Financováno Evropskou unií
NextGenerationEU

Národní plán obnovy

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Analýza sekvenčních dat

Vratislav Peška – vpeska@ibp.cz

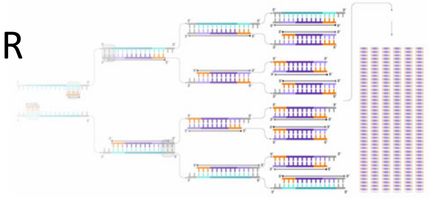
NPO-4 Pokročilé metody v genomice a
proteomice, NPO_MUNI_MSMT-16606/2022

CEITEC
Central European Institute of Technology
BRNO | CZECH REPUBLIC

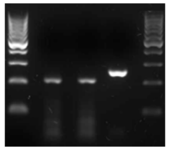
MUNI
SCI

Sangerova metoda sekvenace - výsledky

PCR



ELFO

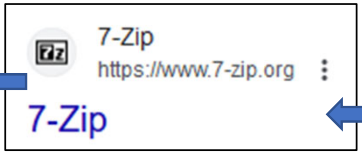


Klonování do vektoru

Sekvenační servis

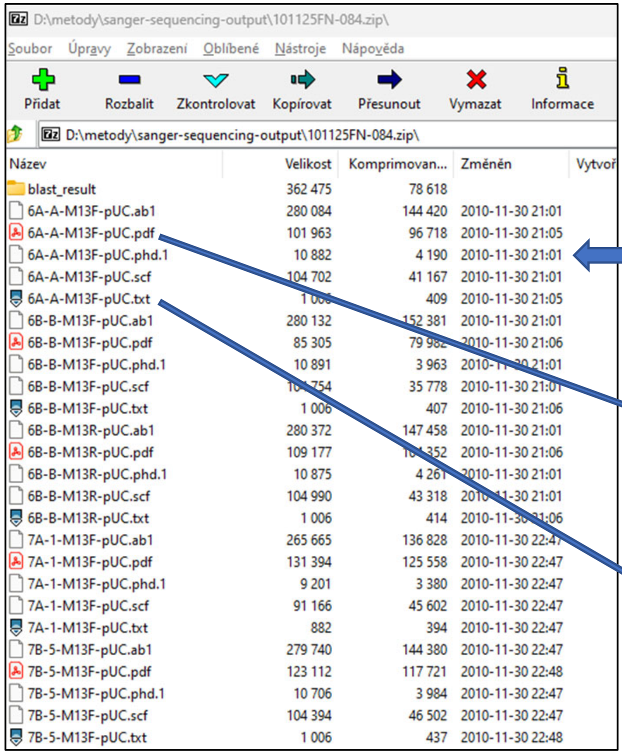
Tento počítač › DATA (D:) › metody › sanger-sequencing-output

Název	Datum změny	Typ	Velikost
101125FN-084.zip	02.12.2010 14:05	Soubor ZIP	4 177 kB



7-Zip
https://www.7-zip.org

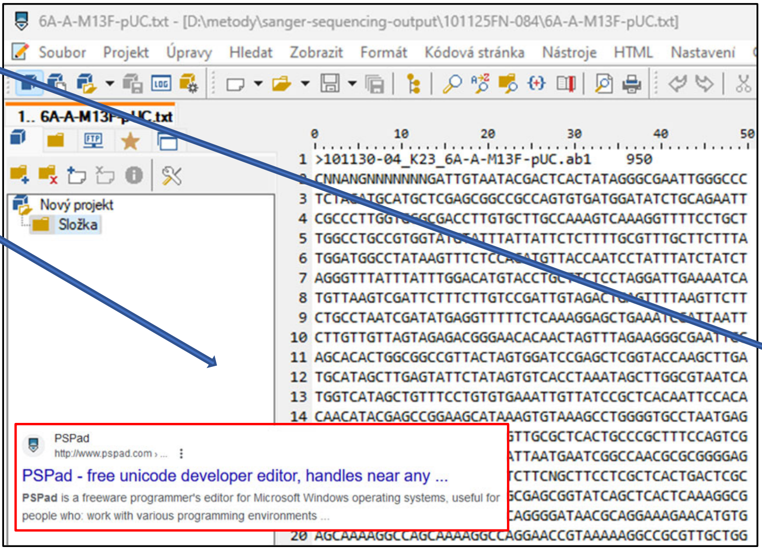
7-Zip



D:\metody\sanger-sequencing-output\101125FN-084.zip\

D:\metody\sanger-sequencing-output\101125FN-084.zip\

Název	Velikost	Komprimovan...	Změněn	Vytvoř
blast_result	362 475	78 618		
6A-A-M13F-pUC.ab1	280 084	144 420	2010-11-30 21:01	
6A-A-M13F-pUC.pdf	101 963	96 718	2010-11-30 21:05	
6A-A-M13F-pUC.phd.1	10 882	4 190	2010-11-30 21:01	
6A-A-M13F-pUC.scf	104 702	41 167	2010-11-30 21:01	
6A-A-M13F-pUC.txt	1 00[illegible]	409	2010-11-30 21:05	
6B-B-M13F-pUC.ab1	280 132	152 381	2010-11-30 21:01	
6B-B-M13F-pUC.pdf	85 305	79 98[illegible]	2010-11-30 21:06	
6B-B-M13F-pUC.phd.1	10 891	3 963	2010-1[illegible]0 21:01	
6B-B-M13F-pUC.scf	10[illegible]754	35 778	2010-11-30 21:01	
6B-B-M13F-pUC.txt	1 006	407	2010-11-30 21:06	
6B-B-M13R-pUC.ab1	280 372	147 458	2010-11-30 21:01	
6B-B-M13R-pUC.pdf	109 177	10[illegible]352	2010-11-30 21:06	
6B-B-M13R-pUC.phd.1	10 875	4 261	2010-11-30 21:01	
6B-B-M13R-pUC.scf	104 990	43 318	2010[illegible]1-30 21:01	
6B-B-M13R-pUC.txt	1 006	414	2010-11-3[illegible]1:06	
7A-1-M13F-pUC.ab1	265 665	136 828	2010-11-30 22:47	
7A-1-M13F-pUC.pdf	131 394	125 558	2010-11-30 22:47	
7A-1-M13F-pUC.phd.1	9 201	3 380	2010-11-30 22:47	
7A-1-M13F-pUC.scf	91 166	45 602	2010-11-30 22:47	
7A-1-M13F-pUC.txt	882	394	2010-11-30 22:47	
7B-5-M13F-pUC.ab1	279 740	144 380	2010-11-30 22:47	
7B-5-M13F-pUC.pdf	123 112	117 721	2010-11-30 22:48	
7B-5-M13F-pUC.phd.1	10 706	3 984	2010-11-30 22:47	
7B-5-M13F-pUC.scf	104 394	46 502	2010-11-30 22:47	
7B-5-M13F-pUC.txt	1 006	437	2010-11-30 22:48	



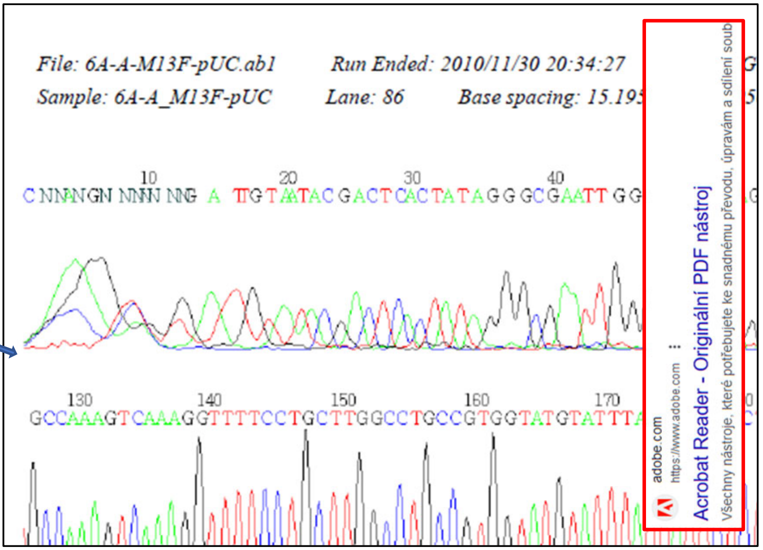
6A-A-M13F-pUC.txt - [D:\metody\sanger-sequencing-output\101125FN-084\6A-A-M13F-pUC.txt]

```
>101130-04_K23_6A-A-M13F-pUC.ab1    950
CNNANGNNNNNNNGATTGTAATACGACTCACTATAGGGCGAATTGGGCCC
TCTA[illegible]ATGCATGCTCGAGCGGCCGCCAGTGTGATGGATATCTGCAGAATT
CGCCCTTGG[illegible]GCGACCTTGTGCTTGCCAAAGTCAAAGGTTTTCCTGCT
TGGCCTGCCGTGGTAT[illegible]ATTTATTATTCTCTTTTGCGTTTGCTTCTTTA
TGGATGGCCTATAAGTTTCTCCA[illegible]ATGTTACCAATCCTATTTATCTATCT
AGGGTTTATTTATTTGGACATGTACCTGC[illegible]CTCCTAGGATTGAAAATCA
TGTTAAGTCGATTCTTTCTTGTCCGATTGTAGACT[illegible]GTTTTAAGTTCTT
CTGCCTAATCGATATGAGGTTTTTCTCAAAGGAGCTGAAATC[illegible]ATTAATT
CTTGTTGTTAGTAGAGACGGGAACACAACTAGTTTAGAAGGGCGAAT[illegible]
AGCACACTGGCGGCCGTTACTAGTGGATCCGAGCTCGGTACCAAGCTTGA
TGCATAGCTTGAGTATTCTATAGTGTCACCTAAATAGCTTGGCGTAATCA
TGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACA
CAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAG
                      GTTGCGCTCACTGCCCGCTTTCCAGTCG
                      ATTAATGAATCGGCCAACGCGCGGGGAG
                      TCTTCNGCTTCCTCGCTCACTGACTCGC
                      GCGAGCGGTATCAGCTCACTCAAAGGCG
                      CAGGGGATAACGCAGGAAAGAACATGTG
AGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGG
```

PSPad
http://www.pspad.com › ...

PSPad - free unicode developer editor, handles near any ...

PSPad is a freeware programmer's editor for Microsoft Windows operating systems, useful for people who: work with various programming environments ...



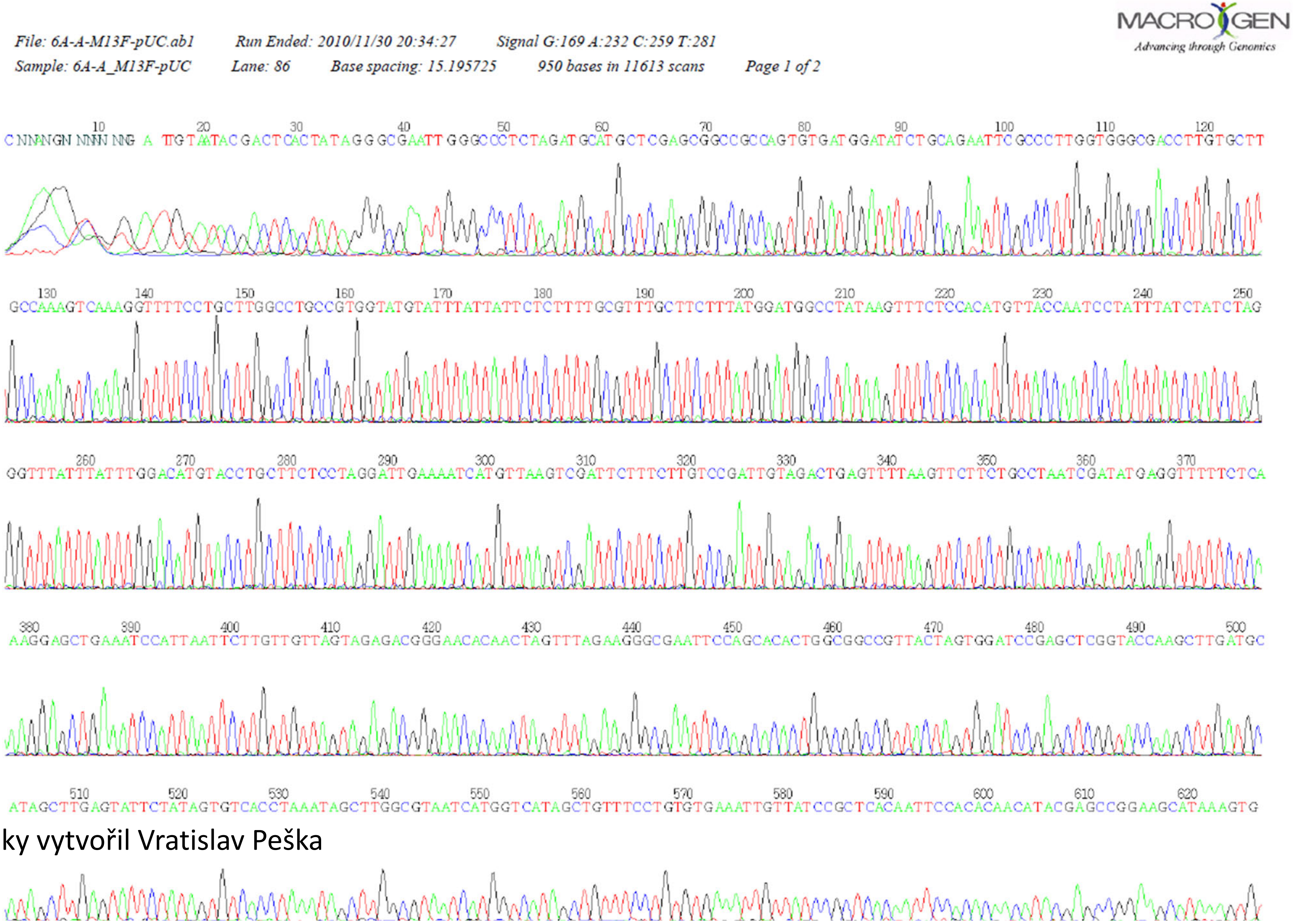
File: 6A-A-M13F-pUC.ab1 *Run Ended: 2010/11/30 20:34:27*
Sample: 6A-A_M13F-pUC *Lane: 86* *Base spacing: 15.195*

adobe.com
https://www.adobe.com

Acrobat Reader - Originální PDF nástroj

Všechny nástroje, které potřebujete ke snadnému převodu, úpravám a sdílení soub

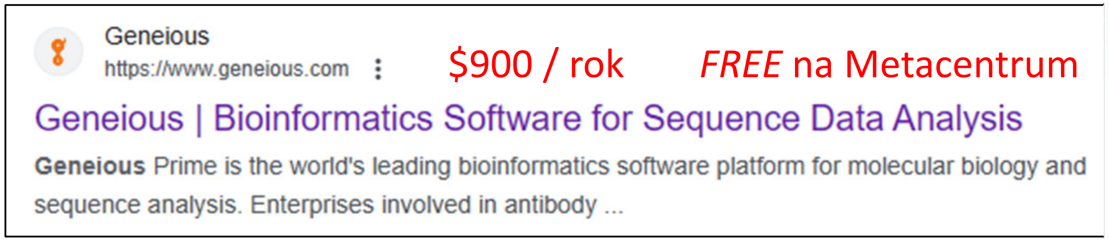
Obrázky vytvořil Vratislav Peška



File: 6A-A-M13F-pUC.ab1 Run Ended: 2010/11/30 20:34:27 Signal G:169 A:232 C:259 T:281

Sample: 6A-A_M13F-pUC Lane: 86 Base spacing: 15.195725 950 bases in 11613 scans Page 1 of 2

Obrázky vytvořil Vratislav Peška



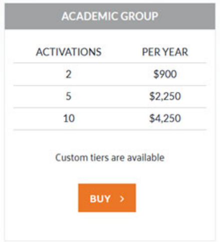
Geneious
https://www.geneious.com

$900 / rok

FREE na Metacentrum

Geneious | Bioinformatics Software for Sequence Data Analysis

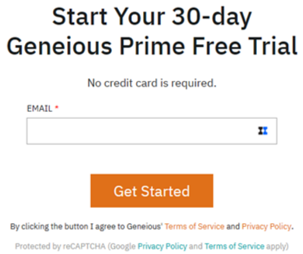
Geneious Prime is the world's leading bioinformatics software platform for molecular biology and sequence analysis. Enterprises involved in antibody ...



ACADEMIC GROUP

ACTIVATIONS	PER YEAR
2	$900
5	$2,250
10	$4,250

Custom tiers are available



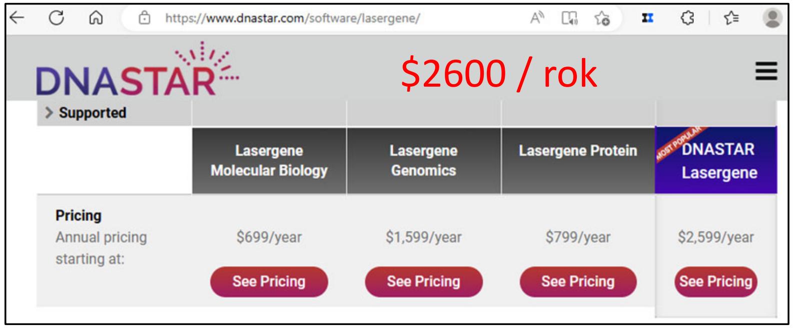
Start Your 30-day
Geneious Prime Free Trial

No credit card is required.

EMAIL *

By clicking the button I agree to Geneious' Terms of Service and Privacy Policy.

Protected by reCAPTCHA (Google Privacy Policy and Terms of Service apply)



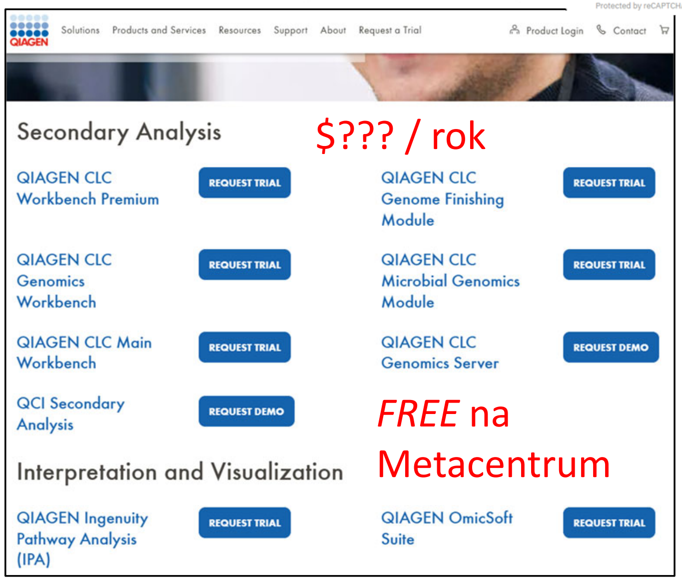
https://www.dnastar.com/software/lasergene/

DNASTAR

$2600 / rok

› Supported

	Lasergene Molecular Biology	Lasergene Genomics	Lasergene Protein	MOST POPULAR DNASTAR Lasergene
Pricing Annual pricing starting at:	$699/year See Pricing	$1,599/year See Pricing	$799/year See Pricing	$2,599/year See Pricing



Secondary Analysis

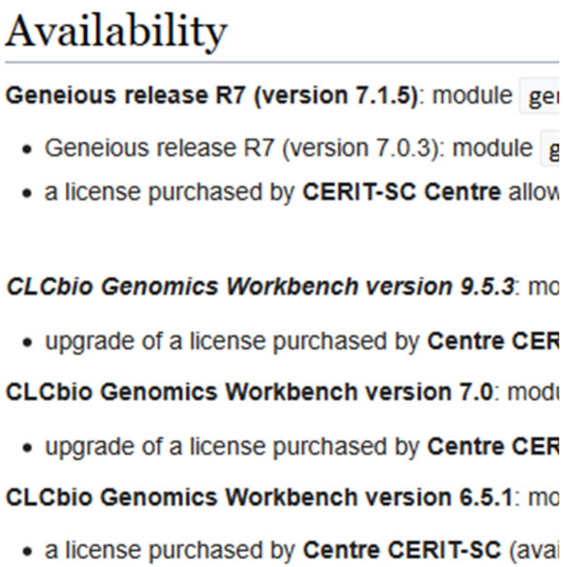
$??? / rok

QIAGEN CLC Workbench Premium	REQUEST TRIAL	QIAGEN CLC Genome Finishing Module	REQUEST TRIAL
QIAGEN CLC Genomics Workbench	REQUEST TRIAL	QIAGEN CLC Microbial Genomics Module	REQUEST TRIAL
QIAGEN CLC Main Workbench	REQUEST TRIAL	QIAGEN CLC Genomics Server	REQUEST DEMO
QCI Secondary Analysis	REQUEST DEMO		

FREE na Metacentrum

Interpretation and Visualization

QIAGEN Ingenuity Pathway Analysis (IPA)	REQUEST TRIAL	QIAGEN OmicSoft Suite	REQUEST TRIAL



Availability

Geneious release R7 (version 7.1.5): module ge

- Geneious release R7 (version 7.0.3): module g
- a license purchased by **CERIT-SC Centre** allow

CLCbio Genomics Workbench version 9.5.3: mo

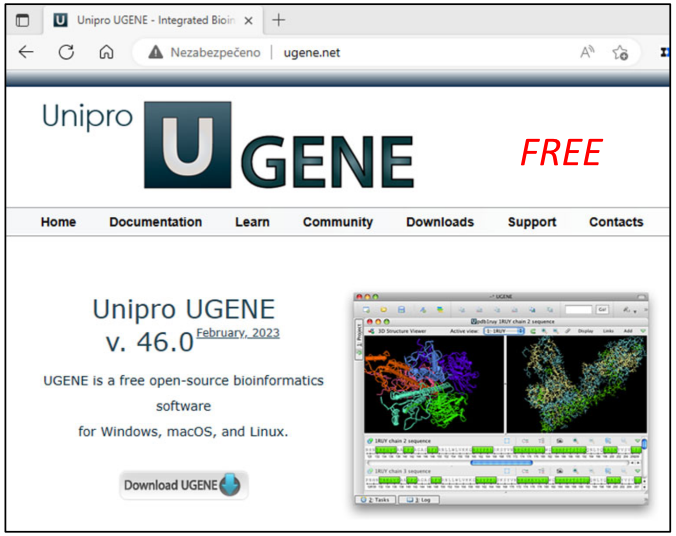
- upgrade of a license purchased by **Centre CER**

CLCbio Genomics Workbench version 7.0: modu

- upgrade of a license purchased by **Centre CER**

CLCbio Genomics Workbench version 6.5.1: mo

- a license purchased by **Centre CERIT-SC** (avai



Unipro UGENE - Integrated Bioin

Nezabezpečeno | ugene.net

Unipro UGENE

FREE

Unipro UGENE
v. 46.0 February, 2023

UGENE is a free open-source bioinformatics software
for Windows, macOS, and Linux.

https://thalljiscience.github.io

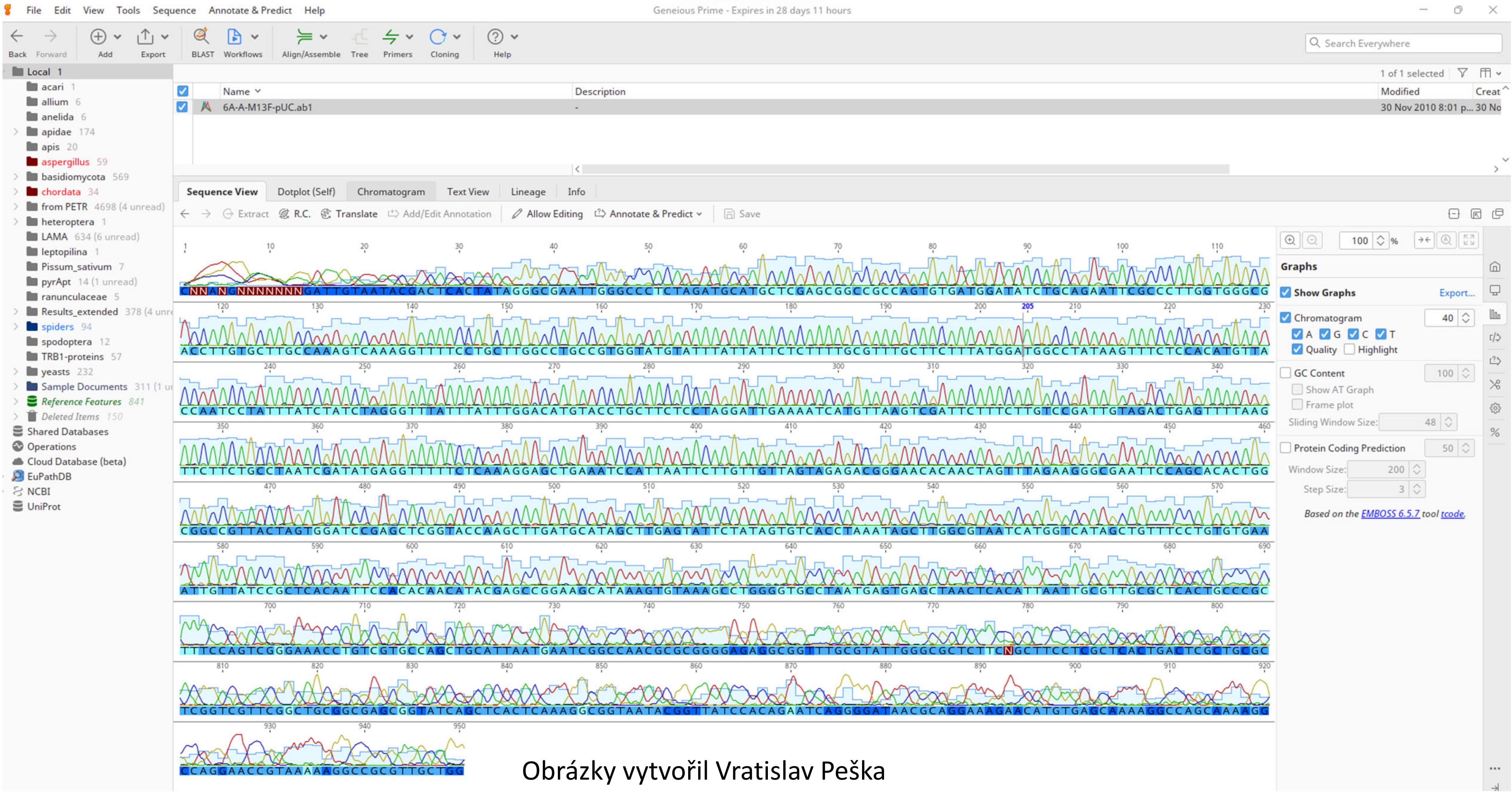
FREE

BioEdit

Copyright © 1997-2017
Tom Hall

Biological sequence alignment editor for Win95/98/NT/2K/XP

Obrázky vytvořil Vratislav Peška



☑ 6A-A-M13F-pUC.ab1 | - | 30 Nov 2010 8:01 p... | 30 No

100 %

☑ Show Graphs

☑ Chromatogram 40

☑ A ☑ G ☑ C ☑ T

☑ Quality ☐ Highlight

☐ GC Content 100

☐ Show AT Graph

☐ Frame plot

Sliding Window Size: 48

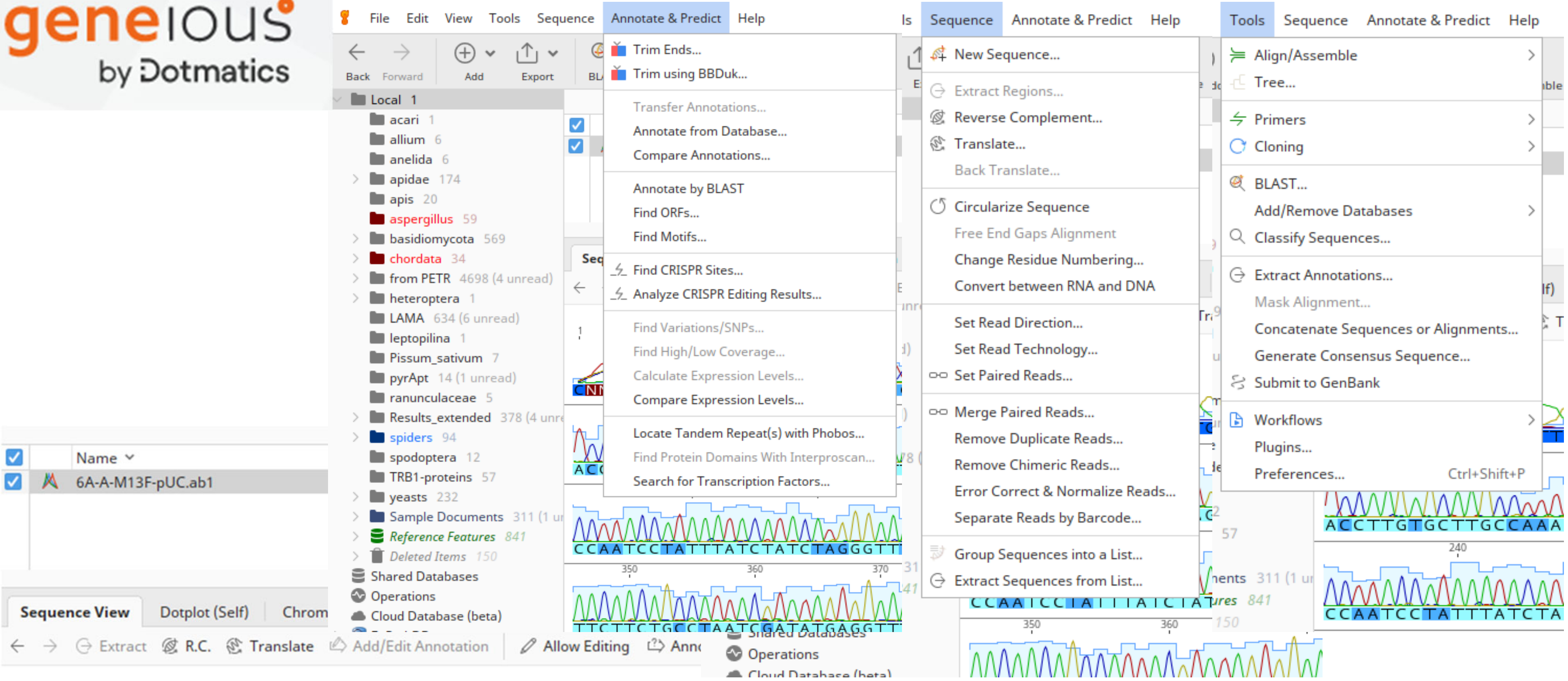
☐ Protein Coding Prediction 50

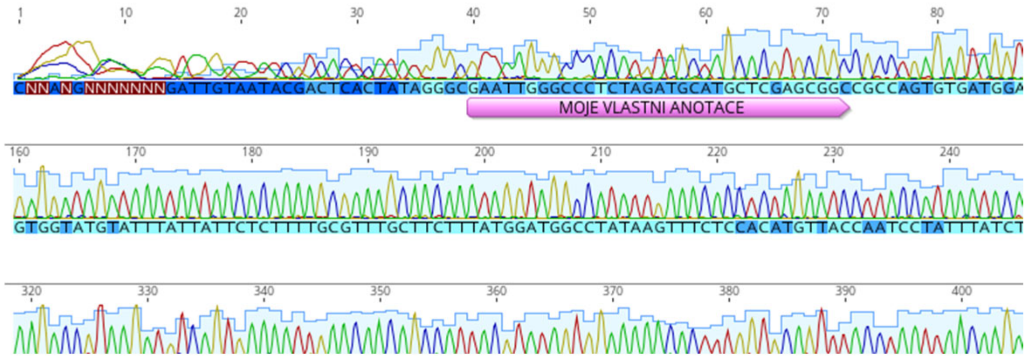
Window Size: 200

Step Size: 3

Obrázky vytvořil Vratislav Peška

geneious by Dotmatics



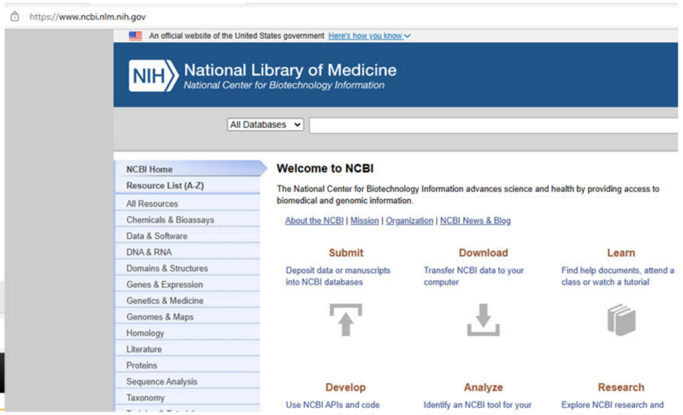


Obrázky vytvořil Vratislav Peška

DATABÁZE - Velká trojice EMBL/NCBI/DDBJ

Obrázky vytvořil Vratislav Peška

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

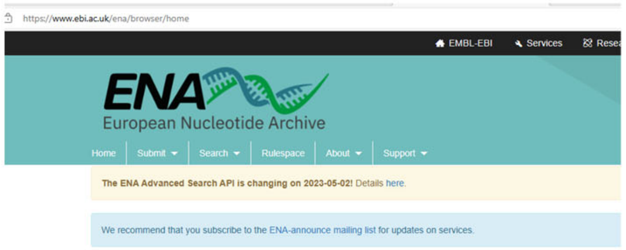
Use NCBI APIs and code

Analyze

Identify an NCBI tool for your

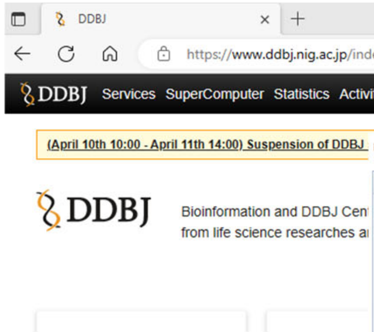
Research

Explore NCBI research and



The ENA Advanced Search API is changing on 2023-05-02! Details here.

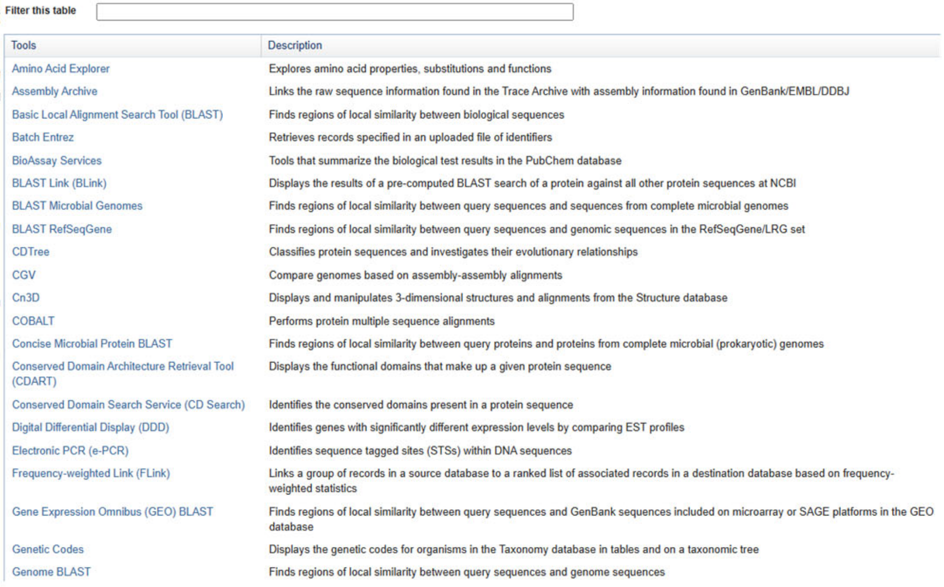
We recommend that you subscribe to the ENA-announce mailing list for updates on services.



(April 10th 10:00 - April 11th 14:00) Suspension of DDBJ

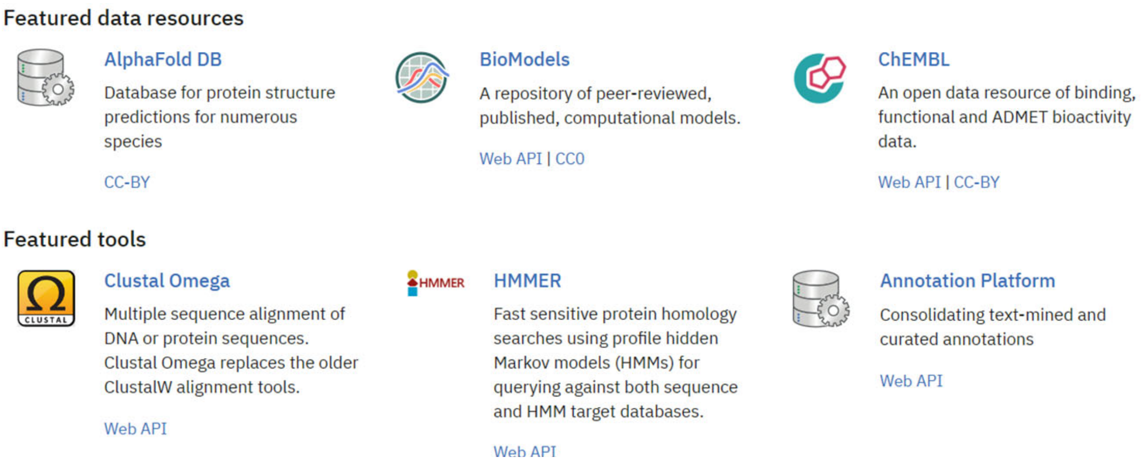
Bioinformation and DDBJ Cen

from life science researches a



Filter this table

Tools	Description
Amino Acid Explorer	Explores amino acid properties, substitutions and functions
Assembly Archive	Links the raw sequence information found in the Trace Archive with assembly information found in GenBank/EMBL/DDBJ
Basic Local Alignment Search Tool (BLAST)	Finds regions of local similarity between biological sequences
Batch Entrez	Retrieves records specified in an uploaded file of identifiers
BioAssay Services	Tools that summarize the biological test results in the PubChem database
BLAST Link (BLink)	Displays the results of a pre-computed BLAST search of a protein against all other protein sequences at NCBI
BLAST Microbial Genomes	Finds regions of local similarity between query sequences and sequences from complete microbial genomes
BLAST RefSeqGene	Finds regions of local similarity between query sequences and genomic sequences in the RefSeqGene/LRG set
CDTree	Classifies protein sequences and investigates their evolutionary relationships
CGV	Compare genomes based on assembly-assembly alignments
Cn3D	Displays and manipulates 3-dimensional structures and alignments from the Structure database
COBALT	Performs protein multiple sequence alignments
Concise Microbial Protein BLAST	Finds regions of local similarity between query proteins and proteins from complete microbial (prokaryotic) genomes
Conserved Domain Architecture Retrieval Tool (CDART)	Displays the functional domains that make up a given protein sequence
Conserved Domain Search Service (CD Search)	Identifies the conserved domains present in a protein sequence
Digital Differential Display (DDD)	Identifies genes with significantly different expression levels by comparing EST profiles
Electronic PCR (e-PCR)	Identifies sequence tagged sites (STSs) within DNA sequences
Frequency-weighted Link (FLink)	Links a group of records in a source database to a ranked list of associated records in a destination database based on frequency-weighted statistics
Gene Expression Omnibus (GEO) BLAST	Finds regions of local similarity between query sequences and GenBank sequences included on microarray or SAGE platforms in the GEO database
Genetic Codes	Displays the genetic codes for organisms in the Taxonomy database in tables and on a taxonomic tree
Genome BLAST	Finds regions of local similarity between query sequences and genome sequences



Featured data resources

AlphaFold DB

Database for protein structure predictions for numerous species

CC-BY

BioModels

A repository of peer-reviewed, published, computational models.

Web API | CC0

ChEMBL

An open data resource of binding, functional and ADMET bioactivity data.

Web API | CC-BY

Featured tools

Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Web API

HMMER

Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

Web API

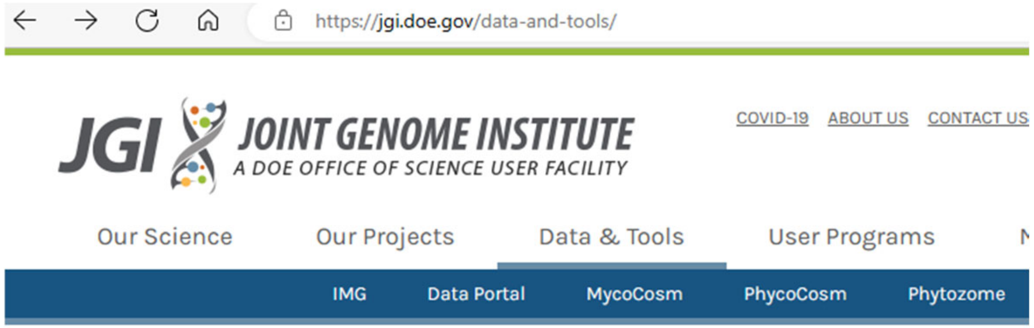
Annotation Platform

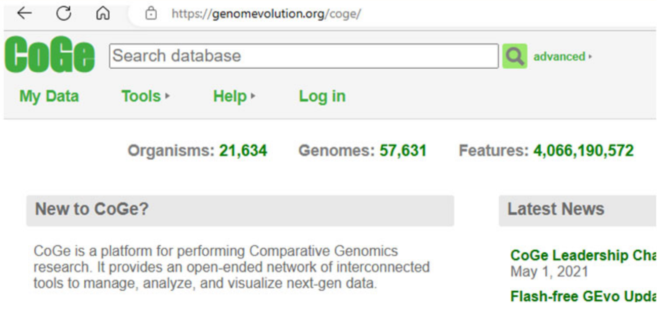
Consolidating text-mined and curated annotations

Web API

NCBI databáze

Assembly
BioCollections
BioProject (formerly Genome Project)
BioSample
Bookshelf
ClinVar
ClinicalTrials.gov
Computational Resources from NCBI's Structure Group
Consensus CDS (CCDS)
Conserved Domain Database (CDD)
Database of Genomic Structural Variation (dbVar)
Database of Genotypes and Phenotypes (dbGaP)
Database of Short Genetic Variations (dbSNP)
GenBank
Gene
Gene Expression Omnibus (GEO) Database
Gene Expression Omnibus (GEO) Datasets
Gene Expression Omnibus (GEO) Profiles
GeneReviews
Genes and Disease
Genetic Testing Registry (GTR)
Genome
Genome Reference Consortium (GRC)
Glycans
HIV-1, Human Protein Interaction Database
Identical Protein Groups
Influenza Virus
Journals in NCBI Databases
MeSH Database
MedGen
NCBI C++ Toolkit Manual
NCBI Education Page
NCBI Glossary
NCBI Handbook
NCBI Help Manual
NCBI Pathogen Detection Project
National Library of Medicine (NLM) Catalog
Nucleotide Database
Online Mendelian Inheritance in Man (OMIM)
PopSet
Protein Clusters
Protein Database
Protein Family Models
PubChem BioAssay
PubChem Compound
PubChem Substance
PubMed
PubMed Central (PMC)
RefSeqGene
Reference Sequence (RefSeq)
Retrovirus Resources
SARS CoV
Sequence Read Archive (SRA)
Structure (Molecular Modeling Database)
Taxonomy
Third Party Annotation (TPA) Database
Trace Archive
Viral Genomes
Virus Variation





CoGe is a platform for performing Comparative Genomics research. It provides an open-ended network of interconnected tools to manage, analyze, and visualize next-gen data.

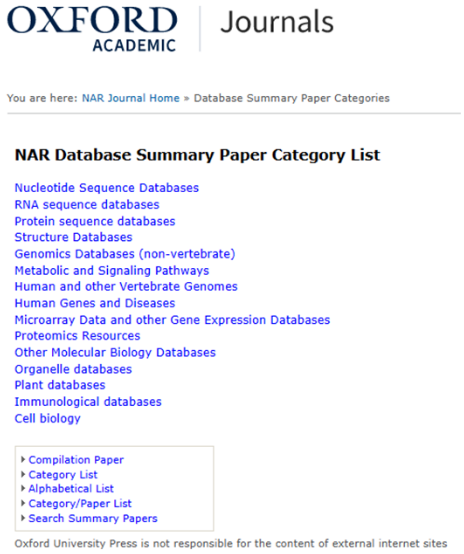
Obrázky vytvořil Vratislav Peška

Obrázky vytvořil Vratislav Peška

KEGG: Kyoto Encyclopedia of Genes and Genomes

UniProtKB/Swiss-Prot

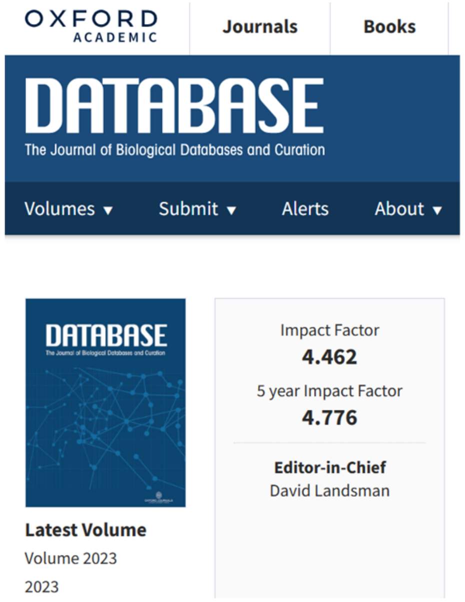
The Arabidopsis Information Resource



NAR Database Summary Paper Category List

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology

Oxford University Press is not responsible for the content of external internet sites



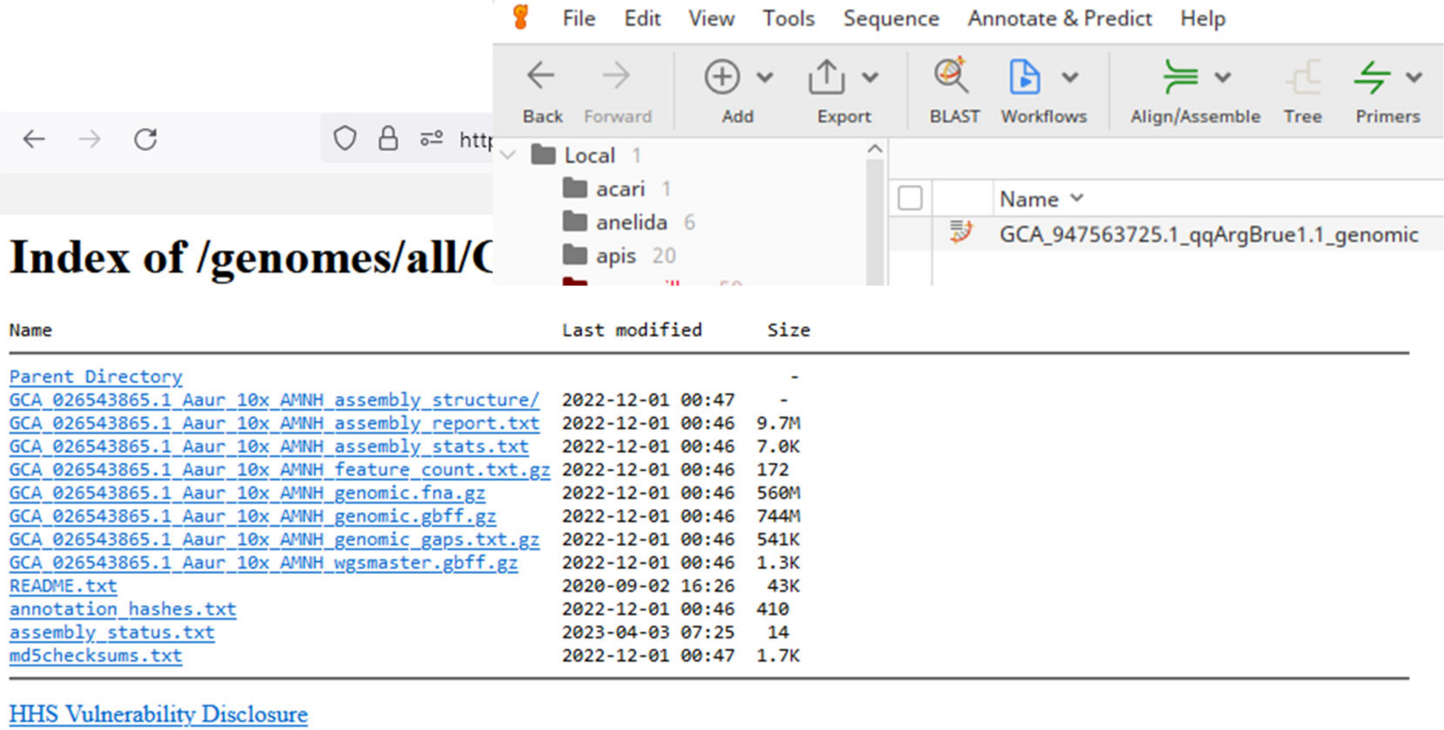
Impact Factor
4.462

5 year Impact Factor
4.776

Editor-in-Chief
David Landsman

Latest Volume
Volume 2023
2023

Příklad stahování dat z NCBI



☐ GCA_947563725.1_qqArgBrue1.1_genomic

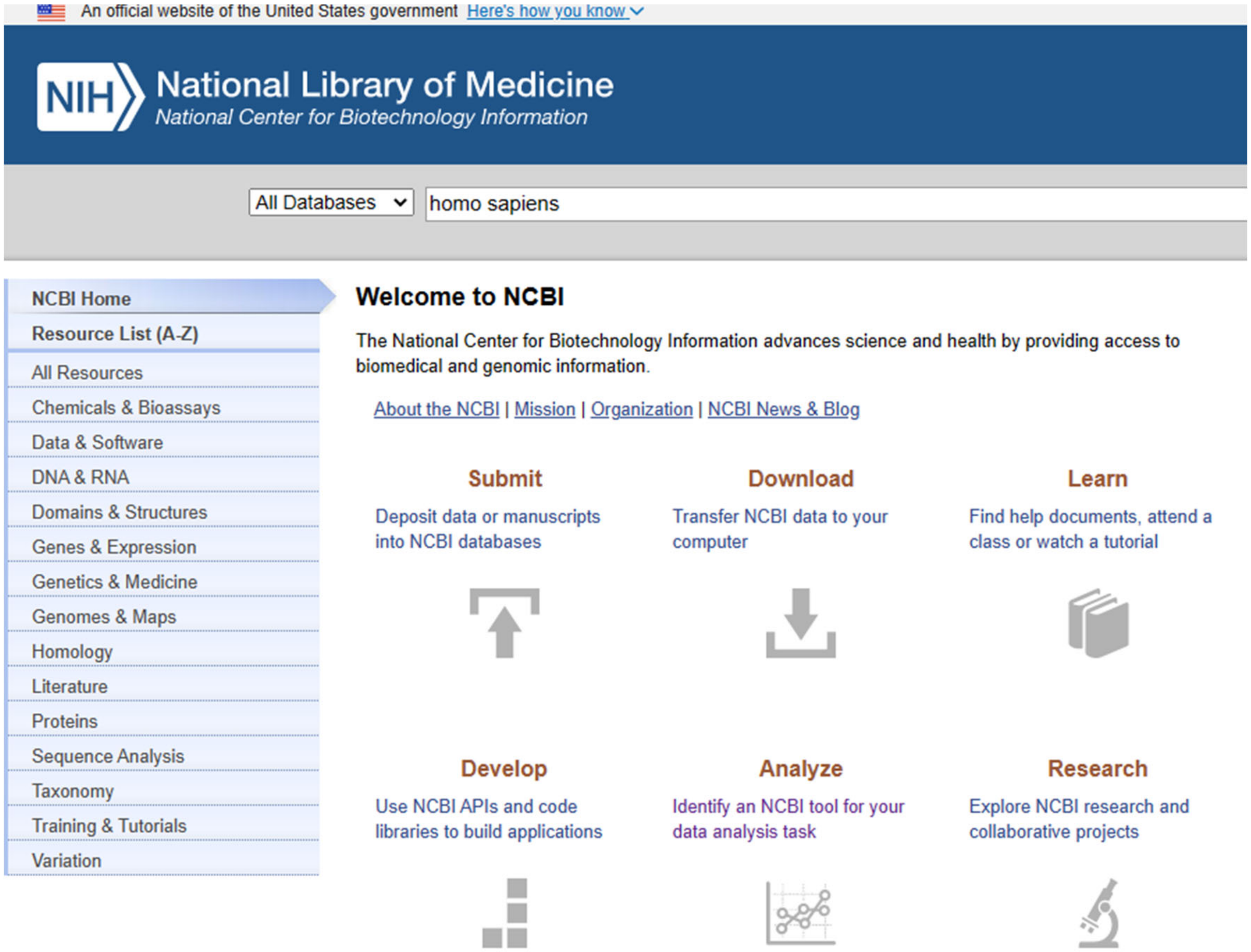
Index of /genomes/all/C

```
Name                                              Last modified      Size

Parent Directory                                                       -
GCA_026543865.1_Aaur_10x_AMNH_assembly_structure/ 2022-12-01 00:47    -
GCA_026543865.1_Aaur_10x_AMNH_assembly_report.txt 2022-12-01 00:46  9.7M
GCA_026543865.1_Aaur_10x_AMNH_assembly_stats.txt  2022-12-01 00:46  7.0K
GCA_026543865.1_Aaur_10x_AMNH_feature_count.txt.gz 2022-12-01 00:46  172
GCA_026543865.1_Aaur_10x_AMNH_genomic.fna.gz      2022-12-01 00:46  560M
GCA_026543865.1_Aaur_10x_AMNH_genomic.gbff.gz     2022-12-01 00:46  744M
GCA_026543865.1_Aaur_10x_AMNH_genomic_gaps.txt.gz 2022-12-01 00:46  541K
GCA_026543865.1_Aaur_10x_AMNH_wgsmaster.gbff.gz   2022-12-01 00:46  1.3K
README.txt                                        2020-09-02 16:26   43K
annotation_hashes.txt                             2022-12-01 00:46  410
assembly_status.txt                               2023-04-03 07:25   14
md5checksums.txt                                  2022-12-01 00:47  1.7K
```

HHS Vulnerability Disclosure

Obrázky vytvořil Vratislav Peška



An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

All Databases | homo sapiens

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

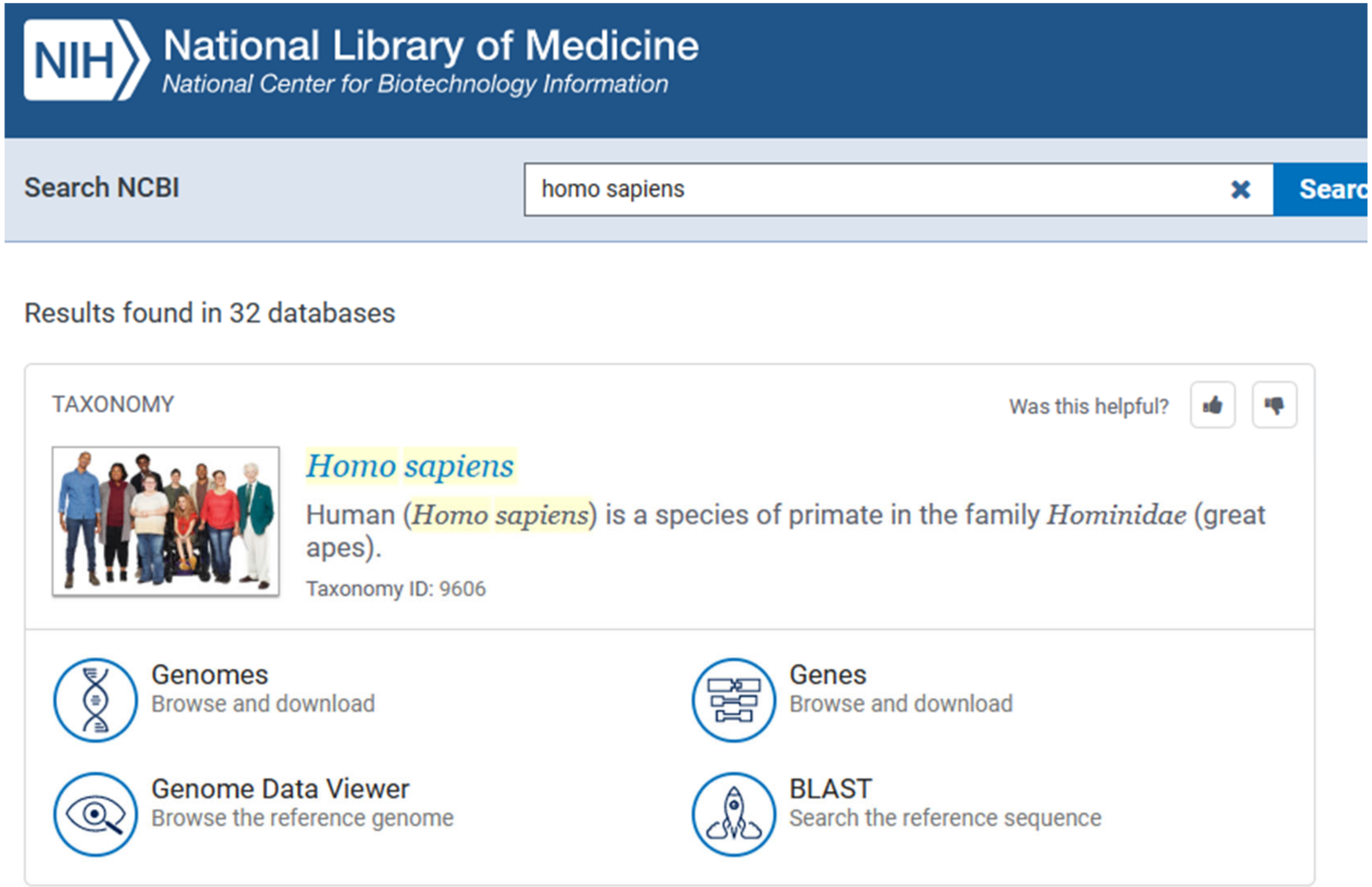
Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Obrázky vytvořil Vratislav Peška



National Library of Medicine
National Center for Biotechnology Information

Search NCBI

homo sapiens

Results found in 32 databases

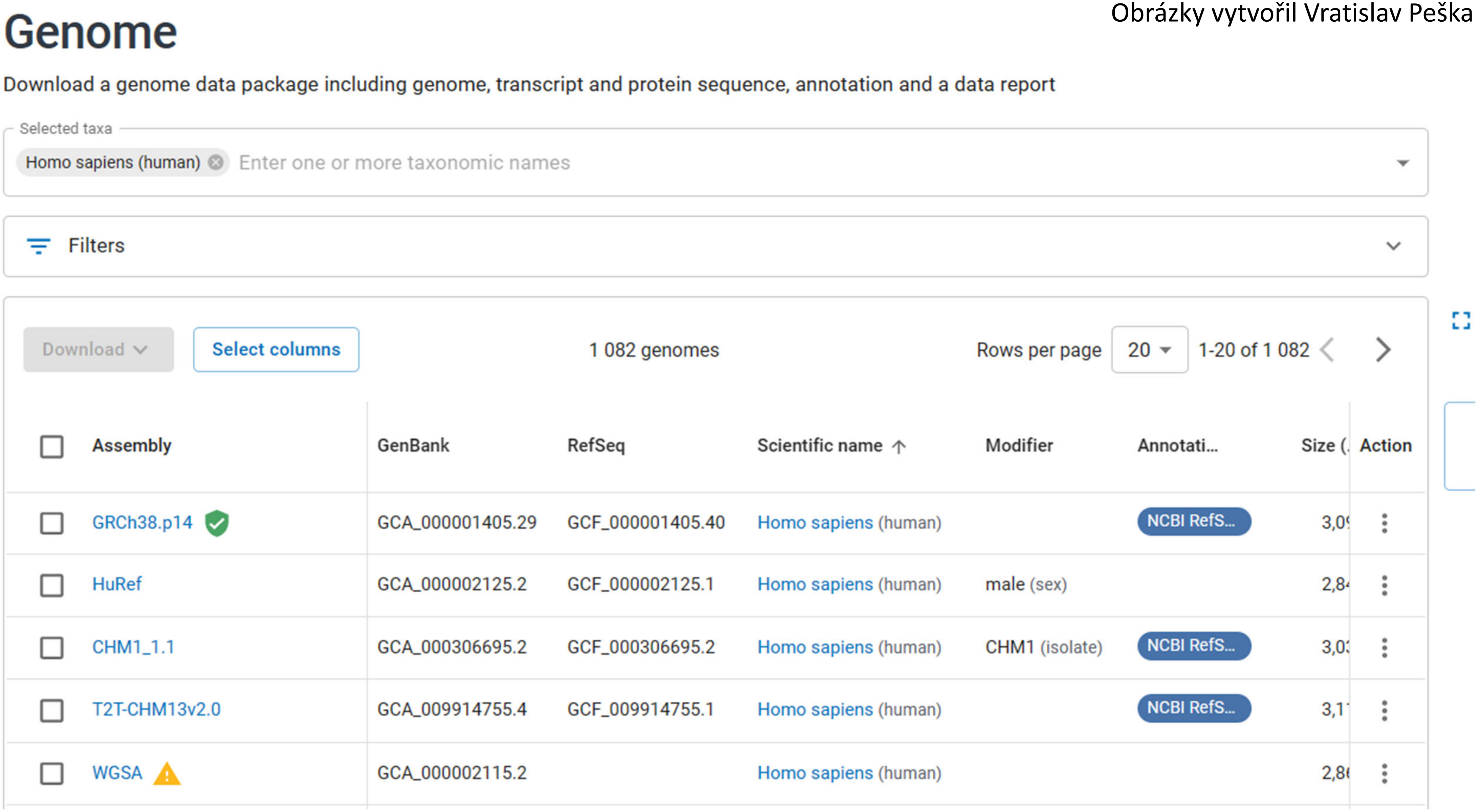
TAXONOMY

Homo sapiens

Human (*Homo sapiens*) is a species of primate in the family *Hominidae* (great apes).

Taxonomy ID: 9606

Obrázky vytvořil Vratislav Peška



Obrázky vytvořil Vratislav Peška

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

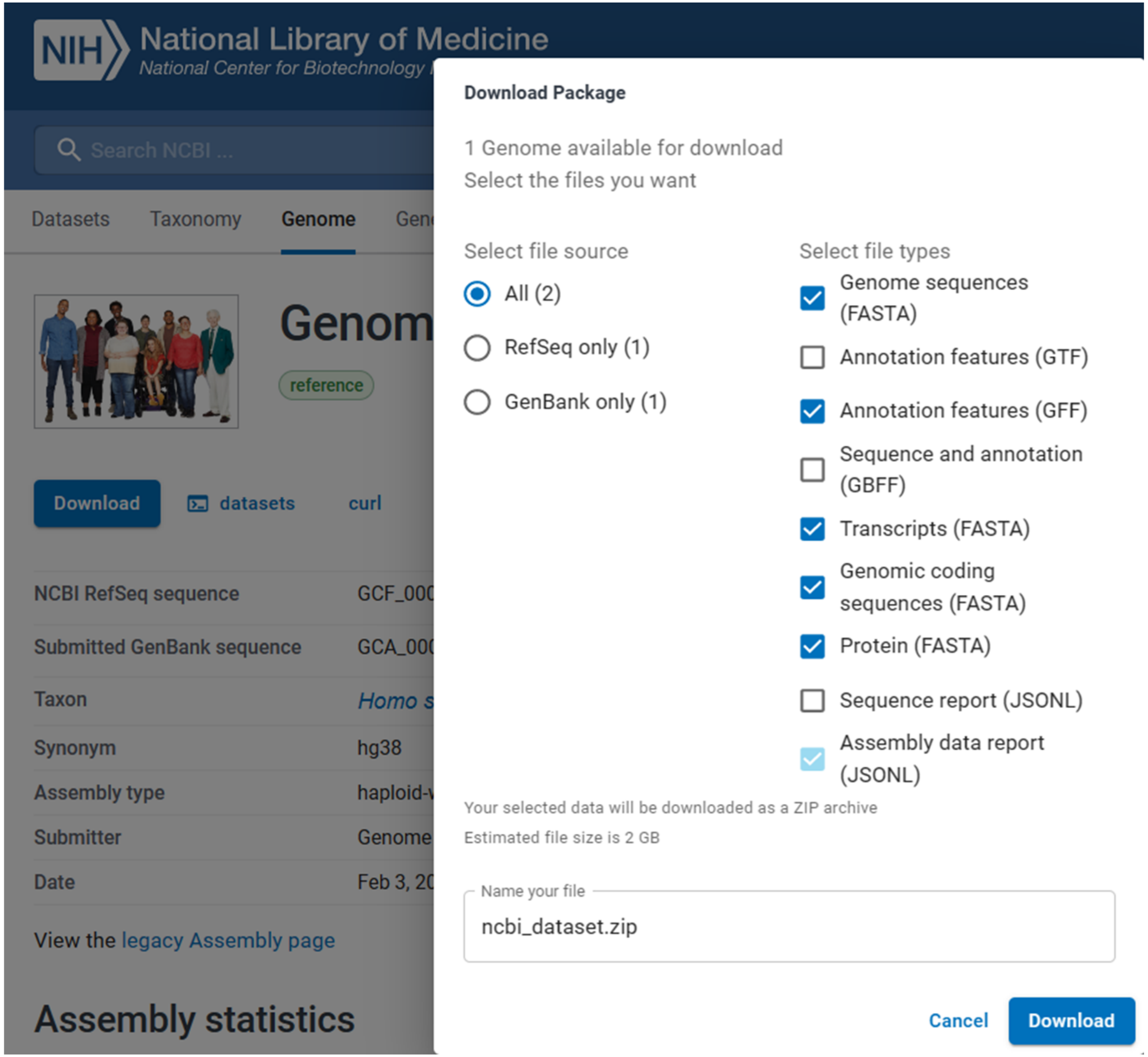
Selected taxa: Homo sapiens (human) ⊗ Enter one or more taxonomic names

1 082 genomes

Rows per page: 20

1-20 of 1 082

☐	Assembly	GenBank	RefSeq	Scientific name ↑	Modifier	Annotati…	Size (	Action
☐	GRCh38.p14 ✓	GCA_000001405.29	GCF_000001405.40	Homo sapiens (human)		NCBI RefS…	3,0	⋮
☐	HuRef	GCA_000002125.2	GCF_000002125.1	Homo sapiens (human)	male (sex)		2,8	⋮
☐	CHM1_1.1	GCA_000306695.2	GCF_000306695.2	Homo sapiens (human)	CHM1 (isolate)	NCBI RefS…	3,0	⋮
☐	T2T-CHM13v2.0	GCA_009914755.4	GCF_009914755.1	Homo sapiens (human)		NCBI RefS…	3,1	⋮
☐	WGSA ⚠	GCA_000002115.2		Homo sapiens (human)			2,8	⋮



NCBI RefSeq sequence GCF_000
Submitted GenBank sequence GCA_000
Taxon *Homo s*
Synonym hg38
Assembly type haploid-
Submitter Genome
Date Feb 3, 20

View the legacy Assembly page

Assembly statistics

Download Package

1 Genome available for download
Select the files you want

Select file source

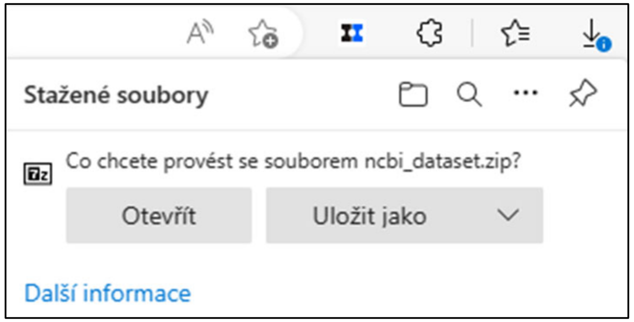
● All (2)
○ RefSeq only (1)
○ GenBank only (1)

Select file types

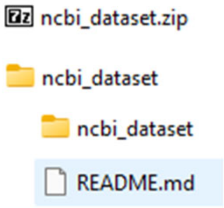
☑ Genome sequences (FASTA)
☐ Annotation features (GTF)
☑ Annotation features (GFF)
☐ Sequence and annotation (GBFF)
☑ Transcripts (FASTA)
☑ Genomic coding sequences (FASTA)
☑ Protein (FASTA)
☐ Sequence report (JSONL)
☑ Assembly data report (JSONL)

Your selected data will be downloaded as a ZIP archive
Estimated file size is 2 GB

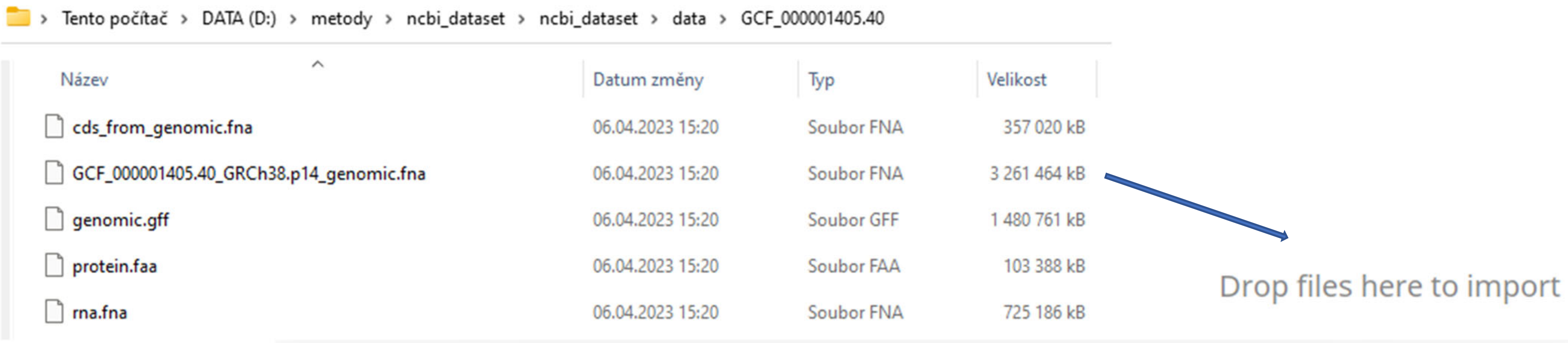
Name your file: ncbi_dataset.zip


Stažené soubory

Co chcete provést se souborem ncbi_dataset.zip?


- ncbi_dataset.zip
- ncbi_dataset
  - ncbi_dataset
  - README.md

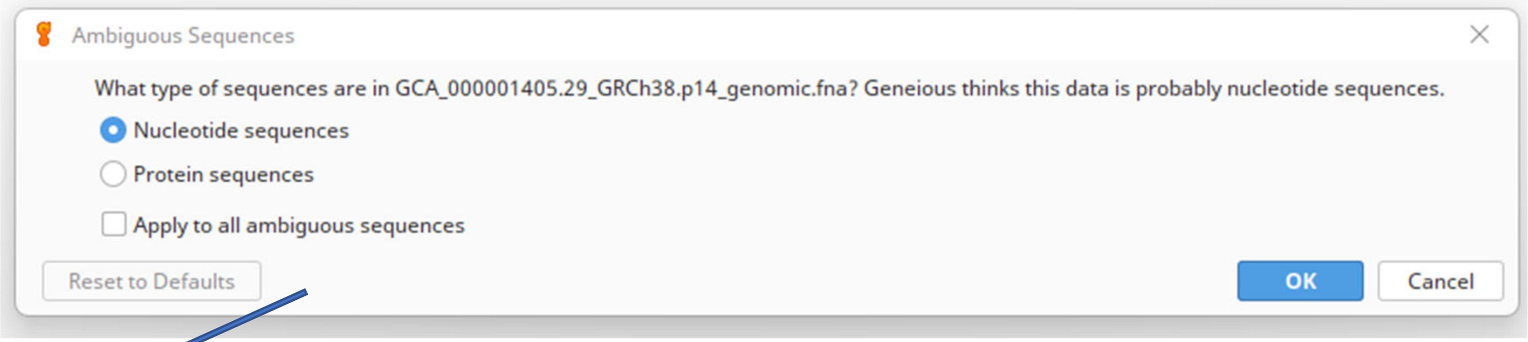
Obrázky vytvořil Vratislav Peška



Tento počítač › DATA (D:) › metody › ncbi_dataset › ncbi_dataset › data › GCF_000001405.40

Název	Datum změny	Typ	Velikost
cds_from_genomic.fna	06.04.2023 15:20	Soubor FNA	357 020 kB
GCF_000001405.40_GRCh38.p14_genomic.fna	06.04.2023 15:20	Soubor FNA	3 261 464 kB
genomic.gff	06.04.2023 15:20	Soubor GFF	1 480 761 kB
protein.faa	06.04.2023 15:20	Soubor FAA	103 388 kB
rna.fna	06.04.2023 15:20	Soubor FNA	725 186 kB

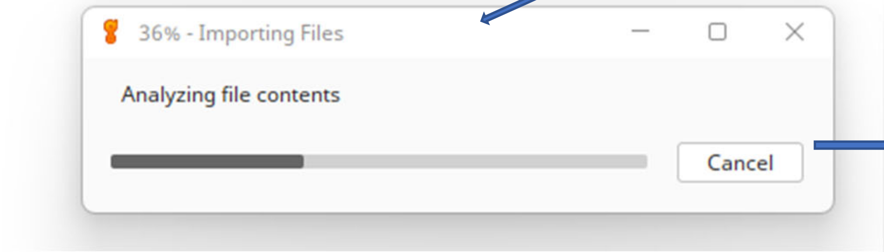
Drop files here to import



Ambiguous Sequences

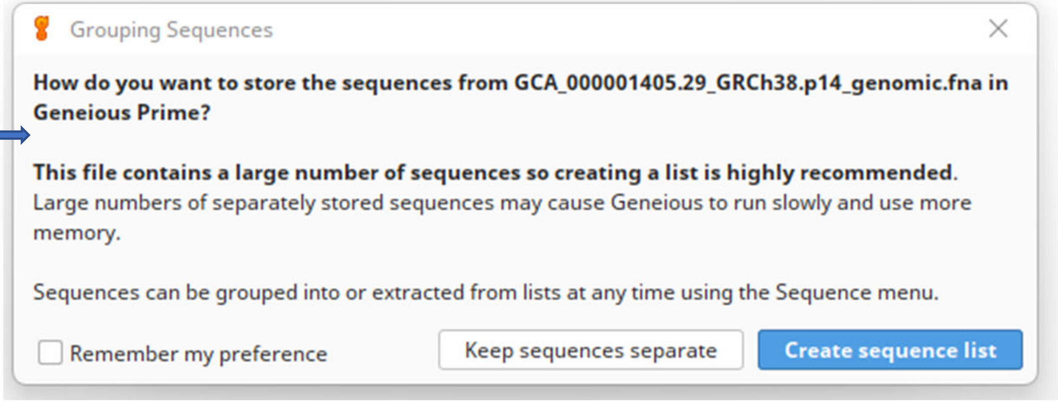
What type of sequences are in GCA_000001405.29_GRCh38.p14_genomic.fna? Geneious thinks this data is probably nucleotide sequences.

● Nucleotide sequences
○ Protein sequences
☐ Apply to all ambiguous sequences



36% - Importing Files

Analyzing file contents



Grouping Sequences

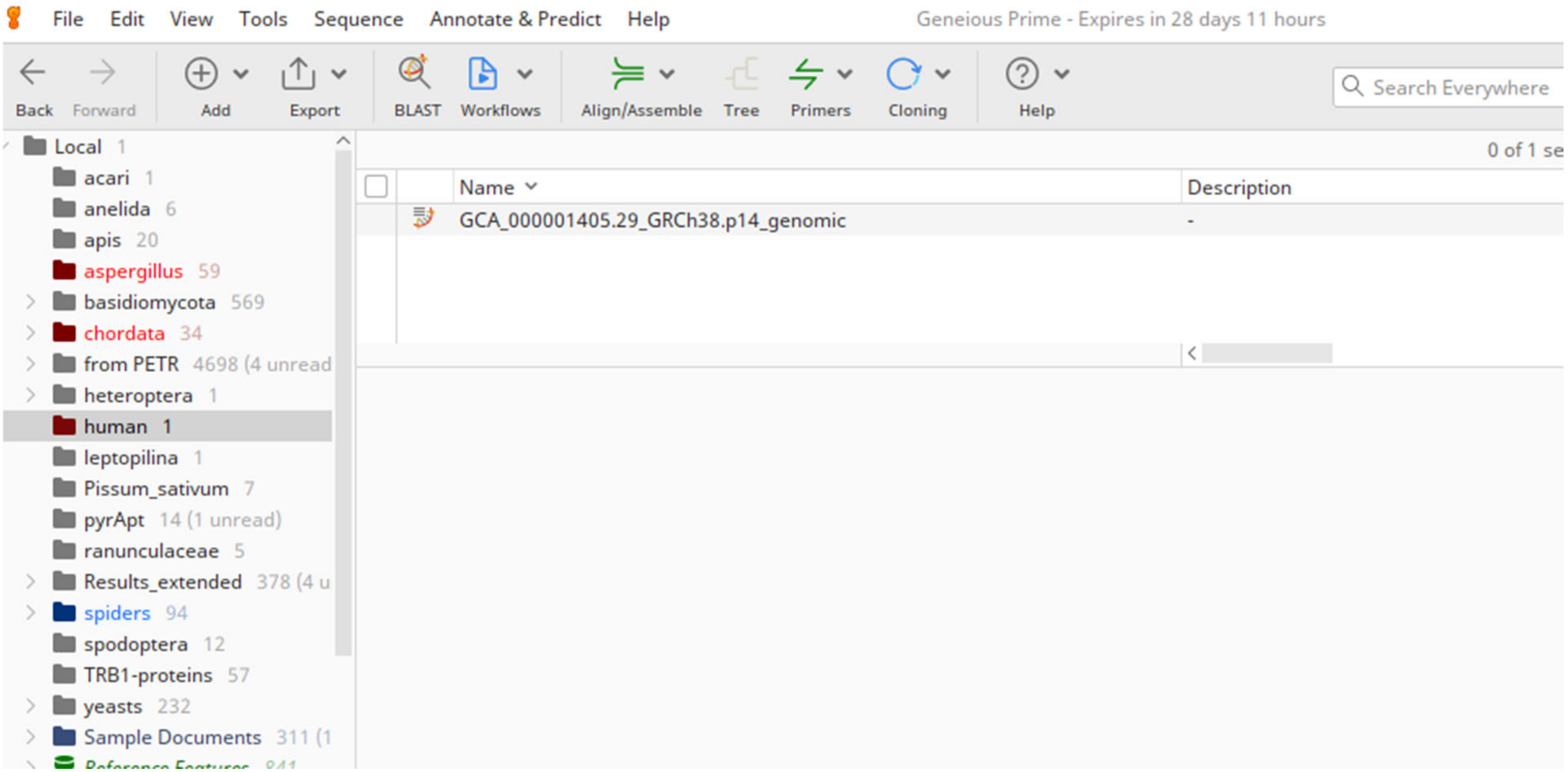
How do you want to store the sequences from GCA_000001405.29_GRCh38.p14_genomic.fna in Geneious Prime?

This file contains a large number of sequences so creating a list is highly recommended. Large numbers of separately stored sequences may cause Geneious to run slowly and use more memory.

Sequences can be grouped into or extracted from lists at any time using the Sequence menu.

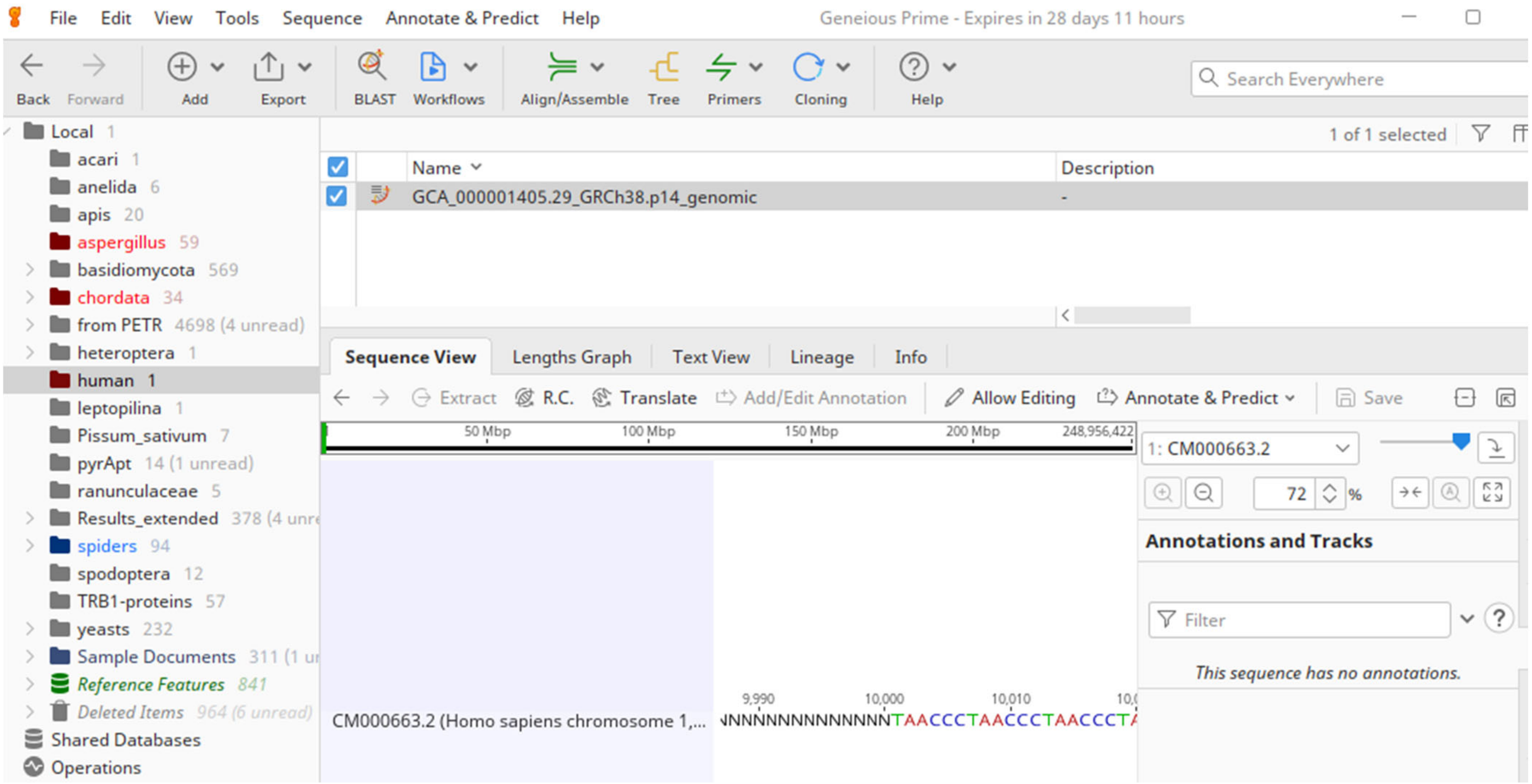
☐ Remember my preference

Obrázky vytvořil Vratislav Peška



Name	Description
GCA_000001405.29_GRCh38.p14_genomic	-

Obrázky vytvořil Vratislav Peška



☑	Name	Description
☑	GCA_000001405.29_GRCh38.p14_genomic	-

This sequence has no annotations.

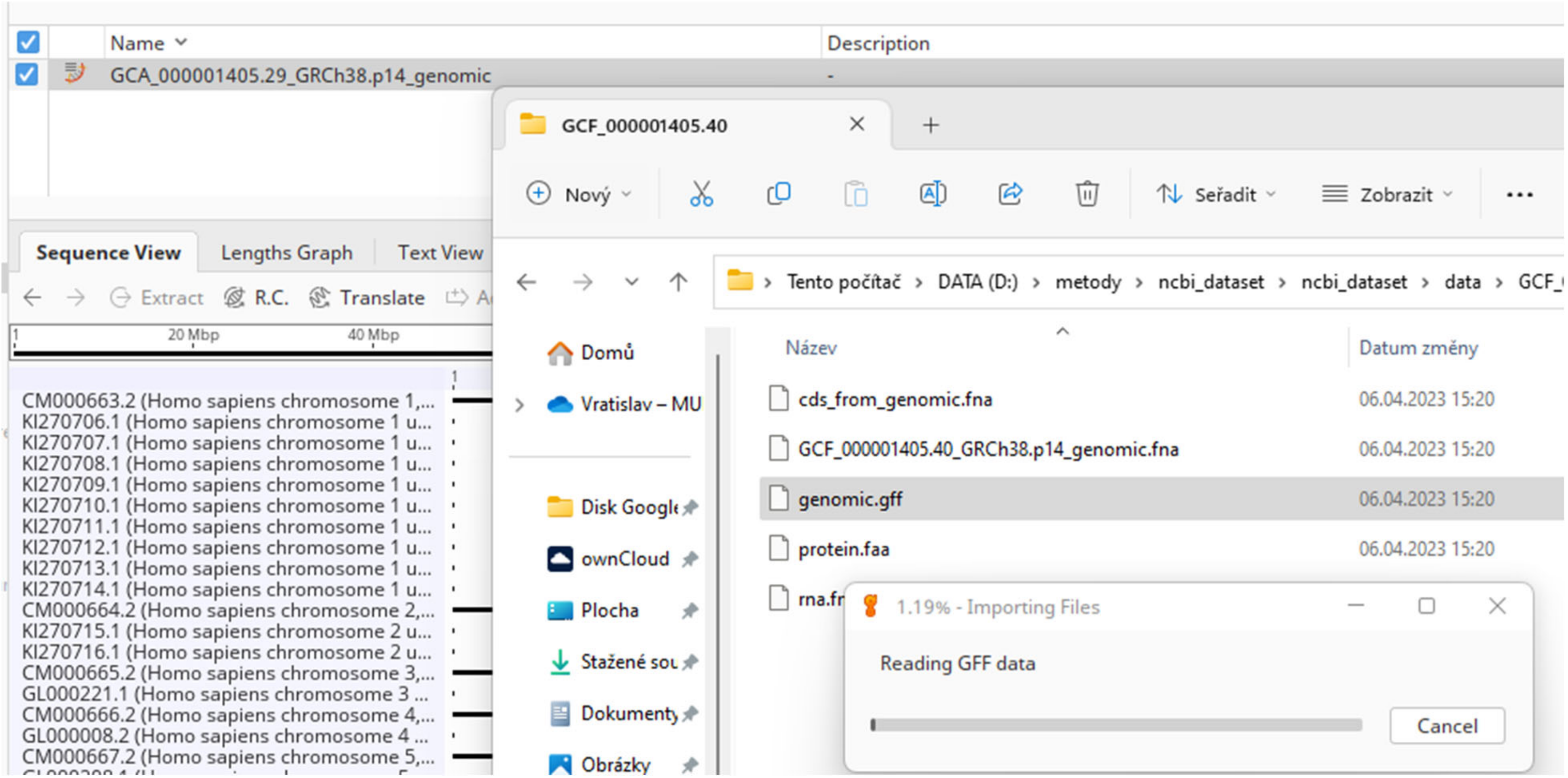
CM000663.2 (Homo sapiens chromosome 1,... NNNNNNNNNNNNNNNTAACCCTAACCCTAACCCTA

Obrázky vytvořil Vratislav Peška

Name	Description
GCA_000001405.29_GRCh38.p14_genomic	-

CM000663.2 (Homo sapiens chromosome 1,...
KI270706.1 (Homo sapiens chromosome 1 u...
KI270707.1 (Homo sapiens chromosome 1 u...
KI270708.1 (Homo sapiens chromosome 1 u...
KI270709.1 (Homo sapiens chromosome 1 u...
KI270710.1 (Homo sapiens chromosome 1 u...
KI270711.1 (Homo sapiens chromosome 1 u...
KI270712.1 (Homo sapiens chromosome 1 u...
KI270713.1 (Homo sapiens chromosome 1 u...
KI270714.1 (Homo sapiens chromosome 1 u...
CM000664.2 (Homo sapiens chromosome 2,...
KI270715.1 (Homo sapiens chromosome 2 u...
KI270716.1 (Homo sapiens chromosome 2 u...
CM000665.2 (Homo sapiens chromosome 3,...
GL000221.1 (Homo sapiens chromosome 3 ...
CM000666.2 (Homo sapiens chromosome 4,...
GL000008.2 (Homo sapiens chromosome 4 ...
CM000667.2 (Homo sapiens chromosome 5,...
GL000208.1 (Homo sapiens chromosome 5 ...
CM000668.2 (Homo sapiens chromosome 6,...
CM000669.2 (Homo sapiens chromosome 7,...
CM000670.2 (Homo sapiens chromosome 8,...
CM000671.2 (Homo sapiens chromosome 9,...
KI270717.1 (Homo sapiens chromosome 9 u...
KI270718.1 (Homo sapiens chromosome 9 u...
KI270719.1 (Homo sapiens chromosome 9 u...
KI270720.1 (Homo sapiens chromosome 9 u...
CM000672.2 (Homo sapiens chromosome 1...
CM000673.2 (Homo sapiens chromosome 1...
KI270721.1 (Homo sapiens chromosome 11 ...
CM000674.2 (Homo sapiens chromosome 1...
CM000675.2 (Homo sapiens chromosome 1...
CM000676.2 (Homo sapiens chromosome 1...
GL000009.2 (Homo sapiens chromosome 14...
GL000225.1 (Homo sapiens chromosome 14...
KI270722.1 (Homo sapiens chromosome 14 ...
GL000194.1 (Homo sapiens chromosome 14...
KI270723.1 (Homo sapiens chromosome 14 ...
KI270724.1 (Homo sapiens chromosome 14 ...
KI270725.1 (Homo sapiens chromosome 14 ...
KI270726.1 (Homo sapiens chromosome 14 ...
CM000677.2 (Homo sapiens chromosome 1...
KI270727.1 (Homo sapiens chromosome 15 ...
CM000678.2 (Homo sapiens chromosome 1...
KI270728.1 (Homo sapiens chromosome 16 ...
CM000679.2 (Homo sapiens chromosome 1...
GL000205.2 (Homo sapiens chromosome 17...
KI270729.1 (Homo sapiens chromosome 17 ...
KI270730.1 (Homo sapiens chromosome 17 ...
CM000680.2 (Homo sapiens chromosome 1...
CM000681.2 (Homo sapiens chromosome 1...
CM000682.2 (Homo sapiens chromosome 2...

Obrázky vytvořil Vratislav Peška



Name	Description
☑ GCA_000001405.29_GRCh38.p14_genomic	-

CM000663.2 (Homo sapiens chromosome 1,...
KI270706.1 (Homo sapiens chromosome 1 u...
KI270707.1 (Homo sapiens chromosome 1 u...
KI270708.1 (Homo sapiens chromosome 1 u...
KI270709.1 (Homo sapiens chromosome 1 u...
KI270710.1 (Homo sapiens chromosome 1 u...
KI270711.1 (Homo sapiens chromosome 1 u...
KI270712.1 (Homo sapiens chromosome 1 u...
KI270713.1 (Homo sapiens chromosome 1 u...
KI270714.1 (Homo sapiens chromosome 1 u...
CM000664.2 (Homo sapiens chromosome 2,...
KI270715.1 (Homo sapiens chromosome 2 u...
KI270716.1 (Homo sapiens chromosome 2 u...
CM000665.2 (Homo sapiens chromosome 3,...
GL000221.1 (Homo sapiens chromosome 3 ...
CM000666.2 (Homo sapiens chromosome 4,...
GL000008.2 (Homo sapiens chromosome 4 ...
CM000667.2 (Homo sapiens chromosome 5,...

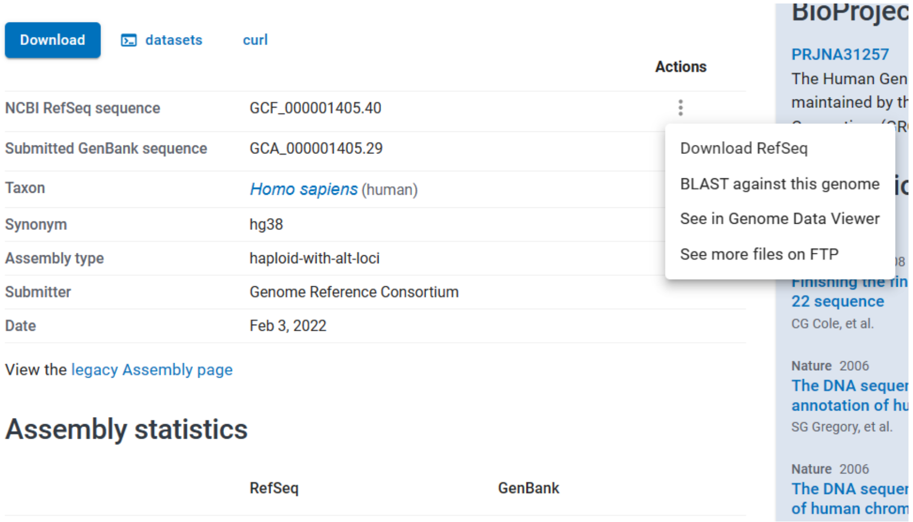
GCF_000001405.40

Název	Datum změny
cds_from_genomic.fna	06.04.2023 15:20
GCF_000001405.40_GRCh38.p14_genomic.fna	06.04.2023 15:20
genomic.gff	06.04.2023 15:20
protein.faa	06.04.2023 15:20
rna.fr	

1.19% - Importing Files

Reading GFF data

Obrázky vytvořil Vratislav Peška



NCBI RefSeq sequence	GCF_000001405.40
Submitted GenBank sequence	GCA_000001405.29
Taxon	*Homo sapiens* (human)
Synonym	hg38
Assembly type	haploid-with-alt-loci
Submitter	Genome Reference Consortium
Date	Feb 3, 2022

View the legacy Assembly page

Assembly statistics

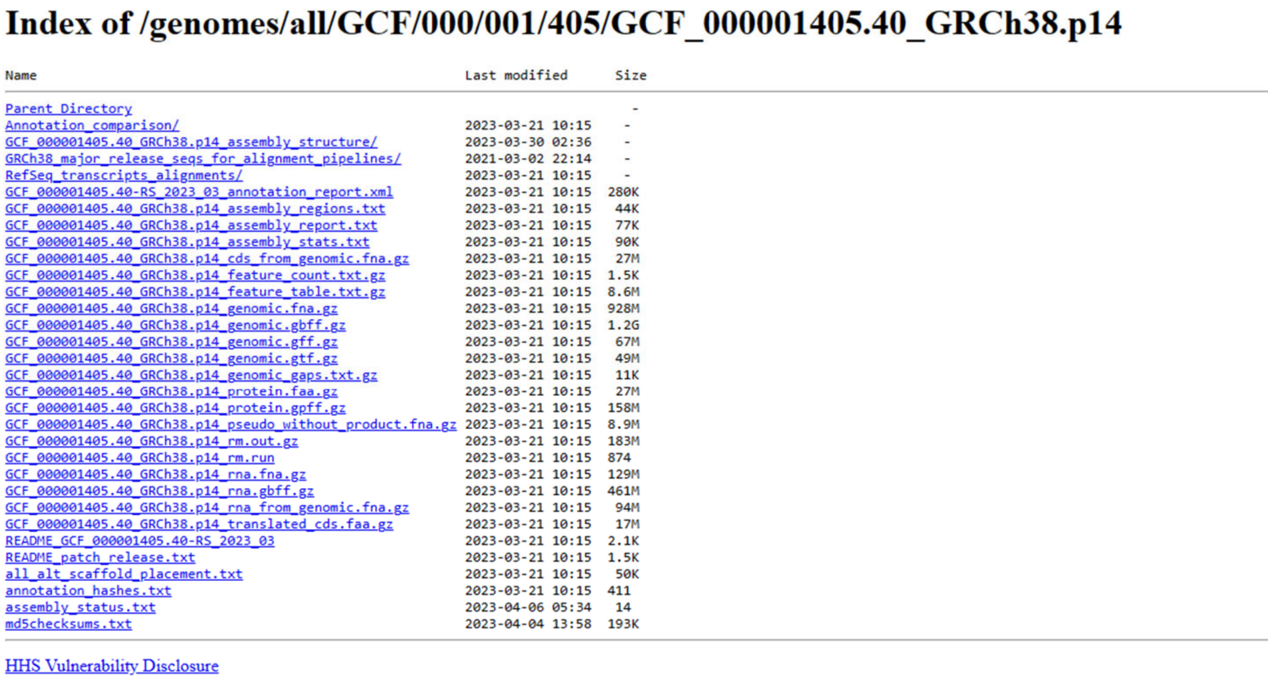
	RefSeq	GenBank

BioProjec
PRJNA31257
The Human Gen
maintained by th

22 sequence
CG Cole, et al.

Nature 2006
The DNA sequer
annotation of hu
SG Gregory, et al.

Nature 2006
The DNA sequer
of human chrom



Index of /genomes/all/GCF/000/001/405/GCF_000001405.40_GRCh38.p14

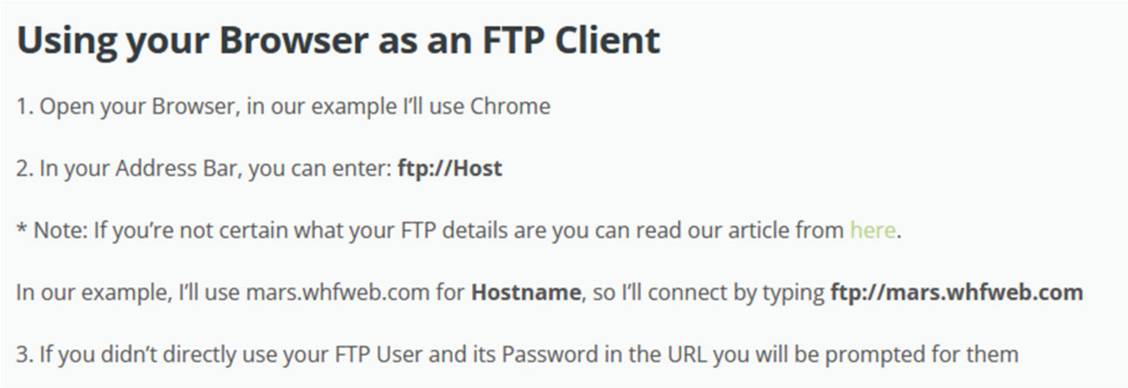
```
Name                                                     Last modified      Size

Parent Directory                                                              -
Annotation_comparison/                                   2023-03-21 10:15    -
GCF_000001405.40_GRCh38.p14_assembly_structure/          2023-03-30 02:36    -
GRCh38_major_release_seqs_for_alignment_pipelines/       2021-03-02 22:14    -
RefSeq_transcripts_alignments/                           2023-03-21 10:15    -
GCF_000001405.40-RS_2023_03_annotation_report.xml        2023-03-21 10:15  280K
GCF_000001405.40_GRCh38.p14_assembly_regions.txt         2023-03-21 10:15   44K
GCF_000001405.40_GRCh38.p14_assembly_report.txt          2023-03-21 10:15   77K
GCF_000001405.40_GRCh38.p14_assembly_stats.txt           2023-03-21 10:15   90K
GCF_000001405.40_GRCh38.p14_cds_from_genomic.fna.gz      2023-03-21 10:15   27M
GCF_000001405.40_GRCh38.p14_feature_count.txt.gz         2023-03-21 10:15  1.5K
GCF_000001405.40_GRCh38.p14_feature_table.txt.gz         2023-03-21 10:15  8.6M
GCF_000001405.40_GRCh38.p14_genomic.fna.gz               2023-03-21 10:15  928M
GCF_000001405.40_GRCh38.p14_genomic.gbff.gz              2023-03-21 10:15  1.2G
GCF_000001405.40_GRCh38.p14_genomic.gff.gz               2023-03-21 10:15   67M
GCF_000001405.40_GRCh38.p14_genomic.gtf.gz               2023-03-21 10:15   49M
GCF_000001405.40_GRCh38.p14_genomic_gaps.txt.gz          2023-03-21 10:15   11K
GCF_000001405.40_GRCh38.p14_protein.faa.gz               2023-03-21 10:15   27M
GCF_000001405.40_GRCh38.p14_protein.gpff.gz              2023-03-21 10:15  158M
GCF_000001405.40_GRCh38.p14_pseudo_without_product.fna.gz 2023-03-21 10:15  8.9M
GCF_000001405.40_GRCh38.p14_rm.out.gz                    2023-03-21 10:15  183M
GCF_000001405.40_GRCh38.p14_rm.run                       2023-03-21 10:15  874
GCF_000001405.40_GRCh38.p14_rna.fna.gz                   2023-03-21 10:15  129M
GCF_000001405.40_GRCh38.p14_rna.gbff.gz                  2023-03-21 10:15  461M
GCF_000001405.40_GRCh38.p14_rna_from_genomic.fna.gz      2023-03-21 10:15   94M
GCF_000001405.40_GRCh38.p14_translated_cds.faa.gz        2023-03-21 10:15   17M
README_GCF_000001405.40-RS_2023_03                       2023-03-21 10:15  2.1K
README_patch_release.txt                                 2023-03-21 10:15  1.5K
all_alt_scaffold_placement.txt                           2023-03-21 10:15   50K
annotation_hashes.txt                                    2023-03-21 10:15  411
assembly_status.txt                                      2023-04-06 05:34   14
md5checksums.txt                                         2023-04-04 13:58  193K
```

HHS Vulnerability Disclosure

Obrázky vytvořil Vratislav Peška

File Transfer Protocol – FTP (FTPS, SFTP)



Using your Browser as an FTP Client

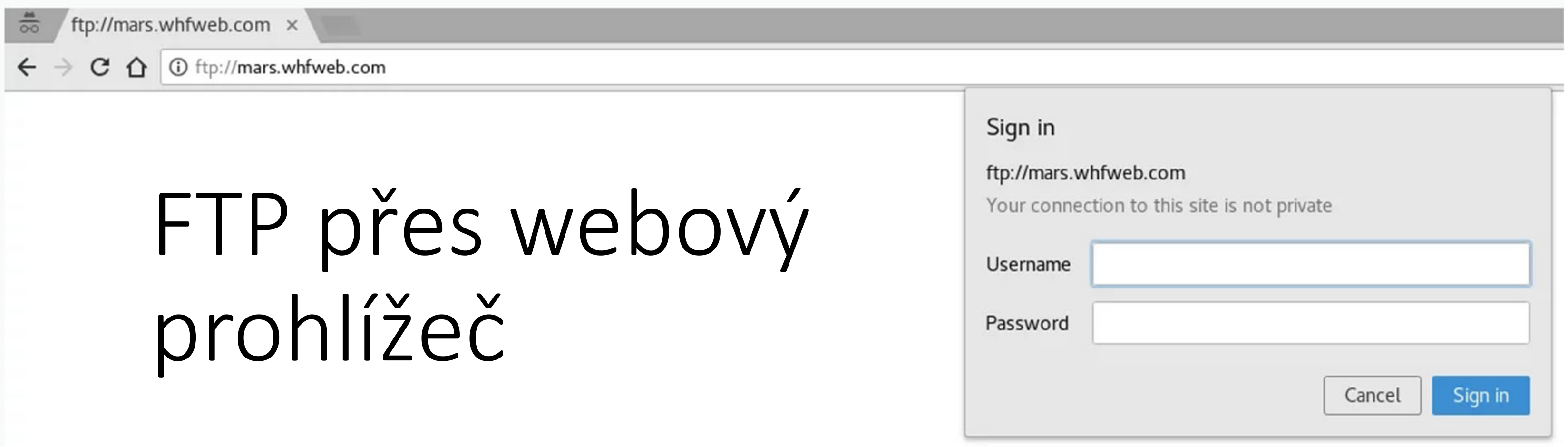
1. Open your Browser, in our example I'll use Chrome

2. In your Address Bar, you can enter: **ftp://Host**

* Note: If you're not certain what your FTP details are you can read our article from here.

In our example, I'll use mars.whfweb.com for **Hostname**, so I'll connect by typing **ftp://mars.whfweb.com**

3. If you didn't directly use your FTP User and its Password in the URL you will be prompted for them



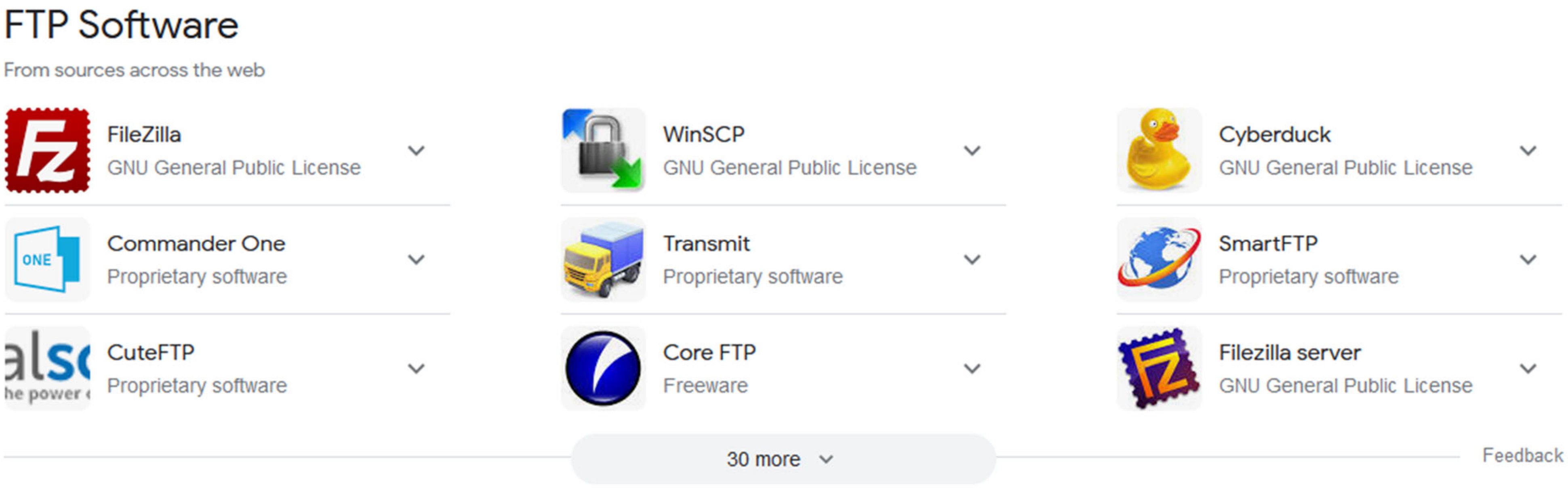
Sign in

ftp://mars.whfweb.com

Your connection to this site is not private

FTP přes webový prohlížeč

Obrázky vytvořil Vratislav Peška

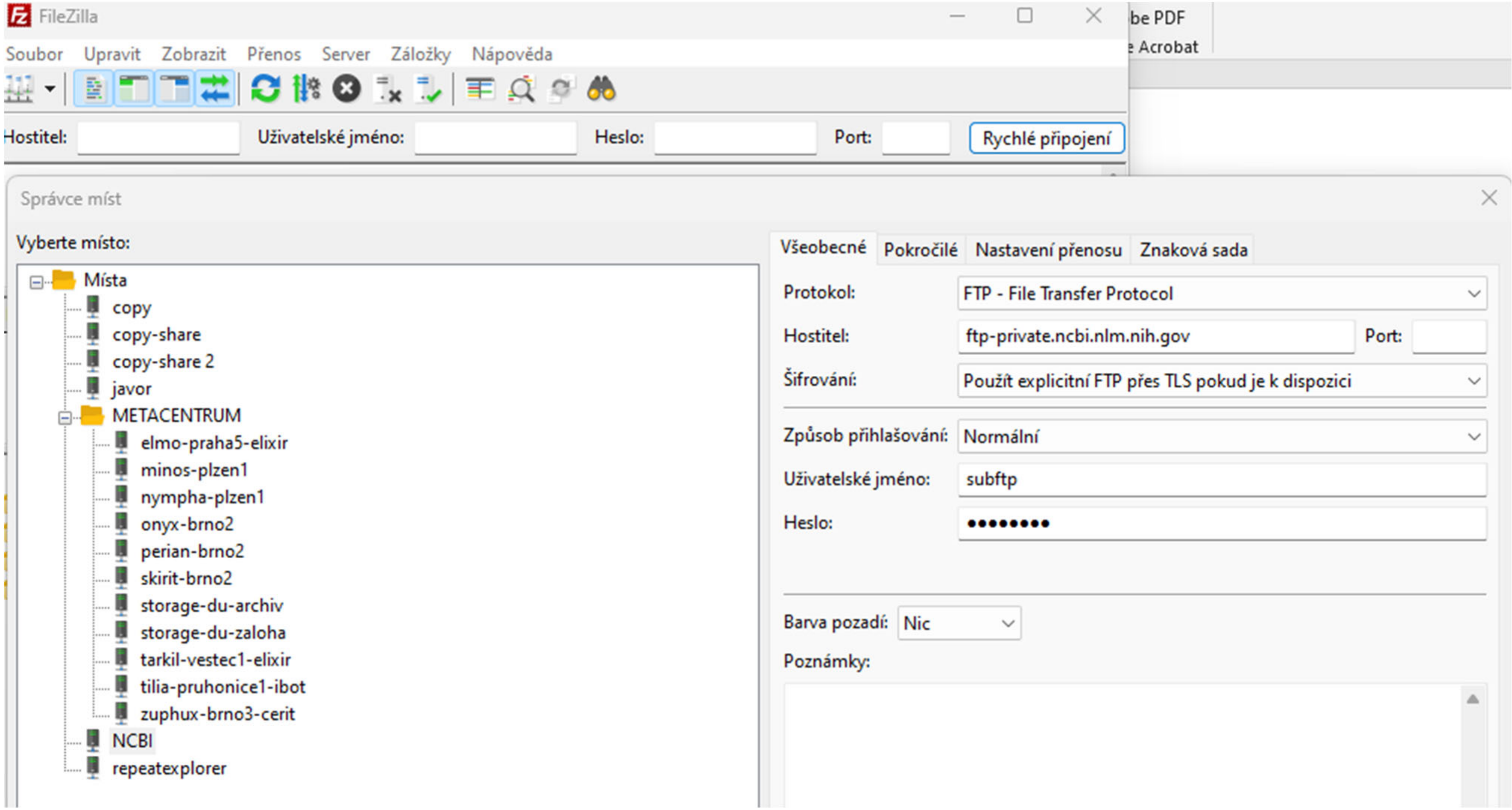

FTP Software

From sources across the web

FileZilla GNU General Public License	WinSCP GNU General Public License	Cyberduck GNU General Public License
Commander One Proprietary software	Transmit Proprietary software	SmartFTP Proprietary software
CuteFTP Proprietary software	Core FTP Freeware	Filezilla server GNU General Public License

Co musím znát:
Hostitel –
Uživatelské jméno –
Heslo –
(port + protokol)

Obrázky vytvořil Vratislav Peška



- Místa
  - copy
  - copy-share
  - copy-share 2
  - javor
  - METACENTRUM
    - elmo-praha5-elixir
    - minos-plzen1
    - nympha-plzen1
    - onyx-brno2
    - perian-brno2
    - skirit-brno2
    - storage-du-archiv
    - storage-du-zaloha
    - tarkil-vestec1-elixir
    - tilia-pruhonice1-ibot
    - zuphux-brno3-cerit
  - NCBI
  - repeatexplorer

Protokol: FTP - File Transfer Protocol

Hostitel: ftp-private.ncbi.nlm.nih.gov Port:

Šifrování: Použít explicitní FTP přes TLS pokud je k dispozici

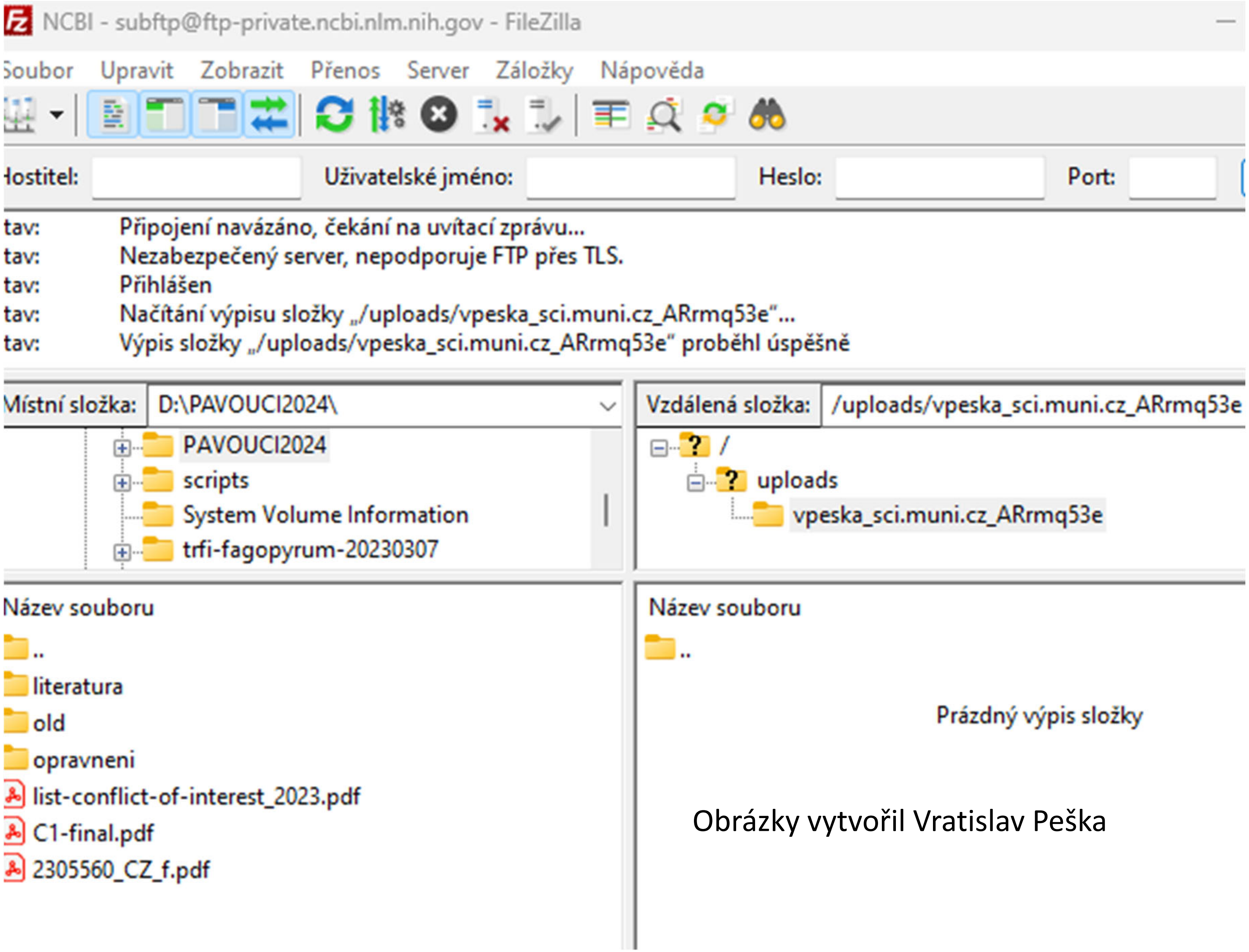
Způsob přihlašování: Normální

Uživatelské jméno: subftp

Heslo: ●●●●●●●●●

Barva pozadí: Nic

Obrázky vytvořil Vratislav Peška



Hostitel: Uživatelské jméno: Heslo: Port:

```
tav:        Připojení navázáno, čekání na uvítací zprávu...
tav:        Nezabezpečený server, nepodporuje FTP přes TLS.
tav:        Přihlášen
tav:        Načítání výpisu složky „/uploads/vpeska_sci.muni.cz_ARrmq53e"...
tav:        Výpis složky „/uploads/vpeska_sci.muni.cz_ARrmq53e" proběhl úspěšně
```

Místní složka: D:\PAVOUCI2024\

- PAVOUCI2024
- scripts
- System Volume Information
- trfi-fagopyrum-20230307

Vzdálená složka: /uploads/vpeska_sci.muni.cz_ARrmq53e

- /
  - uploads
    - vpeska_sci.muni.cz_ARrmq53e

Název souboru

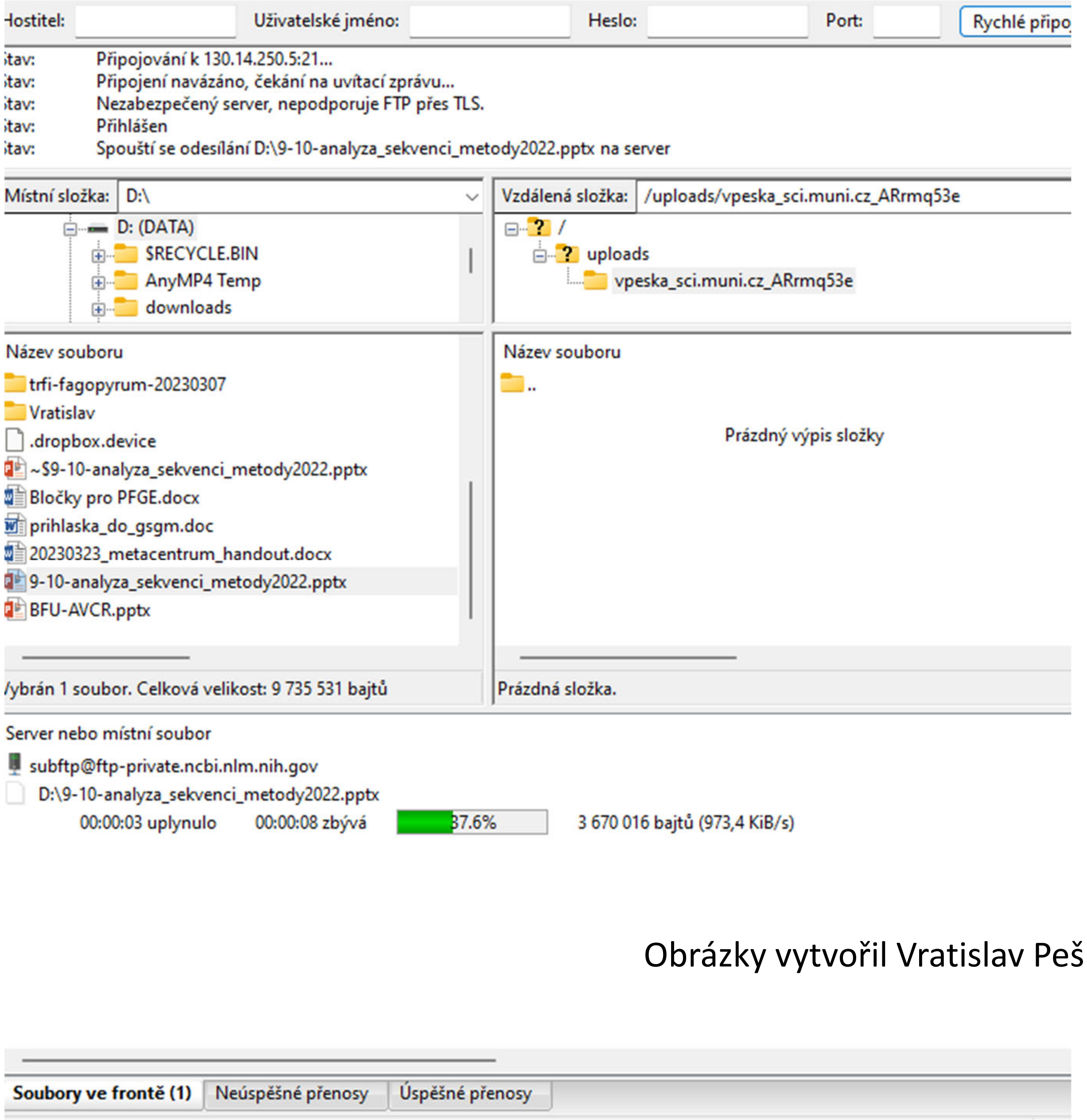
- ..
- literatura
- old
- opravneni
- list-conflict-of-interest_2023.pdf
- C1-final.pdf
- 2305560_CZ_f.pdf

Název souboru

- ..

Prázdný výpis složky

Obrázky vytvořil Vratislav Peška



```
Stav:	Připojování k 130.14.250.5:21...
Stav:	Připojení navázáno, čekání na uvítací zprávu...
Stav:	Nezabezpečený server, nepodporuje FTP přes TLS.
Stav:	Přihlášen
Stav:	Spouští se odesílání D:\9-10-analyza_sekvenci_metody2022.pptx na server
```

Místní složka: D:\

Vzdálená složka: /uploads/vpeska_sci.muni.cz_ARrmq53e

Název souboru
- trfi-fagopyrum-20230307
- Vratislav
- .dropbox.device
- ~$9-10-analyza_sekvenci_metody2022.pptx
- Bločky pro PFGE.docx
- prihlaska_do_gsgm.doc
- 20230323_metacentrum_handout.docx
- 9-10-analyza_sekvenci_metody2022.pptx
- BFU-AVCR.pptx

Název souboru
- ..

Prázdný výpis složky

Vybrán 1 soubor. Celková velikost: 9 735 531 bajtů

Prázdná složka.

Server nebo místní soubor

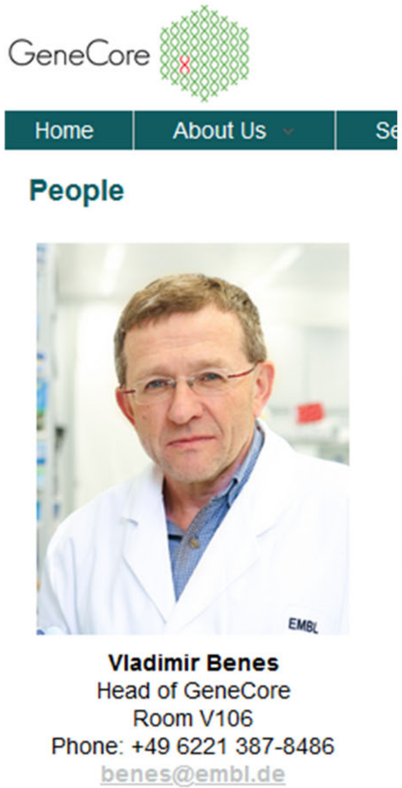
subftp@ftp-private.ncbi.nlm.nih.gov

D:\9-10-analyza_sekvenci_metody2022.pptx

00:00:03 uplynulo 00:00:08 zbývá 37.6% 3 670 016 bajtů (973,4 KiB/s)

Obrázky vytvořil Vratislav Peška

Sekvenační servis nebo centrální laboratoř (CF – Core Facility)



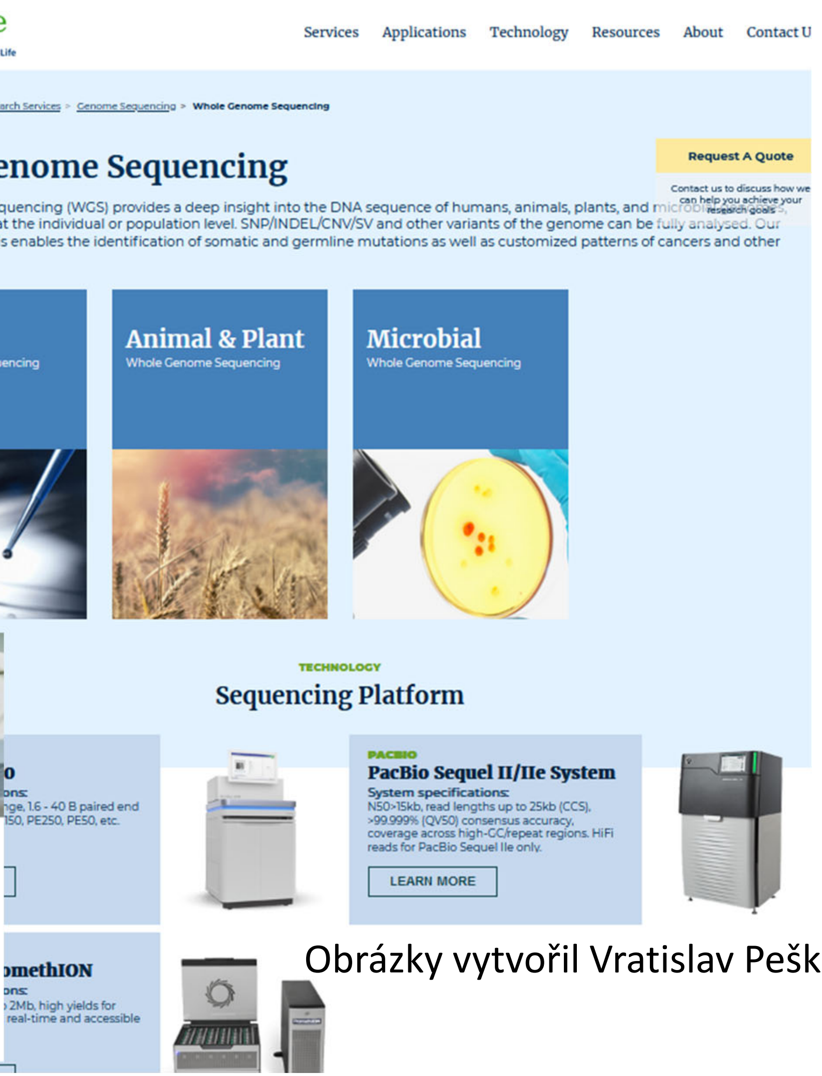
People

Vladimir Benes
Head of GeneCore
Room V106
Phone: +49 6221 387-8486
benes@embl.de

MVDr. Boris Tichý, Ph.D.

Vedoucí sdílené laboratoře

Researcher ID C-1314-2009



Whole Genome Sequencing

Whole Genome Sequencing (WGS) provides a deep insight into the DNA sequence of humans, animals, plants, and microbial genomes, with data analysis at the individual or population level. SNP/INDEL/CNV/SV and other variants of the genome can be fully analysed. Our sequencing analysis enables the identification of somatic and germline mutations as well as customized patterns of cancers and other diseases.

Sequencing Platform

o
ons:
nge, 1.6 - 40 B paired end
150, PE250, PE50, etc.

PACBIO
PacBio Sequel II/IIe System
System specifications:
N50>15kb, read lengths up to 25kb (CCS), >99.999% (QV50) consensus accuracy, coverage across high-GC/repeat regions. HiFi reads for PacBio Sequel IIe only.

omethION
ons:
2Mb, high yields for
real-time and accessible

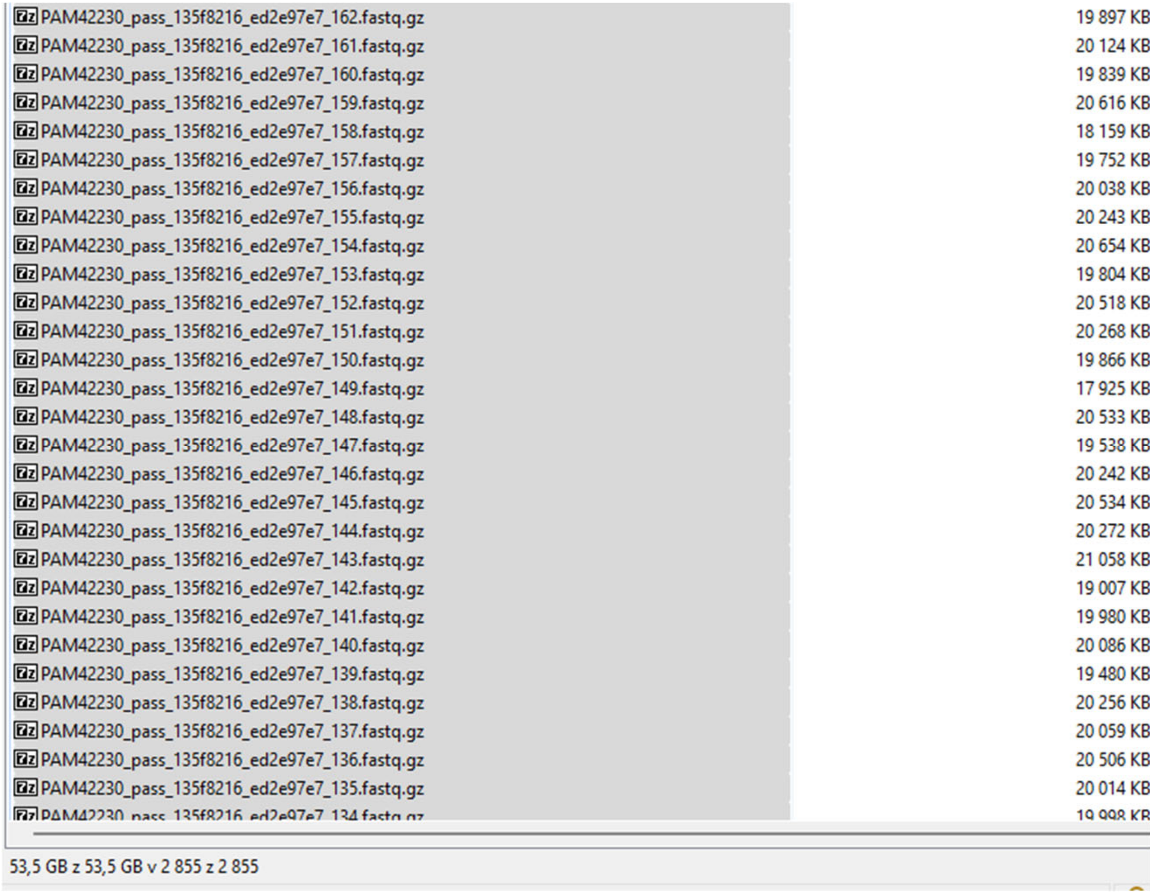
Obrázky vytvořil Vratislav Peška

Jak vypadá produkt poskytnutý sekvenačním servisem nebo centrální laboratoří (CF – core facility)

Obrázky vytvořil Vratislav Peška

..	
TEL103_S11_L002_R2_001.fastq.gz	1 597 274 KB
TEL103_S11_L002_R1_001.fastq.gz	1 506 485 KB
TEL102_S10_L001_R2_001.fastq.gz	3 567 597 KB
TEL102_S10_L001_R1_001.fastq.gz	3 392 515 KB
TEL100_S9_L001_R2_001.fastq.gz	2 088 595 KB
TEL100_S9_L001_R1_001.fastq.gz	2 013 927 KB
TEL97_S14_L002_R2_001.fastq.gz	1 443 726 KB
TEL97_S14_L002_R1_001.fastq.gz	1 385 953 KB
SME_S6_L001_R2_001.fastq.gz	1 980 845 KB
SME_S6_L001_R1_001.fastq.gz	1 881 276 KB
SCENE_S13_L002_R2_001.fastq.gz	1 948 463 KB
SCENE_S13_L002_R1_001.fastq.gz	1 847 953 KB
RACR_S7_L001_R2_001.fastq.gz	1 737 364 KB
RACR_S7_L001_R1_001.fastq.gz	1 662 259 KB
PYR_S20_L002_R2_001.fastq.gz	5 008 736 KB
PYR_S20_L002_R1_001.fastq.gz	4 754 863 KB
PpWT1_S1_L001_R2_001.fastq.gz	2 933 673 KB
PpWT1_S1_L001_R1_001.fastq.gz	2 712 568 KB
PARA_S12_L002_R2_001.fastq.gz	1 945 228 KB
PARA_S12_L002_R1_001.fastq.gz	1 835 490 KB
P_S16_L002_R2_001.fastq.gz	1 987 326 KB
P_S16_L002_R1_001.fastq.gz	1 886 706 KB
L_S17_L002_R2_001.fastq.gz	1 692 458 KB
L_S17_L002_R1_001.fastq.gz	1 609 402 KB
IPS_S19_L002_R2_001.fastq.gz	4 271 811 KB
IPS_S19_L002_R1_001.fastq.gz	4 085 774 KB
CHX302_S8_L001_R2_001.fastq.gz	2 363 943 KB
CHX302_S8_L001_R1_001.fastq.gz	2 226 033 KB
H_S18_L002_R2_001.fastq.gz	1 635 235 KB
H_S18_L002_R1_001.fastq.gz	1 559 576 KB
FTAR_S5_L001_R2_001.fastq.gz	2 784 510 KB
FTAR_S5_L001_R1_001.fastq.gz	2 675 555 KB
FESC_S15_L002_R2_001.fastq.gz	1 756 909 KB
FESC_S15_L002_R1_001.fastq.gz	1 691 095 KB
DKC_MOCK_S4_L001_R2_001.fastq.gz	1 444 264 KB
DKC_MOCK_S4_L001_R1_001.fastq.gz	1 318 614 KB
DKC_IP7_S3_L001_R2_001.fastq.gz	2 597 897 KB
DKC_IP7_S3_L001_R1_001.fastq.gz	2 380 177 KB
DKC_IP6_S2_L001_R2_001.fastq.gz	1 933 748 KB

- Přes webové rozhraní
- Externí disk
- FTP
- Filesender
- Aspera
- Etc.

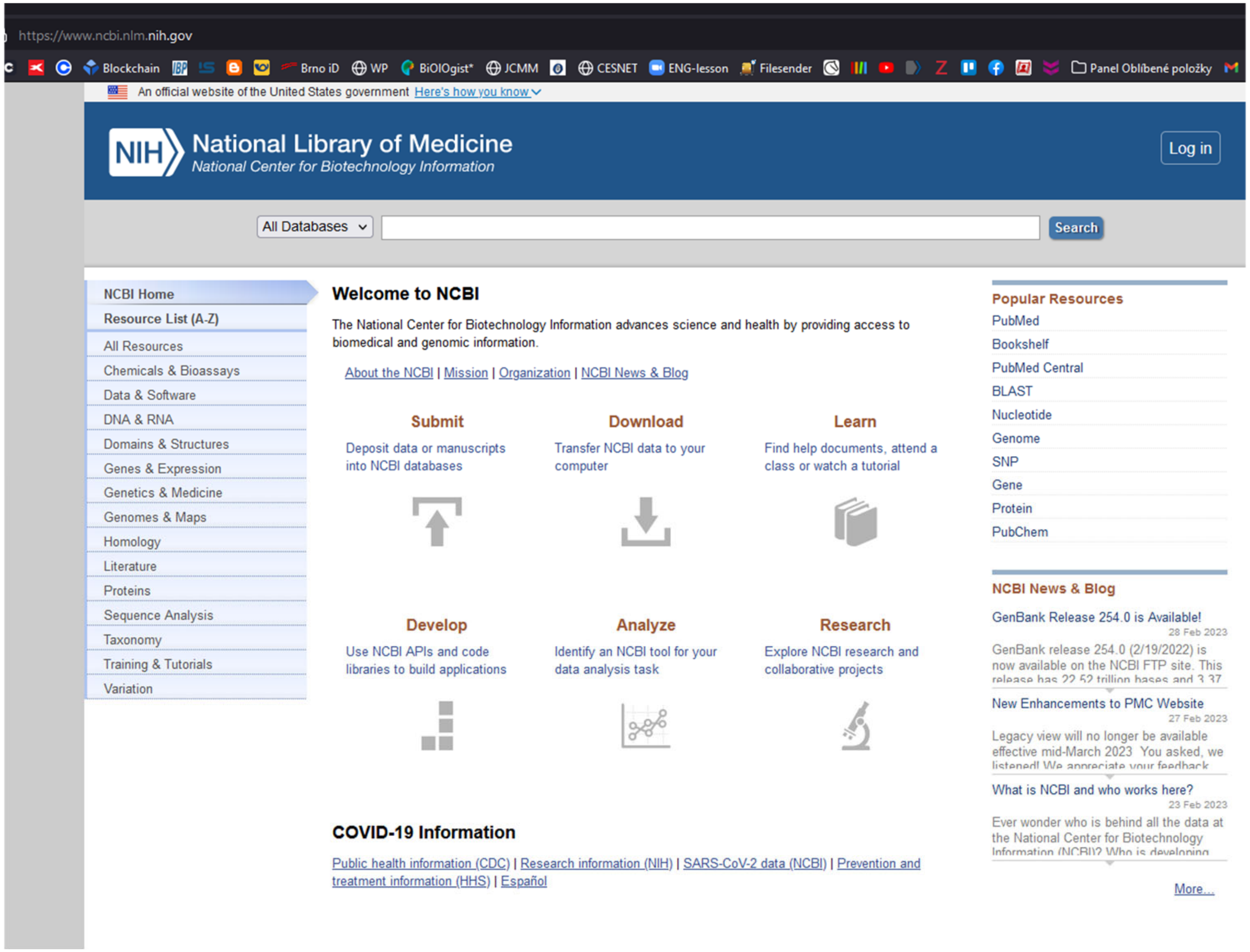


PAM42230_pass_135f8216_ed2e97e7_162.fastq.gz	19 897 KB
PAM42230_pass_135f8216_ed2e97e7_161.fastq.gz	20 124 KB
PAM42230_pass_135f8216_ed2e97e7_160.fastq.gz	19 839 KB
PAM42230_pass_135f8216_ed2e97e7_159.fastq.gz	20 616 KB
PAM42230_pass_135f8216_ed2e97e7_158.fastq.gz	18 159 KB
PAM42230_pass_135f8216_ed2e97e7_157.fastq.gz	19 752 KB
PAM42230_pass_135f8216_ed2e97e7_156.fastq.gz	20 038 KB
PAM42230_pass_135f8216_ed2e97e7_155.fastq.gz	20 243 KB
PAM42230_pass_135f8216_ed2e97e7_154.fastq.gz	20 654 KB
PAM42230_pass_135f8216_ed2e97e7_153.fastq.gz	19 804 KB
PAM42230_pass_135f8216_ed2e97e7_152.fastq.gz	20 518 KB
PAM42230_pass_135f8216_ed2e97e7_151.fastq.gz	20 268 KB
PAM42230_pass_135f8216_ed2e97e7_150.fastq.gz	19 866 KB
PAM42230_pass_135f8216_ed2e97e7_149.fastq.gz	17 925 KB
PAM42230_pass_135f8216_ed2e97e7_148.fastq.gz	20 533 KB
PAM42230_pass_135f8216_ed2e97e7_147.fastq.gz	19 538 KB
PAM42230_pass_135f8216_ed2e97e7_146.fastq.gz	20 242 KB
PAM42230_pass_135f8216_ed2e97e7_145.fastq.gz	20 534 KB
PAM42230_pass_135f8216_ed2e97e7_144.fastq.gz	20 272 KB
PAM42230_pass_135f8216_ed2e97e7_143.fastq.gz	21 058 KB
PAM42230_pass_135f8216_ed2e97e7_142.fastq.gz	19 007 KB
PAM42230_pass_135f8216_ed2e97e7_141.fastq.gz	19 980 KB
PAM42230_pass_135f8216_ed2e97e7_140.fastq.gz	20 086 KB
PAM42230_pass_135f8216_ed2e97e7_139.fastq.gz	19 480 KB
PAM42230_pass_135f8216_ed2e97e7_138.fastq.gz	20 256 KB
PAM42230_pass_135f8216_ed2e97e7_137.fastq.gz	20 059 KB
PAM42230_pass_135f8216_ed2e97e7_136.fastq.gz	20 506 KB
PAM42230_pass_135f8216_ed2e97e7_135.fastq.gz	20 014 KB
PAM42230_pass_135f8216_ed2e97e7_134.fastq.gz	19 998 KB

53,5 GB z 53,5 GB v 2 855 z 2 855

Obrázky vytvořil Vratislav Peška

Nahrávání hrubých NGS dat do SRA NCBI



An official website of the United States government Here's how you know

National Library of Medicine

National Center for Biotechnology Information

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

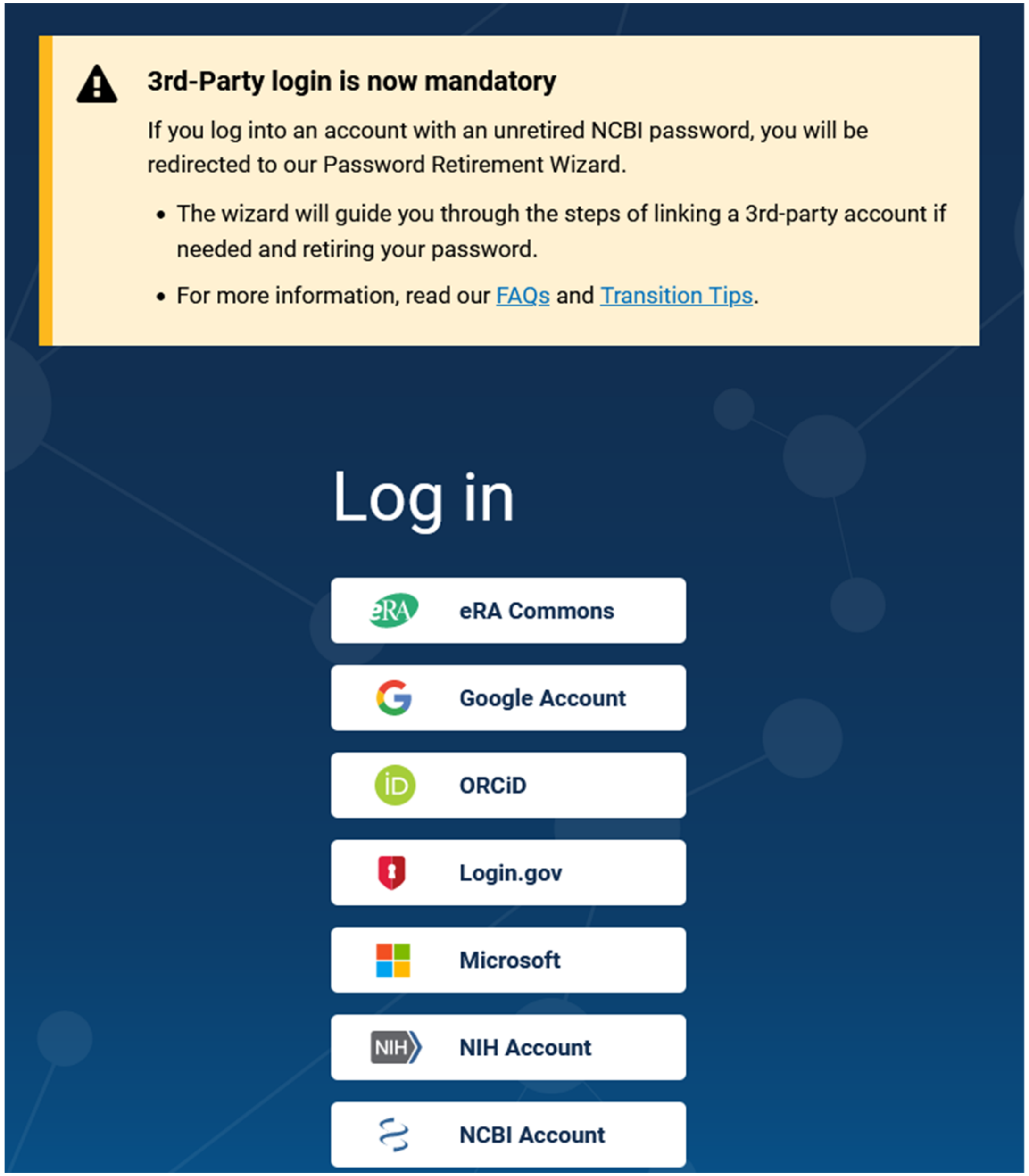
NCBI News & Blog

GenBank Release 254.0 is Available!
28 Feb 2023
GenBank release 254.0 (2/19/2022) is now available on the NCBI FTP site. This release has 22.52 trillion bases and 3.37

New Enhancements to PMC Website
27 Feb 2023
Legacy view will no longer be available effective mid-March 2023. You asked, we listened! We appreciate your feedback

What is NCBI and who works here?
23 Feb 2023
Ever wonder who is behind all the data at the National Center for Biotechnology Information (NCBI)? Who is developing

More...

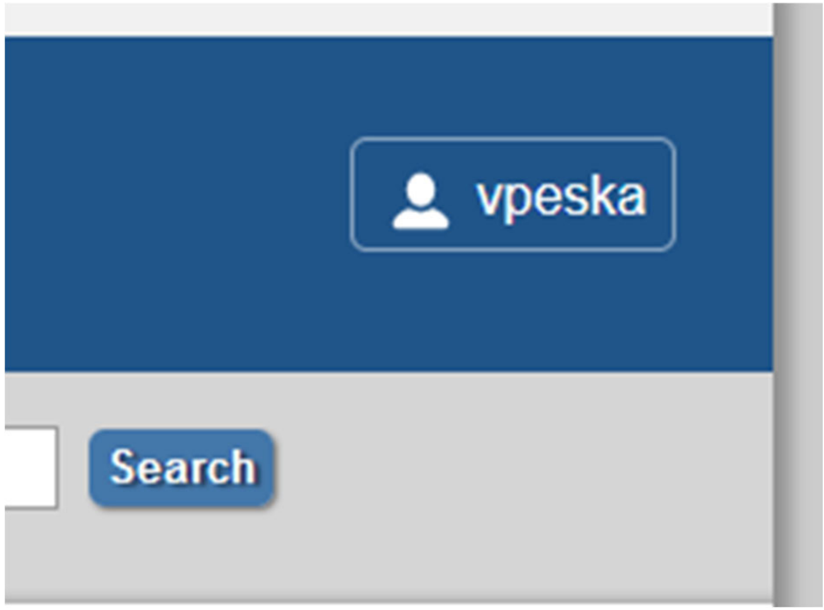


3rd-Party login is now mandatory

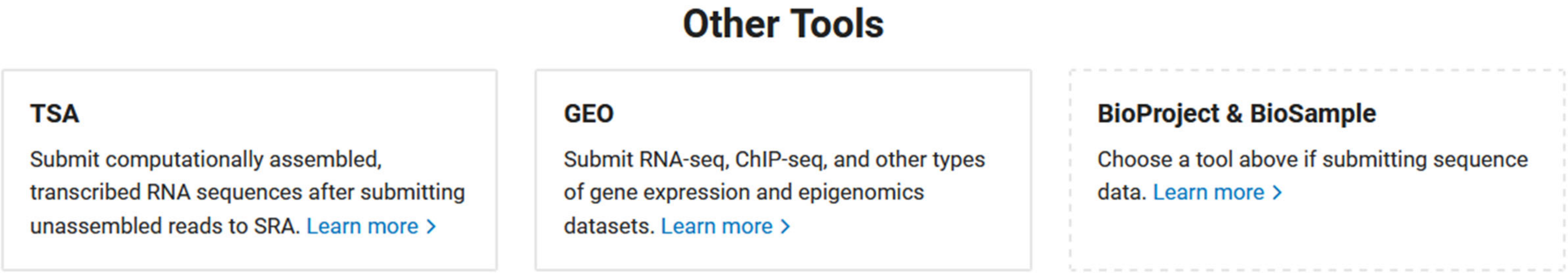
If you log into an account with an unretired NCBI password, you will be redirected to our Password Retirement Wizard.

- The wizard will guide you through the steps of linking a 3rd-party account if needed and retiring your password.
- For more information, read our FAQs and Transition Tips.

Log in



Obrázky vytvořil Vratislav Peška



Other Tools

TSA

Submit computationally assembled, transcribed RNA sequences after submitting unassembled reads to SRA. Learn more ›

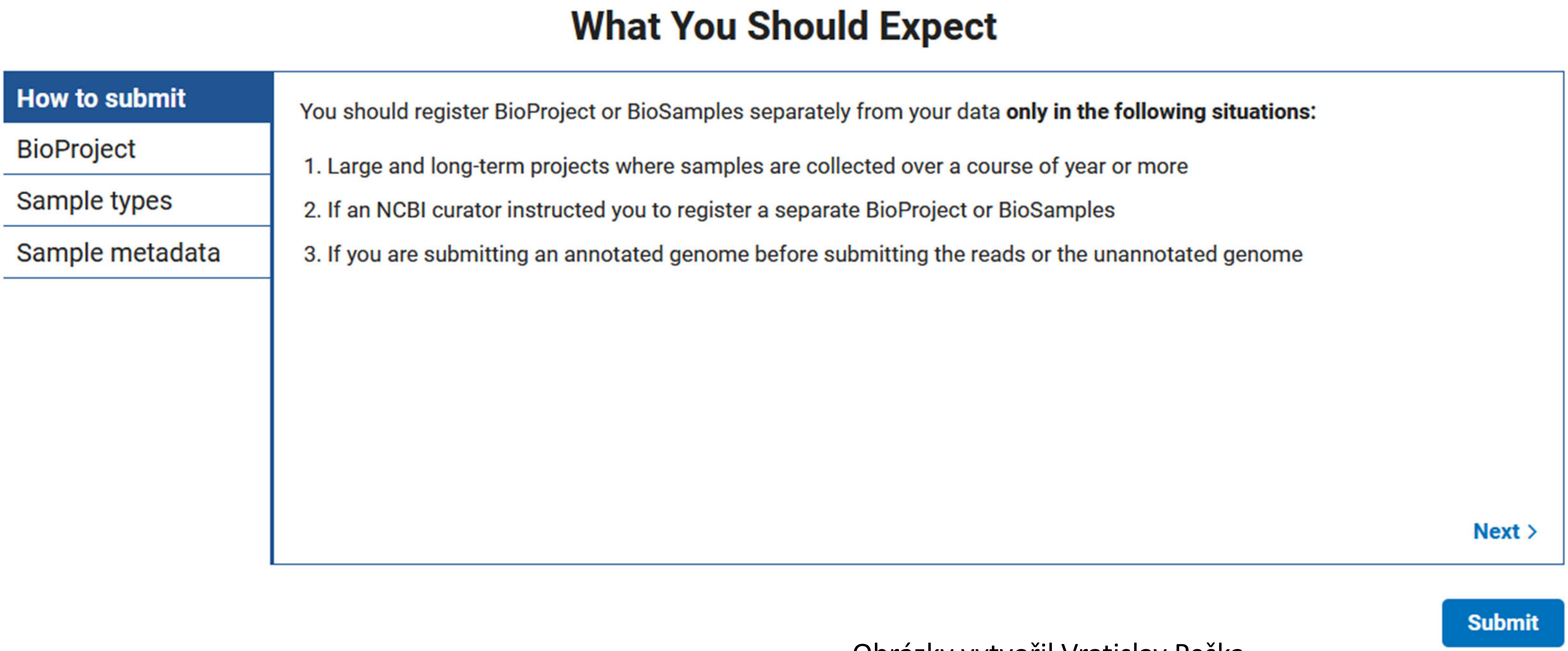
GEO

Submit RNA-seq, ChIP-seq, and other types of gene expression and epigenomics datasets. Learn more ›

BioProject & BioSample

Choose a tool above if submitting sequence data. Learn more ›

Obrázky vytvořil Vratislav Peška

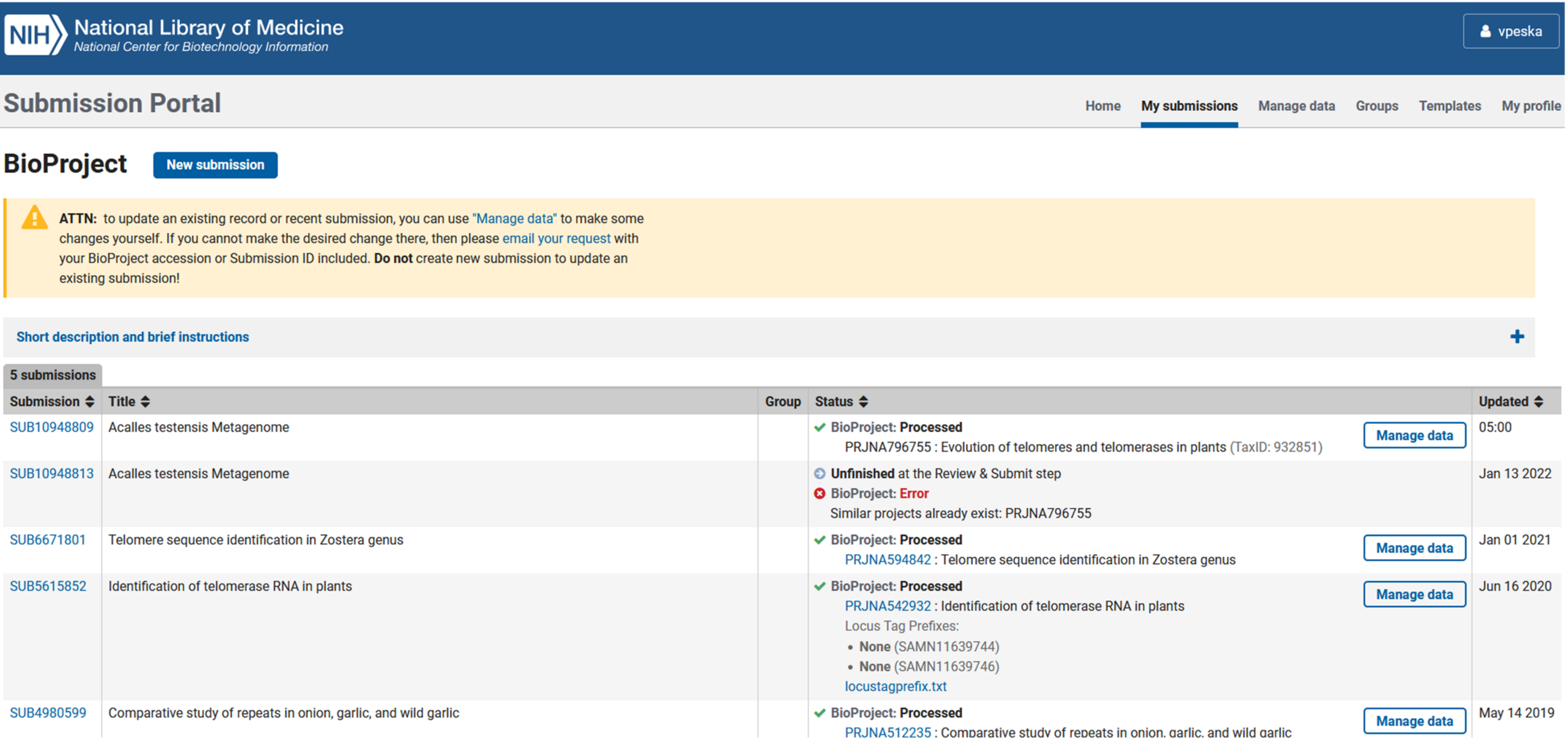


What You Should Expect

You should register BioProject or BioSamples separately from your data **only in the following situations:**

1. Large and long-term projects where samples are collected over a course of year or more
2. If an NCBI curator instructed you to register a separate BioProject or BioSamples
3. If you are submitting an annotated genome before submitting the reads or the unannotated genome

Obrázky vytvořil Vratislav Peška



ATTN: to update an existing record or recent submission, you can use "Manage data" to make some changes yourself. If you cannot make the desired change there, then please email your request with your BioProject accession or Submission ID included. **Do not** create new submission to update an existing submission!

Submission	Title	Group	Status		Updated
SUB10948809	Acalles testensis Metagenome		✓ BioProject: **Processed** PRJNA796755 : Evolution of telomeres and telomerases in plants (TaxID: 932851)	Manage data	05:00
SUB10948813	Acalles testensis Metagenome		**Unfinished** at the Review & Submit step BioProject: Error Similar projects already exist: PRJNA796755		Jan 13 2022
SUB6671801	Telomere sequence identification in Zostera genus		✓ BioProject: **Processed** PRJNA594842 : Telomere sequence identification in Zostera genus	Manage data	Jan 01 2021
SUB5615852	Identification of telomerase RNA in plants		✓ BioProject: **Processed** PRJNA542932 : Identification of telomerase RNA in plants Locus Tag Prefixes: • **None** (SAMN11639744) • **None** (SAMN11639746) locustagprefix.txt	Manage data	Jun 16 2020
SUB4980599	Comparative study of repeats in onion, garlic, and wild garlic		✓ BioProject: **Processed** PRJNA512235 : Comparative study of repeats in onion, garlic, and wild garlic	Manage data	May 14 2019

Obrázky vytvořil Vratislav Peška

Formát sekvenčních dat

- FASTQ
- FASTA
- SAM/BAM
- GFF
- BED
- VCF

Kontrolní otázka:

100 řádků FASTQ, kolik je to readů a co z toho vyplývá ohledně párovosti dat?

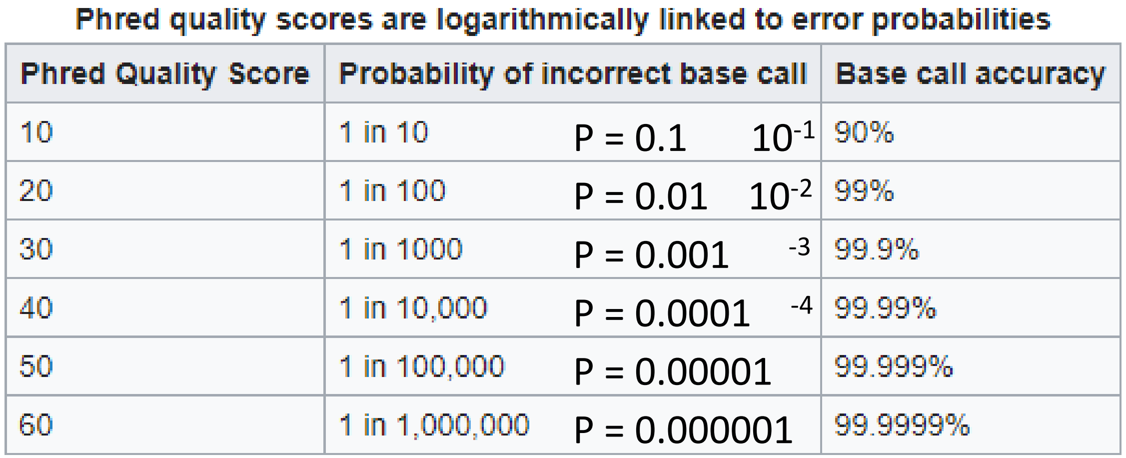
FASTQ [fa:stkju:]

- Nukleotidová sekvence + odpovídající Phred quality scores (*Q*)
- textový formát (4 řádky/1sekvence)
- přípona obvykle *.fq   nebo *.fastq (často GNU zip, *.fastq.gz)
- sekvence + kvalita zastoupena jedním znakem ASCII
- Malý soubor v notepadu, velký soubor nahlížení pomocí less/more, head/tail (linux terminal – lt)
- FASTQC / MULTIQC

Obrázky vytvořil Vratislav Peška

Phred score (*Q*)

- 10x záporný logaritmus o základu 10 pravděpodobnosti chyby



Phred quality scores are logarithmically linked to error probabilities

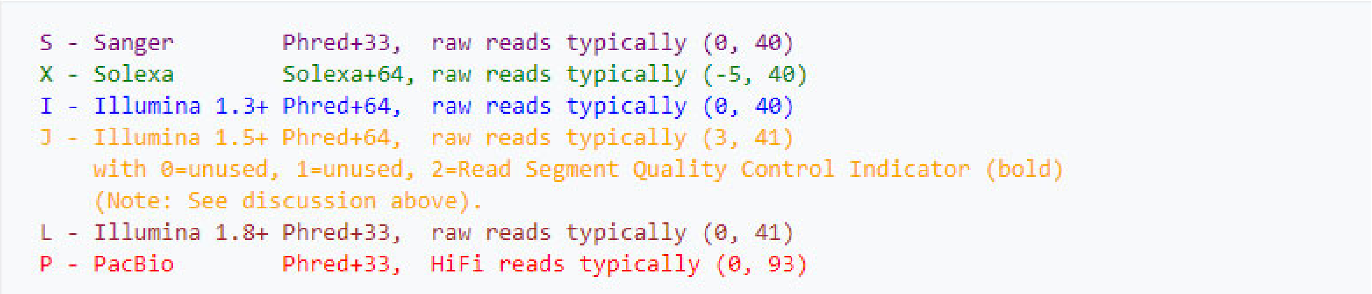
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10 P = 0.1 10^{-1}	90%
20	1 in 100 P = 0.01 10^{-2}	99%
30	1 in 1000 P = 0.001 $^{-3}$	99.9%
40	1 in 10,000 P = 0.0001 $^{-4}$	99.99%
50	1 in 100,000 P = 0.00001	99.999%
60	1 in 1,000,000 P = 0.000001	99.9999%

Obrázky vytvořil Vratislav Peška

$$Q = -10 \log_{10} P \qquad x = \log_a y \qquad a^x = y$$

Převodní tabulka různých *Qs*

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
0.2......................26...31........41
0..................20........30........40........50..........................................93
```



```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Obrázky vytvořil Vratislav Peška

Hlavičky readů

```
@HWI-ST193:542:C2H0GACXX:8:1101:4404:2179 1:Y:0:ACACGA
ATGCNTTTTATAATCAAAGCAGAAGCTTTATGCTAGCTAGCATATAAT
+
<<<<<#2<:5>:944>:??AAAAAAAAABAAAABBBBB??????????
```

`@` – začátek readů (hlavička)

`@HWI-ST193` – Jméno stroje (od výrobce)

`542` – run ID

`C2H0GACXX` – flowcell id

`8` – flowcell lane

`1101` – tile number

`4404` – x-coordinate

`2179` – y-coordinate

`1` – část páru

`Y` – filter status (Y/N)

`0` – status kontrol. Bit. (0 nebo sudá čísla)

`ACACGA` - sekvence indexu

FASTQC

- První náhled na NGS data (short reads, illumina)
- Výstup - HTML formát
- Možnost spouštět z terminalu (linux) i v rámci GALAXY (web)

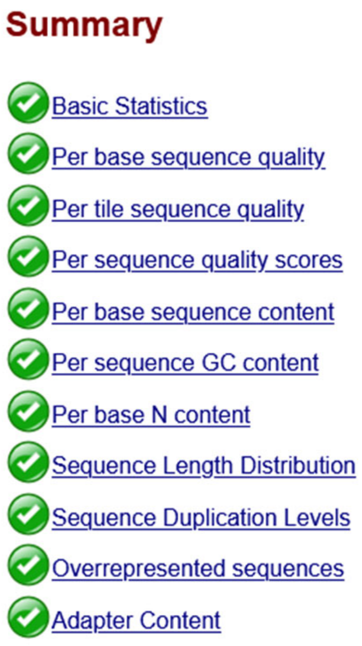
```
@A00180:10:H7NK5DMXX:1:1101:16821:1000 1:N:0:GGACTT+GTCGTTCG
TGCAGCAGCTAATGAGGAACCACTTCCTCCCTCCAGCCGCTCTAAATACCTCAGAACAATAGGATCATCATAATAATCCCCTAGTCTGAACTG
+
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19090:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTTTCATATCGAGATTTATATAGAATCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19325:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATACAGGCGTTCCAACCTATTTTAGTGG
+
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-F-FFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:30897:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAAAGTATTACTCCCAGTCAATATAAGG

+
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19090:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTTTCATATCGAGATTTATATAGAATCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19325:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATACAGGCGTTCCAACCTATTTTAGTGG
+
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-F-FFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:30897:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAAAGTATTACTCCCAGTCAATATAAGG
+
8F888FFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFF
```

Obrázky vytvořil Vratislav Peška

FASTQC summary + basic statistics

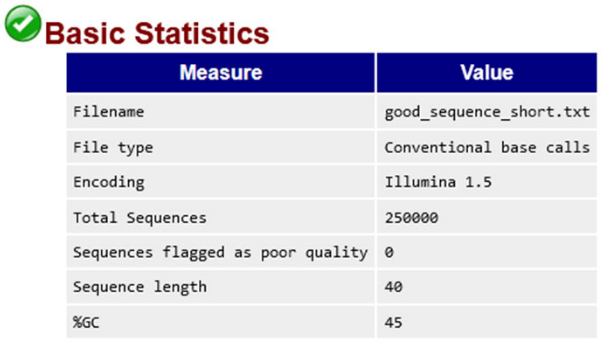
Dobrá data



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

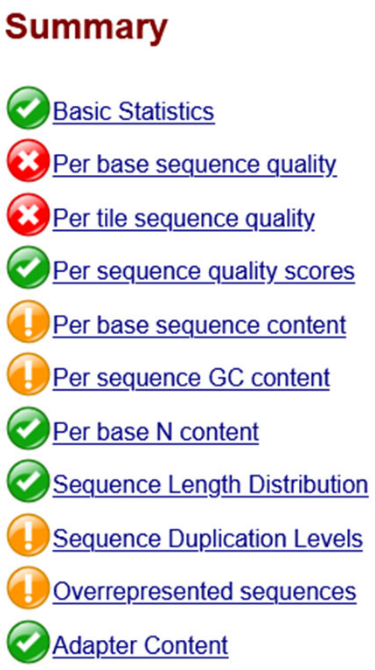
Analýza (mapování, sestavování genomu)



✓ **Basic Statistics**

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

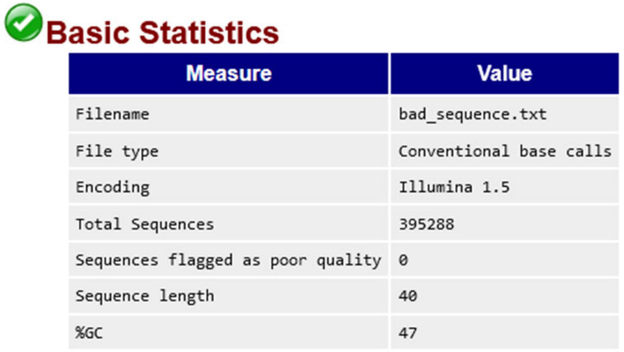
Špatná data



Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content

Pre-processing
Trimming
Filtrování



✓ **Basic Statistics**

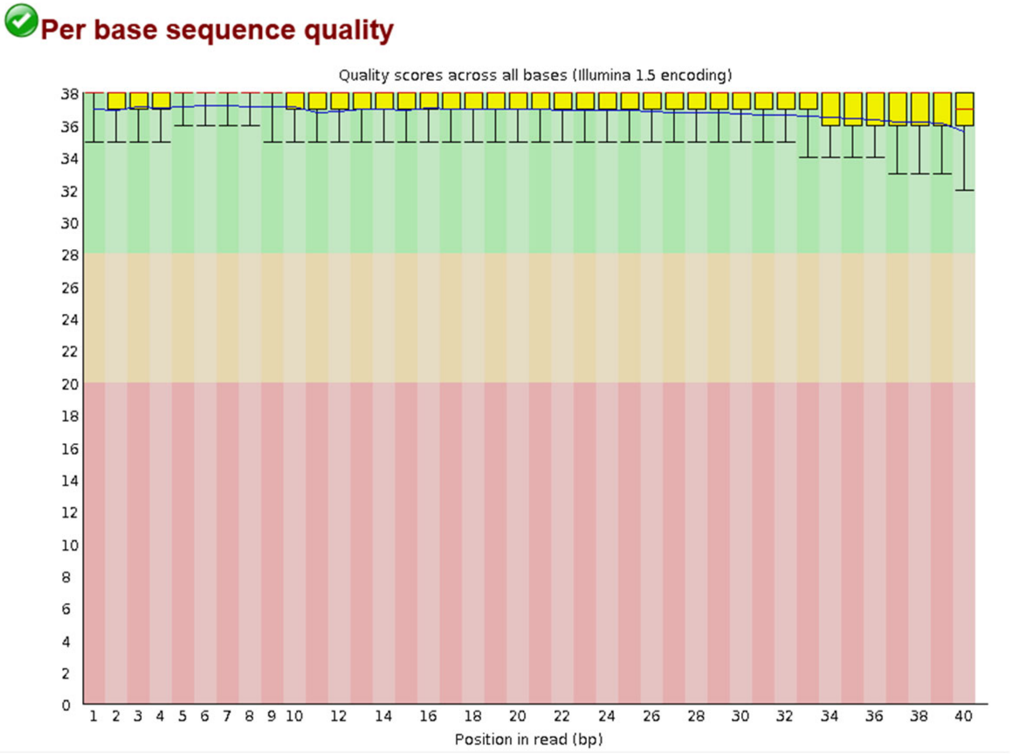
Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

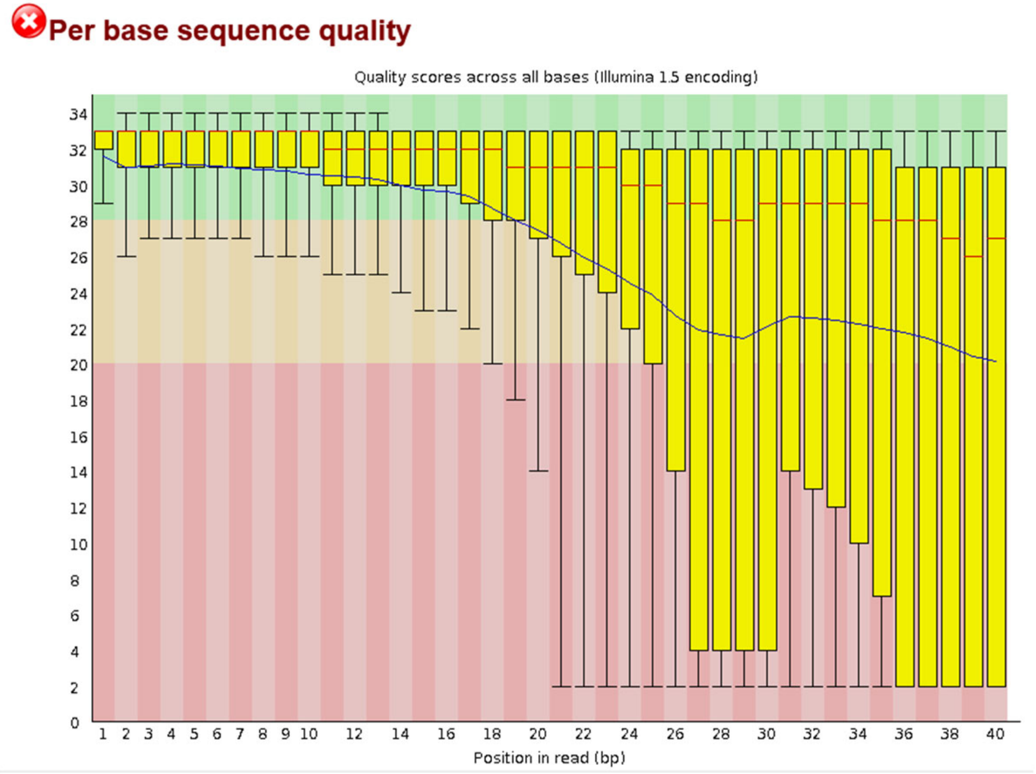
Obrázky vytvořil Vratislav Peška

Per base sequence quality

Dobrá data

Špatná data

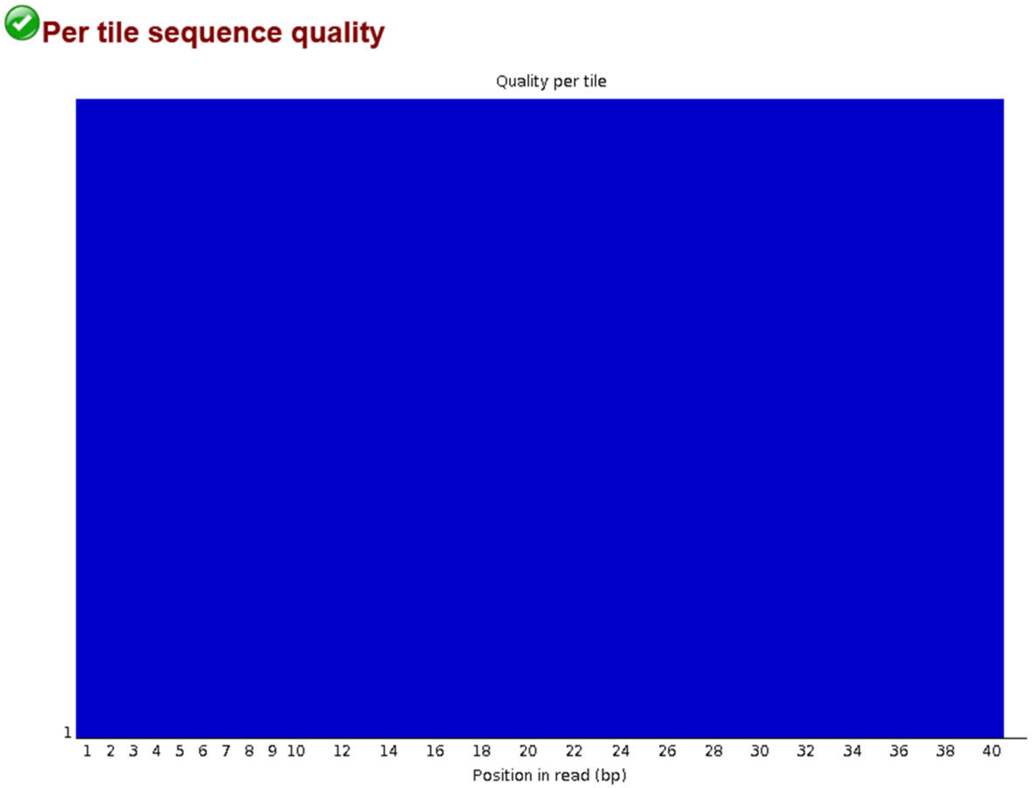




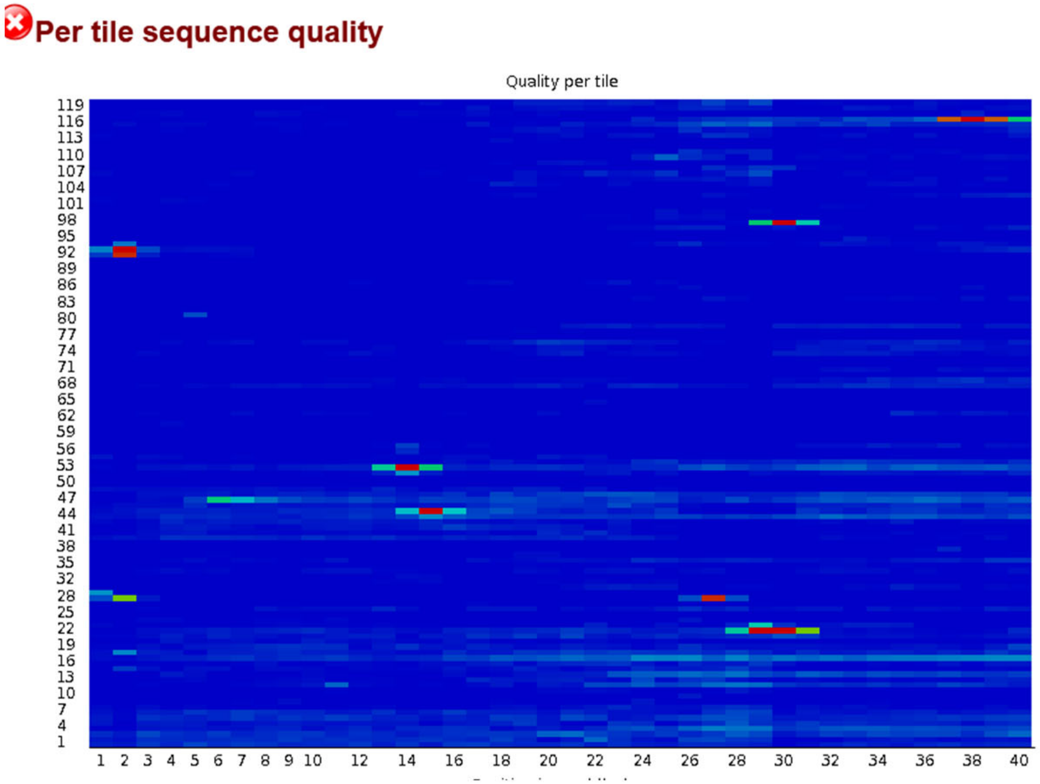
Obrázky vytvořil Vratislav Peška

Per tile sequence quality

Dobrá data



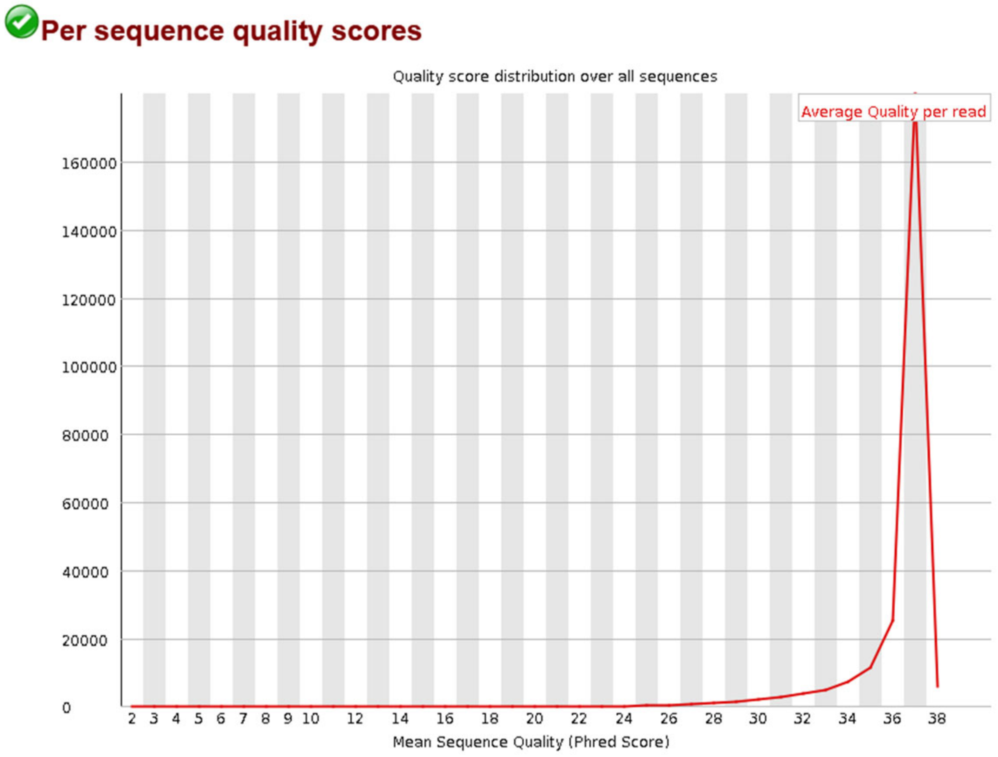
Špatná data



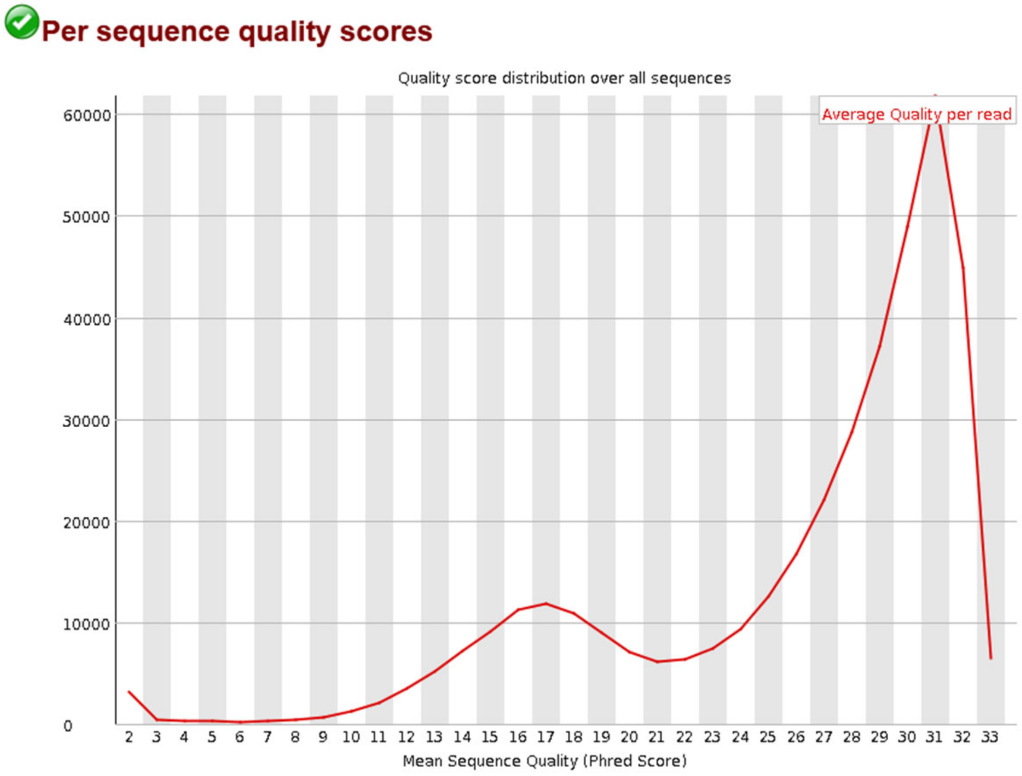
Obrázky vytvořil Vratislav Peška

Per sequence quality scores

Dobrá data



Špatná data



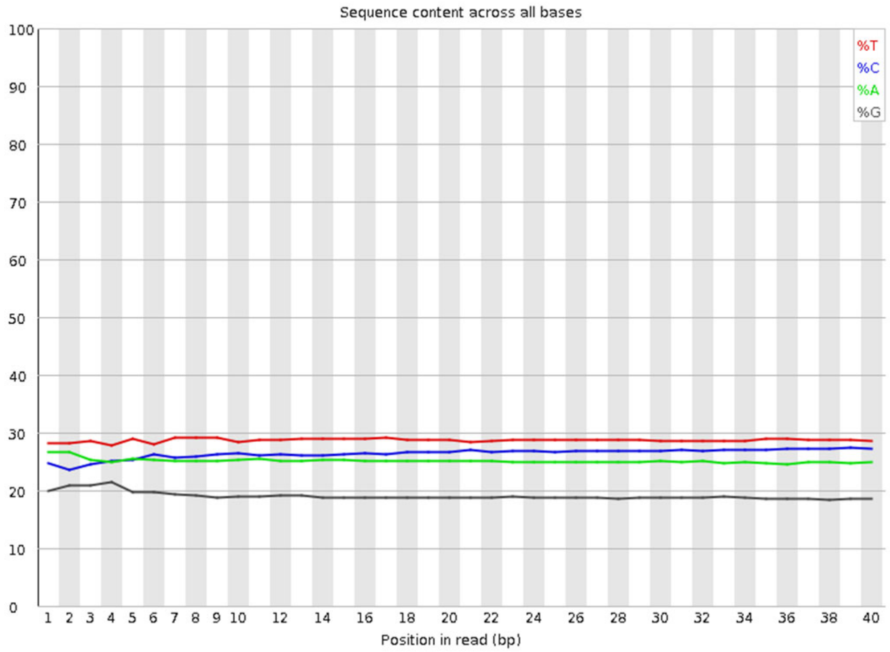
Obrázky vytvořil Vratislav Peška

Per base sequence content

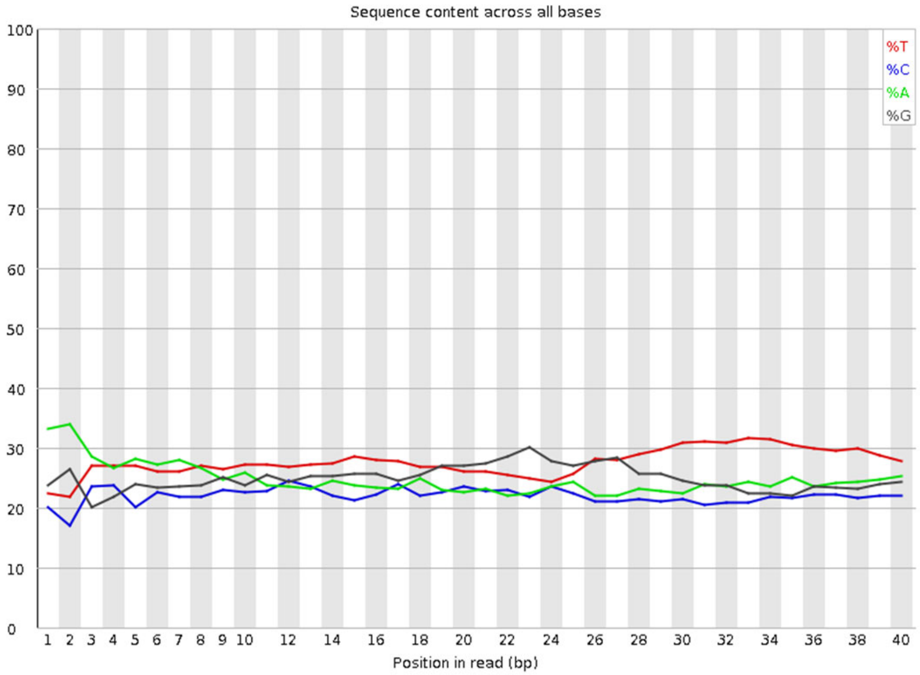
Dobrá data

Špatná data

✓ **Per base sequence content**



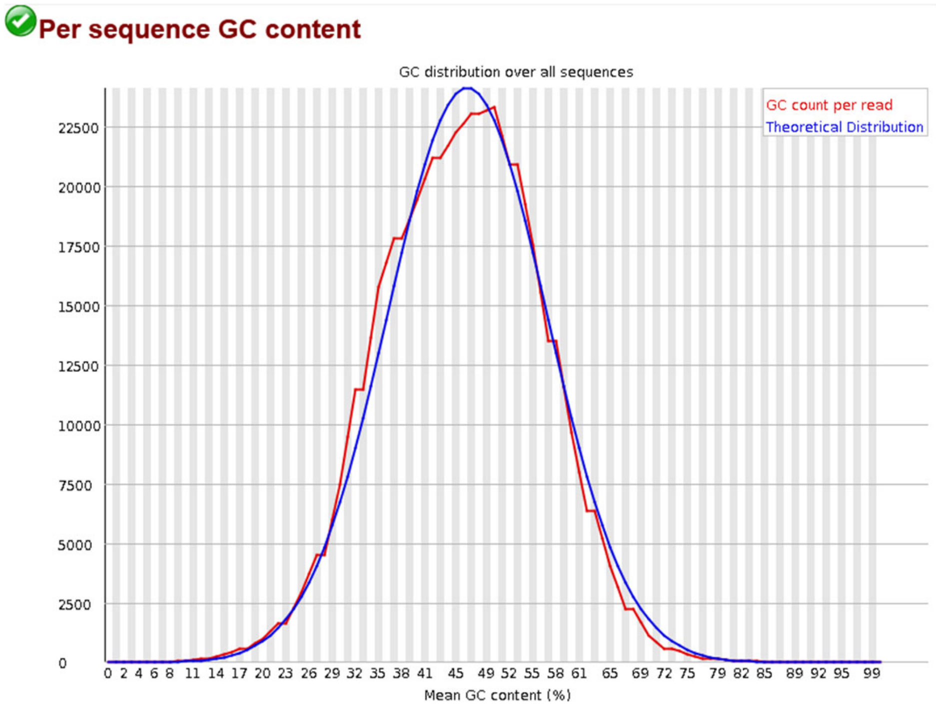
! **Per base sequence content**



Obrázky vytvořil Vratislav Peška

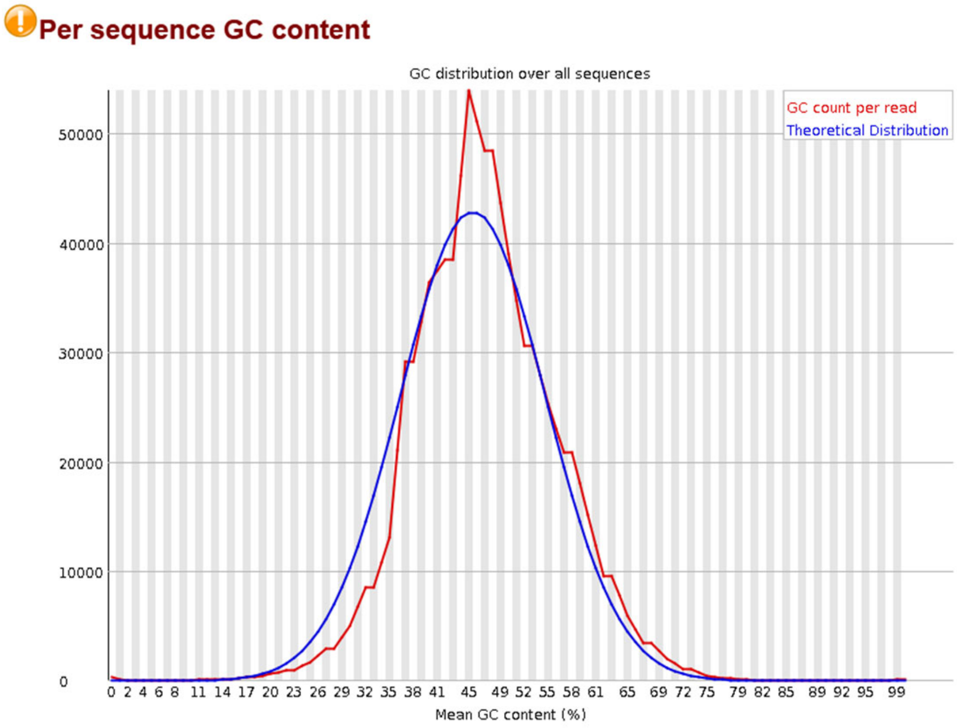
Per sequence GC content

Dobrá data



Per sequence GC content

Špatná data

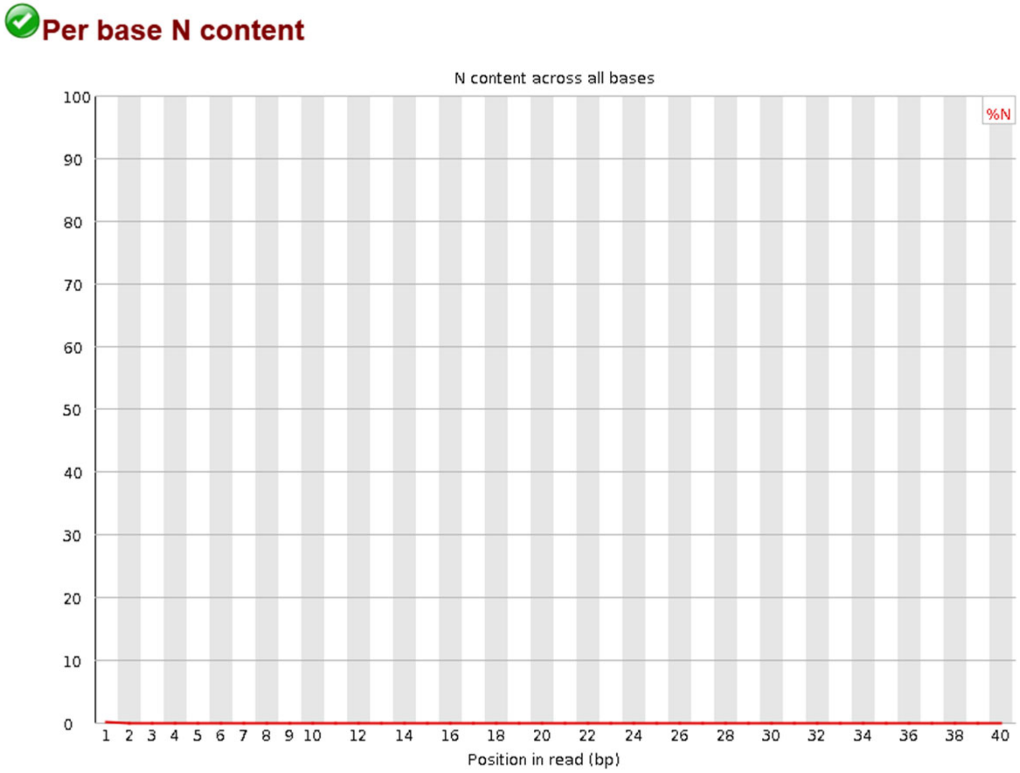


Per sequence GC content

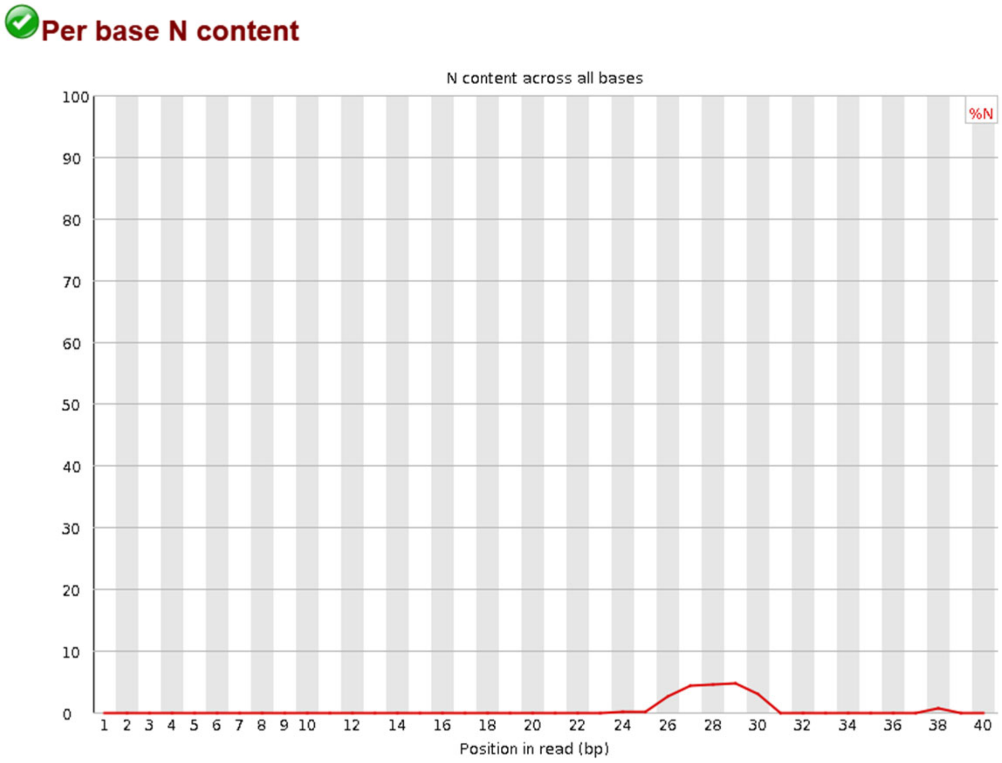
Obrázky vytvořil Vratislav Peška

Per base N content

Dobrá data



Špatná data

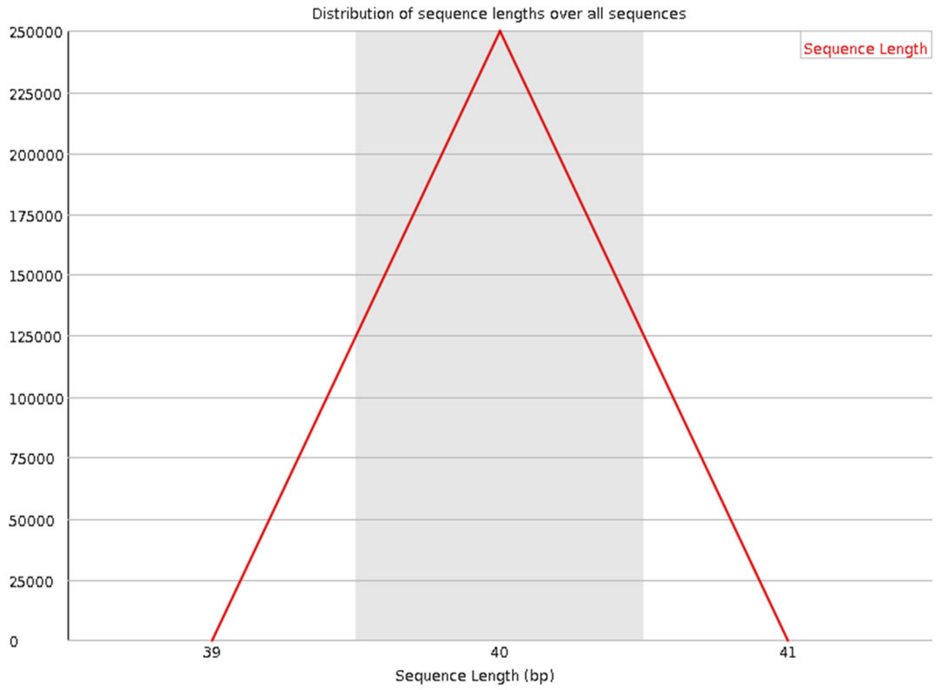


Obrázky vytvořil Vratislav Peška

Sequence Length Distribution

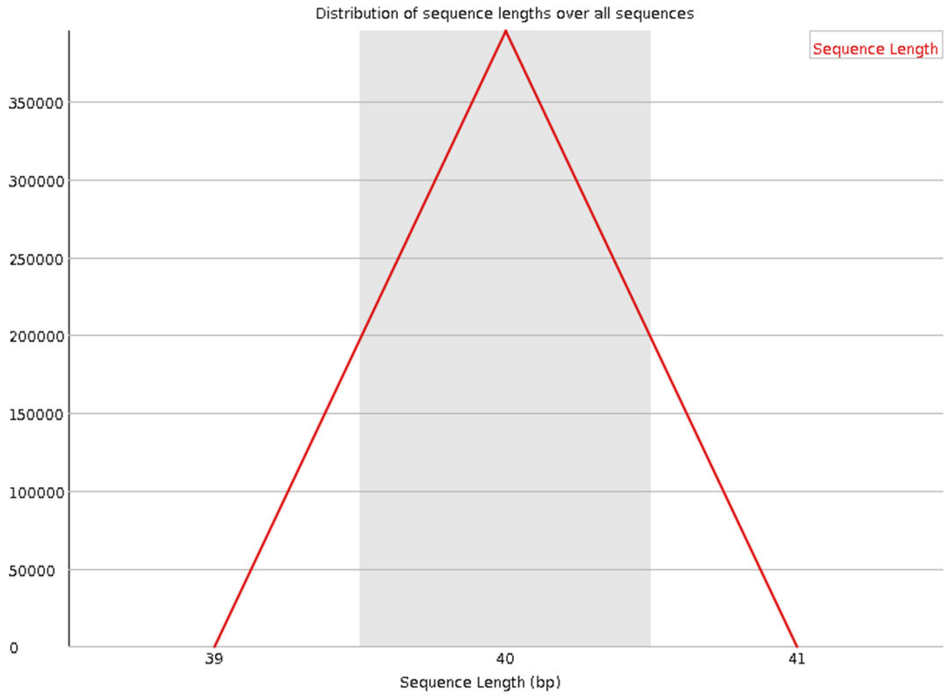
Dobrá data

✓ **Sequence Length Distribution**



Špatná data

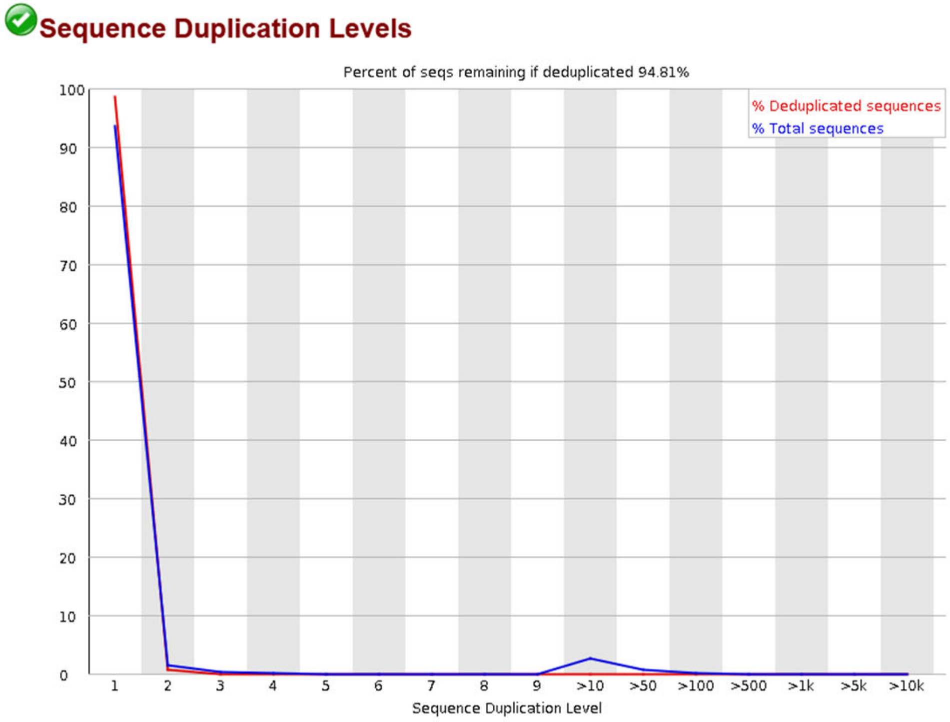
✓ **Sequence Length Distribution**



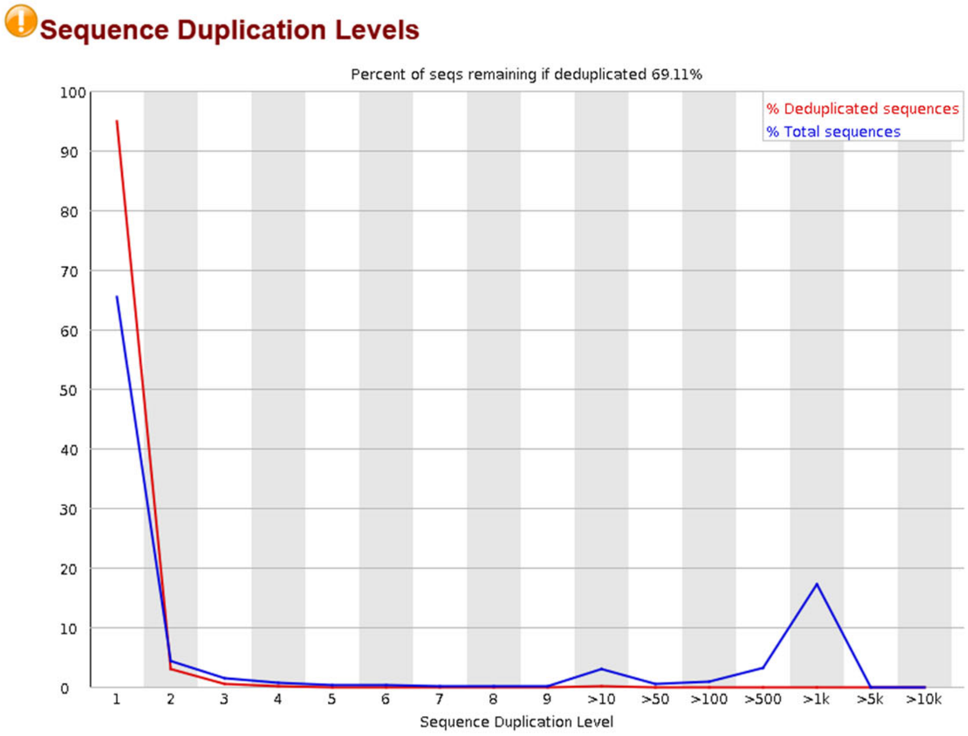
Obrázky vytvořil Vratislav Peška

Sequence Duplication Levels

Dobrá data



Špatná data



Obrázky vytvořil Vratislav Peška

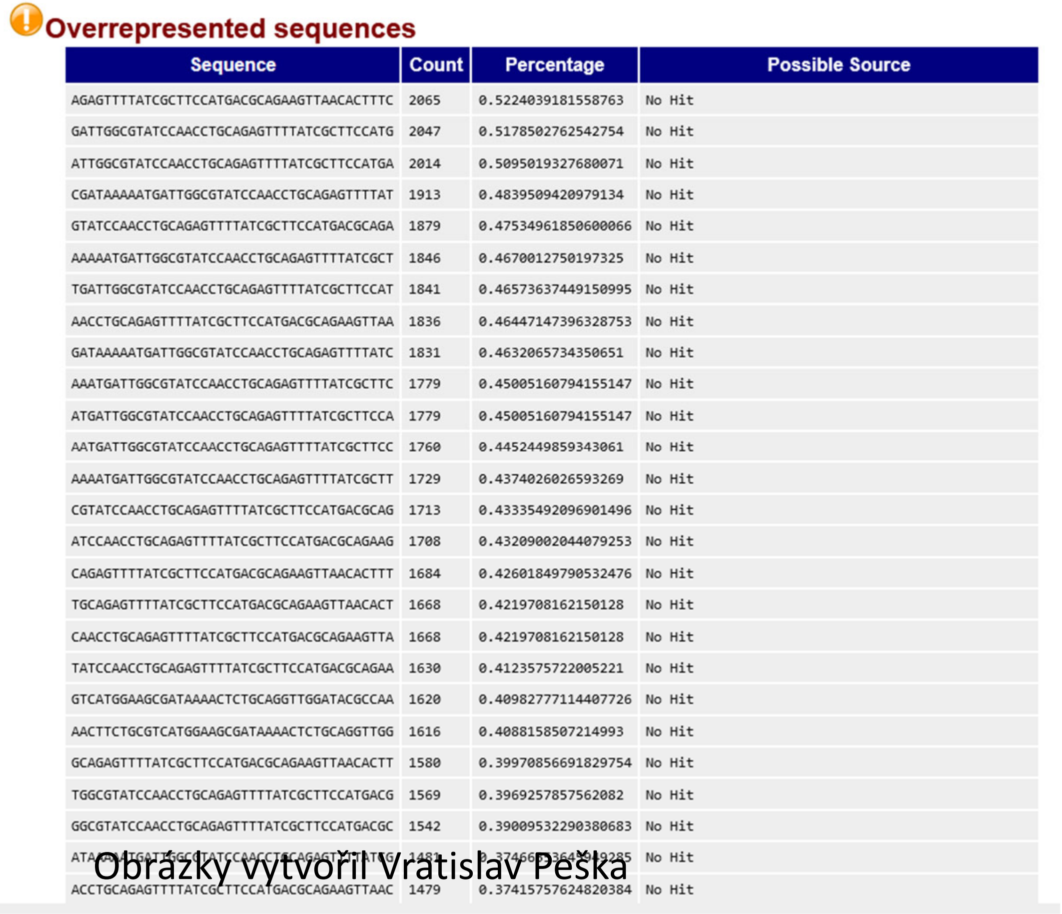
Overrepresented sequences

Dobrá data

Overrepresented sequences

No overrepresented sequences

Špatná data



Overrepresented sequences

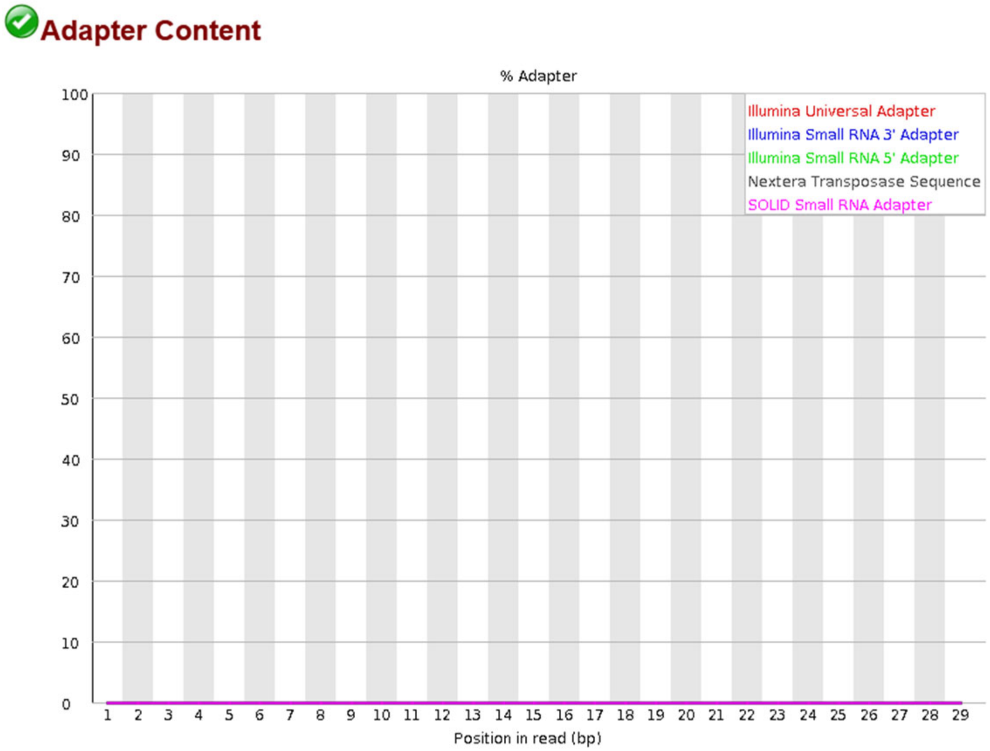
Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	No Hit
GTCATGGAAGCGATAAAACTCTGCAGGTTGGATACGCCAA	1620	0.40982777114407726	No Hit
AACTTCTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGG	1616	0.4088158507214993	No Hit
GCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTT	1580	0.39970856691829754	No Hit
TGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACG	1569	0.3969257857562082	No Hit
GGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGC	1542	0.39009532290380683	No Hit
ATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCG	1481	0.37466533645949285	No Hit
ACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAAC	1479	0.37415757624820384	No Hit

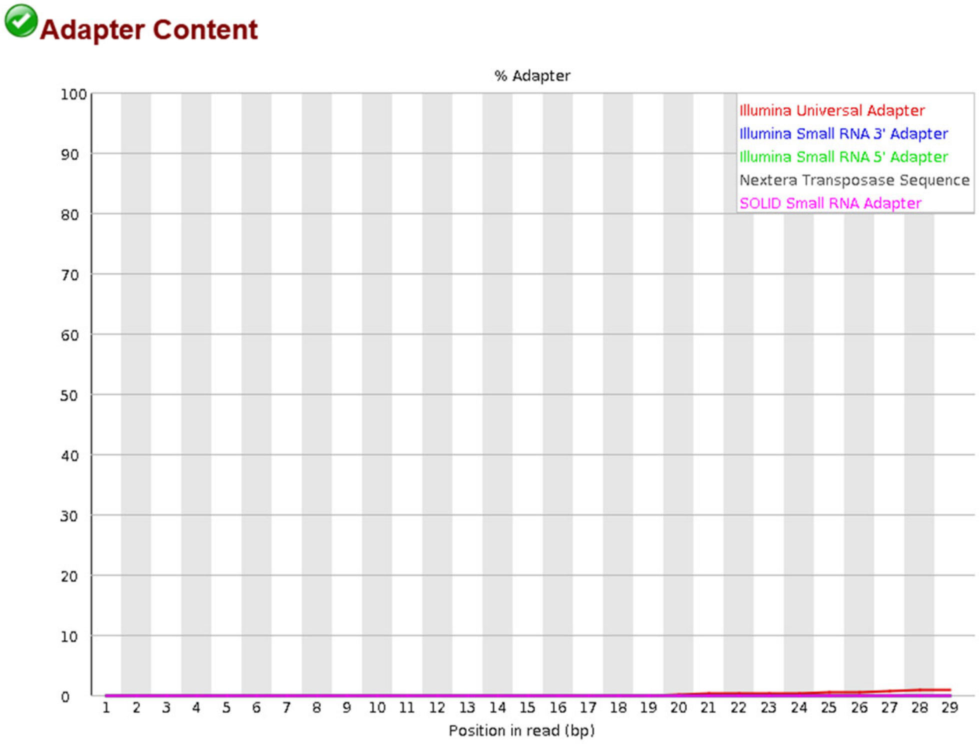
Obrázky vytvořil Vratislav Peška

Adapter Content

Dobrá data

Špatná data





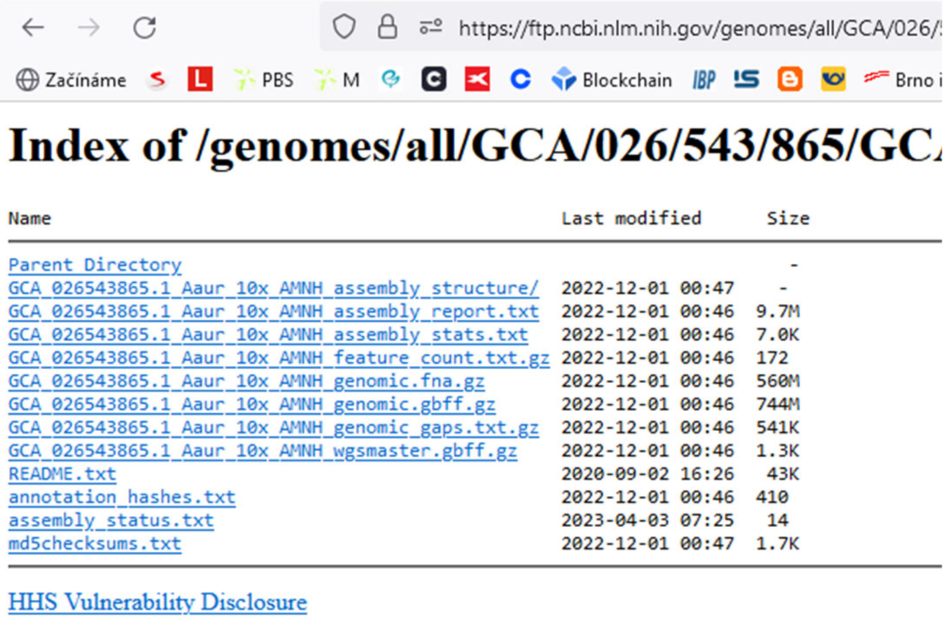
Obrázky vytvořil Vratislav Peška

GALAXY

FASTA [fa:stei]

- Textový soubor, přípony *.fa, *.fas, *.fasta, *.fna
- Řádky – lines – command line
- Zalamování (UNIX/DOS)
- NCBI příklad
- multifasta

```
>hlavička-1
GATCGATCG
>hlavička-2
ATCGATCGATCG
```



```
Index of /genomes/all/GCA/026/543/865/GC

Name                                              Last modified      Size
Parent Directory                                                       -
GCA_026543865.1_Aaur_10x_AMNH_assembly_structure/ 2022-12-01 00:47    -
GCA_026543865.1_Aaur_10x_AMNH_assembly_report.txt 2022-12-01 00:46  9.7M
GCA_026543865.1_Aaur_10x_AMNH_assembly_stats.txt  2022-12-01 00:46  7.0K
GCA_026543865.1_Aaur_10x_AMNH_feature_count.txt.gz 2022-12-01 00:46  172
GCA_026543865.1_Aaur_10x_AMNH_genomic.fna.gz      2022-12-01 00:46  560M
GCA_026543865.1_Aaur_10x_AMNH_genomic.gbff.gz     2022-12-01 00:46  744M
GCA_026543865.1_Aaur_10x_AMNH_genomic_gaps.txt.gz 2022-12-01 00:46  541K
GCA_026543865.1_Aaur_10x_AMNH_wgsmaster.gbff.gz   2022-12-01 00:46  1.3K
README.txt                                        2020-09-02 16:26   43K
annotation_hashes.txt                             2022-12-01 00:46  410
assembly_status.txt                               2023-04-03 07:25   14
md5checksums.txt                                  2022-12-01 00:47  1.7K

HHS Vulnerability Disclosure
```

Obrázky vytvořil Vratislav Peška

SAM (BAM) - sekvence, mapování, kvalita…

- Textový – TAB-delimited/TAB-separated. *.sam *.bam
- SAM = sequence alignment map (BAM – binární SAM tj. zapsaný formou jedniček a nul)

Hlavička je nepovinná a skládá se z řádků:

@HD – (řádek hlavička) VN:(verze formátu) SO:(alignment seřazen podle – unknown/unsorted/queryname/coordinate) GO:(group order – none/query/reference)

@SQ – (slovník referenční sekvence, může být více @SQ řádků) **SN:** jméno referenční sekvence (každá SQ musí mít unikátní SN tag) **LN:** (délka referenční sekvence 1 – $(1^{23}-1)$) **AS:** identifikátor genomového assembly M5:MD5 kontrolní součet sekvence **SP:** (species) **UR:** (URI – http:; ftp: file-systém path)

@RG – (read group information , může být více @RG řádků) **ID:**(unikátní pro každý @RG) **CN:**jméno sekvenačního centra **DS:**description **DT:**date run ISO8601 **FO:**flow order **KS:**key sequence **LB:**library **PG:**program of processing **PI:**median insert size **PL:**platform model **PU:**platform unit unikátní identifikátor **SM:**sample name

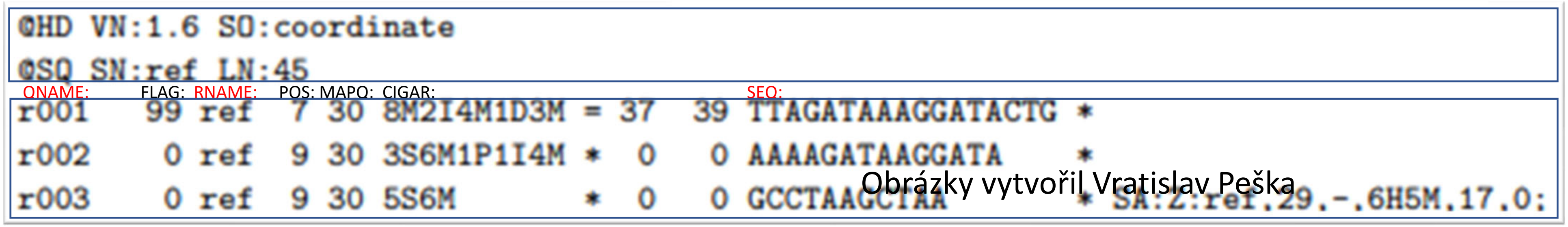
@PG – (info o programu pro vytvoření BAM/SAM) ID:unikátní identifikátor PN:jméno programu CL:command line PP:previous @PG-ID DS:popis programu VN:verze programu

@CO Line – (jednořádkový textový komentář)

Sekce alignmentu

11 povinných polí (sloupců) + další volitelné (pole SEQ * = sekvence není uložena)

QNAME: query name (=FASTQ) FLAG: RNAME:jméno referenční sekvence (jedna z SN z @SQ řádku) POS:první shoda v referenci vlevo MAPQ:kvalita mapování CIGAR:charakter shody RNEXT:jméno reference dalšího namapovaného readů PNEXT:pozice dalšího namapovaného readu TLEN:délka pozorované shody SEQ: sekvence readů (=FASTQ) QUAL:quality scores(=FASTQ)



```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001     99 ref  7 30 8M2I4M1D3M = 37   39 TTAGATAAAGGATACTG *
r002      0 ref  9 30 3S6M1P1I4M *  0    0 AAAAGATAAGGATA    *
r003      0 ref  9 30 5S6M       *  0    0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
```

Obrázky vytvořil Vratislav Peška

GFF/GTF - general feature format/general transfer format (GFF2=GTF, GFF3)

- Textový soubor – TAB-delimited, přípony *.gff *.gtf
- Popisuje geny a jiné >>features<< v DNA, RNA, proteinových sekvencích
- Každý feature má svůj řádek (8 povinných polí/sloupců, které nesmí být prázdné – >> . <<)

seqname source feature start end score strand frame {attributes comments}

seqname – jméno sekvence, může být shoda s EMBL/Genbank/DDBJ

source – program, který identifikoval feature

feature – gen, exon, repeat, transcript…

start – pozice 1. nt

end – pozice posledního nt

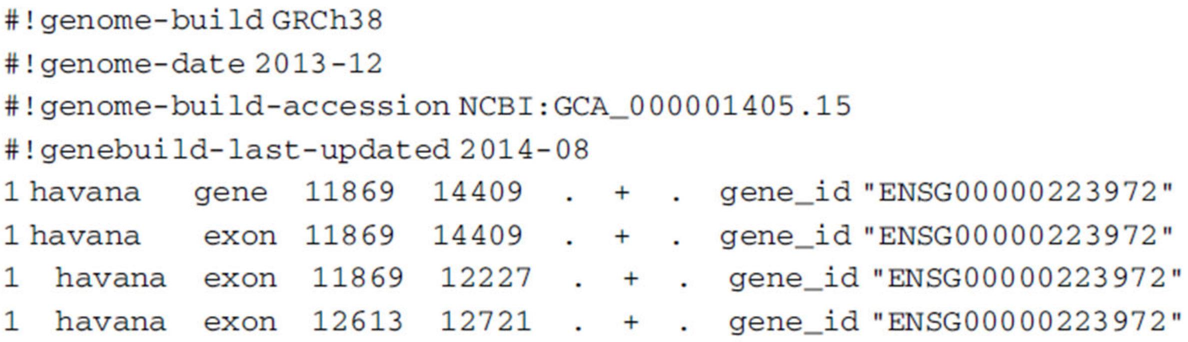
score – floating point value

strand – + (forward) – (reverse)

frame – 0/1/2/.

attributes

comments



```
#!genome-build GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.15
#!genebuild-last-updated 2014-08
1 havana    gene   11869   14409   .   +   .   gene_id "ENSG00000223972"
1 havana     exon  11869   14409   .   +   .   gene_id "ENSG00000223972"
1   havana   exon   11869   12227   .   +   .   gene_id "ENSG00000223972"
1   havana   exon   12613   12721   .   +   .   gene_id "ENSG00000223972"
```

Obrázky vytvořil Vratislav Peška

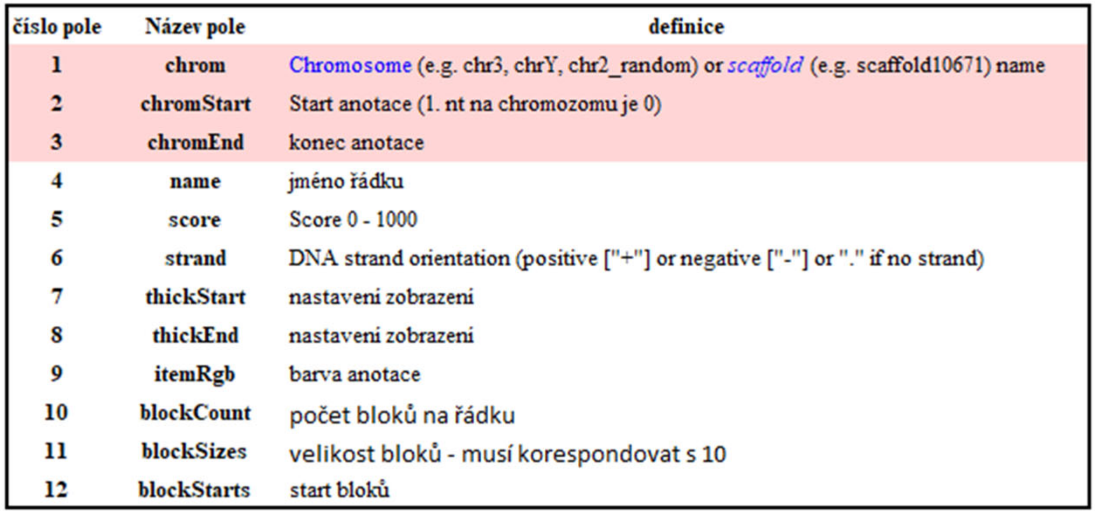
BED - Browser Extensible Data

- Textový soubor – TAB-delimited, *.bed
- Primární účel - anotace genomových dat
- Mohou mít hlavičku s řádky:

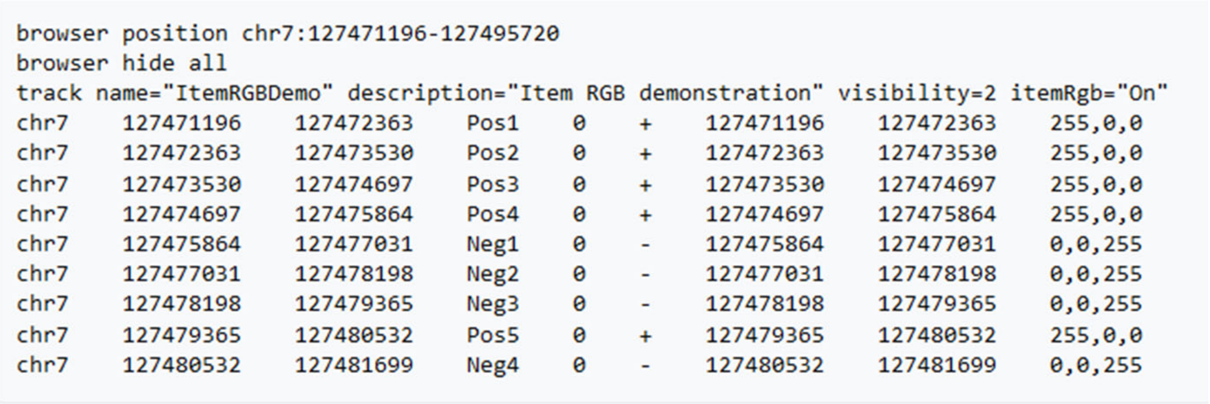
browser

track

#



číslo pole	Název pole	definice
1	chrom	Chromosome (e.g. chr3, chrY, chr2_random) or *scaffold* (e.g. scaffold10671) name
2	chromStart	Start anotace (1. nt na chromozomu je 0)
3	chromEnd	konec anotace
4	name	jméno řádku
5	score	Score 0 - 1000
6	strand	DNA strand orientation (positive ["+"] or negative ["-"] or "." if no strand)
7	thickStart	nastaveni zobrazeni
8	thickEnd	nastaveni zobrazeni
9	itemRgb	barva anotace
10	blockCount	počet bloků na řádku
11	blockSizes	velikost bloků - musí korespondovat s 10
12	blockStarts	start bloků



```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    -    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    -    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    -    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    -    127480532    127481699    0,0,255
```

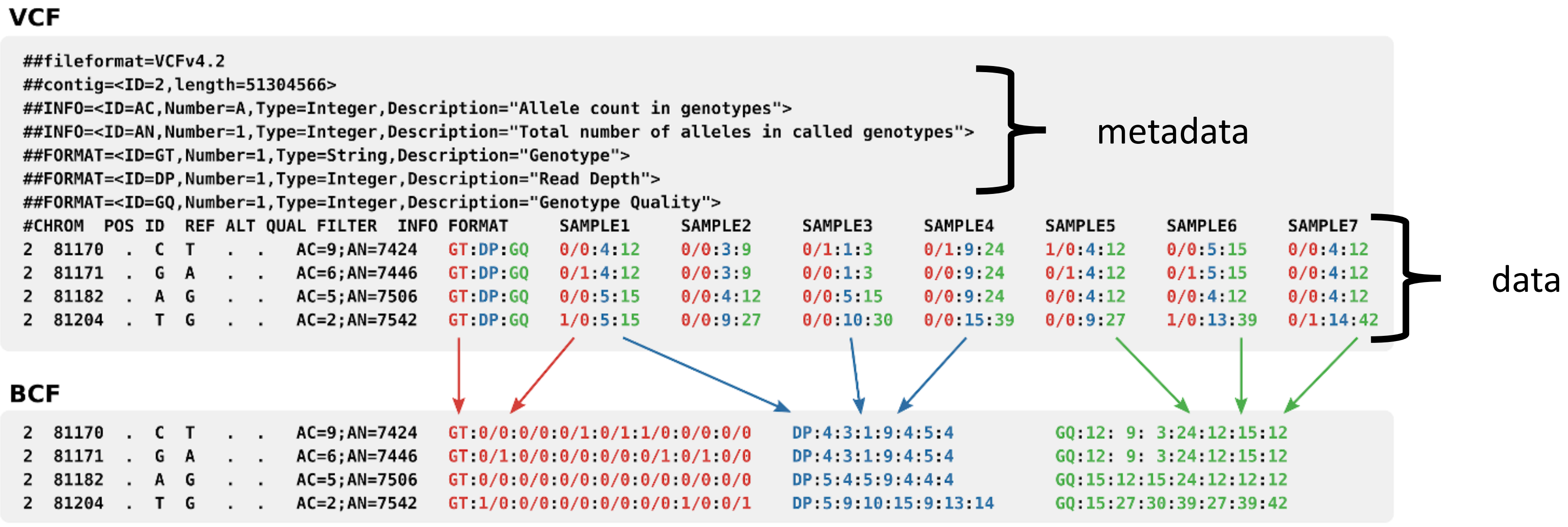
Obrázky vytvořil Vratislav Peška

VCF (BCF) - Variant Call Format

Kontrolní otázka:

Jaký je rozdíl mezi SNP a SNV?

- Textový soubor, TAB-delimited, rozšíření GFF3
- Primárně pro anotaci včetně varianility (nejčastější SNPs, malé INDELs)
- gVCF



VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:DP:GQ   0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:DP:GQ   0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:DP:GQ   0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:DP:GQ   1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

BCF

```
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0   DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0   DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0   DP:5:4:5:9:4:4:4        GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:1/0:0/0:0/0:0/0:0/0:1/0:0/1   DP:5:9:10:15:9:13:14    GQ:15:27:30:39:27:39:42
```

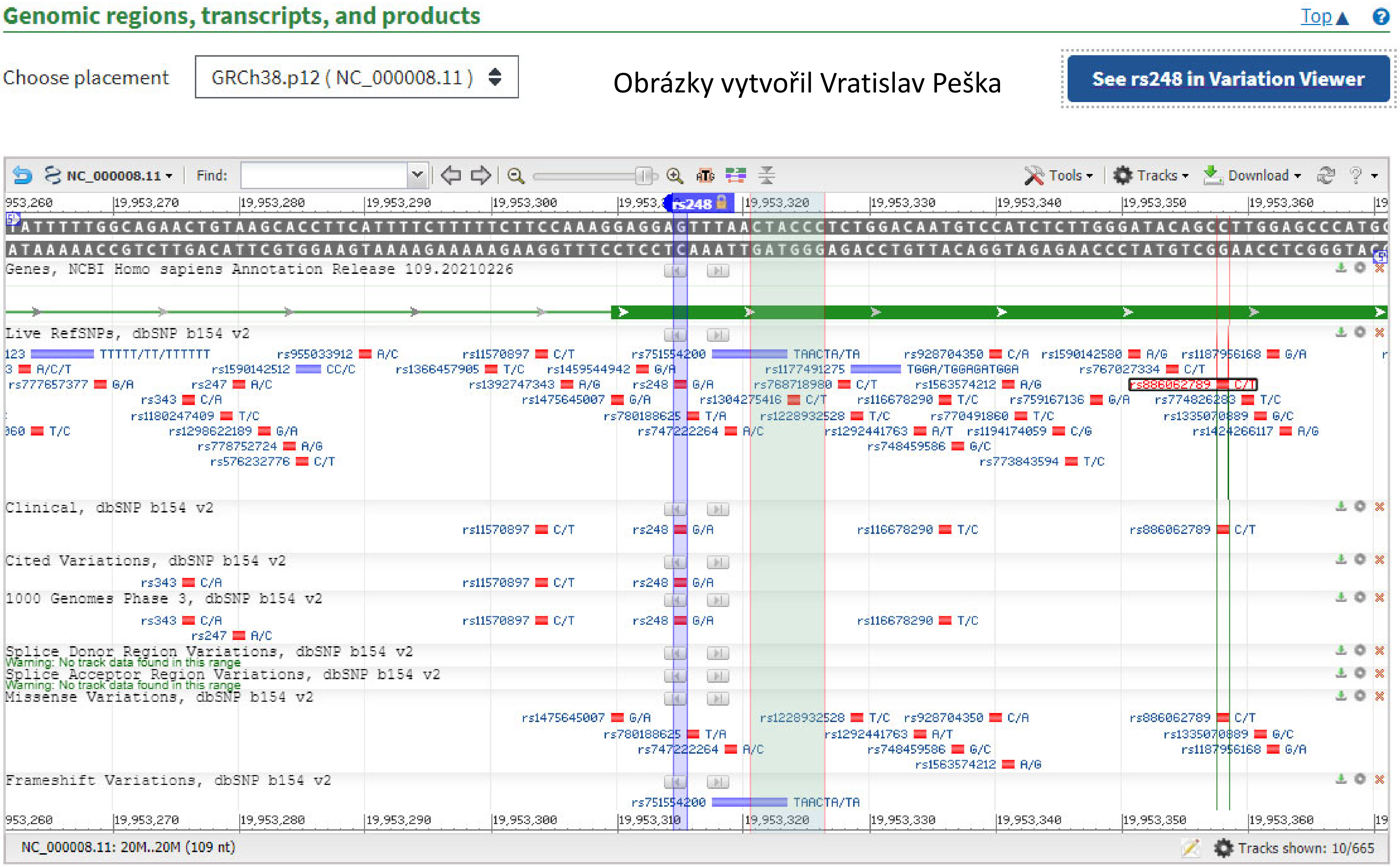
Obrázky vytvořil Vratislav Peška

Jaký je rozdíl mezi SNP a SNV?

- single nucleotide variant (SNV)
- single nucleotide polymorphism (SNP)

- single-base substituce (real-time PCR, microarrays, NGS)
- SNV variatna na úrovni 1 nukleotidu v genomu populace
- SNP variatna na úrovni 1 nukleotidu v genomu zárodečné linie alespoň u 1% populace.

Obrázky vytvořil Vratislav Peška

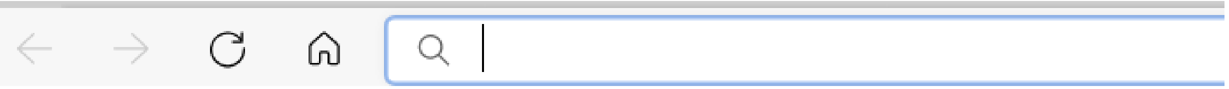


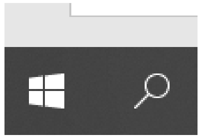
Genomic regions, transcripts, and products

Choose placement: GRCh38.p12 (NC_000008.11)

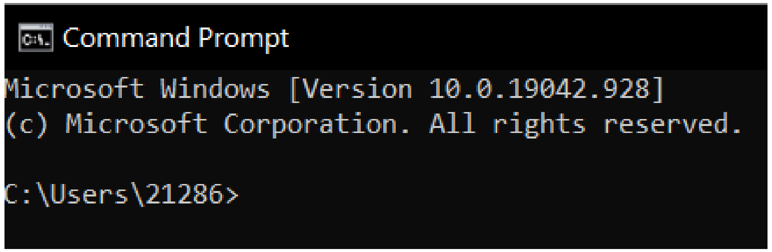
Obrázky vytvořil Vratislav Peška

Command line introduction





Type cmd and press enter



```
Microsoft Windows [Version 10.0.19042.928]
(c) Microsoft Corporation. All rights reserved.

C:\Users\21286>
```

dir (+ enter)
cd
cd ..
G:
del
mkdir
rmdir
ipconfig
netstat
ping
systeminfo
cls
color 0A
title

šipky nahoru dolu
tabulator

Is there a Windows command line with smart bash-like autocompletions / command history? - Super User

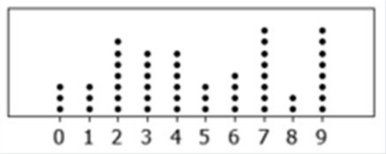
key F7

označ text + right-click

Obrázky vytvořil Vratislav Peška

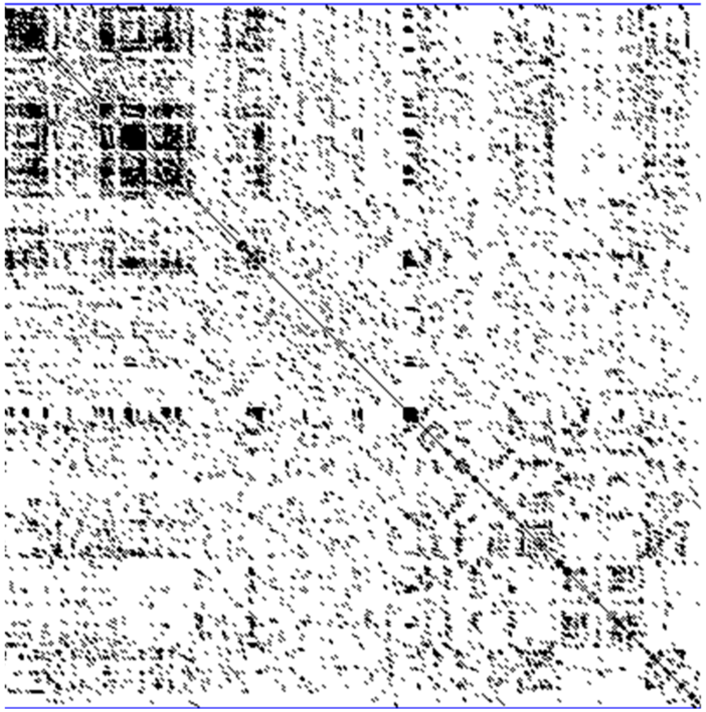
Dot plot (dotter)

STATISTIKA



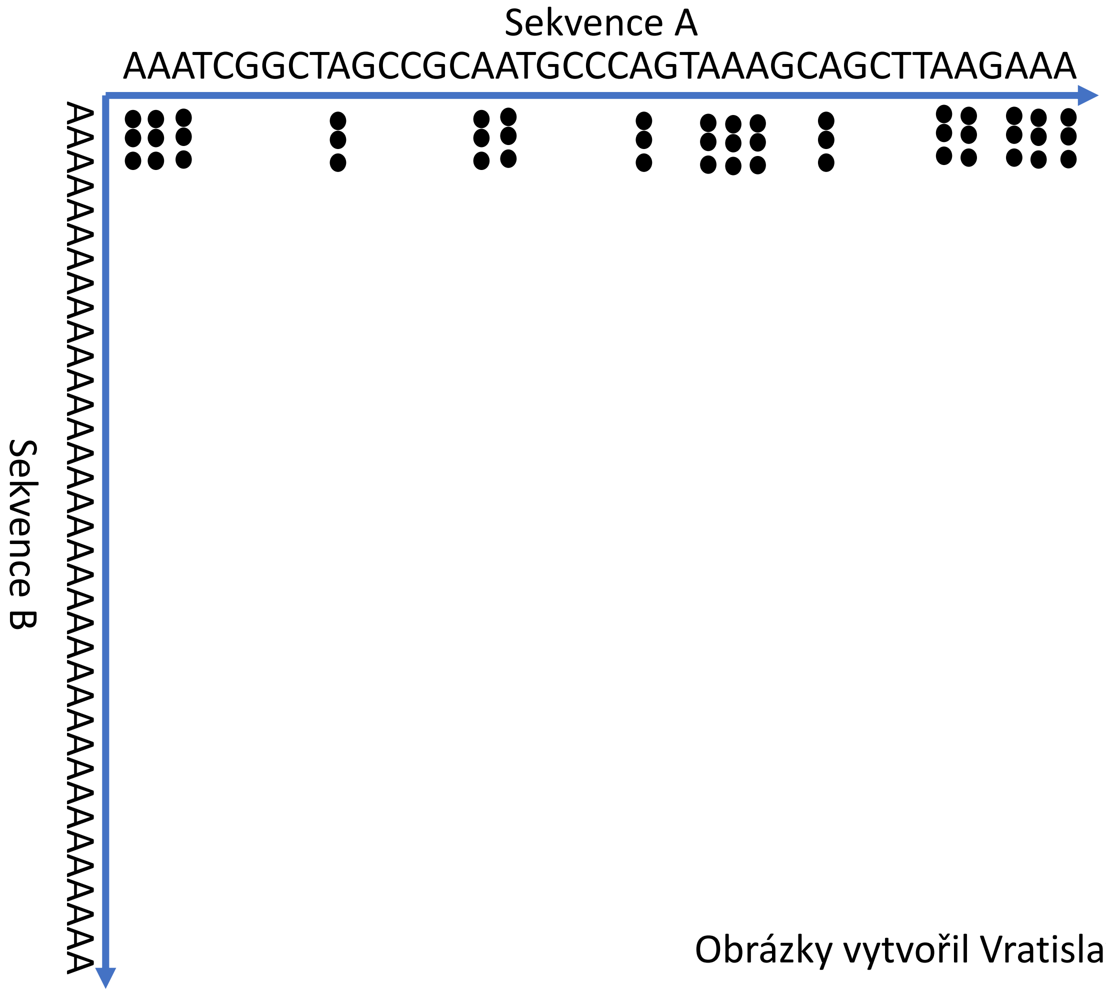
A dot plot of 50 random values from 0 to 9.

BIOINFORMATIKA



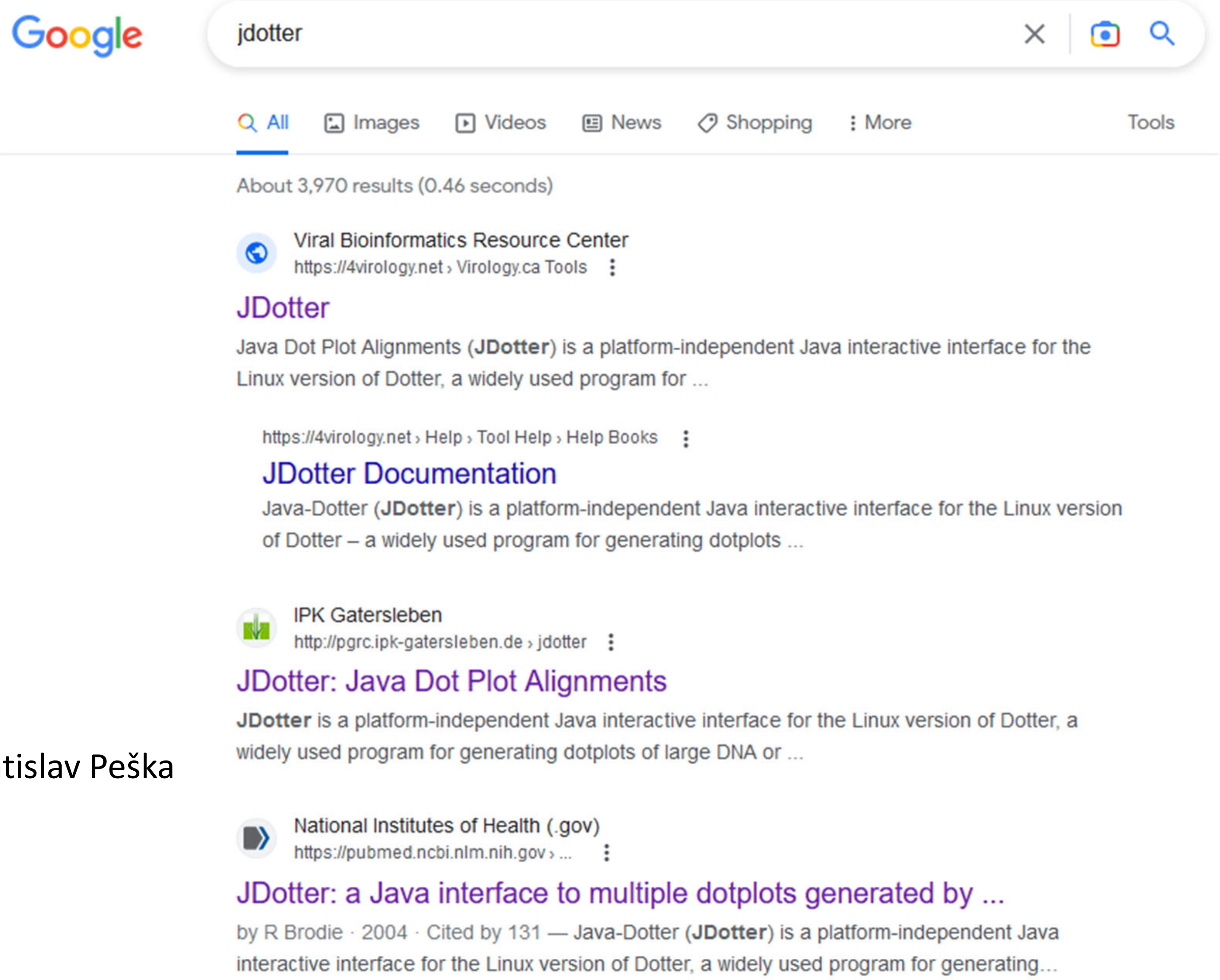
Obrázky vytvořil Vratislav Peška

Dotplot



Obrázky vytvořil Vratislav Peška

JDotter



jdotter

About 3,970 results (0.46 seconds)

Viral Bioinformatics Resource Center
https://4virology.net › Virology.ca Tools

JDotter

Java Dot Plot Alignments (**JDotter**) is a platform-independent Java interactive interface for the Linux version of Dotter, a widely used program for ...

https://4virology.net › Help › Tool Help › Help Books

JDotter Documentation

Java-Dotter (**JDotter**) is a platform-independent Java interactive interface for the Linux version of Dotter – a widely used program for generating dotplots ...

IPK Gatersleben
http://pgrc.ipk-gatersleben.de › jdotter

JDotter: Java Dot Plot Alignments

JDotter is a platform-independent Java interactive interface for the Linux version of Dotter, a widely used program for generating dotplots of large DNA or ...

National Institutes of Health (.gov)
https://pubmed.ncbi.nlm.nih.gov › ...

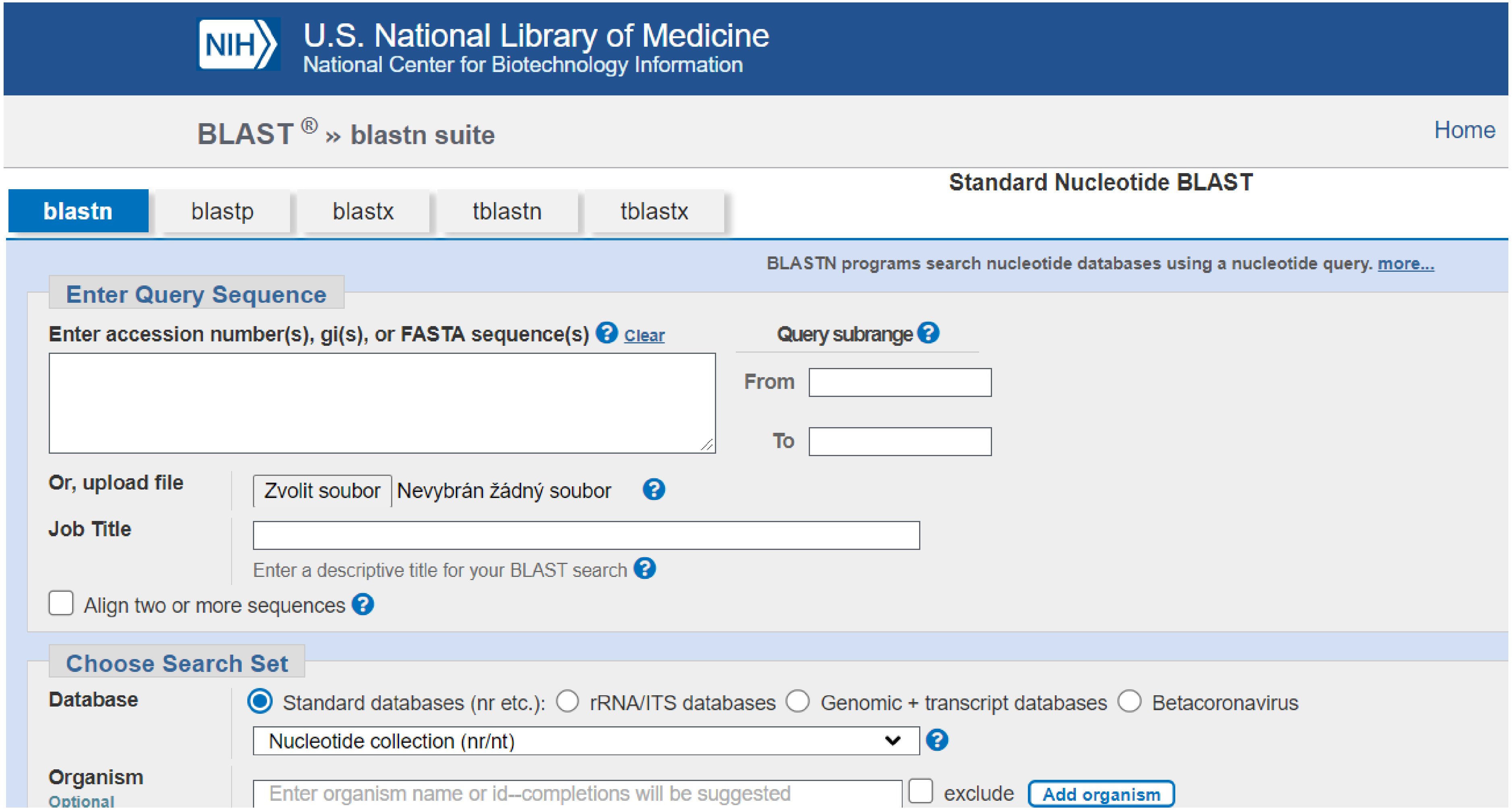
JDotter: a Java interface to multiple dotplots generated by ...

by R Brodie · 2004 · Cited by 131 — Java-Dotter (**JDotter**) is a platform-independent Java interactive interface for the Linux version of Dotter, a widely used program for generating...

Obrázky vytvořil Vratislav Peška

BLAST

Obrázky vytvořil Vratislav Peška



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file: Nevybrán žádný soubor

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

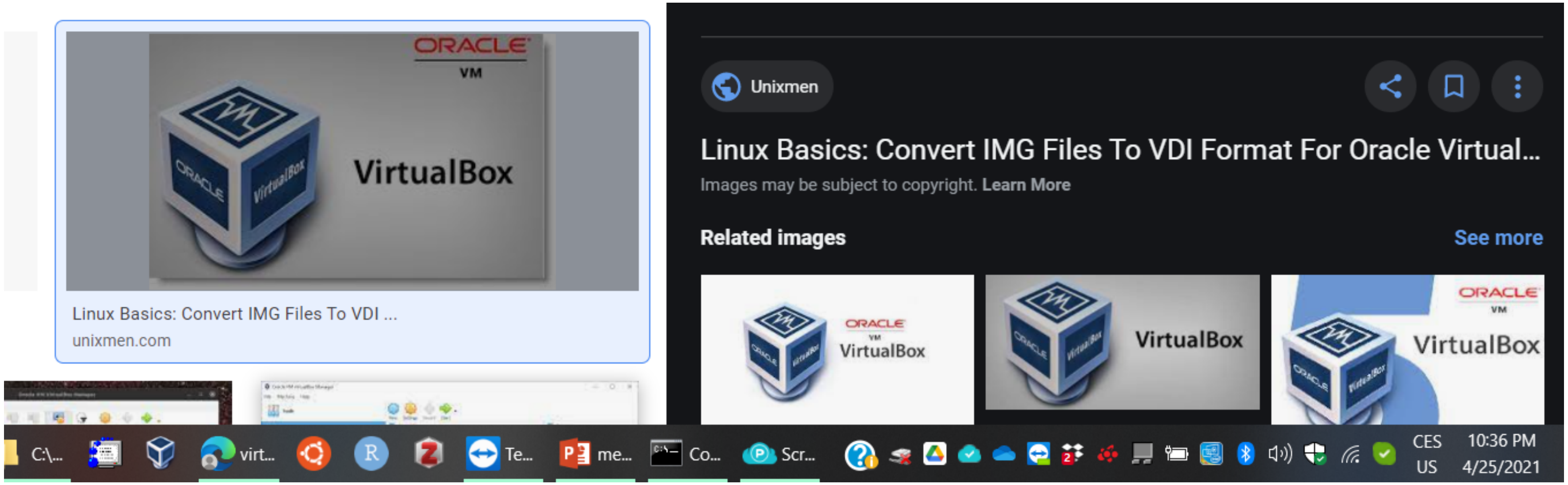
Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested ☐ exclude

Virtualizace – cesta k linuxu (windows)


Linux Basics: Convert IMG Files To VDI ...
unixmen.com

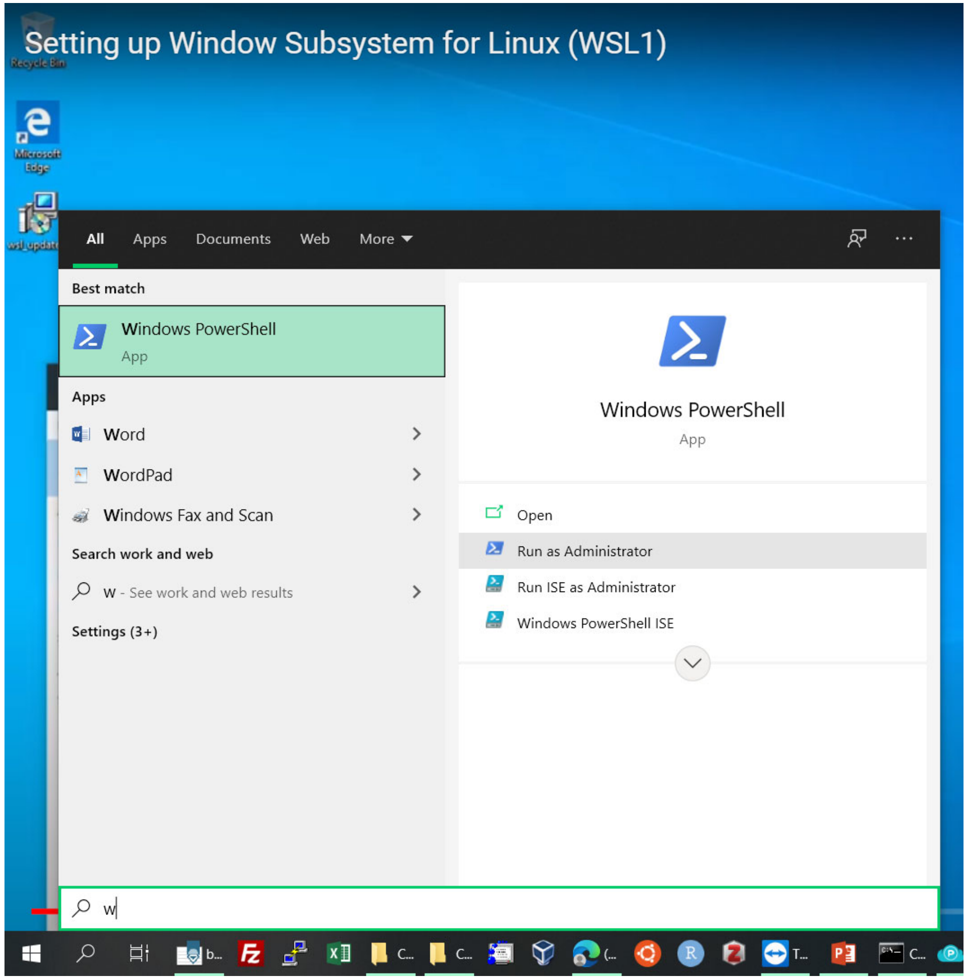
Linux Basics: Convert IMG Files To VDI Format For Oracle Virtual...

Images may be subject to copyright. **Learn More**

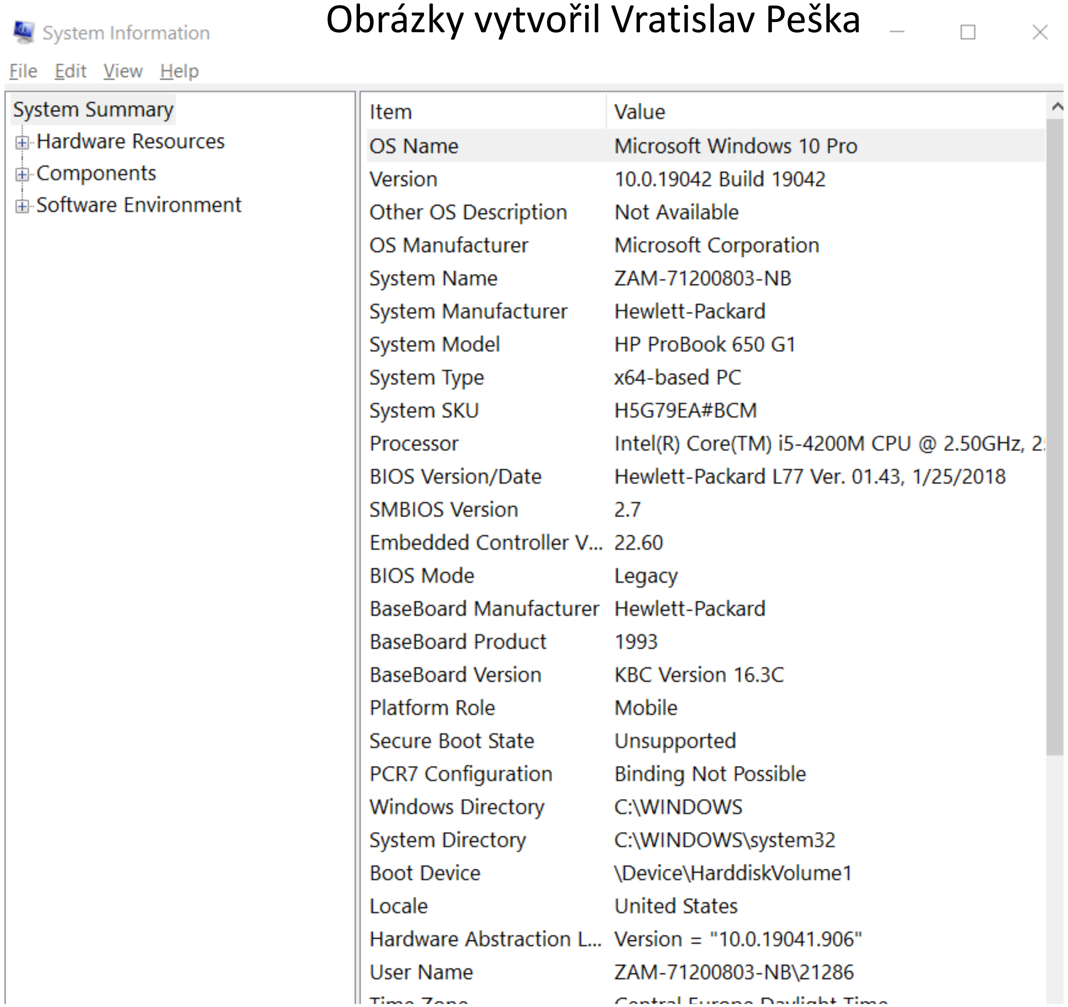
Obrázky vytvořil Vratislav Peška

WSL – window for subsystem linux

Obrázky vytvořil Vratislav Peška

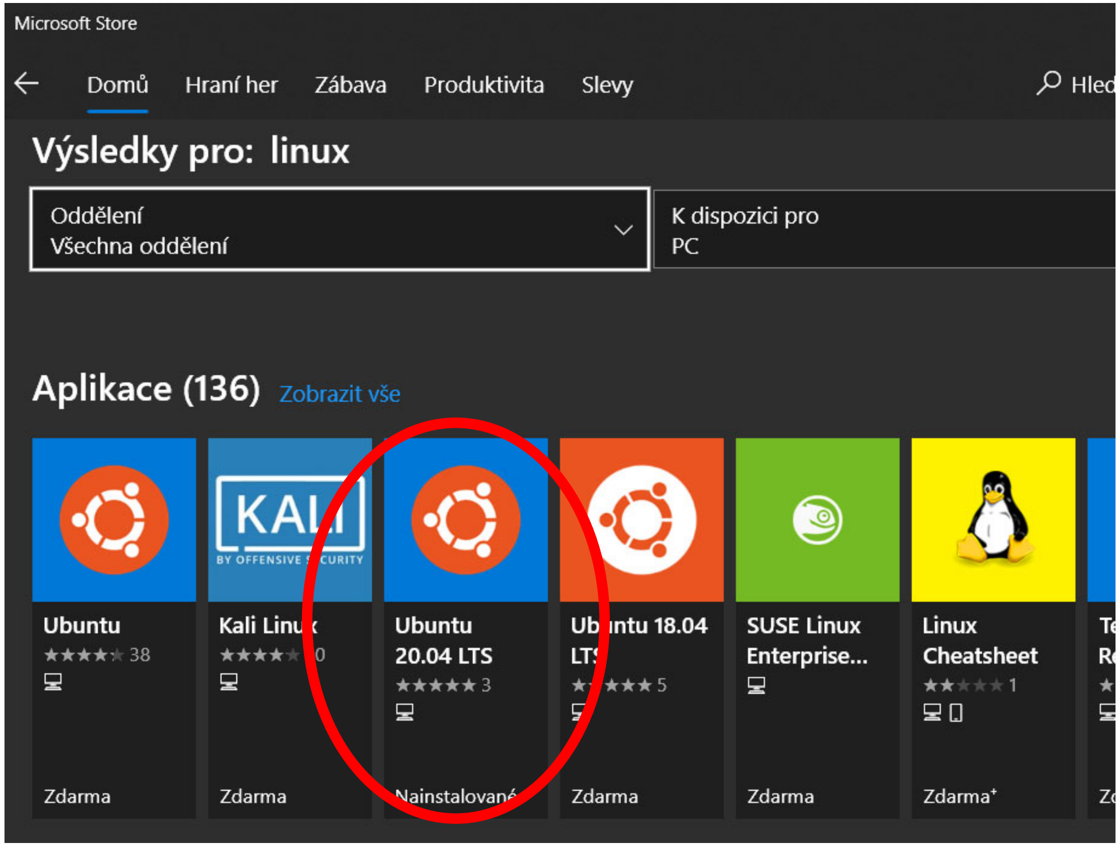


Setting up Window Subsystem for Linux (WSL1)



Item	Value
OS Name	Microsoft Windows 10 Pro
Version	10.0.19042 Build 19042
Other OS Description	Not Available
OS Manufacturer	Microsoft Corporation
System Name	ZAM-71200803-NB
System Manufacturer	Hewlett-Packard
System Model	HP ProBook 650 G1
System Type	x64-based PC
System SKU	H5G79EA#BCM
Processor	Intel(R) Core(TM) i5-4200M CPU @ 2.50GHz, 2
BIOS Version/Date	Hewlett-Packard L77 Ver. 01.43, 1/25/2018
SMBIOS Version	2.7
Embedded Controller V...	22.60
BIOS Mode	Legacy
BaseBoard Manufacturer	Hewlett-Packard
BaseBoard Product	1993
BaseBoard Version	KBC Version 16.3C
Platform Role	Mobile
Secure Boot State	Unsupported
PCR7 Configuration	Binding Not Possible
Windows Directory	C:\WINDOWS
System Directory	C:\WINDOWS\system32
Boot Device	\Device\HarddiskVolume1
Locale	United States
Hardware Abstraction L...	Version = "10.0.19041.906"
User Name	ZAM-71200803-NB\21286

nainstalovat aplikaci (virtuální systém)



Výsledky pro: linux

Oddělení
Všechna oddělení

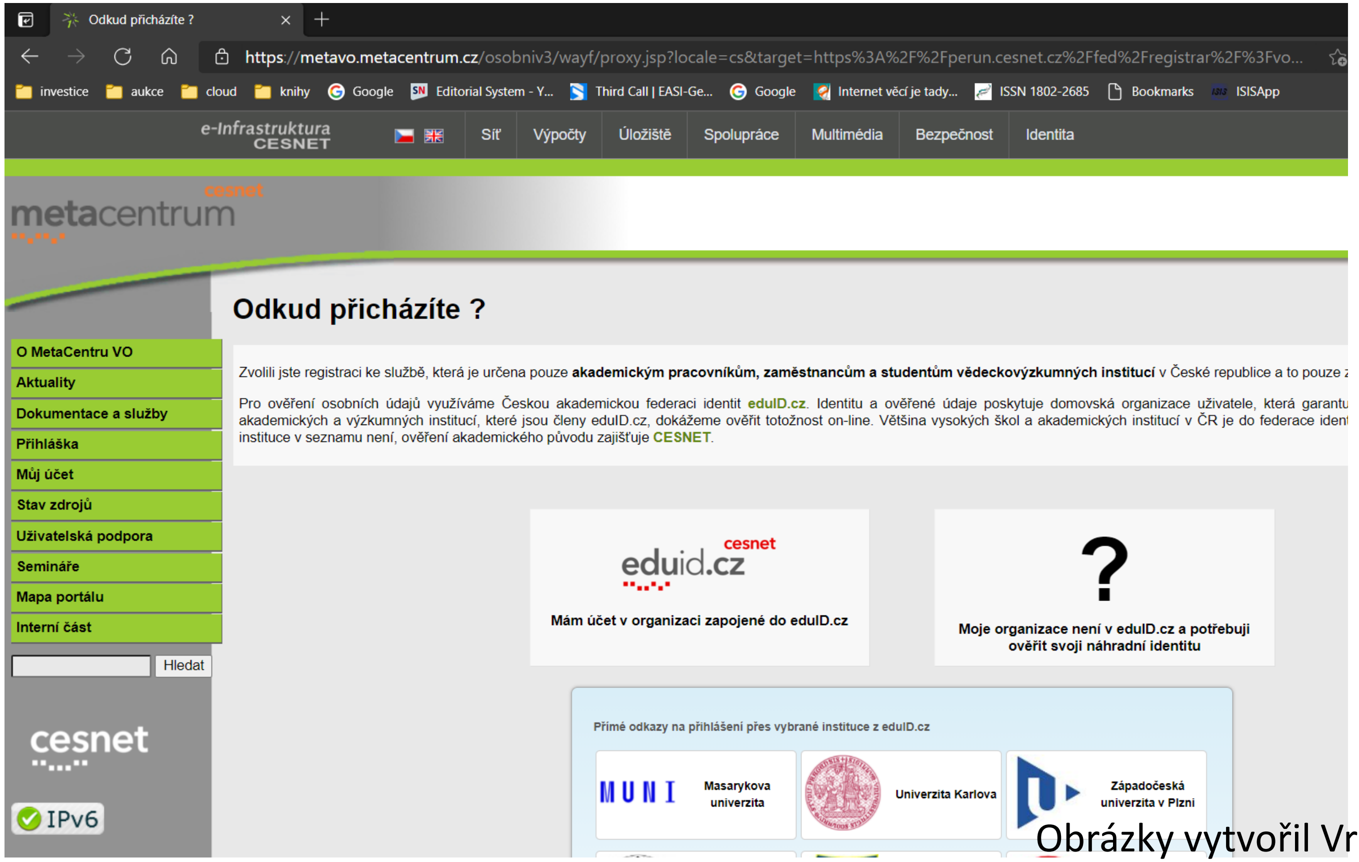
K dispozici pro
PC

Aplikace (136) Zobrazit vše

Ubuntu 20.04 LTS
download VS + instalace VS + restart celého počítače

Obrázky vytvořil Vratislav Peška

MetaCentrum



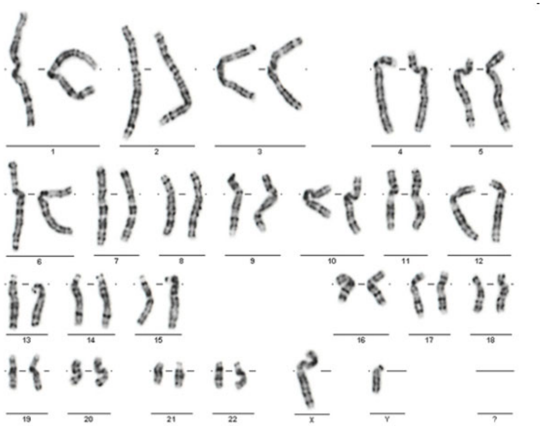
Odkud přicházíte ?

Zvolili jste registraci ke službě, která je určena pouze **akademickým pracovníkům, zaměstnancům a studentům vědeckovýzkumných institucí** v České republice a to pouze z

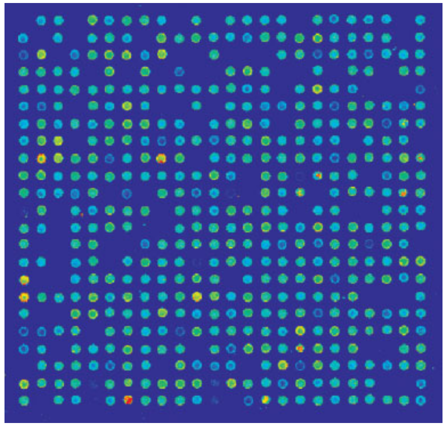
Pro ověření osobních údajů využíváme Českou akademickou federaci identit **eduID.cz**. Identitu a ověřené údaje poskytuje domovská organizace uživatele, která garantu akademických a výzkumných institucí, které jsou členy eduID.cz, dokážeme ověřit totožnost on-line. Většina vysokých škol a akademických institucí v ČR je do federace ident instituce v seznamu není, ověření akademického původu zajišťuje **CESNET**.

Obrázky vytvořil Vratislav Peška

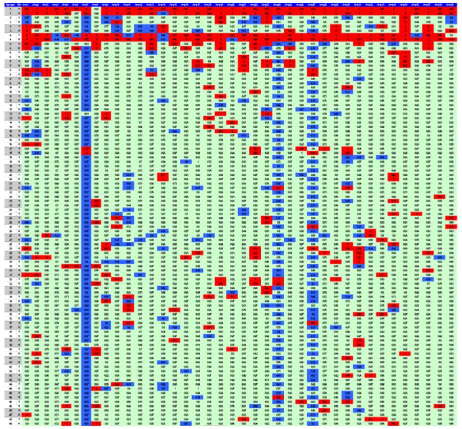
Prenatální diagnostika



Klasický karyotyp (rozlišení ~ 5 Mb)



Array-CGH (rozlišení ~ 0,1 Mb)

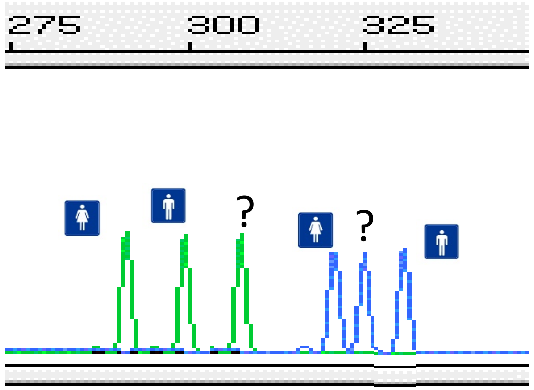


NGS (rozlišení- jednotlivé báze)

Obrázky vytvořil Vratislav Peška

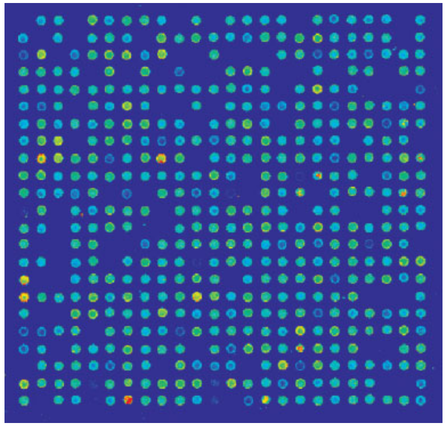
Doba odezvy (TAT)

v den odběru



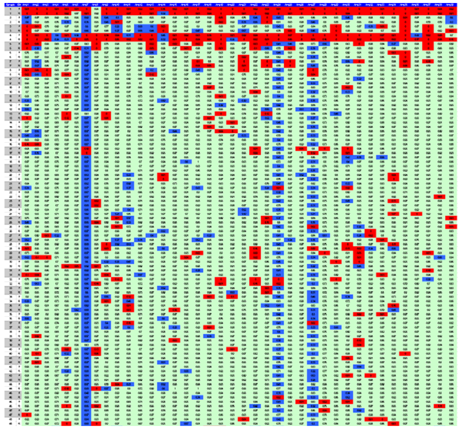
QF-PCR

1-2 týdny



Array-CGH (rozlišení ~ 0,1 Mb)

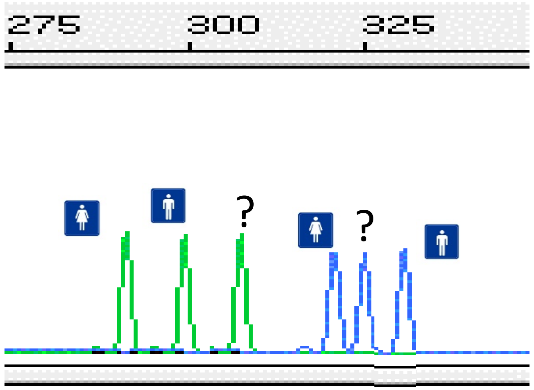
2-3 týdny



NGS (rozlišení- jednotlivé báze)

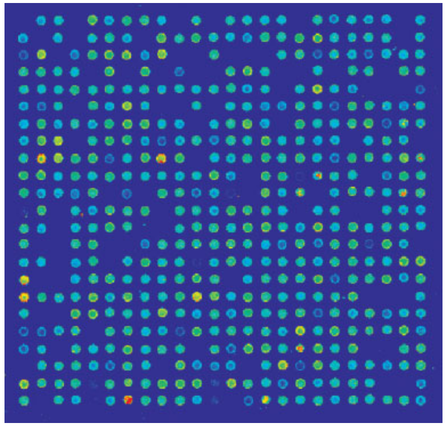
Cena $$$

$



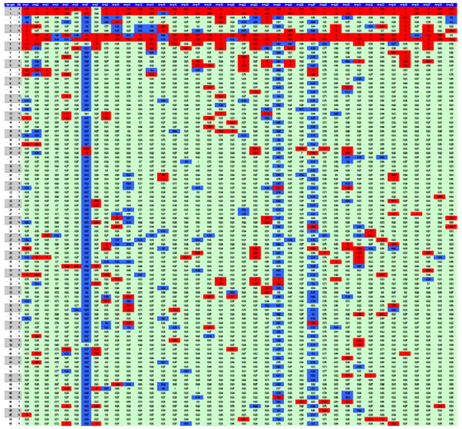
QF-PCR

$$



Array-CGH (rozlišení ~ 0,1 Mb)

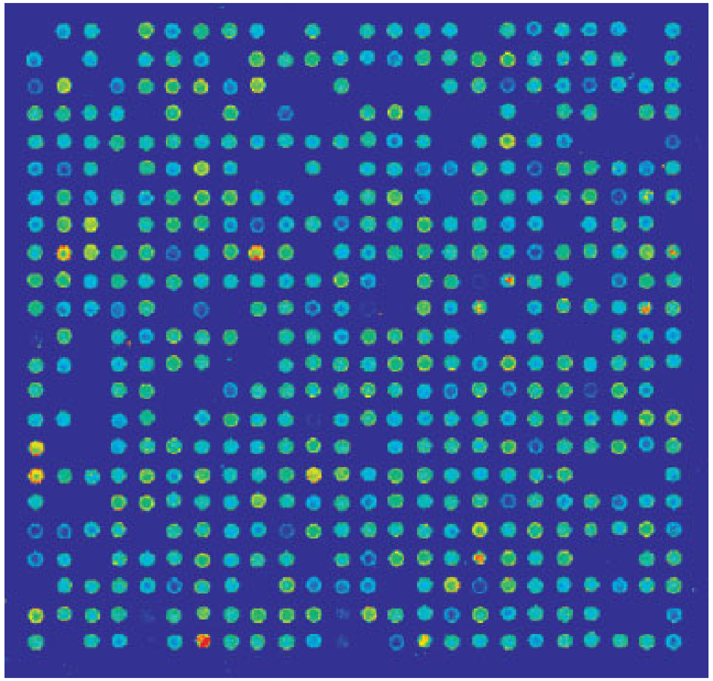
$$$



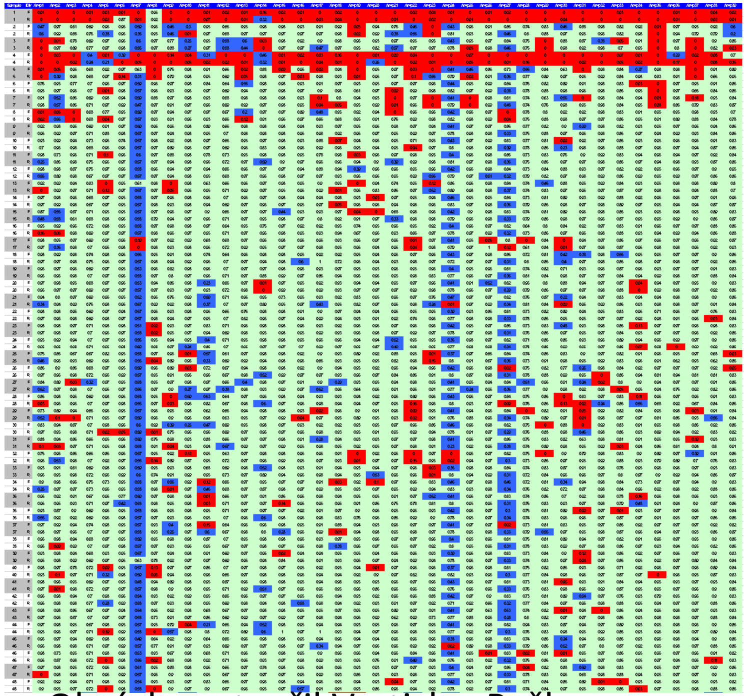
NGS (rozlišení- jednotlivé báze)

„Netušíme, kde a co hledat“

array CGH



NGS



Obrázky vytvořil Vratislav Peška

NGS v prenatální diagnostice

- **Panely genů pro specifické fenotypové skupiny**
  - např. vrozené srdeční vady, RASopatie, skeletální vady
- WES (popř. WGS) – u nás Heredity panel (klinický exom 3332 genů)
- trend WES + CNV (2v1)
- **NIPT** – neinvazivní prenatální testování

Obrázky vytvořil Vratislav Peška

RASopatie – sy Noonanové

Syndrom Noonanové

Noonan-like (CBL syndrom)

Noonan-like sy with loose anagen hair

Costello syndrom

Hereditární gingivální fibromatóza

Jaffe Campanacci syndrom

Kardio-facio-kutánní syndrom (CFC)

Legius syndrom (NF1-like)

LEOPARD syndrom (Multiple lentigines)

Neurofibromatóza typu 1

Neurofibromatóza typu 2

Neurofibromatosa-Noonan (NFNS)

Watson syndrom

RASopatie – sy Noonanové

Rasopathy genes	
PTPN11	*SHOC2*
SOS1	*A2ML1*
RAF1	*LZTR1*
RIT1	*RASA2*
BRAF	*SOS2*
KRAS	*MAP2K2*
NRAS	*HRAS*
MAP2K1	*SPRED1*
RRAS	*NF1*
CBL	*NF2*

Indikace pro prenatální diagnostiku of RASopatií:

- zvýšené šíjové projasnění (NT) nebo cystická hygroma (karyotyp, aCGH v normě) v kombinaci s:
  - » *hydrops fetalis*
  - » srdeční anomálie
  - » polyhydramnion a/nebo pleurální výpotek
  - » specifické faciální anomálie (hypertelorismus and micrognathia)
  - » ledvinové anomálie

RASopatie – sy Noonanové

Klinický popis

matka 32 let

UZ:

- 13/14/16/19 tg NT:2,3/3,1/1,8 mm
- 14/16 tg jugular lymfatické vaky
- 19 tg anomálie renálního duplexu a

  hypertrofie pravé komory

Noonan syndrome

IVF
13 gw - NT 6 mm
17 gw - total fetal hydrops with subcutaneous infiltration, ascites

Mutation in *RIT1*: c.319A>G, p.Met107Val *de novo*

RIT1 **(RIT1)** ~5 % cases (Aoki *et al.*, 2016), causal gene since 2013

Milosavljević *et al.*, 2016:

UZ: zvýšené NT, lymfatické vaky, duplikace renálního systému, polyhydramnion, hydrops s lymfedémem, hydrothorax, ascites.

Ve 26. t. g. intrauterinní úmrtí plodu.

Pitva plodu prokázala edém, kraniofaciální abnormality charakteristické pro sy Noonanové (hypertelorismus, nízko posazené uši, široký krk).

- 14 gw - NT 8 mm, lymphatic sacs, hygroma colli, renal pelvis bilat 2.5 mm, susp. CHD
- 16 gw - NT 5 mm, lymphatic sacs, agenesis ductus venosus,
- 19 gw NT 4 mm, dilatation of renal pelvis 8 mm, hypertelorism, low set earlobes, CHD
- ***RAF1*** c.770C>T ***de novo*** (several publications)

RASopatie – sy Noonanové

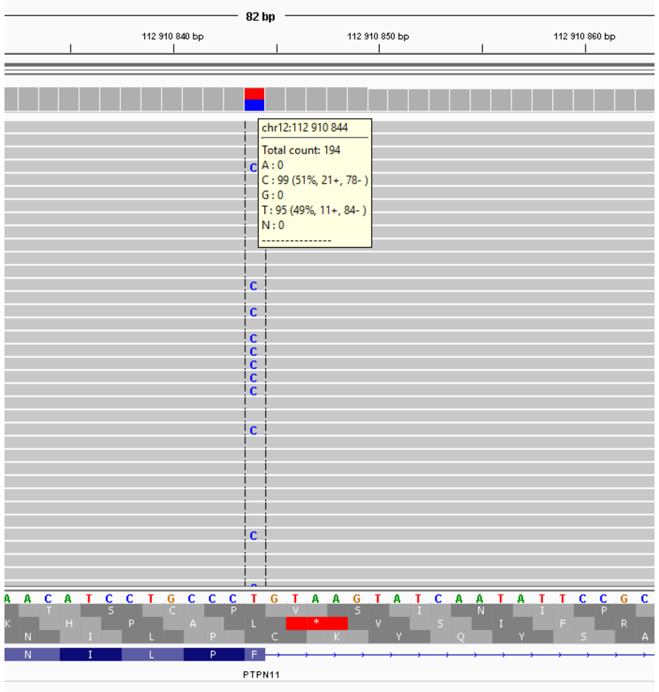
Molekulární diagnostika:

QF-PCR norma (13, 18, 21, X a Y)

Karyotyp norma

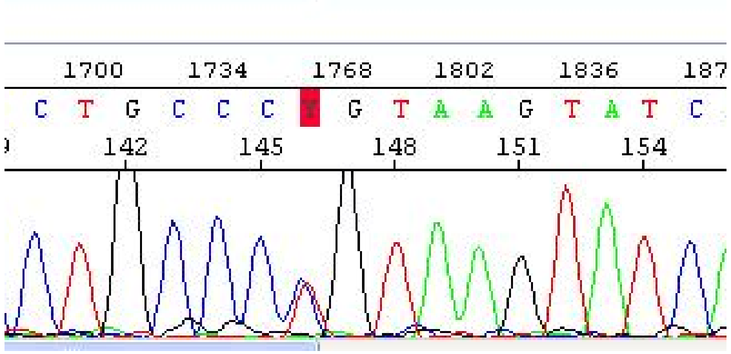
array-CGH – žádná patogenní varianta

RASopatie NGS panel



chr12:112 910 844

Total count: 194
A : 0
C : 99 (51%, 21+, 78-)
G : 0
T : 95 (49%, 11+, 84-)
N : 0



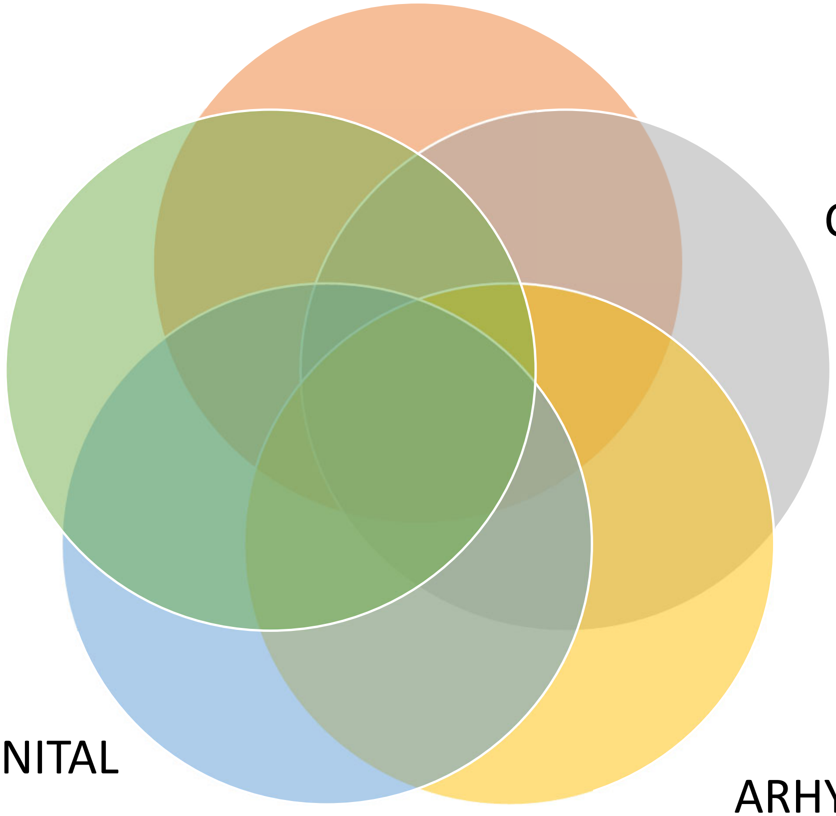
Kauzální mutace:

***PTPN11:* c.853T>C (p.Phe285Leu)** – *de novo*

Obrázky vytvořil Vratislav Peška

Cardio panel

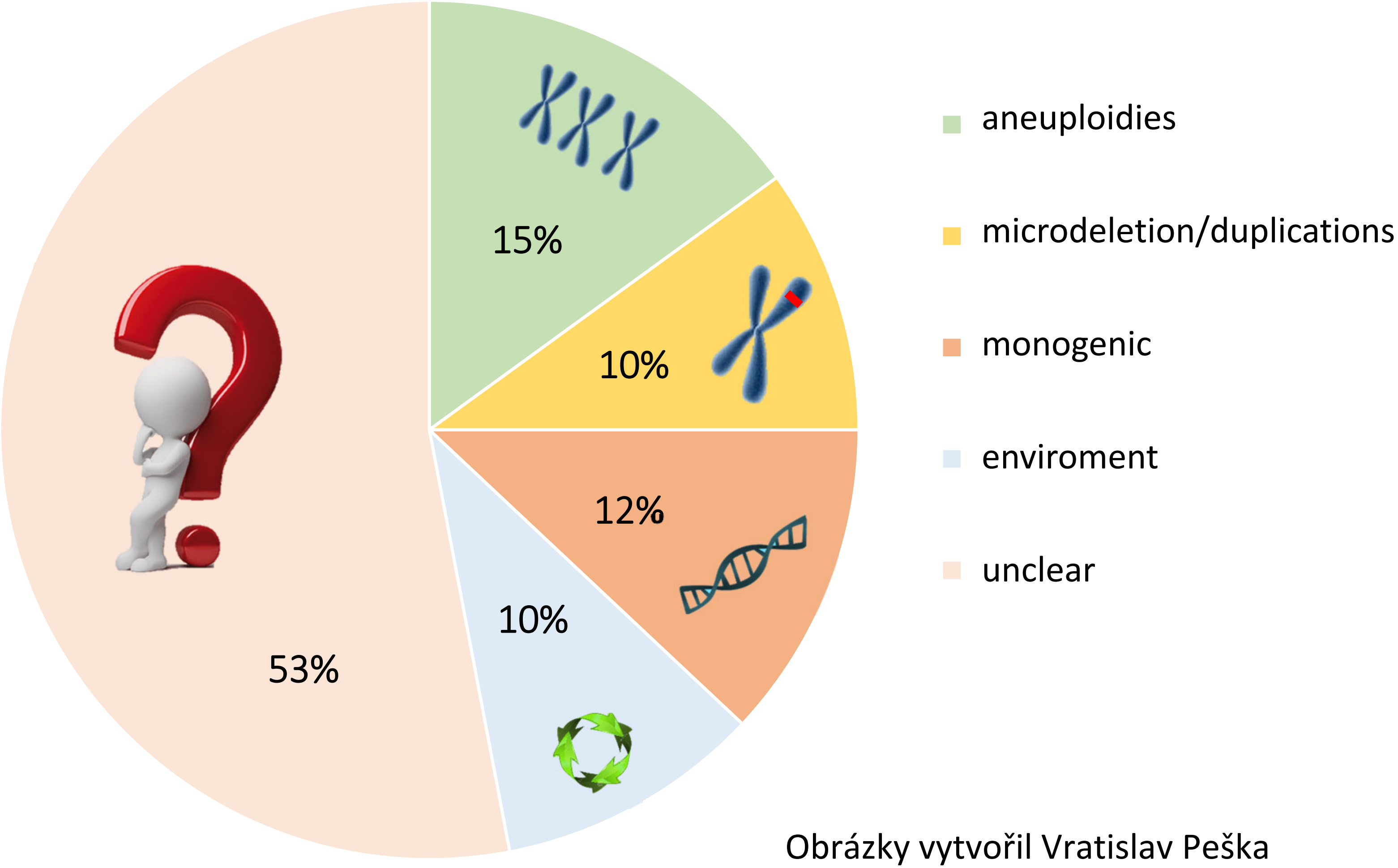
230 genes
CHD ~ 90%



Obrázky vytvořil Vratislav Peška

RASopathies

Ethiology of CHD



Obrázky vytvořil Vratislav Peška

WES – celoexomové sekvenování

WES – celoexomové sekvenování

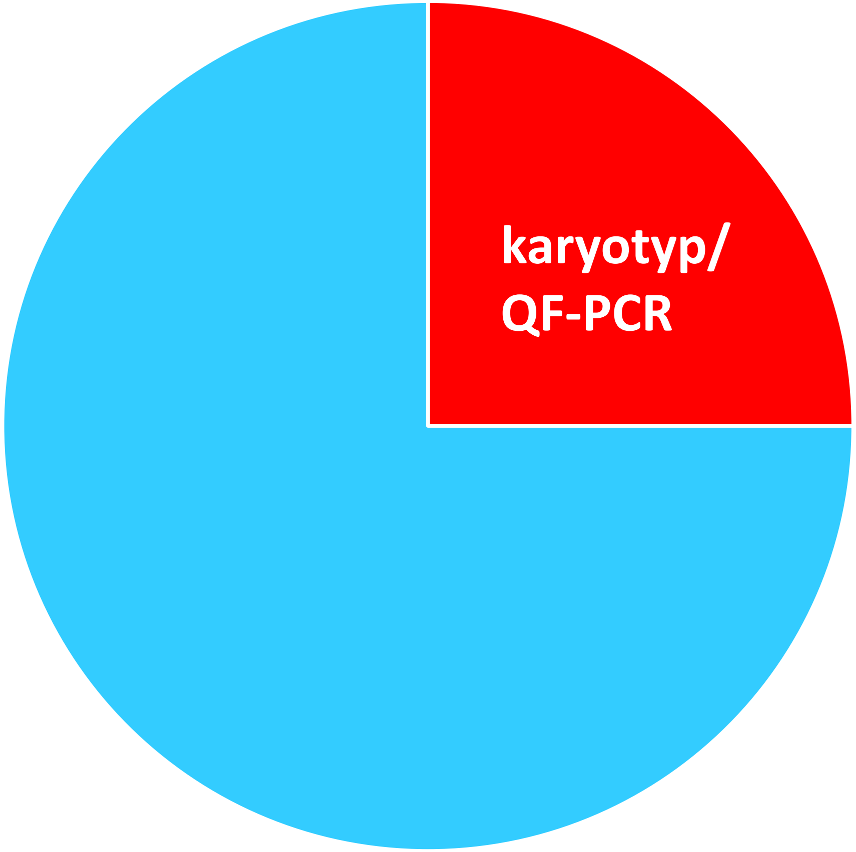
Milroy syndrom

Mutace v genu ***FLT4*** c.3075G>A p.(Met1025Ile)

vascular endothelial growth factor 3

- regulace vývoje struktur lymfatického systému

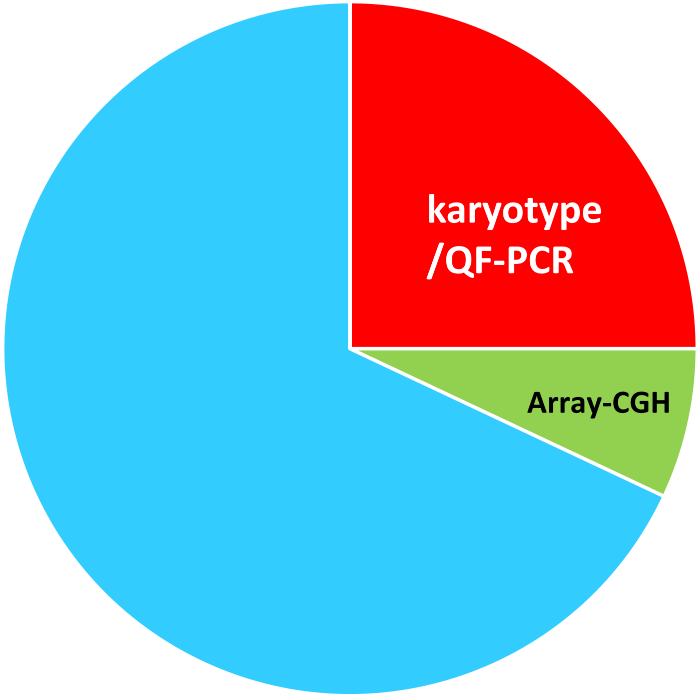
KARYOTYP/QF-PCR diagnostická výtěžnost



- Downův syndrom
- Patauův syndrom
- Edwardsův syndrom
- Turnerův syndrom
- Triploidie
- další aneuploidie
- nebalancované přestavby (5-10 Mb)

Obrázky vytvořil Vratislav Peška

KARYOTYP/QF-PCR + array-CGH
diagnostická výtěžnost



Downův syndrom

Pataův syndrom

Edwardsův syndrom

Turnerův syndrom

Triploidie

další aneuploidie

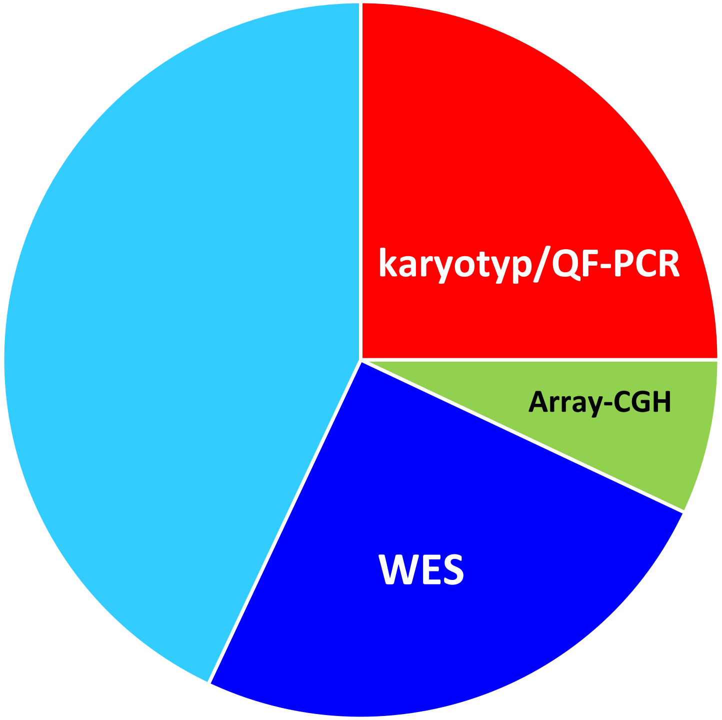
nebalancované přestavby (5-10 Mb)

Mikrodelece a mikroduplikace (10 kb)

UPD, AOH

Obrázky vytvořil Vratislav Peška

KARYOTYP/QF-PCR + array-CGH + WES diagnostická výtěžnost



Downův syndrom

Pataův syndrom

Edwardsův syndrom

Turnerův syndrom

Triploidie

další aneuploidie

nebalancované přestavby (5-10 Mb)

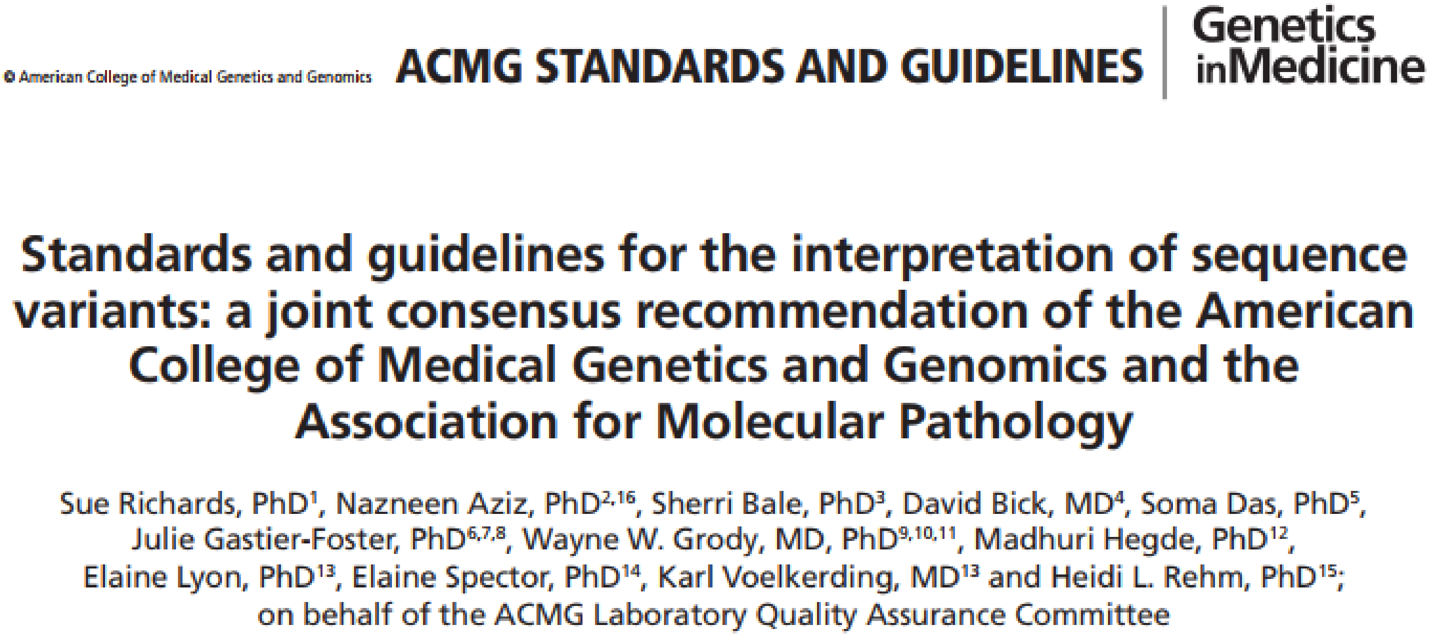
Mikrodelece a mikroduplikace (10 kb)

UPD, AOH

Monogenní nemoci

Obrázky vytvořil Vratislav Peška

Interpretace SNP variant



© American College of Medical Genetics and Genomics **ACMG STANDARDS AND GUIDELINES** | Genetics inMedicine

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5],
Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12],
Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15];
on behalf of the ACMG Laboratory Quality Assurance Committee

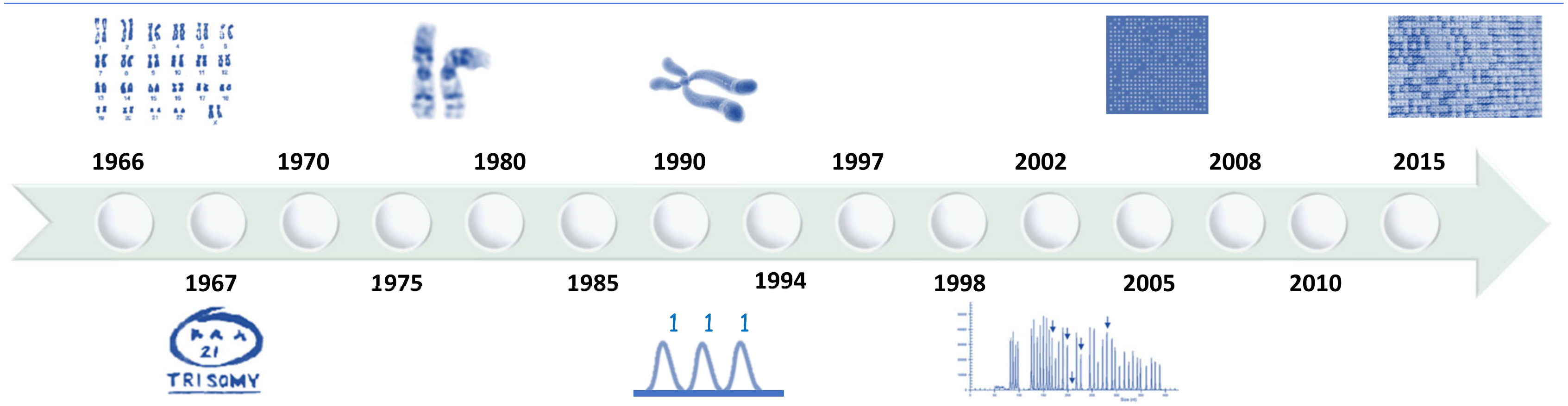
Obrázky vytvořil Vratislav Peška

Databáze

- aktuální verze!!! (hg38 a GRCh38)
- **UCSC, NCBI, Ensembl** – Genome Browser

- **LOVD**
- **HGMD**
- **ClinVar**
- **OMIM + dbSNP**
- **UniProt**

Obrázky vytvořil Vratislav Peška



Další zdroje fetálního materiálu pro prenatální diagnostiku

Volná fetální DNA v maternální cirkulaci

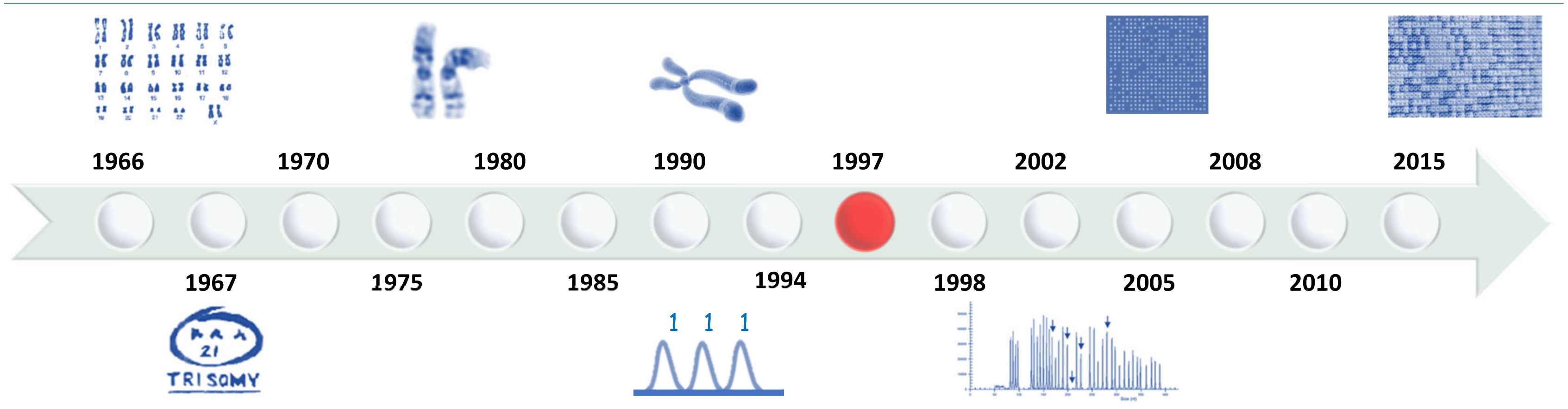
Detekovatelná od 5. tg

Po porodu mizí za 30-60 minut

3 – 25 % celkové cirkulující DNA

Původ z trofoblastu

Obrázky vytvořil Vratislav Peška



Volná fetální DNA cirkulující v krvi matky

určení pohlaví

- pro X vázaná onemocnění, př. hemofilie, DMD
- při nejasnosti genitálu na UZ

RhD genotypizace plodu

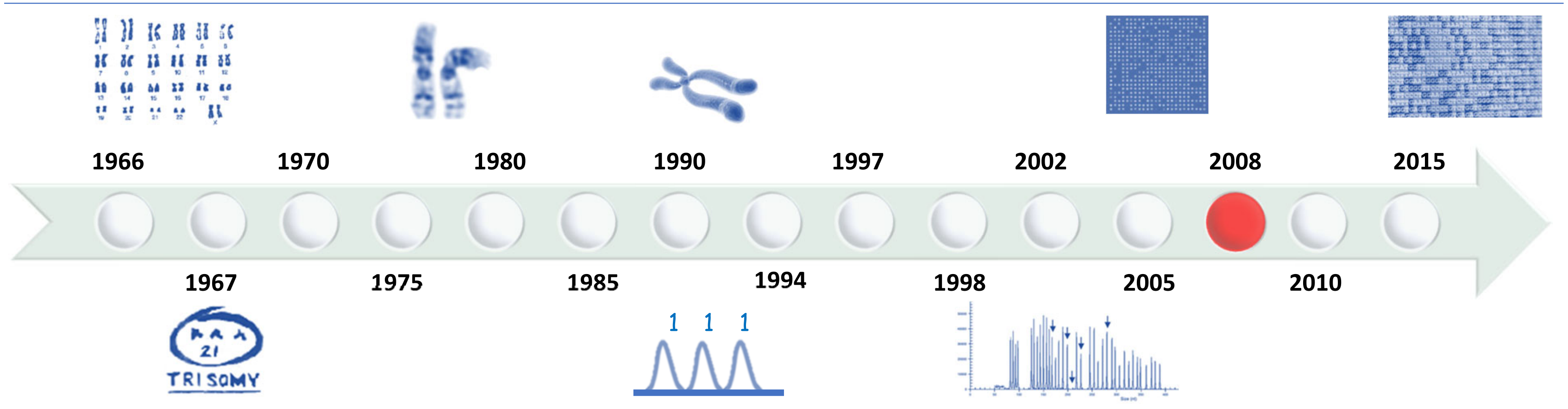
určování otcovství

monogenní onemocnění

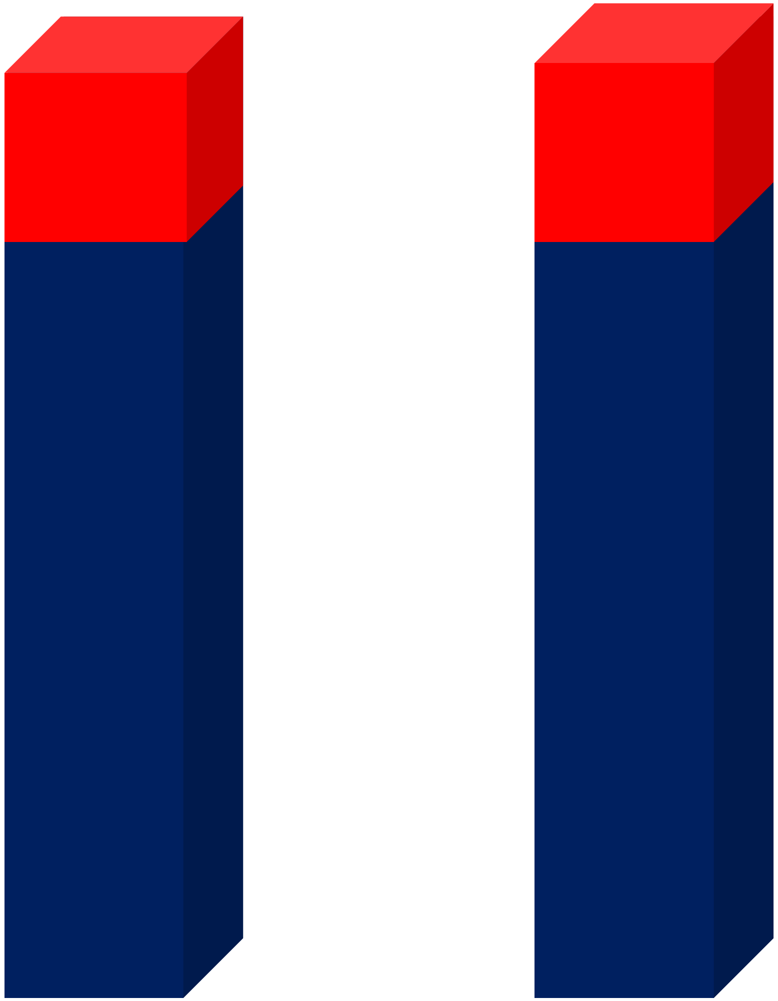
AD: achondroplázie, thanatoformní dysplázie, Huntingtonova chorea, myotonická dystrofie
AR: beta-talasemie, cystická fibróza, kongenitální adrenální hyperplázie
X vázené: hemofilie, retinitis pigmentosa

NIPT aneuploidií

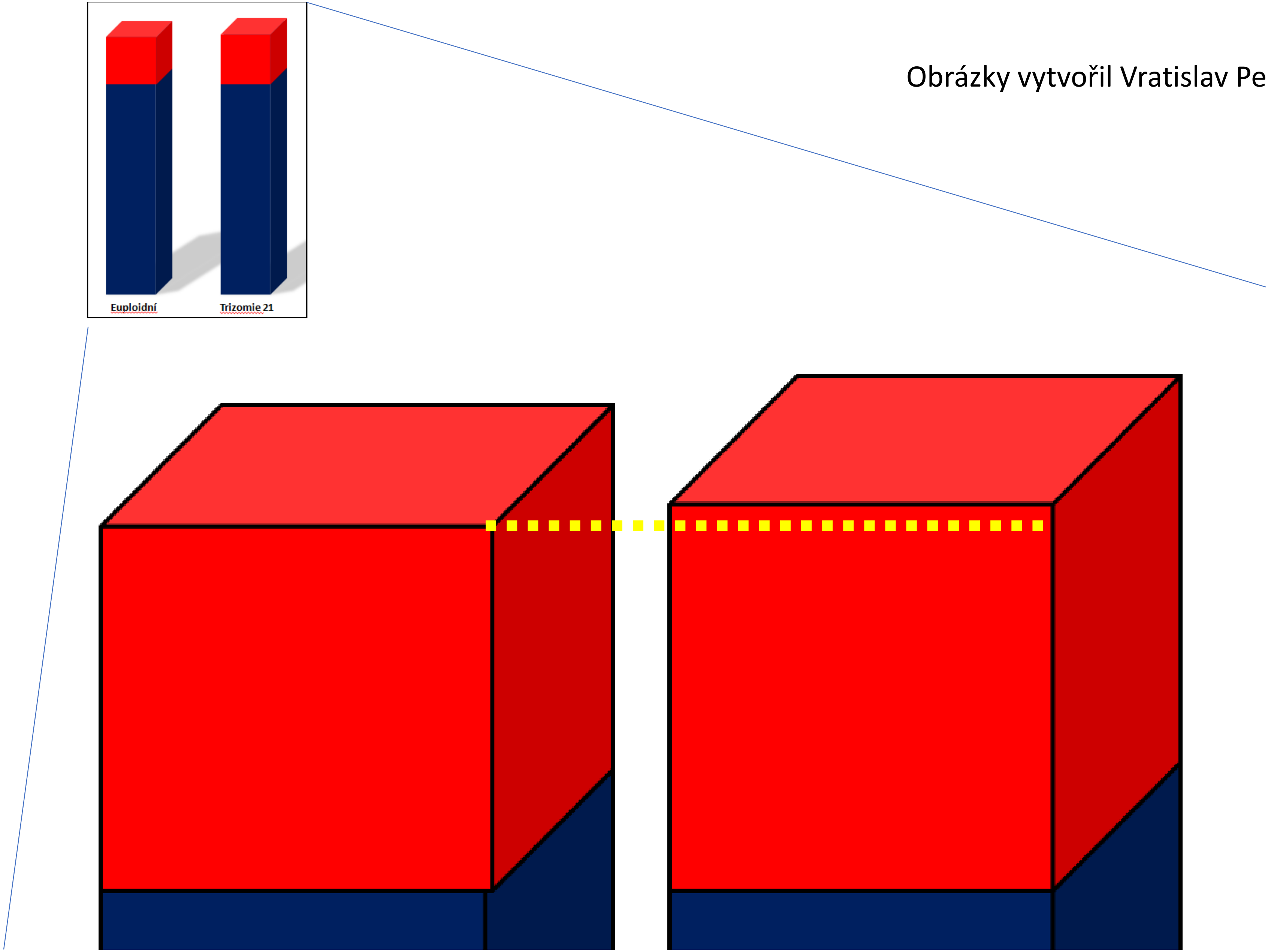
Obrázky vytvořil Vratislav Peška



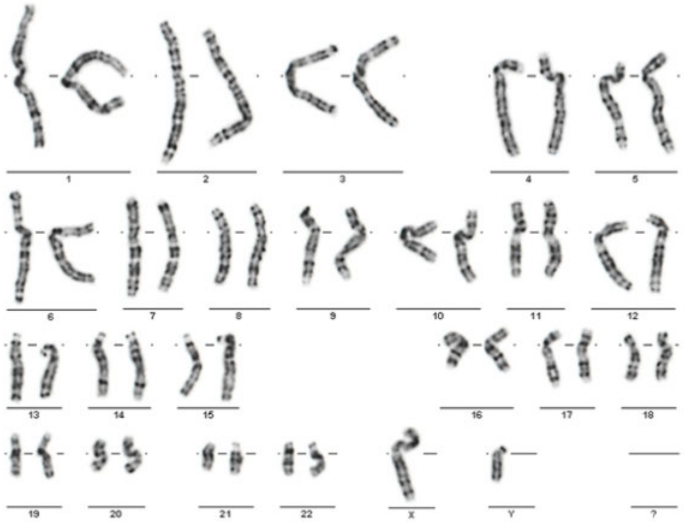
1966
1967
1970
1975
1980
1985
1990
1994
1997
1998
2002
2005
2008
2010
2015



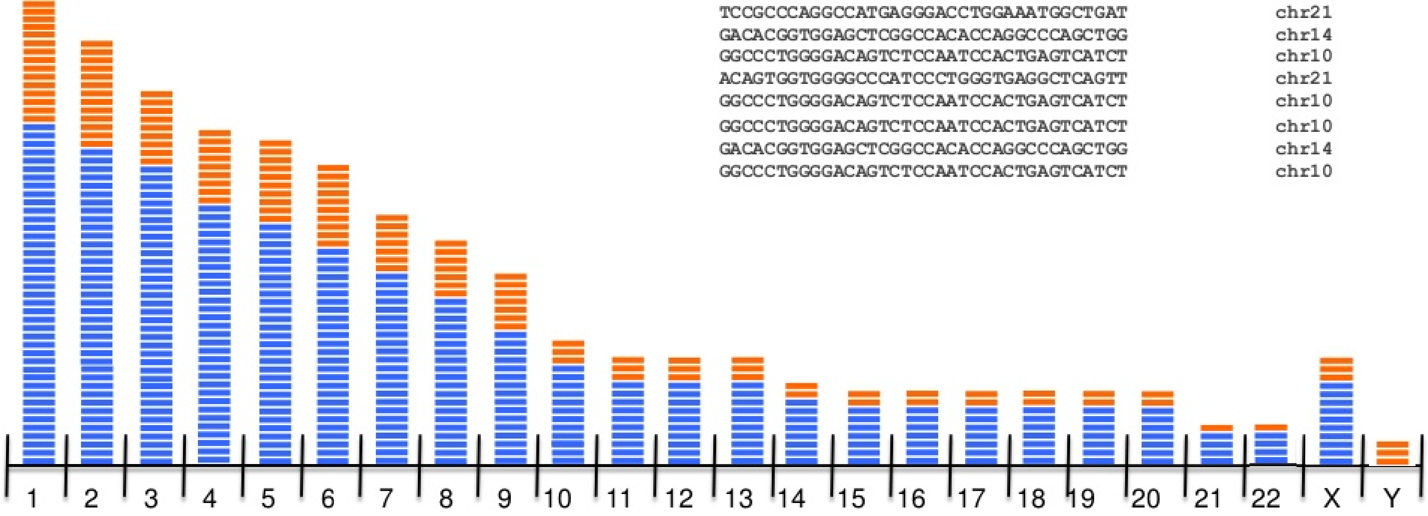
Obrázky vytvořil Vratislav Peška



Obrázky vytvořil Vratislav Peška



Obrázky vytvořil Vratislav Peška

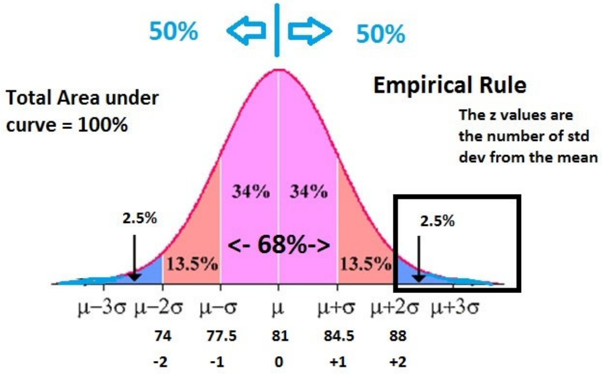


GGCCCTGGGGACAGTCTCCAATCCACTGAGTCATCT	chr10
TCCGCCCAGGCCATGAGGGACCTGGAAATGGCTGAT	chr21
GACACGGTGGAGCTCGGCCACACCAGGCCCAGCTGG	chr14
GGCCCTGGGGACAGTCTCCAATCCACTGAGTCATCT	chr10
ACAGTGGTGGGGCCCATCCCTGGGTGAGGCTCAGTT	chr21
GGCCCTGGGGACAGTCTCCAATCCACTGAGTCATCT	chr10
GGCCCTGGGGACAGTCTCCAATCCACTGAGTCATCT	chr10
GACACGGTGGAGCTCGGCCACACCAGGCCCAGCTGG	chr14
GGCCCTGGGGACAGTCTCCAATCCACTGAGTCATCT	chr10

Chromozóm 21

Euploidní

Trizomie 21



z- score

NCV – normalised chromosome value

Obrázky vytvořil Vratislav Peška

Falešně pozitivní / negativní / zkreslené výsledky NIPT

- feto-placentární mozaicismus

  1-2% všech těhotenství

Obrázky vytvořil Vratislav Peška

Falešně pozitivní / negativní / zkreslené výsledky NIPT

- feto-placentární mozaicismus
- „mizející“ dvojčata (8 týdnů; 0,23%)
- rozdílný karyotyp u vícerčat
- maternální karyotyp, mozaicismus
- terapie (transfuze, transplantace…)
- malignita (1 na 1000 těhotenství)

Forenzní analýza DNA

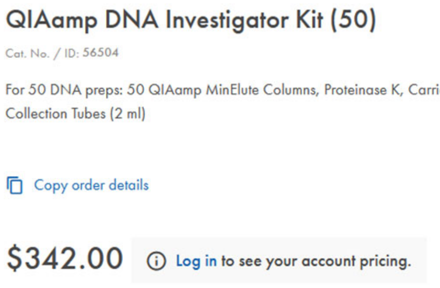
Obrázky vytvořil Vratislav Peška

Forenzní analýza DNA – vzorky

- Stěry (FLOQ) – bukální stěry úplně izolováno od všeho ostatního
- Věcné stopy – laboratoř (kousek látky, případně dodatečný stěr) – krev vata, jinak FLOQ

– izolace

- Izolační kity např. QIAGEN



QIAamp DNA Investigator Kit (50)

Cat. No. / ID: 56504

For 50 DNA preps: 50 QIAamp MinElute Columns, Proteinase K, Carrier RNA, Buffers, Collection Tubes (2 ml)

Copy order details

$342.00 Log in to see your account pricing.

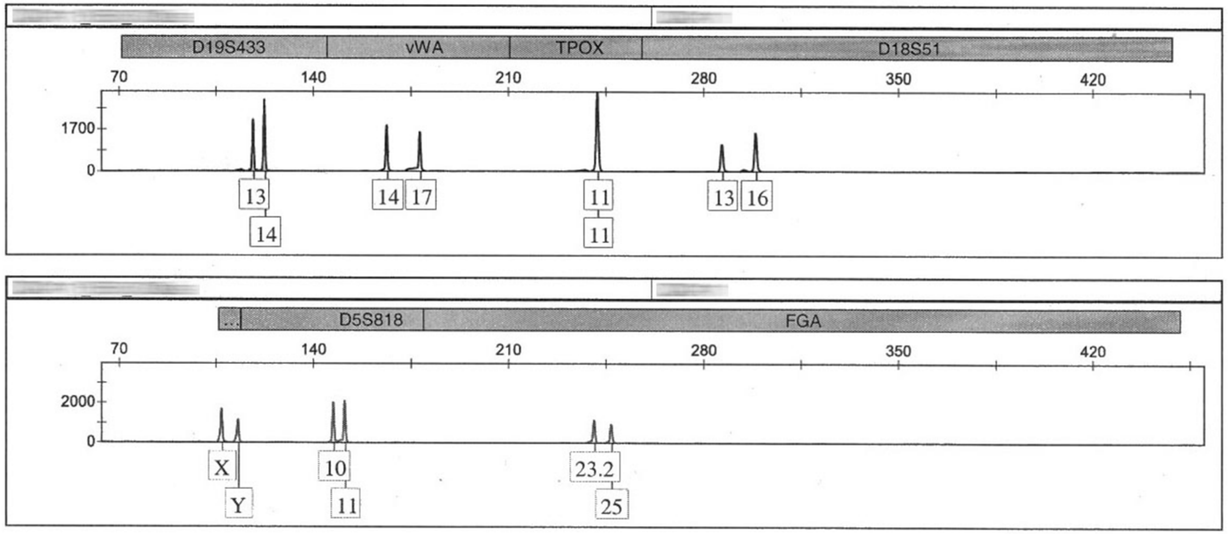
– kvantifikace, amplifikace, vyhodnocení

Forenzní analýza DNA - kvantifikace

- Je potřeba 5pg/ul (17ul do reakce)
- Stanovení pomocí qPCR (Silné vzorky, tj. krev, sliny, sperma – více pod kontrolou ruční příprava vzorků, Slabé vzorky, tj. ze zbraní, páčidel, klik atd. pomocí pipetovacích robotů)
- Celková lidská DNA + Y DNA
- Když pod limit -> konec
- Když nad limit -> amplifikace 23+1 STR markerů

Forenzní analýza DNA - amplifikace

- Analýza 23+1 STR markerů
- Znásilnění – vylučovací metoda Y
- mtDNA – vylučovací metoda v ženské linii



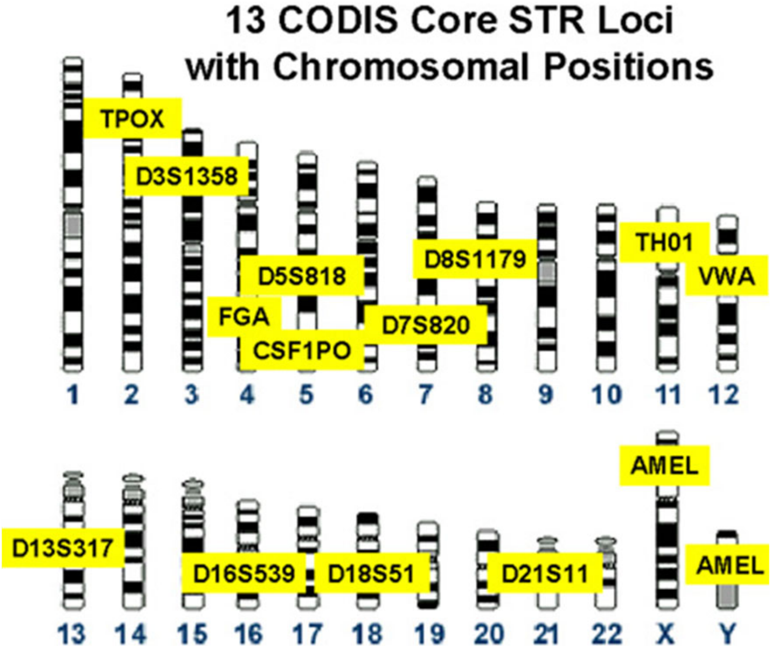
Obrázky vytvořil Vratislav Peška

Forenzní analýza DNA - vyhodnocení

- geneMapper
- CODIS – combined DNA index system (FBI)
- Databáze DNA PČR (jména nezávisle na CODIS)

CODIS

CODIS is an acronym for Combined DNA Index System, which is a computer software program that operates local, state, and national databases of DNA profiles from convicted offenders, unsolved crime scene evidence, and missing persons.


13 CODIS Core STR Loci with Chromosomal Positions

Obrázky vytvořil Vratislav Peška

Závěr:

Seznámili jsme se s metodami analýzy (např. genomových) sekvencí v základním i aplikovaném výzkumu.

Uvedli jsme si, kde genomové sekvence získat (sekvenací/stahováním z databází) a jak pracovat s velkými objemy datových souborů

Probrané databáze: GenBank, SRA, a přehled několika dalších

Vysvětlili jsme si funkci a strukturu formátů, ve kterých jsou sekvenační data uchovávána a zpracovávána

Probrali jsme sw Geneious, FASTQC, JDotter, RepeatExplorer, úvod práce v linuxu a příkazové řádce

Zběžně jsme se seznámili s forenzní analýzou DNA a sekvenační prenatální diagnostikou