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’s cancer?

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

Molecular biology of cancer

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

DNA

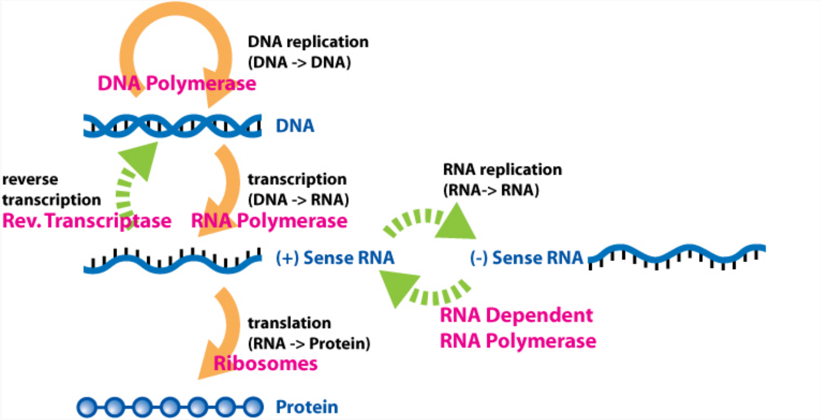

DNA replication
(DNA -> DNA)
DNA Polymerase
DNA
reverse
transcription
Rev. Transcriptase
transcription
(DNA -> RNA)
RNA Polymerase
RNA replication
(RNA-> RNA)
(+) Sense RNA
(-) Sense RNA
RNA Dependent
RNA Polymerase
translation
(RNA -> Protein)
Ribosomes
Protein

Figure 1:

Applying NMF to mutational processes

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.”

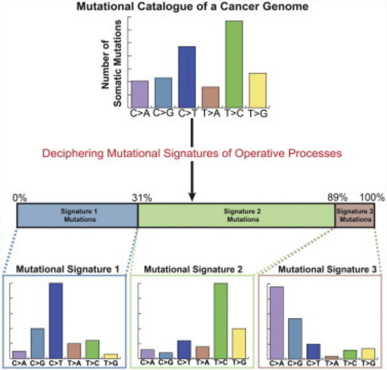
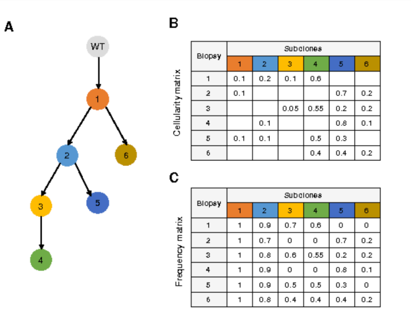


Figure 2:

How is the work of Alexandrov et al. 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.



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

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. related to inferring clonal evolution of tumors?

Different clonal mutations will have different signatures.

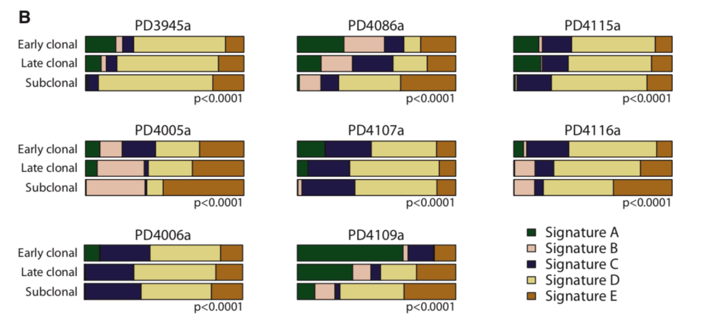


Figure 4:

Both works want to uncover driver mutations

Inferring clonal evolution of tumors

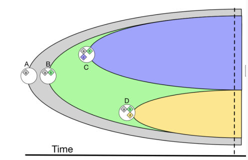
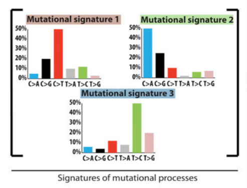


Figure 5:

Deciphering Signatures of mutational processes



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

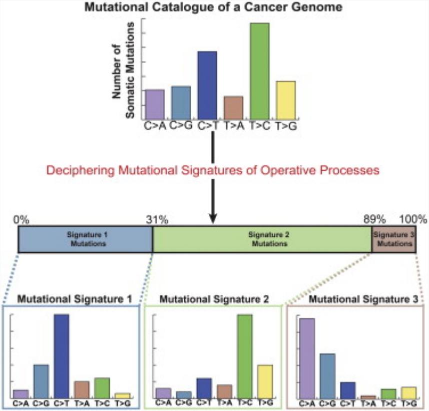


Figure 7:

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).

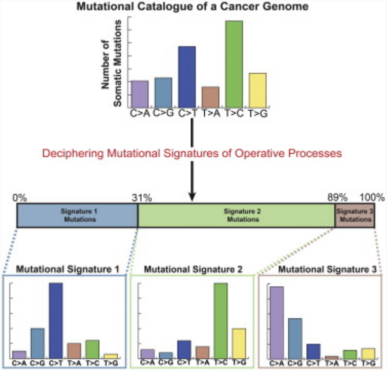
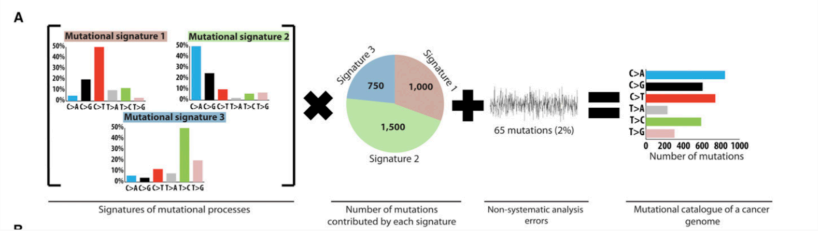


Figure 8:

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?

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

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

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

The parts that make up the whole in mutational processes.

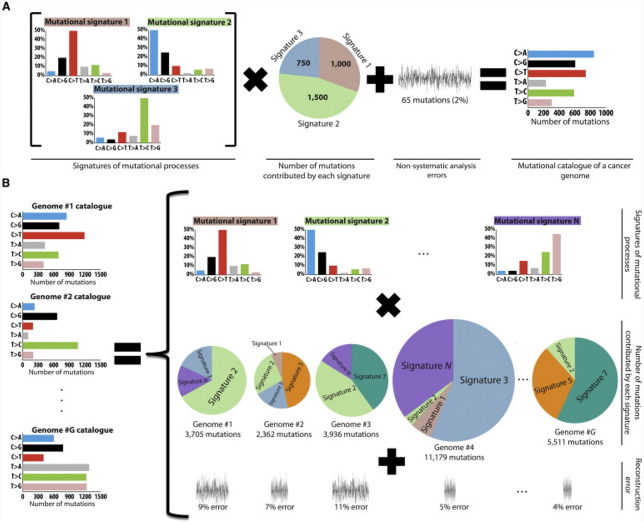


Figure 9:

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 $I = 400 - 500$

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

Figure 10:

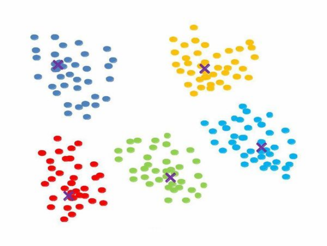
$$e_G^N \leftarrow e_G^N \frac{\left[P^T \breve{M}\right]_{N,G}}{\left[P^T P E\right]_{N,G}}$$

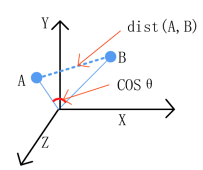
$$p_N^{\dot{K}} \leftarrow p_N^{\dot{K}} \frac{\left[\breve{M} E^T\right]_{\dot{K},N}}{\left[P E E^T\right]_{\dot{K},N}}$$

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.





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

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.

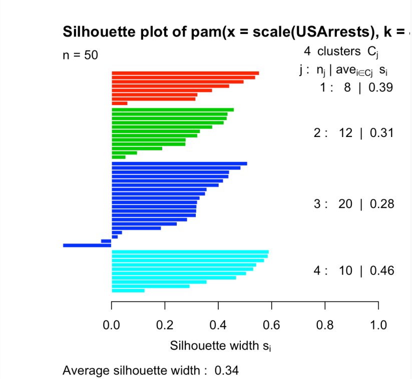
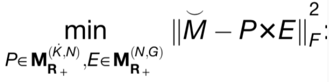


Figure 13

Method for deciphering signatures of mutational processes

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



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

Figure 13:

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.

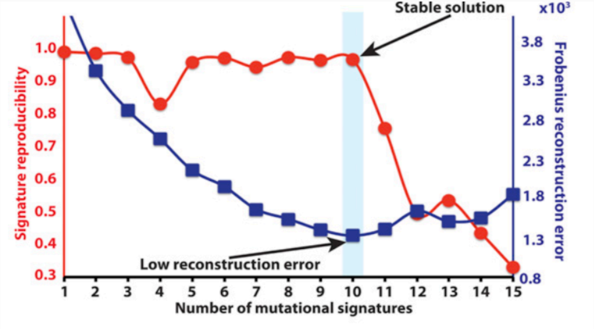


Figure 14:

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

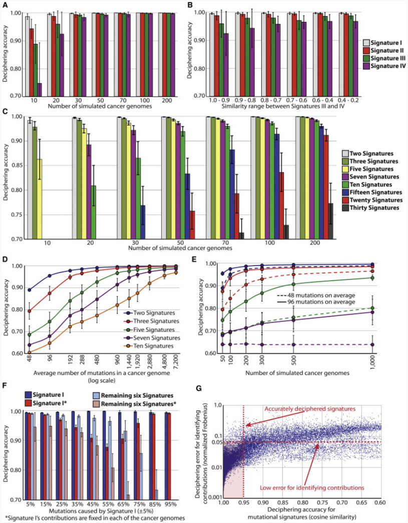


Figure 15:

The method recovers 10 signatures in a simulated cancer genome dataset

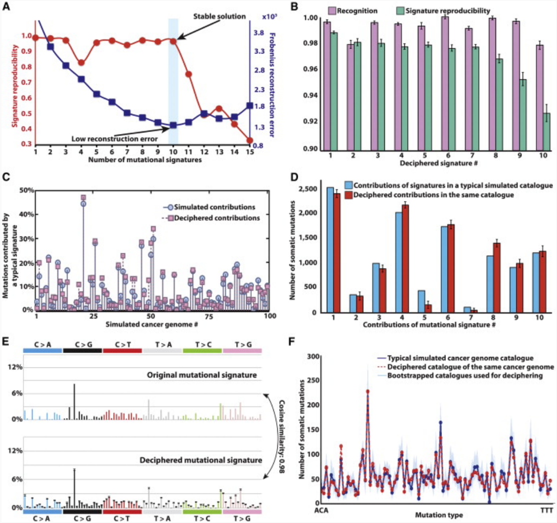


Figure 16:

Findings (Amanda)

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

We’ve talked enough (Amanda)

Discussion

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

References