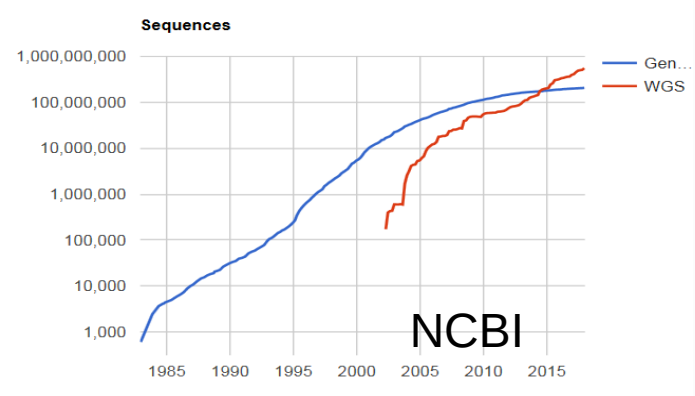
UNIVERSITY OF GOTHENBURG

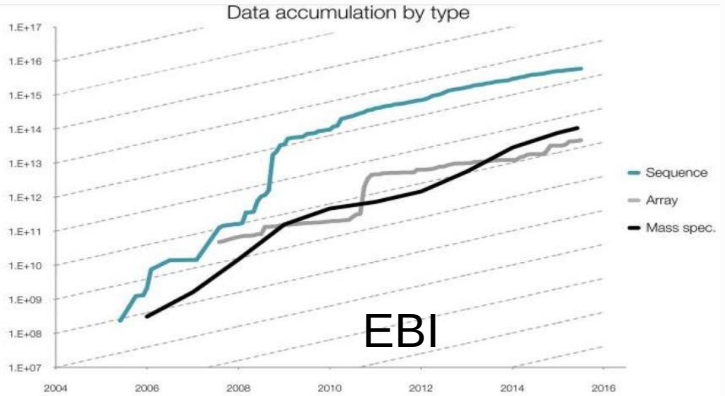
BMA Molekylärbiologisk metodik

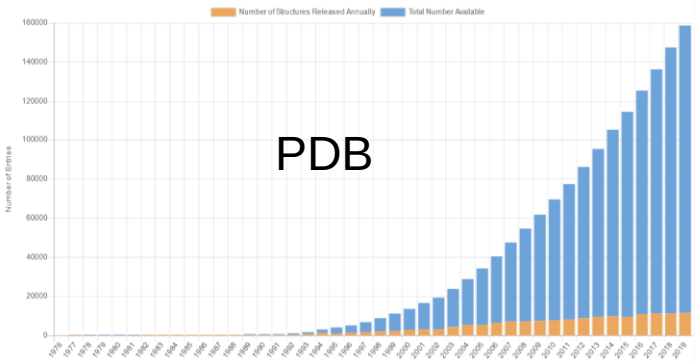
Bioinformatik

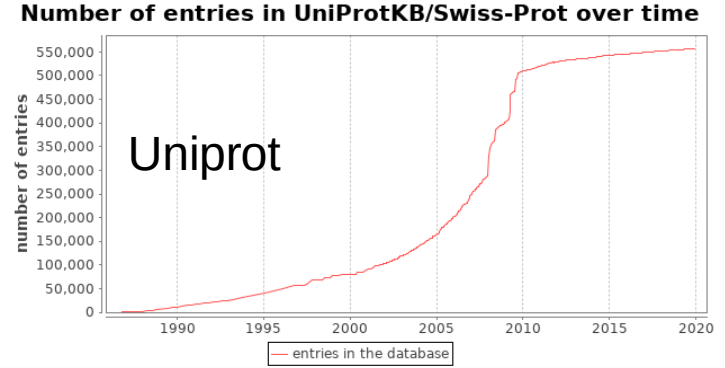
Mars/2021

Explosiv tillväxt av sekvensinformation



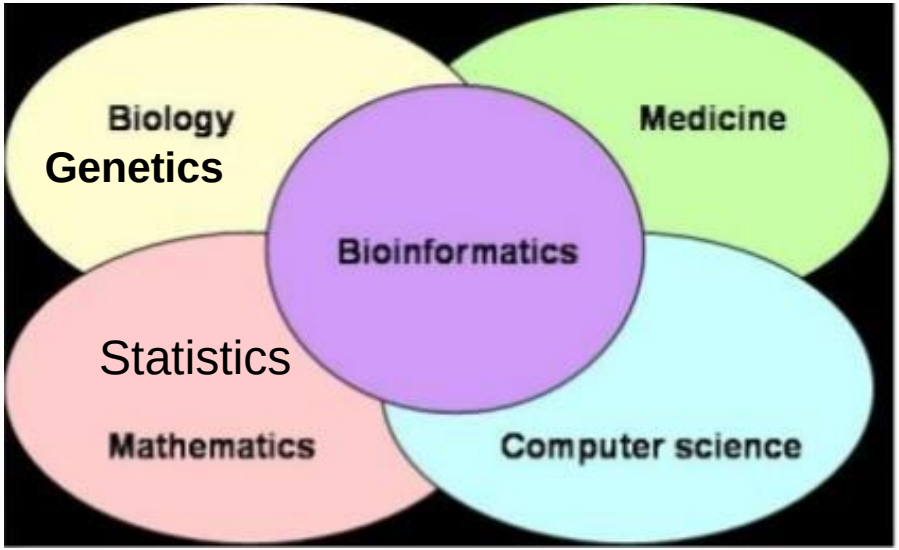






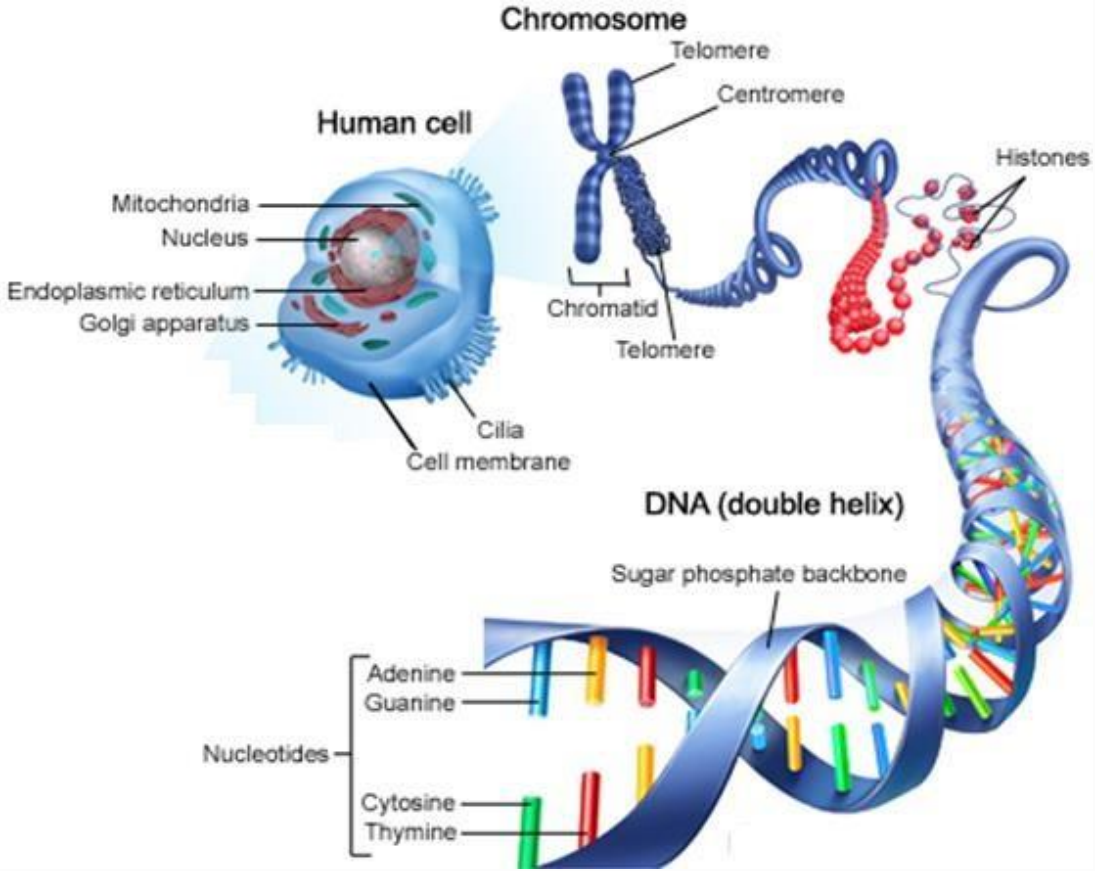
Bioinformatik

Hantering och analys av information från genteknologisk forskning. Tvärvetenskaplig disciplin

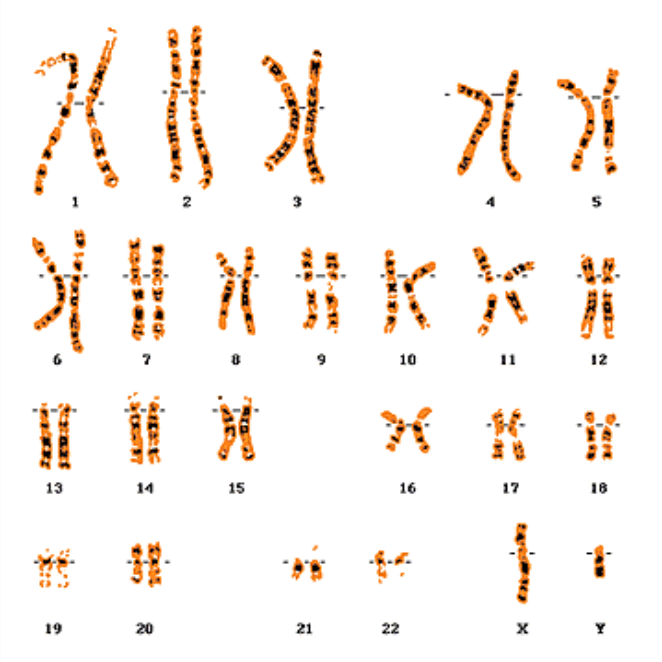


Genom

Genetiskt material som bestämmer de ärvda egenskaperna för en organism



Genomet är uppdelat i -



22 autosomala par & 1 allosom par
Diploid

Species	# Chr	Ploidy
Thale cress	10	Diploid
Bread wheat	42	Hexaploid
Tobacco	48	Tetraploid
Fruit fly	8	Diploid
Earth worm	36	Diploid
Mouse	40	Diploid
Human	46 (2x23)	Diploid
Dog	78	Diploid
Goldfish	100-104	Diploid

Jumper jack

XY -> 1chr
XX -> 2 chr

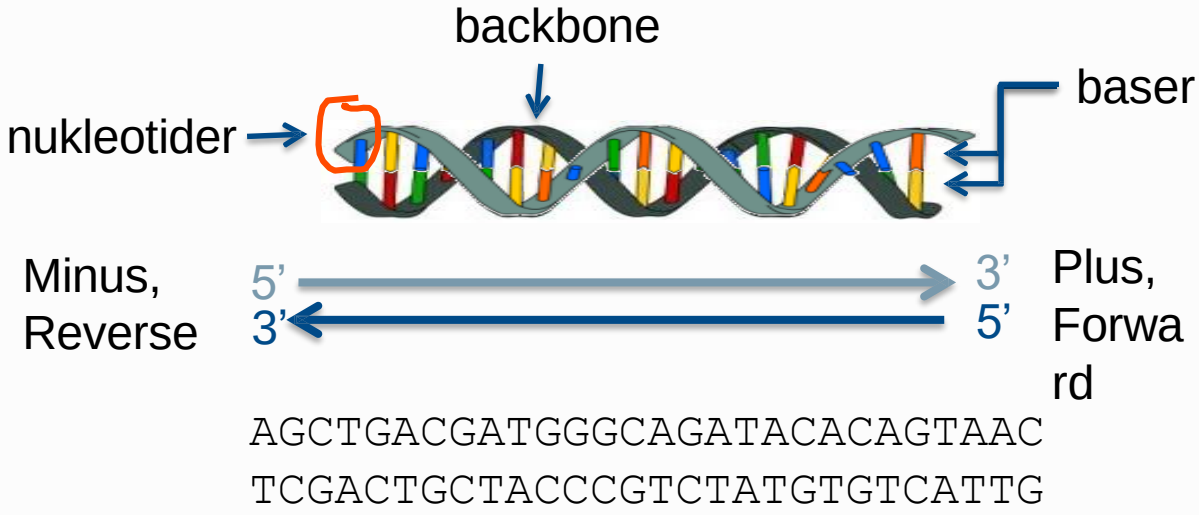
Adder's Tongue

1260 chr

Andra

Type	Organism	Genome size
Virus	P. circovirus	1.8Kb
Virus	Megavirus	48Kb
Bacterium	N. deltocephalinicola	112Kb
Bacterium	E. coli	4.6Mb
Plant	G. margaretae	63Mb
Plant	P. japonica	150GB
Moss	P. patens	480Mb
Yeast	S. cerevisiae	12.1Mb
Nematode	C. elegans	100Mb
Insect	D. melanogaster	130Mb
Fish	T. nigroviridis	390Mb
Fish	P. aethiopicus	130Gb
Mammal	M. musculus	2.7Gb
Mammal	H. sapiens	3.2Gb

DNA



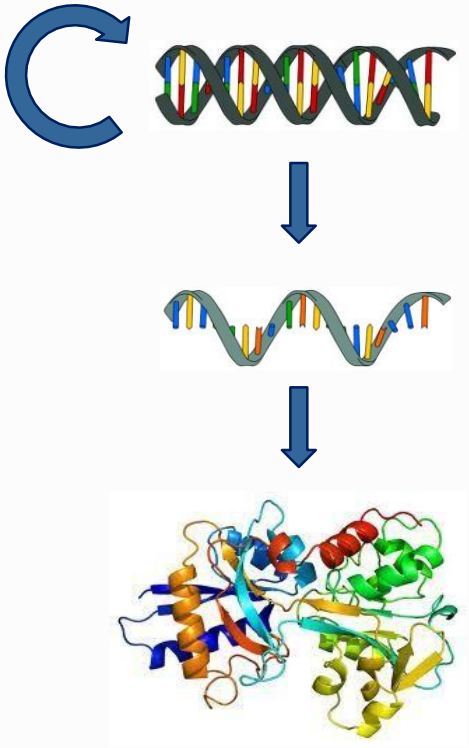
Gener kan ligga på båda strängarna. mRNA sekvensen motsvarar alltid 5- 3’ kodningen av en gen

- Dubbelsträngat
- Antiparralellt
- Komplimentärt

A=T adenine=thyamine

C≡G cytosine≡guanine

Central Dogma: Flödet av genetisk information

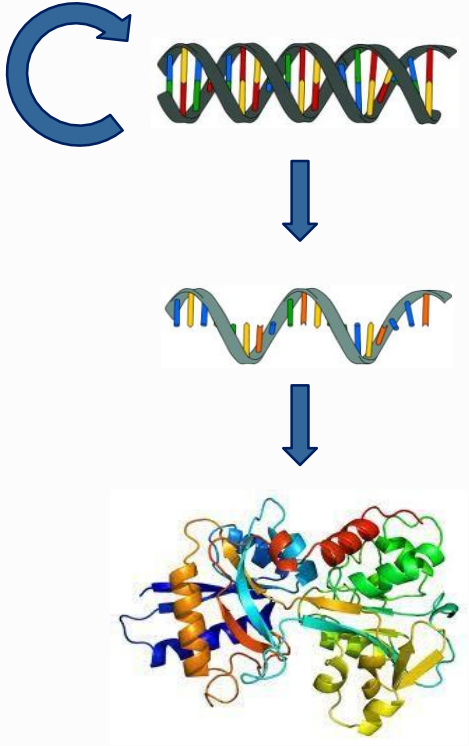


Replikation: DNA till DNA

Transkription: DNA till mRNA

Translation: mRNA till protein

Från mRNA till protein



5′ AGTACCATACTTCTGGCCTGGGCGAAAGAATAA 3′

3′ TCATGGTATGAAGACCGGACCCGCTTTCTTATT 5′

5′ AGUACCAUACUUCUGGCCUGGGCGAAAGAAUAA 3′

S T I L L A W A K E *

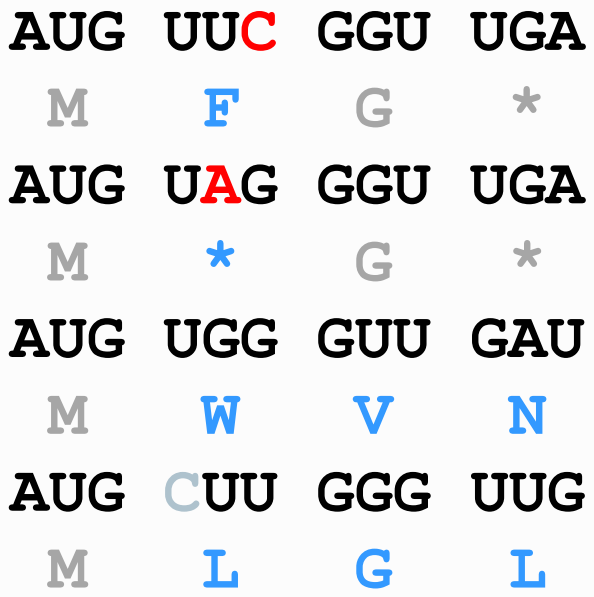
CODON: 3 baser av ett reads som läses av tRNA för att bli translaterat till en aminosyra

Den genetiska koden

UUU Phe F UUC Phe F UUA Leu L UUG Leu L	UCU Ser S UCC Ser S UCA Ser S UCG Ser S	UAU Tyr Y UAC Tyr Y UAA Stop UAG Stop	UGU Cys C UGU Cys C UGA Stop UGG Trp W
CUU Leu L CUC Leu L CUA Leu L CUG Leu L	CCU Pro P CCC Pro P CCA Pro P CCG Pro P	CAU His H CAC His H CAA Gln Q CAG Gln Q	CGU Arg R CGC Arg R CGA Arg R CGG Arg R
AUU Ile I AUC Ile I AUA Ile I AUG Met M	ACU Thr T ACC Thr T ACA Thr T ACG Thr T	AAU Asn N AAC Asn N AAA Lys K AAG Lys K	AGU Ser S AGC Ser S AGC Arg R AGG Arg R
GUU Val V GUC Val V GUA Val V GUG Val V	GCU Ala A GCC Ala A GCA Ala A GCG Ala A	GAU Asp D GAC Asp D GAA Glu E GAG Glu E	GGU Gly G GGC Gly G GGA Gly G GGG Gly G

Translation

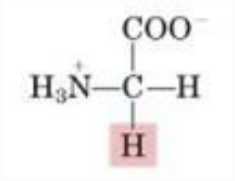
- Diff. nt sekvenser kan översättas till identiska aa sekvenser
- Substitutioner, deletioner, insertioner



Proteiner

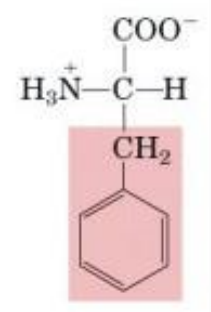
Lång kedja av aminosyror som är nödvändiga för levande celler

G (Gly)



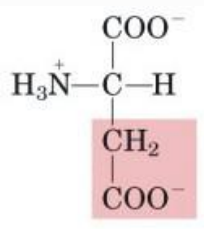
Glycine

F (Phe)



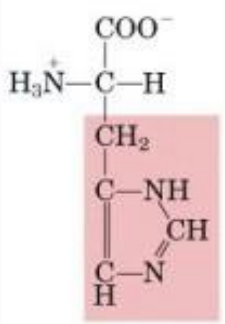
Phenylalanine

D (Asp)



Aspartate

H (Hist)



Histidine

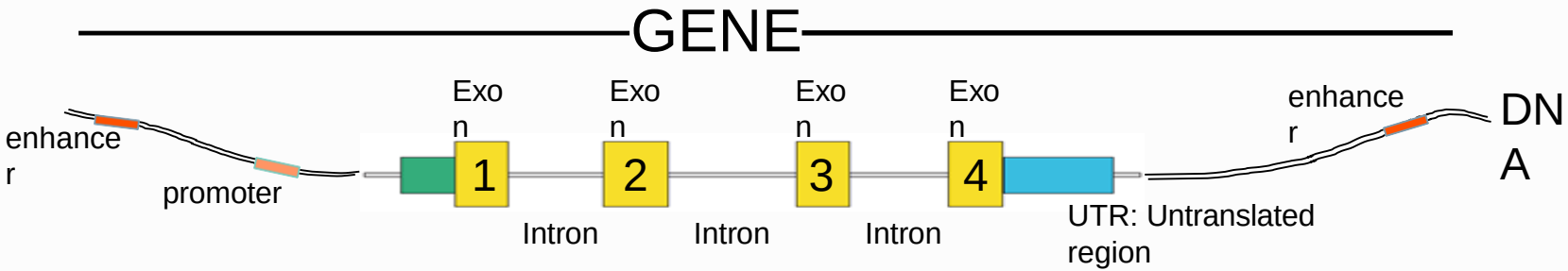
Side chains

20 standard aminosyror

Olika cellulära funktioner: Regulering, struktur, rörelse, katalys, transport, signalering

Gener

Delen av DNAt som innehåller all information för att producera ett protein.

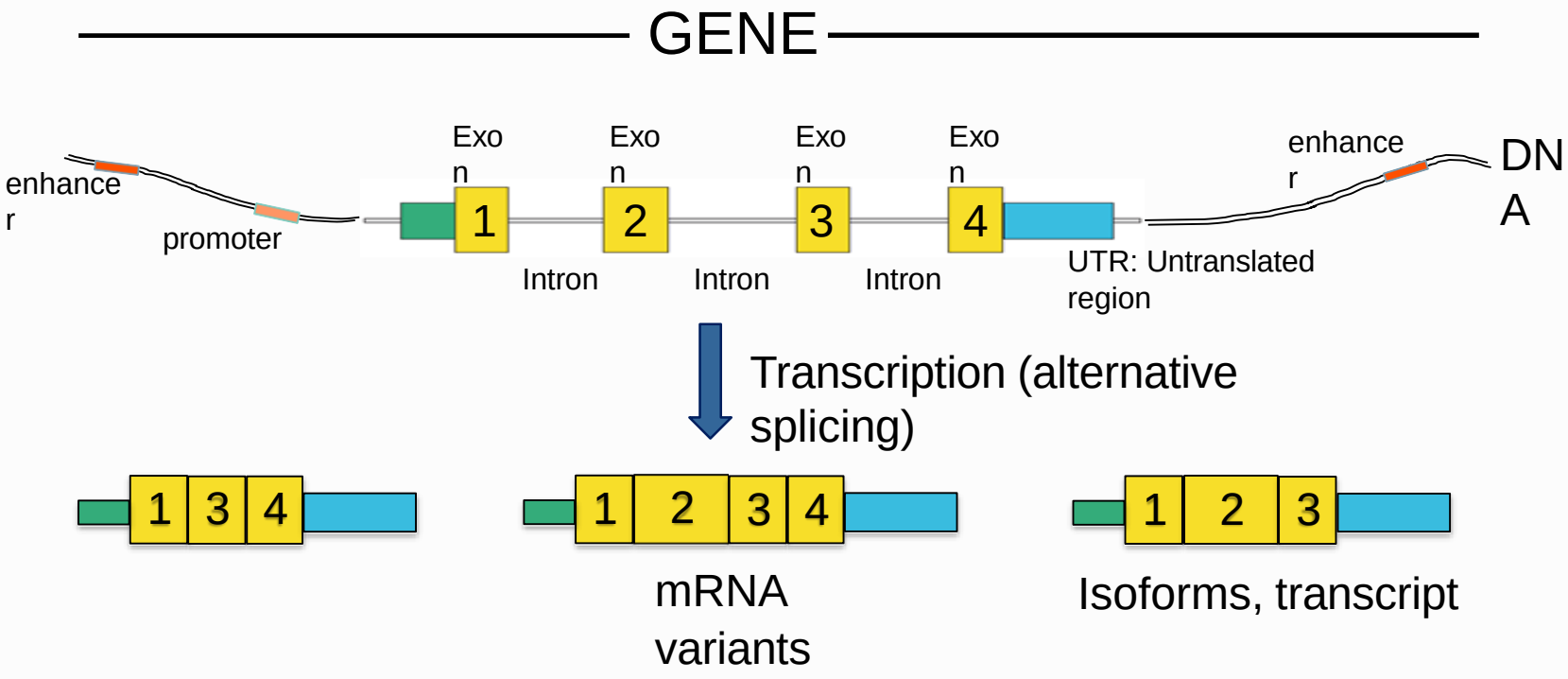


20,000 protein-kodande gener (~1.5%)

Ojämnt distruberade över alla kromosomer

HIST1HIA	781 nt	brist på introner
DMD	2.2 Mb (79 exons)	största genen
TTN	80 Kb (364 exons)	längsta kodande sekvensen

Kodande sekvensen



Davila Lopez, M., Rosenblad, MA and T. Samuelsson, (2008)

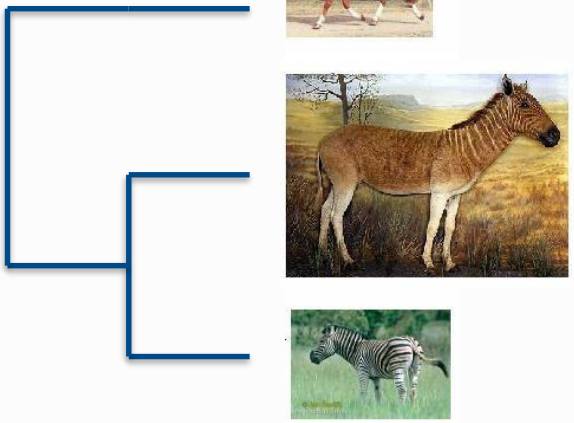
Komparativ genomik

Är vi verkligen så olika?

Leonard, J.A., et al., (2005)

Multiple sequence alignment

```
quagga          GGATTCGTTCACTGATTCCCTCTATTCTCAGGATACACACTCAACCAAACCTGAGCAAAA
zebra           GGATTCGTTCACTGATTCCCTCTATTCTCAGGATACACACTCAACCAAACCTGAGCAAAA
horse           GGATTTGTACACTGATTCCCTCTATTCTCAGGATACACACTCAACCAAACCTGAGCAAAA
                ***** ** ************************************************

quagga          ATTCACTTTACAATTATATTCGTAGGGGTCAACATAATTTTCTTCCCA
zebra           ATTCACTTTACAATTATATTCGTAGGGGTCAATATAACTTTCTTCCCA
horse           ATCCACTTTACAATTATATTCGTAGGGGTAAATATAACCTTCTTCCCA
                ** ************************** ** ****  *********
```



Viktiga data repos

Nukleotid:

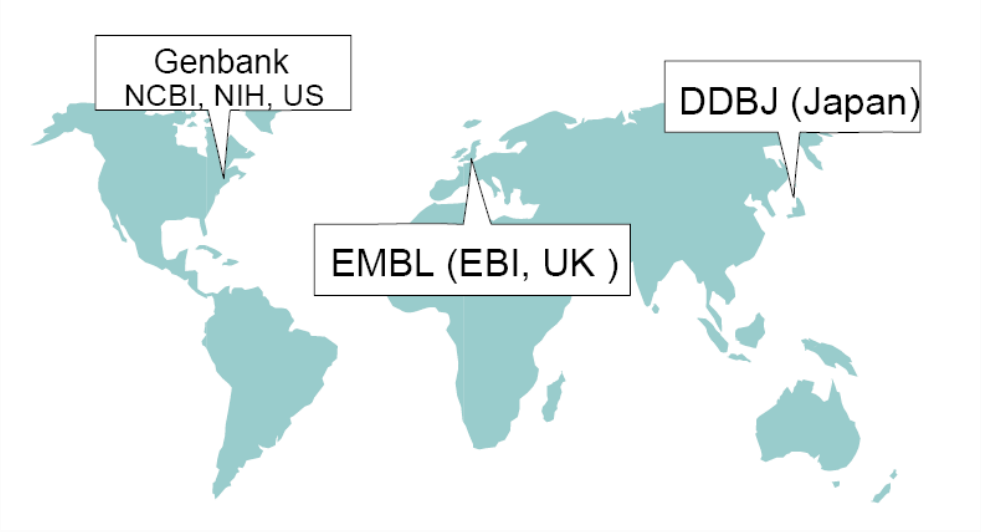
- GenBank - NCBI
- EMBL - EBI
- DDBJ - NIG

Protein:

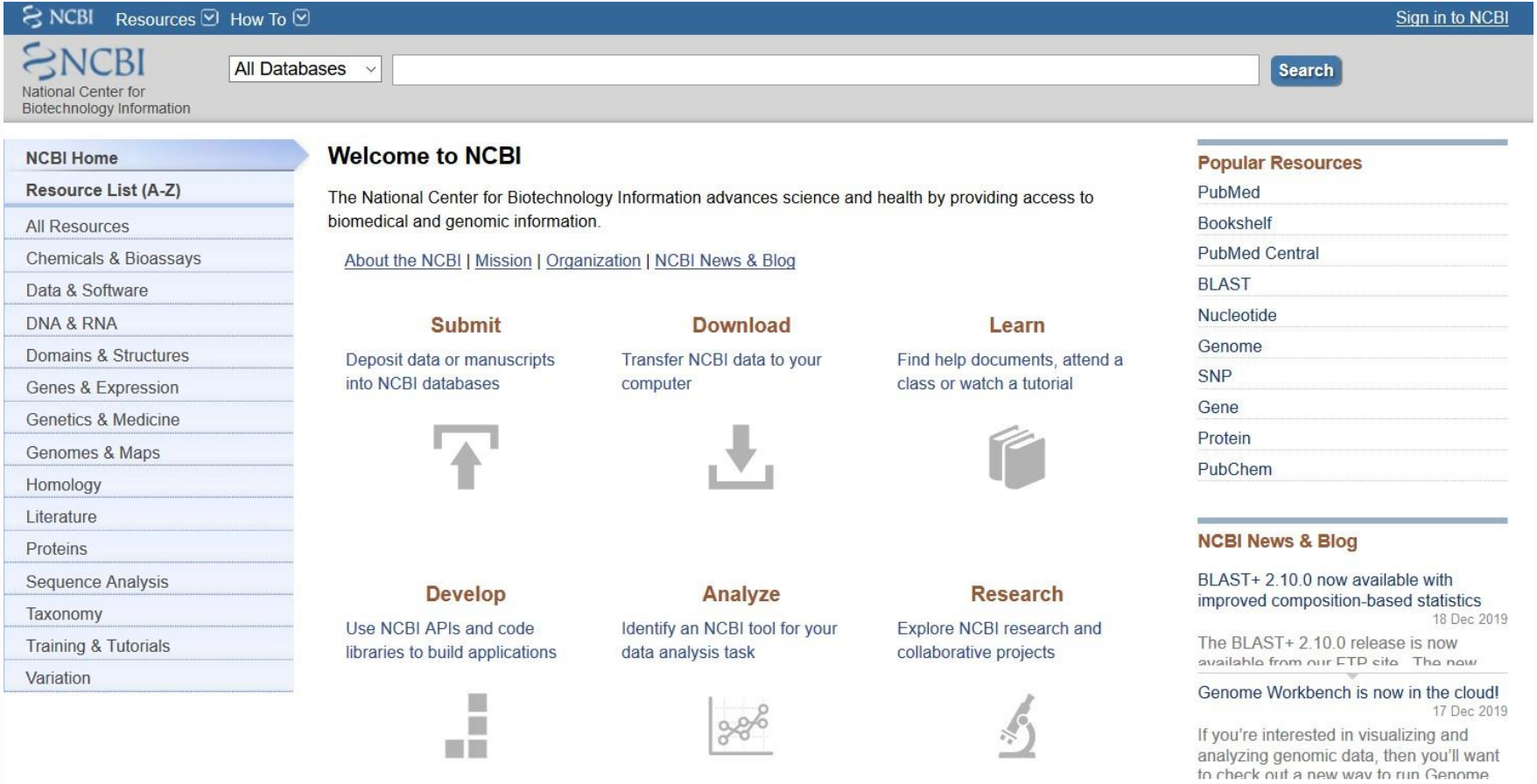
- UniProt – SIB, EBI and PIR

3D strukturer

- PDB - RCSB



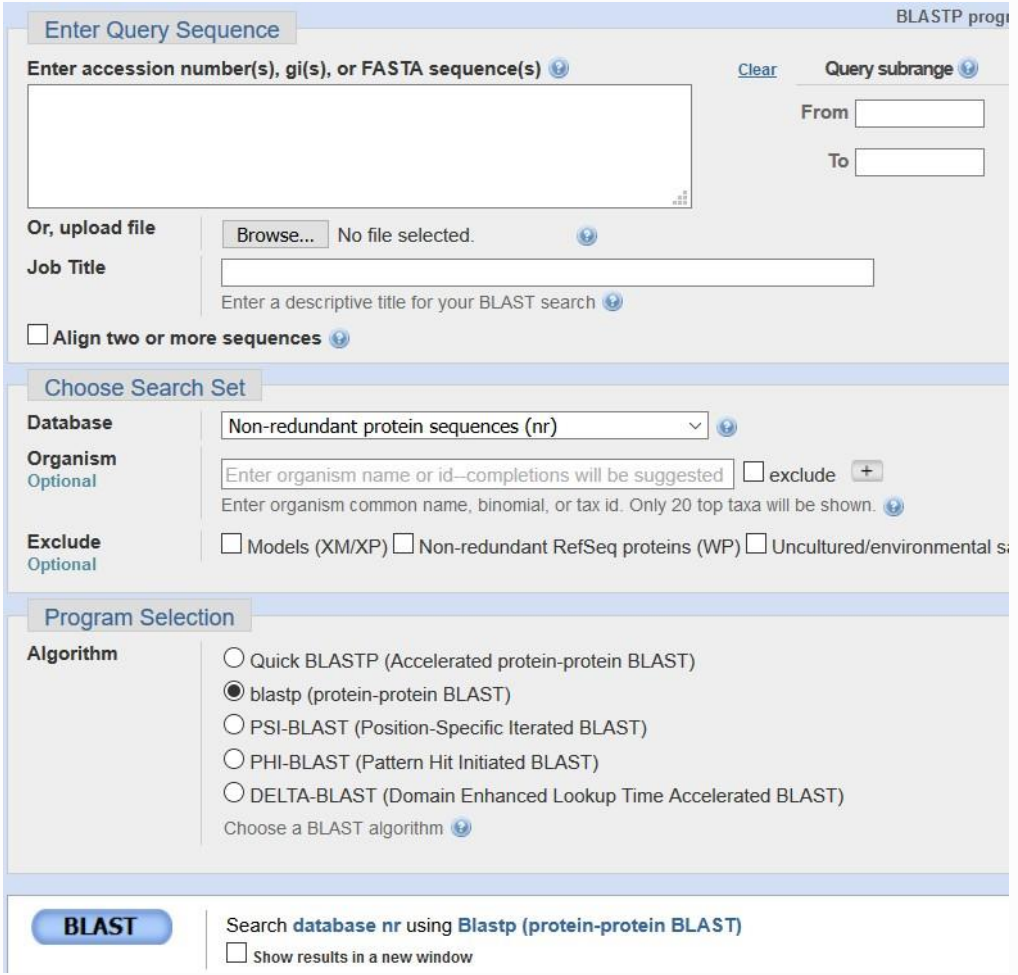
Ensembl
MGI, SGD, PlasmoDB,
JGI, …. TCGA, GEO, 1000GP,
…


NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build applications
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI News & Blog
BLAST+ 2.10.0 now available with improved composition-based statistics
18 Dec 2019
The BLAST+ 2.10.0 release is now available from our FTP site. The new
Genome Workbench is now in the cloud!
17 Dec 2019
If you're interested in visualizing and analyzing genomic data, then you'll want to check out a new way to run Genome

BLAST –Basic Local Alignment Search Tool

Ett verktyg som används för att jämföra sekvensinformation.

Klistra in en sekvens och hitta liknande sekvenser



BLAST

Descriptions | Graphic Summary | Alignments | Taxonomy

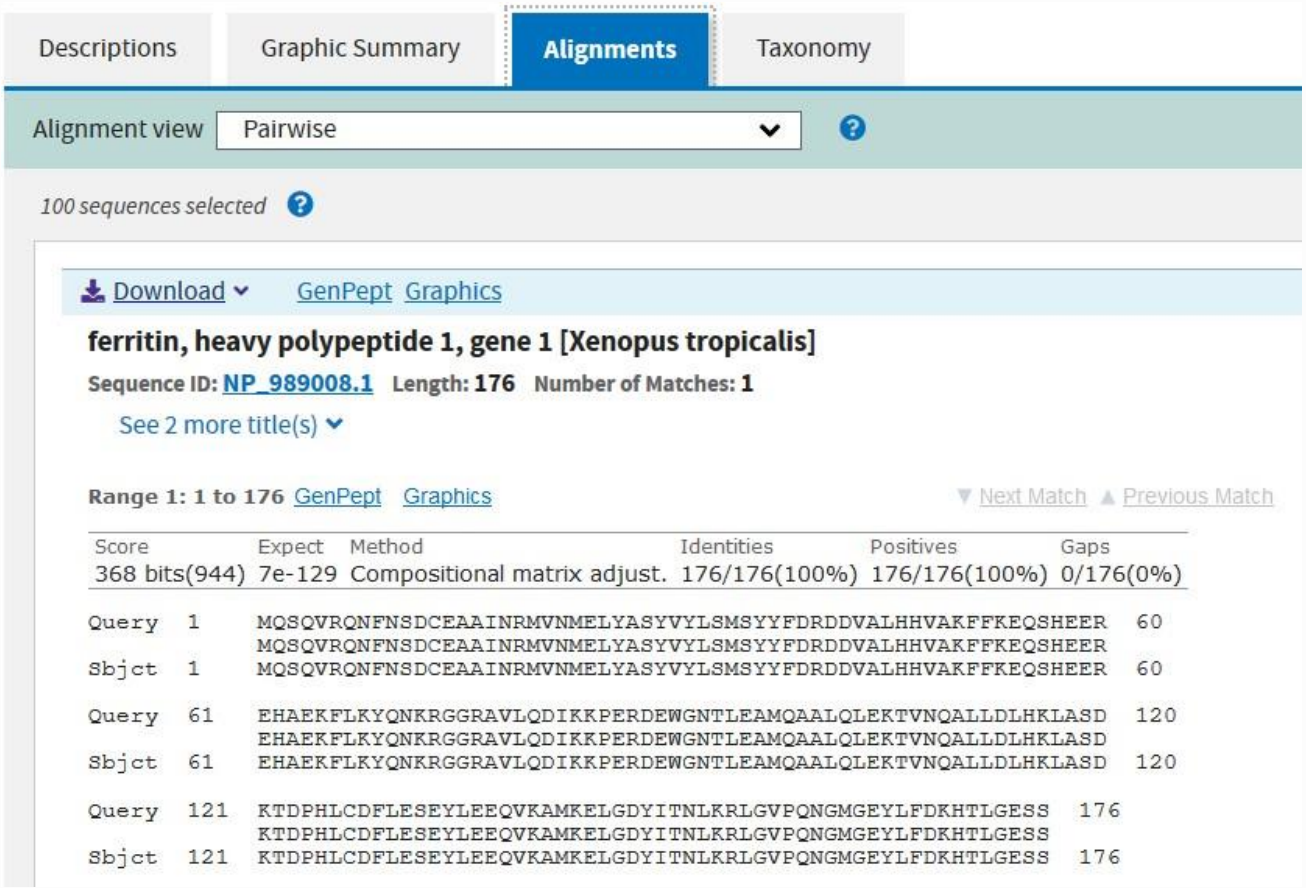
Sequences producing significant alignments — Download — Manage Columns — Show 100

select all *100 sequences selected* — GenPept — Graphics — Distance tree of results — Multiple alignment

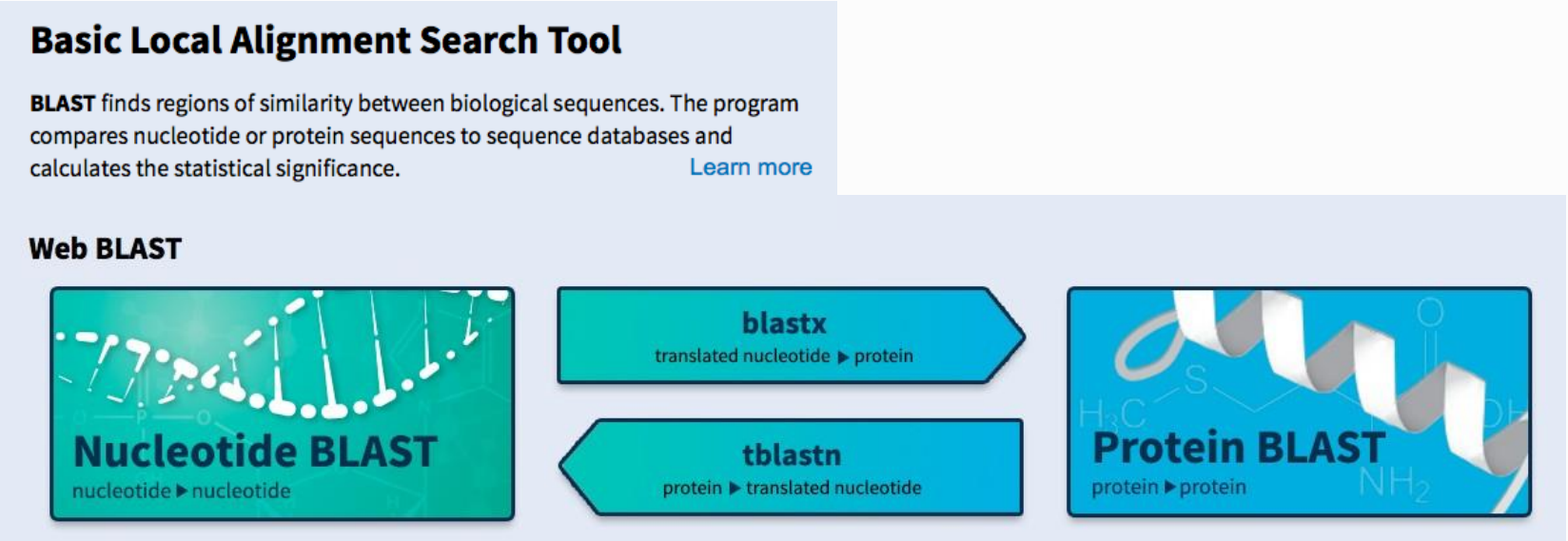
Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
ferritin, heavy polypeptide 1, gene 1 [Xenopus tropicalis]	368	368	100%	7e-129	100.00%	NP_989008.1
hypothetical protein XENTR_v10019363 [Xenopus tropicalis]	360	360	100%	1e-125	95.65%	KAE8593888.1
hypothetical protein XELAEV_18037841mg [Xenopus laevis]	358	358	100%	3e-125	97.16%	OCT70916.1
ferritin heavy chain B [Xenopus laevis]	357	357	100%	1e-124	96.59%	NP_001079580.1
ferritin heavy chain 1 a [Xenopus laevis]	356	356	100%	3e-124	96.02%	NP_001090207.1
ferritin heavy chain A [Xenopus laevis]	355	355	100%	1e-123	96.02%	NP_001084057.1
unnamed protein product [Xenopus laevis]	342	342	100%	8e-119	94.32%	CAA35760.1
ferritin heavy chain A [Microcaecilia unicolor]	342	342	100%	1e-118	91.48%	XP_030053805.1

Mest lik sekvens i toppen av listan. Hög score och E-värde nära 0 betyder att det är liten risk att sekvensen av ren slump är lik ditt query

BLAST resultat



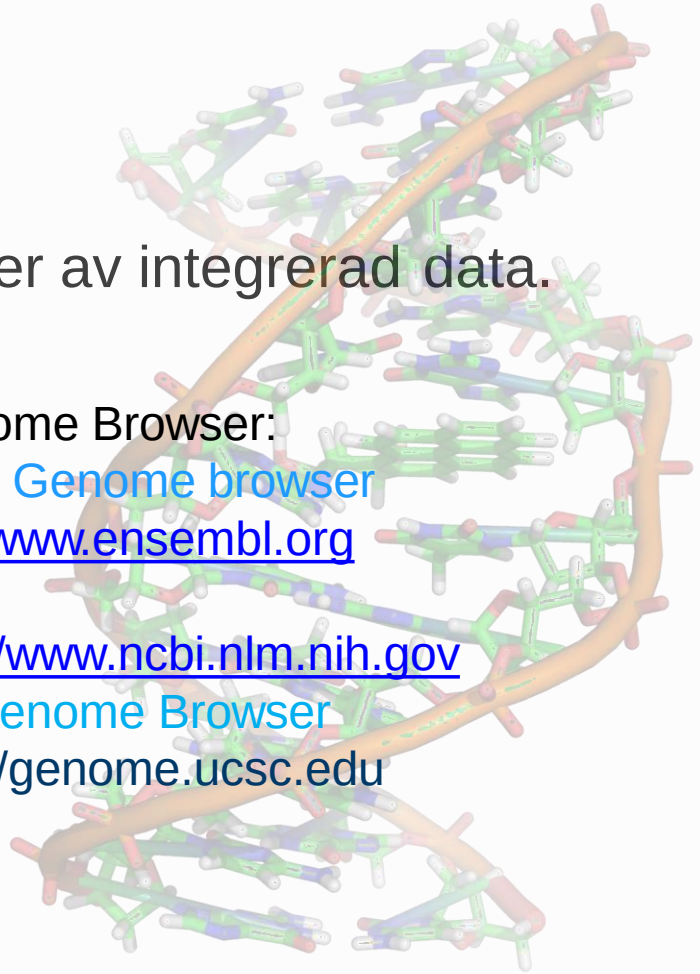
Olika typer av BLAST



Genome Browsers

- Genome browsers visualiserar olika typer av integrerad data.

  - Gener
  - Proteiner
  - Expression
  - Regulering
  - Variation
  - Komperativ genomik



Big Genome Browser:

Ensembl Genome browser
http://www.ensembl.org

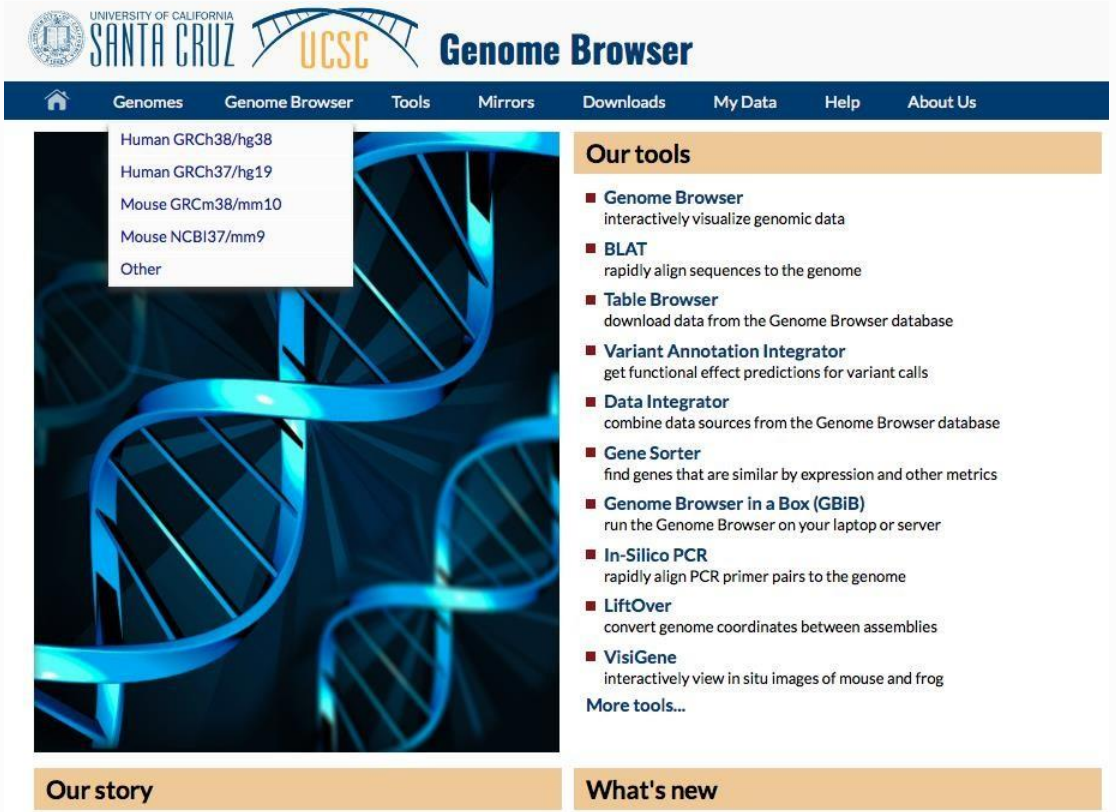
NCBI
https://www.ncbi.nlm.nih.gov

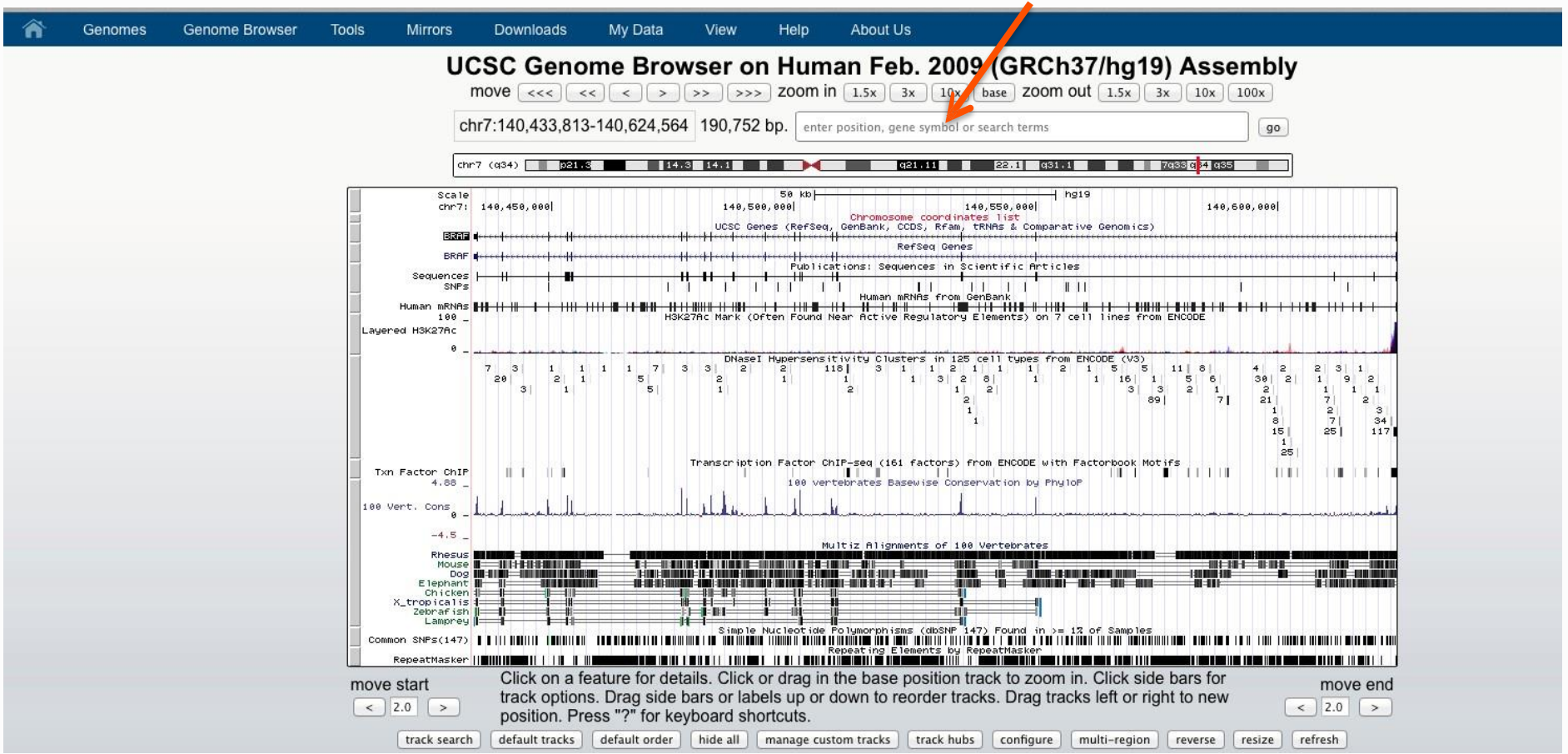
UCSC Genome Browser
https://genome.ucsc.edu

UCSC Genome Browser

NOTE! Genom coordinater skiljer sig mellan versioner

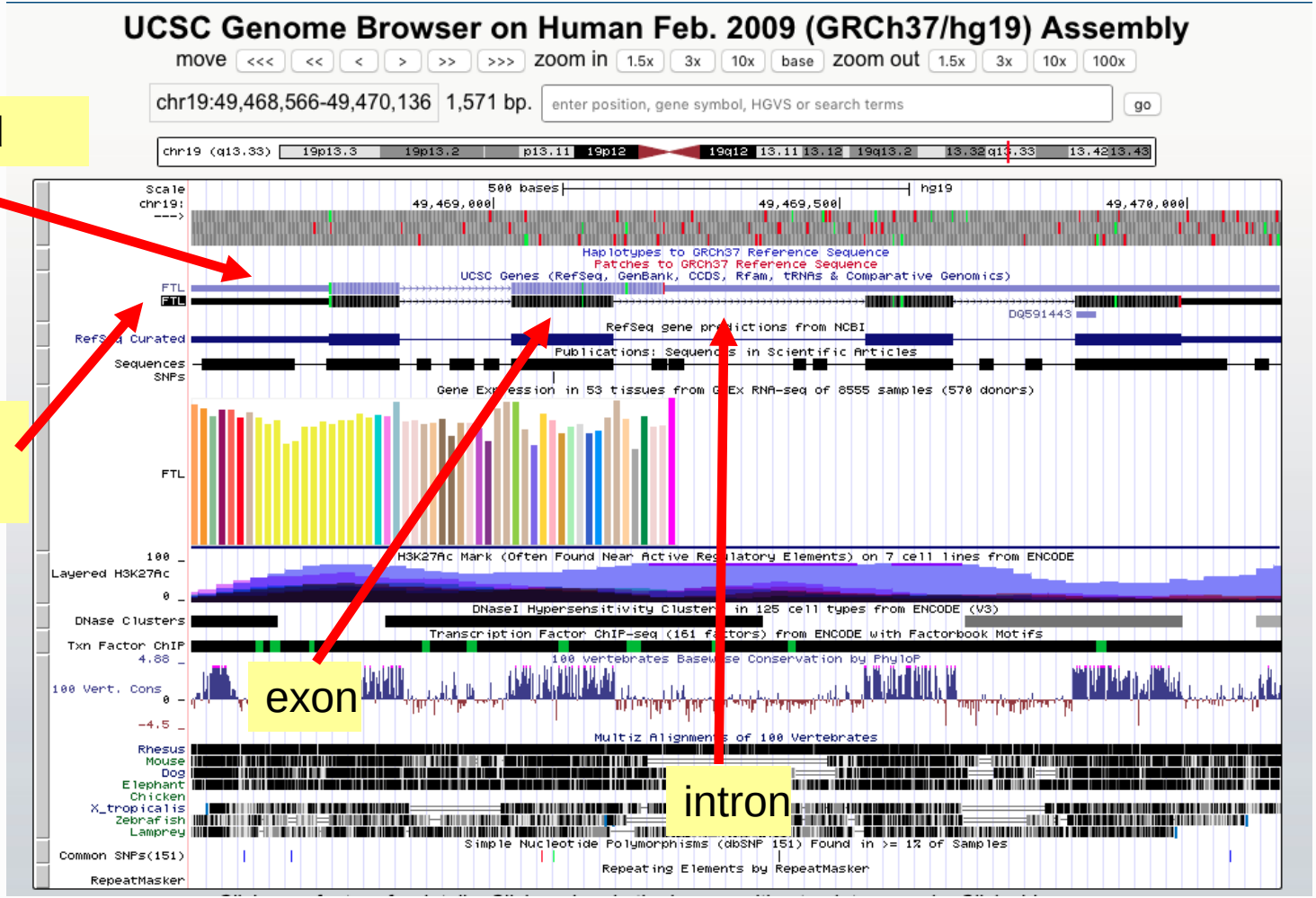
Online genome browser med access till genomsekvensdata ifrån många olika organismer.



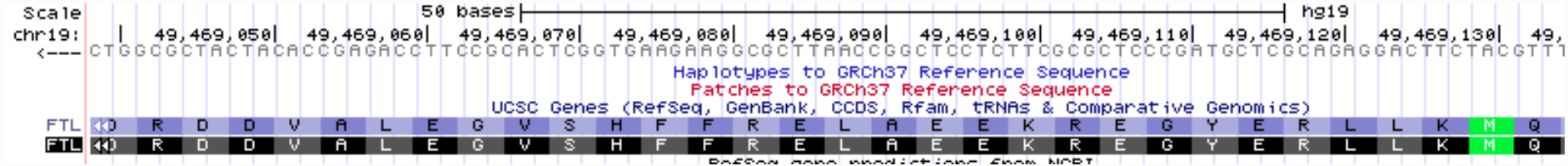

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr7:140,433,813-140,624,564 190,752 bp.
enter position, gene symbol or search terms
go
chr7 (q34)
Scale
chr7:
50 kb
hg19
140,450,000
140,500,000
140,550,000
140,600,000
Chromosome coordinates list
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
BRAF
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 147) Found in >= 1% of Samples
Common SNPs(147)
Repeating Elements by RepeatMasker
RepeatMasker
move start
move end
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.
track search
default tracks
default order
hide all
manage custom tracks
track hubs
configure
multi-region
reverse
resize
refresh

Untranslated region

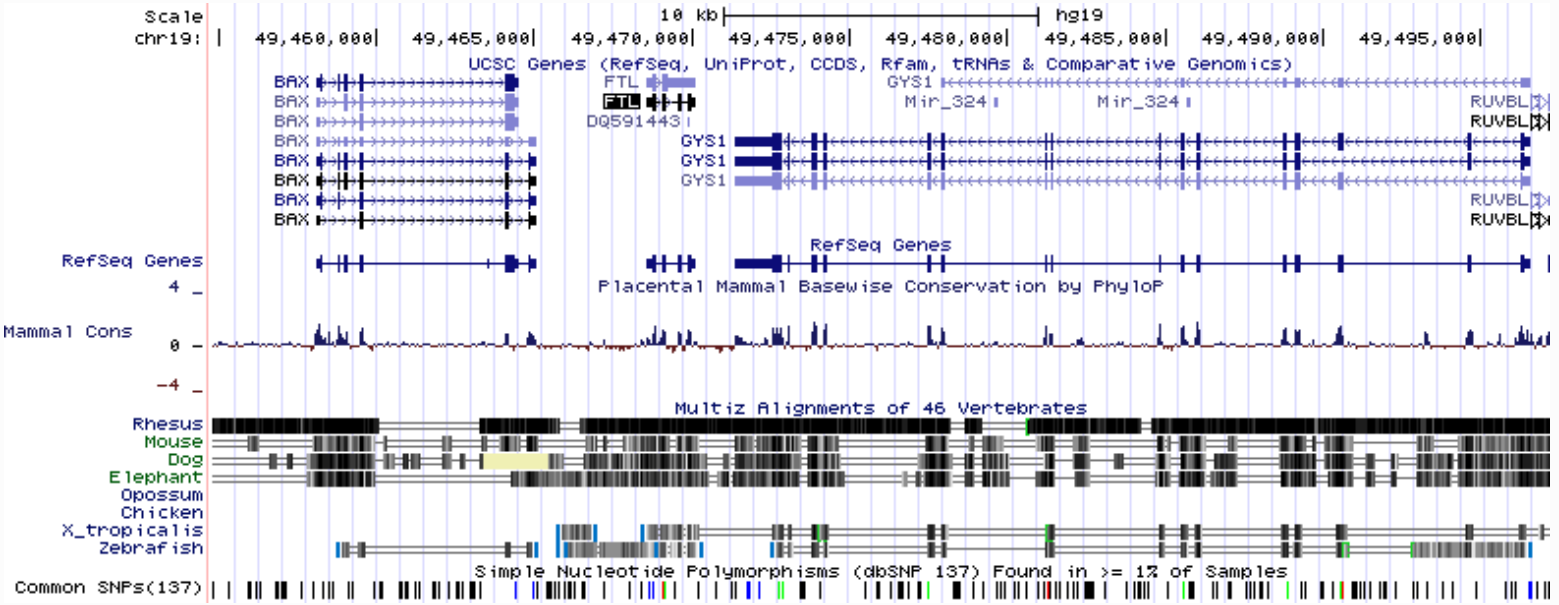
Gene(s) (splice variants)

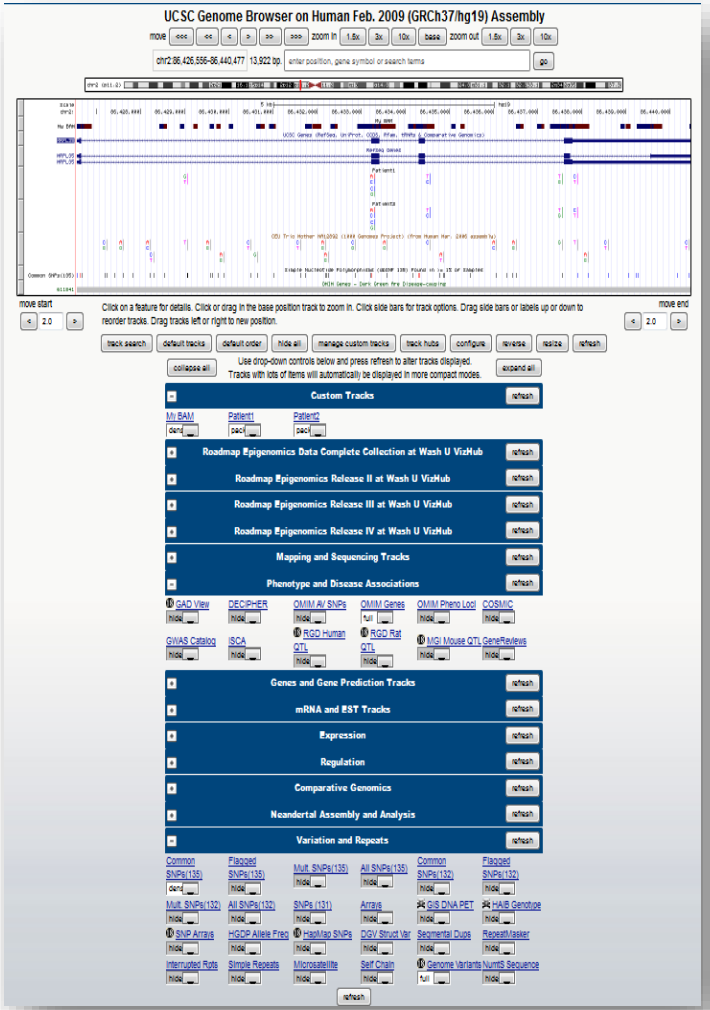


Zooma in:



Zooma ut:

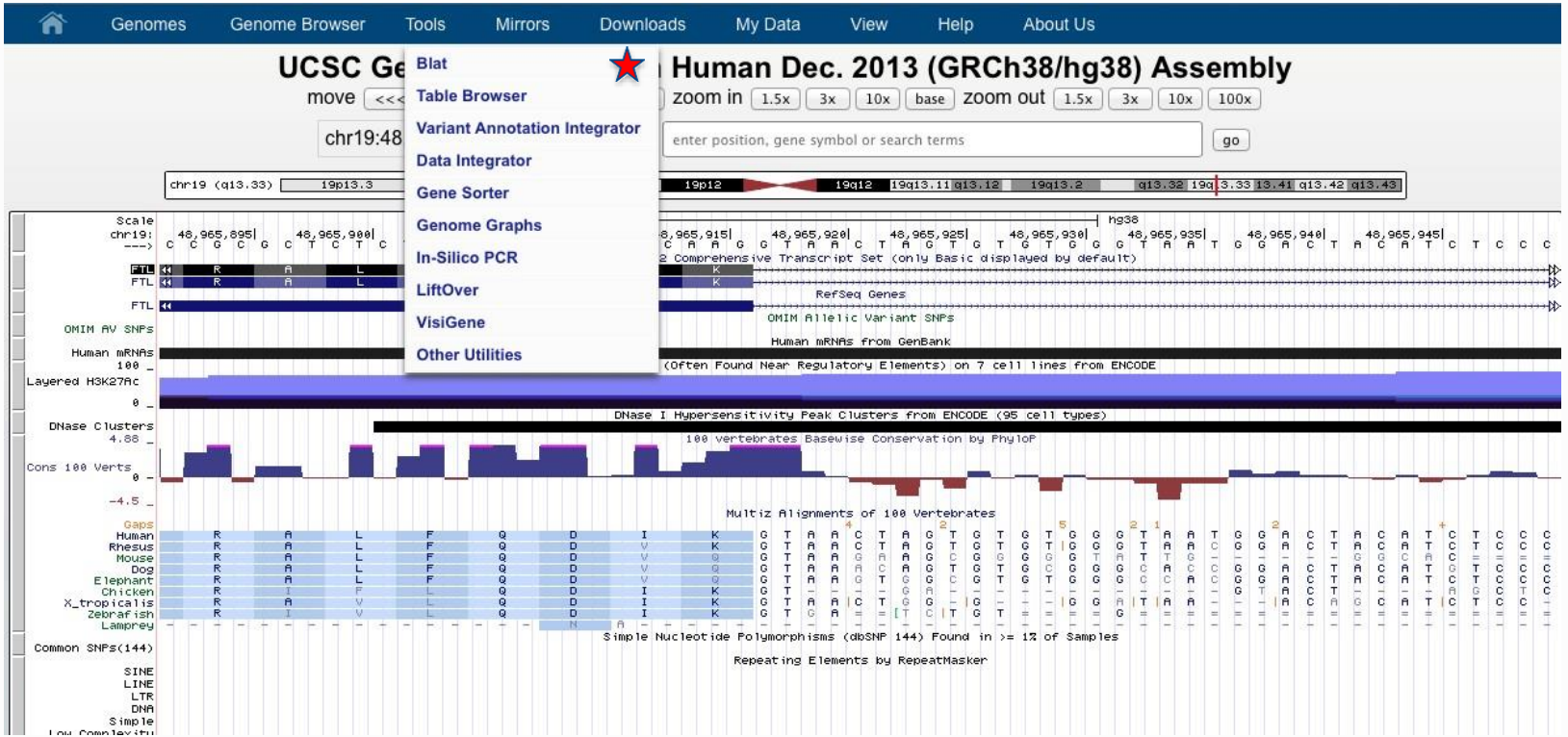




Genome viewer

Custom Annotation tracks (grupper av data)

- Mappning och sekvenser
- Genprediktioner
- Variationer och repeats
- komparativ genomik
- Custom tracks (BAM, VCF)



BLAT har en liknande funktion som BLAST, men behöver en nästan exakt match för att hitta rätt. BLAT passar inte om man vill hitta avlägset relaterade sekvenser.

Variation mellan individer

- Single nucleotide polymorphism (SNP)
- Två mänskliga genom är i genomsnitt 99,9% lika
- Genom är stora så innebär ca 3 miljoner skillander, men de flesta av dessa har ingen funktion
- De flesta varianterna 90% är i form av förändring av en nukleotid SNP

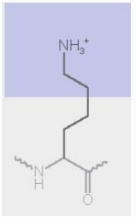
Single nucleotide variants (SNV)

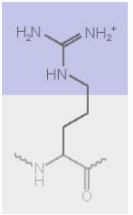
* Substitution
* Insertion

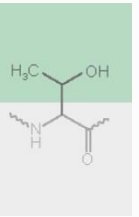
* Deletion

* Non-coding
* Coding
    - Synonymous
    - Non-synonymous
        - Misense
        - Nonsense

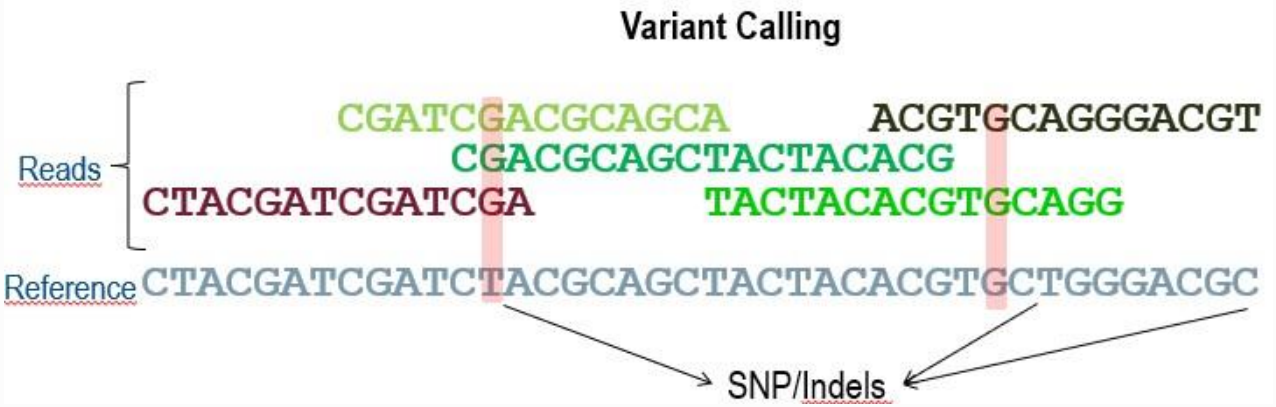
```
CTGAAG   CTGA---AG   (ref)
CTGATG   CTGACTGAG

         CTGAAG      (ref)
         CTG--G

AAG   Lys
AAA   Lys

AGG   Arg
ACG   Thr
UAG   STOP
```







Identifiering av SNVs



Variant Annotation

- Är det en känd SNP?
- Hur vanlig är den i befolkningen?
- Ligger varianten i en gen?
- Vilken påverkan kan varianten ha på genens funktion?
- Vad vet vi om den genens funktion idag?

VCF: Variant Call format

Header

```
##fileformat=VCFv4.1
##FILTER=<ID=FSindel,Description="FS > 200.0">
##FILTER=<ID=FSsnp,Description="FS > 60.0">
##FILTER=<ID=LowDP,Description="DP < 10.0">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseCounts,Number=4,Type=Integer,Description="Counts of each base">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr2,length=243199373,assembly=hg19>
##contig=<ID=chr3,length=198022430,assembly=hg19>
##reference=file:///jumbo/db/GATK/Minusy/ucsc.hg19.minusMY.fasta
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	C2463 S1
chr1	1671852	.	G	C	143.77	LowDP;MQD	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:0,1:99:172,0,296
chr1	1671853	.	C	T	143.77	LowDP;MQD	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:0,1:99:172,0,296
chr1	1671858	.	TCTCGC	T	134.73	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,1:99:172,0,296
chr1	1671866	.	A	AGAGAT	134.73	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,1:99:172,0,296
chr1	1671880	.	G	C	143.77	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,1:99:172,0,296
chr1	1684347	.	C	CCCT	391.79	PASS	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,9:19:429,0,19
chr1	1695462	.	C	A	537.77	PASS	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:21,21:99:566,0,618
chr1	1695574	.	G	T	532.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,16:48:561,48,0
chr1	1695742	.	C	T	34.78	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,2:28:63,0,28
chr1	1695831	.	C	T	61.28	LowDP	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,3:9:89,9,0
chr1	1710035	.	C	T	75.28	LowDP	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,3:9:103,9,0
chr2	123332	.	C	G	521.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,16:48:550,48,0
chr2	128345	.	C	T	4286.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,127:99:4315,378,0
chr2	130618	.	G	C	340.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,12:36:369,36,0
chr2	132469	.	C	A	4888.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,82:99:4917,256,0
chr2	132490	.	C	G	4888.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,62:99:4917,256,0
chr2	132696	.	A	G	144.90	LowDP	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,5:15:173,15,0

Body

VCF: Filter

Header

```
##fileformat=VCFv4.1
##FILTER=<ID=FSindel,Description="FS > 200.0">
##FILTER=<ID=FSsnp,Description="FS > 60.0">
##FILTER=<ID=LowDP,Description="DP < 10.0">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseCounts,Number=4,Type=Integer,Description="Counts of each base">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr2,length=243199373,assembly=hg19>
##contig=<ID=chr3,length=198022430,assembly=hg19>
##reference=file:///jumbo/db/GATK/Minusy/ucsc.hg19.minusMY.fasta
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  C2463_S1
chr1    1671852 .       G       C       143.77  LowDP;MQD       AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:0,1:99:172,0,296
chr1    1671853 .       C       T       143.77  LowDP;MQD       AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:0,1:99:172,0,296
chr1    1671858 .       TCTCGC  T       134.73  LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,1:99:172,0,296
chr1    1671866 .       A       AGAGAT  134.73  LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,1:99:172,0,296
chr1    1671880 .       G       C       143.77  LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,1:99:172,0,296
chr1    1684347 .       C       CCCT    391.79  PASS    AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,9:19:429,0,19
chr1    1695462 .       C       A       537.77  PASS    AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:21,21:99:566,0,618
chr1    1695574 .       G       T       532.77  PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,16:48:561,48,0
chr1    1695742 .       C       T       34.78   LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,2:28:63,0,28
chr1    1695831 .       C       T       61.28   LowDP   AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,3:9:89,9,0
chr1    1710035 .       C       T       75.28   LowDP   AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,3:9:103,9,0
chr2    123332  .       C       G       521.77  PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,16:48:550,48,0
chr2    128345  .       C       T       4286.77 PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,127:99:4315,378,0
chr2    130618  .       G       C       340.77  PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,12:36:369,36,0
chr2    132469  .       C       A       4888.77 PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,82:99:4917,256,0
chr2    132490  .       C       G       4888.77 PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,62:99:4917,256,0
chr2    132696  .       A       G       144.90  LowDP   AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,5:15:173,15,0
```

Body

VCF: Info

```
##fileformat=VCFv4.1
##FILTER=<ID=FSindel,Description="FS > 200.0">
##FILTER=<ID=FSsnp,Description="FS > 60.0">
##FILTER=<ID=LowDP,Description="DP < 10.0">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseCounts,Number=4,Type=Integer,Description="Counts of each base">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr2,length=243199373,assembly=hg19>
##contig=<ID=chr3,length=198022430,assembly=hg19>
##reference=file:///jumbo/db/GATK/Minusy/ucsc.hg19.minusMY.fasta
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  C2463_S1
chr1    1671852 .       G       C       143.77  LowDP;MQD       AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:0,1:99:172,0,296
chr1    1671853 .       C       T       143.77  LowDP;MQD       AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:0,1:99:172,0,296
chr1    1671858 .       TCTCGC  T       134.73  LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,1:99:172,0,296
chr1    1671866 .       A       AGAGAT  134.73  LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,1:99:172,0,296
chr1    1671880 .       G       C       143.77  LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,1:99:172,0,296
chr1    1684347 .       C       CCCT    391.79  PASS    AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,9:19:429,0,19
chr1    1695462 .       C       A       537.77  PASS    AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:21,21:99:566,0,618
chr1    1695574 .       G       T       532.77  PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,16:48:561,48,0
chr1    1695742 .       C       T       34.78   LowDP   AC=1;AF=0.500;AN=2      GT:AD:GQ:PL     0/1:1,2:28:63,0,28
chr1    1695831 .       C       T       61.28   LowDP   AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,3:9:89,9,0
chr1    1710035 .       C       T       75.28   LowDP   AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,3:9:103,9,0
chr2    123332  .       C       G       521.77  PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,16:48:550,48,0
chr2    128345  .       C       T       4286.77 PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,127:99:4315,378,0
chr2    130618  .       G       C       340.77  PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,12:36:369,36,0
chr2    132469  .       C       A       4888.77 PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,82:99:4917,256,0
chr2    132490  .       C       G       4888.77 PASS    AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,62:99:4917,256,0
chr2    132696  .       A       G       144.90  LowDP   AC=2;AF=1.00;AN=2       GT:AD:GQ:PL     1/1:0,5:15:173,15,0
```

Header

Body

VCF: Format

Header

```
##fileformat=VCFv4.1
##FILTER=<ID=FSindel,Description="FS > 200.0">
##FILTER=<ID=FSsnp,Description="FS > 60.0">
##FILTER=<ID=LowDP,Description="DP < 10.0">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseCounts,Number=4,Type=Integer,Description="Counts of each base">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr2,length=243199373,assembly=hg19>
##contig=<ID=chr3,length=198022430,assembly=hg19>
##reference=file:///jumbo/db/GATK/Minusy/ucsc.hg19.minusMY.fasta
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	C2463	S1
chr1	1671852	.	G	C	143.77	LowDP;MQD	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:0,1:99:172,0,296	
chr1	1671853	.	C	T	143.77	LowDP;MQD	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:0,1:99:172,0,296	
chr1	1671858	.	TCTCGC	T	134.73	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,1:99:172,0,296	
chr1	1671866	.	A	AGAGAT	134.73	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,1:99:172,0,296	
chr1	1671880	.	G	C	143.77	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,1:99:172,0,296	
chr1	1684347	.	C	CCCT	391.79	PASS	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,9:19:429,0,19	
chr1	1695462	.	C	A	537.77	PASS	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:21,21:99:566,0,618	
chr1	1695574	.	G	T	532.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,16:48:561,48,0	
chr1	1695742	.	C	T	34.78	LowDP	AC=1;AF=0.500;AN=2	GT:AD:GQ:PL	0/1:1,2:28:63,0,28	
chr1	1695831	.	C	T	61.28	LowDP	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,3:9:89,9,0	
chr1	1710035	.	C	T	75.28	LowDP	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,3:9:103,9,0	
chr2	123332	.	C	G	521.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,16:48:550,48,0	
chr2	128345	.	C	T	4286.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,127:99:4315,378,0	
chr2	130618	.	G	C	340.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,12:36:369,36,0	
chr2	132469	.	C	A	4888.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,82:99:4917,256,0	
chr2	132490	.	C	G	4888.77	PASS	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,62:99:4917,256,0	
chr2	132696	.	A	G	144.90	LowDP	AC=2;AF=1.00;AN=2	GT:AD:GQ:PL	1/1:0,5:15:173,15,0	

Body

Övning

https://github.com/bcfgothenburg/VT21/wiki/Molekylarbiologisk-metodik

- Svara på frågorna på övningarna
- Fastnar du finns svaren uppe på github
- Fråga oss
- Ni behöver inte lämna in något

VISIT OUR WEBSITE

https://cf.gu.se/english/Bioinformatics