Wprowadzenie do analizy danych RNA-Seq

Michał Szcześniak

Wprowadzenie, źródła i formaty danych

Genomy, transkryptomy, przeglądarki genomowe,
źródła danych NGS

Technologie i platformy NGS

SOLiD®

illumina®

PACIFIC BIOSCIENCES™

life

ion torrent

ThermoFisher SCIENTIFIC

Roche 454 SEQUENCING

Oxford NANOPORE Technologies

Sekwencjonowanie genomów

Sekwencjonowanie transkryptomów (RNA-Seq)

Interakcje białko – RNA (CLIP-seq)

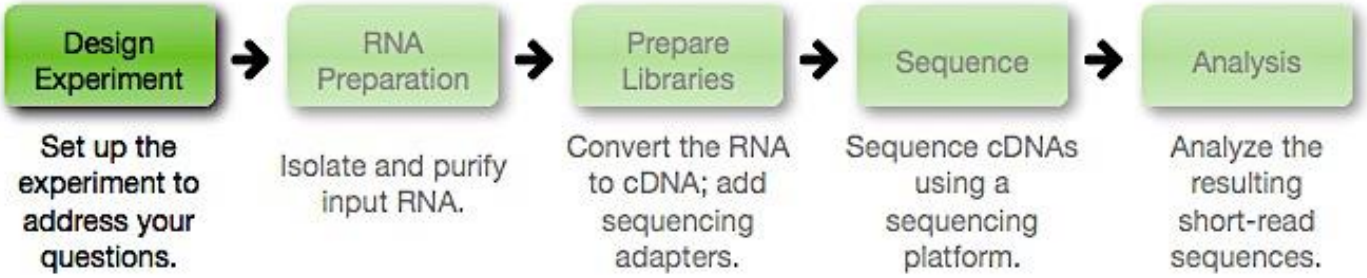
Interakcje białko – DNA (ChIP-seq)

Badanie metylacji DNA

Metagenomika

Sekwencjonowanie transkryptomów RNA-Seq

- ✓ Kodujący mRNA
- ✓ Niekodujący RNA → miRNA, małe RNA, długie niekodujące RNA
- ✓ Cały transkryptom



http://allseq.com/

http://rnaseq.uoregon.edu

Jakich danych potrzebujemy?

- ✓ Sekwencje genomowe
- ✓ Sekwencje transkryptomów
- ✓ Adnotacje
- ✓ Odczyty z sekwencjonowania RNA-seq

Adnotacje genomów

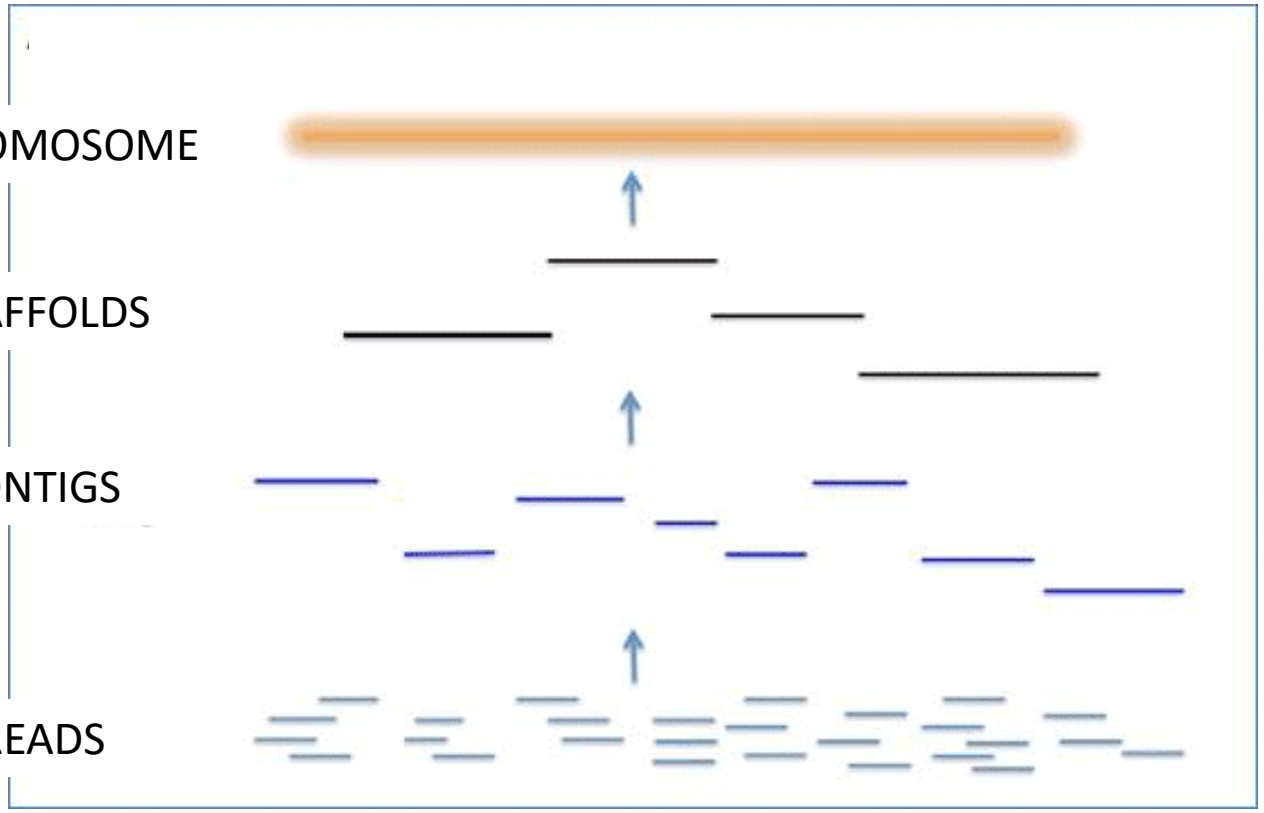
- W przypadku genomów, które nie są w pełni zsekwencjonowane lub złożone (ang. *assembled*), dysponujemy sekwencjami **kontigów** (ang. *contigs*) i **skafoldów** (ang. *scaffolds*).

- Dodatkowo, w przypadku dobrze zaadnotowanych genomów (np. człowiek, mysz) do **chromosomów** dodawane są **łatki** (ang. *patches*), zawierające sekwencje, których nie włączono do chromosomów.

- Niekiedy dysponujemy nie tylko genomem **referencyjnym** ale również **alternatywnymi** złożeniami genomów (inne grupy badawcze, konsorcja).

Składanie genomu - genome assembly

ASSEMBLING = SKŁADANIE



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

PRZYDATNE DEFINICJE: https://www.ncbi.nlm.nih.gov/grc/help/definitions

Wybrane formaty danych

1. FASTA → sekwencje

2. FASTQ → odczyty NGS

3. BAM, SAM → zmapowania NGS

4. BED → koordynaty

5. GTF, GFF3 → adnotacje

6. Plik tekstowy tabularny (TSV)→ np. wyniki działania programu BLAST, wyniki analiz ekspresji

7. PSL → wynik działania programu BLAT

Wybrane formaty danych

FASTA

Plik FASTA składa się z **nagłówka** (zaczyna się od znaku '**>**') po którym następuje **sekwencja.** Może zawierać wiele sekwencji (MULTIFASTA). Rozszerzenie **.fasta** albo **.fa**

>gi|453055269|gb|KC207570.1| Homo sapiens transient receptor potential cation channel subfamily C member 3 variant c (TRPC3) mRNA
GGATATAGAAATGGGAATGGGTAACTCAAAGTCCAGGCAGATAATGAAAAGACTTATAAAGCGGTATGTTTTGAAAGCACAAGTAGACAA
AGAAAATGATGAAGTTAATGAAGGTGAATTAAAAGAAATCAAGCAAGATATCTCCAGCCTTCGTTATGAACTTTTGGAAGACAAGAGCCAA
GCAACTGAGGAATTAGCCATTCTAATTCATAAACTTAGTGAGAAACTGAATCCCAG

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

Wybrane formaty danych

FASTA

Kod IUPAC

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

Wybrane formaty danych

BED

Plik z rozszerzeniem **.bed** zawiera koordynaty czyli informację o lokalizacji jakiegoś rejonu (genu, egzonu, intronu) w genomie. Wymagane pola to chromosom, start i koniec rejonu. Może zawierać więcej kolumn.

```
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

Wybrane formaty danych

GTF, GFF3

Pliki z rozszerzeniem **.gtf .gff3** zawierają informacje o elementach w genomie i ich adnotacjach. Danemu elementowi przypisana jest jedna linia w pliku składająca się z dziewięciu kolumn.

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=OTTH
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.

Przegląd podstawowych źródeł danych

dotyczących genomów i ich adnotacji

Trzy metabazy i przeglądarki

National Center for Biotechnology Information

Ensembl Genome Browser

European Bioinformatics Institute, Wellcome Trust Sanger Institute

University of California, Santa Cruz Genome Browser

RefSeq - sekwencje – identyfikatory

RefSeq Database – zbiór sprawdzonych, niepowtarzalnych sekwencji, często weryfikowanych manualnie.

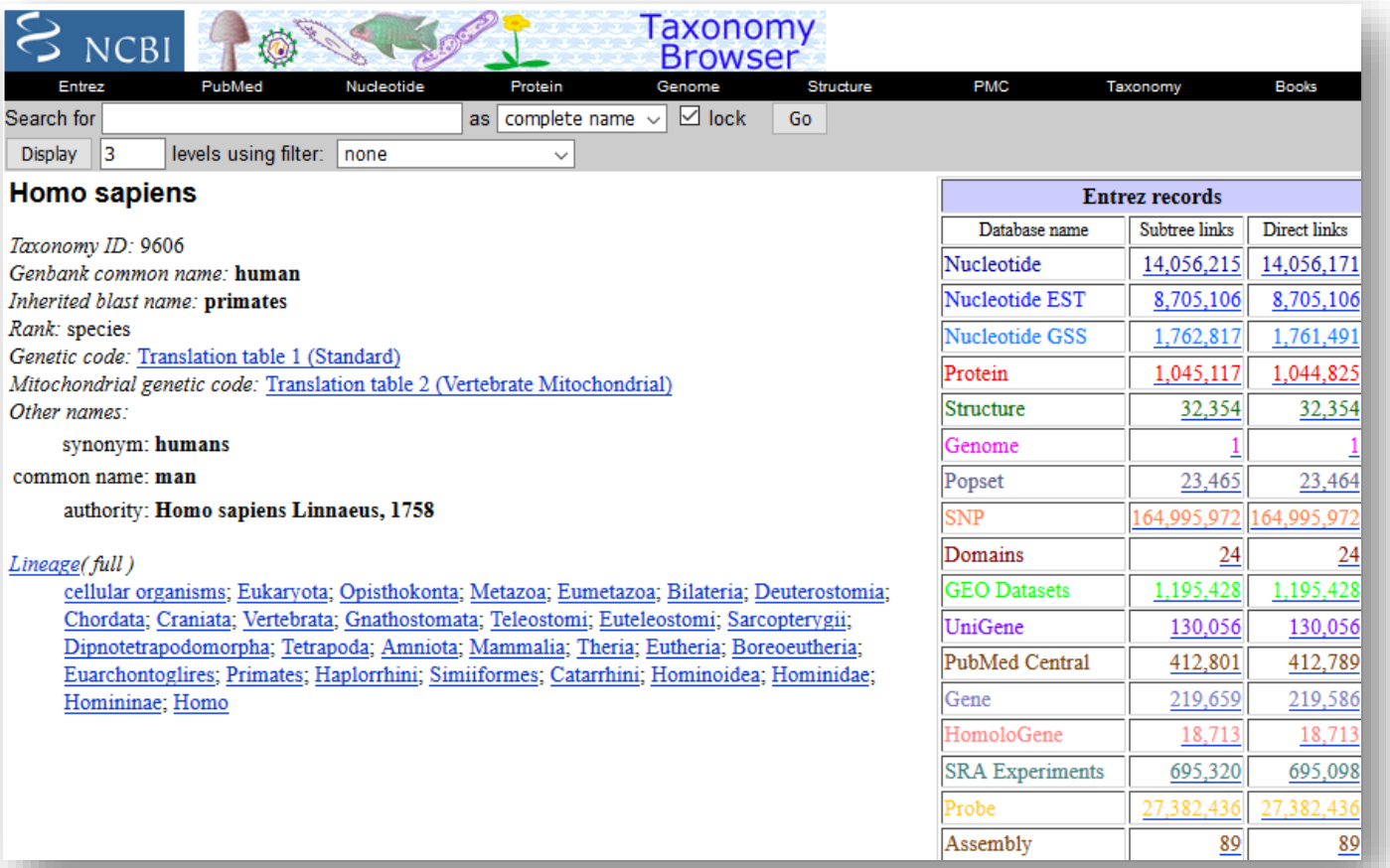
NM_..., XM_... → mRNA
NP_..., XP_... → białka
NR_..., XR_... → niekodujące RNA
NC_..., NG_..., NT…., NW…, AC... → kontigi, sekwencje genomowe

N oznacza sekwencje uzyskane z wyników eksperymentów

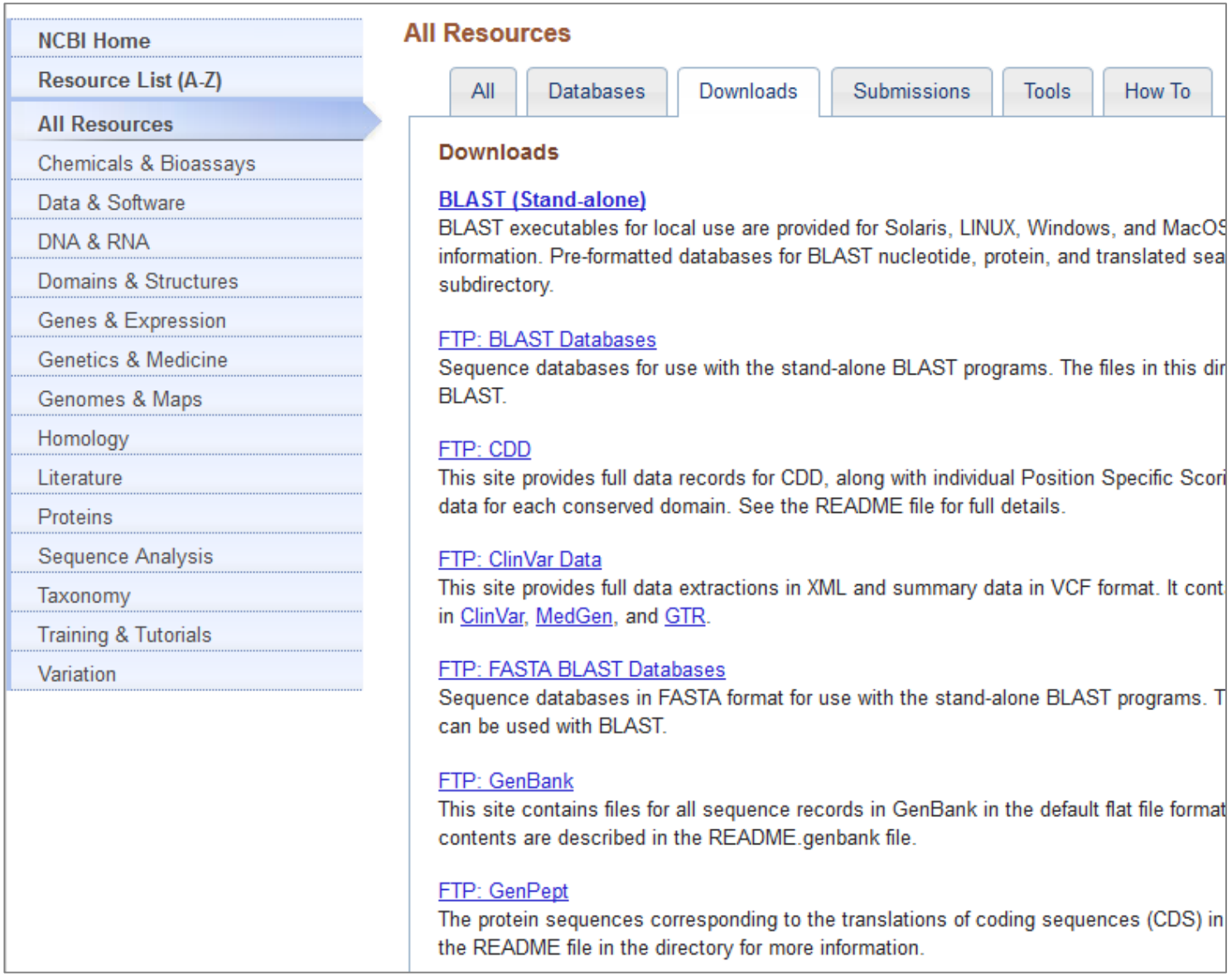
X oznacza, że cząsteczka została przewidziana bioinformatycznie np. przez zmapowanie białka do genomu spokrewnionego organizmu.

NCBI (Taxonomy)

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



NCBI (Downloads FTP)



NCBI (Downloads FTP)

Proszę poszukać sekwencję genomową chromosomu 22 u człowieka i pobrać ją spakowaną w formacie FASTA (.fa) dla najnowszej wersji genomu referencyjnego.

Trzy metabazy i przeglądarki

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 - identyfikatory

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 – pobieranie danych

1. Pojedyncze rekordy w bazie danych

(Uwaga! Na różnym poziomie – różne dane!)

2. Download → FTP

3. BioMart

Ensembl: **ćwiczenie**

W przeglądarce Ensembl proszę znaleźć **gen *BRCA1* u człowieka**.

1. Ile ma form splicingowych?
2. Na którym chromosomie i na jakiej nici się znajduje?

Proszę wybrać jeden transkrypt tego genu, a następnie:

1. Znaleźć kodowane przez nie białko
2. Obejrzeć strukturę intronowo-egzonową
3. Znaleźć sekwencję pierwszego egzonu
4. Pobrać jego sekwencję mRNA
5. Pobrać sekwencję genomową wraz z 100 nt rejonami otaczającymi

Ensembl – pobieranie danych (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 – pobieranie danych

1. Proszę poszukać pliki zawierające sekwencje:

- **Genomową (DNA) chromosomu 22**
- **białkowe** w formacie FASTA
- **Adnotacje** w formacie GTF **człowieka**

- *Jak nazywają się te pliki?*
- *Jak nazywa się plik z całą sekwencją genomową w formacie FASTA?*

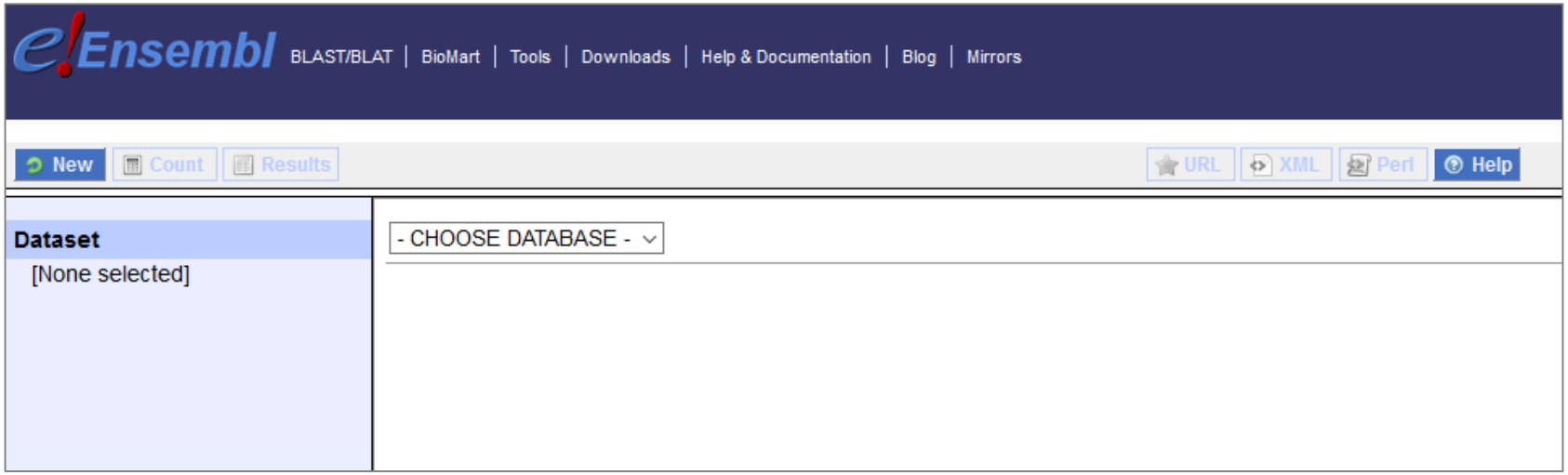
2. Proszę poszukać sekwencje **cDNA** w formacie FASTA dla bakterii ***Bacillus subtilis*** *[Ensembl Bacteria]*

3. Proszę poszukać sekwencje **ncRNA** dla myszy ***Mus musculus*** w bazie Ensembl *[wersja archiwalna 85]*

Ensembl – pobieranie danych

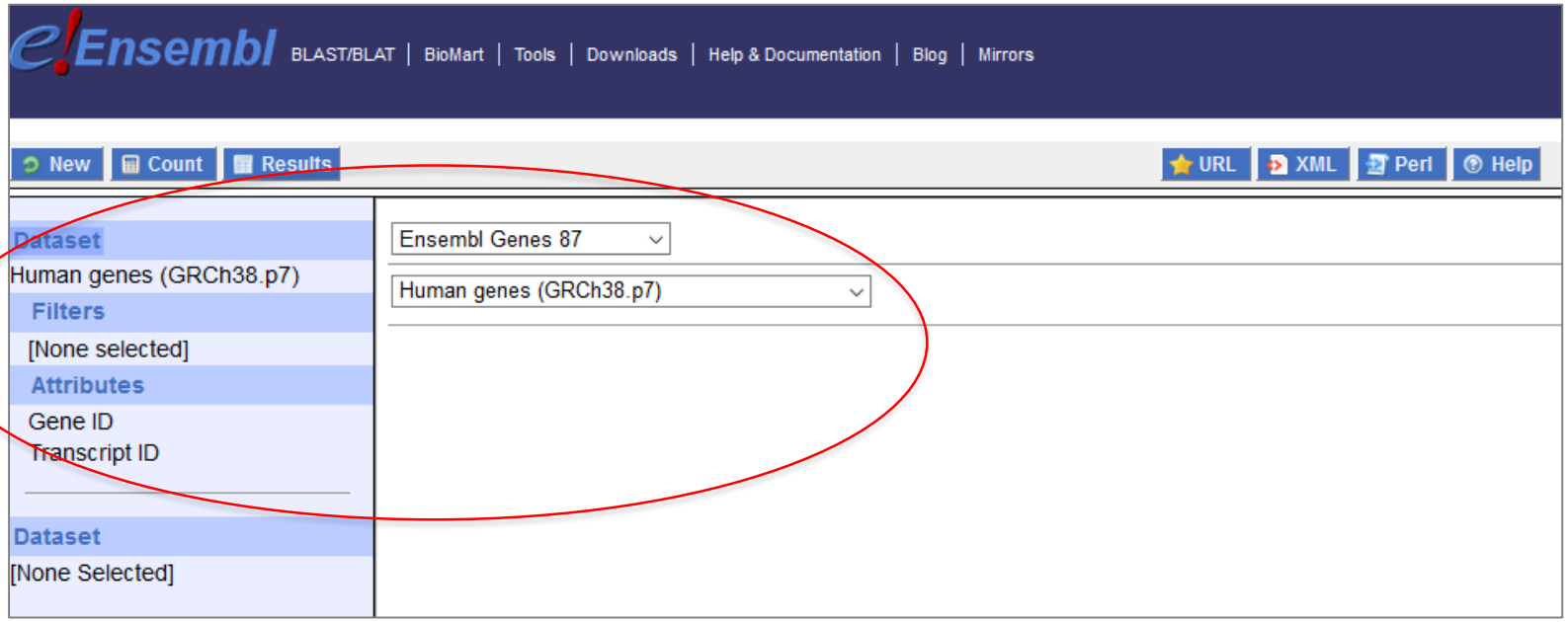
UWAGA! Operujemy nie tylko na identyfikatorach charakterystycznych dla Ensembl

Ensembl BioMart



Ensembl – pobieranie danych

BioMart



- **Dataset** → wybieramy wersję bazy Ensembl oraz organizm
- **Filters** → możliwość zawężania przeszukiwania do określonych cech, identyfikatorów, zbiorów danych itp.
- **Attributes** → określamy jak mają wyglądać dane wyjściowe, jakie cechy, identyfikatory, informacje są dla nas ważne.
- **Count** → informacja o tym ile rekordów o określonych parametrach znaleziono w porównaniu z całym zestawem genów dla danego organizmu.
- **Results** → informacja o tym jak wygląda fragment danych wynikowych, opcje zapisu i eksportu danych

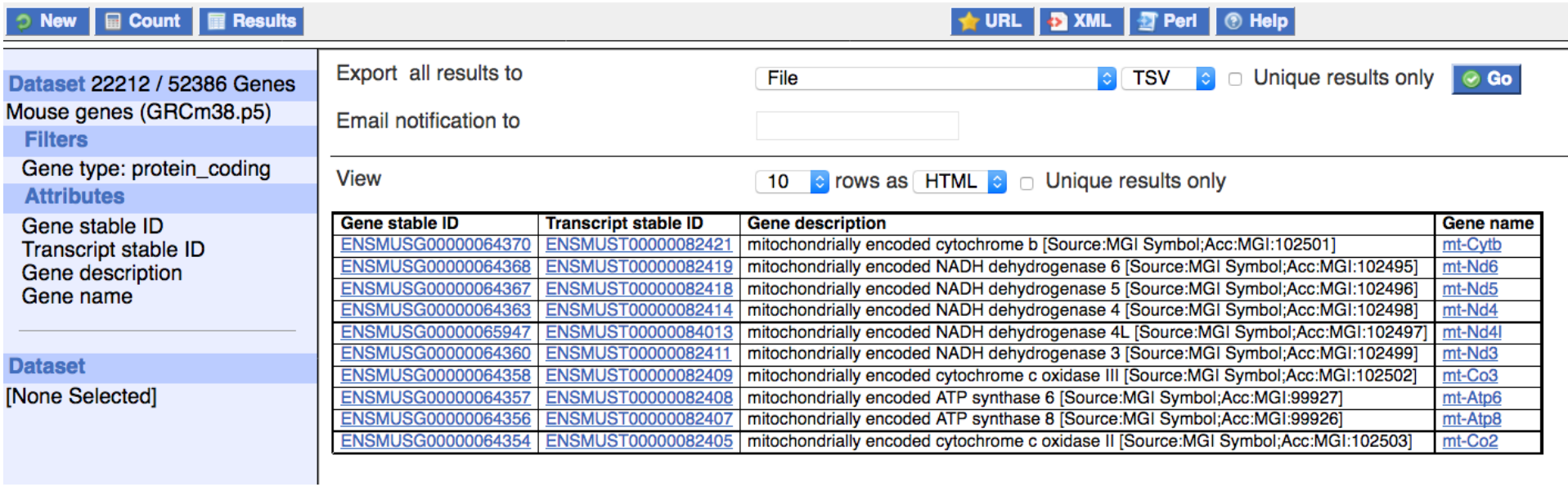
Ensembl – pobieranie danych

BioMart

ĆWICZENIE

Proszę sprawdzić **ile** jest genów kodujących białka u myszy (*Mus musculus*) i **pobrać** listę identyfikatorów Gene ID i Transcript ID, nazwę genu i opis

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



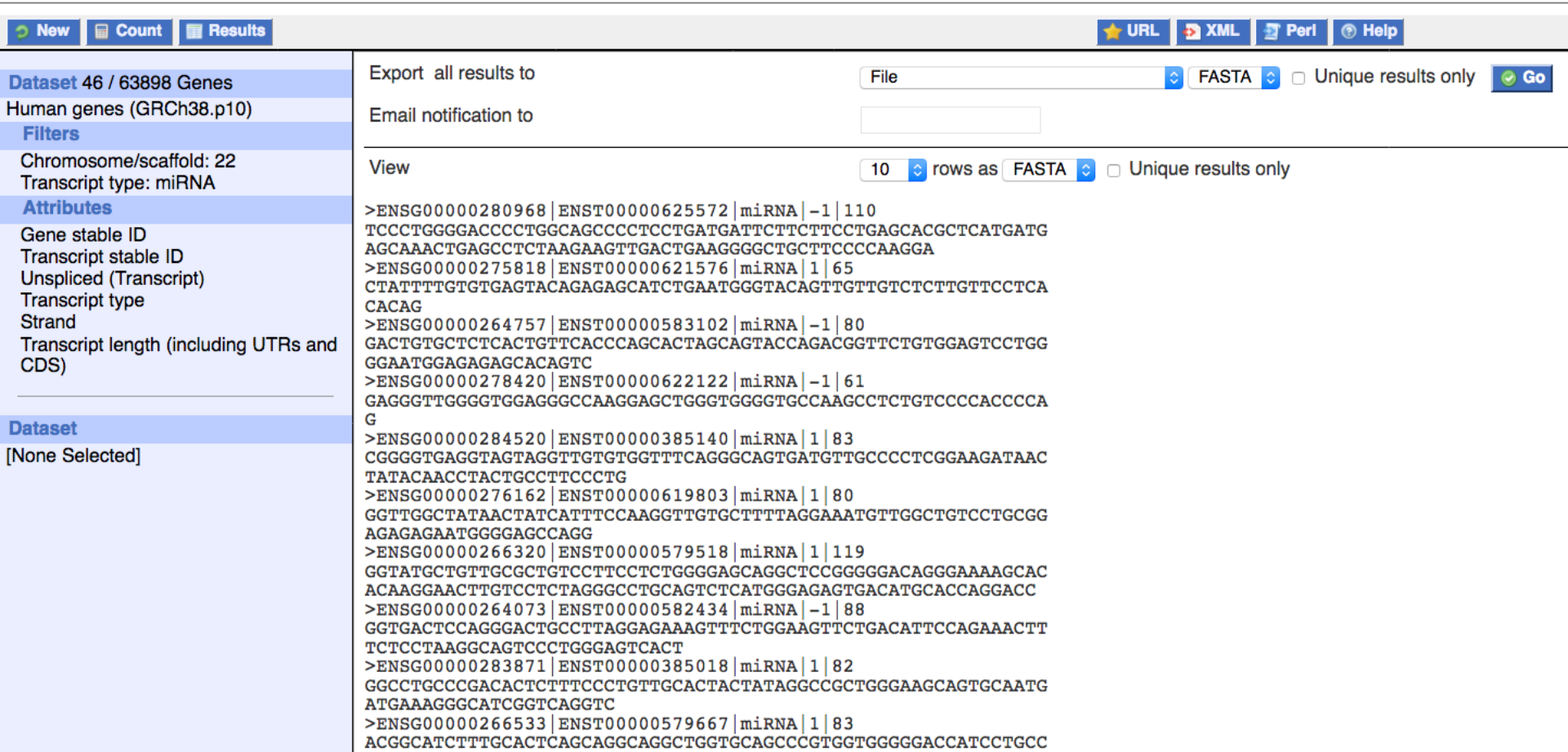
Gene stable ID	Transcript stable ID	Gene description	Gene name
ENSMUSG00000064370	ENSMUST00000082421	mitochondrially encoded cytochrome b [Source:MGI Symbol;Acc:MGI:102501]	mt-Cytb
ENSMUSG00000064368	ENSMUST00000082419	mitochondrially encoded NADH dehydrogenase 6 [Source:MGI Symbol;Acc:MGI:102495]	mt-Nd6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

Ensembl – pobieranie danyc

BioMart

ĆWICZENIE

- Proszę pobrać sekwencje nukleotydowe (Unspliced transcript) miRNA u człowieka (Transcript type) pochodzące z chromosomu 22 wraz z informacjami w linii definicji na temat: typu transkryptu, długości transkryptu i nici (strand).

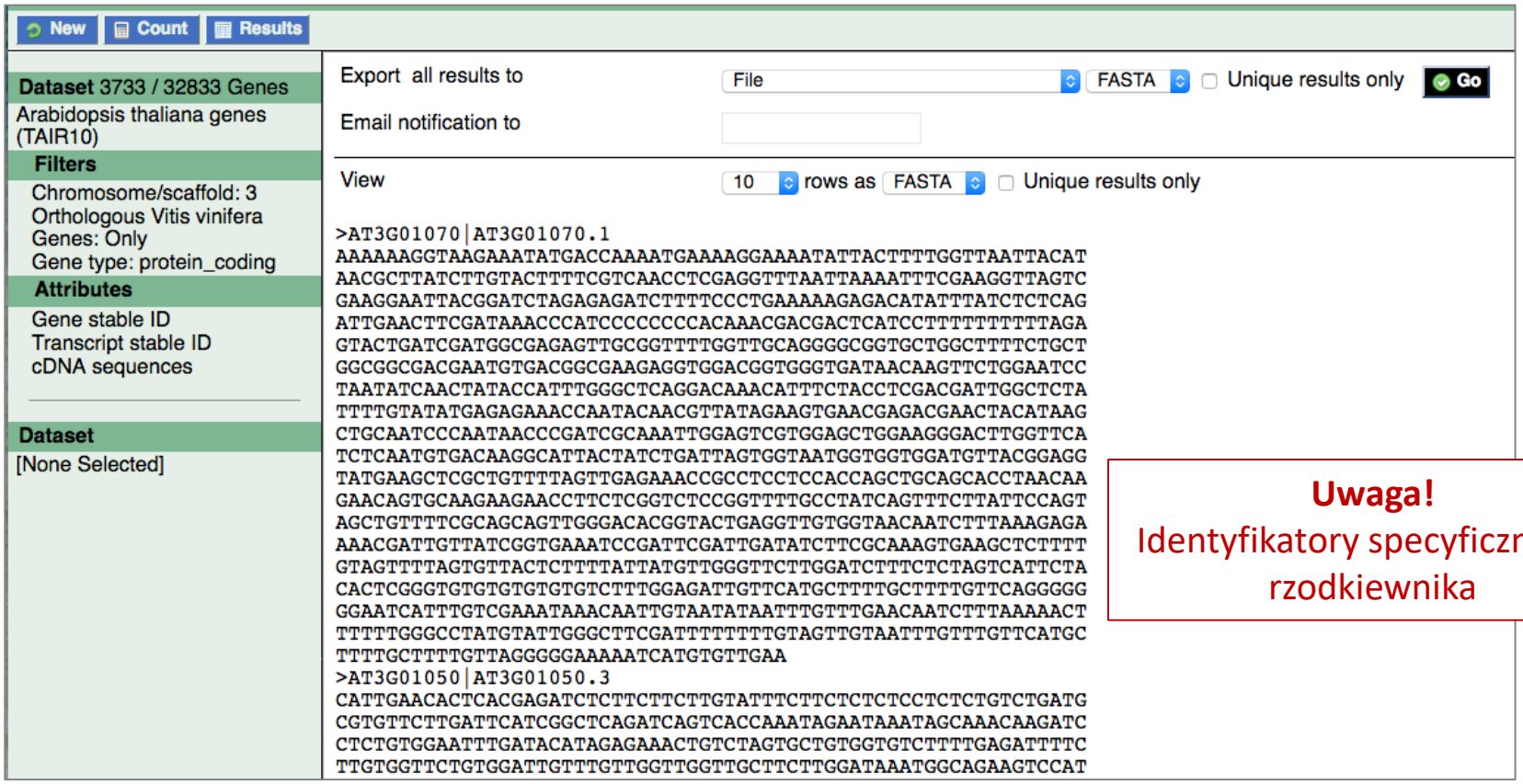


Ensembl – pobieranie danych

BioMart

ĆWICZENIE

Proszę pobrać sekwencje nukleotydowe (cDNA) wszystkich genów kodujących białka u *Arabidopsis thaliana*, które zlokalizowane są na chromosomie 3 i mają swoje ortologiczne geny u *Vitis vinifera.*



Uwaga!
Identyfikatory specyficzne dla rzodkiewnika

Trzy metabazy i przeglądarki

National Center for Biotechnology Information

Ensembl Genome Browser

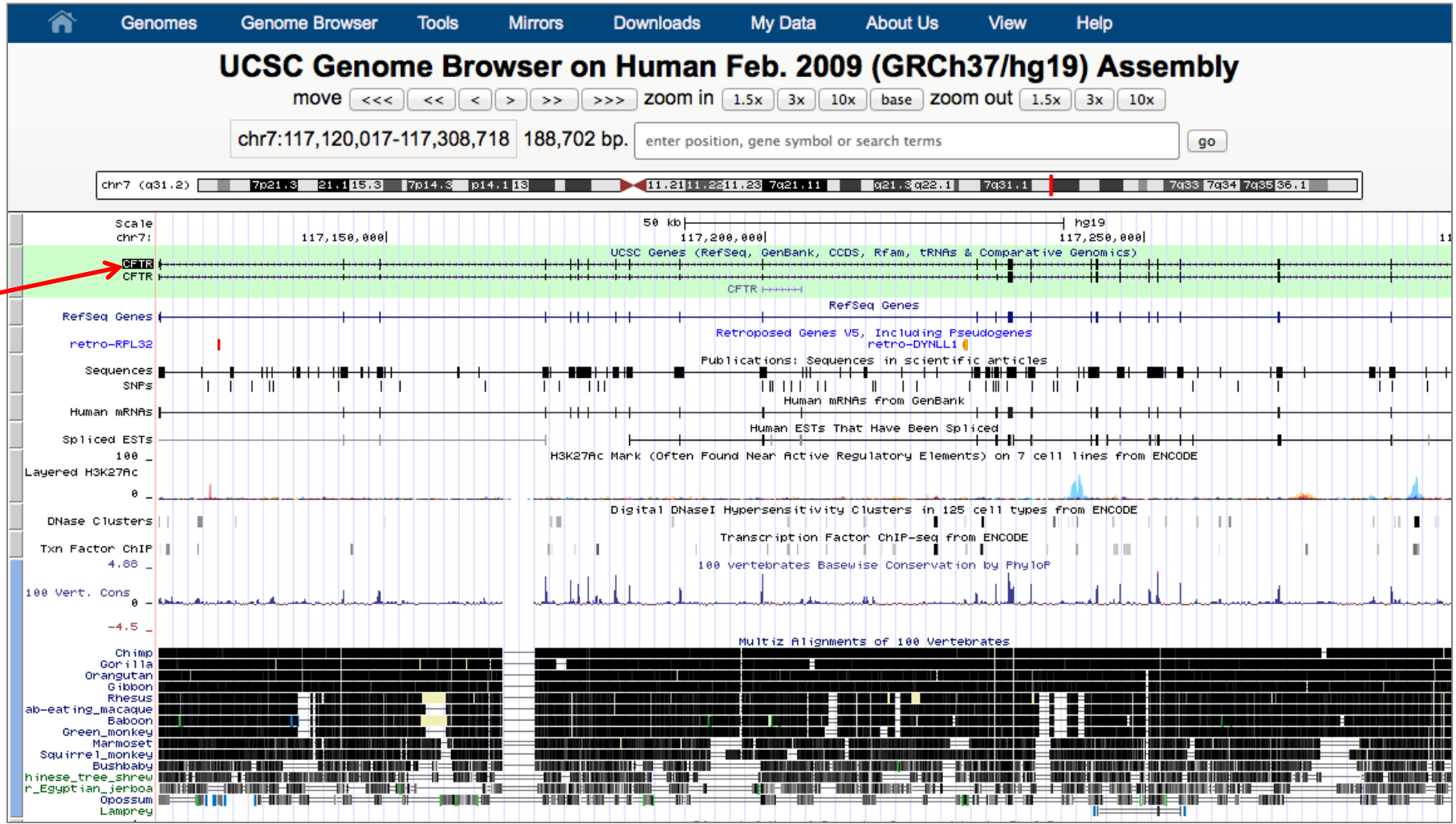
European Bioinformatics Institute, Wellcome Trust Sanger Institute

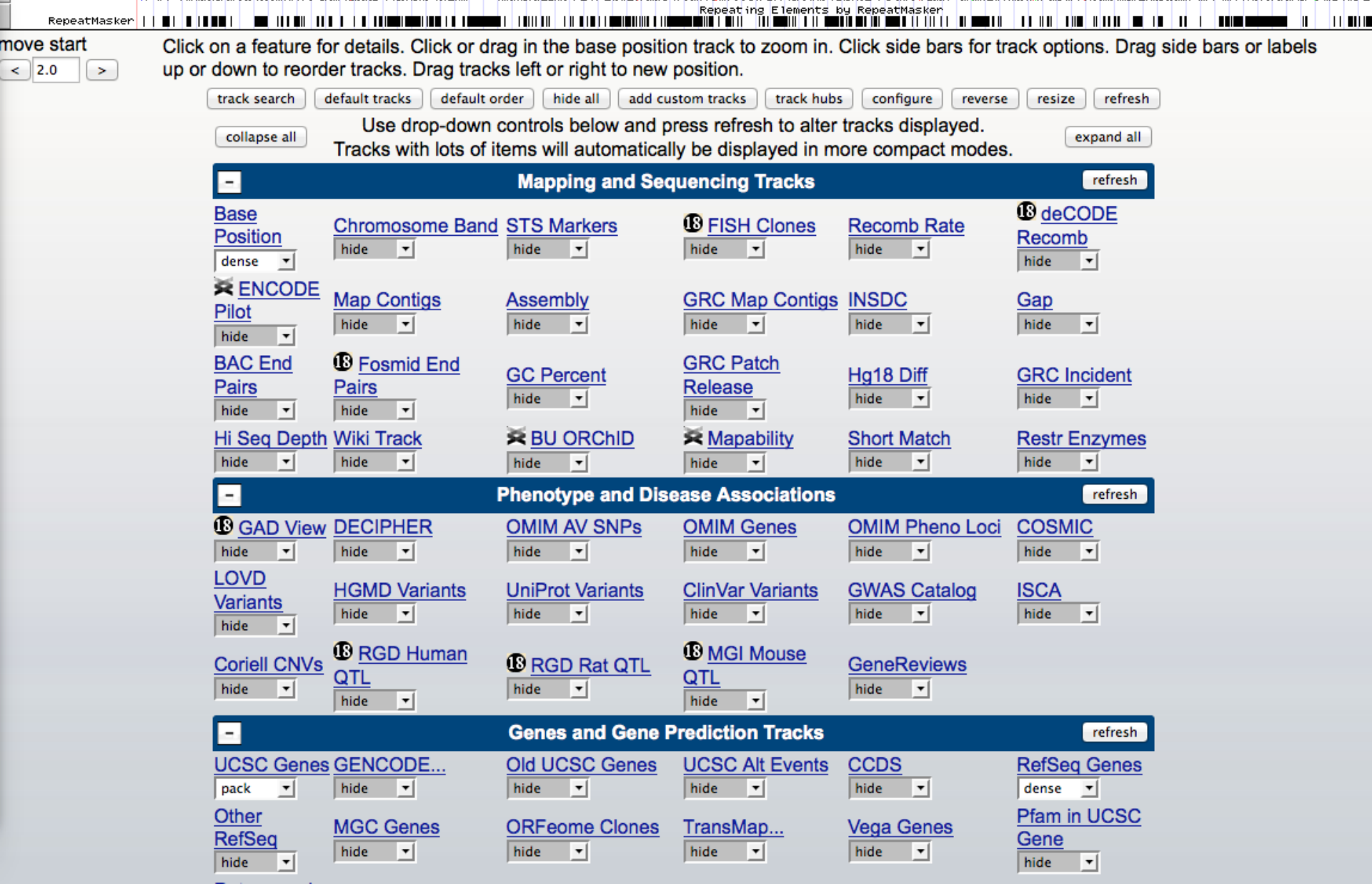
University of California, Santa Cruz Genome Browser

UCSC Genes

UCSC Transcript ID ale bardzo często ID zaczerpnięte jest z innej bazy

CFTR (uc011knq.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





Zmiany ustawień w widoku poszczególnych „ścieżek” w przeglądarce (ang. *tracks*)

UCSC - przeszukiwanie bazy

ĆWICZENIE

Proszę poszukać gen kodujący insulinę **(*INS)* u człowieka** w najnowszej wersji genomu i odpowiedzieć na pytania:

1. Na którym chromosomie jest zlokalizowany?
2. Jakiej długości jest sekwencja genomowa?
3. Ile ma wariantów splicingowych w bazie RefSeq?
4. Ile ma egzonów?
5. Proszę poszukać identyfikator z RefSeq dla sekwencji nt
6. Gdzie ulega najwyższej ekspresji?
7. Czy są jakieś zidentyfikowane elementy powtarzalne w obrębie tego genu?

UCSC – pobieranie danych

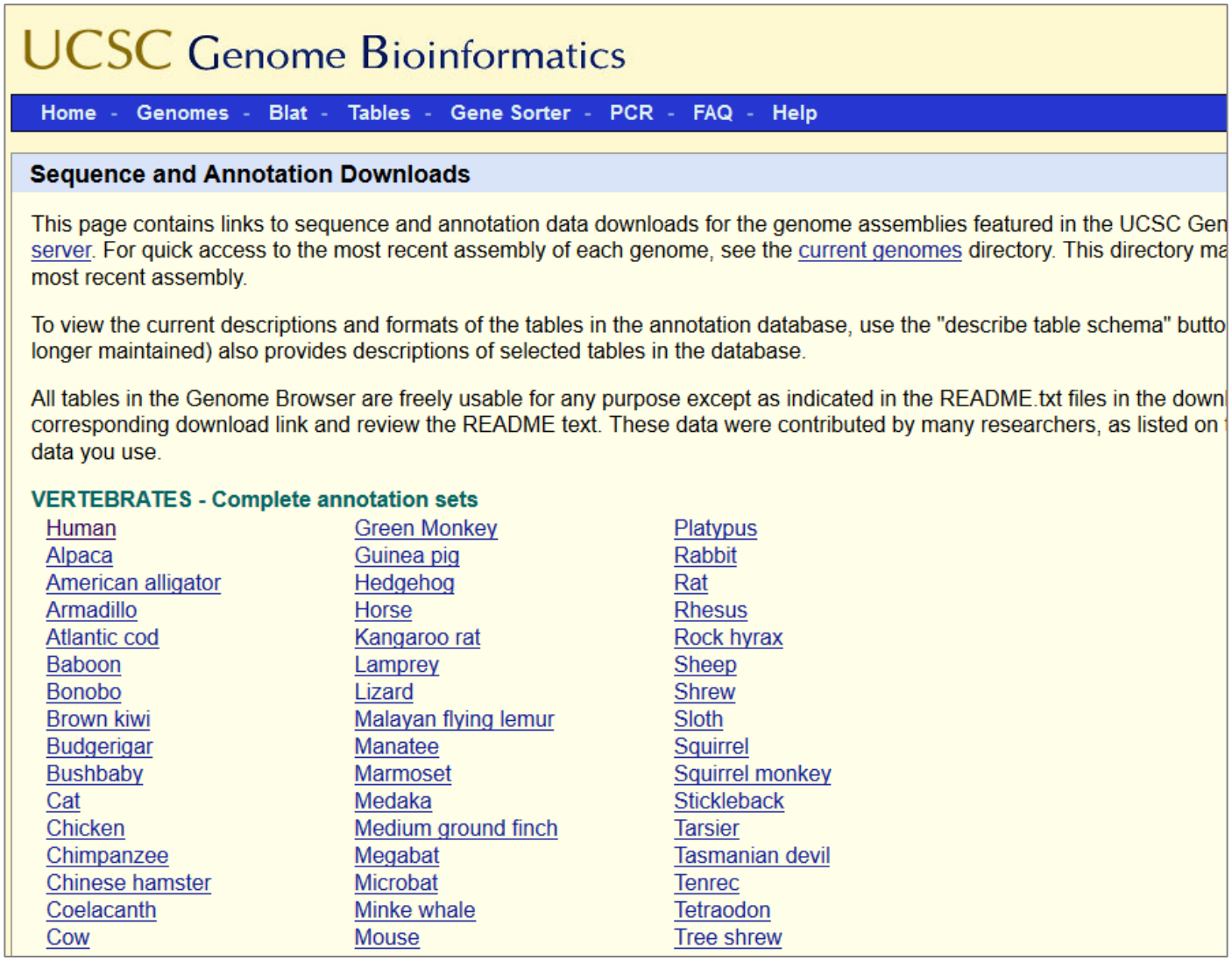
1. Pojedyncze rekordy w bazie danych,
   z poziomu przeglądarki genomowej

2. Download → FTP

3. TableBrowser

UCSC – Pobieranie danych (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 – Pobieranie danych (FTP)

ĆWICZENIE

Proszę **poszukać** zbiór sekwencji mRNA (1) oraz plik z całą sekwencją genomową (2) dla makaka *(Rhesus macaca)* – UWAGA! wersja genomu z października 2010. **[Full dataset]**

- Jak oznaczona jest ta wersja genomu makaka i gdzie go sekwencjonowano?
- Jak nazywają się poszczególne pliki?
- Proszę się zapoznać z opisem danych, nazw plików.

UCSC – Pobieranie danych (FTP)

ĆWICZENIE

Proszę **poszukać** sekwencję genomową chromosomu 22 człowieka *(Homo sapiens)* w formacie FASTA.

UWAGA! Wersja genomu hg38

[Data set by chromosome]

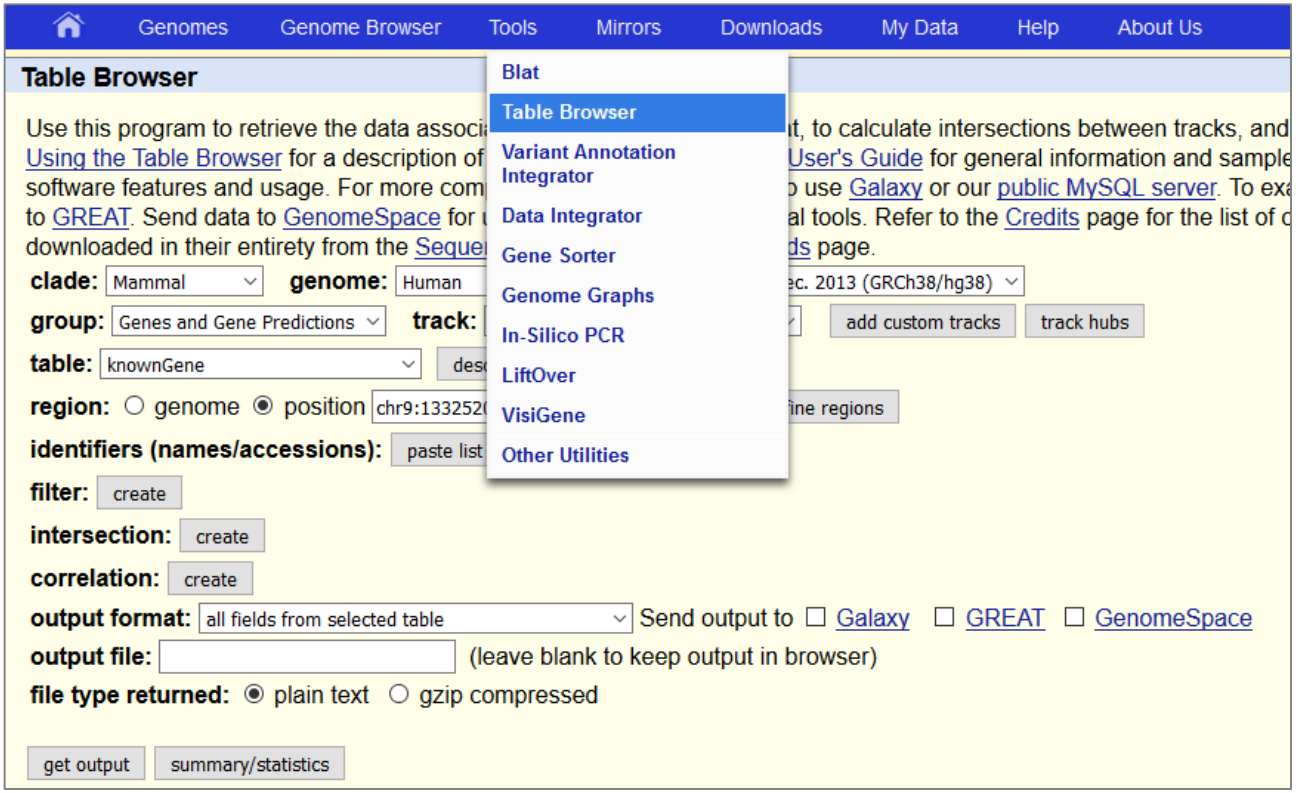
- Jak nazywa się pobrany plik?
- Skąd pochodzą inne nazwy plików?
- Proszę się zapoznać z opisem danych, nazw plików.

UCSC – pobieranie danych

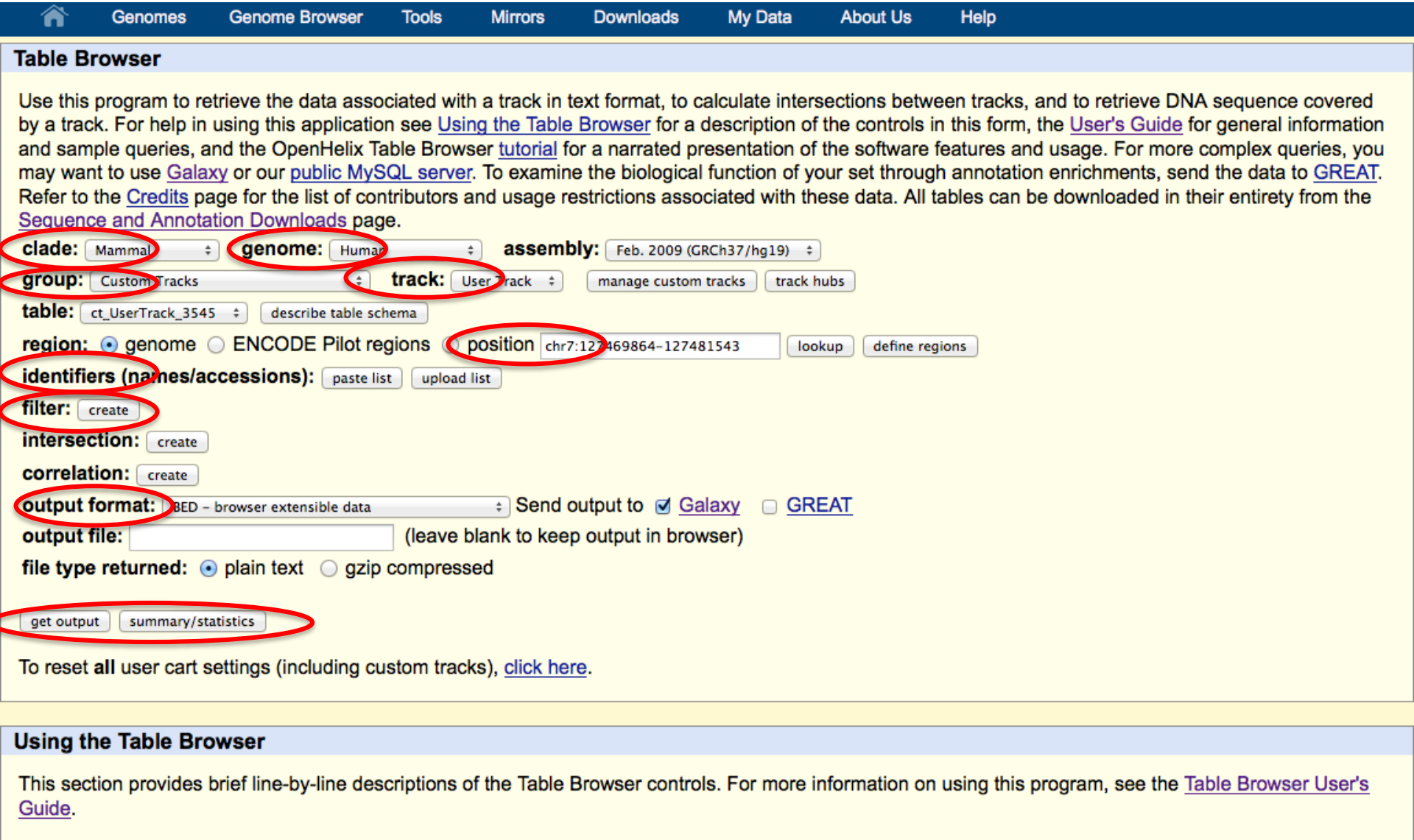
UWAGA! Operujemy nie tylko na identyfikatorach charakterystycznych dla UCSC

UCSC Table Browser

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



UCSC – Pobieranie danych
Table Browser

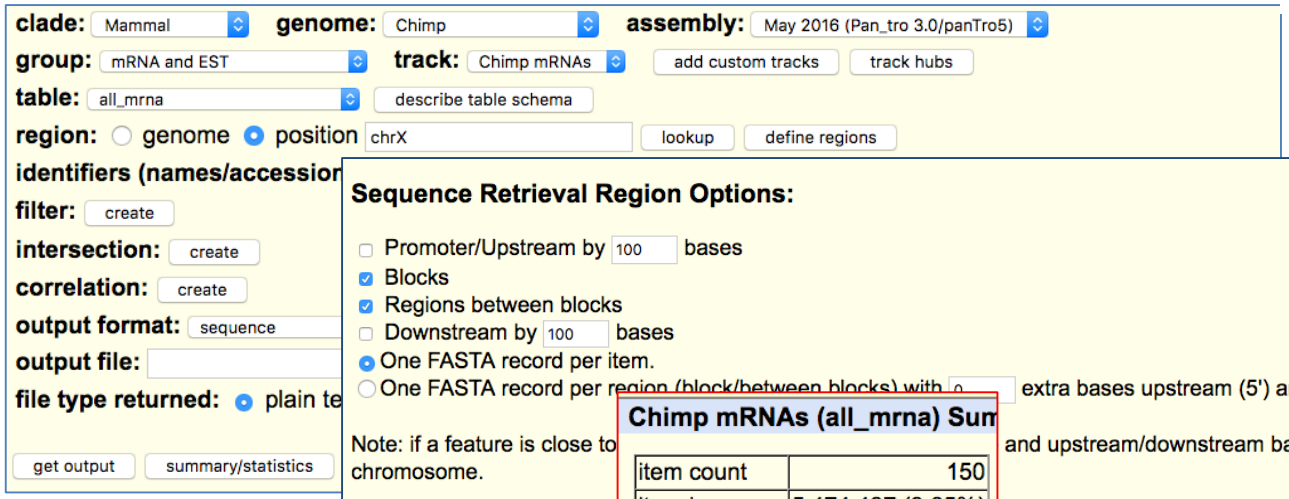


UCSC – Pobieranie danych Table Browser

ĆWICZENIE

Proszę pobrać sekwencje mRNA szympansa w formacie **FASTA**, które pochodzą z chromosomu X.

Ile jest tych sekwencji?



Chimp mRNAs (all_mrna) Sum

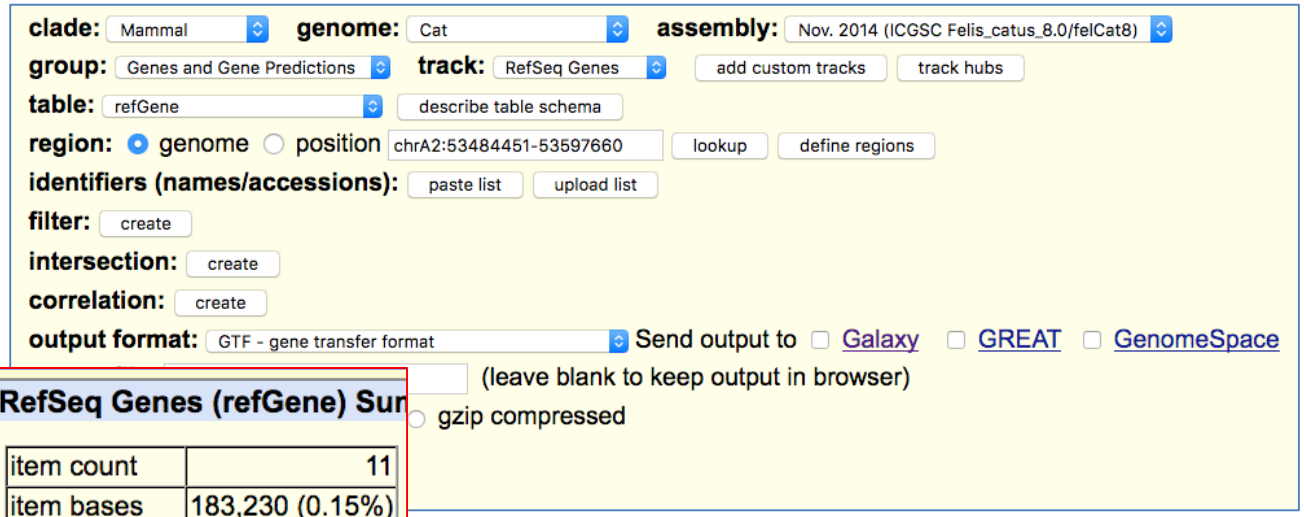
item count	150
item bases	5,474,437 (3.65%)
item total	5,552,068 (3.70%)
smallest item	20
average item	37,014
biggest item	1,157,324
block count	746
block bases	138,497 (0.09%)
block total	153,985 (0.10%)
smallest block	4
average block	206
biggest block	3,148

```
>panTro5_all_mrna_AK305659 range=chrX:1147656-1177541 5'pad=0 3'pad=0 strand=-
GGGTCTCAGGTTTGAGGTGGTCCCCTCCAAGTTTAAAGAGAAGCTGGACA
AAGCCTCCTTCGCTACTCCGTATGGGTACGCCATGGAGACCGCCAAGCAG
AAGGCCCTGGAGGTGGCCAACCGGATGCACCAGgcaaggggcccctgctc
tttggcccttatgaaaattcctttgctgtgatgaggggatttgcaggcag
taaatatacatgatacaaataaacatcattcagaagatgcagttggtggc
acagcttagctattgctatgttataatgtaatgcactattgcggtaacct
aataataccattaatgatacagaattaatgtagcttatctgttactacgt
gttataatattctcttgcaattgatacagtatcattgatgatatagaatt
aatgtagcttatctattactatgtgttataatattctcttgcaattgatt
tcatagtatcattcatgatatagaattaatgtagcttatctgttactatg
ttataatattctcacaattgatacagtatcattgatgatacagaattaat
gtagttatctattactatgcgttgtaatattctcttgcaattgatacagt
atcattgatgatatacaattaatgtagcttatctgttactatgtgttata
atattctcacaattgatatcatagtatcattgatgatatagaattaatgt
agttatctattactatgtgttataatattctcttgcaattgatatagtat
cattgatgatatacaattaatgtagcttatctgttactatgtgttataat
attctcacaattgatatcacagtatcattgatgatatagaattaatgtag
```

UCSC – Pobieranie danych
Table Browser

ĆWICZENIE

Proszę pobrać informacje w formacie **GTF** o genach z chromosomu X z bazy RefSeq u kota.



RefSeq Genes (refGene) Sur

item count	11
item bases	183,230 (0.15%)
item total	183,230 (0.15%)
smallest item	
average item	
biggest item	
block count	
block bases	1
block total	1
smallest block	
average block	
biggest block	

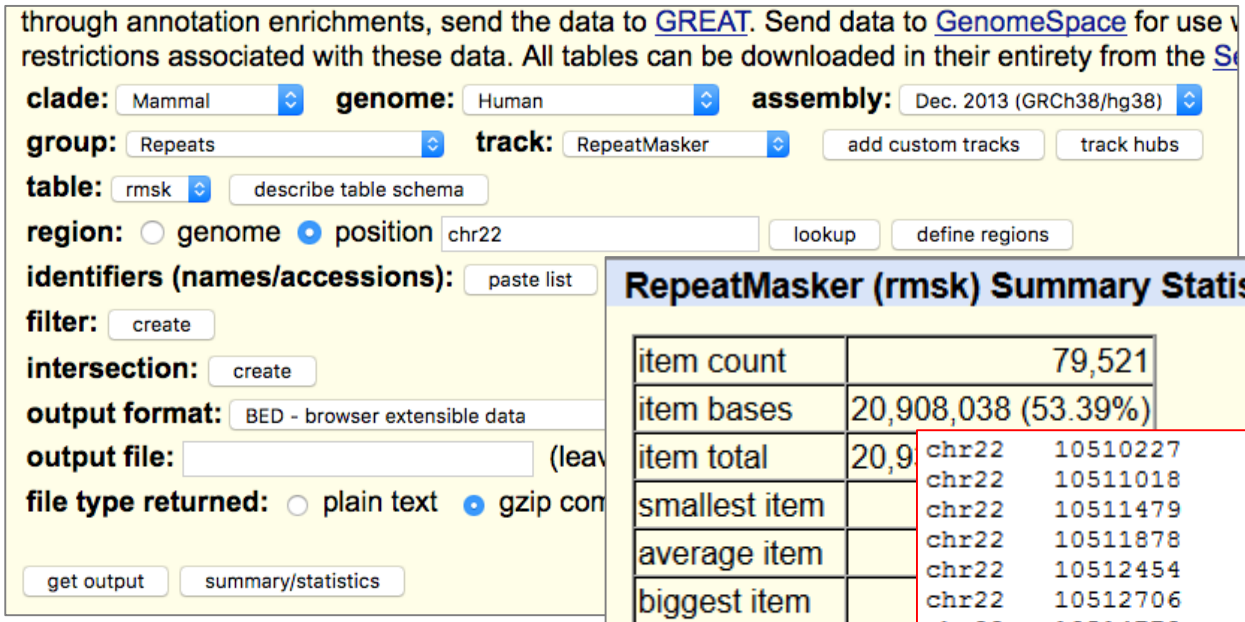
```
chrX    felCat8_refGene exon    21883079        21883169        0.000000        +       .       gene_id "NM_001009262"; transcript_id "NM_001009262";
chrX    felCat8_refGene start_codon     21890310        21890312        0.000000        +       .       gene_id "NM_001009262"; transcript_id "NM_001009262";
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 – Pobieranie danych
Table Browser

ĆWICZENIE

Proszę pobrać wszystkie elementy powtarzalne (repeats) w formacie **BED** dla człowieka pochodzące z chromosomu 22

(UWAGA! najnowsza wersja genomu)



RepeatMasker (rmsk) Summary Statistics

item count	79,521
item bases	20,908,038 (53.39%)
item total	20,9
smallest item	
average item	
biggest item	
smallest score	
average score	
biggest score	

```
chr22   10510227   10510528   AluSx1    2021   +
chr22   10511018   10511332   L1MC5a    781    -
chr22   10511479   10511791   L1MB1     524    +
chr22   10511878   10512212   L1MB1     313    +
chr22   10512454   10512692   L1MB1     656    +
chr22   10512706   10514778   L1MB1     11092  +
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    -
chr22   10516287   10516630   L1MB1     1504   +
```

Inne bazy danych sekwencji i adnotacji

Genomes Online Database

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

Wellcome Trust Sanger Institute

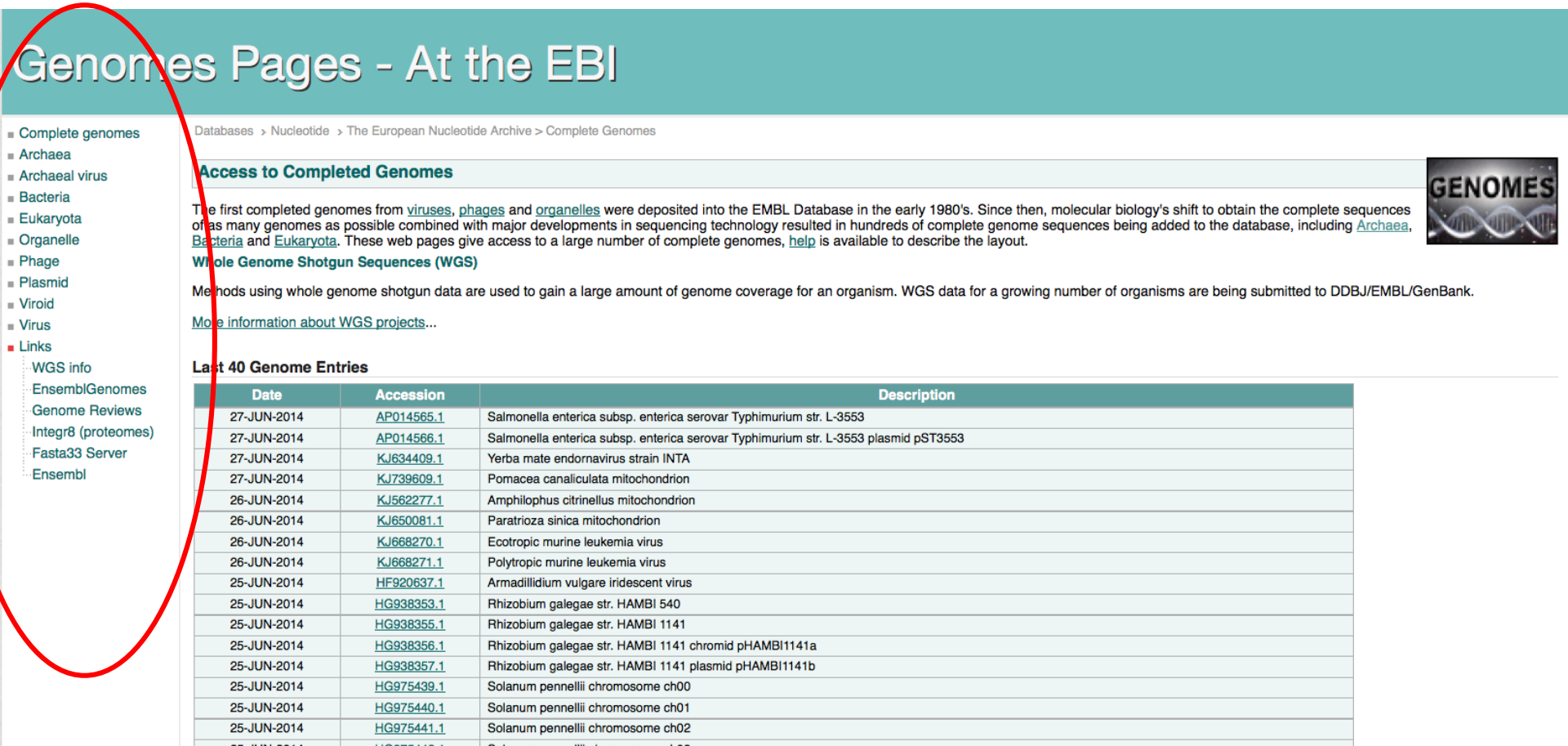
https://www.sanger.ac.uk/resources/downloads/

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

Inne bazy danych sekwencji i adnotacji

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
  - WGS info
  - EnsemblGenomes
  - Genome Reviews
  - Integr8 (proteomes)
  - Fasta33 Server
  - Ensembl

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...

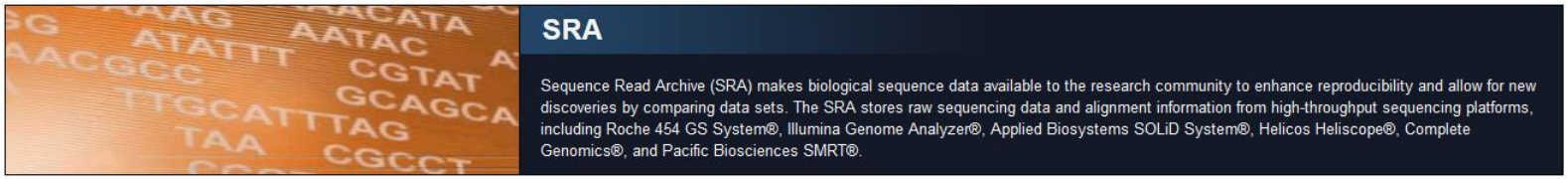
Last 40 Genome Entries

Date	Accession	Description
27-JUN-2014	AP014565.1	Salmonella enterica subsp. enterica serovar Typhimurium str. L-3553
27-JUN-2014	AP014566.1	Salmonella enterica subsp. enterica serovar Typhimurium str. L-3553 plasmid pST3553
27-JUN-2014	KJ634409.1	Yerba mate endornavirus strain INTA
27-JUN-2014	KJ739609.1	Pomacea canaliculata mitochondrion
26-JUN-2014	KJ562277.1	Amphilophus citrinellus mitochondrion
26-JUN-2014	KJ650081.1	Paratrioza sinica mitochondrion
26-JUN-2014	KJ668270.1	Ecotropic murine leukemia virus
26-JUN-2014	KJ668271.1	Polytropic murine leukemia virus
25-JUN-2014	HF920637.1	Armadillidium vulgare iridescent virus
25-JUN-2014	HG938353.1	Rhizobium galegae str. HAMBI 540
25-JUN-2014	HG938355.1	Rhizobium galegae str. HAMBI 1141
25-JUN-2014	HG938356.1	Rhizobium galegae str. HAMBI 1141 chromid pHAMBI1141a
25-JUN-2014	HG938357.1	Rhizobium galegae str. HAMBI 1141 plasmid pHAMBI1141b
25-JUN-2014	HG975439.1	Solanum pennellii chromosome ch00
25-JUN-2014	HG975440.1	Solanum pennellii chromosome ch01
25-JUN-2014	HG975441.1	Solanum pennellii chromosome ch02

Przegląd podstawowych źródeł danych

dotyczących projektów sekwencjonowania nowej generacji

SRA NCBI
ENA EBI
GEO NCBI



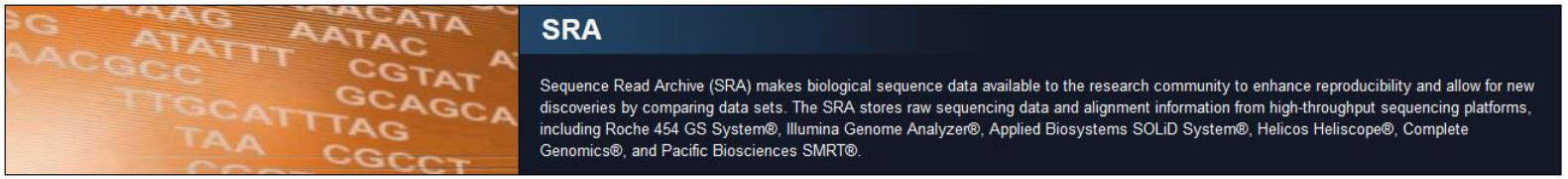
Sequence Read Archive (SRA)

- **Główne źródło danych** z sekwencjonowania nowej generacji
- Wygodny w użyciu sposób pobierania oraz przeszukiwania danych
- Możliwość wysyłania własnych danych
- Dane przechowywane w formacie SRA

Sequence Read Archive przechowuje dane z wielu platform:
454, IonTorrent, Illumina, SOLiD, Helicos, Complete Genomics,
Pacific Biosciences oraz Oxford Nanopore Technologies.

SRA przechowuje również przyrównania w postaci
pozycji odczytów na sekwencji referencyjnej a także dane z bazy ENA.

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



Sequence Read Archive (SRA)

Study/Project

Ogólny cel badań / projektu, może zawierać wiele eksperymentów.

Numery akcesyjne zaczynać się mogą od **ERP** lub **SRP**

Experiment

Szczegółowa informacja co i jaką metodą zostało zsekwencjonowane.

Informacja o bibliotece, platformie i metodzie obróbki danych (data processing)

Numery akcesyjne zaczynają się od liter **SRX**

Run

Wyniki lub odczyty z poszczególnych uruchomień platformy do sekwencjonowania

Numery akcesyjne zaczynają się od liter **SRR**

Sample

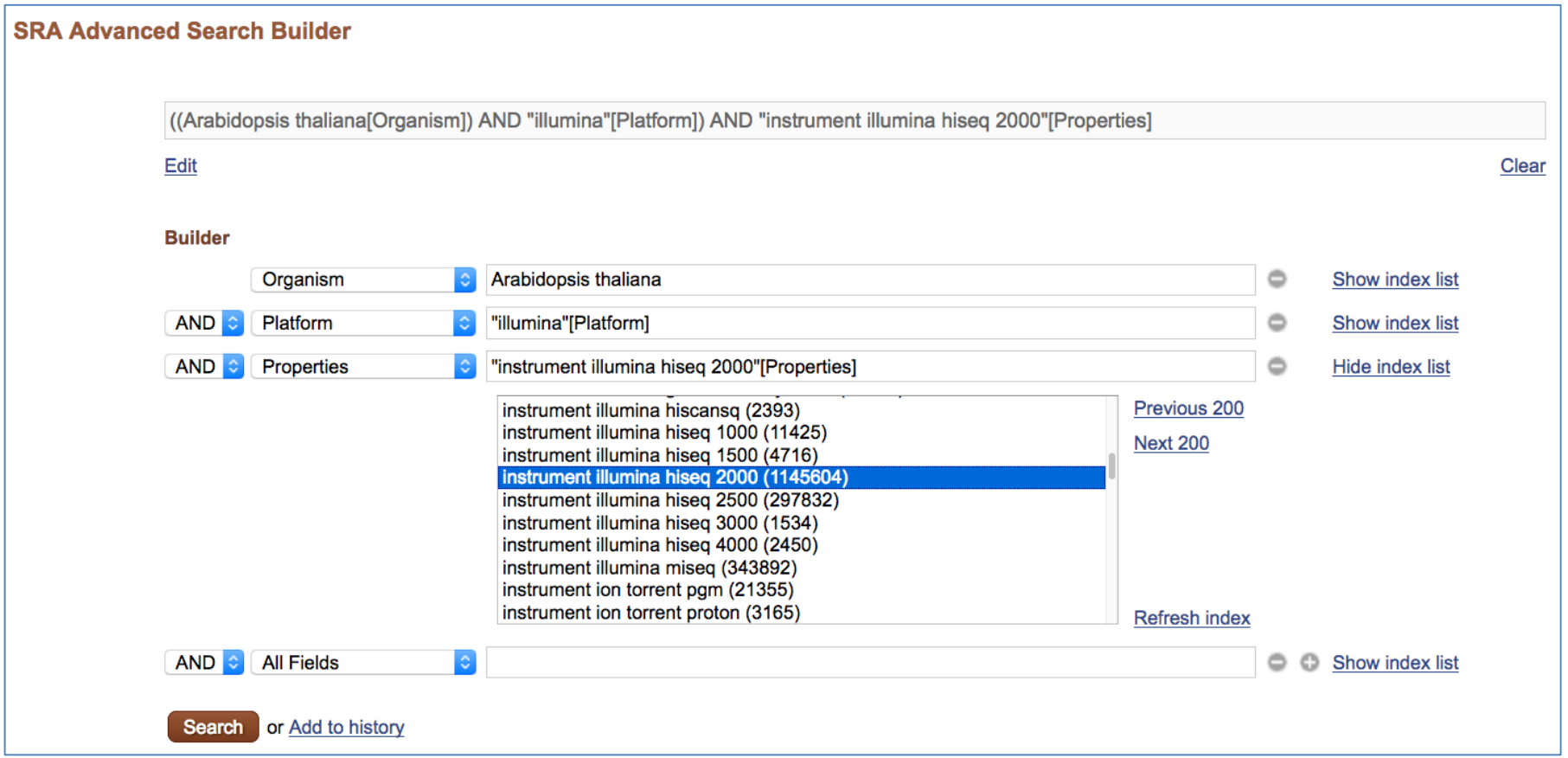
Opis próbki która była sekwencjonowana

Numery akcesyjne zaczynają się od liter **SRS**

SRA: search builder

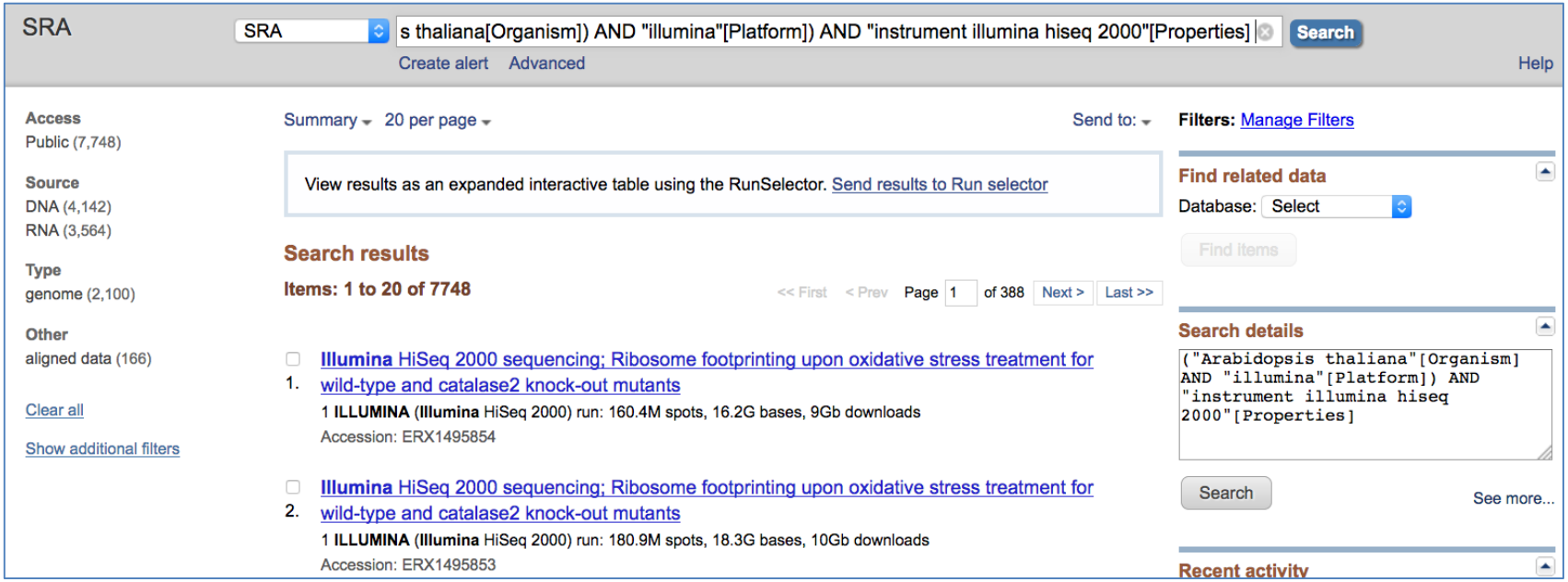
http://www.ncbi.nlm.nih.gov/sra/

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

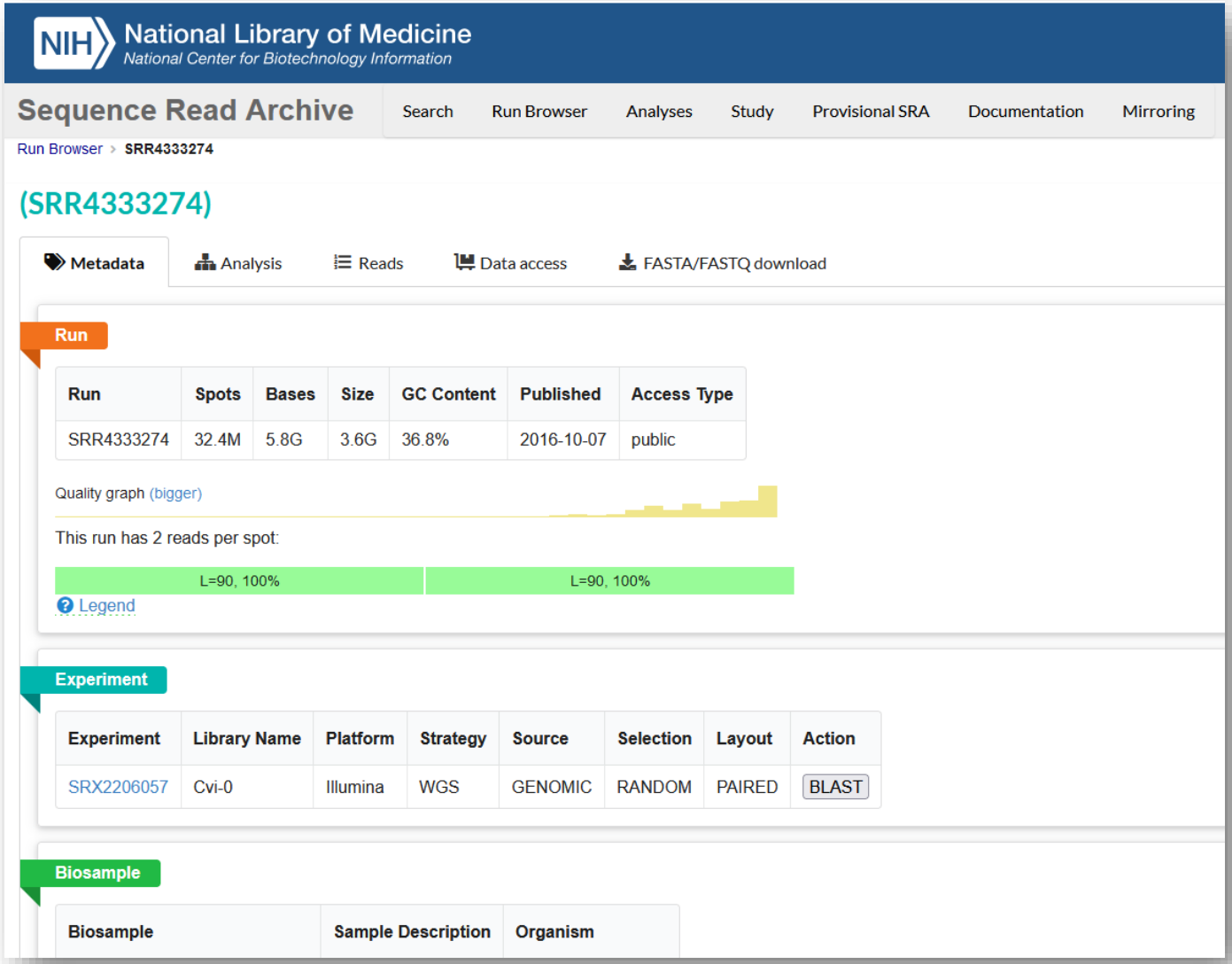


SRA: search builder

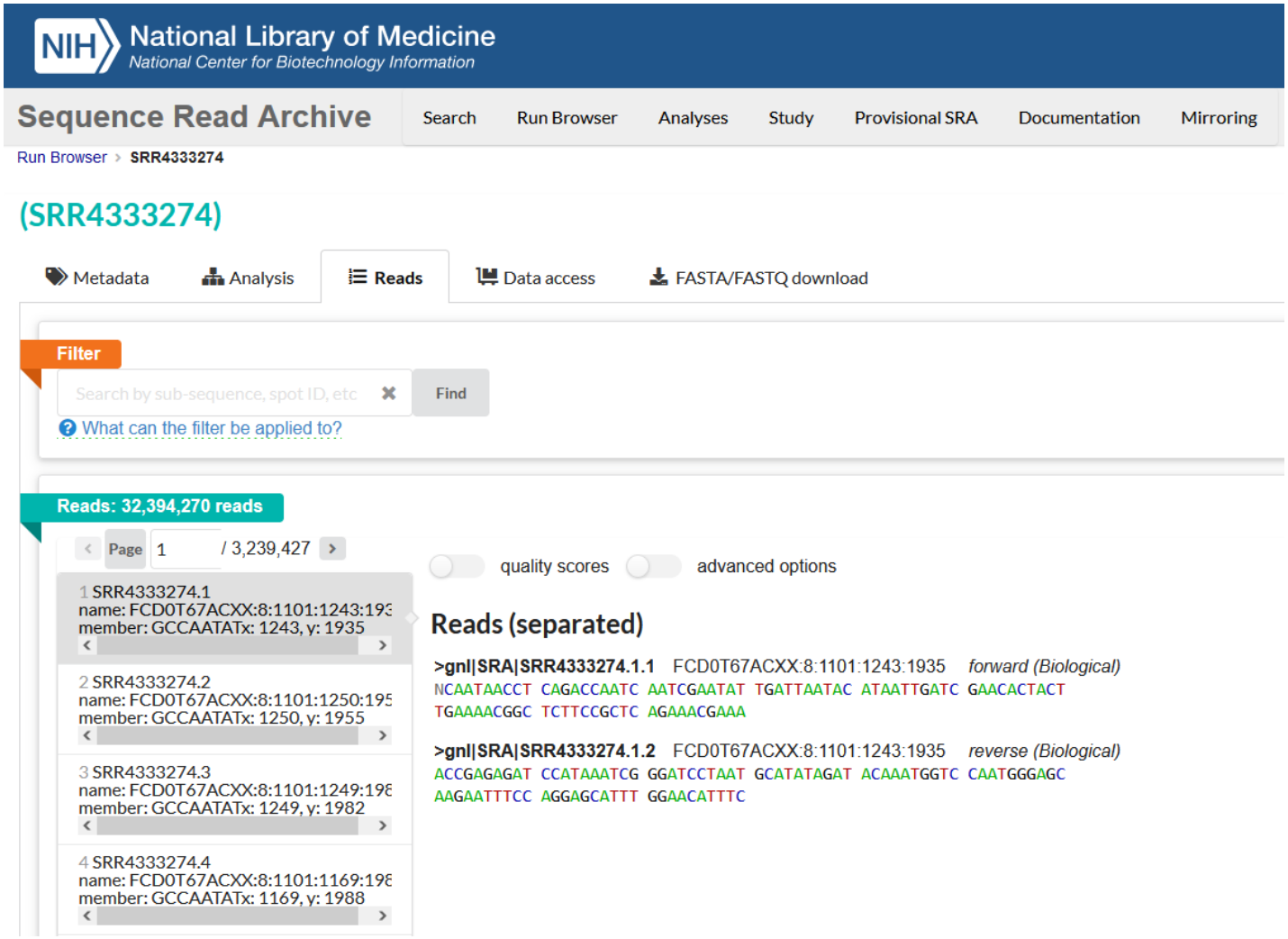
http://www.ncbi.nlm.nih.gov/sra/



SRA: Run


National Library of Medicine
National Center for Biotechnology Information
Sequence Read Archive
Search
Run Browser
Analyses
Study
Provisional SRA
Documentation
Mirroring
Run Browser › SRR4333274
(SRR4333274)
Metadata
Analysis
Reads
Data access
FASTA/FASTQ download
Run
Run | Spots | Bases | Size | GC Content | Published | Access Type
SRR4333274 | 32.4M | 5.8G | 3.6G | 36.8% | 2016-10-07 | public
Quality graph (bigger)
This run has 2 reads per spot:
L=90, 100%
L=90, 100%
Legend
Experiment
Experiment | Library Name | Platform | Strategy | Source | Selection | Layout | Action
SRX2206057 | Cvi-0 | Illumina | WGS | GENOMIC | RANDOM | PAIRED | BLAST
Biosample
Biosample | Sample Description | Organism

SRA: Reads


National Library of Medicine
National Center for Biotechnology Information
Sequence Read Archive
Search
Run Browser
Analyses
Study
Provisional SRA
Documentation
Mirroring
Run Browser › SRR4333274
(SRR4333274)
Metadata
Analysis
Reads
Data access
FASTA/FASTQ download
Filter
Search by sub-sequence, spot ID, etc
Find
What can the filter be applied to?
Reads: 32,394,270 reads
Page 1 / 3,239,427
quality scores
advanced options
1 SRR4333274.1
name: FCD0T67ACXX:8:1101:1243:193
member: GCCAATATx: 1243, y: 1935
2 SRR4333274.2
name: FCD0T67ACXX:8:1101:1250:195
member: GCCAATATx: 1250, y: 1955
3 SRR4333274.3
name: FCD0T67ACXX:8:1101:1249:198
member: GCCAATATx: 1249, y: 1982
4 SRR4333274.4
name: FCD0T67ACXX:8:1101:1169:198
member: GCCAATATx: 1169, y: 1988
Reads (separated)
>gnl|SRA|SRR4333274.1.1 FCD0T67ACXX:8:1101:1243:1935 forward (Biological)
NCAATAACCT CAGACCAATC AATCGAATAT TGATTAATAC ATAATTGATC GAACACTACT
TGAAAACGGC TCTTCCGCTC AGAAACGAAA
>gnl|SRA|SRR4333274.1.2 FCD0T67ACXX:8:1101:1243:1935 reverse (Biological)
ACCGAGAGAT CCATAAATCG GGATCCTAAT GCATATAGAT ACAAATGGTC CAATGGGAGC
AAGAATTTCC AGGAGCATTT GGAACATTTC

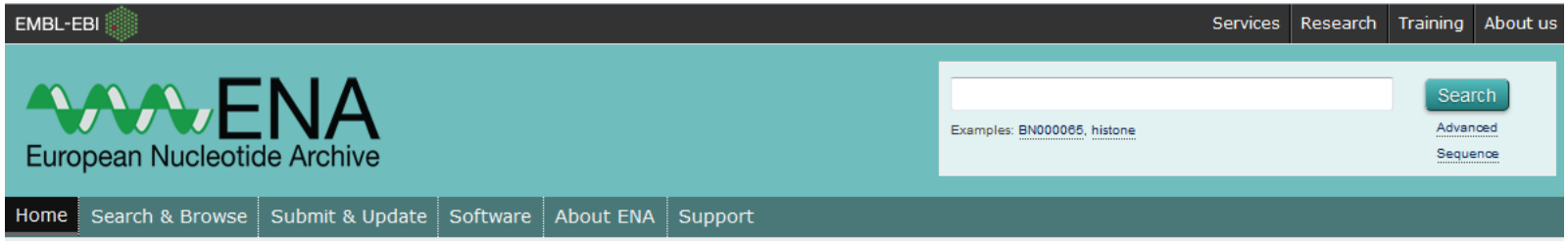
SRA – NCBI
ĆWICZENIE

- Proszę poszukać w bazie SRA danych dla człowieka z sekwencjonowania transkryptomu (Properties→ strategy rna-seq), gdzie wynikiem są odczyty sparowane (Layout → paired)

- Dla jednego wybranego rekordu proszę poszukać identyfikatory Study, Experiment, Run, Sample oraz krótki opis eksperymentu.

ENA
European Nucleotide Archive

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



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 though 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

- Submit and update
- Sequence submissions
- Genome assembly submissions
- Submitting environmental sequences
- Citing ENA data
- Rest URLs for data retrieval
- 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.

ENA
European Nucleotide Archive

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

Identyfikatory – numery akcesyjne

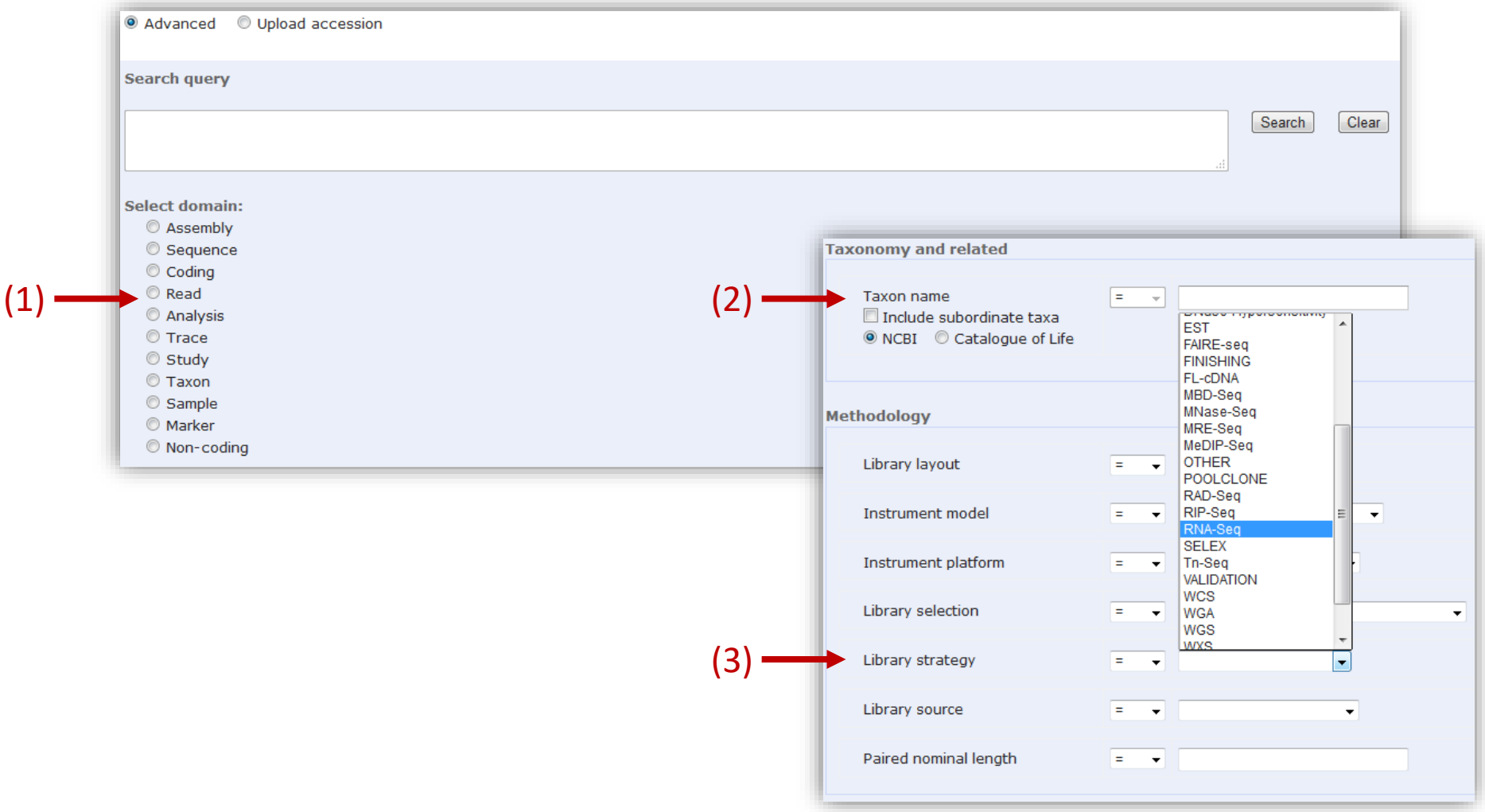
Metadata object	Accession prefix	Number of digits	Example
Submission	ERA, SRA, DRA	6	ERA000092
Sample	ERS, SRS, DRS	6	ERS000081
Study	ERP, SRP, DRP	6	ERP000016
Experiment	ERX, SRX, DRX	6	ERX000398
Run	ERR, SRR, DRR	6	ERR003990
Analysis	ERZ, SRZ, DRZ	6	ERZ000001
Project/Study	PRJ		

ER - EBI
SR – SRA NCBI
DR - DDBJ

European Nucleotide Archive

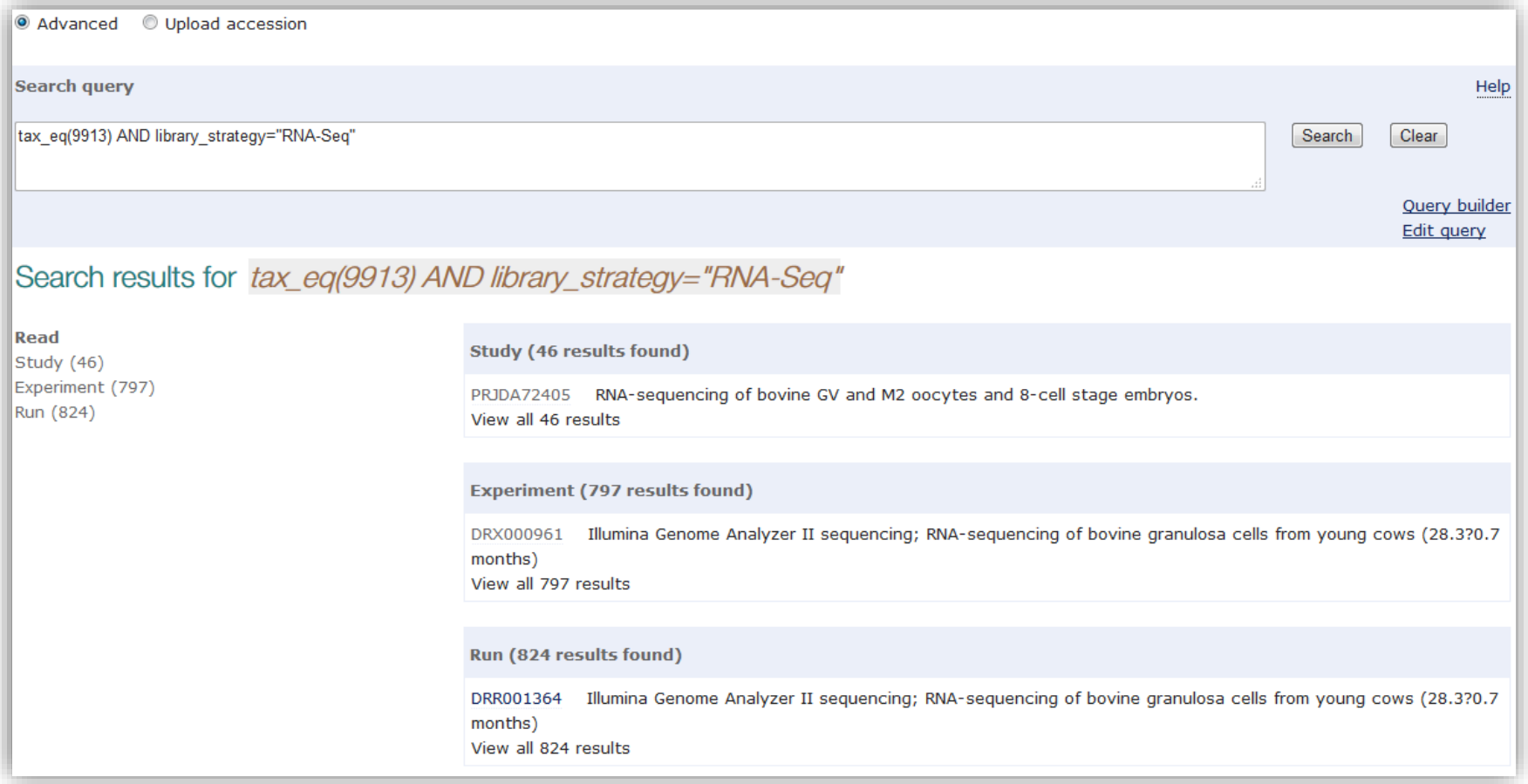
http://www.ebi.ac.uk/ena

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		

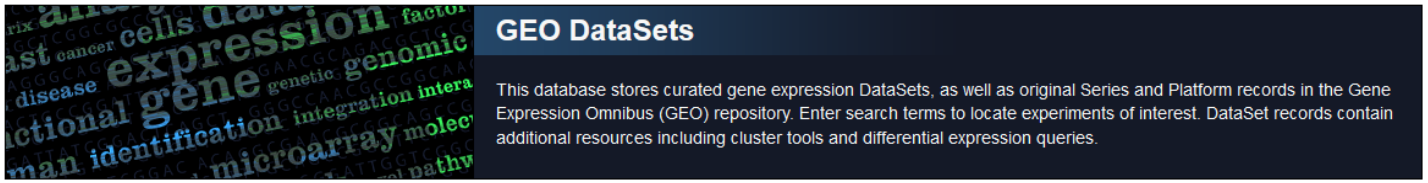
ENA – EBI
ĆWICZENIE

- Proszę poszukać w bazie ENA danych dla człowieka, z sekwencjonowania transkryptomu (Library strategy → RNA-seq), gdzie wynikiem są odczyty sparowane (Library layout → paired)

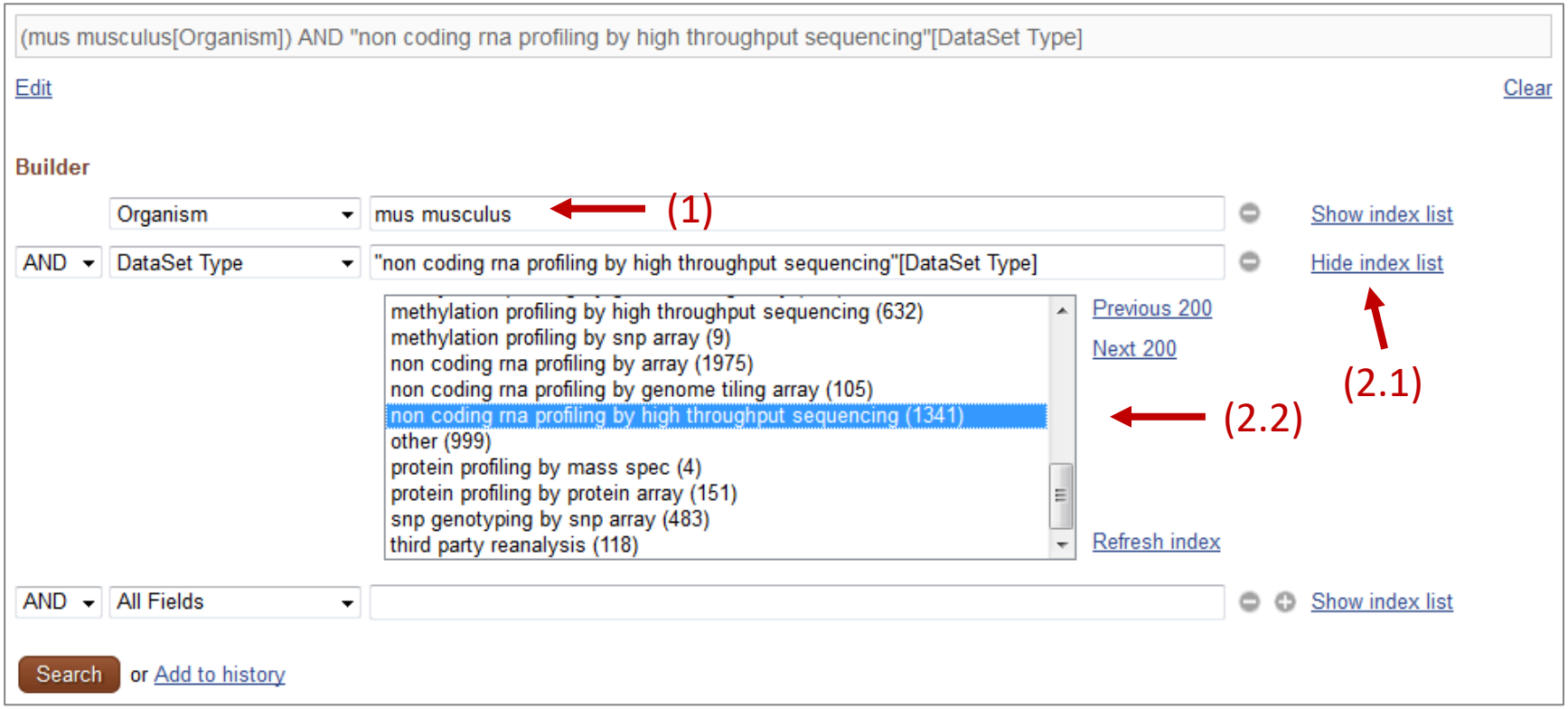
- Dla jednego wybranego rekordu (eksperymentu) proszę poszukać identyfikatory Study, Experiment, Run, Sample oraz krótki opis eksperymentu.

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)**

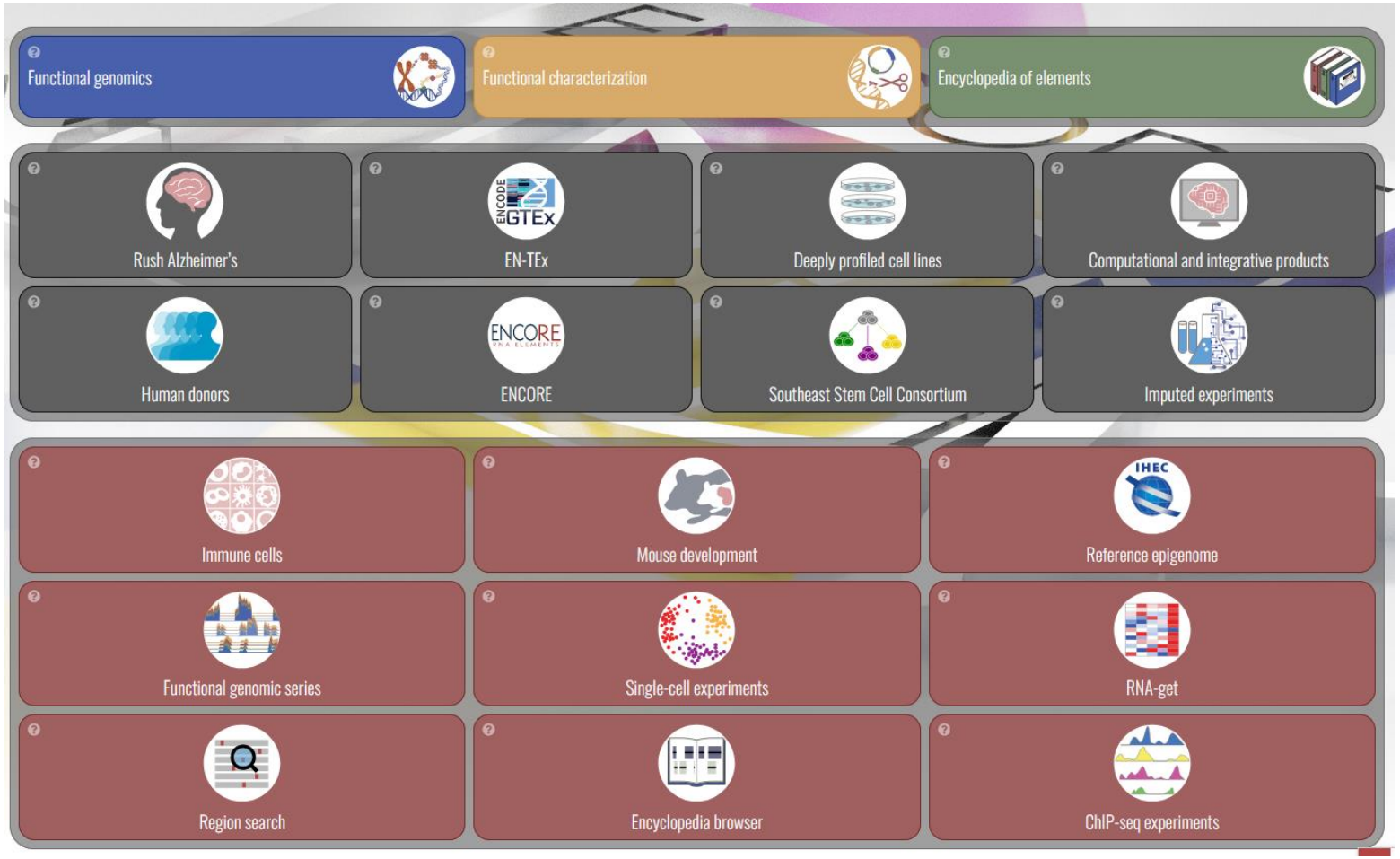


Inne bazy danych

- ENCODE
  https://www.encodeproject.org/
- The Cancer Genome Atlas:
  https://tcga-data.nci.nih.gov/tcga/
- Expression Atlas
  https://www.ebi.ac.uk/gxa/home
- GENT2:
  http://gent2.appex.kr/gent2/
- SCDevDB
  https://scdevdb.deepomics.org/

ENCODE

https://www.encodeproject.org/



RNA-Seq Blog

http://www.rna-seqblog.com/

RNA-seqlopedia https://rnaseq.uoregon.edu

RNA-seqlopedia

RNA-seq produces millions of sequences from complex RNA samples. With this powerful approach, you can:

1. Measure gene expression.
2. Discover and annotate complete transcripts.
3. Characterize alternative splicing and polyadenylation.

The RNA-seqlopedia provides an overview of RNA-seq and of the choices necessary to carry out a successful RNA-seq experiment.

RNA-seqlopedia is written by the Cresko Lab of the University of Oregon and was funded by grant R24 RR032670 (NIH, National Center for Research Resources). Credits.

SEQanswers http://seqanswers.com/

Genohub

https://genohub.com

Technical Guides & Resources

Beginner's Handbook of Next Generation Sequencing
Designing Your Sequencing Runs
Sequencing Coverage Guide
NGS Library Preparation Kit Guide
NGS Instrument Guide
Best Practices for Shipping RNA/DNA/Tissue for NGS
Cluster Density Optimization
External Next-Generation Sequencing Resources

Application Guides

Whole Genome Sequencing (WGS) and Re-sequencing
Whole Exome Sequencing
Whole Transcriptome and mRNA Sequencing
Shotgun Metagenomics Sequencing
Small RNA (miRNA) Sequencing
Whole Genome Sequencing on HiSeq X Ten
Sequencing Applications on HiSeq 3000, HiSeq 4000

1. Find and select matching services (optional):

Sequencing & library prep | Sequencing only | Library prep only

Project type: DNA (whole genome)

Number of samples: 1

Match by coverage | Match by number of reads | Help me

Targeted genome size: 3,200,000,000

Minimum coverage per sample: e.g. 30

Find Services

2. Submit your request:

We typically respond within 2 business hours (US Central Time)

First name | Last name | Your email

Tell us about your project

Submit Request

Dziękuję za uwagę