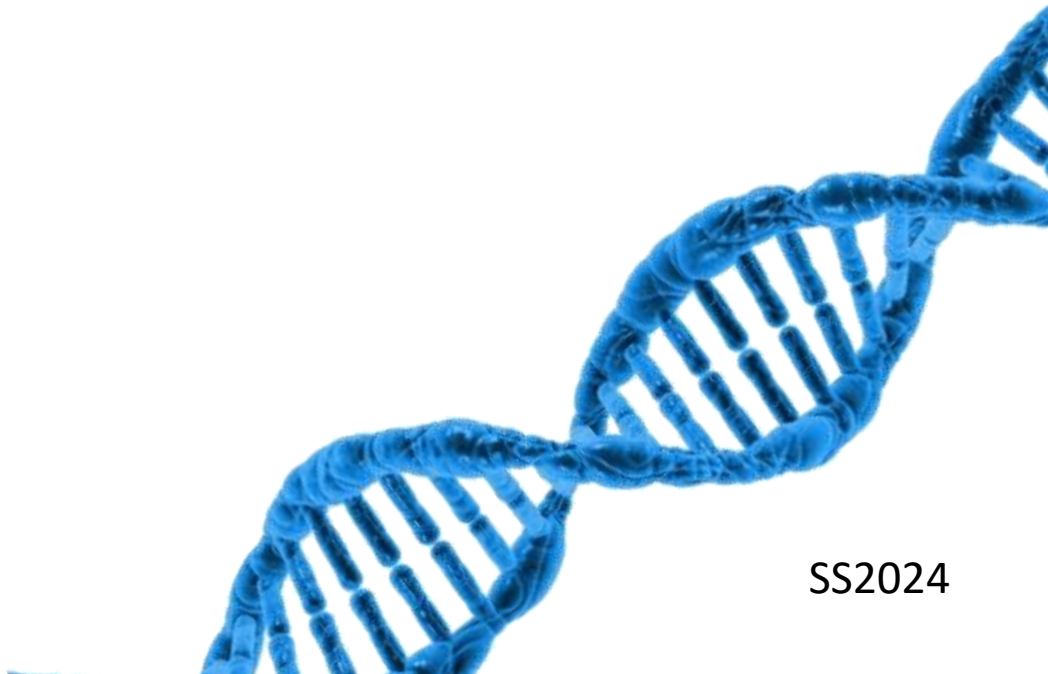
Algorithms and Tools in Bioinformatics

Data, Tools and Technologies in Bioinformatics

Julia Vetter

julia.vetter@fh-hagenberg.at



SS2024

Course Content

(1) Overview
(2) Standard Datasets/Modern File Formats
(3) Databases/Platforms
(4) Data (Pre-) Processing
(5) Tools
(6) Machine Learning

Recap: *Alu* sequence („Real-Life“ Example I) I

- Are the primer suitable to extract the *Alu* sequence from the PLAT gene?
- Primer: **GTGAAAAGCAAGGTCTACCAG** and **GACACCGAGTTCATCTTGAC**

```
# 1. go to: https://www.omim.org/ and search for PLAT
# 2. choose "Plasminogen Activator, Tissue; PLAT"
# 3. click on "DNA" and navigate to "Ensembl (MANE Select)"
# 4. click on "Download sequence" > "Preview" > search (CTRL+F) for
     "Intron 8" > copy the sequence to file (or as string to python
     script)
# 5. close this window and click on "Show transcript table" > click on
     "NM_000930.5" > you can see all genetic relevant information about
     the PLAT gene
# 6. search next to "Nucleotide" for "PLAT Alu sequence" in NCBI
     Nucleotides > select first entry (GenBank: K03021.1)
# 7. download FASTA sequence (use "Send to" + "File" + "FASTA" + "Create
     File") > save as "PLATwithALUsequence.fasta"
# 8. process raw sequences and extract sequence between primers:
     GTGAAAAGCAAGGTCTACCAG and GACACCGAGTTCATCTTGAC
     Hint: remember DNA strands - you won't find the second primer if you
           don't use the reverse complementary version: GTCAAGATGAACTCGGTGTC
# 9. go to https://dotlet.vital-it.ch/ > add both sequences > take a
     screenshot
# 10. go to https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle > enter
      the two sequences > submit
# 11. Note the following parameters: Number of Gaps,  Alignment Score,
      Start position of the Gap
# 12. (optional) search for ORFs (using ORFfinder:
      https://www.ncbi.nlm.nih.gov/orffinder/) > insert
      sequence with Alu > how many ORFs are found?
# 13. (optional) go to https://alfred.med.yale.edu/ and enter PLAT >
      choose "Plasminogen activator, tissue" > select "TPA25 Alu
      insertion" > click on "Frequency Display Formats: Graph" >
      learn something about human migration ;)
```

Mentimeter: Quiz (Alignment)

EXAMPLE I: PLAT

Databases

Name	Type	Usage for our example
OMIM	Secondary Database Bibliographic Database	= Online Mendelian Inheritance in Man Institute: NCBI Information acquisition PLAT gene
Ensembl	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
NCBI Nucleotides	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
ALFRED	Secondary Database Allele Frequency Database	Institute: Yale Center for Medical Informatics Population genetics information

Tools

Name	Type	Usage for our example
In-Silico PCR	PCR simulation	Institute: UCSC Genomics Institute Get primer and sequence information
Dotlet JS *beta*	Sequence alignment	Institute: SIB Vital-IT Sequence comparison/alignment
ORFfinder	Open reading frame finder	Institute: NCBI Identification of open reading frames

OMIM

https://omim.org/

About | Statistics | Downloads | Contact Us | MIMmatch | Donate | Help

Search OMIM... | Options | Display: Highlights

*173370
Table of Contents

* 173370

PLASMINOGEN ACTIVATOR, TISSUE; PLAT

Alternative titles; symbols

TPA

HGNC Approved Gene Symbol: *PLAT*

Cytogenetic location: *8p11.21* ***Genomic coordinates (GRCh38):*** *8:42,174,718-42,207,565* (from NCBI)

TEXT

▼ Description

Tissue plasminogen activator (tPA; PLAT; EC 3.4.21.68) is a serine protease. One of the main functions of tPA is to cleave and activate the proenzyme plasminogen to plasmin (PLG; 173350), which in turn is responsible for fibrinolytic activity.

PLAT is synthesized in vascular endothelial cells as a single polypeptide chain that undergoes proteolytic cleavage by plasmin or trypsin (PRSS1; 276000) at a centrally located arginine-isoleucine bond, resulting in a 2-chain disulfide-linked form composed of the N-terminally derived heavy chain and the C-terminal light chain. The light chain contains the active site (Ny et al., 1984).

▶ Cloning and Expression

▶ Gene Structure

▶ Mapping

▶ Gene Function

▼ Population Genetics

Ludwig et al. (1992) described an insertion/deletion polymorphism of a 311-bp Alu sequence in intron 8 of the PLAT gene. In all populations studied, the frequency of each of 2 alleles varied between 0.40 and 0.60. The similar frequencies among different ethnic groups suggested that the insertion or subsequent deletion of this Alu sequence occurred early in human evolution.

▼ External Links

- ▶ Genome
- ▶ DNA
- ▶ Protein
- ▶ Gene Info
- ▶ Clinical Resources
- ▼ Variation
  - ClinVar
  - gnomAD
  - GWAS Catalog
  - GWAS Central
  - HGMD
  - NHLBI EVS
  - PharmgKB
- ▶ Animal Models
- ▶ Cellular Pathways

EXAMPLE I: PLAT

Databases

Name	Type	Usage for our example
OMIM	Secondary Database Bibliographic Database	= Online Mendelian Inheritance in Man Institute: NCBI Information acquisition PLAT gene
Ensembl	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
NCBI Nucleotides	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
ALFRED	Secondary Database Allele Frequency Database	Institute: Yale Center for Medical Informatics Population genetics information

Tools

Name	Type	Usage for our example
In-Silico PCR	PCR simulation	Institute: UCSC Genomics Institute Get primer and sequence information
Dotlet JS *beta*	Sequence alignment	Institute: SIB Vital-IT Sequence comparison/alignment
ORFfinder	Open reading frame finder	Institute: NCBI Identification of open reading frames

Recap: *Alu* sequence II

- Primer: **GTGAAAAGCAAGGTCTACCAG** and **GACACCGAGTTCATCTTGAC**

```
>K03021.1 Human tissue plasminogen activator (PLAT) gene | with Alu Sequence
...GTCTTGGCAGAACGTGGGATTAGGGTGTGAGACGGGGGAAGATCCAATGTCTCAAGTTGCATGACAGACCCAGTGCGTGGG
AAGCACCCATGGATATTATCTAATCCAACCTCTTCACTTGCTAGAATAACACATATTGTGAAAAGCAAGGTCTACCAGTTTTCC
AACCTAAATCCCAAGTTAAGGGTCCTGGCCTGTAACCATTTAGTCCTCAGCTGTTCTCCTGACATCTTTATTGCAATGATTTGT
AAGAGTTCCGTAACAGGACAGCTCACAGTTCTGTCTGACAACCCTATGAGATTAGAACACTACGGCCGGGCGCGGTGGCTCACG
CCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAA
CCCCGTCTCTACTAAAACTACAAAAAATAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCTGGCTACTTGGGAGGCTGAGGCAGG
AGAATGGCATGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACT
CCGTCTCAAAAAAAAAAAAAAAAAAAAAAGAACACTACATTACTGACTGGGTAACAAAGTTAAAGAGAAGTTCTCCTAGGGTGG
GGGTGTGCTGCAAGGTCAAGATGAACTCGGTGTCCTCCCTCCCAGCTCAGTGGTTTTCATTGGTTGACTGAGTCTCCTTCTACT
CTTACATGGCCTGTGATGTGGCTGAAAATGGGATTGAAAATCTTAAACTCCTGGCCTGGTGTGGTGGTGCATGCCTGTAATCCC
AGCACTTTGGGAGGCTGAGGCAGGAGGATTGCTTGAGCCCAGGAGTTCAAGACCAGCCTGGGCAACATGGCAAGACCCC...
```

EXAMPLE I: PLAT

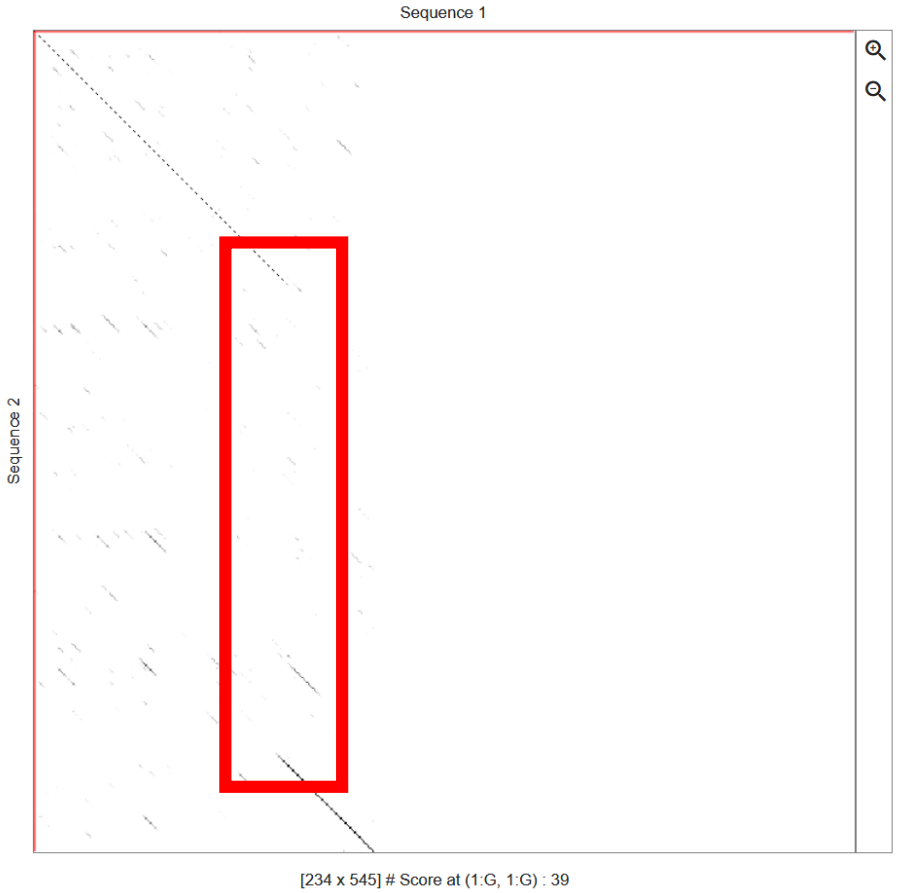
Databases

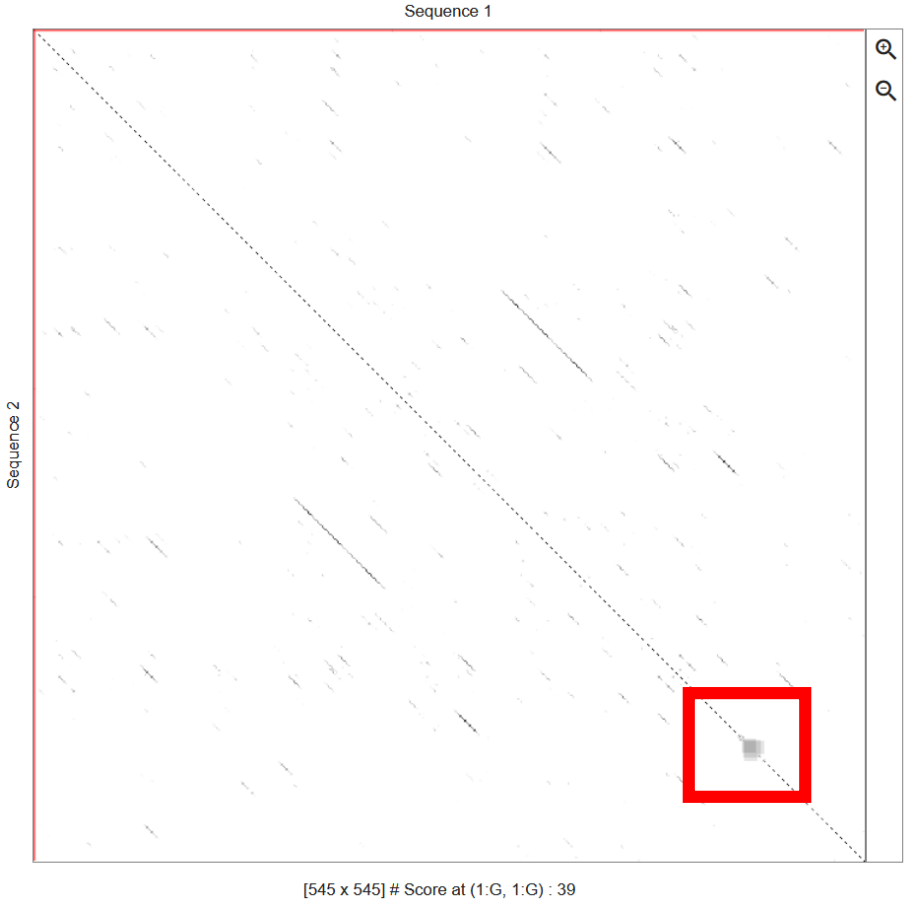
Name	Type	Usage for our example
OMIM	Secondary Database Bibliographic Database	= Online Mendelian Inheritance in Man Institute: NCBI Information acquisition PLAT gene
Ensembl	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
NCBI Nucleotides	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
ALFRED	Secondary Database Allele Frequency Database	Institute: Yale Center for Medical Informatics Population genetics information

Tools

Name	Type	Usage for our example
Dotlet JS *beta*	Sequence alignment	Institute: SIB Vital-IT Sequence comparison/alignment
EMBOSS Needle	Sequence alignment	Institute: EMBL-EBI Sequence alignment using Needleman-Wunsch
ORFfinder	Open reading frame finder	Institute: NCBI Identification of open reading frames

Dotlet https://dotlet.vital-it.ch/





Recap: *Alu* sequence III

- Primer: **GTGAAAAGCAAGGTCTACCAG** and **GACACCGAGTTCATCTTGAC**

```
>K03021.1 Human tissue plasminogen activator (PLAT) gene | with Alu Sequence
...GTCTTGGCAGAACGTGGGATTAGGGTGTGAGACGGGGGAAGATCCAATGTCTCAAGTTGCATGACAGACCCAGTGCGTGGG
AAGCACCCATGGATATTATCTAATCCAACCTCTTCACTTGCTAGAATAACACATATTGTGAAAAGCAAGGTCTACCAGTTTTCC
AACCTAAATCCCAAGTTAAGGGTCCTGGCCTGTAACCATTTAGTCCTCAGCTGTTCTCCTGACATCTTTATTGCAATGATTTGT
AAGAGTTCCGTAACAGGACAGCTCACAGTTCTGTCTGACAACCCTATGAGATTAGAACACTACGGCCGGGCGCGGTGGCTCACG
CCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAAA
CCCCGTCTCTACTAAAACTACAAAAAATAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCTGGCTACTTGGGAGGCTGAGGCAGG
AGAATGGCATGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACT
CCGTCTCAAAAAAAAAAAAAAAAAAAAAAGAACACTACATTACTGACTGGGTAACAAAGTTAAAGAGAAGTTCTCCTAGGGTGG
GGGTGTGCTGCAAGGTCAAGATGAACTCGGTGTCCTCCCTCCCAGCTCAGTGGTTTTCATTGGTTGACTGAGTCTCCTTCTACT
CTTACATGGCCTGTGATGTGGCTGAAAATGGGATTGAAAATCTTAAACTCCTGGCCTGGTGTGGTGGTGCATGCCTGTAATCCC
AGCACTTTGGGAGGCTGAGGCAGGAGGATTGCTTGAGCCCAGGAGTTCAAGACCAGCCTGGGCAACATGGCAAGACCCC...
```

EXAMPLE I: PLAT

Databases

Name	Type	Usage for our example
OMIM	Secondary Database Bibliographic Database	= Online Mendelian Inheritance in Man Institute: NCBI Information acquisition PLAT gene
Ensembl	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
NCBI Nucleotides	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
ALFRED	Secondary Database Allele Frequency Database	Institute: Yale Center for Medical Informatics Population genetics information

Tools

Name	Type	Usage for our example
Dotlet JS *beta*	Sequence alignment	Institute: SIB Vital-IT Sequence comparison/alignment
EMBOSS Needle	Sequence alignment	Institute: EMBL-EBI Sequence alignment using Needleman-Wunsch
ORFfinder	Open reading frame finder	Institute: NCBI Identification of open reading frames

EMBOSS Needle

https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle

```
withALU          1 GTGAAAAGCAAGGTCTACCAGTTTTCCAACCTAAATCCCAAGTTAAGGGT     50
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
withoutALU       1 GTGAAAAGCAAGGTCTACCAGTTTTCCAACCTAAATCCCAAGTTAAGGGT     50

withALU         51 CCTGGCCTGTAACCATTTAGTCCTCAGCTGTTCTCCTGACATCTTTATTG    100
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
withoutALU      51 CCTGGCCTGTAACCATTTAGTCCTCAGCTGTTCTCCTGACATCTTTATTG    100

withALU        101 CAATGATTTGTAAGAGTTCCGTAACAGGACAGCTCACAGTTCTGTCTGAC    150
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
withoutALU     101 CAATGATTTGTAAGAGTTCCGTAACAGGACAGCTCACAGTTCTGTCTGAC    150

withALU        151 AACCCTATGAGATTAGAACACTACGGCCGGGCGCGGTGGCTCACGCCTGT    200
                   ||||||||||||||
withoutALU     151 AACCCTATGAGATT----------------------------------    164

withALU        201 AATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGA    250

withoutALU     165 --------------------------------------------------    164

withALU        251 TCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACTAAAACTAC    300

withoutALU     165 --------------------------------------------------    164

withALU        301 AAAAAATAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCTGGCTACTTGGG    350

withoutALU     165 --------------------------------------------------    164

withALU        351 AGGCTGAGGCAGGAGAATGGCATGAACCCGGGAGGCGGAGCTTGCAGTGA    400

withoutALU     165 --------------------------------------------------    164

withALU        401 GCCGAGATCCCGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCG    450

withoutALU     165 --------------------------------------------------    164

withALU        451 TCTCAAAAAAAAAAAAAAAAAAAAAAAGAACACTACATTACTGACTGGGTA    500
                                              |||||||||||||||||||||||||
withoutALU     165 -------------------------AGAACACTACATTACTGACTGGGTA    189

withALU        501 ACAAAGTTAAAGAGAAGTTCTCCTAGGGTGGGGGTGTGCTGCAAG    545
                   |||||||||||||||||||||||||||||||||||||||||||
withoutALU     190 ACAAAGTTAAAGAGAAGTTCTCCTAGGGTGGGGGTGTGCTGCAAG    234
```

EMBOSS Needle https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle II

Results for Job ID: emboss_needle-I20240227-181057-

Tool Output

```
#######################################
# Program: needle
# Rundate: Tue 27 Feb 2024 18:11:10
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20240227-181057-0979-632480-p1m.asequence
#    -bsequence emboss_needle-I20240227-181057-0979-632480-p1m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
#######################################
```

```
#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 545
# Identity:     234/545 (42.9%)
# Similarity:   234/545 (42.9%)
# Gaps:         311/545 (57.1%)
# Score: 1005.0
#
#
#=======================================
```

EMBOSS Needle

https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle

```
withALU          1 GTGAAAAGCAAGGTCTACCAGTTTTCCAACCTAAATCCCAAGTTAAGGGT     50
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
withoutALU       1 GTGAAAAGCAAGGTCTACCAGTTTTCCAACCTAAATCCCAAGTTAAGGGT     50

withALU         51 CCTGGCCTGTAACCATTTAGTCCTCAGCTGTTCTCCTGACATCTTTATTG    100
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
withoutALU      51 CCTGGCCTGTAACCATTTAGTCCTCAGCTGTTCTCCTGACATCTTTATTG    100

withALU        101 CAATGATTTGTAAGAGTTCCGTAACAGGACAGCTCACAGTTCTGTCTGAC    150
                   ||||||||||||||||||||||||||||||||||||||||||||||||||
withoutALU     101 CAATGATTTGTAAGAGTTCCGTAACAGGACAGCTCACAGTTCTGTCTGAC    150

withALU        151 AACCCTATGAGATTAGAACACTACGGCCGGGCGCGGTGGCTCACGCCTGT    200
                   ||||||||||||||
withoutALU     151 AACCCTATGAGATT----------------------------------      164

withALU        201 AATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGA    250

withoutALU     165 ------------------------------------------------      164

withALU        251 TCGAGACCATCCCGGCTAAAACGGTGAAACCCCGTCTCTACTAAAACTAC    300

withoutALU     165 ------------------------------------------------      164

withALU        301 AAAAAATAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCTGGCTACTTGGG    350

withoutALU     165 ------------------------------------------------      164

withALU        351 AGGCTGAGGCAGGAGAATGGCATGAACCCGGGAGGCGGAGCTTGCAGTGA    400

withoutALU     165 ------------------------------------------------      164

withALU        401 GCCGAGATCCCGCCACTGCACTCCAGCCTGGGCAACAGAGCGAGACTCCG    450

withoutALU     165 ------------------------------------------------      164

withALU        451 TCTCAAAAAAAAAAAAAAAAAAAAAAAGAACACTACATTACTGACTGGGTA    500
                                              |||||||||||||||||||||||||
withoutALU     165 -------------------------AGAACACTACATTACTGACTGGGTA    189

withALU        501 ACAAAGTTAAAGAGAAGTTCTCCTAGGGTGGGGGTGTGCTGCAAG    545
                   |||||||||||||||||||||||||||||||||||||||||||
withoutALU     190 ACAAAGTTAAAGAGAAGTTCTCCTAGGGTGGGGGTGTGCTGCAAG    234
```

EXAMPLE I: PLAT

Databases

Name	Type	Usage for our example
OMIM	Secondary Database Bibliographic Database	= Online Mendelian Inheritance in Man Institute: NCBI Information acquisition PLAT gene
Ensembl	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
NCBI Nucleotides	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
ALFRED	Secondary Database Allele Frequency Database	Institute: Yale Center for Medical Informatics Population genetics information

Tools

Name	Type	Usage for our example
In-Silico PCR	PCR simulation	Institute: UCSC Genomics Institute Get primer and sequence information
Dotlet JS *beta*	Sequence alignment	Institute: SIB Vital-IT Sequence comparison/alignment
ORFfinder	Open reading frame finder	Institute: NCBI Identification of open reading frames

ORFfinder https://www.ncbi.nlm.nih.gov/orffinder/

National Library of Medicine
National Center for Biotechnology Information

Log in

< ORFfinder submitting page

> Help

Open Reading Frame Viewer

Sequence

ORFs found: 4 Genetic code: 1 Start codon: 'ATG' only

Find: Tools Tracks

(U) ORFfinder_3.5.171138323

ORF1 ORF2 ORF3 ORF4

1: 1..545 (545 nt) Tracks shown: 2/5

Six-frame translation...

ORF1 (63 aa) Display ORF as... Mark

```
>lcl|ORF1
MICKSSVTGQLTVLSDNPMRLEHYGRARWLTPVIPALWEAEAGGSRGQEI
ETIPAKTVKPRLY
```

Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF1	+	1	103	294	192 | 63
ORF2	+	3	372	>545	174 | 57
ORF4	-	2	373	266	108 | 35
ORF3	-	1	92	15	78 | 25

ORF1

SmartBLAST

BLAST

Marked set (0)

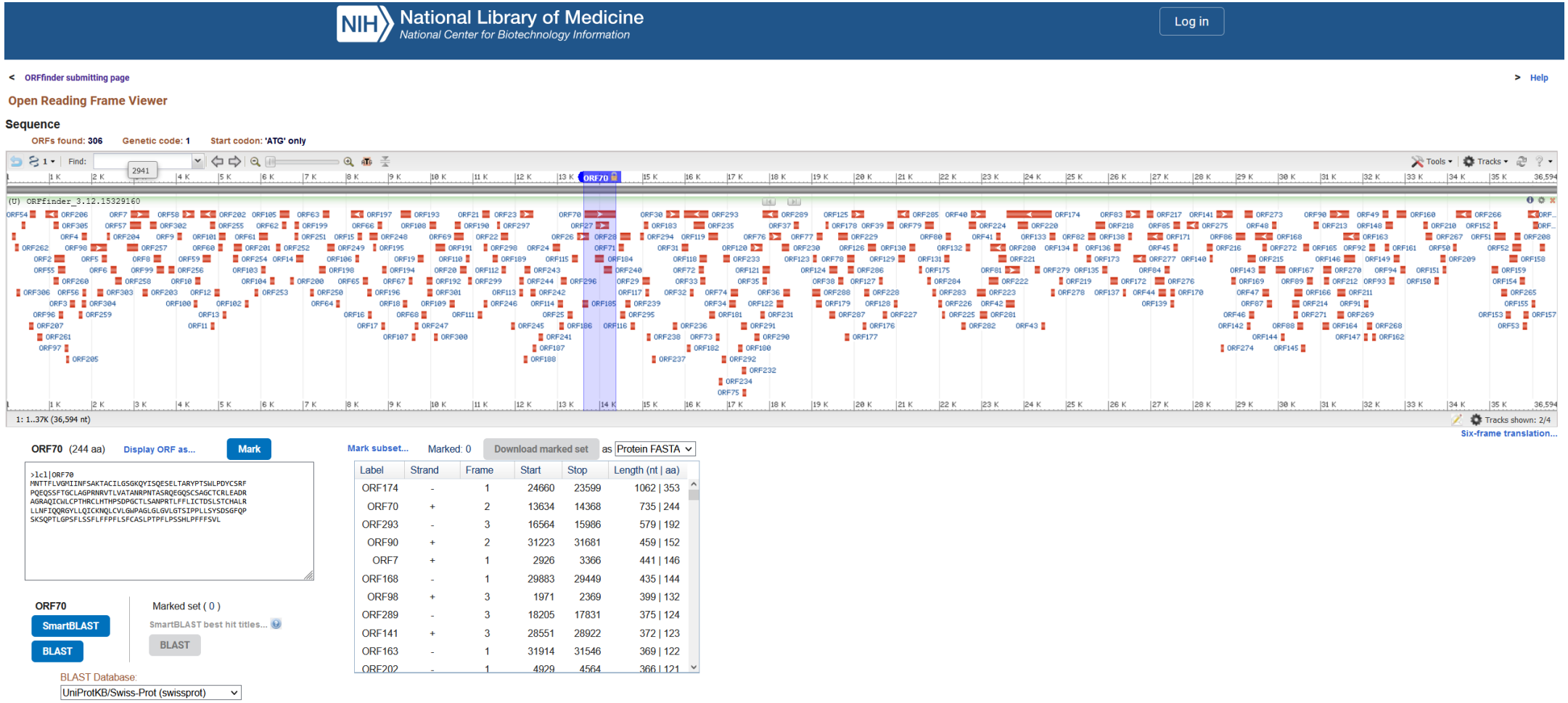
SmartBLAST best hit titles...

BLAST

BLAST Database:

UniProtKB/Swiss-Prot (swissprot)

ORFfinder https://www.ncbi.nlm.nih.gov/orffinder/



Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF174	-	1	24660	23599	1062 | 353
ORF70	+	2	13634	14368	735 | 244
ORF293	-	3	16564	15986	579 | 192
ORF90	+	2	31223	31681	459 | 152
ORF7	+	1	2926	3366	441 | 146
ORF168	-	1	29883	29449	435 | 144
ORF98	+	3	1971	2369	399 | 132
ORF289	-	3	18205	17831	375 | 124
ORF141	+	3	28551	28922	372 | 123
ORF163	-	1	31914	31546	369 | 122
ORF202	-	1	4929	4564	366 | 121

EXAMPLE I: PLAT

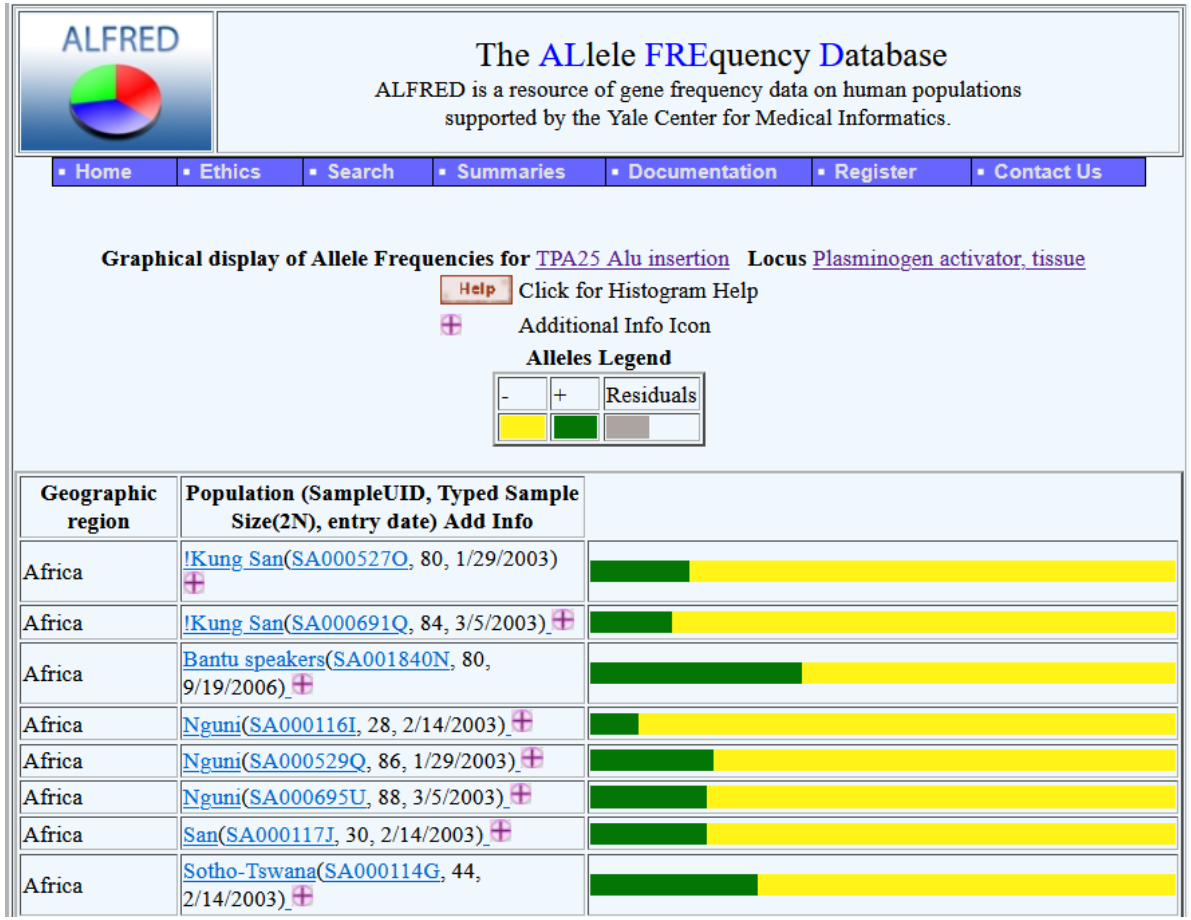
Databases

Name	Type	Usage for our example
OMIM	Secondary Database Bibliographic Database	= Online Mendelian Inheritance in Man Institute: NCBI Information acquisition PLAT gene
Ensembl	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
NCBI Nucleotides	Secondary Database Genome Database	Institute: EMBL (Reference) sequence data acquisition
ALFRED	Secondary Database Allele Frequency Database	Institute: Yale Center for Medical Informatics Population genetics information

Tools

Name	Type	Usage for our example
In-Silico PCR	PCR simulation	Institute: UCSC Genomics Institute Get primer and sequence information
Dotlet JS *beta*	Sequence alignment	Institute: SIB Vital-IT Sequence comparison/alignment
ORFfinder	Open reading frame finder	Institute: NCBI Identification of open reading frames

ALFRED https://alfred.med.yale.edu/alfred/index.asp



ALFRED

The ALlele FREquency Database

ALFRED is a resource of gene frequency data on human populations supported by the Yale Center for Medical Informatics.

Home | Ethics | Search | Summaries | Documentation | Register | Contact Us

Graphical display of Allele Frequencies for TPA25 Alu insertion **Locus** Plasminogen activator, tissue

Help Click for Histogram Help

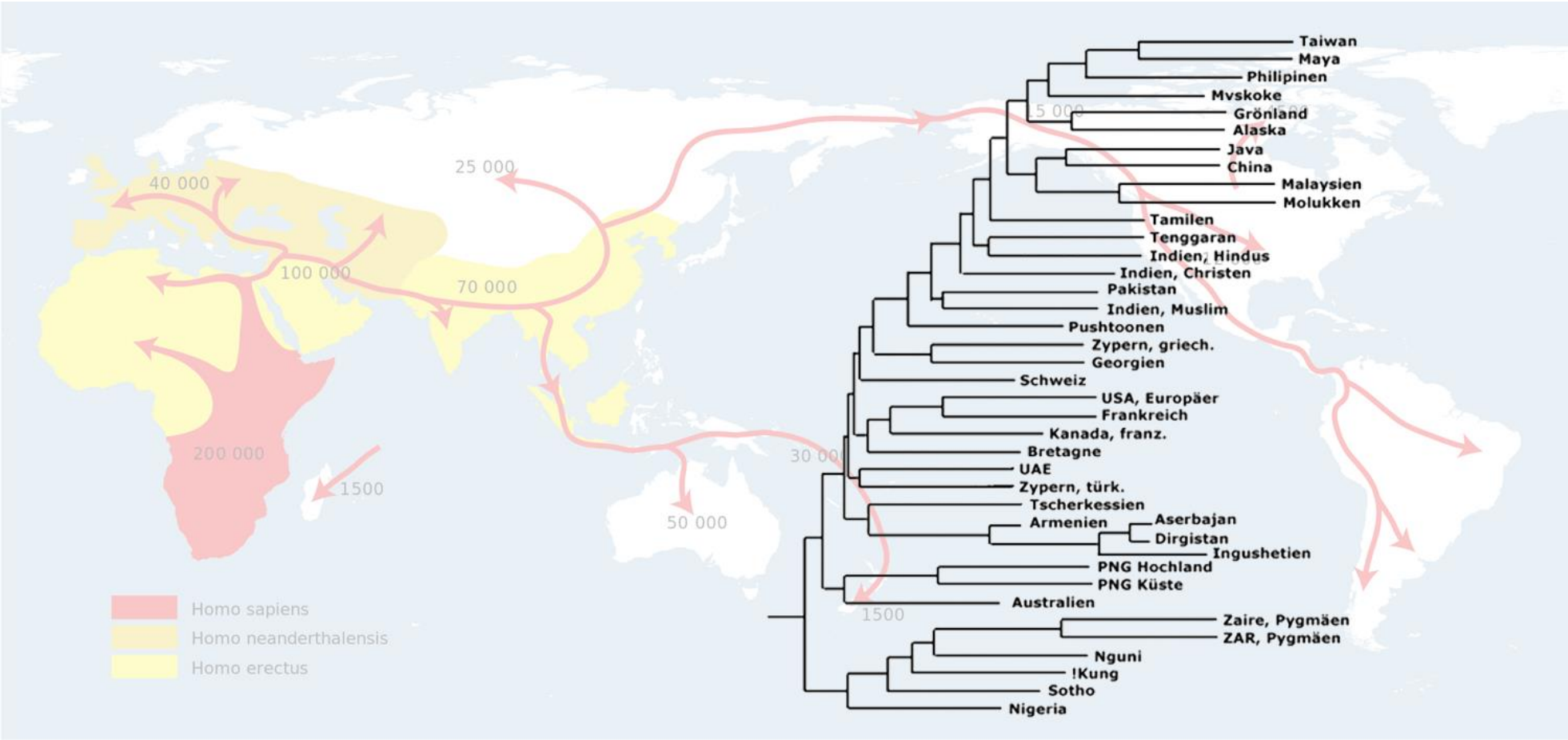
Additional Info Icon

Alleles Legend

-	+	Residuals

Geographic region	Population (SampleUID, Typed Sample Size(2N), entry date) Add Info
Africa	!Kung San(SA000527O, 80, 1/29/2003)
Africa	!Kung San(SA000691Q, 84, 3/5/2003)
Africa	Bantu speakers(SA001840N, 80, 9/19/2006)
Africa	Nguni(SA000116I, 28, 2/14/2003)
Africa	Nguni(SA000529Q, 86, 1/29/2003)
Africa	Nguni(SA000695U, 88, 3/5/2003)
Africa	San(SA000117J, 30, 2/14/2003)
Africa	Sotho-Tswana(SA000114G, 44, 2/14/2003)

ALFRED https://alfred.med.yale.edu/alfred/index.asp

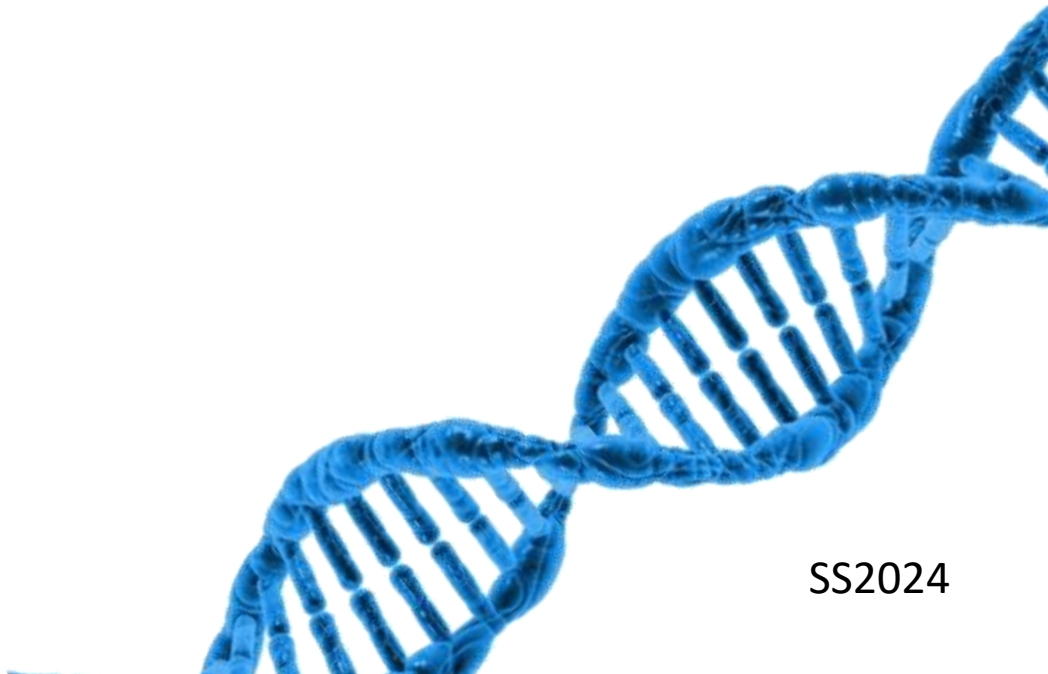


Algorithms and Tools in Bioinformatics

Data, Tools and Technologies in Bioinformatics

Julia Vetter

julia.vetter@fh-hagenberg.at



SS2024

„Real-Life“ Example II

Research Area: Molecular Biology
Research Focus: Chronic Myeloid Leukemia (CML)

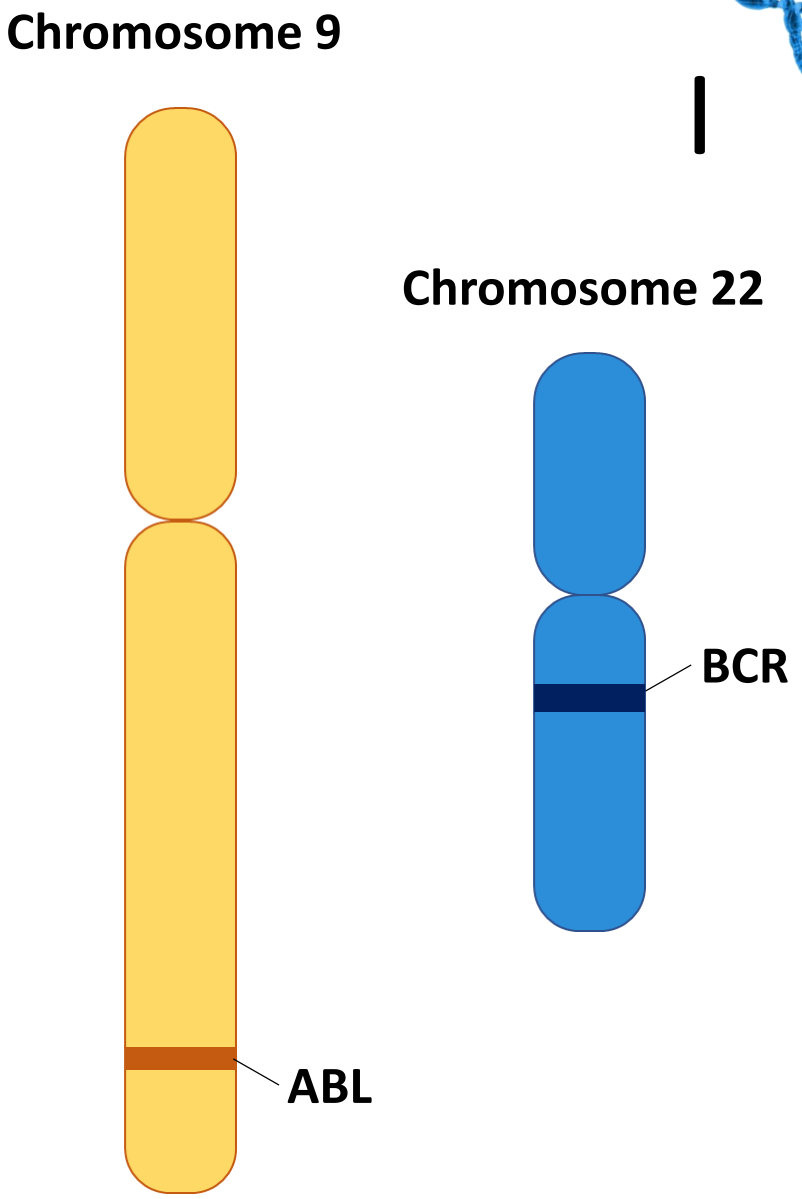
Research Question:
„We have synthesized DNA imitating **CML patients‘** ***BCR-ABL*** **fusion gene** and established a custom-designed NGS workflow. These samples should have a **specific allele frequency** at specific loci in the sequence. Are we able to achieve the defined **target allele frequencies**?“

„Real-Life“ Example II - Questions

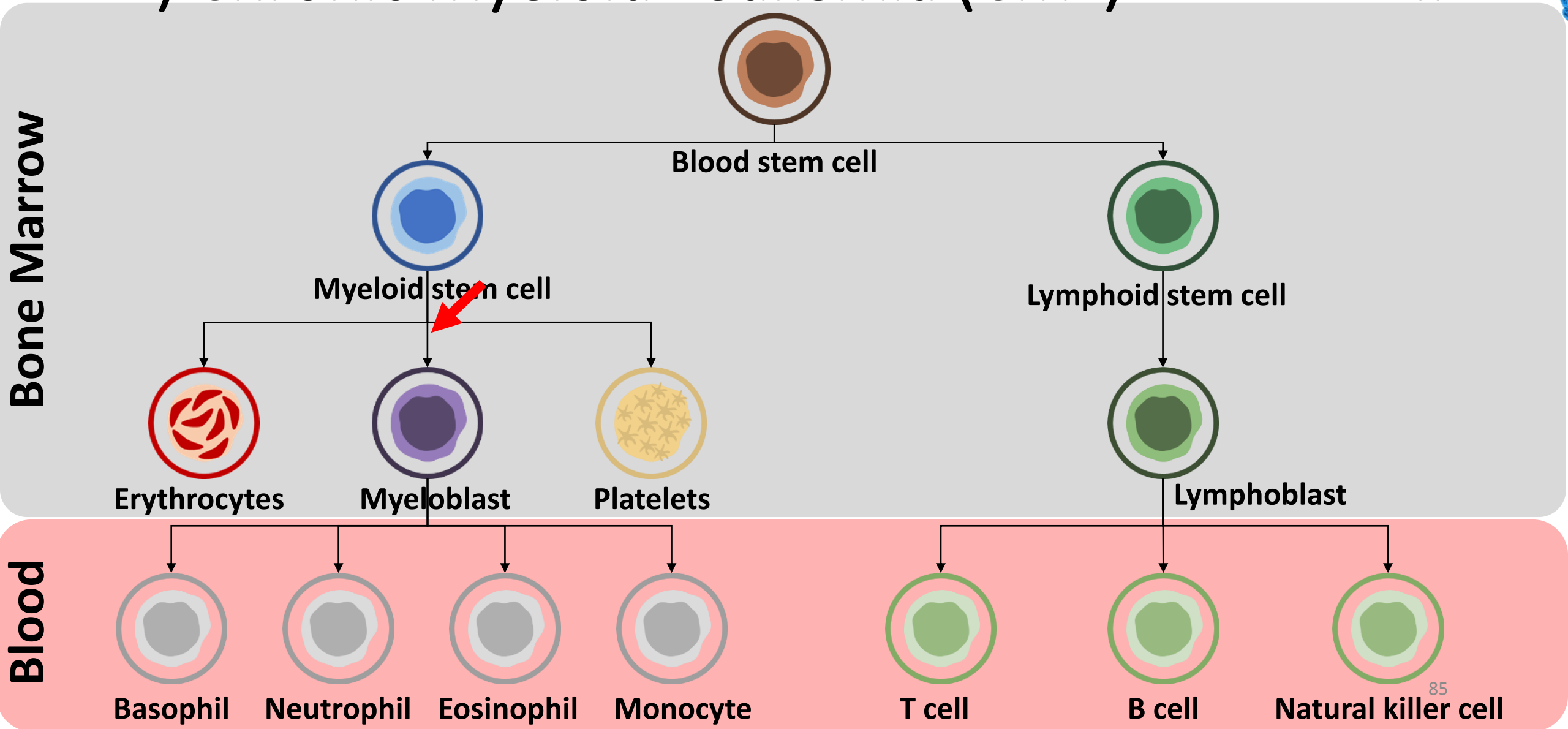
1) What is **chronic myeloid leukemia** (CML)?
2) What **data source/file formats** do we have to work with?
3) Is **additional information** about the received data provided?
4) How can we **evaluate the custom-designed NGS workflow**?

1) Chronic Myeloid Leukemia

- ~ blood cancer
- = hematopoietic neoplasm leading to granulocytic precursor cells‘ uncontrolled proliferation
- Caused by a **translocation** leading to the ***BCR-ABL1* fusion gene** – also called „**Philadelphia chromosome**“

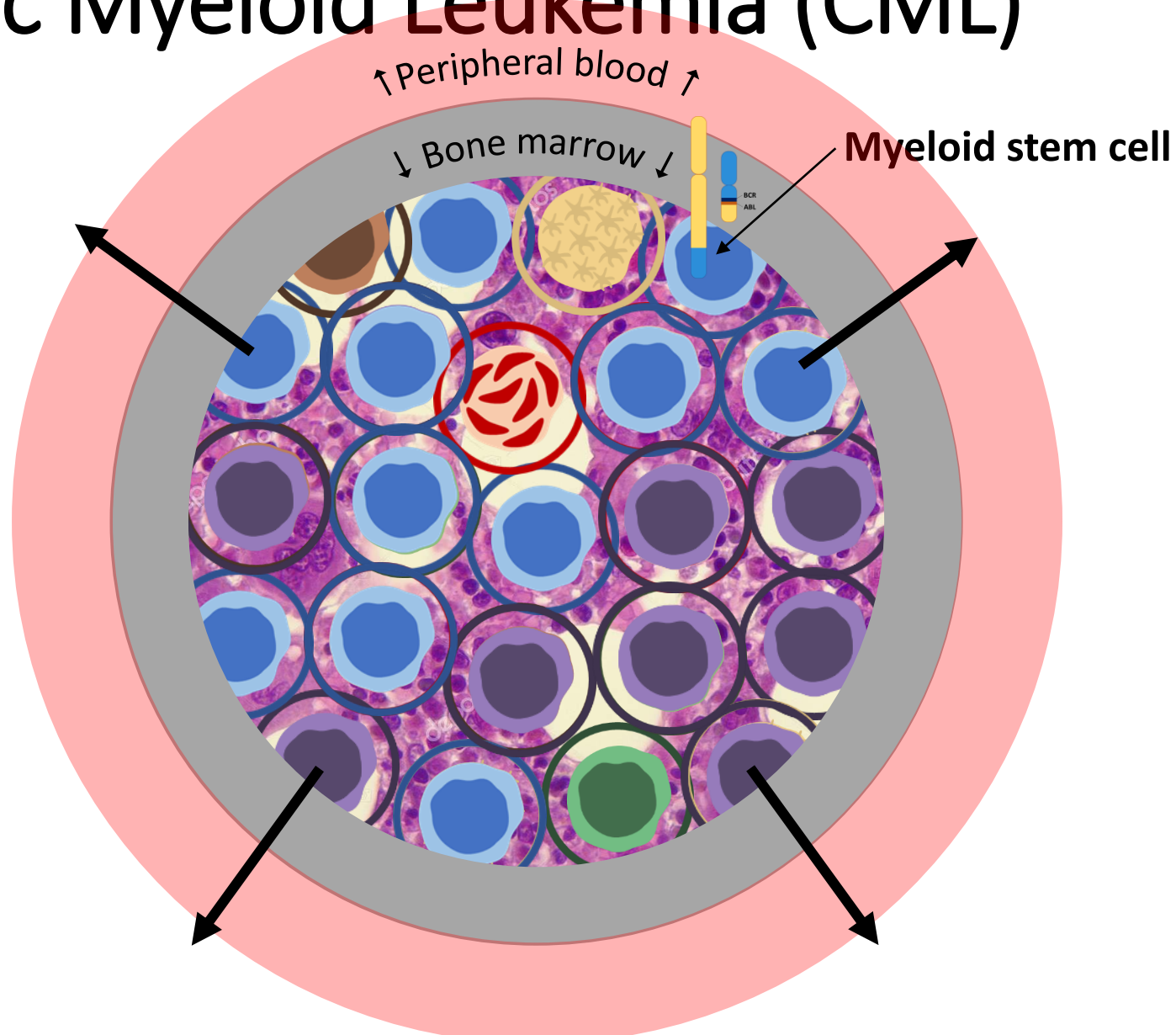


1) Chronic Myeloid Leukemia (CML) II



1) Chronic Myeloid Leukemia (CML)

III



1) Chronic Myeloid Leukemia (CML) IV

- Symptoms
  - Anemia
  - Weight loss
  - Fever
  - Enlarged spleen
- Treatment
  - Tyrosine Kinase Inhibitors (TKIs)
  - (Chemotherapy)
  - (Radiation)
  - (Stem cell transplantation)

1) Chronic Myeloid Leukemia (CML) V

CML Relapse Event

Protein	DNA
M244V	c.730A>G
L248R	c.743T>G
L248V	c.742C>G
G250E	c.749G>A
Q252H	c.756G>C
Y253F	c.758A>T
E255K	c.763G>A
E255V	c.764A>T
D276G	c.827A>G
E279K	c.835G>A
V299L	c.895G>T
T315A	c.943A>G
T315I	c.944C>T
T315V	c.943 944AC>GT
F317L	c.951 c>A
F317R	c.949 950TT>CG
F317V	c.949T>G
M343T	c.1028T>C
M351T	c.1052T>C
F359I	c.1075T>A
F359V	c.1075T>G
L384M	c.1150C>A
H396P	c.1187A>C
H396R	c.1187A>G
F486S	c.1457T>C

2) Data Source/File Formats?

- Synthesized CML NGS data
- in FASTQ file format

3) Additional Information?

- Reference sequence (*ABL*)
- Primer
- Target Mutation Rates: **E255K G > A – 100 %**

G>A 100%

forward primer

reverse primer

ACTACGACAAGTGGGAGATGGAACGCACGGACATCACCATGAAGCACAAGCT…….. AGTTCCCCATCAAATGGACTGCACCCGAGAGCCTGGCCTACAACATGGAGCTGAG

4) Evaluate the Custom-Designed NGS Workflow?

- Are we able to reach the „target allele frequency“?

Hands-on…

```
Colab: ATBI_2

# 1. perform quality check
# 2. clean data
# 3. perform mapping (using bowtie2 or bwa - e.g. on https://usegalaxy.org/ or install bowtie2/bwa)
# 4. perform consensus sequence (using e.g., ivar consensus on https://usegalaxy.org/)
# 5. perform multiple sequence alignment (using e.g., https://www.ebi.ac.uk/Tools/msa/clustalo/)
# 6. merge sequences (using e.g., Samtools merge)
# 7. load the generated BAM and BAI files into IGV
```

EXAMPLE II: CML

Tools

Name	Type	Usage for our example
Galaxy	Tools collection	Provides various tools for biological data (pre-) processing
FastQC	FASTQ quality control tool	Institute: Babraham Institute Quality control tool for high throughput sequencing data
Bowtie2	Sequence mapping tool to reference sequence	Institute: John Hopkins University Mapping of FASTQ files to reference sequence
BWA	Sequence mapping tool to reference sequence	= Burrows-Wheeler Aligner Mapping of FASTQ files to reference sequence
Clustal Omega	Multiple sequence alignment tool	Institute: EMBL-EBI Alignment of two or more sequences – identification of differences/mutations/insertions/deletions – provides information of sequence similarities (phylogenetic tree)
IGV	Integrative genomics viewer	Institute: Broad Institute, University of California Provides information about allele frequencies (and mutations/mismatches/insertions/deletions)

„Real-Life“ Example III

Research Area: Immunology
Research Focus: Gene Expression Analysis

Research Question:
„We are currently analyzing various **gene expressions** of patients suffering from **acute myeloid leukemia (AML)** and comparing them to expressions in healthy individuals. Are you able to **reproduce our results** and do you have some **suggestions for further gene targets**?“

„Real-Life“ Example III - Questions

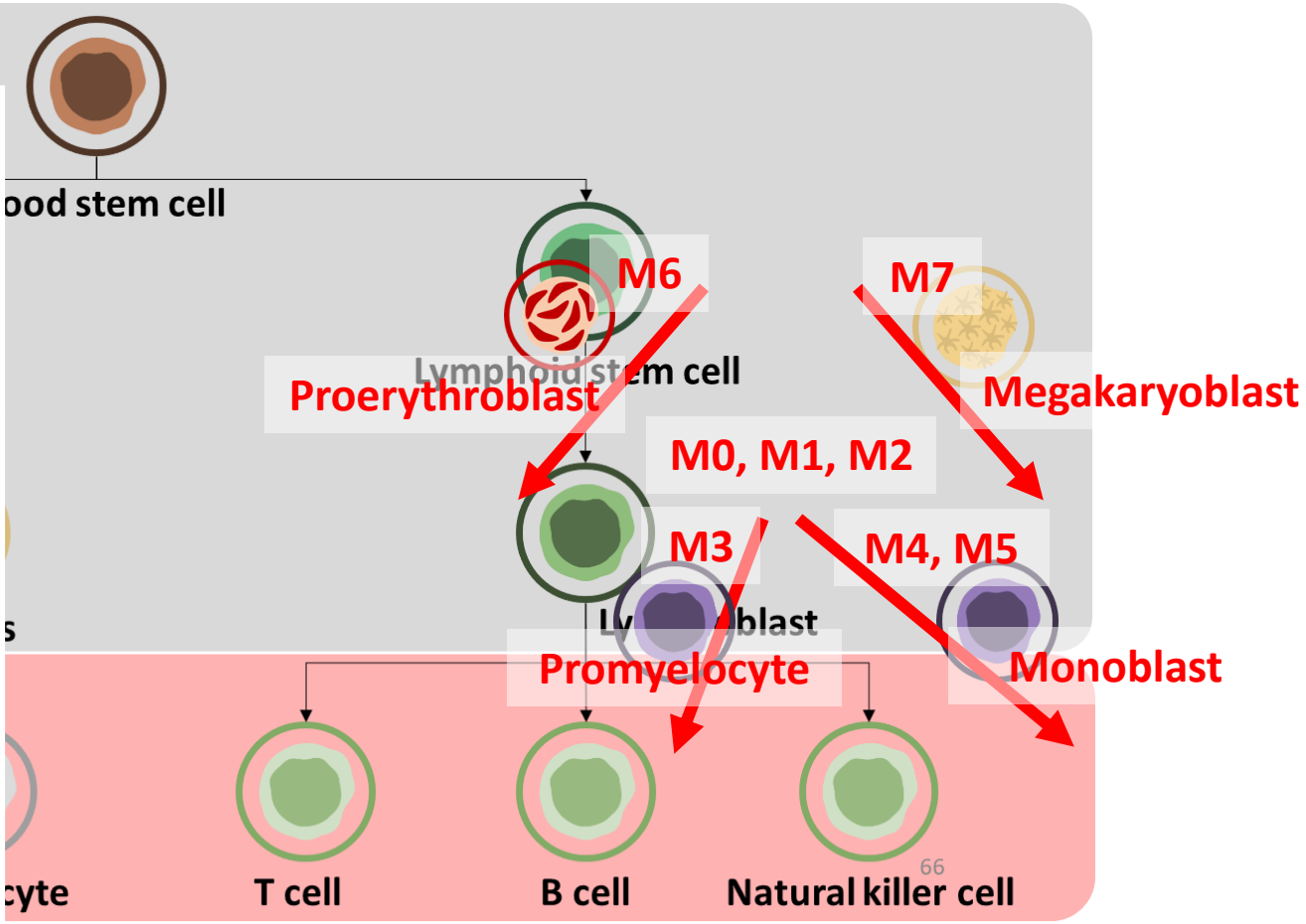
1) What is **acute myeloid leukemia** (AML)?
2) What **data sources** do we have to work with?
3) How do the **results of our “collaboration partners”** look like?
4) Can we **suggest some target genes** for further analysis?

1) Acute Myeloid Leukemia (AML) I

- ~ blood cancer
- = hematopoietic neplasm leading to uncontrolled proliferation of haemotopoietic cells
- 8 subtypes – FAB Classification:
  - M0 – M7
  - Depends on the affected cell types

1) Acute Myeloid Leukemia (AML) II

FAB CLASSIFICATION SYSTEM OF ACUTE MYELOID LEUKAEMIA

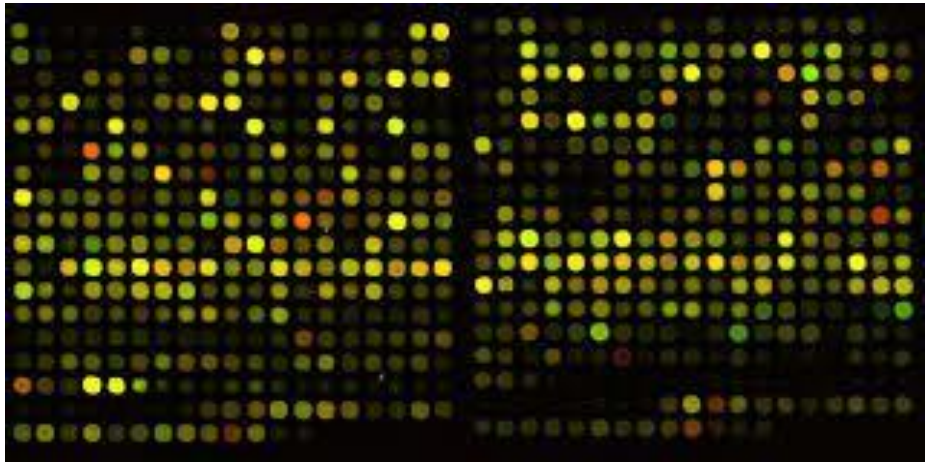


1) Acute Myeloid Leukemia (AML) III

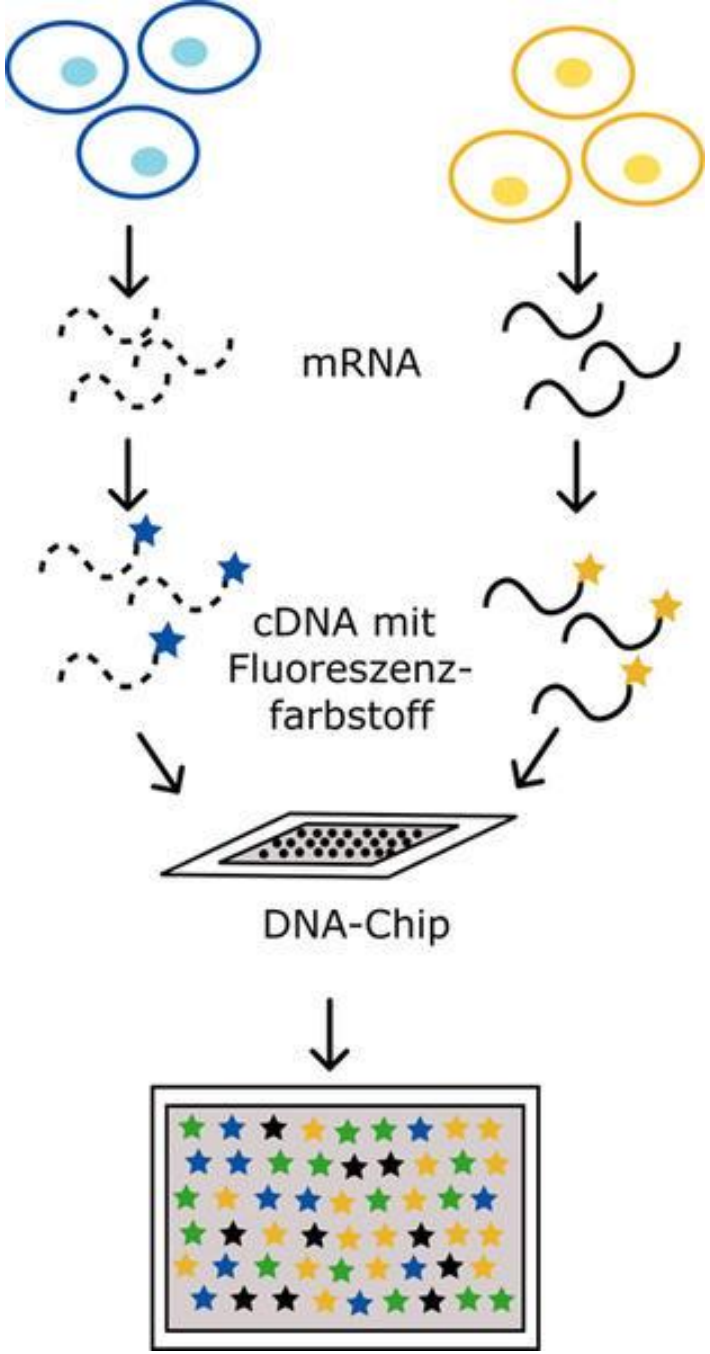
- Symptoms
  - Weight loss
  - Fatigue
  - Fever
  - Night sweats
  - Loss of appetite
- Treatment
  - Chemotherapy
  - Radiation
  - Stem cell transplantation

2) Data Sources

- Public available datasets
- Gene expression data
- Microarray data

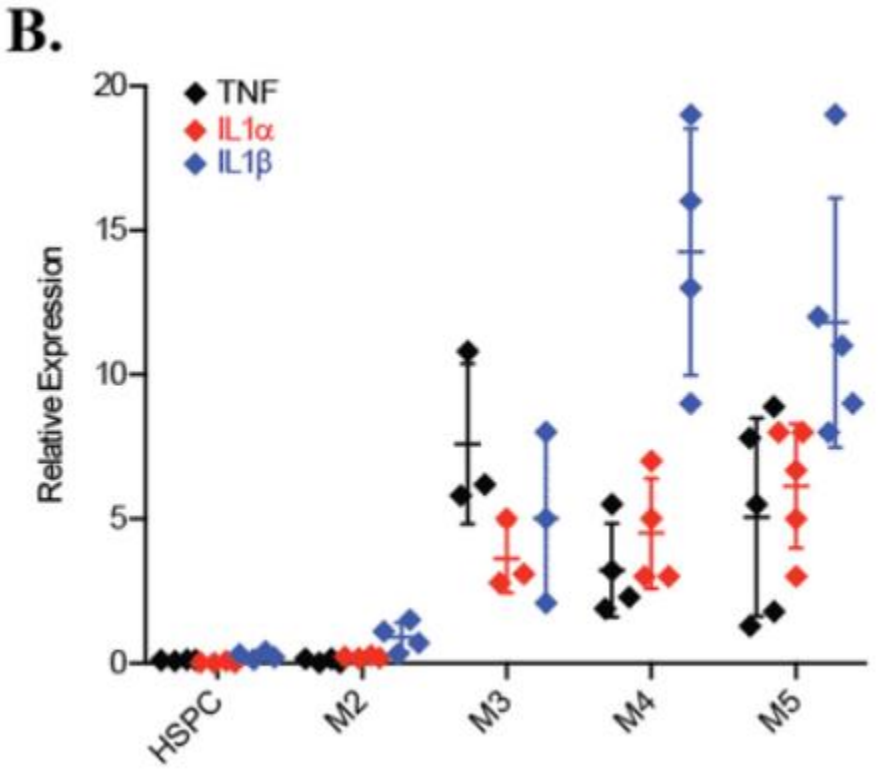


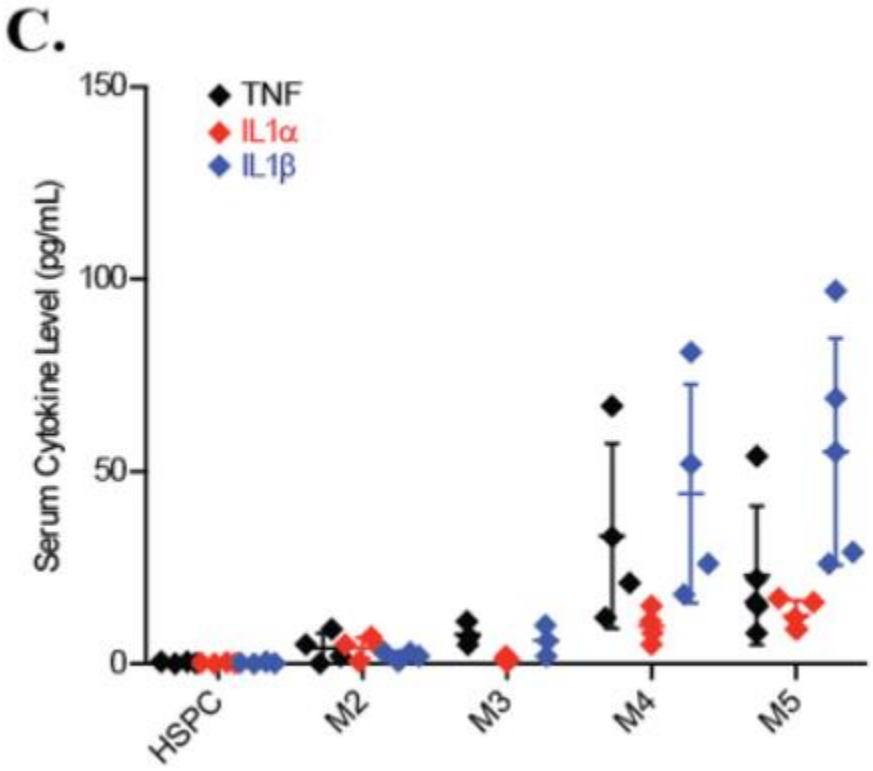
wikipedia



https://www.transgen.de/lexikon/1753.dna-microarray.html

3) Results of „Collaboration Partners“



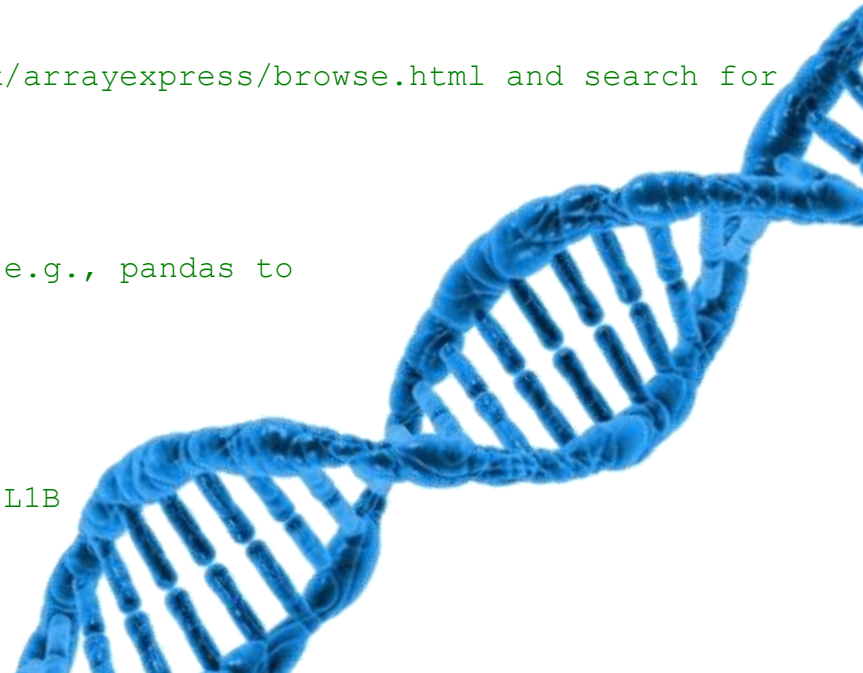


Li et al. (2016) Oncotarget

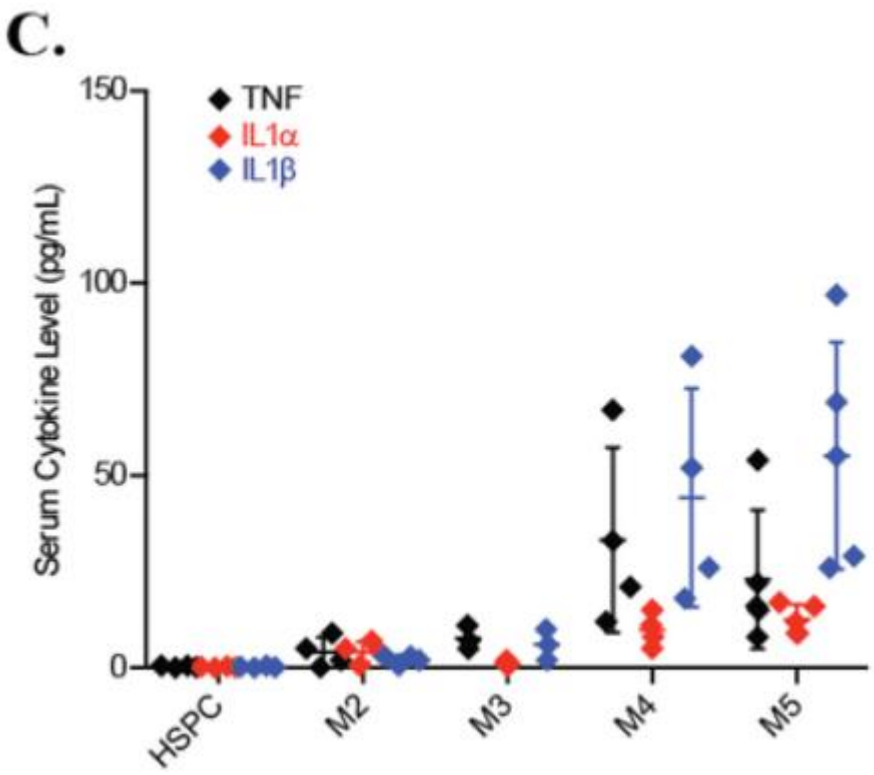
Hands-on…

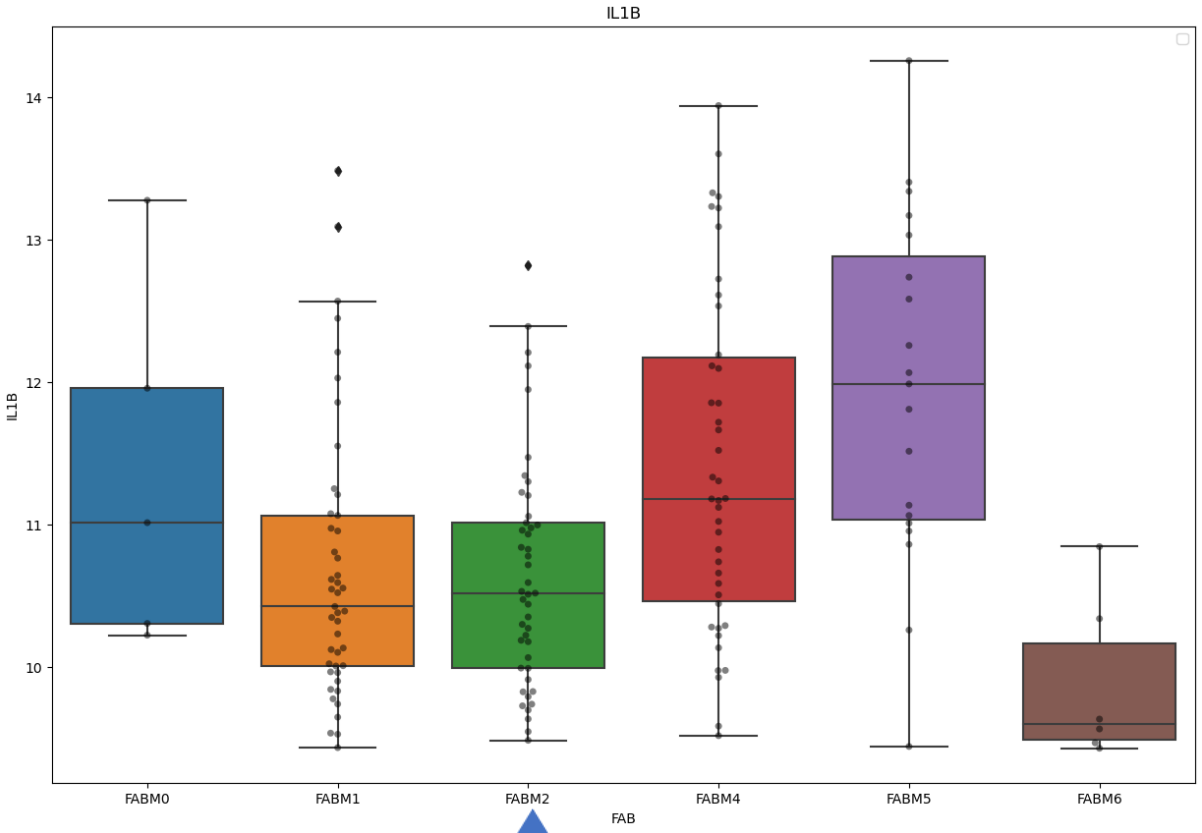
```
Colab: ATBI_2

# 1. get data > go to https://www.ncbi.nlm.nih.gov/geo/browse/ or https://www.ebi.ac.uk/arrayexpress/browse.html and search for
     "AML" > sort by "Assays" > use GSE12417
# 2. use GEOparse to download GSE12417
# 3. get target gene IDs
# 3a. go to http://biogps.org/#goto=welcome and search for the required genes
# 3b. download the platform file (e.g., GPL96-57554.txt) from your GEO dataset and use e.g., pandas to
      find the IDs
# 4. define target genes
# 5. prepare and collect data
# 6. compare FAB classes and perform statistics
# 7. compare AML vs Healthy individuals using GSE13159
# 8. go to https://www.genome.jp/kegg/ and search for additional targets > search for IL1B
     > select TNF SIGNALING PATHWAY
# 9. perform ML (classification AML or healthy control (HC) using GSE13159 data)
```



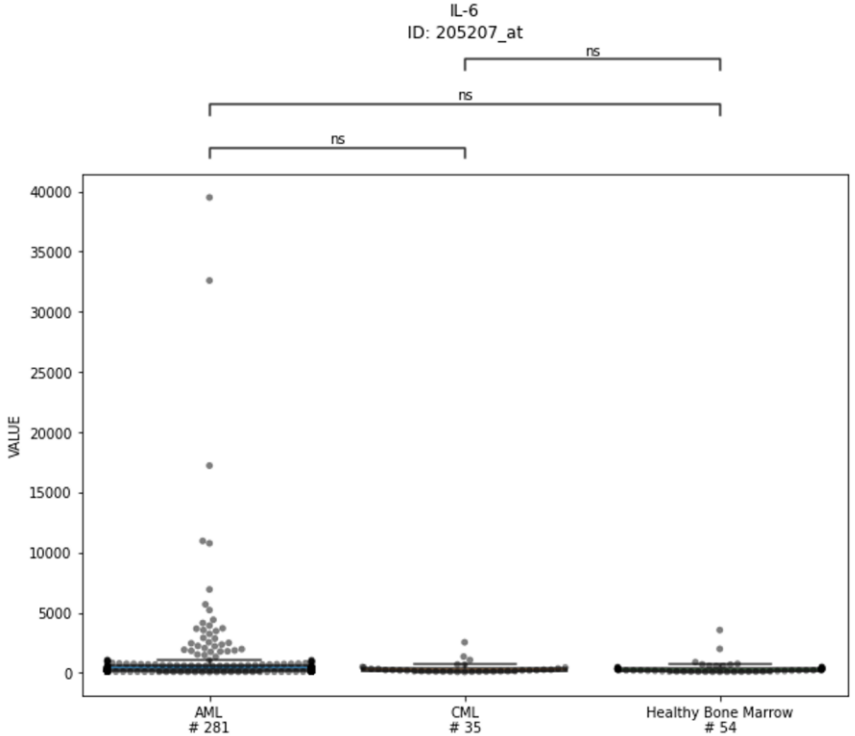
Results comparison

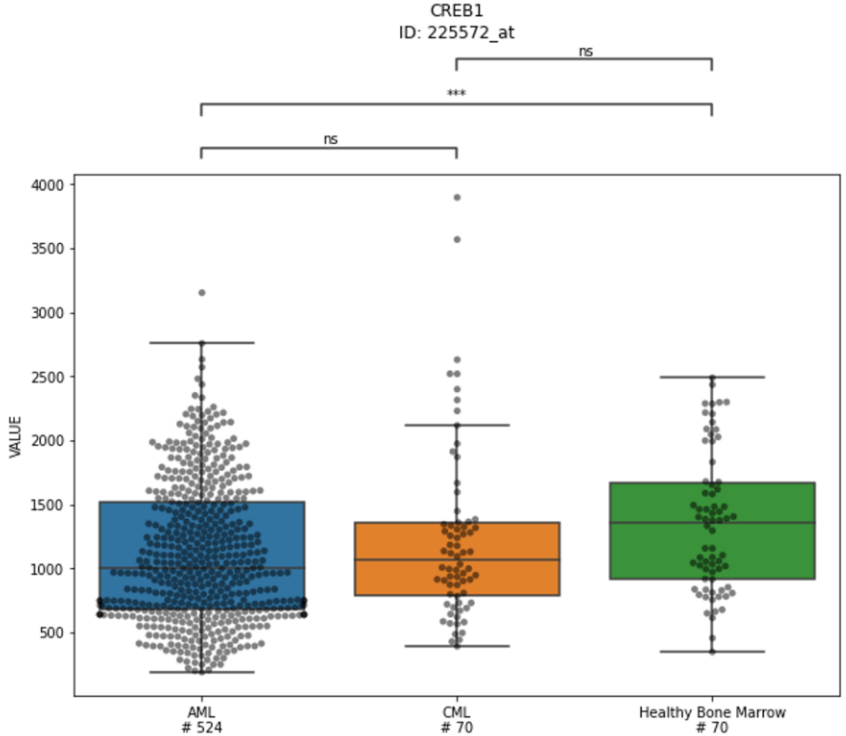


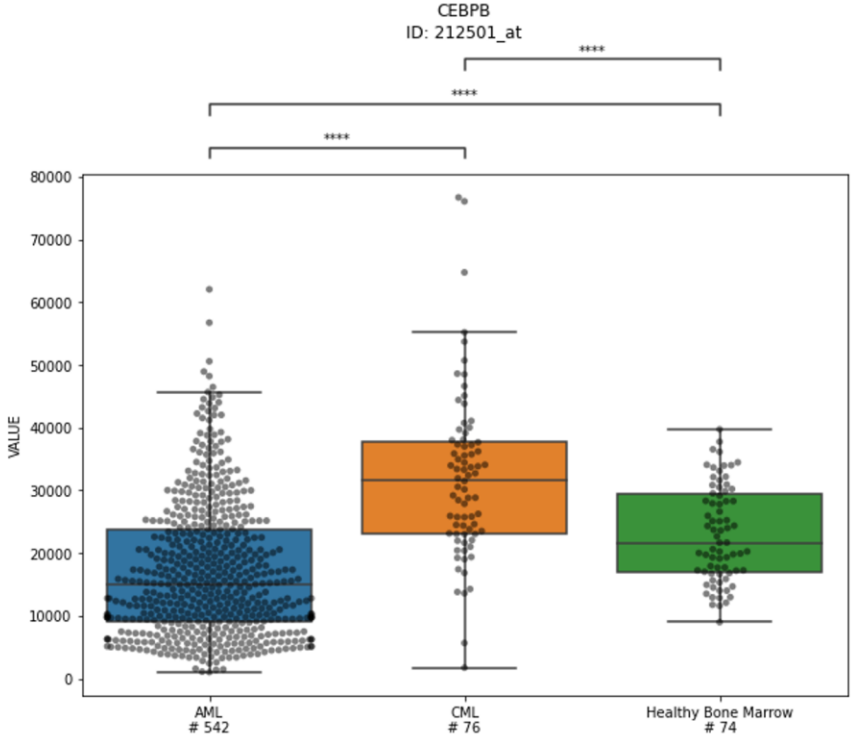


4) Gene Target Suggestions

- CREB1
- CEBPB







EXAMPLE III: AML

Databases

Name	Type	Usage for our example
GEO	Secondary Database Genome Database	= Gene Expression Omnibus Institute: NCBI Gene expression data acquisition
ArrayExpress	Secondary Database Genome Database	Institute: EMBL Gene expression data acquisition
KEGG	Secondary Database Enzymes and Metabolic Pathways	Kyoto Encyclopedia of Genes and Genomes Institute: Kyoto University, Bioinformatics Center, Kanehisa Laboratories Pathway informations
BioGPS	Composite Database	Gene expression data information collection

Tools

Name	Type	Usage for our example
GEOparse	Python package	Collect GEO data
scikit-learn	Python package	Machine Learning for feature importances