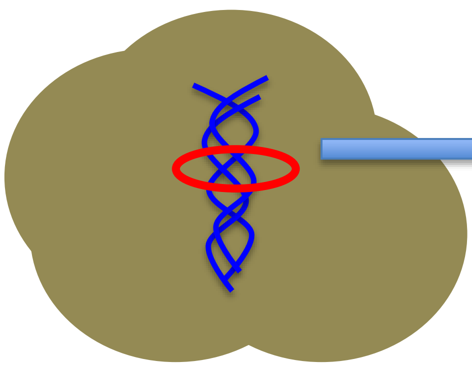
Deciphering Signatures of Mutational Processes Operative in Human Cancer

Tumor Cells Carry Somatic Mutations

Tumor



Sequence

gcttcgctagcgccccttttaatcgatcccgatcg
cccacgatcggatagctagatcgactgttttaatt
agcccacatcactatctccctttttgggagacgatc
atgccccggtttcgaatgctaaaatgctaaagttt
cccacgatcggatagctagatcgactgttttaatt
cagctactgatcgttttgccggcccccgggagat
atgccccggtttcgaatgctaaaatgctaaagttt

<u>Catalog</u>

1. acgatcg
2. ctcccttt
3. tcggata
4. gactgttt
5. gccccgg

….. 500

Motivation

- Catalogs have heterogeneity
  - Different mutation types: Substitution, missense, nonsense, indels
  - DNA Repair mechanisms
  - Passenger mutations

- Many different cancer signatures

Aim to create computational framework to bridge the gap between the catalogs and signatures

Catalog

1. acgatcg
2. ctcccttt
3. tcggata
4. gactgttt
5. gccccgg

….. 500

Lung Cancer Signature

1. Gcgta (G:C > T:A)
2. Cttccg Deletion
3. tcggata

Feature of Signatures

$$P_1 = [p_1^1, p_1^2, \dots p_1^K]^T$$

P = Mutational Signature

$p_{1...k}$ = probability P causes a certain mutation

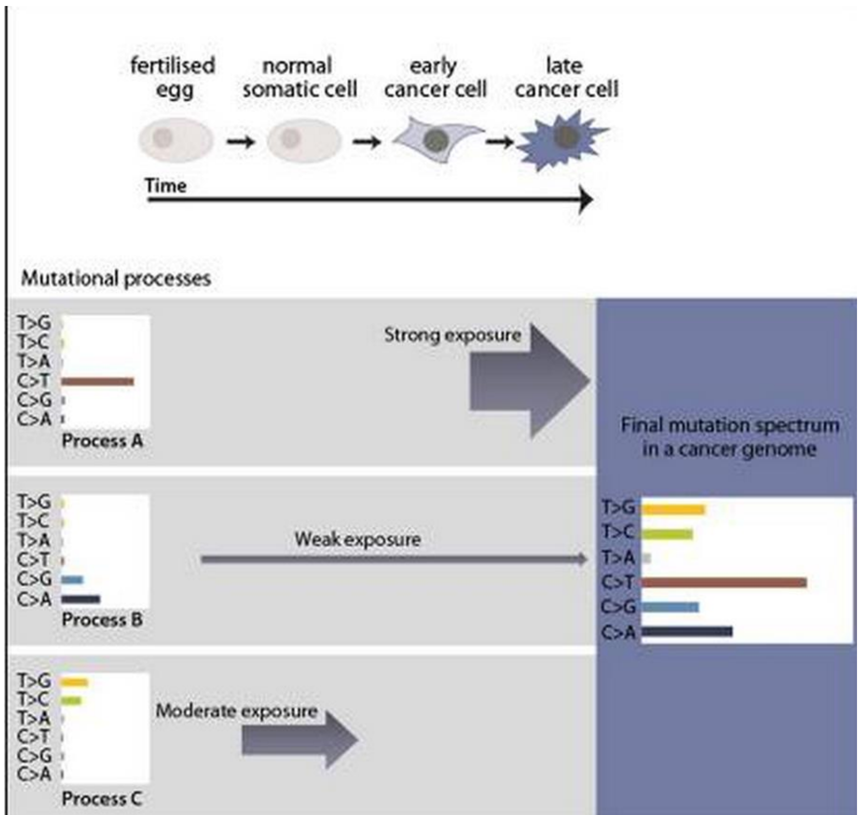
K = 96 (6 types of substitutions * 4 types of 5' bases * 4 types of 3' bases)

Mapping of a Genome

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

P = process/mutation

e = exposure/weight



What we end up with

$$P = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$$

X

$$E = \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$$

$$= \quad M = \begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix}$$

Non-Negative Matrix Factorization

- Want to extract "P" and "e" from M

Step 1 and 2

Reduce Matrix Dimensions $$\sum_{r \in R} \sum_{g=1}^{G} m_g^r \leq 0.01 \times \sum_{k=1}^{K} \sum_{g=1}^{G} m_g^k,$$

Use bootstrap resampling

Step 3&4: Non Negative Matrix Factorization

- All inputs must be non-negative
- Aims to recreate P and e from M

Iterate until convergence

Minimize

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

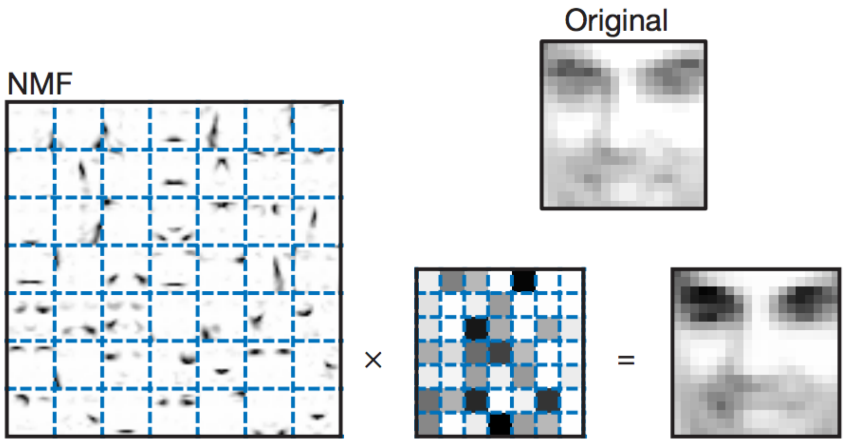
Cost Function

$$\left\| \breve{M} - P \times E \right\|_F^2$$

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

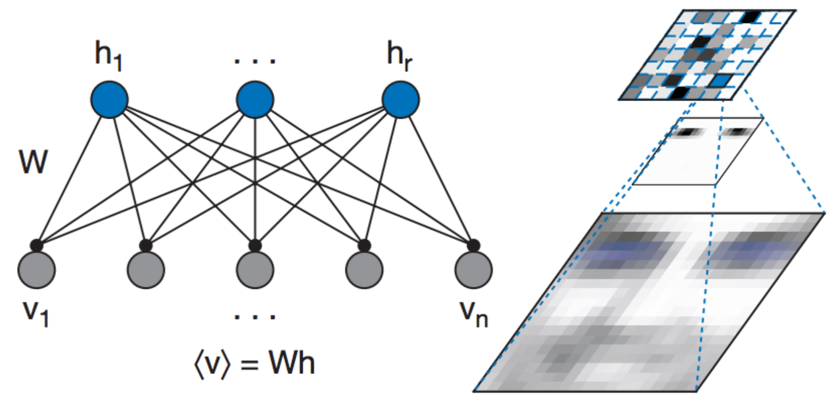
Equivalent to $(K,N)^{th}$ element of matrix

NMF: Faces



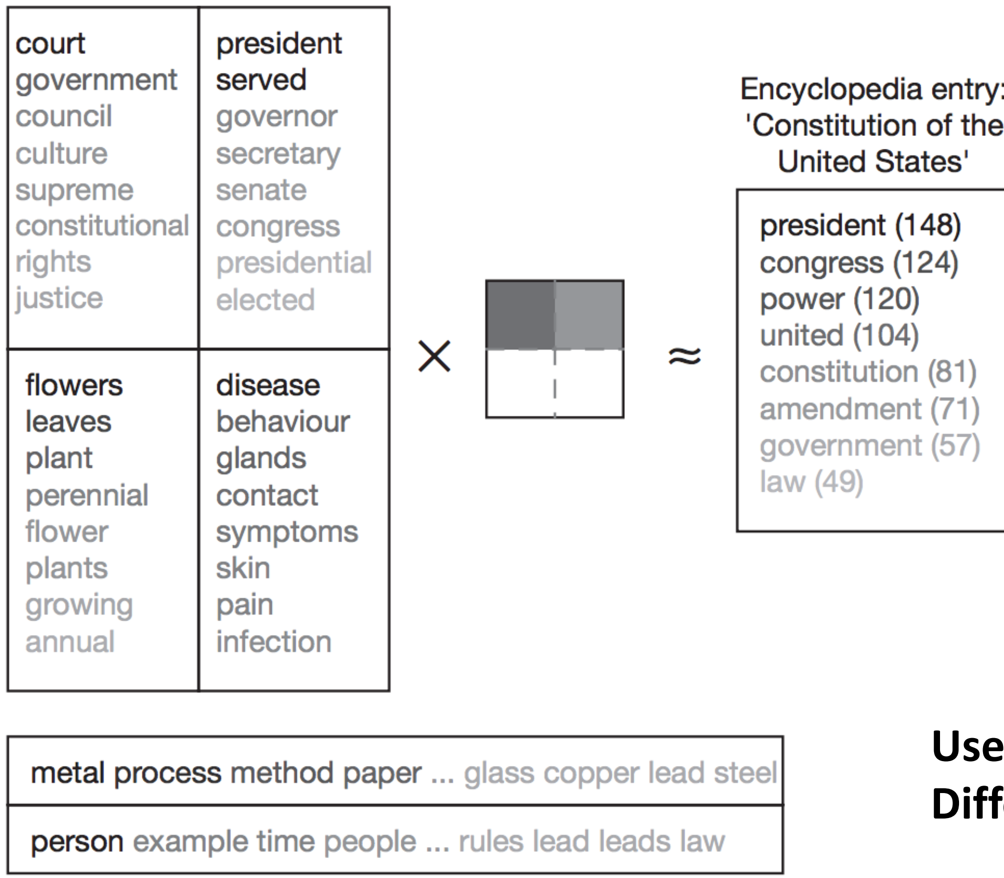
W Basis

H Encodings



From *Lee and Seung, 1999*

NMF: Encyclopedia



Breaks topics into Related words

Uses context to Differentiate

From *Lee and Seung, 1999*

Step 5: Clustering

- Partition-clustering algorithm was applied to cluster data into N clusters

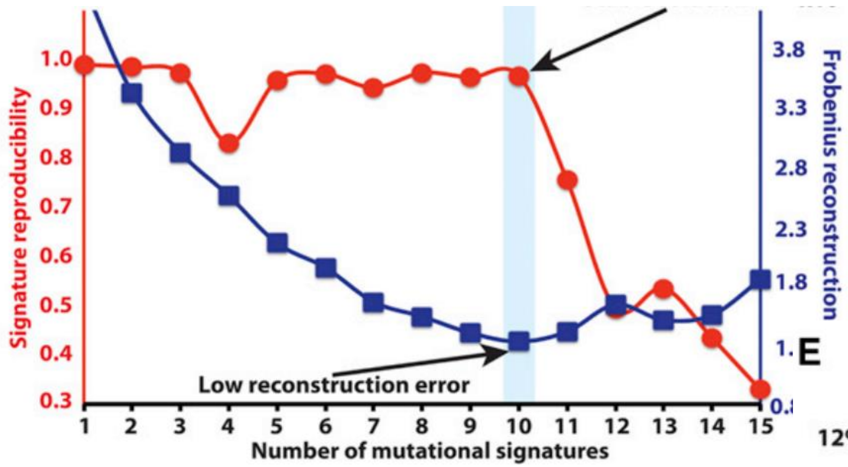
Step 6: Evaluate

- Look at Frobenius reconstruction error to evaluate for accuracy
- Compare mutational signatures:

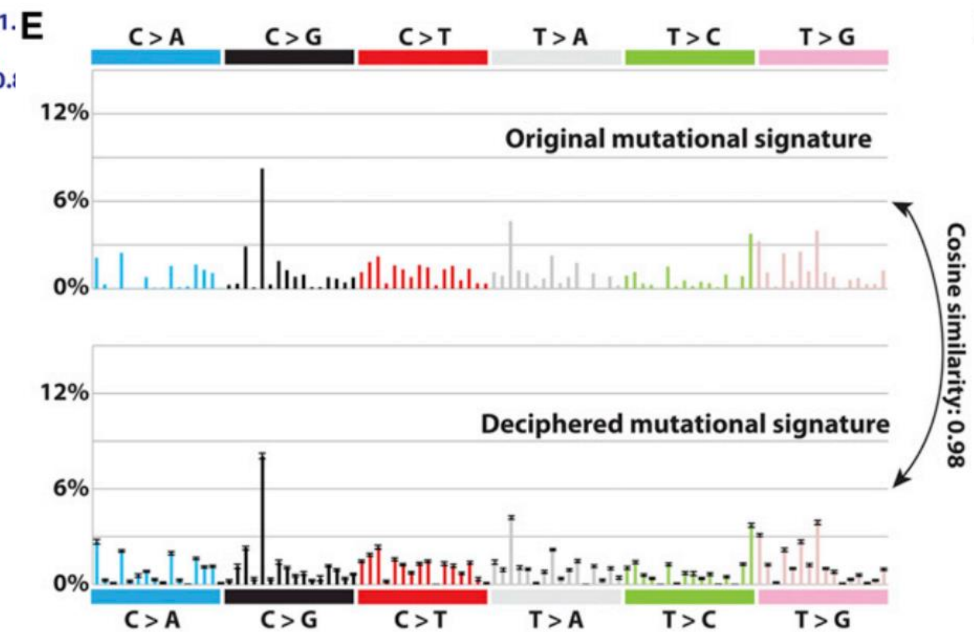
$$\mathrm{sim}(A,B)=\frac{\sum_{k=1}^{K} A_k B_k}{\sqrt{\sum_{k=1}^{K}(A_k)^2}\sqrt{\sum_{k=1}^{K}(B_k)^2}}.$$

Sim(A,B) = 1 means same signature

Does it work?

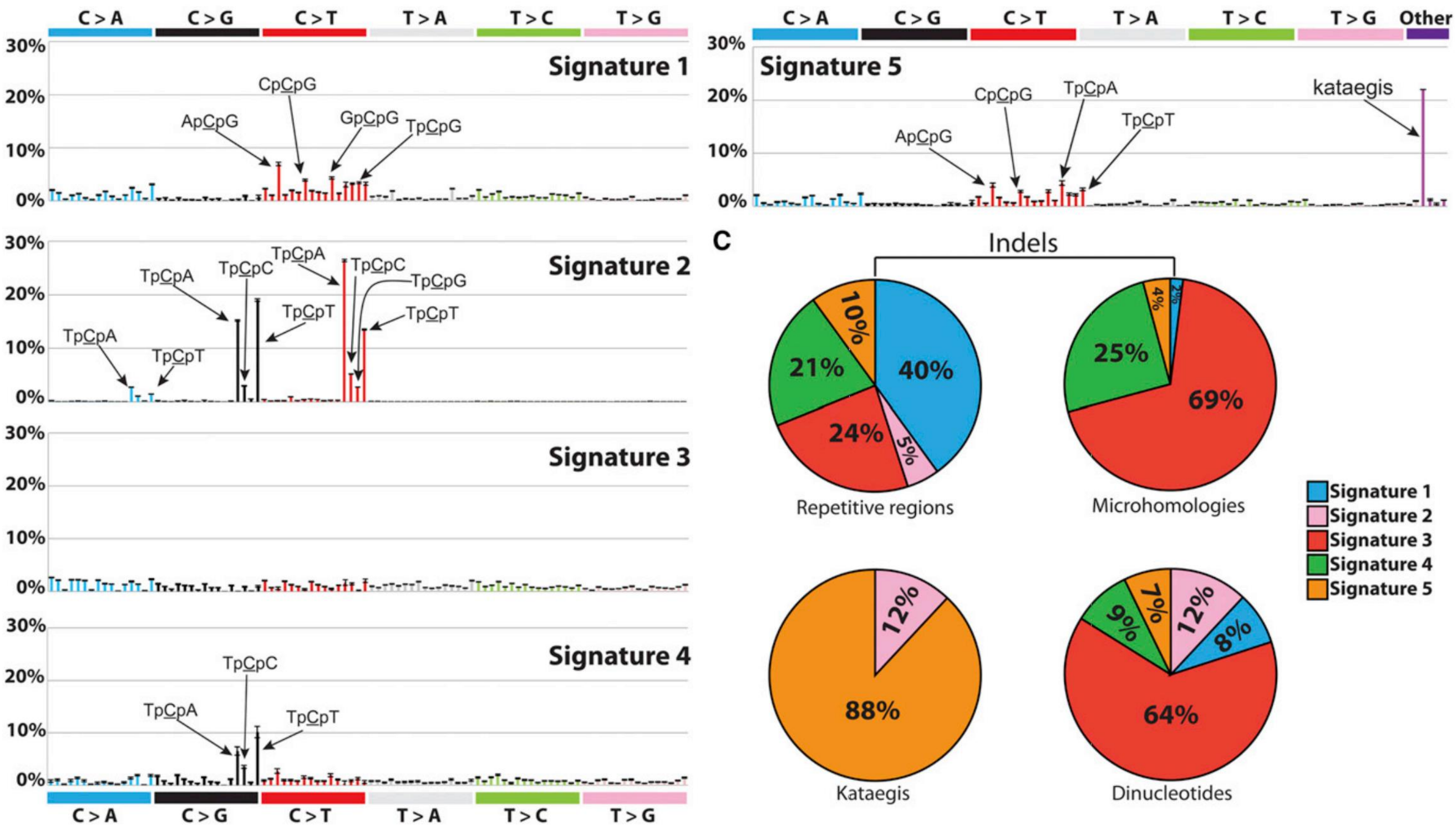


Signature reproducibility
Frobenius reconstruction
Low reconstruction error
Number of mutational signatures
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
3.8
3.3
2.8
2.3
1.8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15



E
C > A
C > G
C > T
T > A
T > C
T > G
12%
6%
0%
Original mutational signature
Deciphered mutational signature
Cosine similarity: 0.98

Breast Cancer Example



Impact

- Ability to generate cancer signatures from comprehensive ‘omic data
- Opens the door for further work. Eg. Sparsity constraint to use a minimum number of signatures