MUNI
SCI

MASARYK UNIVERSITY
FACULTY OF SCIENCE
DEPARTMENT OF EXPERIMENTAL BIOLOGY
LABORATORY OF MICROBIAL MOLECULAR DIAGNOSTICS

Bi5000 – Bioinformatika

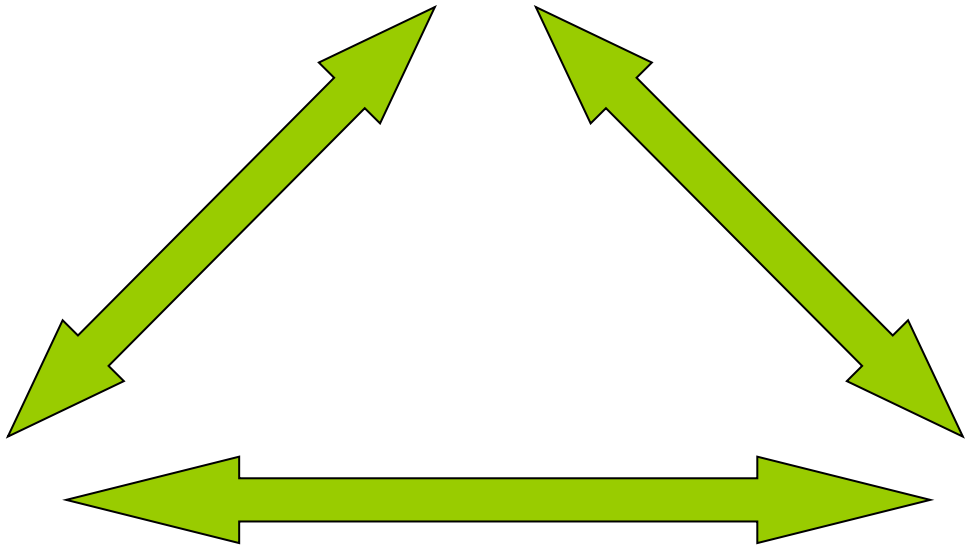
Zpracování sekvenčních dat pro zaslání do nukleotidových databází

Mezinárodní spolupráce primárních sekvenčních databází

Všechny tři hlavní bioinformatická centra (databáze) sídlejí stejná data

NIH
NCBI
ENTREZ
GenBank

Zakladatel
Instituce
Prohlížeč
Databáze



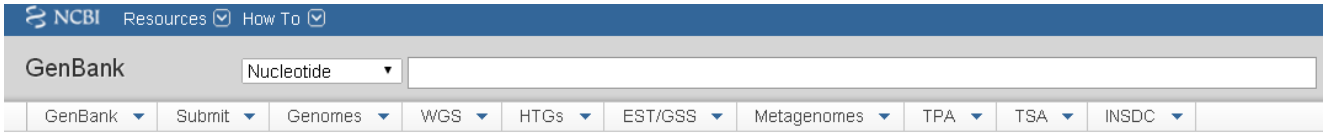
NIG
CIB
Get Entry
DDBJ

EMBL
EBI
SRS
ENA

Divize GenBank

https://www.ncbi.nlm.nih.gov/genbank/htgs/divisions/

ftp://ftp.ncbi.nlm.nih.gov/genbank



GenBank Database Divisions

GenBank divisions are divided into two general categories and were described in an (Genome Research (1997) 7(10)) article by Ouellette and Boguski; the full-text article is available (Database Divisions and Homology Search Files: A Guide for the Perplexed). The "Organismal" category includes databases pertaining to sequences derived from specific organisms and the "Functional" databases pertain to different types of sequence data being collected. Sequence records exist only in one GenBank division. For example, the HTG division includes unfinished sequences (phases 0, 1, and 2) being generated from several different organisms. As a sequence is updated to phase 3, it is moved into the appropriate organismal division. For instance, human phase 3 (finished) HTG sequences are located in the PRI division. The GenBank divisions listed here represent the location of the annotated sequence records; for homology search purposes the records are reformatted and stored in the BLAST databases. The different database divisions currently available, as well as the related BLAST database, are listed below. An example of a submission (one accession number) that has progressed through phase 1, phase 2, and phase 3 is available (Examples).

HTGs Resource

About HTGs
Submitting HTGs
Processing HTGs
HTGs FAQ

Organismal Divisions:

Database	Division	BLAST	Example
BCT	Bacterial sequences	nr, month	
PRI	Primate sequences	nr, month	Human Phase 3
ROD	Rodent sequences	nr, month	
MAM	Other mammalian sequences	nr, month	
VRT	Other vertebrate sequences	nr, month	
INV	Invertebrate sequences	nr, month	Drosophila, C. elegans Phase 3
PLN	Plant and Fungal sequences	nr, month	Arabidopsis Phase 3
VRL	Viral sequences	nr, month	
PHG	Phage sequences	nr, month	
RNA	Structural RNA sequences	nr, month	
SYN	Synthetic and chimeric sequences	nr, month	
UNA	Unannotated sequences	nr, month	

← Divize podle organizmů

Functional Divisions:

Database	Division	BLAST	Example
EST	Expressed Sequence Tags	dbest, month	
STS	Sequence Tagged Sites	dbsts, month	
GSS	Genome Survey Sequences	dbgss, month	
HTG	High Throughput Genomic sequences	htgs, month	All Organisms: Phase 0, 1, and 2

← Funkční divize

Postup zpracování sekvenčních dat

- **Primární analýza dat ze sekvenátoru**
  - Zachycení obrazu
  - Zpracování obrazu
  - Base calling
  - Provedení kontroly kvality
- **Sekundární analýza dat**
  - Alingment (RefSeq data)
    - Mismatches, validace, vizualizace
  - Assembly -> kontigy/scaffoldy
- **Terciární analýza - automatická anotace**
  - Anotace, Geny, Variace, Diferenciální exprese, Metylace, atd.

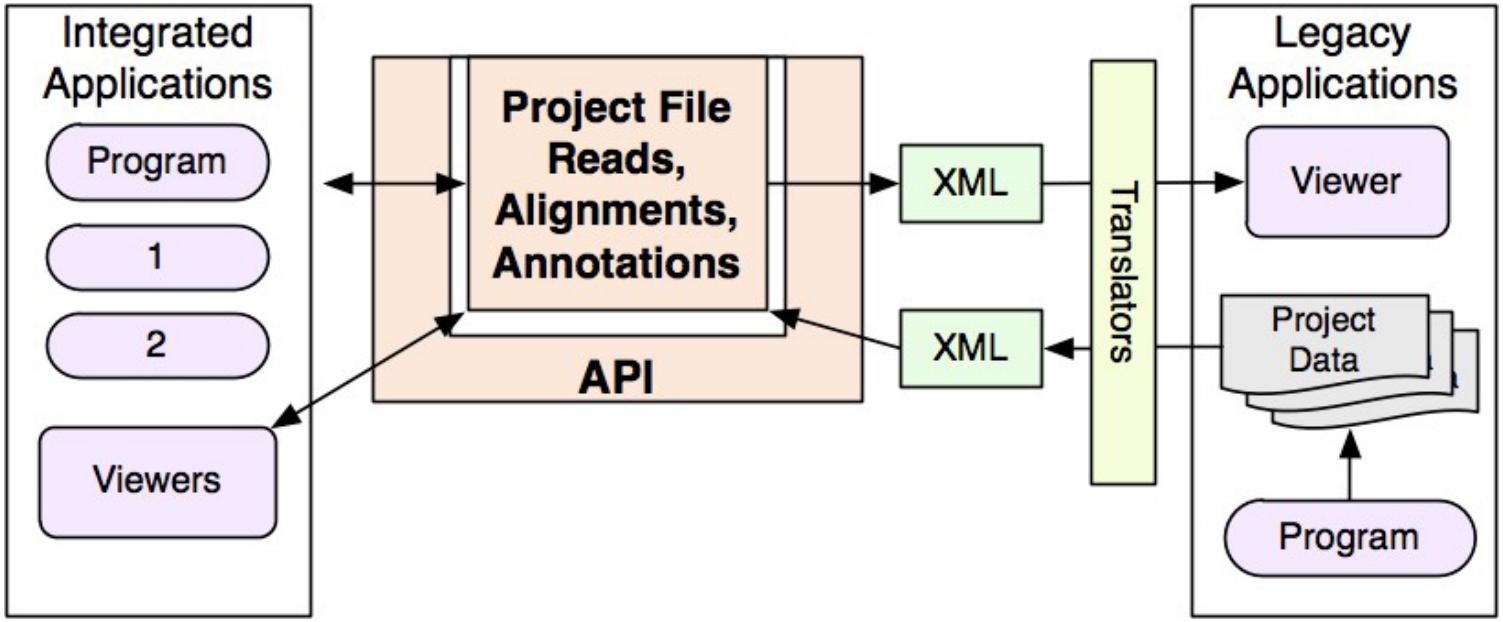
Sequence Read Archive: Nezpracovaná sekvenační data pro zvýšení reprodukovatelnosti a usnadnění nových objevů.

Whole Genome Shotgun: sestavy neúplných nebo kompletních genomů a chromozomů prokaryot a eukaryot získané NGS.

Řada specializovaných databází: GenBank, Genomy, SNP, GEO

Komplexní přístup vs. manuální zpracování

Komplexní přístupy vyžadují strukturovaná, indexovaná a anotovaná data

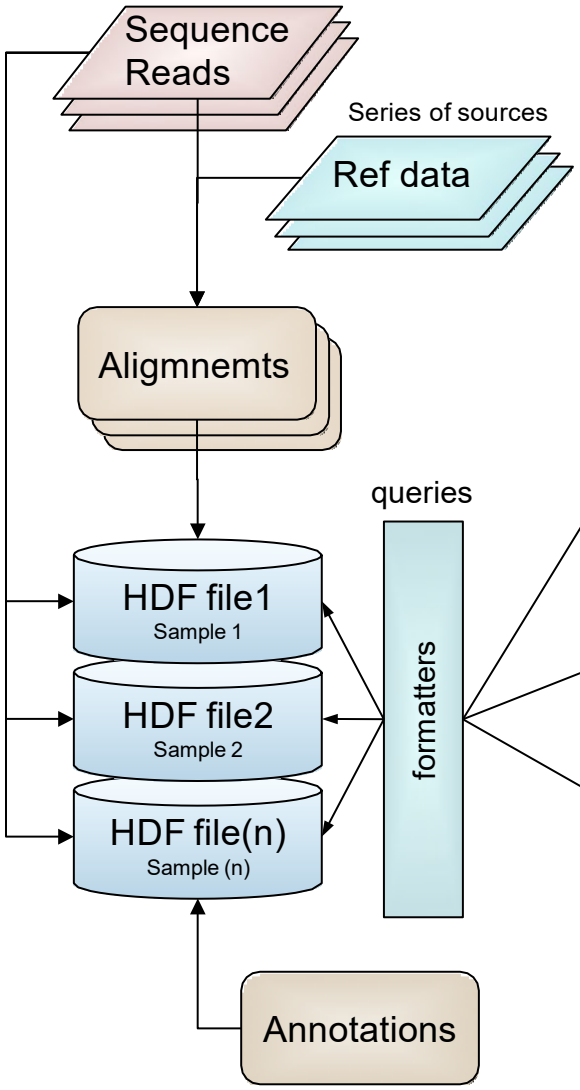


- Zpracování velkého množství dat
- Integrace více typů dat
- Porovnávání vzorků mezi sebou

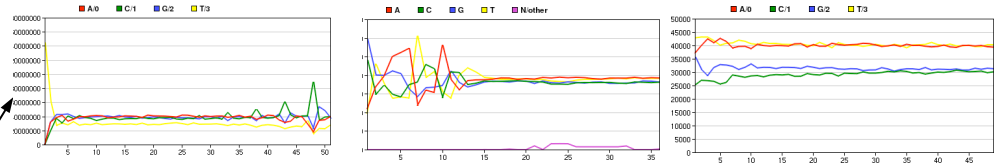
Formáty sekvenačních dat z NGS

- FASTQ
  - Illumina
  - Standardní formát akceptovaný databázemi
- SFF, BAM, qual
  - starší formáty používané 454 a Ion Torrent
  - Obsahují alignment jednotlivých čtení
  - Většinou lze konvertovat do FASTQ
- fast5, HDF5 - Hierarchické formáty dat
  - formáty pro dlouhá čtení
  - PacBio, Nanopore

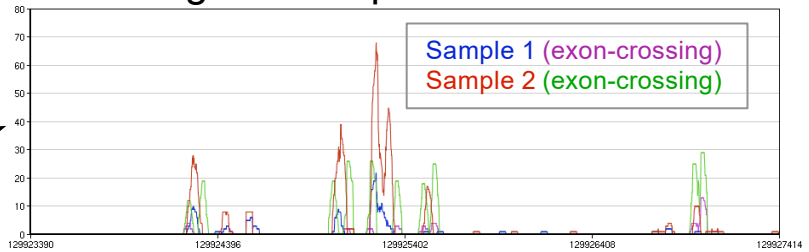
Hlubší integrace dat do komplexních databází (BioHDF)



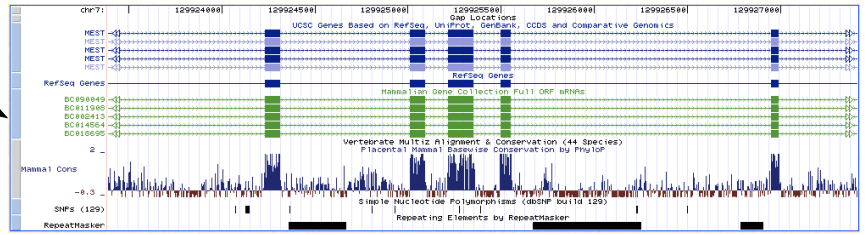
Integrate data across platforms



Integrate samples / annotations

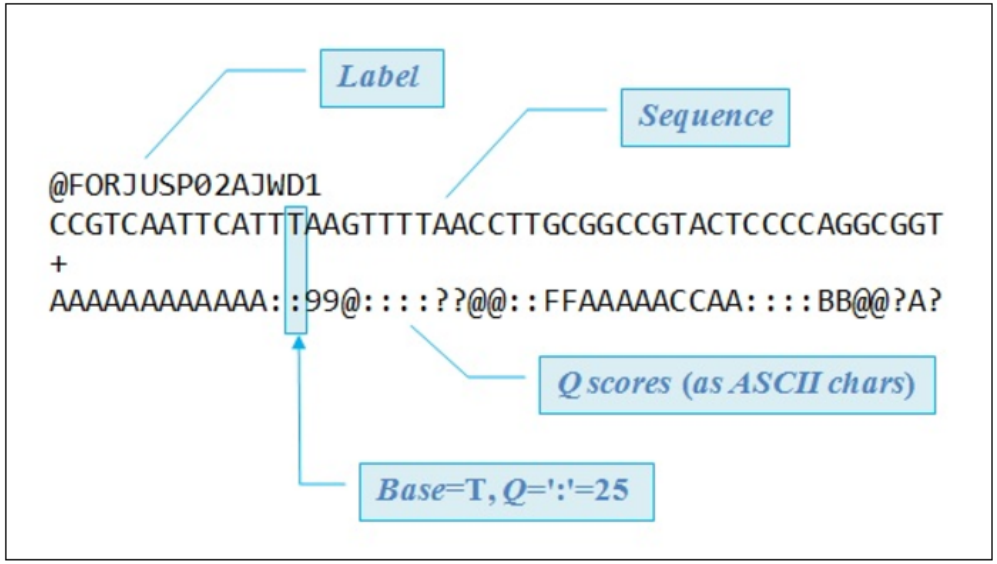


Integrate between systems



FASTQ formát

- Univerzální formát akceptovaný databázemi
- Každý záznam obsahuje 4 řádky
  - řádek 1 začíná hlavičkou '@'ID sekvence
  - Řádek 2 obsahuje primární sekvenci
  - Řádek 3 začíná '+' a může následovat stejné ID a popis
  - Řádek 4 obsahuje zakódované hodnoty o kvalitě sekvence a musí obsahovat stejný počet znaků jako řádek 2



Starší Sequence-alignment-mapping formáty (SAM/BAM) struktura dat

- 454, IonTorrent a starší PacBio
  - Tabulkový textový formát
  - Odvozený z formátu pro sekvenční přiložení
  - Alignment je součástí formátu
    - má volnou definici
    - http://genome.sph.umich.edu/wiki/SAM#What_is_SAM
- BAM je zkomprimovaná binární verze nekomprimovaného formátu SAM (sequence, alignment, mapping)
  - SAM Tools
- Nese informace
  - Kvalita bází (QUAL+33)
  - Mapování k referenční sekvenci

Example Header Lines

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta        M5:fdfd811849cc2fadebc929bb925902e5
@RG     ID:UM0098:1     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L001 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@RG     ID:UM0098:2     PL:ILLUMINA     PU:HWUSI-EAS1707-615LHAAXX-L002 LB:80   DT:2010-05-05T20:00:00-0400     SM:SD37743      CN:UMCORE
@PG     ID:bwa  VN:0.5.4
@PG     ID:GATK TableRecalibration      VN:1.0.3471     CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, CycleCovariate,
DinucCovariate, TileCovariate], default_read_group=null, default_platform=null, force_read_group=null, force_platform=null,
solid_recal_mode=SET_Q_ZERO, window_size_nqs=5, homopolymer_nback=7, exception_if_no_tile=false, ignore_nocall_colorspace=false, pQ=5, maxQ=40,
smoothing=1
```

In the alignment examples below, you will see that the 2nd alignment maps back to the RG line with ID UM0098.1, and all of the alignments point back to the SQ line with SN:1 because their RNAME is 1.

Example Alignments

This is what the alignment section of a SAM file looks like:

```
1:497:R:-272+13M17D24M  113     1       497     37      37M     15      100338662       0       CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-
==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>        XT:A:U  NM:i:0  SM:i:37 AM:i:0  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
19:20389:F:275+18M2D19M 99      1       17644   0       37M     =       17919   314     TATGACTGCTAATAATACCTACACATGTTAGAACCAT
>>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9   RG:Z:UM0098:1   XT:A:R  NM:i:0  SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:37
19:20389:F:275+18M2D19M 147     1       17919   0       18M2D19M        =       17644   -314    GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT
;44999;499<8<8<<<8<<><<<<<<7<;<<<>><<   XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:4  X1:i:0  XM:i:0  XO:i:1  XG:i:2  MD:Z:18^CA19
9:21597+10M2I25M:R:-209 83      1       21678   0       8M2I27M =       21469   -244    CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<>
<<<>><<><>><9>><>>>9>>><>       XT:A:R  NM:i:2  SM:i:0  AM:i:0  X0:i:5  X1:i:0  XM:i:0  XO:i:1  XG:i:2  MD:Z:35
```

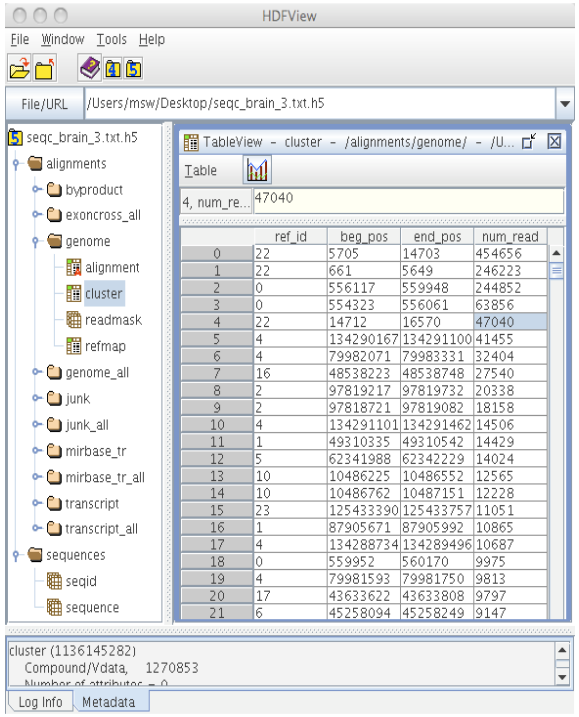
Struktura sekce Illumina Header

```
@HWUSI-
EAS611:34:6669YAAXX:5:1:5069:1159
1:N:0:
```

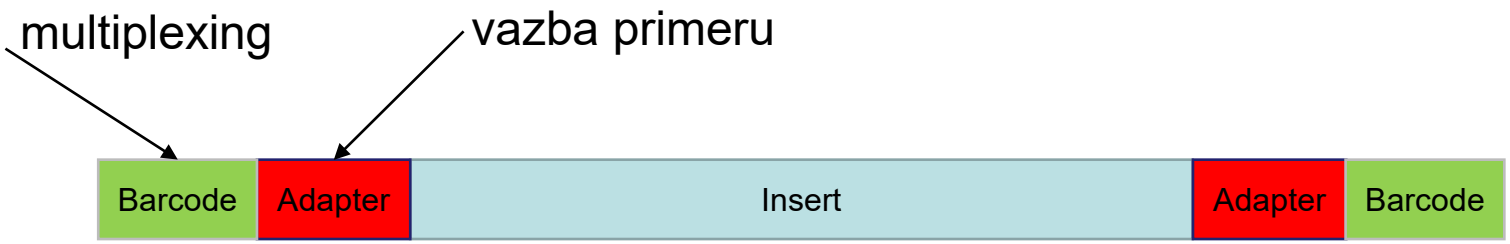
- Starts with @ (required by fastq spec)
- Instrument ID (HWUSI-EAS611)
- Run number (34)
- Flowcell ID (6669YAAXX)
- Lane (5)
- Tile (1)
- X-position (5069)
- Y-position (1159)
- [space]
- Read number (1)
- Was filtered (Y/N) (N) - You wouldn't normally see the Ys
- Control number (0 = no control)
- Sample number (only if demultiplexed using Illumina's software)

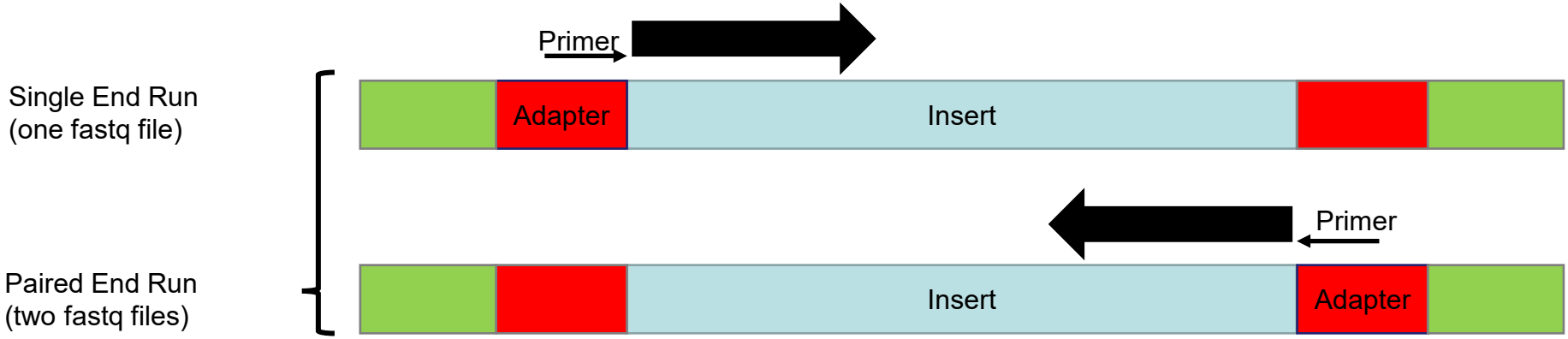
HDF5 a Fast5 formáty

- PacBio a Nanopore
- HDF (hierarchical data format)
  - zkomprimovaná složka s hierarchickou strukturou
    - Skupiny
    - Datové sady
    - Atributy
- Fast5 - komplexní formát
  - odvozený z HDF5
  - má definovanou konkrétní hierarchii souborů
- Obsahuje data pro base calling
  - vyšší informační obsah, base calling může být provedeno opakovaně
  - pro složení sekvence může být převeden do FASTQ



Jak čteme sekvence z NGS?





1. Demultiplexing
2. Base Call Quality
3. Adapter Content (trimming)
4. Mapping Quality

QC Metrics

Kontrola kvality

- Statistiky čtení
  - počet čtení a medián jejich délky
  - obsah G+C
  - skóre kvality bází a jeho distribuce
- Kontaminace
  - adaptory, primery
  - hostitelská DNA u nebuněčných genomů
  - křížové kontaminace
- Hloubka pokrytí
  - Volba vyžadované hloubky závisí na aplikaci
  - pro celogenom je potřeba minimálně 30-50x pokrytí u krátkých čtení a minimálně 100-300x u dlouhých čtení
- FastQC – příklad nástroje pro hodnocení kvality u Illumina
  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Základní statistiky

- Illumina

Filename	CCM9024_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2761796
Sequences flagged as poor quality	0
Sequence length	150
%GC	34

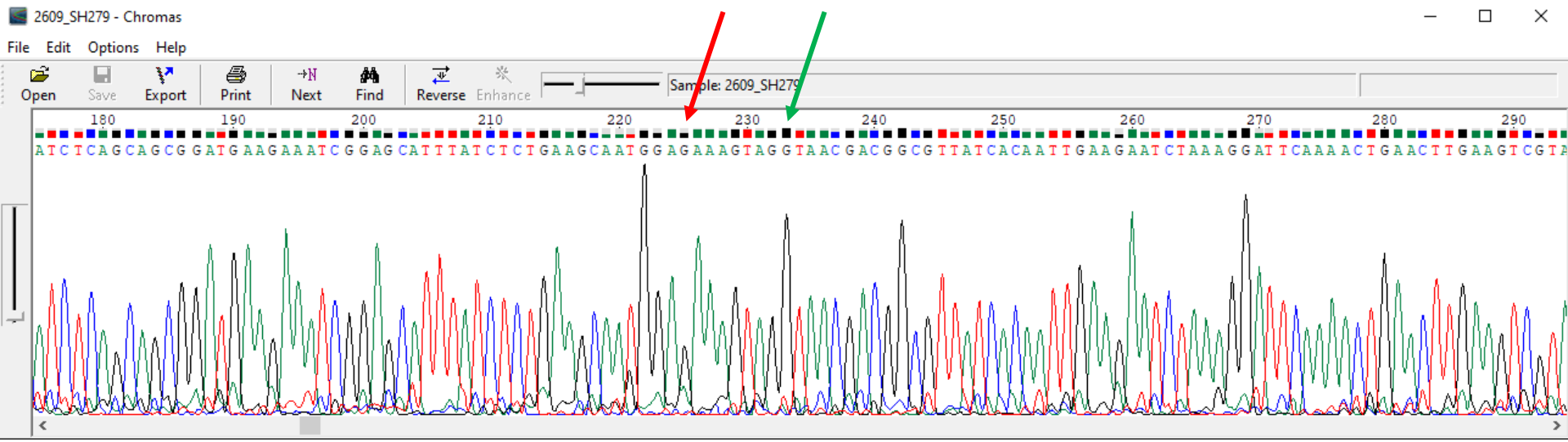
- Nanopore

```
General summary:
Mean read length:              386.9
Mean read quality:              11.8
Median read length:             99.0
Median read quality:            11.5
Number of reads:            11,266.0
Read length N50:             5,927.0
Total bases:            4,358,461.0
```

Hodnotí se především:
- počet čtení
- délka čtení
- distribuce délek čtení

Hodnocení kvality stanovení báze

- Q score nebo Phred skóre
- Celočíselná hodnota představující odhadovanou pravděpodobnost chyby
- Kódování může být specifické dle platformy



Příklad hodnocení prostřednictvím Q-skóre u Sangerova sekvenování

PHRED skóre kvality

- Skóre kvality jsou reprezentována jako znaky ASCII
- Phred+33
  - v současnosti nejpoužívanejší
  - Illumina 1.8+, Sanger, PacBio

  ! = nejnižší kvalita ~ = nejvyšší kvalita

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

- Phred+64
  - lze se s ním setkat u starších dat před r. 2018
  - Illumina 1.3+, Illumina 1.5+, Solexa
  - Před použitím dat může být nutný převod na Phred+33

Výpočet Phred Scores

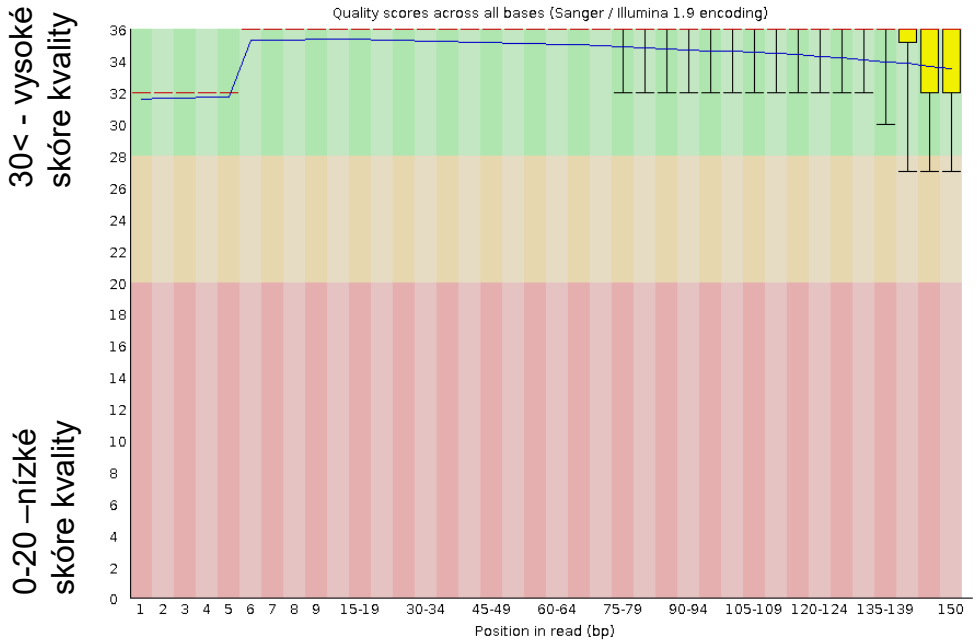
- Stanovení pravděpodobnosti, že uvedená báze je nesprávná: hodnota (p)
- Transformace na skóre Phred z plovoucí desetinné čárky, tedy kladné celé číslo:

  Phred = -10 * (int)$\log_{10}(p)$

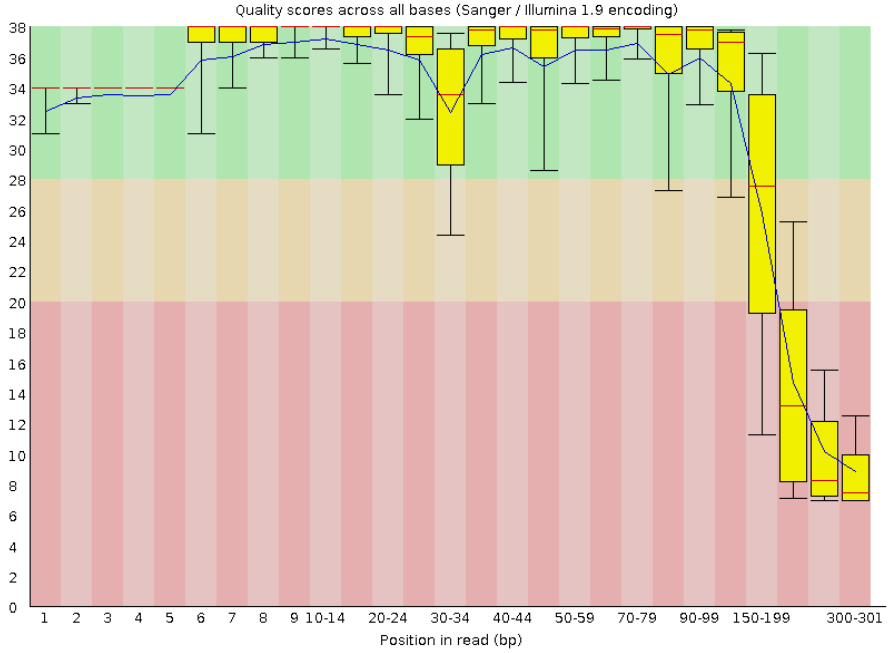
  - p=0.1 Phred = 10
  - p=0.01 Phred = 20
  - p=0.001 Phred = 30

Statistické hodnocení kvality sekvence na bázi z Illumina čtení nástrojem FastQC

Dobrá kvalita

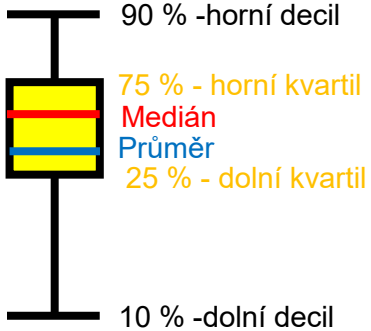


Špatná kvalita



metriky kontroly kvality, jsou hlášeny systémem varování na semaforech, normální (zelená), abnormální (oranžová), špatná (červená)

Distrubce skóre kvality

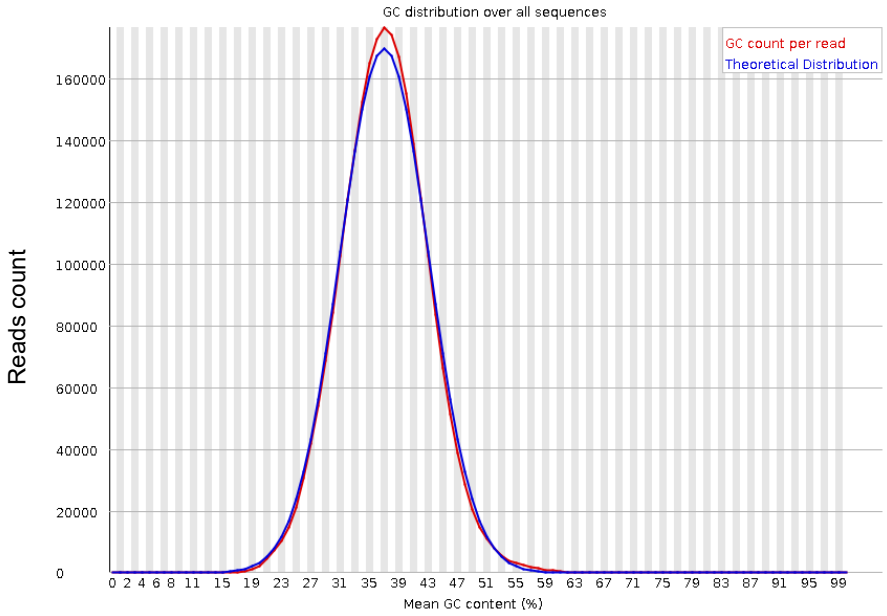


- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

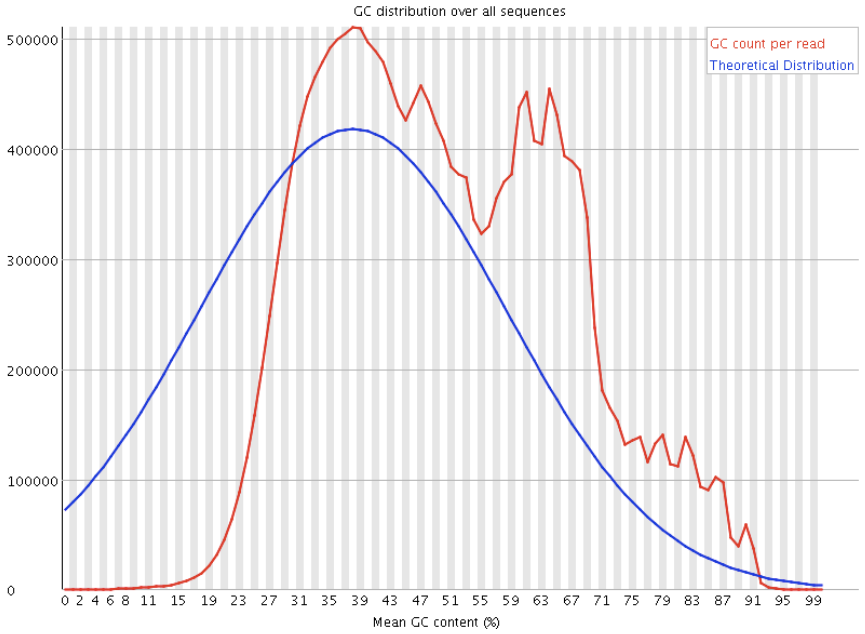
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Hodnocení distribuce G+C pomocí FastQC

- Distrubce CG je první parametr, kde se zjistí kontaminace dat
- Zdroje kontaminace mohou být různé
  - technické sekvence DNA jako primery, adaptory, hostitelská DNA, rRNA
- Pokud je znám zdroj kontaminace, pak je možné tyto sekvence odfiltrovat



Distribuce obsahu CG pro každé čtení, je dobrá a pouze s jedním vrcholem. Data tedy pravděpodobně obsahují DNA pouze z jednoho organismu.



Nerovnoměrná distrubce CG % znamená pravděpodobnou kontaminaci. V tomto případě se jedná RNAseq s nedostatečně odstraněnou rRNA.

FastQC –kontaminace adaptory

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Příliš mnoho adaptorových sekvencí může znamenat chybu při přípravě sekvenační knihovny. Řešením je opět odfiltrovat tyto sekvence, pokud je dostatek čtení. Pokud ne, může být nutné zopakovat sekvenování.

Filtrování podle kvality: co můžeme dělat s nekvalitními daty

- Odstranění nejhorších čtení podle skóre kvality
- Odstranění kontaminant
- Oříznutí (trimming)
  - Na začátku (adaptory, primery)
  - Na konci (klesající kvalita)
    - Automatické nebo manuální stanovení cutoff limitů
    - Při dostatečné hloubce pokrytí se upřednostňuje méně čtení o vyšší kvalitě
- Sloučení párových čtení
  - snížení redunance dat, pokud mají čtení velký překryv
- Obecně se podle typu dat a typu experimentu provádí odlišné kroky

Oříznutí čtení sekvence (trimming)

báze 1->5

zbytek adaptoru

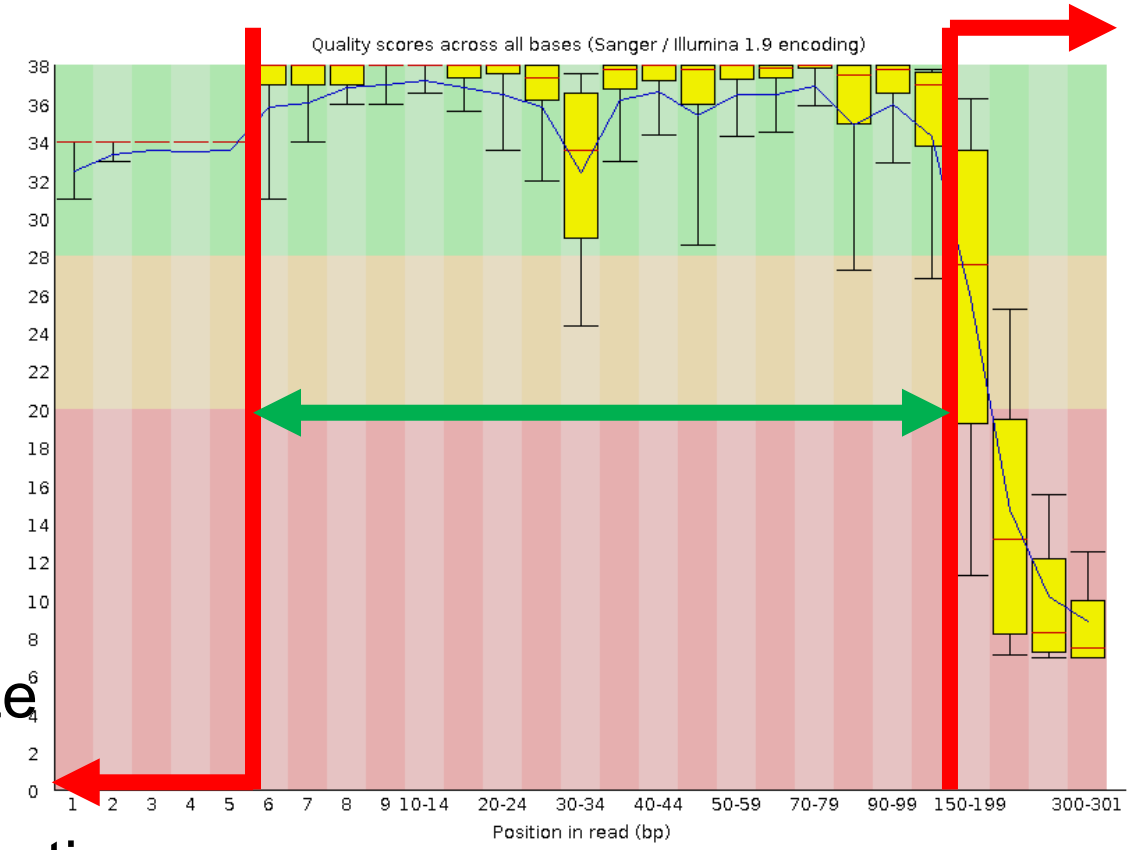
báze 6->150

zachovaný úsek

báze 150->300

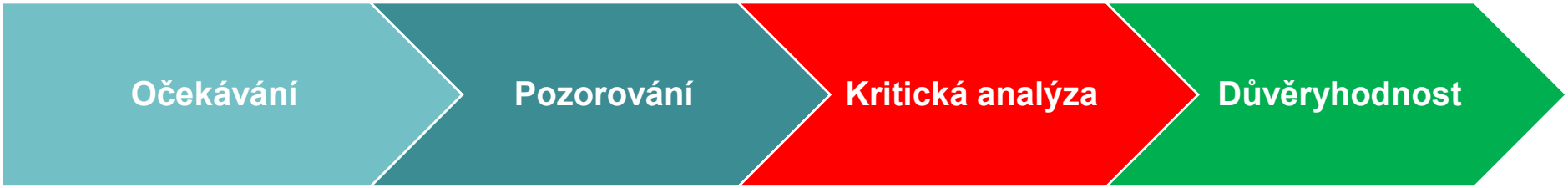
ztráta kvality

ořez cca od 200. báze



- cutadapt, trimmomatic
  - command line programy se širokými možnostmi filtrování a úprav čtení
- filtlong
  - command line, pro filtrování kvality u dlouhých čtení

Závěr: Kontext je klíčem pro hodnocení kvality



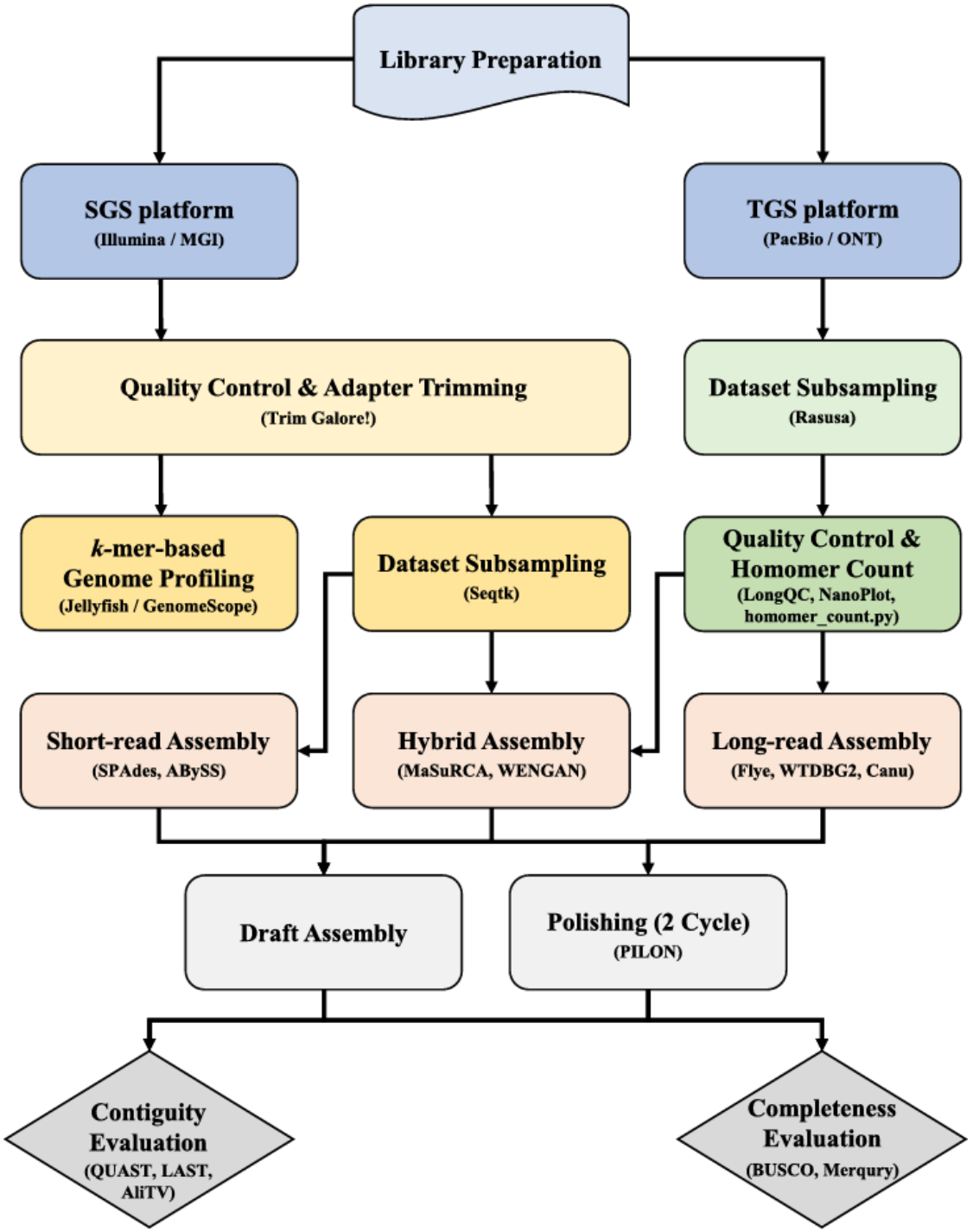
QC by měla být o tom, co očekáváte a co vidíte

Vývojový diagram sekvenování a assembly

Příklady programů

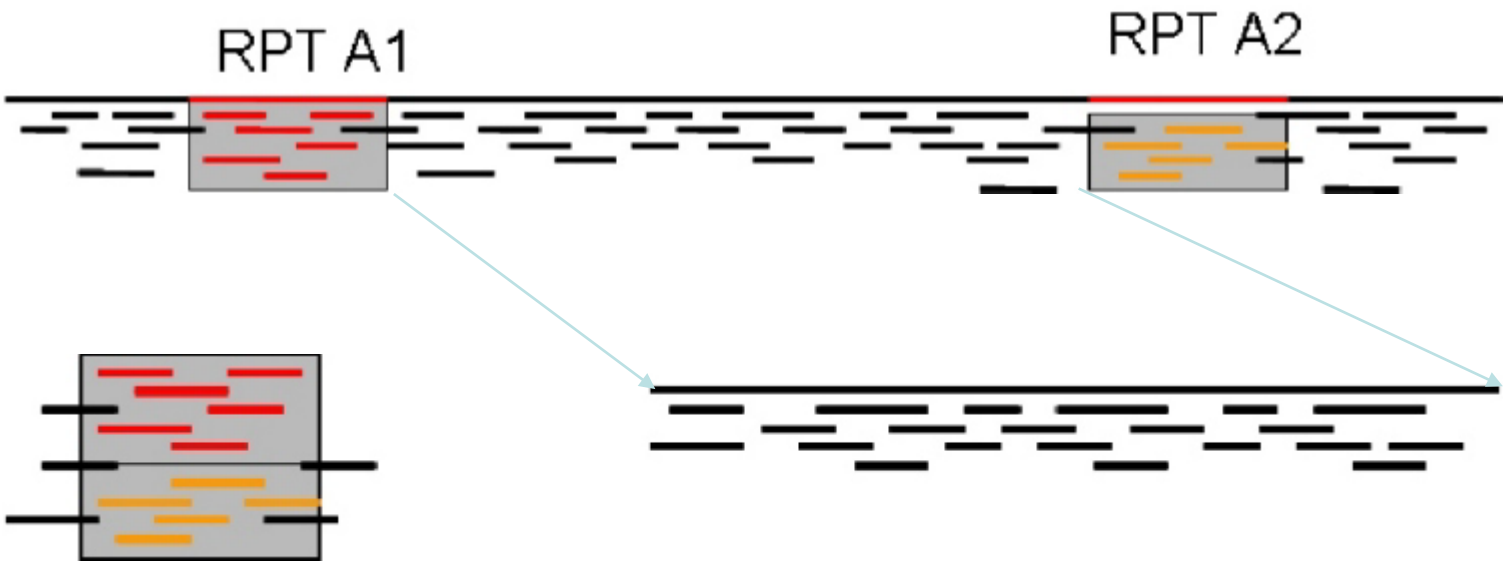
Závislé na:

- Sekvenační platformě
- Získaném pokrytí
- Typu genomu



Repetice jsou příčinou rozdělení genomů do kontigů

Jestliže čtení je kratší než repetice→ nemožnost sestavení sekvence



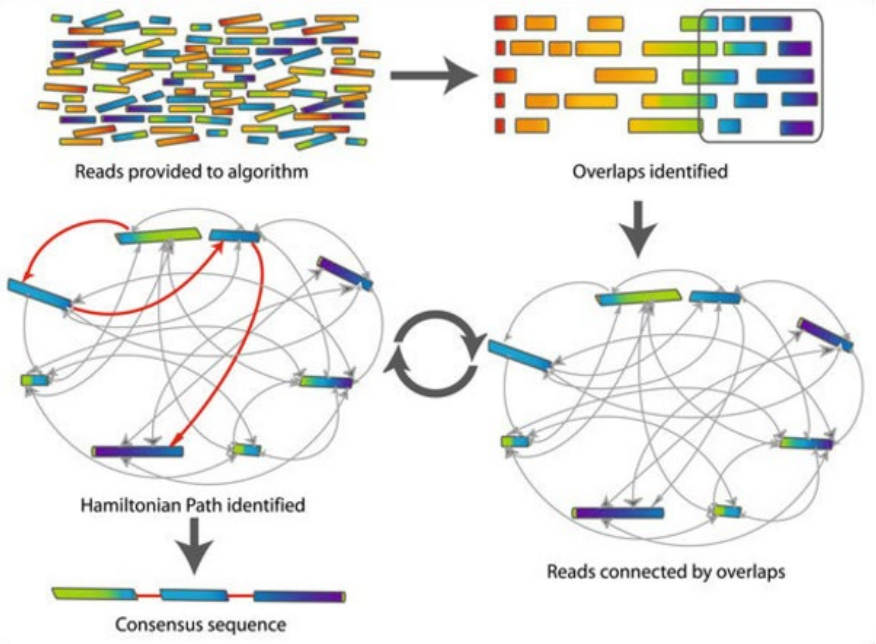
Čtení z mnoha podobných repetic vedou k vytvoření kontigů s pozměněnou strukturou

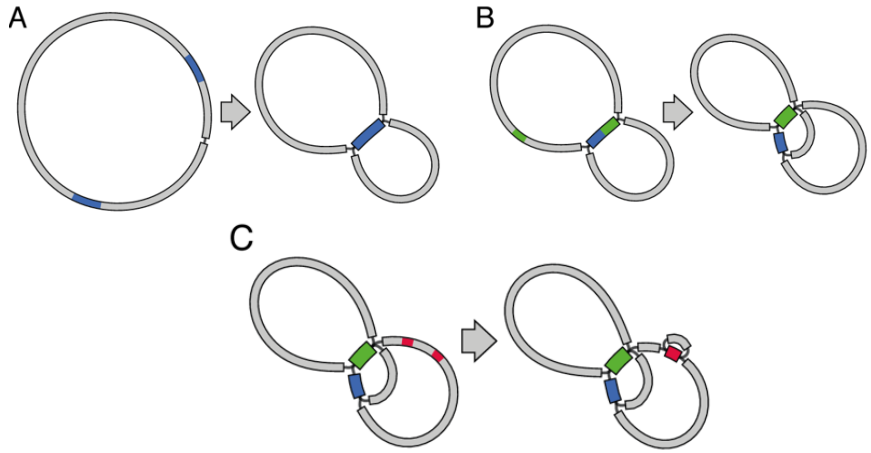
Kontig tvořený jedinečnou sekvencí, ohraničený repetitivními sekvencemi

- Krátká čtení, hlavní příčina omezení kompletního sestavení
- Stejná sekvence se vyskytuje v genomu vícekrát
- Délka čtení není schopna překlenout tuto repetici
- Pokrytí může indikovat multiplicitu

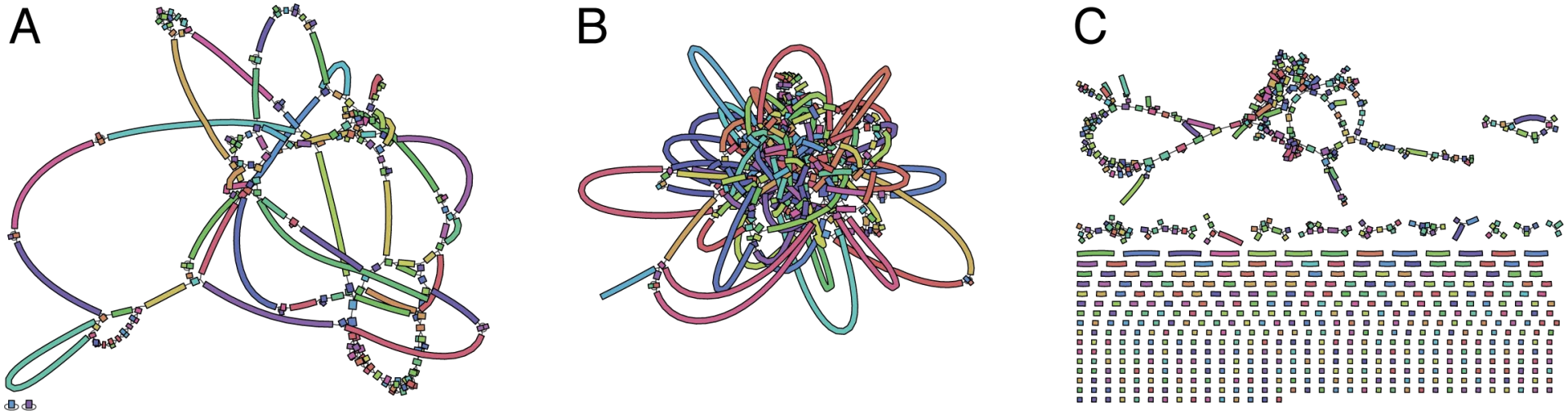
Složení genomu – assembly

- Principy assembly
  - OLC – overlap layout consensus
    - výpočetně náročné
    - vhodnější pro dlouhá čtení
  - De Bruijnův graf
    - ztráta části informace
    - vhodnější pro krátká čtení





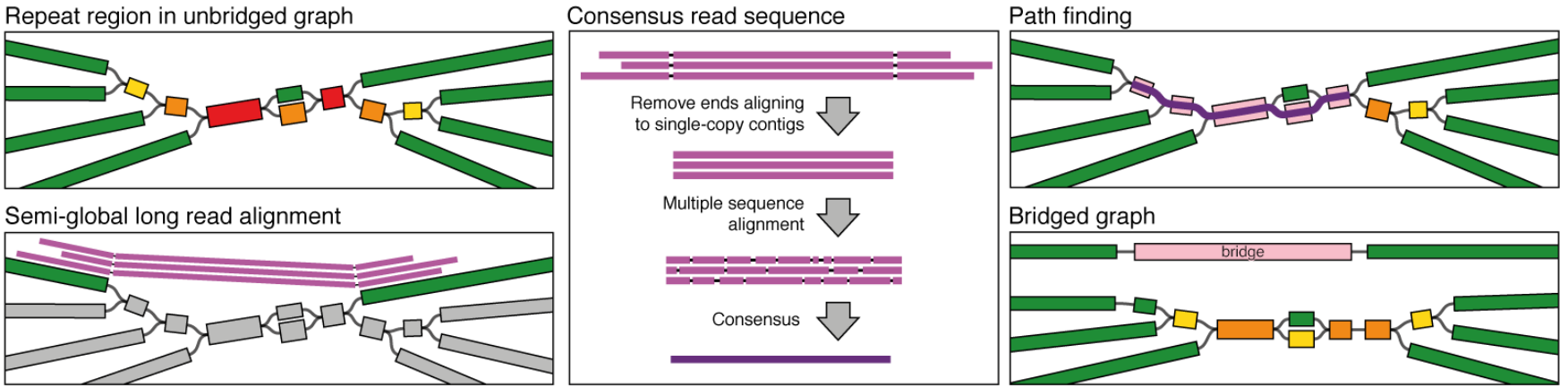
Příklad de Bruijnova grafu u mikrobiálního genomu (Illumina)

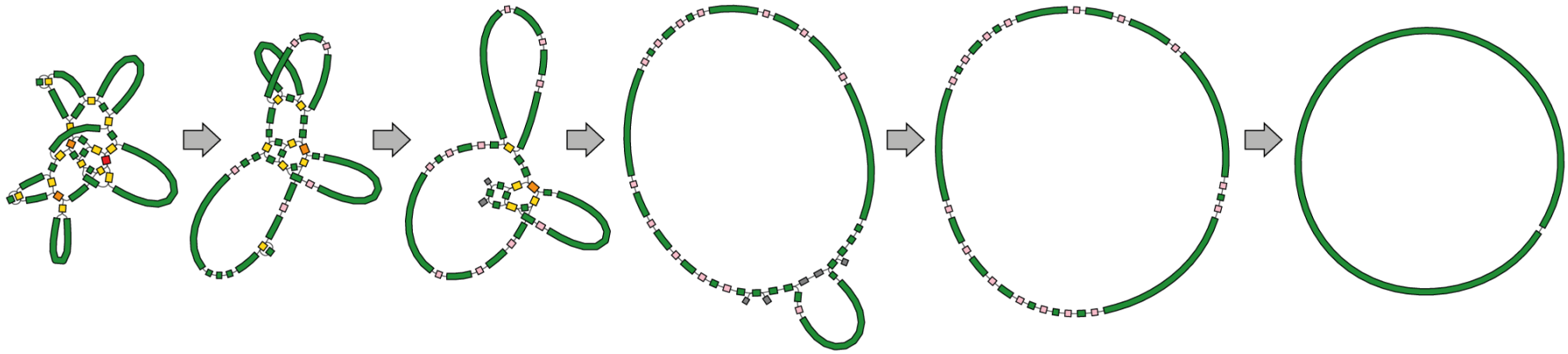


A – kvalitní sestavení
B – sestavení vyžadující optimalizaci, kombinace dlouhých a krátkých kontigů
C – nekvalitní sestavení vycházející z nekvalitních dat, velké množství nezařazených krátkých kontigů

Hybridní assembly a bridging

- Kombinace krátkých čtení (Illumina, IonTorrent) a dlouhých čtení (PacBio, Nanopore) umožňuje hybridní assembly
  - → repetice vyřešeny namapováním dlouhých čtení
  - → chyby vyřešeny namapováním krátkých čtení

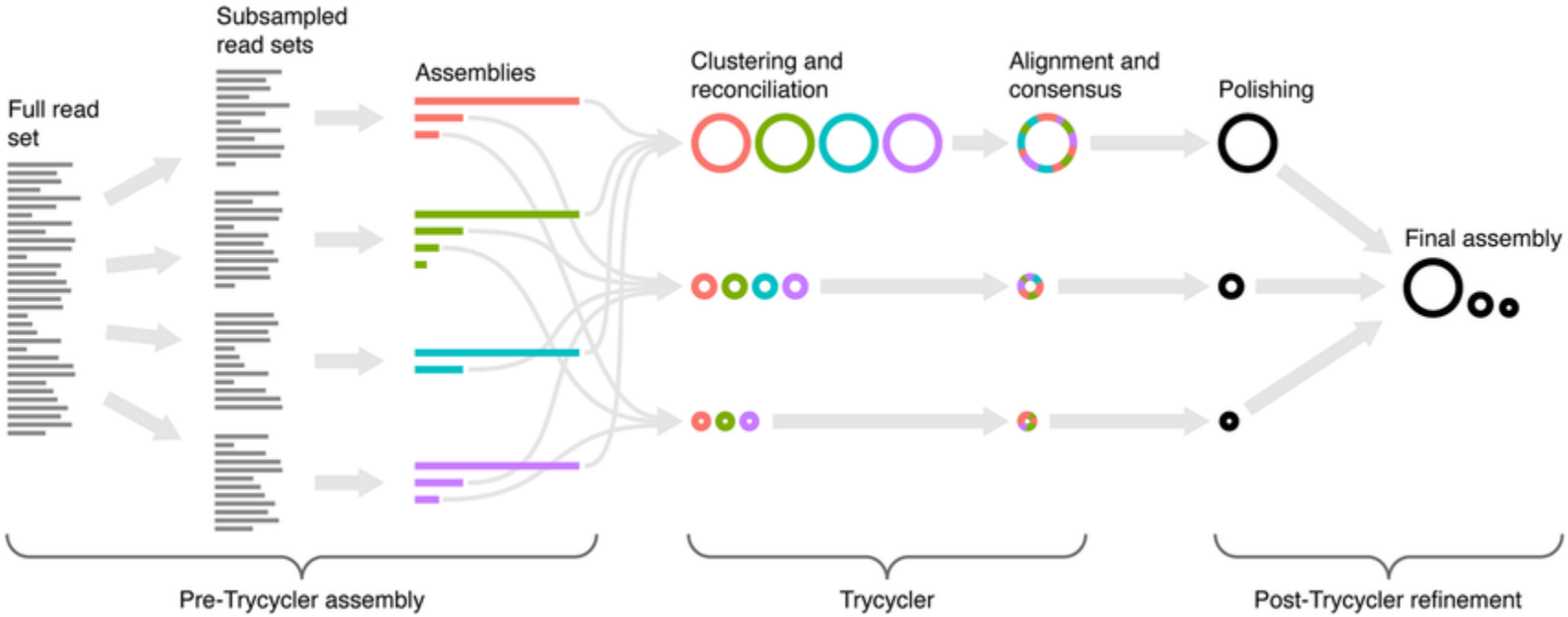




Pro assembly obvykle využíváme připravené pipeline

Mohou být založeny na kombinaci různých assemblerů

- Unicycler: Short reads and hybrid assembly pipeline
- Trycycler: long-read assembly pipeline



Klasifikace assemblerů a využívané strategie

Assembler	Error correction	Contig extension	Complexity reduction
Flye	Direct	Graph-based disjointing correction	Disjointig construction
WTDBG2		Fuzzy Bruijn graph	Hash table based on the *k*-mer block (bin)
		Partial order alignment	Non-redundant *k*-mer removing
Canu	Hierarchical	Best overlap graph	*tf-idf* weighted MHAP
		CABOG	
		PBDAG-CON	
MaSuRCA	Hybrid	de Bruijn	Super-read construction
		CABOG	
WENGAN		de Bruijn	Synthetic scaffolding graph
		Partial order alignment	
SPAdes	Short-read only	Multisized de Bruijn	—
ABySS		de Bruijn	—
		Paired-end–based contig extension	

Statistiky assembly

- N50 a L50 statistiky sady délek kontigů nebo skafoldů, prostřednictvím kterých můžeme srovnat kvalitu assembly
- **N50** definuje kvalitu assembly z hlediska spojitosti. **Délka** sekvence nejkratšího kontigu který přispěl k sestavení 50 % celkové délky.
- **L50** je definován jako nejmenší **počet** kontigů, jejichž součet délek tvoří polovinu velikosti genomu.

Automatická anotace

- Algoritmy- viz lekce Hledání genů
- Servery pro automatickou anotaci
  - **NCBI Prokaryotic Genome Annotation Pipeline** (PGAP) https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/
  - **NCBI Eukaryotic Genome Annotation Pipeline** https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/
  - **RAST** http://rast.nmpdr.org/
- Offline command line aplikace
  - **Prokka**, bakteriální genomy https://github.com/tseemann/prokka
  - **MARK**, eukaryotické genomy https://reslp.github.io/blog/My-MAKER-Pipeline/

NCBI Prokaryotic Genome Annotation Pipeline

NCBI Prokaryotic Genome Annotation Pipeline(PGAP) is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Pipeline was developed in 2001 and is regularly upgraded to improve structural and functional annotation quality (Haft DH et al 2018, Tatusova T et al 2016). Recent improvements utilize curated protein profile hidden Markov models (HMMs), including TIGRFAMS and new HMMs for antimicrobial resistance proteins, and curated complex domain architectures for functional annotation of proteins. NCBI's annotation pipeline depends on several internal databases and is not currently available for download or use outside of the NCBI environment.

Related documentation:

- Annotation process
- Annotation standards
- Assemblies excluded from RefSeq
- Release notes

GenBank

The NCBI prokaryotic annotation pipeline is available as a service for GenBank submitters. The pipeline is capable of annotating both complete genomes and draft WGS genomes consisting of multiple contigs. You can request PGAP annotation when you submit your genome to GenBank.

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are:

non-WGS:

- Each chromosome is in a single sequence and there are no extra sequences
- Each sequence in the genome must be assigned to a chromosome or plasmid or organelle
- Plasmids and organelles can still be in multiple pieces.

WGS:

- One or more chromosomes are in multiple pieces and/or some sequences are not assembled into chromosomes

The NCBI Eukaryotic Genome Annotation Pipeline

The NCBI Eukaryotic Genome Annotation Pipeline provides content for various NCBI resources including Nucleotide, Protein, BLAST, Gene and the Genome Data Viewer genome browser.

This page provides an overview of the annotation process. Please refer to the Eukaryotic Genome Annotation chapter of the NCBI Handbook for algorithmic details.

The pipeline uses a modular framework for the execution of all annotation tasks from the fetching of raw and curated data from public repositories (sequence and Assembly databases) to the alignment of sequences and the prediction of genes, to the submission of the accessioned annotation products to public databases. Core components of the pipeline are alignment programs (Splign and ProSplign) and an HMM-based gene prediction program (Gnomon) developed at NCBI.

Important features of the pipeline include:

- flexibility and speed
- higher weight given to curated evidence than non-curated evidence
- utilization of RNA-Seq for gene prediction
- production of models that compensate for assembly issues
- tracking of gene loci from one annotation to the next
- ability to co-annotate multiple assemblies for the same organism

The products of an annotation run (chromosome, scaffolds and model transcripts and proteins) are labeled with an Annotation Release number. The Annotation Release name is the combination of the organism name and Annotation Release number (e.g. NCBI *Pongo abelii* Annotation Release 103) and is used throughout NCBI as a way to uniquely identify annotation products originating from the same annotation run.

Contents

RAST (Rapid Annotation using Subsystem Technology) Server

http://rast.nmpdr.org/

Upload a Genome

Complete Upload

Please consider the following options for the RAST annotation pipeline:

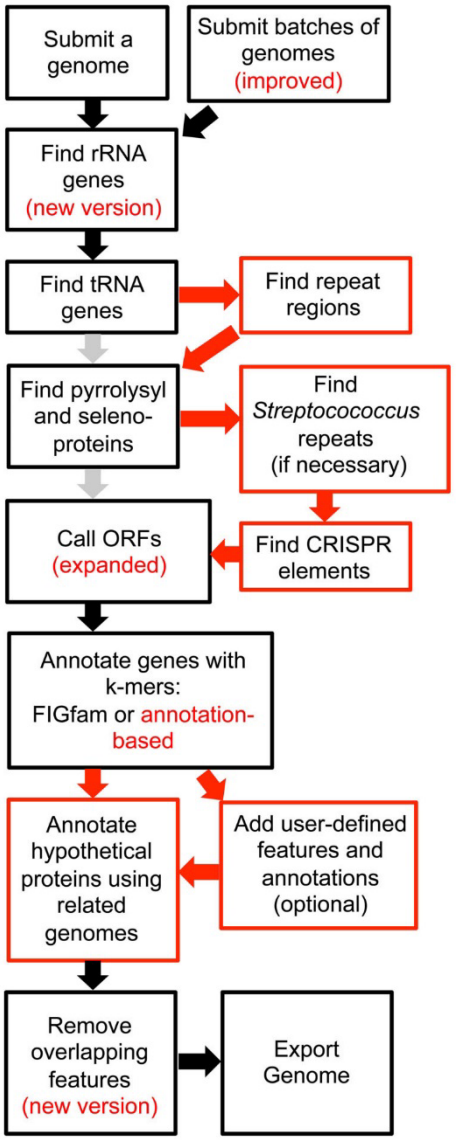
RAST Annotation Settings:

Choose RAST annotation scheme	RASTtk	*Choose "Classic RAST" for the current production RAST, or "RASTtk" for the new modular RAST pipeline currently in testing.*
Customize RASTtk pipeline	☑ Yes	*Customize the RASTtk pipeline*

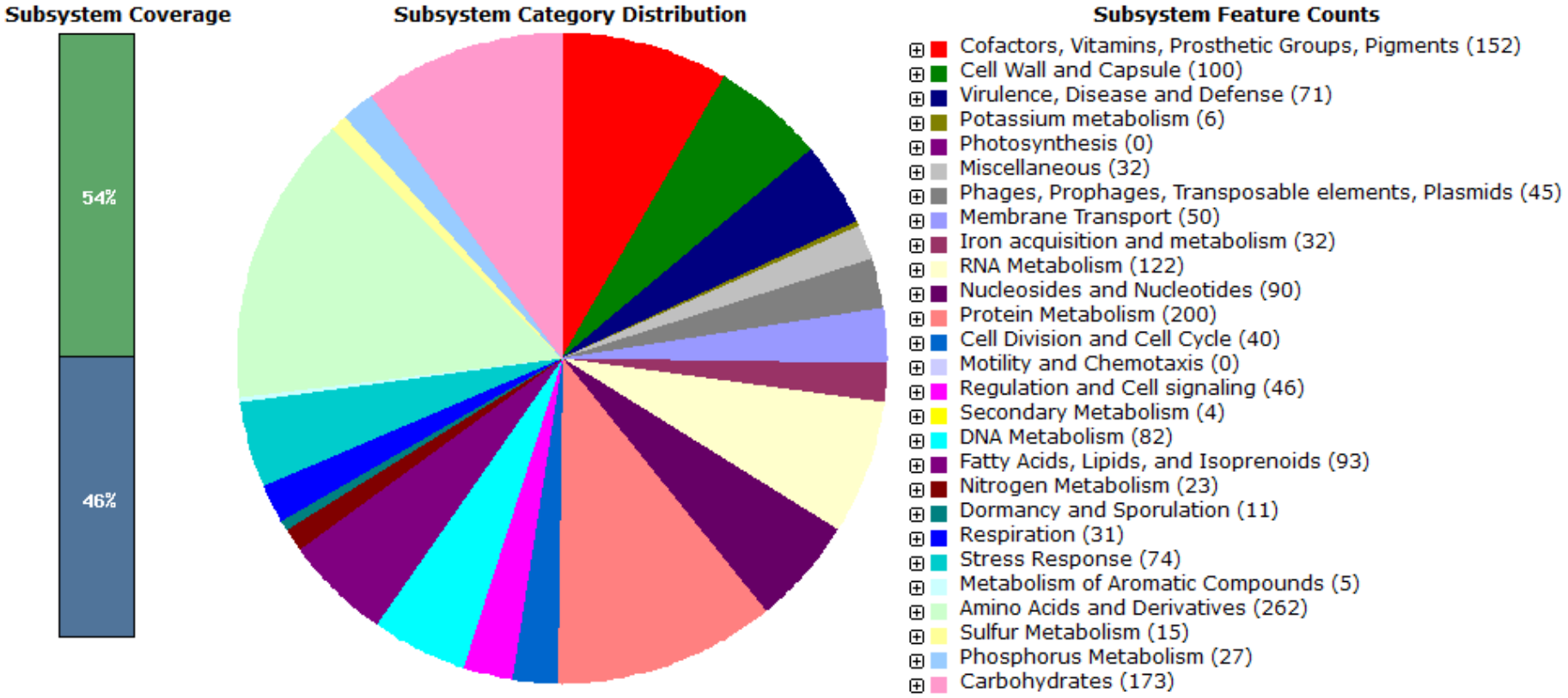
Stage name	Enabled	Parameters	Condition
call-features-rRNA-SEED	☑ Yes		
call-features-tRNA-trnascan	☑ Yes		
call-features-repeat-region-SEED	☑ Yes	Minimum identity 95 Minimum length 100	
call-selenoproteins	☑ Yes		
call-pyrrolysoproteins	☑ Yes		
call-features-insertion-sequences	☐ Yes		
call-features-strep-suis-repeat	☑ Yes		$genome->{scientific_name}
call-features-strep-pneumo-repeat	☑ Yes		$genome->{scientific_name}
call-features-crispr	☑ Yes		
call-features-CDS-glimmer3	☑ Yes	Minimum training length 2000	
call-features-CDS-prodigal	☑ Yes		
call-features-CDS-genemark	☐ Yes		
annotate-proteins-kmer-v2	☑ Yes	Minimum kmer hits required 5 Only annotate hypothetical proteins ☐ Yes	
annotate-proteins-kmer-v1	☑ Yes	Kmer dataset to use ◉ Release70 ○ Release59 Only annotate hypothetical proteins ☑ Yes	
annotate-proteins-similarity	☑ Yes	Only annotate hypothetical proteins ☑ Yes	
resolve-overlapping-features	☑ Yes		
find-close-neighbors	☑ Yes		
call-features-prophage-phispy	☐ Yes		

Automatically fix errors?	☑ Yes	*The automatic annotation process may run into problems, such as gene candidates overlapping RNAs, or genes embedded inside other genes. To automatically resolve these problems (even if that requires deleting some gene candidates), please check this box.*
Fix frameshifts?	☐ Yes	*If you wish for the pipeline to fix frameshifts, check this option. Otherwise frameshifts will not be corrected.*
Build metabolic model?	☐ Yes	*If you wish RAST to build a metabolic model for this genome, check this option.*
Turn on debug?	☐ Yes	*If you wish debug statements to be printed for this job, check this box.*
Set verbose level	0	*Set this to the verbosity level of choice for error messages.*
Disable replication	☐ Yes	*Even if this job is identical to a previous job, run it from scratch.*

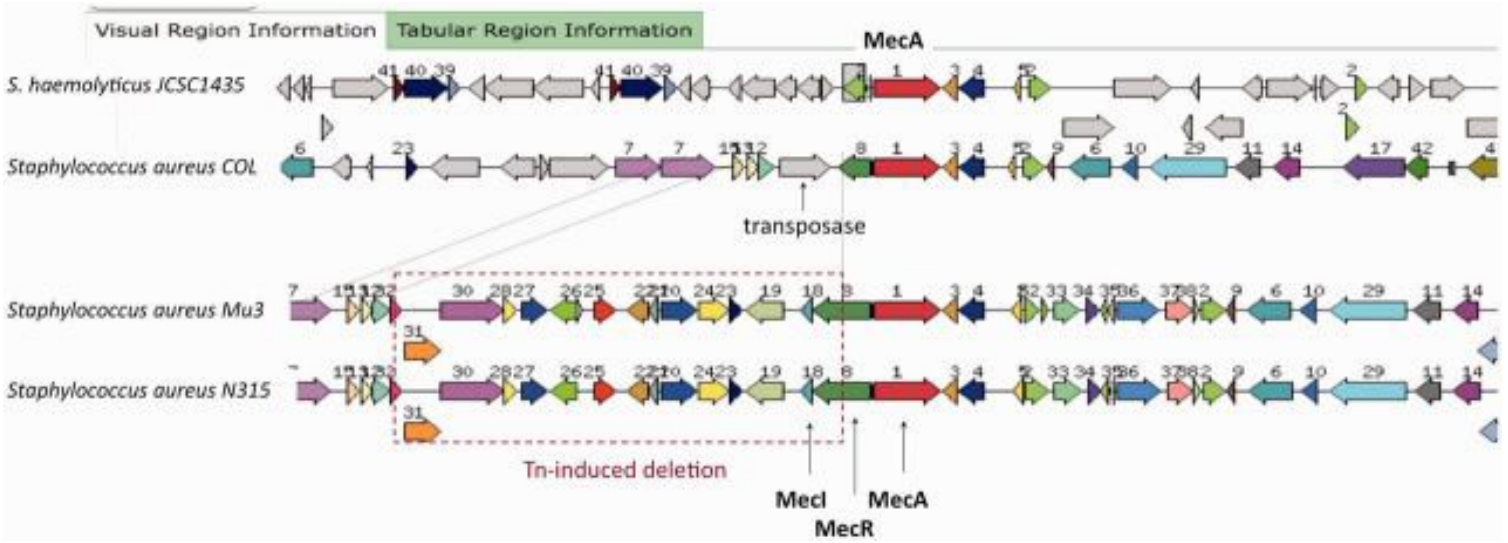
Finish the upload

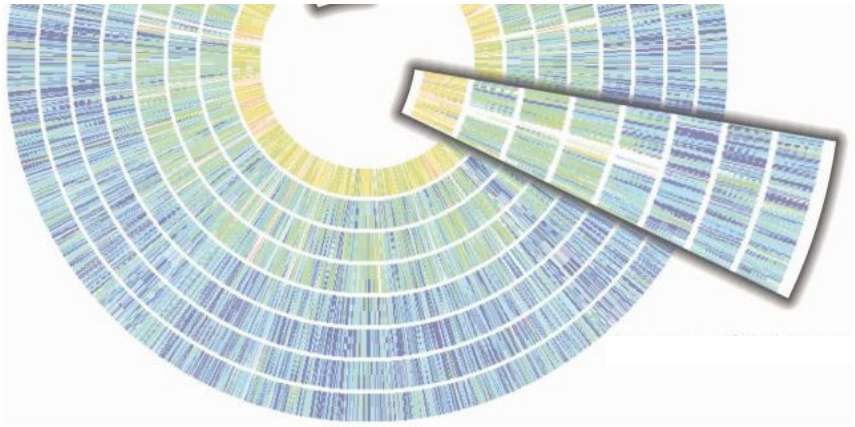


RAST anotace bakteriálho genomu a klasifikace do subsystémů



Vizualizace úseku anotovaného genomu a srovnávací analýzy





Vizualizace lokusů:

- Srovnání anotace genů

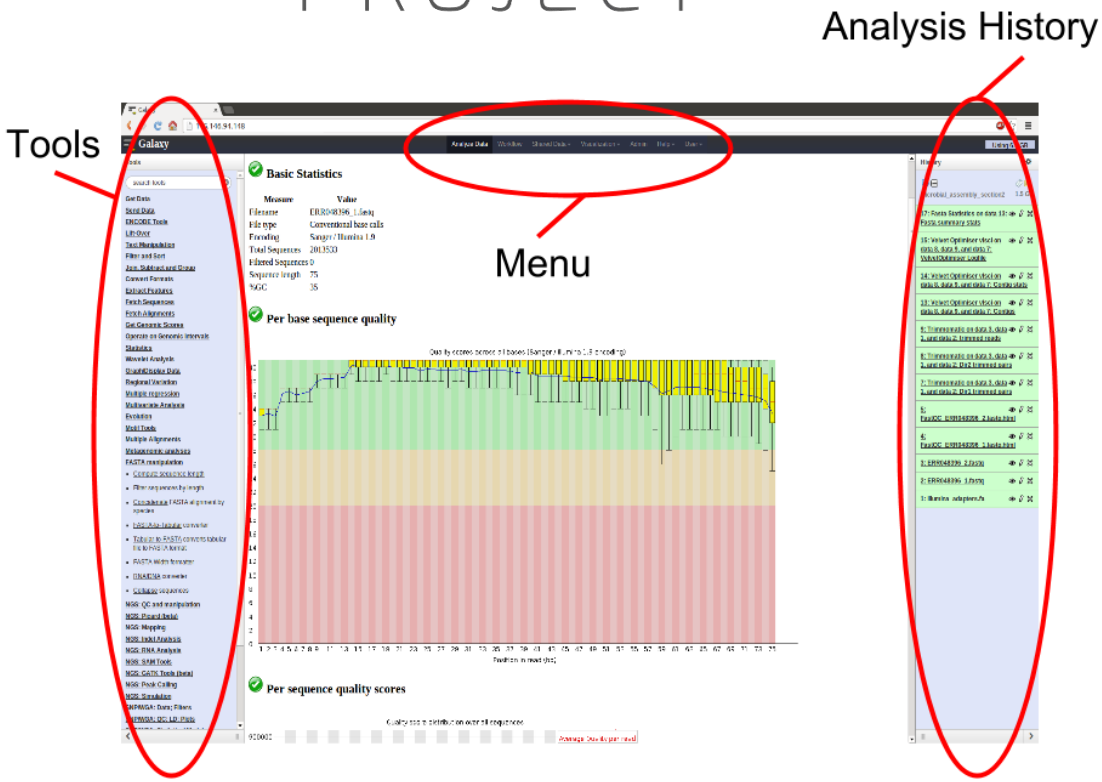
Vizualizace celých genomů:

- Míra konzervovanosti
- Inzerce/ delece

Platformy pro manipulaci sekvencí

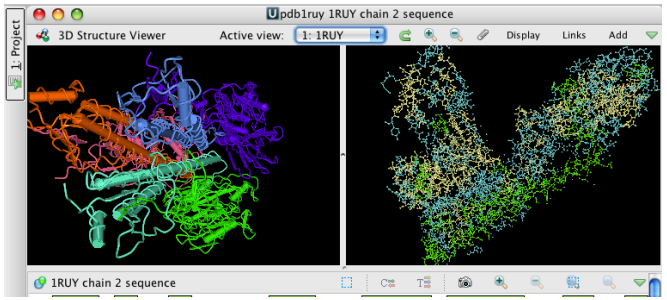
- usnadňují práci uživatelům, kteří nejsou zběhlí v programování a administraci systémů
- Webový server **Galaxy**
  - open source
  - připravené workflows
  - od jednoduchých textových manipulací po analýzu sekvenačních dat, genome assembly, metagenomiku
  - nástroje pro vizualizaci dat
  - https://usegalaxy.org

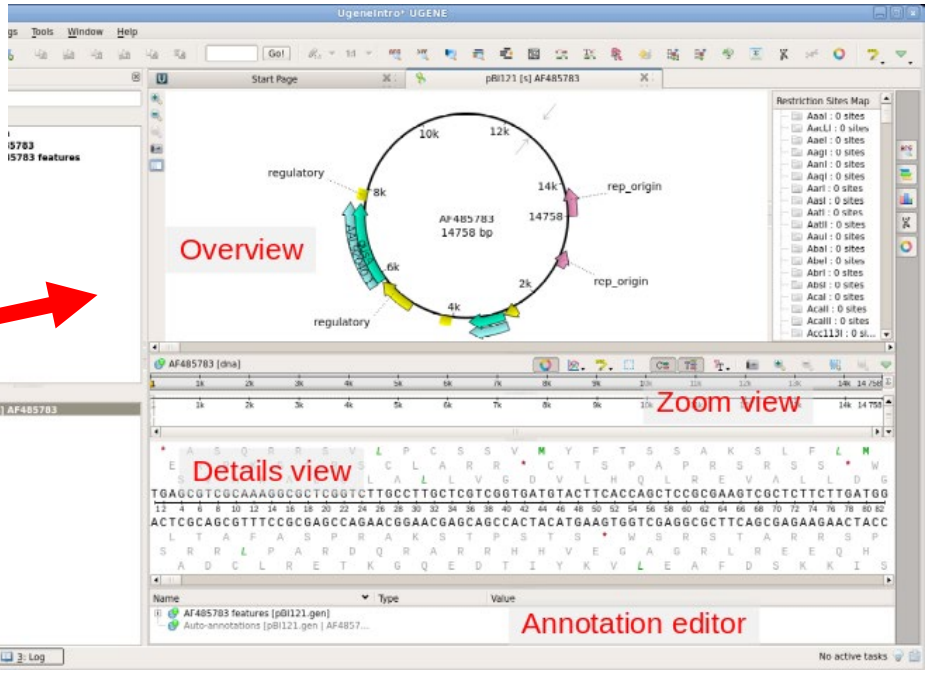
Galaxy PROJECT

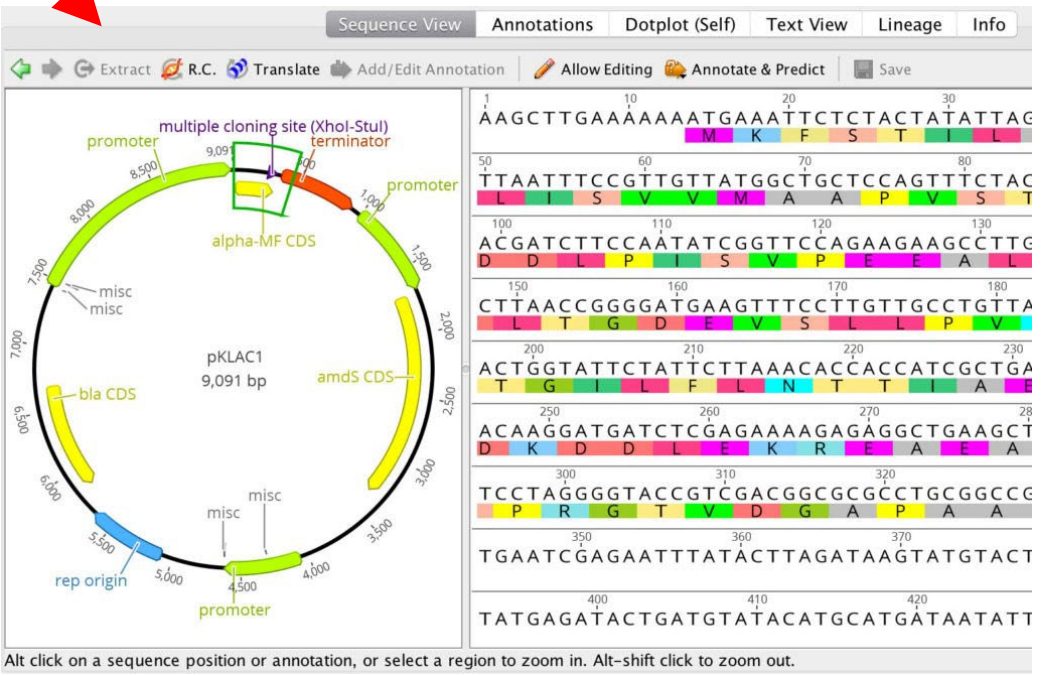


Lokální platformy pro manipulaci sekvencí

- neplacené (např. **Ugene**) nebo licencované (**Geneious**)
- příjemné grafické rozhraní
- implementace nástrojů
  - anotace DNA, proteinů
  - prohledávání databází
  - multiple seqeunce alignment
  - PCR – návrh primerů
  - contig assembly
  - short reads alignment
  - search ORF
  - 3D structure Viewer
  - …







Jak se data dostanou do databází?

Bacteria and Bacteriophage

A Compilation from the Genbank® and EMBL Data Libraries

Book - 1987

```
PHIX174CS:  bacteriophage phix174 gene a protein cleavage site. [ss-DNA]
 EMBL ID: PH174CS                          ACCESSION NUMBERS:  K00813                    DATE:  pre-entry 84-06-01
 REFERENCES:     [1] (bases 1 to 65) Brown,D.R., Schmidt-Glenewinkel,T., Reinberg,D.  and Hurwitz,J.; "dna sequences
           which support activities of the bacteriophage phix174 gene a protein"; J Biol Chem 258, 8402-8412 (1983)
 KEYWORDS:  closing activity; gene A; nicking activity; origin of replication; replication form.
 SOURCE:    bacteriophage phix174 (strain am3) dna, grown in e.coli hf4704.     Bacteriophage phi-X174
 COMMENT:   the gene a protein of phix174 mediates both initiation and termination of viral strand dna synthesis.  it
           nicks and closes within a 30 nucleotide sequence (bases 13-42) which is well conserved among icosohedral
           ss(c) dna phages; this 30 bp region is implicated as the specific target for gene a protein action, and as
           the essential dna sequence required for replication origin function.
 SITES:        key        site span   description
               refnumbr      1    1   sequence not numbered in [1]
               cutss        20    0   gene a protein cleavage site
 ORIGIN:    replication origin of phix174.
 SEQUENCE:     65 bp    23 a    18 c    12 g    12 t
       1 aatgtgctcc cccaacttga tattaataac actatagacc accgccccga  aggggacgaa aaatg
//
PHIX174DE:  bacteriophage phix174 d and e genes.
 EMBL ID: *                                ACCESSION NUMBERS:  J02483                    DATE:  pre-entry 83-03-01
 REFERENCES:     [1] (bases 1 to 521) Barrell,B.G., Air,G.M.  and Hutchison,C.A.III.; "overlapping genes in
           bacteriophage phix174"; Nature 264, 34-41 (1976)
 SOURCE:    phix174.     Bacteriophage phi-X174
 SEQUENCE:    521 bp   125 a   115 c   124 g   157 t
       1 gagtccgatg ctgttcaacc actaataggt aagaaatcat gagtcaagtt  actgaacaat ccgtacgttt ccagaccgct ttggcctcta ttaagctcat
     101 tcaggcttct gccgttttgg atttaaccga agatgatttc gattttctga  cgagtaacaa agtttggatt gctactgacc gctctcgtgc tcgtcgctgc
     201 gttgaggctt gcgtttatgg tacgctggac tttgtgggat accctcgctt  tcctgctcct gttgagttta ttgctgccgt cattgcttat tatgttcatc
     301 ccgtcaacat tcaaacggcc tgtctcatca tggaaggcgc tgaatttacg  gaaaacatta ttaatggcgt cgagcgtccg gttaaagccg ctgaattgtt
     401 cgcgtttacc ttgcgtgtac gcgcaggaaa cactgacgtt cttactgacg  cagaagaaaa cgtgcgtcaa aaattacgtg cggaaggagt gatgtaatgt
     501 ctaaaggtaa aaaacgttct g
//
```

Jak se data dostanou do databází?

- **Submission - Vložení do databáze**
- webový portál
  - **BankIt** (GenBank)
    https://submit.ncbi.nlm.nih.gov/about/bankit/
  - **Submission Portal** (GenBank)
    https://submit.ncbi.nlm.nih.gov/subs/genome/
  - **WebIn** (EMBL/European Nucleotide Archive)
    http://www.ebi.ac.uk/ena/submit
  - **Sakura** (DDBJ)
    http://www.ddbj.nig.ac.jp/sub/websub-e.html
- samostatná aplikace pro PC
  - **Sequin**, delší manuálně anotované sekvence, fylogenetické, populační nebo mutační studie obsahující sekvenční přiložení
    http://www.ncbi.nlm.nih.gov/Sequin/download/seq_download.html
  - **Tbl2asn,** command line program, celé genomy EST, STS a zasílaní velkých dávek sekvencí, automatizuje vytvoření záznamu sekvence
- Minimální požadavky pro vložení

GenBank formát

- Záznam anotovaného genomu
- Skládá se ze tří sekcí
  - **Header**
    Informace o vlastnostech sekvence a jejím zdroji
  - **Feature Table**
    Anotacemi formou deskriptorů; u genů může obsahovat i jejich translaci do proteinu
  - **Sequence**
    Vlastní nukleotidová sekvence

- Historické omezení na 60 znaků na řádek
  - pole v hlavičce tak mají maximální délku
  - v současnosti už neplatí striktně

Tradiční záznam GenBank

```
LOCUS       AY182241               1931 bp    mRNA    linear   PLN 04-MAY-2004
DEFINITION  Malus x domestica (E,E)-alpha-farnesene synthase (AFS1) mRNA,
            complete cds.
ACCESSION   AY182241
VERSION     AY182241.2  GI:32265057
KEYWORDS    .
SOURCE      Malus x domestica (cultivated apple)
  ORGANISM  Malus x domestica
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; eurosids I; Rosales; Rosaceae; Maloideae; Malus.
REFERENCE   1  (bases 1 to 1931)
  AUTHORS   Pechous,S.W. and Whitaker,B.D.
  TITLE     Cloning and functional expression of an (E,E)-alpha-farnesene
            synthase cDNA from peel tissue of apple fruit
  JOURNAL   Planta 219, 84-94 (2004)
REFERENCE   2  (bases 1 to 1931)
  AUTHORS   Pechous,S.W. and Whitaker,B.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-NOV-2002) PSI-Produce Quality and Safety Lab,
            USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
            20705, USA
REFERENCE   3  (bases 1 to 1931)
  AUTHORS   Pechous,S.W. and Whitaker,B.D.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-JUN-2003) PSI-Produce Quality and Safety Lab,
            USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
            20705, USA
  REMARK    Sequence update by submitter
COMMENT     On Jun 26, 2003 this sequence version replaced gi:27804758.
FEATURES             Location/Qualifiers
     source          1..1931
                     /organism="Malus x domestica"
                     /mol_type="mRNA"
                     /cultivar="'Law Rome'"
                     /db_xref="taxon:3750"
                     /tissue_type="peel"
     gene            1..1931
                     /gene="AFS1"
     CDS             54..1784
                     /gene="AFS1"
                     /note="terpene synthase"
                     /codon_start=1
                     /product="(E,E)-alpha-farnesene synthase"
                     /protein_id="AAO22848.2"
                     /db_xref="GI:32265058"
                     /translation="MEFRVHLQADNEQKIFQNQMKPEPEASYLINQRRSANYKPNIWK
                     NDFLDQSLISKYDGDEYRKLSEKLIEEVKIYISAETMDLVAKLELIDSVRKLGLANLF
                     EKEIKEALDSIAAIESDNLGTRDDLYGTALHFKILRQHGYKVSQDIFGRFMDEKGTLE
                     DFLHKNEDLLYNISLIVRLNNDLGTSAAEQERGDSPSSIVCYMREVNASEETARKNIK
                     GMIDNAWKKVNGKCFTTNQVPFLSSFMNNATNMARVAHSLYKDGDGFGDQEKGPRTHI
                     LSLLFQPLVN"
ORIGIN
        1 ttcttgtatc ccaaacatct cgagcttctt gtacaccaaa ttaggtattc actatggaat
       61 tcagagttca cttgcaagct gataatgagc agaaaatttt tcaaaaccag atgaaacccg
      121 aacctgaagc ctcttacttg attaatcaaa gacggtctgc aaattacaag ccaaatattt
      181 ggaagaacga tttcctagat caatctctta tcagcaaata cgatggagat gagtatcgga
      241 agctgtctga gaagttaata gaagaagtta agatttatat atctgctgaa acaatggatt
//
```

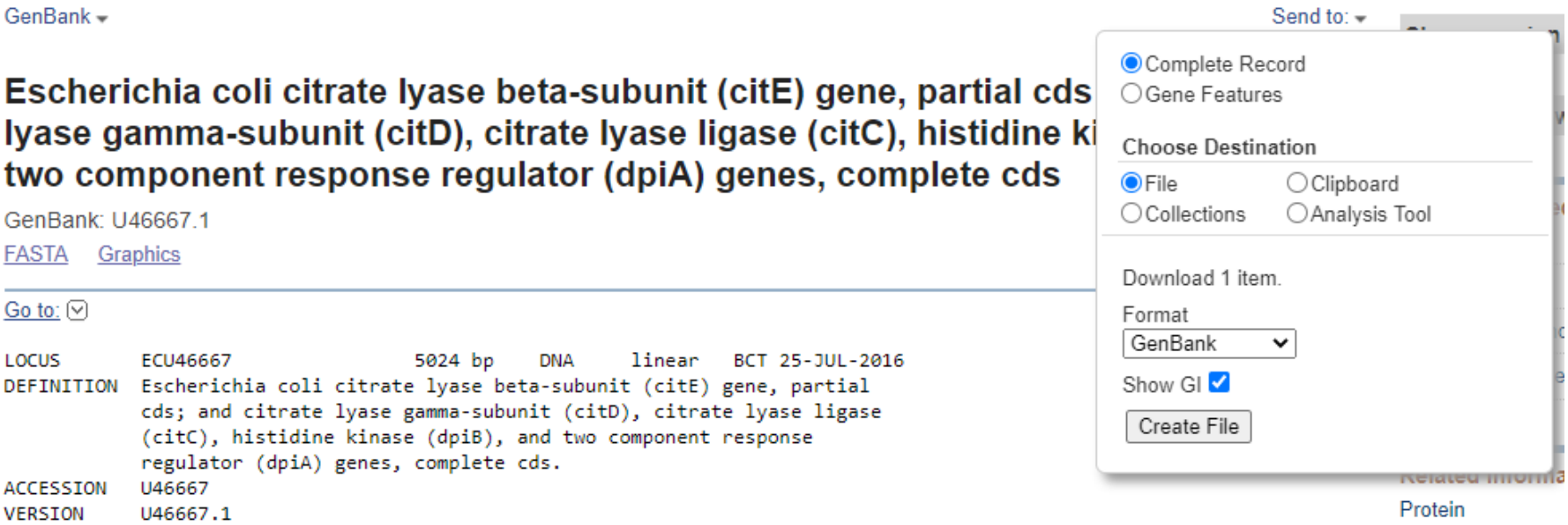
Header

Feature Table

Sequence

Identifikace záznamu v primárních sekvenčních databázích

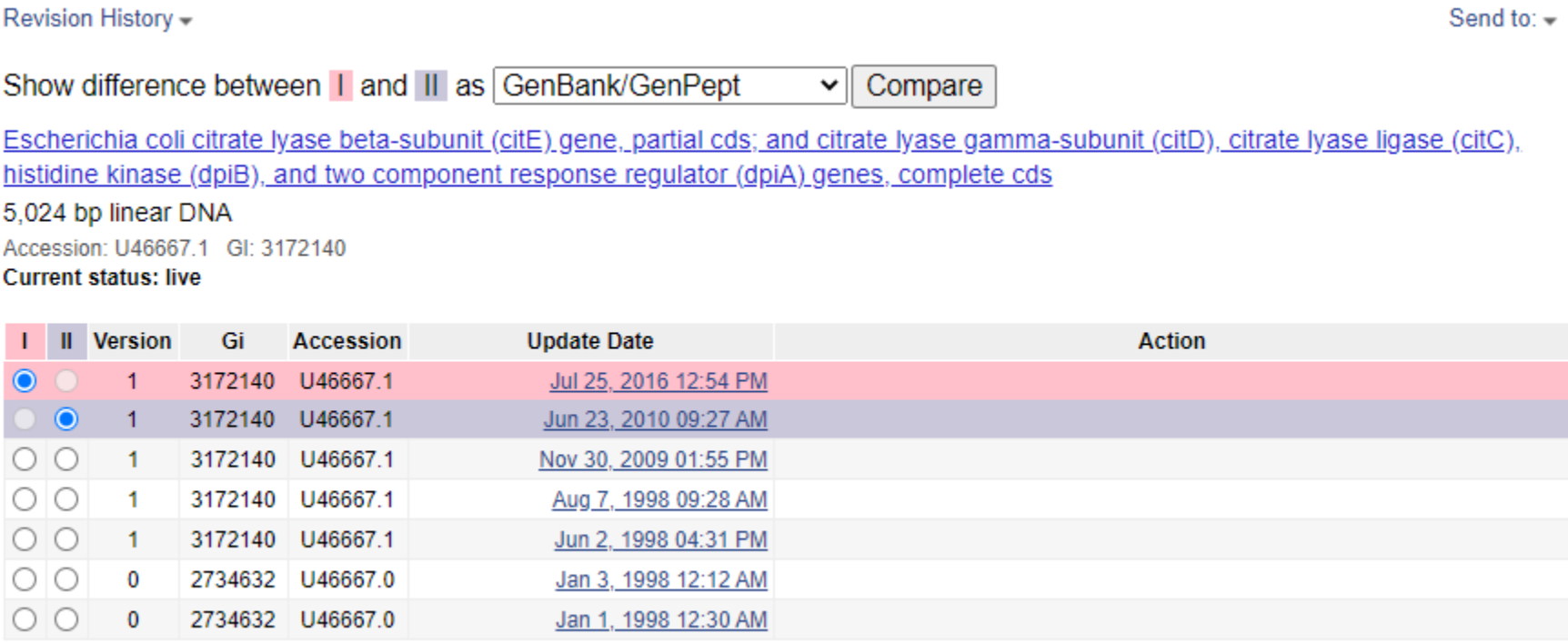
- **Databáze**
  - GenBank
  - EMBL-Bank (European Nucleotide Archive, ENA)
  - DDBJ
- **Identifikátory sekvence**



Historie verzí

- Sequence Revision History tool
  - Struktura zápisu:

    http://www.ncbi.nlm.nih.gov/nuccore/**U46667?report=girevhist**



Revision History

Send to:

Show difference between I and II as GenBank/GenPept Compare

Escherichia coli citrate lyase beta-subunit (citE) gene, partial cds; and citrate lyase gamma-subunit (citD), citrate lyase ligase (citC), histidine kinase (dpiB), and two component response regulator (dpiA) genes, complete cds

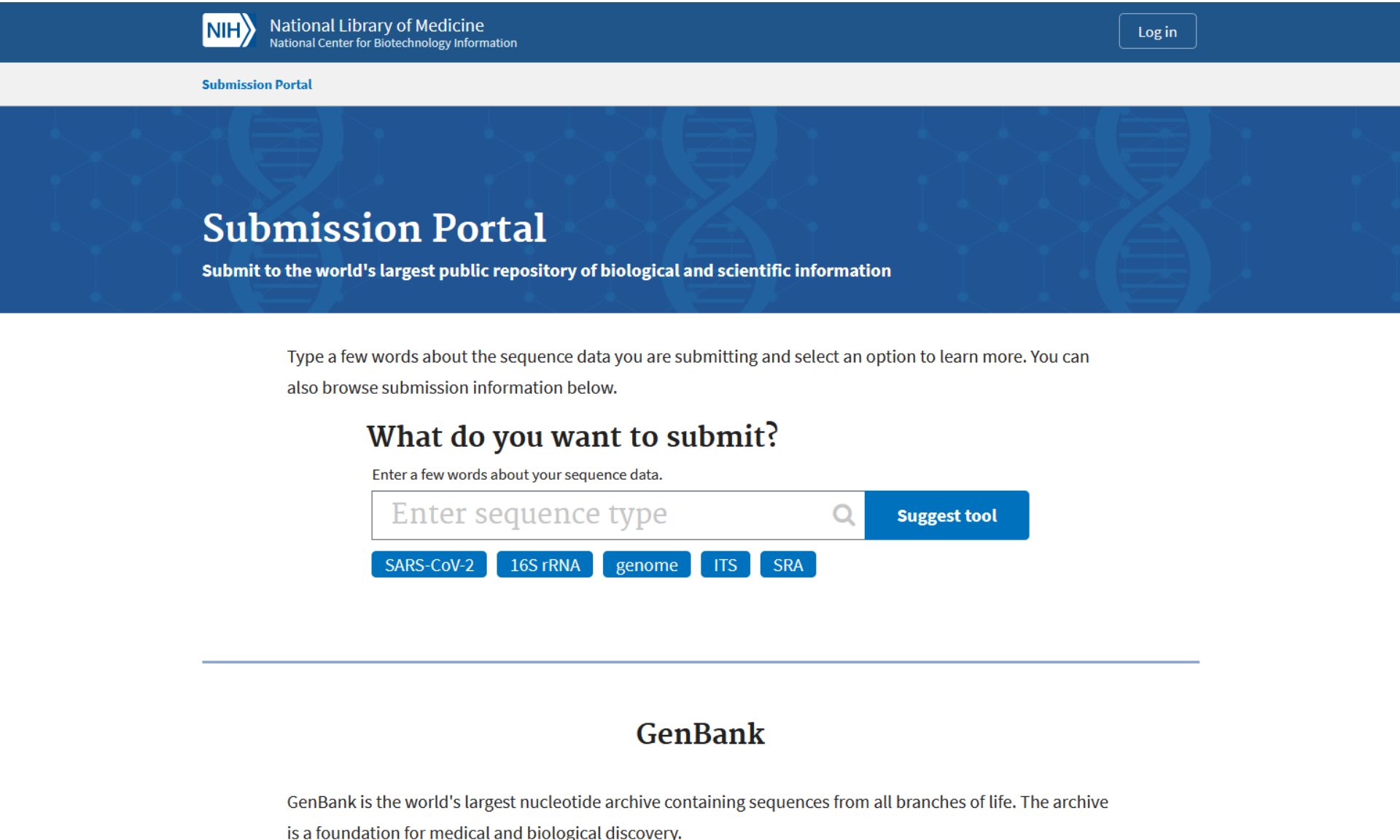
5,024 bp linear DNA

Accession: U46667.1 GI: 3172140

Current status: live

I	II	Version	Gi	Accession	Update Date	Action
◉	○	1	3172140	U46667.1	Jul 25, 2016 12:54 PM	
○	◉	1	3172140	U46667.1	Jun 23, 2010 09:27 AM	
○	○	1	3172140	U46667.1	Nov 30, 2009 01:55 PM	
○	○	1	3172140	U46667.1	Aug 7, 1998 09:28 AM	
○	○	1	3172140	U46667.1	Jun 2, 1998 04:31 PM	
○	○	0	2734632	U46667.0	Jan 3, 1998 12:12 AM	
○	○	0	2734632	U46667.0	Jan 1, 1998 12:30 AM	

https://submit.ncbi.nlm.nih.gov/



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

https://ddbj.nig.ac.jp/submission

DDBJ submission portal

Nucleotide

Submission of small-scale nucleotide sequence data with annotation. In case of project data, please use BioProject, MSS, and DRA.

Create new submission

BioProject

You must obtain BioProject ID and/or locus_tag prefix, before the submission of project data, such as WGS, complete genome, transcriptome project data, DRA, and DTA. The BioProject database collects information about a higher order organization of research projects and its corresponding data. Using BioProject ID make it possible to obtain the same project from various nucleotide sequence databases.

Mass Submission System (MSS)

Please use mass submission system for the submission of following data. WGS, WGS scaffold(s), complete bacterial/eukaryotic genome, HTG, CON, GSS, EST, TSA, and other data includes huge number of sequences.

DDBJ Sequence Read Archive (DRA)

For repository of output data generated by next-generation sequencing machines including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® System, and others.

Podklady pro GenBank

Povinné položky

- Autoři sekvence
- Kontakt na autory
- Publikace (PubMed)
- Použitá sekvenační technologie
- Použitá metoda assembly
- FASTA nukleotidová sekvence
- Název organismu, taxonomické zařazení
- Metadata

  -izolát, kmen, datum odběru, země původu
- Anotace sekvencí

  - buď vlastní nebo automatická

Protokoly pro zaslání do nukleotidové databáze

- Standard
- Whole Genome Shotgun (WGS)
- Complete Microbial or Eukaryotic Genomes
- ESTs (expressed sequence tags) a GSSs (genome survey sequences)
- High-Throughput Genomic Sequences (HTGs)
- Transcriptome Shotgun Assembly (TSA)
- Third Party Annotation (TPA)
  - záznamy, které upřesňují existující sekvence uložené do databází jinými autory
  - striktní požadavek na přímý experimentální důkaz

Typy **standardních** anotovaných sekvencí (nucleotide sequence database)

- prokaryotické geny a části genomu
- eukaryotické geny a části genomu
- mRNA sekvence
- rRNA a nebo ITS
- nekódující RNA
- virové sekvence
- transpozony a inzerční sekvence
- mikrosatelity
- pseudogeny
- klonovací vektroy
- fylogenetické nebo populační studie (alignmenty)

Sekvence, které nejsou akceptovány v primárních databázích

- sekvence bez fyzického (biologického) protějšku – např. konsenzní sekvence
- genomové sekvence více exonů bez údajů o sekvencích intronů
- sekvence <200 bp (vyjma patentových)
- sekvence primerů (mohou být zaslány do NCBI's Probe database)
- pouze sekvence proteinů (mohou být zaslány do UniProt/SwissProt)
- sekvence složené z genomové sekvence a mRNA reprezentované jako jedna sekvence

Nezpracovaná zdrojová data z genomových projektů

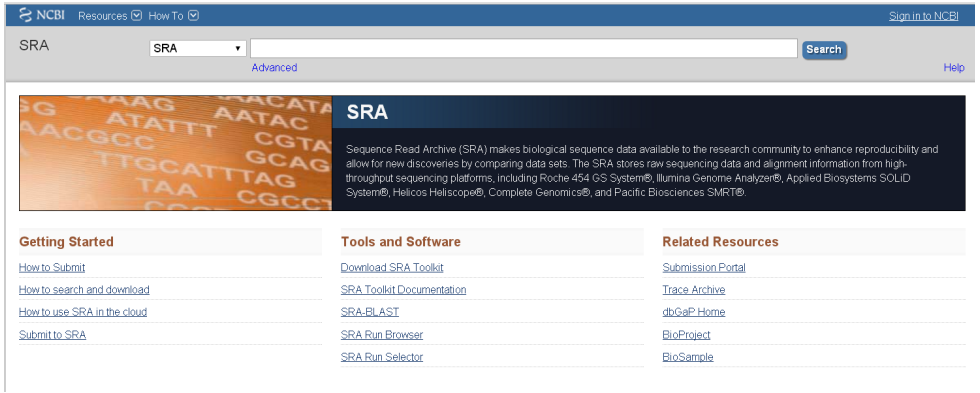
- BioSample & BioProject mohou obsahovat různé typy archivů
  - Trace Archive
    - sekvence získaní Sangerovou technikou sekvenování
    - struktura složek se *.scf nebo *.abi soubory

```
TOP_DIRECTORY/
TOP_DIRECTORY/TRACEINFO.txt
TOP_DIRECTORY/MD5
TOP_DIRECTORY/README
TOP_DIRECTORY/traces
TOP_DIRECTORY/traces/HBBA/
TOP_DIRECTORY/traces/HBBA/HBBAA1U0001.scf
TOP_DIRECTORY/traces/HBBA/HBBAA1U0002.scf
TOP_DIRECTORY/traces/HBBA/HBBAA1U0003.scf
```

  - Sequence Read Archive (SRA)
    - archiv obsahující alignment sekvencí získaných při 454, IonTorrent, Illumina, SOLiD, Helicos, PacBio nebo Complete Genomics
  - The database of Genotypes and Phenotypes (dbGaP)
    - interakce genotypu a fenotypu člověka

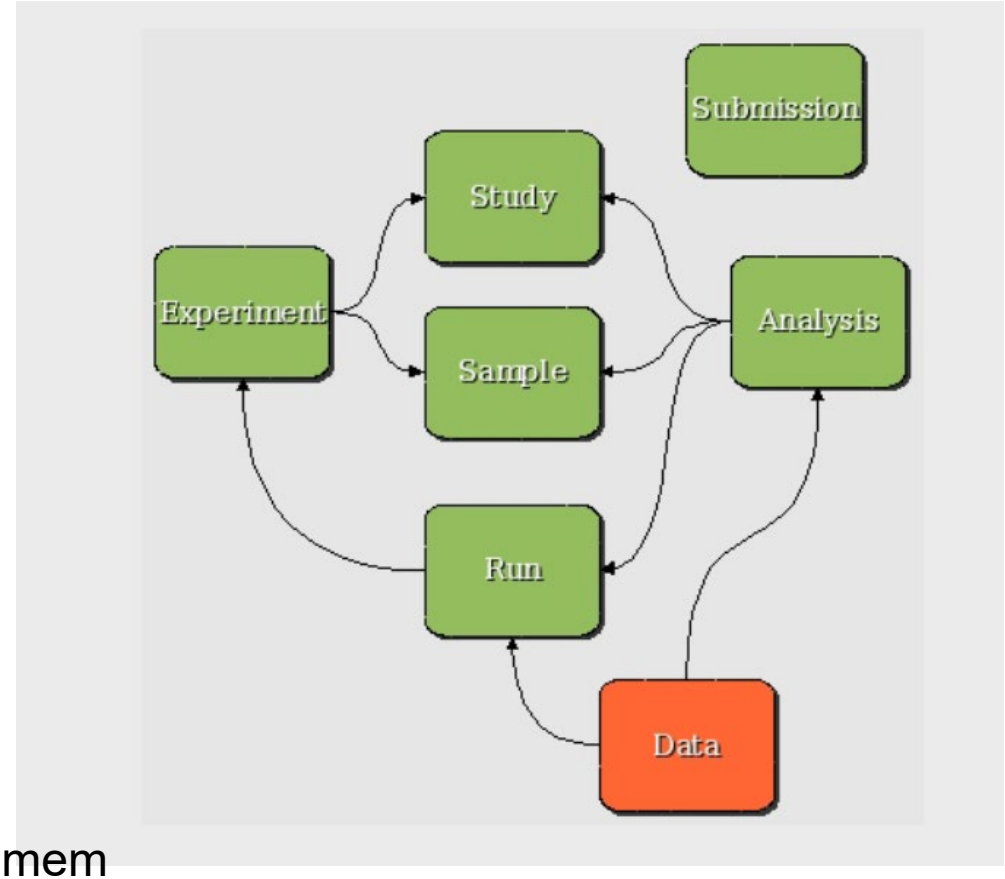
Sequence Read Archive (SRA)
Formát dat a minimální požadavky

- Submission portal
  https://submit.ncbi.nlm.nih.gov/subs/sra/
- Volitelné nahrání primárních sekvenačních dat
- SRA toolkit
  https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
- Minimální požadavek je: primární sekvence (báze) a kvalita = **FASTQ**
- Doporučený formát dat je **BAM** (aligned)
- Další akceptovatelné formáty dat z různých platforem jsou
  - SRF
  - General Fastq
  - SOLiD Fastq
  - Illumina Fastq
  - 454 SFF
  - Ion Torrent SFF
  - PacBio, Nanopore HDF5
  - CompleteGenomics Data Package



Metadata v SRA

- Datové soubory jsou zasílány s metadaty
  - Studie
  - Experiment
  - Vzorek
  - Běh
  - Analýza
  - eticky citlivá data (EGA)



Příklad SRA s mikrobiálním genomem
https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR9600155

Whole Genome Shotgun (WGS)

- WGS sekvenační projekty jsou celé genomy nebo chromozomy sekvenované strategií celogenomového shotgun sekvenování
- Části WGS projektu jsou kontigy, které nesmí obsahovat mezery
- WGS projekty mohou být anotovány, může být zvolena automatická anotace s NCBI pipeline
- Volitelně - soubor AGP ukazuje, jak jsou kontigy oddělené mezerami uspořádány na chromozomu
- Zasílají se přes Genome submissoin portal

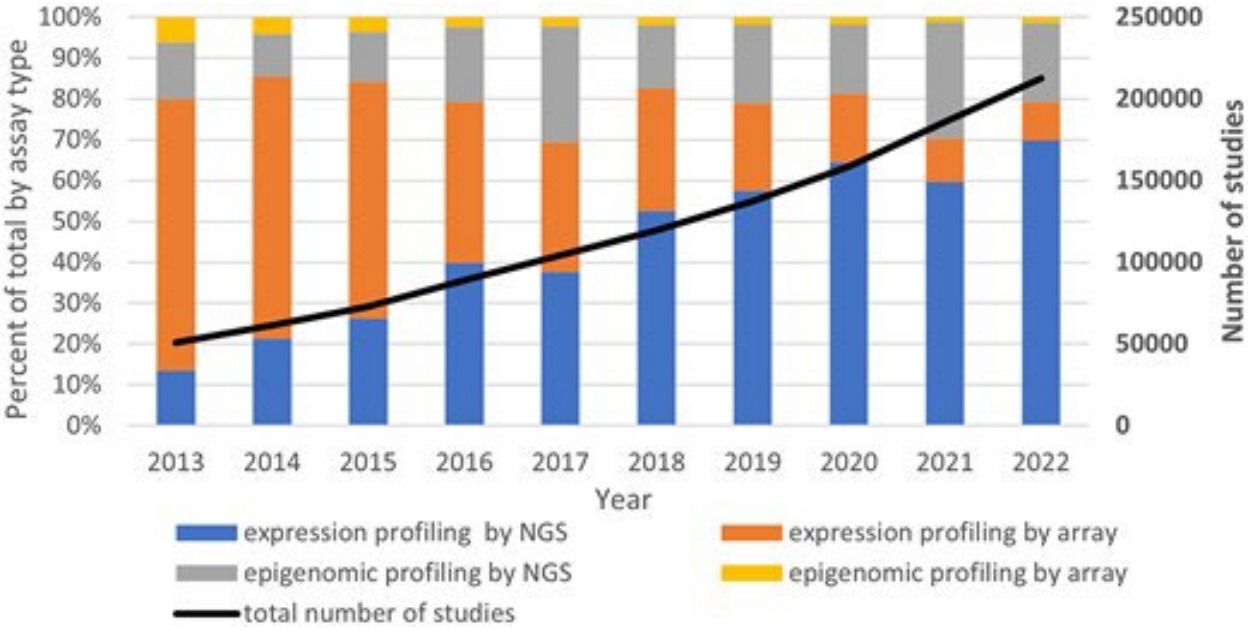
Metagenomy

- Sekvenační projekty analýzy společenstev z určitých ekologických zdrojů nezávislé na kultivaci slouží pro studium
  - genetické diversity, struktury populací, ekologické úlohy
  - metabolických funkcí
  - stanovení kompletních genomů nekultivovatelných organismů
  - izolaci nových mikroorganismů (genetických zdrojů) z prostředí
- **Sekvence jsou vzájemně propojené v rámci BioProject**
- **Bývá vyžadováno nahrání do SRA**
- Metagenomové projekty mohou představovat
  - Neanotované sekvence s převahou informačních sekvencí rRNA
  - Celkové metagenomové projekty sestavené do kontigů
    - obsahují částečné genomy z taxonomicky různých skupin
    - MG-RAST - anotace metagenomu na serveru
- Mothur a QIIME2 obsahují nástroje pro přípravu dat

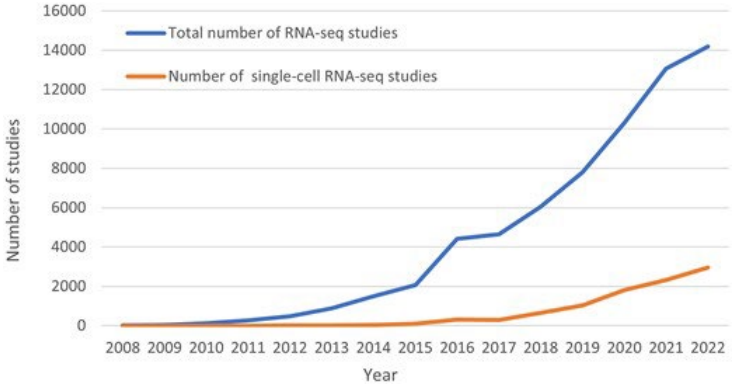
NCBI Gene Expression Omnibus (GEO)

- mezinárodní veřejné úložiště v NCBI
  - https://www.ncbi.nlm.nih.gov/geo
- archivuje zpracované datové soubory a metadata o
  - genové expresi
  - epigenomice
- generované technologiemi
  - sekvenování nové generace
  - Microarray
- nabízí webové nástroje, pro analýzu a vizualizaci diferenciální genové exprese

NCBI Gene Expression Omnibus (GEO)

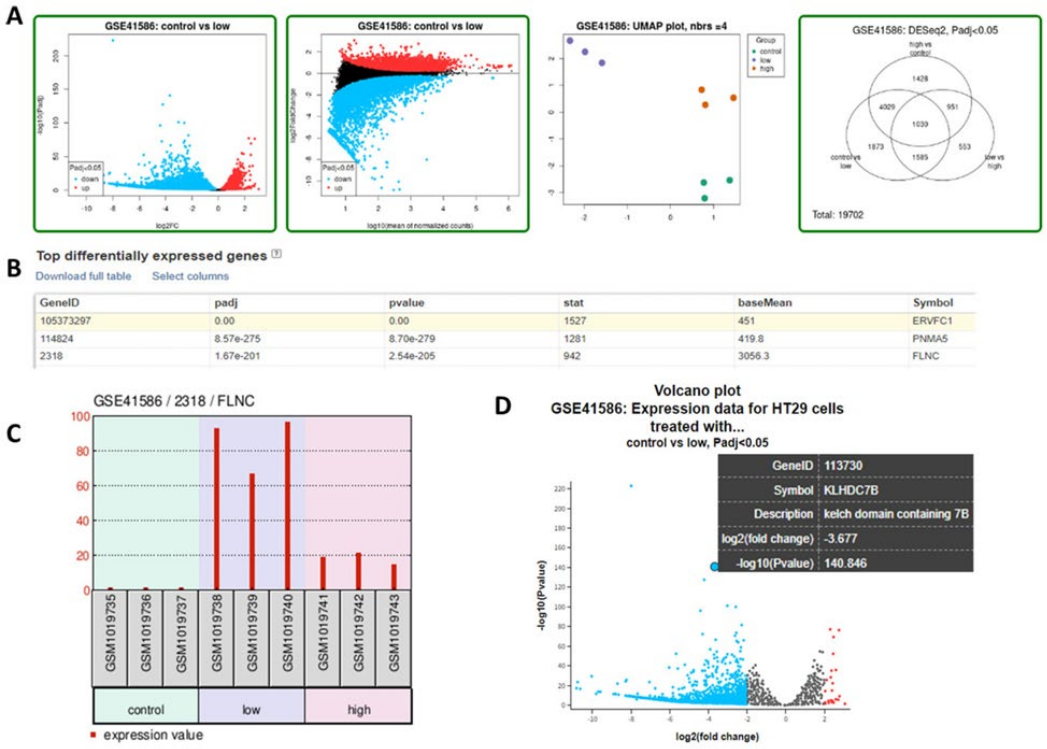


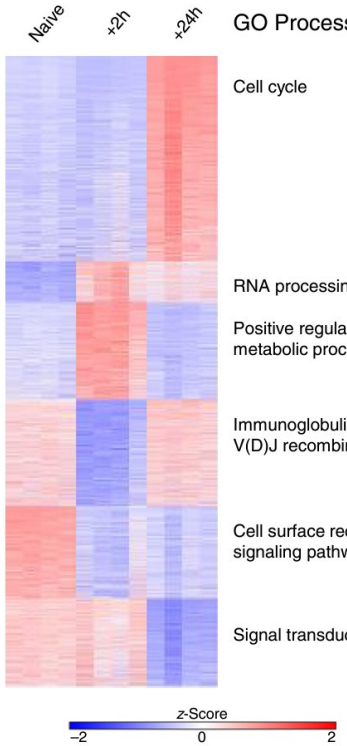
Růst a trendy datových typů za poslední desetiletí v databázi GEO



Z toho počet studií RNA-seq a jednobuněčné RNA-seq v GEO

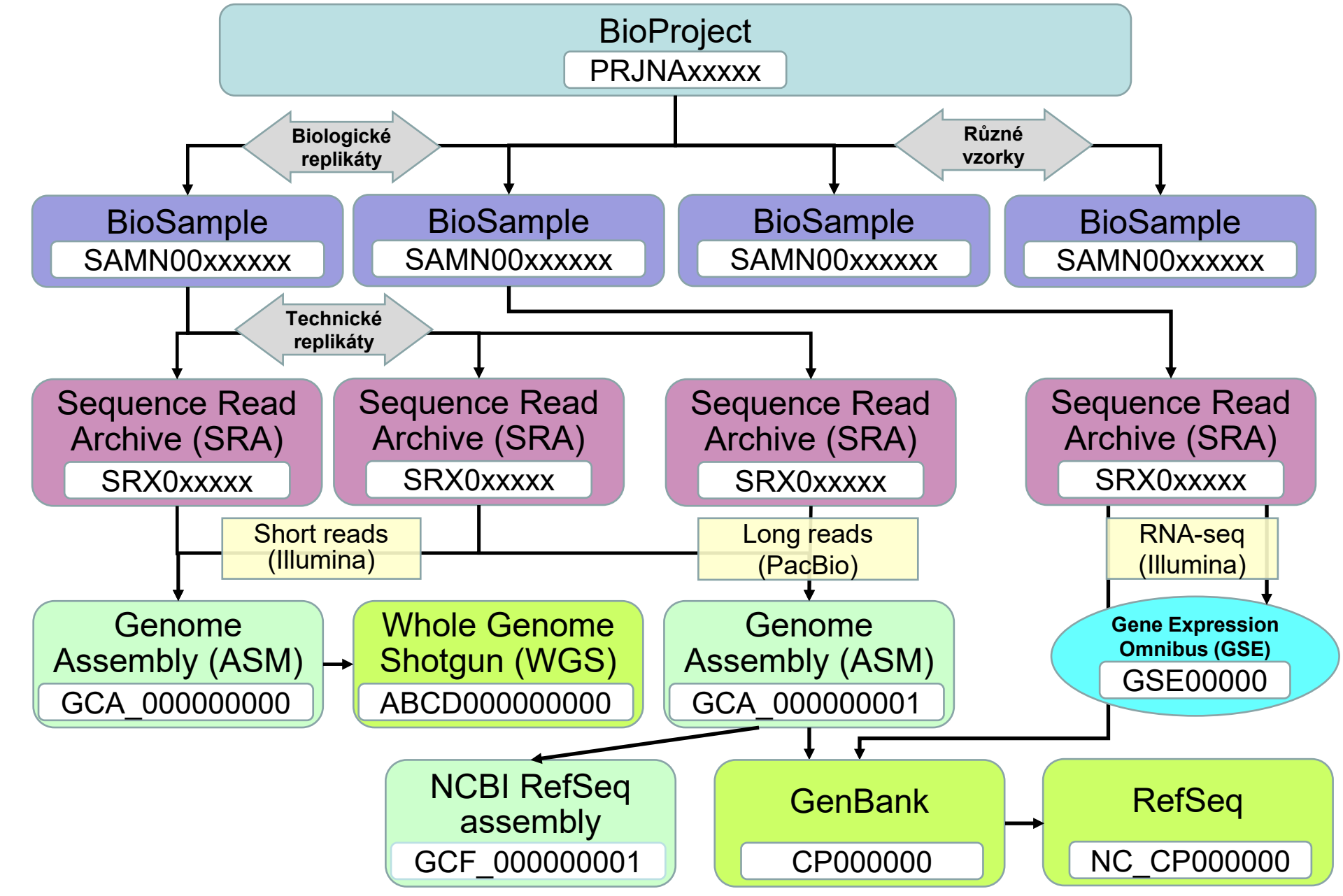
NCBI Gene Expression Omnibus (GEO)





Příklady vizualizace dat z databáze GEO nástrojem GEO2R. Poskytuje interaktivní grafy (Umístěním kurzoru na bod se zobrazí jeho GeneID, Symbol, Popis, $\log_2$ (násobná změna) a $-\log_{10}$ (P -hodnota), tabulky nejlepších odlišně exprimovaných genů se statistikou („Top“ diferenciálně exprimované geny)

Struktura databází (BioSample & BioProject)



Genome submissoin portal

https://www.ncbi.nlm.nih.gov/genbank/genomesubmit/

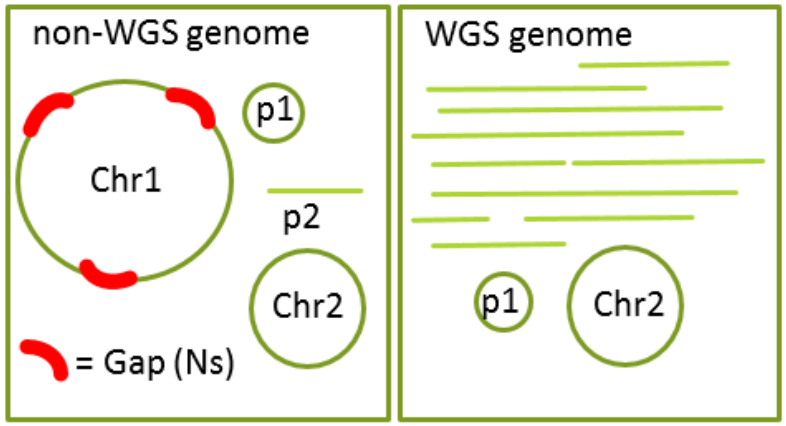
NCBI Resources How To pantucek My NCBI Sign O

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

Prokaryotic and Eukaryotic Genomes Submission Guide

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are:



non-WGS

- Each chromosome is in a single sequence and there are no extra sequences
- Each sequence in the genome must be assigned to a chromosome or plasmid or organelle
- Plasmids and organelles can still be in multiple pieces.

WGS

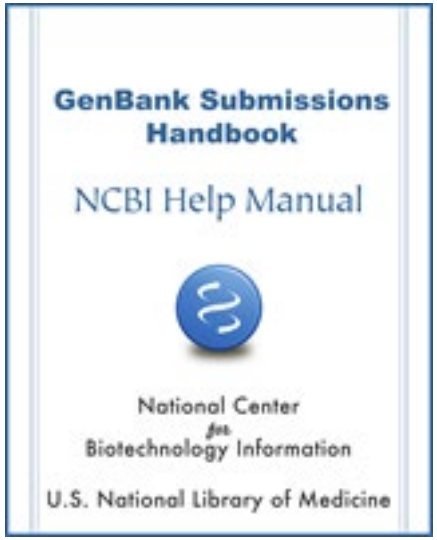
Genome Resources

About WGS
WGS Browser
Genome Submission Guide
Genome Submission Portal
Update Genome Records
FAQ
tbl2asn
Create Submission Template
Eukaryotic Annotation Guide
Prokaryotic Annotation Guide
Annotation Example Files
Discrepancy Report
NCBI Prokaryotic Genome Annotation Pipeline
AGP Format
Complex Assembly Submission Guide
Metagenome Submission Guide
BioProject

Postup zaslání GenBank Standardního typu

- http://www.ncbi.nlm.nih.gov/books/NBK51157/

  The GenBank Submissions Handbook



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

BankIt

Submit new sequences to GenBank

What type of sequence data do you have?

- ○ SARS-CoV-2
- ○ Ribosomal RNA (rRNA) or rRNA-ITS
- ○ Metazoan (multicellular animal) COX1
- ○ Influenza virus
- ○ Norovirus
- ○ Dengue virus
- ○ Eukaryotic and Prokaryotic Genomes (WGS or Complete)
- ○ Transcriptome Shotgun Assembly (TSA)
- ○ Unassembled sequence reads (SRA)

- ○ Sequence data not listed above (through BankIt): mRNA, genomic DNA, organelle, ncRNA, plasmids, other viruses, phages, synthetic constructs

Start

Need help? Contact GenBank user services at info@ncbi.nlm.nih.gov.

Další požadavky na zaslání sekvence

- Informace o datu zveřejnění
- Informace o relevantních publikacích
- Popis zdroje sekvence
- Vlastní sekvence
  - typ a tvar molekuly
  - anotace vlastností sekvence

Popis zdroje sekvence 1

- **organism**
  nezkrácené vědecké jméno
  Příklad: [organism=Drosophila melanogaster]
- **lineage**
  taxonomické zařazení organismu (dle NCBI taxonomy database)
  http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root
- **molecule**
  ve tvaru "DNA" nebo "RNA".
  Příklad : [molecule=DNA]
- **moltype**
  může nabývat následujících hodnot
  Příklad : [moltype=Genomic DNA]
  - Genomic DNA
  - Genomic RNA
  - Precursor RNA
  - mRNA [cDNA]
  - Ribosomal RNA
  - Transfer RNA
  - Small nuclear RNA
  - Small cytoplasmic RNA
  - Other-Genetic
  - cRNA
  - Small nucleolar RNA
- **topology**

Popis zdroje sekvence 2

- **location**
  může nabývat následujících hodnot
  Příklad: [location=mitochondrion]
  - genomic
  - chloroplast
  - kinetoplast
  - mitochondrion
  - plastid
  - macronuclear
  - extrachromosomal
  - plasmid
  - cyanelle
  - proviral
  - virion
  - nucleomorph
  - apicoplast
  - leucoplast
  - proplastid
  - endogenous-virus
  - hydrogenosome
- **Genetic code**
  (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c)

Popis zdroje sekvence 3

Další popisovače ke zdroji sekvence

- acronym
- anamorph
- authority
- biotype
- biovar
- breed
- cell-line
- cell-type
- chemovar
- chromosome
- clone
- clone-lib
- collected-by
- common
- country
- cultivar
- dev-stage
- ecotype
- endogenous-virus-name
- forma
- forma-specialis
- fwd-pcr-primer-name
- fwd-pcr-primer-seq
- genotype
- group
- haplotype
- identified-by
- isolate
- isolation-source
- lab-host
- lat-lon
- map
- note
- pathovar
- plasmid-name
- plastid-name
- pop-variant
- rev-pcr-primer-name
- rev-pcr-primer-seq
- segment
- serogroup
- serotype
- serovar
- sex
- specific-host
- specimen-voucher
- strain
- sub-species
- subclone
- subgroup
- substrain
- subtype
- synonym
- teleomorph
- tissue-lib
- tissue-type
- type
- variety

Formát sekvence

- Sekvence nukleové kyseliny a kódovaných proteinů připravené ve formátu FASTA

```
Nucleotide Sequence:

>ABC-1 [organism=Saccharomyces cerevisiae][strain=ABC][clone=1]
ATTGCGTTATGGAAATTCGAAACTGCCAAATACTATGTCACCATCATTGA
TGCACCTGGACACAGAGATTTCATCAAGAACATGATCACTGGTACTT

Protein Sequences:

>4E-I [gene=eIF4E] [protein=eukaryotic initiation factor 4E-I]
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGN ...
>4E-II [gene=eIF4E] [protein=eukaryotic initiation factor 4E-II]
MVVLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGNTATTTAPAGDD ...
```

Přerušená sekvence

```
>m_gagei [organism=Mansonia gagei] Mansonia gagei NADH dehydrogenase ...
ATGGAGCATACATATCAATATTCATGGATCATACCGTTTGTGCCACTTCCAATTCCTATTTTAATAGGAA
TTGGACTCCTACTTTTTCCGACGGCAACAAAAAATCTTCGTCGTATGTGGGCTCTTCCCAATATTTTATT
GTTAAGTATAGTTATGATTTTTTCGGTCGATCTGTCCATTCAGCAAATAAATAAAAGTTCTATCTATCAA
TATGTATGGTCTTGGACCATCAATAATGATTTTTCTTTCGAGTTTGGCTACTTTATTGATTCGCTTACCT
>?200 ←——— Délka přerušení
GGTATAATAACAGTATTATTAGGGGCTACTTTAGCTCTTGC
TCAAAAAGATATTAAGAGGGGTTTAGCCTATTCTACAATGTCCCAACTGGGTTATATGATGTTAGCTCTA
GGTATGGGGTCTTATCGAGCCGCTTTATTTCATTTGATTACTCATGCTTATTCGAAGGCATTGTTGTTTT
TAGGATCCGGATCCGTTATTCATTCCATGGAAGCTATTGTTGGATATTCTCCAGATAAAAGCCAGAATAT
GGTTTTTATGGGCGGTTTAAGAAAGCATGTGCCAATTACACAAATTGCTTTTTTAGTGGGTACACTTTCT
CTTTGTGGTATTCCACCCCTTGCTTGTTTTTGGTCCAAAGATGAAATTCTTAGTGACAGCTGGTTGT
>?unk100 ←——— Přerušení neznámé délky
TCAATAAAACTATGGGGTAAAGAAGAACAAAAAATAATTAACAGAAATTTTCGTTTATCTCCTTTATTAA
TATTAACGATGAATAATAATGAGAAGCCATATAGAATTGGTGATAATGTAAAAAAAGGGGCTCTTATTAC
TATTACGAGTTTTGGCTACAAGAAGGCTTTTTCTTATCCTCATGAATCGGATAATACTATGCTATTTCCT
ATGCTTATATTGGCTCTATTTACTTTTTTTTGTTGGAGCCATAGCAATTCCTTTTAATCAAGAAGGACTAC
ATTTGGATATATTATCCAAATTATTAACTCCATCTATAAATCTTTTACATCAAAATTCAAATGATTTTGA
GGATTGGTATCAATTTTTAACAAATGCAACTCTTTCAGTGAGTATAGCCTGTTTCGGAATATTTACAGCA
TTCCTTTTATATAAGCCTTTTTATTCATCTTTACAAAATTTGAACTTACTAAATTTATTTTCGAAAGGGG
GTCCTAAAAGAATTTTTTTGGATAAAATAATATACTTGATATACGATTGGTCATATAATCGTGGTTACAT
```

Sekvenční přiložení

- Fasta+GAP

```
>ABC-1 [organism=Saccharomyces cerevisiae][strain=ABC][clone=1]
---ATTGCGTTATGGAAATTCGAAACTGCCAAATACTATGTCACCATCAT
TGATGCACCTGGACACAGAGATTTCATCAAGAACATGATCACTGGTACTT
>ABC-2 [organism=Saccharomyces cerevisiae][strain=ABC][clone=2]
GATATTGCTTTATGGAAATTCGAAACTGCCAAATACTATGTCACCATCAT
TGATGCACCTGGACACAGAAATTTCATCAAGAACATGATCACTGGTACTT
>ABC-3 [organism=Saccharomyces cerevisiae][strain=ABC][clone=3]
---ATTGCTTTATGGAAATTCGAAACTGCCAAATACTATGTTA-------
TGATGCACCTGGACACAGAGATTTCATCAAAAACATGATCACTGGTACTT
```

- PHYLIP

```
 3  100
ABC-1      ---ATTGCGT TATGGAAATT CGAAACTGCC AAATACTATG TCACCATCAT
ABC-2      GATATTGCTT TATGGAAATT CGAAACTGCC AAATACTATG TCACCATCAT
ABC-3      ---ATTGCTT TATGGAAATT CGAAACTGCC AAATACTATG TTA-------

           TGATGCACCT GGACACAGAG ATTTCATCAA GAACATGATC ACTGGTACTT
           TGATGCACCT GGACACAGAA ATTTCATCAA GAACATGATC ACTGGTACTT
           TGATGCACCT GGACACAGAG ATTTCATCAA AAACATGATC ACTGGTACTT

>[organism=Saccharomyces cerevisiae][strain=ABC][clone=1]
>[organism=Saccharomyces cerevisiae][strain=ABC][clone=2]
>[organism=Saccharomyces cerevisiae][strain=ABC][clone=3]
```

Anotace vlastní sekvence

- Kódované proteiny
  - CDS
    interval
    nekompletnost na N- nebo C- konci
  - gene
    interval odpovídající CDS u experimentálně prokázaných genů
  - mRNA
    interval obsahující 5‘-UTR a 3‘-UTR
- Kódované strukturní RNA

Příklady některých dalších modifikací deskriptorů

- Title
  - Informace vyskytující se v databázi v DEFINITION LINE
- Comment
  - Poznámka k různým vlastnostem
- Technique
  - Umožňuje výběr techniky použité pro vytvoření nebo experimentální evidenci vlastností sekvence

Přehled deskriptorů pro popis vlastností sekvence

(http://www.ncbi.nlm.nih.gov/BankIt/help.html)

- attenuator
- C-region
- CAAT_signal
- CDS
- conflict
- D-loop
- D-segment
- enhancer
- exon
- gap
- GC_signal
- gene
- iDNA
- intron
- J_segment
- LTR
- mat_peptide
- misc_binding
- misc_difference
- misc_feature
- misc_recomb
- misc_RNA
- misc_signal
- misc_structure
- modified_base
- mRNA
- N_region
- old_sequence
- operon
- oriT
- polyA_signal
- polyA_site
- precursor_RNA
- prim_transcript
- primer_bind
- promoter
- protein_bind
- RBS
- repeat_region
- repeat_unit
- rep_origin
- rRNA
- S_region
- satellite
- scRNA
- sig_peptide
- snRNA
- snoRNA
- source
- stem_loop
- STS
- TATA_signal
- terminator
- transit_peptide
- tRNA
- unsure
- V_region
- V_segment
- variation
- 3‘clip
- 3‘UTR
- 5‘clip
- 5‘UTR

Příklady sekvencí

Sekvence mRNA nebo cDNA

- Kódující oblasti včetně iniciačního a terminačního kodonu
- Název proteinu
- Název genu
- Sekvence proteinu

```
Homo sapiens prolidase (PEPD) mRNA, complete cds.

FEATURES             Location/Qualifiers
     source          1..1888
                     /organism="Homo sapiens"
                     /chromosome="19"
                     /map="19q12-q13.2"
                     /cell_type="fibroblasts"
     mRNA            1..1888
                     /gene="PEPD"
     gene            1..1888
                     /gene="PEPD"
     CDS             17..1498
                     /gene="PEPD"
                     /EC_number="3.4.13.9"
                     /note="imidodipeptidase"
                     /product="prolidase"
```

Sekvence prokaryotického genu

- Kódující intervaly
- Název proteinu
- Název genu, je-li známý
- Aminokyselinová sekvence

```
Escherichia coli RecA protein (recA) gene, complete cds.

FEATURES             Location/Qualifiers

     source          1..3300
                     /organism="Escherichia coli"
                     /strain="K-12"
     gene            783..1961
                     /gene="recA"
     CDS             783..1961
                     /gene="recA"
                     /function="DNA repair protein"
                     /product="RecA protein"
```

Sekvence eukaryotického genu

- Intervaly kódujících oblastí včetně start- a stop-kodonů a intervaly všech intronů
- Název proteinu
- Název genu, je-li známý
- Aminokyselinová sekvence

```
Caenorhabditis elegans tyrosine kinase PTK-2 (ptk-2) gene,
complete cds.

FEATURES             Location/Qualifiers
     source          1..3180
                     /organism="Caenorhabditis elegans"
     gene            211..3011
                     /gene="ptk-2"
     mRNA            join(211..288,533..703,763..890,940..1024,
                     1084..1380,1838..1962,2018..2099,2301..3011)
                     /gene="ptk-2"
                     /product="protein kinase PTK-2"
     CDS             join(250..288,533..703,763..890,940..1024,
                     1084..1380,1838..1962,2018..2099,2301..2456)
                     /gene="ptk-2"
                     /product="protein kinase PTK-2"
```

Ribosomální RNA a vnitřní přepisované mezerníky

- Názvy jakékoli strukturní RNA (např. tRNA-Ile, 16S ribosomal RNA)
- Názvy mezerníkových oblastí (např., internal transcribed spacer 1, 16S/23S intergenic spacer)
- Nukleotidové pozice

```
Saccharomyces cerevisiae 18S ribosomal RNA gene, partial sequence; internal
   transcribed spacer 1, 5.8S ribosomal RNA gene and internal transcribed
   spacer 2, complete sequence; and 28S ribosomal RNA gene, partial
   sequence.

FEATURES             Location/Qualifiers
     source          1..540
                     /organism="Saccharomyces cerevisiae"
                     /strain="UMD 334"
     rRNA            <1..5
                     /product="18S ribosomal RNA"
     misc_RNA        6..178
                     /product="internal transcribed spacer 1 "
     rRNA            179..377
                     /product="5.8S ribosomal RNA"
     misc_RNA        378..519
                     /product="internal transcribed spacer 2"
     rRNA            520..>540
                     /product="28S ribosomal RNA"
```

Oblast promotoru

- Název proteinu nebo genu, ke kterému patří promotor a jeho 5‘ a 3‘ obklopující sekvence
- Intervaly přepisovaných a kódujících sekvencí, pokud jsou přítomné

```
Homo sapiens enhancer-binding protein 2 (EBP2) gene, promoter region and partial
    cds.

FEATURES             Location/Qualifiers

     source          1..3061
                     /organism="Homo sapiens"
                     /chromosome="15"
                     /map="15q13"
                     /cell_line="H441"
                     /tissue_type="lung"
     gene            1..>3061
                     /gene="EBP2"
     promoter        1..2947
                     /gene="EBP2"
     TATA_signal     2918..2923
                     /gene="EBP2"
     mRNA            2948..>3061
                     /gene="EBP2"
                     /product="enhancer-binding protein 2"
     5'UTR           2948..3010
                     /gene="EBP2"
     CDS             3011..>3061
                     /gene="EBP2"
                     /product="enhancer-binding protein 2"
```

Transpozon nebo inzerční sekvence

Specifické jméno elementu

- Nukleotidoné pozice
- Jména a intervaly kódovaných genových produktů, pokud jsou přítomny (např., transposase)
- Pozice a intervaly dalších vlastností (např. LTRs, repeat regions)

```
Bacillus subtilis transposon BLT transposase (tnpA) gene,
complete cds

FEATURES             Location/Qualifiers
     source          1..1221
                     /organism="Bacillus subtilis"
                     /strain="RS2"
     source          21..1127
                     /organism="Bacillus subtilis"
                     /strain="RS2"
                     /transposon="BLT"
     repeat_region   21..61
                     /rpt_type=inverted
     gene            128..1034
                     /gene="tnpA"
     CDS             128..1034
                     /gene="tnpA"
                     /product="transposase"
     repeat_region   1085..1127
                     /rpt_type=inverted
```

Oblasti repeticí

- Intervaly repetitivních sekvencí
- Rodina repeticí (např., Alu, Mer)
- Typ repetice (tandem, inverted, flanking, terminal, direct, dispersed, or other)
- Jednotka repetice (repeat unit) popis intervalů, jestliže sekvence obsahuje více než jednu repetici

```
Homo sapiens repeat regions

FEATURES             Location/Qualifiers
     source          1..2050
                     /organism="Homo sapiens"
                     /chromosome="6"
                     /map="6q25"
     repeat_region   8..126
                     /rpt_type=dispersed
                     /rpt_family="B2"
     repeat_region   197..344
                     /rpt_type="direct"
                     /rpt_unit="197..220"
     repeat_region   389..673
                     /rpt_family="AluSx"
                     /rpt_type=dispersed
     repeat_region   847..876
                     /note="microsatellite BT21"
                     /rpt_type="tandem"
                     /rpt_unit="ca"
     repeat_region    1000..2000
                     /rpt_family="human endogeneous retrovirus K-10
```

Klonovací vektor

- Jedinečné jméno vektoru
- Kódující intervaly, jména genů a proteinů

```
Cloning vector pRB223, complete sequence

FEATURES             Location/Qualifiers
     source          1..4361
                     /organism="Cloning vector pRB223"
     gene            86..1276
                     /gene="tet"
     CDS             86..1276
                     /gene="tet"
                     /product="tetracycline resistance protein"
     RBS             1905..1909
                     /note="Shine-Dalgarno sequence"
     rep_origin      2535
     gene            complement(3293..4194)
                     /gene="bla"
     CDS             complement(3293..4153)
                     /gene="bla"
                     /product="beta-lactamase"
     misc_feature    4069..4125
                     /note="multiple cloning site"
     RBS             complement(4161..4165)
                     /gene="bla"
                     /note="Shine-Dalgarno sequence"
     promoter        complement(4188..4194)
                     /gene="bla"
```

Bacteriophage lysis module; endolysin and HNH endonuclease genes, complete CDS

```
FEATURES             Location/Qualifiers
     source          1..3165
                     /organism="Staphylococcus bacteriophage 812"
                     /virion
                     /mol_type="genomic DNA"
                     /strain="phi812"
                     /lab_host="Staphylococcus aureus CCM 4028"
                     /type="wild type"
     gene            654..3017
                     /gene="lyt812"
     CDS             join(654..1449,2329..3017)
                     /gene="lyt812"
                     /experiment="peptide sequencing"
                     /note="Lyt812"
                     /codon_start=1
                     /transl_table=11
                     /product="endolysin"
                     /translation="MAKTQAEI…… "
     misc_feature    join(1239..1449,2329..2576)
                     /gene="lyt812"
                     /note="SM00644; Ami_2; This family includes zinc amidases
                     that have N-acetylmuramoyl-L-alanine amidase activity;
                     Region: Ami_2"
     intron          1450..2328
                     /gene="lyt812"
                     /standard_name="lyt812-I1"
                     /experiment="cDNA synthesis and sequencing"
     CDS             1617..2117
                     /gene="lyt812"
                     /note="ORFI-812III"
                     /codon_start=1
                     /transl_table=11
                     /product="putative HNH endonuclease
```