



Signatures of mutational processes in human cancer

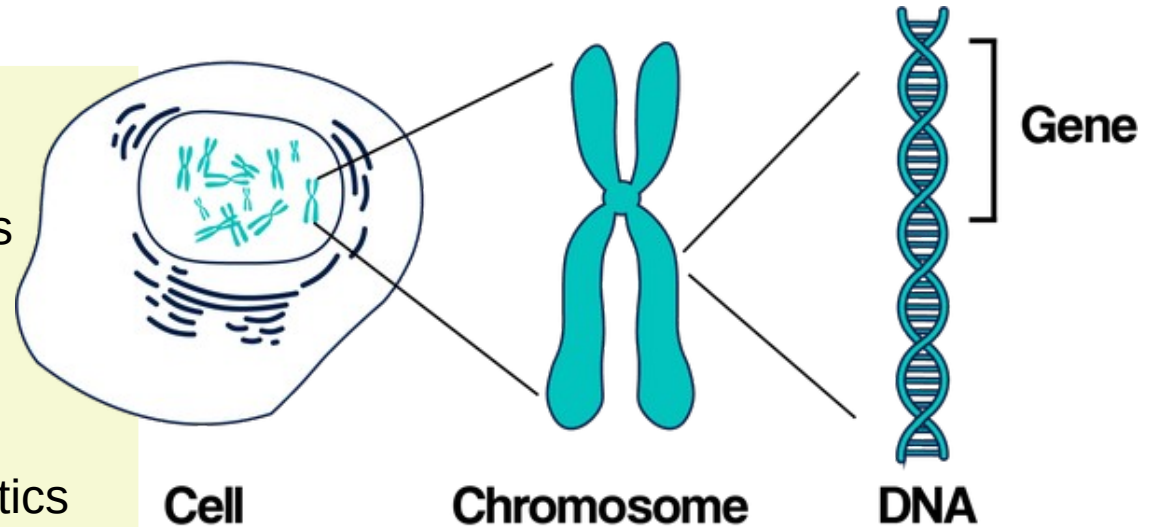
Noel László Plaszkó

Data science laboratory
Supervisor: Orsolya Pipek

Chromosomes, genes and mutations (for dummies...)

Chromosomes

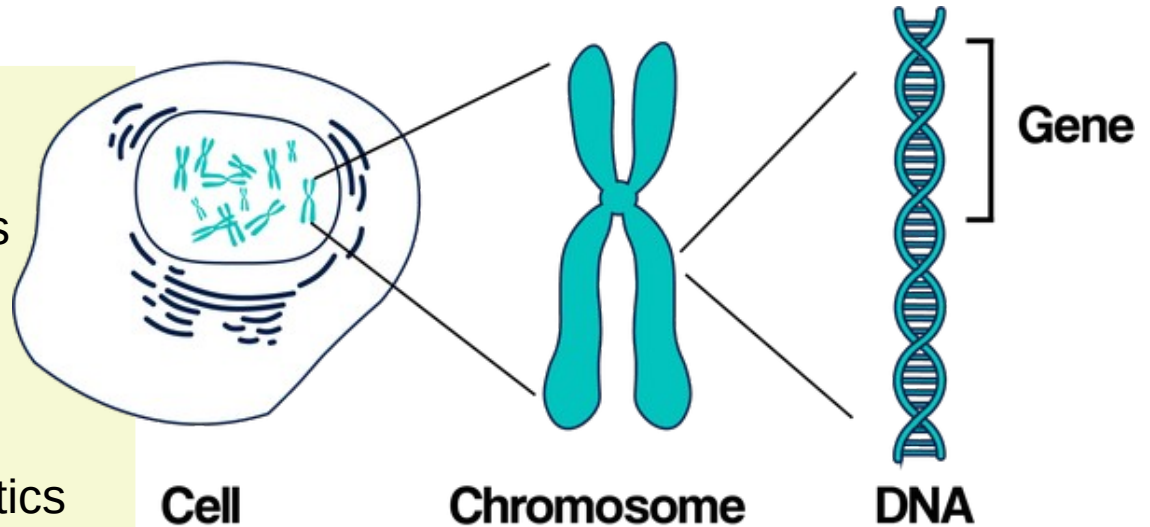
- thread-like structures in cells
 - 46 chromosomes (arranged in 2 sets of 23)
 - one set from your mother, one from your father
- determine physical characteristics
 - contain genes



Chromosomes, genes and mutations (for dummies...)

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



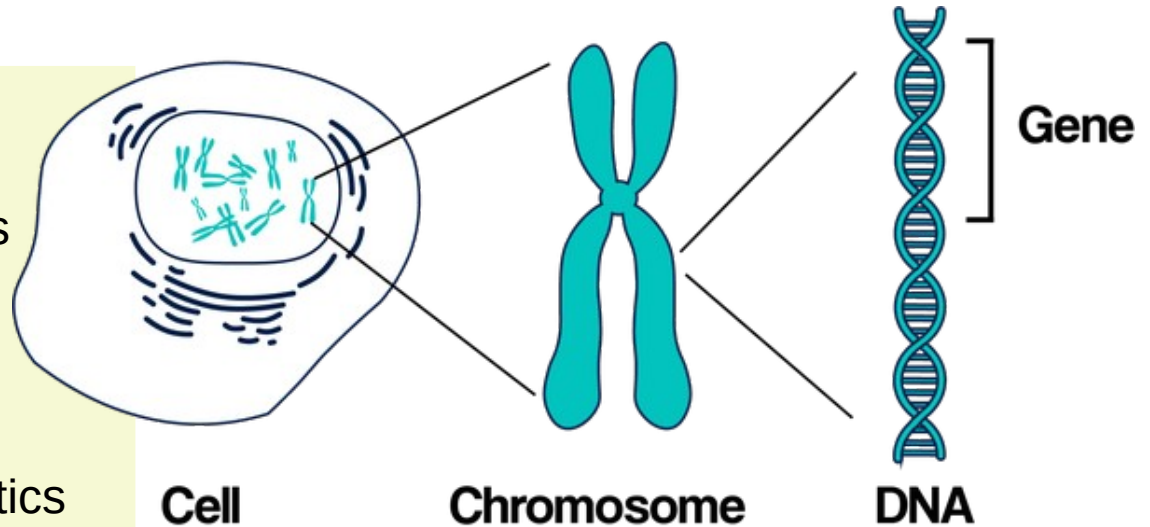
Genes

- 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

Chromosomes, genes and mutations (for dummies...)

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

Proteins

- act as messengers for the cell

Chromosomes, genes and mutations (for dummies...)

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

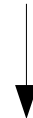


Chromosomes, genes and mutations (for dummies...)

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



Abnormal protein provides different information



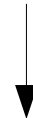
This can cause cells to multiply uncontrollably and become cancerous

Chromosomes, genes and mutations (for dummies...)

- 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.**

DNA sequencing and mutations

DNA sequencing:

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

gctagctataggaccgactatcgaaatgctttcctggcgccgacgcccgcctctggacgtct
atatcgtctgttaaccaagtcggagccgctcagtcataagtaaaacgctacgtttgcttc
tggtcaatgaccagtcgaatttcttatgcagcgaatacctttcgtccctgcgaggaactg
gtccagctacaggtacgcggactagtgtatcacgcgacgtgcagcccccagcgcgtatcgg
ccctcgtcgtgcttctgaatcggtttttgatcgggcgctgcagccatgatacctttgtac
ccggcgcggaatccaatgctctagcaagaggggtgccgatgtcacccggttgcaattatcgt
gcggaatgaacaaaacagtgcaacggcagcgcgttcgtgaggtgacgtcgcggcagctc
tggcaagtgcgtctaaataaagtgctctacatattactattgggcccgcgacccg
ggggaagggttgagaaacgaactcggaggggttaggtacgtgaatggcttgaggtgaggt
gttcaagaggtgtaatatcctgttcagcgaactccagccctcatagagtcgtcaccagtac
gctcataaacaccatcgctaatacctgcctgacaccgaggggttttgcaggtatgcgg
gccaattaacaccacggggataattaacgaacaaatccgtgttatctcaagtaggaacc
aacgtctcggctcgtacacgctatcgagagactcgaatgtaaaccgatacacgaaaagg
actaatagctctactgcctcgttggggtggagggcatgcgggcctagaggggttcgcacggag
gcaacgagcgttgccctctcgcgatcggtcgaaccaaactagggaggtcgaactaagcctg
actcccgccggaattcgactccattccattaggaacatgaagagcgcgactaaattta

DNA sequencing and mutations

DNA sequencing:

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

gctagctataggaccgactatcgaaatgctttcctggcgacgcgcgctctggacgtct
atatcgtctgttaaccaagtcggagccgctcagtcataagtaaaacgctacgtttgcttc
tggtcaatgaccagtcgaatttcttatgcagcgaatacctttcgtccctgcgaggaactg
gtccagctacaggtacgcggactagtgtatcacgcgacgtgcagcccccagcgcgtatcgg
ccctcgtcgtgcttctgaatcgggttttggattgcgggcgctgcagcccccagcgcgtatcgg
ccggcgcgaatccaatgctctagcaagaggggtgccggtgcagcccccagcgcgtatcgg
gcggaatgaacaaacagtcgaacggcagcgcgttcgtgaggtgcagcccccagcgcgtatcgg
tggcaagtcgttaatatgctctaaataaagtgctctacatattactttgggcccgcgaccg
ggggaaggtttgagaaacgaactcggaggggttaggtacgtgaatggcttgaggtgaggt
gttcaagaggtgtaataacgtgttcagcgaactccagccctcatagagtcgtcaccagttac
gcgtcataaacaccatcgtataatccctgcctgacaccgagggaggttttggaggtatgcgg
gccaattaacaccacgggataattaacgaacaaatccgtgttatctcaagtaggaacc
aacgtctcggctcgtacacgctatcgagagactcgaatgtaaaccgatacacgaaaagg
actaatagtcactgcctgttggggtggagggatgcgggcctagaggggttgcgcacggag
gcaacgagcgttgccctctcgcgatcggtcgaacaaactagggaggtcgaactaagcctg
actcccgccggaattcgactccattccattaggaacatgaagagcgcgactaaattta

DNA mutation:



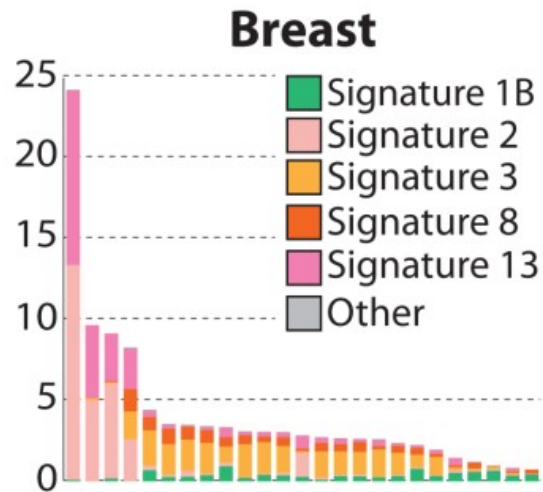


Aim of the project

Investigation of somatic mutations (substitutions)
(somatic: occur from damage to genes in a particular cell)

Aim of the project

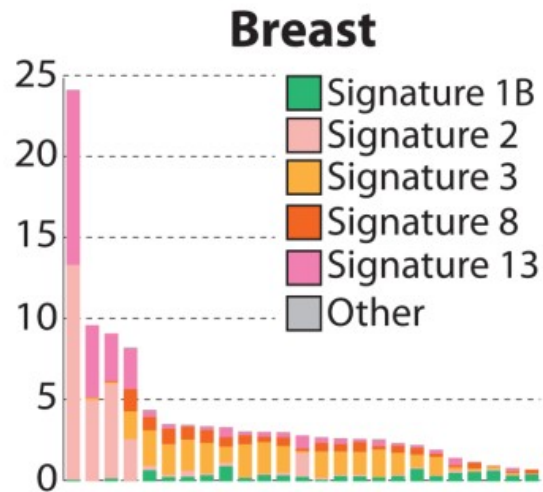
- Investigation of somatic mutations (substitutions)
(somatic: occur from damage to genes in a particular cell)



Decomposing the mutational spectra of each sample to a weighted combination of the operating mutational processes (signatures) in a specific disease

Aim of the project

- Investigation of somatic mutations (substitutions)
(somatic: occur from damage to genes in a particular cell)



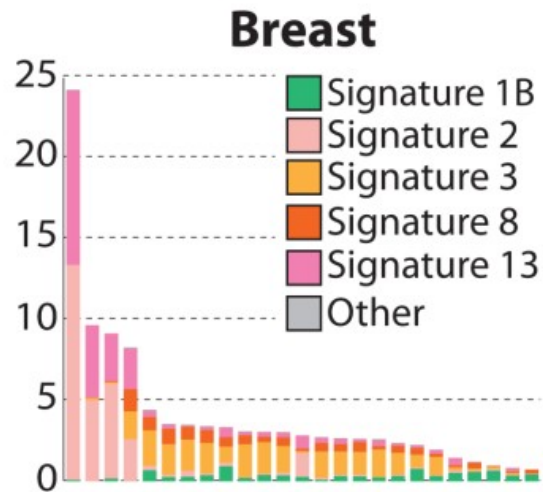
Decomposing the mutational spectra of each sample to a weighted combination of the operating mutational processes (signatures) in a specific disease

Clustering of these signatures across disease types

		ALL	AML	Bladder	Breast	Cervix
		2	2	4	5	2
Signature 1A	7					
Signature 1B	19					
Signature 2	16					
Signature 3	3					
Signature 4	5					
Signature 5	9					

Aim of the project

- Investigation of somatic mutations (substitutions)
(somatic: occur from damage to genes in a particular cell)



Decomposing the mutational spectra of each sample to a weighted combination of the operating mutational processes (signatures) in a specific disease

Clustering of these signatures across disease types

General task:

Perform the analysis described in [4] to identify "mutational signatures" in sample sets of different cancer types

		ALL	AML	Bladder	Breast	Cervix
		2	2	4	5	2
Signature 1A	7					
Signature 1B	19					
Signature 2	16					
Signature 3	3					
Signature 4	5					
Signature 5	9					

Datasets

1) Lists of somatic mutations for different cancer types

- LUAD (lung adenocarcinoma)
- LUSC (lung squamous cell carcinoma)
- KIRC (kidney renal clear cell carcinoma)
- OV (ovarian cancer)
- PRAD (prostate adenocarcinoma)

Task: understand mutational catalogs, Mutation Annotation Format (MAF) data structure

Source: https://github.com/sdam-elte/dslab2021/tree/main/projects/07-mutational_signatures_in_cancer

Datasets

1) Lists of somatic mutations for different cancer types

- LUAD (lung adenocarcinoma)
- LUSC (lung squamous cell carcinoma)
- KIRC (kidney renal clear cell carcinoma)
- OV (ovarian cancer)
- PRAD (prostate adenocarcinoma)

Task: understand mutational catalogs, Mutation Annotation Format (MAF) data structure

Source: https://github.com/sdam-elte/dslab2021/tree/main/projects/07-mutational_signatures_in_cancer

2) Reference genome for each chromosome separately

Task: download data, understand FASTA file format (representing either nucleotide sequences or amino acid sequences)

Source: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/>

Datasets

1) Lists of somatic mutations for different cancer types

- LUAD (lung adenocarcinoma)
- LUSC (lung squamous cell carcinoma)
- KIRC (kidney renal clear cell carcinoma)
- OV (ovarian cancer)
- PRAD (prostate adenocarcinoma)

Task: understand mutational catalogs, Mutation Annotation Format (MAF) data structure

Source: https://github.com/sdam-elte/dslab2021/tree/main/projects/07-mutational_signatures_in_cancer

2) Reference genome for each chromosome separately

Task: download data, understand FASTA file format (representing either nucleotide sequences or amino acid sequences)

Source: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/>

3) 96 mutational categories based on the reference [4]

Datasets

1) Lists of somatic mutations for different cancer types

- LUAD (lung adenocarcinoma)
- LUSC (lung squamous cell carcinoma)
- KIRC (kidney renal clear cell carcinoma)
- OV (ovarian cancer)
- PRAD (prostate adenocarcinoma)

Task: understand mutational catalogs, Mutation Annotation Format (MAF) data structure

Source: https://github.com/sdam-elte/dslab2021/tree/main/projects/07-mutational_signatures_in_cancer

2) Reference genome for each chromosome separately

Task: download data, understand FASTA file format (representing either nucleotide sequences or amino acid sequences)

Source: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/>

3) 96 mutational categories based on the reference [4]

Data cleaning:

- handle headers
- lower case vs upper case letters
- filter the mutations: insertions and deletions are ignored

Datasets

1) Lists of somatic mutations for different cancer types

- LUAD (lung adenocarcinoma)
- LUSC (lung squamous cell carcinoma)
- KIRC (kidney renal clear cell carcinoma)
- OV (ovarian cancer)
- PRAD (prostate adenocarcinoma)

Task: understand mutational catalogs, Mutation Annotation Format (MAF) data structure

Source: https://github.com/sdam-elte/dslab2021/tree/main/projects/07-mutational_signatures_in_cancer

2) Reference genome for each chromosome separately

Task: download data, understand FASTA file format (representing either nucleotide sequences or amino acid sequences)

Source: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/>

3) 96 mutational categories based on the reference [4]

Data cleaning:

- handle headers
- lower case vs upper case letters
- filter the mutations: insertions and deletions are ignored

Data storage:

- Not large (<300MB)
- Huge amount of rows
- Huge amount of character sequences

Datasets

1) Lists of somatic mutations for different cancer types

- LUAD (lung adenocarcinoma)
- LUSC (lung squamous cell carcinoma)
- KIRC (kidney renal clear cell carcinoma)
- OV (ovarian cancer)
- PRAD (prostate adenocarcinoma)

Task: understand mutational catalogs, Mutation Annotation Format (MAF) data structure

Source: https://github.com/sdam-elte/dslab2021/tree/main/projects/07-mutational_signatures_in_cancer

2) Reference genome for each chromosome separately

Task: download data, understand FASTA file format (representing either nucleotide sequences or amino acid sequences)

Source: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/>

3) 96 mutational categories based on the reference [4]

Data cleaning:

- handle headers

- low

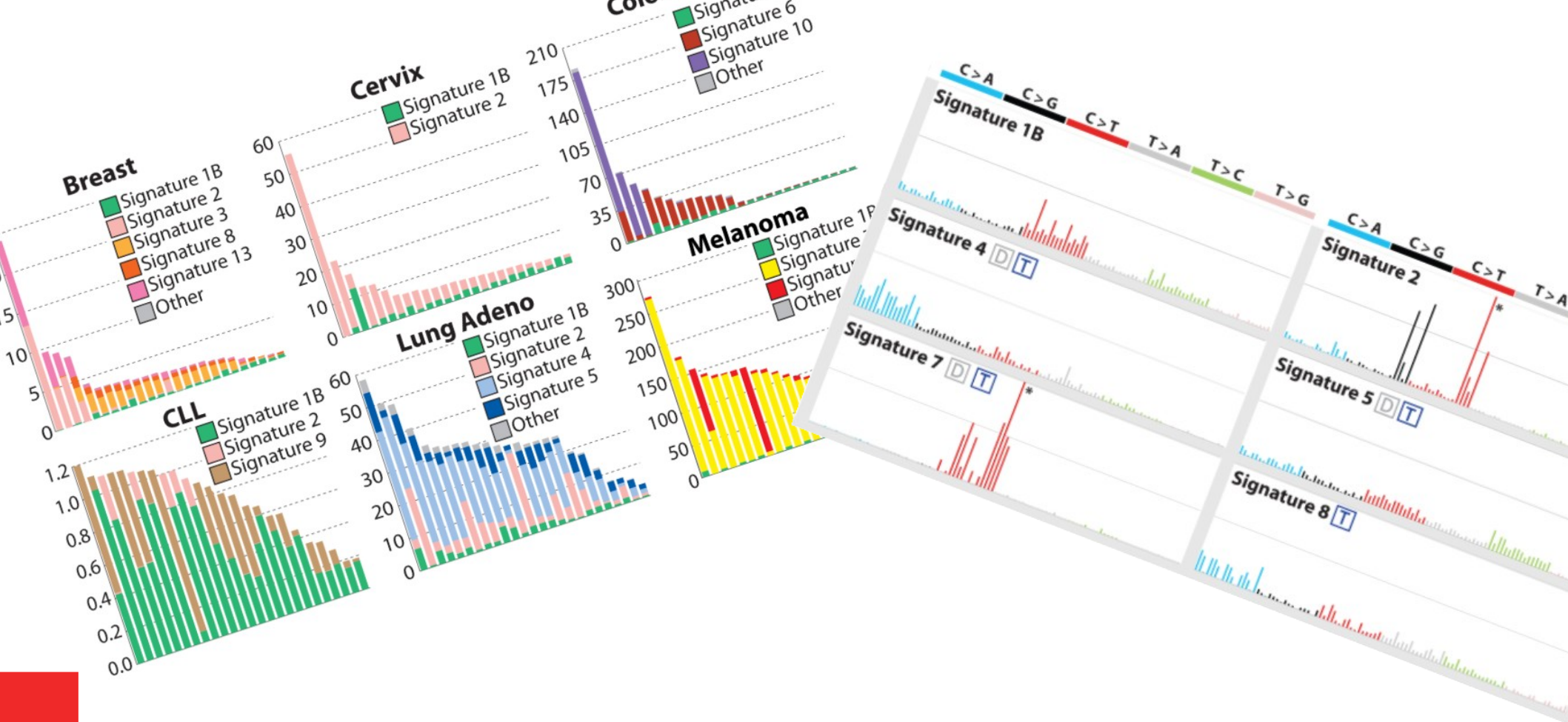
- filter

- delete

Tools:

Python, Pandas, sklearn (non-negative matrix factorization),
Matplotlib/Seaborn/Bokeh/Plotly, numPy

amount of character
sequences



Thanks for the attention