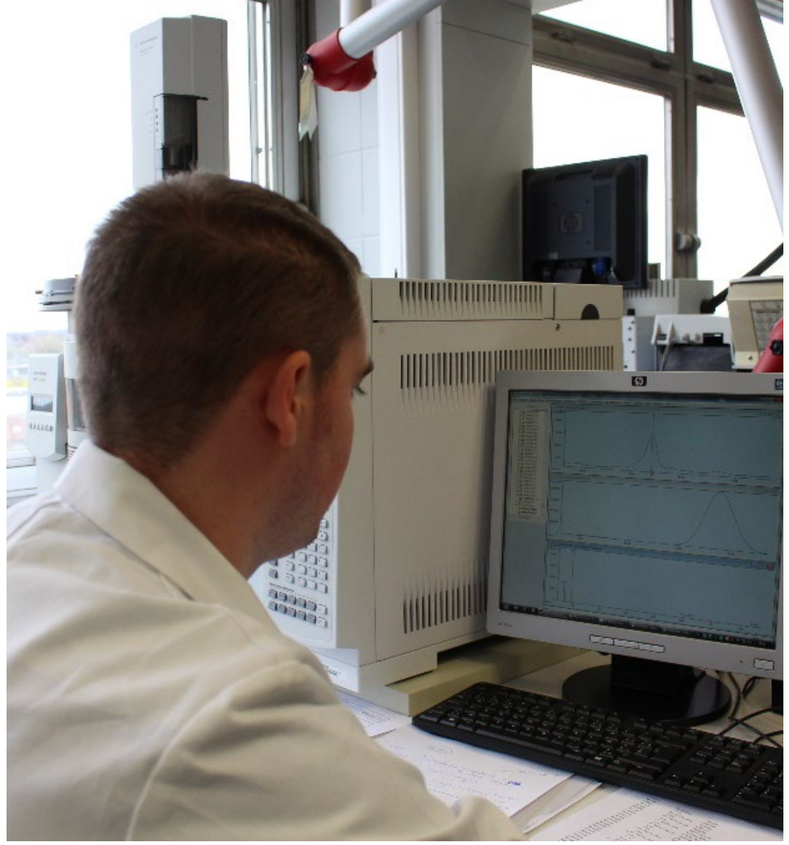
howest
hogeschool

European Association for Professions in Biomedical Science
EPBS

Bio-informatica voor dummies
MLT symposium 2022

Jasper Decuyper

howest
BIOINFORMATICS

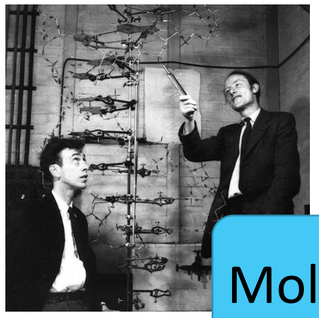


Imagine your workspace without the computers…

Both in research laboratories and in hospitals…

howest hogeschool

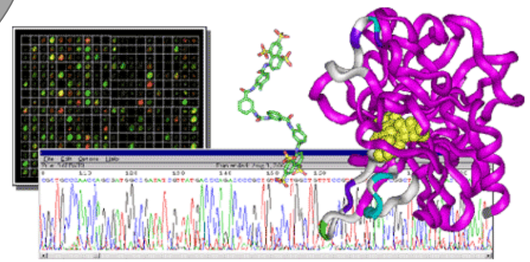
Bio-informatica?



Moleculaire biologie

Informatica

Bio-informatica



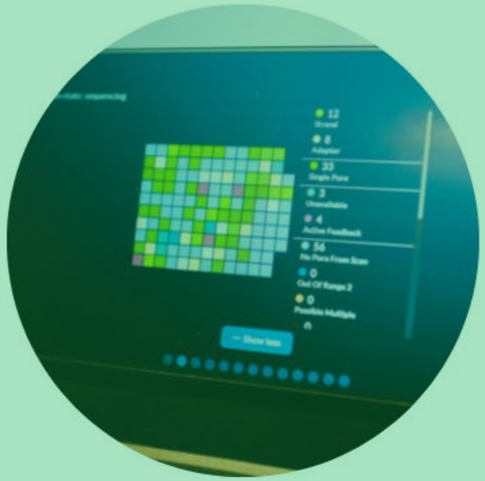
Combinatie van:

- Nieuwe inzichten en technologieën binnen de moleculaire biologie
- Vooruitgang binnen de informatica

howest hogeschool

Bio-informatica?

Moleculair biologische data opslaan, organiseren en delen



Data-verwerking, analyse en visualisatie

Automatisatie en integratie van tools in pipelines

howest hogeschool

howest
hogeschool

Biologische data in online databanken

MLT symposium 2022
Bio-informatica voor dummies

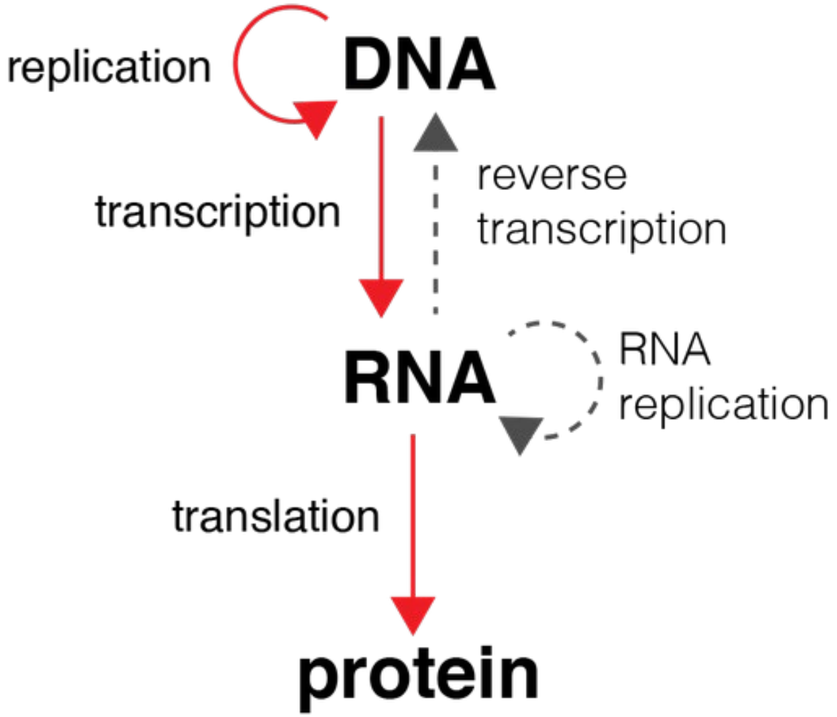
howest
BIOINFORMATICS

Biologische data

Centraal dogma moleculaire biologie

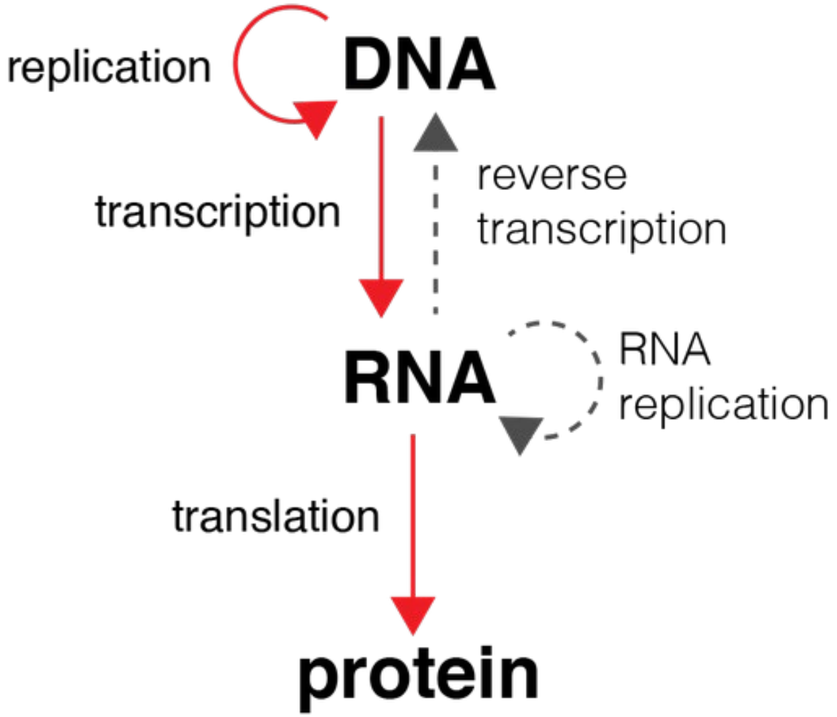
Belangrijke (high-throughput) technologieën:

- Massively parallel sequencing
  - Sequencing en expressie analyse
- Microarray
  - Expressie en genetische variatie analyse
- Massa spectrometrie
  - Eiwit (sequentie) identificatie

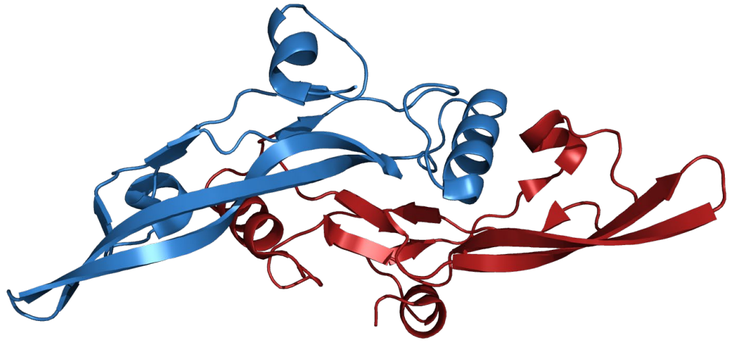


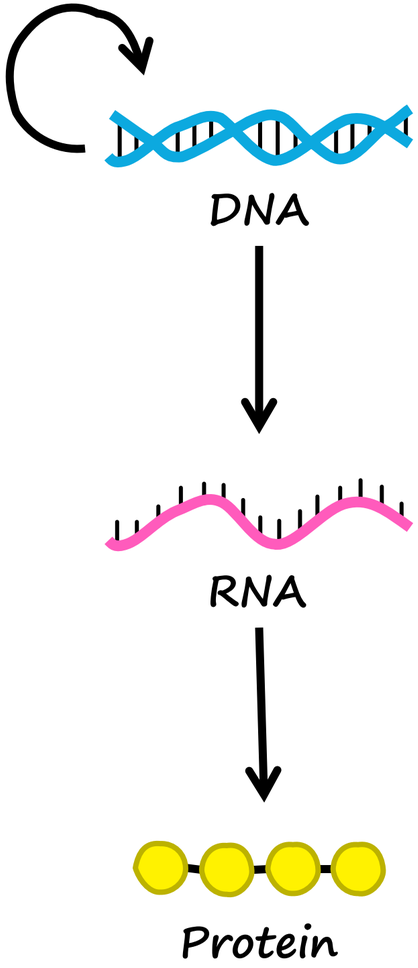
Biologische databanken

- DNA
  - Sequentie en loci
  - (Natuurlijke) genetische variatie
- RNA
  - Transcripten (en varianten)
  - Genexpressie
- Eiwit
  - Sequentie en functie
  - Fenotype (en ziekten)



Biologische databanken





Voorbeeld: **TGF beta 1** (TGFB1)

- **NCBI Gene:** Algemene en geïntegreerde sequentie- en locusinformatie
  - https://www.ncbi.nlm.nih.gov/gene/?term=TGF+beta+1+human
- **NCBI Nucleotide:** Alle beschikbare (partiële) TGF beta 1 nucleotidesequenties ± 137 records (!)
  - https://www.ncbi.nlm.nih.gov/nuccore/?term=TGFB1+AND+%22Homo+sapiens%22%5bOrganism%5d
- **UniProt:** Kwalitatieve gegevens over de sequentie en de functionele eigenschappen van eiwitten
  - https://www.uniprot.org/uniprot/P01137

Genoomprojecten

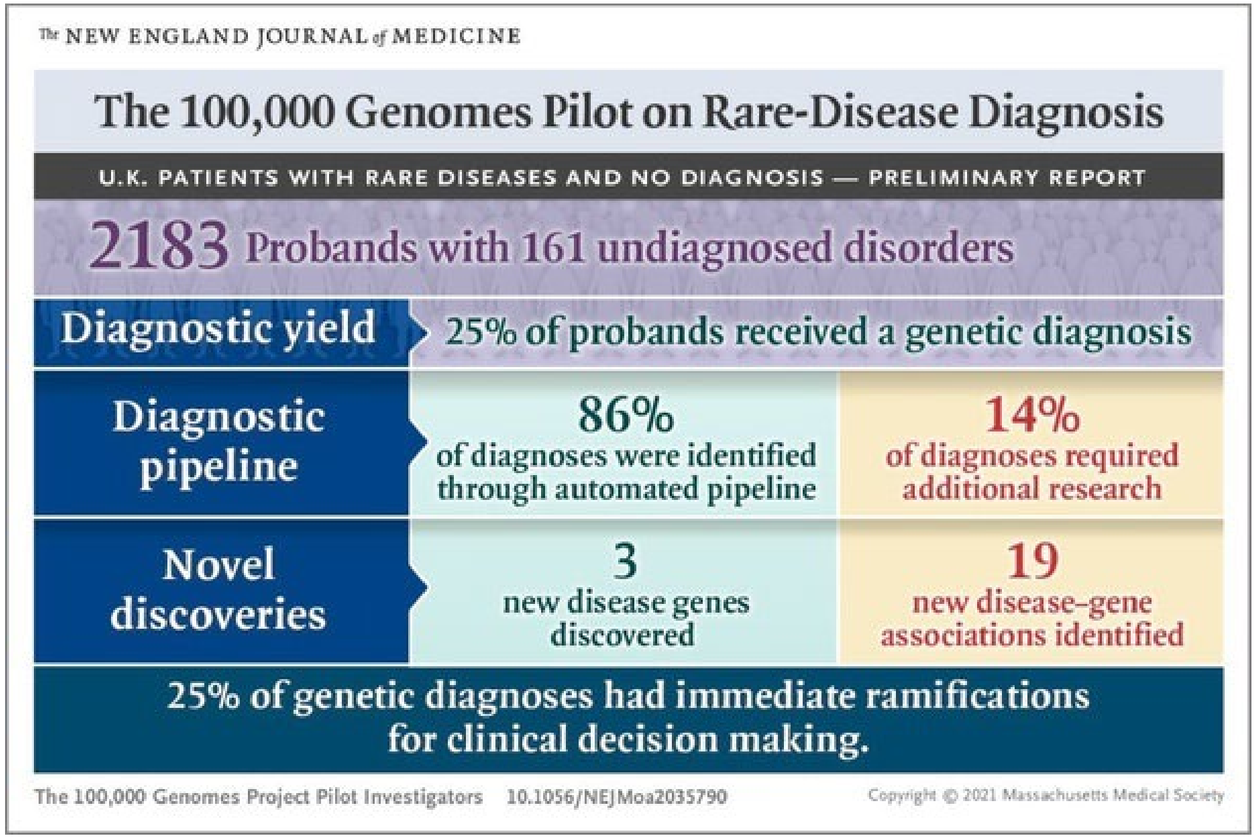
```
ACGTACGTACGTACGTACGTACGT
ACGTACCTACGTACGTACGTACGT
ACGTACCTACGTATGTTCGTACGT
ACGTACGTACGTATGTTCGTACGT
```

1000 Genomes Project (2008-2015)

- Doel: (alle) genetische varianten vinden met een frequentie ≥ 1% in de bestudeerde populaties

100 000 Genomes Project (2013-2018)

- Doel: focus op zeldzame ziekten, voorkomende vormen van kanker en infectieziekten

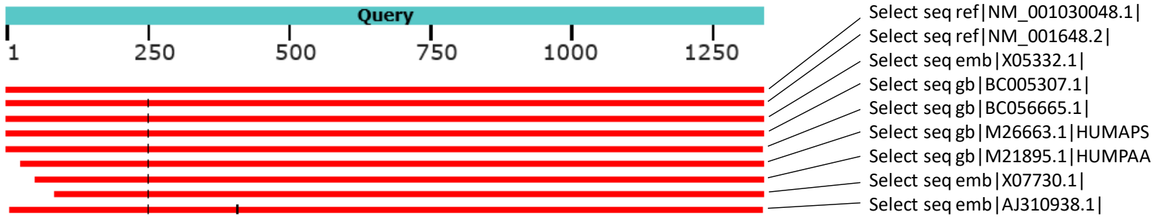


howest hogeschool

Data-tsunami

Nood aan een consensus - 3 initiatieven:

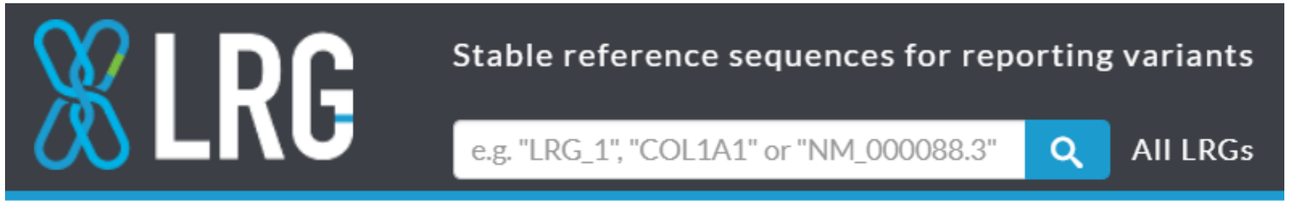
- **Genome Reference Consortium (GRC):** de best mogelijke referentie assembly (voor de mens) maken
  - Laatste major release: GRCh38 (ook hg38 genoemd)
  - https://www.ncbi.nlm.nih.gov/grc/human
- **NCBI Reference Sequence Database (RefSeq):** een niet-redundante, goed geannoteerde reeks referentiesequenties
  - Eén gen = één sequentie
  - https://www.ncbi.nlm.nih.gov/refseq/
- **Locus Reference Genomic (LRG)**



Data-tsunami

Nood aan een consensus - 3 initiatieven:

- **Genome Reference Consortium (GRC)**
- **NCBI Reference Sequence Database (RefSeq)**
- **Locus Reference Genomic (LRG) sequenties:** één stabiele referentie voor klinische toepassingen
  - Bij voorkeur één LRG sequentie per locus
  - Onafhankelijk van wijzigingen aan RefSeq en GENCODE
  - Eén LRG sequentie omvat één genomische sequentie, meerdere transcripten en meerdere eiwitten
  - https://www.lrg-sequence.org/



Data-tsunami

Oefening: Gebruik NCBI Gene om de referentiesequentie(s) van **humaan FLT3** op te zoeken

- Hoeveel referenties vind je terug van elk type (DNA/mRNA/eiwit)?

- Zoek vervolgens de LRG sequentie van FLT3 op

Fixed reference sequences in this record

Number of sequences Genomic: 1 / Transcript: 1 / Protein: 1

Genomic			Transcript				Protein			
Name	Length	Source	Name	Length	Source	MANE type	Name	Length	Source	CCDS
LRG_457	104,953 nt	NG_007066.1	t1	3,842 nt	NM_004119.2	-	p1	993 aa	NP_004110.2 ENSP00000241453.7	CCDS31953.1

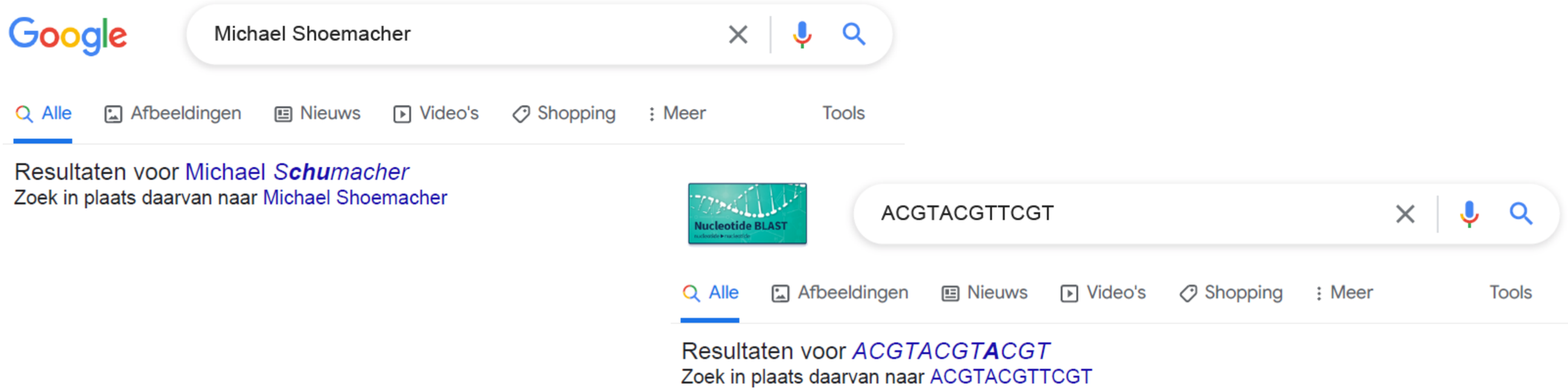
Download LRG_457 data: XML - FASTA

Table of contents

Homology searching

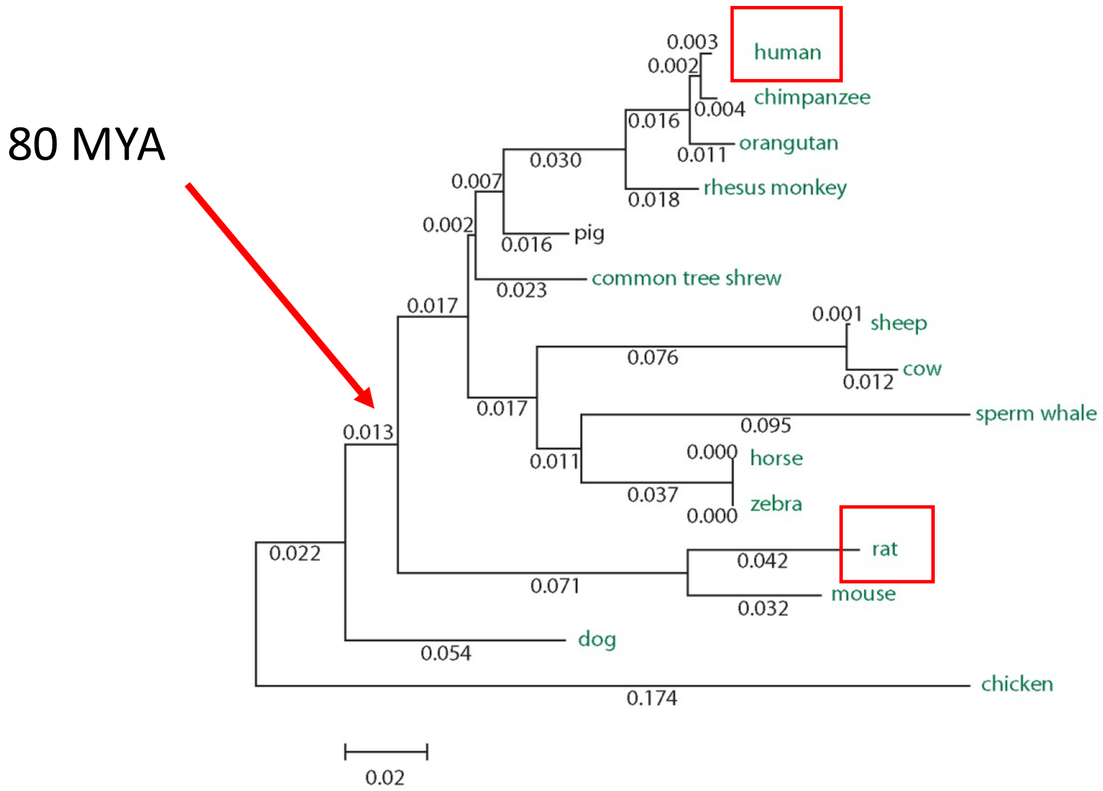
= zoeken met een biologische sequentie in een databank van sequenties

≠ eenvoudige trefwoord zoekstrategie



- Gebruik een model van evolutie om homologe sequenties te vinden/vergelijken

Homology searching



Homologe sequenties?

- Afgeleid van een gemeenschappelijke voorouder
- 2 soorten:
  - Orthologen = ontstaan door speciatie
  - Paralogen = ontstaan door duplicatie

Verwantschap op basis van moleculaire data

= moleculaire fylogenie

Homology searching

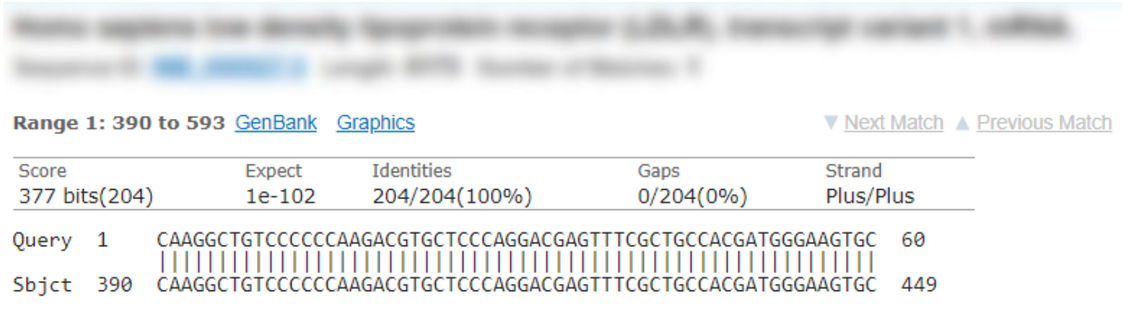
```
CAAGGCTGTCCCCCCAAGACGTGCTCCCAGGACGAGTTTCGCTGCCACGATGGGAAGTGCATCTCTCGGCAGTTCGT
CTGTGACTCAGACCGGGACTGCTTGGACGGCTCAGACGAGGCCTCCTGCCCGGTGCTCACCTGTGGTCCCGCCAGCT
TCCAGTGCAACAGCTCCACCTGCATCCCCCAGCTGTGGGCCTGCGACAAC
```

Gegeven = een onbekende humane nucleotidesequentie

- https://www.bioit.be > “unknown human nucleotide sequence.fasta”

Identiteit bepalen → BLAST

- Zoek in de Genomic + transcript databases (Human G+T), exclude models
- https://blast.ncbi.nlm.nih.gov/Blast.cgi



howest hogeschool

howest
hogeschool

Enkele voorbeelden van bio-informatica tools

MLT symposium 2022
Bio-informatica voor dummies

howest
BIOINFORMATICS

DNA - Genome browsers

Alle informatie van een genomisch locus → NCBI Gene

Letterlijk inzomen en navigeren over het genoom: **Genome browsers:**

- Annotaties zien in en rondom genen
- Chromosomale regio's bekijken
- Zoeken naar informatie op gen, chromosoom en genoom niveau
- Vergelijken van genomen

UCSC Genome Browser

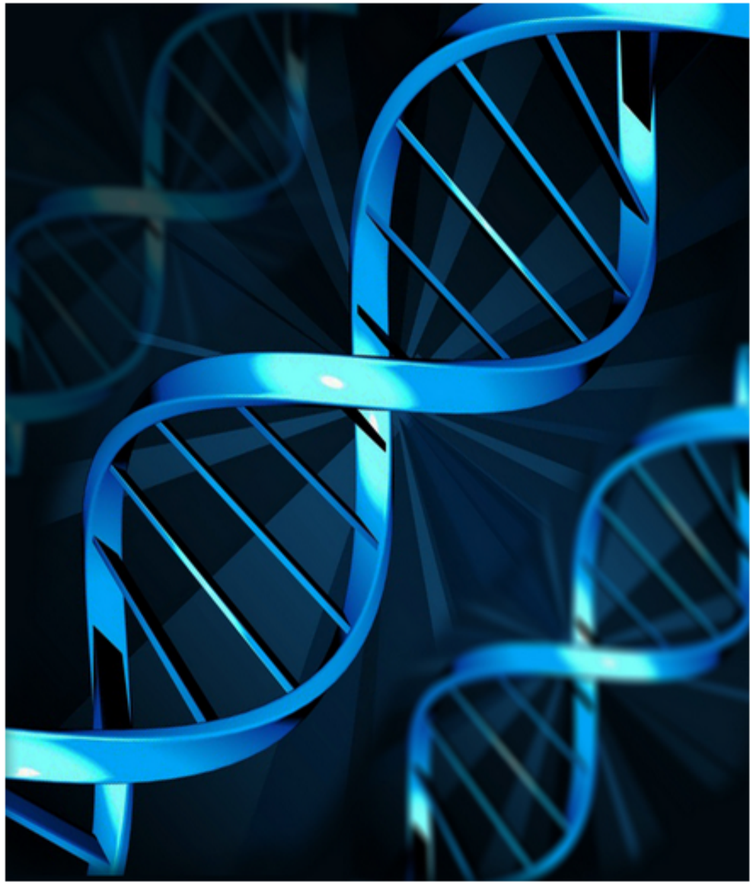
- https://genome.ucsc.edu/

howest hogeschool

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute | UCSC

Genome Browser

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us



Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls
- **Data Integrator**
  combine data sources from the Genome Browser database
- **Gene Sorter**
  find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server
- **In-Silico PCR**
  rapidly align PCR primer pairs to the genome
- **LiftOver**
  convert genome coordinates between assemblies
- **VisiGene**
  interactively view in situ images of mouse and frog

More tools...

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly assembled genome was released on the web at http://genome.ucsc.edu, along with the initial prototype of a graphical viewing tool, the UCSC Genome Browser. In the ensuing years, the website has grown to include a broad

What's new

Aug. 10, 2018 - New interact track type

Jul. 31, 2018 - New Ensembl gene tracks for 60 assemblies

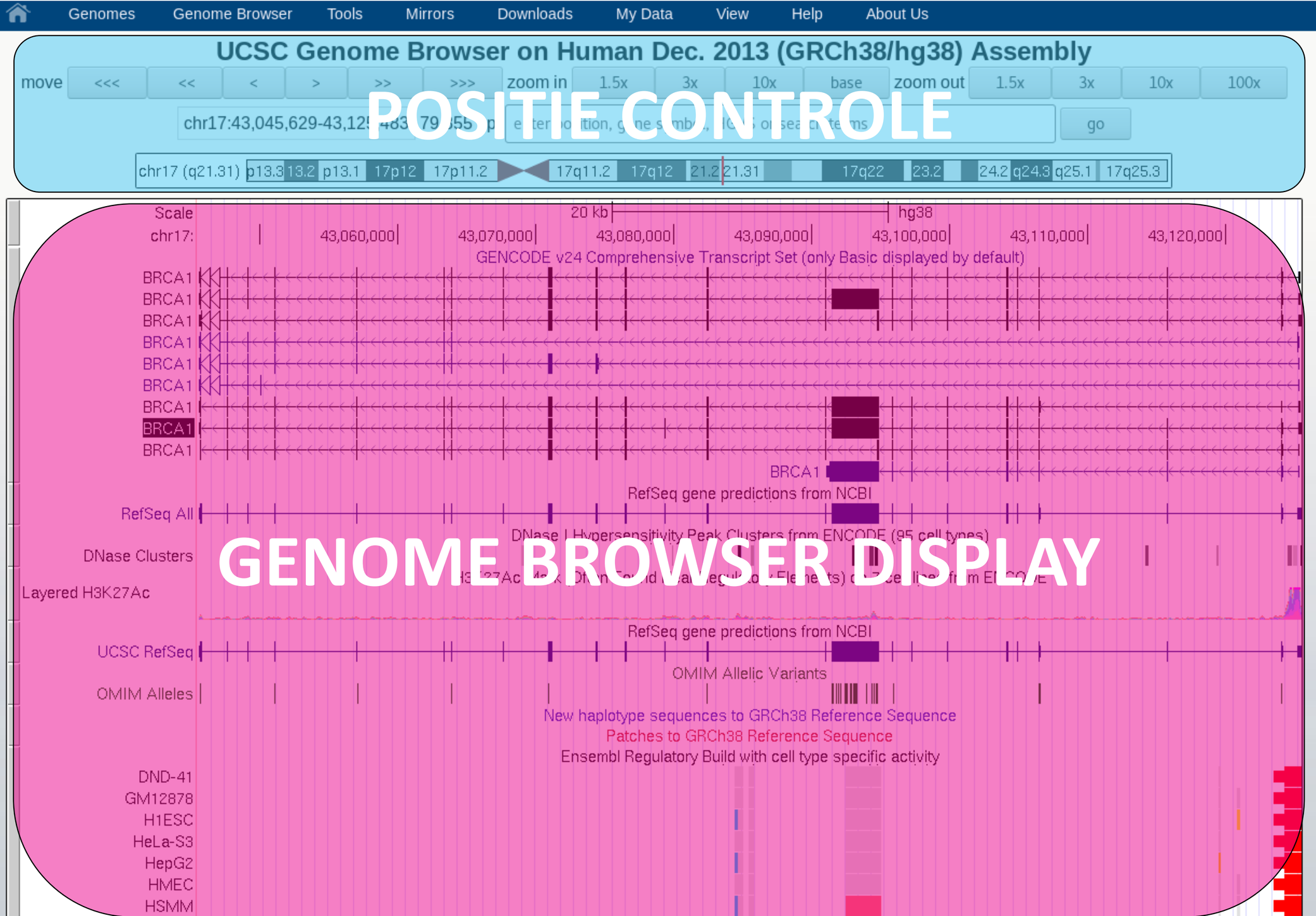
Jul. 17, 2018 - DECIPHER variants track available for human (GRCh37/hg19)

More news...

Subscribe

“Kent’s draft assembly” → 22 juni 2000
Celera Genomics assembly → 25 juni 2000

howest hogeschool


Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move
zoom in
zoom out
go
chr17 (q21.31)
POSITIE CONTROLE
Scale
20 kb
hg38
chr17:
43,060,000
43,070,000
43,080,000
43,090,000
43,100,000
43,110,000
43,120,000
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
BRCA1
RefSeq gene predictions from NCBI
RefSeq All
DNase Clusters
Layered H3K27Ac
UCSC RefSeq
OMIM Allelic Variants
OMIM Alleles
New haplotype sequences to GRCh38 Reference Sequence
Patches to GRCh38 Reference Sequence
Ensembl Regulatory Build with cell type specific activity
DND-41
GM12878
H1ESC
HeLa-S3
HepG2
HMEC
HSMM
GENOME BROWSER DISPLAY

move start

< 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

move end

< 2.0 >

track search | default tracks | default order | hide all | add custom tracks | track hubs | configure | multi-region | reverse | resize

refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

+	Ensembl Regulatory Build	disconnect	refresh
+	FANTOM5	disconnect	refresh
+	Mapping and Sequencing		refresh
+	Genes and Gene Predictions		refresh
+	Phenotype and Literature		refresh
+	mRNA and ESTs		refresh
+	Expression		refresh
+	Regulation		refresh
+	Comparative Genomics		refresh
+	Variation		refresh
+	Repeats		refresh

refresh

TRACK CONTROLE

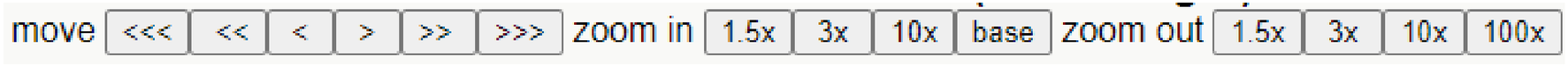
DNA - Genome browsers

Oefening 1:

- Surf naar https://genome.ucsc.edu/ en klik op 

- Zoek naar TGF beta 1 (TGFB1)

- Gebruik de positie controle om in/uit te zomen en/of te navigeren



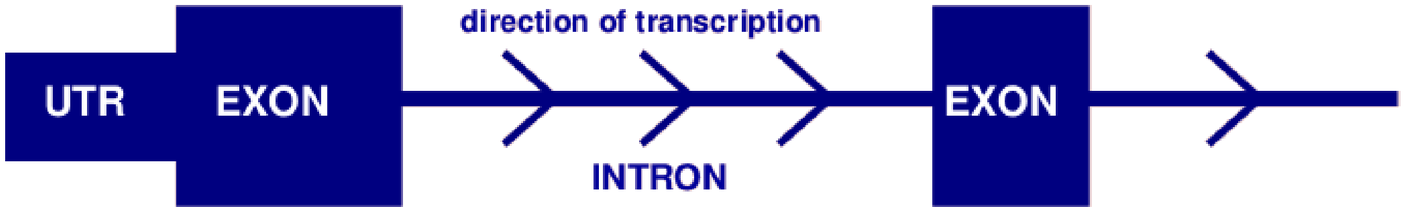
- Waar bevindt het gen zich op het genoom?
- Welke genen liggen er in de buurt van TGFB1?

DNA - Genome browsers

Oefening 2:

- Zoek naar CD27 met behulp van de UCSC Genome browser

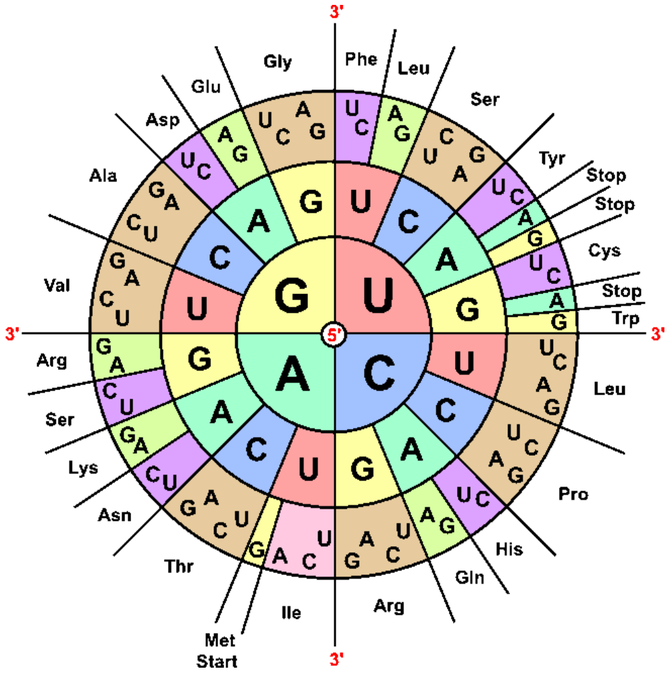
- Zoom in op de translatiestartplaats
- Kan je het startcodon herkennen?



RNA - In silico vertalen

Eiwit = lineair polymeer van aminozuren

- 20 fundamentele aminozuren
- mRNA basen coderen per drie (codon) voor één aminozuur



Op basis van de relatie tussen DNA/RNA en eiwit is het mogelijk om een nucleotidesequentie “in silico” te vertalen:

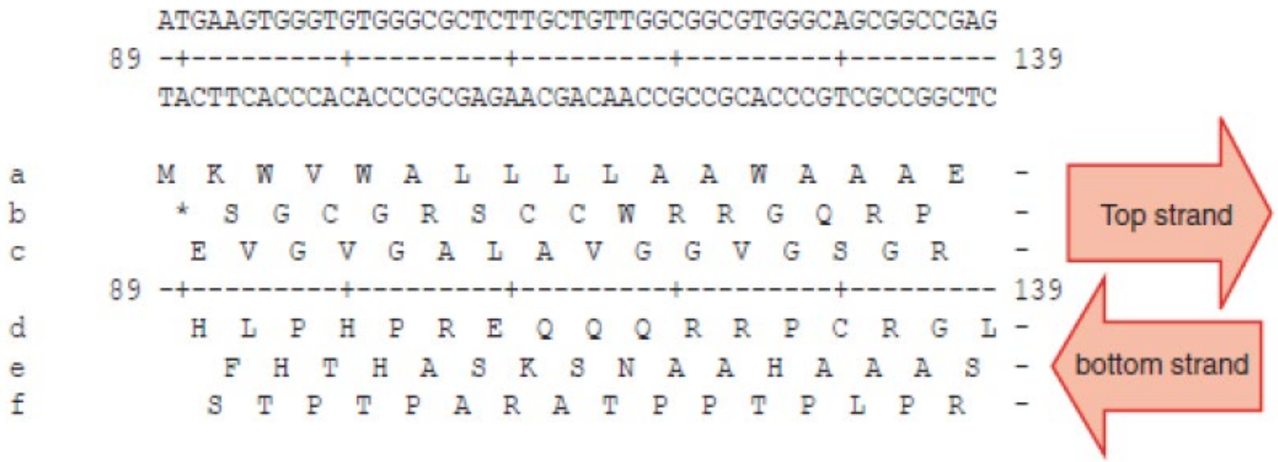
- `CTG TGC TCA GAC CGG`
- `Leu Cys Ser ... Arg`

RNA - In silico vertalen

Expasy translate tool vertaalt een nucleotidesequentie (DNA/RNA) in een eiwitsequentie

- https://web.expasy.org/translate/
- Alle *open reading frames* (ORFs) worden bekeken

- Start met een startcodon (AUG)
- Stopt met een stopcodon (UAA, UAG of UGA)



RNA - In silico vertalen

Gegeven = de humane beta actine (ACTB) mRNA sequentie

- https://www.bioit.be > “human beta actin transcript sequence.fasta”

Oefening:

- Voorspel de mogelijke transcriptie startplaats (Ctrl + F)
  - Inspecteer hiervoor beide strengen (let op: slechts één streng wordt weergegeven)
  - Hoeveel startplaatsen kan je identificeren?
- Gevraagd: gebruik de ExPASy *translate* tool om de mRNA sequentie te vertalen
  - In welk leesraam kan je (waarschijnlijk) de werkelijke eiwitsequentie terugvinden?

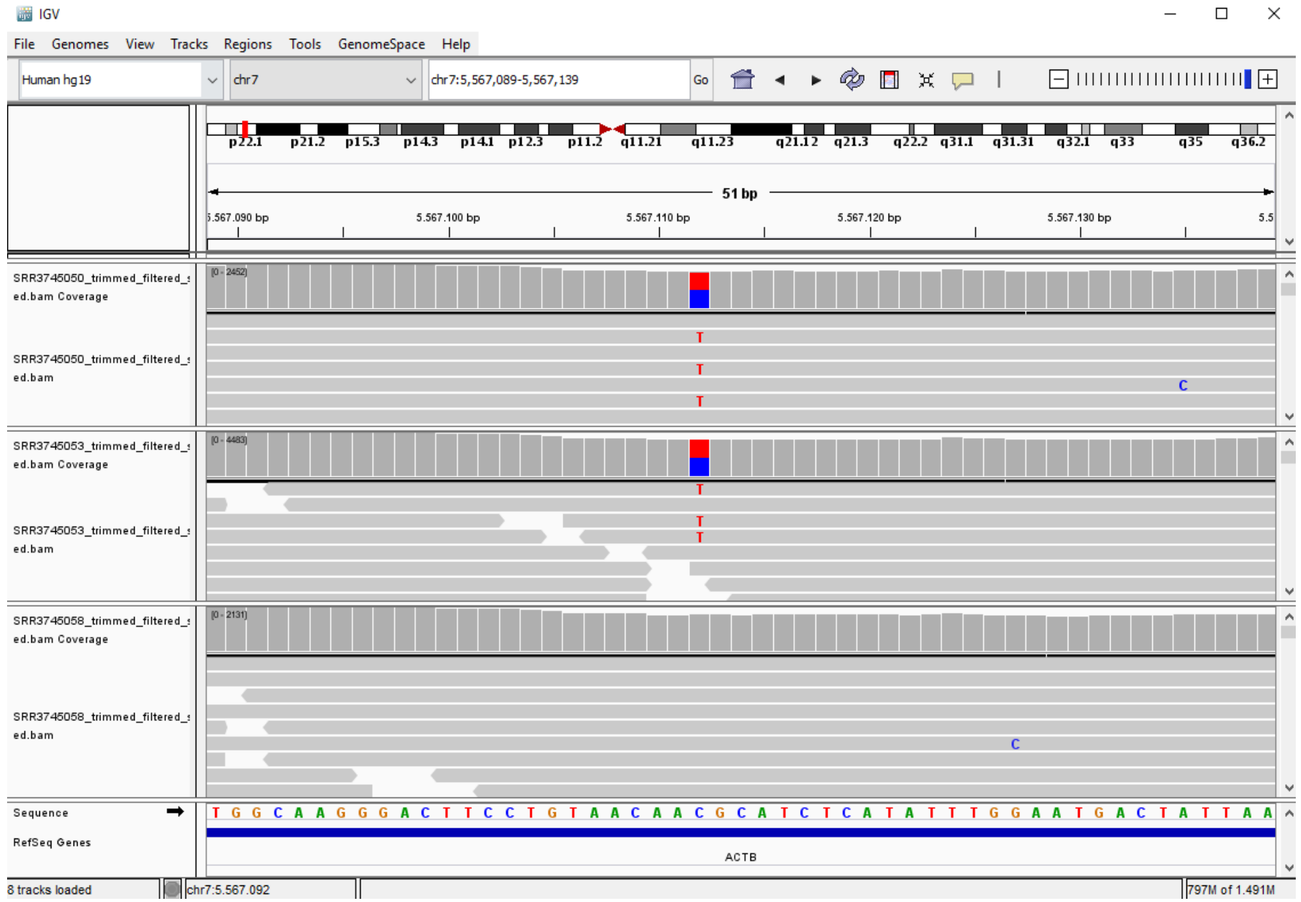
DNA/Eiwit - Variant effect voorspelling

DNA variant analyse:

- Nucleotidesequentie vergelijken met referentiesequentie
- Met NGS data: read mapping

Alle genetische variatie wordt gepubliceerd in databanken:

- NCBI dbSNP (enkel mens)
- https://www.ncbi.nlm.nih.gov/snp/



howest hogeschool

DNA/Eiwit - Variant effect voorspelling

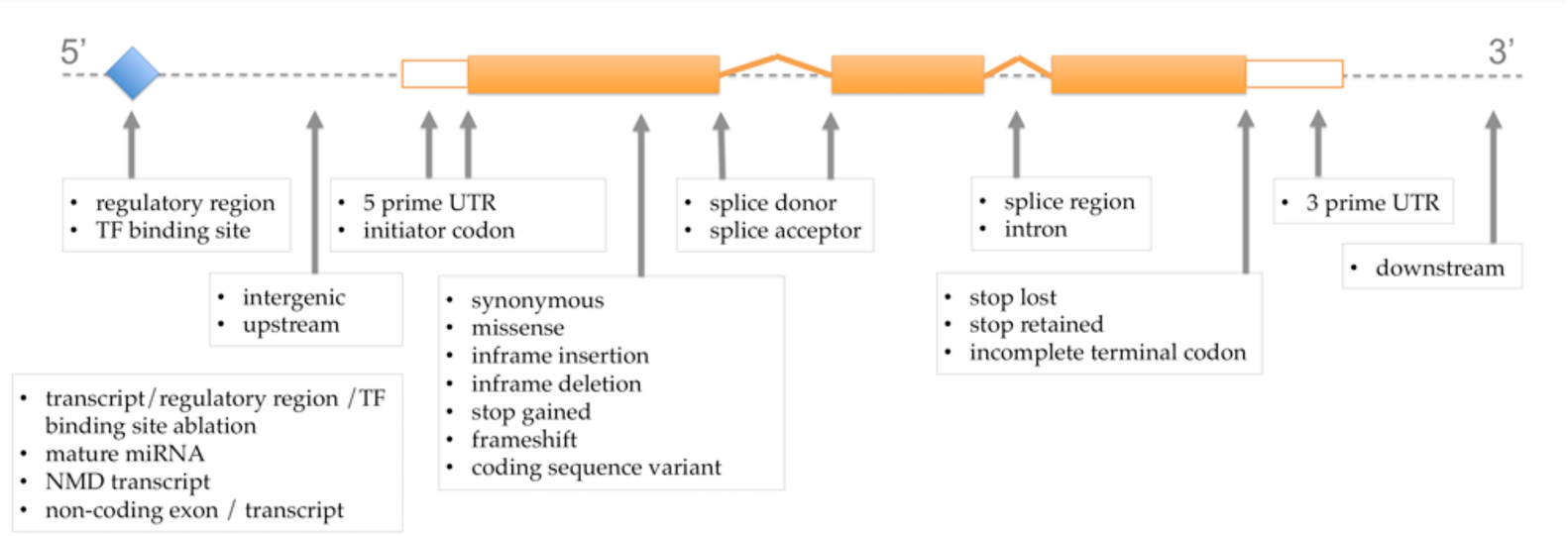
Genetische variatie verkennen:

- Databank van kleine genetische variatie → NCBI dbSNP *(→ BRCA1?)*
- Genome browser voor genetische variatie → Variation Viewer
  - https://www.ncbi.nlm.nih.gov/variation/view/ *(→ BRCA1?)*

DNA/Eiwit - Variant effect voorspelling

Wat is het **effect** van genetische variatie op de structuur/functie van een eiwit?

- Afhankelijk van de locatie van de mutatie/variatie



- Gebruik een fylogenetische methode om consequentie in te schatten

DNA/Eiwit - Variant effect voorspelling

Wat is het **effect** van genetische variatie op de structuur/functie van een eiwit?

- Verschillende tools beschikbaar zoals PROVEAN en SIFT (**s**orts **i**ntolerant **f**rom **t**olerant)

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	**A**AC/**G**AC	0.85
rs33958637	11:5225717-5225717	G	missense_variant	MODERATE	HBB	3043	protein_coding	3/3	-	N/H	**A**AC/**C**AC	0
rs576852971	11:5226131-5226131	G	intron_variant	MODIFIER	HBB	3043	protein_coding	-	2/2	-	-	-

Variant table @ Ensembl genome browser

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

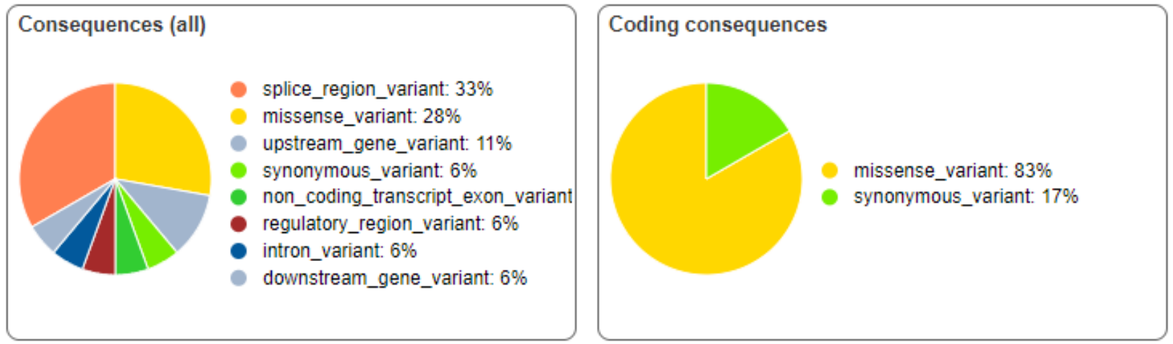
howest hogeschool

DNA/Eiwit - Variant effect voorspelling

Oefening 1: Gebruik VEP om het effect van rs104894719 te onderzoeken:

- Wat is het gevolg van deze *single nucleotide variation* (SNV)?
- Wat is de klinische significantie van de SNV?
- Zoek rs104894719 op in dbSNP, klopt de voorspelling van de VEP?

Oefening 2: Voer dezelfde opdrachten uit voor rs13306510



Slotopmerking

Bio-informatica is veel meer dan databanken, homology searching, genome browsers, *in silico* vertalen en variant effect voorspelling ...

Interesse in meer? **Advanced Bachelor of Bioinformatics (@home)**

ℹ https://howest.be/BIT (2 jaar - 60 ECTS)

Navormingscyclus Toegepaste Bio-informatica in de moleculaire diagnostiek

ℹ https://link.howest.be/toegepasteBIT (1 jaar - 7 volle dagen)

Of iets minder? **Studiedag Excel voor Laboratoriumtechnologen – Basis/Gevorderden**

ℹ https://link.howest.be/excelgevorderden (1 volle dag)

Wordt verwacht… **Data analysis, visualization and biostatistics using R(studio)** (5 ECTS)

howest
hogeschool

European Association for Professions in Biomedical Science
EPBS

Bio-informatica voor dummies
MLT symposium 2022

Jasper Decuyper

howest
BIOINFORMATICS