

Jasper Decuyper

BIOINFORMATICS FOR DUMMIES

MB&C2019 WORKSHOP

howest.be

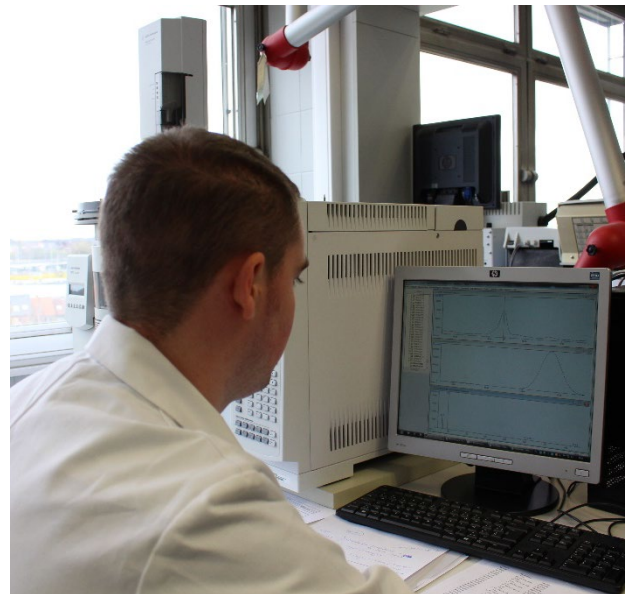


INTRODUCTION

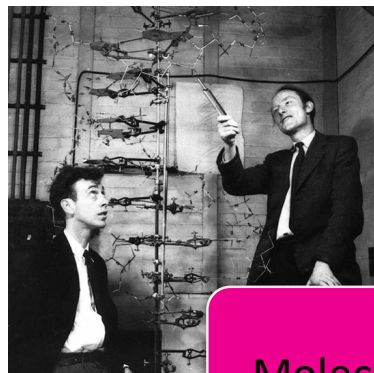


Imagine your workspace
without the computers...

Both in research
laboratories and in
hospitals...



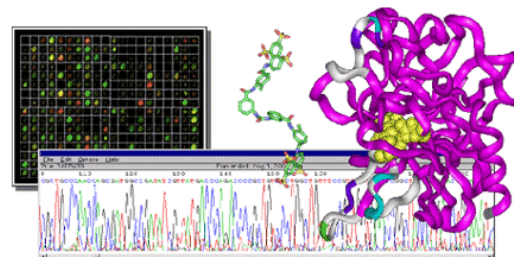
INTRODUCTION



Molecular
Biology

Information
Technologies

Bio-
informatics



Combine:

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

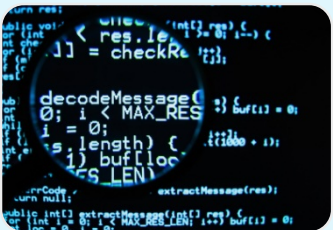
INTRODUCTION



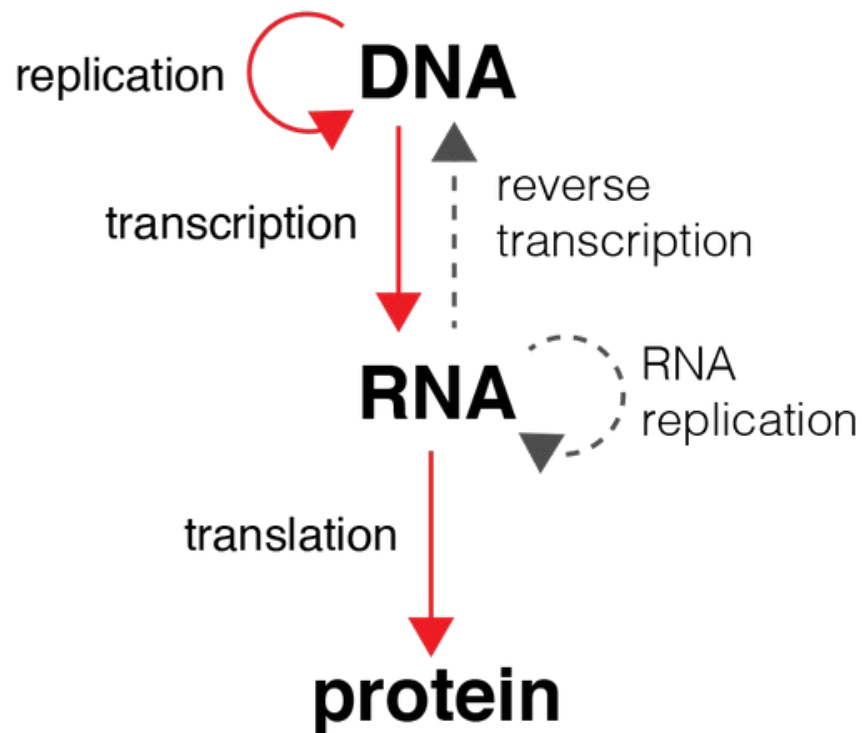
To store, organize and share molecular biological data in database systems



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



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

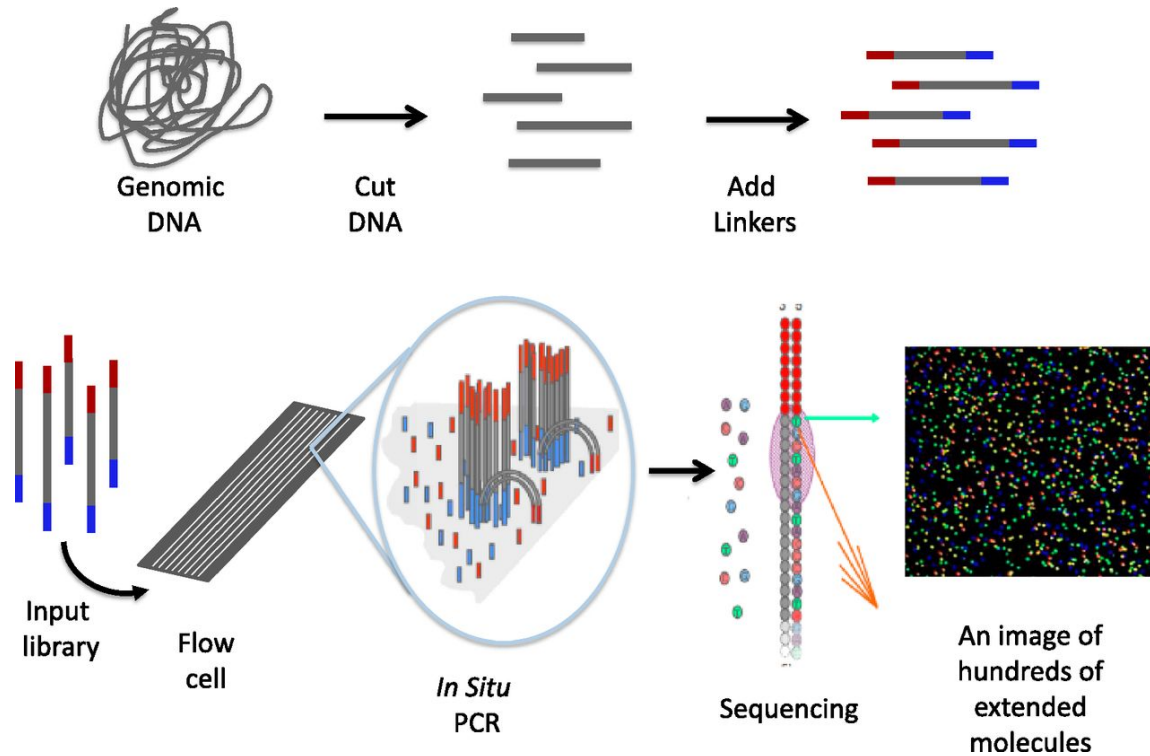
MOLECULAR BIOLOGY AND BIOINFORMATICS

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

Johnsen, J. M., Nickerson, D. A. & Reiner, A. P. (2013). Massively parallel sequencing: the new frontier of hematologic genomics. *Blood*, 122(19), 3268–3275.



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

SMRTbell template

Two hairpin adapters allow continuous circular sequencing



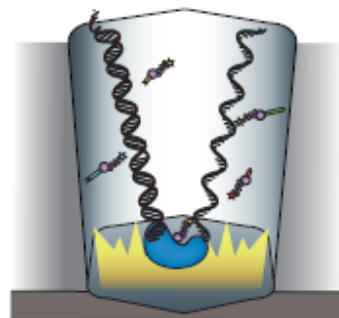
ZMW wells

Sites where sequencing takes place



Labelled nucleotides

All four dNTPs are labelled and available for incorporation

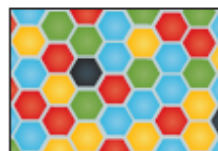


Modified polymerase

As a nucleotide is incorporated by the polymerase, a camera records the emitted light

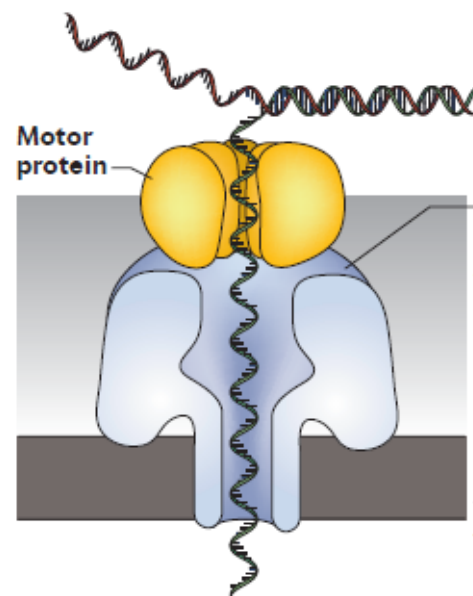
PacBio output

A camera records the changing colours from all ZMWs; each colour change corresponds to one base



Leader-Hairpin template

The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing



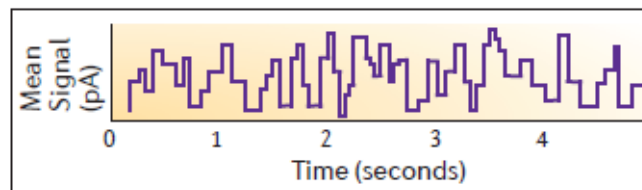
Motor protein

Alpha-hemolysin

A large biological pore capable of sensing DNA

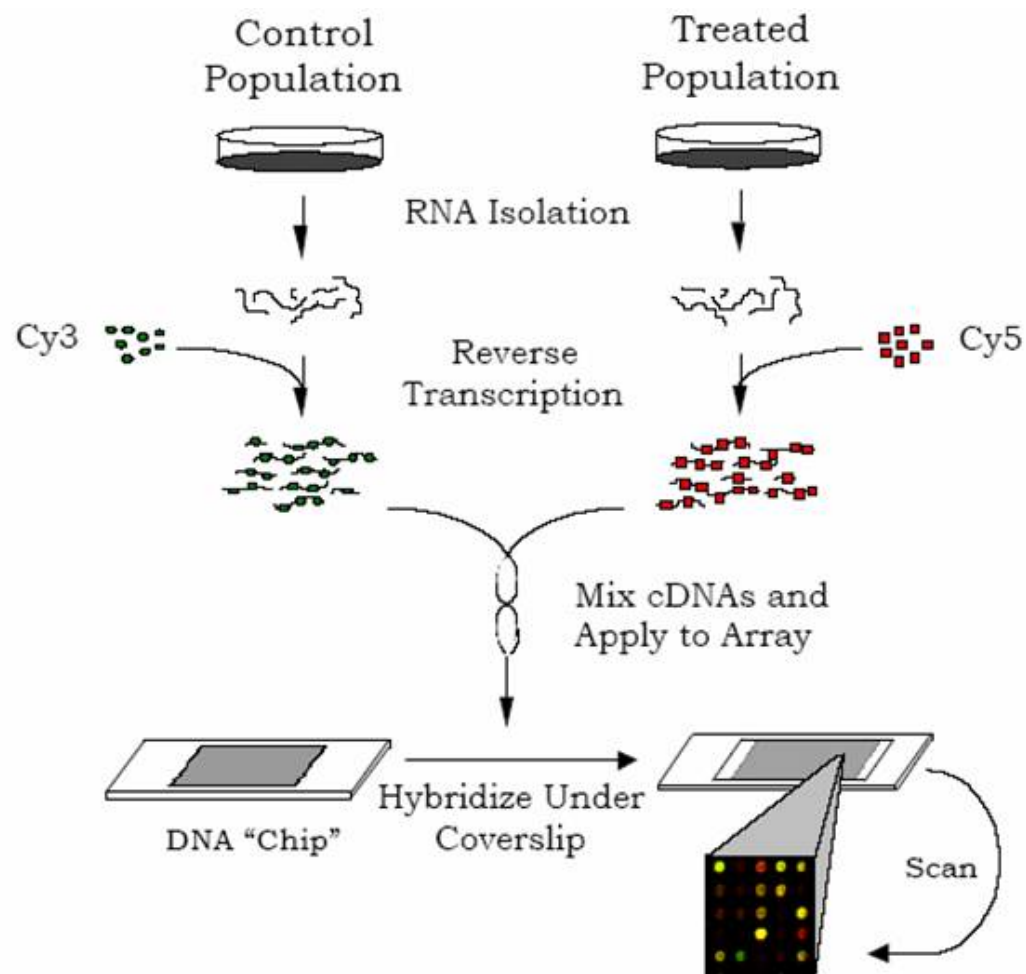
Current

Passes through the pore and is modulated as DNA passes through

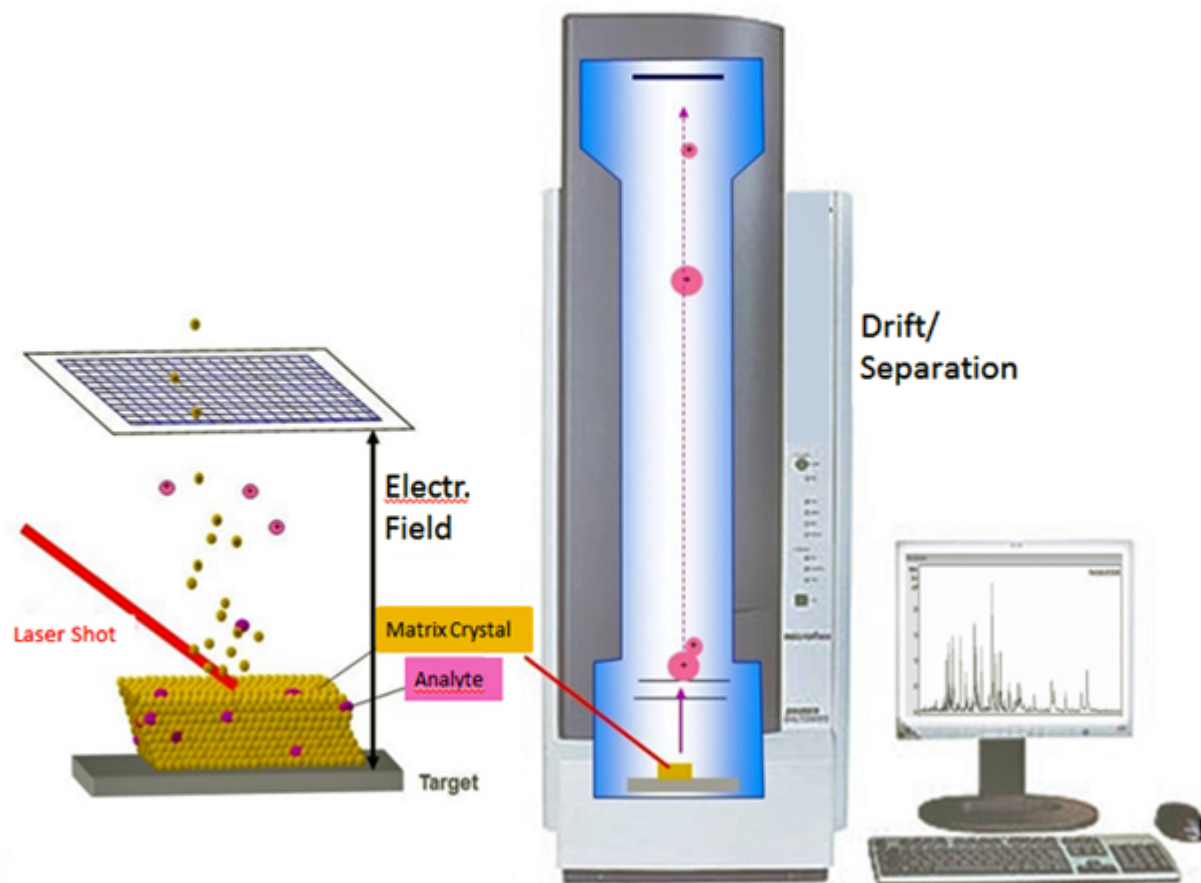


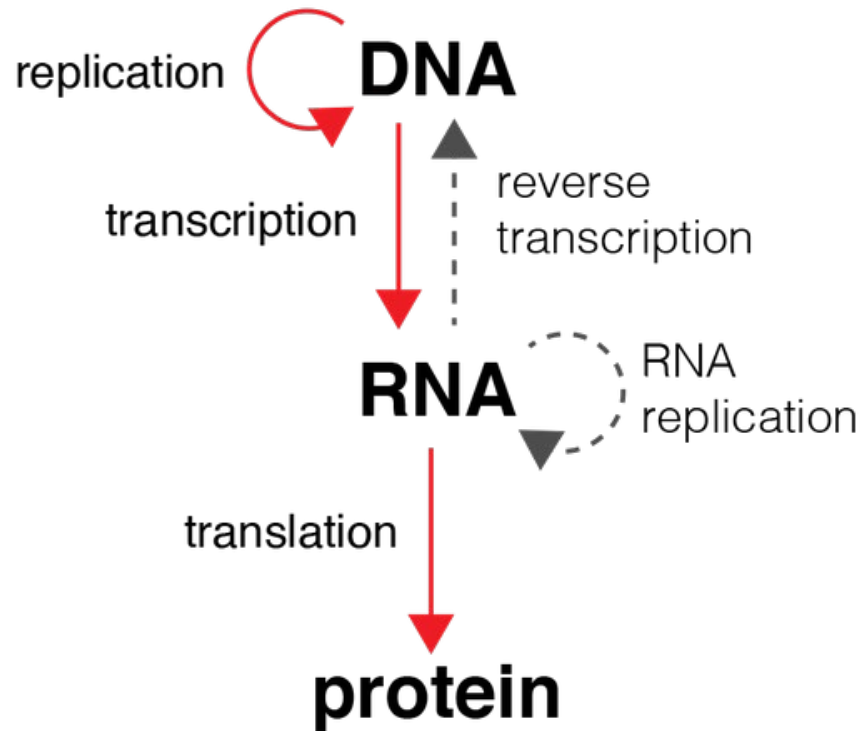
ONT output (squiggles)

Each current shift as DNA translocates through the pore corresponds to a particular k-mer

MICROARRAYS

MASS SPECTROMETRY



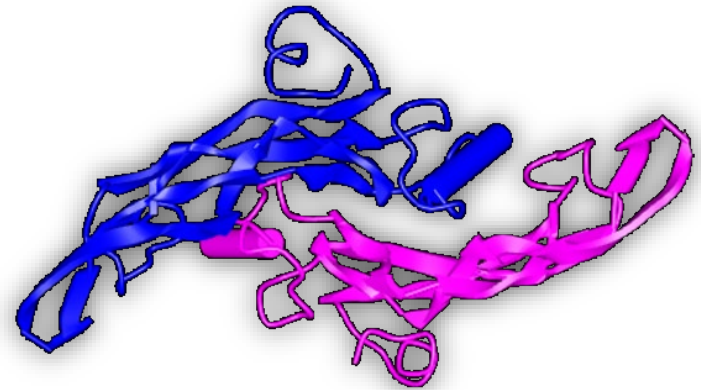
MOLECULAR BIOLOGY AND BIOINFORMATICS

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/nucleotide/?term=TGFB1+AND+\"Homo+sapiens\"\[Organism\]](https://www.ncbi.nlm.nih.gov/nucleotide/?term=TGFB1+AND+\)
- All available (partial) TGF beta 1 nucleotide sequences → ± 135 records (!)

NCBI UniGene

- <https://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?UGID=2394304>
- Transcripts and gene expression (EST Profile) information

UniProt or NCBI Protein

- <http://www.uniprot.org/uniprot/P01137>
- High-quality recourse of protein sequence and functional information

(SEQUENCE) REPOSITORIES

- Example 1: looking for the nucleotide sequence of PSA
- <https://www.ncbi.nlm.nih.gov/nucleotide/>
- NCBI nucleotide query: “(prostate specific antigen)” restricted to humans

The screenshot shows the NCBI Nucleotide search interface. The search query is "(prostate specific antigen) AND \"Homo sapiens\"[porgn: __txid9606]". The results page displays a list of 123 items, with the first four items shown:

- [Human prostate specific antigen gene, complete cds](#)
- [Homo sapiens mRNA for prostate specific antigen \(KLK3 gene\), splice variant RP5](#)
- [Homo sapiens mRNA for prostate specific antigen \(KLK3 gene\), splice variant 2](#)
- [Homo sapiens mRNA for prostate specific antigen \(KLK3 gene\), splice variant 1](#)

The search details section shows the query: "prostate specific antigen[All Fields] AND \"Homo sapiens\"[porgn]". The recent activity section shows the search history: "(prostate specific antigen) AND \"Homo sapiens\"[porgn] (123)" and "(PSA) AND \"Homo sapiens\"[porgn] (477)".

(SEQUENCE) REPOSITORIES

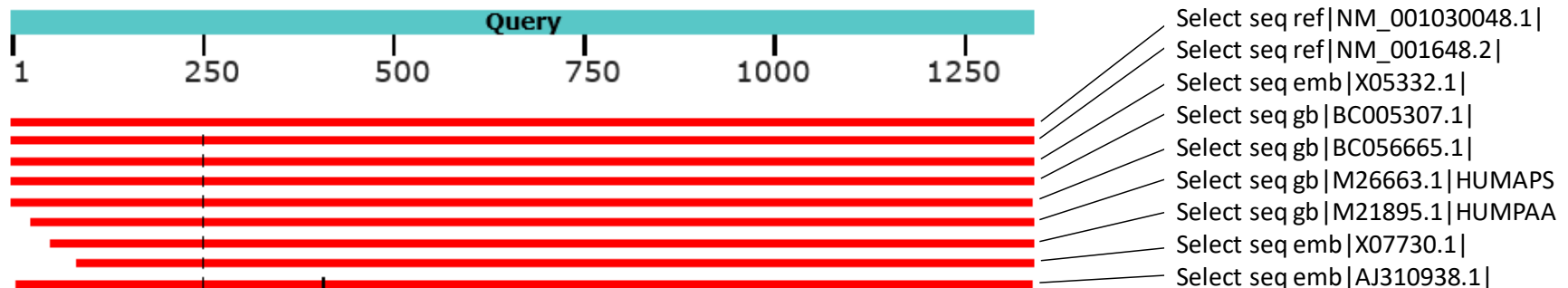


- Example 2: **the 1000 Genomes project** (2008-2015)
- Goal = to find most genetic variants with frequencies of at least 1% in the populations studied
- ACGTACGTACGTACG**C**GTACGTACGT
- ACGTAC**C**TACGTAC**C**GTACGTACGT
- ACGTAC**C**TACGTATG**T****T**CGTACGT
- ACGTACGTACGTATG**T****T**CGTACGT

(SEQUENCE) REPOSITORIES

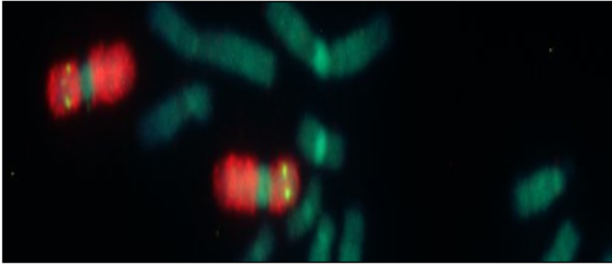
Solution for genetic and sequence diversity:

- **Genome Reference Consortium (GRC):** to create the best possible reference assembly for human → latest major release: GRCh38
 - <https://www.ncbi.nlm.nih.gov/grc/human>
- **NCBI Reference Sequence Database (RefSeq):** a non-redundant, well-annotated set of reference sequences incl. genomic, transcript, and protein
 - <https://www.ncbi.nlm.nih.gov/refseq/>
 - One gene – one sequence



NCBI Resources How To Sign in to NCBI

Gene [Advanced](#) [Help](#)



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](#)

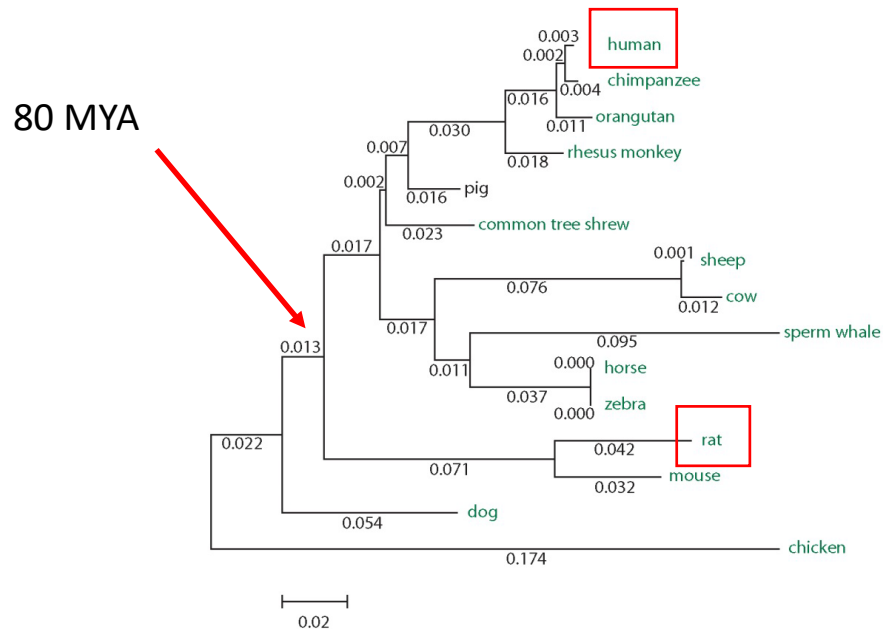
- ### Other Resources
- [HomoloGene](#)
 - [OMIM](#)
 - [RefSeq](#)
 - [RefSeqGene](#)
 - [UniGene](#)
 - [Protein Clusters](#)

Representative queries

Find genes by...	Search text
free text	human muscular dystrophy
chromosome and symbol	(1[chr] OR 2[chr]) AND adh*[sym]

BEST PRACTICE - on how to find a reference sequence

HOMOLOGY SEARCHING



Homology

- Derived from a common ancestor
- 2 types:
 - Orthologs = due to speciation event
 - Paralogs = due to duplication event
- Typically based on morphological characteristics
- Making use of “molecular phylogeny” to determine homology

HOMOLOGY SEARCHING

```
CAAGGCTGTCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGCATCTCTCG
GCAGTTCGTCTGTGACTCAGACCGGGACTGCTTGGACGGCTCAGACGAGGCCTCCTGCCCGGTGCTCA
CCTGTGGTCCCGCCAGCTTCCAGTGCAACAGCTCCACCTGCATCCCCCAGCTGTGGGCCTGCGACAAC
```

- Given = an unknown human nucleotide sequence
 → <https://www.bioit.be> > “unknown human nucleotide sequence.fasta”

- To determine the identity → use **BLAST**
 - Against the *Homo sapiens* RefSeq RNA database, exclude models
 - <https://blast.ncbi.nlm.nih.gov/Blast.cgi>



- Identity?
- Bits score?
- Expect value?

Homo sapiens low density lipoprotein receptor (LDLR), transcript variant 2, mRNA
 Sequence ID: [NM_001107281.1](#) Length: 1200 Number of Matches: 1

Range 1: 491 to 700 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
388 bits(210)	3e-107	210/210(100%)	0/210(0%)	Plus/Plus
Query 1	CAAGGCTGTCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGC	60		
Sbjct 491	CAAGGCTGTCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGC	550		

HOMOLOGY SEARCHING

BLAST

- ≠ 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.

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
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
- Generate similar lists of words spanning the query (e.g. words for **WGW**, **GWG**, **WGK**...).

threshold —————→
 examples of words >= threshold 12
 examples of words below threshold

Phase 2: Scanning and extensions

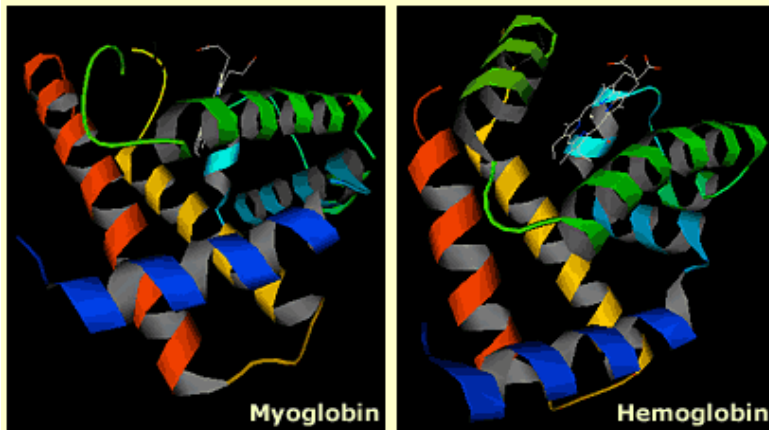
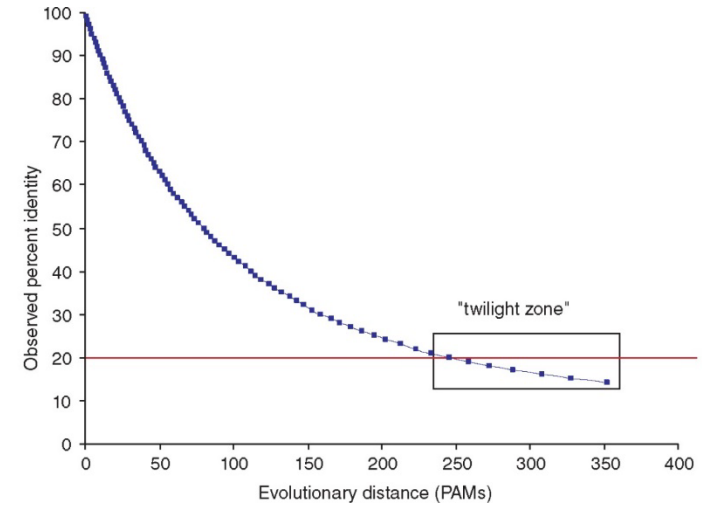
- Select all the words above threshold T (LWG, IWG, MWG, VWG, FWG, AWG, LWS, LWN, LWA, LYG)
- Scan the database for entries (“hits”) that match the compiled list
- Create a hash table index with the locations of all the hits for each word
- Perform gap free extensions
- Perform gapped extensions

```

    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV HBB
    L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F      D G+ +V
    LSPADKTNVKAAWGKGVAHAGEYGAEALERMFSLFPTTKTYFPHF-----DLSHGSAQV HBA
    ←-----extension                      extension-----→
    word pair from
    first phases of search
    "hits" alpha globin,
    triggers extension
    
```

HOMOLOGY SEARCHING

- Are two sequences homologous?
 - Percent identity (quantitative)
 - + Expect value
- While homology = YES or NO question !!



Example: is it possible to predict that human **myoglobin** (NP_005359) and **beta hemoglobin** (NP_000509) are paralogs?

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.bioit.be> > “HBB multiple sequence alignment.fasta”
 - Align sequences using MUSCLE software (www.ebi.ac.uk/Tools/msa/muscle/), output = HTML

Multiple sequence alignment (MSA)

→ Phylogenetic analysis

```

AY136510.1 -----ATGGTGCAC
V00497.1 -ACATTTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAC
AF349114.1 -----ACAACCTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAC
NM_000518.4 -ACATTTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAt
AF181989.1 -----AACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAt
BC007075.1 gACATTTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAC
AF117710.1 -----TGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCAC

AY136510.1 CTGACTCCTGtGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA
V00497.1 CTGACTCCTGAGGAGAAGTCTGCgGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA
AF349114.1 CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA
NM_000518.4 CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA
AF181989.1 CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA
BC007075.1 CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA
AF117710.1 CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA

```

DNA VARIANT ANALYSIS

Browsing genetic variations:

- Natural genetic variation → the 1000 Genomes Browser and/or Variation Viewer (→ *BRCA1*?)
 - <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>
 - <https://www.ncbi.nlm.nih.gov/variation/view/>
- The database of short genetic variation → NCBI dbSNP (→ *BRCA1*?)

The screenshot displays the NCBI 1000 Genomes Browser interface. At the top, the NCBI logo and navigation links are visible. The main header shows the current view: "Homo sapiens: GRCh37.p13 (GCF_000001405.25) Chr 1 (NC_000001.10): 1 - 249.3M". A navigation bar includes "Reset All", "Share this page", "FAQ", "Help", and "Version 3.7". A green attention banner states: "ATTENTION: You are browsing the alignment and genotype data from the Phase 3 May 2013 call set."

The interface is divided into several sections:

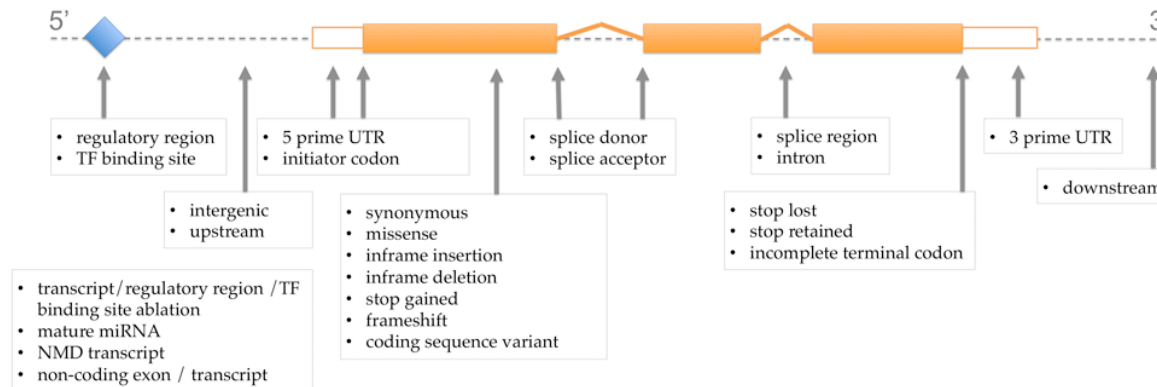
- Ideogram View:** Shows a vertical ideogram of chromosomes 1 through 22, X, Y, and MT. Chromosome 1 is highlighted in green.
- Exon Navigator:** A blue banner with a warning icon stating: "There are too many genes in the region (3817). Please narrow the region to enable exon navigation."
- Genomic Tracks:** A horizontal track showing various genomic features:
 - Segmental Duplications, Eichler Lab:** Shows blue bars representing duplications, with labels like 233.
 - 1000 Genomes Phase 3 Strict Accessibility Mask:** Shows a blue signal representing accessibility, with labels like 1029.
 - Genes, NCBI Homo sapiens Annotation Release ...:** Shows gene models with exons and introns.

The bottom of the interface includes a search bar and navigation controls.

DNA VARIANT ANALYSIS

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

- Depends on the location of the mutation/variation:



- Make use of PROVEAN or SIFT (sorts intolerant from tolerant) score for amino acid substitutions:

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Biotype	Exon	Intron	Amino acids	Codons	SIFT
rs33958637	11:5225717-5225717	C	missense_variant	MODERATE	HBB	3043	protein_coding	3/3	-	N/D	AAC/GAC	0.85
rs33958637	11:5225717-5225717	G	missense_variant	MODERATE	HBB	3043	protein_coding	3/3	-	N/H	AAC/CAC	0
rs576852971	11:5226131-5226131	G	intron_variant	MODIFIER	HBB	3043	protein_coding	-	2/2	-	-	-

Variant table @ Ensembl genome browser

DNA VARIANT ANALYSIS

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

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

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

Consequences (all)



Coding consequences



- 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 and variant calling ...
- Interested in more?

Toegepaste Bio-informatica in de medische moleculaire diagnostiek

(advanced bachelor) Bioinformatics

@home



The programme of the advanced bachelor training in **distance training** is the same as in full-time daytime education, but is built up differently and spread over **two academic years**.

Jasper Decuyper

BIOINFORMATICS FOR DUMMIES

MB&C2019 WORKSHOP

howest.be