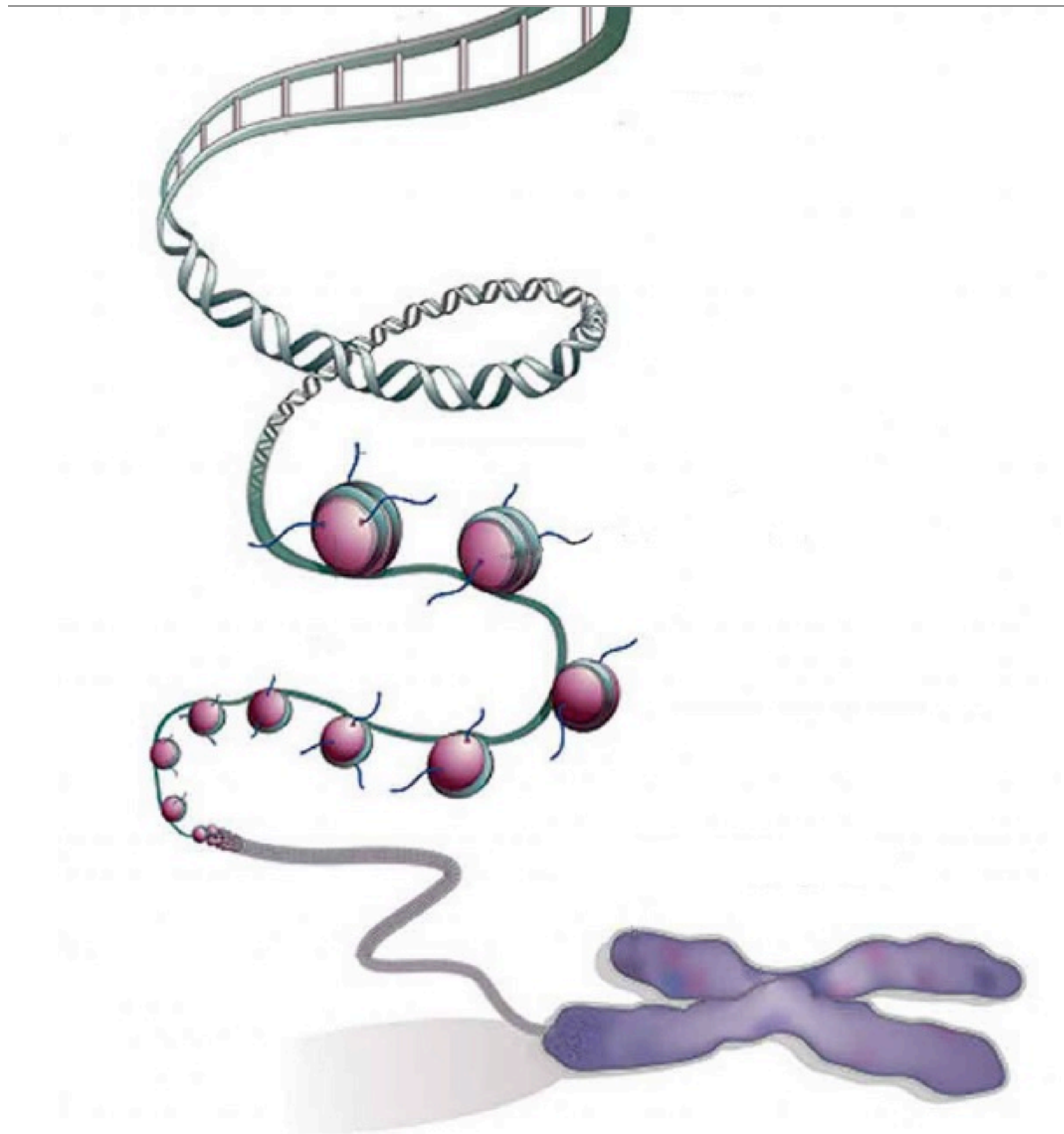
Matrix decomposition in DNA sequence analysis



Peter J Park
Department of Biomedical Informatics
Harvard Medical School

November 11, 2023

CC

Calvin Cunningham

6 hours ago

If **Ted Pick** was a real math whiz, he would stop trying to stuff his pockets and attempt to solve the Navier–Stokes Equation.

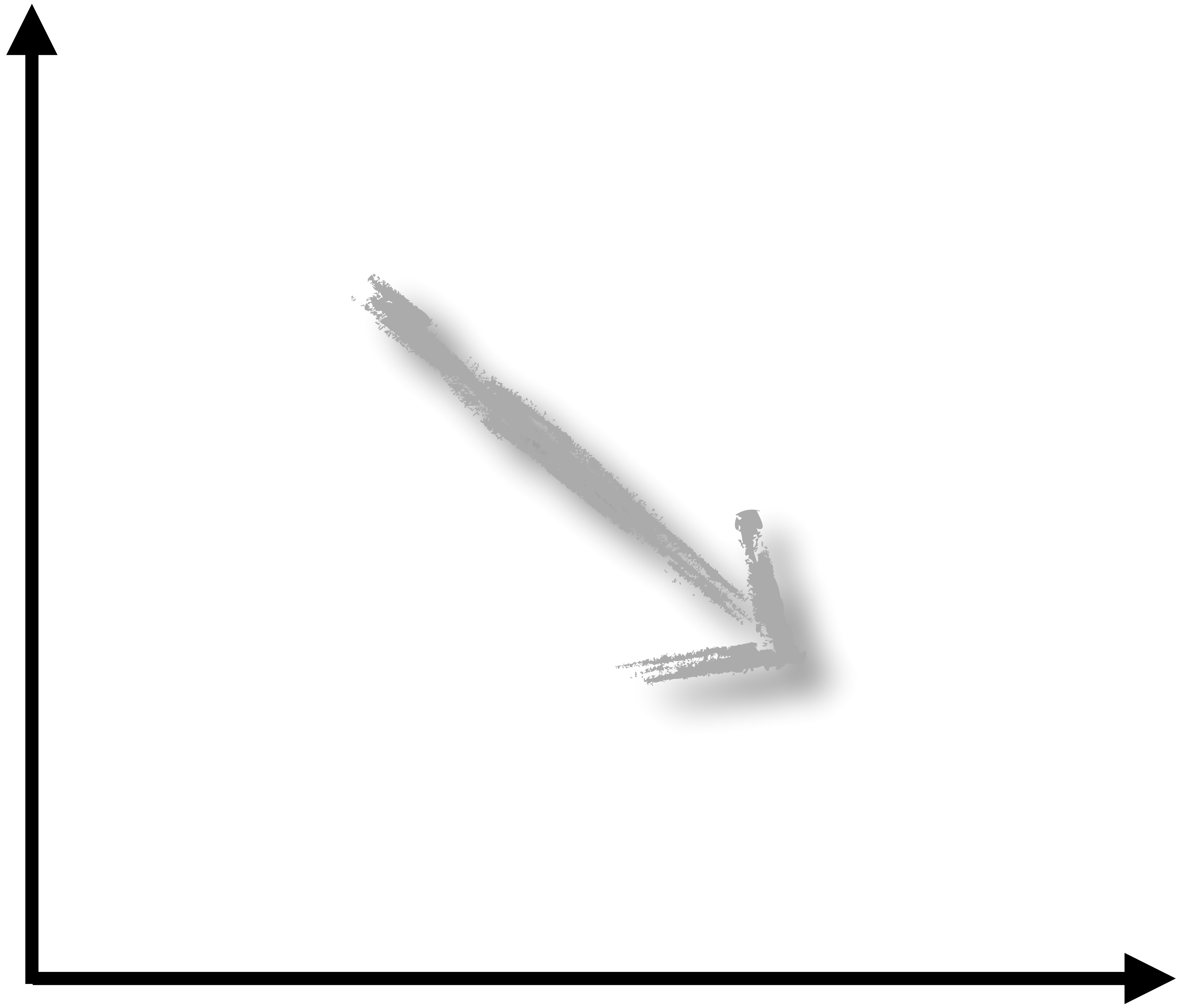
Reply · 👍 1 · Share

Ted Pick Is a Math Whiz Among Math Whizzes. He's the New Morgan Stanley CEO

The company lifer will have to keep the wealth-management unit happy, while doing the same for investment banking and trading

Wall Street Journal, Oct 26, 2023

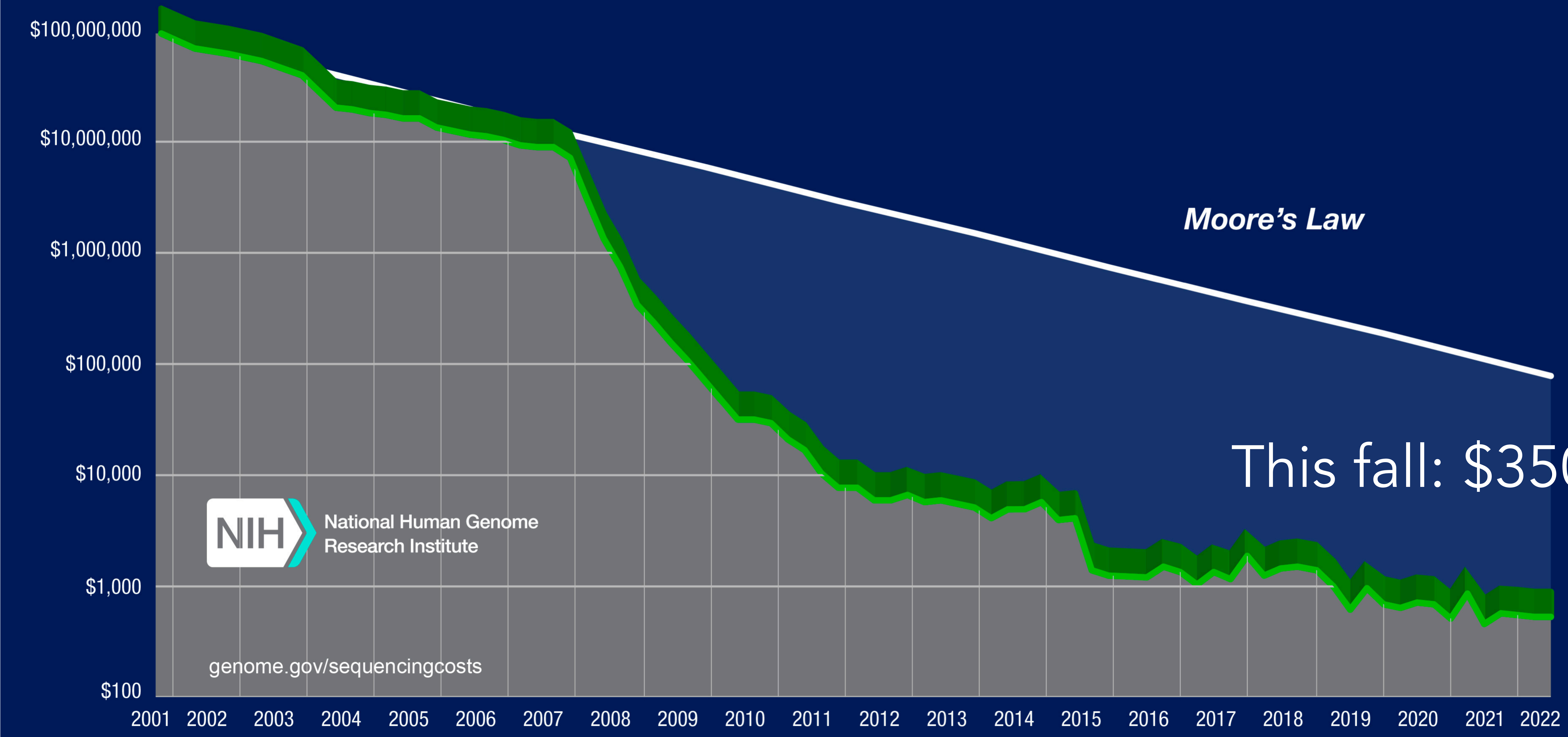
Trade-offs?



beautiful
& deep

messy but useful

Cost per Human Genome



This fall: $350!

NIH National Human Genome Research Institute

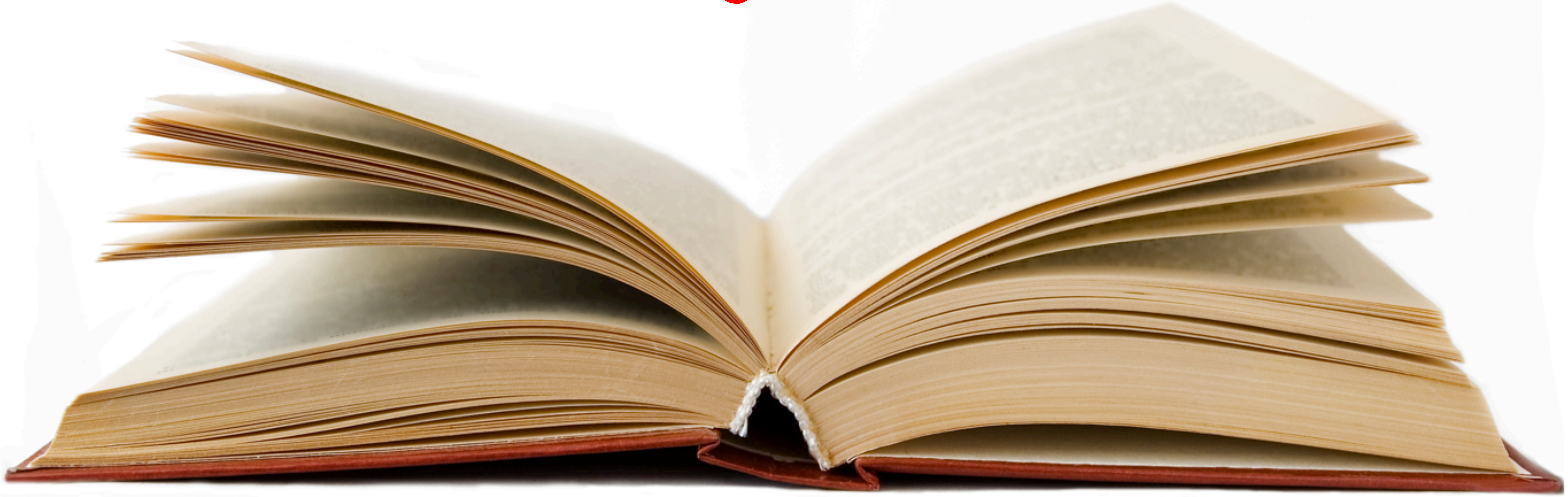
genome.gov/sequencingcosts

Genome sequencing technology

- "Reading" the 3-billion nucleotide sequences in a person's genome
- Four nucleotides: A, C, T, G
- One "whole-genome": ~100 GB
- Latest machine: ~$1M, 128 genomes in 2 days

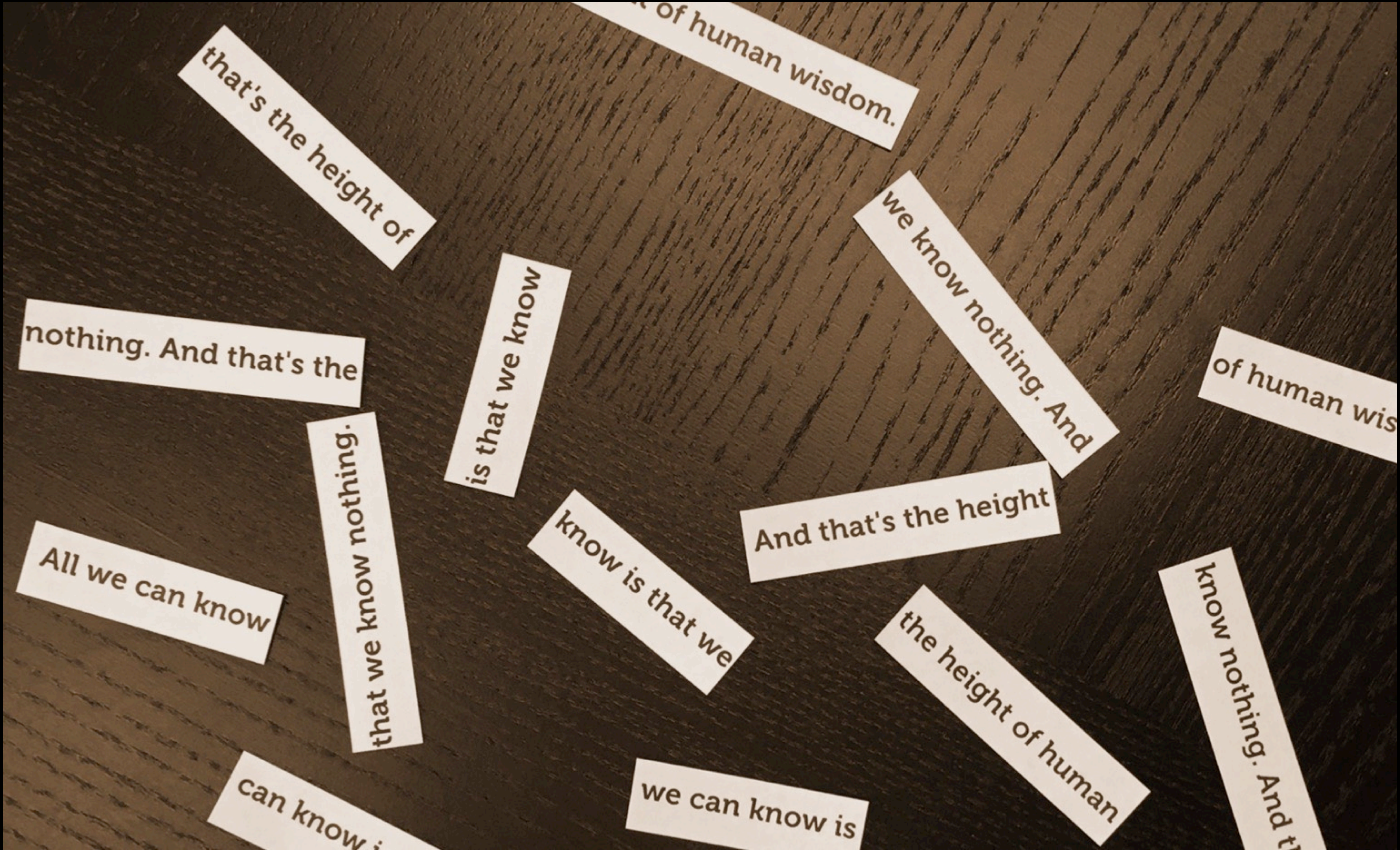
Sequencing the human genome

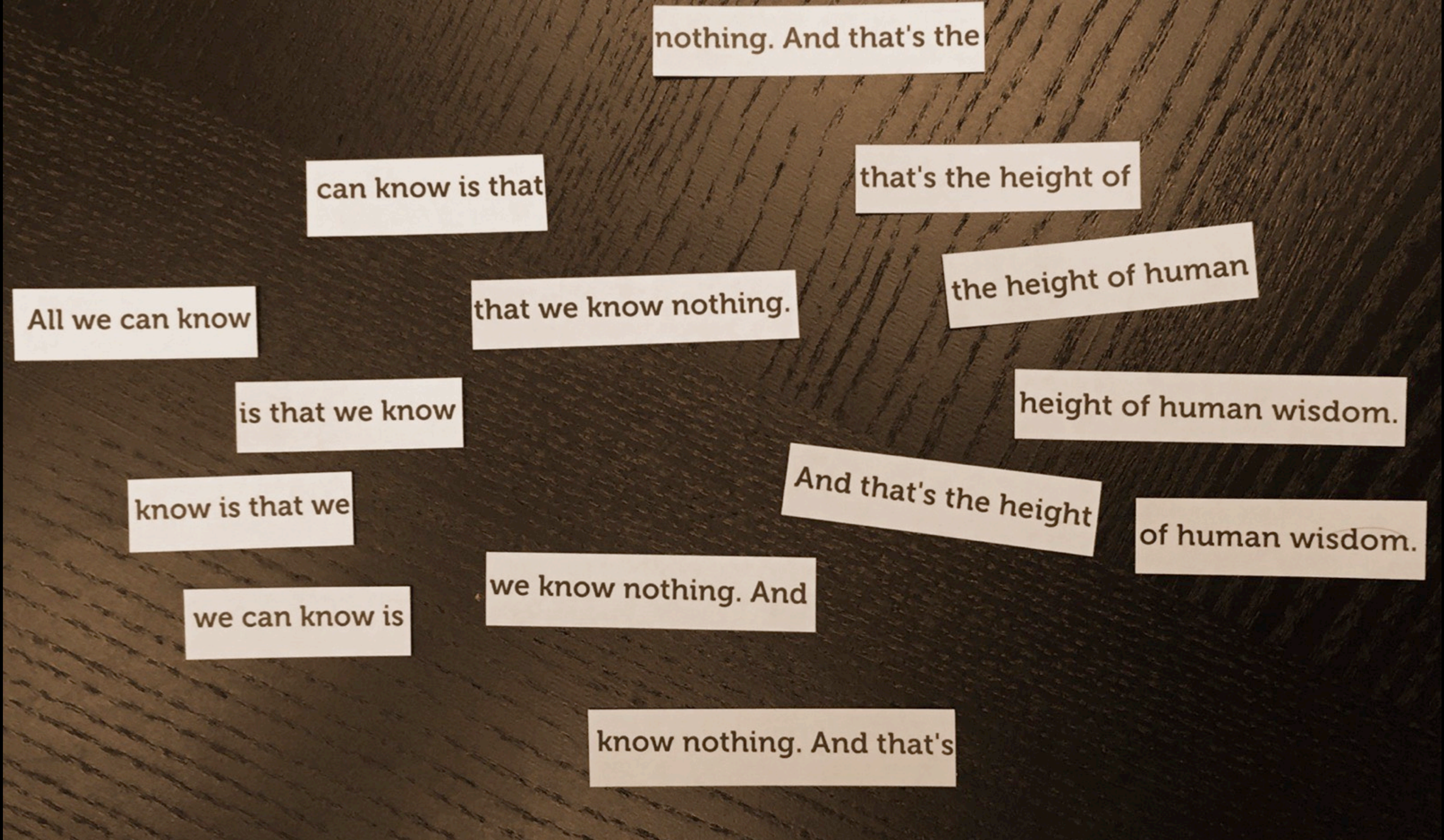
and interpreting



War and Peace: ~3 million characters

Human genome: ~3 billion characters


of human wisdom.
that's the height of
nothing. And that's the
is that we know
we know nothing. And
of human wis
that we know nothing.
And that's the height
know is that we
All we can know
the height of human
know nothing. And th
can know i
we can know is



nothing. And that's the

can know is that

that's the height of

All we can know

that we know nothing.

the height of human

is that we know

height of human wisdom.

know is that we

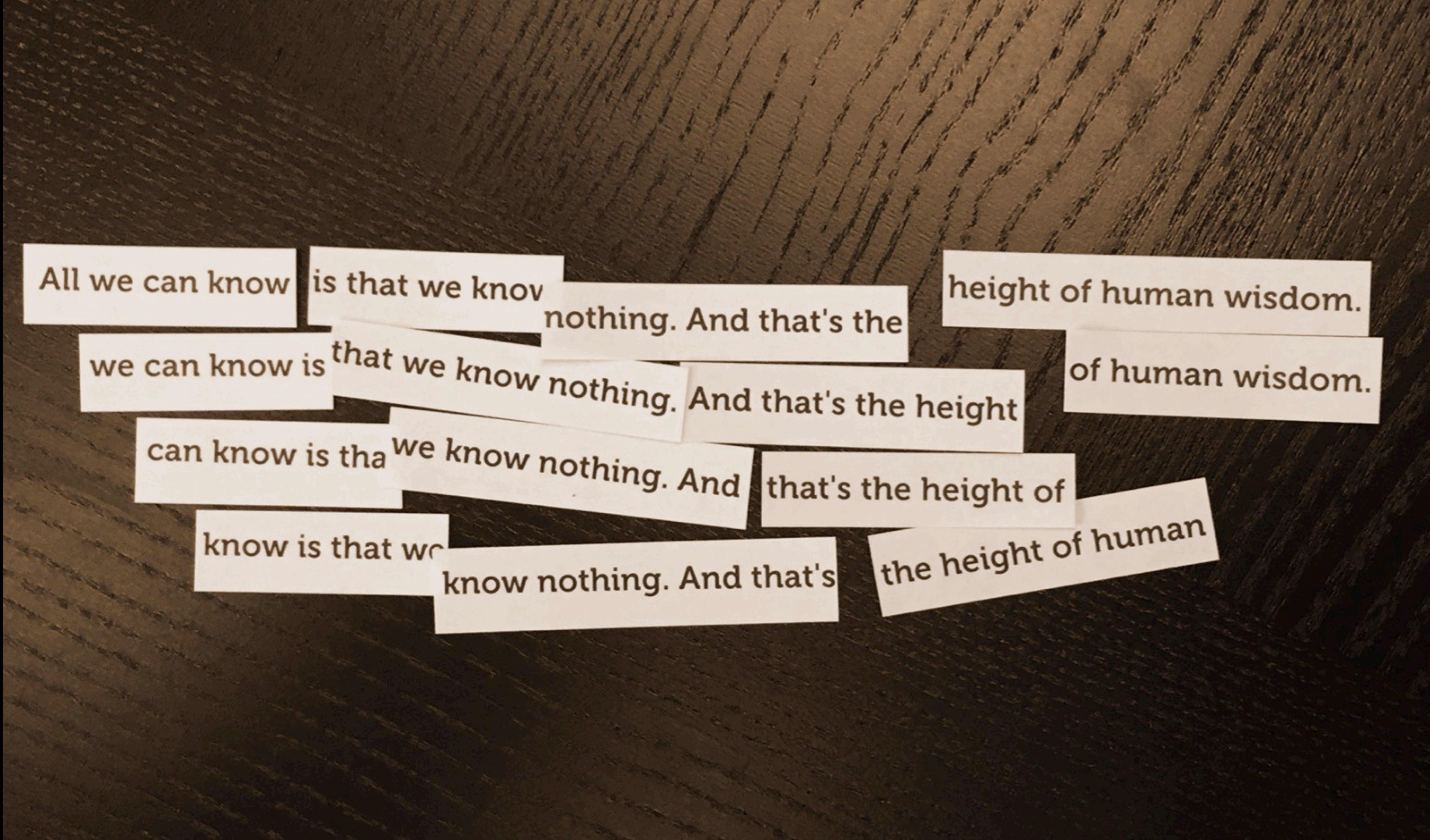
And that's the height

of human wisdom.

we can know is

we know nothing. And

know nothing. And that's

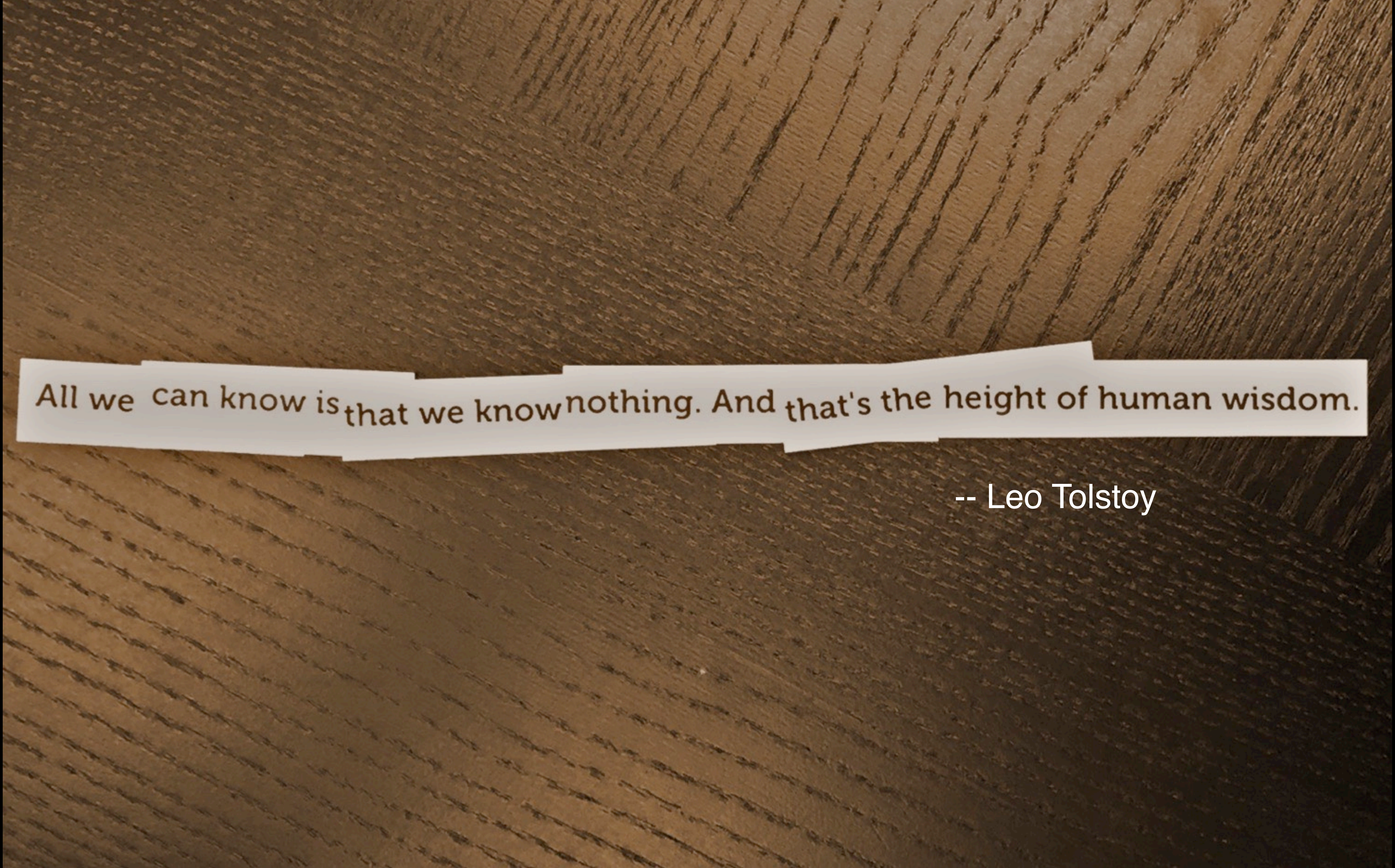


All we can know is that we knov nothing. And that's the height of human wisdom.

we can know is that we know nothing. And that's the height of human wisdom.

can know is tha we know nothing. And that's the height of

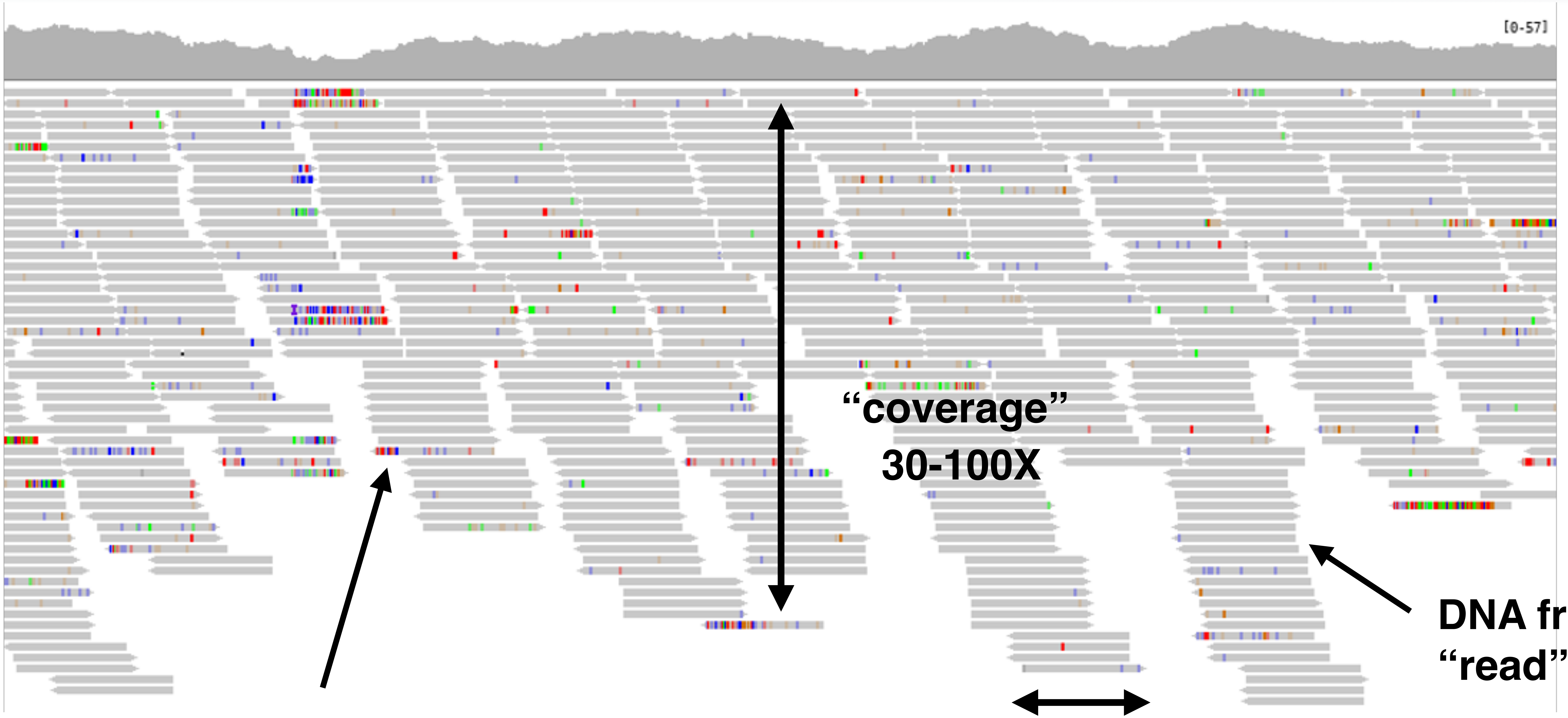
know is that wc know nothing. And that's the height of human



All we can know is that we know nothing. And that's the height of human wisdom.

-- Leo Tolstoy

Genome sequencing data



"**coverage**"
30-100X

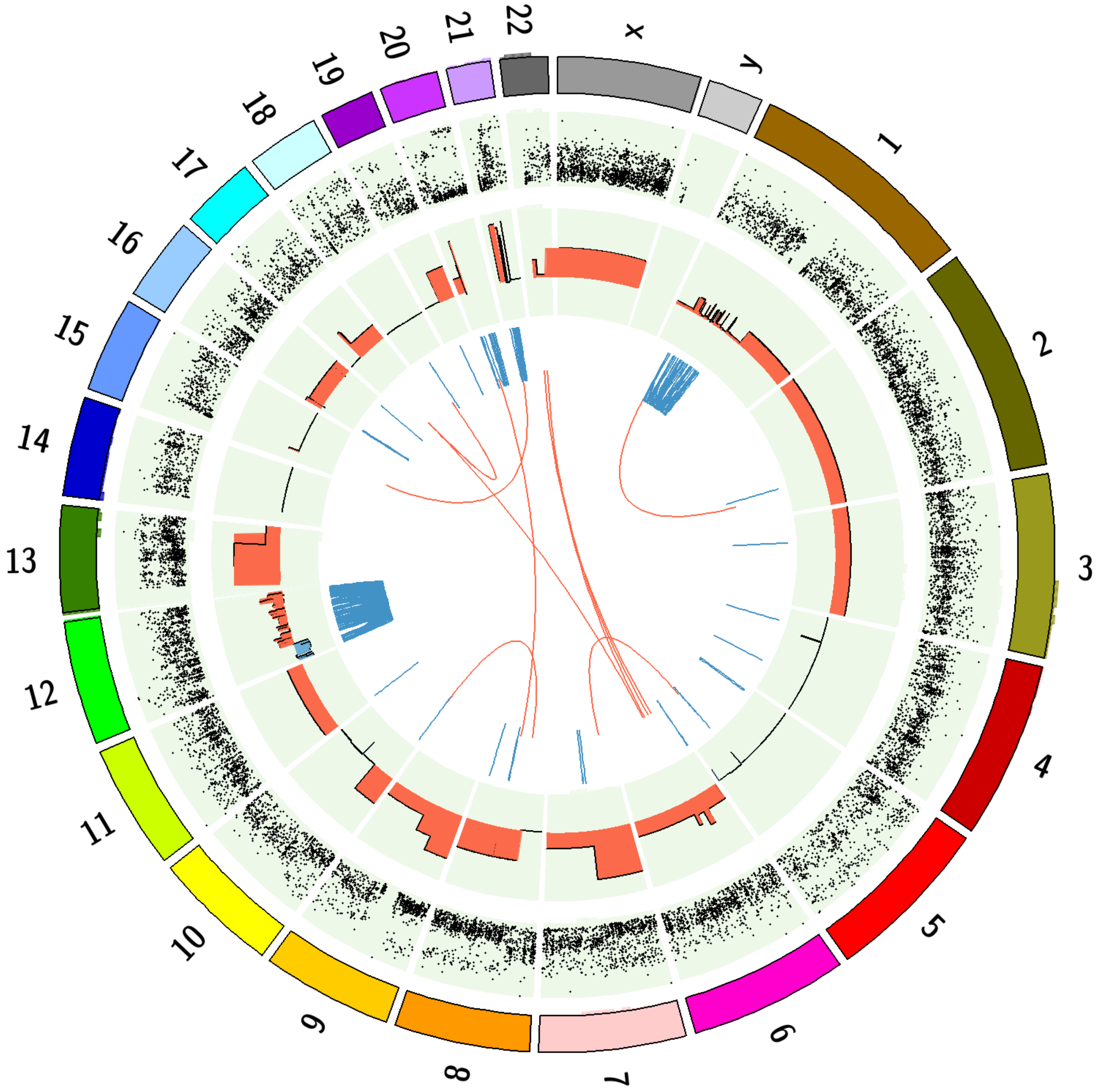
DNA fragment
"read"

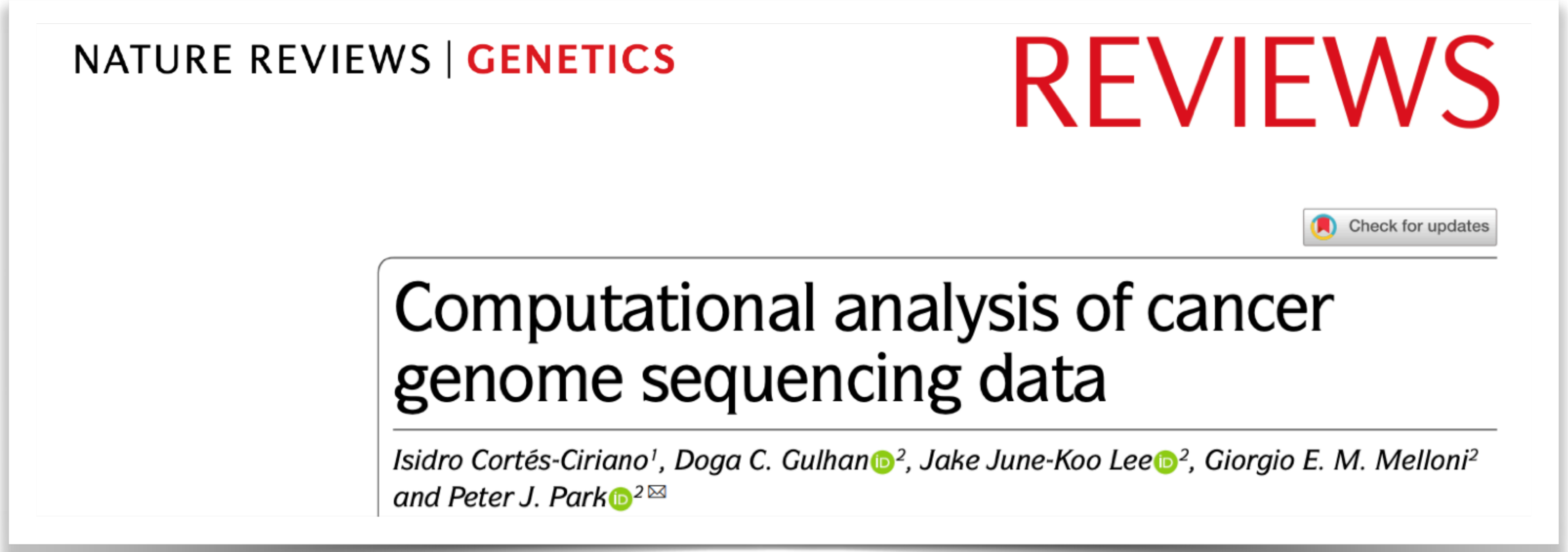
colored dots -
"**mismatch**" to the
reference genome

"read length"
~150bp

Mutational processes in cancer and normal cells

- *How is your genome mutated when you have cancer?*
- *What are the mechanisms generating the mutations?*
- *With whole-genome sequencing, many types of genomic alterations can be detected*





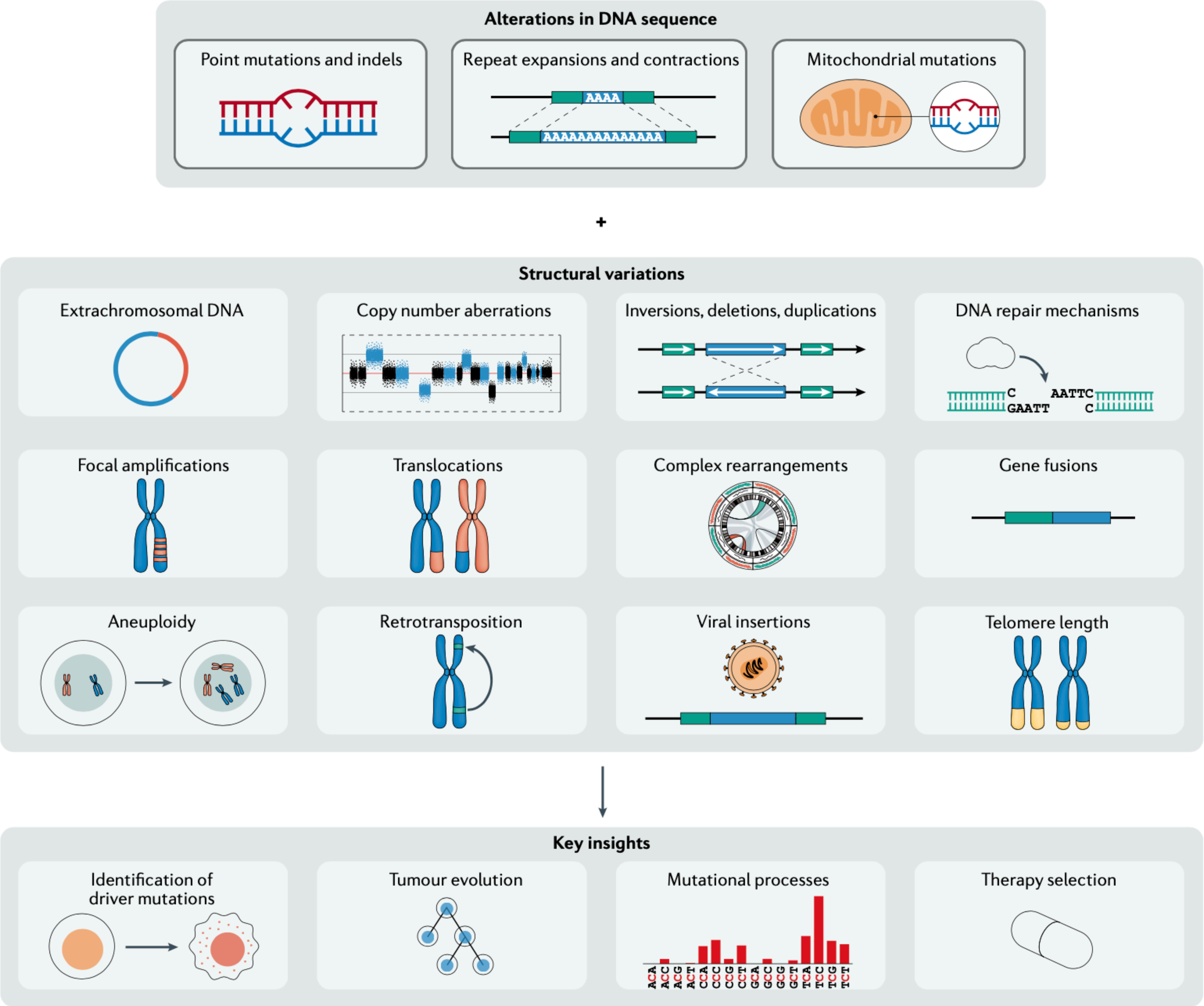
NATURE REVIEWS | GENETICS

REVIEWS

Check for updates

Computational analysis of cancer genome sequencing data

Isidro Cortés-Ciriano[1], Doga C. Gulhan[2], Jake June-Koo Lee[2], Giorgio E. M. Melloni[2] and Peter J. Park[2]✉



Alterations in DNA sequence

Point mutations and indels

Repeat expansions and contractions

Mitochondrial mutations

+

Structural variations

Extrachromosomal DNA

Copy number aberrations

Inversions, deletions, duplications

DNA repair mechanisms

Focal amplifications

Translocations

Complex rearrangements

Gene fusions

Aneuploidy

Retrotransposition

Viral insertions

Telomere length

Key insights

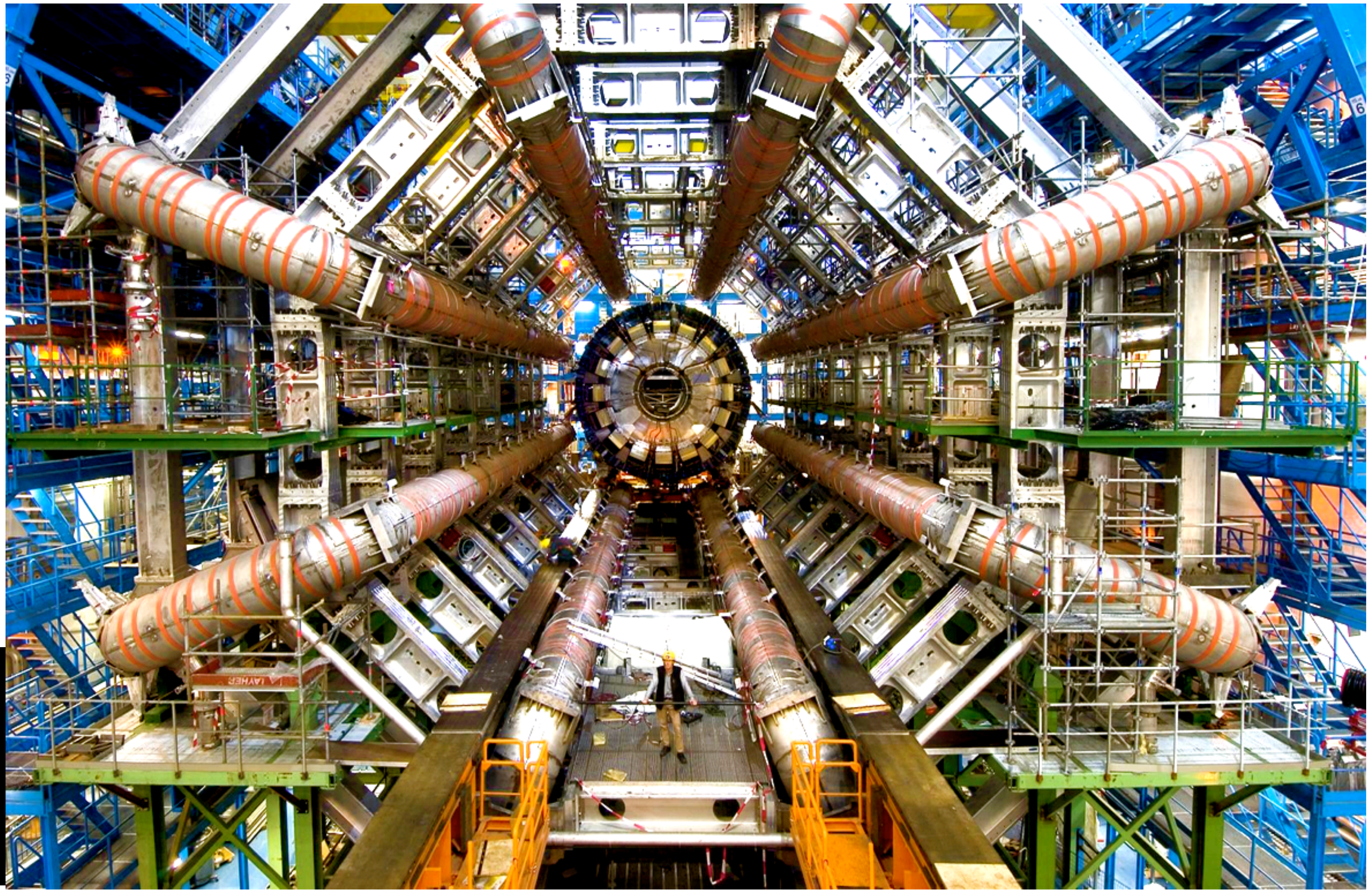
Identification of driver mutations

Tumour evolution

Mutational processes

Therapy selection

Hubble Telescope
50 Terabytes in 20 years



Large Hadron Collider
15 Petabytes in 1 year

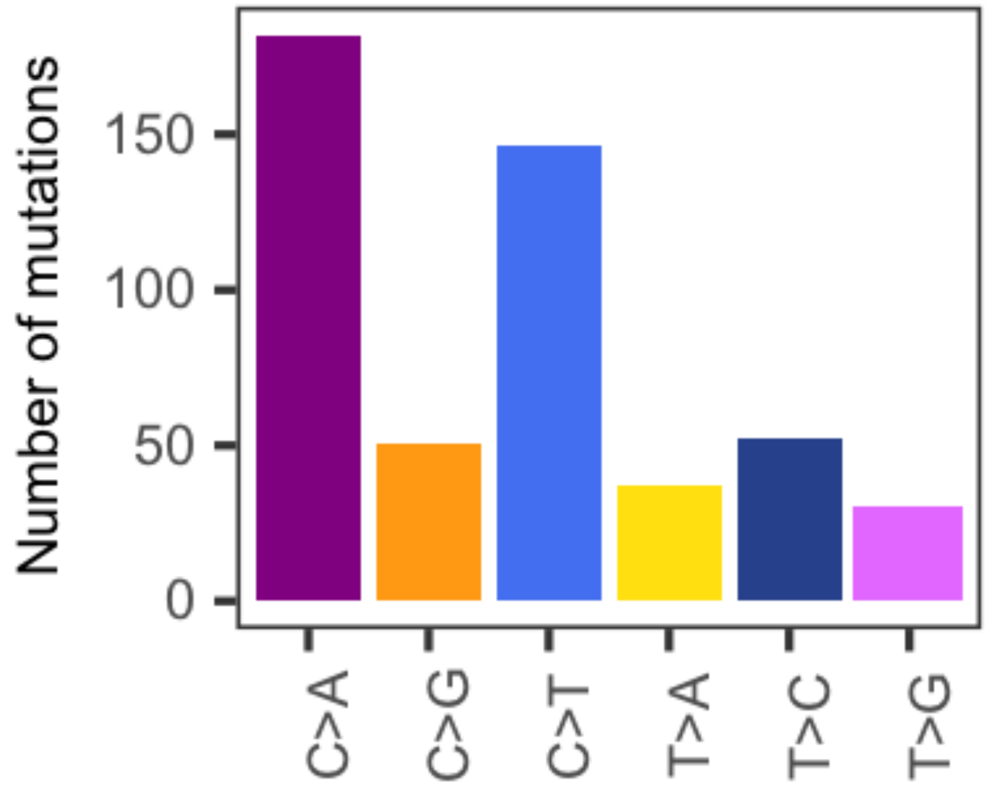
My lab's data: 2.7 PB

Mutational signature analysis - introduction

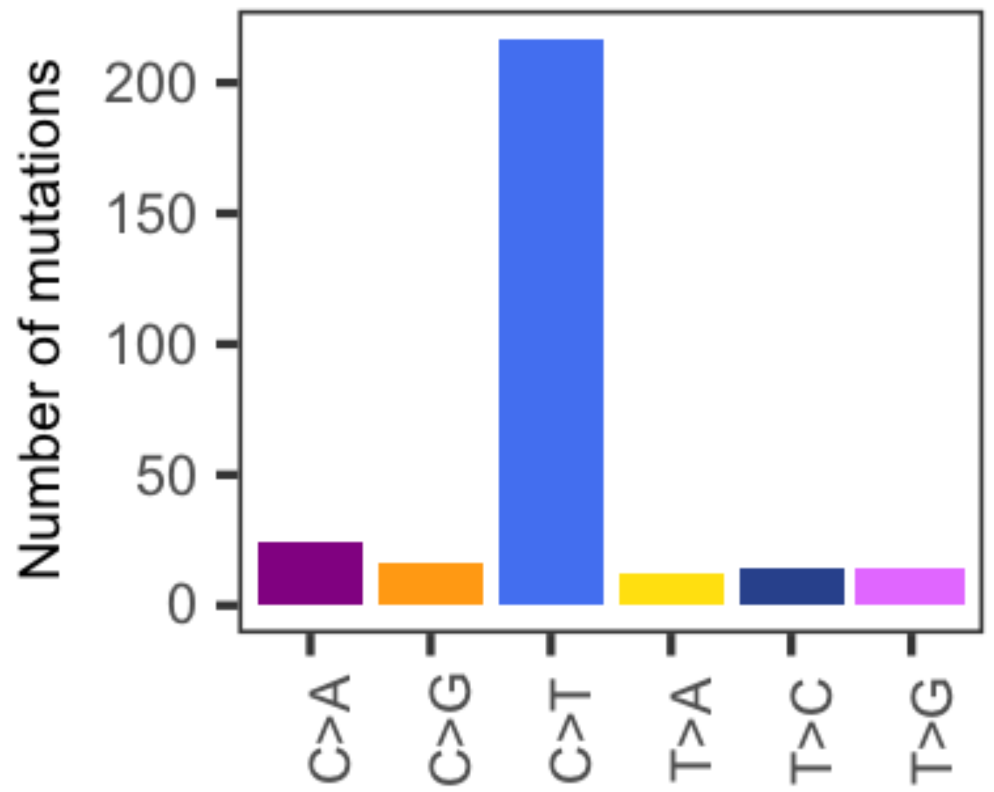
Different mutagenic mechanisms generate different errors on the DNA

smoking

Lung cancer

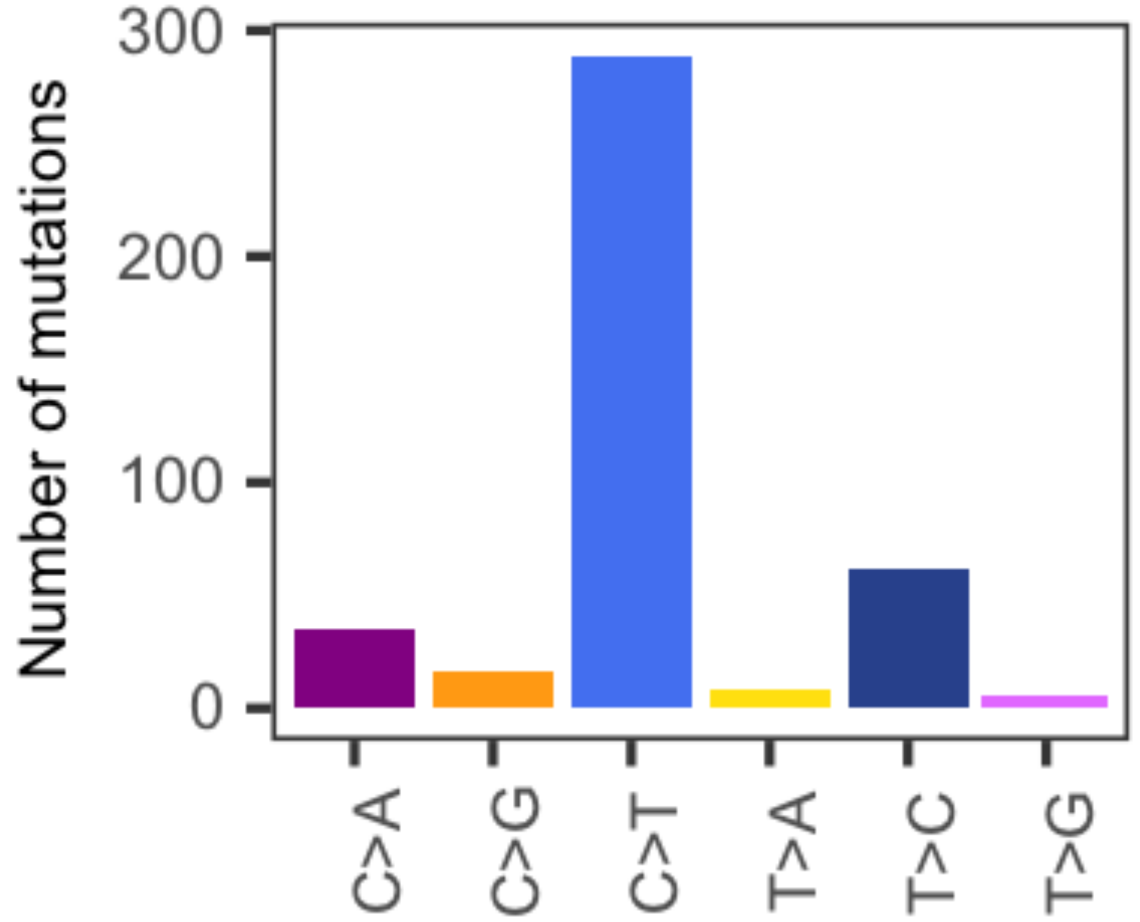


Melanoma, skin



sunlight

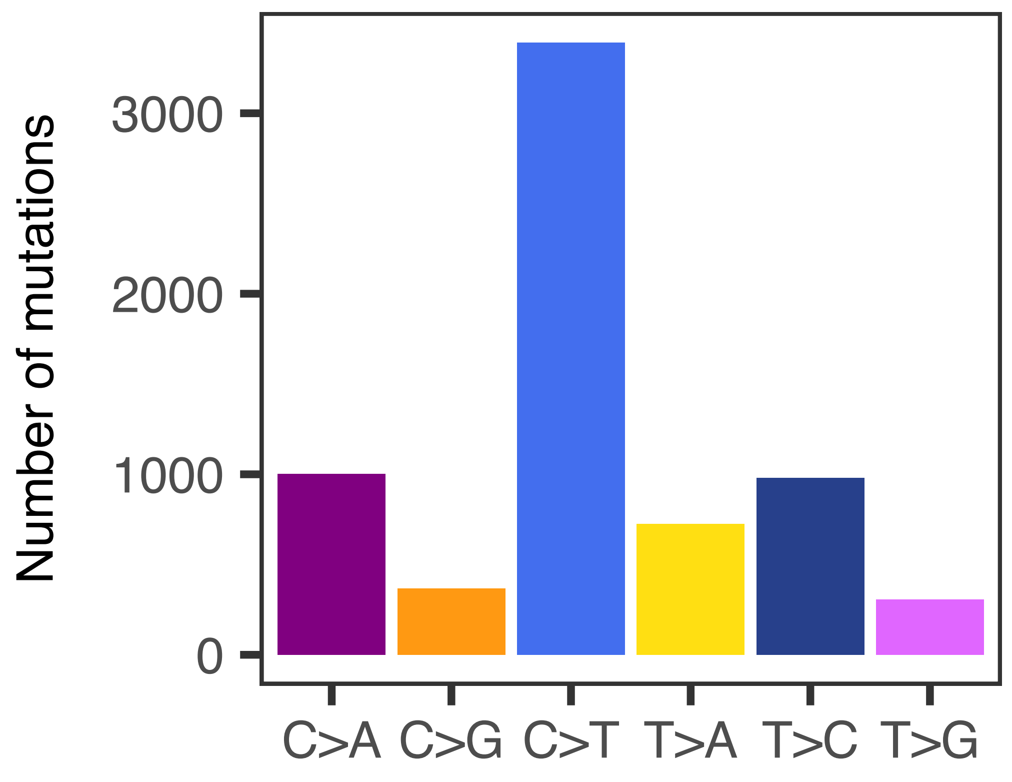
Glioma, brain

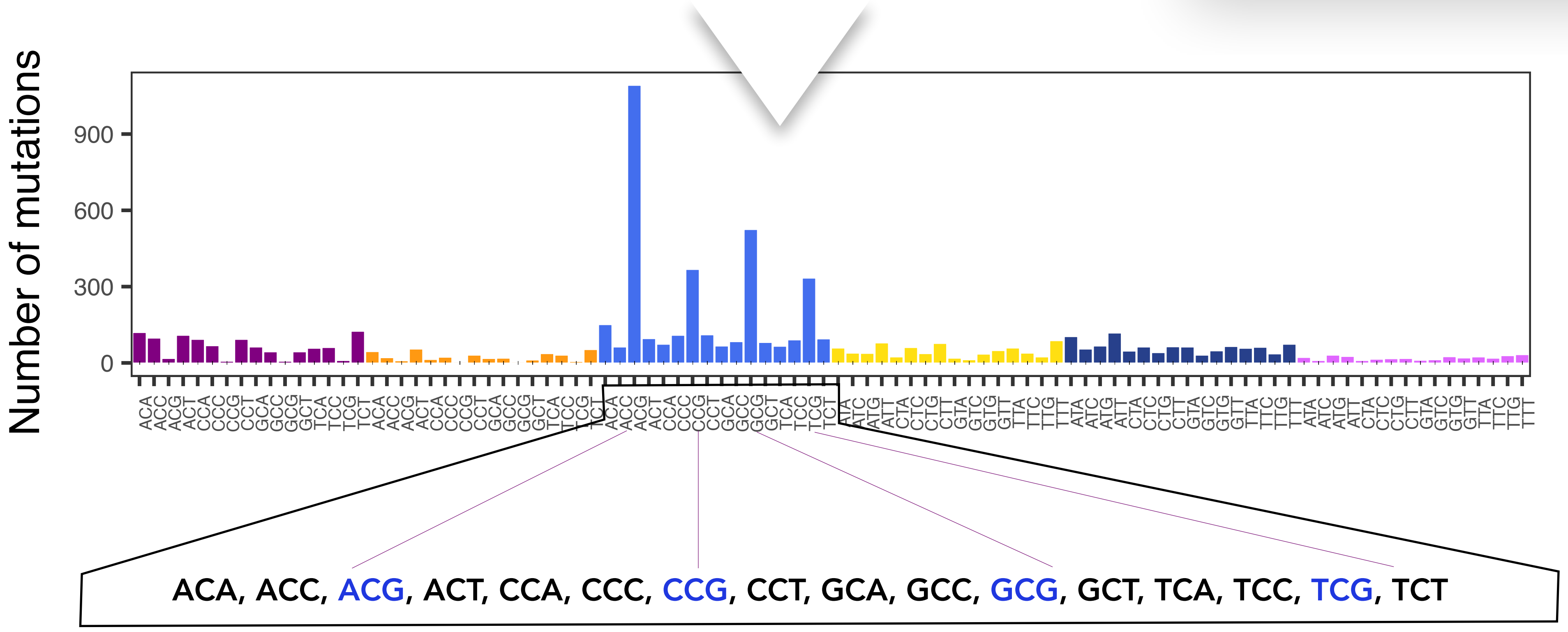


Very similar in 6 dimensions although the source is definitely different

Neighboring nucleotides are informative

Brain cancer sample





ACA, ACC, ACG, ACT, CCA, CCC, CCG, CCT, GCA, GCC, GCG, GCT, TCA, TCC, TCG, TCT

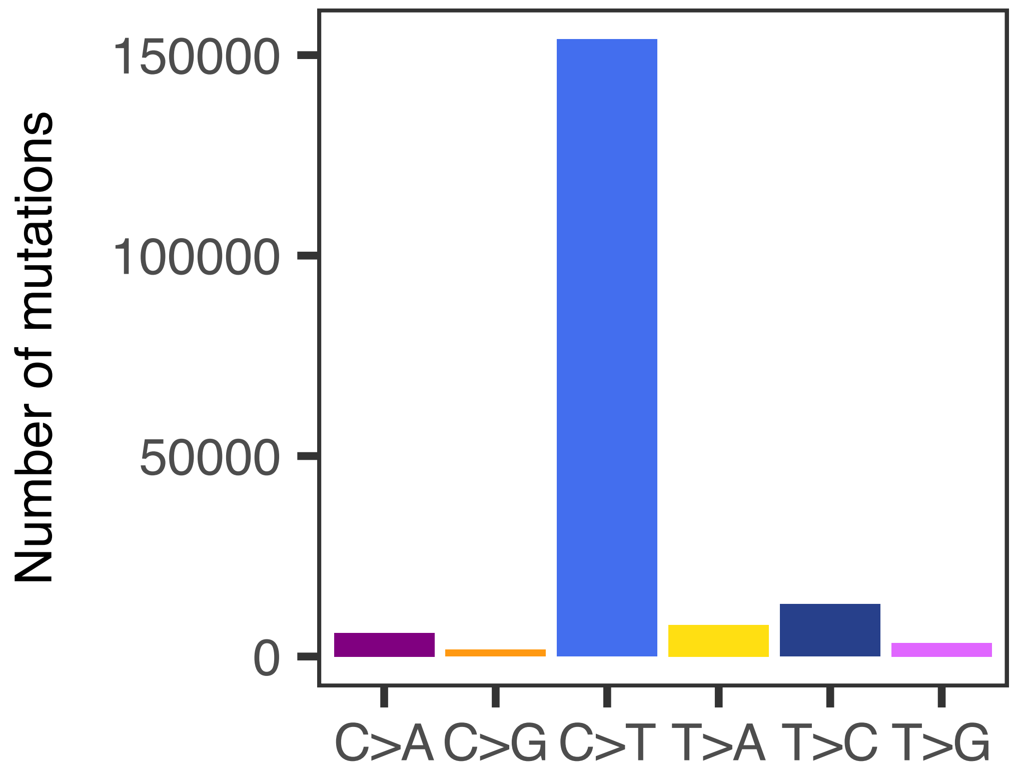
- Spontaneous deamination of 5-methylcytosine

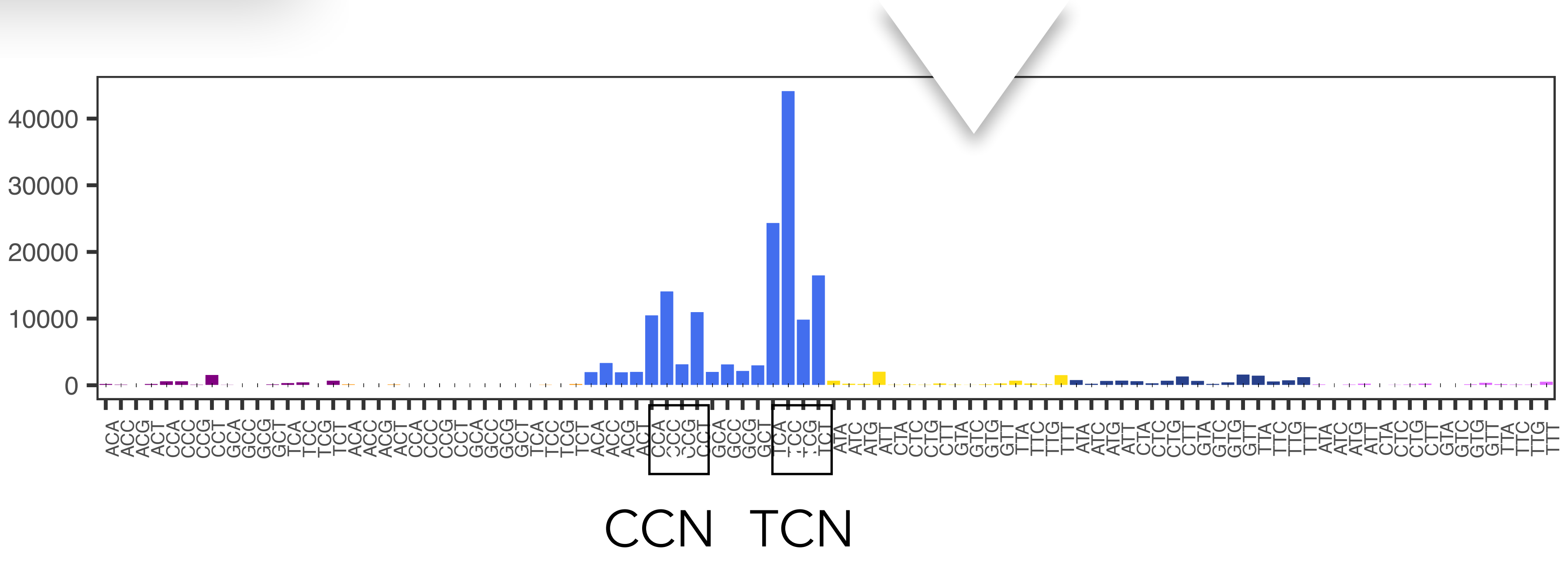
Trinucleotide context

N[N>X]N

4 x 6 x 4 = 96

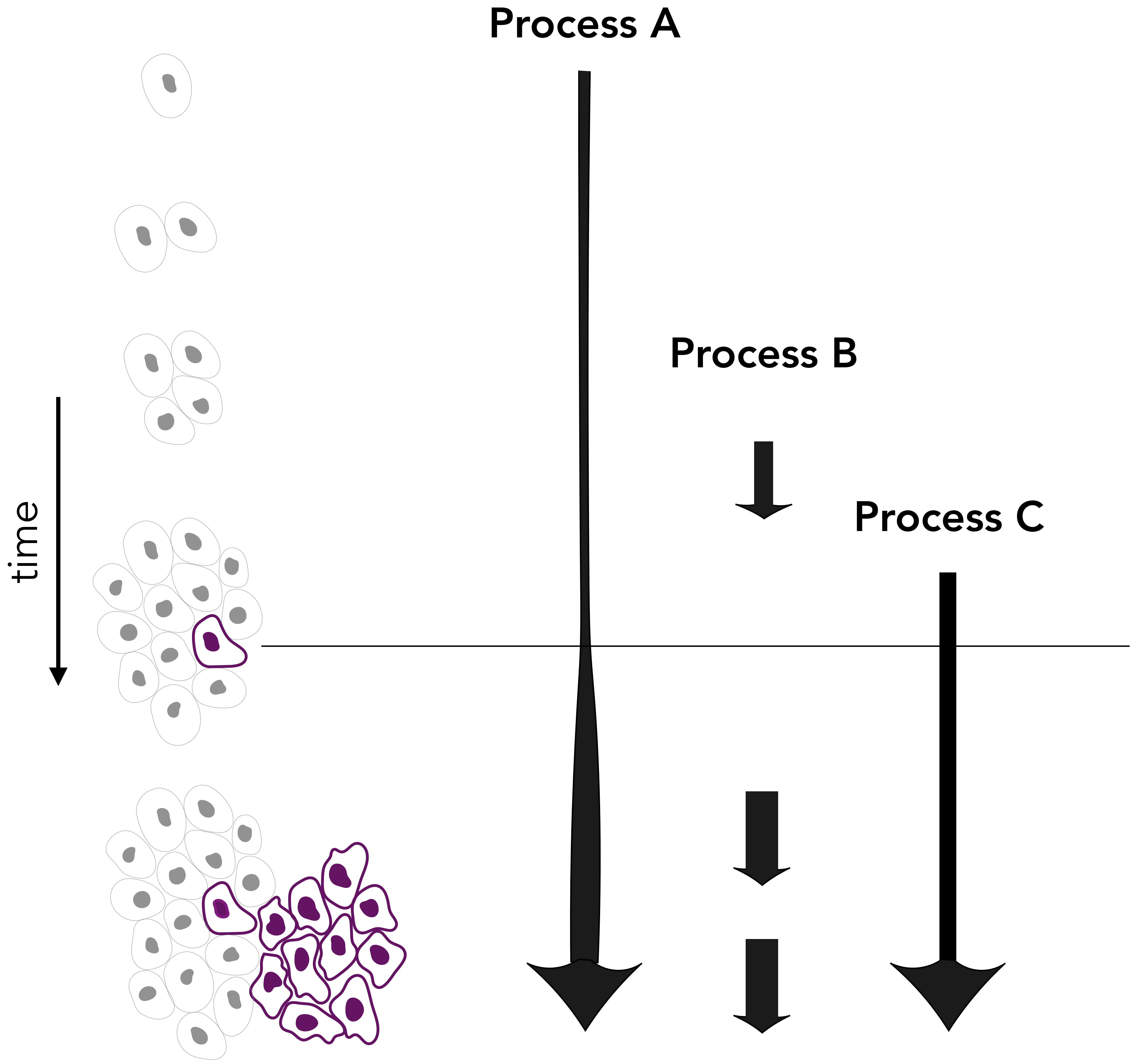
Melanoma sample



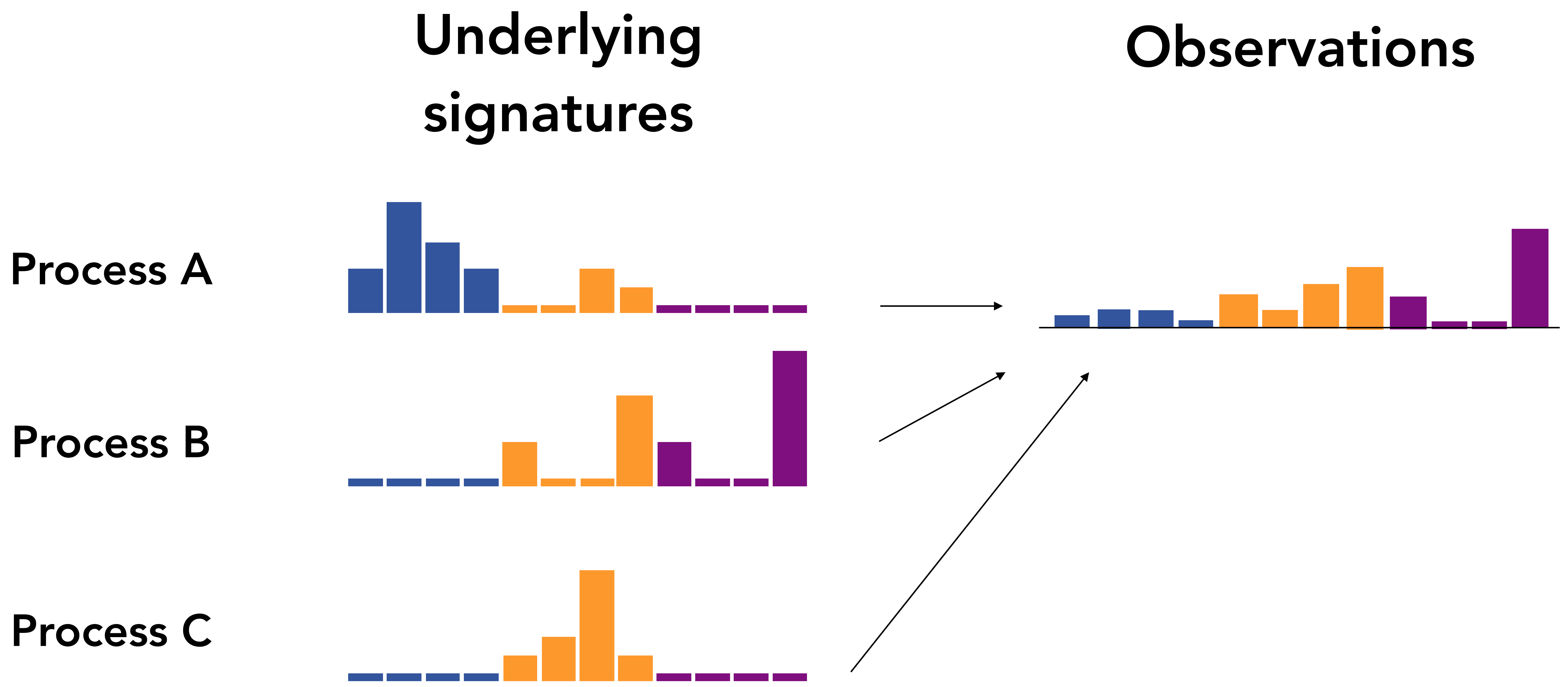


- C>T at CCN and TCN
- UV radiation is known to cause CC > TT

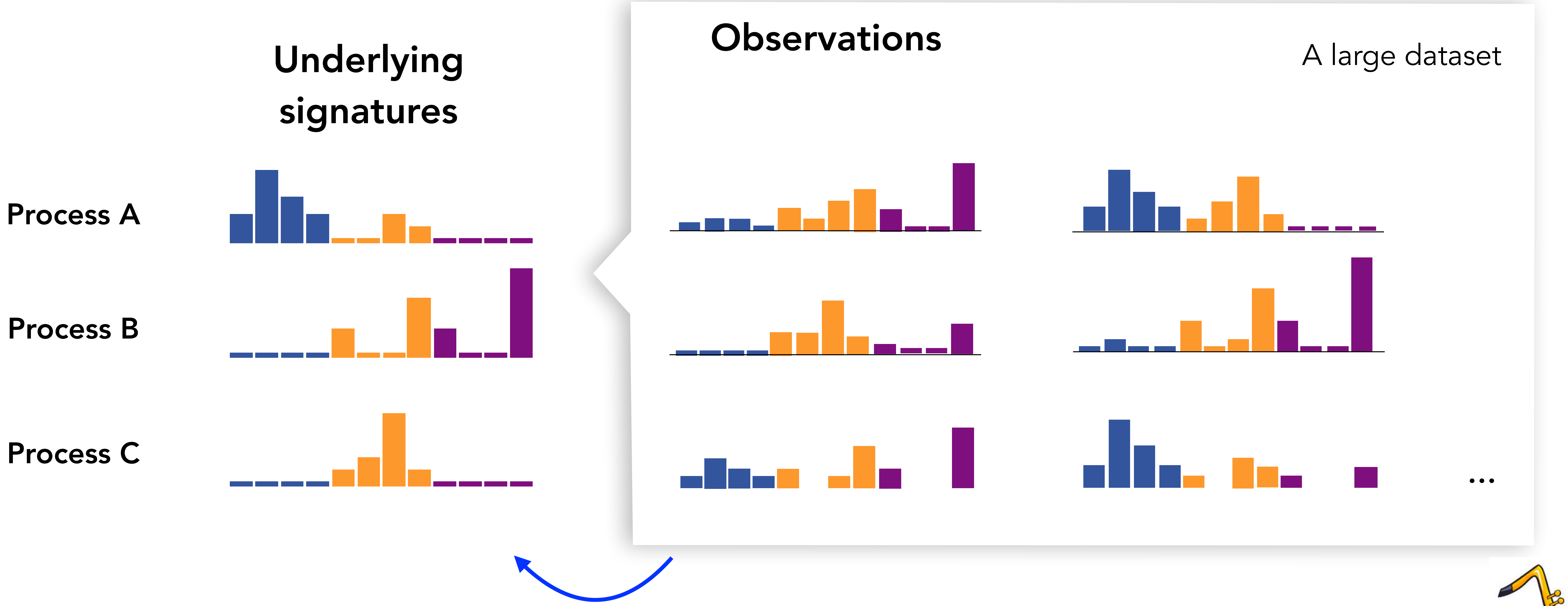
More than one mutational process in a single tumor



Signal decomposition into multiple processes



Signal decomposition into multiple processes



Underlying signatures

Process A

Process B

Process C

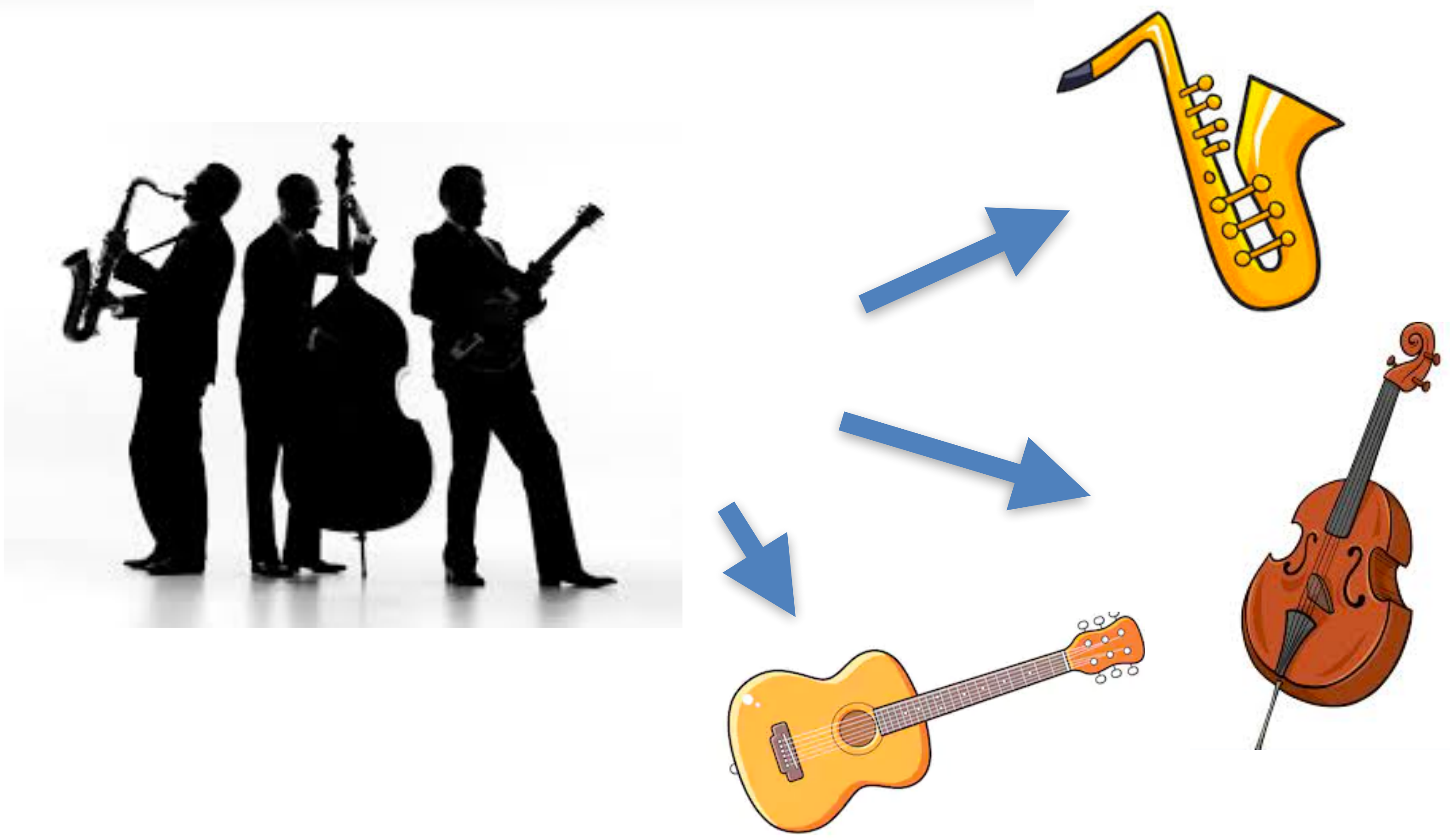
Observations

A large dataset

...

NMF

Non-negative matrix factorization



When enough observations are present pattern recognition algorithms can be used to discover the underlying signatures

Non-negative matrix factorization

Mutation count matrix

row: mutation type

column: sample

$$X \in \mathbb{R}^{V \times D}$$

Mutation matrix

Signature matrix

Exposure matrix

$$X \approx W H$$

$$W \in \mathbb{R}^{V \times K} \qquad H \in \mathbb{R}^{K \times D}$$

K mutational signatures $w_1, \cdots, w_K$

Non-negativity constraint (element-wise): $W, H \geq 0$

NP-hard, requires an iterative algorithm for finding local minima

A probabilistic interpretation -> maximum likelihood approach

Non-negative matrix factorization

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

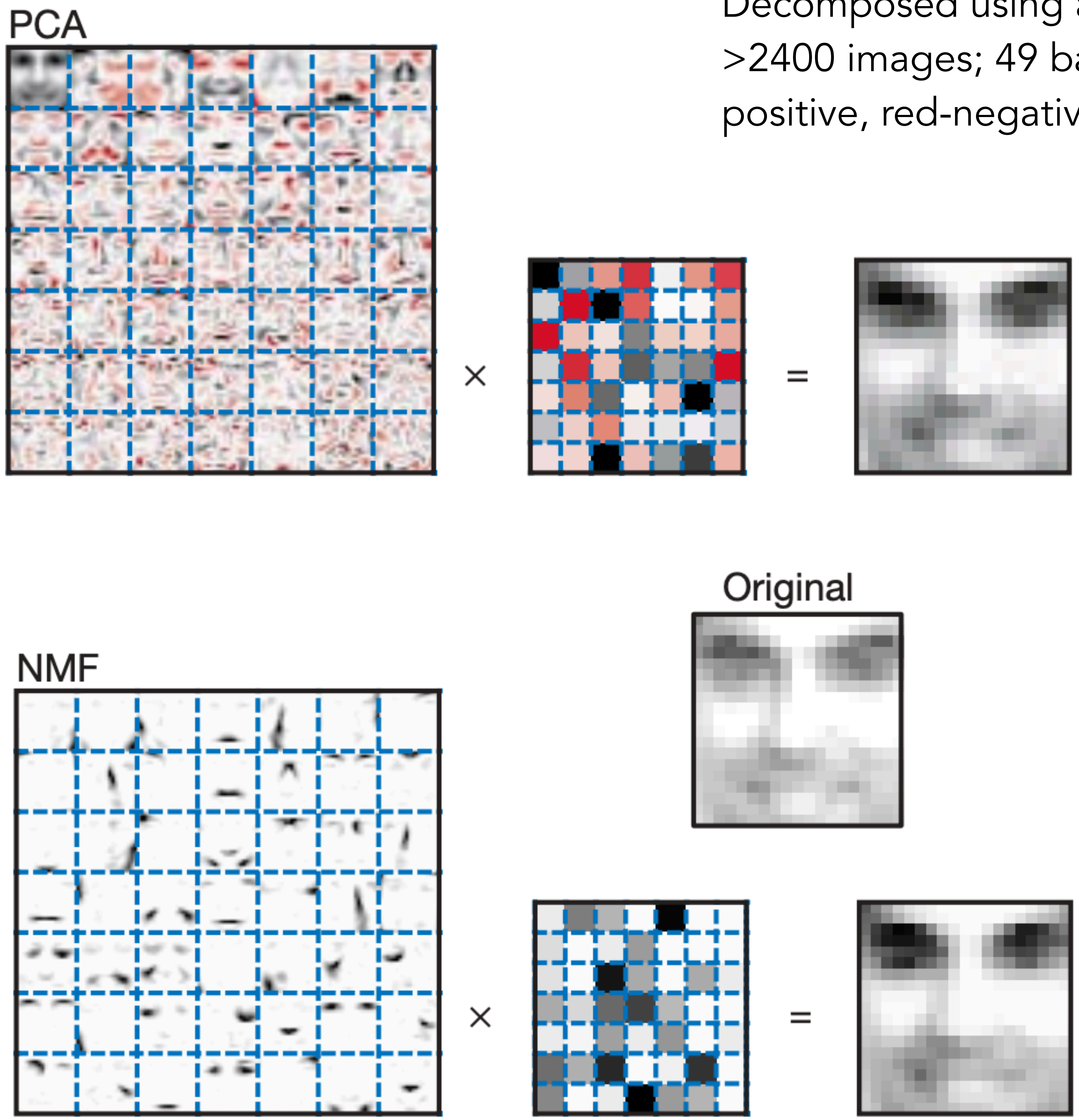
Nature, 1999; cited 14000 times

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia}H_{a\mu}$$

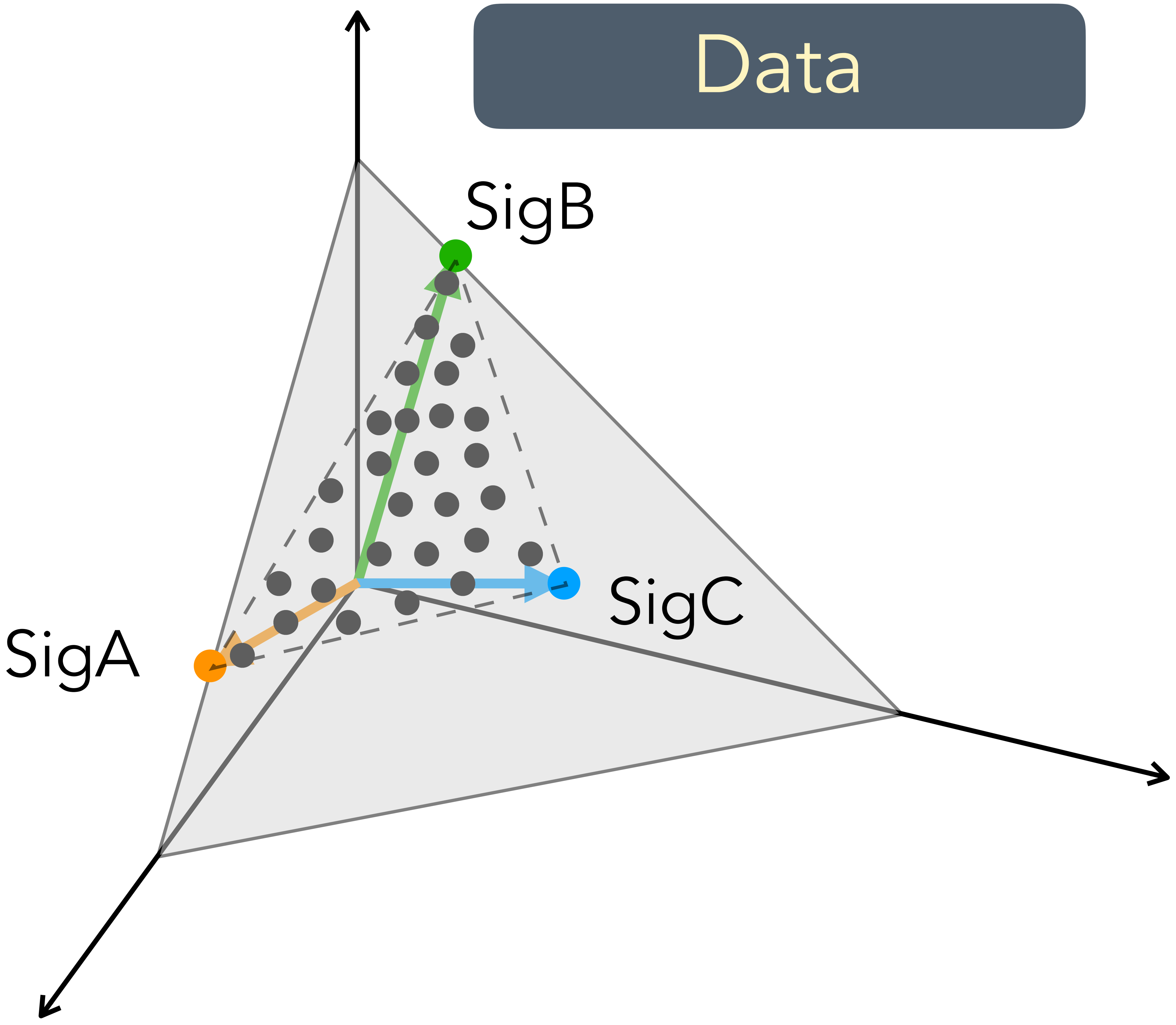
$$\boxed{\begin{aligned} W_{ia} &\leftarrow W_{ia}\sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}}H_{a\mu}\\ W_{ia} &\leftarrow \frac{W_{ia}}{\sum_{j} W_{ja}}\end{aligned}}$$

$$\boxed{H_{a\mu} \leftarrow H_{a\mu}\sum_{i} W_{ia}\frac{V_{i\mu}}{(WH)_{i\mu}}}$$

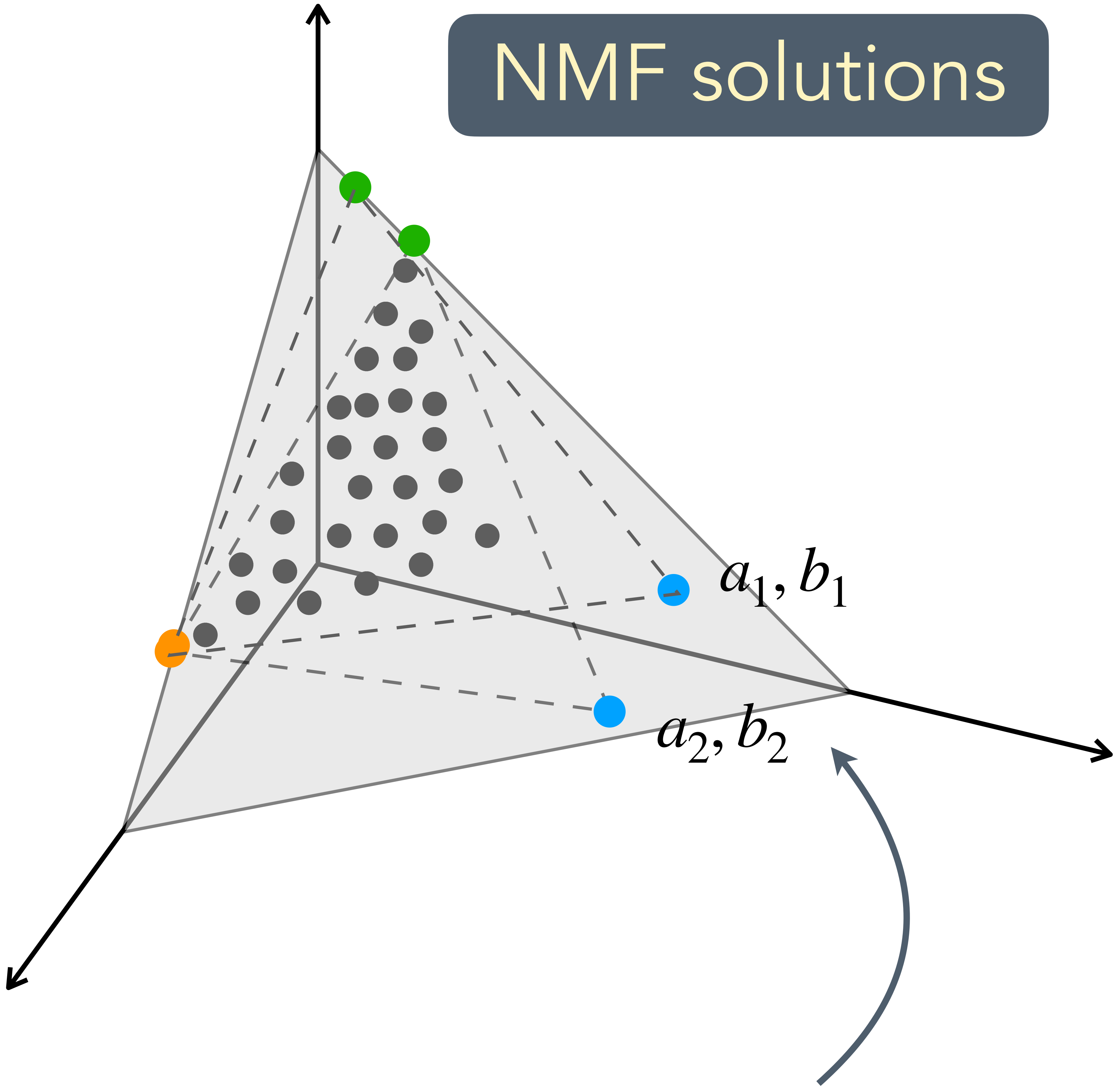
Decomposed using a dataset of >2400 images; 49 basis; black-positive, red-negative;



Non-uniqueness of NMF solutions



Data

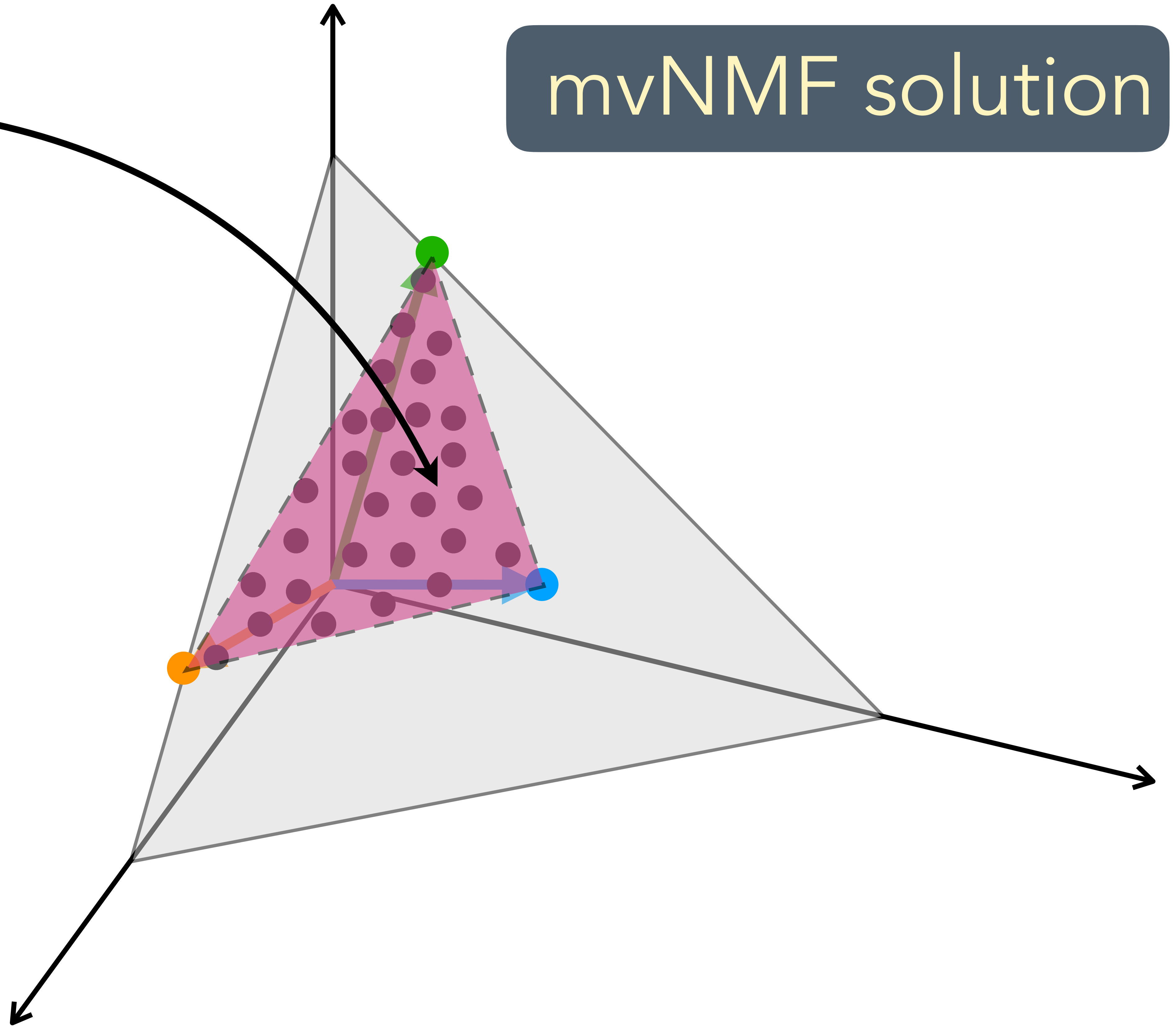


NMF solutions

Many possible solutions because of non-uniqueness

Minimum-volume NMF (mvNMF)

mvNMF penalizes the volume spanned by the signatures and induces a unique solution.

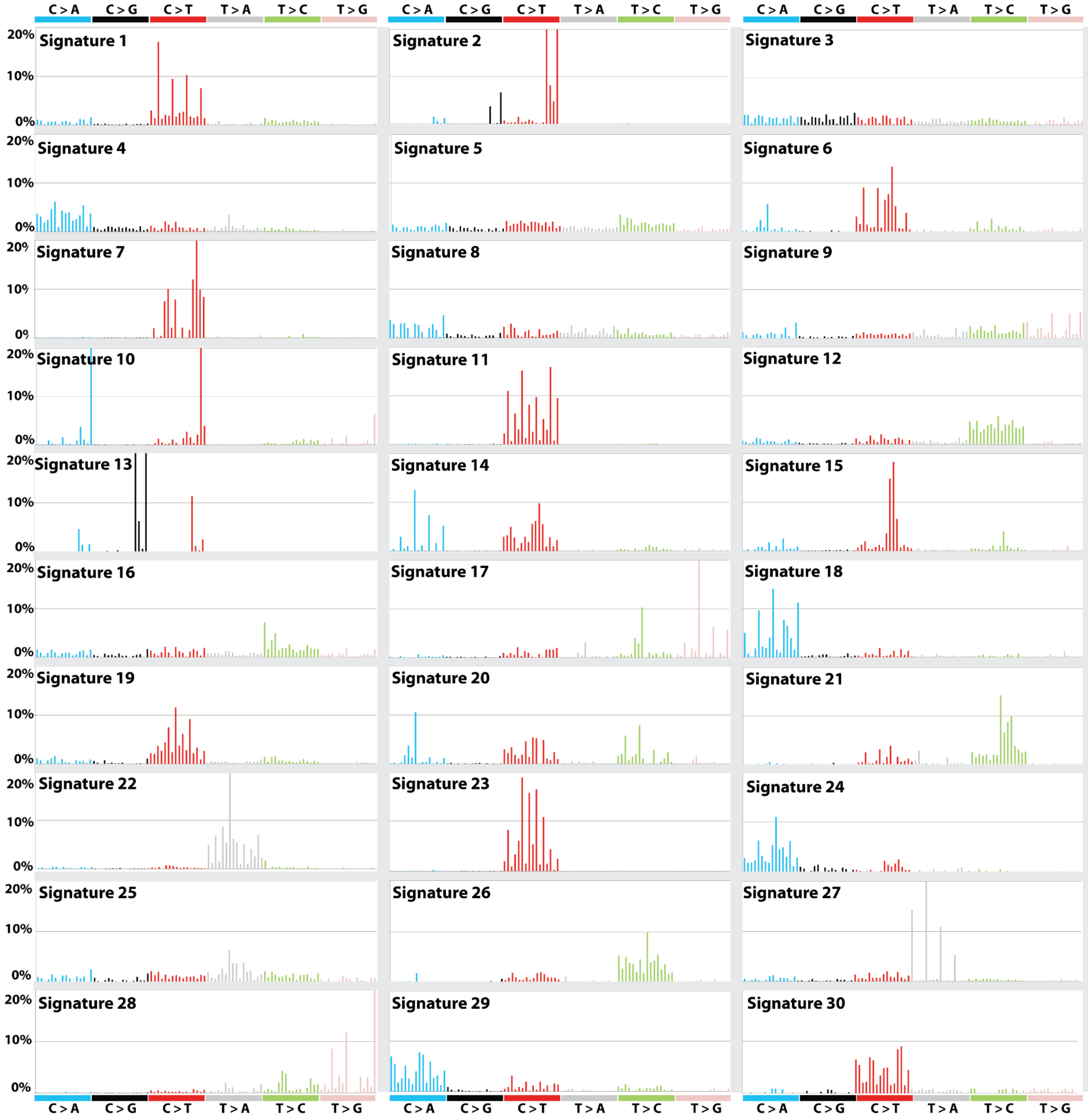

mvNMF solution

mvNMF:
Craig et al., *IEEE Transactions on Geoscience and Remote Sensing*, 1994
Miao et al., *IEEE Transactions on Geoscience and Remote Sensing*, 2007
Leplat et al., *IEEE Transactions on Signal Processing*, 2020

But the signatures are correlated…

$$\mathcal{L}(L, U, W, \alpha, \sigma^2) = -D_{\mathrm{KL}}(X||WH) - \frac{m}{2}(K + D)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_k \|\ell_k\|^2 + \sum_d \|u_d\|^2\right),$$

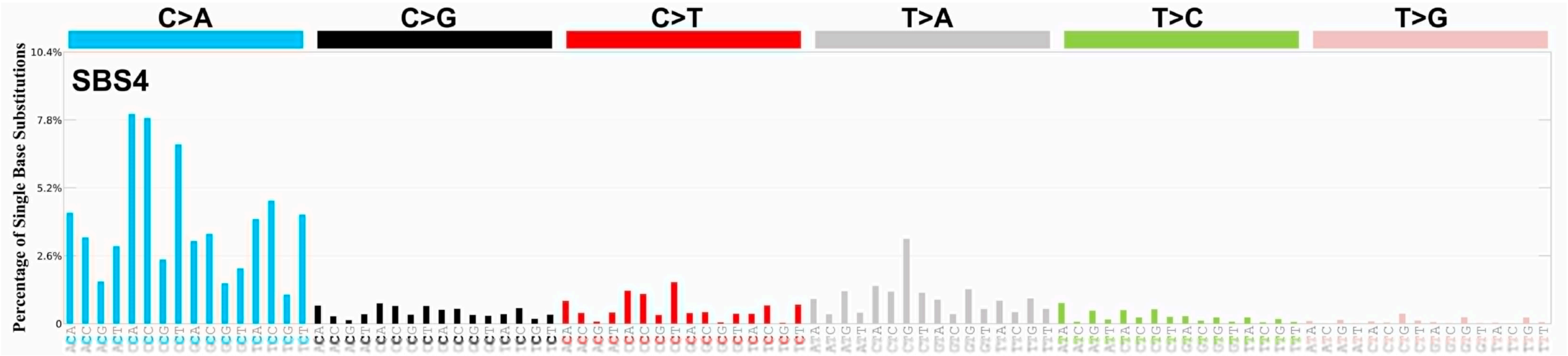
Catalog of mutational signatures



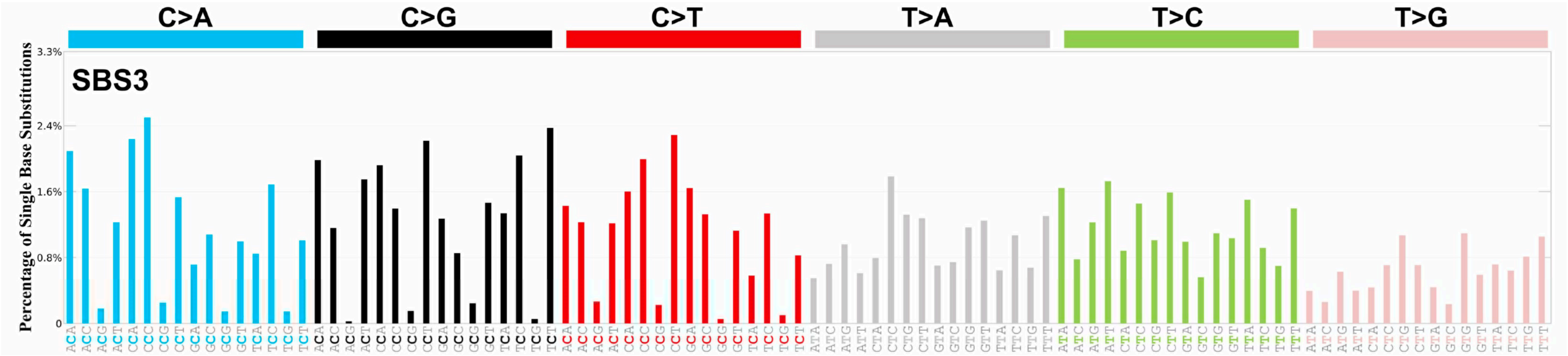
Alexandrov et al, *Nature* (2013)

Mutational signatures - examples

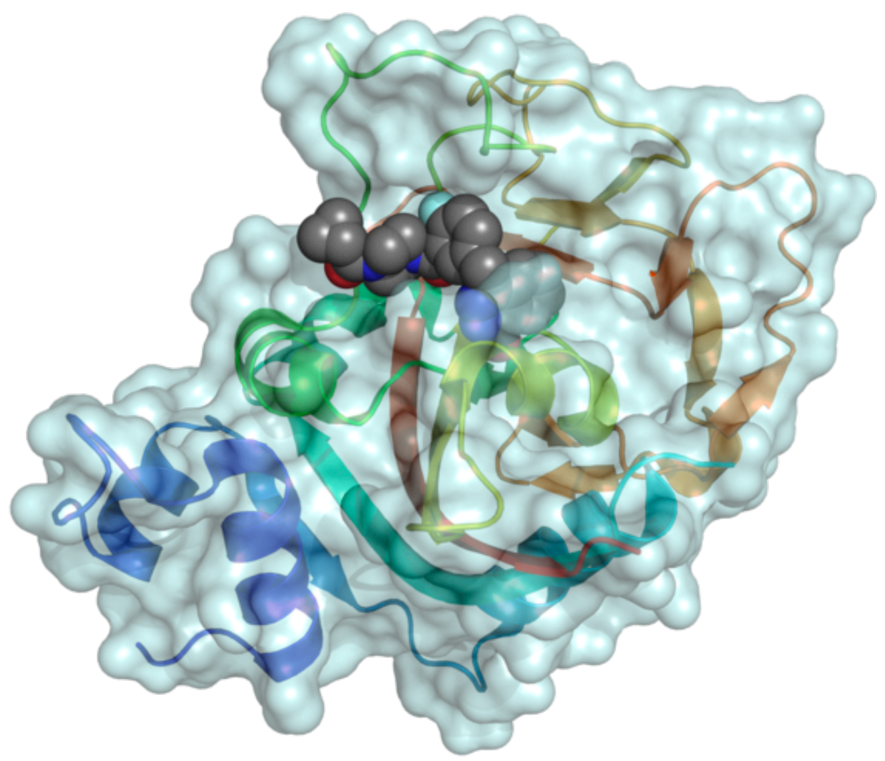
Smoking



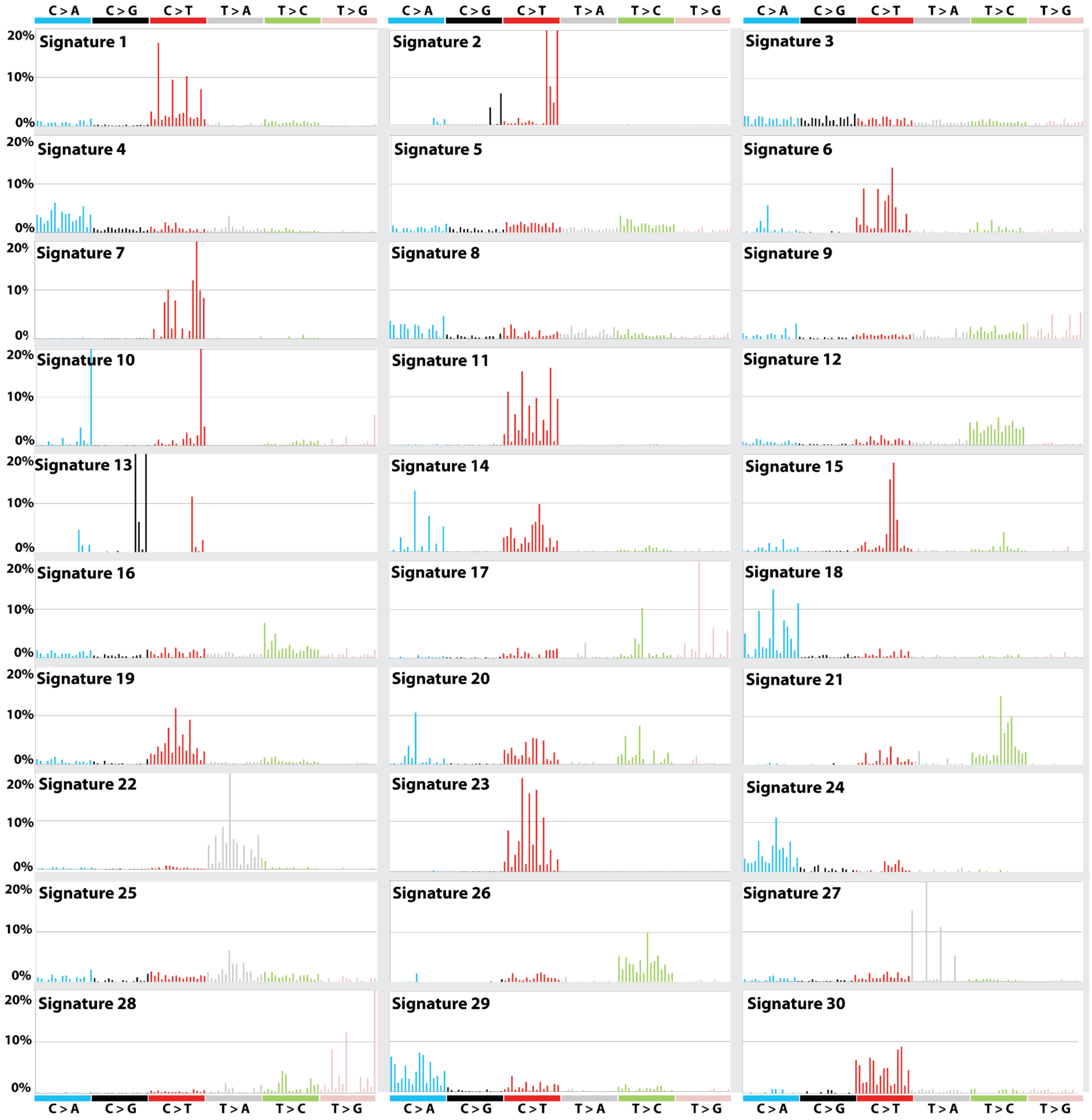
Homologous recombination deficiency



If your tumor genome shows SBS3, you should be considered for PARP inhibitor treatment.



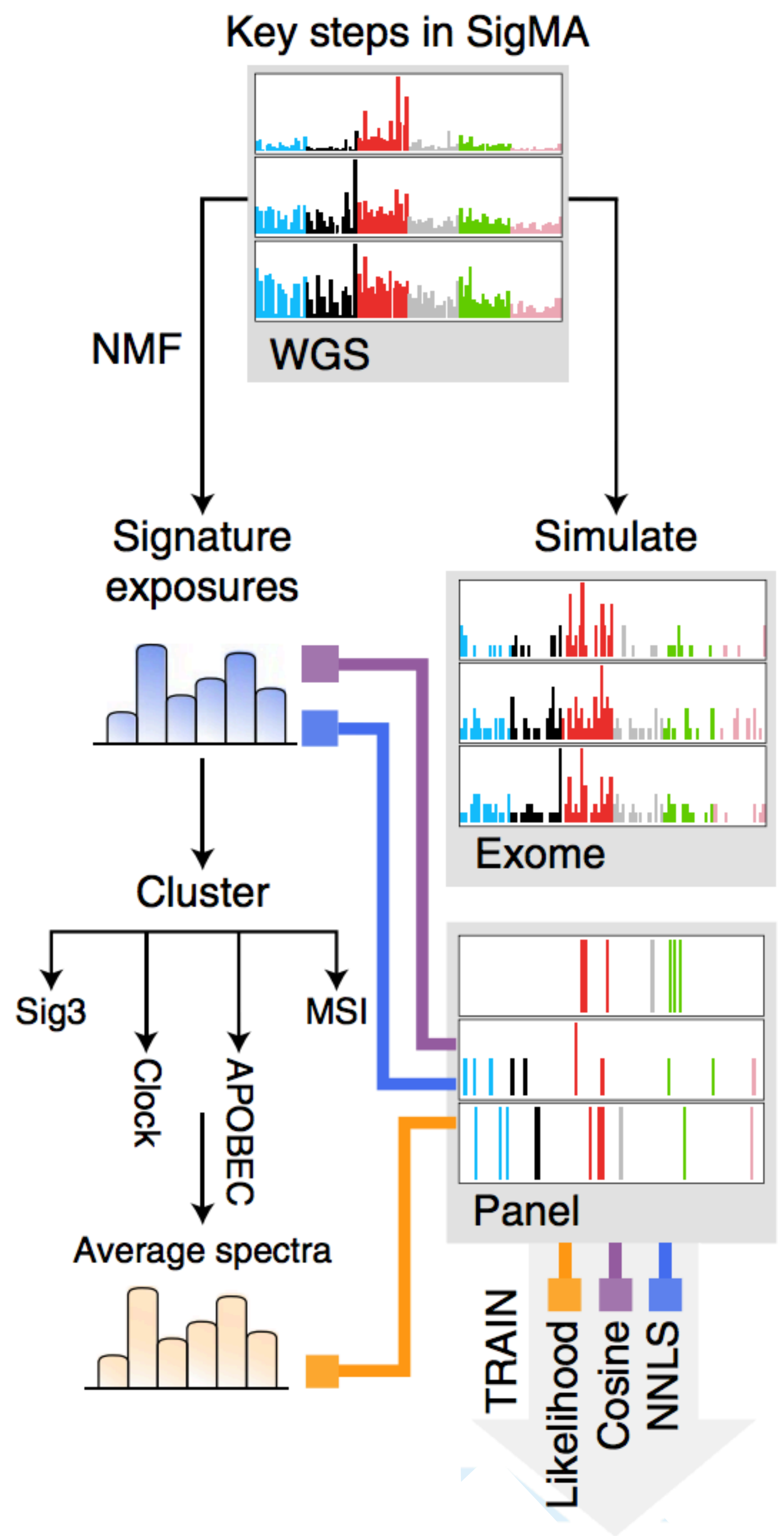
Catalog of mutational signatures

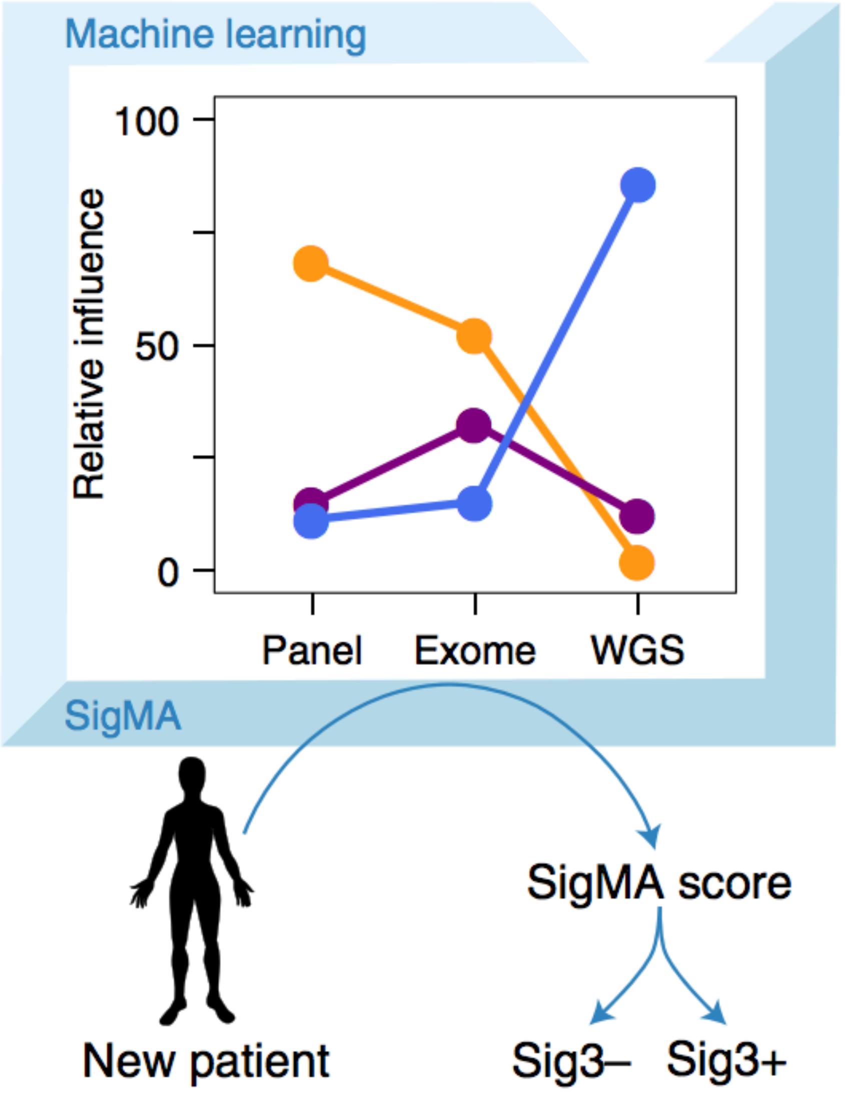


- How many signatures are there?
- More data -> more signatures?
- What is the mechanism behind each signature?
- What is the best way to determine whether a given patient has a specific signature?
- Are there signatures for other types of mutations?
- Can we identify signatures from blood DNA?

Alexandrov et al, *Nature* (2013)

Mutational signature analysis methods





Can we find patients who should receive PARP inhibitor?

SigMA

Signature Multivariate Analysis

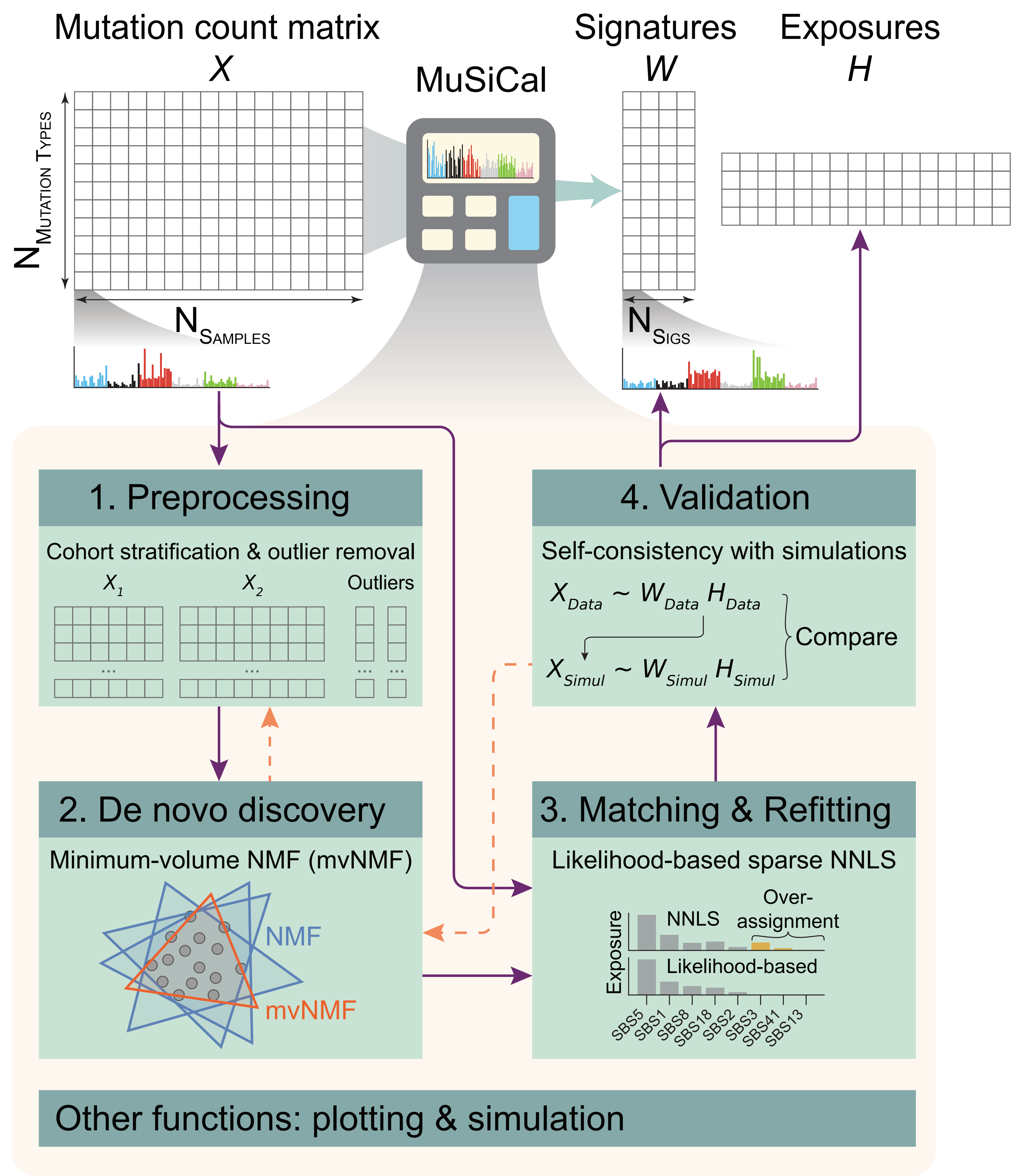
Gulhan et al, *Nature Genetics*, 2019

MuSiCal:

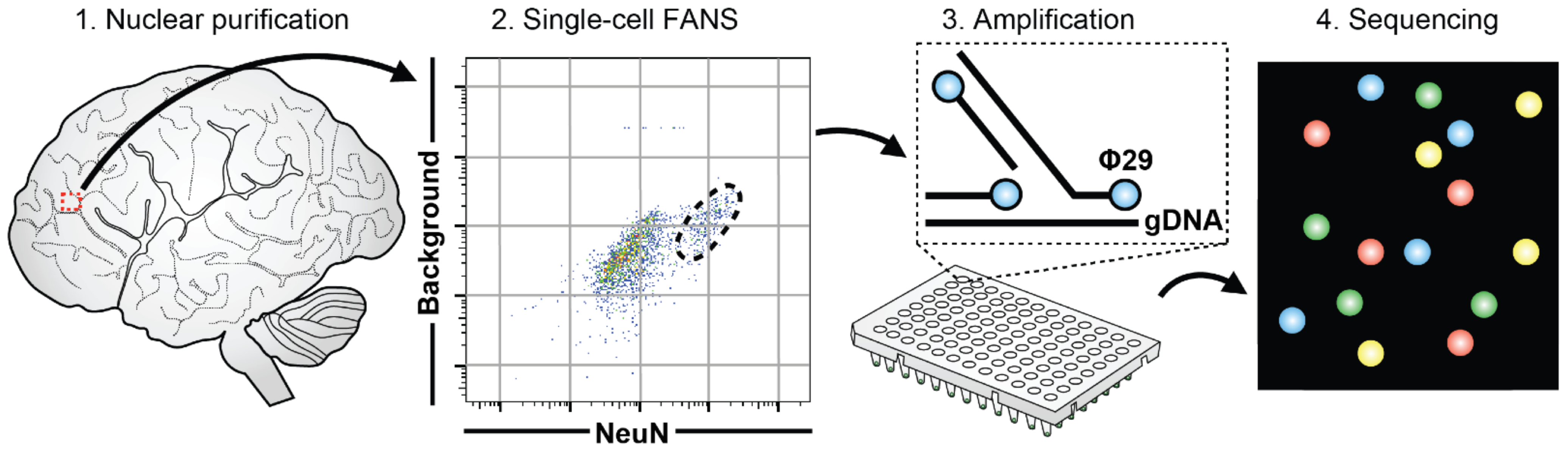
Mutational Signature Calculator

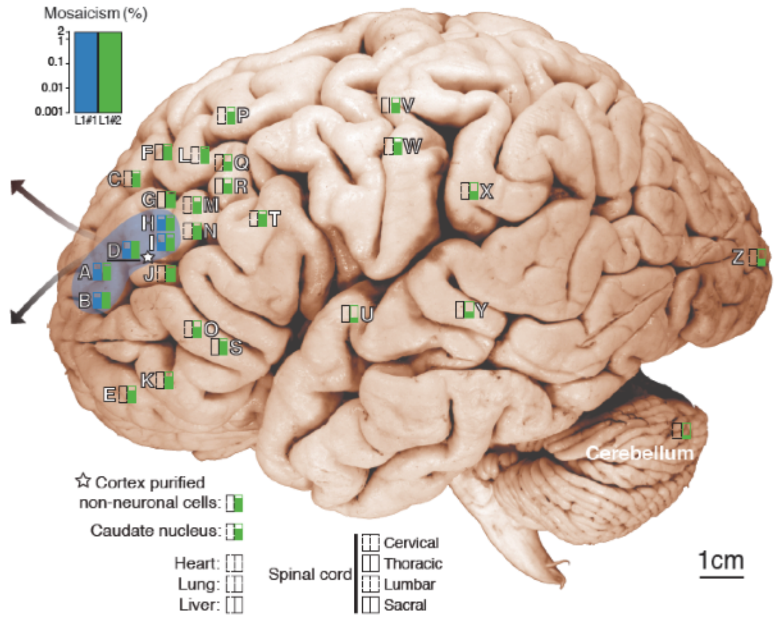
Hu et al, *Nature Genetics*, in press

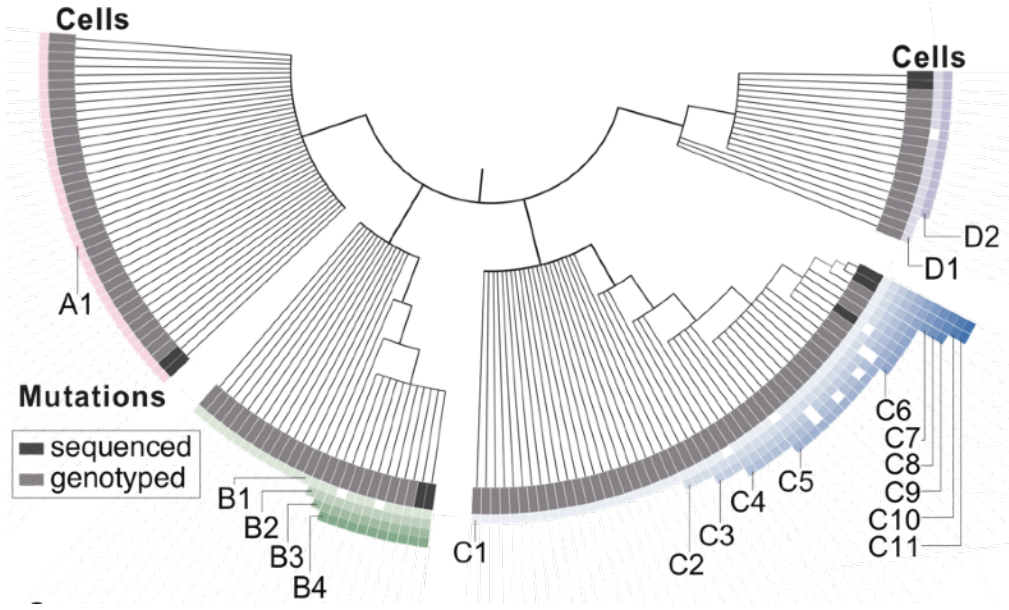
Can we find signatures more accurately?

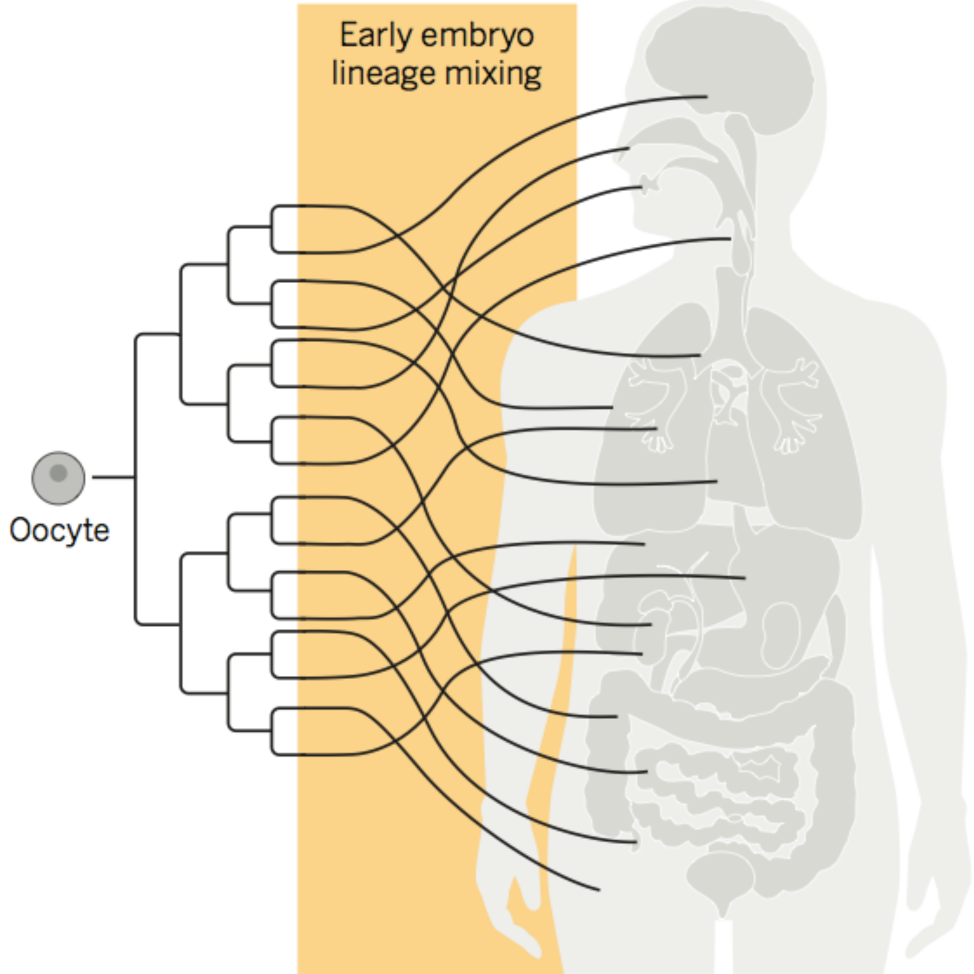


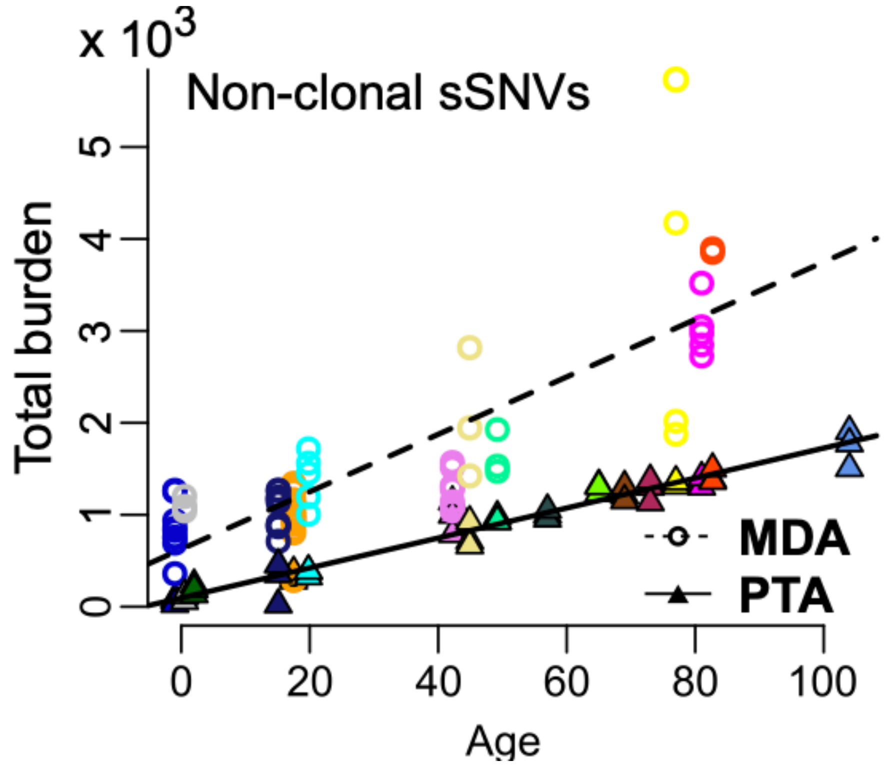
Studying mutations in single cells and in the brain



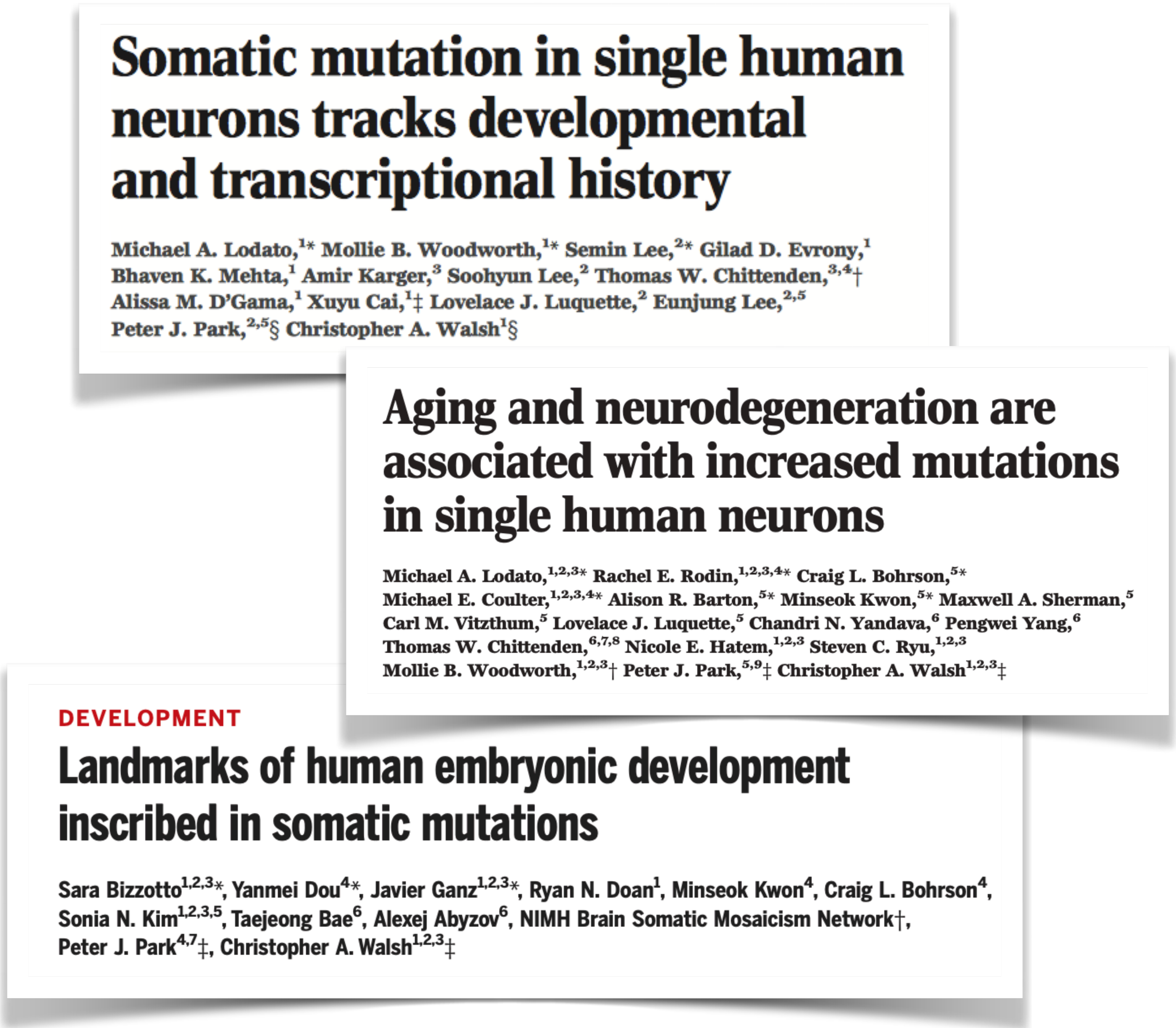








~17 somatic SNVs/year per neuron



Somatic mutation in single human neurons tracks developmental and transcriptional history

Michael A. Lodato,[1*] Mollie B. Woodworth,[1*] Semin Lee,[2*] Gilad D. Evrony,[1] Bhaven K. Mehta,[1] Amir Karger,[3] Soohyun Lee,[2] Thomas W. Chittenden,[3,4†] Alissa M. D'Gama,[1] Xuyu Cai,[1‡] Lovelace J. Luquette,[2] Eunjung Lee,[2,5] Peter J. Park,[2,5§] Christopher A. Walsh[1§]

Aging and neurodegeneration are associated with increased mutations in single human neurons

Michael A. Lodato,[1,2,3*] Rachel E. Rodin,[1,2,3,4*] Craig L. Bohrson,[5*] Michael E. Coulter,[1,2,3,4*] Alison R. Barton,[5*] Minseok Kwon,[5*] Maxwell A. Sherman,[5] Carl M. Vitzthum,[5] Lovelace J. Luquette,[5] Chandri N. Yandava,[6] Pengwei Yang,[6] Thomas W. Chittenden,[6,7,8] Nicole E. Hatem,[1,2,3] Steven C. Ryu,[1,2,3] Mollie B. Woodworth,[1,2,3†] Peter J. Park,[5,9‡] Christopher A. Walsh[1,2,3‡]

DEVELOPMENT

Landmarks of human embryonic development inscribed in somatic mutations

Sara Bizzotto[1,2,3*], Yanmei Dou[4*], Javier Ganz[1,2,3*], Ryan N. Doan[1], Minseok Kwon[4], Craig L. Bohrson[4], Sonia N. Kim[1,2,3,5], Taejeong Bae[6], Alexej Abyzov[6], NIMH Brain Somatic Mosaicism Network†, Peter J. Park[4,7‡], Christopher A. Walsh[1,2,3‡]

Brain projects come together pp. 11 & 42
The hidden force behind this year's El Niño p. 22
Reducing fuel fire hazards pp. 32 & 72
Science
$10 2 OCTOBER 2015 sciencemag.org
AAAS
Brain neurons diverge
Mutations reveal developmental pathways pp. 37 & 94

Evrony et al, *Neuron*, 2015
Lodato et al, *Science*, 2015
Lodato et al, *Science*, 2018
Bohrson et al, *Nature Genetics*, 2019
Dou et al, *Nature Biotechnology*, 2020
Rodin et al, *Nature Neuroscience*, 2021
Bizzotto et al, *Science*, 2021
Sherman et al, *Nature Neuroscience*, 2021
Luquette et al, *Nature Genetics*, 2022

Acknowledgement

NIH
National Human Genome Research Institute

LUDWIG CENTER
HARVARD MEDICAL SCHOOL

CANCER RESEARCH UK GRAND CHALLENGE

The Mark Foundation® for Cancer Research

SIMONS FOUNDATION

Park Lab

- Clara Bakker
- Michele Berselli
- Joseph Brew
- Craig Bohrson
- Greg Brunette
- Elizabeth Chun
- Shannon Ehmsen
- Niklas Engel
- William Feng
- Cesar Ferreyra-Mansilla
- Beverly Fu
- Teng Gao
- Benedikt Geiger
- Dominik Glodzik
- Doga Gulhan
- Hu Jin
- Clara Kim
- Yoo-Na Kim
- Joe Luquette
- Julia Markowski
- Dominika Maziec
- Bianca Morris
- Rahi Navelkar
- Cassidy Perry
- Kent Pitman
- Doug Rioux
- Will Ronchetti
- Andy Schroeder
- Alex Veit
- Vinay Viswanadham
- Yifan Zhao
- Yuwei Zhang

Recent alumni:

- Katerina Chatzipli
- Simon Chu
- Josh Cook
- Andrea Cosolo
- Jake Lee
- Viktor Ljungstrom
- Catherine Song
- Antuan Tran
- Dana Vuzman

Collaborators

- Christopher Walsh (BCH)
- Steve Elledge (BWH)
- David Ting (MGH)
- Kevin Haigis (DFCI)
- Charles Roberts (St Jude)
- Mark Johnson (UMass)
- Peter Kharchenko (BCH)
- Mitzi Kuroda (BWH)
- Fred Winston (HMS)

- Colleagues from TCGA, ENCODE, ICGC, 4D Nucleome, Brain Somatic Mosaicism Network

I enjoy learning new things. When you start in a new field you have to ask dumb questions. I often say I'm paid for my ability to tolerate feeling stupid.

- Persi Diaconis