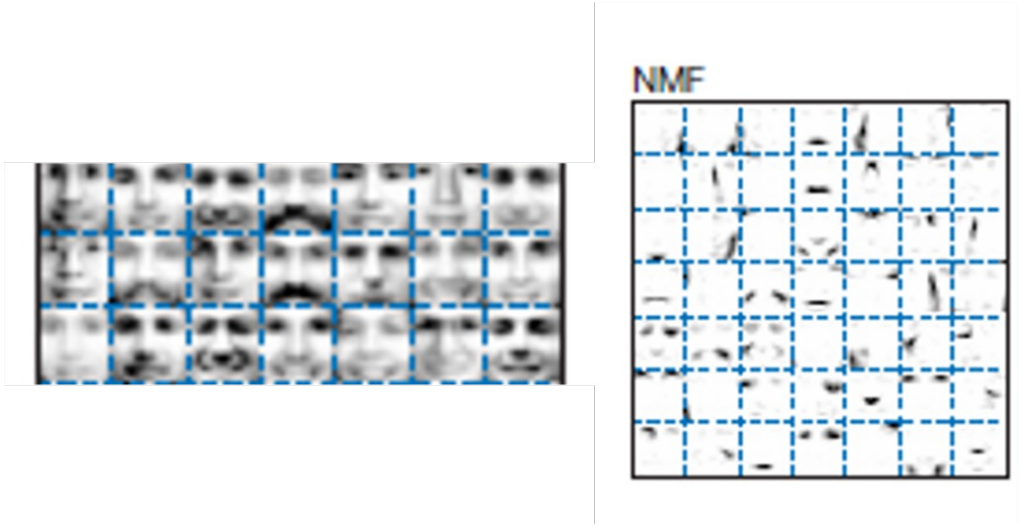
Non-negative matrix factorization

Mathematical models allows the *un-mixing* and the extraction of mutational signatures

Learning the parts of objects by non-negative matrix factorization

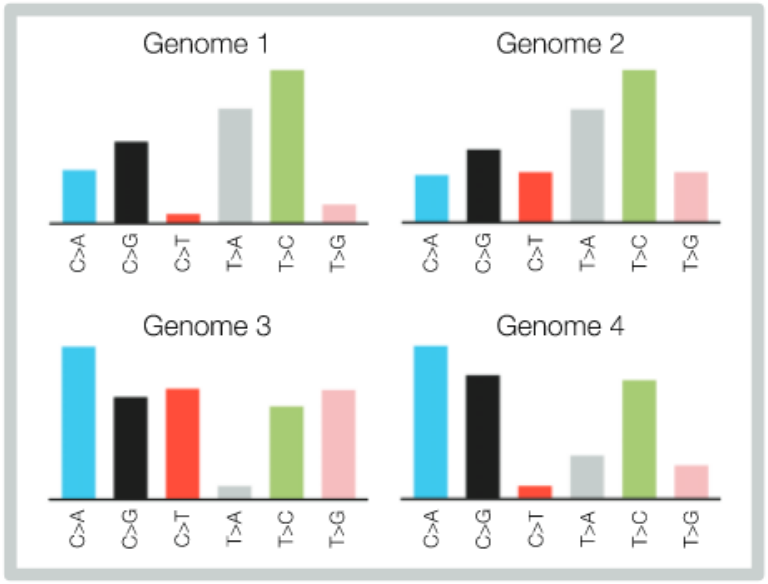
Daniel D. Lee* & H. Sebastian Seung*†

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



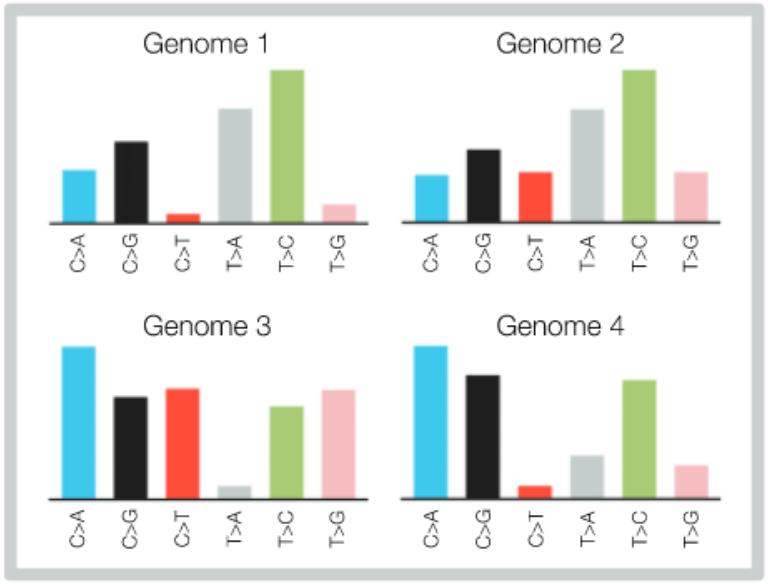
Lee & Seung 1999 Nature

Mathematical models allows the *un-mixing* and the extraction of mutational signatures

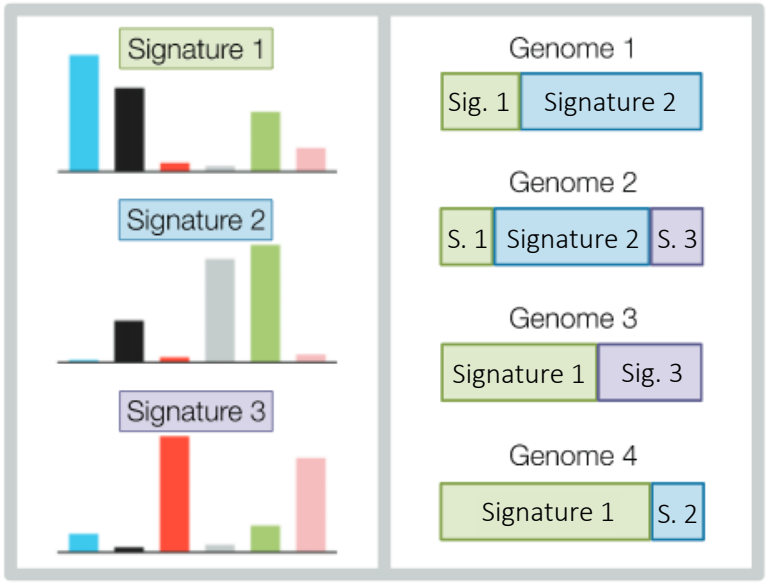


Large number of samples

Mathematical models allows the *un-mixing* and the extraction of mutational signatures

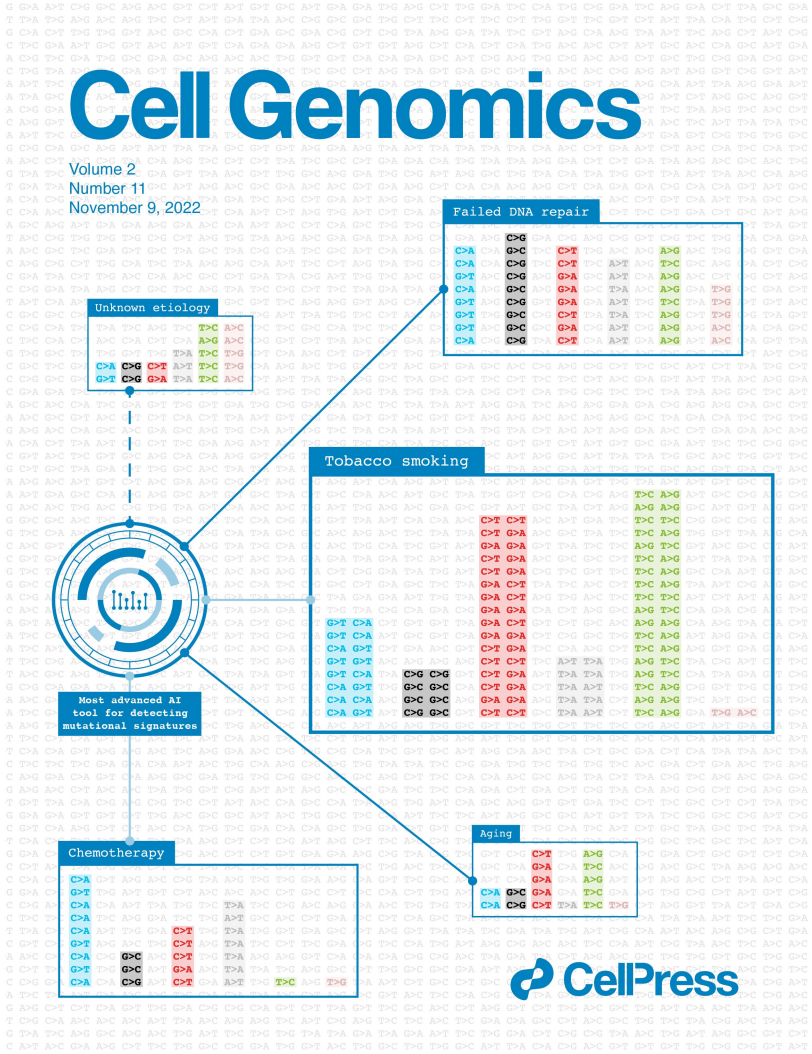


Large number of samples

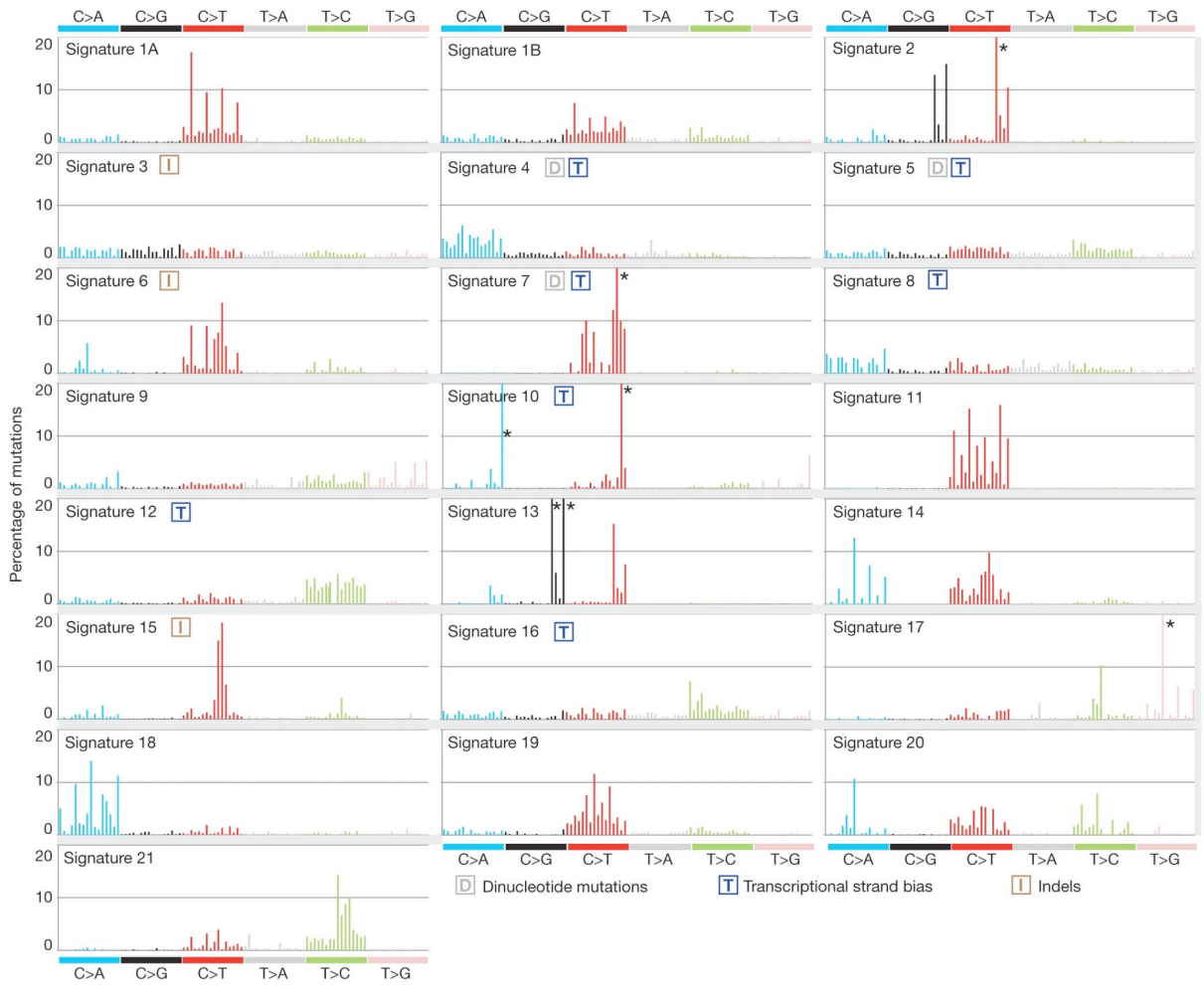


Set of consensus signatures

Mathematical models allows the *un-mixing* and the extraction of mutational signatures



Reference mutational signatures have been extracted from thousands of samples



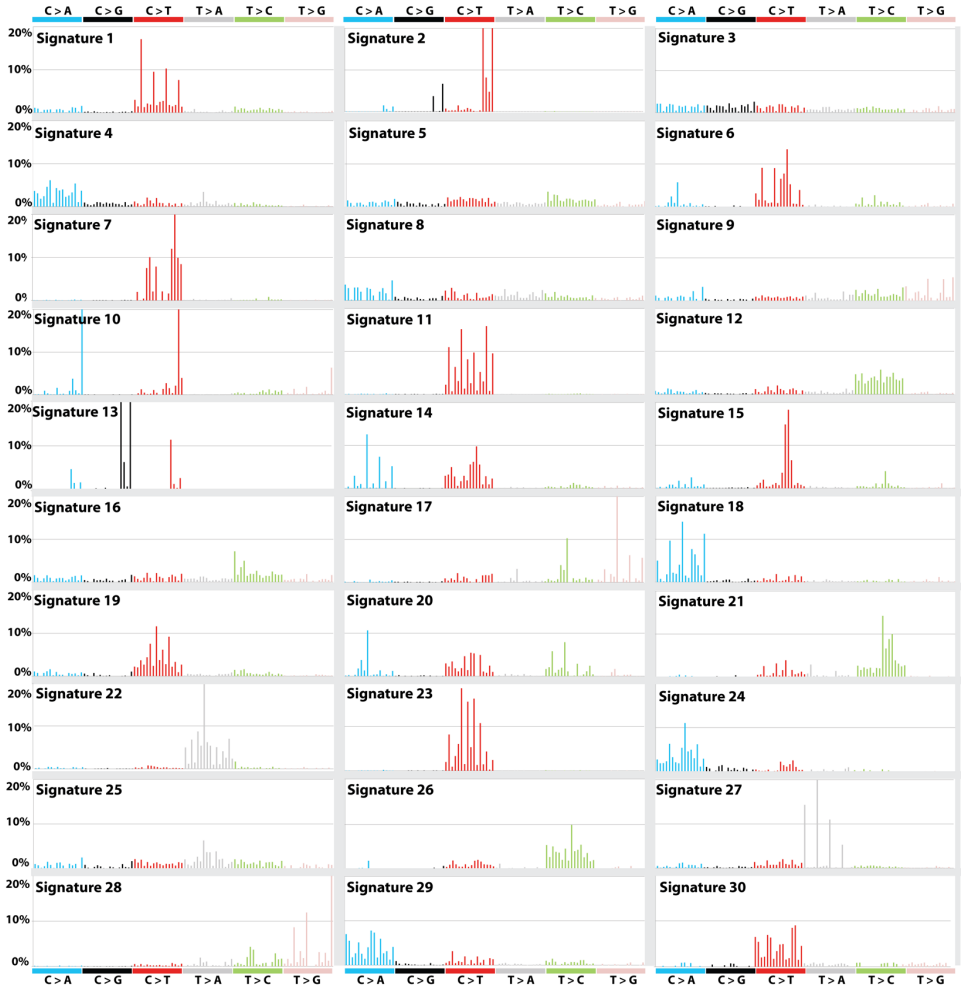
TCGA

v1 (August 2013)

- 21 SBS signatures

Alexandrov *et al.* 2013 Nature

Reference mutational signatures have been extracted from thousands of samples



TCGA

v1 (August 2013)

- 21 SBS signatures

v2 (March 2015)

- 30 SBS signatures

Reference mutational signatures have been extracted from thousands of samples

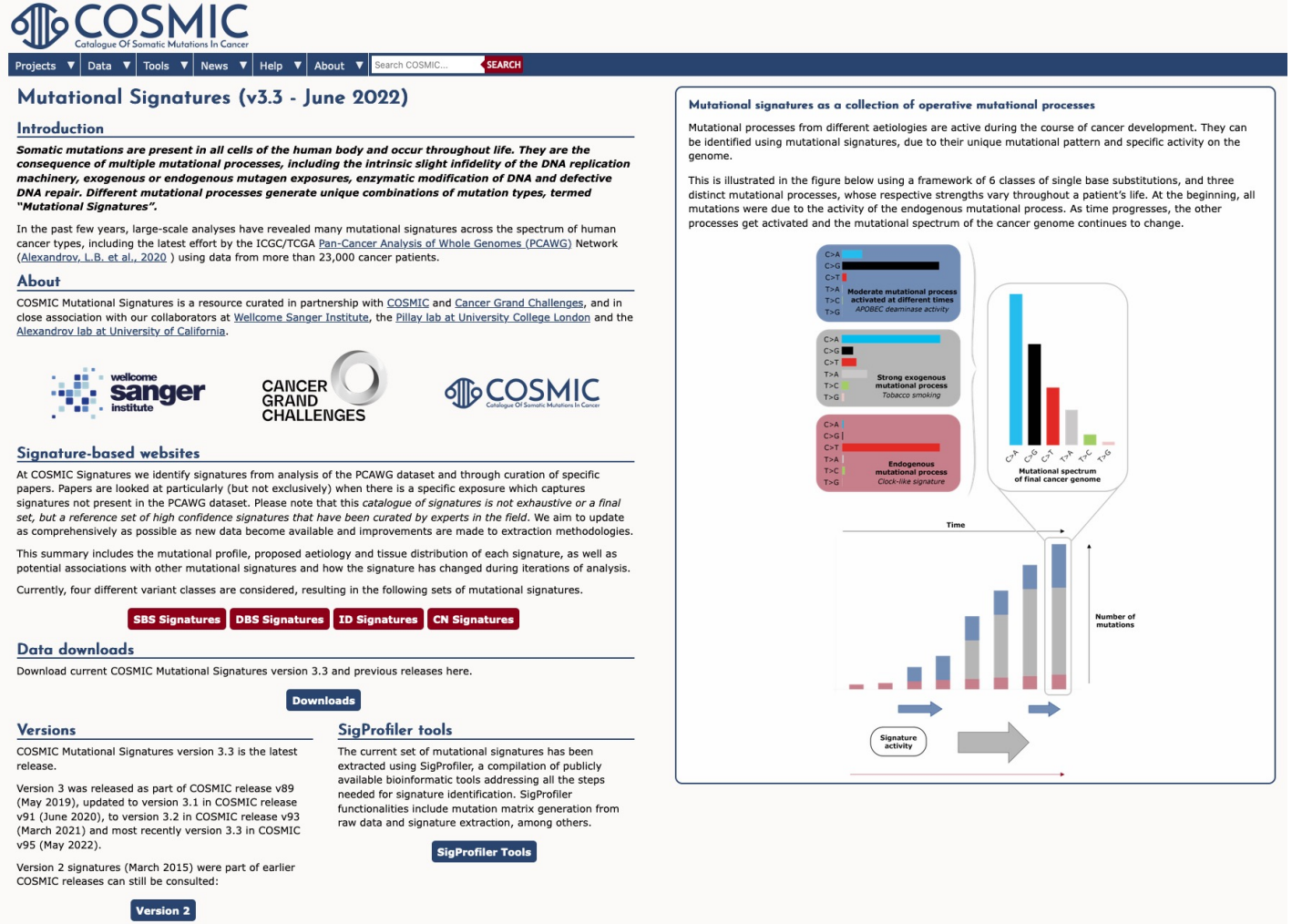
PCAWG
PanCancer Analysis of WHOLE GENOMES

v3 (May 2019)

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

Alexandrov *et al.* 2020 Nature

Reference mutational signatures have been extracted from thousands of samples



COSMIC
Catalogue Of Somatic Mutations In Cancer

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.

wellcome sanger institute | CANCER GRAND CHALLENGES | COSMIC

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

SigProfiler tools

The current set of mutational signatures has been extracted using SigProfiler, a compilation of publicly available bioinformatic tools addressing all the steps needed for signature identification. SigProfiler functionalities include mutation matrix generation from raw data and signature extraction, among others.

SigProfiler Tools

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
Catalogue Of Somatic Mutations In Cancer

Current set (v3.3)

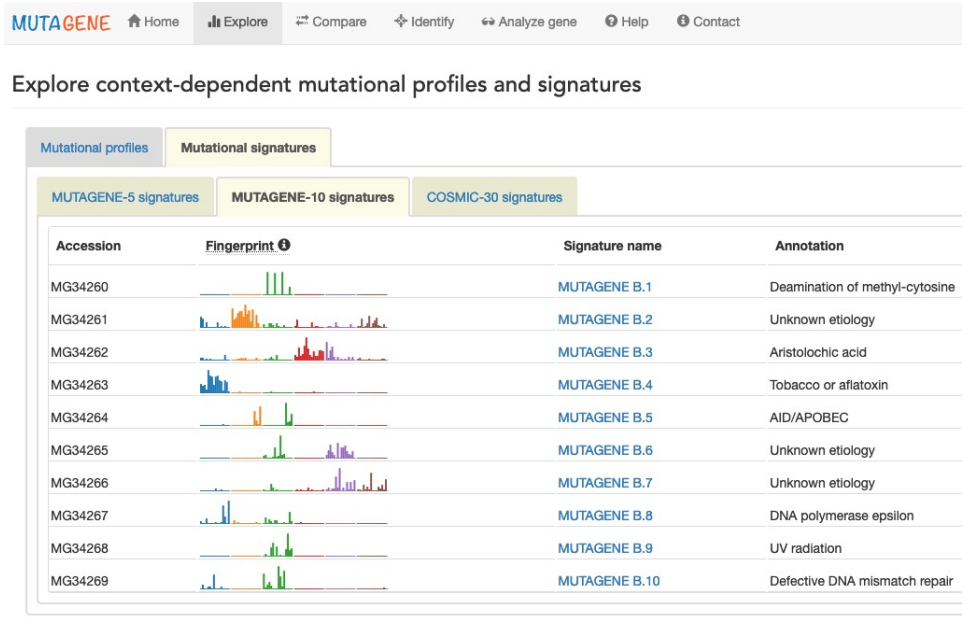
- 79 SBS signatures
- 11 DBS signatures
- 18 ID signatures
- 21 CN signatures

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

Reference mutational signatures have been extracted from thousands of samples

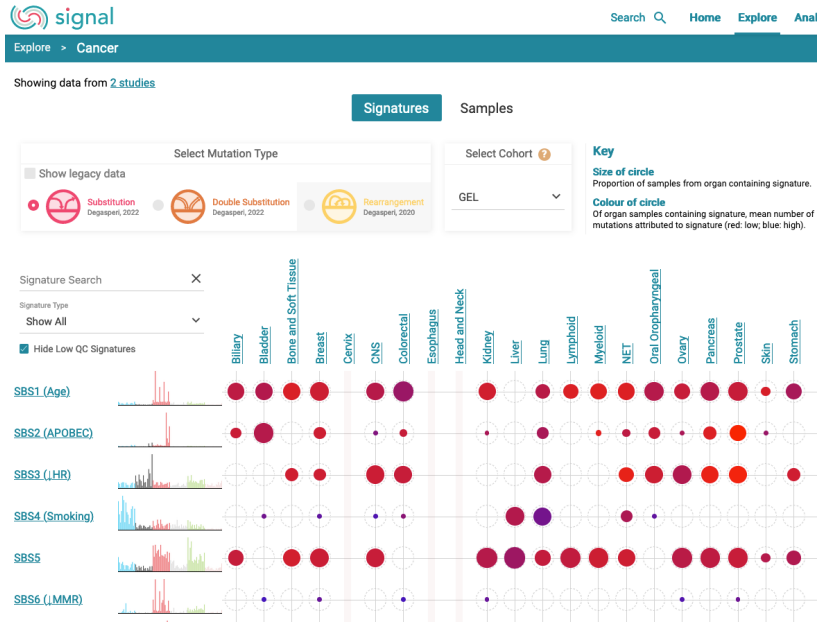
Other reference databases exist that include different variant classes

MUTAGENE



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

signal



https://signal.mutationalsignatures.com/

Day 4. Cancer Genome Analysis - Latin America and the Caribbean

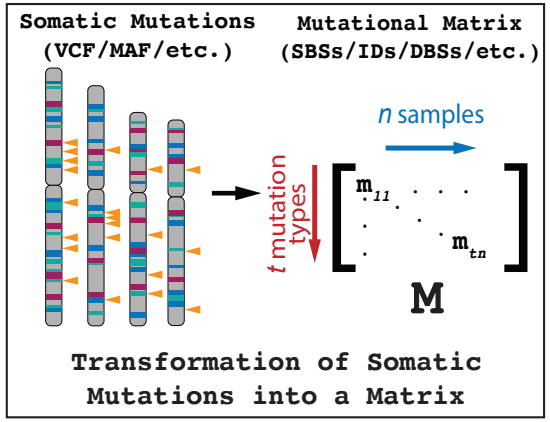
SigProfilerExtractor

Marcos Díaz-Gay

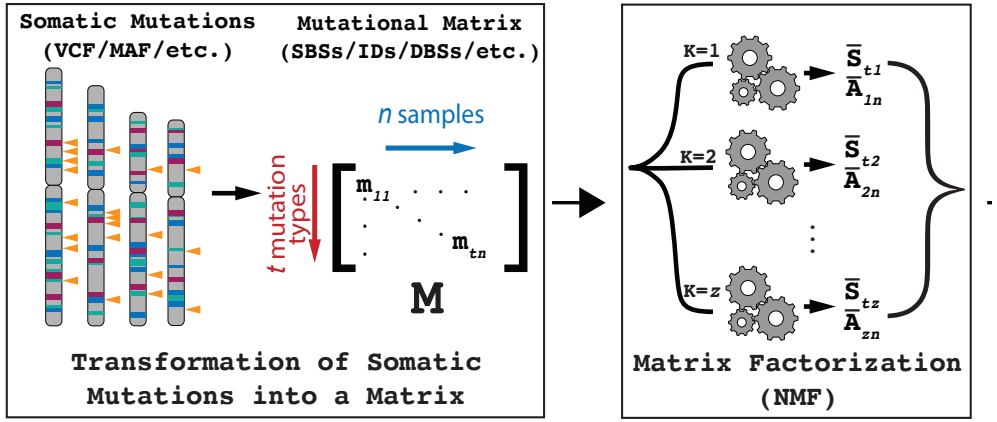
UC San Diego

wellcome connecting science

SigProfilerExtractor uses NMF for identifying *de novo* mutational signatures

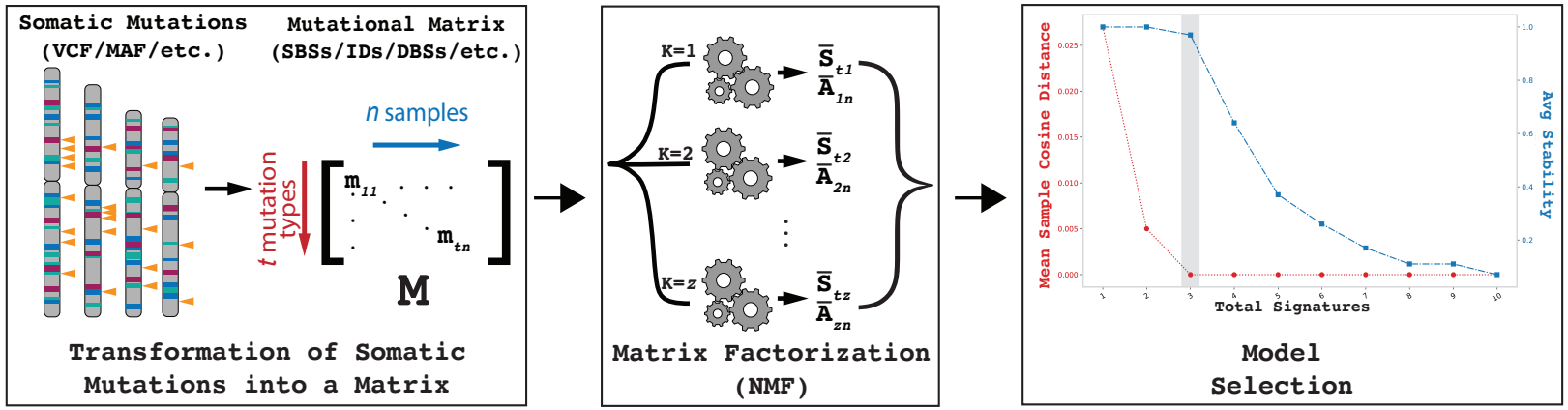


SigProfilerExtractor uses NMF for identifying *de novo* mutational signatures



$$\underset{\text{txn}}{\mathbf{M}} = \underset{\text{txk}}{\mathbf{S}} \times \underset{\text{kxn}}{\mathbf{A}}$$

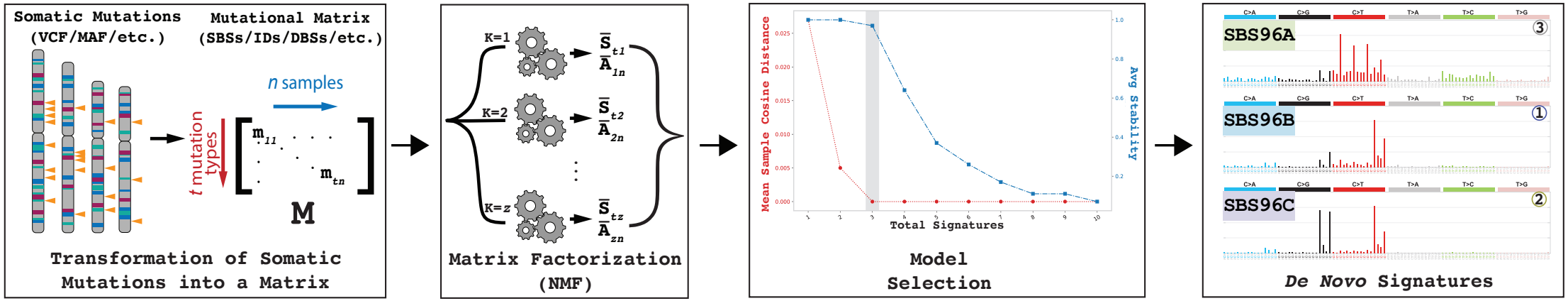
SigProfilerExtractor uses NMF for identifying *de novo* mutational signatures



$$\mathbf{M} = \mathbf{S} \times \mathbf{A}$$

t x n t x k k x n

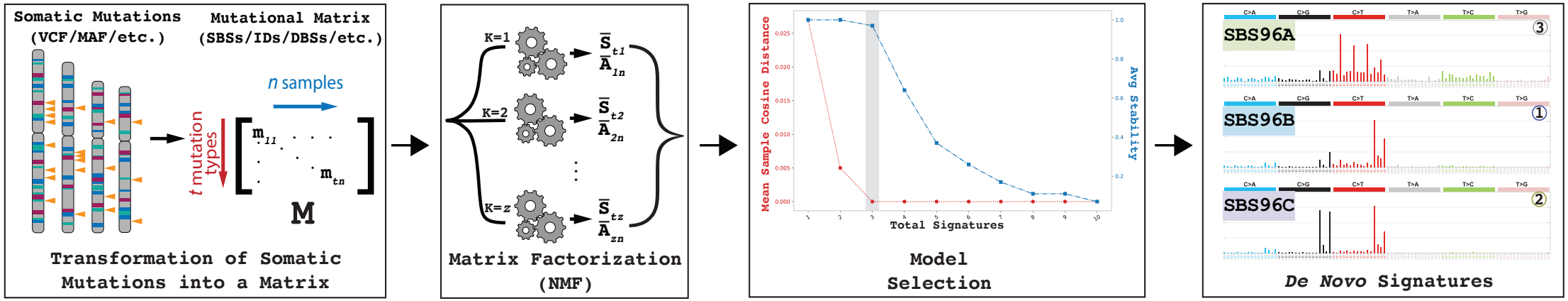
SigProfilerExtractor uses NMF for identifying *de novo* mutational signatures



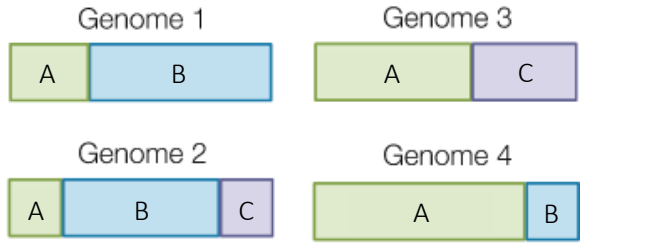
$$\mathbf{M} = \mathbf{S} \times \mathbf{A}$$

txn txk kxn

SigProfilerExtractor uses NMF for identifying *de novo* mutational signatures

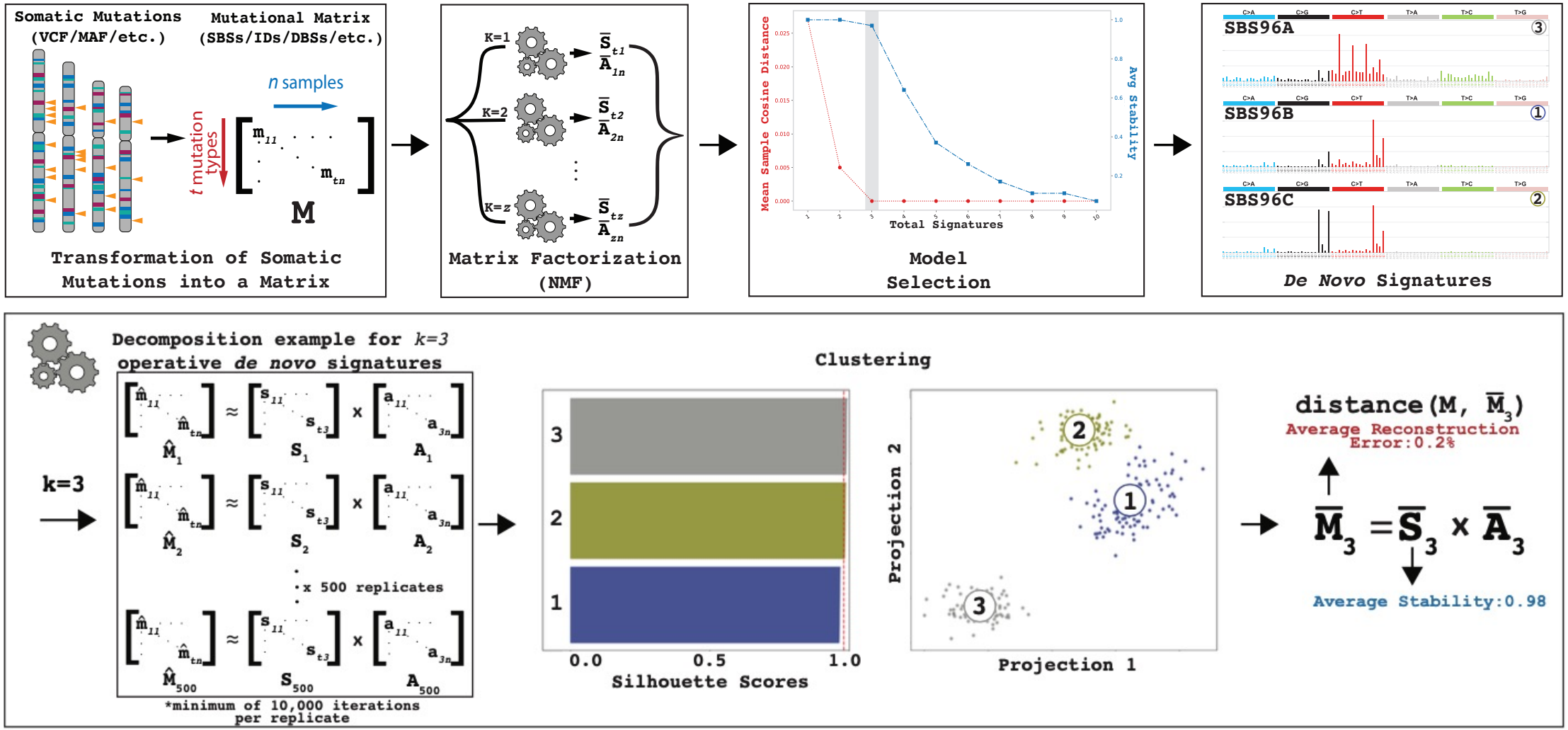


$$\underset{txn}{\mathbf{M}} = \underset{txk}{\mathbf{S}} \times \underset{kxn}{\mathbf{A}}$$

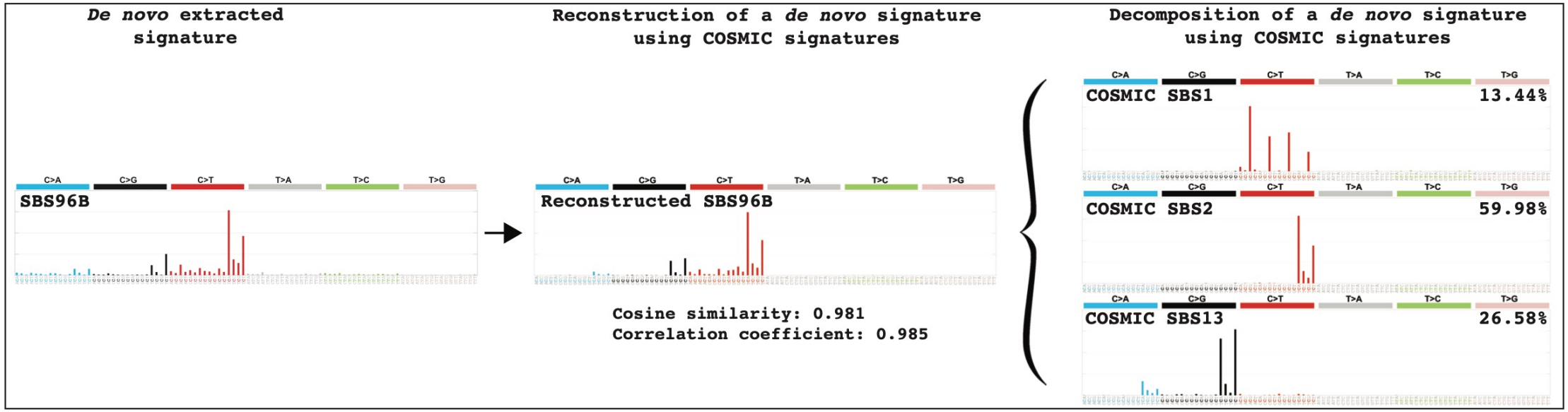


SigProfiler Assignment

SigProfilerExtractor uses NMF for identifying *de novo* mutational signatures



SigProfilerExtractor decomposes *de novo* mutational signatures using COSMIC reference signatures



Different tools are currently available to perform *de novo* signature extraction

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

Different tools are currently available to perform *de novo* signature extraction

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

Different tools are currently available to perform *de novo* signature extraction

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

Ground truth signatures allow evaluation of tool performance on synthetic data

	Extracted Signature A	Extracted Signature B	Extracted Signature C	Extracted Signature D
Ground Truth Signature 1	0.14	0.98	0.56	0.36
Ground Truth Signature 2	0.35	0.29	0.93	0.46
Ground Truth Signature 3	0.31	0.56	0.78	0.66
Ground Truth Signature 4	0.34	0.08	0.57	0.67
Ground Truth Signature 5	0.95	0.15	0.81	0.39
Ground Truth Signature 6	0.23	0.74	0.48	0.26

Ground truth signatures allow evaluation of tool performance on synthetic data

	Extracted Signature A	Extracted Signature B	Extracted Signature C	Extracted Signature D
Ground Truth Signature 1	0.14	0.98	0.56	0.36
Ground Truth Signature 2	0.35	0.29	0.93	0.46
Ground Truth Signature 3	0.31	0.56	0.78	0.66
Ground Truth Signature 4	0.34	0.08	0.57	0.67
Ground Truth Signature 5	0.95	0.15	0.81	0.39
Ground Truth Signature 6	0.23	0.74	0.48	0.26

Cosine similarity between Extracted Signature C and Ground Truth Signature 6

Ground truth signatures allow evaluation of tool performance on synthetic data

	Extracted Signature A	Extracted Signature B	Extracted Signature C	Extracted Signature D
Ground Truth Signature 1	0.14	**0.98**	0.56	0.36
Ground Truth Signature 2	0.35	0.29	**0.93**	0.46
Ground Truth Signature 3	0.31	0.56	0.78	0.66
Ground Truth Signature 4	0.34	0.08	0.57	0.67
Ground Truth Signature 5	**0.95**	0.15	0.81	0.39
Ground Truth Signature 6	0.23	0.74	0.48	0.26

True Positives (TP; ≥ 0.90)

Extracted Signature A
Extracted Signature B
Extracted Signature C

Signatures correctly extracted from the *dataset*

False Positives (FP)

Extracted Signature D

Signatures extracted but absent in the *dataset*

False Negatives (FN)

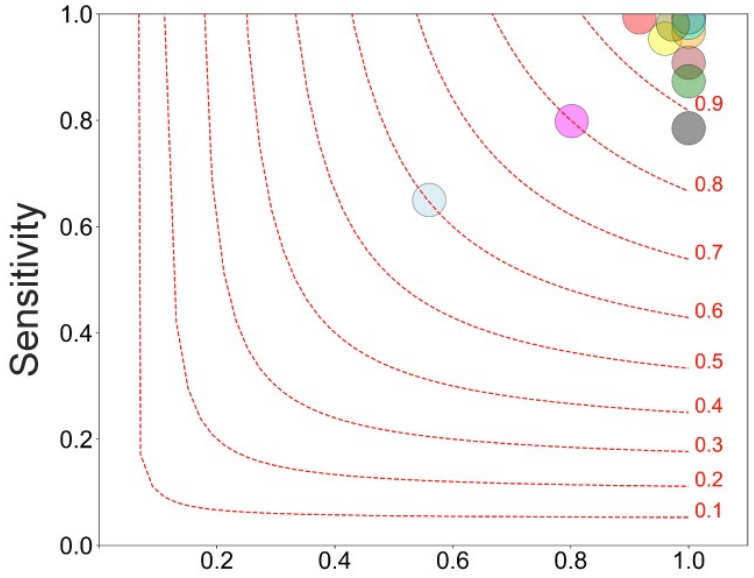
Ground Truth Signature 3
Ground Truth Signature 4
Ground Truth Signature 6

Signatures not extracted but used in simulating the *dataset*

SigProfilerExtractor outperformed other approaches in WGS noiseless data, specially in hard scenarios

Easy (3-5 signatures)

~7% human cancer types

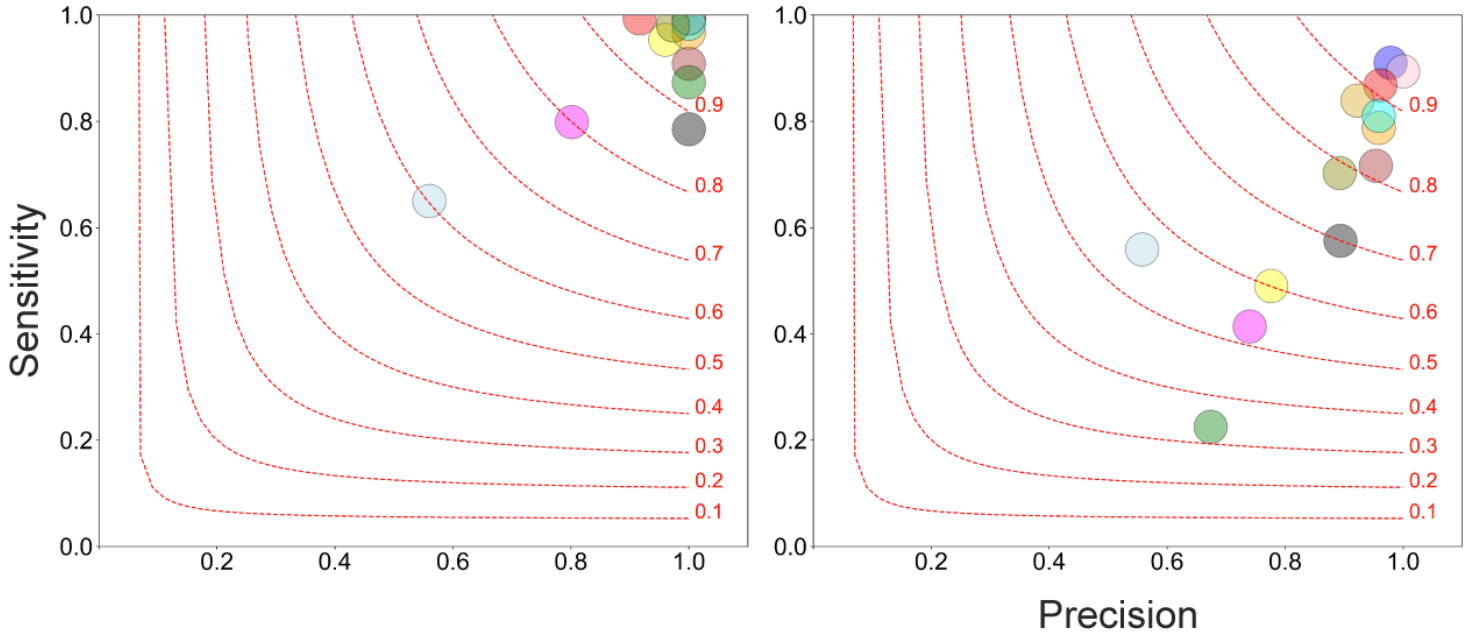


SigProfilerExtractor
SignatureAnalyzer
SigProfiler_PCAWG
SigneR
MutationalPatterns
SomaticSignatures
MutSpec
SignatureToolsLib
SigMiner
MutSignatures
Maftools
EMu
SigFit
TensorSignatures

SigProfilerExtractor outperformed other approaches in WGS noiseless data, specially in hard scenarios

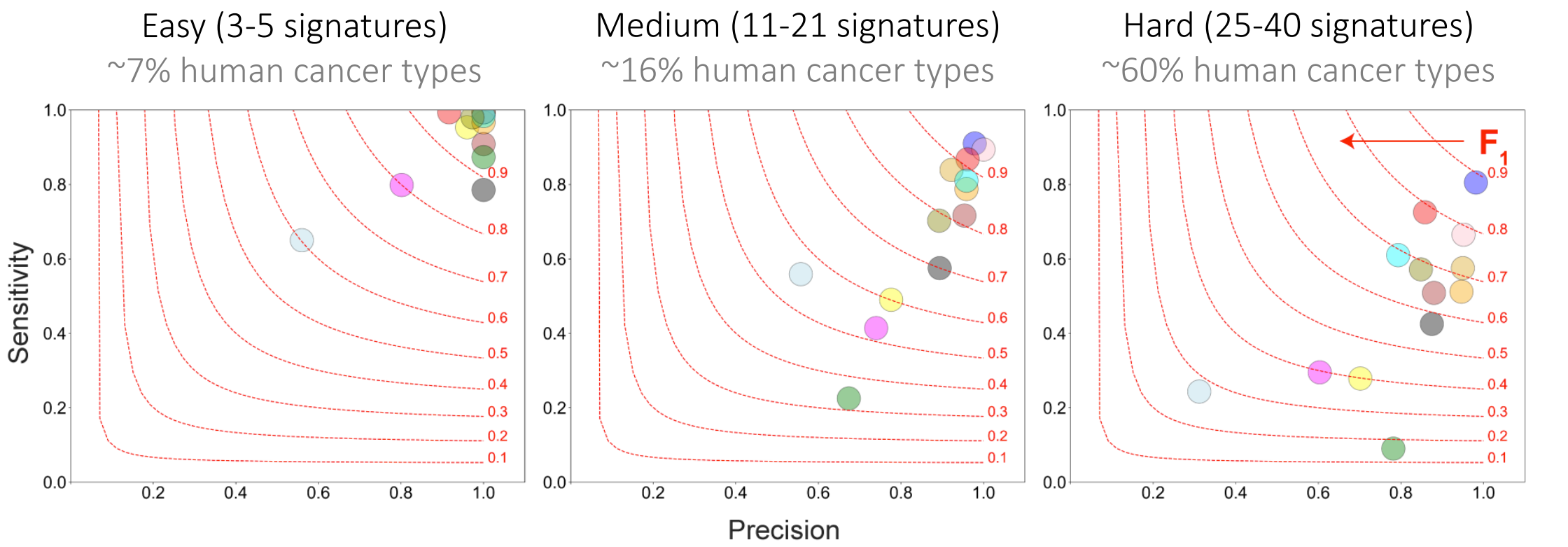
Easy (3-5 signatures)
~7% human cancer types

Medium (11-21 signatures)
~16% human cancer types



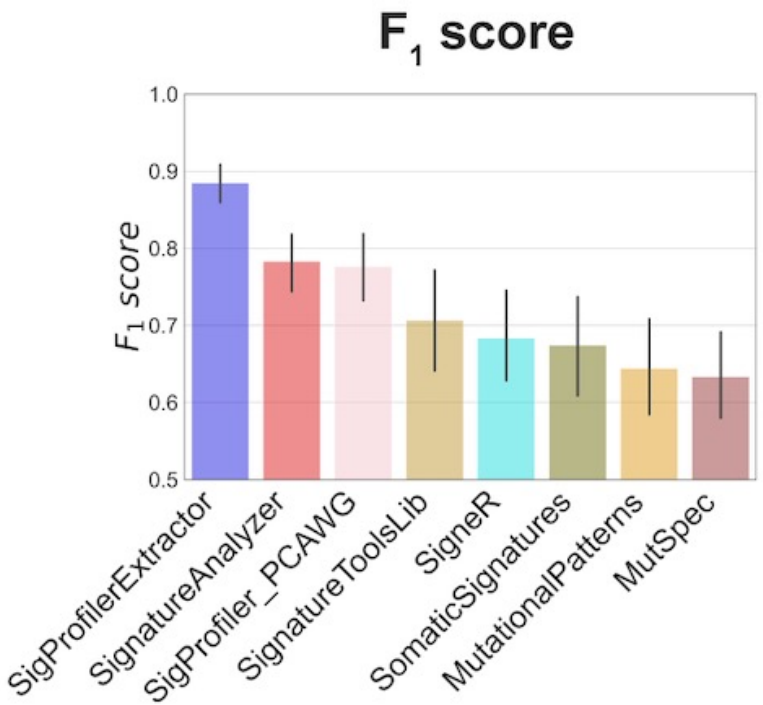
SigProfilerExtractor
SignatureAnalyzer
SigProfiler_PCAWG
SigneR
MutationalPatterns
SomaticSignatures
MutSpec
SignatureToolsLib
SigMiner
MutSignatures
Maftools
EMu
SigFit
TensorSignatures

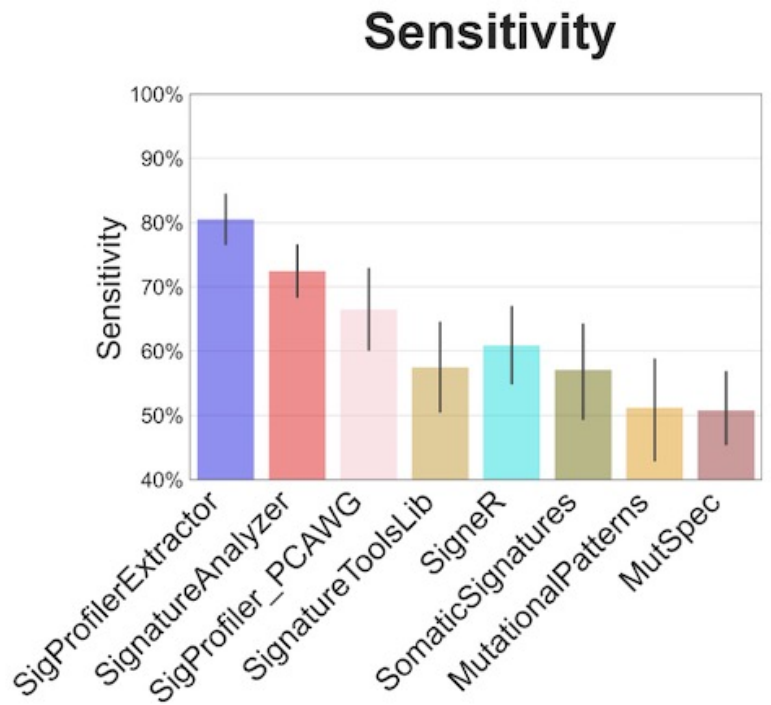
SigProfilerExtractor outperformed other approaches in WGS noiseless data, specially in hard scenarios

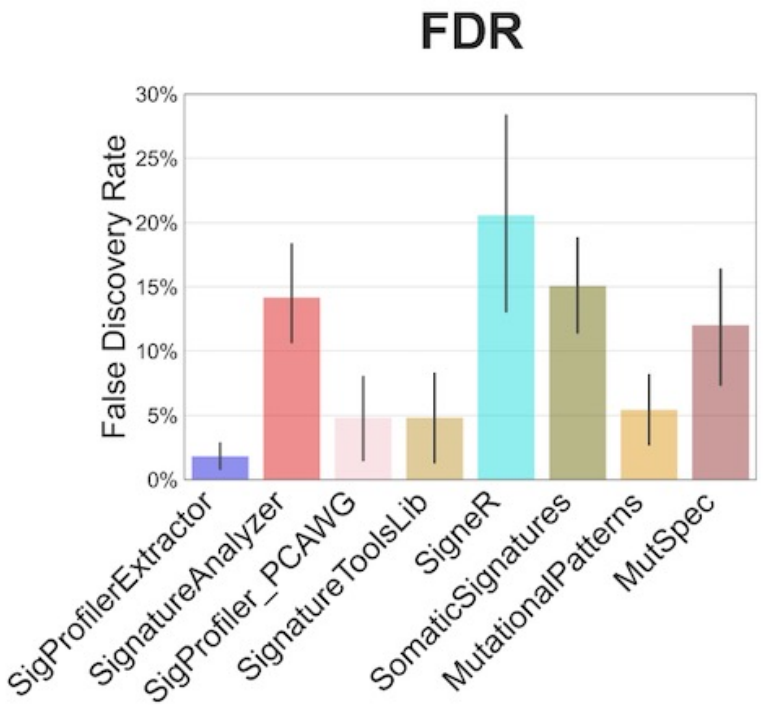


SigProfilerExtractor
SignatureAnalyzer
SigProfiler_PCAWG
SigneR
MutationalPatterns
SomaticSignatures
MutSpec
SignatureToolsLib
SigMiner
MutSignatures
Maftools
EMu
SigFit
TensorSignatures

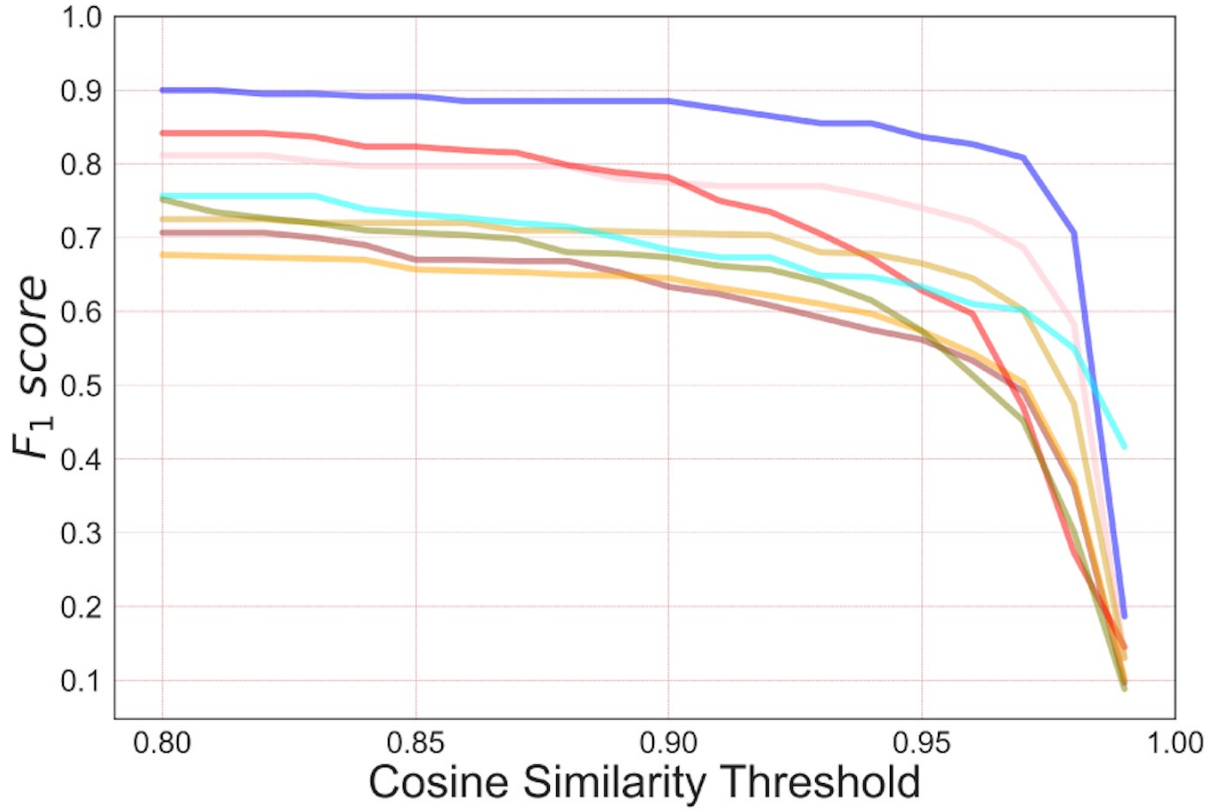
SigProfilerExtractor outperformed other approaches in **WGS noiseless** data, specially in **hard** scenarios





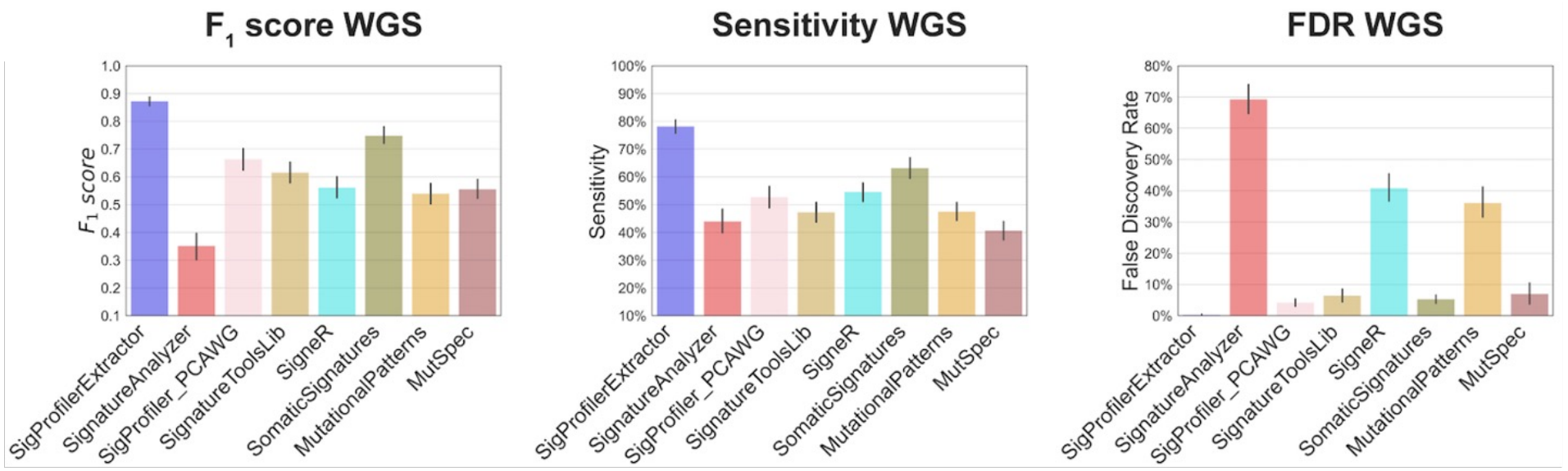


SigProfilerExtractor outperformed other approaches in **WGS noiseless** data, specially in **hard** scenarios



SigProfilerExtractor
SigProfiler_PCAWG
SigneR
MutationalPatterns
SignatureAnalyzer
SignatureToolsLib
SomaticSignatures
MutSpec

SigProfilerExtractor outperformed other tools in both WGS and WES realistic scenarios (5% noise)



Useful links – SigProfilerExtractor

- Publication
  http://dx.doi.org/10.1016/j.xgen.2022.100179
- GitHub repository (python package)
  https://github.com/AlexandrovLab/SigProfilerExtractor
- GitHub repository (R wrapper)
  https://github.com/AlexandrovLab/SigProfilerExtractorR
- Wiki page (usage instructions)
  https://osf.io/t6j7u/wiki/home/