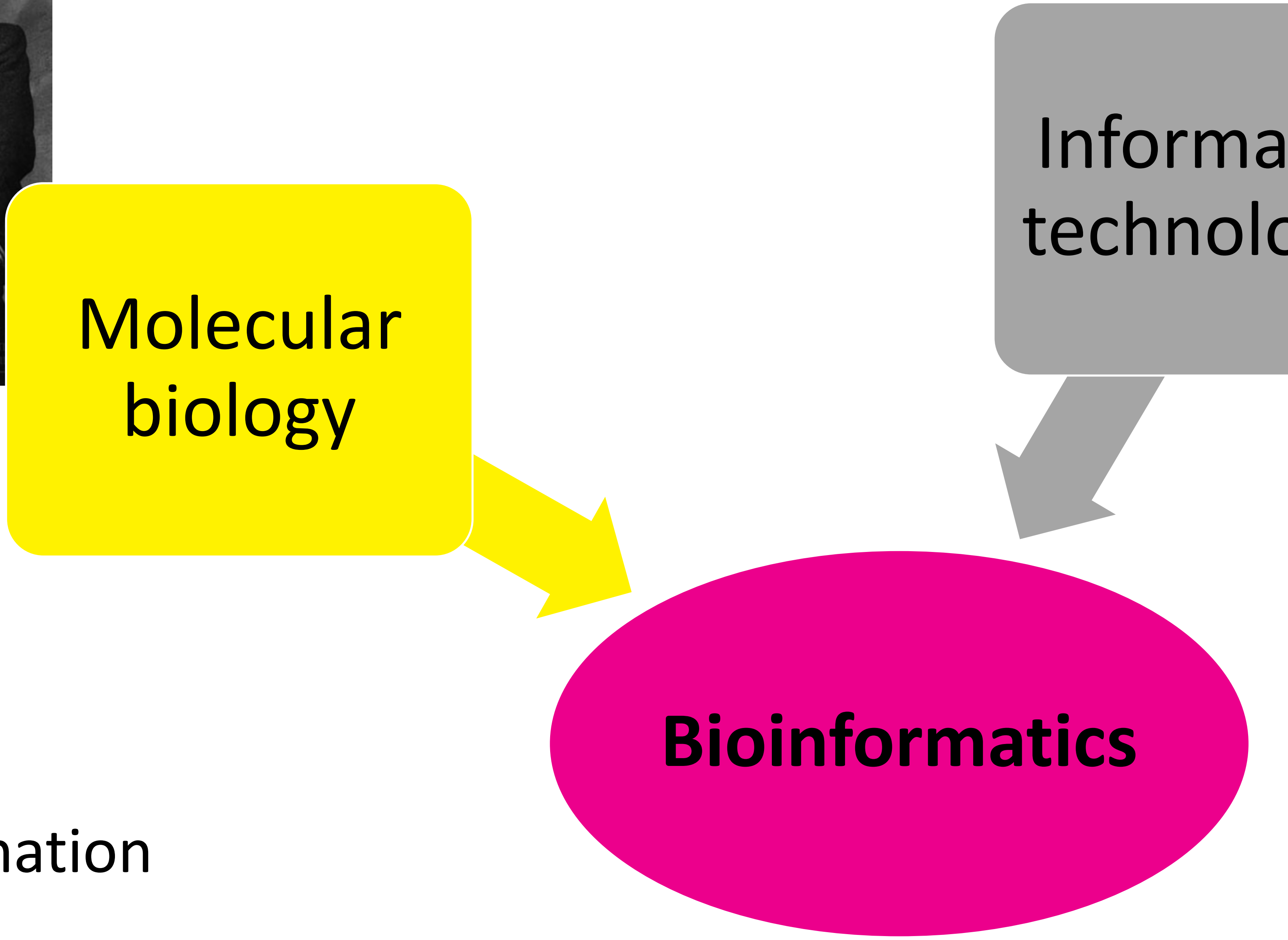
howest
university of applied sciences

Bioinformatics for dummies
MB&C2024 Workshop

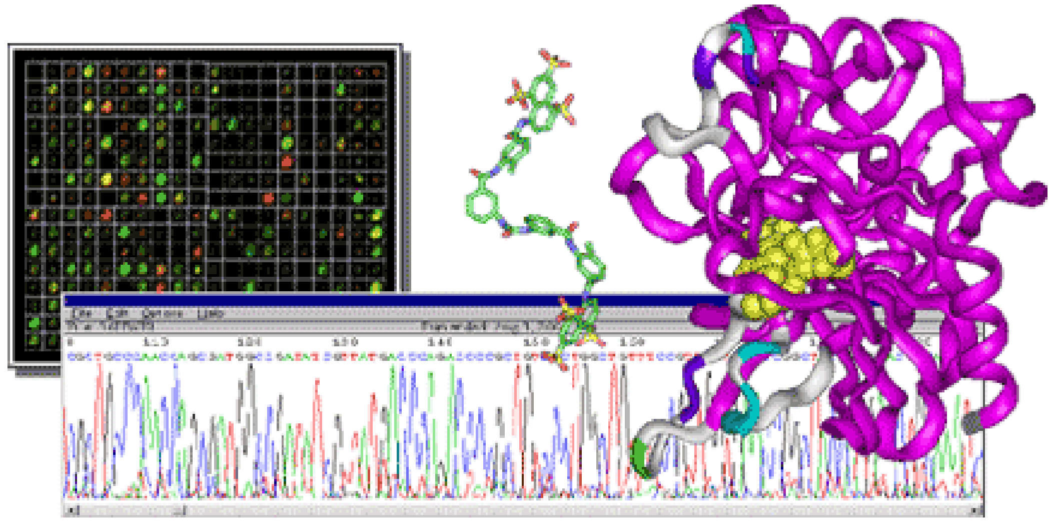
Cedric Hermans
Paco Hulpiau

Introduction



Combine:

- New insights and technologies in molecular biology
- Advances in information technologies



Introduction

Informatics

To store, organize and share molecular biological data in database systems

Bioinformatics

To process and analyse biological data by using bioinformatics tools in a “dry lab”

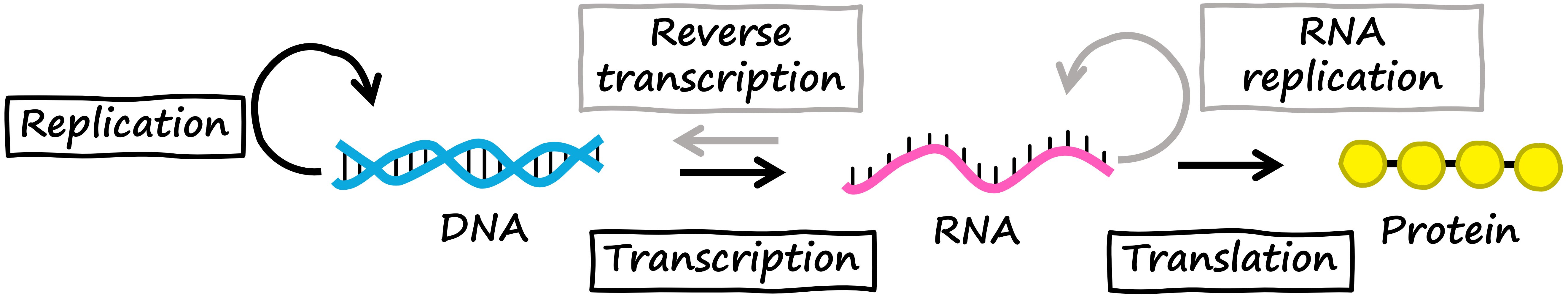
Programming

To integrate the different tools by means of scripting into a bioinformatics pipeline

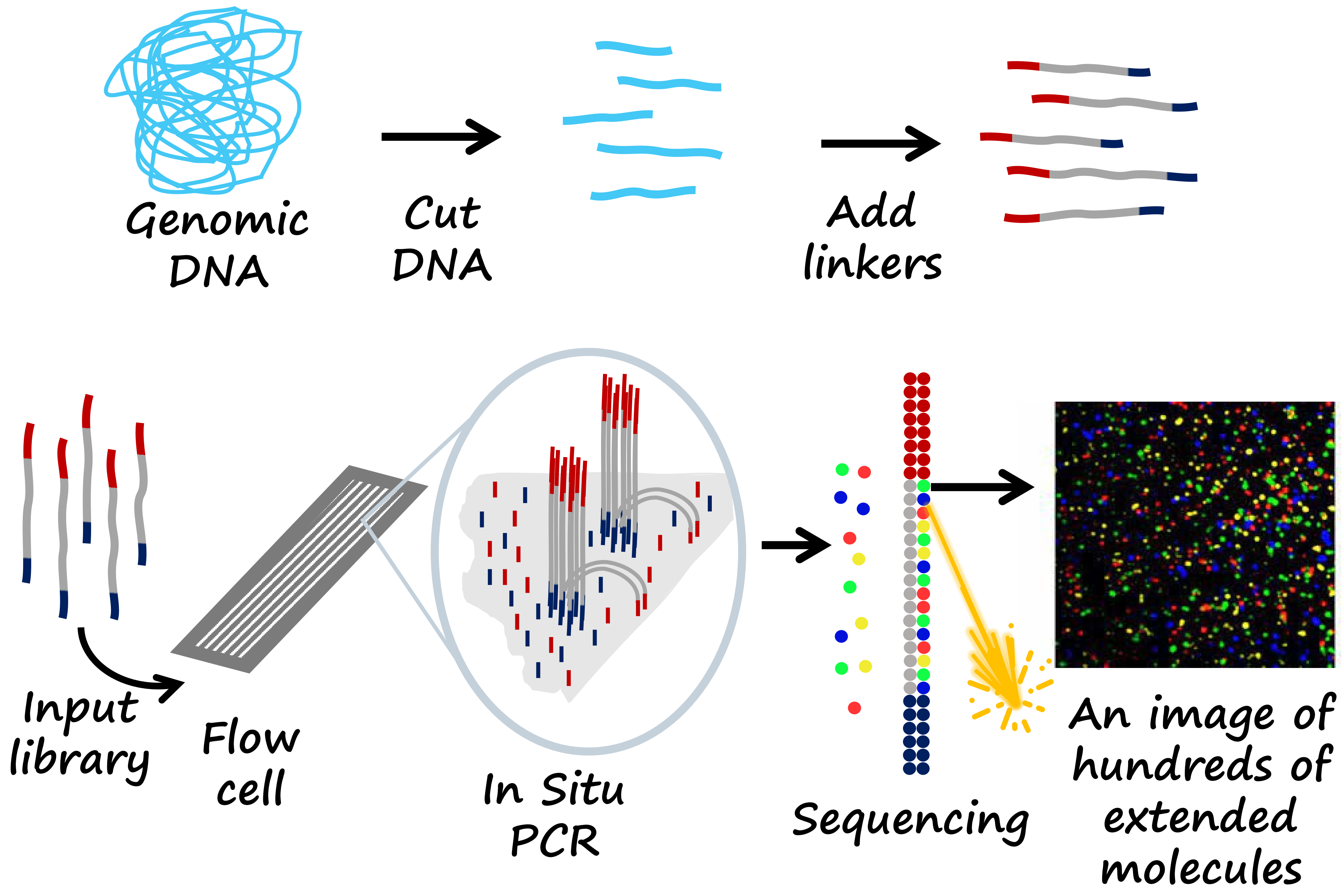
Molecular biology and bioinformatics

Important (high-throughput) technologies:

- Next Generation Sequencing
  - ➢ Sequencing and expression analysis
- Microarray
  - ➢ Expression and genetic variation analysis
- Mass spectrometry
  - ➢ Protein (sequence) identification



Next generation sequencing

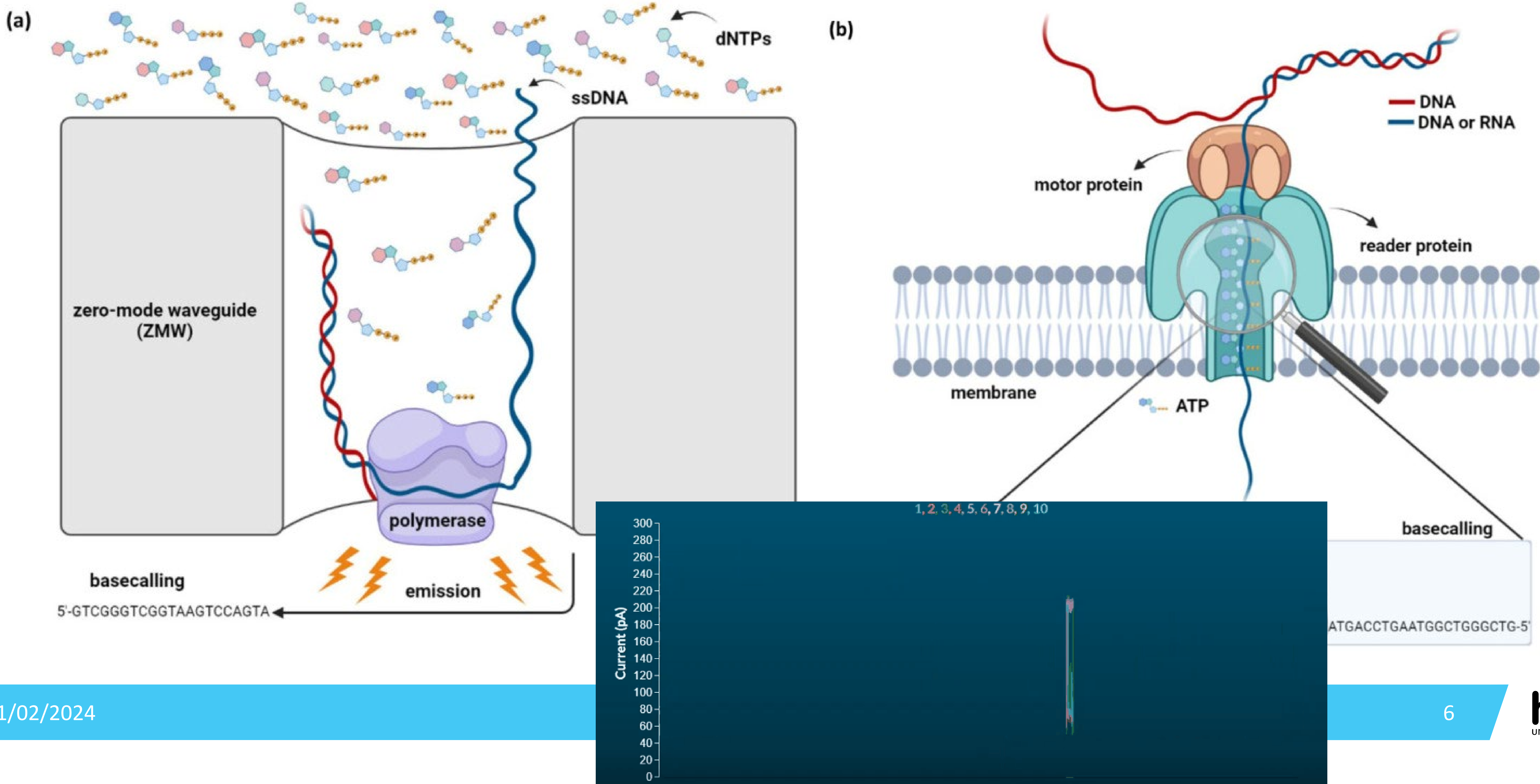


Short-read NGS

- 2 approaches:
  - ➢Sequencing by synthesis
  - ➢Sequencing by ligation
- 35-700 bp read length
- High accuracy (~ 99,99%)
- Complex assembly

Next next generation sequencing

Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A.
Third-Generation Sequencing: The Spearhead towards the Radical
Transformation of Modern Genomics. Life (Basel). 2021 Dec 26;12(1):30.
doi: 10.3390/life12010030. PMID: 35054423; PMCID: PMC8780579.



Microarrays

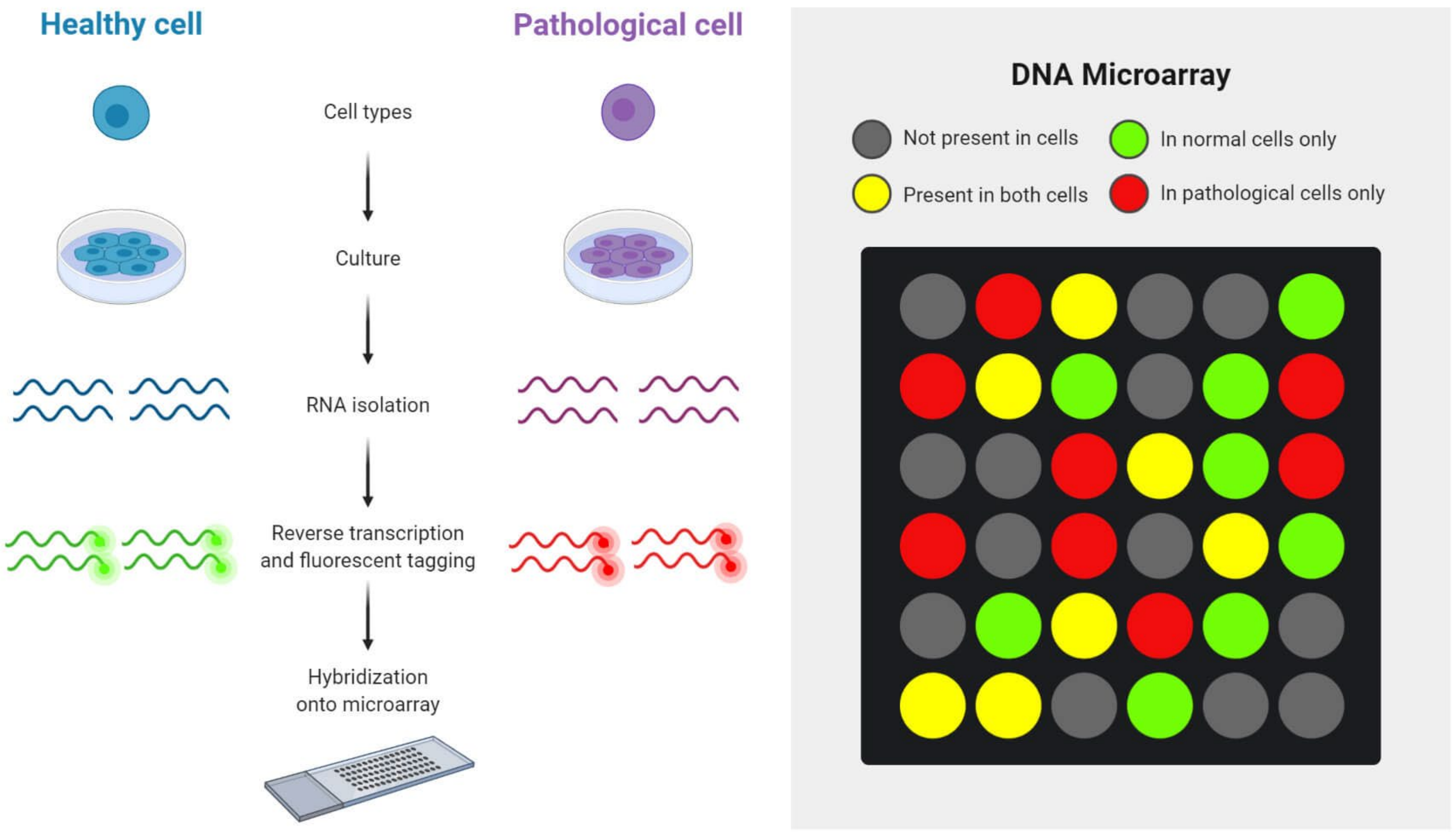
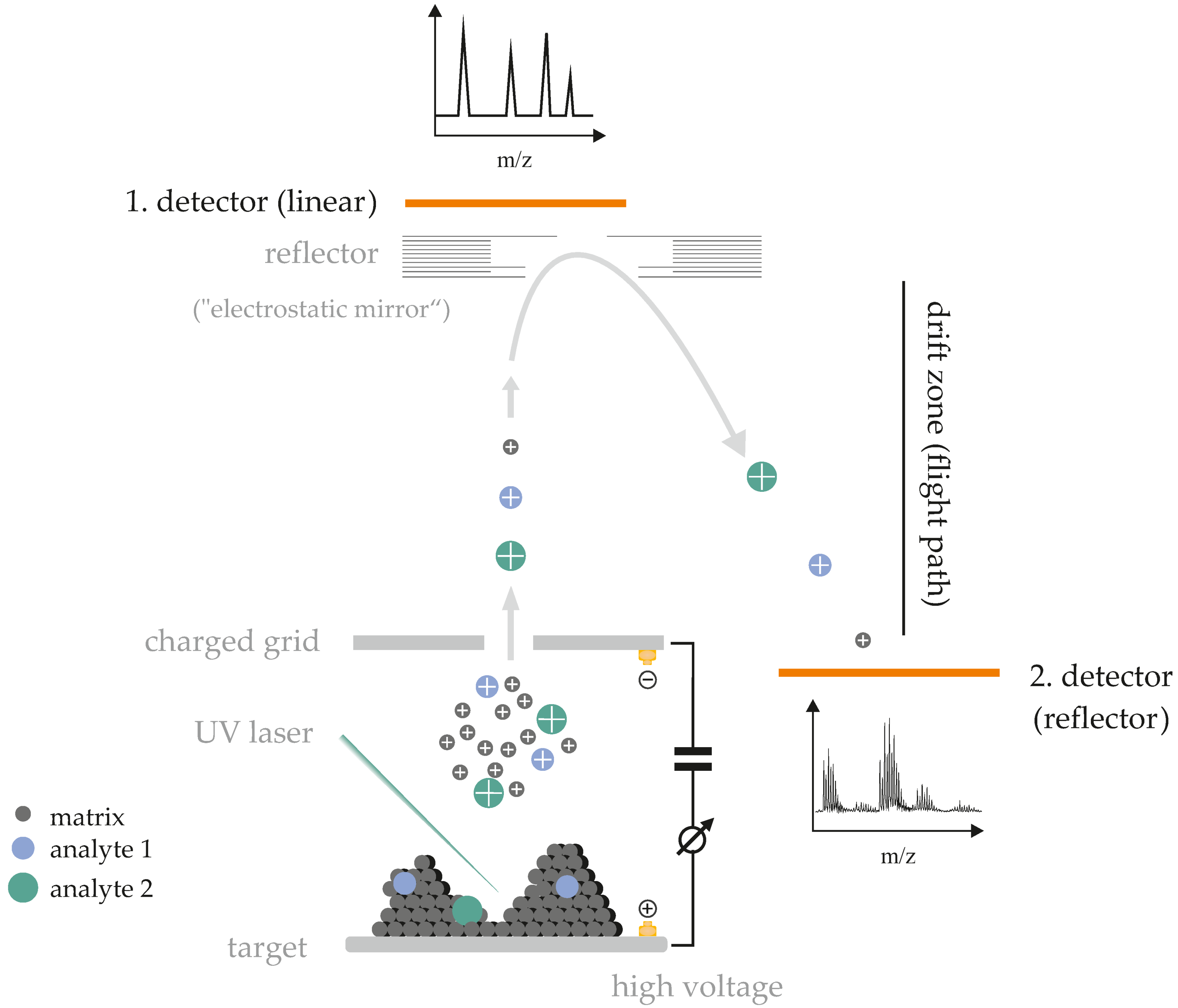


Image By Sagar Aryal, created using biorender.com

Leopold, J., Popkova, Y., Engel, K. M., & Schiller, J. (2018). Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Lipids. *Biomolecules*, *8*(4), 173. https://doi.org/10.3390/biom8040173

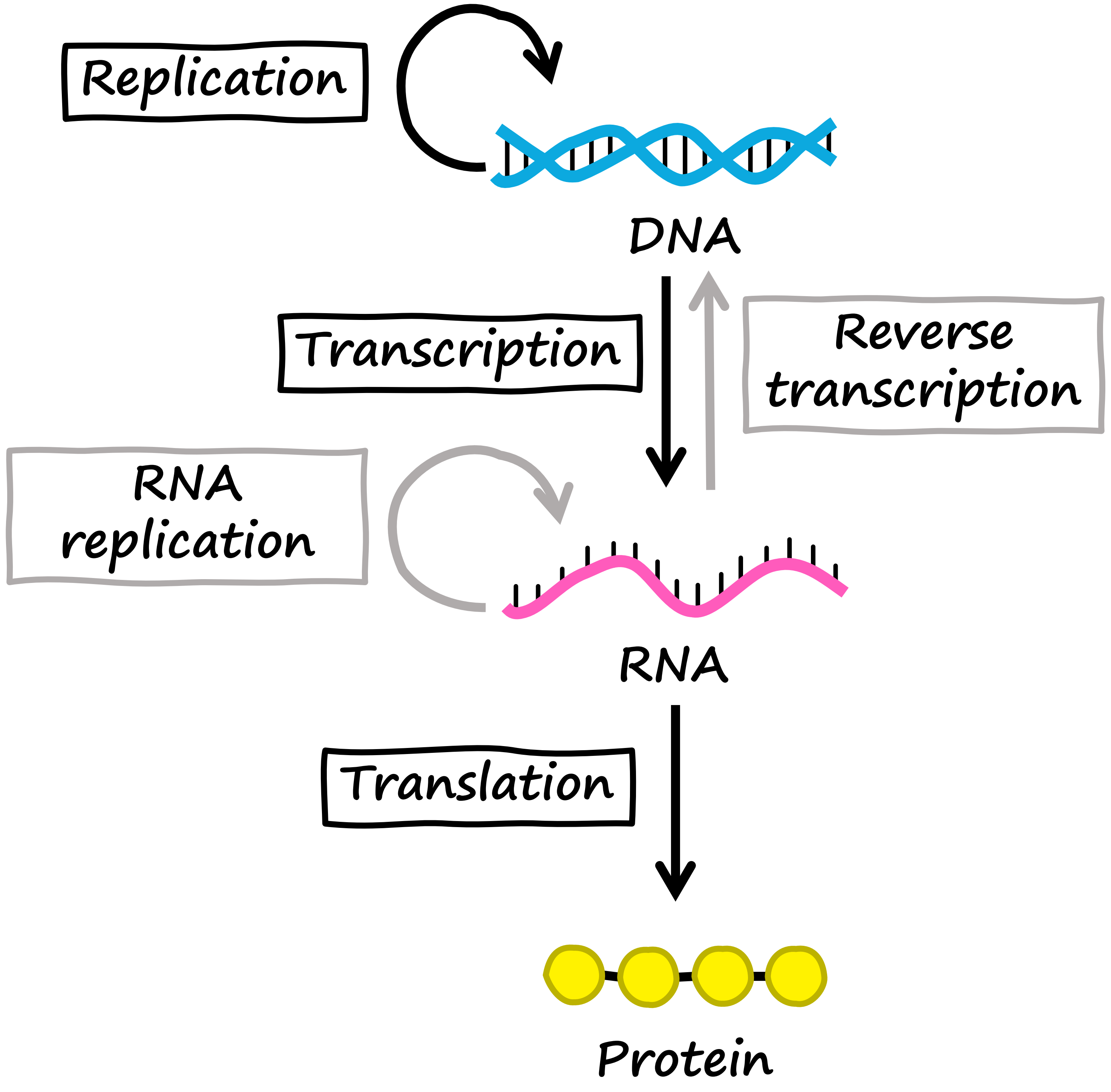
Mass spectrometry



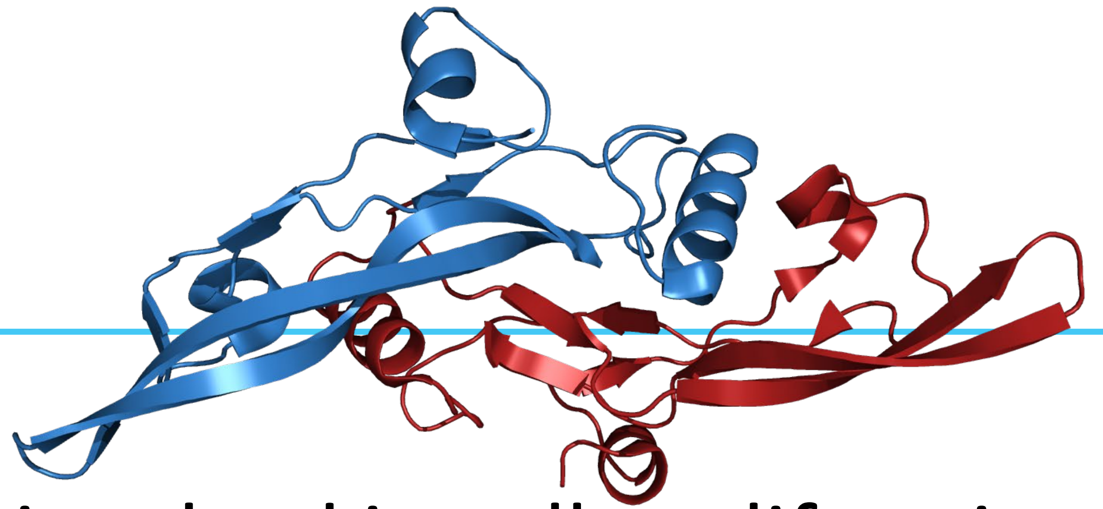
Molecular biology and bioinformatics

Biological databases:

- DNA
  - Sequence and loci
  - (Natural) genetic variation
- RNA
  - Transcripts (and variants)
  - Gene expression
- Protein
  - Sequence and function
  - Phenotype (and diseases)



(Sequence) repositories



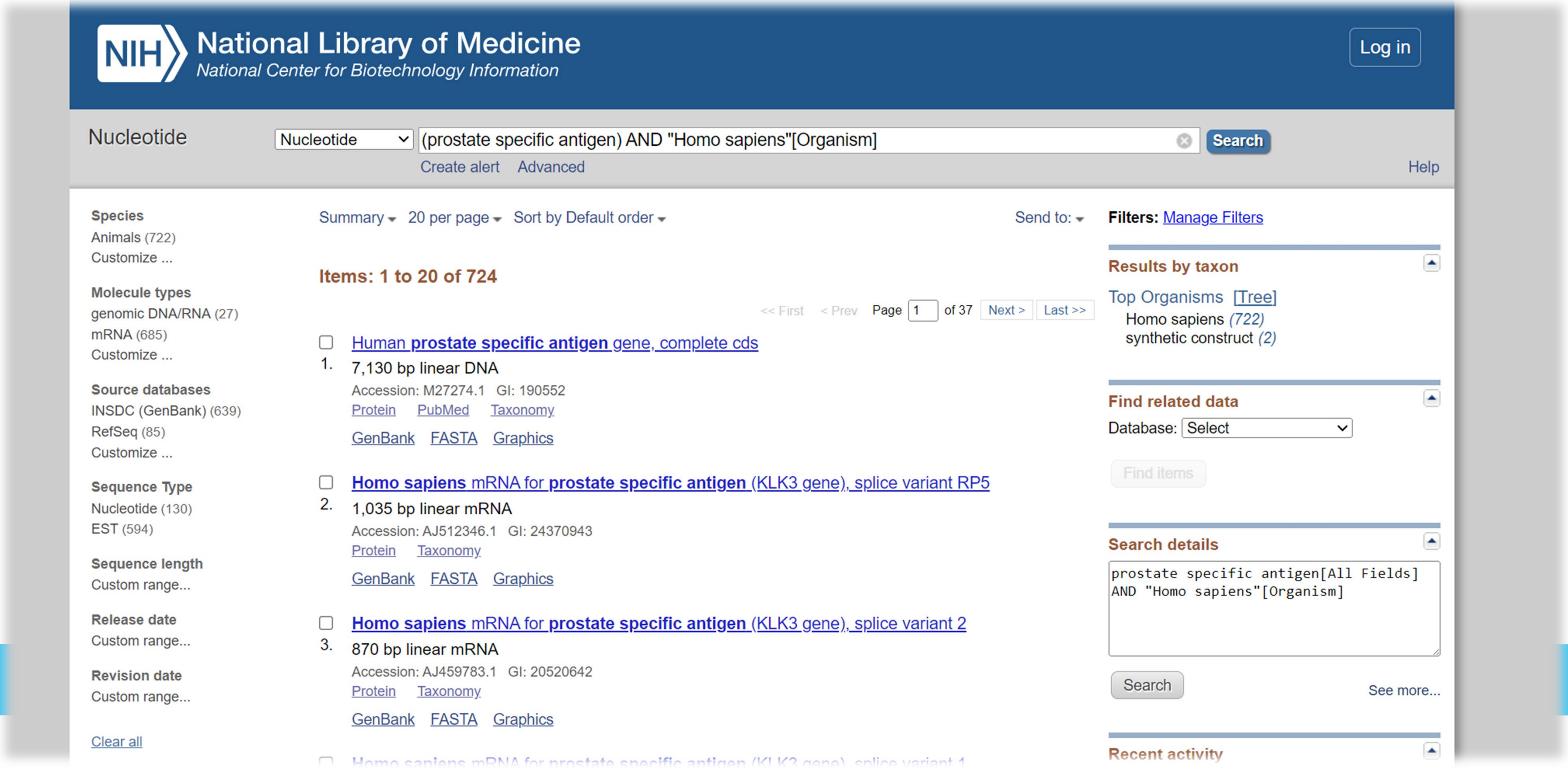
Exploratory example: **TGF beta 1** – an important protein involved in cell proliferation, differentiation and growth

NCBI Gene	https://www.ncbi.nlm.nih.gov/gene/7040 General and integrated sequence and locus information
NCBI Nucleotide	https://www.ncbi.nlm.nih.gov/nuccore/?term=TGFB1+AND+"Homo+sapiens"[Organism] All available (partial) TGF beta 1 nucleotide sequences → ± 138 records (!)
Ensembl	https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000105329 General information + detailed transcripts and gene expression
UniProt or NCBI Protein	http://www.uniprot.org/uniprot/P01137 High-quality recourse of protein sequence and functional information

(Sequence) repositories

Example 1: Look for the nucleotide sequence of PSA

- https://www.ncbi.nlm.nih.gov/nucleotide/
- NCBI nucleotide query: “(prostate specific antigen)” restricted to humans



(prostate specific antigen) AND "Homo sapiens"[Organism]

Species
Animals (722)
Customize ...

Molecule types
genomic DNA/RNA (27)
mRNA (685)
Customize ...

Source databases
INSDC (GenBank) (639)
RefSeq (85)
Customize ...

Sequence Type
Nucleotide (130)
EST (594)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Items: 1 to 20 of 724

Page 1 of 37

1. □ Human **prostate specific antigen** gene, complete cds
   7,130 bp linear DNA
   Accession: M27274.1 GI: 190552
   Protein PubMed Taxonomy
   GenBank FASTA Graphics

2. □ **Homo sapiens** mRNA for **prostate specific antigen** (KLK3 gene), splice variant RP5
   1,035 bp linear mRNA
   Accession: AJ512346.1 GI: 24370943
   Protein Taxonomy
   GenBank FASTA Graphics

3. □ **Homo sapiens** mRNA for **prostate specific antigen** (KLK3 gene), splice variant 2
   870 bp linear mRNA
   Accession: AJ459783.1 GI: 20520642
   Protein Taxonomy
   GenBank FASTA Graphics

□ Homo sapiens mRNA for prostate specific antigen (KLK3 gene), splice variant 1

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(722)*
synthetic construct *(2)*

Find related data

Database: Select

Search details

```
prostate specific antigen[All Fields]
AND "Homo sapiens"[Organism]
```

Recent activity

(Sequence) repositories

Example 2: the Genome Projects

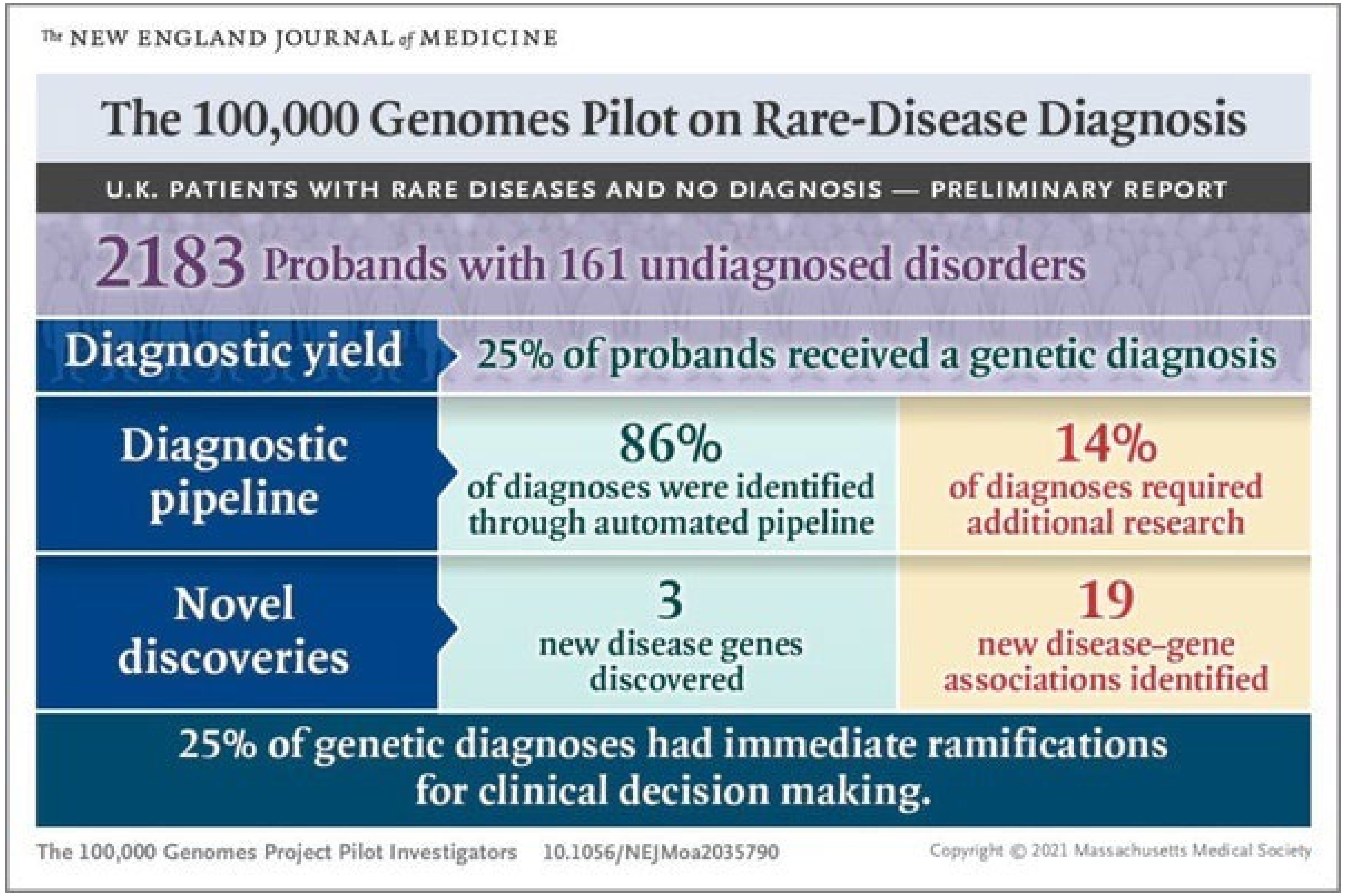
- 1000 Genomes Project (2008-2015)
  - Goal: to find most genetic variants with frequencies of at least 1% in the populations studied
- 100 000 Genomes Project (2013-2018)
  - Goal: focus on rare diseases, some common types of cancer and infectious diseases
- 1+ Million Genomes (2018-2027)

```
ACGTACGTACGTACGTACGTACGT
ACGTACCTACGTACGTACGTACGT
ACGTACCTACGTATGTTCGTACGT
ACGTACGTACGTATGTTCGTACGT
```

(Sequence) repositories

Example 2: the Genome Projects

- 100 000 Genomes Project (2013-2018)



The NEW ENGLAND JOURNAL of MEDICINE

The 100,000 Genomes Pilot on Rare-Disease Diagnosis

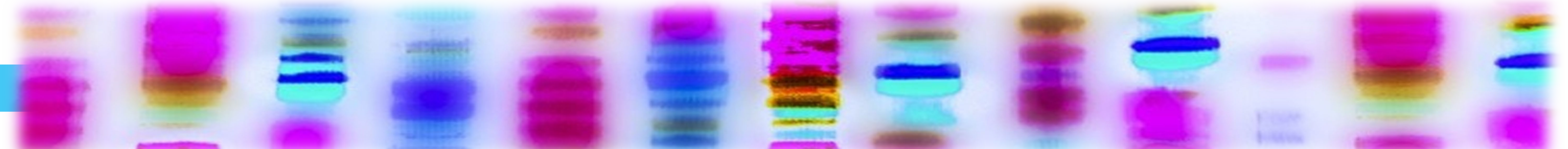
U.K. PATIENTS WITH RARE DISEASES AND NO DIAGNOSIS — PRELIMINARY REPORT

2183 Probands with 161 undiagnosed disorders

Diagnostic yield	25% of probands received a genetic diagnosis	
Diagnostic pipeline	86% of diagnoses were identified through automated pipeline	14% of diagnoses required additional research
Novel discoveries	3 new disease genes discovered	19 new disease–gene associations identified

25% of genetic diagnoses had immediate ramifications for clinical decision making.

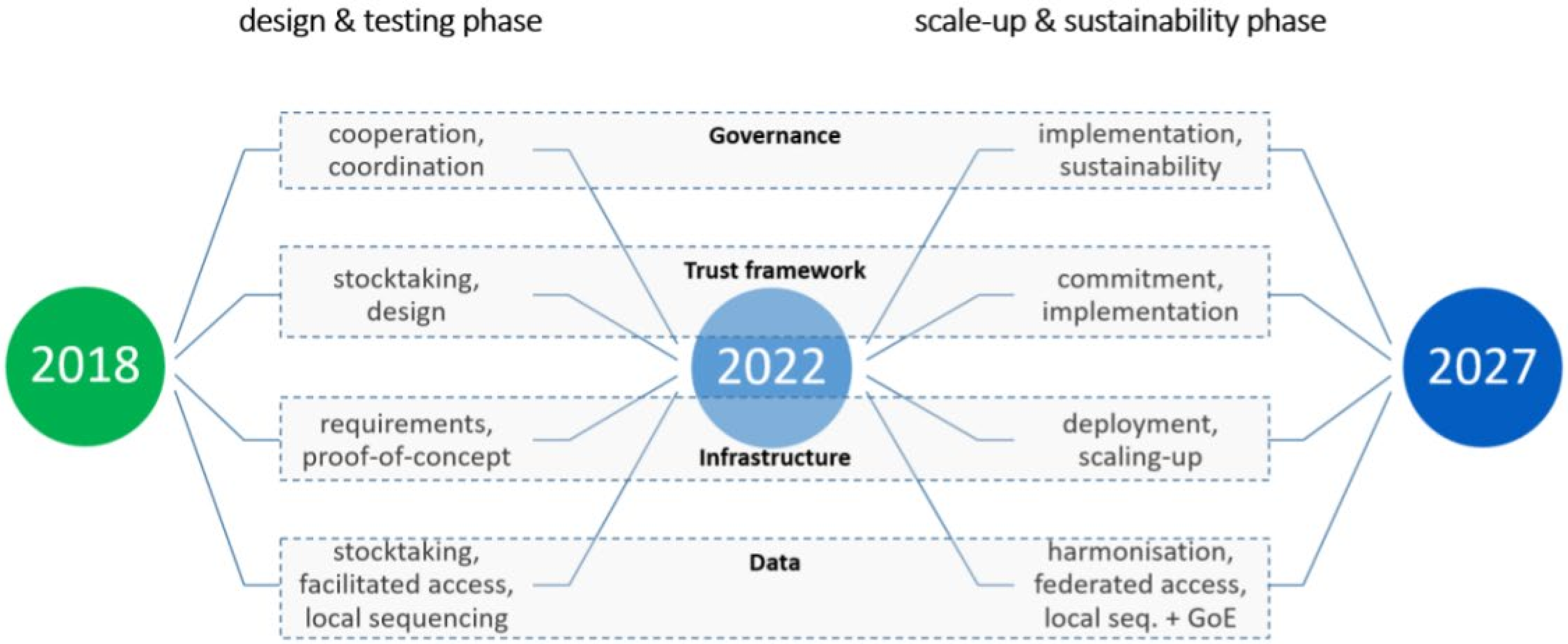
The 100,000 Genomes Project Pilot Investigators 10.1056/NEJMoa2035790



(Sequence) repositories

Example 2: the Genome Projects

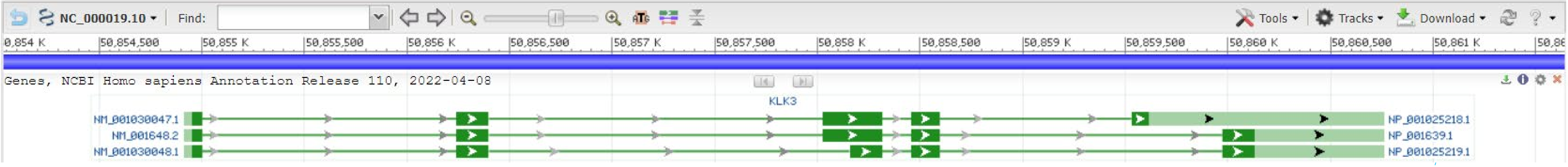
- 1+ Million Genomes


design & testing phase
scale-up & sustainability phase
2018
cooperation, coordination
stocktaking, design
requirements, proof-of-concept
stocktaking, facilitated access, local sequencing
Governance
Trust framework
2022
Infrastructure
Data
implementation, sustainability
commitment, implementation
deployment, scaling-up
harmonisation, federated access, local seq. + GoE
2027

(Sequence) repositories

Lots of diversity between genomes => solution is needed

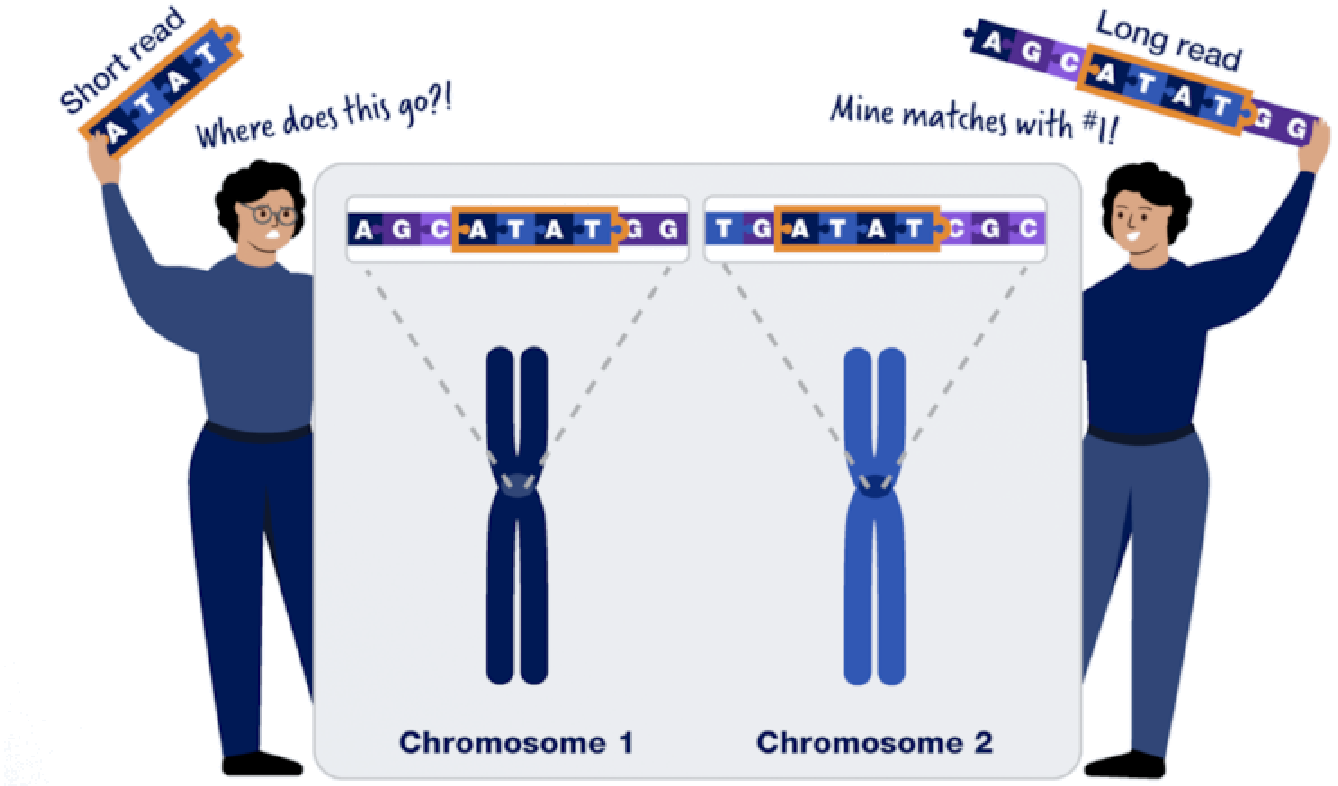
- **Genome Reference Consortium (GRC)**
  - Goal: create the best possible reference assembly for humans
  - -> latest major release: GRCh38 (also know as hg38) version 14
  - https://www.ncbi.nlm.nih.gov/grc/human
- **NCBI Reference Sequence Database (RefSeq)**
  - Non-redundant, well annotated set of reference sequences including genomes, transcripts and proteins
  - https://www.ncbi.nlm.nih.gov/refseq/
  - One gene/transcript/protein = one sequence

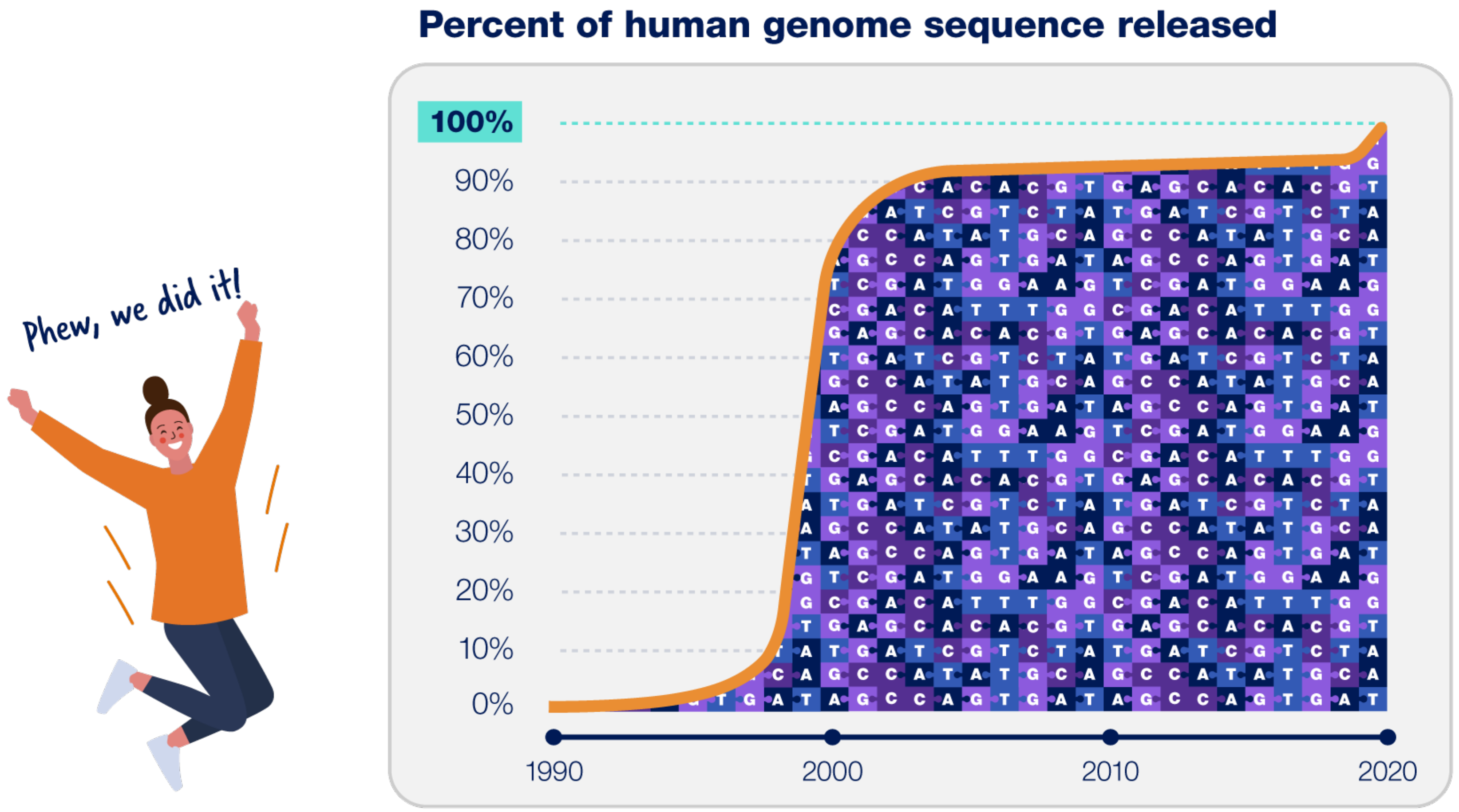


(Sequence) repositories

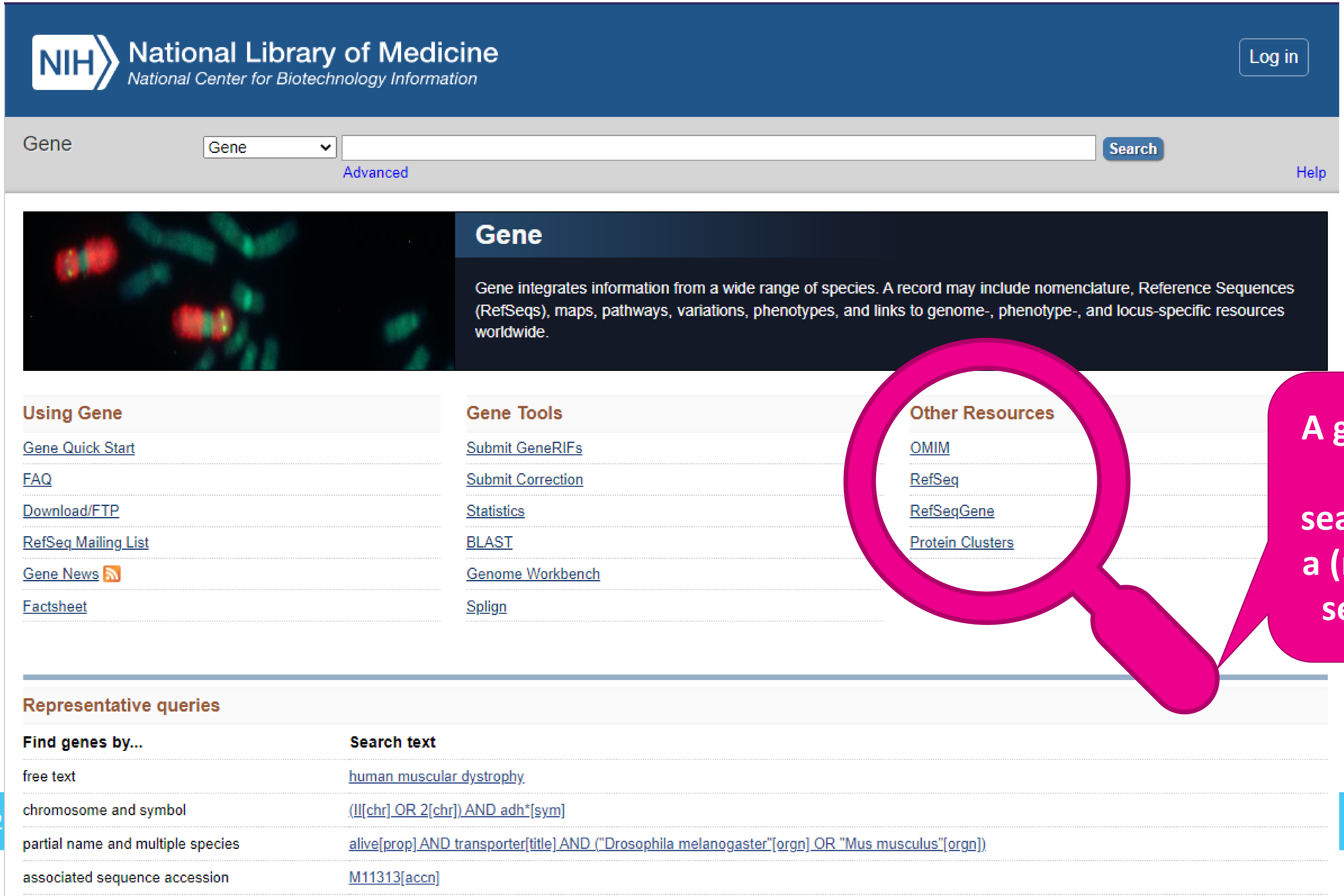
Note: GRCh38 is not complete…

- Telomere-to-Telomere (T2T) consortium: https://www.genome.gov/t2t
- Data: https://github.com/marbl/CHM13





(Sequence) repositories



NIH National Library of Medicine
National Center for Biotechnology Information

Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Using Gene

- Gene Quick Start
- FAQ
- Download/FTP
- RefSeq Mailing List
- Gene News
- Factsheet

Gene Tools

- Submit GeneRIFs
- Submit Correction
- Statistics
- BLAST
- Genome Workbench
- Splign

Other Resources

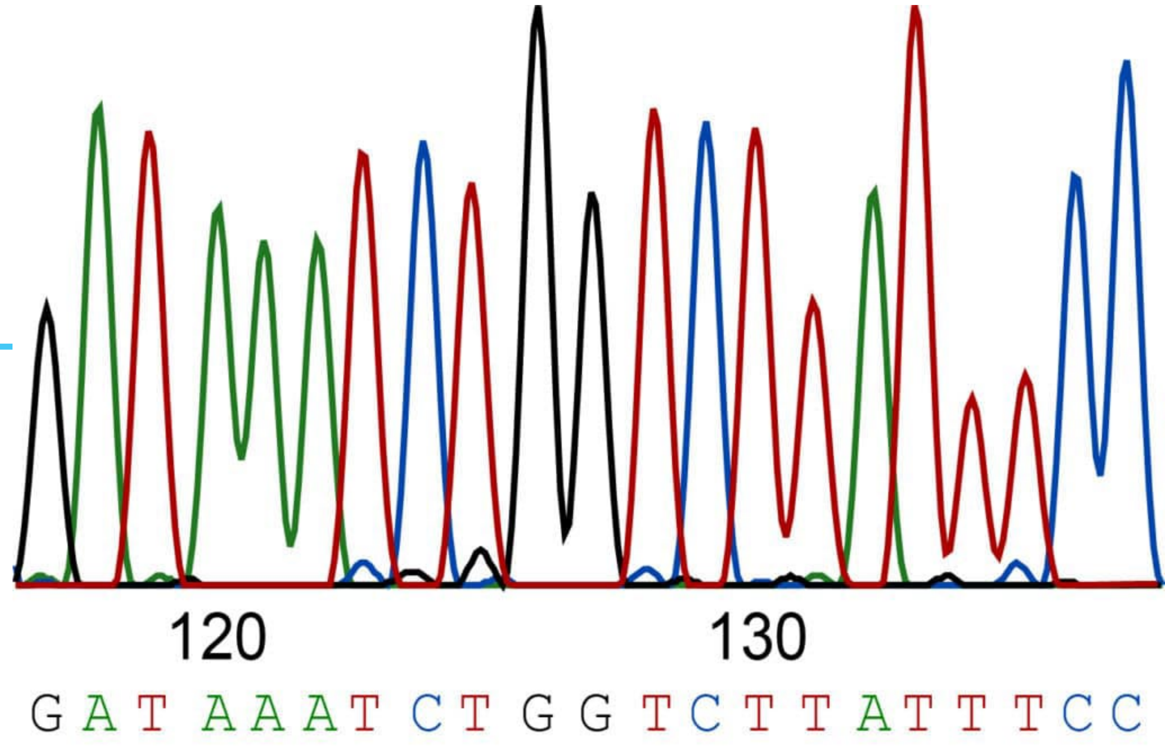
- OMIM
- RefSeq
- RefSeqGene
- Protein Clusters

Representative queries

Find genes by...	Search text
free text	human muscular dystrophy
chromosome and symbol	(II[chr] OR 2[chr]) AND adh*[sym]
partial name and multiple species	alive[prop] AND transporter[title] AND ("Drosophila melanogaster"[orgn] OR "Mus musculus"[orgn])
associated sequence accession	M11313[accn]

A good place to start searching for a (reference) sequence!

Homology searching

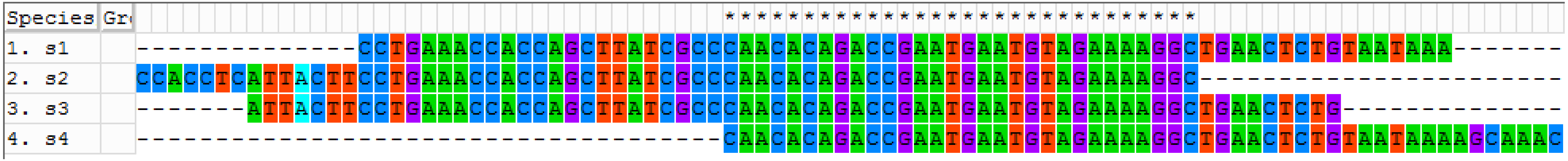


Next (next) Generation Sequencing

- Result: unknown nucleotide sequences

Determination of sequence ≠ simple keyword search strategy

⇒ Usage of <u>evolutionary models</u> to determine <u>homology</u> between nucleotide (or protein) sequences



```
Species Gr                                      ******************************
1. s1      --------------CCTGAAACCACCAGCTTATCGCCCAACACAGACCGAATGAATGTAGAAAAGGCTGAACTCTGTAATAAA-------
2. s2      CCACCTCATTACTTCCTGAAACCACCAGCTTATCGCCCAACACAGACCGAATGAATGTAGAAAAGGC-----------------------
3. s3      -------ATTACTTCCTGAAACCACCAGCTTATCGCCCAACACAGACCGAATGAATGTAGAAAAGGCTGAACTCTG--------------
4. s4      -------------------------------------CAACACAGACCGAATGAATGTAGAAAAGGCTGAACTCTGTAATAAAAGCAAAC
```

- Based on sequence alignment
- **BLAST**: <u>B</u>asic <u>L</u>ocal <u>A</u>lignment <u>S</u>earch <u>T</u>ool

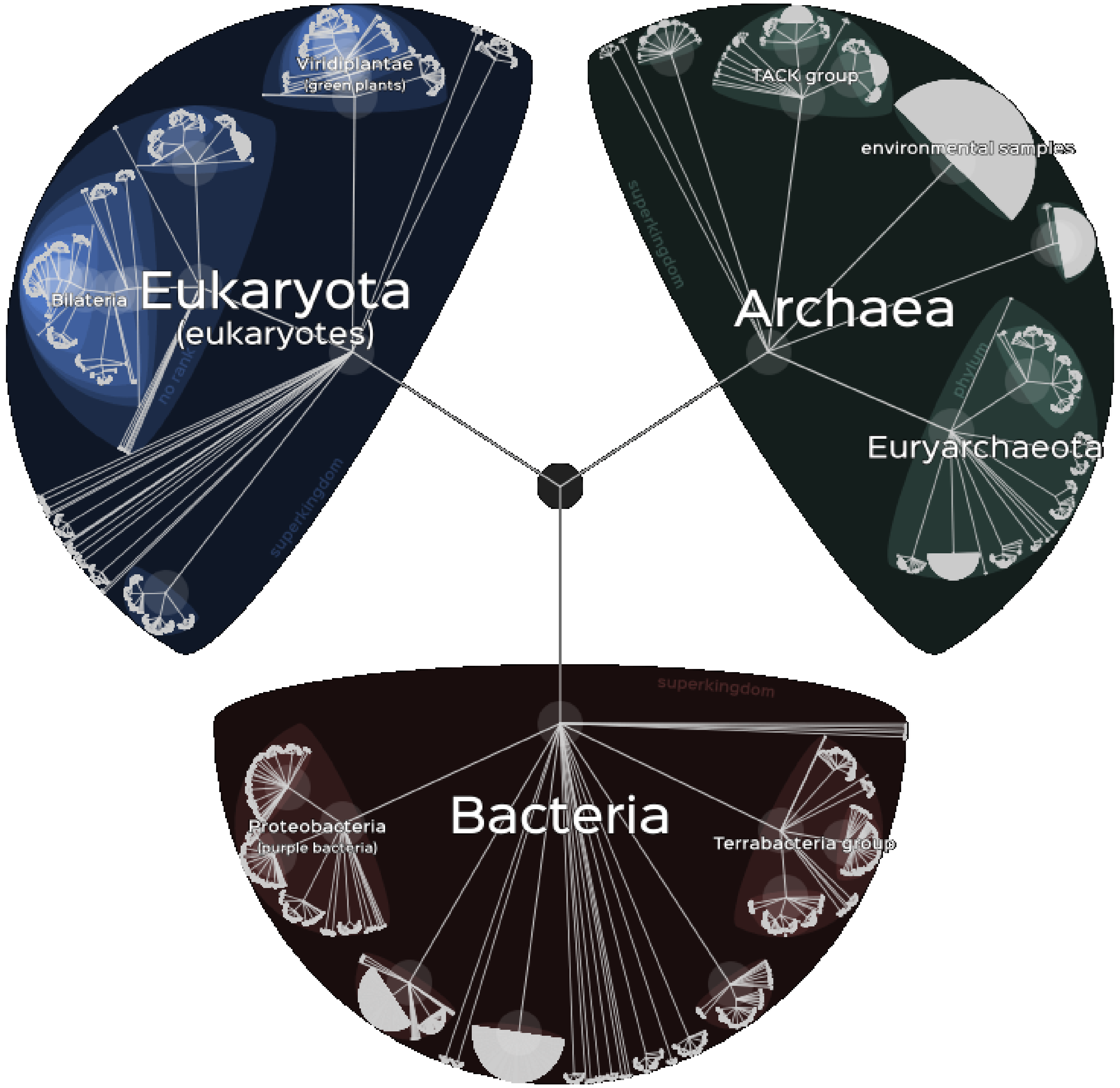
Homology searching

Homology

- Derived from a common ancestor
- 2 types:
  - Orthologs = speciation event (different species)
  - Paralogs = duplication event (same species)
- Typically based on morphological characteristics

- Use “molecular phylogeny” to determine homology

⇒ Phylogenetic tree



https://lifemap-ncbi.univ-lyon1.fr/#

Homology searching

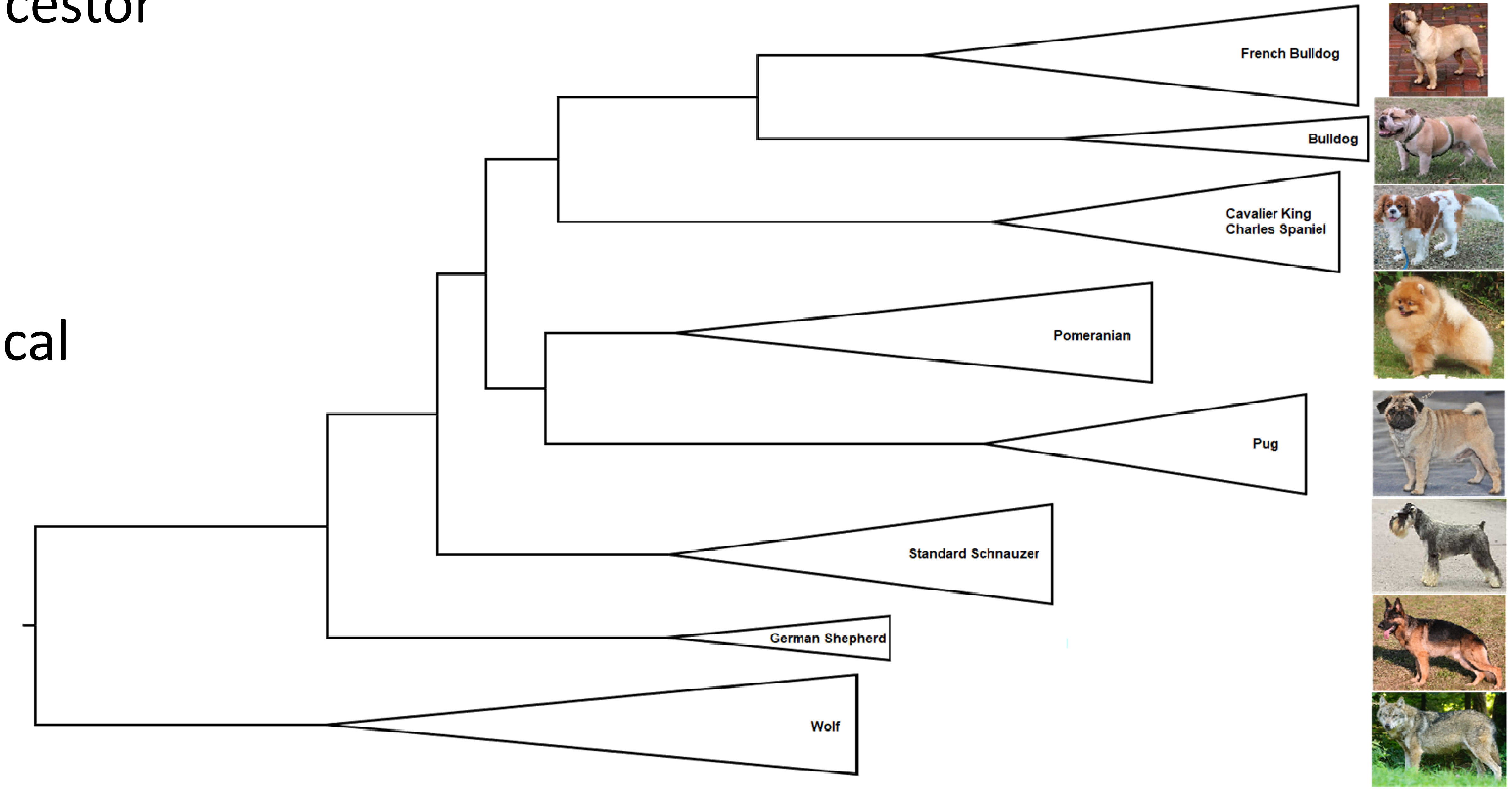
Homology

- Derived from a common ancestor
- 2 types:
  - Orthologs
  - Paralogs

Typically based on morphological characteristics

- Use "molecular phylogeny" to determine homology

⇒ Phylogenetic tree



Homology searching

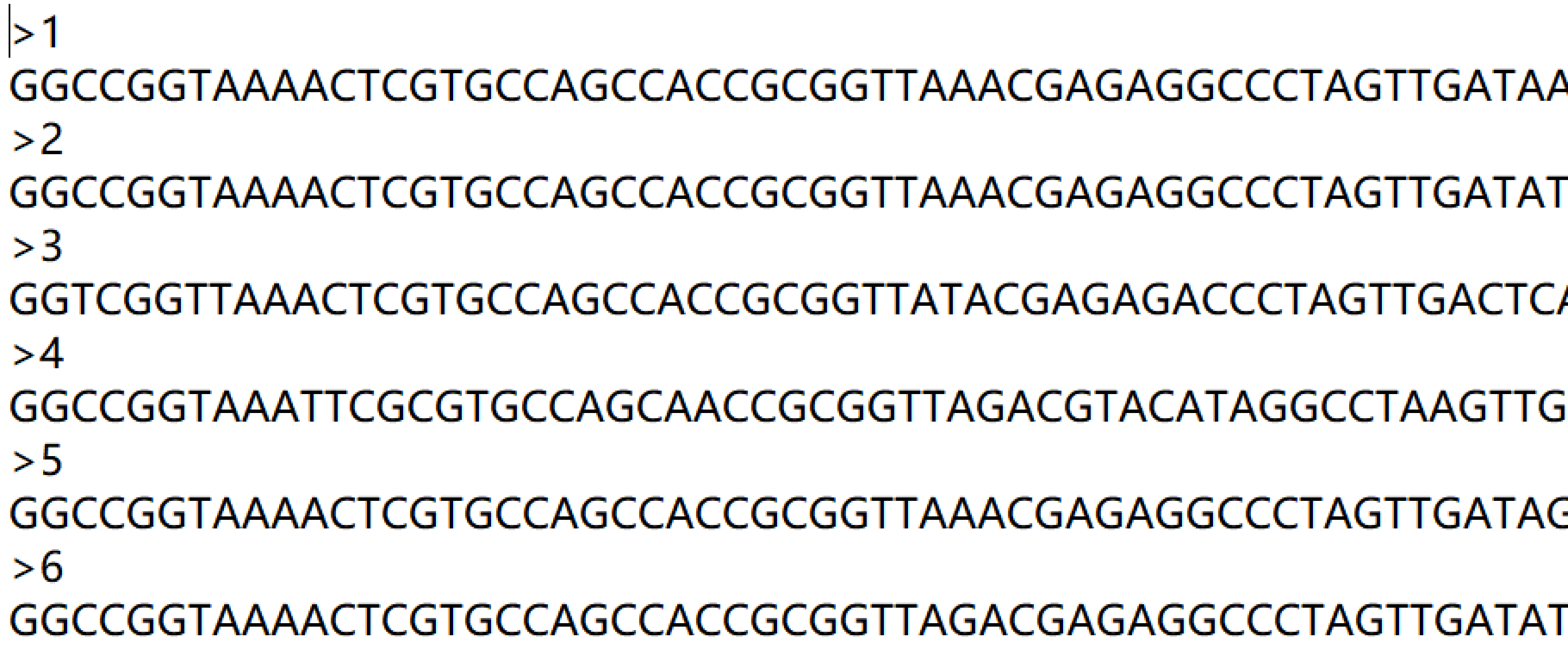
File format

Fasta-file:

- Header line starting with “>”
- One or more lines containing the sequence

Multifasta-file:

- Mutiple sequences in Fasta format below one another.
- A new sequence is recognized by the “>” in front of each header



```
>1
GGCCGGTAAAACTCGTGCCAGCCACCGCGGTTAAACGAGAGGCCCTAGTTGATAA
>2
GGCCGGTAAAACTCGTGCCAGCCACCGCGGTTAAACGAGAGGCCCTAGTTGATAT
>3
GGTCGGTTAAACTCGTGCCAGCCACCGCGGTTATACGAGAGACCCTAGTTGACTCA
>4
GGCCGGTAAATTCGCGTGCCAGCAACCGCGGTTAGACGTACATAGGCCTAAGTTG
>5
GGCCGGTAAAACTCGTGCCAGCCACCGCGGTTAAACGAGAGGCCCTAGTTGATAG
>6
GGCCGGTAAAACTCGTGCCAGCCACCGCGGTTAGACGAGAGGCCCTAGTTGATAT
```

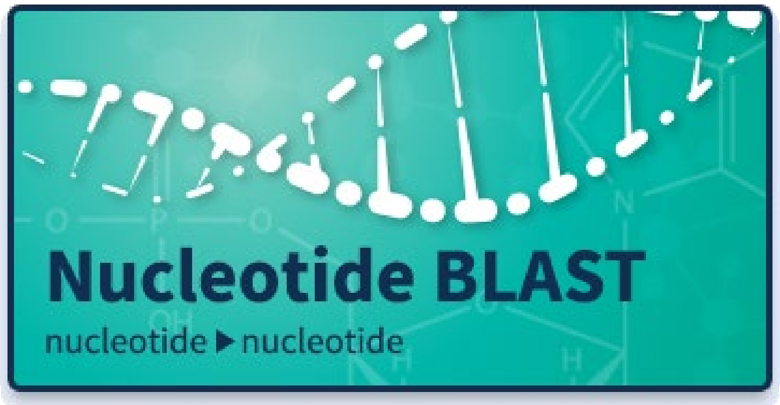
Homology searching

```
>unknown human nucleotide sequence
CAAGGCTGTCCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGCATCTCTCGGCAGTTCGTCTGTGACTC
AGACCGGGACTGCTTGGACGGCTCAGACGAGGCCTCCTGCCCGGTGCTCACCTGTGGTCCCGCCAGCTTCCAGTGCAACAGCTCCA
CCTGCATCCCCCAGCTGTGGGCCTGCGACAAC
```

- Given: an unknown human nucleotide Fasta sequence

-> https://www.bio-informatica.be/workshops/

> "unknown human nucleotide sequence.fasta"

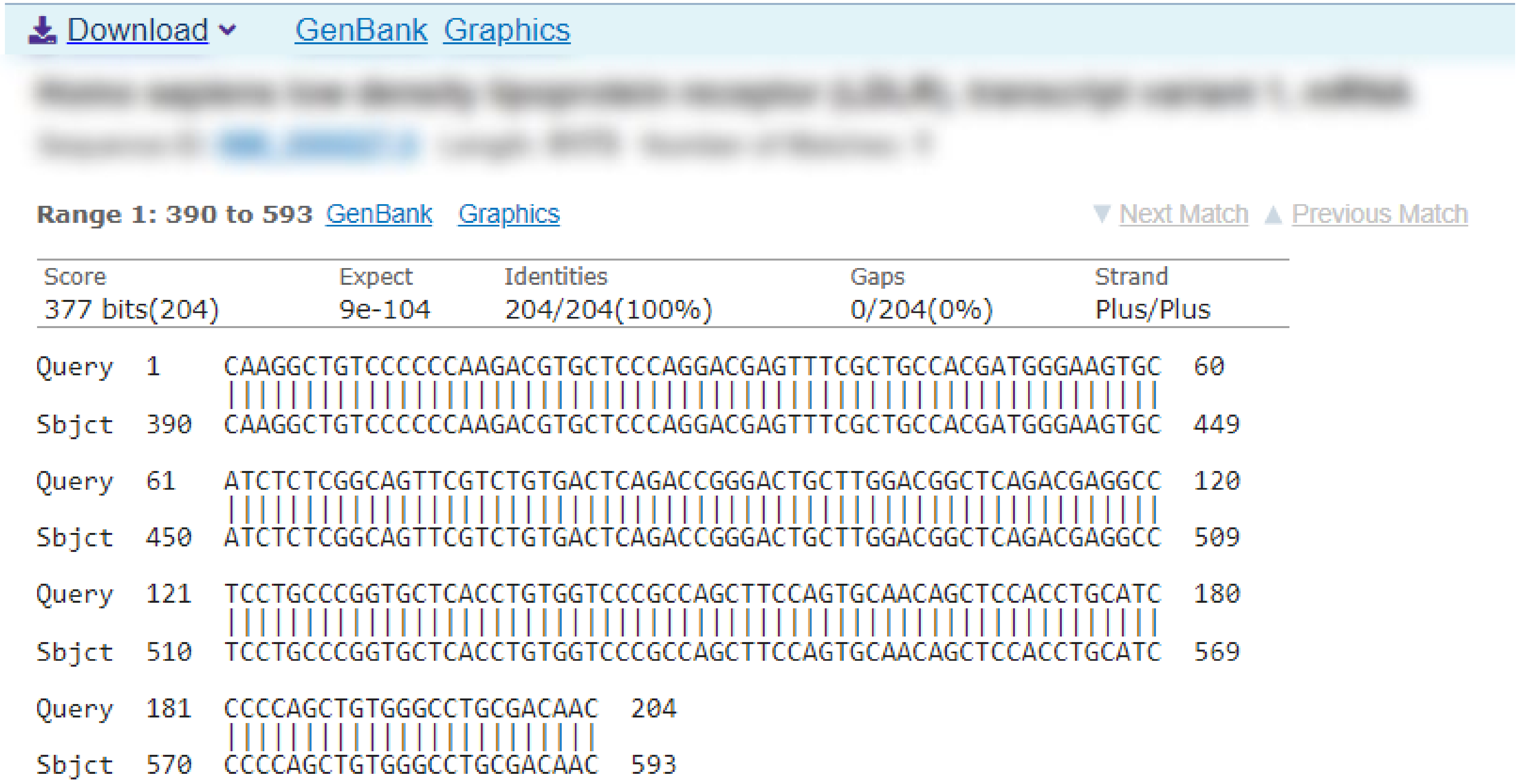


- To determine the identity -> use BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
  - Settings:
    - Organism: *Homo sapiens*
    - Database: refseq_rna
    - Exclude: models

Homology searching

Results:

- Identity
- Bits score
- Expect value
- Gaps



[illegible]

[illegible]

Range 1: 390 to 593 GenBank Graphics

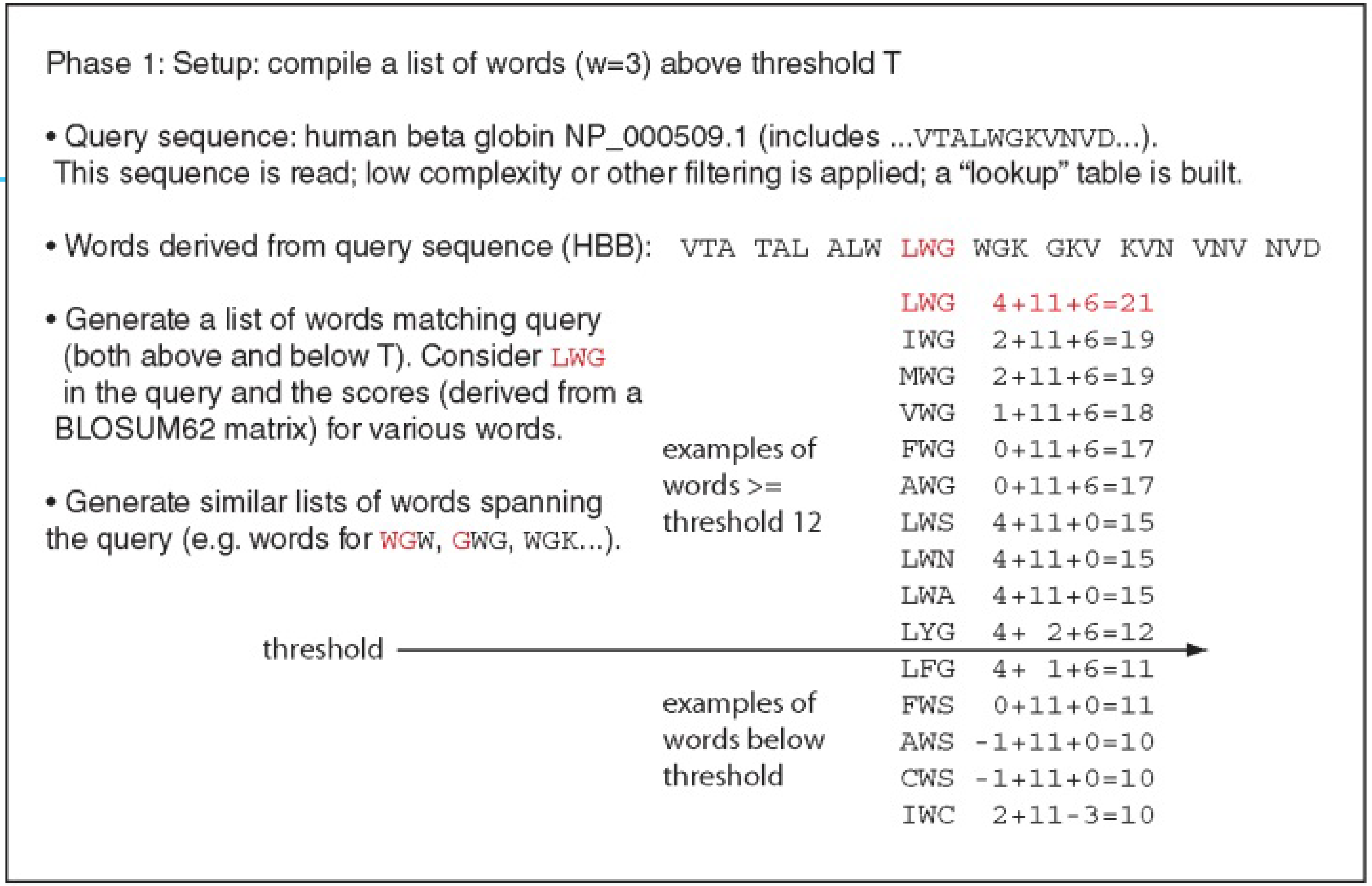
Score	Expect	Identities	Gaps	Strand
377 bits(204)	9e-104	204/204(100%)	0/204(0%)	Plus/Plus

```
Query  1    CAAGGCTGTCCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  390  CAAGGCTGTCCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGC  449

Query  61   ATCTCTCGGCAGTTCGTCTGTGACTCAGACCGGGACTGCTTGGACGGCTCAGACGAGGCC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  450  ATCTCTCGGCAGTTCGTCTGTGACTCAGACCGGGACTGCTTGGACGGCTCAGACGAGGCC  509

Query  121  TCCTGCCCGGTGCTCACCTGTGGTCCCGCCAGCTTCCAGTGCAACAGCTCCACCTGCATC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  510  TCCTGCCCGGTGCTCACCTGTGGTCCCGCCAGCTTCCAGTGCAACAGCTCCACCTGCATC  569

Query  181  CCCCAGCTGTGGGCCTGCGACAAC  204
            ||||||||||||||||||||||||
Sbjct  570  CCCCAGCTGTGGGCCTGCGACAAC  593
```

Homology searching

BLAST

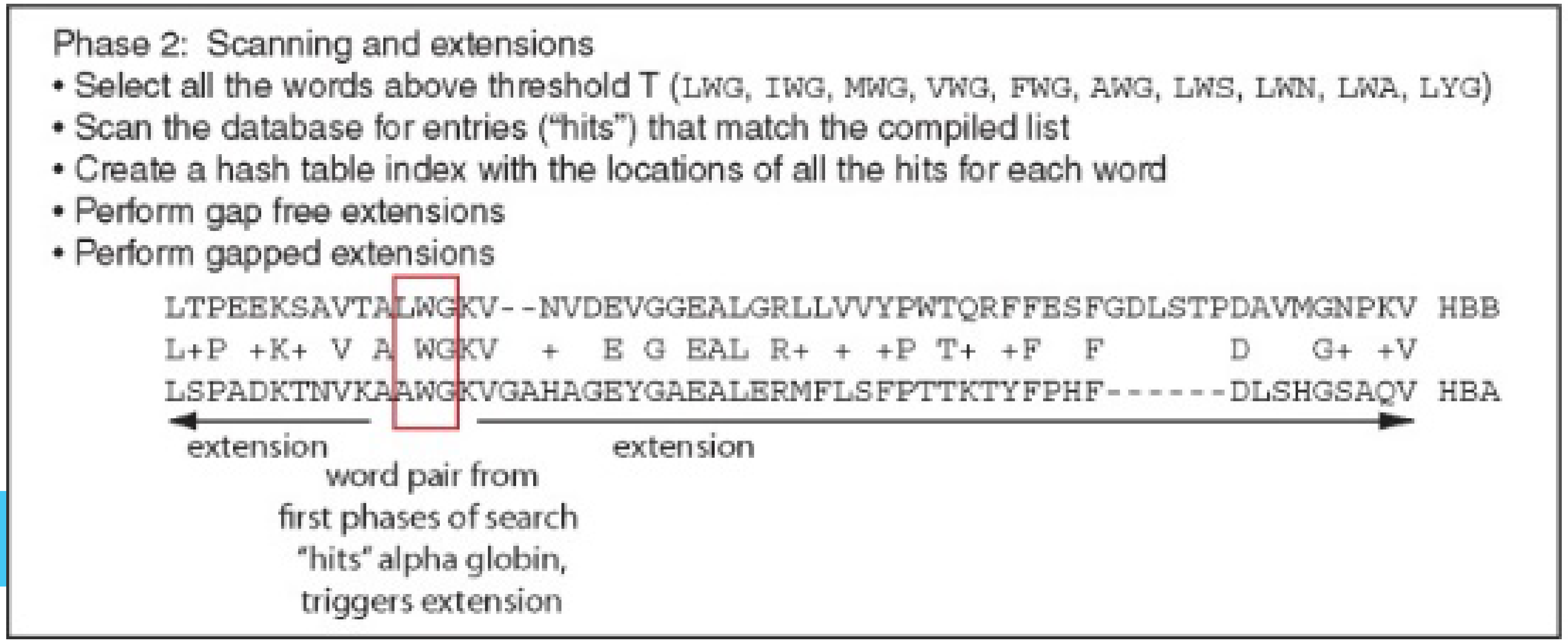
- Not a simple keyword search strategy
- 3 steps
  - LIST
  - SCAN
  - EXTEND
- Based on a model of evolution and scoring system



Phase 1: Setup: compile a list of words (w=3) above threshold T

- Query sequence: human beta globin NP_000509.1 (includes ...VTALWGKVNVD...). This sequence is read; low complexity or other filtering is applied; a "lookup" table is built.
- Words derived from query sequence (HBB): VTA TAL ALW LWG WGK GKV KVN VNV NVD
- Generate a list of words matching query (both above and below T). Consider LWG in the query and the scores (derived from a BLOSUM62 matrix) for various words.
- Generate similar lists of words spanning the query (e.g. words for WGW, GWG, WGK...).

examples of words >= threshold 12	LWG	4+11+6=21
	IWG	2+11+6=19
	MWG	2+11+6=19
	VWG	1+11+6=18
	FWG	0+11+6=17
	AWG	0+11+6=17
	LWS	4+11+0=15
	LWN	4+11+0=15
	LWA	4+11+0=15
	LYG	4+ 2+6=12
threshold		
examples of words below threshold	LFG	4+ 1+6=11
	FWS	0+11+0=11
	AWS	-1+11+0=10
	CWS	-1+11+0=10
	IWC	2+11-3=10



Phase 2: Scanning and extensions

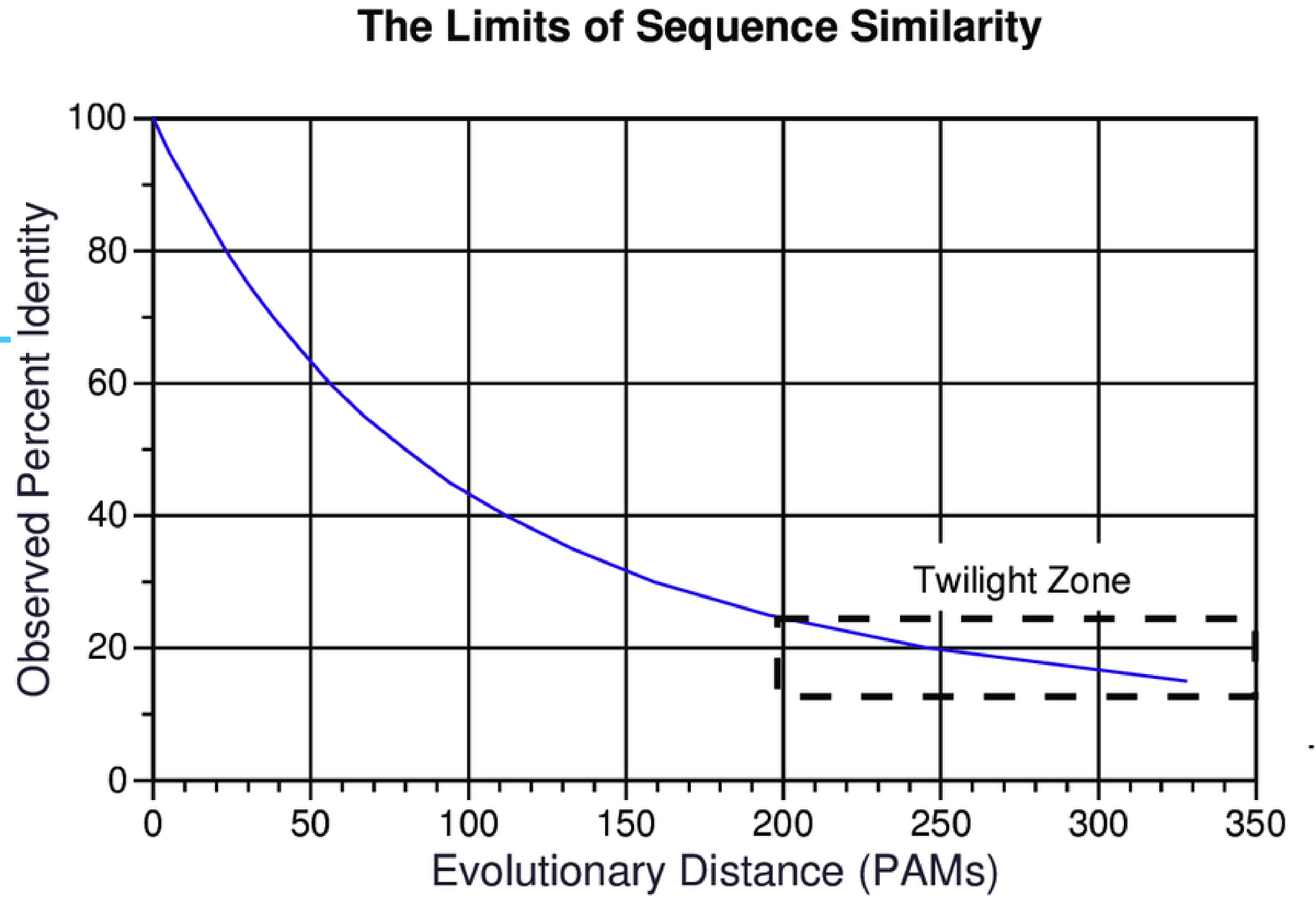
- Select all the words above threshold T (LWG, IWG, MWG, VWG, FWG, AWG, LWS, LWN, LWA, LYG)
- Scan the database for entries ("hits") that match the compiled list
- Create a hash table index with the locations of all the hits for each word
- Perform gap free extensions
- Perform gapped extensions

```
LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV HBB
L+P +K+ V A WGKV   + E G EAL R+ + +P T+ +F  F      D   G+ +V
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV HBA
```

extension
extension
word pair from first phases of search "hits" alpha globin, triggers extension

Homology searching

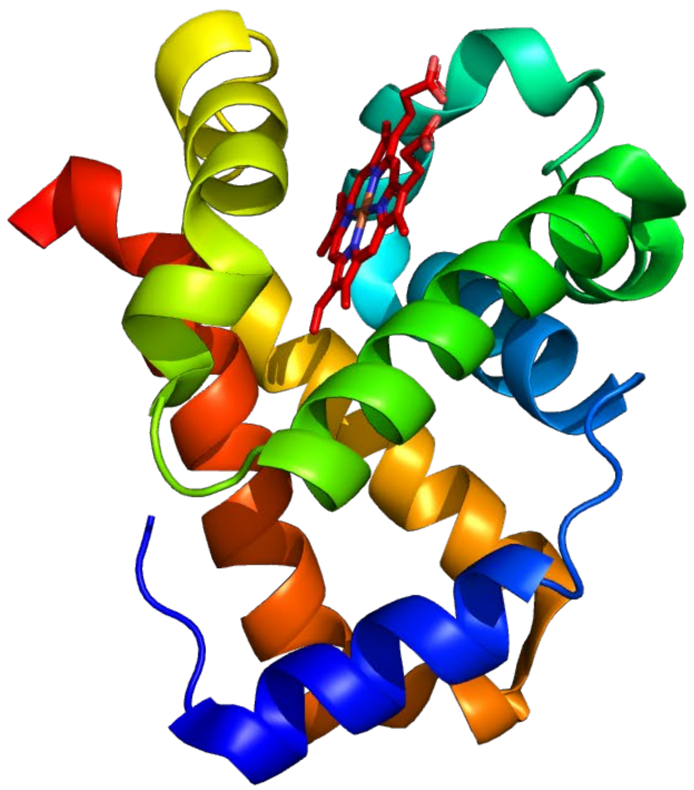
- Are two sequences homologues?
  - Look at percent identity (quantitative) + expect value
- Problem: homology = YES/NO question



Example case:

Is it possible to predict that human **myoglobin** (NP_05359) and **beta hemoglobin** (NP_000509) are paralogs?

Myoglobin



Hemoglobin

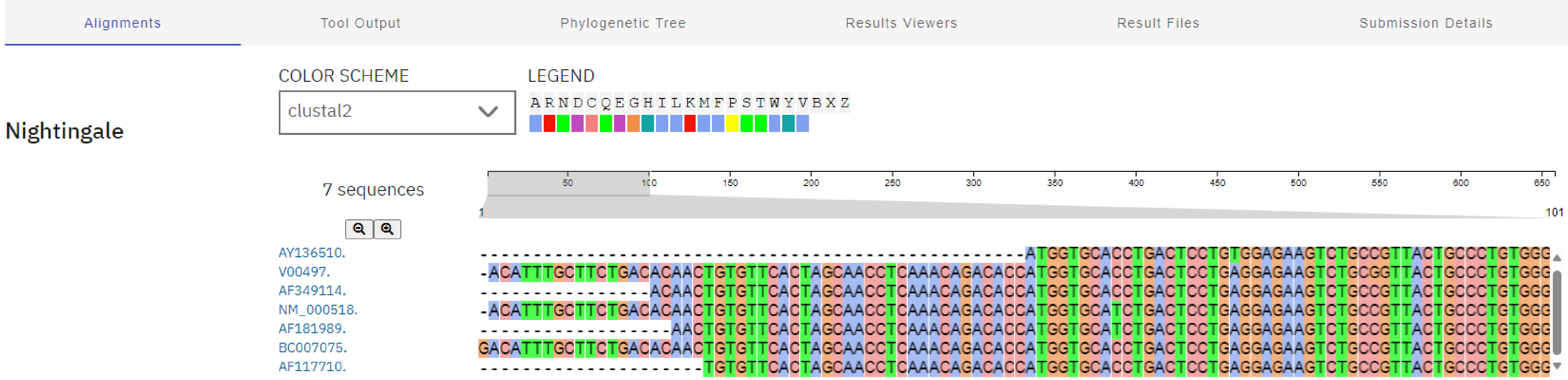
DNA variant analysis

Compare nucleotide sequence with a reference sequence

- Nucleotide diversity -> DNA variant identification
- Example: nucleotide diversity in multiple hemoglobin beta variants
  - https://www.bio-informatica.be/workshops/
    > "HBB multiple sequence alignment.fasta"
  - Align sequences using MUSCLE software (Muscle < EMBL-EBI)
    - -> output: HTML

Multiple sequence alignment (MSA)

-> phylogenetic analysis



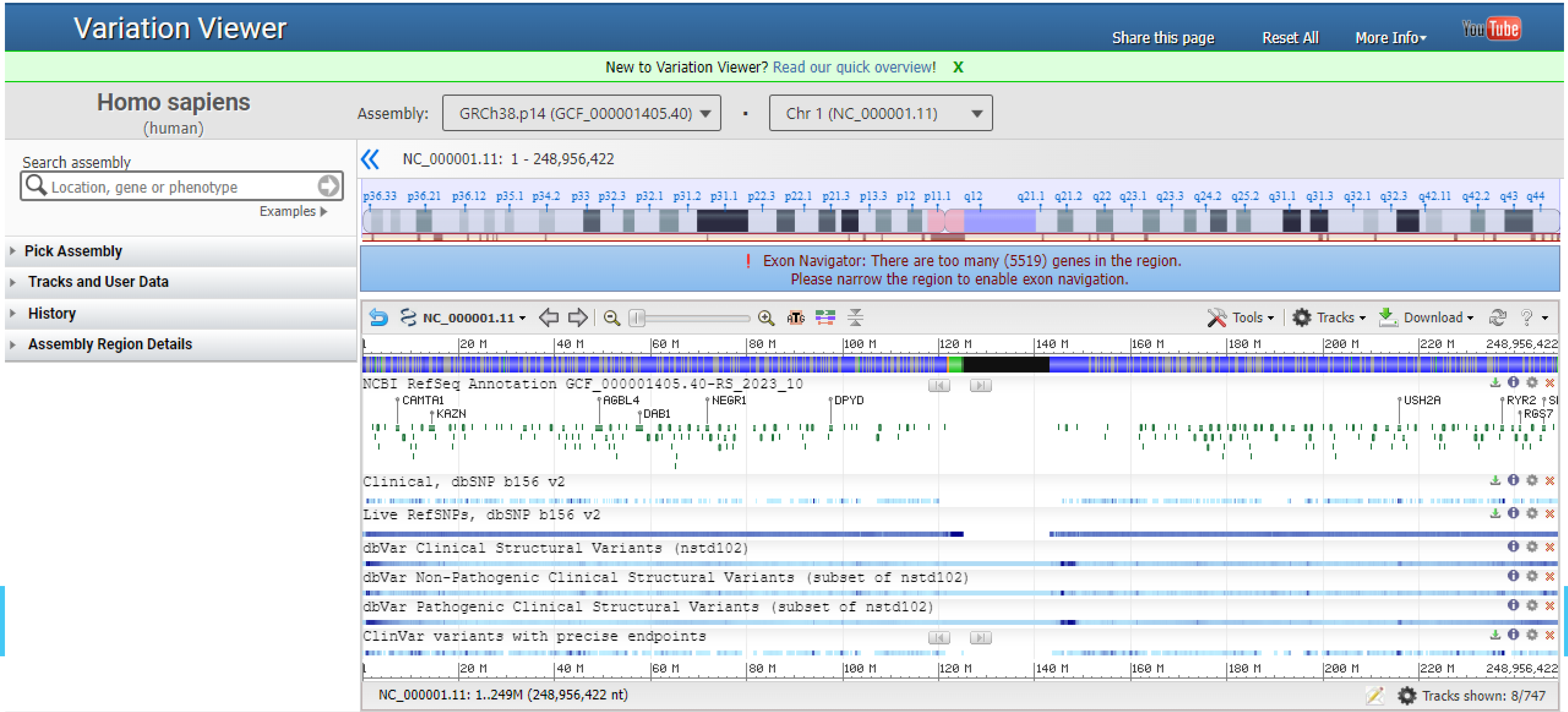
COLOR SCHEME: clustal2

```
AY136510.   ---------------------------------------------------ATGGTGCACCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGG
V00497.     -ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCGGTTACTGCCCTGTGGG
AF349114.   ----------------ACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
NM_000518.  -ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
AF181989.   ------------------AACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
BC007075.   GACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
AF117710.   ---------------------TGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGG
```

DNA variant analysis

Browsing genetic variations

- Natural genetic variation -> Variation Viewer (https://www.ncbi.nlm.nih.gov/variation/view)
- Database of short genetic variations -> NCBI dbSNP (https://www.ncbi.nlm.nih.gov/snp/)



New to Variation Viewer? Read our quick overview! X

Homo sapiens (human)

Assembly: GRCh38.p14 (GCF_000001405.40) · Chr 1 (NC_000001.11)

Search assembly: Location, gene or phenotype

NC_000001.11: 1 - 248,956,422

p36.33 p36.21 p36.12 p35.1 p34.2 p33 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.3 p13.3 p12 p11.1 q12 q21.1 q21.2 q22 q23.1 q23.3 q24.2 q25.2 q31.1 q31.3 q32.1 q32.3 q42.11 q42.2 q43 q44

! Exon Navigator: There are too many (5519) genes in the region. Please narrow the region to enable exon navigation.

NC_000001.11

NCBI RefSeq Annotation GCF_000001405.40-RS_2023_10

Clinical, dbSNP b156 v2

Live RefSNPs, dbSNP b156 v2

dbVar Clinical Structural Variants (nstd102)

dbVar Non-Pathogenic Clinical Structural Variants (subset of nstd102)

dbVar Pathogenic Clinical Structural Variants (subset of nstd102)

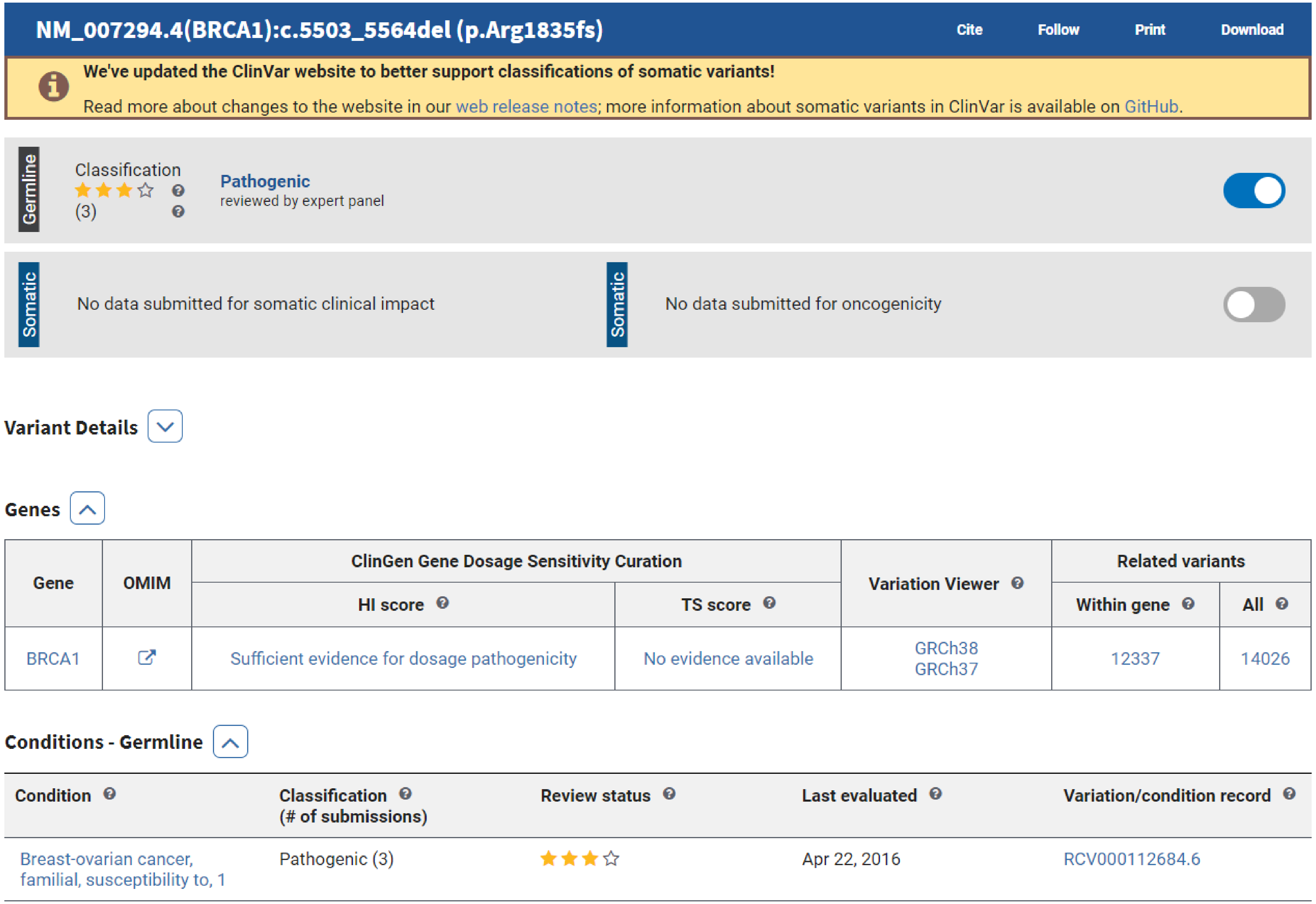
ClinVar variants with precise endpoints

NC_000001.11: 1..249M (248,956,422 nt)

Tracks shown: 8/747

DNA variant analysis

Database of variants with clinical significance: **ClinVar** (https://www.ncbi.nlm.nih.gov/clinvar/)



NM_007294.4(BRCA1):c.5503_5564del (p.Arg1835fs)

We've updated the ClinVar website to better support classifications of somatic variants!

Read more about changes to the website in our web release notes; more information about somatic variants in ClinVar is available on GitHub.

Germline

Classification ★★★☆ (3) — **Pathogenic** reviewed by expert panel

Somatic — No data submitted for somatic clinical impact

Somatic — No data submitted for oncogenicity

Variant Details

Genes

Gene	OMIM	ClinGen Gene Dosage Sensitivity Curation		Variation Viewer	Related variants	
		HI score	TS score		Within gene	All
BRCA1	↗	Sufficient evidence for dosage pathogenicity	No evidence available	GRCh38 GRCh37	12337	14026

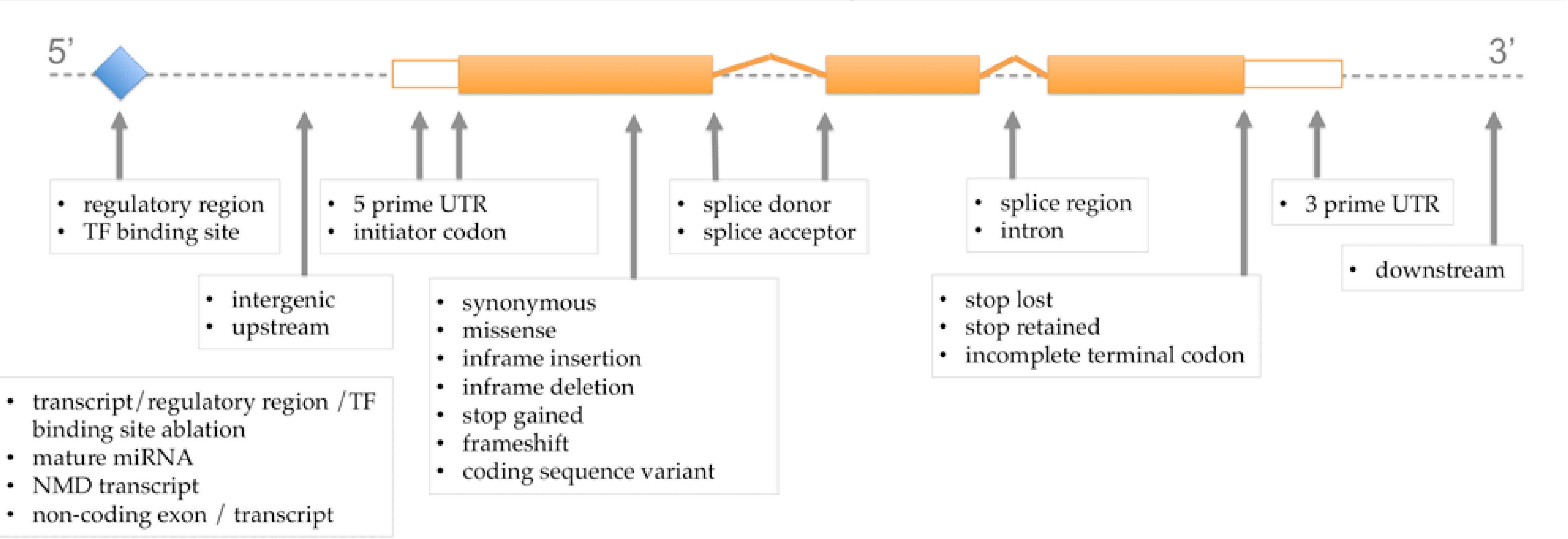
Conditions - Germline

Condition	Classification (# of submissions)	Review status	Last evaluated	Variation/condition record
Breast-ovarian cancer, familial, susceptibility to, 1	Pathogenic (3)	★★★☆	Apr 22, 2016	RCV000112684.6

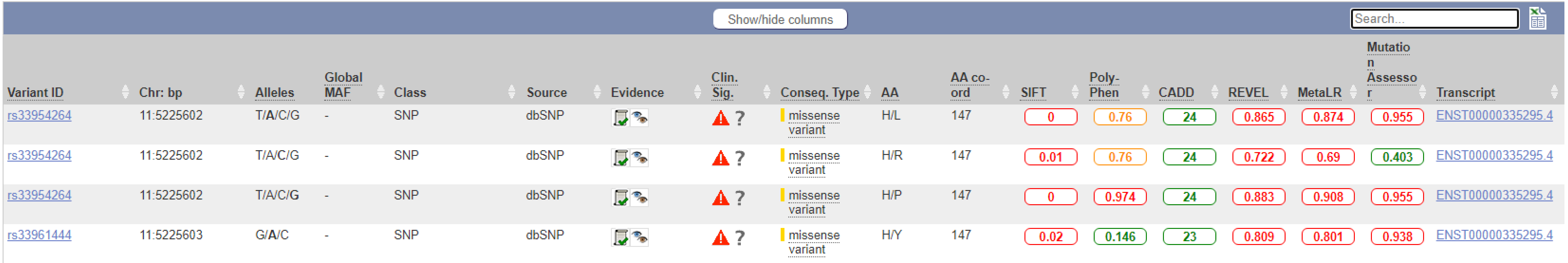
DNA variant analysis

Genetic variation -> effect on protein structure/function?

- Depends on the location of the mutation/variation



- Use PROVEAN or SIFT (**S**orts **I**ntolerant **F**rom **T**olerant) score for amino acid substitutions

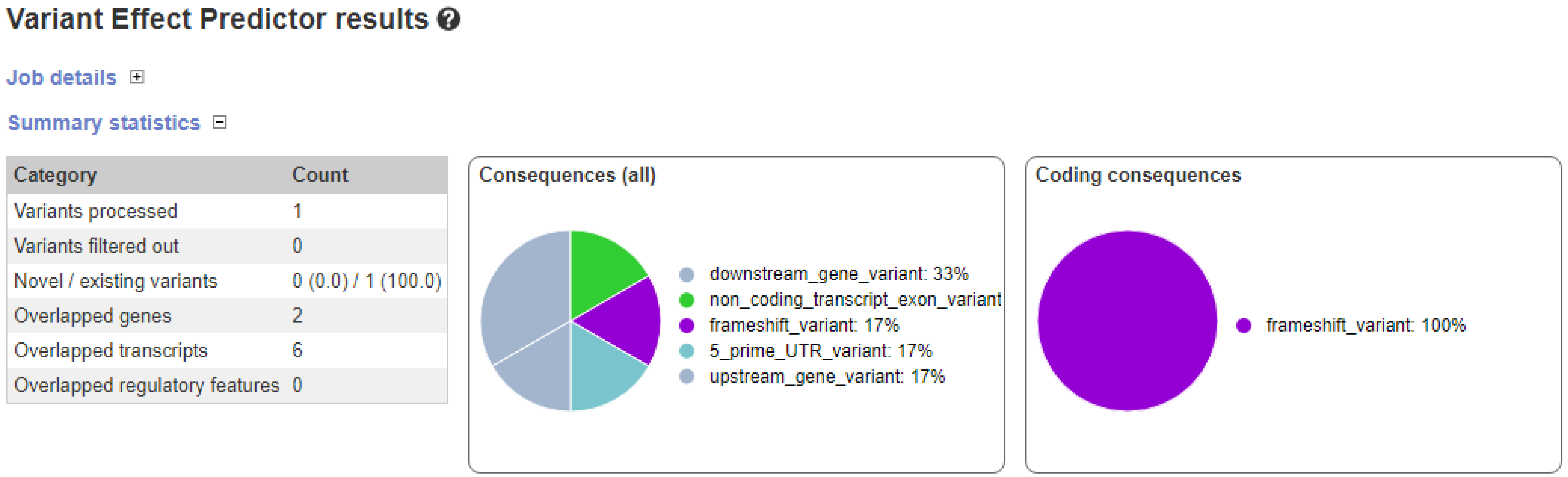


Variant ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA co-ord	SIFT	Poly-Phen	CADD	REVEL	MetaLR	Mutation Assessor	Transcript
rs33954264	11:5225602	T/A/C/G	-	SNP	dbSNP		⚠ ?	missense variant	H/L	147	0	0.76	24	0.865	0.874	0.955	ENST00000335295.4
rs33954264	11:5225602	T/A/C/G	-	SNP	dbSNP		⚠ ?	missense variant	H/R	147	0.01	0.76	24	0.722	0.69	0.403	ENST00000335295.4
rs33954264	11:5225602	T/A/C/G	-	SNP	dbSNP		⚠ ?	missense variant	H/P	147	0	0.974	24	0.883	0.908	0.955	ENST00000335295.4
rs33961444	11:5225603	G/A/C	-	SNP	dbSNP		⚠ ?	missense variant	H/Y	147	0.02	0.146	23	0.809	0.801	0.938	ENST00000335295.4

DNA variant analysis

Genetic variation → **effect** on protein structure/function?

- Variant Effect Predictor (https://www.ensembl.org/Homo_sapiens/Tools/VEP)



Variant Effect Predictor results

Job details ⊞

Summary statistics ⊟

Category	Count
Variants processed	1
Variants filtered out	0
Novel / existing variants	0 (0.0) / 1 (100.0)
Overlapped genes	2
Overlapped transcripts	6
Overlapped regulatory features	0

Consequences (all)
● downstream_gene_variant: 33%
● non_coding_transcript_exon_variant
● frameshift_variant: 17%
● 5_prime_UTR_variant: 17%
● upstream_gene_variant: 17%
Coding consequences
● frameshift_variant: 100%

- Example: investigate rs13306510
  - Look up the SNP in the dbSNP database
  - Examine the SNP with the Variant Effect Predictor

Concluding remarks

Bioinformatics is more than sequence alignment, BLAST, variant calling…

- Interested in more? Be sure to check our offers of further training!

(advanced bachelor) Bioinformatics

- Also in @Home version!

Click us!

R for Data Analysis and Visualisation

- Also in @Home version!

howest
university of applied sciences

Bioinformatics for dummies
MB&C2024 Workshop

Cedric Hermans
Paco Hulpiau