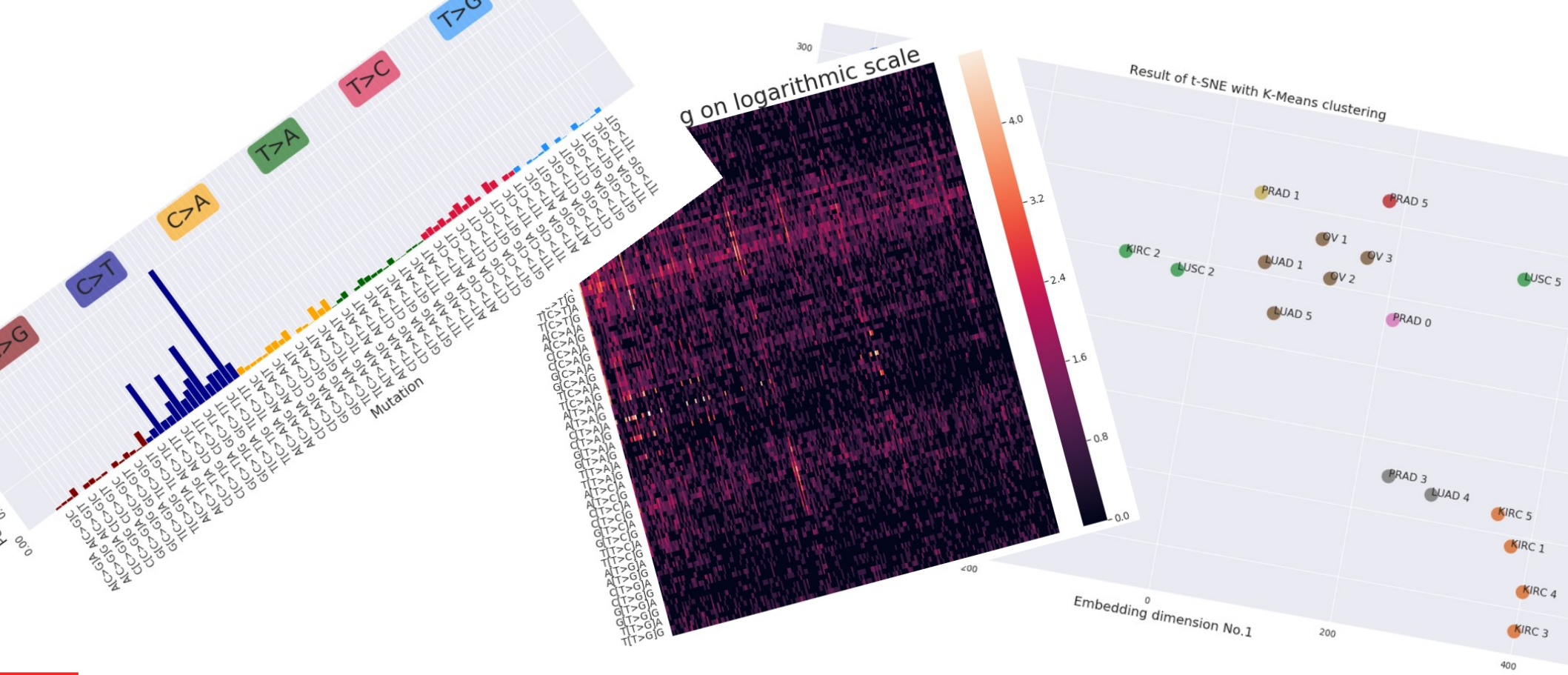


Signatures of mutational processes in human cancer

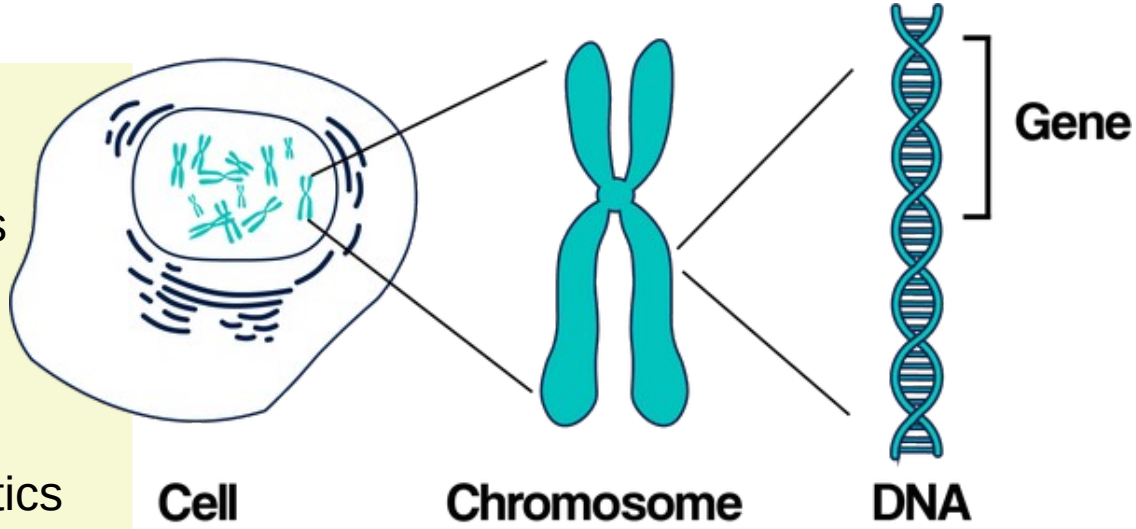
Noel László Plaszkó

Data Science Laboratory
Supervisor: Orsolya Pipek

Introduction: chromosomes, genes

Chromosomes

- thread-like structures in cells
- 46 chromosomes (arranged in 2 sets of 23)
- one set from your mother, one from your father
- determine physical characteristics
- contain genes

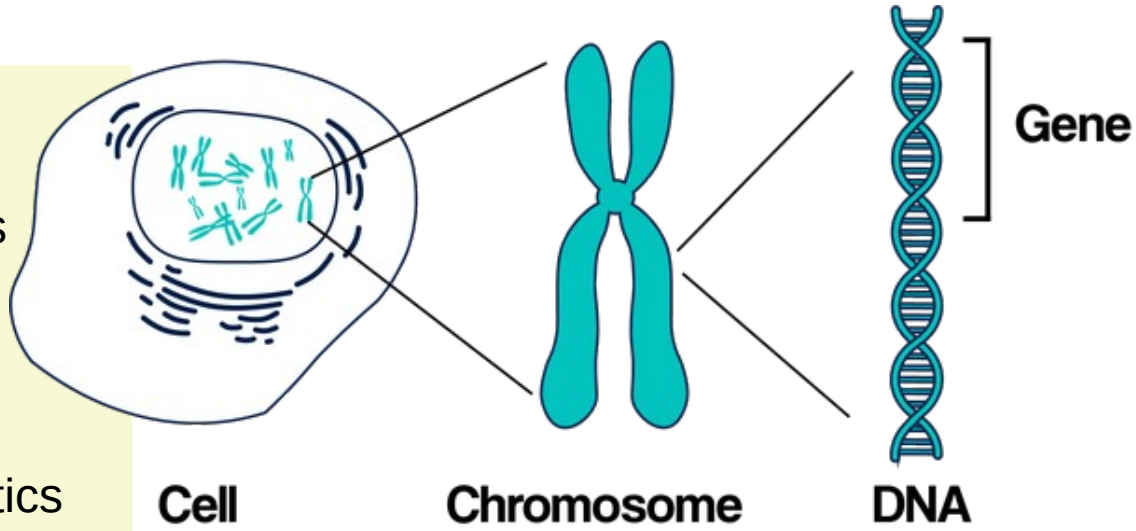


[1] https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer

Introduction: chromosomes, genes

Chromosomes

- thread-like structures in cells
- 46 chromosomes (arranged in 2 sets of 23)
- one set from your mother, one from your father
- determine physical characteristics
- contain genes



Genes

- the basic physical unit of inheritance
- control how your cells work by making proteins
- must have the correct instructions for making its protein
- allows the protein to perform the correct function for the cell

[1] https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer

Sorry, something went wrong…

- Genes in cells can mutate
- Due to mutation:
  - ➔ Abnormal protein is created
  - ➔ Or protein's formation is prevented

Abnormal protein provides different information

↓

This can cause cells to multiply uncontrollably and become cancerous

Cancer is caused
by somatic mutations of the DNA in the tumor cells.

[2] https://www.nature.com/articles/d41586-019-03536-x

DNA sequencing and mutations

DNA sequence:

- the order of the four bases in DNA (adenine, guanine, cytosine, and thymine)

DNA mutation:

Where my story begins...

[3] https://www.singerinstruments.com/resource/what-are-genetic-mutation/

Somatic mutations for different cancer types

File name	Type of cancer
KIRC.maf	kidney renal clear cell carcinoma
LUAD.maf	lung adenocarcinoma
LUSC.maf	lung squamous cell carcinoma
OV.maf	ovarian cancer
PRAD.maf	prostate adenocarcinoma

Somatic mutations for different cancer types

File name	Type of cancer
KIRC.maf	kidney renal clear cell carcinoma
LUAD.maf	lung adenocarcinoma
LUSC.maf	lung squamous cell carcinoma
OV.maf	ovarian cancer
PRAD.maf	prostate adenocarcinoma

Files have many columns, but only a few will be necessary

↓

For what?!

↓

Different mutational processes (“signatures”) exist

↓

Generate different combinations of mutation types

Somatic mutations for different cancer types

File name	Type of cancer	Number of Samples	Number of mutations
KIRC.maf	kidney renal clear cell carcinoma	235	26245
LUAD.maf	lung adenocarcinoma	561	232492
LUSC.maf	lung squamous cell carcinoma	497	173223
OV.maf	ovarian cancer	142	6174
PRAD.maf	prostate adenocarcinoma	499	36805

Files have many columns, but only a few will be necessary

↓

For what?!

↓

Different mutational processes (“signatures”) exist

↓

Generate different combinations of mutation types

What is in the .maf files?

- Hugo_Symbol
- Entrez_Gene_Id
- Center
- Ncbi_Build
- **Chrom**
- **Start_Position**
- **End_Position**
- Strand
- Variant_Classification
- Variant_Type
- **Reference_Allele**
- Tumor_Seq_Allele1
- **Tumor_Seq_Allele2**
- Dbsnp_Rs
- Dbsnp_Val_Status
- Tumor_Sample_Barcode
- Matched_Norm_Sample_Barcode
- Match_Norm_Seq_Allele1
- Match_Norm_Seq_Allele2
- Tumor_Validation_Allele1
- Tumor_Validation_Allele2
- Match_Norm_Validation_Allele1
- Match_Norm_Validation_Allele2
- Verification_Status
- Validation_Status
- Mutation_Status
- Sequencing_Phase
- Sequence_Source
- Validation_Method
- Score
- Bam_File
- Sequencer
- **Tumor_Sample_UUID**
- Matched_Norm_Sample_UUID
- File_Name
- Archive_Name
- Line_Number

Chrom
Start_Position
End_Position
Reference_Allele
Tumor_Seq_Allele2
Tumor_Sample_UUID

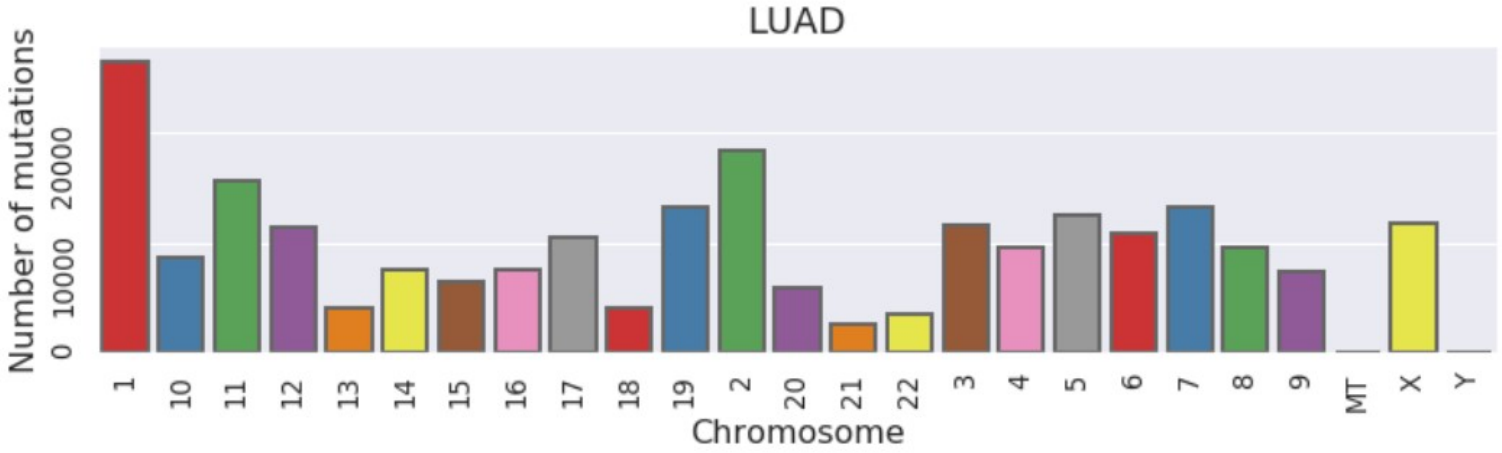
What is in the .maf files?

- Hugo_Symbol
- Entrez_Gene_Id
- Center
- Ncbi_Build
- **Chrom**
- **Start_Position**
- **End_Position**
- Strand
- Variant_Classification
- Variant_Type
- **Reference_Allele**
- Tumor_Seq_Allele1
- **Tumor_Seq_Allele2**
- Dbsnp_Rs
- Dbsnp_Val_Status
- Tumor_Sample_Barcod
- Matched_Norm
- Match
- _Method
- ore
- Bam_File
- Sequencer
- **Tumor_Sample_UUID**
- Matched_Norm_Sample_UUID
- File_Name
- Archive_Name
- Line_Number

Chrom	Start_Position	End_Position	Reference_Allele	Tumor_Seq_Allele2	Tumor_Sample_UUID
10	52587996	52587996	C	G	d7107990-0f77-4440-b760-a19d4a920c04
12	8990969	8990969	T	A	90f1ea49-ae0d-4f4d-9df0-cb993ec99642
12	8990997	8990997	G	A	b4b3281b-5dfc-434d-a9a6-2c2d94123220
12	9016477	9016477	C	G	ca5ac140-34e2-4684-ad12-65ae699172fc
12	9020925	9020925	C	T	ae2cb5e3-cd36-4d7c-97dc-2f72a08dddb1

ence_Allele
umor_Seq_Allele2

What else do we need?

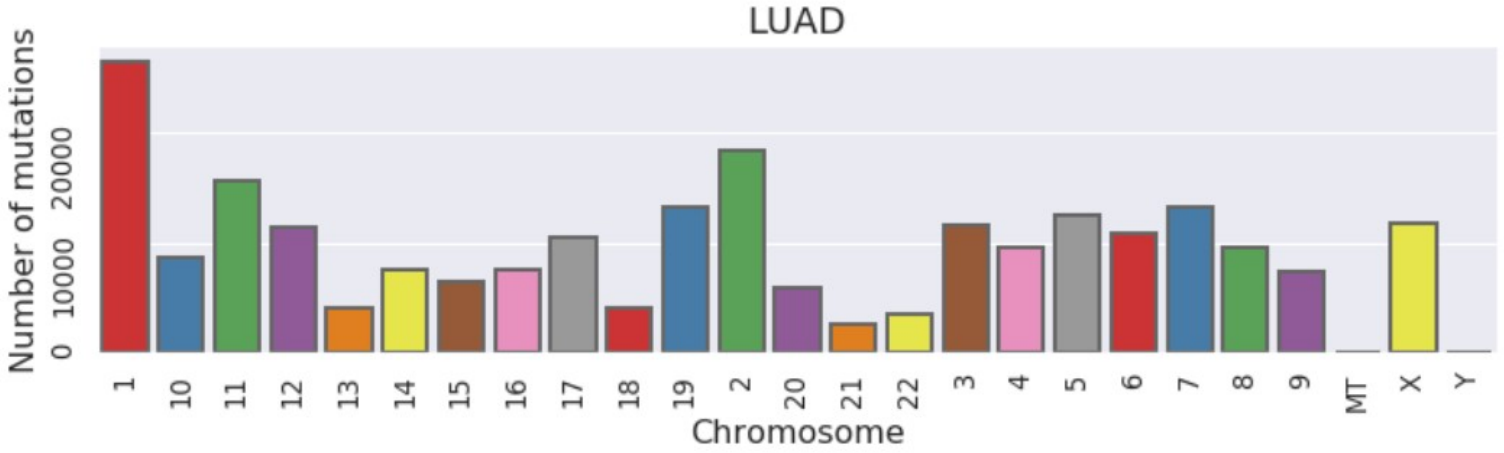


Info from substitution is not enough

↓

Take the neighbours into account

What else do we need?



Info from substitution is not enough

↓

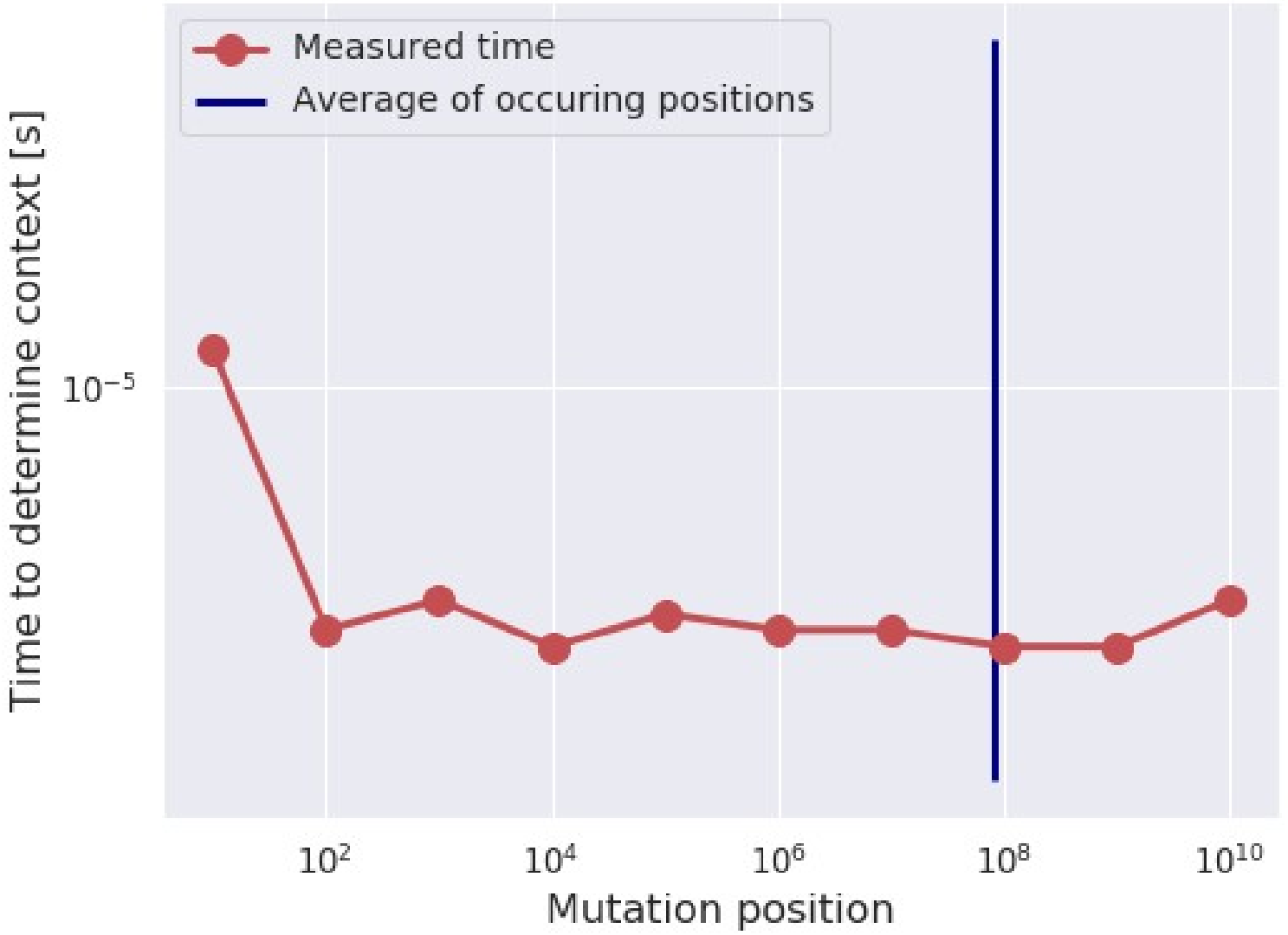
Take the neighbours into account

↓

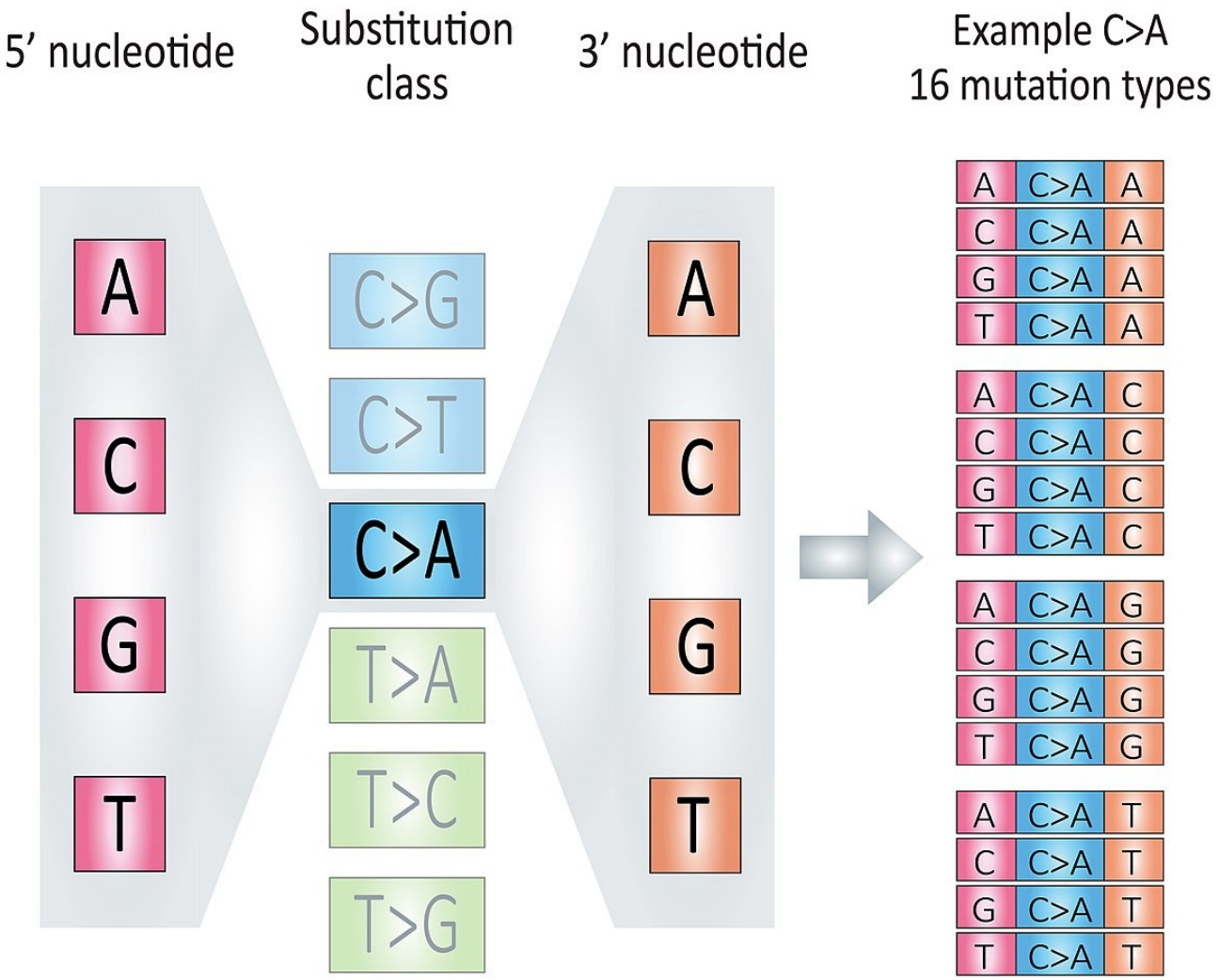
For this purpose **hg19 human reference genome** is used

```
...TGTGTGTGTCCACACTTCCTCTCATGAGAACAG
CAGGTTGCTTTAGGGCCCACCCTGACAGCCTCGTTC
TAATACTATGAGGCCAAATACACTCACGTTCT...
```

What else do we need?



Mutational catalogs



Go to the notebook, to see catalogs...

[4] https://en.wikipedia.org/wiki/Mutational_signatures

Non-negative matrix factorization

Goal: Find two non-negative matrices (W, H) whose product approximates the non-negative matrix X

Objective:

$$
\begin{aligned}
&0.5 * ||X - WH||^2_{loss} \\
&+alpha\_W * l1_{ratio} * n\_features * ||vec(W)||_1 \\
&+alpha\_H * l1_{ratio} * n\_samples * ||vec(H)||_1 \\
&+0.5 * alpha\_W * (1 - l1_{ratio}) * n\_features * ||W||^2_{Fro} \\
&+0.5 * alpha\_H * (1 - l1_{ratio}) * n\_samples * ||H||^2_{Fro}
\end{aligned}
$$

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

Non-negative matrix factorization

Goal: Find two non-negative matrices (W, H) whose product approximates the non-negative matrix X

Objective:

$$
\begin{aligned}
&0.5 * ||X - WH||^2_{loss} \\
&+alpha\_W * l1_{ratio} * n\_features * ||vec(W)||_1 \\
&+alpha\_H * l1_{ratio} * n\_samples * ||vec(H)||_1 \\
&+0.5 * alpha\_W * (1 - l1_{ratio}) * n\_features * ||W||^2_{Fro} \\
&+0.5 * alpha\_H * (1 - l1_{ratio}) * n\_samples * ||H||^2_{Fro}
\end{aligned}
$$

Hyperparameter tuning:

alpha_W = alpha_H = alpha l1 ratio number of Components

normalization of X

Mutation Max Norm Sample Max Norm Mutation Sum Norm Sample Sum Norm

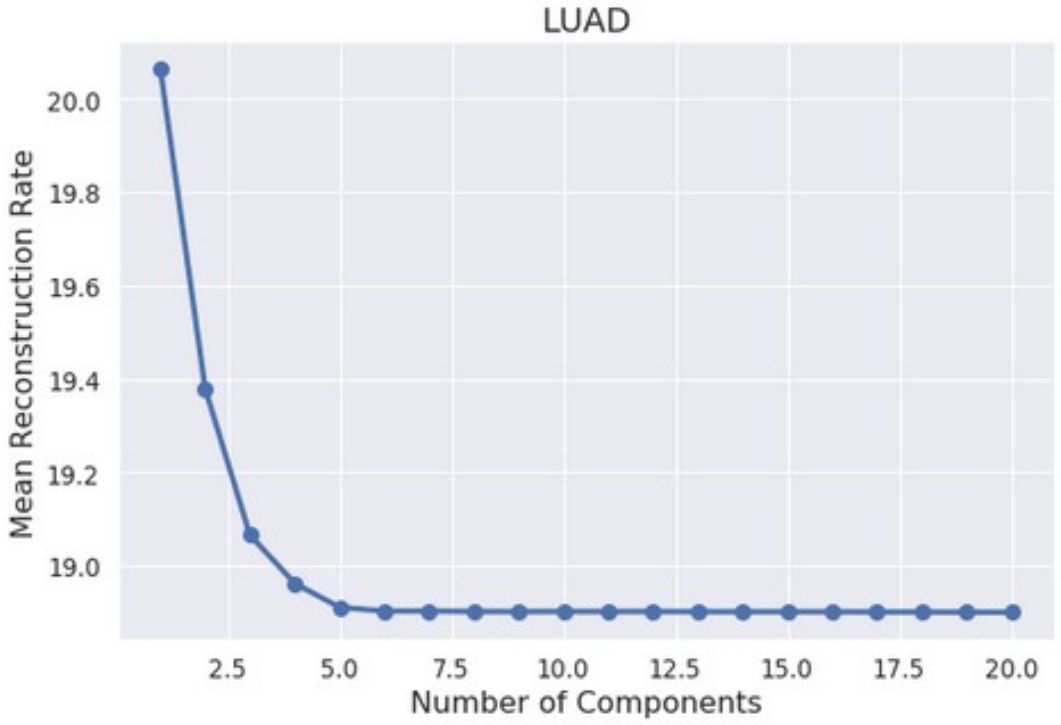
[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

Non-negative matrix factorization

Choosing the best parameters:

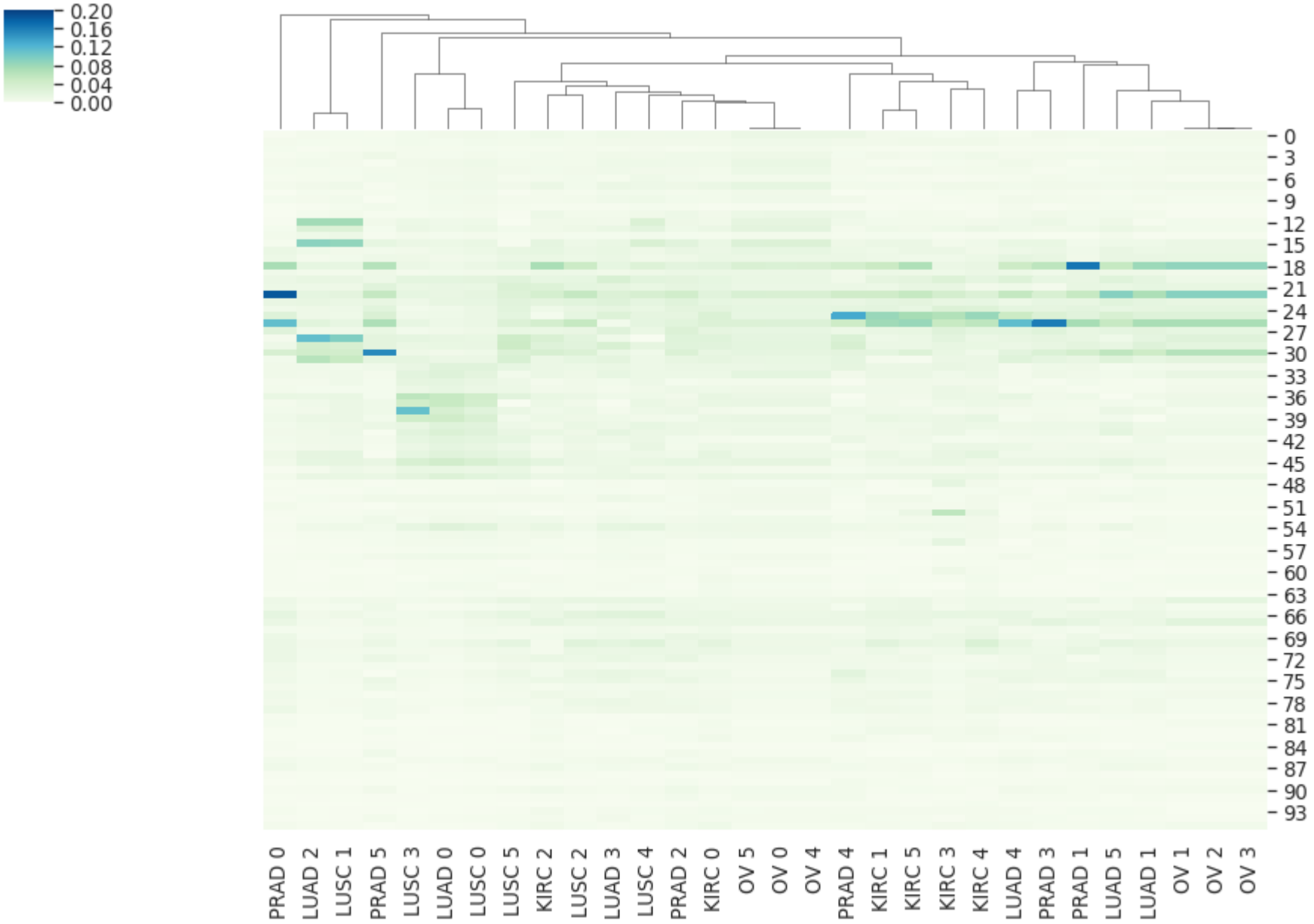
Reconstruction error is calculated

(Frobenius norm of the matrix difference between the training data X and the reconstructed data WH from the fitted model)

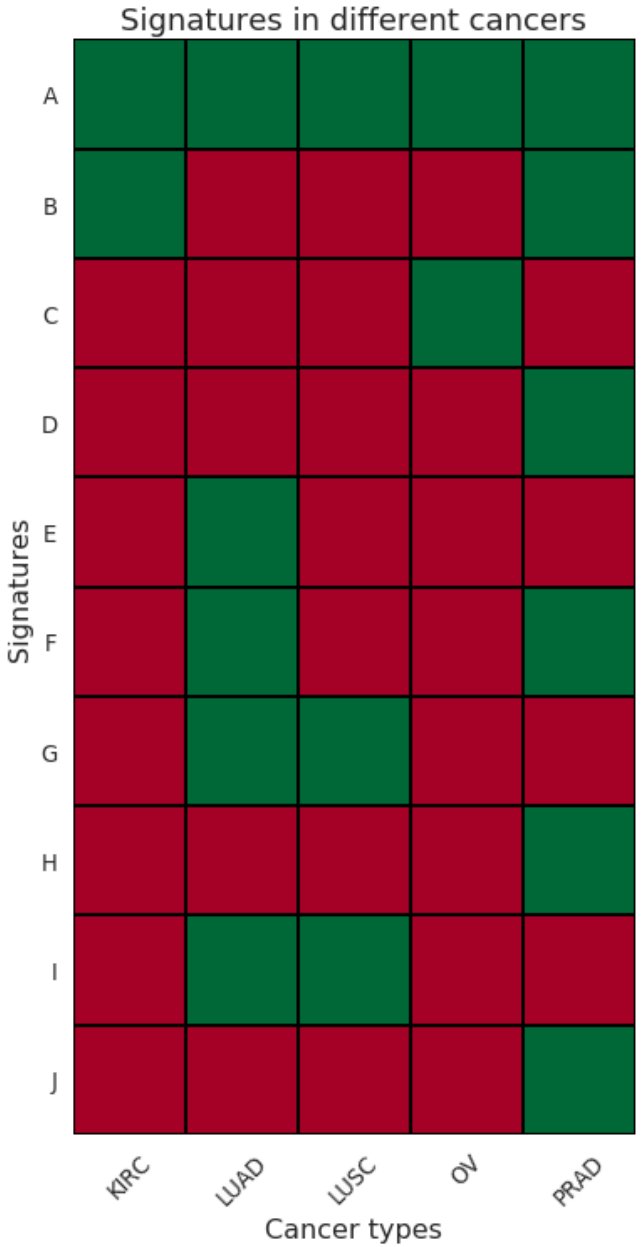


6 number of components are kept ⟶ totally 30 signatures

Clustering signatures

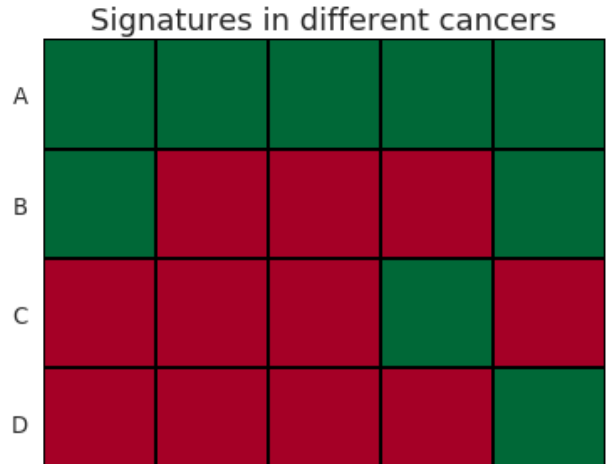


Clustering signatures



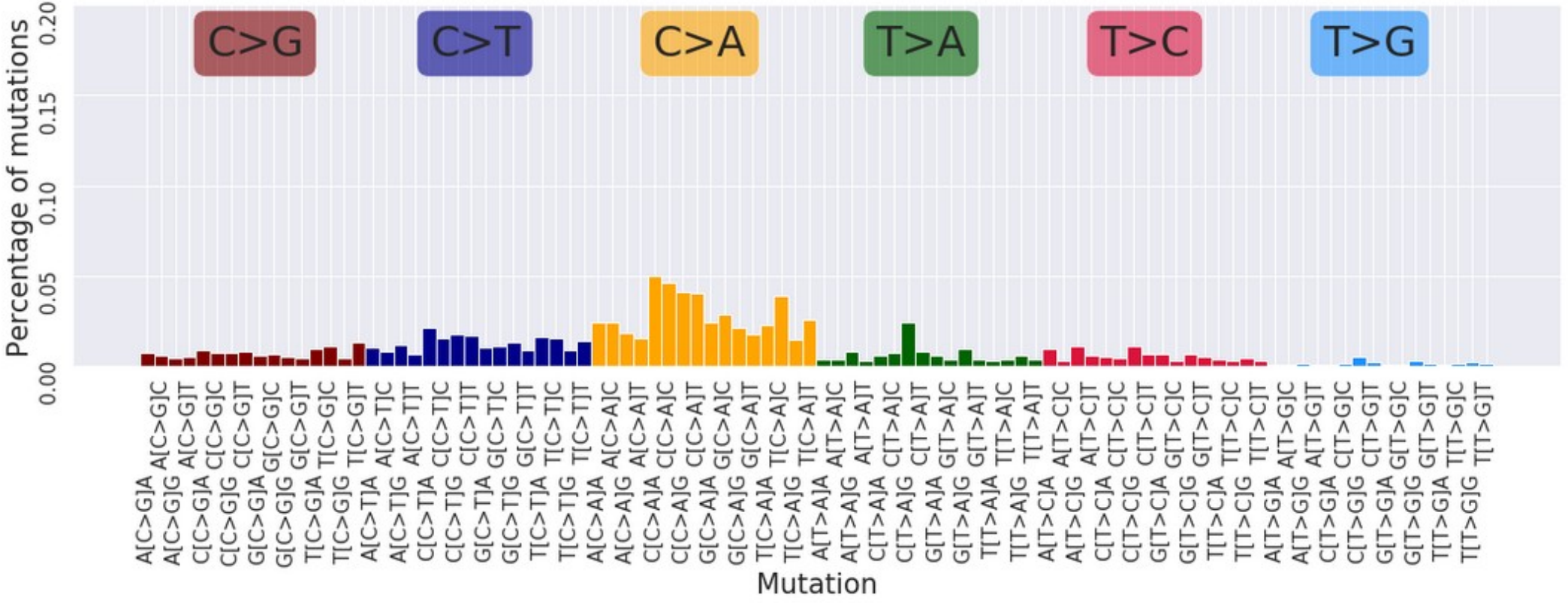
- Hierarchical clustering and K-means give similar result

Clustering signatures



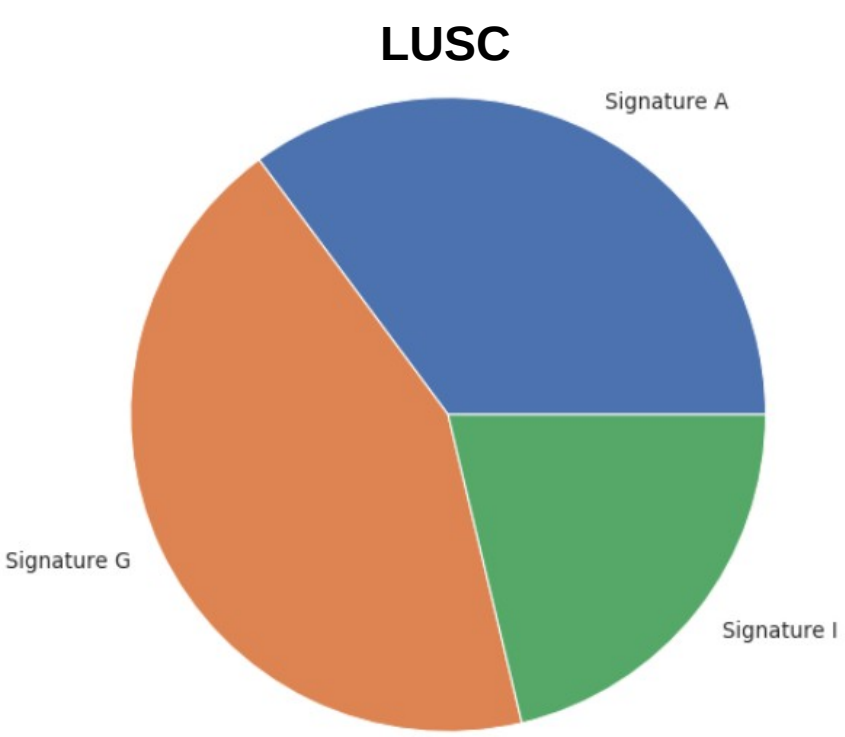
- Hierarchical clustering and K-means give similar result
- **See more figures in the notebook!**

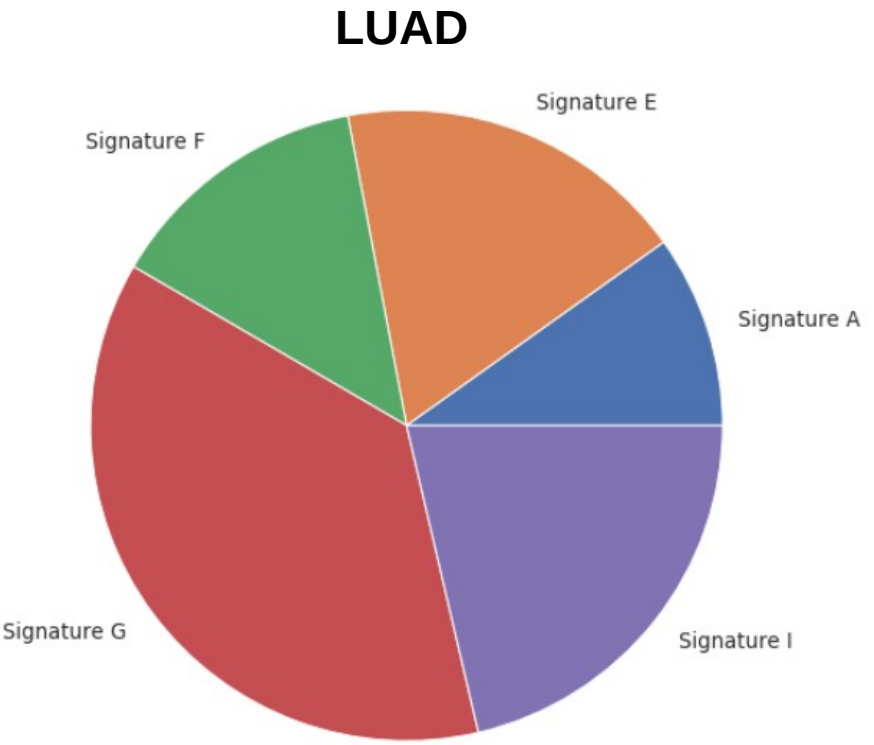
Signature G

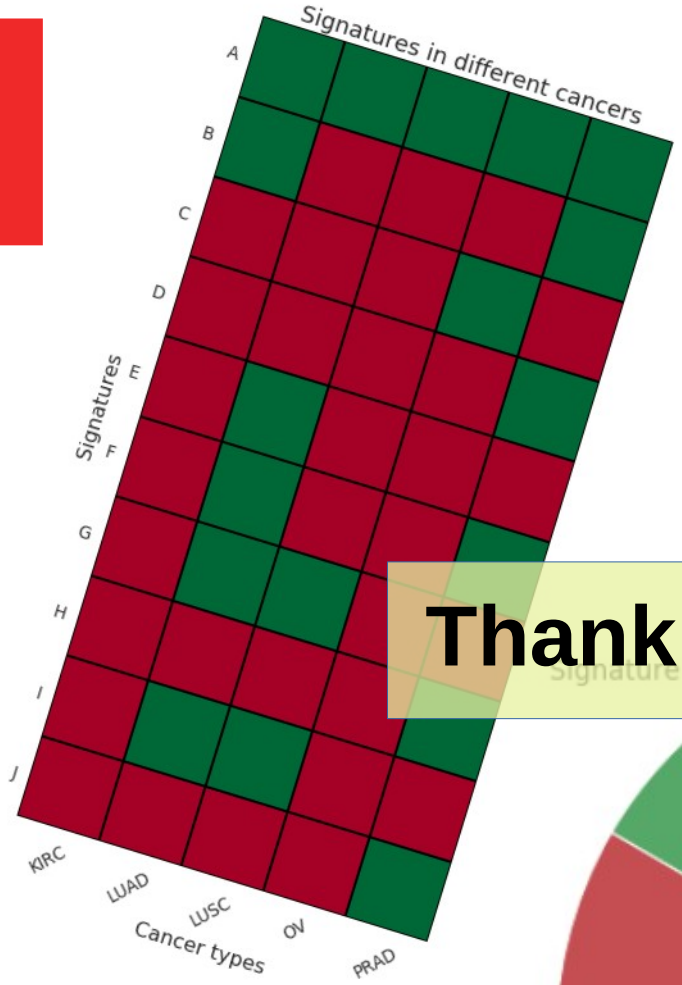


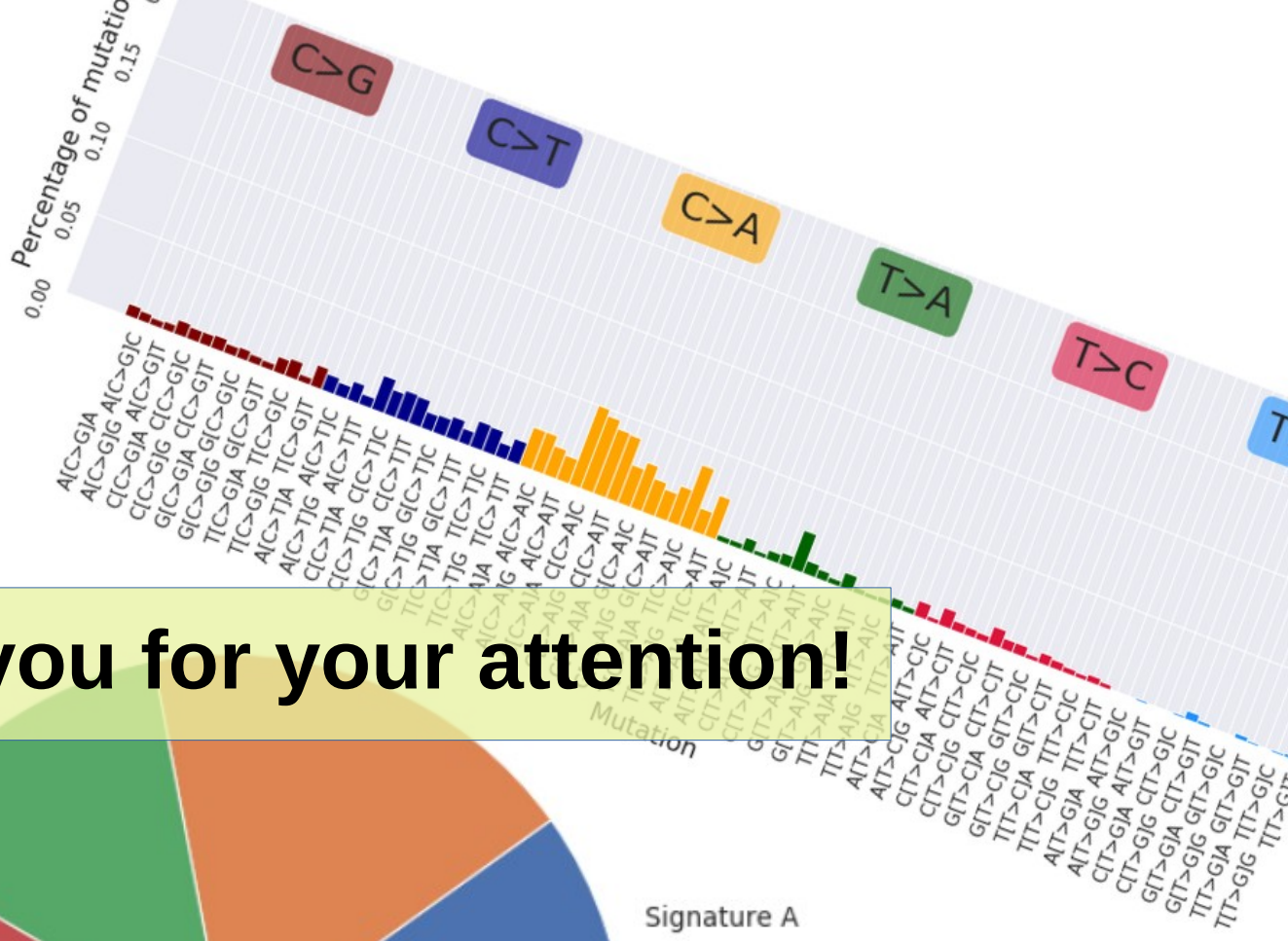
Conclusion, further questions

- Mutational signatures are identified
- How are they related to biological processes?
- How are they presented on different samples?









Thank you for your attention!