Discovering Cancer Signatures via Non-Negative Matrix Factorization

Nima Hejazi, Amanda Mok, Courtney Schiffman

2018-10-02

Introduction (Nima)

Overview

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

Overview

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

Non-Negative Matrix Factorization (Nima)

Why NMF?

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

What is NMF?

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

Alternative factorizations?

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

NMF versus PCA

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

Some fun with NMF

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

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

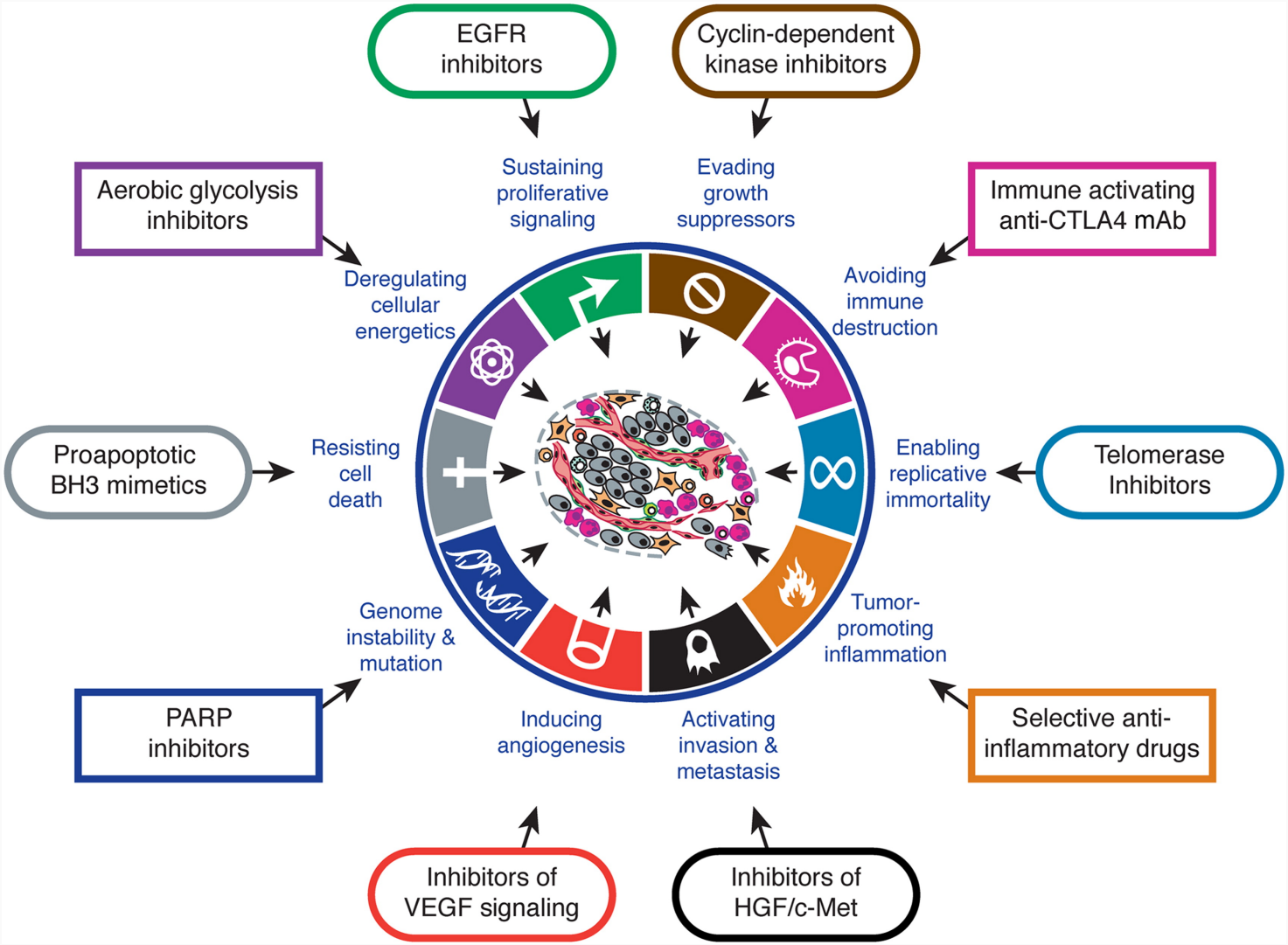


Figure 1: Hallmarks of Cancer

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 a biological challenge

Alexandrov et al. 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."

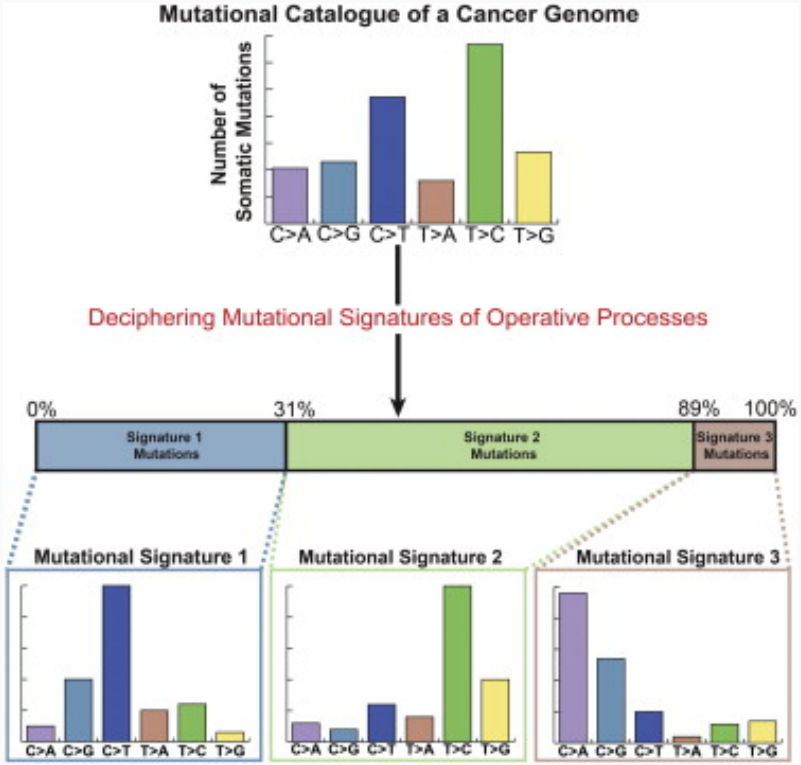


Figure 2:

NMF is a natural method for handling the BSS problem

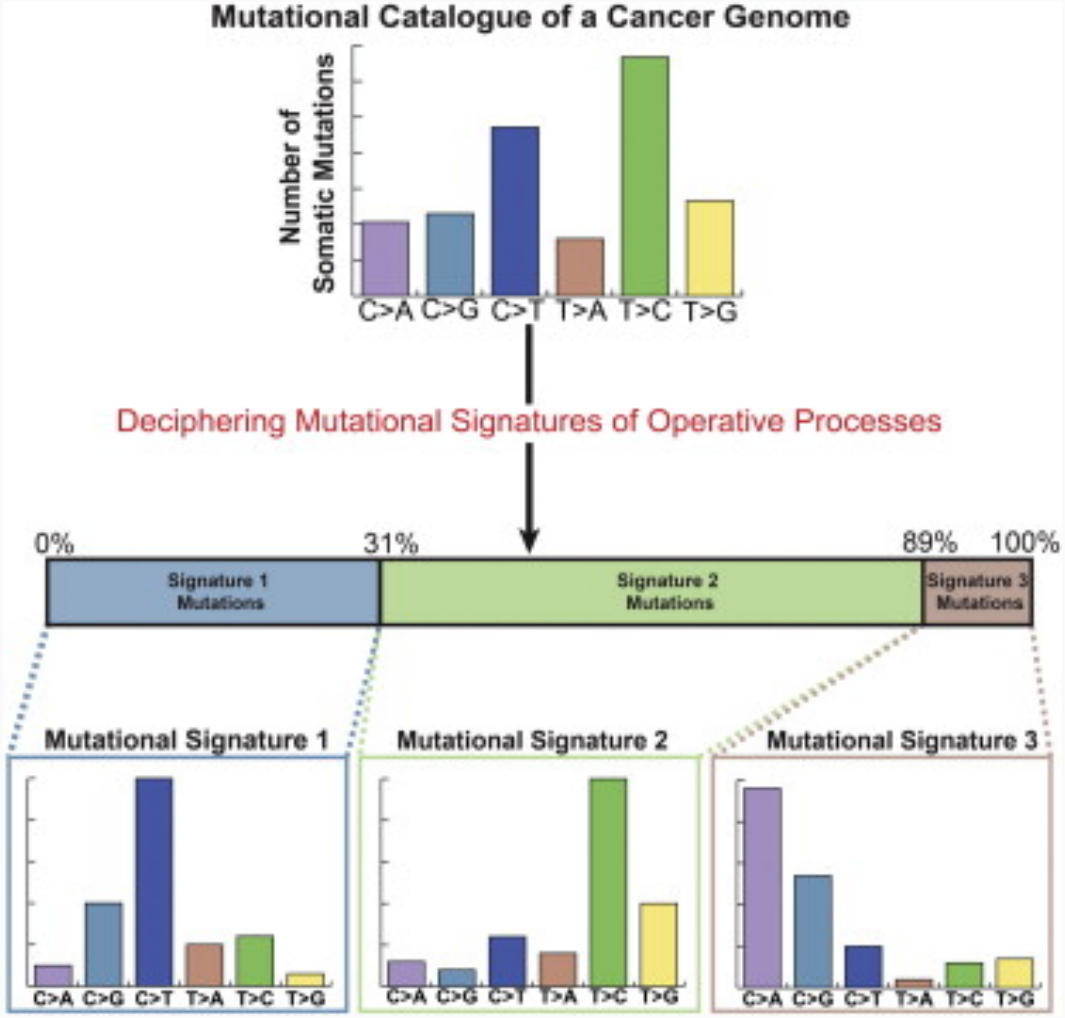
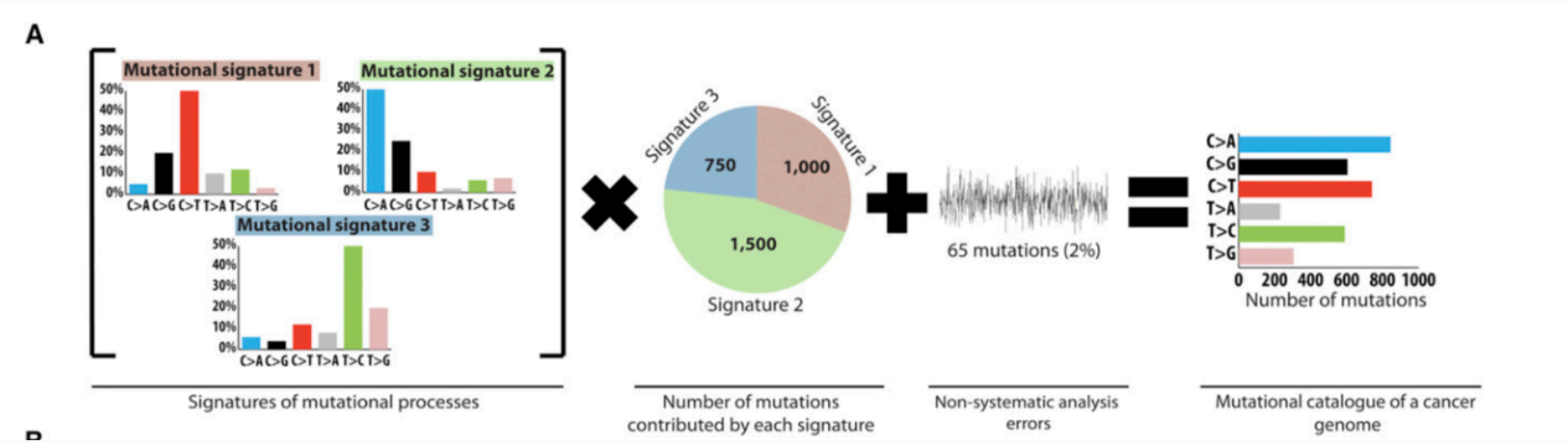


Figure 3:

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



$$M \approx P \times E$$

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?

$$\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix} \approx \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix} \times \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$$

Figure 4:

$$m_g^i \approx \sum_{n=1}^{N} p_n^i e_g^n.$$

Figure 5:

The parts that make up the whole in mutational processes

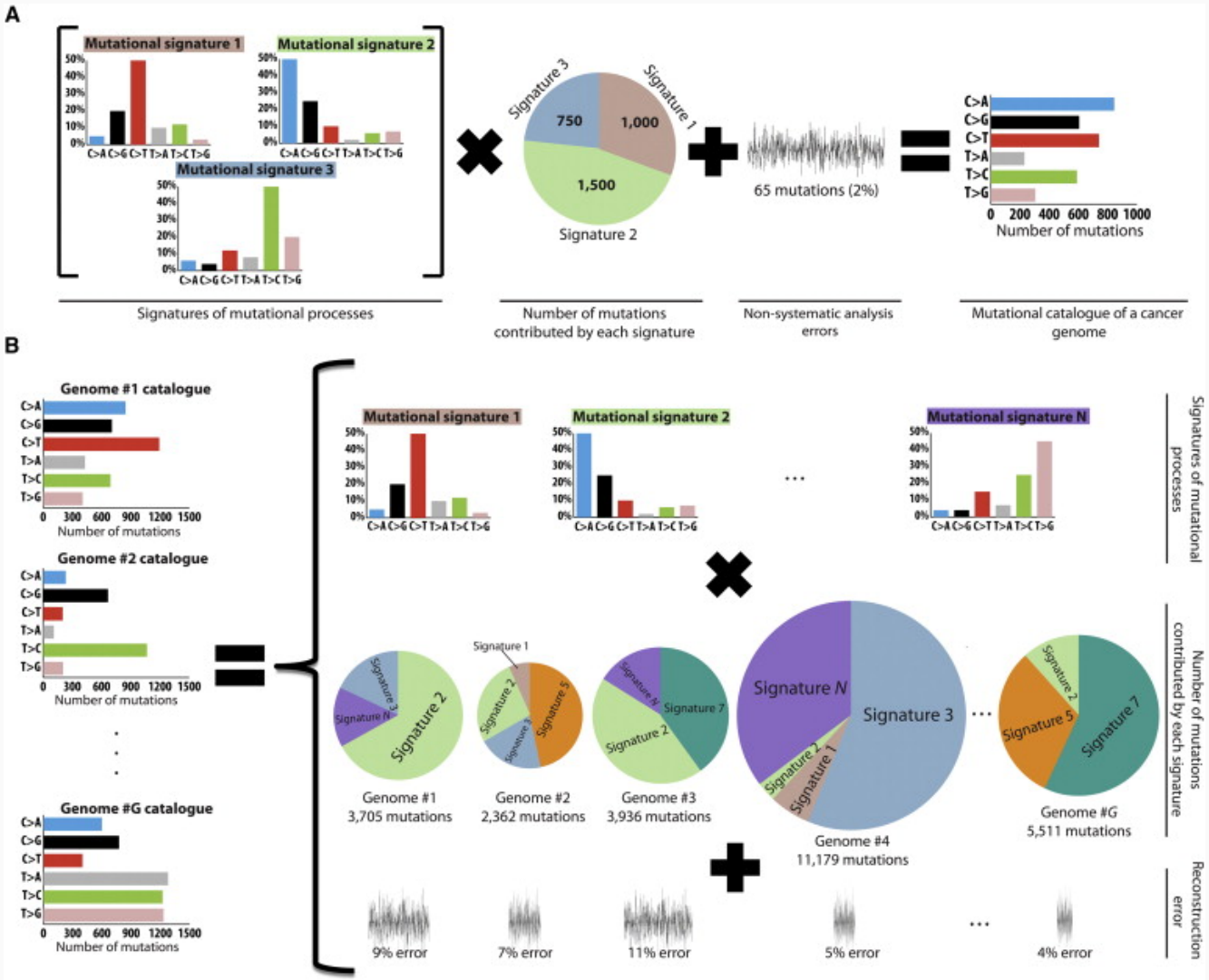


Figure 6:

Method for deciphering signatures of mutational processes

1. Input matrix M of dimension K (mutation types) by G (genomes).
2. Remove rare mutations ($< 1\%$).
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 $I = 400 - 500$

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

Figure 7:

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

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

Method for deciphering signatures of mutational processes

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.

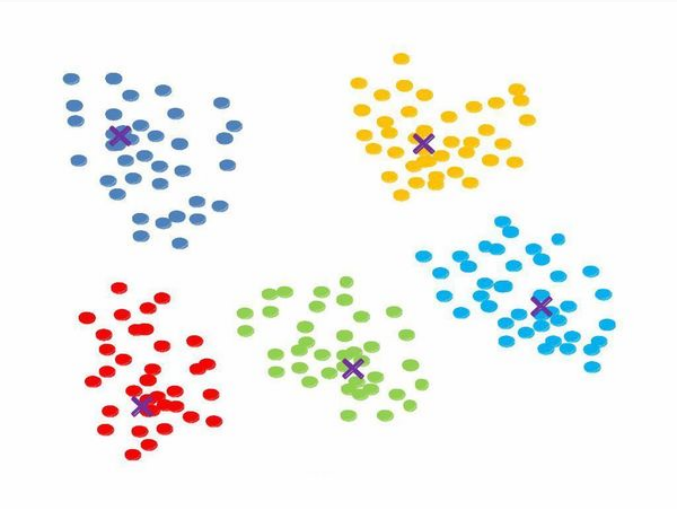


Figure 9:

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.

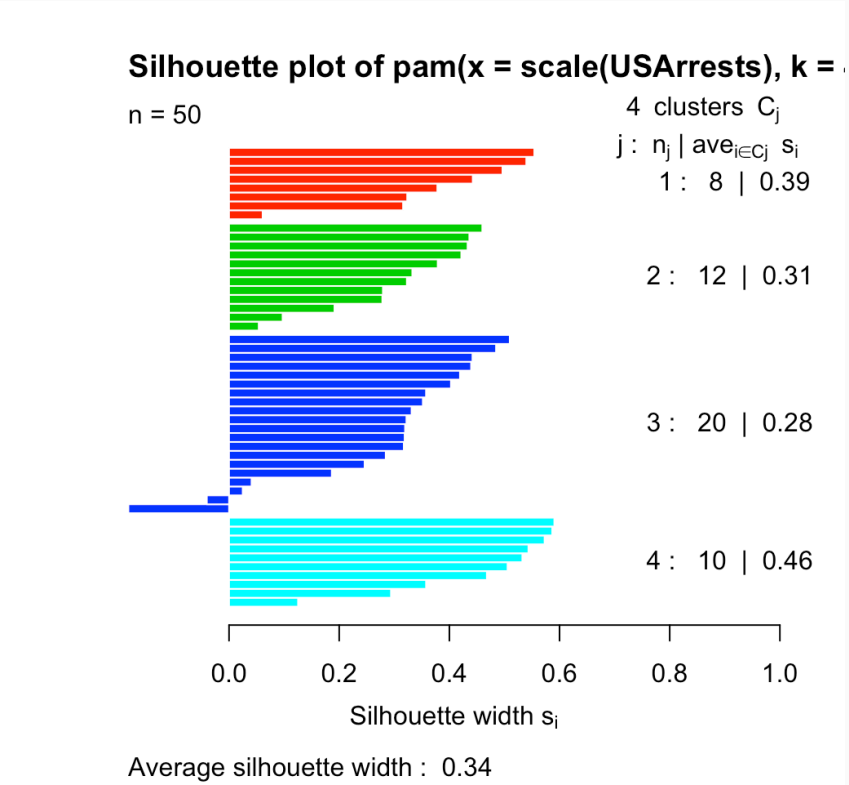


Figure 10:

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 11:

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.

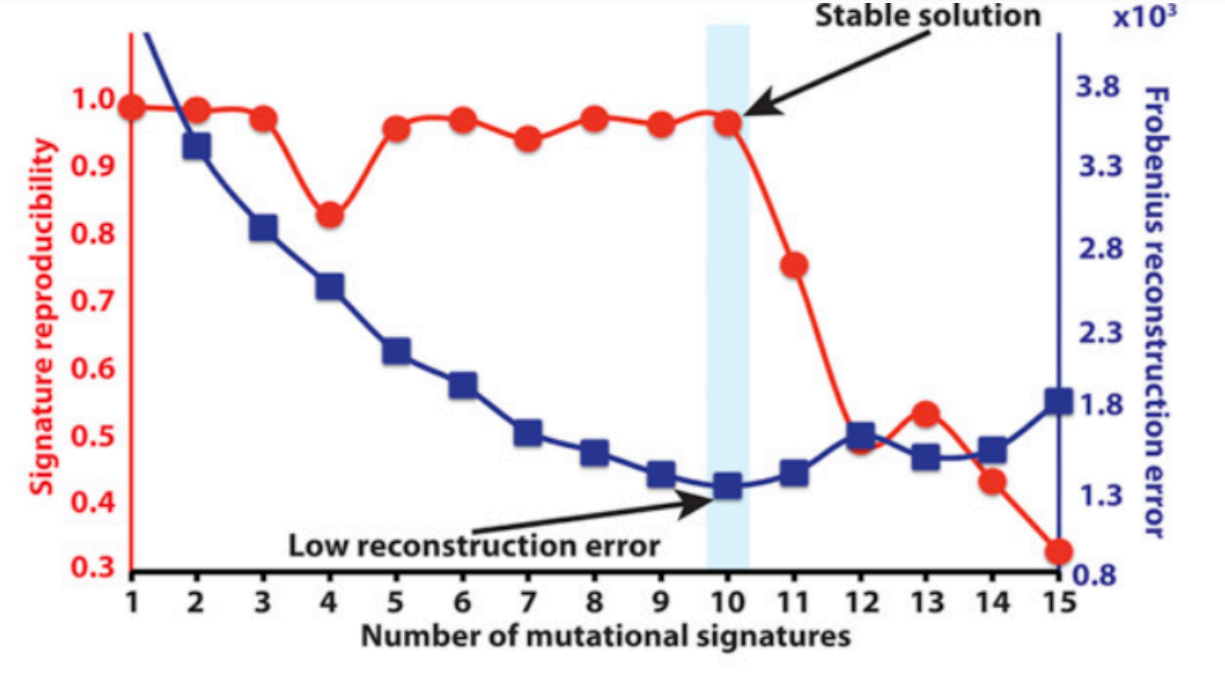


Figure 12:

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

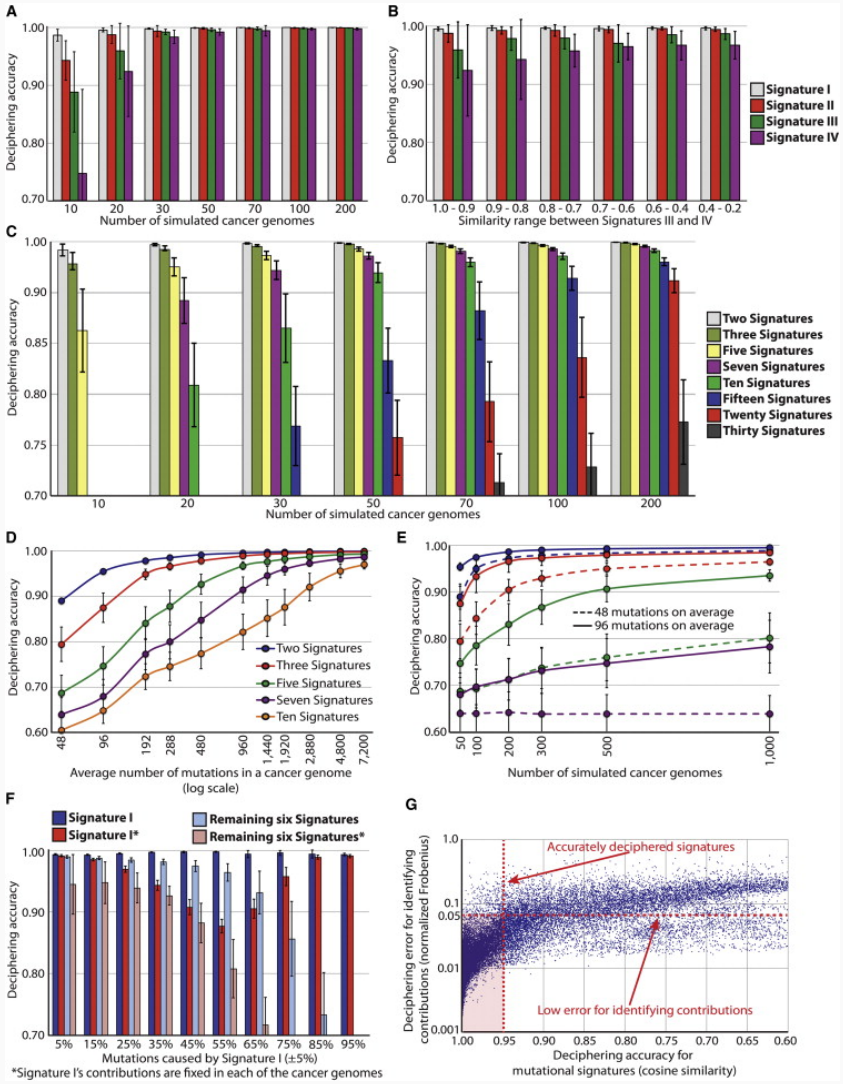


Figure 13:

The method recovers 10 signatures in a simulated cancer genome dataset

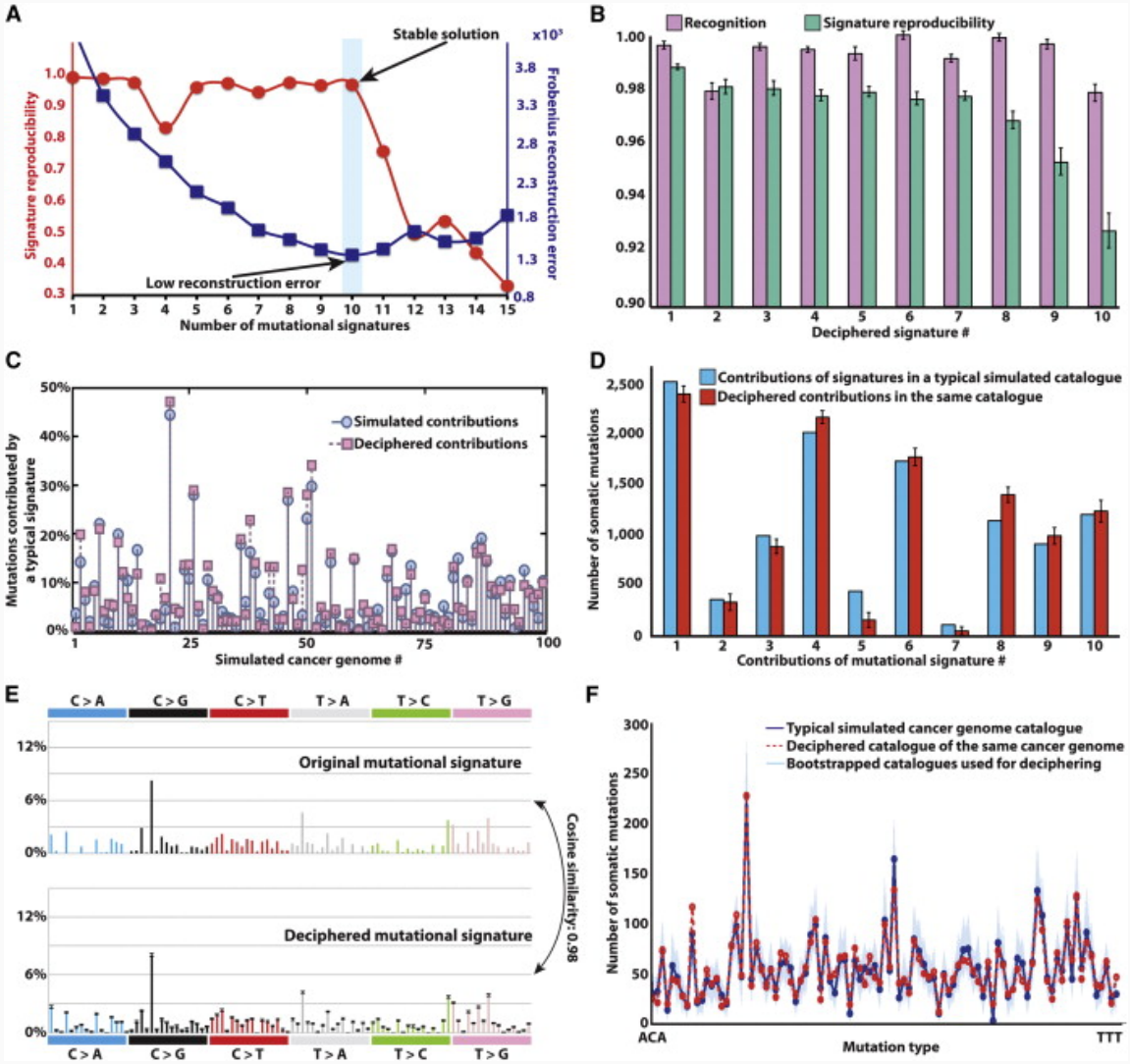


Figure 14:

Findings (Amanda)

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

We've talked enough (Amanda)

Discussion

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

References

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