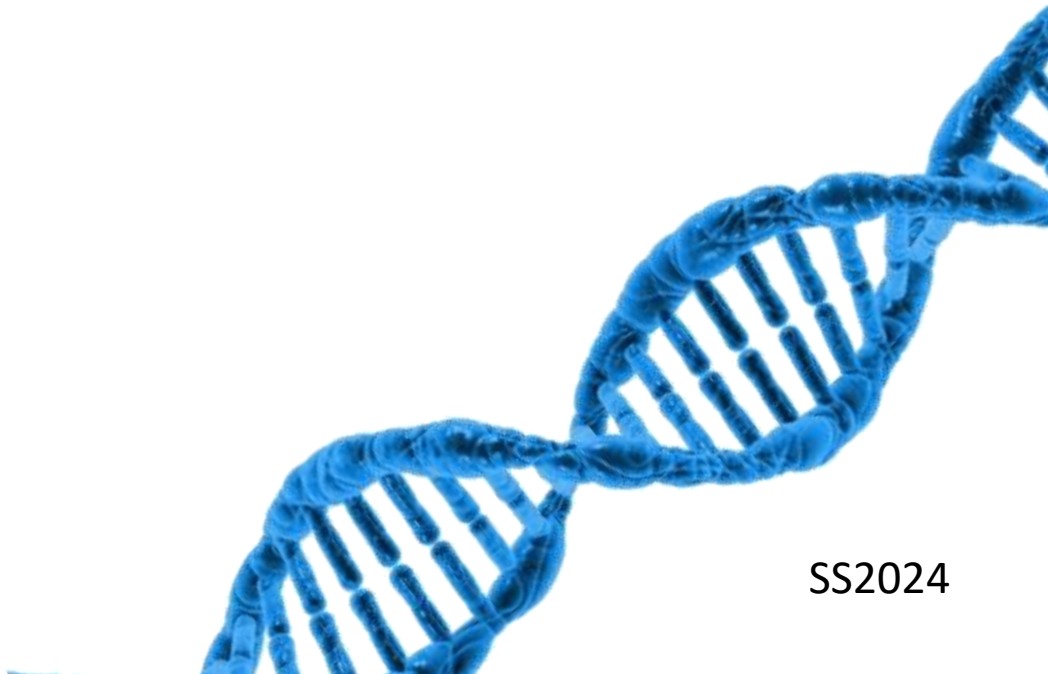
Algorithms and Tools in Bioinformatics

Julia Vetter

julia.vetter@fh-hagenberg.at



SS2024

Course Content

Part I

Data, Tools, and Technologies

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

Part II

Algorithms: Sequence Alignment

(1) Motivation
(2) Similarity of Sequences/ Scoring matrices
(3) Global/Local Alignments
(4) Heuristic Methods
(5) Multiple Sequence Alignment
(6) Phylogenetic Trees

Course Exam

- 2x Exercise Sheets
- Oral exam

(1) Overview

Definitions

Terminology

Why algorithms/tools in bioinformatics?

Data sources

Definition: Bioinformatics

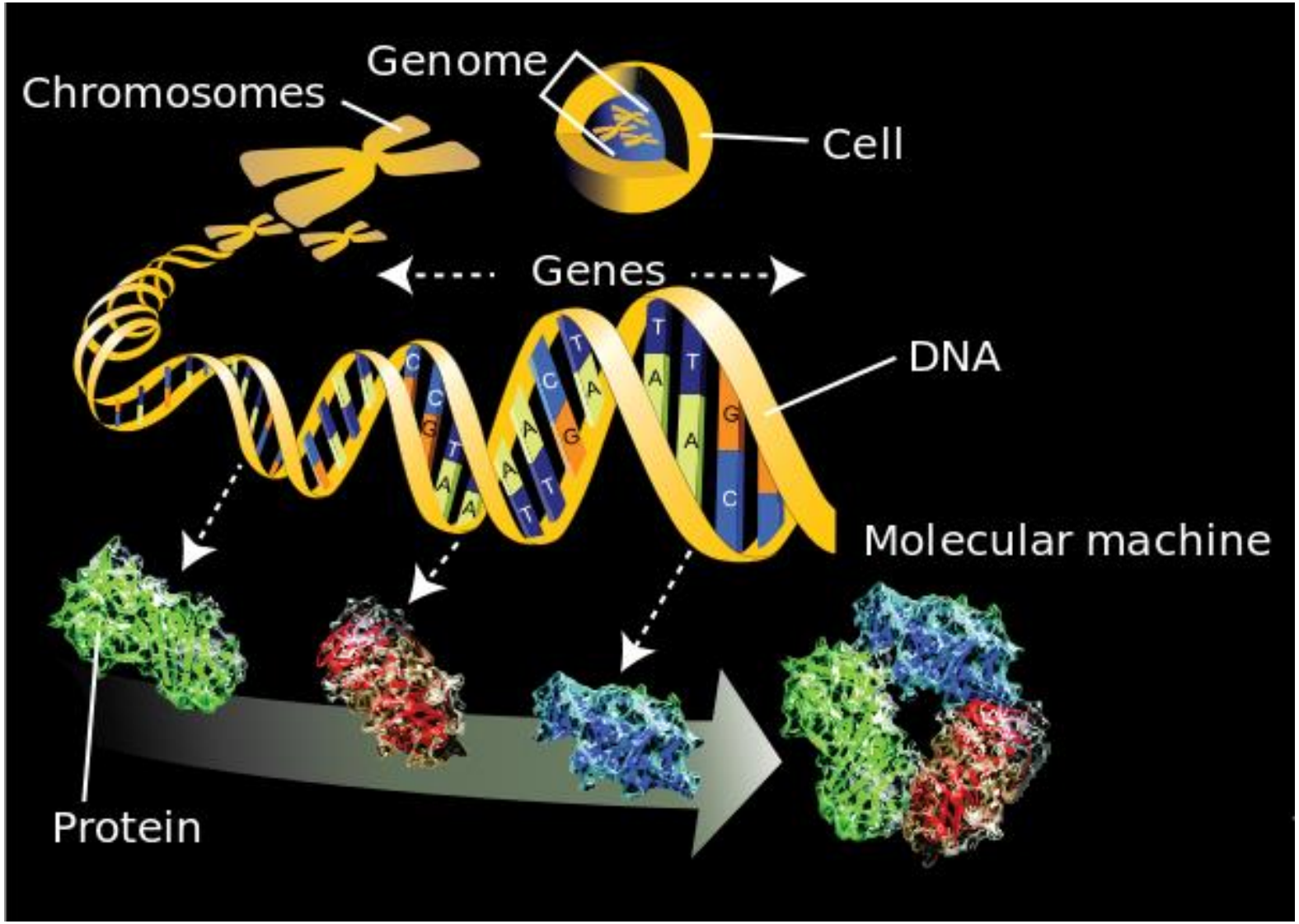
National Human Genome Research Institute (NIH):

"Bioinformatics is a **subdiscipline of biology and computer science** concerned with the **acquisition**, **storage**, **analysis**, and **dissemination of biological data**, most often **DNA and amino acid sequences**.

Bioinformatics uses computer programs for a variety of applications, including **determining gene and protein functions**, **establishing evolutionary relationships**, and **predicting the three-dimensional shapes of proteins**."

Terminology

- Genome
- Gene
- Genotype
- Phenotype
- Nucleic Acid
- Proteome
- Protein/Peptides
- Amino Acid



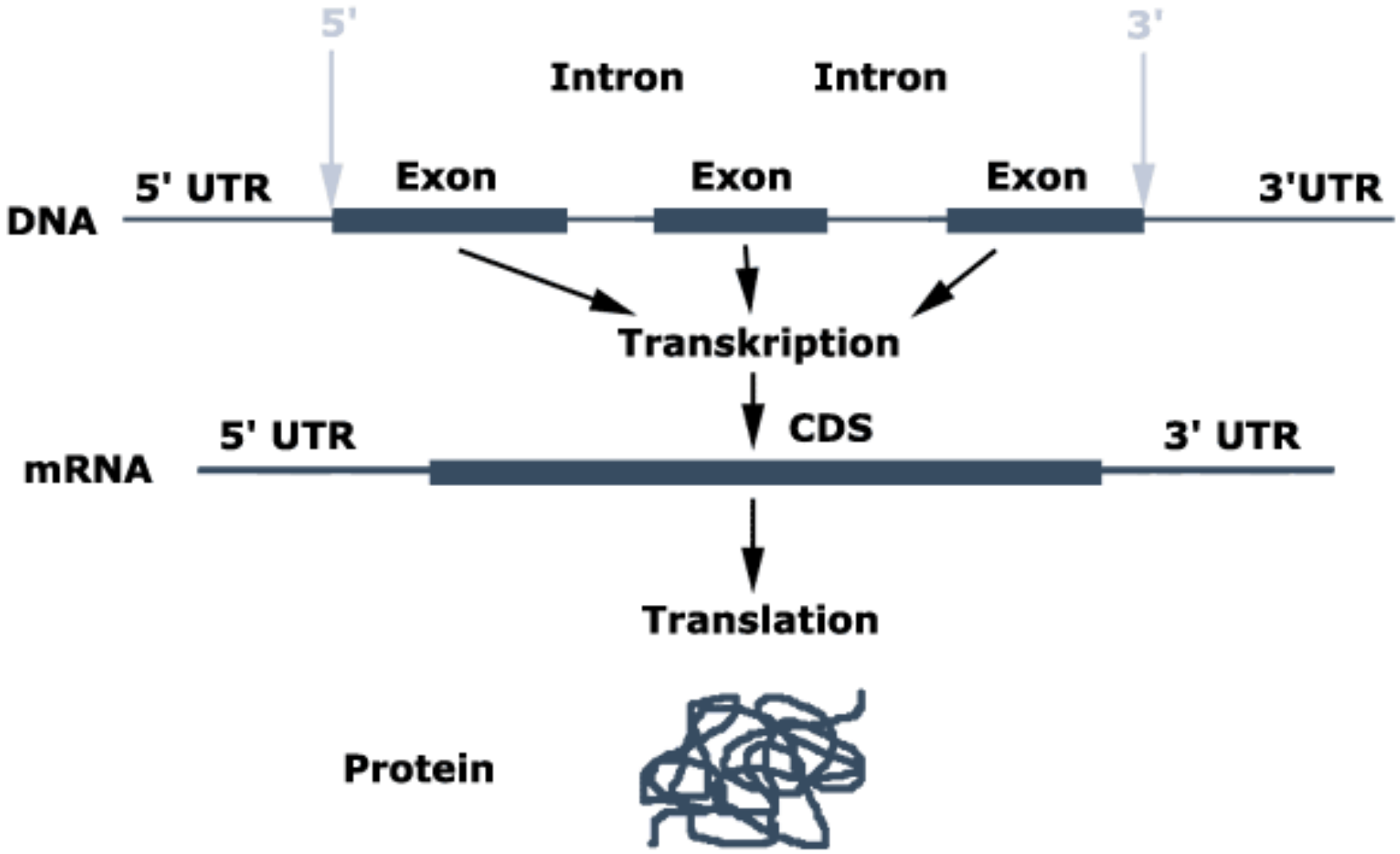
Terminology II

Term	Definition
Genome	Whole genetic material of an organism
Gene	Discrete unit of hereditary information located on the chromosomes and consisting of DNA
Genotype	An individual‘s collection of genes
Phenotype	The physical traits of an organism
Nucleic Acid	= biologic molecules (DNA, RNA) that consists of bases (A, T/U, C, G); they are made of polymers of strings of repeating units; nucleic acids in the cell act to store information and allow organisms to reproduce

Terminology III

Term	Definition
Proteome	Entire set of proteins within an organism
Protein/Peptides	Large biologic molecules which comprise one or more long chains of amino acid residues
Amino Acid	Simple organic compound consisting of an amino group (-NH_2) and and acidic carboxyl group (-COOH); 20 (21) different AA

Protein Biosynthesis



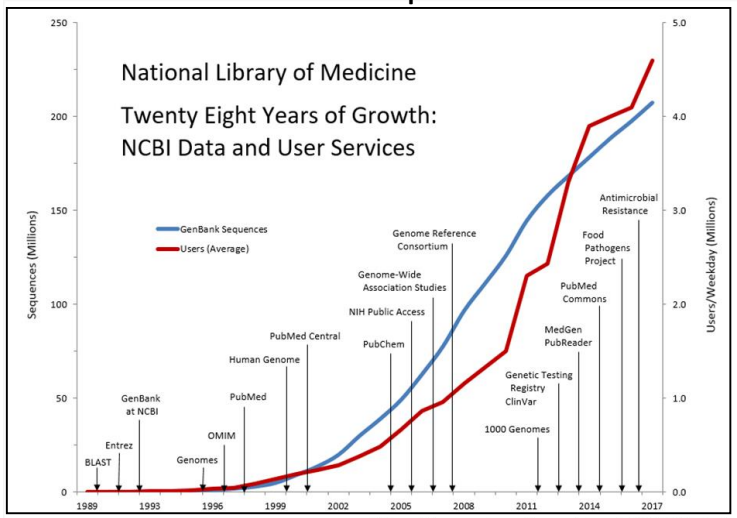
Bioinformatics

- Implementation and design of algorithms/tools for:
  - Biological data analysis (*nt* or *aa* sequence analysis)
  - Knowledge extraction from biological databases
  - Biological data categorization
  - Molecular modelling
  - Protein analysis
  - *In-silico* drug design

Why algorithms and tools are required?

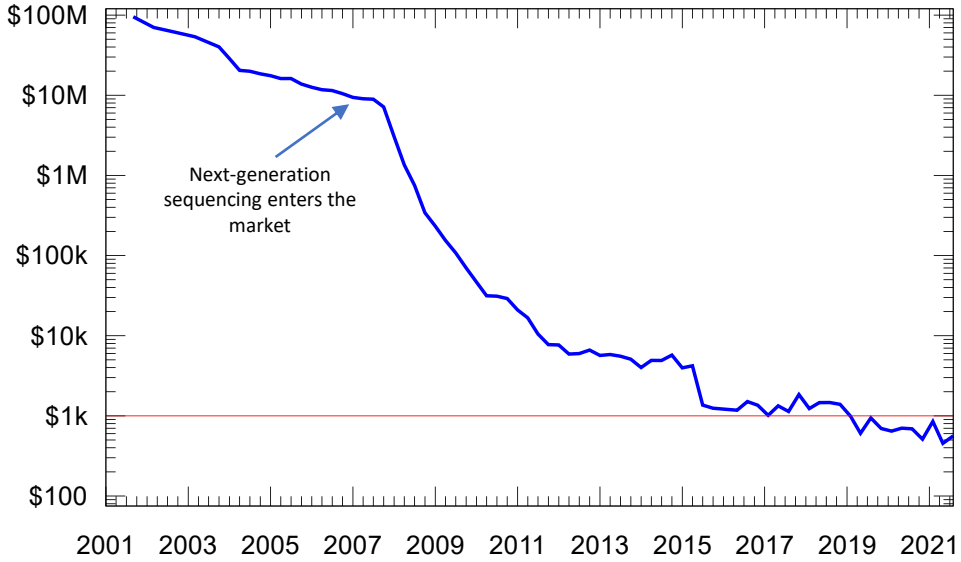
- Increasing data volume
  - Complete genome sequencing
  - Improved sequencing methods
- Public databases
- Performance possibilities
  - High performance data analytics (HPDA)
  - Big data
    - Cloud as a big data source
- Lower data generation costs

GenBank Sequences



https://www.nlm.nih.gov/about/2019CJ.html

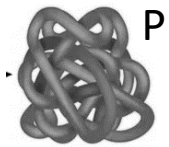
Cost to sequence a human genome (USD)



https://commons.wikimedia.org/wiki/File:Historic_cost_of_sequencing_a_human_genome.svg

Data Sources

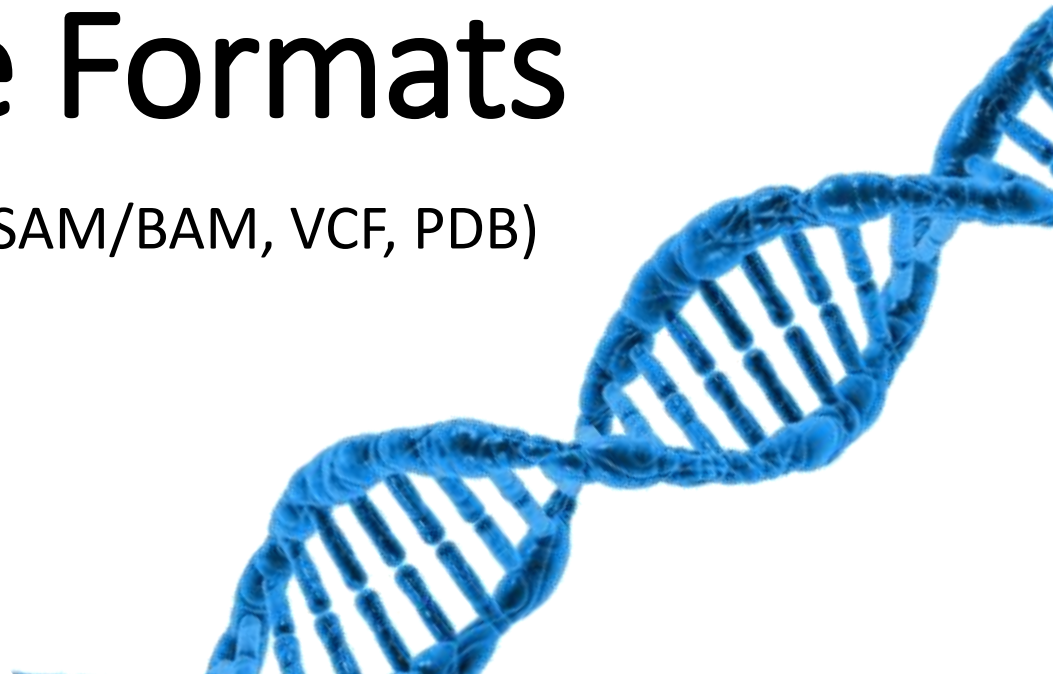
- All life depends on 3 critical molecules: DNA, RNA, Proteins



Nucleic Acid Sequences	Amino Acid Sequences
Source: • DNA & RNA strands • Genes & Gene expression • Genome	**Source:** • Raw AA sequences • Protein/Peptides (spatial information) • Proteom
Analyses: • (Point) Mutations • Open reading frame (ORF) • Phylogenetic analyses • Drug development • Primer design • Virtual translation	**Analyses:** • Missense mutations • Protein structure (prediction) • Protein-protein interactions • Protein-ligand interactions • Metabolism models

(2) Standard Datasets and Modern File Formats

Standard File Formats (FASTA, FASTQ, SAM/BAM, VCF, PDB)



1. FASTA

- Store **nucleotide sequence** or **amino acid** biological information
- First line: >
- Can contain multiple sequences
- No spaces allowed within sequence(s)

FASTA file identifier

Label

Comment

DNA

RNA

```
>NucleotideSequence|Proto-oncogene tyrosine-protein kinase ABL1
CACGGACATCACCATGAAGCACAAGCTGGGCGGGGGCCAGTACGGGG…TTGATGACAGGGGACACCTACACAGCCCATGCTGG
AGCCAAGTTCCCCATCAAATGGACTGCACC
```

Data lines

Protein

```
>2GQG_1|Chains A, B|Proto-oncogene tyrosine-protein kinase ABL1|Homo sapiens (9606)
GAMDPSSPNYDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLVQLLGVC
TREPPFYIITEFMTYGNLLDYLREC…KSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRAC
WQWNPSDRPSFAEIHQAFETMFQES
```

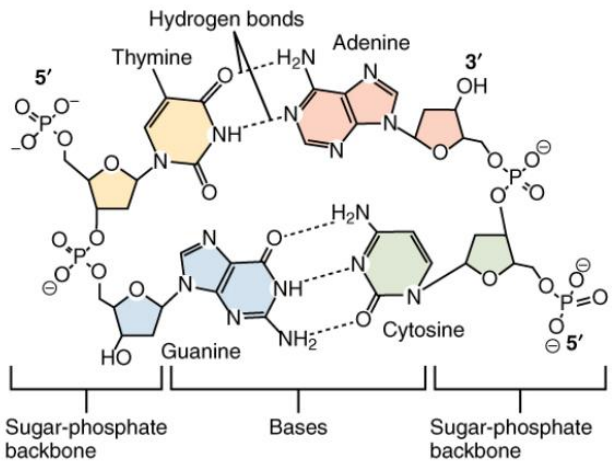
Nucleic Acid Notation: IUPAC Code

A Adenosine
C Cytidine
G Guanine
T Thymidine
U Uridine

R G or A (purine)
Y T or C (pyrimidine)
K G or T (keto)
M A or C (amino)

S C or G (strong)
W A or T (weak)
B G or T or C
D G or A or T
H A or C or T
V G or C or A

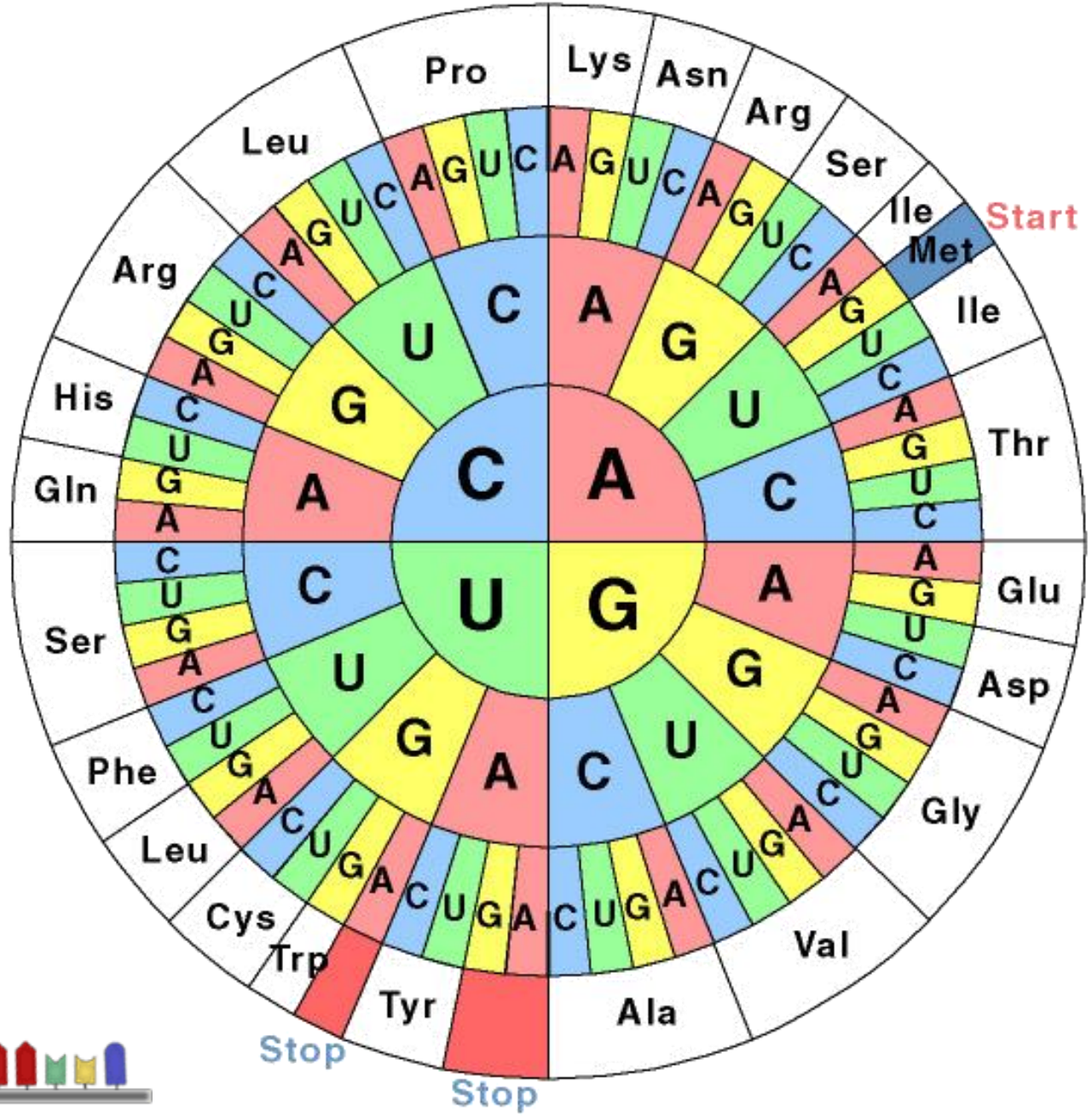
N A or G or C or T
- Gap

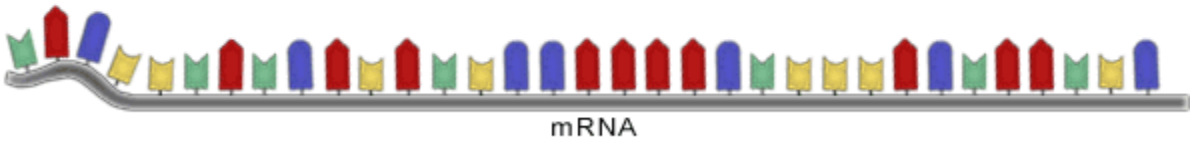


Amino Acid Notation: IUPAC Code

A	Alanine
B	Aspartate/Asparagine
C	Cystine
D	Aspartate
E	Glutamate
F	Phenylalanin
G	Glycin
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
P	Proline
Q	Glutamine
P	Arginine
S	Serine
T	Threonine
U	Selenocysteine
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamate/Glutamine
X	any
*****	Translation stop
-	Gap

Genetic Code





2. FASTQ

- Store **nucleotide sequence** biological information with **their quality** (Quality Score)
- First line: @
- Can contain multiple sequences
- Blocks are called „Reads“
- No spaces allowed within sequence(s)

FASTQ file identifier **Label** **Comment**

```
@NucleotideSequence|Proto-oncogene tyrosine-protein kinase ABL1
CACGGACATCACCATGAAGCACAAGCTGGGCGGGGGCCAGTACGGGG…TTGATGACAGGGGACACCTACACAGCCCATGCTGG
AGCCAAGTTCCCCATCAAATGGACTGCACC
+
FGDADC5CFFFFFFFGFBDGEFFFFFGGGGGGFD86GGGGFGFFFEG…GGGGGGG!#GGGGGGFGGGGGGGGGGFDGGGGGGG
GGGGGGGGGGGGGFGGGG2GGG#GGGGGGG
```

Data lines

Phred Quality Score

Phred Quality Score

- = a measure of the **quality of the identification of the nucleobases** generated by automated DNA sequencing
- **Conversion**:
  - Based on ASCII code (ASCII_BASE 33 (or ASCII_BASE 64))

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

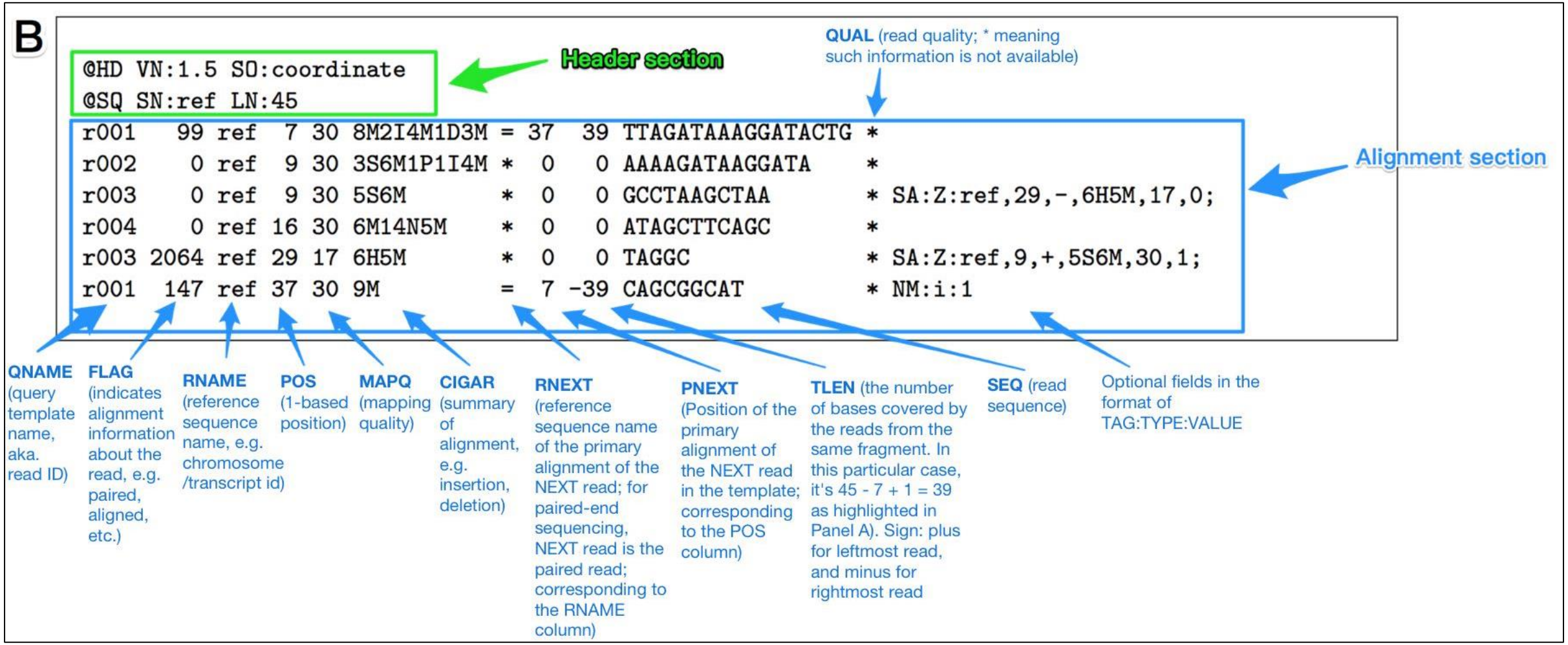
ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

3. SAM/BAM

- SAM = **S**equence **A**lignment/**M**ap
- BAM = **B**inary **A**lignment/**M**ap
- text-based format for storing **biological sequences aligned to a reference sequence**

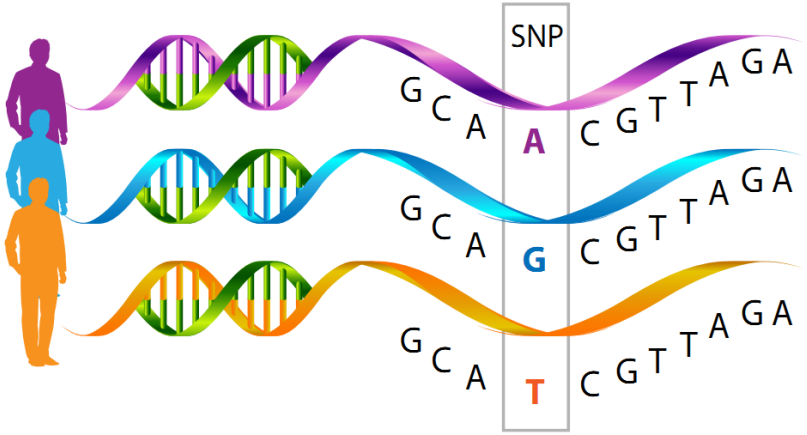
```
Ref:  AAAAATTTAATACGACAAGTGGGAGATGGAACGCACGGACATCACCATGAAGCACAAGCTGGGCGGGGGCCAGTAC...
      | ||| || |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||...
Seq1: ATAAAATTTATAC-ACAAGTGGGAGATGGAACGCACGGACATCACCATGAAGCACAAGCTGGGCGGGGGCCAGTAC...
       *   *  *    *
```



http://zyxue.github.io/2017/09/26/sam-format-example.html

4. VCF

- = Variant Call Format
- Store gene sequence variations
  - SNVs = Single nucleotide variants
  - SNPs = Single-nucleotide polymorphisms
  - Insertions
  - Deletions
  - Mismatches

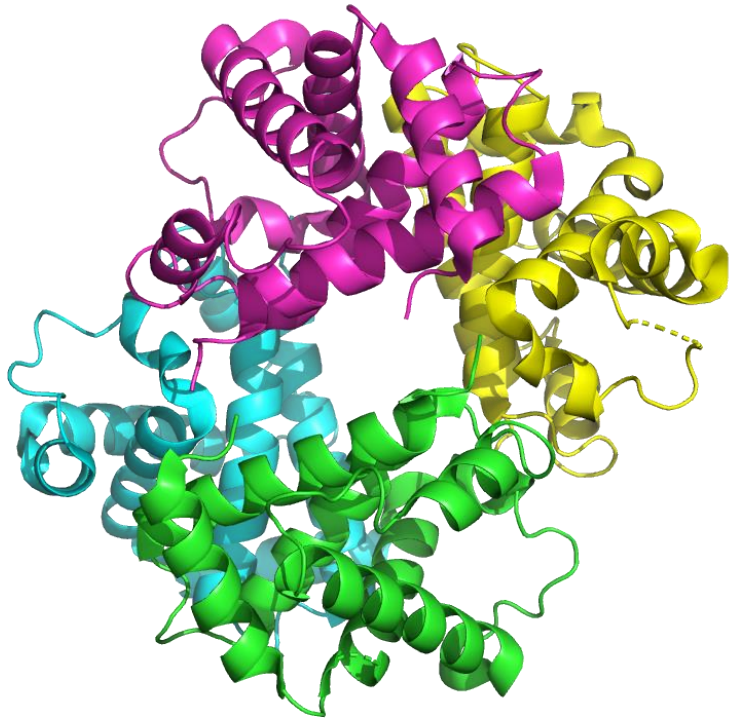



```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1	Sample2	Sample3
2	4370	rs6057	G	A	29	.	NS=2;DP=13;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:52,51	1|0:48:8:51,51	1/1:43:5:.,.
2	7330	.	T	A	3	q10	NS=5;DP=12;AF=0.017	GT:GQ:DP:HQ	0|0:46:3:58,50	0|1:3:5:65,3	0/0:41:3
2	110696	rs6055	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
2	130237	.	T	.	47	.	NS=2;DP=16;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:56,51	0/0:61:2
2	134567	microsat1	GTCT	G,GTACT	50	PASS	NS=2;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3
chr1	45796269 .	G	C
chr1	45797505 .	C	G
chr1	45798555 .	T	C
chr1	45798901 .	C	T
chr1	45805566 .	G	C
chr2	47703379 .	C	T
chr2	48010488 .	G	A
chr2	48030838 .	A	T
chr2	48032875 .	CTAT	-
chr2	48032937 .	T	C
chr2	48033273 .	TTTTTGTTTTA	-
chr2	48033551 .	C	G
chr2	48033910 .	A	T
chr2	215632048 .	G	T
chr2	215632125 .	TT	-
chr2	215632155 .	T	C
chr2	215632192 .	G	A
chr2	215632255 .	CA	TG
chr2	215634055 .	C	T
```

5. PDB

- = Protein Data Bank (pdb) file format
- Textual file format describing three-dimensional structures of molecules of the pdb

```
HEADER    EXTRACELLULAR MATRIX                    22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```



Hands-on…

Modern File Formats

Quality control

File conversion tools

Tools

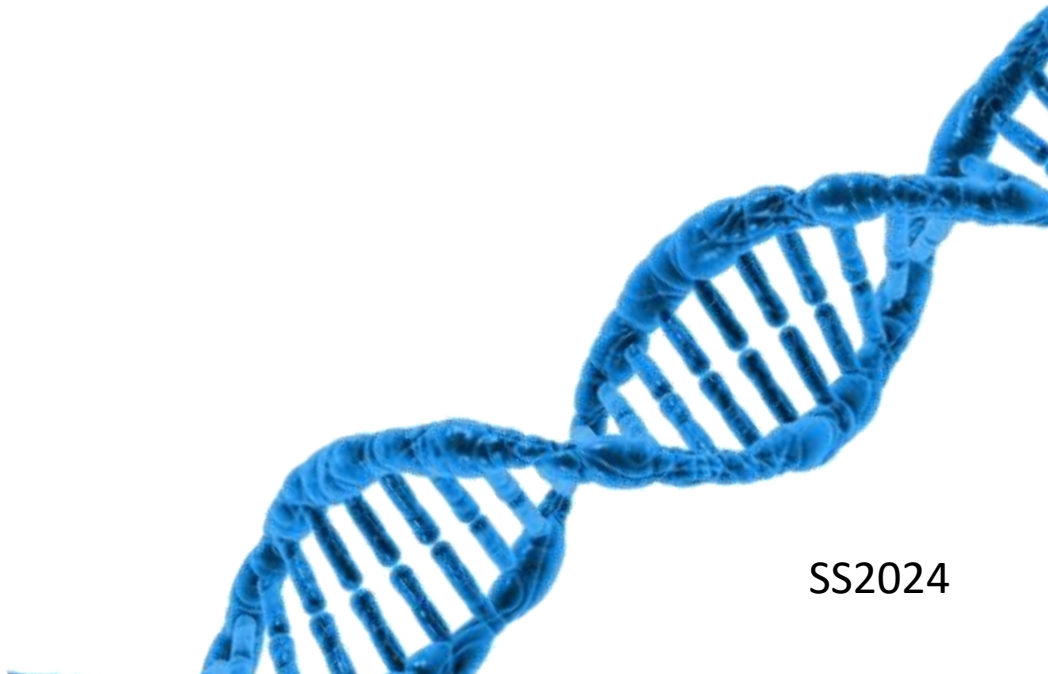
- FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- IGV: https://software.broadinstitute.org/software/igv/download
- PyMol: https://pymol.org/2/
- Galaxy: https://usegalaxy.org/
- samtools: http://www.htslib.org/download/

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

(3) Standard Databases

Institutions

Database Types

Institutions

- NCBI: https://www.ncbi.nlm.nih.gov/ 
- EMBL-EBI: https://www.ebi.ac.uk/   
- Sanger Center: https://www.sanger.ac.uk/ 
- Swiss Institute of Bioinformatics: https://www.sib.swiss/ 
- DDBJ: https://www.ddbj.nig.ac.jp 
- etc.

Institutions: NCBI

- **N**ational **C**enter for **B**iotechnology **I**nformation
- Department of the "National Library of Medicine" (NLM) at the "National Institute of Health" (NIH)
- Founded: 1988
- Tasks:
  - U.S. national resource for molecular biology information
  - Creating public databases
  - Conducts research in computational biology
  - Developing software for analyzing genome data
  - Disseminating biomedical information

Institutions: NCBI

- **Databases**:
  - **PubMed** (Literature)
  - **OMIM** (Online Mendelian Inheritance in Man)
  - **Taxonomy** Browser
  - **GenBank**
  - SNP
  - MMDB (Molecular Modeling Database)
  - UniGene (Unique Human Gene Sequence Collection)
  - **Gene Expression Omnibus (GEO)**

Institutions: NCBI

- **Tools**:
  - **Basic Local Alignment Search Tool (BLAST)**
  - **Open Reading Frame Finder (ORFfinder)**
  - 1000 Genome Browser
  - CDTree

Institutions: EMBL

- **E**uropean **M**olecular **B**iology **L**aboratory
- Founded: 1974
- 21 + 3 + 2 Countries (Europe, Israel, Argentina, Australia)
- 5 main departments:
  - Heidelberg (D)
  - Hamburg (D)
  - Grenoble (F)
  - Hinxton (UK)
  - External „Research Programs“ in Monterotondo (I)

Institutions: EMBL

- Main Tasks:
  - Basic research in molecular biology
  - Service provider for scientists in member states
  - Training provider for employees, students and visitors
  - Development of new methods for research
- Tools:
  - **BLAST**
  - **Clustal Omega** (multiple sequence alignment tool)
- Databases:
  - **EMBL-ENA**
  - **Ensembl**
  - **Protein Data Bank (PDB)**
  - **ArrayExpress**
  - **UniProt**
- **EBI** (**E**uropean **B**ioinformatics **I**nstitute)

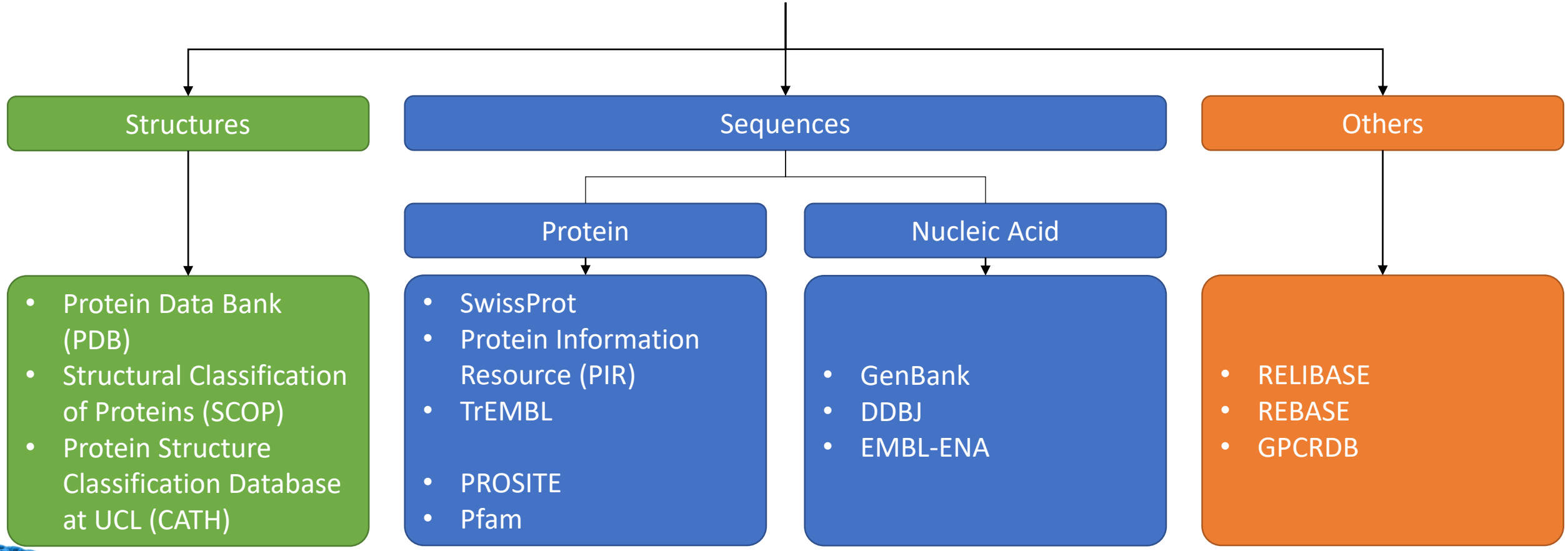
Institutions: Sanger Center

- = Wellcome Trust Sanger Institute
- Research center by Wellcome Genome Campus in Hinxton (UK)
- Main Tasks:
  - Mapping and sequencing of genomes
  - 1998: whole genome of *C. elegans*
  - Participation Human genome project

Database Types: based on Source Types

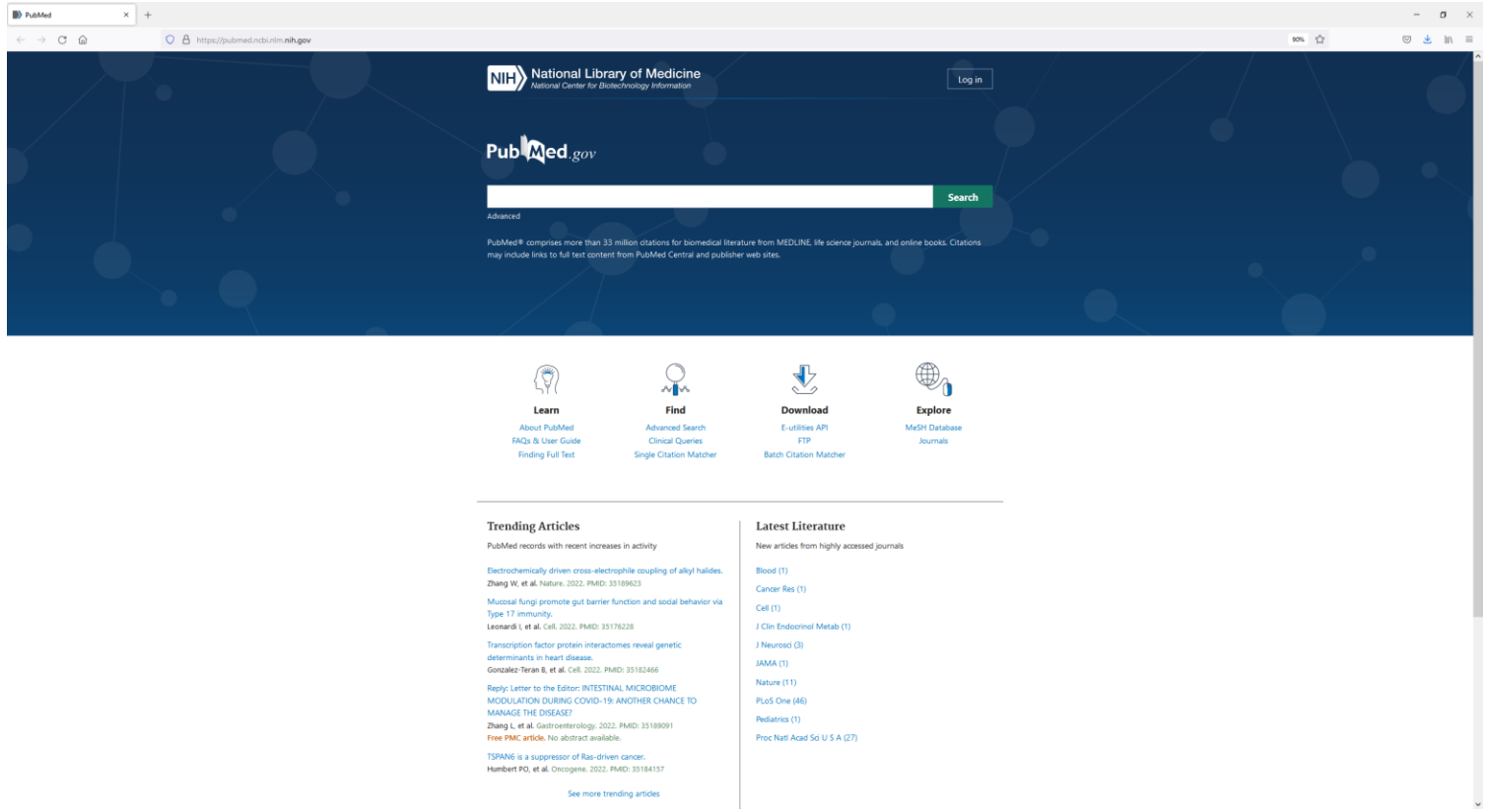
- **Primary Databases:**
  - = archival databases
  - Store experimental results submitted by scientists
    - Sequencing data
    - Macromolecule structure
  - Data have accession numbers
  - e.g.: Protein Data Bank (PDB), GenBank, EMBL-EBI Nucleotide Sequence Database (EMBL-ENA), DNA Data Bank of Japan (DDBJ)
- **Secondary Databases:**
  - Analyzed results of primary databases -> computational algorithms have been applied
  - Contain more valuable knowledge
  - e.g.: UniProt Knowledgebase, InterPro
- **Composite Databases:**
  - Data is filtered and compared before
  - Data is taken from primary database and then merged together
  - BioGPS, OWL, NRDB, BioSilico

Database Types: based on Data Types



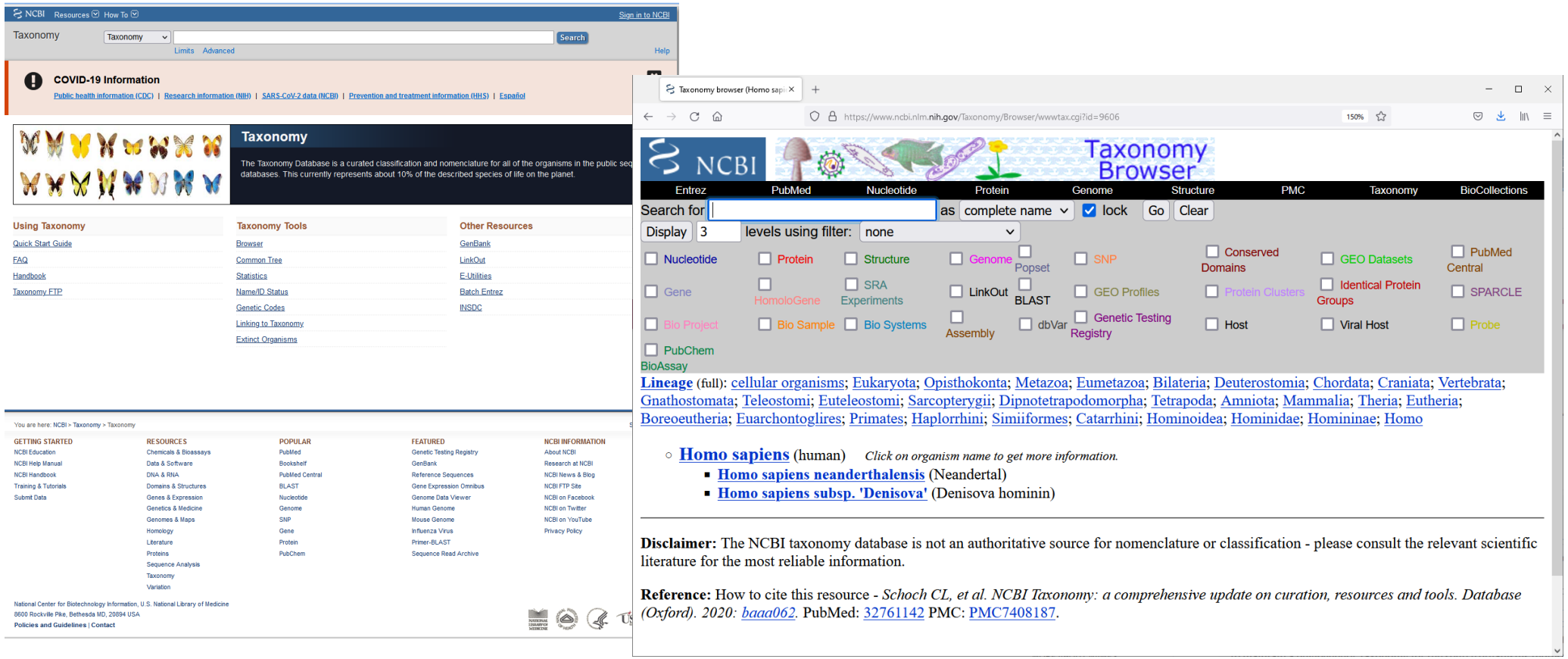
1. Bibliographic Databases

- NCBI – PubMed: http://www.ncbi.nlm.nih.gov/PubMed/
- Google Scholar: https://scholar.google.com/



2. Taxonomic Databases

- NCBI – Taxonomy: https://www.ncbi.nlm.nih.gov/taxonomy



3. Nucleic Acid Databases

- **NCBI – GenBank:** https://www.ncbi.nlm.nih.gov/genbank/
- **EMBL-ENA**: https://www.ebi.ac.uk/ena/browser/
- **DNA Data Bank of Japan (DDBJ)**: https://www.ddbj.nig.ac.jp/index-e.html
- NDB: http://ndbserver.rutgers.edu/

4. Genome Databases

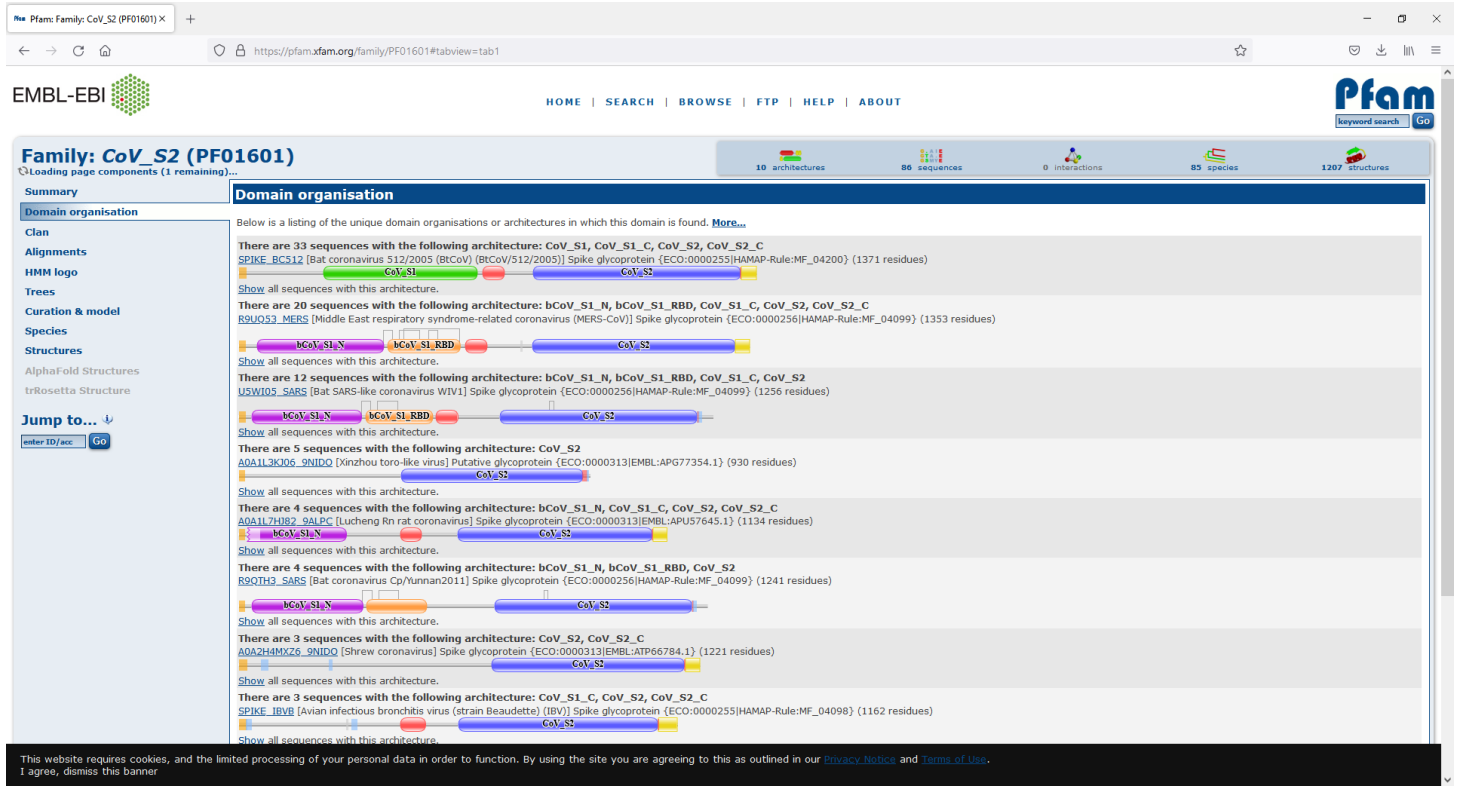
- Gene Expression Omnibus (**GEO**): https://www.ncbi.nlm.nih.gov/geo/
- EMBL-EBI – **ArrayExpress**: https://www.ebi.ac.uk/arrayexpress/
- **Ensemble**: http://www.ensembl.org/index.html
- NCBI – Genome: https://www.ncbi.nlm.nih.gov/genome/
- NCBI – dbVAR: https://www.ncbi.nlm.nih.gov/dbvar
- NCBI – dbSNP: https://www.ncbi.nlm.nih.gov/snp/
- 1000Genomes: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

5. Protein Structure Databases

- RSCB Protein Data Bank (PDB): https://www.rcsb.org/

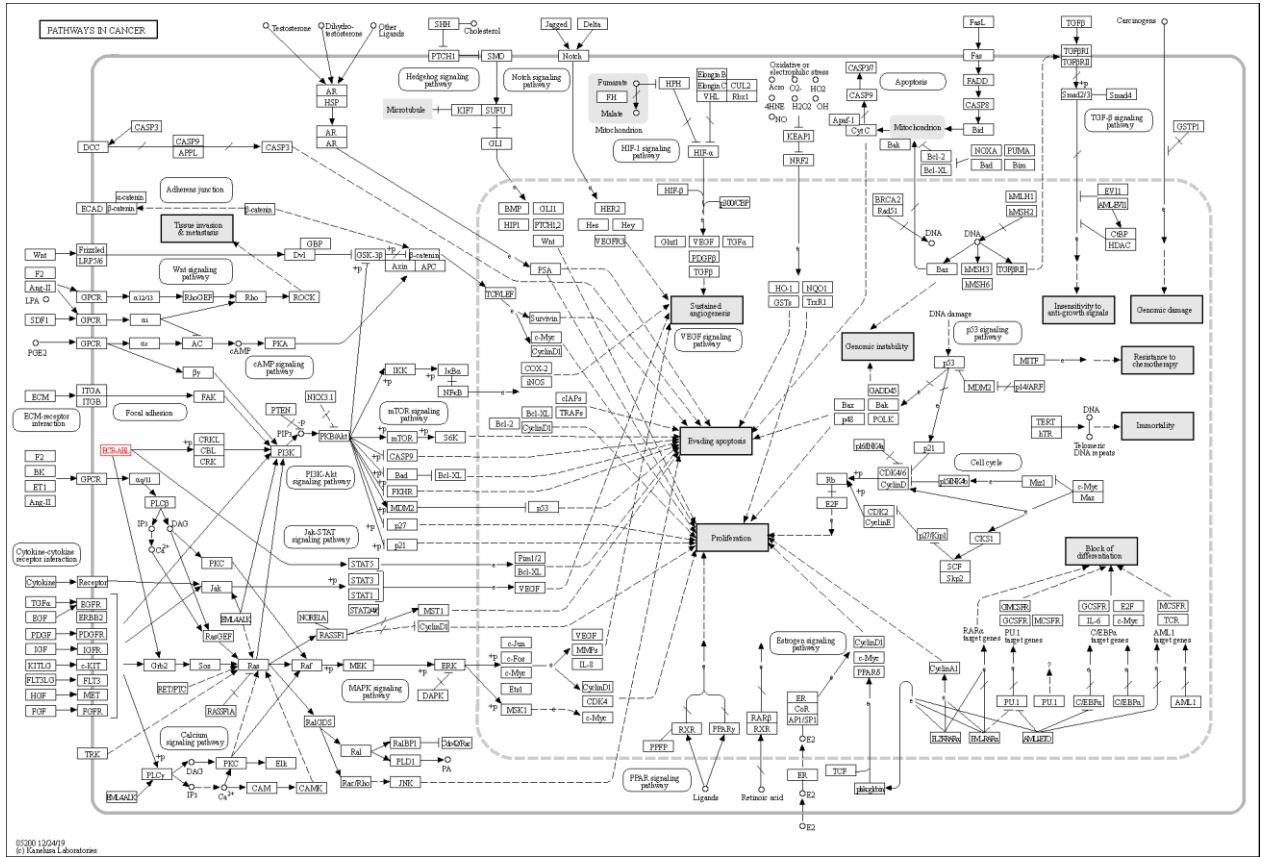
6. Protein Families, Domains, Functionality

- **PFAM**: http://pfam.xfam.org/
- EMBL-EBI – **InterPro**: https://www.ebi.ac.uk/interpro/



7. Enzymes and Metabolic Pathways

- WikiPathways: https://www.wikipathways.org/index.php/WikiPathways
- KEGG: https://www.genome.jp/kegg/



8. … and many more

- RNA specific databases:
  - Collection of different types of RNA data (e.g., tRNA)
  - e.g., Rfam: https://rfam.xfam.org/
- Cancer databases:
  - Collection of characterized cancer genomic data of different cancer types
  - e.g., **The Cancer Genome Atlas (TCGA)**: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
    or Catalogue Of Somatic Mutations In Cancer (COSMIC): https://cancer.sanger.ac.uk/cosmic
- Phenotype databases
- Immunological databases
- Plant databases
- …

Summary

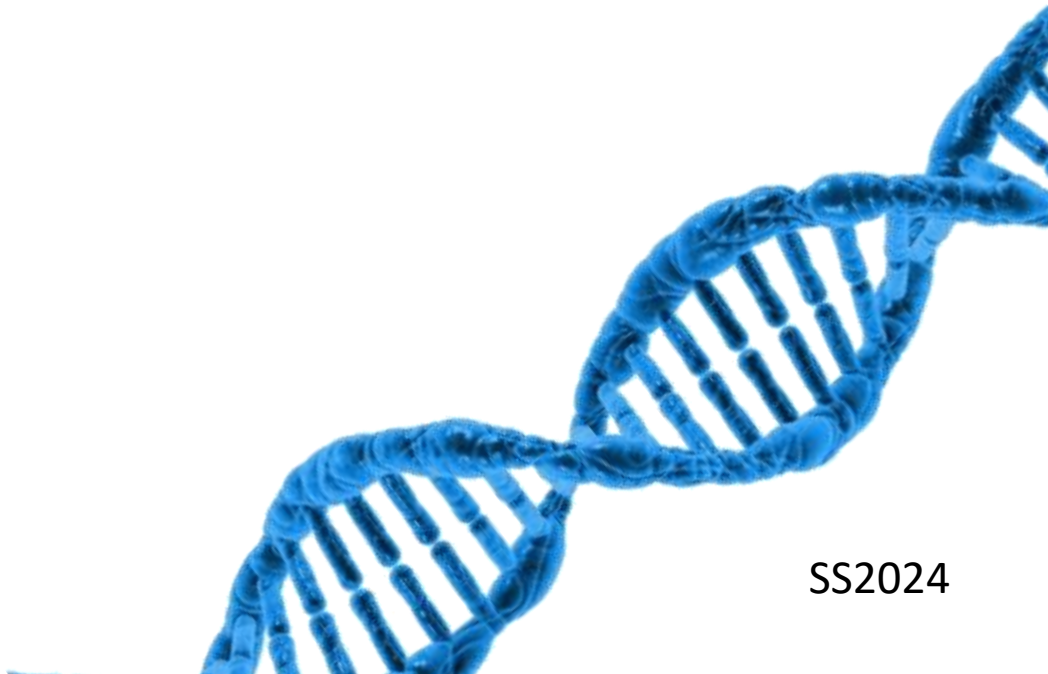
Database	Description	Examples
1. Bibliographic Database	Collection of published literature such as journal and newspaper articles, conference proceedings, (case) reports, books etc.	PubMed, Google Scholar
2. Taxonomic Database	Provides information about biological taxa – groups by species name. Important for biodiversity analyses.	NCBI Taxonomy
3. Nucleic Acid Database	Provides sequencing data in FASTA and FASTQ file format – can be split into DNA and RNA databases	GenBank, EBI-ENA, DDBJ
4. Genome Database	Collection of genome sequences – mostly annotated and analyzed	GEO, ArrayExpress, Ensembl
5. Protein Structure Database	Collection of protein structure information	PDB
6. Protein Families, Domains, Functionality	Provides information about (distant) protein relationships and for protein function analysis	InterPro, Pfam
7. Enzymes and Metabolic Database	Database for understanding biological systems and pathways	Kegg, WikiPathways

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

(4) Data (Pre-) Processing

“Real-Life” Examples:

Population Genetics

NGS Data

Genome Data

Protein Data

Most important…

… know your data!

… know the research question!

„Real-Life“ Example I

Research Area: Genomics
Research Focus: Population genetics

Research Question:
„We have the following primer sequences: **GTGAAAAGCAAGGTCTACCAG** and **GACACCGAGTTCATCTTGAC**. We want to find a so-called *Alu* sequence within the PLAT gene. Are the primers suitable for this problem?“

„Real-Life“ Example I - Questions

1) What is the **PLAT gene**?
2) What (the hell) is a ***Alu* sequence**?
3) What are **primers**?
4) Where do **the primers bind**?
5) Evaluation: Is the desired ***Alu* sequence covered by using these primers**?

1) PLAT Gene

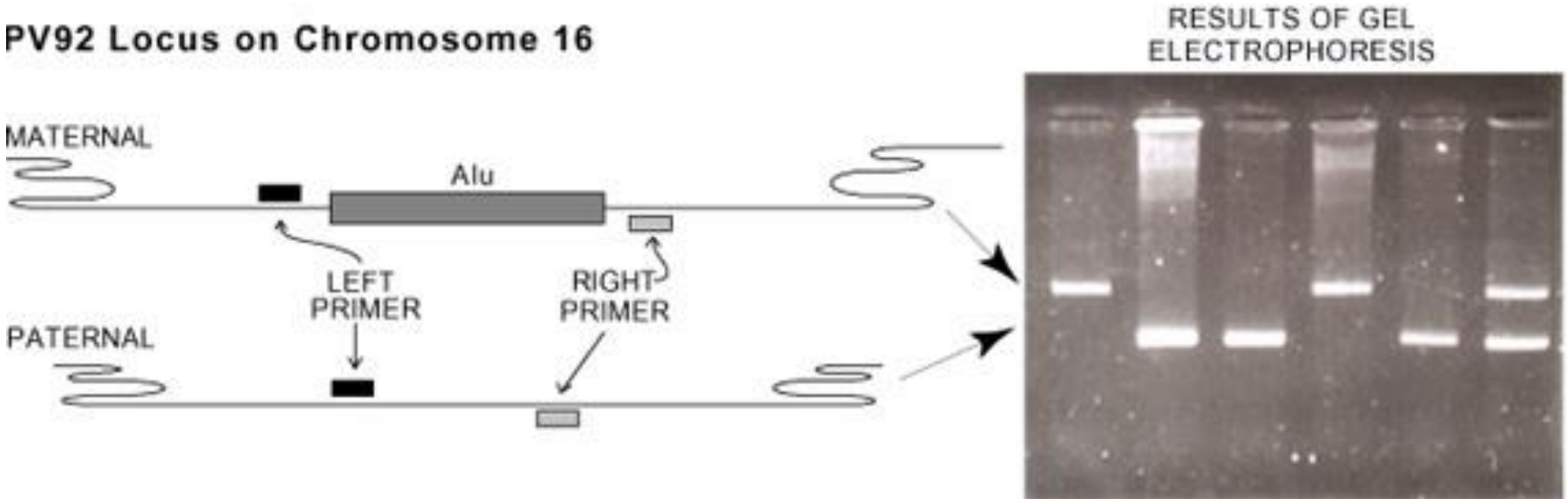
I

- OMIM: https://www.omim.org/

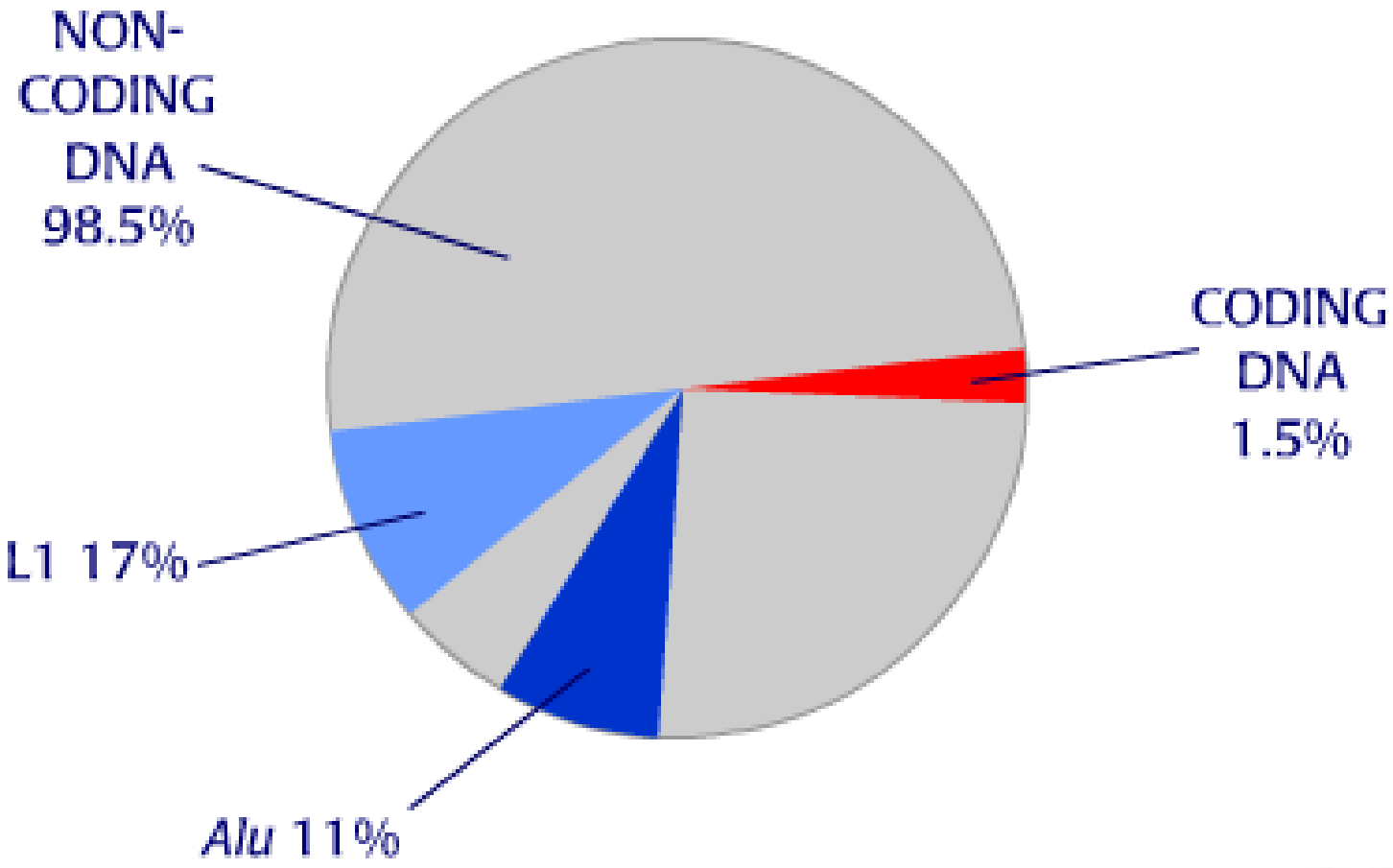
- **Gene:** ✓ PLASMINOGEN ACTIVATOR, TISSUE; PLAT
- **Location**: ✓ Chromosome 8
- **Intron or Exon:** ✓ Intron 8
- **Population genetics:** ✓ insertion/deletion polymorphism of a 311-bp *Alu* sequence

1) PLAT Gene

II



2) *Alu* Sequence



2) *Alu* Sequence II

NCBI Resources How To

MeSH MeSH alu sequence

Create alert Limits Advanced

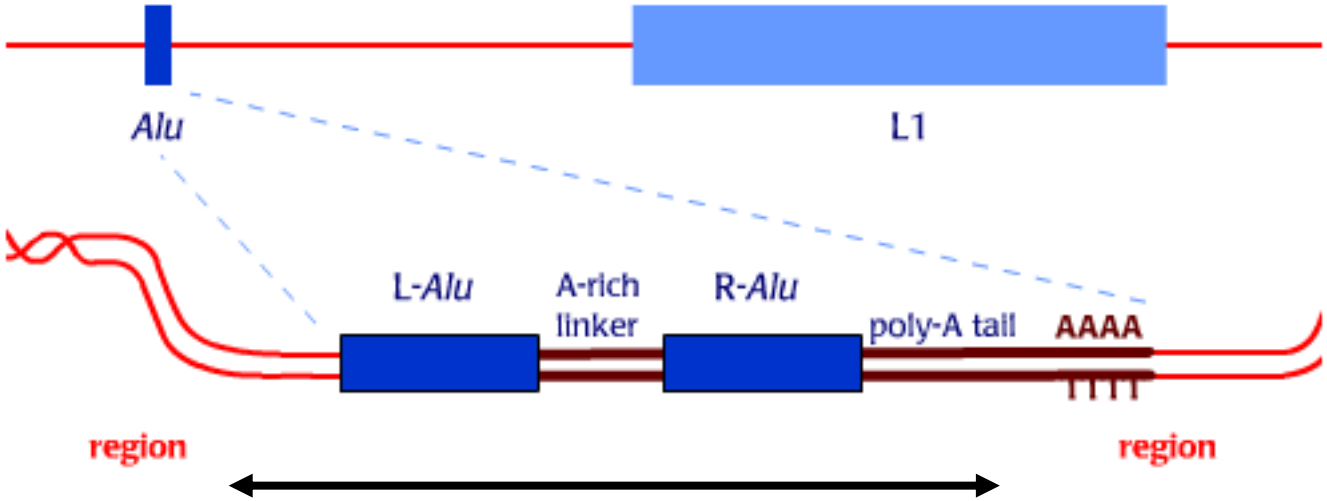
Full Send to:

Alu Elements

The Alu sequence family (name repeat element
in humans (over a million copie ns an RNA
polymerase III promoter. Trans diseases.
Year introduced: 1999

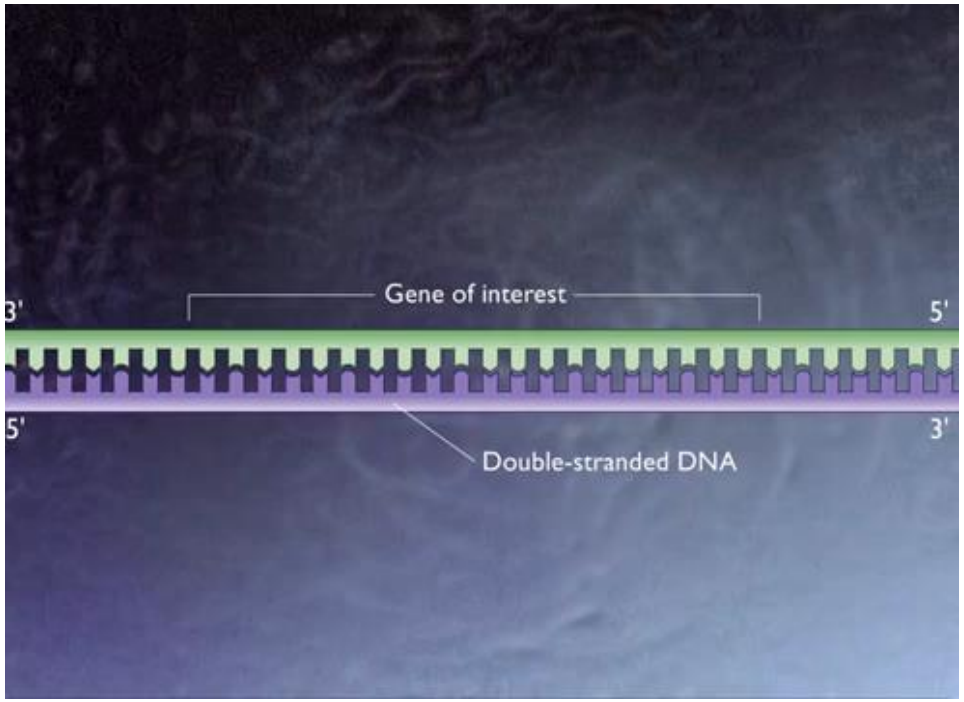
PubMed search builder options
Subheadings:

- drug effects
- etiology



6.000 bp

3) Primer



4) Primer Analysis

In-Silico PCR: https://genome.ucsc.edu/cgi-bin/hgPcr

Ensembl: https://www.ensembl.org/Multi/Tools/Blast?db=core

- Hits: ✓ 2 & 1
- Melting temperature: ✓ 56,6 °C & 55,4 °C

eLearning Task

```
# 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 ;)
```