Monday – part 2

Introduction to NGS data analysis

Joanna Ciomborowska-Basheer, PhD

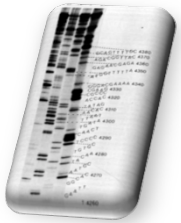
Faculty of Biology , Adam Mickiewicz University, Poznań

ideas4biology Ltd.

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

5′--TGGAATTGTGAGCGGATAACAATT3′
3′--ACCTTAACACTCGCCTATTGTTAA5′

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

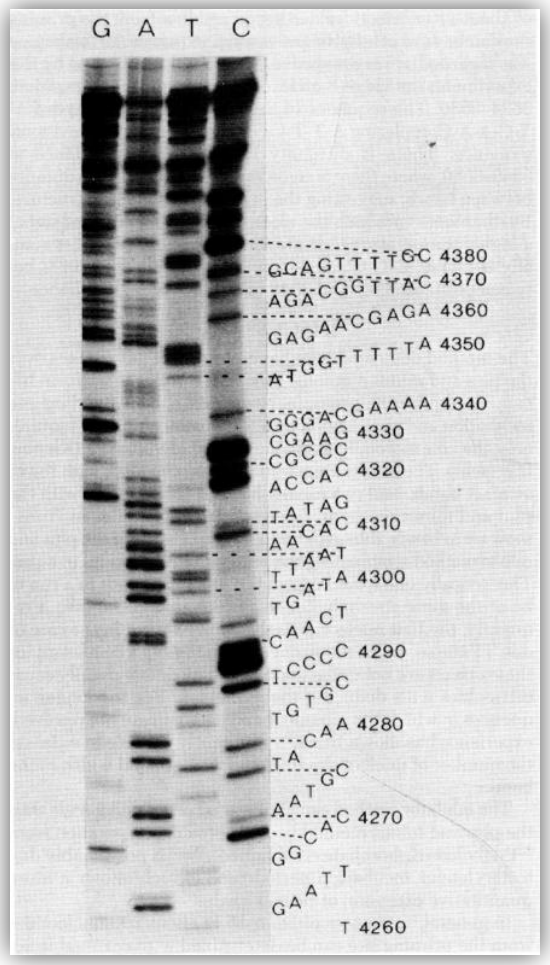
2009 : Helicos – single cell sequencing

2011 : Ion Torrent : PGM

2011 : Pacific Biosciences – PacBio RS

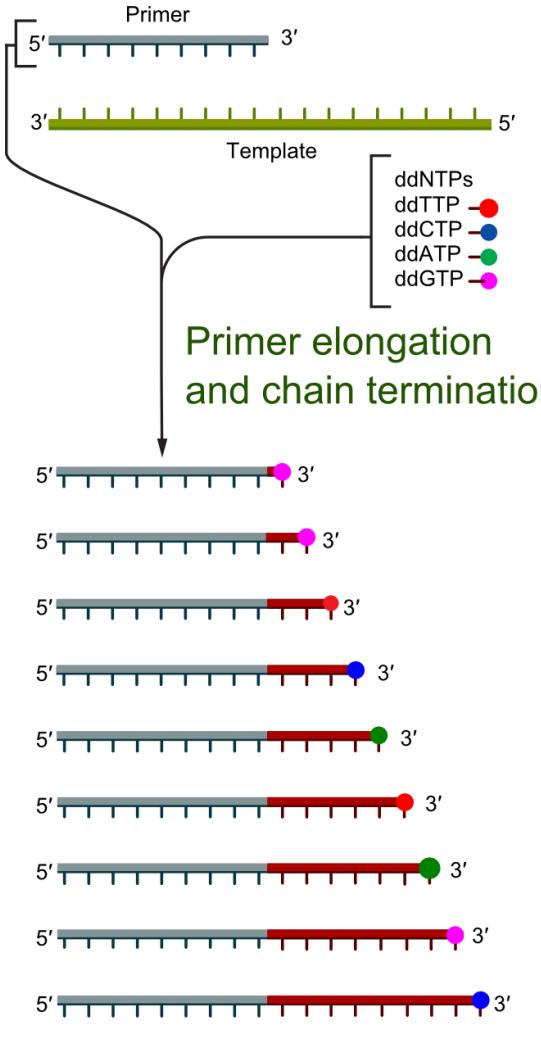
2012 : Oxford Nanopore Technologies

Sanger sequencing method



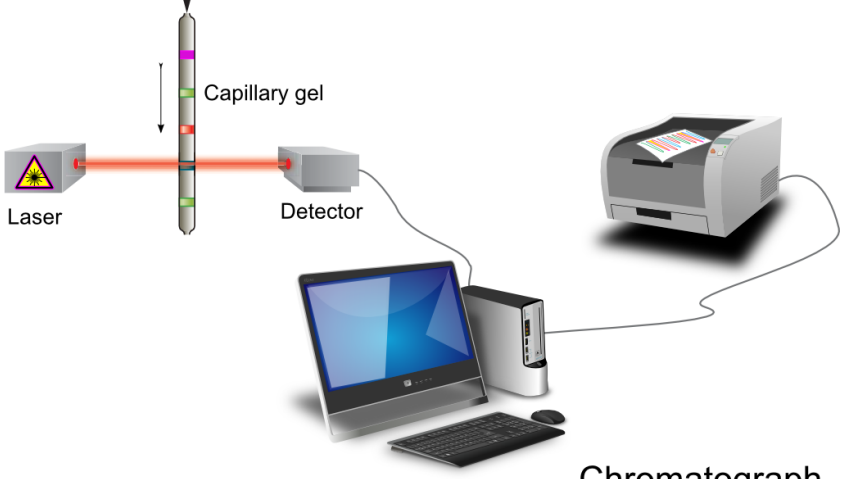
Sanger et al. 1977

First generation sequencing

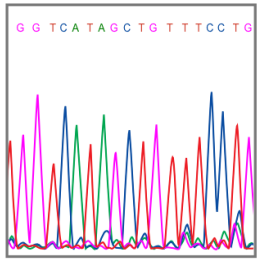


Primer elongation and chain termination

Capillary gel electrophoresis separation of DNA fragments



Chromatograph

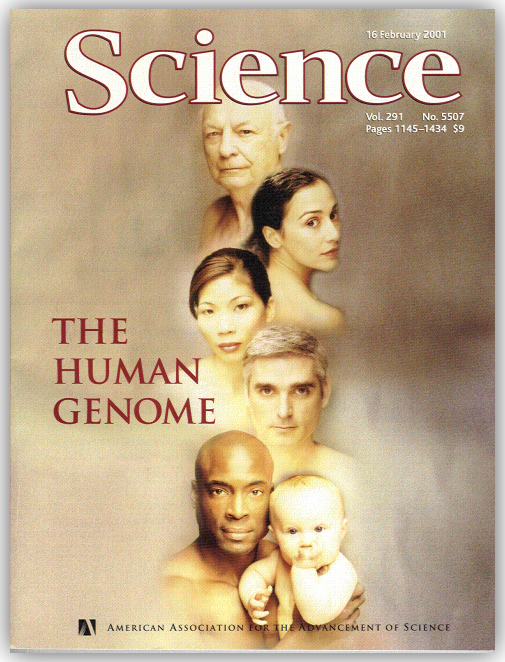


Laser detection of flourochromes and computational sequence analysis

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.sciencemag.org/

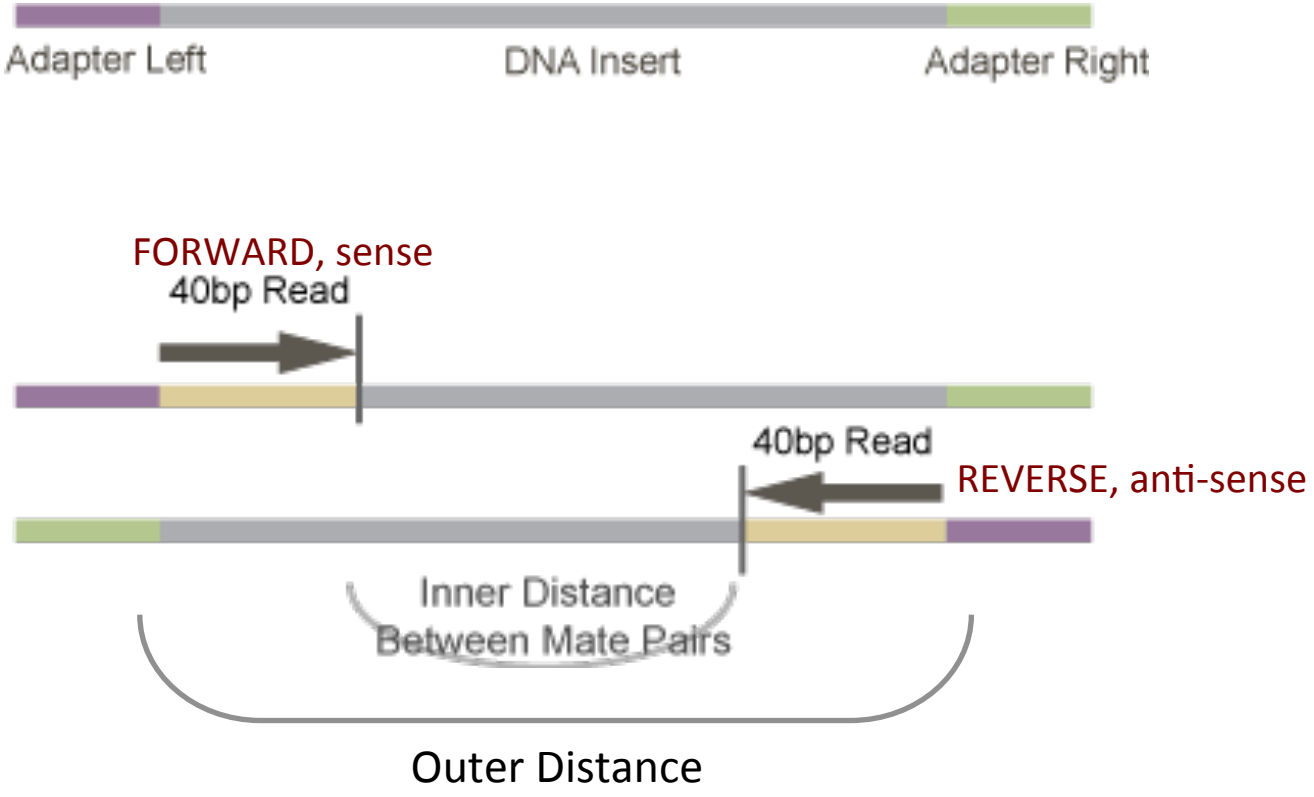
http://www.numismondo.net/

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**

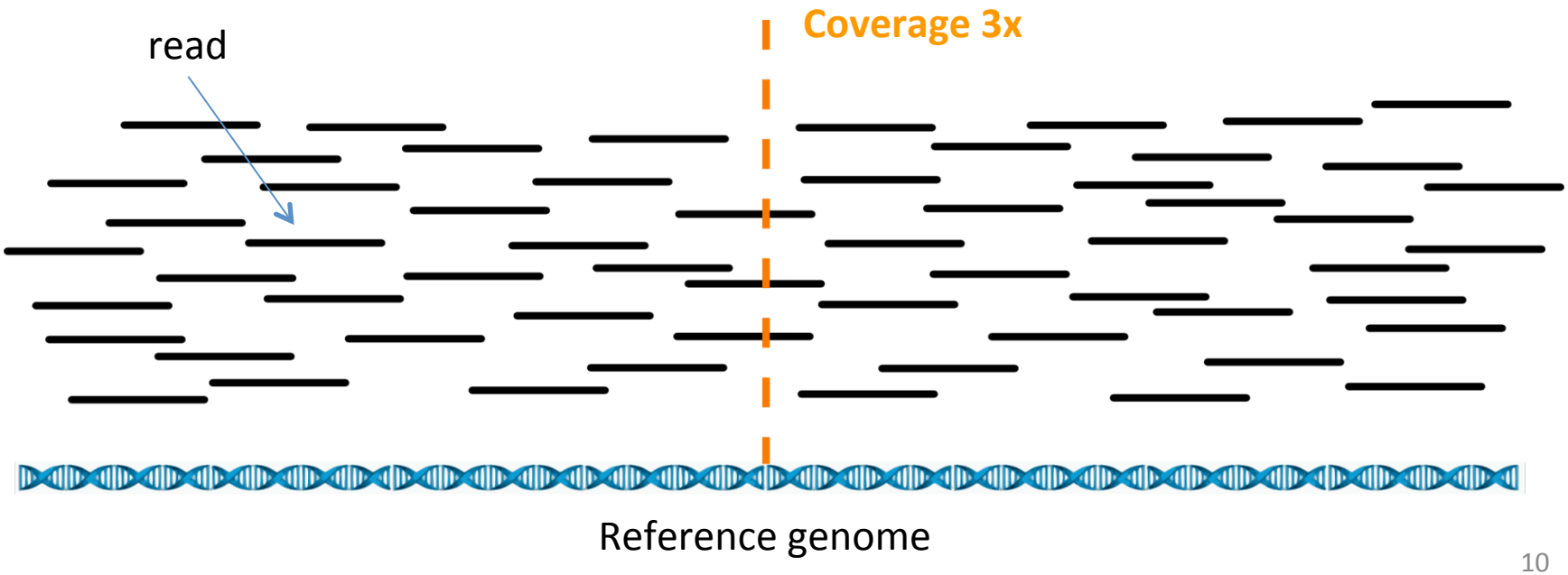
Next generation sequencing
GLOSSARY

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

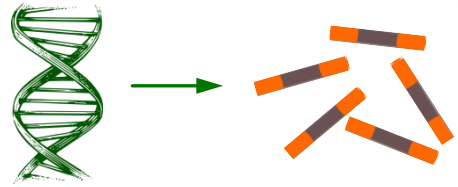
- **Coverage level** – The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was sequenced 30 times.



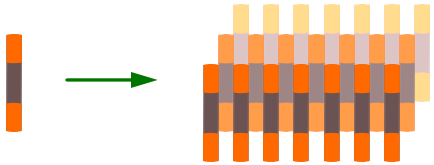
Next generation sequencing

Most of NGS protocols consist of:

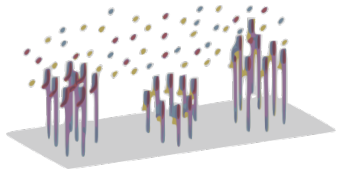
1. • Library construction



2. • PCR amplification



3. • Massive parallel sequencing



NGS platforms and technologies

SOLiD®

illumina®

PACIFIC BIOSCIENCES™

ion torrent

ThermoFisher SCIENTIFIC

Roche 454 SEQUENCING

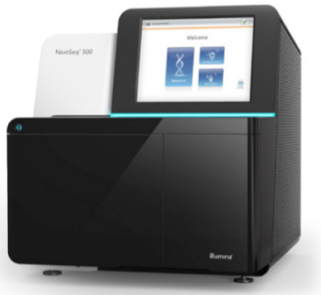
Oxford NANOPORE Technologies®

illumina®

HiSeq X Ten

HiSeq 2500

NextSeq 500



MiSeq

illumina®

	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

http://allseq.com/ http://www.illumina.com/

Roche

454 SEQUENCING

Genome Sequencer FLX+

Genome Sequencer Junior

Roche

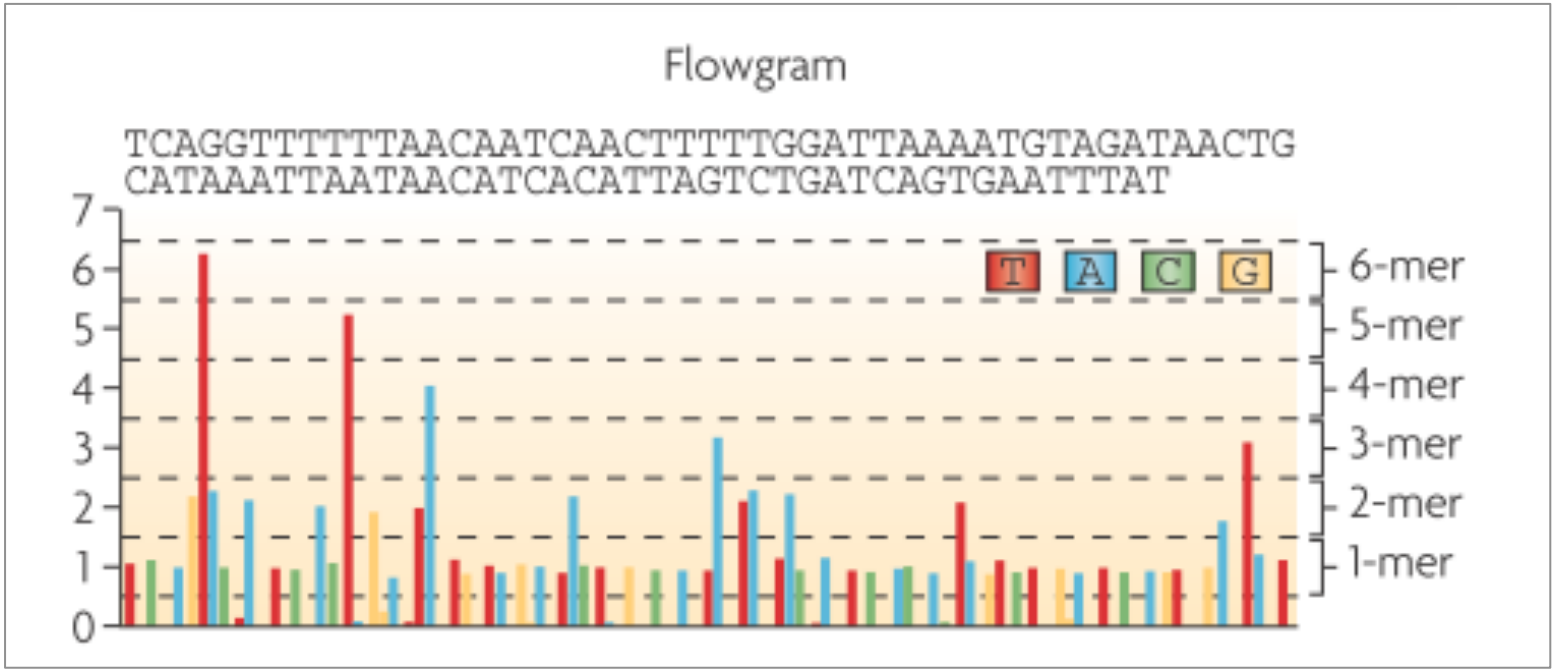
454 SEQUENCING

	GS FLX+	GS Jr.
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

http://454.com/ http://allseq.com/

Roche 454 SEQUENCING

Higher error rate because of HOMOPOLYMERS !!!



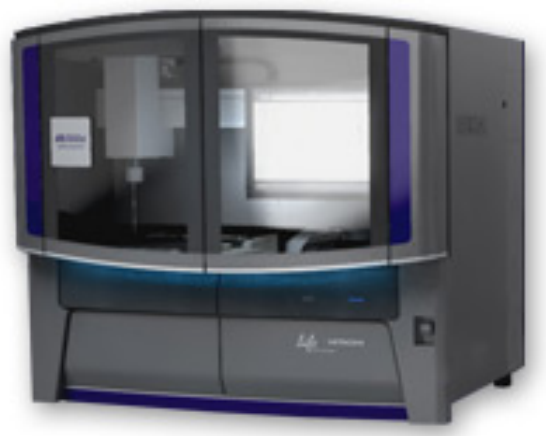
Metzker 2010

SOLiD®

by life technologies™
ThermoFisher SCIENTIFIC

SOLiD 5500

SOLiD 5500xl



http://allseq.com/

SOLiD®

by life technologies™

ThermoFisher SCIENTIFIC

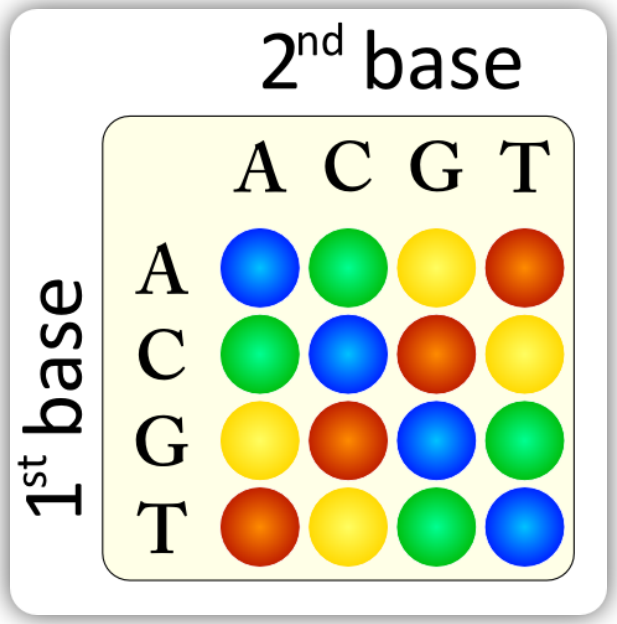
	SOLiD 5500xl	SOLiD 5500xl Wildfire	SOLiD 5500	SOLiD 5500 Wildfire
Total output/run	95 Gb	240 Gb	48 Gb	120 Gb
Run Time	6 days	10 days	6 days	10 days
Output/day	16 Gb	24 Gb	8 Gb	12 Gb
Read length	2 X 60	2 X 50	2 X 60	2 X 50
# of single reads	800M	2.4B	400M	1.2B
Instrument Price	$595k	$70k upgrade	$349kk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

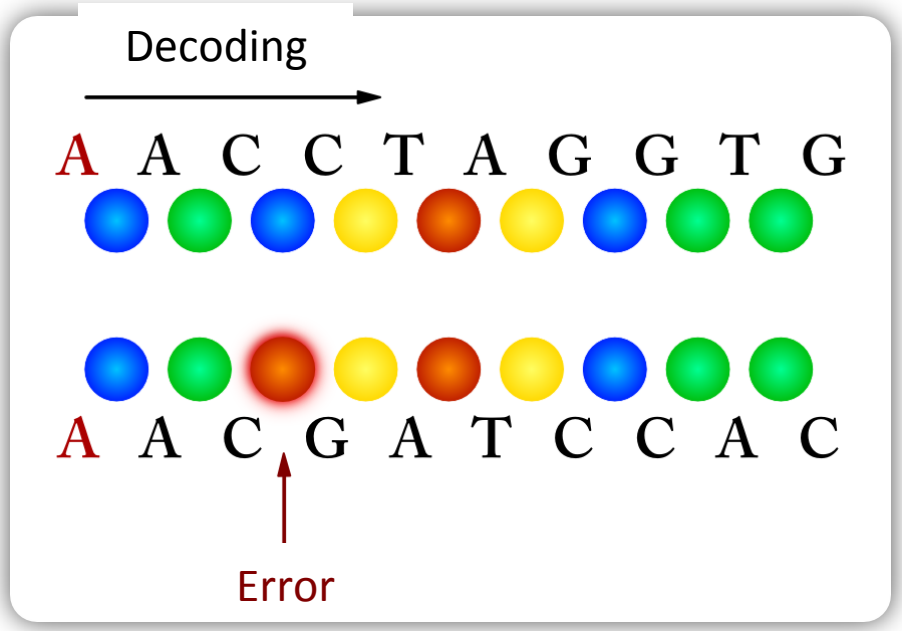
http://allseq.com/

SOLiD®

by life technologies™
ThermoFisher SCIENTIFIC

Color space





ion torrent

by life technologies™

ThermoFisher SCIENTIFIC

Ion PGM
(**P**ersonal **G**enome **M**achine)

Ion Proton

http://allseq.com/ http://www.invitrogen.com

ion torrent

by life technologies™

ThermoFisher SCIENTIFIC

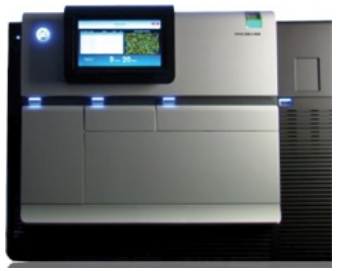
	PGM 314	PGM 316	PGM 318	PI	PII (est. early 2015)
# of sensors	1.2M	6.1M	11M	165M	660M
Total output	up to 100Mb	up to 1Gb	up to 2Gb	~10Gb	~32Gb(at launch)
Run Time	2-4h rs	3-5 hrs	4-7 hrs	2-4 hrs	2-4 hrs
Output/day*	up to 200Mb	up to 2Gb	up to 4Gb	~20Gb	~64Gb
Avg read length	up to 400b	up to 400b	up to 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™



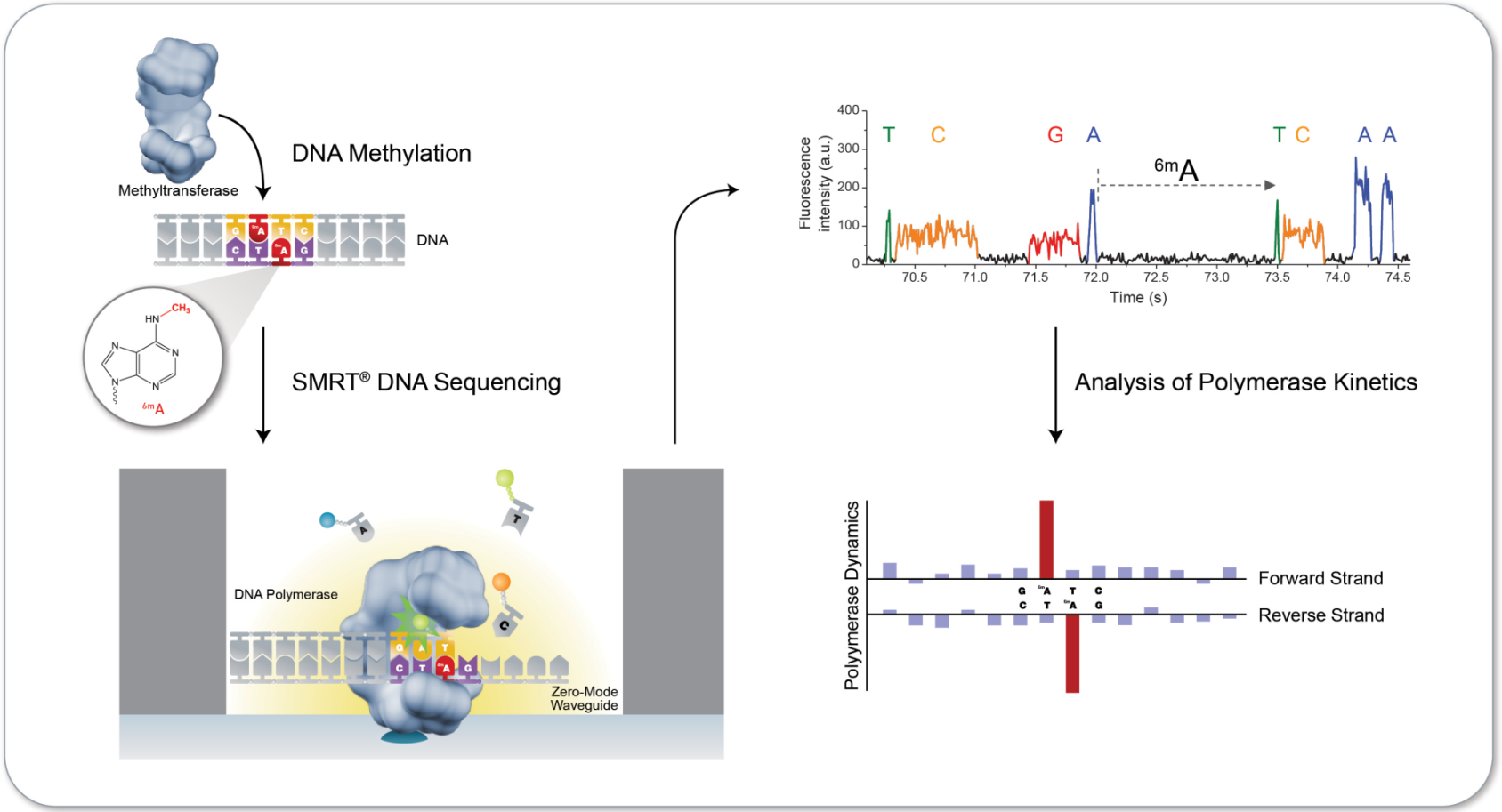
Single molecule real-time sequencing !

	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/

PACIFIC BIOSCIENCES™

DNA methylation

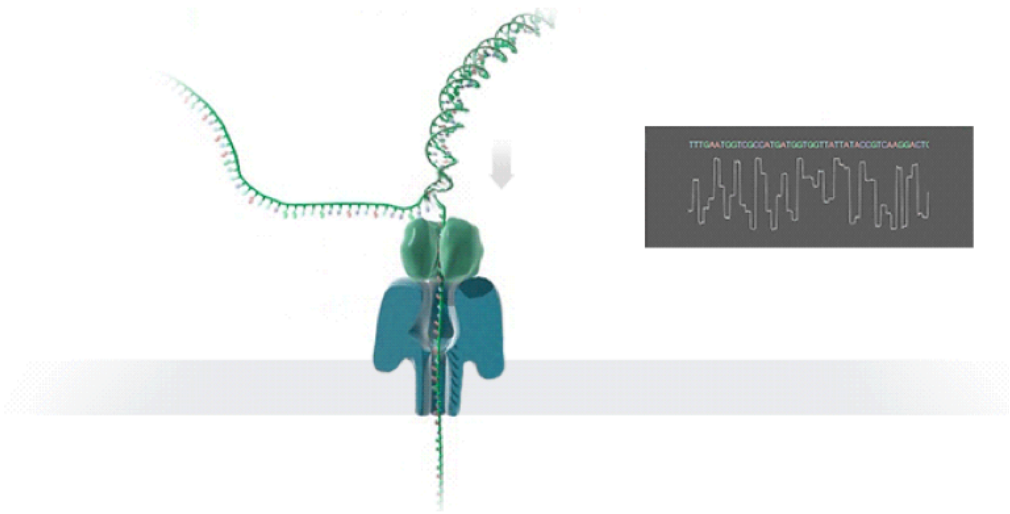


http://www.pacificbiosciences.com/applications/base_modification/

Single molecule real-time sequencing !

Oxford NANOPORE Technologies®

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 Mkl Flow cell

The MinION Mkl 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 Mkl device.

https://nanoporetech.com/science-technology/how-it-works

Next generation sequencing
Applications

Genome sequencing

Transcriptome sequencing (RNA-Seq)

Protein – RNA interactions (CLIP-seq)

Protein – DNA interactions (ChIP-seq)

DNA methylation

Metagenomics

Next generation sequencing projects

http://www.1000genomes.org/

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

http://cancergenome.nih.gov/

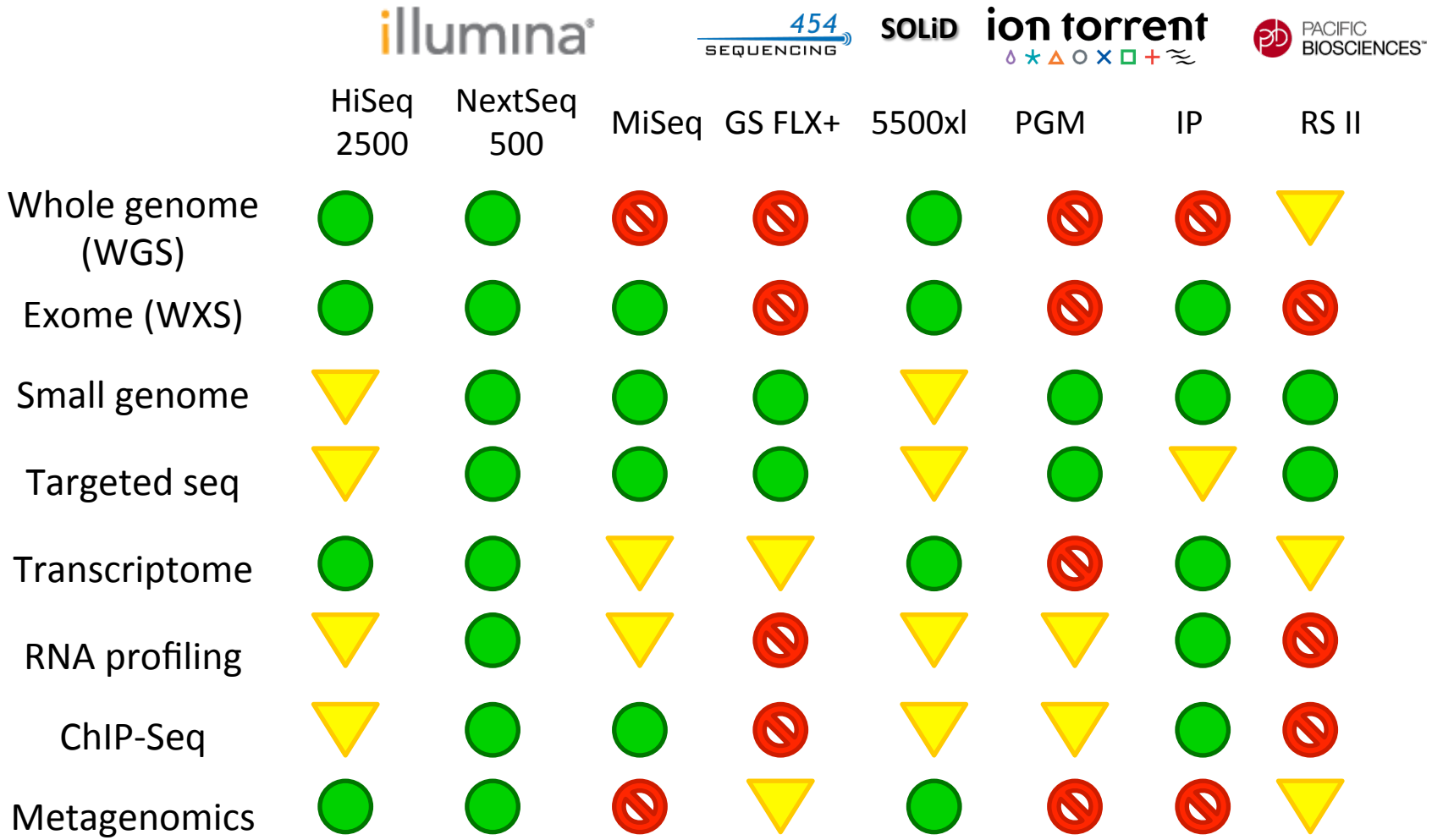
RNA Seq Atlas

http://medicalgenomics.org/rna_seq_atlas

Genomics england

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

Examples of applications



Let's start to analyse…
What data usually do we need?

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

Genome annotation
GLOSSARY

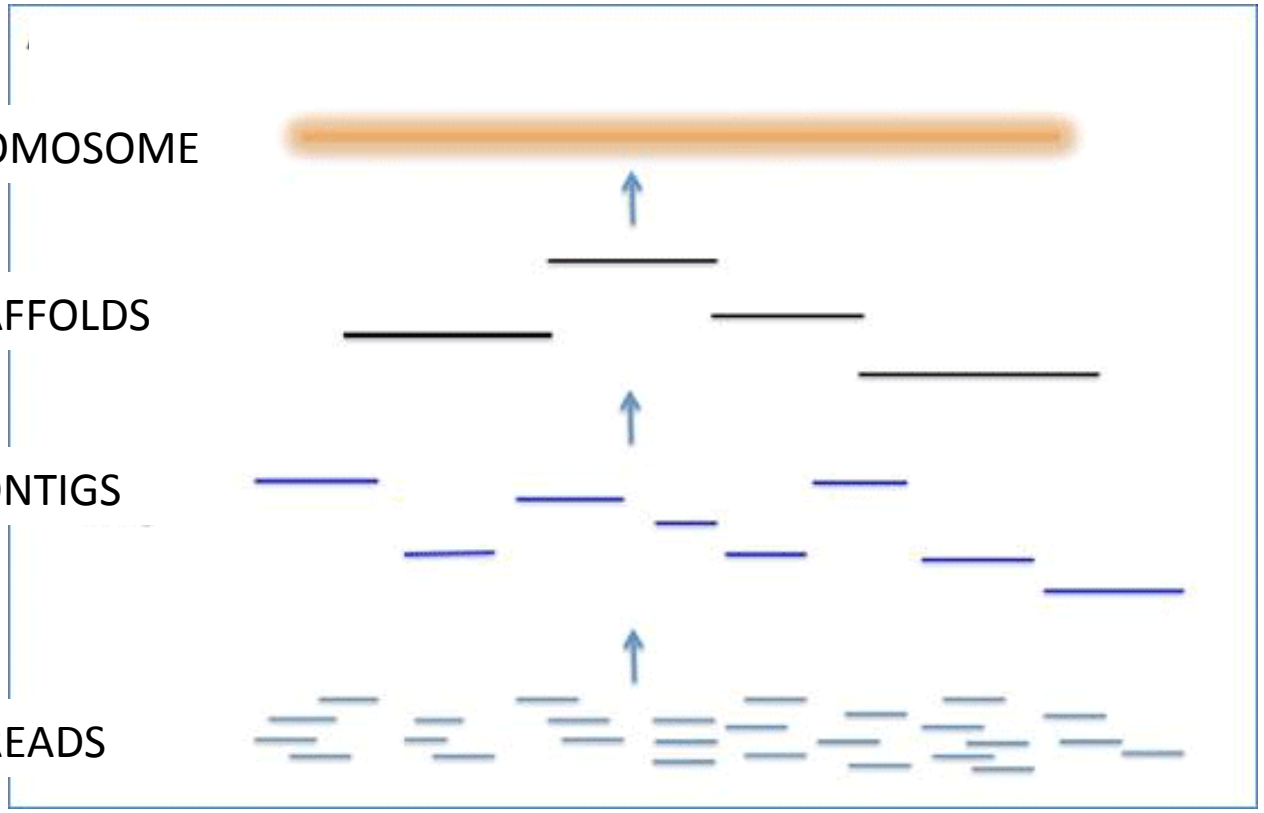
- **Annotation as a PROCESS** in which structural and functional elements of genomes are identified and being described.
- **Annotation as a DATASET** which contains and describes all structural and functional elements of sequences (i.e. strand, chromosome, coordintes). Basic sources of genome annotations 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 asigned 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
GGATATAGAAATGGGAATGGGTAACTCAAAGTCCAGGCAGATAATGAAAAGACTTATAAAGCGGTATGTTTTGAAAGCACAAGTAGAC
AAAGAAAATGATGAAGTTAATGAAGGTGAATTAAAAGAAATCAAGCAAGATATCTCCAGCCTTCGTTATGAACTTTTGGAAGACAAGAG
CCAAGCAACTGAGGAATTAGCCATTCTAATTCATAAACTTAGTGAGAAACTGAATCCCAG
```

```
>XP_007907608.1 PREDICTED: keratin, type I cytoskeletal 19-like [Callorhinchus milii]
MSRSVYSANIGGSVIVSSNQRRSFASTSSSLFGSGAPSRRAISVYNIGANRGKRISSAGGSWNASYASLG
GDAGILCNDEKQTMQNLNARLSSYMEKVRSLEKSNRQLEFQIHEFYEKKAPVSTKDLTVYEGNISDCHLQ
IYAISLHNAKLMLQIDNARLAADDFRIKYESELAIRKGVEADIQGLRKVMDELSLTKRGLESQVTALKED
LVYLHRSHKEELLSLRTGMGGSVTVDLDSTPATDLNKILSNLRIEYETIAEKNRKDVEAWYLEKCHTLNQ
```

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

http://davinci.mimuw.edu.pl/biologia/enzymy-restrykcyjne

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
```

https://genome.ucsc.edu/FAQ/FAQformat#format1

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 anotated elements in the genome. For each element there is a separated line in the file consits 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 "
```

https://genome.ucsc.edu/FAQ/FAQformat#format1

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.

ftp://ftp.ensemblgenomes.org/pub/release-29/bacteria//gtf/bacteria_20_collection/escherichia_coli_pa4/README

GFF3 - General Feature Format

1	2	3	4	5	6	7	8	9
1	ensembl_havana	lincRNA_gene	89295	133723	.	-	.	ID=gene:ENSG00000238009;Name=RP11-34P13.7;biotype=lincRNA;gene_id=ENSG00000238009;havana_gene=OTT
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 – accesion numbers

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

NM_..., XM_... → mRNA
NP_..., XP_... → protein
NR_..., XR_... → noncoding RNA

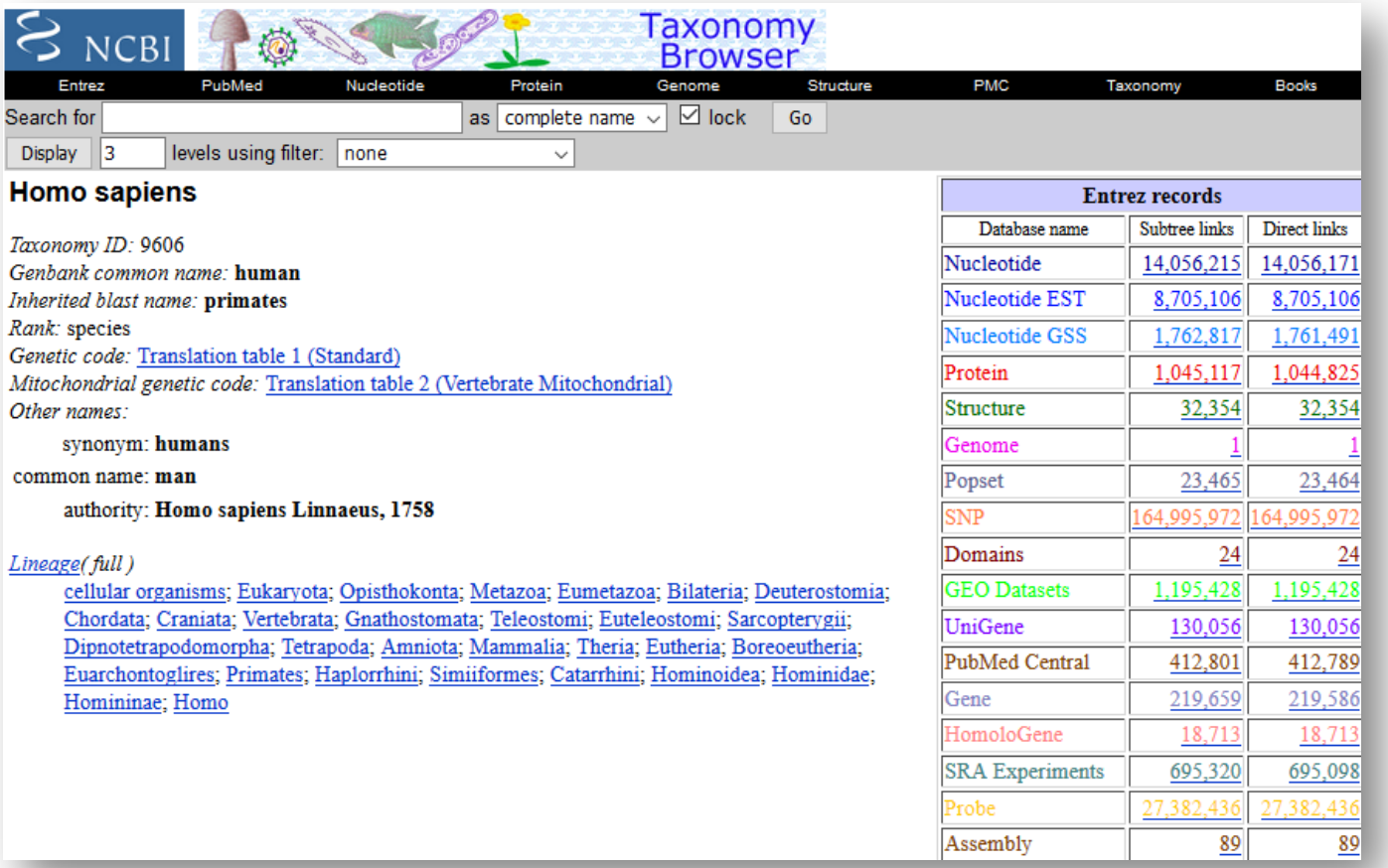
NC_..., NG_..., NT…., NW…, AC... → contigs, genomic sequences

N - sequences from experiments

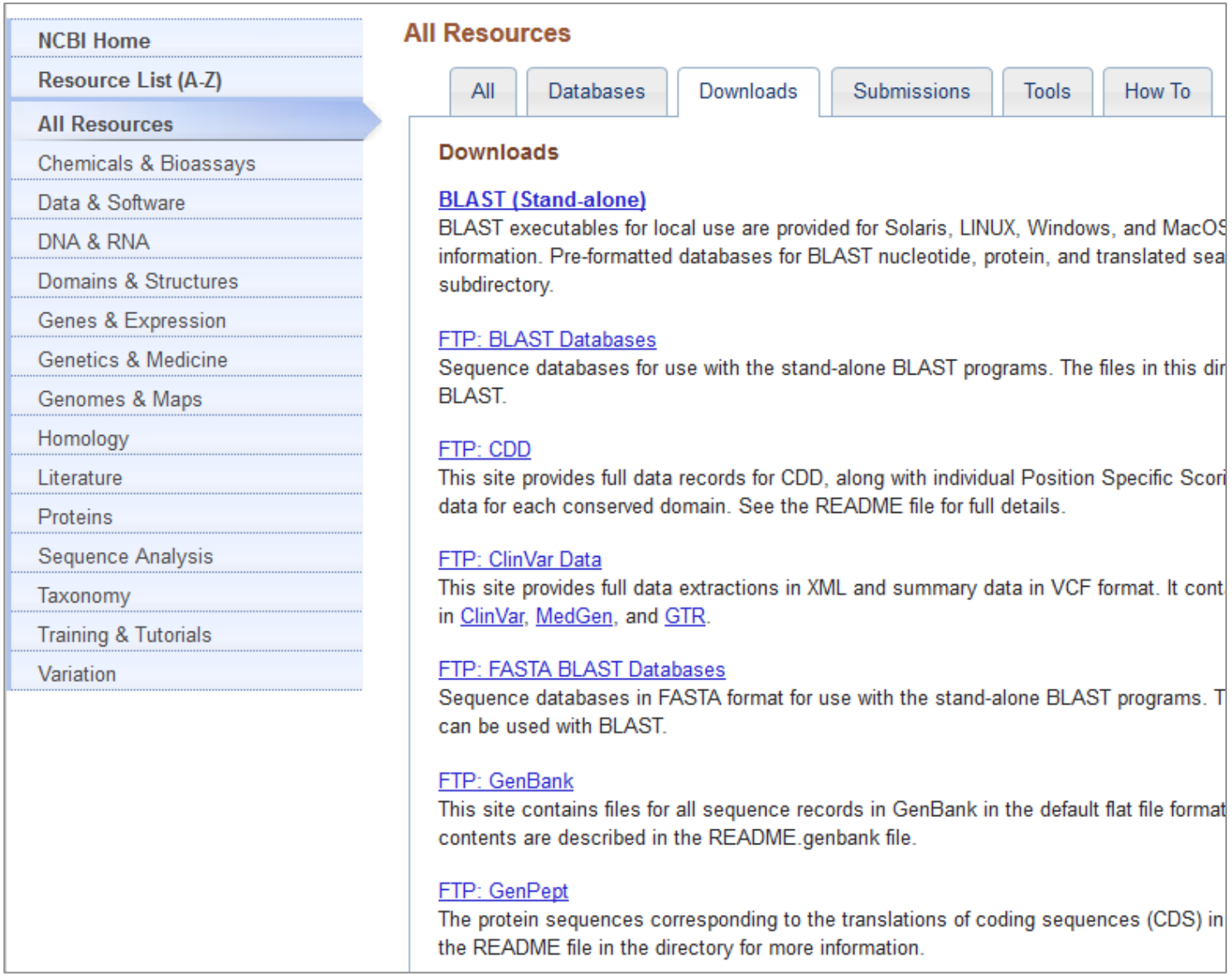
X – sequences based on comparison with existing sequences or prediction

NCBI (Taxonomy)

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



NCBI (Downloads FTP)



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/

Ensembl – accession numbers

Show/hide columns (1 hidden) Filter

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

ENS**G** → **Gene**
ENS**T** → **Transcript (.ver)**
ENS**P** → **Protein (.ver)**
ENS**E** → **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

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 chromsome 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)

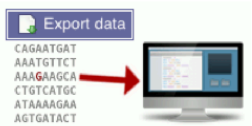
Downloads

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species...

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

Ensembl release 73 - September 2013 © WTSI / EBI

About Ensembl | Privacy Policy | Contact Us

Multi-species data

Database						
Comparative genomics	MySQL	EMF	MAF	BED	XML	Ancestral Alleles
BioMart	MySQL	-	-	-	-	-
Stable ids	MySQL	-	-	-	-	-

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/BigW
Y	Human *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BigW
Y	Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BigW

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 ?*

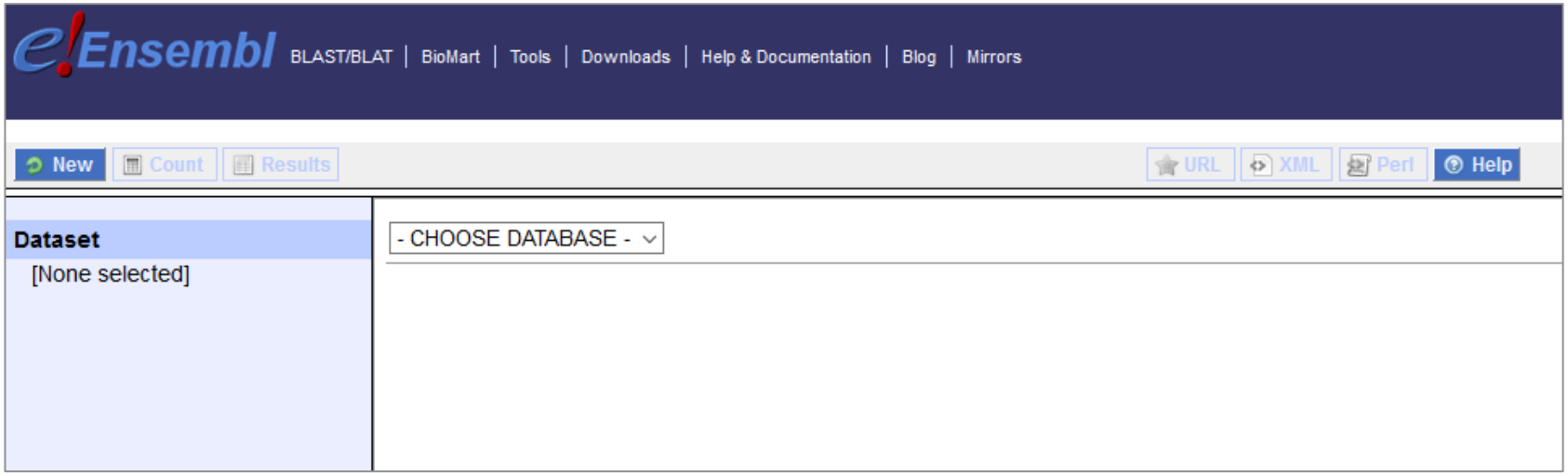
2. Please find noncoding RNAs (**ncRNA**) for **mouse (*Mus musculus*)**

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

Ensembl – download

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

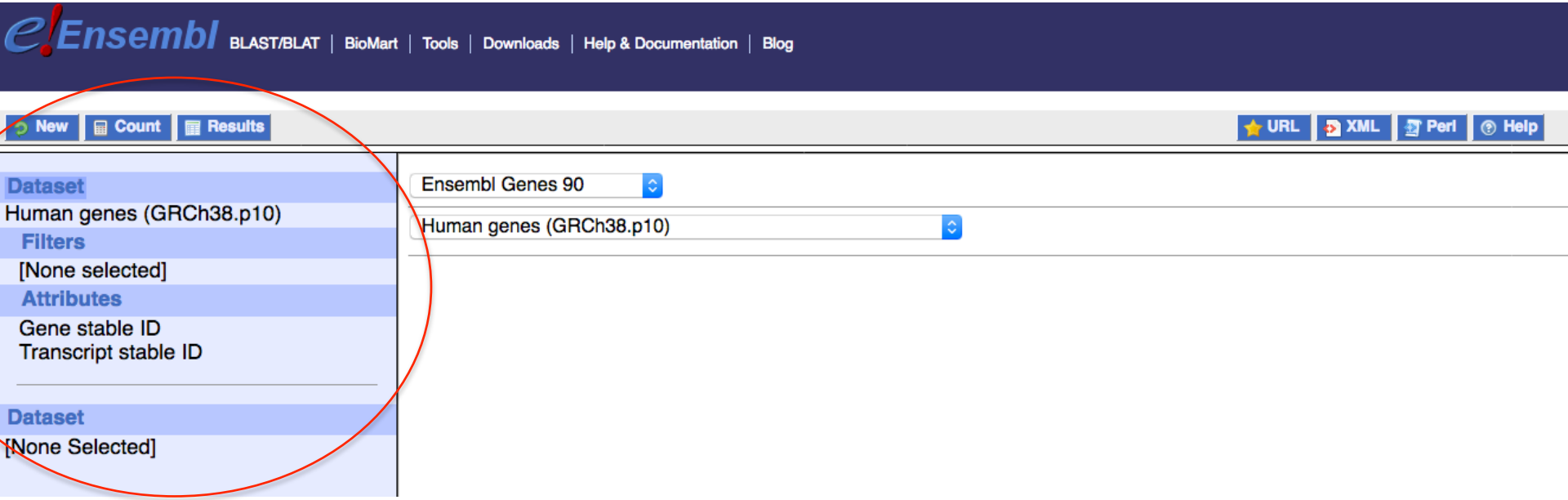
Ensembl BioMart



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

Ensembl – download

BioMart



- **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

BioMart

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.

New | Count | Results | URL | XML | Perl | Help

Dataset 22214 / 53110 Genes
Mouse genes (GRCm38.p5)
Filters
Gene type: protein_coding
Attributes
Gene stable ID
Transcript stable ID
Gene name
Gene description

Dataset
[None Selected]

Export all results to File TSV Unique results only Go
Email notification to
View 10 rows as HTML Unique results only

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

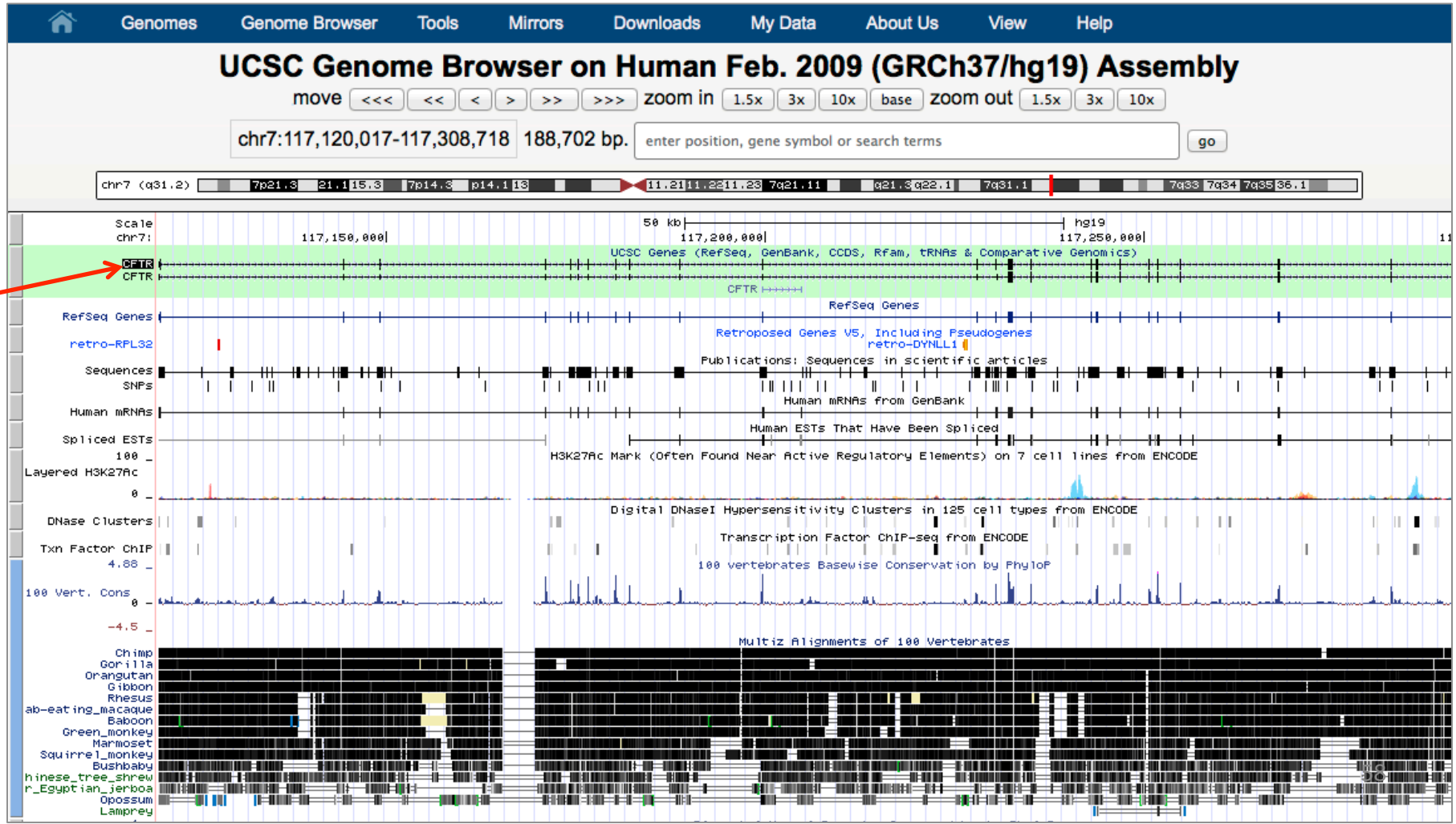
Ensembl Genome Browser
European Bioinformatics Institute, Wellcome Trust Sanger Institute

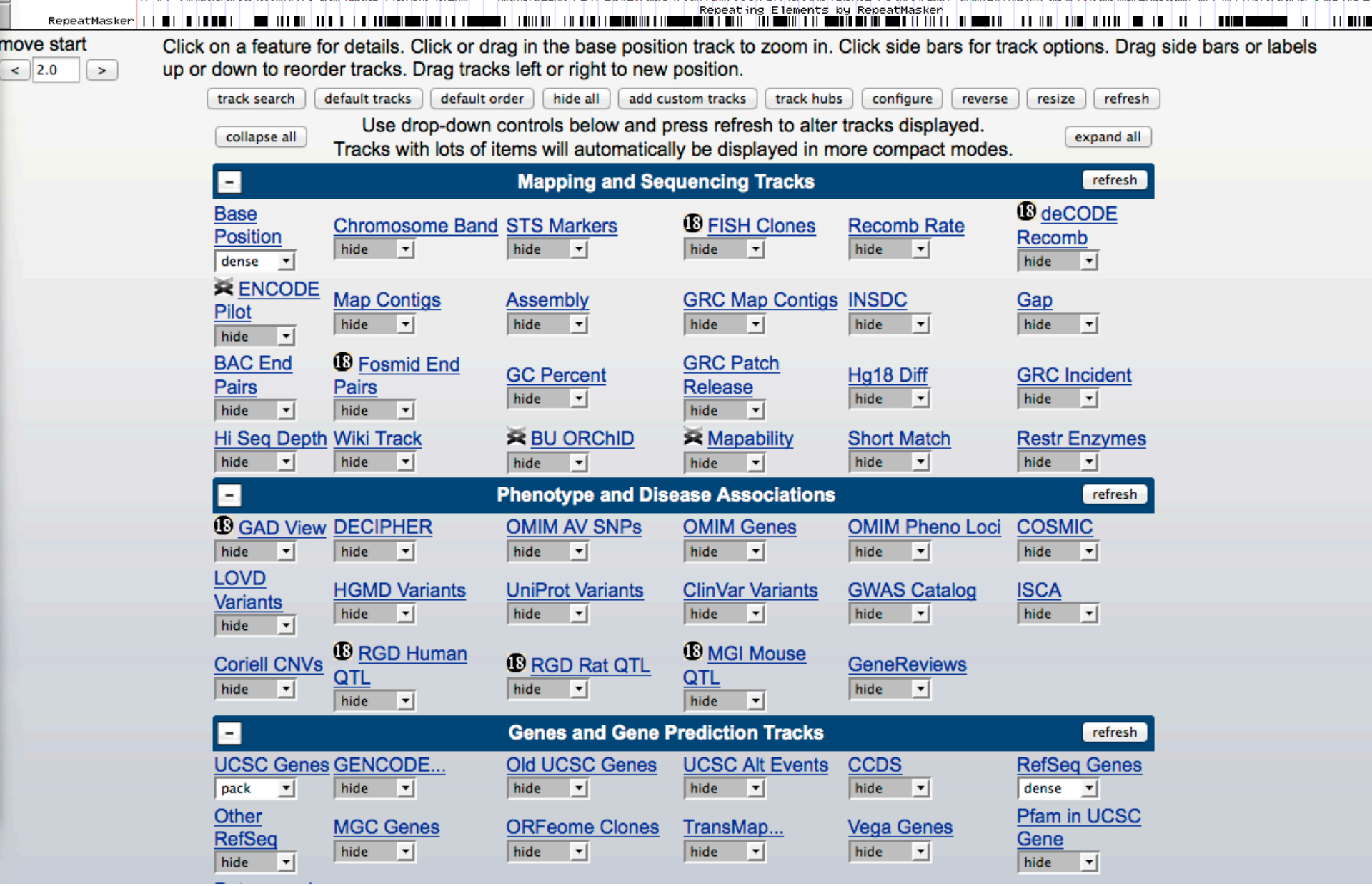
University of California, Santa Cruz Genome Browser

UCSC Genes

UCSC Transcript ID but often ID is from different database

CFTR (uc011kng.2) at chr7:117120017-117308718 - Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family
CFTR (uc003vje.1) at chr7:117199518-117204842 - Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family
CFTR (uc003vjd.3) at chr7:117120017-117308718 - Homo sapiens cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family
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 (uc021tdq.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





Possibilities of changes in tracks → you can easily castomise 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 doeas 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 allels from OMIM database you can find here? (OMIM allelic variants) → disease associations

UCSC - download

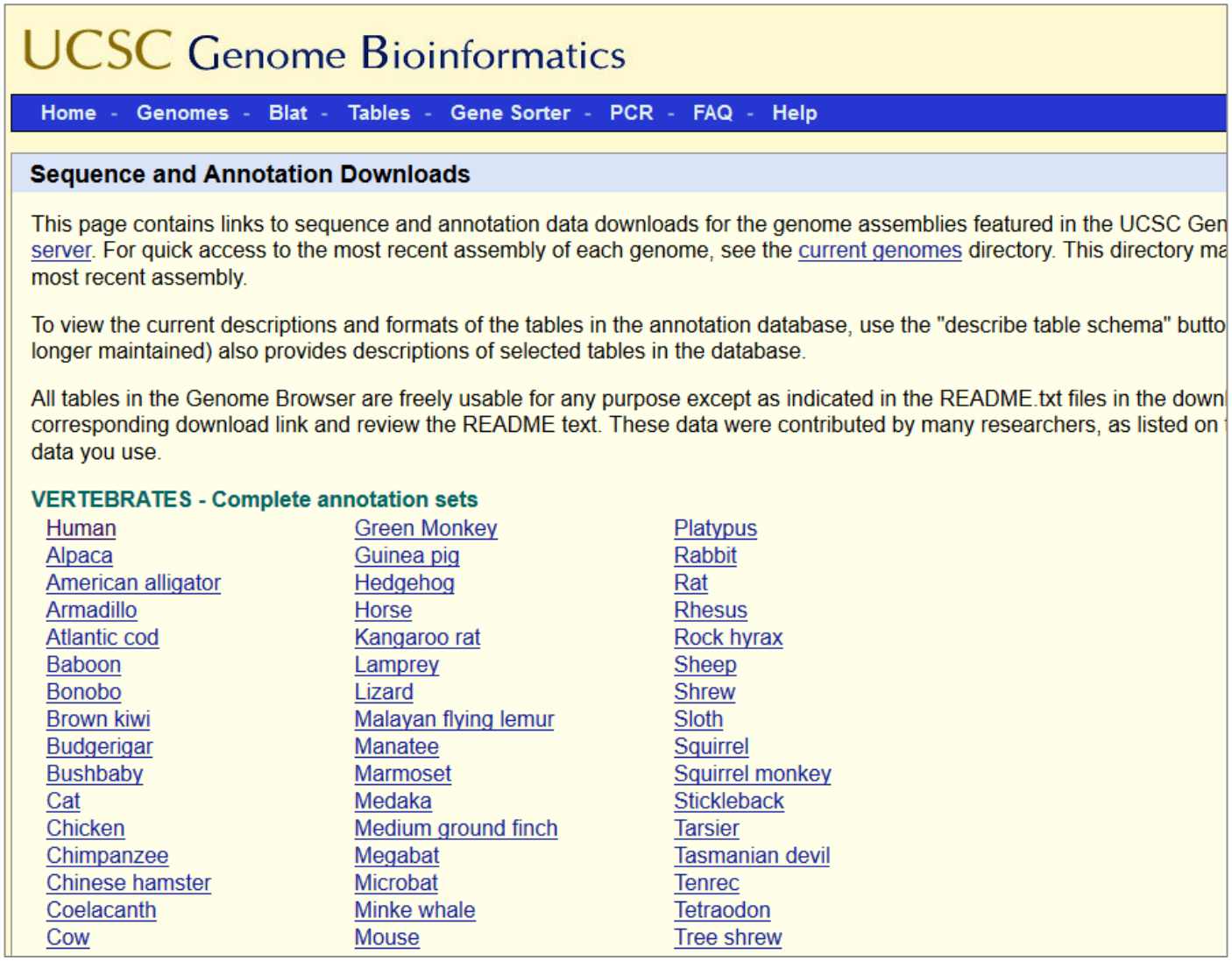
1. For single records and regions from the genome browser (View → DNA)

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 Gen
server. For quick access to the most recent assembly of each genome, see the current genomes directory. This directory ma
most recent assembly.

To view the current descriptions and formats of the tables in the annotation database, use the "describe table schema" butto
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 down
corresponding download link and review the README text. These data were contributed by many researchers, as listed on
data you use.

VERTEBRATES - Complete annotation sets

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

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.

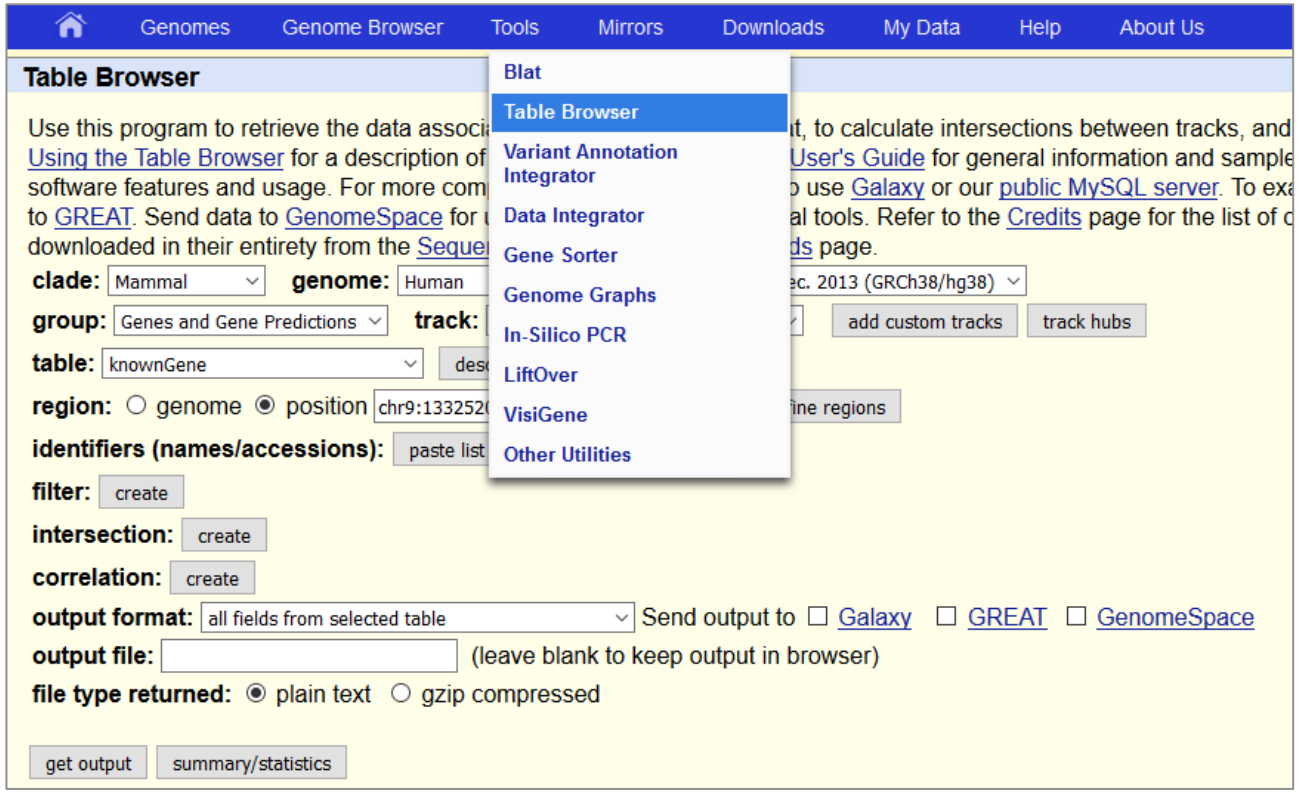
What we can find using option **[Full dataset] ?**

UCSC – download

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

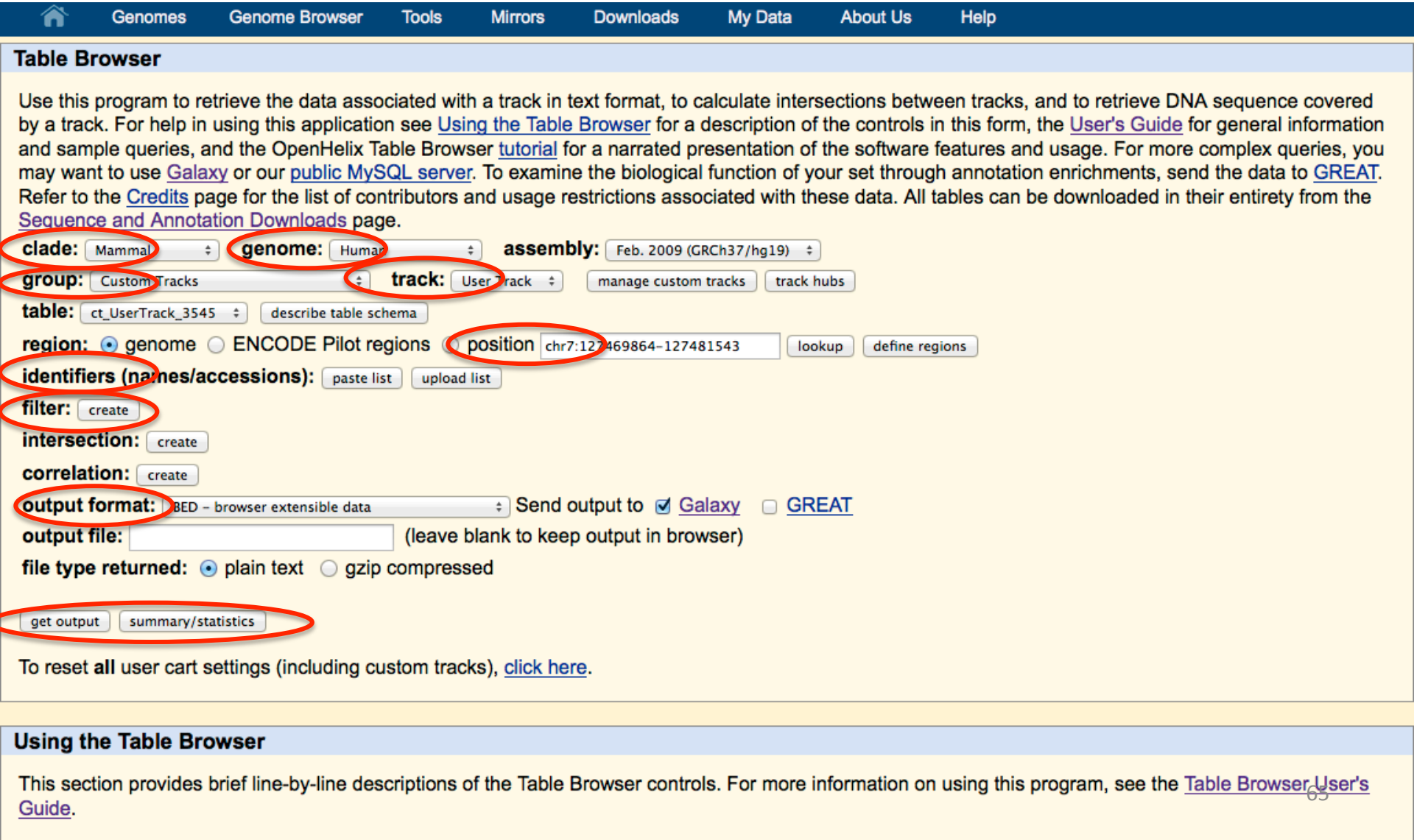
UCSC Table Browser

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



UCSC – download

Table Browser



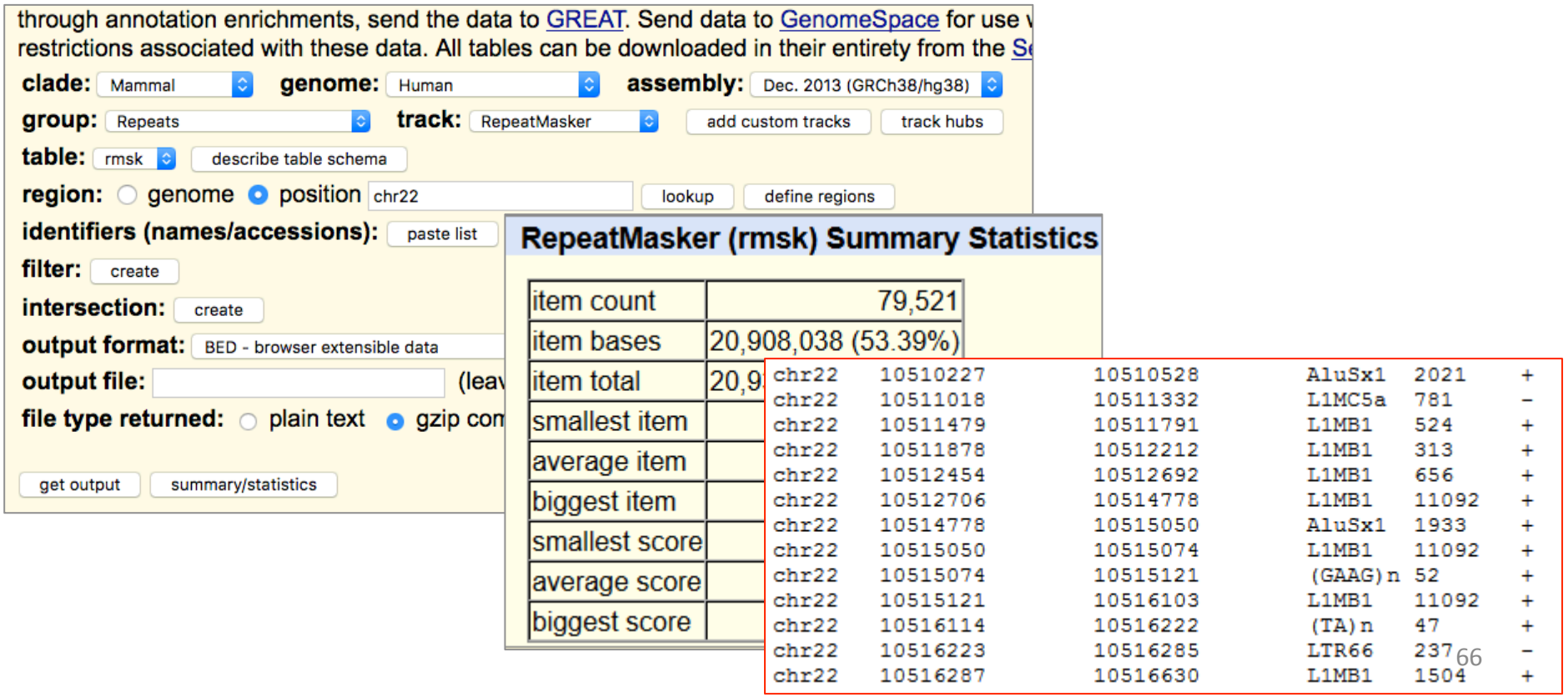
UCSC – download

Table Browser

TASK

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

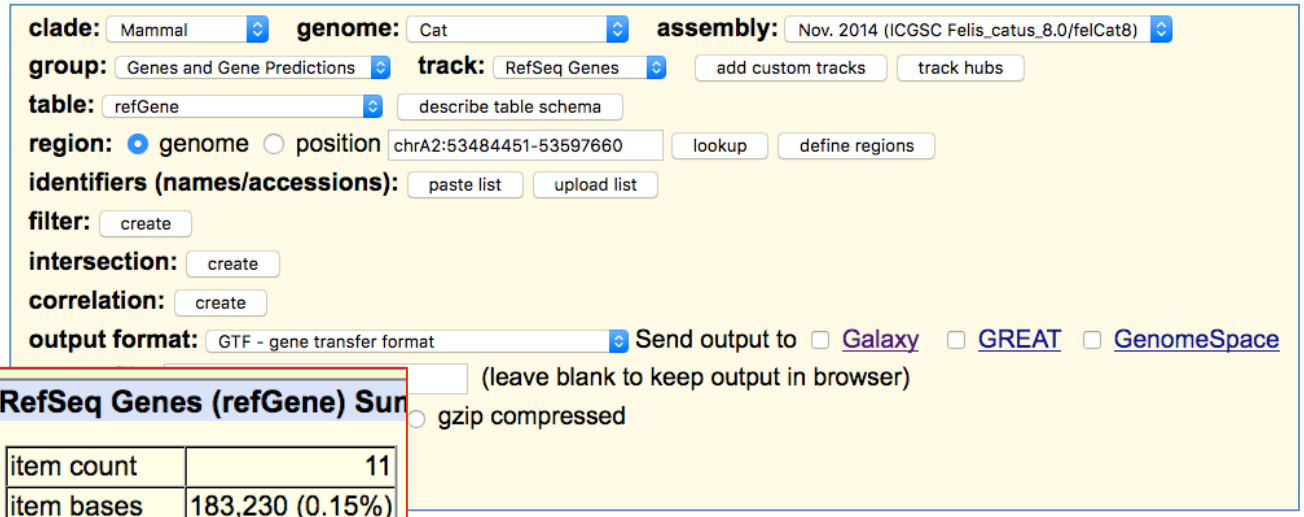
(ATTENTION! Recent genome version)

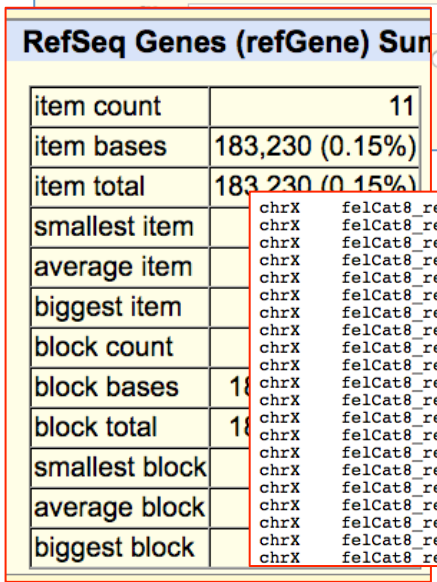


UCSC – download
Table Browser

TASK

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





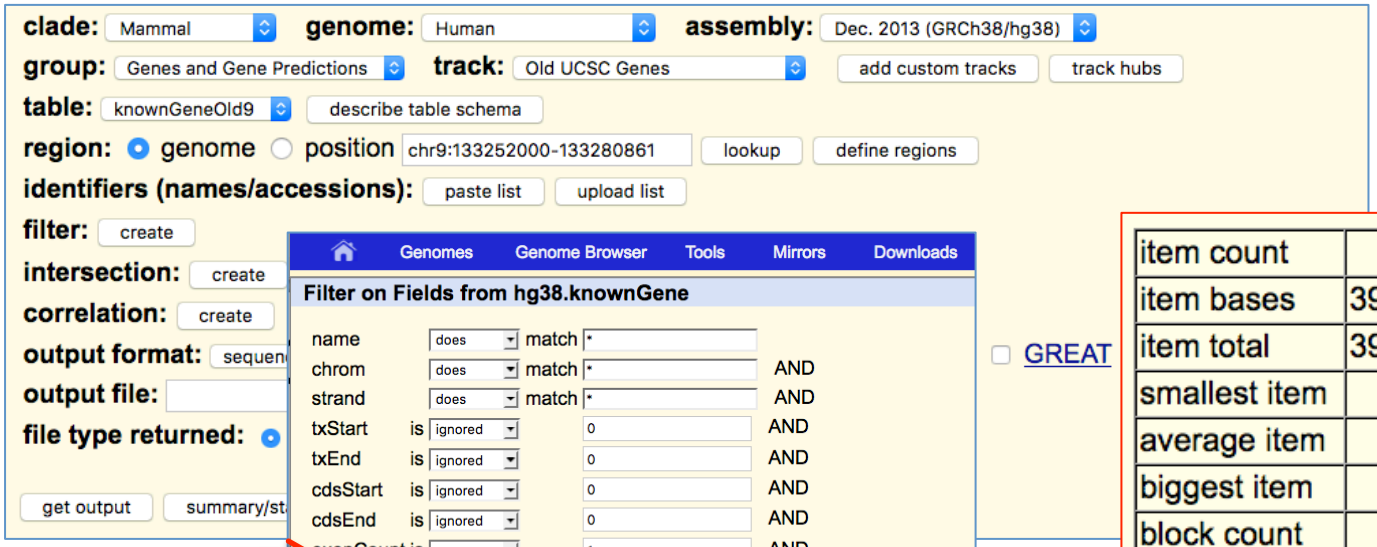

```
chrX    felCat8_refGene exon    21883079        21883169        0.000000        +       .       gene_id "NM_001009262"; transcript_id "NM_001009262";
chrX    felCat8_refGene start_codon     21890310        21890312        0.000000        +       .       gene_id "NM_001009262"; transcript_id "NM_001009262";
chrX    felCat8_refGene CDS     21890310        21891434        0.000000        +       0       gene_id "NM_001009262"; transcript_id "NM_001009262";
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";
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";
chrX    felCat8_refGene exon    42849493        42849594        0.000000        -       .       gene_id "NM_001083952"; transcript_id "NM_001083952";
chrX    felCat8_refGene CDS     42850334        42850410        0.000000        -       2       gene_id "NM_001083952"; transcript_id "NM_001083952";
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";
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";
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";
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.**



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

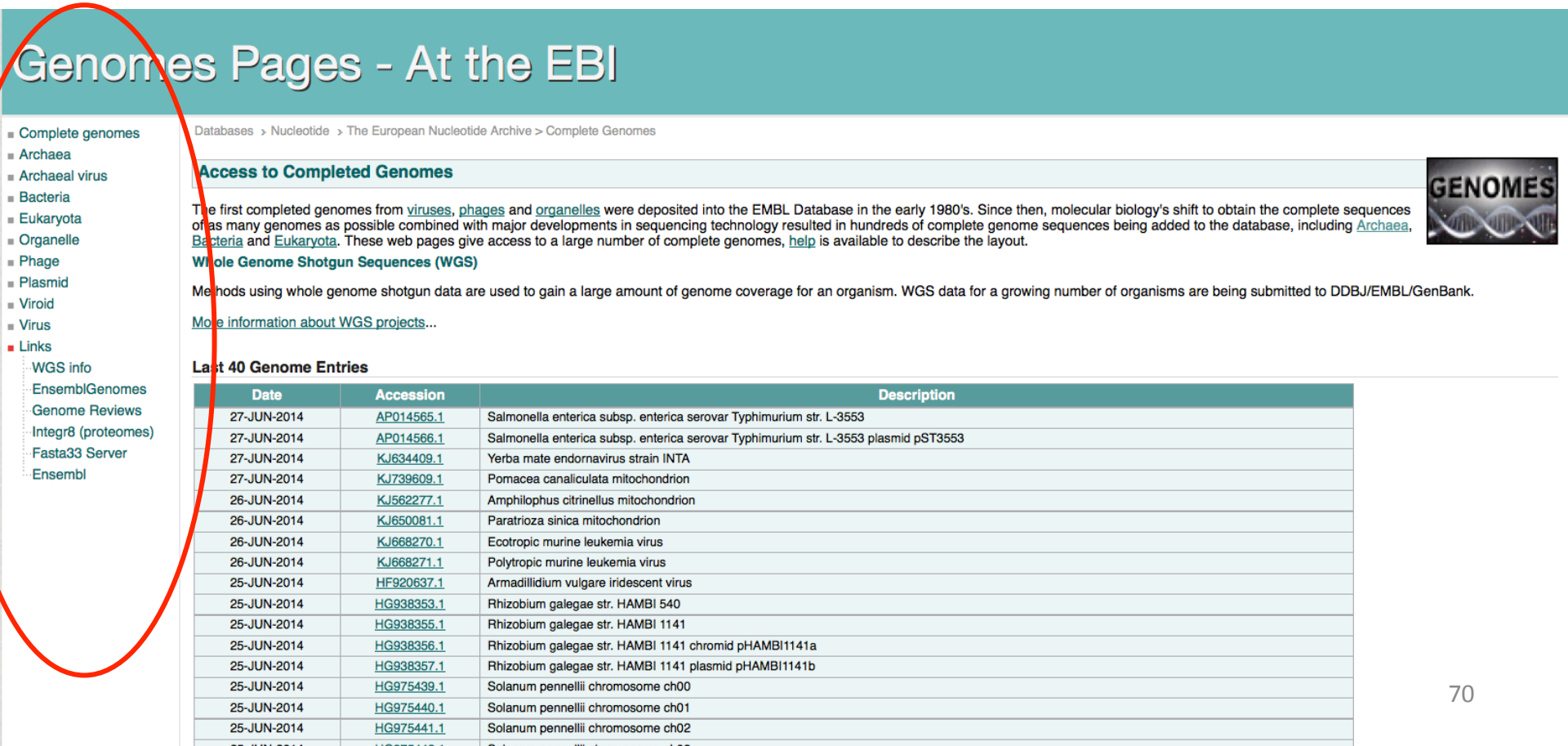
https://www.sanger.ac.uk/resources/downloads/

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

Other genomic and annotation resources

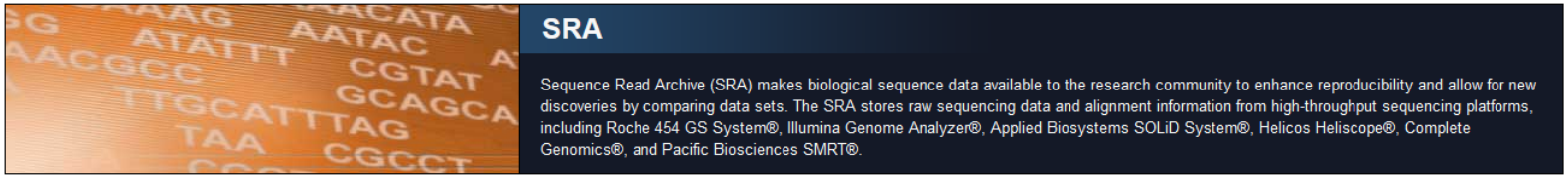
Genomes – EMBL - EBI

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



Basic resources related to the next generation sequencing data

SRA NCBI
ENA EBI
GEO NCBI
ENCODE



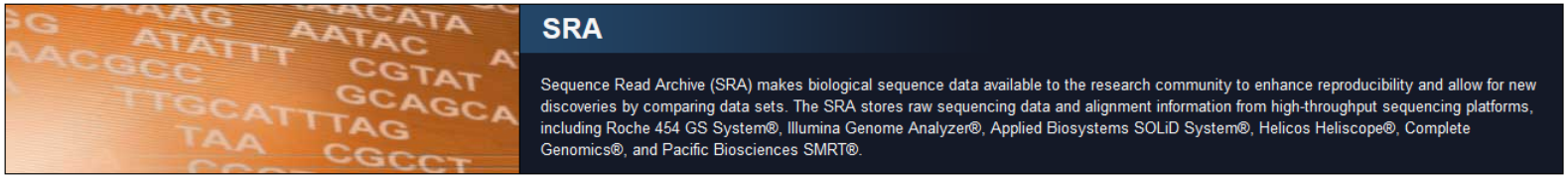
Sequence Read Archive (SRA)

- **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.

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



Sequence Read Archive (SRA)

Study/Project

General aim of the project. May contain many experiments.

SRA accession numbers starting from **SRP**

Experiment

More details about sequencing method and experimental design.

SRA accession numbers starting from **SRX**

Run

Results from particular runs of the machine

SRA accession numbers starting from **SRR**

Sample

Sample which was sequenced.

SRA accession numbers starting from **SRS**

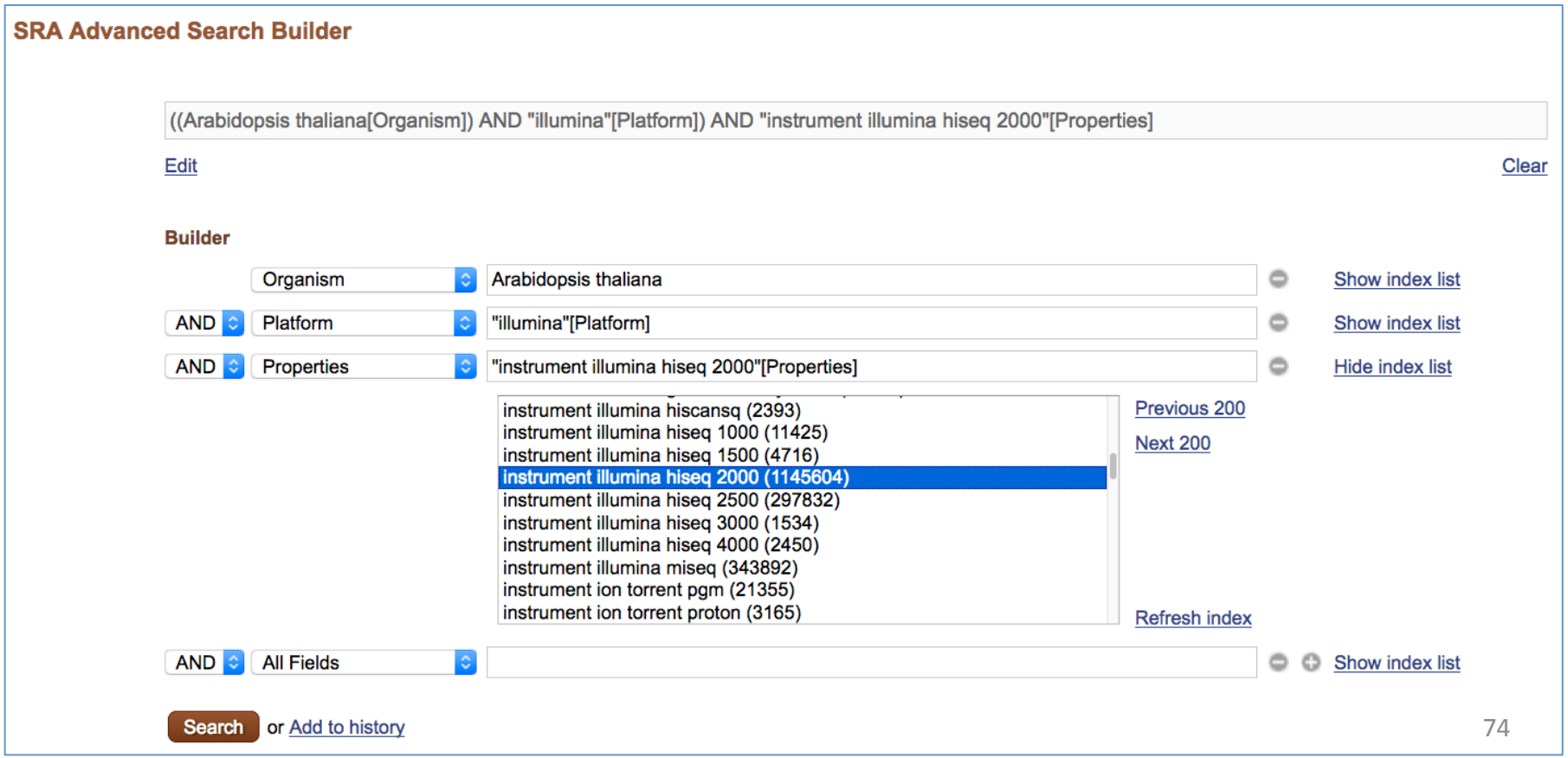
Data originating from ENA (EMBL-EBI)
ERP, ERX, ERR, ERS

Data originating from DDBJ
DRP, DRX, DRR, DRS

SRA: search builder

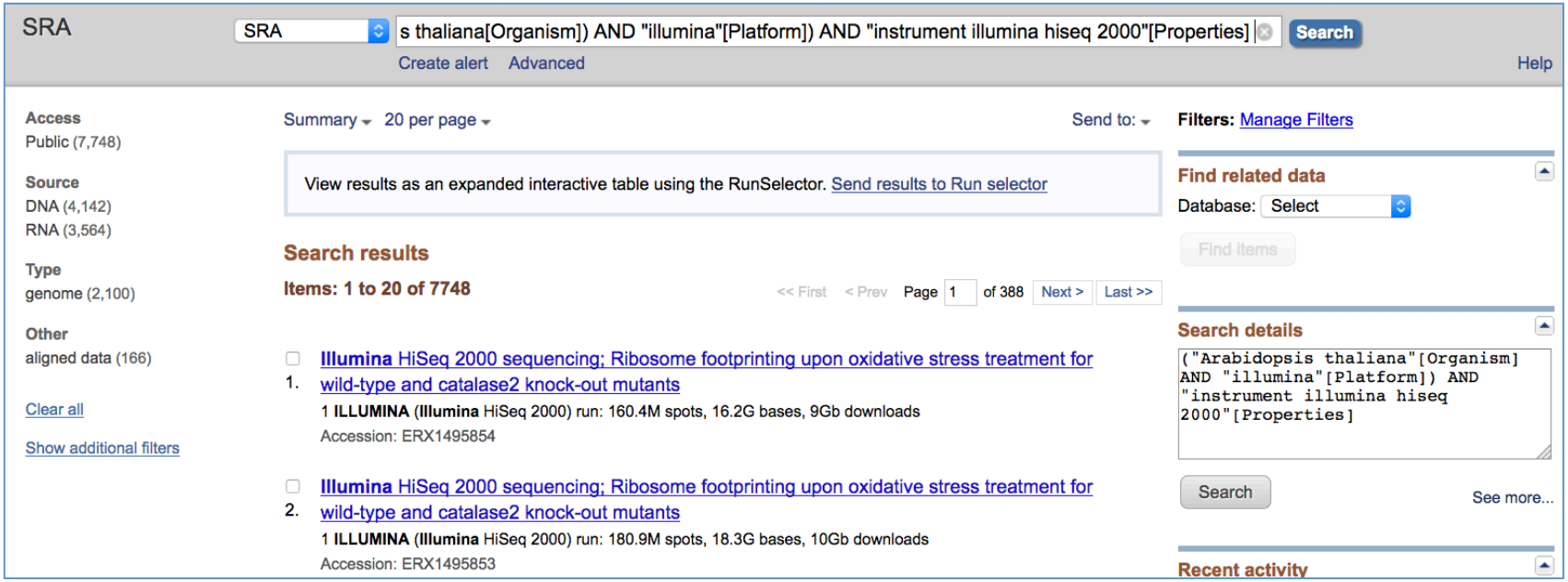
http://www.ncbi.nlm.nih.gov/sra/

Advanced Search > Organism: **Arabidopsis thaliana** (1)
Platform: **„ illumina”** (2), Properties: **„instrument illumina hiseq 2000”** (3)



SRA: search builder

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



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

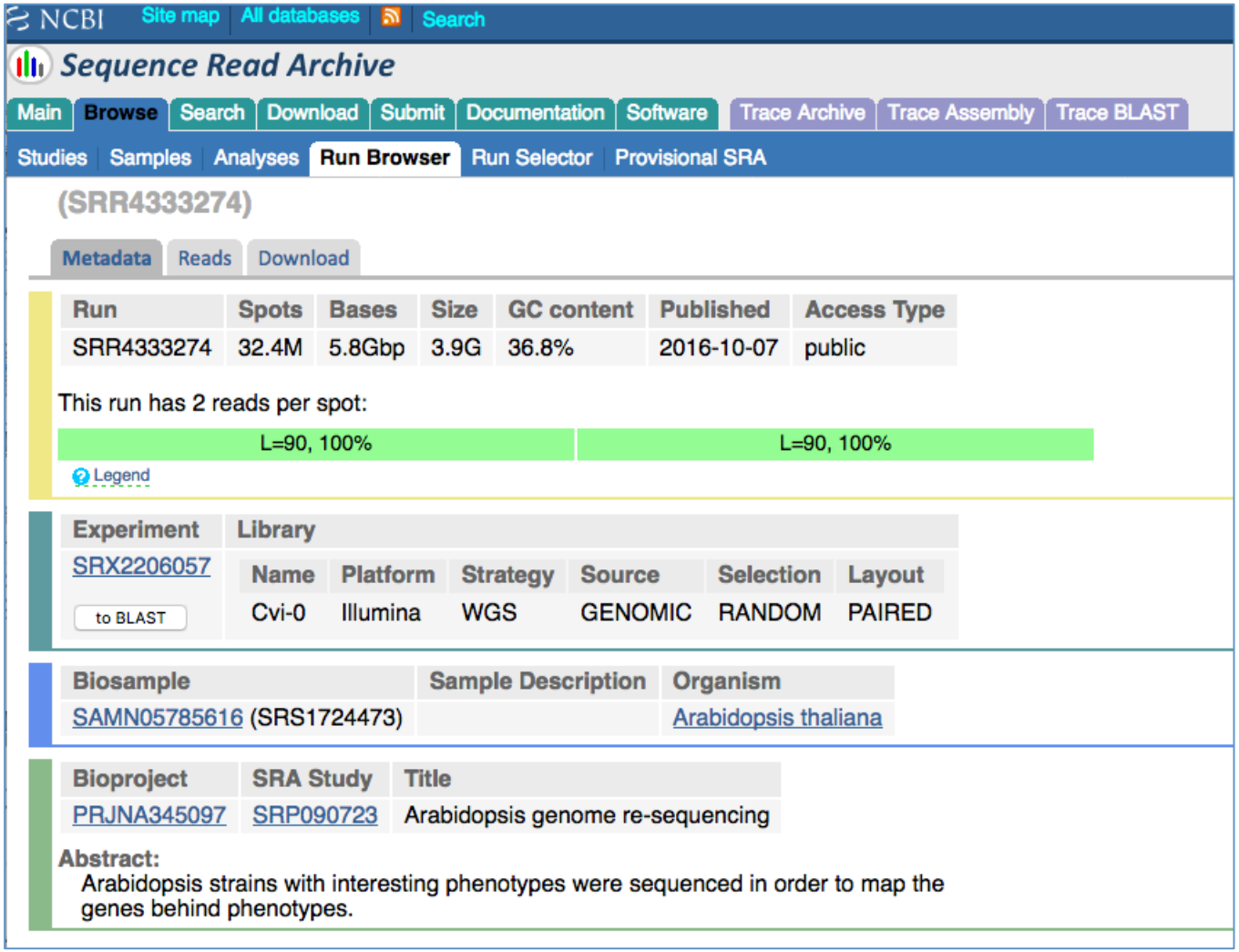
Library:
Name: Cvi-0
Instrument: Illumina HiSeq 2000
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Runs: 1 run, 32.4M spots, 5.8G bases, 3.6Gb

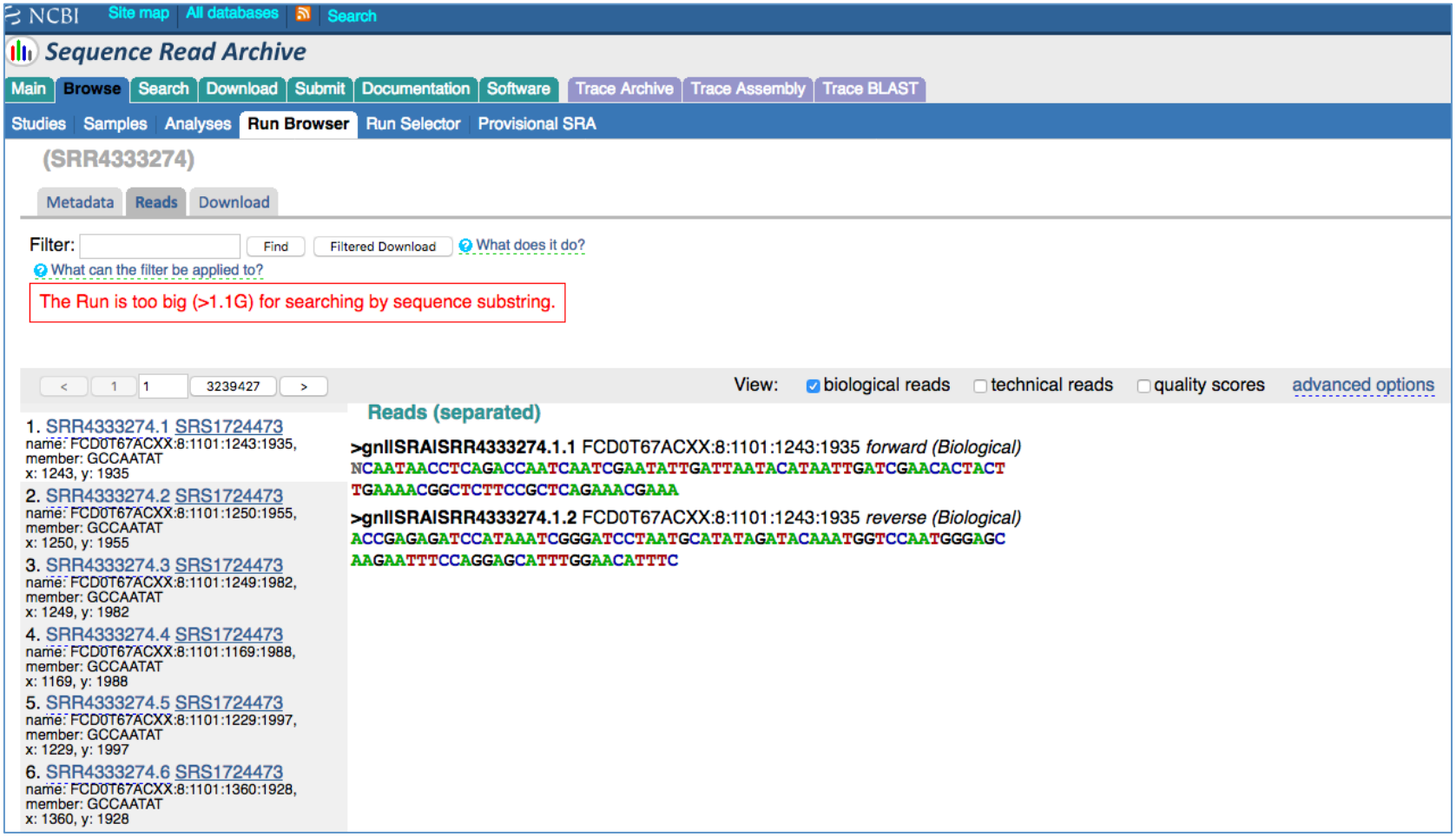
Run	# of Spots	# of Bases	Size	Published
SRR4333274	32,394,270	5.8G	3.6Gb	2016-10-07

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
SRX2206057
to BLAST
Name
Platform
Strategy
Source
Selection
Layout
Cvi-0
Illumina
WGS
GENOMIC
RANDOM
PAIRED
Biosample
Sample Description
Organism
SAMN05785616 (SRS1724473)
Arabidopsis thaliana
Bioproject
SRA Study
Title
PRJNA345097
SRP090723
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)
1. SRR4333274.1 SRS1724473
name: FCD0T67ACXX:8:1101:1243:1935,
member: GCCAATAT
x: 1243, y: 1935
2. SRR4333274.2 SRS1724473
name: FCD0T67ACXX:8:1101:1250:1955,
member: GCCAATAT
x: 1250, y: 1955
3. SRR4333274.3 SRS1724473
name: FCD0T67ACXX:8:1101:1249:1982,
member: GCCAATAT
x: 1249, y: 1982
4. SRR4333274.4 SRS1724473
name: FCD0T67ACXX:8:1101:1169:1988,
member: GCCAATAT
x: 1169, y: 1988
5. SRR4333274.5 SRS1724473
name: FCD0T67ACXX:8:1101:1229:1997,
member: GCCAATAT
x: 1229, y: 1997
6. SRR4333274.6 SRS1724473
name: FCD0T67ACXX:8:1101:1360:1928,
member: GCCAATAT
x: 1360, y: 1928
>gnl|SRA|SRR4333274.1.1 FCD0T67ACXX:8:1101:1243:1935 forward (Biological)
NCAATAACCTCAGACCAATCAATCGAATATTGATTAATACATAATTGATCGAACACTACT
TGAAAACGGCTCTTCCGCTCAGAAACGAAA
>gnl|SRA|SRR4333274.1.2 FCD0T67ACXX:8:1101:1243:1935 reverse (Biological)
ACCGAGAGATCCATAAATCGGGATCCTAATGCATATAGATACAAATGGTCCAATGGGAGC
AAGAATTTCCAGGAGCATTTGGAACATTTC

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

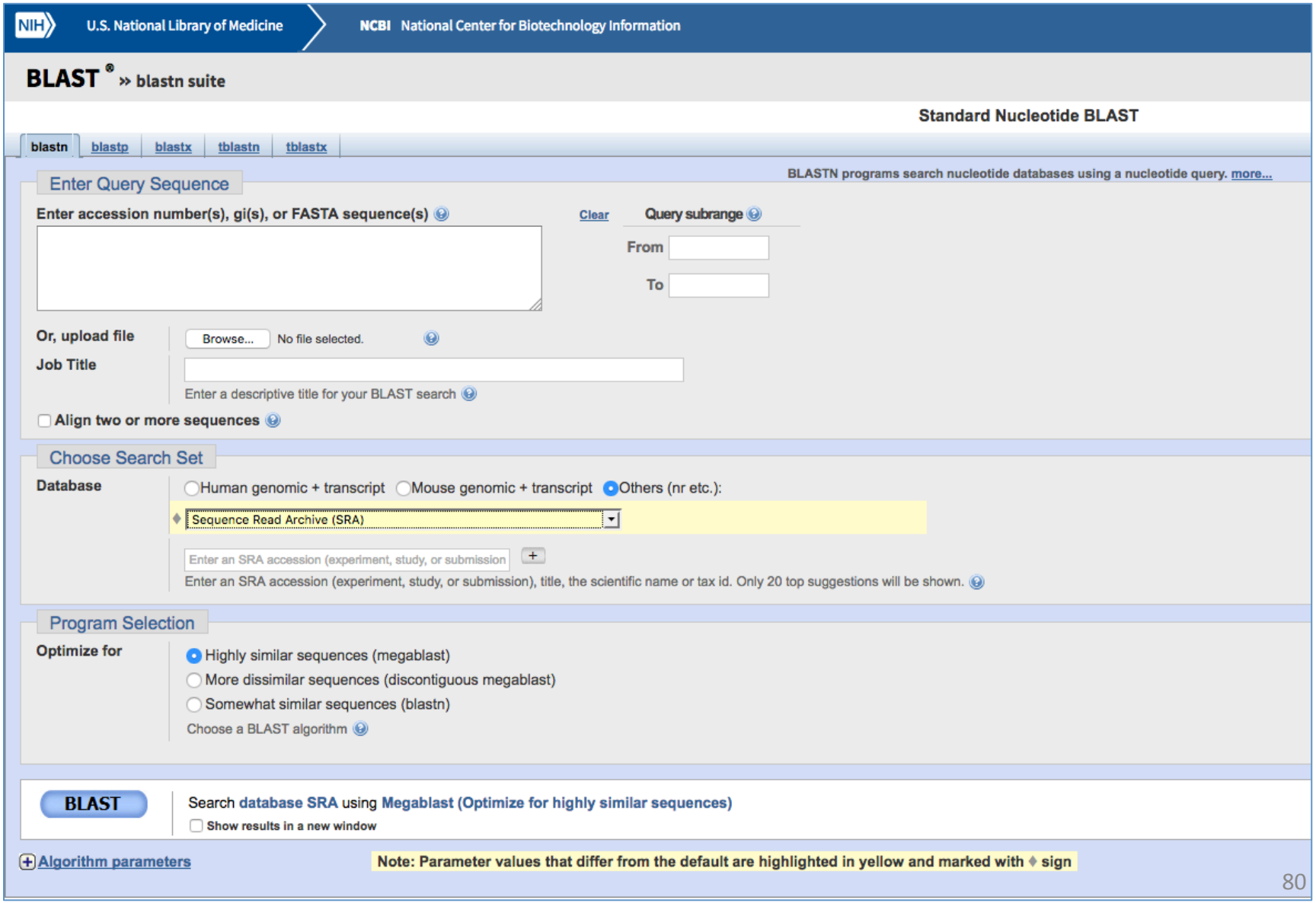
Frequently Used Tools:

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

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
Browse...
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

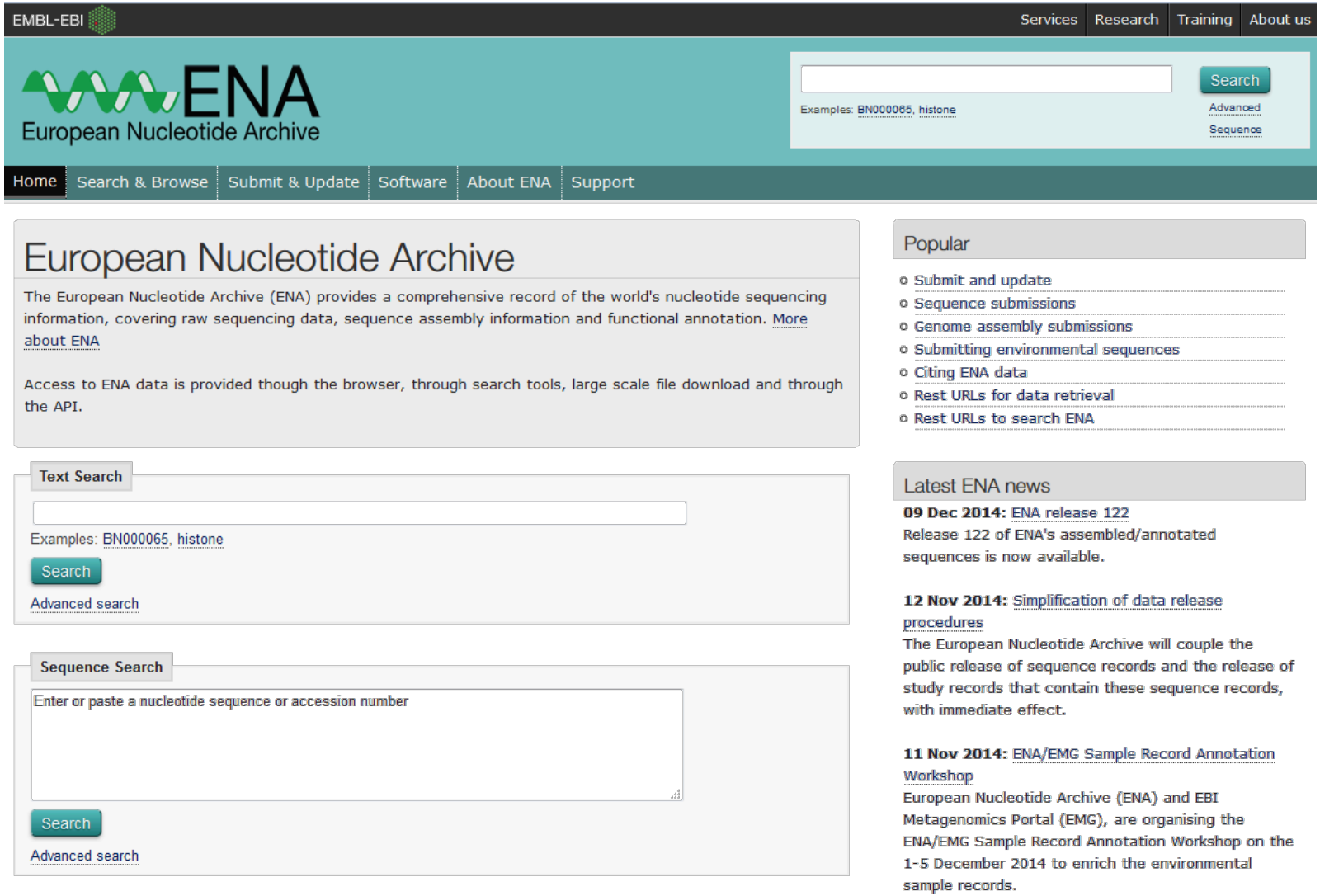
SRA: TASK

- Please find in SRA some reads from experiments for human perfomed 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 accestion numbers (Study, Experiment, Run, Sample) and short description of experiment.

ENA
European Nucleotide Archive

http://www.ebi.ac.uk/ena



ENA
European Nucleotide Archive

http://www.ebi.ac.uk/ena

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

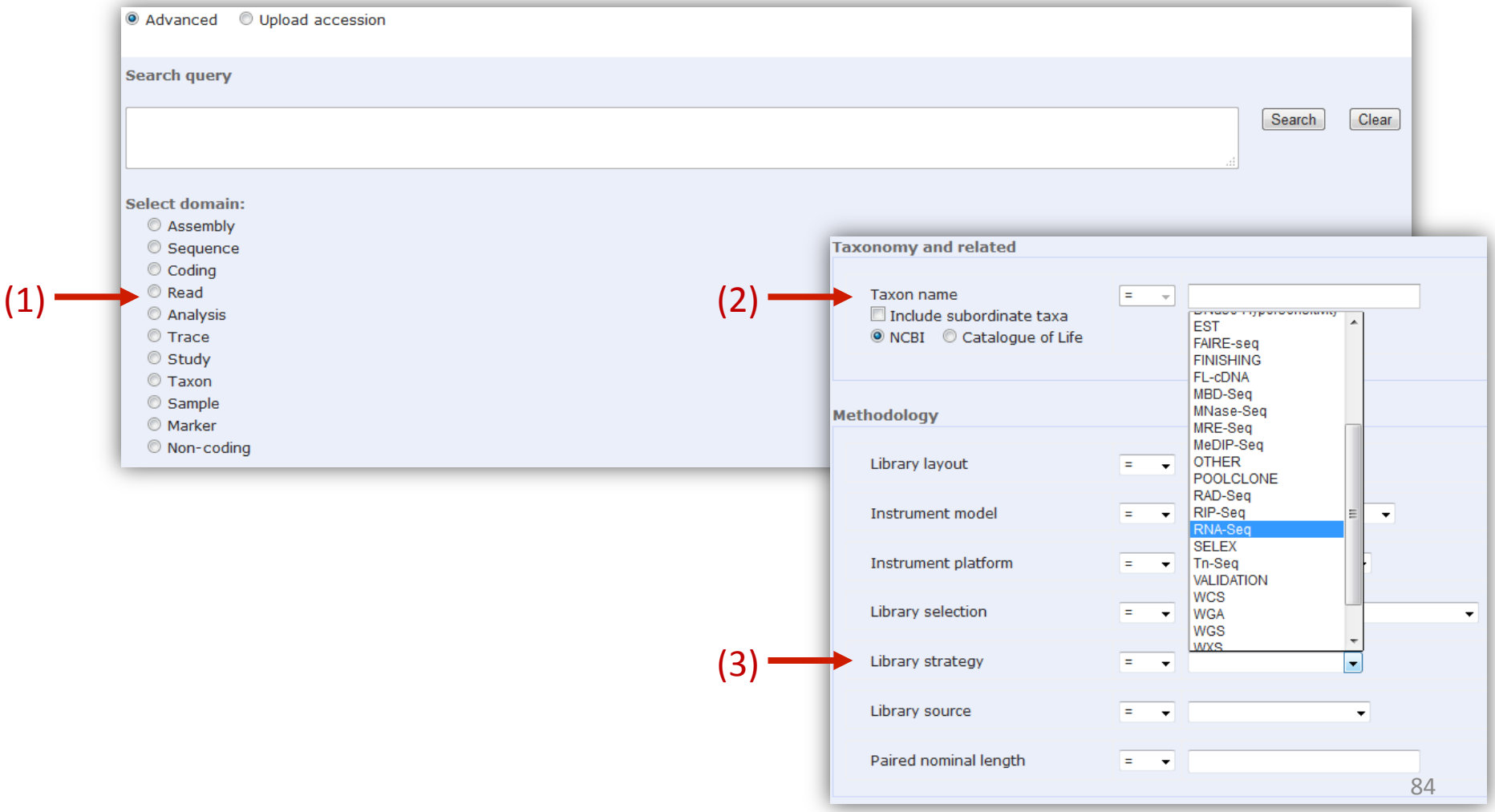
Project/Study	PRJ

ER - EBI
SR – SRA NCBI
DR - DDBJ

European Nucleotide Archive

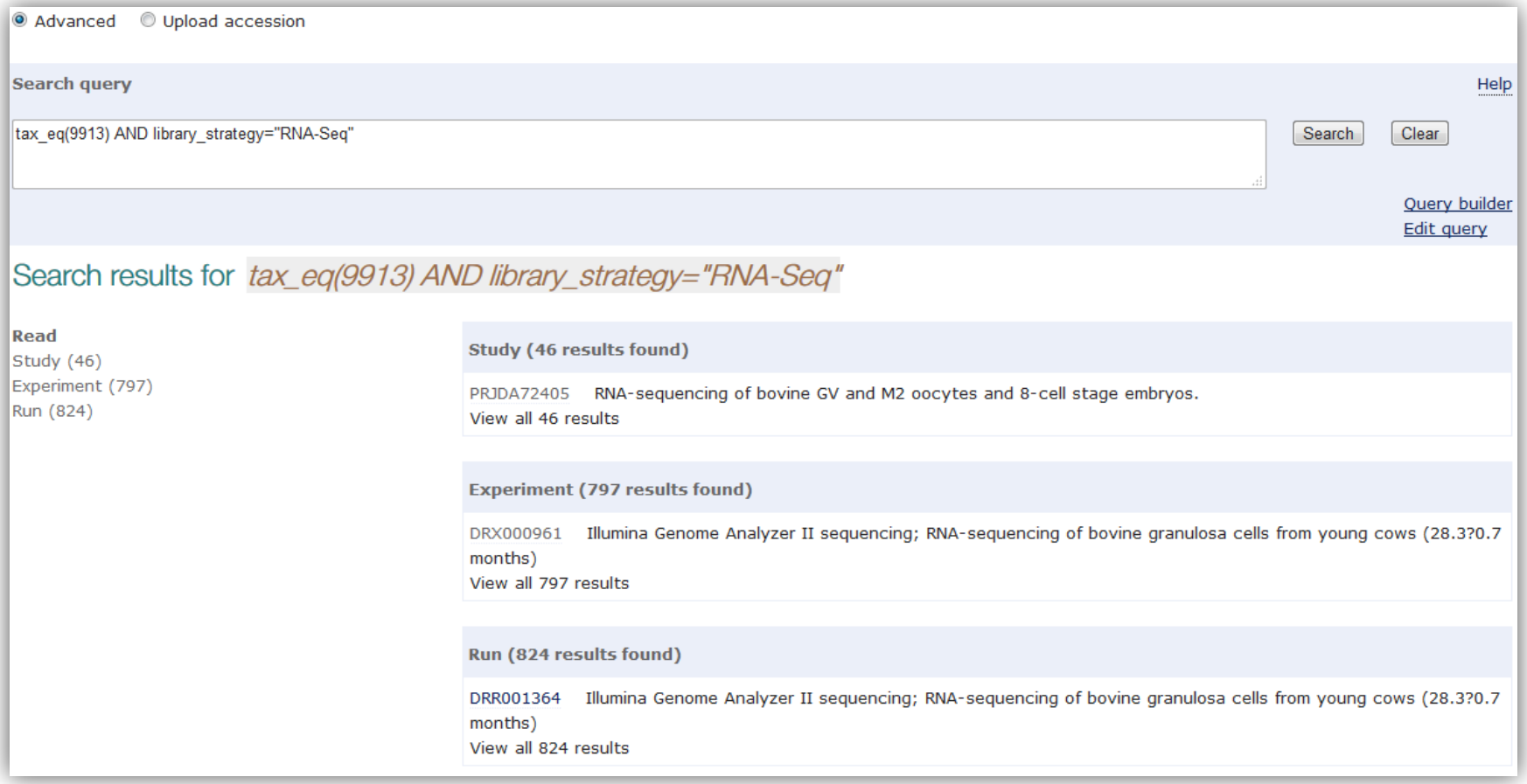
http://www.ebi.ac.uk/ena

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



ENA
European Nucleotide Archive

http://www.ebi.ac.uk/ena



ENA
European Nucleotide Archive

http://www.ebi.ac.uk/ena

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)
PRJDA72405	DRP000449	SAMD00000728	DRS000859	DRX000961	DRR001364	9913	Bos taurus	Illumina Genome Analyzer II	SINGLE	File 1	File 1		
PRJDA72405	DRP000449	SAMD00000728	DRS000859	DRX000962	DRR001365	9913	Bos taurus	Illumina Genome Analyzer II	SINGLE	File 1	File 1		
PRJDA72405	DRP000449	SAMD00000728	DRS000859	DRX000962	DRR001366	9913	Bos taurus	Illumina Genome Analyzer II	SINGLE	File 1	File 1		
PRJDA72405	DRP000556	SAMD00011319	DRS001301	DRX001352	DRR001889	9913	Bos taurus	Illumina Genome Analyzer II	SINGLE	File 1	File 1		
PRJDA72405	DRP000556	SAMD00011318	DRS001302	DRX001354	DRR001890	9913	Bos taurus	Illumina Genome Analyzer II	SINGLE	File 1	File 1		
PRJDA72405	DRP000556	SAMD00011318	DRS001302	DRX001354	DRR001891	9913	Bos taurus	Illumina Genome Analyzer II	SINGLE	File 1	File 1		

ENA
European Nucleotide Archive

http://www.ebi.ac.uk/ena

Download options:

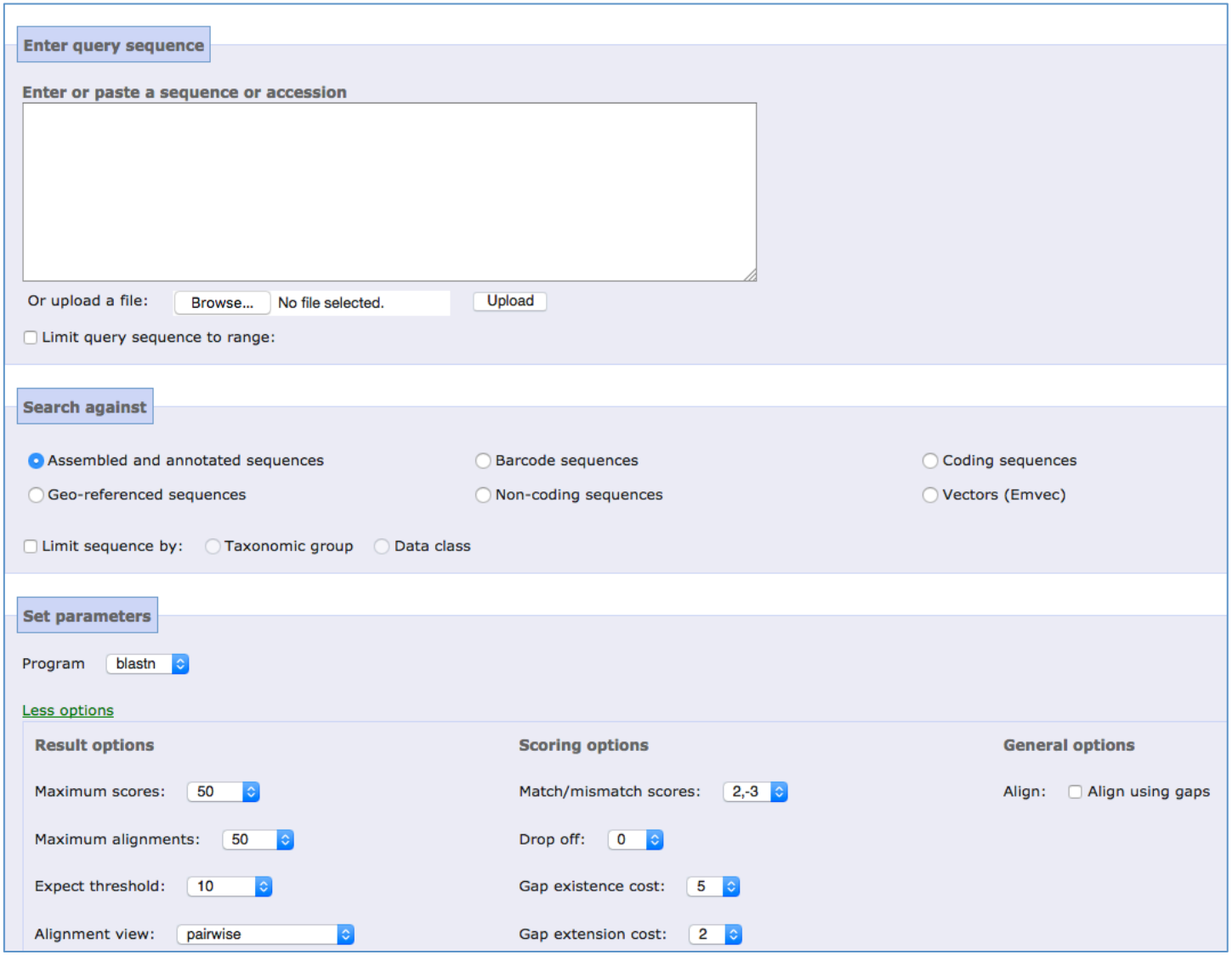
1. FTP files
2. ENA Browser
3. Aspera

Attention! Sequeces 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

ENA
European Nucleotide Archive

Search by sequence



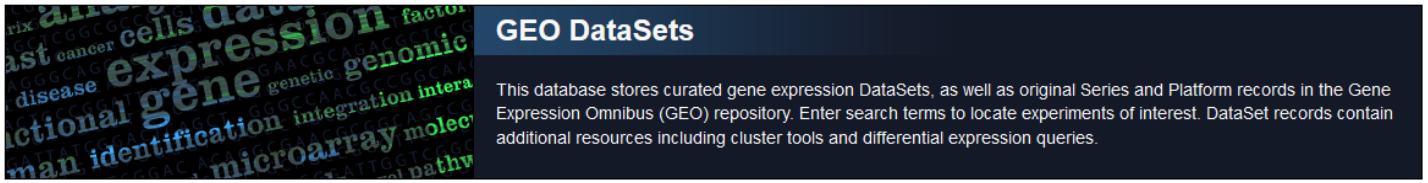
ENA: TASK

- Please find in ENA some reads from experiments for human perfomed on selected platform, for example from transcriptome sequencing (Library strategy → RNA-seq), where as a result paired reads were obtained (Library layout → paired).

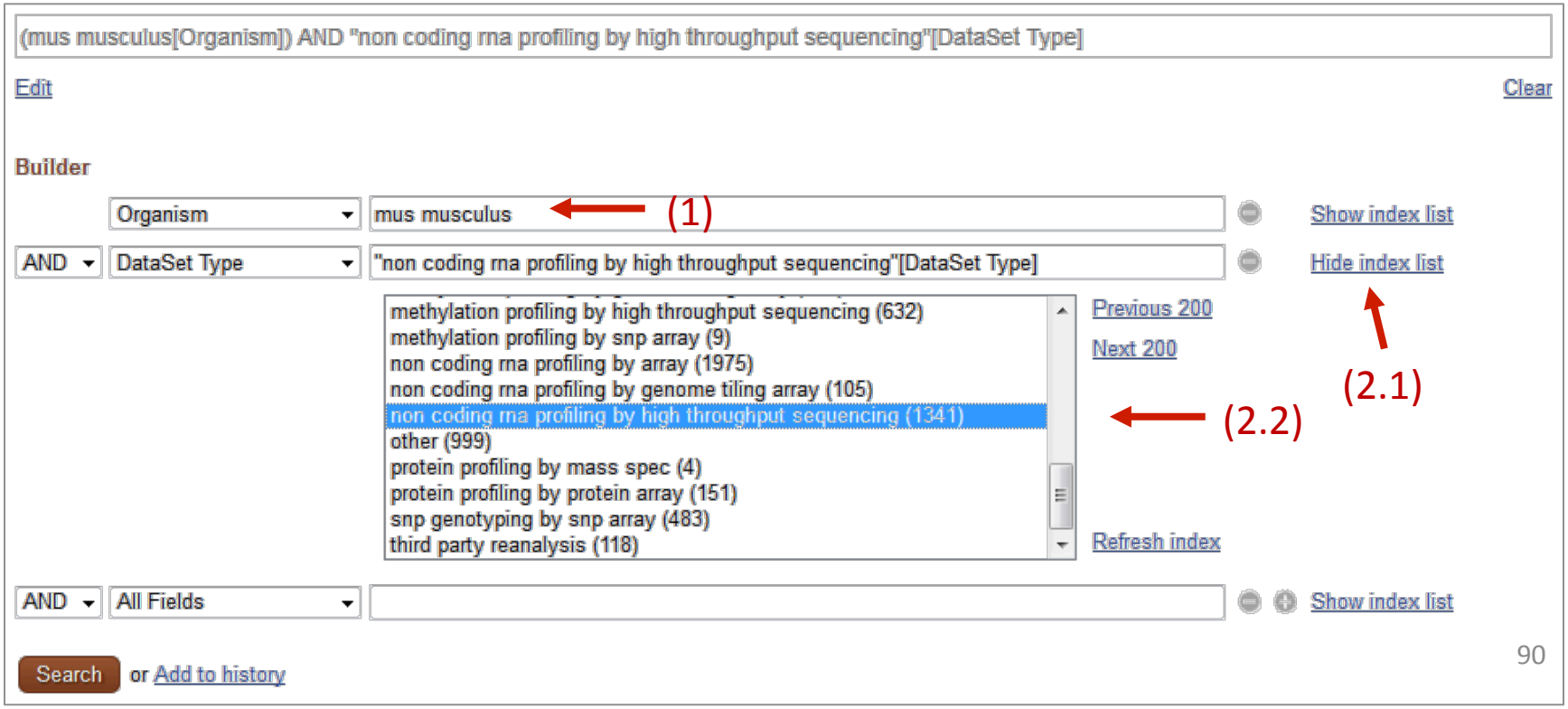
- For some selected record, please find all possible accestion numbers (Study, Experiment, Run, Sample) and short description of experiment.

GEO – Gene Expression Omnibus (NCBI)

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



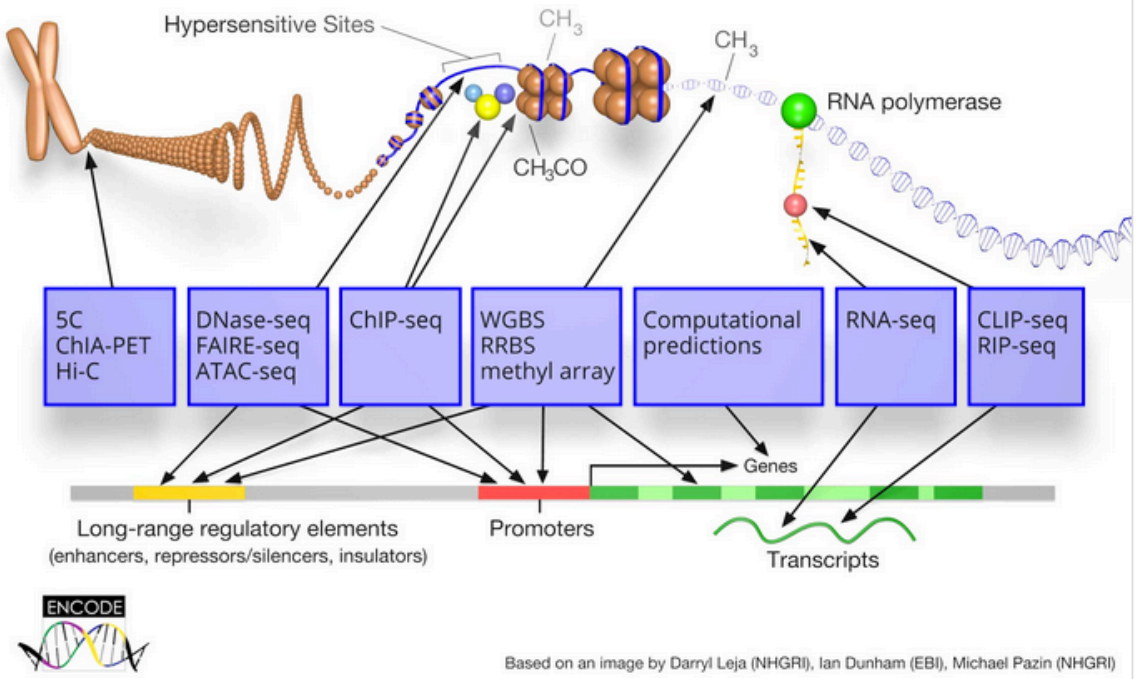
Advanced Search > Organism: **Mus musculus (1)**, DataSet Type: **„non coding rna profiling by high throughput sequencing" (2)**



ENCODE

https://www.encodeproject.org/

ENCODE: Encyclopedia of DNA Elements

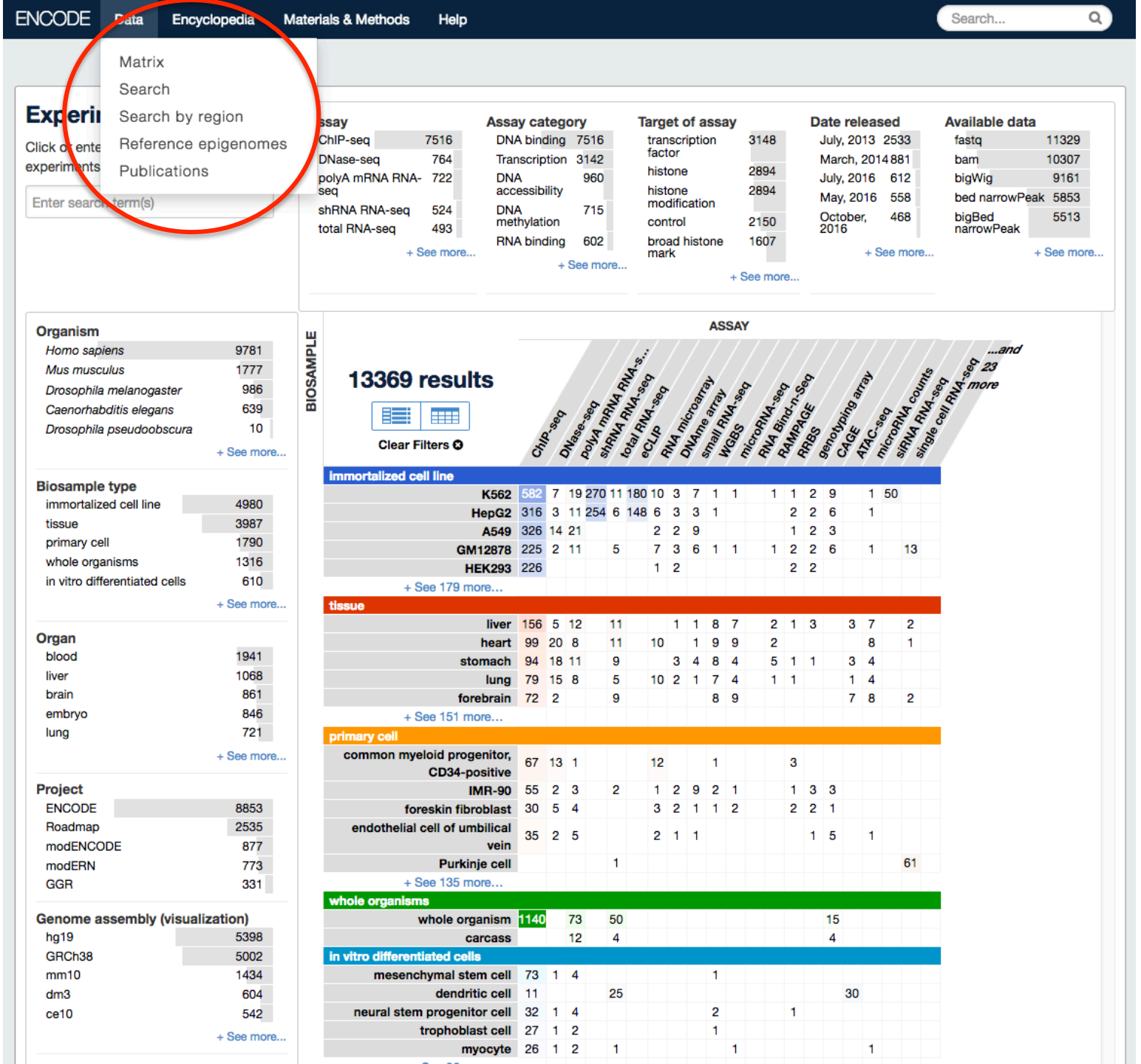


The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at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

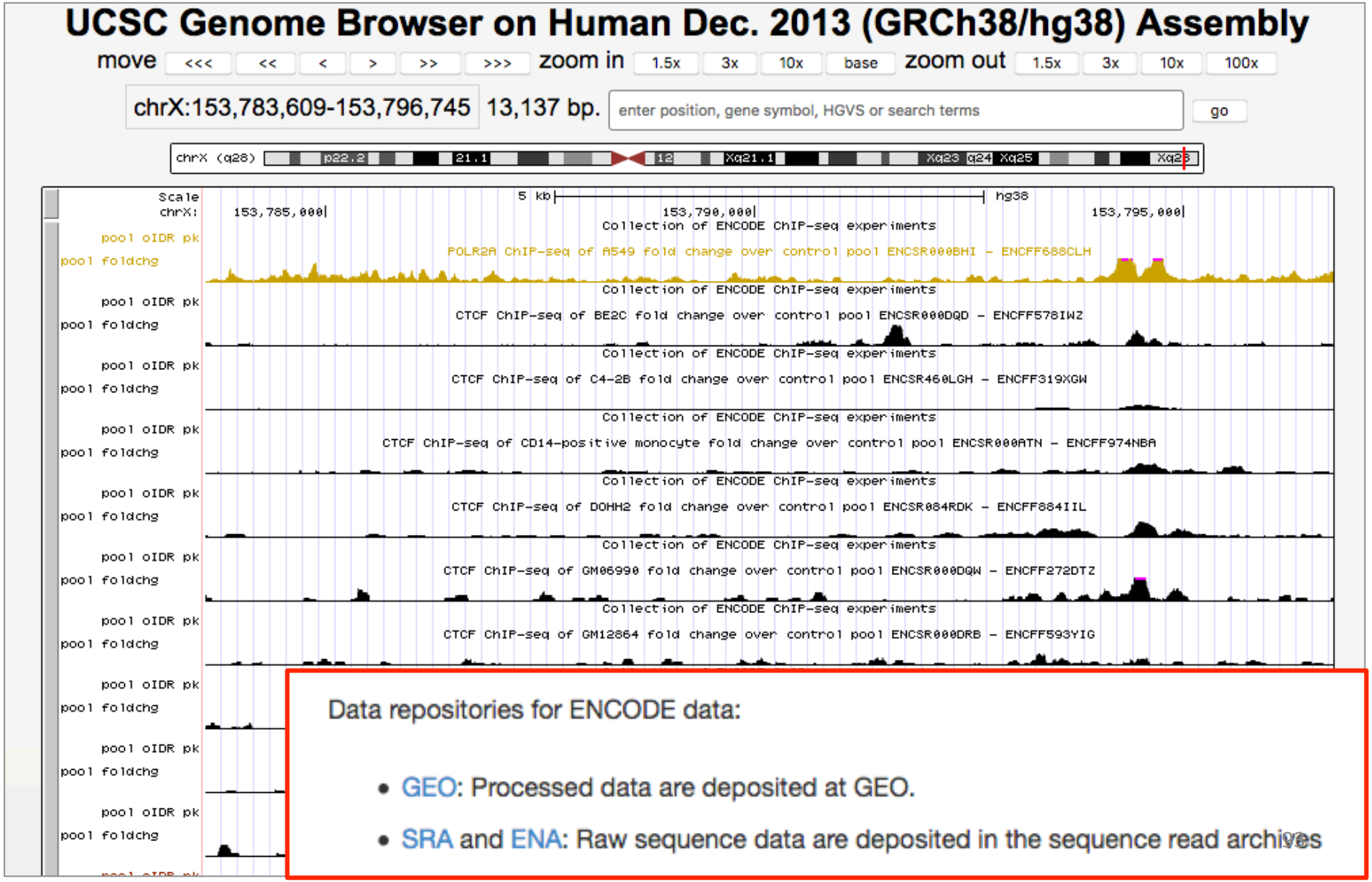
Get Started

HUMAN | MOUSE | WORM | FLY

View Assay Matrix


ENCODE
Data
Encyclopedia
Materials & Methods
Help
Search...
Matrix
Search
Search by region
Reference epigenomes
Publications
Enter search term(s)
Assay
ChIP-seq 7516
DNase-seq 764
polyA mRNA RNA-seq 722
shRNA RNA-seq 524
total RNA-seq 493
+ See more...
Assay category
DNA binding 7516
Transcription 3142
DNA accessibility 960
DNA methylation 715
RNA binding 602
+ See more...
Target of assay
transcription factor 3148
histone 2894
histone modification 2894
control 2150
broad histone mark 1607
+ See more...
Date released
July, 2013 2533
March, 2014 881
July, 2016 612
May, 2016 558
October, 2016 468
+ See more...
Available data
fastq 11329
bam 10307
bigWig 9161
bed narrowPeak 5853
bigBed narrowPeak 5513
+ See more...
Organism
Homo sapiens 9781
Mus musculus 1777
Drosophila melanogaster 986
Caenorhabditis elegans 639
Drosophila pseudoobscura 10
+ See more...
Biosample type
immortalized cell line 4980
tissue 3987
primary cell 1790
whole organisms 1316
in vitro differentiated cells 610
+ See more...
Organ
blood 1941
liver 1068
brain 861
embryo 846
lung 721
+ See more...
Project
ENCODE 8853
Roadmap 2535
modENCODE 877
modERN 773
GGR 331
Genome assembly (visualization)
hg19 5398
GRCh38 5002
mm10 1434
dm3 604
ce10 542
+ See more...
13369 results
Clear Filters
BIOSAMPLE
ASSAY
ChIP-seq
DNase-seq
polyA mRNA RNA-s...
shRNA RNA-seq
total RNA-seq
eCLIP
RNA microarray
DNAme array
small RNA-seq
WGBS
microRNA-seq
RNA Bind-n-Seq
RAMPAGE
RRBS
genotyping array
CAGE
ATAC-seq
microRNA counts
siRNA RNA-seq
single cell RNA-seq
...and 23 more
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 1 13
HEK293 226 1 2 2 2
+ See 179 more...
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
+ See 151 more...
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
+ See 135 more...
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

ENCODE → UCSC



Data repositories for ENCODE data:

- GEO: Processed data are deposited at GEO.
- SRA and ENA: Raw sequence data are deposited in the sequence read archives

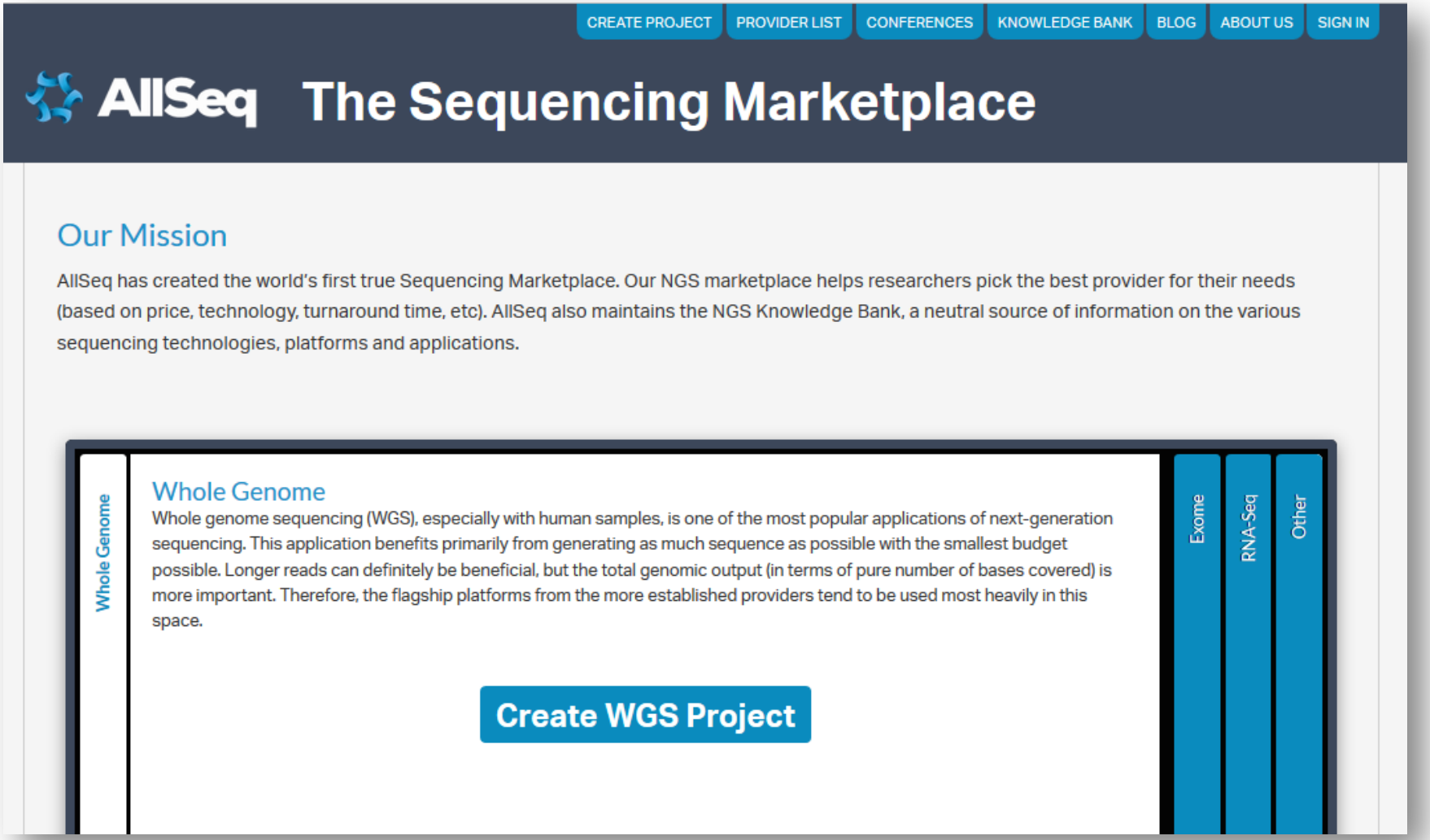
Other selected databases

- NONCODE
  http://noncode.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.medplantrnaseq.org/
- 1000 Genomes Project:
  http://www.internationalgenome.org/data

Where to look for help?

AllSeq

http://allseq.com/

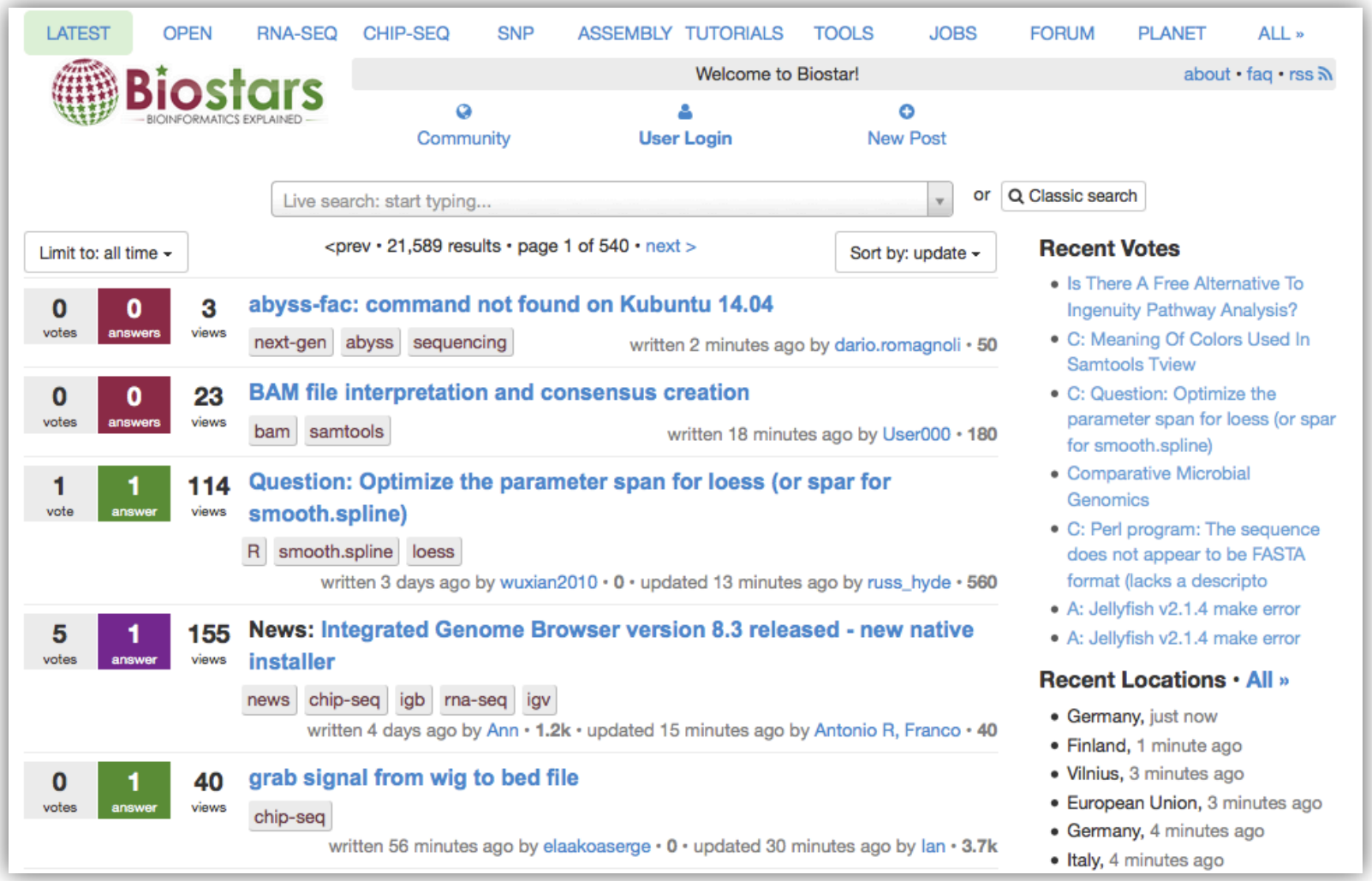


SEQanswers

http://seqanswers.com/

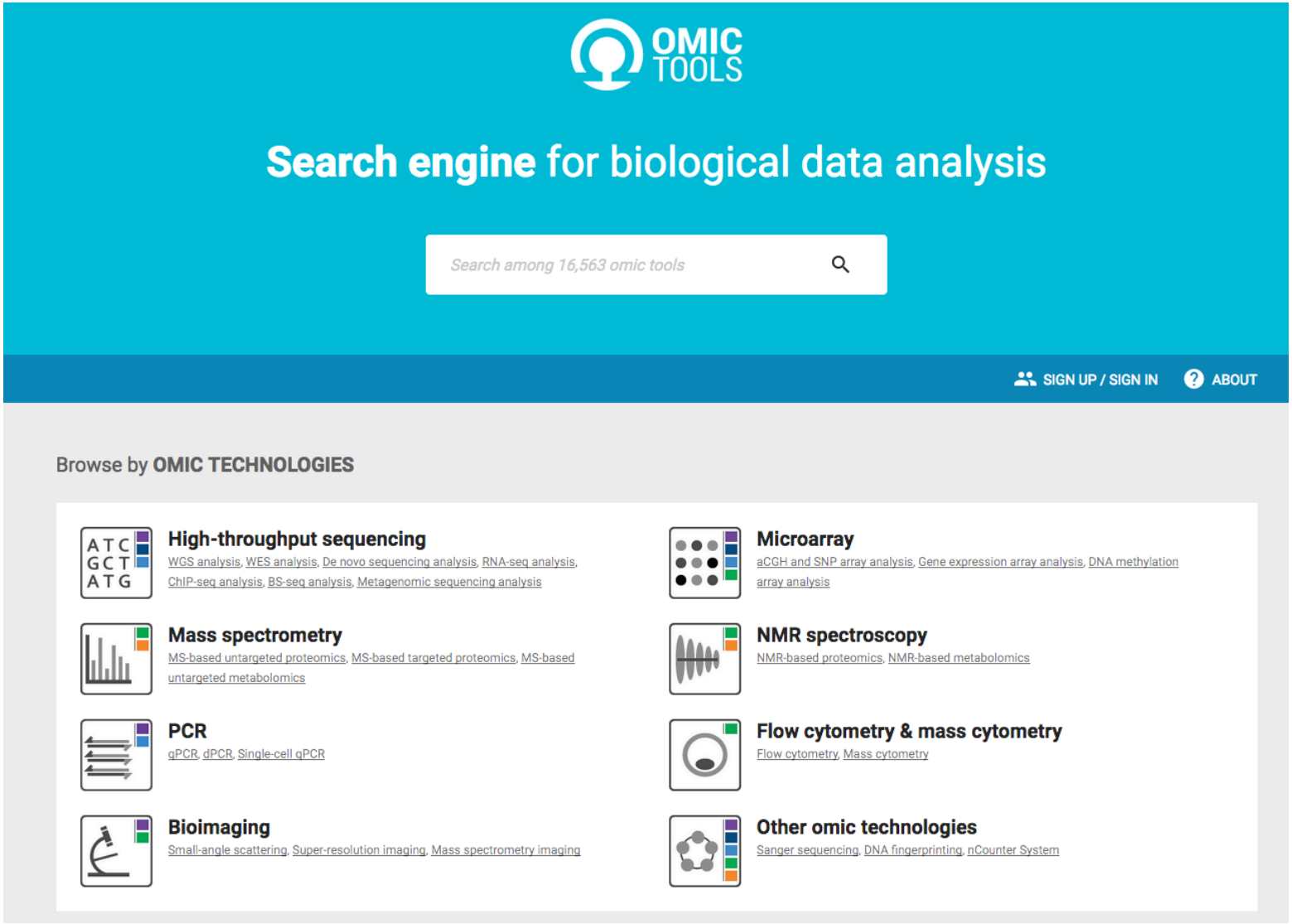
BioStars

https://www.biostars.org/



OMICtools

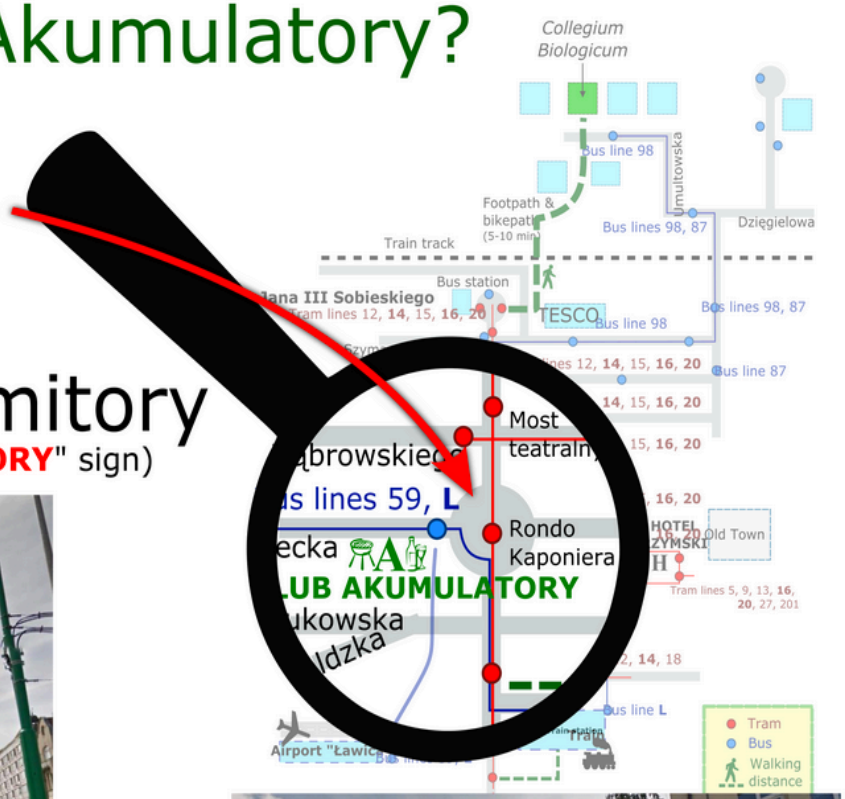
https://omictools.com/



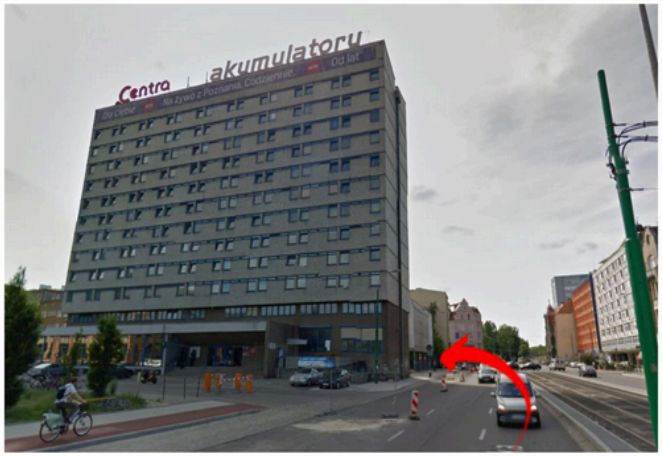
Thank you for your attention ☺

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)