

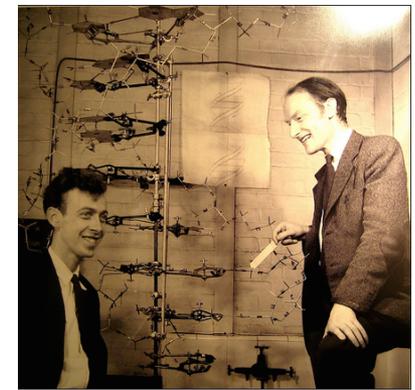
Monday – part 2

# Introduction to NGS data analysis

**Joanna Ciomborowska-Basheer, PhD**

**Faculty of Biology , Adam Mickiewicz University, Poznań  
ideas4biology Ltd.**

1953 **Discovery of DNA structure (Watson & Crick)**



5'--TGG AATTGTGAGCGGATAACAATT3'  
3'--ACCTTAACA CTCGCCTATTGTTAA5'

1973 : **First sequences published (24 bp)**

1977 : **Sequencing method published by Sanger**

1980 : The Nobel Prize for Gilbert and Sanger

1982 : The beginning of GenBank

1983 : Development of PCR

1987 : First automatic sequencer – Applied Biosystems Prism 373



1996 : Capilar sequencer ABI 310

1998 : Sequencing of *C.elegans* genome

2000 : **The Human Genome Project**

2005 : **First NGS machine – GS 20 System (454)**

2006 : First machine by Solexa – Genome Analyzer

2007 : First machine by Applied Biosystems

2009 : Helicos – single cell sequencing

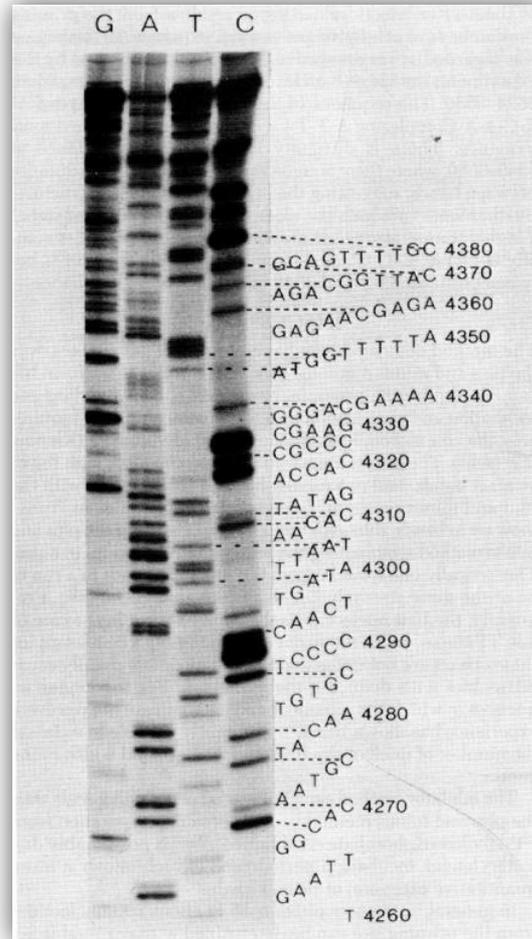
2011 : Ion Torrent : PGM

2011 : Pacific Biosciences – PacBio RS

2012 : Oxford Nanopore Technologies

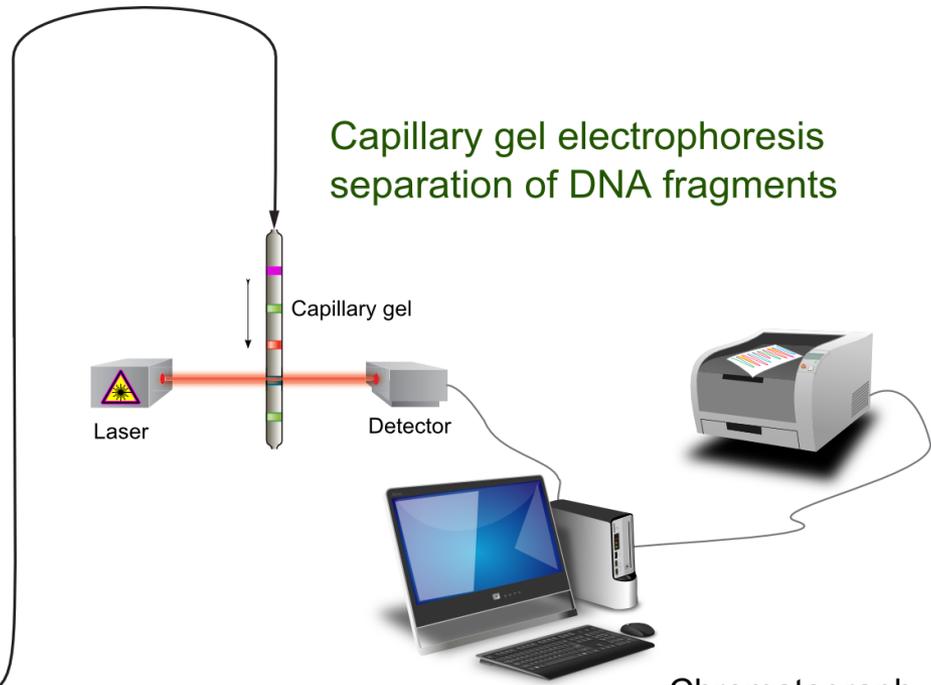
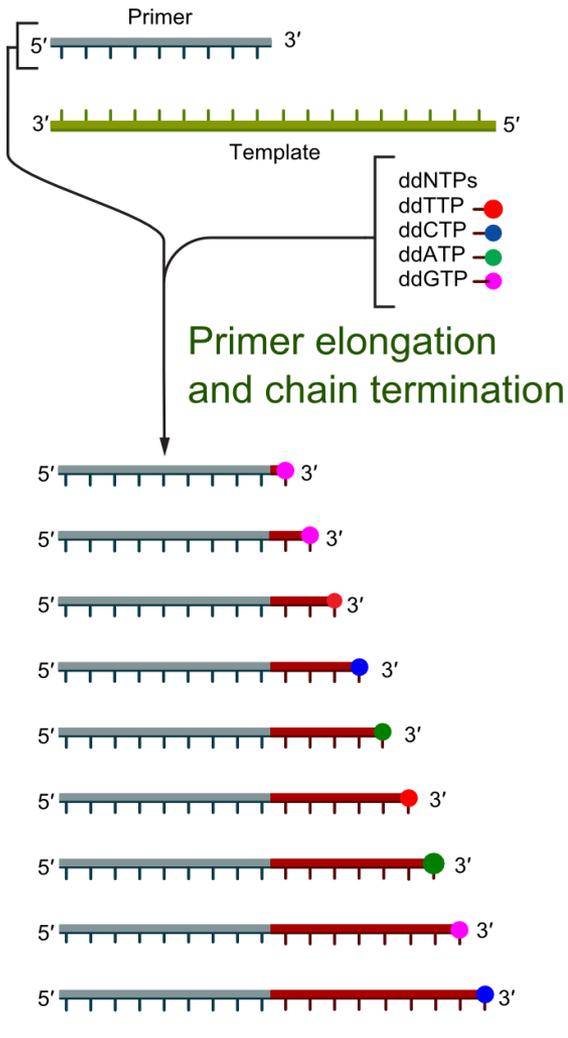


# Sanger sequencing method

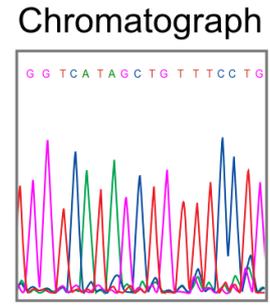


Sanger et al. 1977

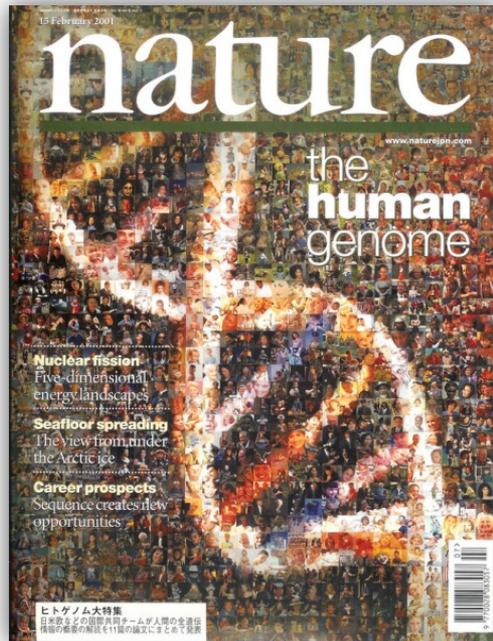
# First generation sequencing



Laser detection of flouochromes and computational sequence analysis

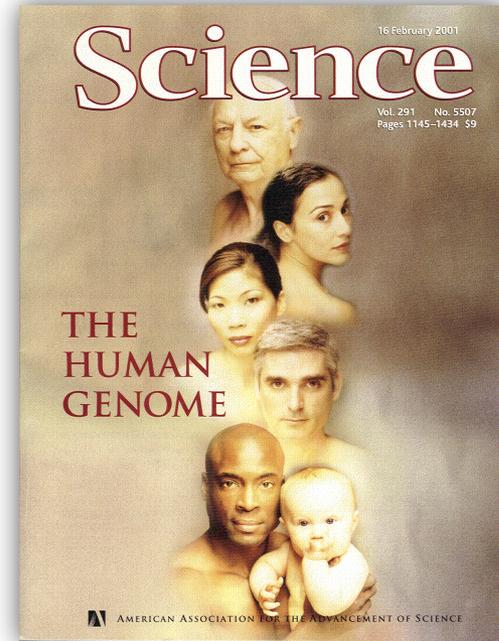


# First generation sequencing



<http://www.nature.com/>

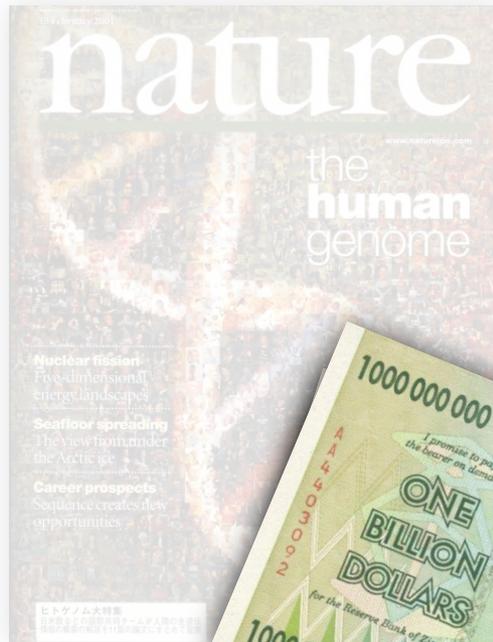
F. Collins - NIH



<http://www.sciencemag.org/>

C. Venter - Celera Genomics

# First generation sequencing



**10** years ...



[http://www](http://www.numismo.net/)

[/www.sciencemag.org/](http://www.sciencemag.org/)

<http://www.numismo.net/>

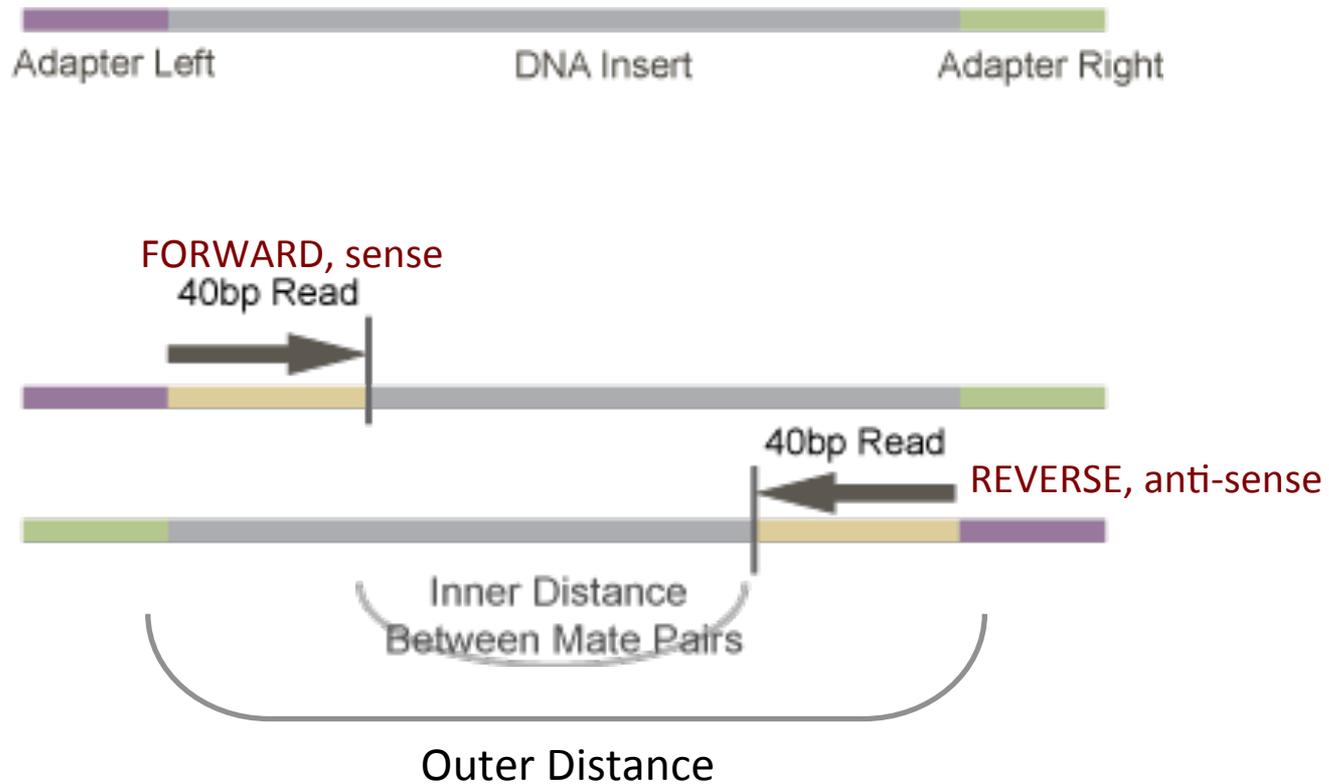
# Next generation sequencing

## GLOSSARY

- **Read** - refers to the data string of A,T, C, and G bases corresponding to the sample DNA or RNA.
- **Library** – set of prepared and sequenced fragments of DNA/RNA.
- **Adapters** – short oligos bound to the 5' and 3' end of each DNA fragment in a sequenced library. They are often removed right after sequencing but sometimes it has to be done by us.
- **Inserts** - fragments of a specific size are ligated or “inserted” in between two oligo adapters. The original sample DNA fragments are also referred to as “inserts.”

# Next generation sequencing GLOSSARY

- **Pair end reads – PE**



- **Single reads – SE / SR**

# Next generation sequencing GLOSSARY

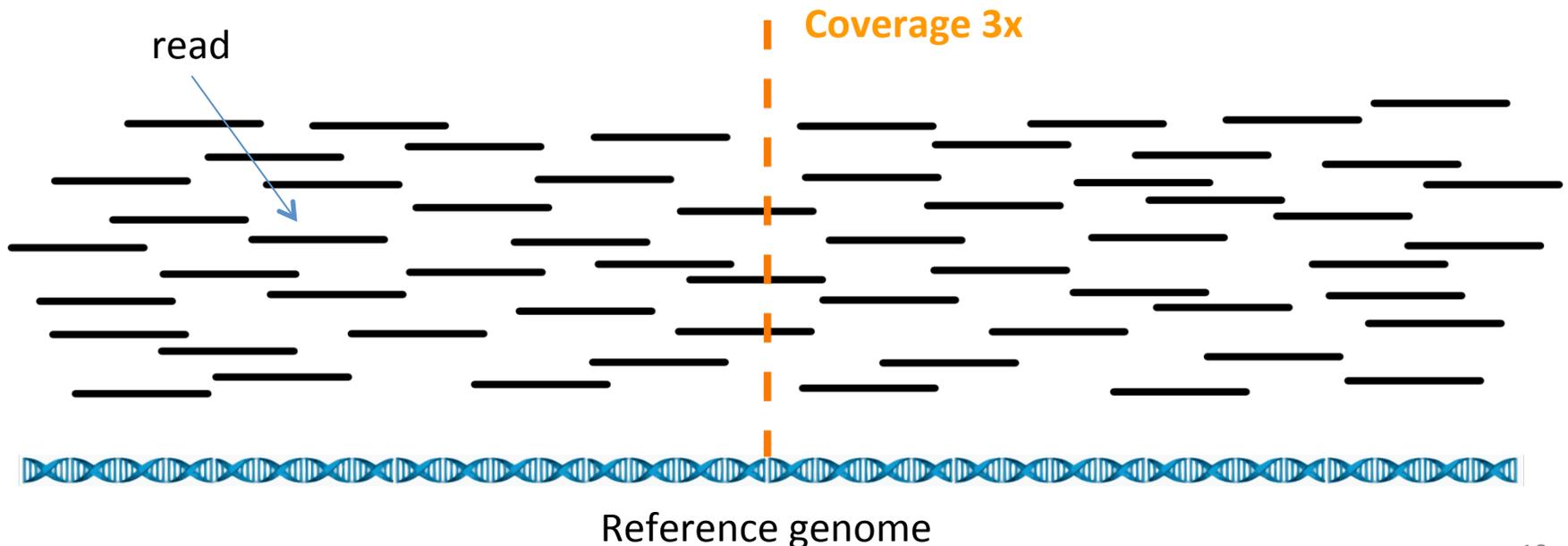
## Strand specific reads



	--SS_lib_type	Read 1	Read 2	
SE cases	F		—	
	R		—	
PE cases	FR			
	RF			e.g., dUTP/UDG protocol

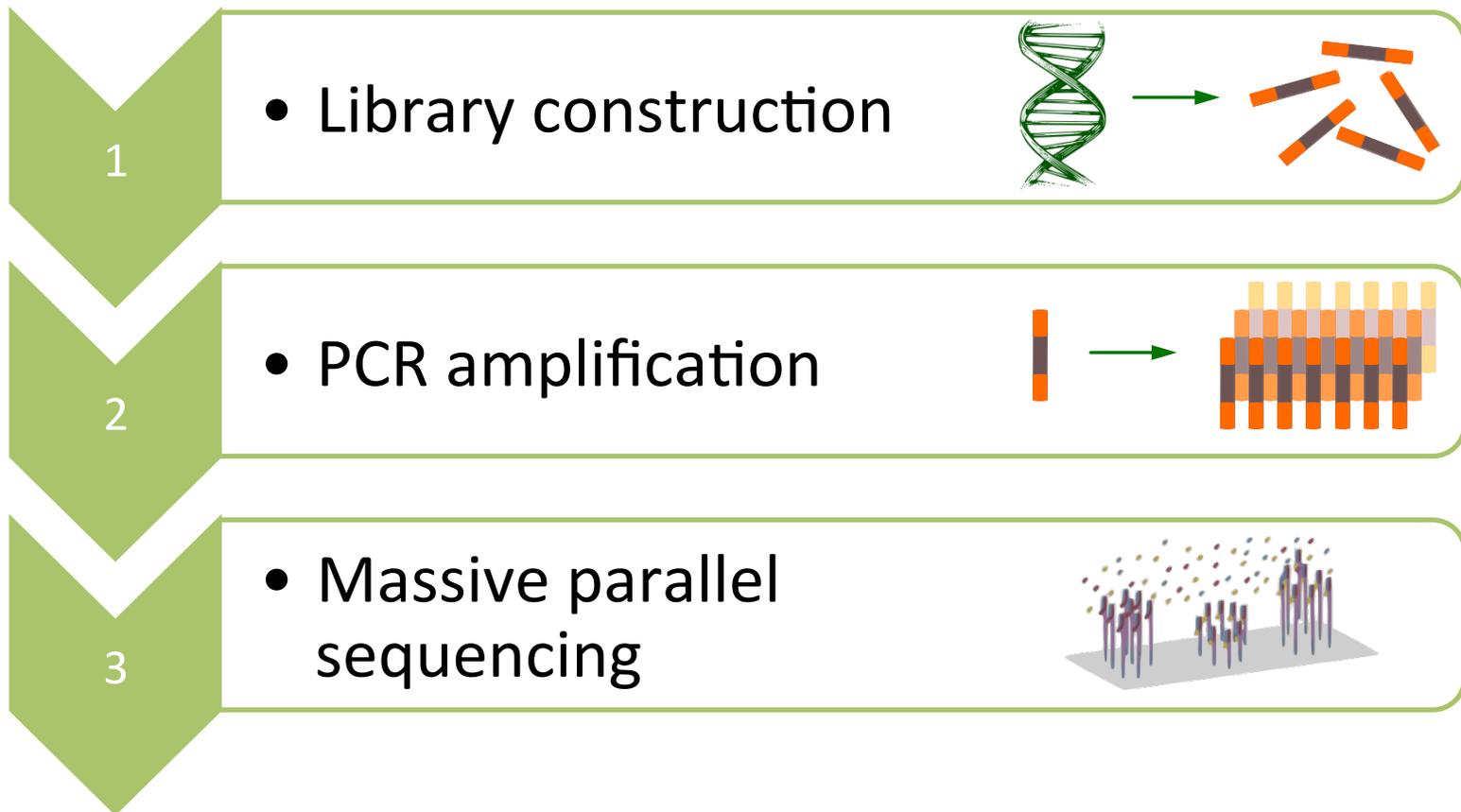
# Next generation sequencing GLOSSARY

- **Coverage level** – The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was sequenced 30 times.



# Next generation sequencing

Most of NGS protocols consist of:



# NGS platforms and technologies

illumina®



SOLiD®



ion torrent  
⬇ \* ⬆ ○ × □ + ≈

ThermoFisher  
SCIENTIFIC





## HiSeq X Ten



## HiSeq 2500



## NextSeq 500



## MiSeq





	HiSeq X Ten*	Hi Seq 2500			NextSeq 500		MiSeq
		HT v4	HT v3	Rapid	High	Mid	
Total output	1.8 Tb	1 Tb	600 Gb	180 Gb	129 Gb	39 Gb	15 Gb
Run time	3 days	6 days	11 days	40 hrs	29 hrs	26 hrs	~65 hrs
Output/day	600 Gb	167 Gb	55 Gb	~110 gb	~100 Gb	~36 Gb	~5.5 Gb
Read length	2 X 150	2 X 125	2 X 100	2 X 150	2 X 150	2 X 150	2 X 300
# of single reads	6B	4B	3B	600M	400M	130M	25M
Instrument price	\$1M*	\$740K	\$740K	\$740K	\$250K	\$250K	\$125K
Run price	~\$12k	~\$29k	~\$26k	~\$8k	\$4k	?	~\$1.4k
\$/Gb	\$7	\$29	\$43	\$44	\$33	?	\$93



## Genome Sequencer FLX+



## Genome Sequencer Junior





# 454

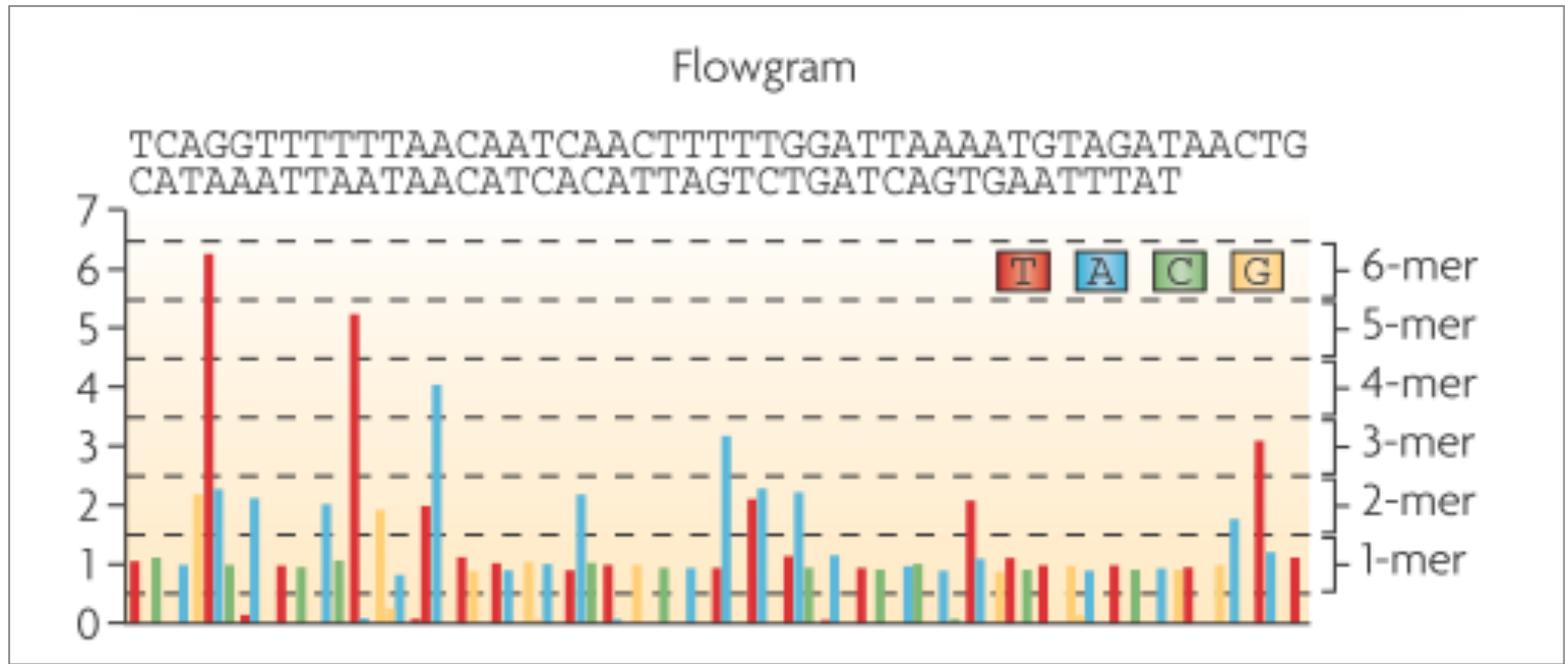
## SEQUENCING

	<b>GS FLX+</b>	<b>GS Jr.</b>
Total output/run	700 Mb	35 Mb
Run Time	23 hrs	10 hrs
Output/day	700 Mb	35 Mb
Read length	up to 1 Kb	~700b
# of single reads	1 M	0.1 M
Instrument Price	~\$500 k	\$125 k
Run price	~\$6 k	~\$1 k

# Roche 454 SEQUENCING



Higher error rate because of HOMOPOLYMERS !!!



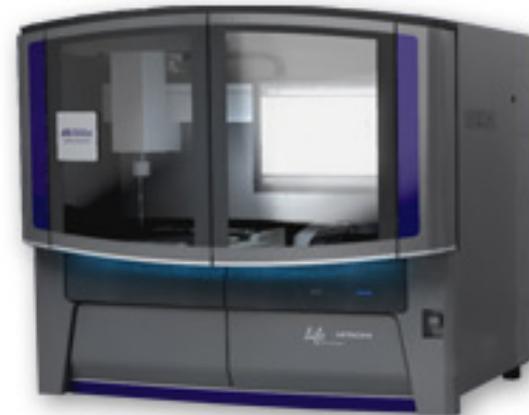
**SOLiD®**

by *life* technologies™  
**ThermoFisher**  
SCIENTIFIC

**SOLiD 5500**



**SOLiD 5500xl**



<http://allseq.com/>

# SOLiD<sup>®</sup>

by *life* technologies™  
**ThermoFisher**  
SCIENTIFIC

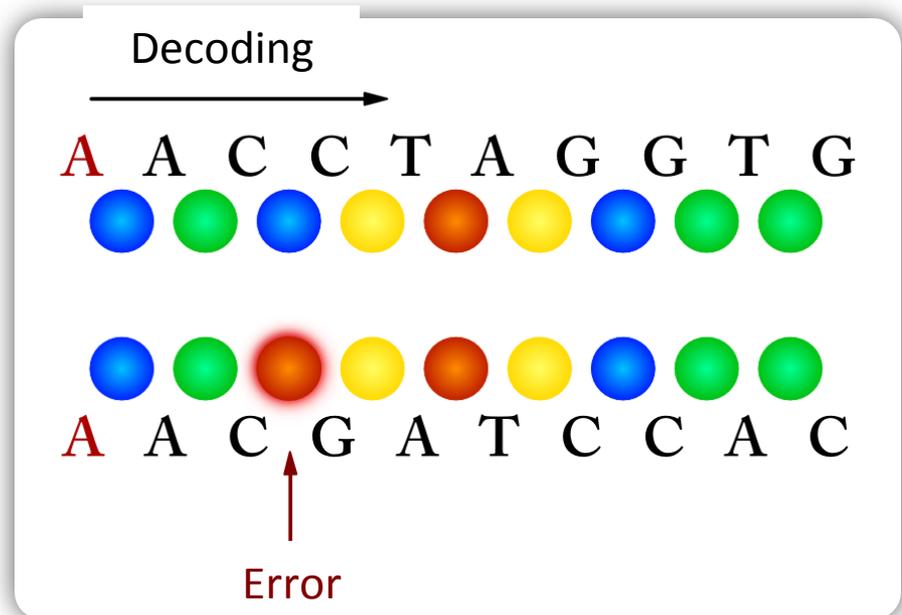
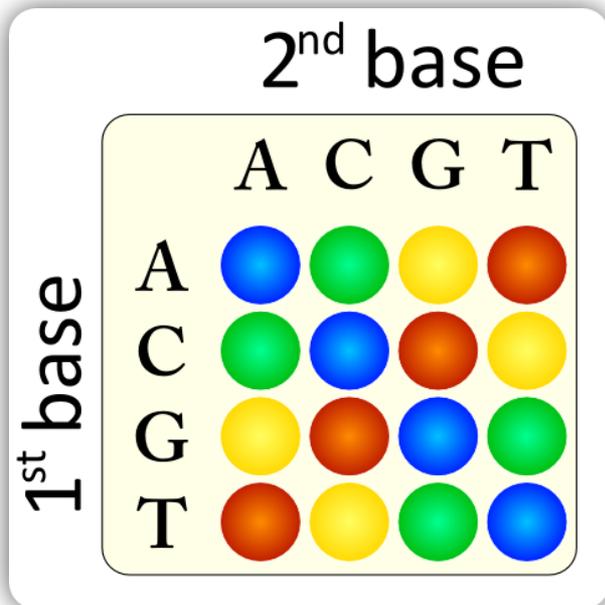
	SOLiD 5500xl	SOLiD 5500xl Wildfire	SOLiD 5500	SOLiD 5500 Wildfire
Total output/run	95 Gb	240 Gb	48 Gb	120 Gb
Run Time	6 days	10 days	6 days	10 days
Output/day	16 Gb	24 Gb	8 Gb	12 Gb
Read length	2 X 60	2 X 50	2 X 60	2 X 50
# of single reads	800M	2.4B	400M	1.2B
Instrument Price	\$595k	\$70k upgrade	\$349kk	\$70k upgrade
Run Price	~\$10k	~\$5k	~\$5k	~\$2.05k

<http://allseq.com/>

SOLiD<sup>®</sup>

by *life* technologies™  
**ThermoFisher**  
SCIENTIFIC

## Color space



# ion torrent



by *life* technologies™  
**ThermoFisher**  
SCIENTIFIC



**Ion PGM**  
(Personal Genome Machine)

**Ion Proton**



	PGM 314	PGM 316	PGM 318	PI	PII (est. early 2015)
# of sensors	1.2M	6.1M	11M	165M	660M
Total output	up to 100Mb	up to 1Gb	up to 2Gb	~10Gb	~32Gb(at launch)
Run Time	2-4h rs	3-5 hrs	4-7 hrs	2-4 hrs	2-4 hrs
Output/day*	up to 200Mb	up to 2Gb	up to 4Gb	~20Gb	~64Gb
Avg read length	up to 400b	up to 400b	up to 400b	up to 200b	100b
# of single reads	up to 0.6M	up to 3M	up to 5.5M	up to 82M	up to 330M
Chip price	\$99	\$299	\$499	\$699	~\$699
Reagent price**	\$250	\$250	\$250	\$300	~\$300
Instrument price	\$50k	\$50k	\$50k	\$149k	\$1419k

\* assumes two runs/day

\*\* template prep plus sequencing per chip



Higher error rate because of  
**HOMOPOLYMERS !!!**



PACIFIC  
BIOSCIENCES™



## Single molecule real-time sequencing !

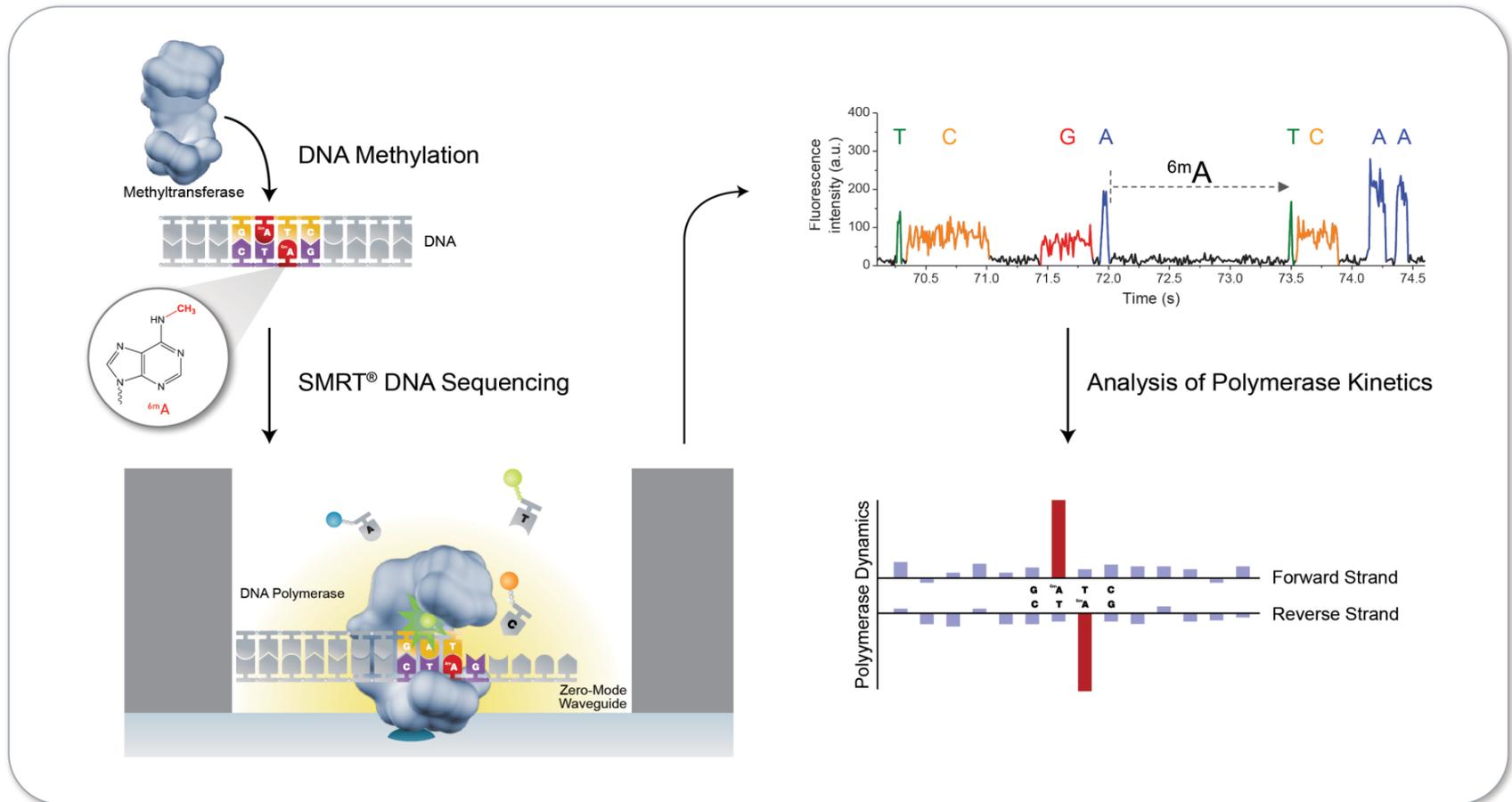
	RS II (P6-C4)	Sequel
Run time	up to 240 min	up to 240 min
Total output	~500 Mb - 1 Gb	5 Gb - 10 Gb
Output/day	~2 Gb	20 Gb
Mean read length	10 -15 kb	10 -15 kb
Single pass accuracy	~86%	~86%
Consensus (30X) accuracy	>99.999%	>99.999%
# of reads	~50k	~500k
Instrument price	~\$700k	\$350k
Run price	~\$400	~\$850

<http://allseq.com/>

<http://www.pacificbiosciences.com/>

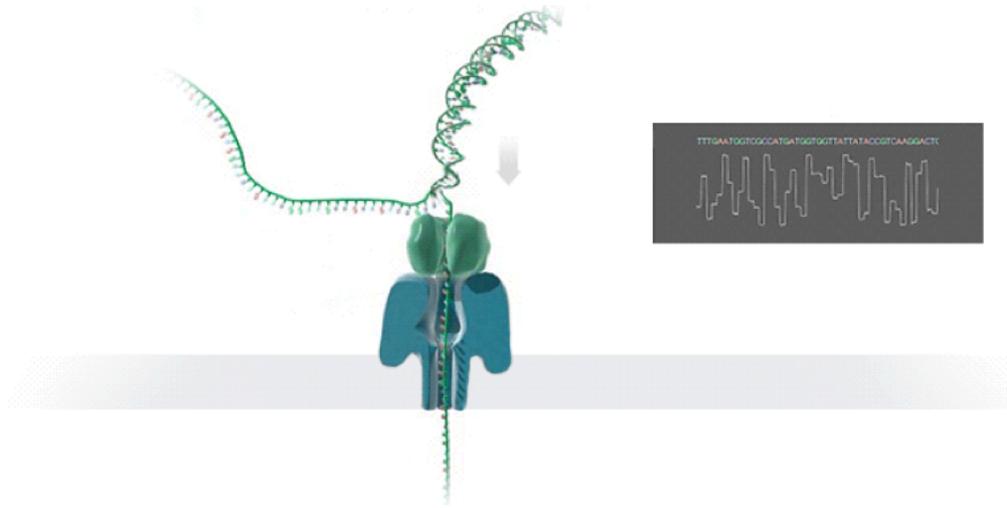


# DNA methylation



# Single molecule real-time sequencing !

How it works: the MinION for nanopore DNA sequencing



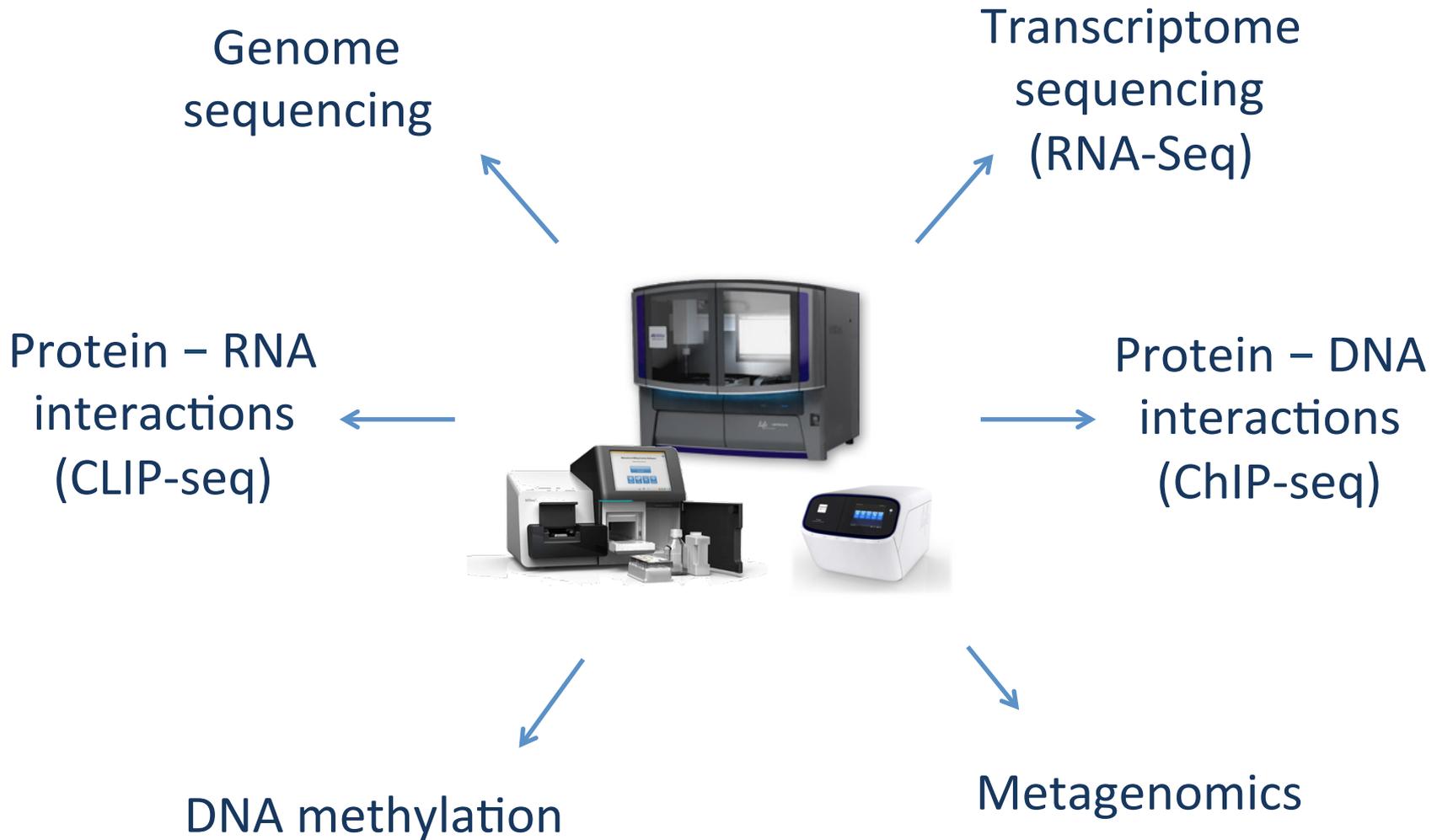
A nanopore is a nano-scale hole. In its devices, Oxford Nanopore passes an ionic current through nanopores and measures the changes in current as biological molecules pass through the nanopore or near it. The information about the change in current can be used to identify that molecule.

MinION MkI Flow cell

The MinION MkI flow cell contains a sensor array of several hundred channels, to enable multiple nanopore experiments to be performed in parallel. The flow cell is compatible with the MinION MkI device.



# Next generation sequencing Applications



# Next generation sequencing projects



<http://www.1000genomes.org/>



<http://cancergenome.nih.gov/>



[http://medicalgenomics.org/rna\\_seq\\_atlas](http://medicalgenomics.org/rna_seq_atlas)



<http://www.genomicsengland.co.uk/>

# Examples of applications

	illumina		454 SEQUENCING	SOLID	ion torrent	PACIFIC BIOSCIENCES		
	HiSeq 2500	NextSeq 500	MiSeq	GS FLX+	5500xl	PGM	IP	RS II
Whole genome (WGS)	●	●	⊘	⊘	●	⊘	⊘	▼
Exome (WXS)	●	●	●	⊘	●	⊘	●	⊘
Small genome	▼	●	●	●	▼	●	●	●
Targeted seq	▼	●	●	●	▼	●	▼	●
Transcriptome	●	●	▼	▼	●	⊘	●	▼
RNA profiling	▼	●	▼	⊘	▼	▼	●	⊘
ChIP-Seq	▼	●	●	⊘	▼	▼	●	⊘
Metagenomics	●	●	⊘	▼	●	⊘	⊘	▼

# Let's start to analyse...

## What data usually do we need?



- ✓ Reads
- ✓ Genomic sequences
- ✓ Transcriptomes
- ✓ Annotations

# Genome annotation

## GLOSSARY

- **Annotation as a PROCESS** in which structural and functional elements of genomes are identified and being described.
- **Annotation as a DATASET** which contains and describes all structural and functional elements of sequences (i.e. strand, chromosome, coordintes). Basic sources of genome annotations are: NCBI, Ensembl, UCSC Genome Browser.

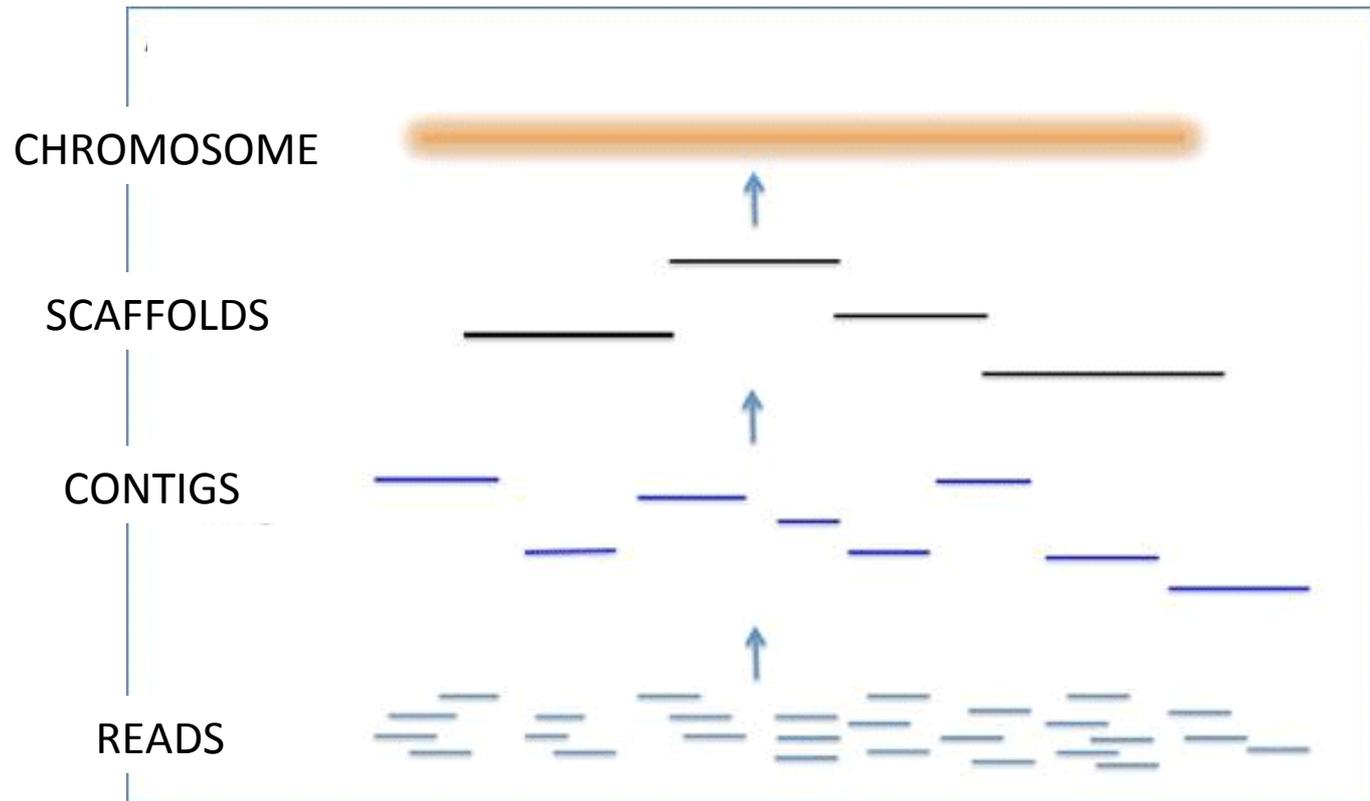
# Genome annotation

## GLOSSARY

- In case of genomes which are not fully sequenced or assembled, we usually have **contigs** and **scaffolds** (instead of chromosomes).
- Well annotated genomes (i.e. human, mouse) can have additional sequences assigned to **patches**, which contain fragments not included in chromosomes.
- Sometimes, apart from **reference genome**, we can find **alternative assemblies** provided by various sequencing groups.

# Genome assembly

ASSEMBLING



<http://www.ddbj.nig.ac.jp/sub/assembly-e.html>

Useful definitions: <https://www.ncbi.nlm.nih.gov/grc/help/definitions>

# Selected data formats

1. FASTA → sequences
2. FASTQ → raw reads
3. BAM, SAM → mapped reads
4. BED → coordinates
5. GTF, GFF3 → annotations
6. Tabular text file (TSV) → i.e. BLAST results, expression estimation results

# Selected data formats

## FASTA

FASTA format consists of **header** (begins with '>') after which there is a **sequence**. File can contain more sequences.

Extension: **.fasta** or **.fa**

```
>gi|453055269|gb|KC207570.1| Homo sapiens transient receptor potential cation channel  
subfamily C member 3 variant c (TRPC3) mRNA  
GGATATAGAAATGGGAATGGGTA ACTCAAAGTCCAGGCAGATAATGAAAAGACTTATAAAGCGGTATGTTTTGAAAGCACAAGTAGAC  
AAAGAAAATGATGAAGTTAATGAAGGTGAATTAAGAAATCAAGCAAGATATCTCCAGCCTTCGTTATGAACTTTTGGAAAGACAAGAG  
CCAAGCAACTGAGGAATTAGCCATTCTAATTCATAAACTTAGTGAGAACTGAATCCCAG
```

```
>XP_007907608.1 PREDICTED: keratin, type I cytoskeletal 19-like [Callorhinchus milii]  
MSRSVYSANIGGSVIVSSNQRRSFASTSSSLFGSGAPSRRAISVYNIGANRGKRISAGGSWNASYASLG  
GDAGILCNDEKQTMQNLNARLSSYMEKVRSLKSNRQLEFQIHEFYEKKAPVSTKDLTVYEGNISDCHLQ  
IYAISLHNAKLMLQIDNARLAADDFRIKYESELAIRKGVEADIQGLRKVMDELSLTKRGLESQVTALKED  
LVYLHRSHKEELSLRTGMGGSVTVDL DSTPATDLNKILSNLRIEYETIAEKNRKDVEAWYLEKCHTLNQ
```

# Selected data formats

## FASTA

### IUPAC code

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

# Selected data formats

## BED

**.bed** file contains coordinates so information about localisation of some regions (gene, exon, intron) in the genome. Obligatory columns are: chromosome/region, start, stop.

```
chr7 127474697 127475864
chr7 127475864 127477031
chr7 127477031 127478198
chr7 127478198 127479365
chr7 127479365 127480532
chr7 127480532 127481699
```

↙  
chromosome

↓  
START

↓  
END

# Selected data formats

Coordinates can be shown in different ways

chrX	151073054	151173000
chrX	151183000	151190000
chrX	151283000	151290000
chrX	151383000	151390000

## UCSC Genome Browser

chr5:62,797,383-63,627,669  
chr6:61,274,683-63,627,459  
chr7:38,797,383-82,640,939  
chr8:11,838,483-11,773,546

## Ensembl Genome Browser

5:62797383-63627669  
6:61274683-63627459  
7:38797383-82640939  
8:11838483-11773546

## NCBI

NC\_000017.11 (43044295..43125483)  
NC\_000017.10 (41196312..41277500)

# Selected data formats

## GTF, GFF3

**.gtf .gff3** files contain information about annotated elements in the genome. For each element there is a separated line in the file consists of 9 columns.

## GTF - General Transfer Format

```
#!genome-build GRCh38.p7
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.22
#!genebuild-last-updated 2016-06
```

**1 2 3 4 5 6 7 8 9**

```
1 havana gene 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_u
1 havana transcript 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name
1 havana exon 11869 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "
1 havana exon 12613 12721 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "
1 havana exon 13221 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "
1 havana transcript 12010 13670 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; gene_name
1 havana exon 12010 12057 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 12179 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 12613 12697 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 12975 13052 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 13221 13374 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 13453 13670 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
```

# GTF - General Transfer Format

1 2 3 4 5 6 7 8 9

```
1 havana gene 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_u
1 havana transcript 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name
1 havana exon 11869 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "
1 havana exon 12613 12721 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "
1 havana exon 13221 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "
1 havana transcript 12010 13670 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; gene_name
1 havana exon 12010 12057 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 12179 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 12613 12697 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 12975 13052 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 13221 13374 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
1 havana exon 13453 13670 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000450305"; transcript_version "2"; exon_number "
```

- 1. seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix (the convention in Ensembl is to omit the 'chr' prefix).
- 2. source** - name of the program that generated this feature, the data source (database or project name) or gene status
- 3. feature** - feature type name, e.g. Gene, Variation, Similarity
- 4. start** - start position of the feature, with sequence numbering starting at 1.
- 5. end** - end position of the feature, with sequence numbering starting at 1.
- 6. score** - a floating point value.
- 7. strand** - defined as + (forward) or - (reverse).
- 8. frame** - one of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
- 9. attribute** - a semicolon-separated list of tag-value pairs, providing additional information about each feature.

# GFF3 - General Feature Format

```
1 2 3 4 5 6 7 8 9
1 ensembl_havana lincRNA_gene 89295 133723 . - . ID=transcript:ENST00000466430;Parent=gene:ENSG00000238009;Name=RP11-34P13.7;biotype=lincRNA;gene_id=ENSG00000238009;havana_gene=UII
1 havana lincRNA 89295 120932 . - . ID=transcript:ENST00000466430;Parent=gene:ENSG00000238009;Name=RP11-34P13.7-001;biotype=lincRNA;havana_transcript:
1 havana exon 89295 91629 . - . Parent=transcript:ENST00000466430;Name=ENSE00001846804;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 92091 92240 . - . Parent=transcript:ENST00000466430;Name=ENSE00001944529;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 112700 112804 . - . Parent=transcript:ENST00000466430;Name=ENSE00001957285;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 120775 120932 . - . Parent=transcript:ENST00000466430;Name=ENSE00001606755;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana lincRNA 92230 129217 . - . ID=transcript:ENST00000477740;Parent=gene:ENSG00000238009;Name=RP11-34P13.7-003;biotype=lincRNA;havana_transcript:
1 havana exon 92230 92240 . - . Parent=transcript:ENST00000477740;Name=ENSE00001896976;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 112700 112804 . - . Parent=transcript:ENST00000477740;Name=ENSE00001957285;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 120721 120932 . - . Parent=transcript:ENST00000477740;Name=ENSE00001171005;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 129055 129217 . - . Parent=transcript:ENST00000477740;Name=ENSE00001919246;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana lincRNA 110953 129173 . - . ID=transcript:ENST00000471248;Parent=gene:ENSG00000238009;Name=RP11-34P13.7-002;biotype=lincRNA;havana_transcript:
1 havana exon 110953 111357 . - . Parent=transcript:ENST00000471248;Name=ENSE00001879696;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 112700 112804 . - . Parent=transcript:ENST00000471248;Name=ENSE00001957285;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
1 havana exon 129055 129173 . - . Parent=transcript:ENST00000471248;Name=ENSE00001934975;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_
```

- 1. seqid** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. Important note: the seq ID must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly
- 2. source** - name of the program that generated this feature, or the data source (database or project name)
- 3. type** - type of feature. Must be a term or accession from the SOFA sequence ontology
- 4. start** - start position of the feature, with sequence numbering starting at 1.
- 5. end** - end position of the feature, with sequence numbering starting at 1.
- 6. score** - a floating point value.
- 7. strand** - defined as + (forward) or - (reverse).
- 8. phase** - one of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
- 9. attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature. Some of these tags are predefined, e.g. ID, Name, Alias, Parent etc.

Basic resources related to  
genomes and annotations

# Three databases and browsers



National Center for Biotechnology Information



Ensembl Genome Browser

European Bioinformatics Institute, Wellcome Trust Sanger Institute



University of California, Santa Cruz Genome Browser

# RefSeq - sequences – accession numbers

**N**M\_..., **X**M\_... → mRNA

**N**P\_..., **X**P\_... → protein

**N**R\_..., **X**R\_... → noncoding RNA

**N**C\_..., **N**G\_..., **N**T\_..., **N**W\_..., **A**C\_... → contigs, genomic sequences

**N** - sequences from experiments

**X** – sequences based on comparison with existing sequences or prediction

**RefSeq Database** – set of checked, non-redundant and unique sequences which are often manually curated.

# NCBI (Taxonomy)

<https://www.ncbi.nlm.nih.gov/taxonomy/>

The screenshot shows the NCBI Taxonomy Browser interface. At the top, there are navigation tabs for Entrez, PubMed, Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and Books. Below the navigation is a search bar with the text "Search for" and a dropdown menu set to "complete name". There is also a "lock" checkbox and a "Go" button. Below the search bar, there is a "Display" field set to "3" and a "levels using filter" dropdown set to "none".

The main content area displays the results for "Homo sapiens". It includes the following information:

- Taxonomy ID:** 9606
- Genbank common name:** human
- Inherited blast name:** primates
- Rank:** species
- Genetic code:** [Translation table 1 \(Standard\)](#)
- Mitochondrial genetic code:** [Translation table 2 \(Vertebrate Mitochondrial\)](#)
- Other names:**
  - synonym: humans
  - common name: man
  - authority: **Homo sapiens Linnaeus, 1758**
- Lineage(full)**
  - [cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Dipnotetrapodomorpha](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Boreoeutheria](#); [Euarchontoglires](#); [Primates](#); [Haplorrhini](#); [Simiiformes](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#); [Homininae](#); [Homo](#)

On the right side of the interface, there is a table titled "Entrez records" with three columns: Database name, Subtree links, and Direct links. The table contains the following data:

Database name	Subtree links	Direct links
Nucleotide	<a href="#">14,056,215</a>	<a href="#">14,056,171</a>
Nucleotide EST	<a href="#">8,705,106</a>	<a href="#">8,705,106</a>
Nucleotide GSS	<a href="#">1,762,817</a>	<a href="#">1,761,491</a>
Protein	<a href="#">1,045,117</a>	<a href="#">1,044,825</a>
Structure	<a href="#">32,354</a>	<a href="#">32,354</a>
Genome	<a href="#">1</a>	<a href="#">1</a>
Popset	<a href="#">23,465</a>	<a href="#">23,464</a>
SNP	<a href="#">164,995,972</a>	<a href="#">164,995,972</a>
Domains	<a href="#">24</a>	<a href="#">24</a>
GEO Datasets	<a href="#">1,195,428</a>	<a href="#">1,195,428</a>
UniGene	<a href="#">130,056</a>	<a href="#">130,056</a>
PubMed Central	<a href="#">412,801</a>	<a href="#">412,789</a>
Gene	<a href="#">219,659</a>	<a href="#">219,586</a>
HomoloGene	<a href="#">18,713</a>	<a href="#">18,713</a>
SRA Experiments	<a href="#">695,320</a>	<a href="#">695,098</a>
Probe	<a href="#">27,382,436</a>	<a href="#">27,382,436</a>
Assembly	<a href="#">89</a>	<a href="#">89</a>

# NCBI (Downloads FTP)

<a href="#">NCBI Home</a>	<h2>All Resources</h2> <p><a href="#">All</a> <a href="#">Databases</a> <a href="#">Downloads</a> <a href="#">Submissions</a> <a href="#">Tools</a> <a href="#">How To</a></p> <h3>Downloads</h3> <p><a href="#">BLAST (Stand-alone)</a> BLAST executables for local use are provided for Solaris, LINUX, Windows, and MacOS information. Pre-formatted databases for BLAST nucleotide, protein, and translated sea subdirectory.</p> <p><a href="#">FTP: BLAST Databases</a> Sequence databases for use with the stand-alone BLAST programs. The files in this dir BLAST.</p> <p><a href="#">FTP: CDD</a> This site provides full data records for CDD, along with individual Position Specific Score data for each conserved domain. See the README file for full details.</p> <p><a href="#">FTP: ClinVar Data</a> This site provides full data extractions in XML and summary data in VCF format. It cont in <a href="#">ClinVar</a>, <a href="#">MedGen</a>, and <a href="#">GTR</a>.</p> <p><a href="#">FTP: FASTA BLAST Databases</a> Sequence databases in FASTA format for use with the stand-alone BLAST programs. T can be used with BLAST.</p> <p><a href="#">FTP: GenBank</a> This site contains files for all sequence records in GenBank in the default flat file format contents are described in the README.genbank file.</p> <p><a href="#">FTP: GenPept</a> The protein sequences corresponding to the translations of coding sequences (CDS) in the README file in the directory for more information.</p>
<a href="#">Resource List (A-Z)</a>	
<b>All Resources</b>	
<a href="#">Chemicals &amp; Bioassays</a>	
<a href="#">Data &amp; Software</a>	
<a href="#">DNA &amp; RNA</a>	
<a href="#">Domains &amp; Structures</a>	
<a href="#">Genes &amp; Expression</a>	
<a href="#">Genetics &amp; Medicine</a>	
<a href="#">Genomes &amp; Maps</a>	
<a href="#">Homology</a>	
<a href="#">Literature</a>	
<a href="#">Proteins</a>	
<a href="#">Sequence Analysis</a>	
<a href="#">Taxonomy</a>	
<a href="#">Training &amp; Tutorials</a>	
<a href="#">Variation</a>	

# NCBI (Downloads FTP) - **TASK**

Please, find human genomic sequence from chromosome 22 and try to download it in packed FASTA format (.fa).

**Attention: human reference genome – version GRCh38.p7**

# Three databases and browsers



National Center for Biotechnology Information



Ensembl Genome Browser

European Bioinformatics Institute, Wellcome Trust Sanger Institute



University of California, Santa Cruz Genome Browser

# Ensembl

<http://www.ensembl.org/>  
<http://ensemblgenomes.org/>

 **e!EnsemblGenomes**

---

 **e!Ensembl**

 **e!EnsemblFungi**

 **e!EnsemblMetazoa**

 **e!EnsemblProtists**

 **e!EnsemblPlants**

 **e!EnsemblBacteria**

 **Pre!Ensembl**

# Ensembl – accession numbers

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
CFTR-001	<a href="#">ENST00000003084.10</a>	6132	<a href="#">1480aa</a>	Protein coding	<a href="#">CCDS5773</a>	<a href="#">A0A024R730</a> <a href="#">P13569</a>	<a href="#">NM_000492</a> <a href="#">NP_000483</a>	TSL:1 GENCODE basic APPRIS P1
CFTR-005	<a href="#">ENST00000426809.5</a>	4316	<a href="#">1438aa</a>	Protein coding	-	<a href="#">E7EPB6</a>	-	CDS 3' incomplete TSL:5
CFTR-002	<a href="#">ENST00000468795.1</a>	682	<a href="#">190aa</a>	Protein coding	-	<a href="#">H0Y8A9</a>	-	CDS 5' incomplete TSL:5
CFTR-004	<a href="#">ENST00000446805.1</a>	575	<a href="#">36aa</a>	Protein coding	-	<a href="#">C9J6L5</a>	-	CDS 3' incomplete TSL:4
CFTR-008	<a href="#">ENST00000600166.1</a>	559	<a href="#">156aa</a>	Protein coding	-	<a href="#">M0QYZ3</a>	-	CDS 5' incomplete TSL:5
CFTR-009	<a href="#">ENST00000608965.5</a>	896	No protein	Processed transcript	-	-	-	TSL:5
CFTR-010	<a href="#">ENST00000610149.1</a>	519	No protein	Processed transcript	-	-	-	TSL:5
CFTR-007	<a href="#">ENST00000429014.1</a>	423	No protein	Processed transcript	-	-	-	TSL:5
CFTR-003	<a href="#">ENST00000546407.1</a>	222	No protein	Processed transcript	-	-	-	TSL:1
CFTR-006	<a href="#">ENST00000472848.1</a>	148	No protein	Processed transcript	-	-	-	TSL:5
CFTR-011	<a href="#">ENST00000621535.1</a>	657	No protein	Retained intron	-	-	-	TSL:5

ENSG → Gene

ENST → Transcript (.ver)

ENSP → Protein (.ver)

ENSE → Exon

} *Homo sapiens*

Protein\_coding

Processed\_transcript

Processed\_pseudogene

miRNA

rRNA, scRNA

snoRNA, snRNA ...

ENSXETG00000003158

*Xenopus tropicalis*

ENSMUST00000044620.1

*Mus musculus*

ENSDARG00000079015

*Danio rerio*

ENSECAT00000016282.5

*Equus caballus*

<http://www.ensembl.org/Help/Faq?id=468>

[http://vega.sanger.ac.uk/info/about/gene\\_and\\_transcript\\_types.html](http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html)

# Ensembl – download

1. From particular record in the database  
(Attention! Different level – different data!)
2. Download → FTP
3. BioMart

# Ensembl – TASK

Please, find in Ensembl Genome Browser ***BRCA1 in human***

1. How many isoforms does this gene have?
2. On which chromosome and strand it is located?

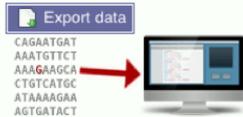
Please, choose the longest transcript of this gene and:

1. Find encoded protein
2. Have a look at exon-intron structure
3. Find sequence of the first exon
4. Download mRNA sequence
5. Download genomic sequence with 100 bp flanking regions

# Ensembl – download (FTP)

Downloads

## Download a sequence or region



Click on the 'Export data' button in the lefthand menu of most pages to export:

- FASTA sequence
- GTF or GFF features

...and more!

## Customise your download



Custom datasets can be retrieved using the BioMart data-mining tool.

You may find exploring this web-based query tool easier than extracting information direct from our databases.

## Fetch data programmatically



Write your own Perl scripts to retrieve small-to-medium datasets. All our data, as well as added functionality, is available through the Ensembl Perl API.

Use the API to retrieve gene and transcript sets, fetch alignments between sequences, compare allele frequencies and much more!

You can also use our [REST API](#) (currently in Beta) to retrieve data to process in the programming language of your choice.

## Download databases & software



All of our data and software, including pipelines and web code, is available free.

- [Download data via FTP](#)
- [Ensembl pipeline in CVS](#)
- [Set up your own Ensembl website](#)

## Multi-species data

Database						
Comparative genomics	<a href="#">MySQL</a>	<a href="#">EMF</a>	<a href="#">MAF</a>	<a href="#">BED</a>	<a href="#">XML</a>	<a href="#">Ancestral Alleles</a>
BioMart	<a href="#">MySQL</a>	-	-	-	-	-
Stable ids	<a href="#">MySQL</a>	-	-	-	-	-

## Single species data

Popular species are listed first. You can customise this list via our [home page](#).

Show 10 entries

Show/hide columns

Filter

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM/BigV
Y	<a href="#">Human</a> <i>Homo sapiens</i>	<a href="#">FASTA</a>	<a href="#">EMBL</a>	<a href="#">GenBank</a>	<a href="#">GTF</a> <a href="#">GFF3</a>	<a href="#">MySQL</a>	<a href="#">GVF</a>	<a href="#">VCF</a>	<a href="#">VEP</a>	<a href="#">Regulation</a> (GFF)	<a href="#">Regulation on data files</a>	<a href="#">BAM/BigV</a>				
Y	<a href="#">Mouse</a> <i>Mus musculus</i>	<a href="#">FASTA</a>	<a href="#">EMBL</a>	<a href="#">GenBank</a>	<a href="#">GTF</a> <a href="#">GFF3</a>	<a href="#">MySQL</a>	<a href="#">GVF</a>	<a href="#">VCF</a>	<a href="#">VEP</a>	<a href="#">Regulation</a> (GFF)	<a href="#">Regulation on data files</a>	<a href="#">BAM/BigV</a>				

# Ensembl – download – TASK

1. Please find files which contain:

- **Chromosome 22 genomic sequence (DNA)**
- **Proteins in FASTA format**
- **Annotations in GTF format for human.**

- *How these files are called?*

- *What's the name of file with whole genomic sequence in FASTA ?*

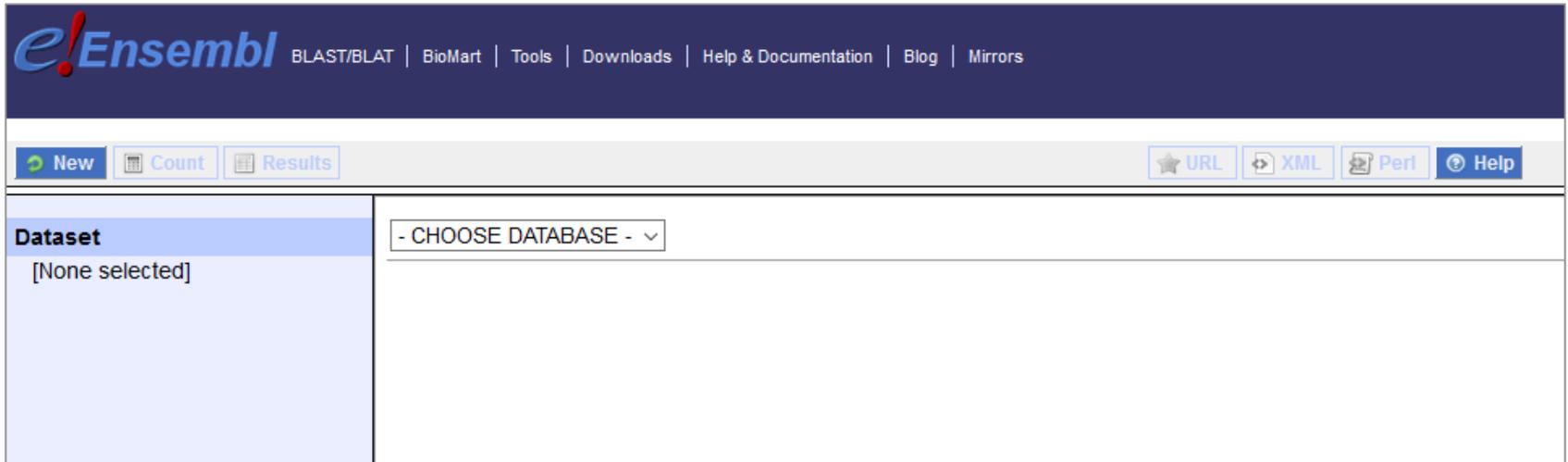
2. Please find noncoding RNAs (**ncRNA**) for **mouse (*Mus musculus*)**

in Ensembl but **archive!** [version 85]

# Ensembl – download

**ATTENTION! We can work with accession numbers not only from Ensembl**

## Ensembl BioMart



The screenshot displays the Ensembl BioMart web interface. At the top, the Ensembl logo is followed by navigation links: BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. Below the navigation bar is a toolbar with buttons for 'New', 'Count', and 'Results'. On the right side of the toolbar are icons for 'URL', 'XML', 'Perl', and 'Help'. The main content area is divided into two sections. The left section, titled 'Dataset', shows '[None selected]'. The right section contains a dropdown menu with the text '- CHOOSE DATABASE -' and a downward arrow.

Each version of Ensembl database and most of taxonomic groups have devoted BioMart tool.

# Ensembl – download

The screenshot shows the Ensembl BioMart interface. The top navigation bar includes links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, and Blog. Below this, there are buttons for 'New', 'Count', and 'Results'. On the right side, there are buttons for 'URL', 'XML', 'Perl', and 'Help'. The main content area is divided into sections: 'Dataset' (Ensembl Genes 90), 'Filters' ([None selected]), and 'Attributes' (Gene stable ID, Transcript stable ID). A red circle highlights the 'Dataset', 'Filters', and 'Attributes' sections.

- **Dataset** → we have to choose Ensembl version and organism
- **Filters** → allow to search for particular features, filter the data
- **Attributes** → describe output data format
- **Count** → summary showing how many genes/elements fulfill our filtering criteria
- **Results** → preview of our output data and options for export

# Ensembl – download

## TASK

Please check how many **protein coding genes** you can find in **mouse** (*Mus musculus*) and download the list of accession numbers (Gene ID, Transcript ID), as well as gene names and descriptions.

(Filters → Gene → Gene type → 'protein coding' → Count)

Then, try to search for their cDNA sequences.

Export all results to    Unique results only

Email notification to

View  rows as   Unique results only

Gene stable ID	Transcript stable ID	Gene name	Gene description
<a href="#">ENSMUSG00000064370</a>	<a href="#">ENSMUST00000082421</a>	<a href="#">mt-Cytb</a>	mitochondrially encoded cytochrome b [Source:MGI Symbol;Acc:MGI:102501]
<a href="#">ENSMUSG00000064368</a>	<a href="#">ENSMUST00000082419</a>	<a href="#">mt-Nd6</a>	mitochondrially encoded NADH dehydrogenase 6 [Source:MGI Symbol;Acc:MGI:102495]
<a href="#">ENSMUSG00000064367</a>	<a href="#">ENSMUST00000082418</a>	<a href="#">mt-Nd5</a>	mitochondrially encoded NADH dehydrogenase 5 [Source:MGI Symbol;Acc:MGI:102496]
<a href="#">ENSMUSG00000064363</a>	<a href="#">ENSMUST00000082414</a>	<a href="#">mt-Nd4</a>	mitochondrially encoded NADH dehydrogenase 4 [Source:MGI Symbol;Acc:MGI:102498]
<a href="#">ENSMUSG00000065947</a>	<a href="#">ENSMUST00000084013</a>	<a href="#">mt-Nd4l</a>	mitochondrially encoded NADH dehydrogenase 4L [Source:MGI Symbol;Acc:MGI:102497]
<a href="#">ENSMUSG00000064360</a>	<a href="#">ENSMUST00000082411</a>	<a href="#">mt-Nd3</a>	mitochondrially encoded NADH dehydrogenase 3 [Source:MGI Symbol;Acc:MGI:102499]
<a href="#">ENSMUSG00000064358</a>	<a href="#">ENSMUST00000082409</a>	<a href="#">mt-Co3</a>	mitochondrially encoded cytochrome c oxidase III [Source:MGI Symbol;Acc:MGI:102502]
<a href="#">ENSMUSG00000064357</a>	<a href="#">ENSMUST00000082408</a>	<a href="#">mt-Atp6</a>	mitochondrially encoded ATP synthase 6 [Source:MGI Symbol;Acc:MGI:99927]
<a href="#">ENSMUSG00000064356</a>	<a href="#">ENSMUST00000082407</a>	<a href="#">mt-Atp8</a>	mitochondrially encoded ATP synthase 8 [Source:MGI Symbol;Acc:MGI:99926]
<a href="#">ENSMUSG00000064354</a>	<a href="#">ENSMUST00000082405</a>	<a href="#">mt-Co2</a>	mitochondrially encoded cytochrome c oxidase II [Source:MGI Symbol;Acc:MGI:102503]

Attention! Take care about the name of output file!

# Three databases and browsers



National Center for Biotechnology Information



Ensembl Genome Browser

European Bioinformatics Institute, Wellcome Trust Sanger Institute

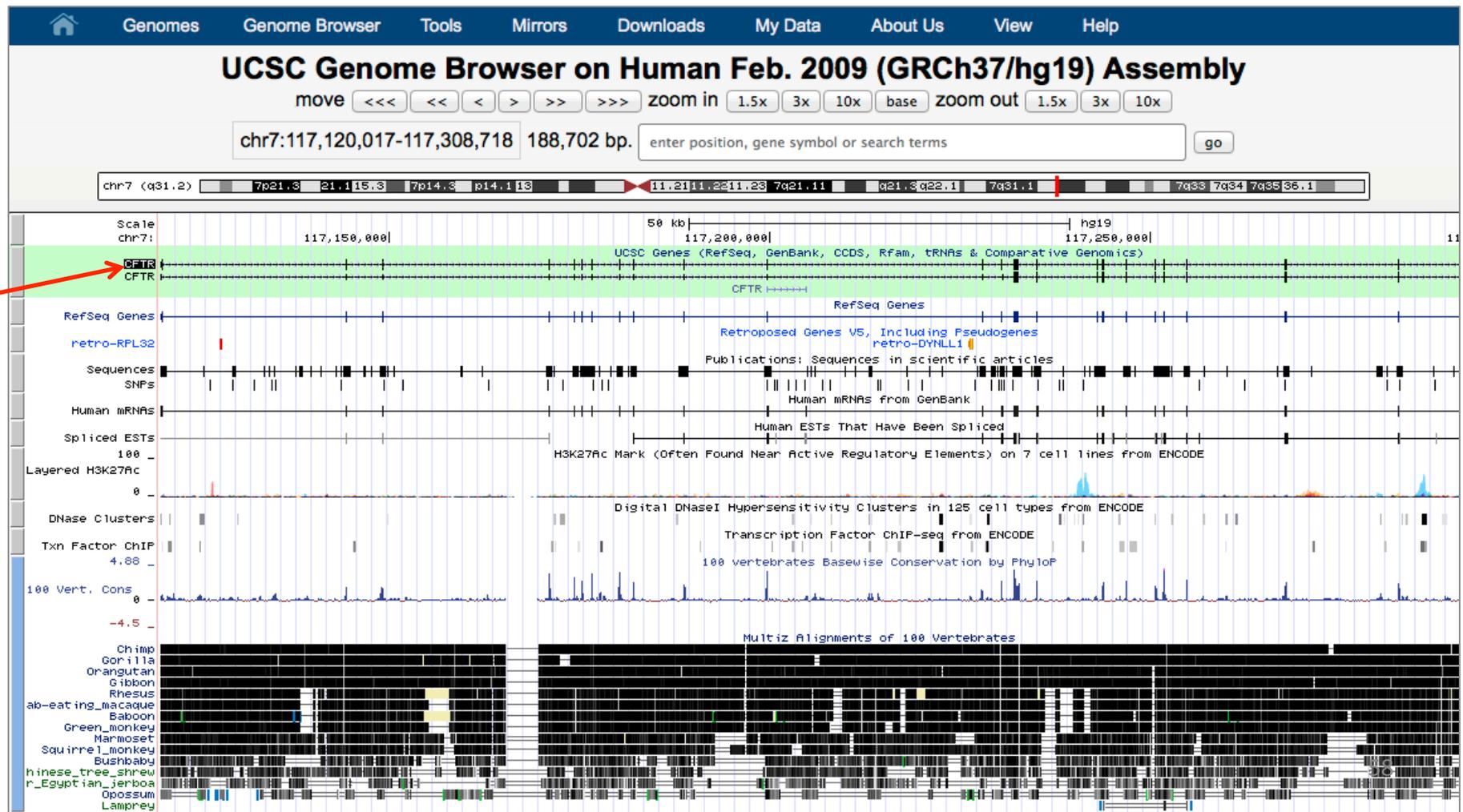


University of California, Santa Cruz Genome Browser

# UCSC Genes

→ UCSC Transcript ID but often ID is from different database

CFTR	(uc011knq.2)	at chr7:117120017-117308718	- Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
CFTR	(uc003vje.1)	at chr7:117199518-117204842	- Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
CFTR	(uc003vjd.3)	at chr7:117120017-117308718	- Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
ABCC1	(uc021tdt.1)	at chr16:16208623-16225792	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
ABCC1	(uc021tds.1)	at chr16:16208623-16225792	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
ABCC1	(uc021tdr.1)	at chr16:16149949-16170258	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
ABCC1	(uc021tdg.1)	at chr16:16146581-16170258	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 1 (ABCC1), mRNA.
ABCC8	(uc021qej.1)	at chr11:17491648-17498449	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 8 (ABCC8), mRNA.
ABCC5	(uc011bqt.2)	at chr3:183637724-183732235	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 5 (ABCC5), transcript variant 2 (ABCC5), mRNA.
ABCC11	(uc010vxl.1)	at chr16:48250025-48281478	- Homo sapiens ATP-binding cassette, sub-family C (CFTR/MRP), member 11 (ABCC11), transcript variant 1 (ABCC11), mRNA.



move start

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.

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

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

Mapping and Sequencing Tracks refresh

<a href="#">Base Position</a> dense	<a href="#">Chromosome Band</a> hide	<a href="#">STS Markers</a> hide	<a href="#">FISH Clones</a> hide	<a href="#">Recomb Rate</a> hide	<a href="#">deCODE Recomb</a> hide
<a href="#">ENCODE Pilot</a> hide	<a href="#">Map Contigs</a> hide	<a href="#">Assembly</a> hide	<a href="#">GRC Map Contigs</a> hide	<a href="#">INSDC</a> hide	<a href="#">Gap</a> hide
<a href="#">BAC End Pairs</a> hide	<a href="#">Fosmid End Pairs</a> hide	<a href="#">GC Percent</a> hide	<a href="#">GRC Patch Release</a> hide	<a href="#">Hg18 Diff</a> hide	<a href="#">GRC Incident</a> hide
<a href="#">Hi Seq Depth</a> hide	<a href="#">Wiki Track</a> hide	<a href="#">BU ORChID</a> hide	<a href="#">Mapability</a> hide	<a href="#">Short Match</a> hide	<a href="#">Restr Enzymes</a> hide

Phenotype and Disease Associations refresh

<a href="#">GAD View</a> hide	<a href="#">DECIPHER</a> hide	<a href="#">OMIM AV SNPs</a> hide	<a href="#">OMIM Genes</a> hide	<a href="#">OMIM Pheno Loci</a> hide	<a href="#">COSMIC</a> hide
<a href="#">LOVD Variants</a> hide	<a href="#">HGMD Variants</a> hide	<a href="#">UniProt Variants</a> hide	<a href="#">ClinVar Variants</a> hide	<a href="#">GWAS Catalog</a> hide	<a href="#">ISCA</a> hide
<a href="#">Coriell CNVs</a> hide	<a href="#">RGD Human QTL</a> hide	<a href="#">RGD Rat QTL</a> hide	<a href="#">MGI Mouse QTL</a> hide	<a href="#">GeneReviews</a> hide	

Genes and Gene Prediction Tracks refresh

<a href="#">UCSC Genes</a> pack	<a href="#">GENCODE...</a> hide	<a href="#">Old UCSC Genes</a> hide	<a href="#">UCSC Alt Events</a> hide	<a href="#">CCDS</a> hide	<a href="#">RefSeq Genes</a> dense
<a href="#">Other RefSeq</a> hide	<a href="#">MGC Genes</a> hide	<a href="#">ORFeome Clones</a> hide	<a href="#">TransMap...</a> hide	<a href="#">Vega Genes</a> hide	<a href="#">Pfam in UCSC Gene</a> hide

Possibilities of changes in tracks → you can easily customise your view!

# UCSC - browsing

## TASK

Please, find insulin coding gene (*INS*) in human in recent genome version and try to answer these questions:

1. Where it is located?
2. What is the length of genomic sequence?
3. How many isoforms does it have in RefSeq database?
4. How many exons does it have?
5. Find the accession number from RefSeq database
6. Where does it perform the highest expression level?
7. Do you see any repetitive elements in this region?
8. How many alleles from OMIM database you can find here? (OMIM allelic variants) → disease associations

# UCSC - download

1. For single records and regions from the genome browser (View → DNA)
2. Download → FTP
3. TableBrowser

# UCSC – download (FTP)

Home → Downloads → Genome Data

## UCSC Genome Bioinformatics

[Home](#) - [Genomes](#) - [Blat](#) - [Tables](#) - [Gene Sorter](#) - [PCR](#) - [FAQ](#) - [Help](#)

### Sequence and Annotation Downloads

This page contains links to sequence and annotation data downloads for the genome assemblies featured in the UCSC Genome Browser. For quick access to the most recent assembly of each genome, see the [current genomes](#) directory. This directory may not always contain the most recent assembly.

To view the current descriptions and formats of the tables in the annotation database, use the "describe table schema" button (no longer maintained) also provides descriptions of selected tables in the database.

All tables in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directory. Please use the corresponding download link and review the README text. These data were contributed by many researchers, as listed on the [Genome Browser](#) data you use.

#### VERTEBRATES - Complete annotation sets

<a href="#">Human</a>	<a href="#">Green Monkey</a>	<a href="#">Platypus</a>
<a href="#">Alpaca</a>	<a href="#">Guinea pig</a>	<a href="#">Rabbit</a>
<a href="#">American alligator</a>	<a href="#">Hedgehog</a>	<a href="#">Rat</a>
<a href="#">Armadillo</a>	<a href="#">Horse</a>	<a href="#">Rhesus</a>
<a href="#">Atlantic cod</a>	<a href="#">Kangaroo rat</a>	<a href="#">Rock hyrax</a>
<a href="#">Baboon</a>	<a href="#">Lamprey</a>	<a href="#">Sheep</a>
<a href="#">Bonobo</a>	<a href="#">Lizard</a>	<a href="#">Shrew</a>
<a href="#">Brown kiwi</a>	<a href="#">Malayan flying lemur</a>	<a href="#">Sloth</a>
<a href="#">Budgerigar</a>	<a href="#">Manatee</a>	<a href="#">Squirrel</a>
<a href="#">Bushbaby</a>	<a href="#">Marmoset</a>	<a href="#">Squirrel monkey</a>
<a href="#">Cat</a>	<a href="#">Medaka</a>	<a href="#">Stickleback</a>
<a href="#">Chicken</a>	<a href="#">Medium ground finch</a>	<a href="#">Tarsier</a>
<a href="#">Chimpanzee</a>	<a href="#">Megabat</a>	<a href="#">Tasmanian devil</a>
<a href="#">Chinese hamster</a>	<a href="#">Microbat</a>	<a href="#">Tenrec</a>
<a href="#">Coelacanth</a>	<a href="#">Minke whale</a>	<a href="#">Tetraodon</a>
<a href="#">Cow</a>	<a href="#">Mouse</a>	<a href="#">Tree shrew</a>

# UCSC – download (FTP)

## TASK

Please find **human** genomic sequence from **chromosome 22** in FASTA format.

**Attention! Genome version: hg38**

### [Data set by chromosome]

- What is the name of this file?
- What other files you can find here?
- Please, read more about the content.

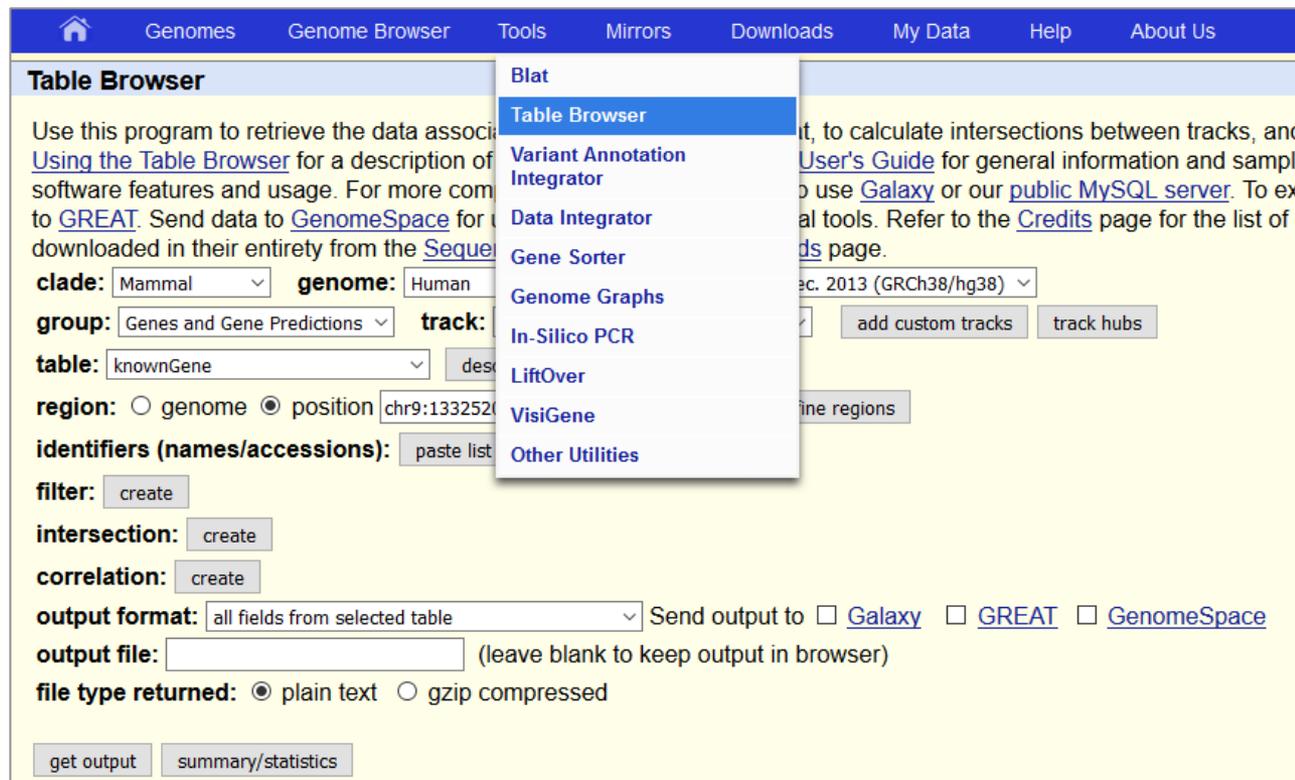
What we can find using option **[Full dataset]** ?

# UCSC – download

**ATTENTION! We can work with accession numbers not only from UCSC**

## UCSC Table Browser

<http://genome.ucsc.edu/cgi-bin/hgTables>



The screenshot displays the UCSC Table Browser interface. At the top, a navigation bar includes links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The main content area is titled "Table Browser" and contains a detailed description of the tool's purpose: to retrieve data associated with a genomic region. Below the description, several configuration options are visible: "clade" (Mammal), "genome" (Human), "group" (Genes and Gene Predictions), "track" (set to "knownGene"), "region" (position: chr9:133252...), "identifiers (names/accessions)" (with a "paste list" button), "filter" (create), "intersection" (create), "correlation" (create), "output format" (all fields from selected table), "output file" (leave blank to keep output in browser), and "file type returned" (plain text). A dropdown menu is open over the "Tools" link, listing various utilities such as Blat, Table Browser, Variant Annotation Integrator, Data Integrator, Gene Sorter, Genome Graphs, In-Silico PCR, LiftOver, VisiGene, and Other Utilities. The "Table Browser" option in the menu is highlighted.

# UCSC – download Table Browser

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

## Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

**clade:** Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

**group:** Custom Tracks **track:** User track [manage custom tracks](#) [track hubs](#)

**table:** ct\_UserTrack\_3545 [describe table schema](#)

**region:**  genome  ENCODE Pilot regions **position:** chr7:127469864-127481543 [lookup](#) [define regions](#)

**identifiers (names/accessions):** [paste list](#) [upload list](#)

**filter:** [create](#)

**intersection:** [create](#)

**correlation:** [create](#)

**output format:** BED – browser extensible data  Send output to  [Galaxy](#)  [GREAT](#)

**output file:**  (leave blank to keep output in browser)

**file type returned:**  plain text  gzip compressed

[get output](#) [summary/statistics](#)

To reset all user cart settings (including custom tracks), [click here](#).

## Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the [Table Browser User's Guide](#).

# UCSC – download Table Browser

## TASK

Please, find all repetitive elements in **BED** format from human chromosome 22.

(ATTENTION! Recent genome version)

through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use v restrictions associated with these data. All tables can be downloaded in their entirety from the [S](#)

**clade:** Mammal  **genome:** Human  **assembly:** Dec. 2013 (GRCh38/hg38)

**group:** Repeats  **track:** RepeatMasker

**table:** rmsk

**region:**  genome  position chr22

**identifiers (names/accessions):**

**filter:**

**intersection:**

**output format:** BED - browser extensible data

**output file:**  (leave blank)

**file type returned:**  plain text  gzip compressed

### RepeatMasker (rmsk) Summary Statistics

item count	79,521						
item bases	20,908,038 (53.39%)						
item total	20,9	chr22	10510227	10510528	AluSx1	2021	+
smallest item		chr22	10511018	10511332	L1MC5a	781	-
average item		chr22	10511479	10511791	L1MB1	524	+
biggest item		chr22	10511878	10512212	L1MB1	313	+
smallest score		chr22	10512454	10512692	L1MB1	656	+
average score		chr22	10512706	10514778	L1MB1	11092	+
biggest score		chr22	10514778	10515050	AluSx1	1933	+
		chr22	10515050	10515074	L1MB1	11092	+
		chr22	10515074	10515121	(GAAG)n	52	+
		chr22	10515121	10516103	L1MB1	11092	+
		chr22	10516114	10516222	(TA)n	47	+
		chr22	10516223	10516285	LTR66	237 <sup>66</sup>	-
		chr22	10516287	10516630	L1MB1	1504	+

# UCSC – download Table Browser

## TASK

Please download **GTF** file for all **cat** genes located on X chromosome but only from RefSeq database. (**Recent genome version**).

clade: Mammal genome: Cat assembly: Nov. 2014 (ICGSC Felis\_catus\_8.0/felCat8)  
 group: Genes and Gene Predictions track: RefSeq Genes add custom tracks track hubs  
 table: refGene describe table schema  
 region:  genome  position chrA2:53484451-53597660 lookup define regions  
 identifiers (names/accessions): paste list upload list  
 filter: create  
 intersection: create  
 correlation: create  
 output format: GTF - gene transfer format Send output to  Galaxy  GREAT  GenomeSpace  
 (leave blank to keep output in browser)  
 gzip compressed

### RefSeq Genes (refGene) Summary

item count	11
item bases	183,230 (0.15%)
item total	183,230 (0.15%)

smallest item	chrX	felCat8_refGene	exon	21883079	21883169	0.000000	+	.	gene_id	"NM_001009262"; transcript_id	"NM_001009262";
	chrX	felCat8_refGene	start_codon	21890310	21890312	0.000000	+	.	gene_id	"NM_001009262"; transcript_id	"NM_001009262";
	chrX	felCat8_refGene	CDS	21890310	21891434	0.000000	+	0	gene_id	"NM_001009262"; transcript_id	"NM_001009262";
average item	chrX	felCat8_refGene	stop_codon	21891435	21891437	0.000000	+	.	gene_id	"NM_001009262"; transcript_id	"NM_001009262";
	chrX	felCat8_refGene	exon	21890249	21891487	0.000000	+	.	gene_id	"NM_001009262"; transcript_id	"NM_001009262";
	chrX	felCat8_refGene	exon	21891501	21891669	0.000000	+	.	gene_id	"NM_001009262"; transcript_id	"NM_001009262";
biggest item	chrX	felCat8_refGene	stop_codon	42849134	42849136	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	CDS	42849137	42849283	0.000000	-	0	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	exon	42849134	42849283	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	CDS	42849493	42849594	0.000000	-	0	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
block count	chrX	felCat8_refGene	exon	42849493	42849594	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
block bases	chrX	felCat8_refGene	CDS	42850334	42850410	0.000000	-	2	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
block total	chrX	felCat8_refGene	exon	42850334	42850410	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	CDS	42851084	42851234	0.000000	-	0	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	exon	42851084	42851234	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
smallest block	chrX	felCat8_refGene	CDS	42852193	42852273	0.000000	-	0	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	exon	42852193	42852273	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	CDS	42852462	42852549	0.000000	-	1	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
average block	chrX	felCat8_refGene	exon	42852462	42852549	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
	chrX	felCat8_refGene	CDS	42853442	42853546	0.000000	-	1	gene_id	"NM_001083952"; transcript_id	"NM_001083952";
biggest block	chrX	felCat8_refGene	exon	42853442	42853546	0.000000	-	.	gene_id	"NM_001083952"; transcript_id	"NM_001083952";

# UCSC – download Table Browser

## TASK

Let's check how many **single exon genes** there is annotated in UCSC (Old UCSC Genes) in the most recent version of **human** genome but only on **chromosome 4**.

clade: Mammal genome: Human assembly: Dec. 2013 (GRCh38/hg38)  
group: Genes and Gene Predictions track: Old UCSC Genes  
table: knownGeneOld9  
region: genome position chr9:133252000-133280861  
identifiers (names/accessions):  
filter: create  
intersection: create  
correlation: create  
output format: sequen  
output file:  
file type returned:   
get output summary/st

Genomes Genome Browser Tools Mirrors Downloads

Filter on Fields from hg38.knownGene

name	does	match	*	
chrom	does	match	*	AND
strand	does	match	*	AND
txStart	is	ignored	0	AND
txEnd	is	ignored	0	AND
cdsStart	is	ignored	0	AND
cdsEnd	is	ignored	0	AND
exonCount	is	=	1	AND
exonStarts	does	match	*	
exonEnds	does	match	*	
proteinID	does	match	*	AND
alignID	does	match	*	AND

Free-form query:   
submit cancel

GREAT

item count	665
item bases	391,984 (0.21%)
item total	395,125 (0.21%)
smallest item	43
average item	594
biggest item	9,848
block count	665
block bases	391,984 (0.21%)
block total	395,125 (0.21%)
smallest block	43
average block	594
biggest block	9,848

# Other genomic and annotation resources



Genomes Online Database

<http://genomesonline.org/cgi-bin/GOLD/index.cgi>



Wellcome Trust Sanger Institute

<https://www.sanger.ac.uk/resources/downloads/>



<http://www.cbs.dtu.dk/services/GenomeAtlas/>

# Other genomic and annotation resources



Genomes – EMBL - EBI

<http://www.ebi.ac.uk/genomes/index.html>

## Genomes Pages - At the EBI

- Complete genomes
- Archaea
- Archaeal virus
- Bacteria
- Eukaryota
- Organelle
- Phage
- Plasmid
- Viroid
- Virus
- Links

Databases > Nucleotide > The European Nucleotide Archive > Complete Genomes

### Access to Completed Genomes

The first completed genomes from [viruses](#), [phages](#) and [organelles](#) were deposited into the EMBL Database in the early 1980's. Since then, molecular biology's shift to obtain the complete sequences of as many genomes as possible combined with major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including [Archaea](#), [Bacteria](#) and [Eukaryota](#). These web pages give access to a large number of complete genomes, [help](#) is available to describe the layout.

### Whole Genome Shotgun Sequences (WGS)

Methods using whole genome shotgun data are used to gain a large amount of genome coverage for an organism. WGS data for a growing number of organisms are being submitted to DDBJ/EMBL/GenBank.

[More information about WGS projects...](#)

### Last 40 Genome Entries

Date	Accession	Description
27-JUN-2014	<a href="#">AP014565.1</a>	Salmonella enterica subsp. enterica serovar Typhimurium str. L-3553
27-JUN-2014	<a href="#">AP014566.1</a>	Salmonella enterica subsp. enterica serovar Typhimurium str. L-3553 plasmid pST3553
27-JUN-2014	<a href="#">KJ634409.1</a>	Yerba mate endornavirus strain INTA
27-JUN-2014	<a href="#">KJ739609.1</a>	Pomacea canaliculata mitochondrion
26-JUN-2014	<a href="#">KJ562277.1</a>	Amphilophus citrinellus mitochondrion
26-JUN-2014	<a href="#">KJ650081.1</a>	Paratritioza sinica mitochondrion
26-JUN-2014	<a href="#">KJ668270.1</a>	Ecotropic murine leukemia virus
26-JUN-2014	<a href="#">KJ668271.1</a>	Polytropic murine leukemia virus
25-JUN-2014	<a href="#">HF920637.1</a>	Armadillidium vulgare iridescent virus
25-JUN-2014	<a href="#">HG938353.1</a>	Rhizobium galegae str. HAMB1 540
25-JUN-2014	<a href="#">HG938355.1</a>	Rhizobium galegae str. HAMB1 1141
25-JUN-2014	<a href="#">HG938356.1</a>	Rhizobium galegae str. HAMB1 1141 chromid pHAMB11141a
25-JUN-2014	<a href="#">HG938357.1</a>	Rhizobium galegae str. HAMB1 1141 plasmid pHAMB11141b
25-JUN-2014	<a href="#">HG975439.1</a>	Solanum pennellii chromosome ch00
25-JUN-2014	<a href="#">HG975440.1</a>	Solanum pennellii chromosome ch01
25-JUN-2014	<a href="#">HG975441.1</a>	Solanum pennellii chromosome ch02



# Basic resources related to the next generation sequencing data

SRA NCBI

ENA EBI

GEO NCBI

ENCODE



## SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

# Sequence Read Archive (SRA)

- **Basic source of NGS data**
- User-friendly browsing and filtering system
- It is possible to submit your own data

Sequence Read Archive contains reads generated by many platforms: 454, IonTorrent, Illumina, SOLiD, Pacific Biosciences oraz Oxford Nanopore Technologies.

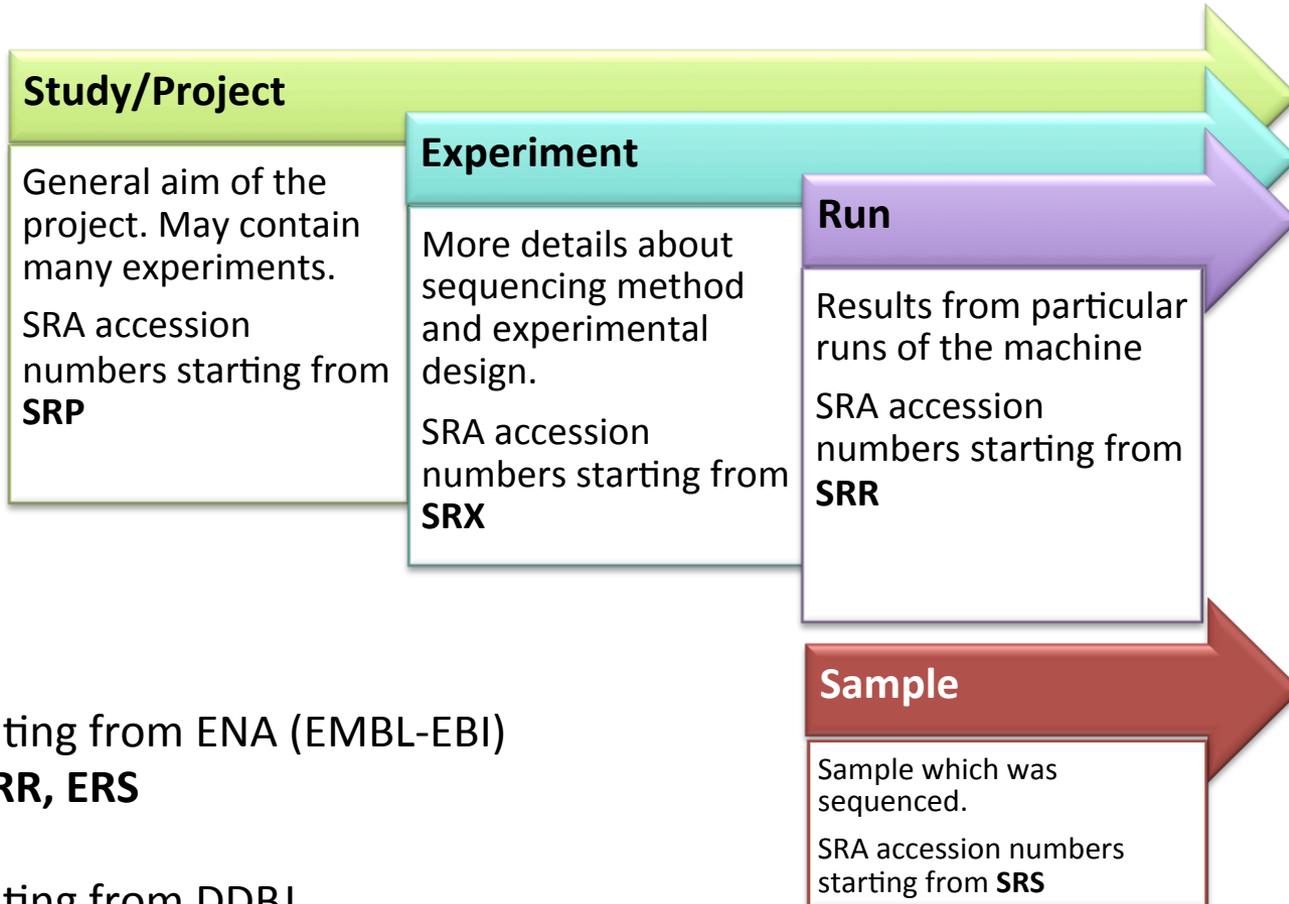
SRA cooperates with ENA and vice versa, as well as DDBJ.



## SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

# Sequence Read Archive (SRA)



Data originating from ENA (EMBL-EBI)  
**ERP, ERX, ERR, ERS**

Data originating from DDBJ  
**DRP, DRX, DRR, DRS**

# SRA: search builder



<http://www.ncbi.nlm.nih.gov/sra/>

Advanced Search > Organism: **Arabidopsis thaliana (1)**

Platform: „ **illumina**” (2), Properties: „**instrument illumina hiseq 2000**” (3)

SRA    
Advanced

## SRA Advanced Search Builder

((Arabidopsis thaliana[Organism]) AND "illumina"[Platform]) AND "instrument illumina hiseq 2000"[Properties]

[Edit](#)

[Clear](#)

### Builder

Organism   [Show index list](#)

AND  Platform  [Platform]  [Show index list](#)

AND  Properties  [Properties]  [Hide index list](#)

- instrument illumina hiscansq (2393)
- instrument illumina hiseq 1000 (11425)
- instrument illumina hiseq 1500 (4716)
- instrument illumina hiseq 2000 (1145604)**
- instrument illumina hiseq 2500 (297832)
- instrument illumina hiseq 3000 (1534)
- instrument illumina hiseq 4000 (2450)
- instrument illumina miseq (343892)
- instrument ion torrent pgm (21355)
- instrument ion torrent proton (3165)

[Previous 200](#)

[Next 200](#)

[Refresh index](#)

AND     [Show index list](#)

or [Add to history](#)

# SRA: search builder

<http://www.ncbi.nlm.nih.gov/sra/>

SRA    [Create alert](#) [Advanced](#) [Help](#)

Access: Public (7,748)  
Source: DNA (4,142), RNA (3,564)  
Type: genome (2,100)  
Other: aligned data (166)  
[Clear all](#)  
[Show additional filters](#)

Summary ▾ 20 per page ▾ Send to: ▾ **Filters:** [Manage Filters](#)

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

**Search results**  
Items: 1 to 20 of 7748 << First < Prev Page 1 of 388 Next > Last >>

- [Illumina HiSeq 2000 sequencing: Ribosome footprinting upon oxidative stress treatment for wild-type and catalase2 knock-out mutants](#)  
1 ILLUMINA (Illumina HiSeq 2000) run: 160.4M spots, 16.2G bases, 9Gb downloads  
Accession: ERX1495854
- [Illumina HiSeq 2000 sequencing: Ribosome footprinting upon oxidative stress treatment for wild-type and catalase2 knock-out mutants](#)  
1 ILLUMINA (Illumina HiSeq 2000) run: 180.9M spots, 18.3G bases, 10Gb downloads  
Accession: ERX1495853

**Find related data** Database:

**Search details**  
{"Arabidopsis thaliana"[Organism] AND "illumina"[Platform]) AND "instrument illumina hiseq 2000"[Properties]}  [See more...](#)

**Recent activity**

# SRA: experiment

## [SRX2206057](#): Other Sequencing of thale cress

1 ILLUMINA (Illumina HiSeq 2000) run: 32.4M spots, 5.8G bases, 3.6Gb downloads

**Submitted by:** University of Helsinki

**Study:** Arabidopsis genome re-sequencing

[PRJNA345097](#) • [SRP090723](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

**Sample:**

[SAMN05785616](#) • [SRS1724473](#) • [All experiments](#) • [All runs](#)

*Organism:* [Arabidopsis thaliana](#)

**Library:**

*Name:* Cvi-0

*Instrument:* Illumina HiSeq 2000

*Strategy:* WGS

*Source:* GENOMIC

*Selection:* RANDOM

*Layout:* PAIRED

**Runs:** 1 run, 32.4M spots, 5.8G bases, [3.6Gb](#)

Run	# of Spots	# of Bases	Size	Published
<a href="#">SRR4333274</a>	32,394,270	5.8G	3.6Gb	2016-10-07

ID: 3240811

# SRA: Run

NCBI [Site map](#) [All databases](#) [Search](#)

## Sequence Read Archive

[Main](#) [Browse](#) [Search](#) [Download](#) [Submit](#) [Documentation](#) [Software](#) [Trace Archive](#) [Trace Assembly](#) [Trace BLAST](#)

[Studies](#) [Samples](#) [Analyses](#) **Run Browser** [Run Selector](#) [Provisional SRA](#)

### (SRR4333274)

[Metadata](#) [Reads](#) [Download](#)

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR4333274	32.4M	5.8Gbp	3.9G	36.8%	2016-10-07	public

This run has 2 reads per spot:

L=90, 100%	L=90, 100%
------------	------------

[Legend](#)

Experiment	Library												
<a href="#">SRX2206057</a>	<table border="1"><thead><tr><th>Name</th><th>Platform</th><th>Strategy</th><th>Source</th><th>Selection</th><th>Layout</th></tr></thead><tbody><tr><td>Cvi-0</td><td>Illumina</td><td>WGS</td><td>GENOMIC</td><td>RANDOM</td><td>PAIRED</td></tr></tbody></table>	Name	Platform	Strategy	Source	Selection	Layout	Cvi-0	Illumina	WGS	GENOMIC	RANDOM	PAIRED
Name	Platform	Strategy	Source	Selection	Layout								
Cvi-0	Illumina	WGS	GENOMIC	RANDOM	PAIRED								

[to BLAST](#)

Biosample	Sample Description	Organism
<a href="#">SAMN05785616</a> (SRS1724473)		<a href="#">Arabidopsis thaliana</a>

Bioproject	SRA Study	Title
<a href="#">PRJNA345097</a>	<a href="#">SRP090723</a>	Arabidopsis genome re-sequencing

**Abstract:**  
Arabidopsis strains with interesting phenotypes were sequenced in order to map the genes behind phenotypes.

# SRA: Reads

NCBI Site map All databases Search

Sequence Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

(SRR4333274)

Metadata Reads Download

Filter:  Find Filtered Download [What does it do?](#)

[What can the filter be applied to?](#)

The Run is too big (>1.1G) for searching by sequence substring.

< 1 1 3239427 >

View:  biological reads  technical reads  quality scores [advanced options](#)

**Reads (separated)**

- SRR4333274.1 SRS1724473**  
name: FCD0T67ACXX:8:1101:1243:1935,  
member: GCCAATAT  
x: 1243, y: 1935  
>gnlISRAISRR4333274.1.1 FCD0T67ACXX:8:1101:1243:1935 forward (Biological)  
NCAATAACCTCAGACCAATCAATCGAATATTGATTAATACATAATTGATCGAACACTACT  
TGAAAACGGCTCTTCCGCTCAGAAACGAAA
- SRR4333274.2 SRS1724473**  
name: FCD0T67ACXX:8:1101:1250:1955,  
member: GCCAATAT  
x: 1250, y: 1955  
>gnlISRAISRR4333274.1.2 FCD0T67ACXX:8:1101:1243:1935 reverse (Biological)  
ACCGAGAGATCCATAAATCGGGATCCTAATGCATATAGATACAAATGGTCCAATGGGAGC  
AAGAAATTCAGGAGCATTTGGAACATTTTC
- SRR4333274.3 SRS1724473**  
name: FCD0T67ACXX:8:1101:1249:1982,  
member: GCCAATAT  
x: 1249, y: 1982
- SRR4333274.4 SRS1724473**  
name: FCD0T67ACXX:8:1101:1169:1988,  
member: GCCAATAT  
x: 1169, y: 1988
- SRR4333274.5 SRS1724473**  
name: FCD0T67ACXX:8:1101:1229:1997,  
member: GCCAATAT  
x: 1229, y: 1997
- SRR4333274.6 SRS1724473**  
name: FCD0T67ACXX:8:1101:1360:1928,  
member: GCCAATAT  
x: 1360, y: 1928

# SRA: download

1. Filtered download (Run Browser → Reads → Filter)
2. Download → FASTA/FASTQ
3. SRA Toolkit
4. FTP (difficult access)
5. Aspera
6. <http://sradownload.com/>

## SRA Toolkit Documentation

[SRA Toolkit Installation and Configuration Guide](#)

[Protected Data Usage Guide](#)

### Frequently Used Tools:

[fastq-dump](#): Convert SRA data into fastq format

[prefetch](#): Allows command-line downloading of SRA, dbGaP, and ADSP data

[sam-dump](#): Convert SRA data to sam format

[sra-pileup](#): Generate pileup statistics on aligned SRA data

[vdb-config](#): Display and modify VDB configuration information

[vdb-decrypt](#): Decrypt non-SRA dbGaP data ("phenotype data")

### Additional Tools:

[abi-dump](#): Convert SRA data into ABI format (csfasta / qual)

[illumina-dump](#): Convert SRA data into Illumina native formats (qseq, etc.)

[sff-dump](#): Convert SRA data to sff format

[sra-stat](#): Generate statistics about SRA data (quality distribution, etc.)

[vdb-dump](#): Output the native VDB format of SRA data.

[vdb-encrypt](#): Encrypt non-SRA dbGaP data ("phenotype data")

[vdb-validate](#): Validate the integrity of downloaded SRA data

## Download NCBI SRA files

Enter the Run Number and Download SRA file

ByRunName

Submit

Example SRA-run number : SRR1149591

Find the SRA run number by Study

Submit

Steps to get SRA run number

- Enter the Study name a window will open in new tab.
- Find desired study and select the run number.
- Enter the selected run number in the input of "ByRunName"

# SRA: BLAST

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

**BLAST** ® » blastn suite

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

**Enter Query Sequence**

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#) Query subrange [?](#)

From

To

Or, upload file  No file selected. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

**Choose Search Set**

Database  Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):

◆ Sequence Read Archive (SRA)

Enter an SRA accession (experiment, study, or submission)

Enter an SRA accession (experiment, study, or submission), title, the scientific name or tax id. Only 20 top suggestions will be shown. [?](#)

**Program Selection**

Optimize for  Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

**BLAST** Search database SRA using Megablast (Optimize for highly similar sequences)

Show results in a new window

[+ Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted in yellow and marked with ◆ sign**

80

# SRA: TASK

- Please find in SRA some reads from experiments for human performed on selected platform, for example from transcriptome sequencing (Properties → strategy rna-seq), where as a result paired reads were obtained (Layout → paired).
- For some selected record, please find all possible accession numbers (Study, Experiment, Run, Sample ) and short description of experiment.

EMBL-EBI 

Services | Research | Training | About us



European Nucleotide Archive

Search

Examples: [BN000065](#), [histone](#)

[Advanced](#)  
[Sequence](#)

Home | Search & Browse | Submit & Update | Software | About ENA | Support

## European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)

Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.

### Text Search

Examples: [BN000065](#), [histone](#)

Search

[Advanced search](#)

### Sequence Search

Enter or paste a nucleotide sequence or accession number

Search

[Advanced search](#)

### Popular

- o [Submit and update](#)
- o [Sequence submissions](#)
- o [Genome assembly submissions](#)
- o [Submitting environmental sequences](#)
- o [Citing ENA data](#)
- o [Rest URLs for data retrieval](#)
- o [Rest URLs to search ENA](#)

### Latest ENA news

#### 09 Dec 2014: [ENA release 122](#)

Release 122 of ENA's assembled/annotated sequences is now available.

#### 12 Nov 2014: [Simplification of data release procedures](#)

The European Nucleotide Archive will couple the public release of sequence records and the release of study records that contain these sequence records, with immediate effect.

#### 11 Nov 2014: [ENA/EMG Sample Record Annotation Workshop](#)

European Nucleotide Archive (ENA) and EBI Metagenomics Portal (EMG), are organising the ENA/EMG Sample Record Annotation Workshop on the 1-5 December 2014 to enrich the environmental sample records.

## Accession numbers

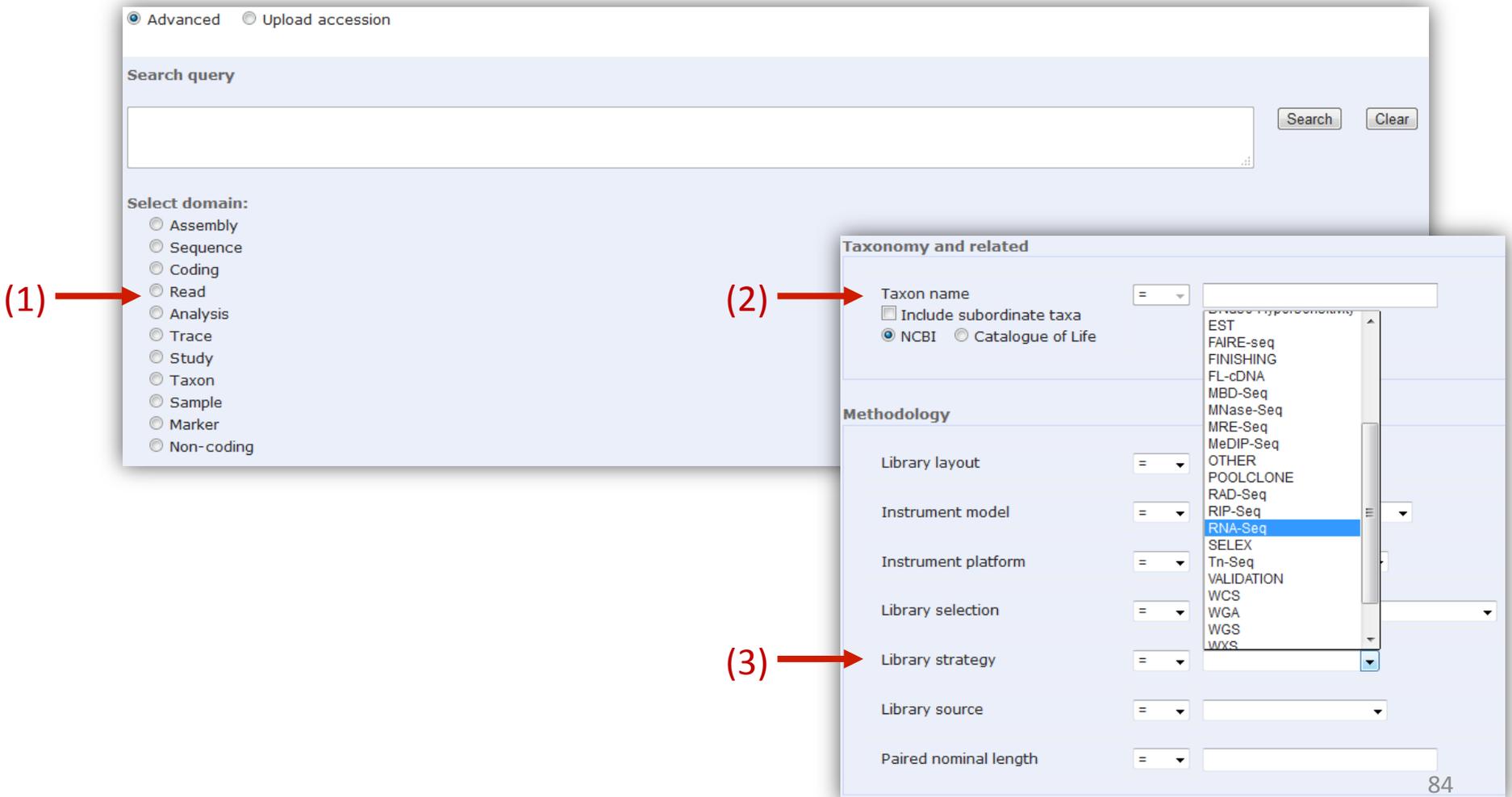
Metadata object	Accession prefix	Number of digits	Example
Submission	ERA, SRA, DRA	6	ERA000092
Sample	ERS, SRS, DRS	6	ERS000081
Study	ERP, SRP, DRP	6	ERP000016
Experiment	ERX, SRX, DRX	6	ERX000398
Run	ERR, SRR, DRR	6	ERR003990
Analysis	ERZ, SRZ, DRZ	6	ERZ000001
Project/Study	PRJ		

ER - EBI

SR – SRA NCBI

DR - DDBJ

Select domain: **read** (1), Taxon name: **Bos taurus** (2), Library strategy = **RNA-Seq** (3)



The screenshot shows the ENA search interface with the following elements and annotations:

- Advanced** (selected) / Upload accession
- Search query: [Empty text box]
- Buttons: Search, Clear
- Select domain:**
  - Assembly
  - Sequence
  - Coding
  - Read** (1)
  - Analysis
  - Trace
  - Study
  - Taxon
  - Sample
  - Marker
  - Non-coding
- Taxonomy and related**
  - Taxon name: [Empty text box] (2)
  - Include subordinate taxa
  - NCBI /  Catalogue of Life
- Methodology**
  - Library layout: [=]
  - Instrument model: [=]
  - Instrument platform: [=]
  - Library selection: [=]
  - Library strategy: [=] **RNA-Seq** (3)
  - Library source: [=]
  - Paired nominal length: [=]

The dropdown menu for Library strategy is open, showing the following options: EST, FAIRE-seq, FINISHING, FL-cDNA, MBD-Seq, MNase-Seq, MRE-Seq, MeDIP-Seq, OTHER, POOLCLONE, RAD-Seq, RIP-Seq, **RNA-Seq**, SELEX, Trn-Seq, VALIDATION, WCS, WGA, WGS, WXS.

Advanced  Upload accession

Search query [Help](#)

[Query builder](#)  
[Edit query](#)

Search results for *tax\_eq(9913) AND library\_strategy='RNA-Seq'*

**Read**  
Study (46)  
Experiment (797)  
Run (824)

**Study (46 results found)**  
PRJDA72405 RNA-sequencing of bovine GV and M2 oocytes and 8-cell stage embryos.  
[View all 46 results](#)

**Experiment (797 results found)**  
DRX000961 Illumina Genome Analyzer II sequencing; RNA-sequencing of bovine granulosa cells from young cows (28.3±0.7 months)  
[View all 797 results](#)

**Run (824 results found)**  
DRR001364 Illumina Genome Analyzer II sequencing; RNA-sequencing of bovine granulosa cells from young cows (28.3±0.7 months)  
[View all 824 results](#)

Read Files

Portal

Attributes

Publications

[Download files](#)

View: [TEXT](#)

Download: [TEXT](#)

[Select columns](#)

Showing results 1 - 10 of 16 results

[Next](#)

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)
<a href="#">PRJDA72405</a>	<a href="#">DRP000449</a>	<a href="#">SAMD00000728</a>	<a href="#">DRS000859</a>	<a href="#">DRX000961</a>	<a href="#">DRR001364</a>	9913	<a href="#">Bos taurus</a>	Illumina Genome Analyzer II	SINGLE	<a href="#">File 1</a>	<a href="#">File 1</a>		
<a href="#">PRJDA72405</a>	<a href="#">DRP000449</a>	<a href="#">SAMD00000728</a>	<a href="#">DRS000859</a>	<a href="#">DRX000962</a>	<a href="#">DRR001365</a>	9913	<a href="#">Bos taurus</a>	Illumina Genome Analyzer II	SINGLE	<a href="#">File 1</a>	<a href="#">File 1</a>		
<a href="#">PRJDA72405</a>	<a href="#">DRP000449</a>	<a href="#">SAMD00000728</a>	<a href="#">DRS000859</a>	<a href="#">DRX000962</a>	<a href="#">DRR001366</a>	9913	<a href="#">Bos taurus</a>	Illumina Genome Analyzer II	SINGLE	<a href="#">File 1</a>	<a href="#">File 1</a>		
<a href="#">PRJDA72405</a>	<a href="#">DRP000556</a>	<a href="#">SAMD00011319</a>	<a href="#">DRS001301</a>	<a href="#">DRX001352</a>	<a href="#">DRR001889</a>	9913	<a href="#">Bos taurus</a>	Illumina Genome Analyzer II	SINGLE	<a href="#">File 1</a>	<a href="#">File 1</a>		
<a href="#">PRJDA72405</a>	<a href="#">DRP000556</a>	<a href="#">SAMD00011318</a>	<a href="#">DRS001302</a>	<a href="#">DRX001354</a>	<a href="#">DRR001890</a>	9913	<a href="#">Bos taurus</a>	Illumina Genome Analyzer II	SINGLE	<a href="#">File 1</a>	<a href="#">File 1</a>		
<a href="#">PRJDA72405</a>	<a href="#">DRP000556</a>	<a href="#">SAMD00011318</a>	<a href="#">DRS001302</a>	<a href="#">DRX001354</a>	<a href="#">DRR001891</a>	9913	<a href="#">Bos taurus</a>	Illumina Genome Analyzer II	SINGLE	<a href="#">File 1</a>	<a href="#">File 1</a>		

## Download options:

1. FTP files
2. ENA Browser
3. Aspera

Attention! Sequences from archive, older version of the database can have shorter accession numbers.

[http://www.ebi.ac.uk/ena/browse/read-download#downloading\\_files\\_ena\\_browser](http://www.ebi.ac.uk/ena/browse/read-download#downloading_files_ena_browser)

### Enter query sequence

Enter or paste a sequence or accession

Or upload a file:  No file selected.

Limit query sequence to range:

### Search against

Assembled and annotated sequences       Barcode sequences       Coding sequences  
 Geo-referenced sequences       Non-coding sequences       Vectors (Emvec)

Limit sequence by:     Taxonomic group     Data class

### Set parameters

Program

[Less options](#)

Result options	Scoring options	General options
Maximum scores: <input type="button" value="50"/>	Match/mismatch scores: <input type="button" value="2,-3"/>	Align: <input type="checkbox"/> Align using gaps
Maximum alignments: <input type="button" value="50"/>	Drop off: <input type="button" value="0"/>	
Expect threshold: <input type="button" value="10"/>	Gap existence cost: <input type="button" value="5"/>	
Alignment view: <input type="button" value="pairwise"/>	Gap extension cost: <input type="button" value="2"/>	

# ENA: TASK

- Please find in ENA some reads from experiments for human performed on selected platform, for example from transcriptome sequencing (Library strategy → RNA-seq), where as a result paired reads were obtained (Library layout → paired).
- For some selected record, please find all possible accession numbers (Study, Experiment, Run, Sample ) and short description of experiment.

# GEO – Gene Expression Omnibus (NCBI)

<http://www.ncbi.nlm.nih.gov/gds>



**GEO DataSets**

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Advanced Search > Organism: **Mus musculus (1)**, DataSet Type: „non coding rna profiling by high throughput sequencing” (2)

(mus musculus[Organism]) AND "non coding rna profiling by high throughput sequencing"[DataSet Type]

[Edit](#) [Clear](#)

**Builder**

Organism  ← (1) [Show index list](#)

AND  [DataSet Type] [Hide index list](#)

methylation profiling by high throughput sequencing (632)

methylation profiling by snp array (9)

non coding rna profiling by array (1975)

non coding rna profiling by genome tiling array (105)

non coding rna profiling by high throughput sequencing (1341) ← (2.2)

other (999)

protein profiling by mass spec (4)

protein profiling by protein array (151)

snp genotyping by snp array (483)

third party reanalysis (118)

[Previous 200](#)  
[Next 200](#)  
↑ (2.1)  
[Refresh index](#)

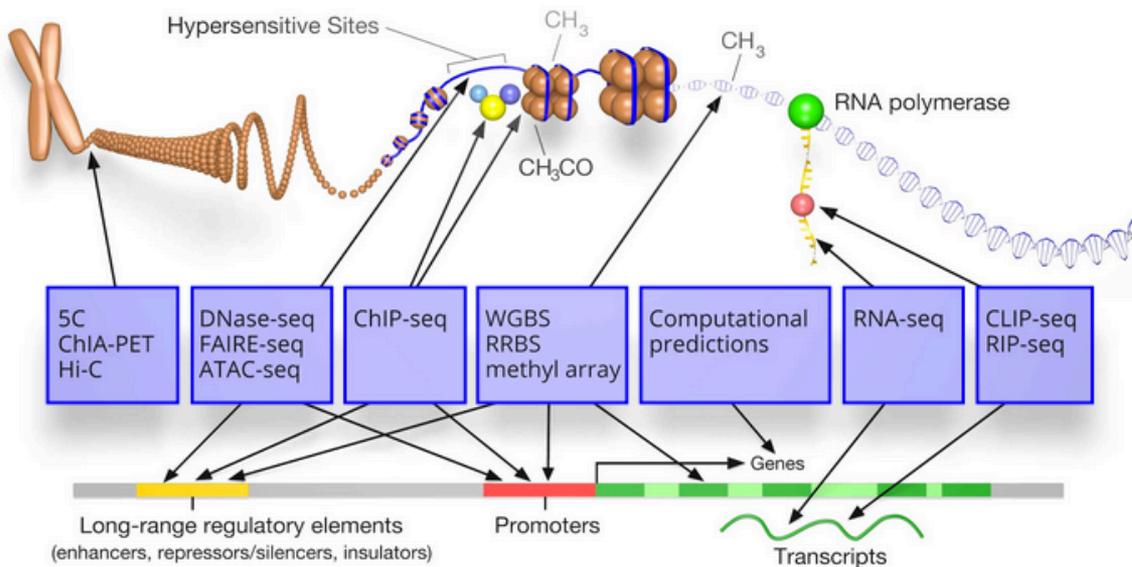
AND  [Show index list](#)

or [Add to history](#)

# ENCODE

<https://www.encodeproject.org/>

## ENCODE: Encyclopedia of DNA Elements



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

[Get Started](#)

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

HUMAN

MOUSE

WORM

FLY

[View Assay Matrix](#)

- Matrix
- Search
- Search by region
- Reference epigenomes
- Publications

**Experi**

Click on the  
experiments

Enter search term(s)

Assay	Assay category	Target of assay	Date released	Available data
ChIP-seq 7516	DNA binding 7516	transcription factor 3148	July, 2013 2533	fastq 11329
DNase-seq 764	Transcription 3142	histone 2894	March, 2014 881	bam 10307
polyA mRNA RNA-seq 722	DNA accessibility 960	histone modification 2894	July, 2016 612	bigWig 9161
shRNA RNA-seq 524	DNA methylation 715	control 2150	May, 2016 558	bed narrowPeak 5853
total RNA-seq 493	RNA binding 602	broad histone mark 1607	October, 2016 468	bigBed narrowPeak 5513

**Organism**

<i>Homo sapiens</i>	9781
<i>Mus musculus</i>	1777
<i>Drosophila melanogaster</i>	986
<i>Caenorhabditis elegans</i>	639
<i>Drosophila pseudoobscura</i>	10

**Biosample type**

immortalized cell line	4980
tissue	3987
primary cell	1790
whole organisms	1316
in vitro differentiated cells	610

**Organ**

blood	1941
liver	1068
brain	861
embryo	846
lung	721

**Project**

ENCODE	8853
Roadmap	2535
modENCODE	877
modERN	773
GGR	331

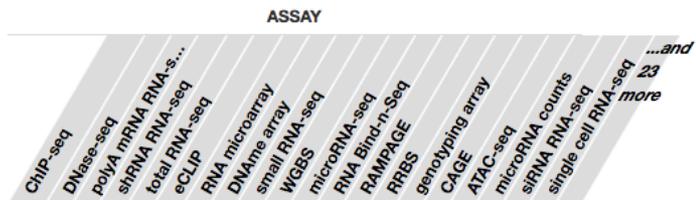
**Genome assembly (visualization)**

hg19	5398
GRCh38	5002
mm10	1434
dm3	604
ce10	542

BIOSAMPLE

**13369 results**

Clear Filters



**immortalized cell line**

K562	582	7	19	270	11	180	10	3	7	1	1	1	1	2	9	1	50
HepG2	316	3	11	254	6	148	6	3	3	1				2	2	6	1
A549	326	14	21					2	2	9				1	2	3	
GM12878	225	2	11	5		7	3	6	1	1			1	2	2	6	13
HEK293	226							1	2					2	2		

**tissue**

liver	156	5	12	11		1	1	8	7	2	1	3	3	7	2		
heart	99	20	8	11	10		1	9	9	2				8	1		
stomach	94	18	11	9		3	4	8	4	5	1	1		3	4		
lung	79	15	8	5	10	2	1	7	4	1	1			1	4		
forebrain	72	2		9				8	9					7	8	2	

**primary cell**

common myeloid progenitor, CD34-positive	67	13	1			12		1					3				
IMR-90	55	2	3	2		1	2	9	2	1			1	3	3		
foreskin fibroblast	30	5	4			3	2	1	1	2			2	2	1		
endothelial cell of umbilical vein	35	2	5			2	1	1						1	5	1	
Purkinje cell						1											61

**whole organisms**

whole organism	1140	73	50												15		
carcass		12	4												4		

**in vitro differentiated cells**

mesenchymal stem cell	73	1	4						1								
dendritic cell	11			25											30		
neural stem progenitor cell	32	1	4						2				1				
trophoblast cell	27	1	2						1								
myocyte	26	1	2	1						1							1



# Other selected databases

- NONCODE  
<http://noncode.org/>
- RNA-Seq Atlas:  
[http://medicalgenomics.org/rna\\_seq\\_atlas/](http://medicalgenomics.org/rna_seq_atlas/)
- The Cancer Genome Atlas:  
<https://tcga-data.nci.nih.gov/tcga/>
- MedPlant RNA-Seq Database:  
<http://www.medplanrnaseq.org/>
- 1000 Genomes Project:  
<http://www.internationalgenome.org/data>

# Where to look for help?





## AllSeq The Sequencing Marketplace

### Our Mission

AllSeq has created the world's first true Sequencing Marketplace. Our NGS marketplace helps researchers pick the best provider for their needs (based on price, technology, turnaround time, etc). AllSeq also maintains the NGS Knowledge Bank, a neutral source of information on the various sequencing technologies, platforms and applications.

Whole Genome

#### Whole Genome

Whole genome sequencing (WGS), especially with human samples, is one of the most popular applications of next-generation sequencing. This application benefits primarily from generating as much sequence as possible with the smallest budget possible. Longer reads can definitely be beneficial, but the total genomic output (in terms of pure number of bases covered) is more important. Therefore, the flagship platforms from the more established providers tend to be used most heavily in this space.

Create WGS Project

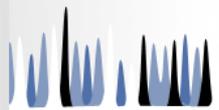
Exome

RNA-Seq

Other

# SEQanswers

<http://seqanswers.com/>



**SEQanswers**  
the next generation sequencing community

**SEQanswers Home** User Name   Remember Me?  
Password

[Register](#)   
 [FAQ](#)   
 [Community](#) ▼   
 [Calendar](#)   
 [Today's Posts](#)   
 [Search](#)

You are currently viewing the SEQanswers forums as a guest, which limits your access. [Click here to register now](#), and join the discussion

» **Site Navigation** ⌵

- » [About SEQanswers](#)
- » [Next Gen Summaries](#)
- » **Forums**
- » [Wiki](#)
- » [Instrument Map](#)

» **Log in** ⌵

User Name:

Password:

Remember Me?

Not a member yet?  
[Register Now!](#)

» **Online Users: 222** ⌵

11 members and 211 guests

[AndrewO](#), [boucceerb](#),  
[Chipper](#), [dsenalik](#), [Michael L. Altshuler](#), [netpumber](#),  
[nrd](#), [vasha](#), [wen vuan](#),  
[vzhang](#), [zirouc](#)

Most users ever online was  
1,120 - 06-26-2014 at 02:21

» **New Posts** ⌵

Title, Username, & Date	Last Post	Replies	Views	Forum
<a href="#">Short RNAseq results for metabolic pathways</a> netpumber	Today 02:22 PM by <a href="#">netpumber</a> ⌵	0	1	<a href="#">RNA Sequencing</a>
<a href="#">BLASTp parameter -dbsize problems (blastall -z)</a> fireLog2	Today 01:24 PM by <a href="#">dsenalik</a> ⌵	5	1,125	<a href="#">Bioinformatics</a>
<a href="#">Wood Frog Genome ver 0.1</a> cement_head	Today 01:10 PM by <a href="#">cement_head</a> ⌵	8	405	<a href="#">General</a>
<a href="#">NCBI taxid download</a> Pol8	Today 12:56 PM by <a href="#">GenoMax</a> ⌵	1	45	<a href="#">Bioinformatics</a>
<a href="#">Lotsa new toys from Illumina: HiSeq X Five, 3000, 4000, NextSeq 550</a> GW_OK	Today 12:56 PM by <a href="#">kcchan</a> ⌵	42	3,233	<a href="#">Illumina/Solexa</a>
<a href="#">DESeq2 Error</a> nw328	Today 12:52 PM by <a href="#">nw328</a> ⌵	0	35	<a href="#">Bioinformatics</a>

» **Lotsa new toys from Illumina: HiSeq X Five, 3000, 4000, NextSeq 550** ⌵

Jan 12, 2015 - 11:36 AM - by [GW\\_OK](#)

Hiseq 4000  
Hiseq 3000  
Nextseq 550  
HiseqX 5

<http://www.illumina.com/company/news...newsid=2006979>

42 Replies | 3,233 Views

» **Our Sponsors** ⌵

» **Recent Job Postings** ⌵

[Quality Engineer - Manufacturing](#)  
01-21-2015 04:50 PM  
by [Pacific Biosciences](#)

[Quality Engineer - R&D](#)  
01-21-2015 04:37 PM  
by [Pacific Biosciences](#)

[Scientist, Next-Generation Applications and...](#)  
01-15-2015 10:00 AM  
by [Pacific Biosciences](#)

[Senior Scientist, Applications](#)  
01-13-2015 05:18 PM  
by [Pacific Biosciences](#)

[Senior Bioinformatics Test Engineer](#)  
01-09-2015 04:45 PM  
by [Pacific Biosciences](#)

LATEST OPEN RNA-SEQ CHIP-SEQ SNP ASSEMBLY TUTORIALS TOOLS JOBS FORUM PLANET ALL »

 Welcome to Biostar! [about](#) • [faq](#) • [rss](#)

[Community](#) [User Login](#) [New Post](#)

Live search: start typing... or

Limit to: all time <prev • 21,589 results • page 1 of 540 • next > Sort by: update ▾

**0** votes **0** answers **3** views **abyss-fac: command not found on Kubuntu 14.04**  
next-gen abyss sequencing written 2 minutes ago by dario.romagnoli • 50

**0** votes **0** answers **23** views **BAM file interpretation and consensus creation**  
bam samtools written 18 minutes ago by User000 • 180

**1** vote **1** answer **114** views **Question: Optimize the parameter span for loess (or spar for smooth.spline)**  
R smooth.spline loess written 3 days ago by wuxian2010 • 0 • updated 13 minutes ago by russ\_hyde • 560

**5** votes **1** answer **155** views **News: Integrated Genome Browser version 8.3 released - new native installer**  
news chip-seq igb rna-seq igv written 4 days ago by Ann • 1.2k • updated 15 minutes ago by Antonio R, Franco • 40

**0** votes **1** answer **40** views **grab signal from wig to bed file**  
chip-seq written 56 minutes ago by elaakoaserge • 0 • updated 30 minutes ago by lan • 3.7k

**Recent Votes**

- [Is There A Free Alternative To Ingenuity Pathway Analysis?](#)
- [C: Meaning Of Colors Used In Samtools Tview](#)
- [C: Question: Optimize the parameter span for loess \(or spar for smooth.spline\)](#)
- [Comparative Microbial Genomics](#)
- [C: Perl program: The sequence does not appear to be FASTA format \(lacks a descripto](#)
- [A: Jellyfish v2.1.4 make error](#)
- [A: Jellyfish v2.1.4 make error](#)

**Recent Locations • All »**

- [Germany](#), just now
- [Finland](#), 1 minute ago
- [Vilnius](#), 3 minutes ago
- [European Union](#), 3 minutes ago
- [Germany](#), 4 minutes ago
- [Italy](#), 4 minutes ago



## Search engine for biological data analysis

Search among 16,563 omic tools



SIGN UP / SIGN IN ABOUT

### Browse by OMIC TECHNOLOGIES



#### High-throughput sequencing

[WGS analysis](#), [WES analysis](#), [De novo sequencing analysis](#), [RNA-seq analysis](#), [ChIP-seq analysis](#), [BS-seq analysis](#), [Metagenomic sequencing analysis](#)



#### Mass spectrometry

[MS-based untargeted proteomics](#), [MS-based targeted proteomics](#), [MS-based untargeted metabolomics](#)



#### PCR

[qPCR](#), [dPCR](#), [Single-cell qPCR](#)



#### Bioimaging

[Small-angle scattering](#), [Super-resolution imaging](#), [Mass spectrometry imaging](#)



#### Microarray

[aCGH and SNP array analysis](#), [Gene expression array analysis](#), [DNA methylation array analysis](#)



#### NMR spectroscopy

[NMR-based proteomics](#), [NMR-based metabolomics](#)



#### Flow cytometry & mass cytometry

[Flow cytometry](#), [Mass cytometry](#)



#### Other omic technologies

[Sanger sequencing](#), [DNA fingerprinting](#), [nCounter System](#)

Thank you for your attention 😊



**KEEP  
CALM  
AND  
GOOD  
LUCK**

# How to get to Club Akumulatory?

**1** Go to Kaponiera  
(jakdojade.pl might be useful)

**2** Locate Jowita dormitory  
(The building with a huge "**AKUMULATORY**" sign)



**3** Pass it by and you will find  
**AKUMULATORY CLUB**  
(the exact address is Zwierzyniecka 7, 60-813 Poznań)

