Discovering Cancer Signatures via Non-Negative Matrix Factorization

Nima Hejazi, Amanda Mok, Courtney Schiffman

2018-10-03

Introduction (Nima)

Overview and Motivations

- ...
- ...
- ...

Overview of Matrix Factorization

- Matrix factorization as unsupervised learning
- What can we learn about objects by matrix factorization?
- A general formulation of matrix factorization
- Various forms of matrix factorization: NMF, PCA, VQ
- Applications of matrix factorization: images, text
- Biological applications of matrix factorization

Non-Negative Matrix Factorization (Nima)

What is Matrix Factorization?

- Suppose we have a *data matrix* V of dimension $n \times m$, each column of which is an n-vector of observations of a given variable.
- A factorization of V produces two matrices $\{W, H\}$ that approximately capture the information present in V.
- From linear algebra, we have $V_{ij} \approx (WH)_{ij} = \sum_{a=1}^{r} W_{ia} H_{aj}$.
- The dimensionality of the induced matrix factors is reduced wrt V – that is, let W be $n \times r$ and H be $r \times m$.
- This can be viewed as a form of data compression when the rank r is small in comparison to n and m.
- In particular, r is often chosen such that $(n + m)r \leq nm$.

What is Matrix Factorization?

- With the general factorization $V_{ij} \approx \sum_{a=1}^{r} W_{ia}H_{aj}$, W and H each pick up different important aspects of V.
- When V is a $n \times m$ matrix of images of faces, where each row corresponds to a pixel and each column an image:
- the r columns of W may be thought of as basis images,
- and each of the j columns of H is termed an encoding (coefficients to be applied to basis images).
- Various forms of matrix factorization place different types of constraints on the manner in which W and H are generated.

Vector Quantization (VQ)

- **Constraint:** each column of H has a single entry equal to unity, with all other entries being set to zero.
- Since this is a constraint on the *encoding* columns, this results in each column of W representing some distortion of the target image.
- Equivalently, each column of V is approximated by a single basis (column of W).
- In terms of image learning, this results in the VQ decomposition learning *prototypical* faces.

VQ: Prototypical Faces

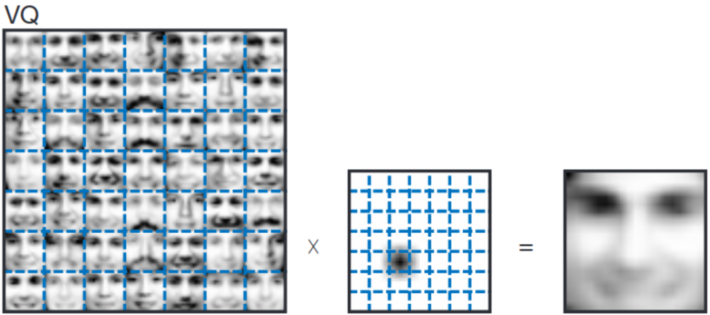


Figure 1:

Principal Components Analysis (PCA)

- **Constraint:** columns of W are set to be orthonormal; rows of H are set to be orthogonal to one another.
- Relaxation of the constraint of VQ in the sense that each face in our data set may be represented by a linear combination of the basis images in W.
- This results in a distributed encoding of each of the face images contained in V; basis images are referred to as *eigenfaces*.
- Statistical interpretation: each eigenface represents the direction of largest variance within the sample data.
- Intuitive interpretation: ??? (Complex cancellations make eigenfaces very difficult to interpret.)

PCA: *Eigenfaces*

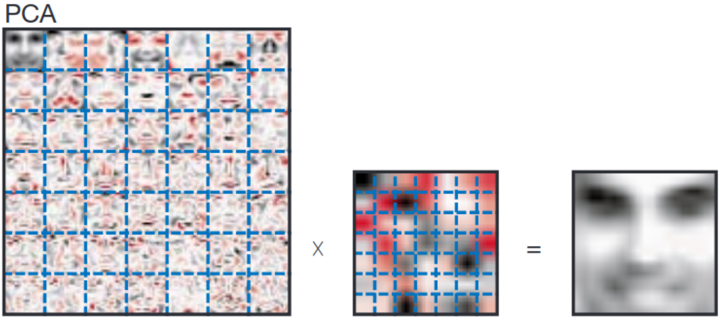


Figure 2:

PCA in Biology

- Obligatory example: John Novembre's European populations

What is NMF?

- **Constraint:** similar decomposition to PCA, but any nonzero entries in W and H must be *positive*.
- Multiple basis images may be used to reconstruct a face by linear combination; however, there are no possible cancellations (unlike in PCA).
- Since the basis images and encodings are all positive, each basis image may be intuitively thought of as picking up a *part of a face*.

What does non-negativity buy us?

- In practice, NMF produces sparse basis and encoding matrices.
- The basis images are *non-global* – that is, picking up variation in parts of a face.
- The encoding are also spare, resulting in ...
- ...

NMF: Parts of Faces

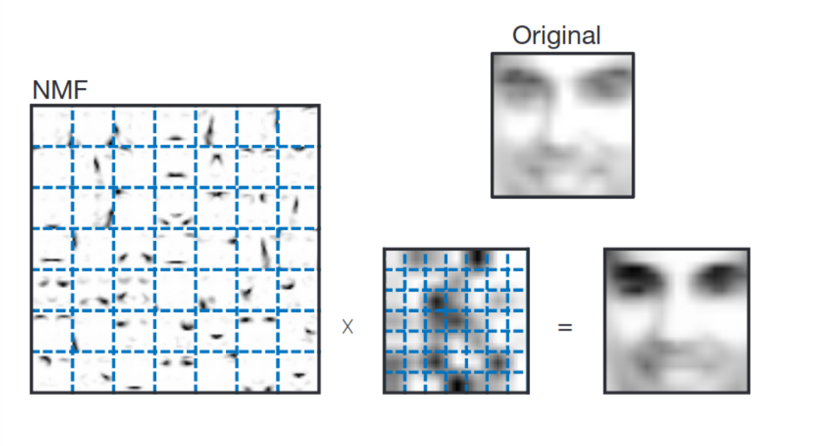


Figure 3:

Implementing NMF

...

- ...
- ...

Some fun with NMF

- ...
- ...
- ...

NMF in biology

- example from Bioconductor?
- pretty plot goes here

NMF in cancer biology

- So, we've now established that NMF finds *parts* of the input matrix through the non-negativity constraint it imposes on the matrix factors.
- This has important applications for exploring cancer biology; namely, applying NMF could help us detect *parts of tumors*.
- Interpretation is challenging: does this mean we're detecting subclonal populations?
- There's a whole lot more to come.

A bit of biology (Amanda)

What is cancer?

- Complex tissues with multiple cell types and interactions
- Characterized by unchecked somatic cell proliferation
- Normal cells acquire hallmark traits that enable them to become tumorigenic[1]

[1] Hanahan and Weinberg (2011)

Hallmarks of Cancer

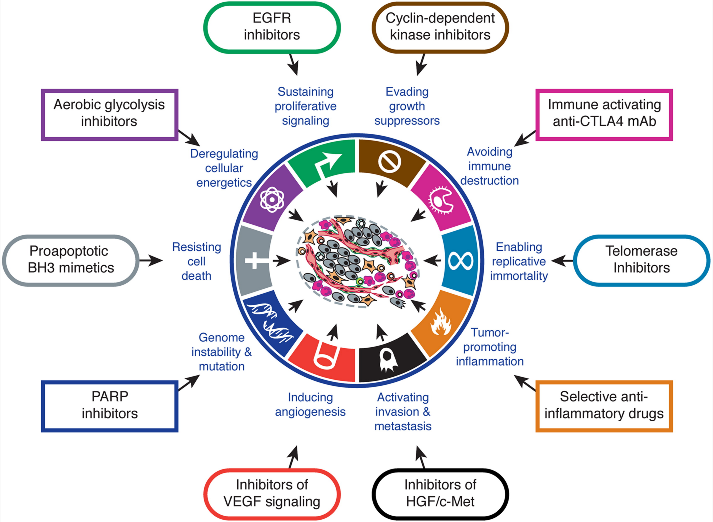


Figure 4: Hallmarks of Cancer

Cancer is a genetic disease

- Germline mutations: inherited from parents
  - Mutations in tumor suppressor genes or oncogenes can predispose someone to develop cancer
- Somatic mutations: acquired over time in somatic cells
  - Endogenous: DNA damage as a result of metabolic byproducts
  - Exogenous: DNA damage as a result of mutagenic exposure
- Epigenetic modifications: no change to DNA sequence
  - DNA methylation
  - Histone modification
  - MicroRNA gene silencing

Somatic mutations

- Rearrangements
- Copy number changes
- Indels
- Base substitutions
  - 6 types of substitutions (C>G, C>T, C>A, G>T, G>A, T>A)
  - 4 types of 5' base nucleotide
  - 4 types of 3' base nucleotide
  - Transcriptional strand

Clonal evolution in cancer

Applying NMF to mutational processes

Alexandrov et al. (2013) characterize mutational processess as a blind source separation problem

Mutational catalogs "are the cumulative result of all the somatic mutational mechanisms …that have been operative during the cellular lineage starting from the fertilized egg…to the cancer cell."

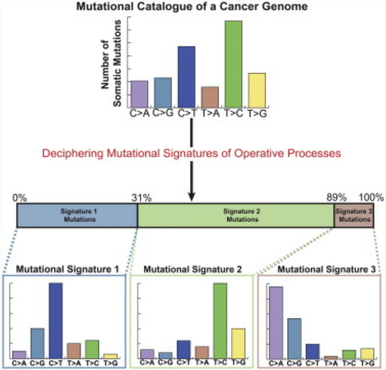


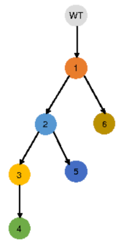
Figure 5:

How is the work of Alexandrov et al. (2013) related to inferring clonal evolution of tumors?

Goal: learn the "evolutionary history and population frequency of the subclonal lineages of tumor cells."

- From SNV frequency measurements, try to infer the phylogeny and genotype of the major subclonal lineages.

A



B

Cellularity matrix

Biopsy	Subclones 1	2	3	4	5	6
1	0.1	0.2	0.1	0.6		
2	0.1				0.7	0.2
3			0.05	0.55	0.2	0.2
4		0.1			0.8	0.1
5	0.1	0.1		0.5	0.3	
6				0.4	0.4	0.2

C

Frequency matrix

Biopsy	Subclones 1	2	3	4	5	6
1	1	0.9	0.7	0.6	0	0
2	1	0.7	0	0	0.7	0.2
3	1	0.8	0.6	0.55	0.2	0.2
4	1	0.9	0	0	0.8	0.1
5	1	0.9	0.5	0.5	0.3	0
6	1	0.8	0.4	0.4	0.4	0.2

How is the work of Alexandrov et al. (2013) related to inferring clonal evolution of tumors?

Different clonal mutations will have different signatures.

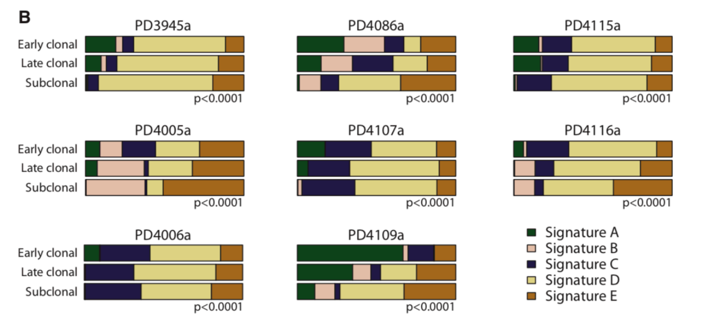


Figure 7:

Both works want to uncover driver mutations

Inferring clonal evolution of tumors

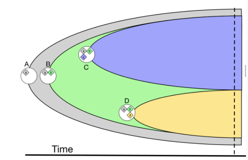
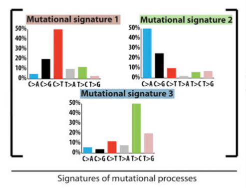


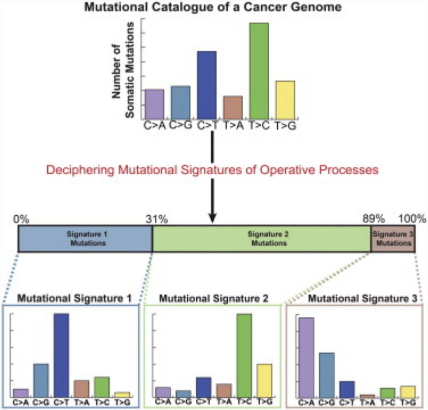
Figure 8:

Deciphering Signatures of mutational processes



Alexandrov et al. (2013) focus more on uncovering the cumulative mutational processes that make up a cancer genome, rather than the evolution of the tumor.

Goal: unscramble the latent signals from a mixture of a set of these signals.



NMF is a natural method for handling the BSS problem.

- Non-negative matrix entries.
- Want to learn the parts (mutational signatures of mutational processes) that add to the whole (mutational catalog).

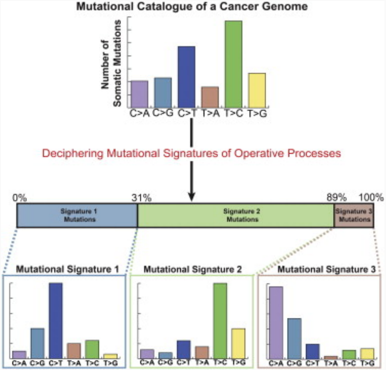


Figure 11:

What are the basis vectors and encodings in the context of mutational processes?

- A signature of a mutational process is defined as a probability mass function with a domain of preselected mutation types.
- The exposure of a mutational process is the mutation intensity

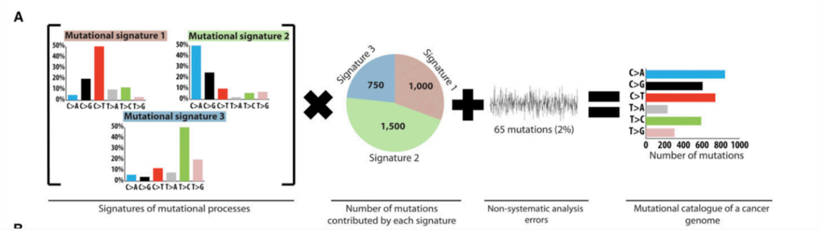
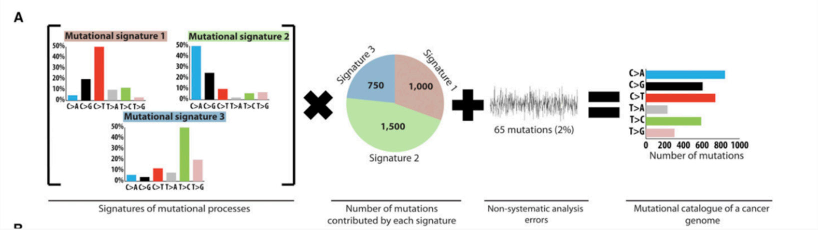


Figure 12:

What are the basis vectors and encodings in the context of mutational processes?



$$P \times E \approx M$$

M: K mutation types by G genomes

P: K mutation types by N mutation signatures

E: N mutation signatures by G genomes

What are the basis vectors and encodings in the context of mutational processes?

$$\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix} \approx \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$$

$$\times \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix} \qquad m_g^i \approx \sum_{n=1}^{N} p_n^i e_g^n.$$

- $K =$ number of mutation types.
- $N =$ number of signatures.
- $G =$ number of genomes.

The parts that make up the whole in mutational processes.

A somatic mutation catalog can be thought of as “a linear superposition of the signatures and intensities of exposure of mutational processes.”

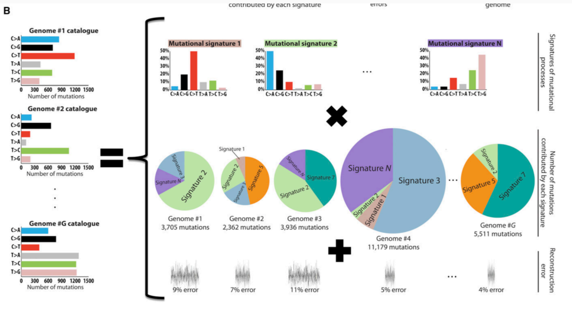


Figure 13:

Method for deciphering signatures of mutational processes

1. Input matrix M of dimension K (mutation types) by G (genomes).
2. Remove rare mutations ($\leq 1\%$).
3. Monte Carlo bootstrap resampling.

Method for deciphering signatures of mutational processes.

4. Apply the multiplicative update algorithm until convergence.

- Repeat steps 3 and 4 I times, each time storing P and E.
- Typical values $I = 400 - 500$

$$\min_{P \in \mathbf{M}_{\mathbf{R}_+}^{(\dot{K},N)},\, E \in \mathbf{M}_{\mathbf{R}_+}^{(N,G)}} \left\| \breve{M} - P \times E \right\|_F^2$$

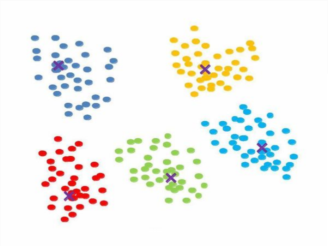
Figure 14:

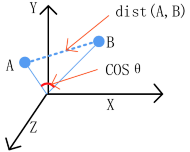
$$e_G^N \leftarrow e_G^N \frac{\left[P^T \breve{M}\right]_{N,G}}{\left[P^T P E\right]_{N,G}}$$

$$p_N^{\dot{K}} \leftarrow p_N^{\dot{K}} \frac{\left[\breve{M} E^T\right]_{\dot{K},N}}{\left[P E E^T\right]_{\dot{K},N}}$$

5. Cluster the signatures (columns of P matrix) from the I iterations into N clusters, one signature per cluster for each of the I matrices.

- This automatically clusters the exposures.
- Use cosine similarity for clustering.

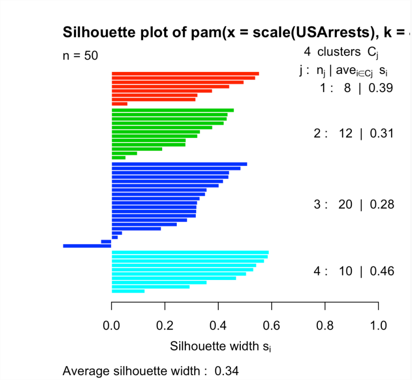




$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Method for deciphering signatures of mutational processes

6. Create the iteration averaged centroid matrix, $\overline{\mathrm{P}}$, by averaging the signatures within each cluster.
7. Evaluate the reproducibility of the signatures by calculating the average silhouette width over the N clusters.



Method for deciphering signatures of mutational processes

8. Evaluate the accuracy of the approximation of M by calculating the Frobenius reconstruction errors.

$$\min_{P \in \mathbf{M}_{\mathbf{R}_+}^{(K,N)},\, E \in \mathbf{M}_{\mathbf{R}_+}^{(N,G)}} \|\breve{M} - P \times E\|_F^2:$$

Figure 17:

9. Repeat steps 1-8 for different values of $N = 1, \ldots, min(K, G) - 1$.

Method for deciphering signatures of mutational processes

10. Choose an N corresponding to highly reproducible mutational signatures and low reconstruction error.

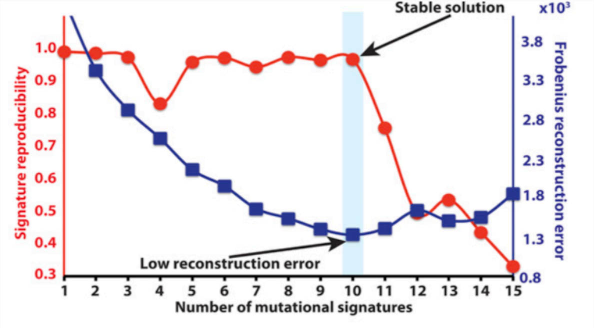


Figure 18:

The method recovers 10 signatures in a simulated cancer genome dataset

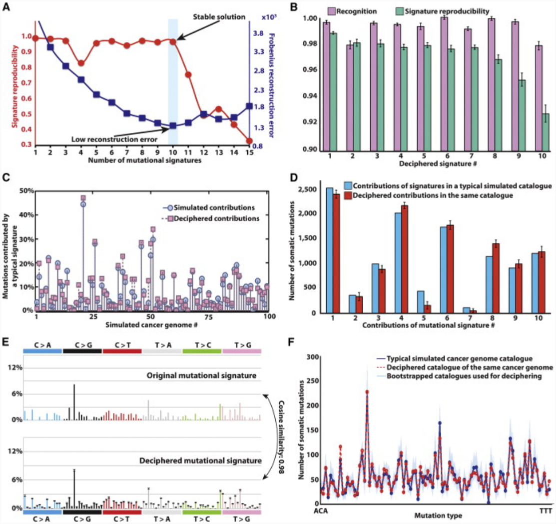


Figure 19:

The method recovers 10 signatures in a simulated cancer genome dataset

- 100 simulated cancer genome mutational catalogs
- 10 mutational processes with distinct signatures over 96 mutation types
- Add Poisson noise

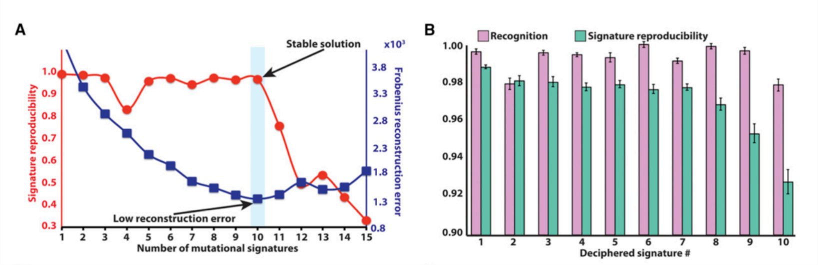


Figure 20:

The method recovers 10 signatures in a simulated cancer genome dataset

Deciphered and simulated contributions of the mutational signatures are similar.

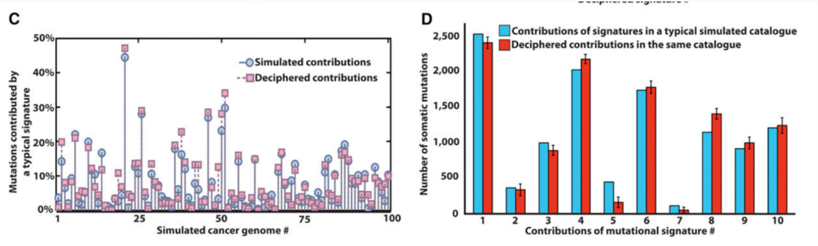


Figure 21:

The method recovers 10 signatures in a simulated cancer genome dataset

Deciphered and simulated mutation signatures and catalogs are similar.

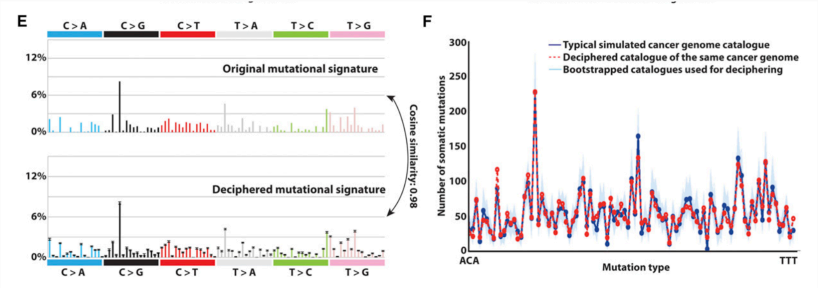


Figure 22:

The method is affected by the number of genomes, uniqueness of signatures, and number of mutations

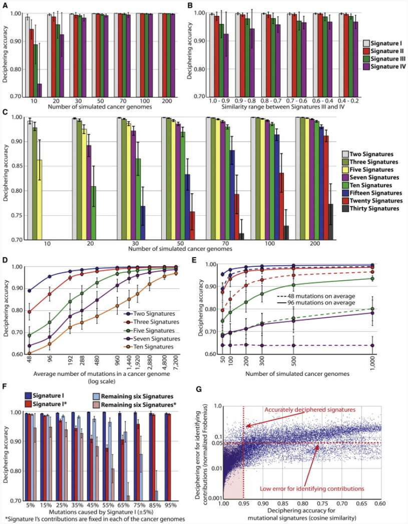


Figure 23:

The method is affected by the number of genomes, uniqueness of signatures, and number of mutations

The similarity of mutational signatures affects the method performance.

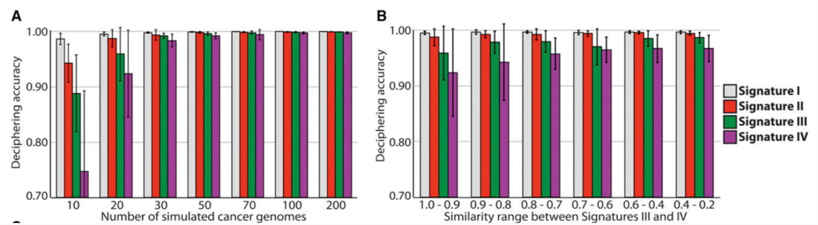


Figure 24:

The method is affected by the number of genomes, uniqueness of signatures, and number of mutations

More mutational signatures require more genomes.

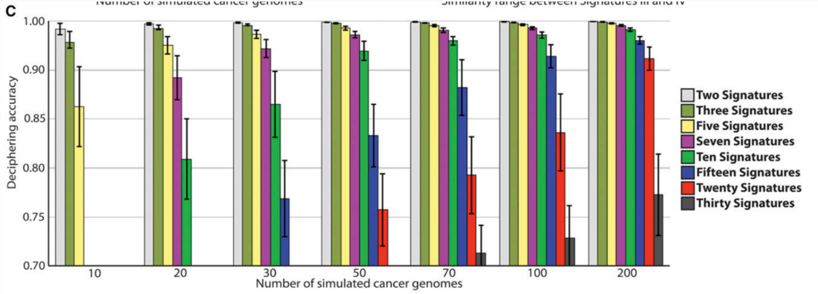


Figure 25:

The method is affected by the number of genomes, uniqueness of signatures, and number of mutations

The method performs better when there are more mutations in the cancer genomes.

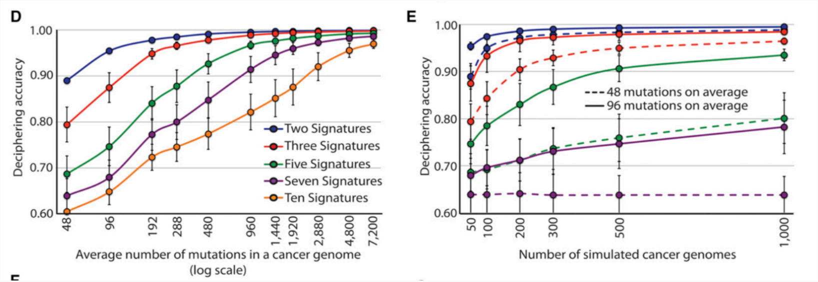


Figure 26:

The method is affected by the number of genomes, uniqueness of signatures, and number of mutations

Accurately deciphered mutational signatures correspond to accurate exposure estimates.

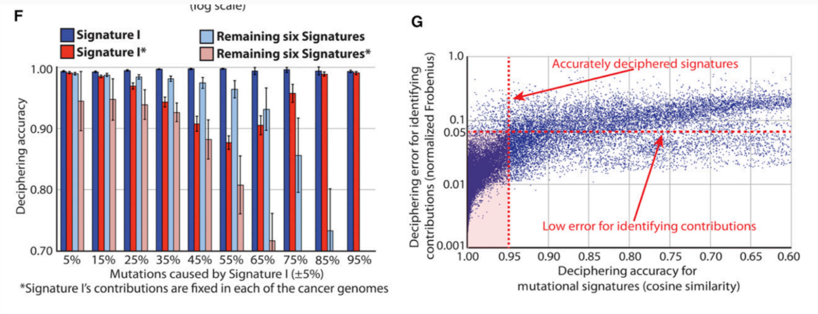


Figure 27:

Findings (Amanda)

- ...
- ...
- ...

We've talked enough (Amanda)

Discussion

- ...
- ...
- ...

References

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. 2013. "Deciphering Signatures of Mutational Processes Operative in Human Cancer." *Cell Reports* 3 (1). Elsevier: 246–59.

Hanahan, Douglas, and Robert A Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5). Elsevier: 646–74.