Data v mikrobiologii

CORE042: Data – odpověď na základní otázku života, vesmíru a vůbec...
6. přednáška

Stanislava Bezdíček Králová <kralova.s@sci.muni.cz> a kol.
Externista – Přírodovědecká fakulta MU

Přednášející a spoluautoři

- **Stanislava Bezdíček Králová**
  - University of Vienna, Center of Microbial Ecology a nd System Science
  - Přírodovědecká fakulta MU, externista
  - **Life science projekty, FAIR Data v mikrobiologii**

- **Václav Šeda**
  - CEITEC
  - Mikroprostředí imunitních buněk
  - **FAIR data v klinickém a univerzitním výzkumu**

- **Matěj Bezdíček**
  - Fakultní nemocnice Brno
  - Interní hematologická a onkologická klinika
  - **FAIR data v klinické mikrobiologii**

MUNI

Osnova

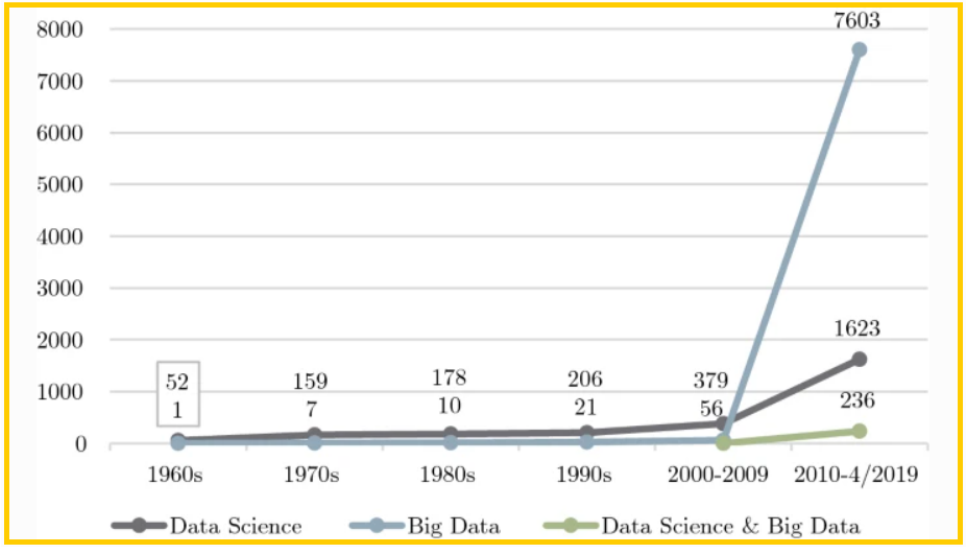
- Life cycle
- Historie vs. současnost
- Plánování
- Sběr/získávání dat
- Zpracování dat
- Analýza dat
- Uchování dat
- Poskytování dat
- Opakované využití dat

MUNI

Historie

V minulosti

- méně dat
- menší výpočetní náročnost
- jednoduší na uchování, organizaci, vyhodnocení



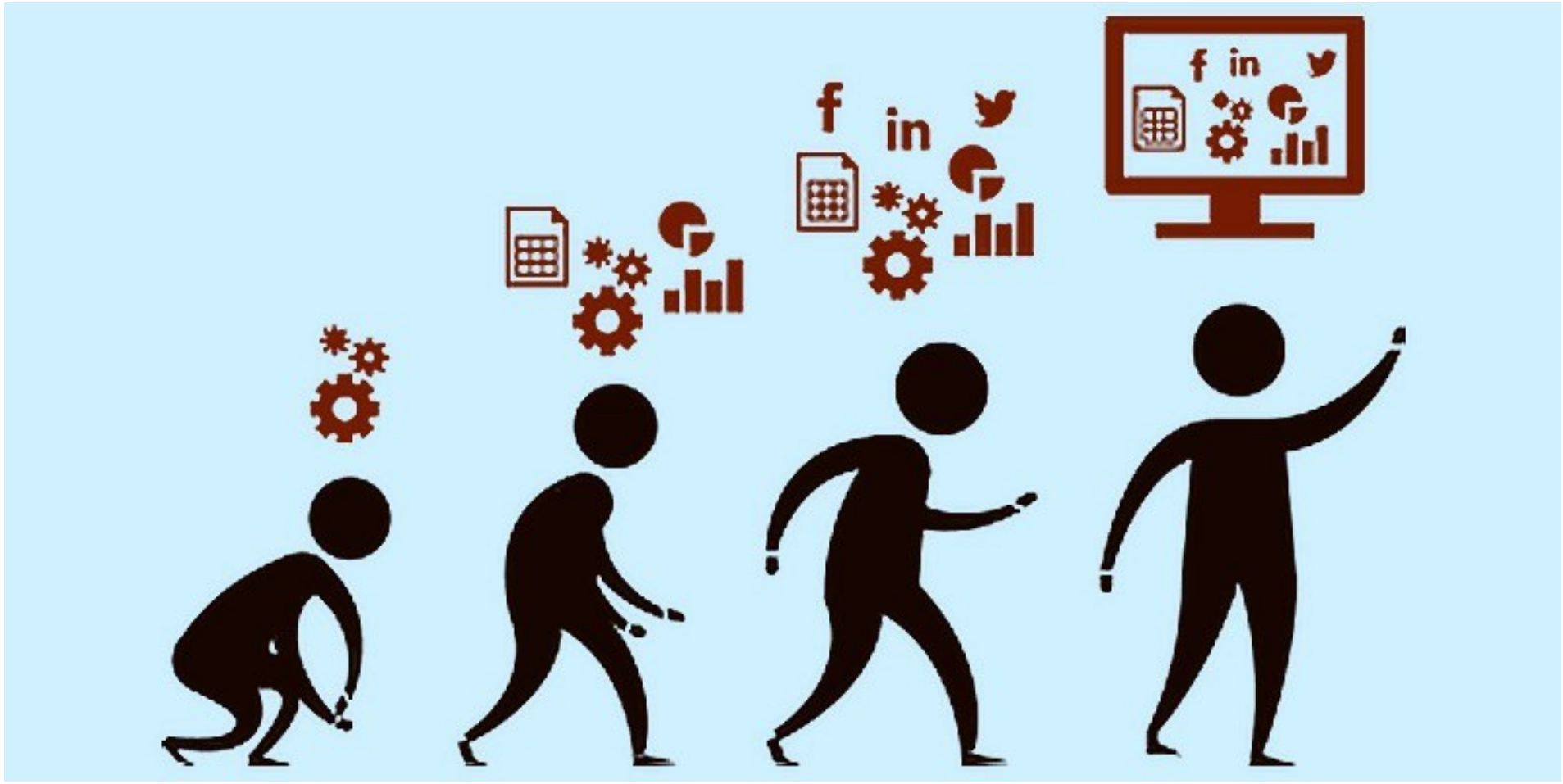
Evolutionary trend in the number of publications covering data science and big data

Raban & Gordon, Scientometrics (2020)

Současnost

- změna instrumentace
- změna charakteru dat (pozorování vs. sekvenace)
- enormní nárust množství dat – digitální forma
- vyšší nároky na rutinní procesovaní – zisk, organizace dat, jejich sdílení, uchovávaní, procesovaní (úpravy), analýzy

MUNI



MUNI

MUNI

Plánování

Co?

- stanovení cíle

Jak?

- určení metodologického postupu
  (máme vše co potřebujeme?)
- určení znalostního základu
  (dokážeme uskutečnit všechny kroky?)
- dohodnutí spolupráce
- rozložení finančních prostředků

Kdo?

- jeden výzkumník?
- tým?
- studenti?
- kdo to zaplatí?!

Kde?

- získat vzorek + logistika
- máme potřebnou instrumentaci?
- dílčí časti u spolupracujícího pracoviště? (tuzemsko/zahraničí)

Kdy?

- časové plánovaní je extrémně důležité
- kdy odeberu vzorky
- stihnu je pak zpracovat?
- kolik vzorků zpracuji?
- potřebuji více odběrů v různých časech?
- dílčí kroky vs. celkový projekt

MUNI

Plánování – specifika klinického výzkumu

Etická komise

- Lidská DNA ve vzorcích = etický problém → nutné schválení
- Ochrana osobních údajů
- Informovaný souhlas

Původ vzorků?

- ambulantní vs. hospitalizovaný pacient
- transport
- krátkodobé uchování
- dlouhodobé uchování
- jeden výzkumník?
- tým?
- studenti?
- kdo to zaplatí?!

Kdy?

- časové plánovaní je extrémně důležité
- jak často budeme odebírat vzorky?
- hledám něco typické nebo spíš raritní?? → někdy extrémně dlouhá doba do vytvoření dostačujícího souboru vzorků

Kdo?

- interpretace: často nutný klinik
- často nutné zapojení více pracovišť

Financování

- nemocnice
- granty → klinický vs základní výzkum

MUNI

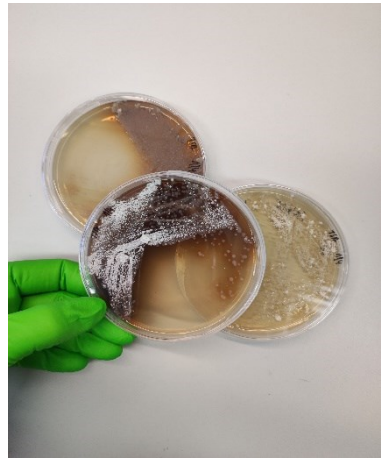
Plánování – příklady

Terénní výzkum (Antarktida)

- chystání materiálu (6 měsíců před expedicí)
- chystání sensitivního materiálu (1 měsíc před expedicí)
- jak budou vzorky cestovat?
- jak budou vzorky cestovat?
- jak dokážu zpracovat vzorky?

Přílet na Antarktidu:

1. příprava kultivačních médií
2. sterilizace materiálu po cestě
3. odběry **na začátku** expedice (kultivace 6–8 týdnů)
4. zpracovávaní vzorků
5. odběry **na konci** expedice (přenos do ČR)
   - ✓ nesmí roztát
   - ✓ nesmí se zahřát
   - ✓ nesmí se střídat teploty
   - ✓ kolik gramů
   - ✓ kolik replikátů
   - ✓ původ vzorků



Když plánovaní experimentu nevyjde – příklad z klinické praxe

Studium invazivních mykóz

- Zahájení nového projektu současně se sběrem vzorků
- Spolupráce s klinikou v rámci nemocnice
- Za rok sběru vzorků nasbírám 2 potvrzené a 3 suspektní vzorky na invazivní mykózu

- výsledky → ***z výsledků u 5 vzorků neudělám relevantní závěry, nic nepublikuji***

Studium invazivních mykóz

- Zahájení spolupráce s dalšími centry a po dlouhodobém bankování vzorků
- Zahájení projektu s již získaným větším množstvím vzorků

- výsledky → ***signifikantní počet vzorků k vyšetření umožňuje výsledky hodnotit, korelovat a dělat statisticky významné závěry, publikaci práce založené na takovém souboru nic nebrání***

MUNI

Když plánovaní experimentu nevyjde – příklad z praxe

Myší experiment – nepromyšlené

- ráno: podaní sulfoquinolózy žrádle (10mg/l)
- kontrolní skupina – bez přídavku
- po 12 hodinách odběr stolice
- sekvenování

- výsledky → ***žádný rozdíl*** kontrola vs. testovaná skupina

Myší experiment – kvalitní plán

- večer podání sulfoquinolózy žrádle (10mg/l)
- kontrolní skupina – bez přídavku
- po 3, 6,12 a 24 hodinách odběr stolice
- sekvenování každé vzorky

- výsledky → ***rozdíl*** mezi skupinami ***pouze v 6 h***

MUNI

MUNI

Získávání dat

Terén

- pozorování + zápis
- měření + zápis

Laboratoř

- pozorování + zápis
- měření + zápis
- generace dat přístrojem → hrubá data

experimentálně (původce = vy)

Sdílená data

- použití dat v databázích s open access
- použití poskytnutých dat
- „share and re-use“

Placená data

- placené databáze

původce je někdo jiný

MUNI

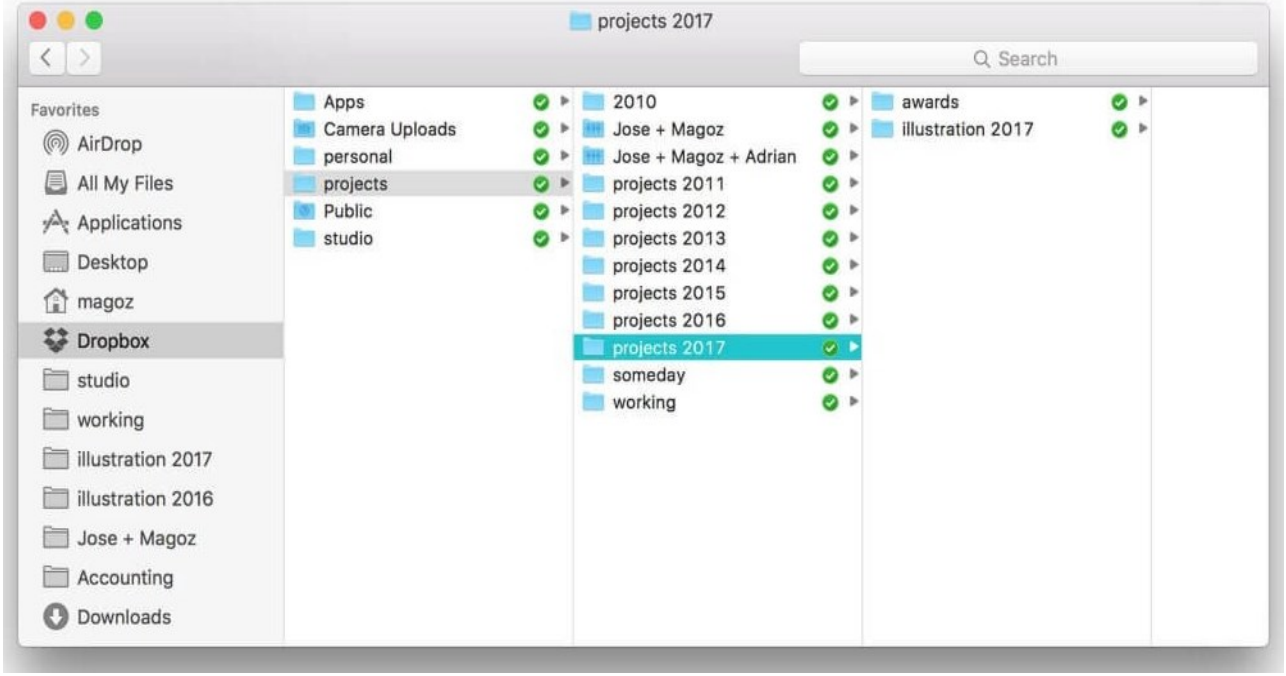
Organizace dat

Složky → podsložky → soubory

- konzistentní pojmenování (název projektu, název analýzy, datum provedení analýzy)
- hierarchický systém
- oddělení probíhajících a uzavřených projektů
- najít si vlastní systém který jste schopní **dlouhodobě sledovat a udržet**

Pojmenování souborů

- konzistentně
- datumy v určitém sledu
- interpunkce (_ / . mezery???)
- číselné pořadí na začátku / na konci



Organizace dat

Zlatý standard – tabulky

- jasně
- stručně
- jedna poznámka jedno místo
- stejný formát v danem sloupci
- jeden nadpis na sloupec
- žádné slučování buněk
- žádné prázdné buňky
- kde to jde dělat vybírací seznamy
  (víc lidí pracuje se stejnou tabulkou)
- žádné výpočty v raw data souboru

	A	B	C	D	E	F
1	Region	Retailer Type	Qtr1	Qtr2	Qtr3	Qtr4
2	**Mid West**					
3		Food & Staples	19,47,100	26,93,000	19,50,300	19,11,400
4		Multiline	9,77,200	7,29,700	3,67,800	3,22,600
5	**North East**					
6		Specialty	14,55,100	14,22,700	4,98,200	17,86,900
7	**South**					
8		Food & Staples	16,91,700	21,39,800	18,30,000	17,21,800
9		Multiline	18,37,500	18,80,500	17,20,500	24,89,200
10		Specialty	11,55,100	21,00,200	16,52,100	9,07,700
11	**West**					
12		Food & Staples	12,50,900	23,68,200	15,55,100	14,55,500
13		Multiline	13,74,300	7,93,200	12,54,900	15,17,400

BAD DESIGN

	A	B	C	D	E
1	Region	Qtr1	Qtr2	Qtr3	Qtr4
2	Mid West	2924300	3422700	2318100	2234000
3	North East	1455100	1422700	498200	1786900
4	South	4684000	6220500	5202600	5118700
5	West	2625200	3161400	2810000	2972900

BAD DESIGN

	A	B	C	D	E	F	G	H	I
1	Region	Qtr1	Qtr2	Qtr3	Qtr4		Region	Quarter	Sales
2	Mid West	2924300	3422700	2318100	2234000		Mid West	Qtr1	2924300
3	North East	1455100	1422700	498200	1786900		Mid West	Qtr2	3422700
4	South	4684000	6220500	5202600	5118700		Mid West	Qtr3	2318100
5	West	2625200	3161400	2810000	2972900		Mid West	Qtr4	2234000
6							North East	Qtr1	1455100
7							North East	Qtr2	1422700
8							North East	Qtr3	498200
9							North East	Qtr4	1786900
10							South	Qtr1	4684000
11							South	Qtr2	6220500
12							South	Qtr3	5202600
13							South	Qtr4	5118700
14							West	Qtr1	2625200
15							West	Qtr2	3161400
16							West	Qtr3	2810000
17							West	Qtr4	2972900

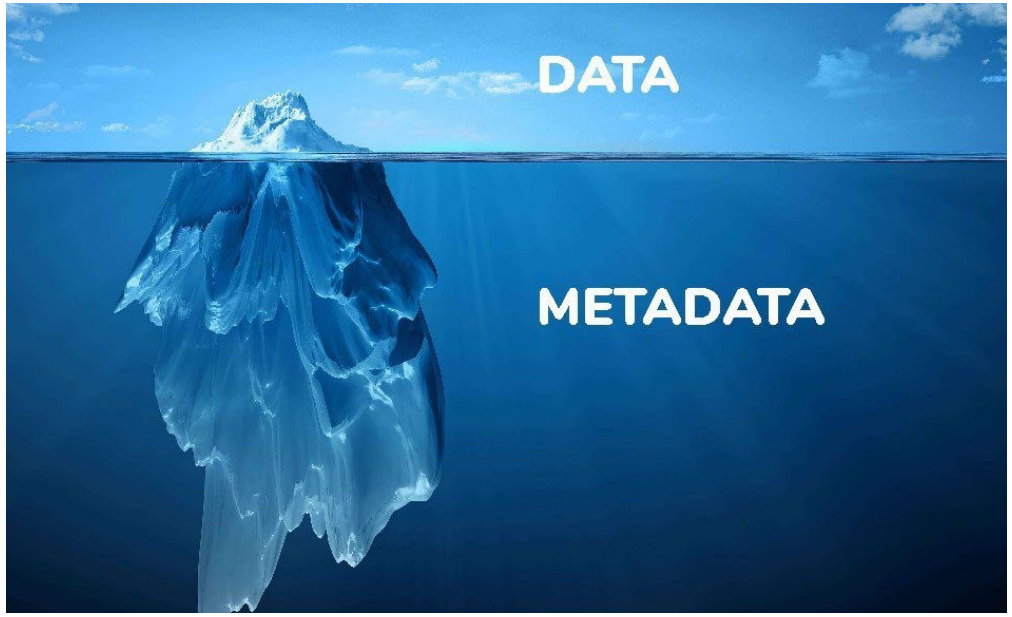
Organizace dat a metadat

Doplňkové informace

- mohou být uloženy separátně od metadat nebo dat
  - dotazníky (pacienti, probandi)
  - **laboratorní deníky**
  - **metadata**
  - reporty
  - publikace

Rozšiřující informace

- přesné popisky míst z tabulky
- přesný popis odběru vzorků z míst v tabulce



Doplňkové informace → nutnost?!

Laboratorní deníky

- novinka grantových agentur i některých pracovišť (zejména zahraničí)
- každý člen týmu vede laboratorní deník – přesné záznamy
- skladovaní i po skončení grantu
- varianty: papírová i elektronická

Příklad: Rakousko – přísná pravidla

- denně podrobné zápisy o průběhu experimentů
- žádné škrtání
- žádné trhání listu
- elektronická verze musí byt opatřena časovým zámkem

→

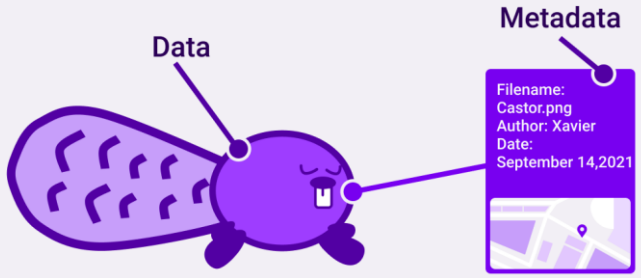
- naučit se zaznamenávat práci v laboratoři
- případné změny = lehčí implementace
- nutnost u projektů i grantů, i když to nevyžadují

MUNI

Metadata v biologii

Metadata = informace o nasbíraných datech

- **Fotky:** jak-kde-kdy vznikli, čím byli pořízeny, barevný profil, čočky, zoom…
- **Vzorek:** kde se odebral, GPS, kolik, typ zdrojového materiálu, jak se skladoval, jak se transportoval, jak se zpracoval
- **Bakteriální izolát:** Z jaké vzorky pochází, za jaké teploty roste, na jakých mikrobiologických půdách roste, do jaké BSL kategorie spadá
- **Experimentální data:** design jednotlivých experimentů, teploty, pH, živiny, obsah vody, salinita, objemy, hmotnosti, trvání, jakékoliv změny



Metadata v biologii

Sekvenační data:

Links from BioProject

Flavobacterium flabelliformis sp. nov. P4023T

Identifiers	BioSample: SAMN18262075; Sample name: Flavobacterium flabelliformis sp. nov. P4023T
Organism	Flavobacterium flabelliforme cellular organisms; Bacteria; FCB group; Bacteroidetes/Chlorobi group; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Flavobacterium
Package	MIGS: cultured bacteria/archaea, soil; version 5.0

Attributes

strain	P4023
collection date	2011-02-02
depth	0
elevation	1630 m
broad-scale environmental context	polar environment [ENVO:01001703]
local-scale environmental context	polar biome [ENVO:01000339]
environmental medium	nest of bird [ENVO:00005805]
geographic location	Antarctica
isolation and growth condition	R2A agar, 20 degrees C
latitude and longitude	63.77 S 57.78 W
number of replicons	1
reference for biomaterial	CCM 9062T
isolation source	sample taken from an organic matter in an abandoned nest of a bird
type-material	type strain of Flavobacterium flabelliforme

Description	Keywords: GSC:MIxS;MIGS:5.0
BioProject	PRJNA713758 Flavobacterium flabelliforme strain:P4023 Retrieve all samples from this project
Submission	Masaryk University, Stanislava Kralova; 2021-03-11

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

Metadata v biologii

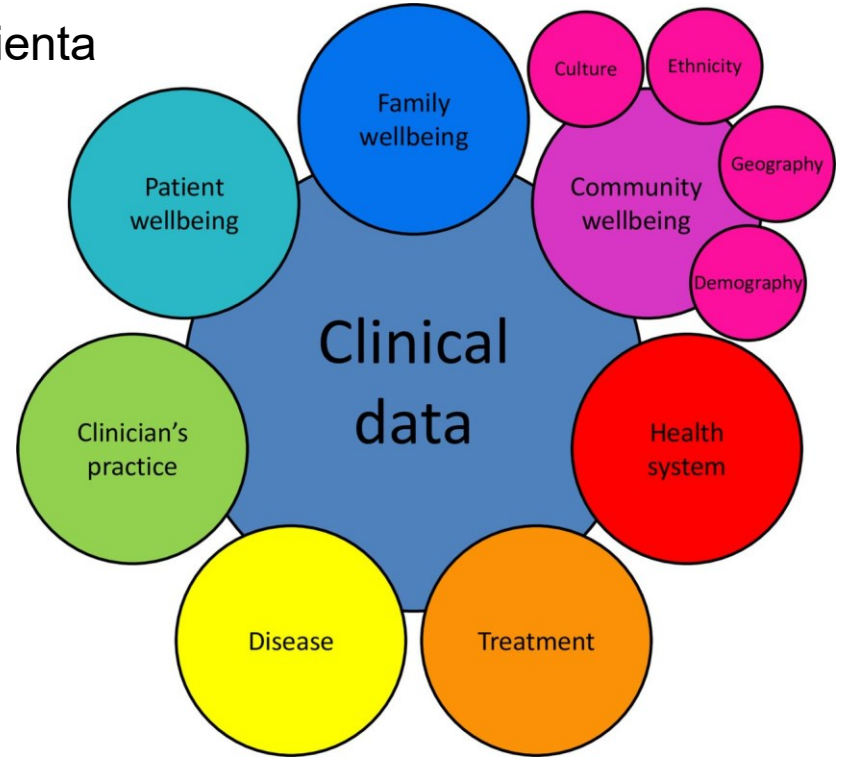
Tvorba a uchování metadat – jedno pravidlo: „There is no good reason not to save metadata“

EU projekty

- musíte uchovávat **vše**
  (musíte mít odhad o velikosti a typu dat)
- po dobu projektu **+ 5 let** po ukončení
  (nutno přemyslet dopředu, kdo bude odpovědná osoba?)
- aktuálně: metadata musíte nahrát **do veřejně dostupných** databází a spárovat s daty

A co klinická data ve výzkumu?

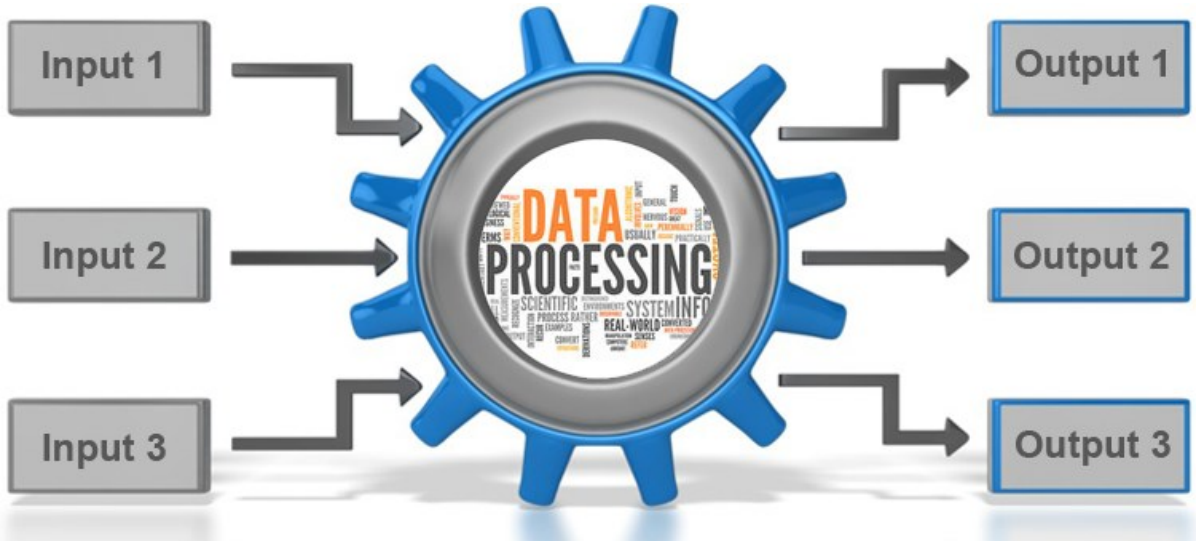
- Specifické
- **Přísná pravidla** na sdílení informaci
- **Pseudonymizace** – identifikace pacienta + dat možná dle klíče (př. náhodné čísla jako ID)
- **Anonymizace (de-identifikace)** → oddělení dat od identifikace pacienta
- Neanonymizované data nesmí opustit nemocnici



MUNI

Data processing

- **Vložená nasbíraná data „input“**
- **Příprava dat** (filtrace, kontrola kvality)
- **Zpracovaní** (z raw, tedy tzv. surových dat)
- **Interpretace „output“**



raw data
často nečitelná člověkem

jednodušší
zobrazení

MUNI

Krok 1: Filtrace dat

RAW DATA → ZPRACOVANÁ DATA

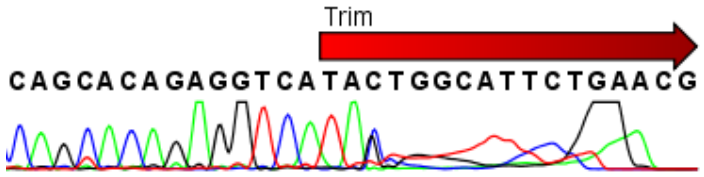
- **Kvalitativní vyhodnocení dat**

- **Normalizace** získaných dat pro možnost porovnání
- **Odstranění** chybných/nevyhovujících dat
- **Organizace**
- **Anotace**

Příklady – filtrace dat

Genomická data

- **sekvenace genomu**
  - výsledek tisíce čtení (readů) – nejsou stejné
  - jiná délka
  - jiná kvalita bází
  - přítomnost adaptorů
  - přítomnost kontaminací
  - nekvalitní koncové sekvence (jiné báze než v skutečnosti)



Kontrola kvality
→ filtrace

Phred skóre kvality –
kvalita jednotlivých bází

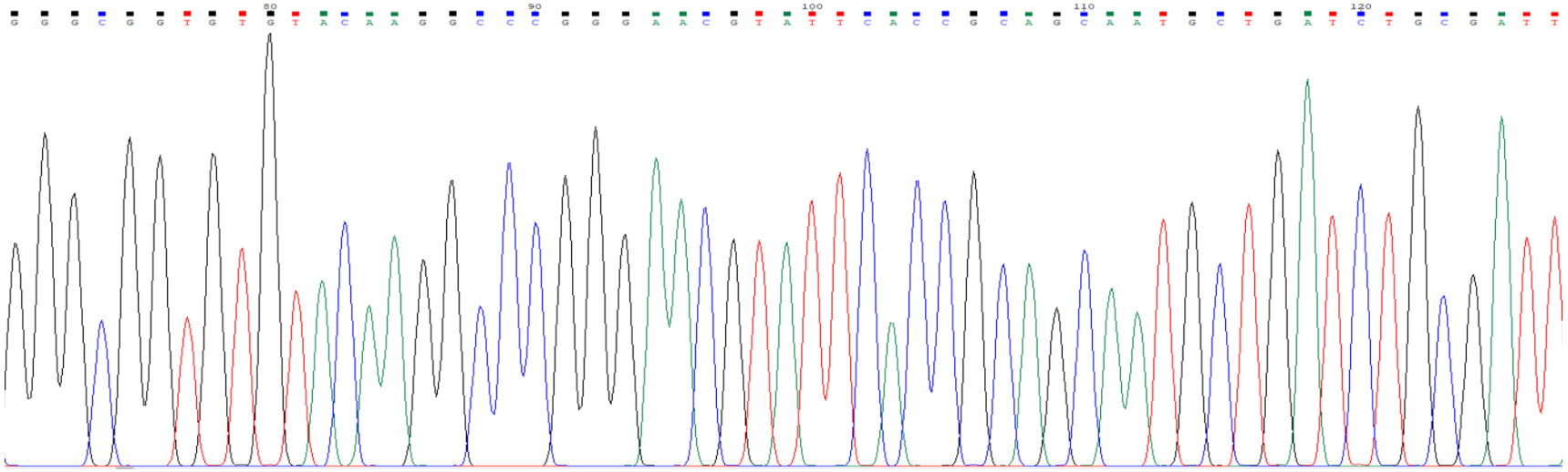
Minimální délka readů

Výsledný soubor

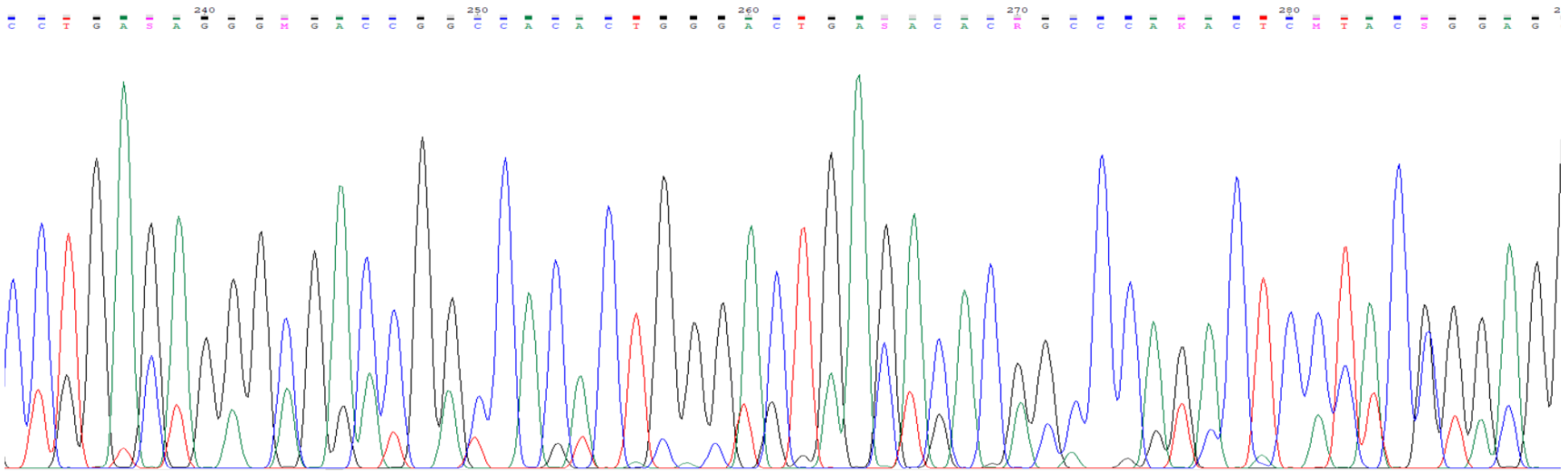
MUNI

Příklady – filtrace dat

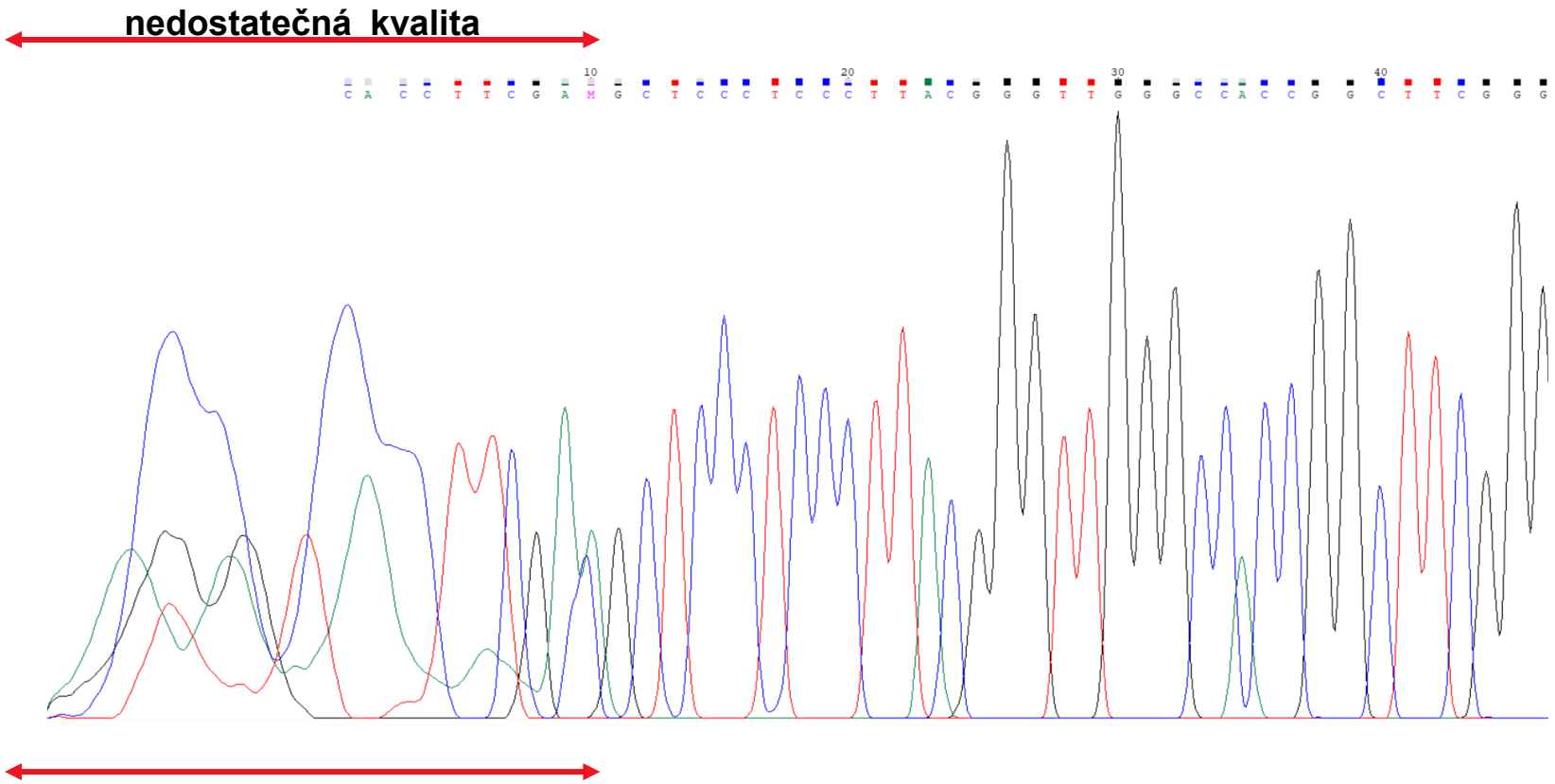
čistá sekvence



smíšená sekvence



Příklady – filtrace dat

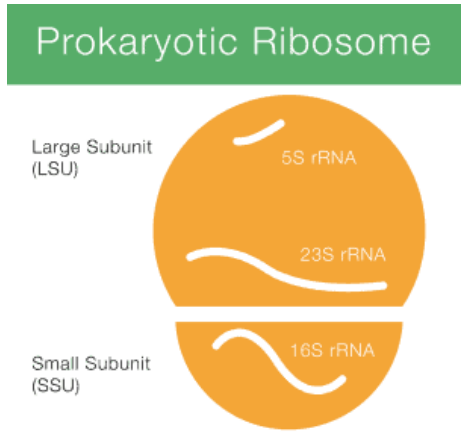


MUNI

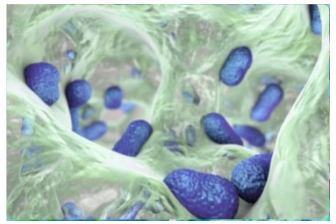
Příklady – filtrace dat

Metagenom – lidský střevní mikrobiom

- PCR – výběr pouze jednoho genu (16S rRNA)
- sekvenace
- výsledek je nutné **filtrovat** (obdobně jako příklad 1)
- výsledek je nutné **normalizovat**

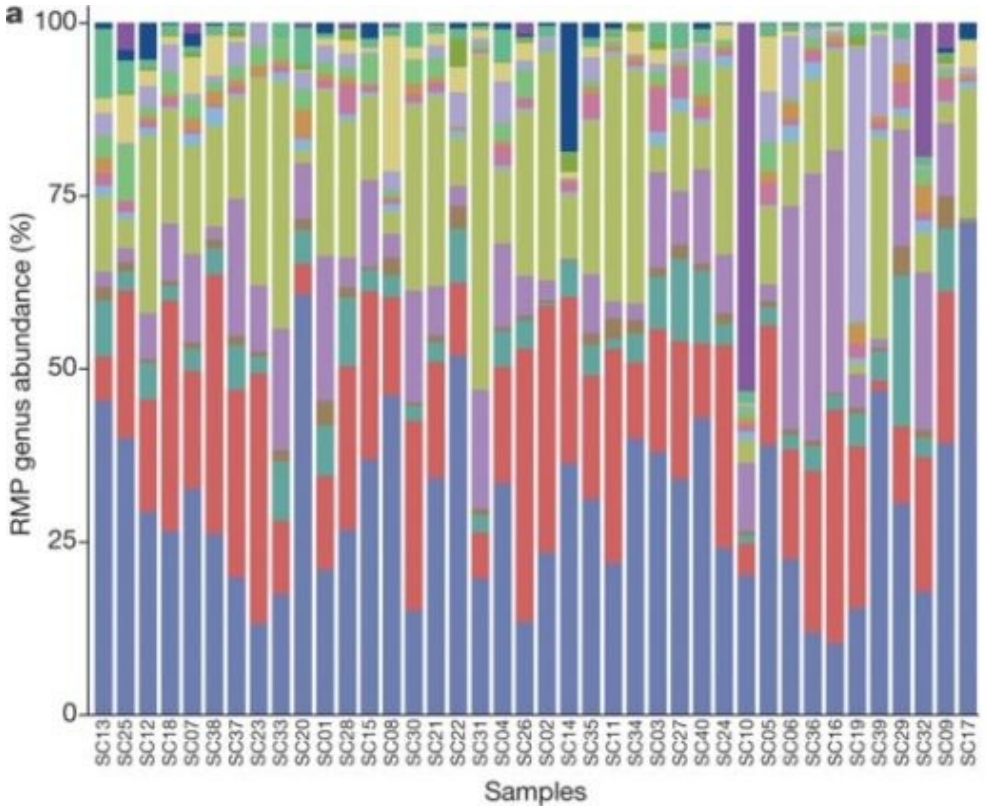


příklad – výsledek tvrdí že máme v stolici:

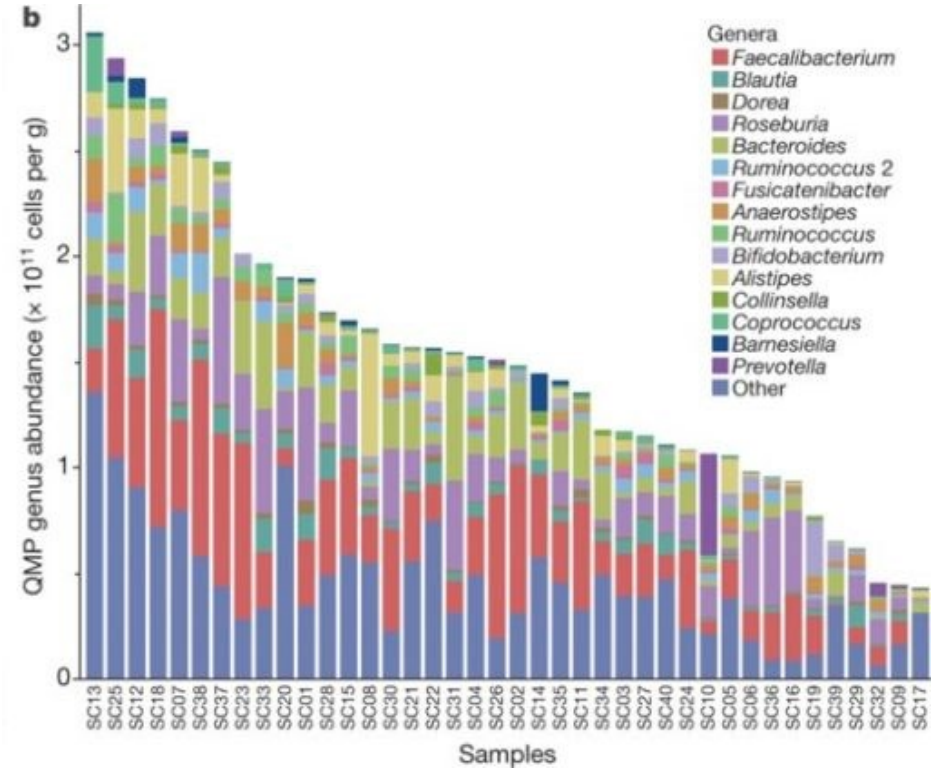


- 100× *Bacteroides fragilis* a 100× *Prevotella copri* → 1:1
- normalizace genem 16S rRNA – *B.fragilis* 6 kopií X *P. copri* 4 kopie → 3:2

Příklad normalizace sekvenačních dat



bez normalizace
tzv „**relativní početnosti**“



po normalizaci vzhledem k velikosti vzorku
tzv „**absolutní početnosti**“

Michaelsen, Albertsen Lab, 2018

Kontrola kvality dat

- **Stanovení postupu kontroly kvality dat**

  (kvalita fotek, kvalita sekvencí, úplnost metadat, kompatibilita dat, odstranění chybějících hodnot)

- **Parametry kontroly vždy dle daných dat**
  - **sekvenace:** hloubka čtení, délka čtení, překrytí, kontaminace…
  - **experimentální data:** pozitivní a negativní kontroly, minimální počet opakování, reprodukovatelnost, statistické vyhodnocení…

- kvalita kontroly dat je nutná prakticky v každém kroku – sběr, generace dat, zpracování dat
- možná automatizace

Problémy se zpracováváním dat

- Zejména BIOINFORMATIKA → tj. sekvenační data
- neexistují robustní univerzální pipeliny
- každý si vytváří vlastní systém:

  - **reprodukovatelnost = těžká až nemožná**
  - **jedna data = různé výsledky!!!**

MUNI

MUNI

Analýza dat

- biochemické a fyziologické data + způsoby vyhodnocení
- genomická data – bioinformatika (různé cíle = různé zpracování + příklady)
- metagenomická data (dtto)
- analýza dat od pacientu a asociace s výzkumnými daty

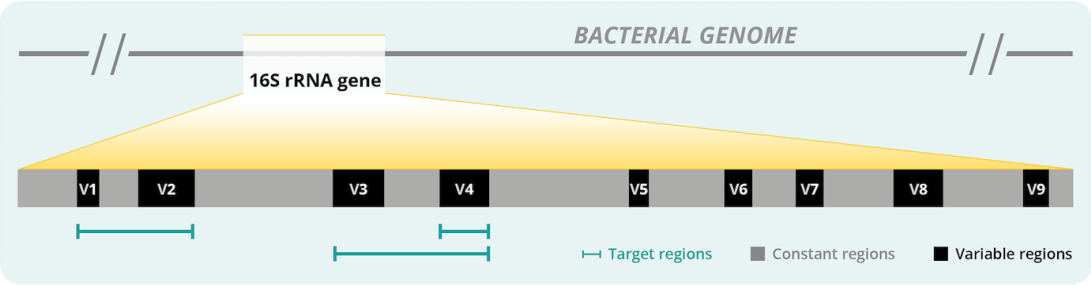
Příklady:

- popisy nových bakteriálních druhů, fylogeneze, genomika
- sledování epidemiologie klinických kmenů
- hledání nových antibiotik

MUNI

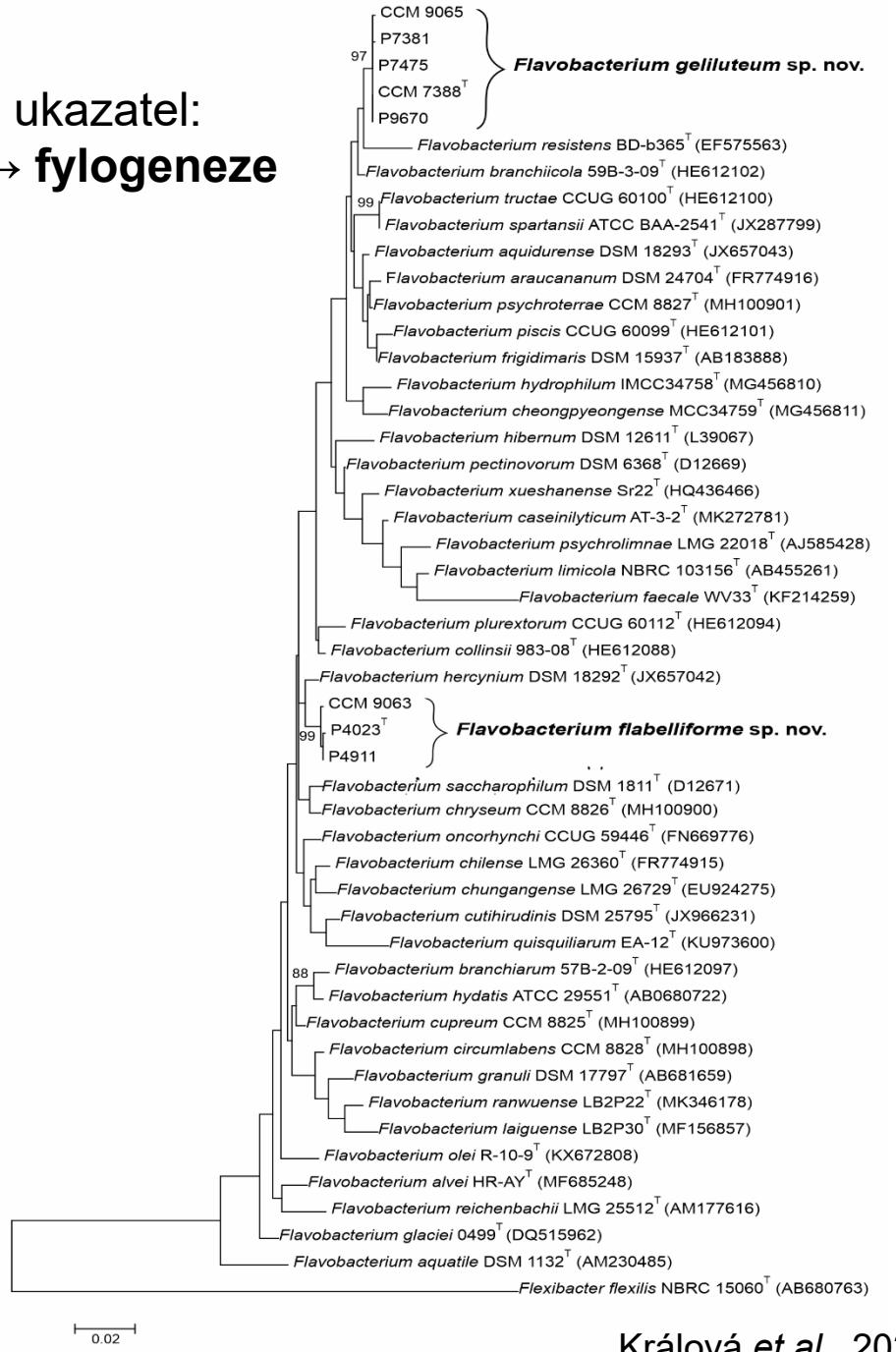
Příklad analýzy 1 (nové bakteriální druhy – fylogeneze)

– Vstup: filtrovaná sekvenační data
jednotlivé geny



1. ukazatel:
→ **Cut-off value:** 98.65% podobnosti genu 16S rRNA

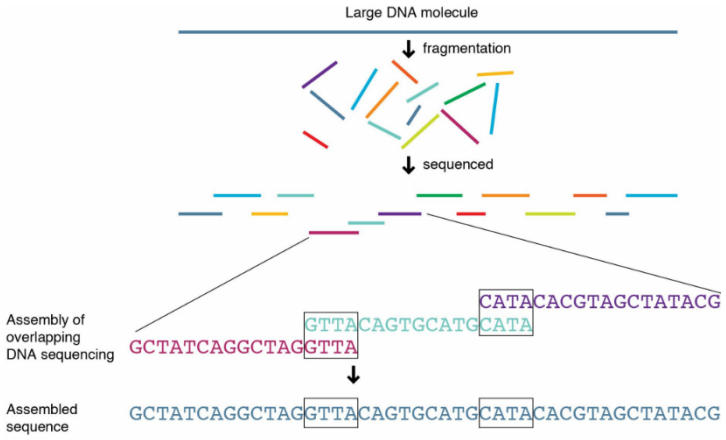
2. ukazatel:
→ **fylogeneze**

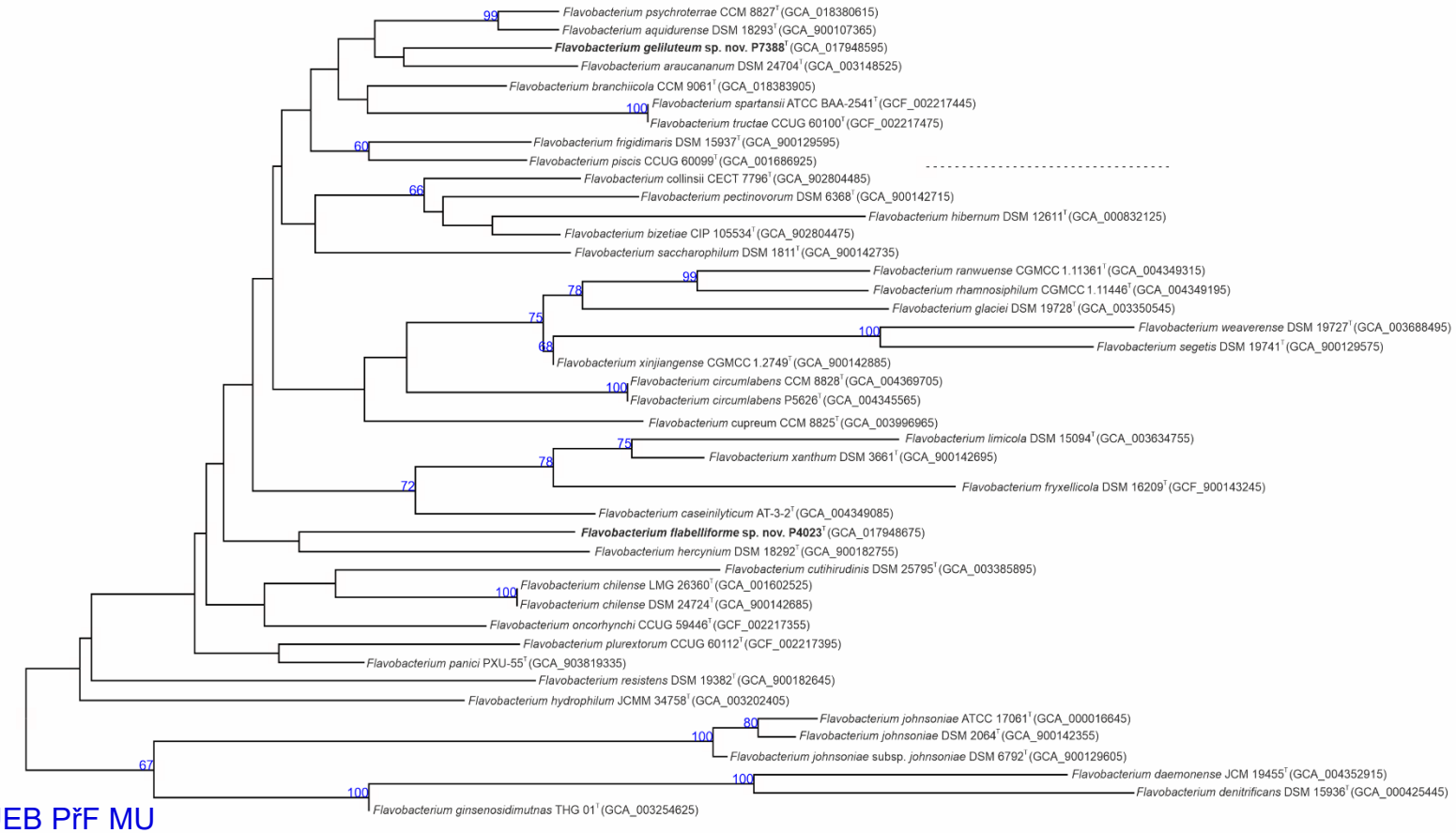


Králová *et al.*, 2021

Příklad analýzy 1 (genomika)

– Vstup: filtrovaná sekvenační data, po kvalitě kontroly

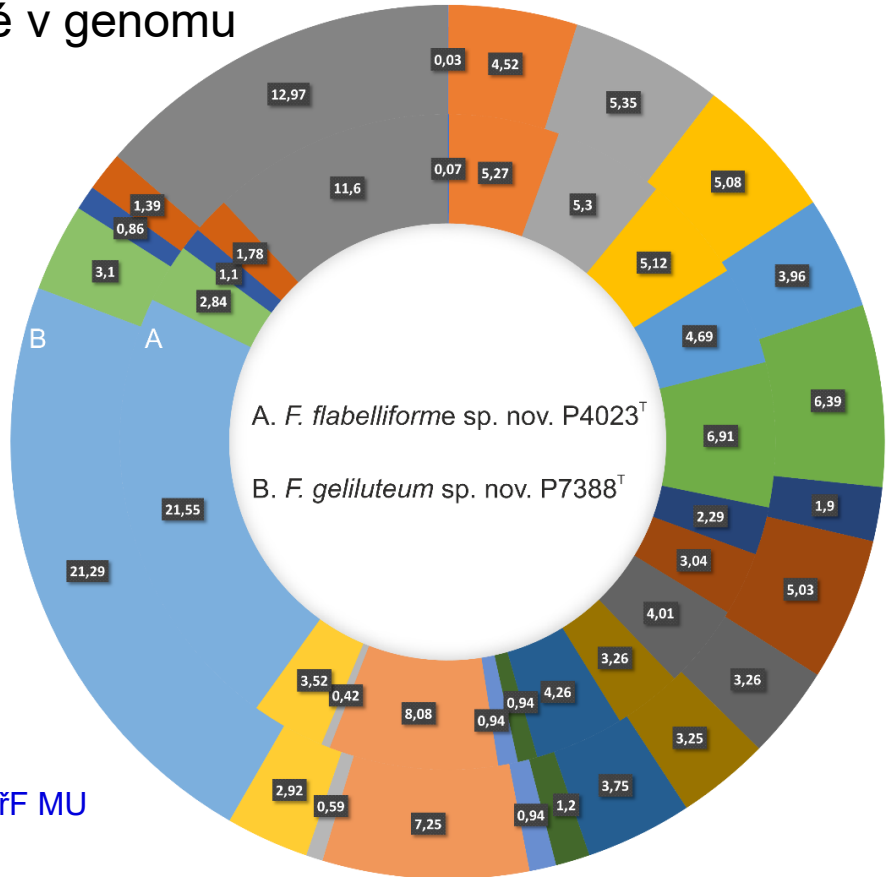




1. ukazatel:
→ **Cut-off value:** 95% podobnosti genomu
2. ukazatel:
→ **fylogeneze**

Příklad analýzy 1 (genomika)

- Vstup: filtrovaná sekvenační data, po kvalitě kontroly
  - nejenom, jestli je bakterie „nová“
  - můžeme přidat vlastnosti kódované v genomu



(A) RNA processing and modification
(J) Translation, ribosomal structure and biogenesis
(K) Transcription
(L) Replication, recombination and repair
(C) Energy production and conversion
(E) Amino acid transport and metabolism
(F) Nucleotide transport and metabolism
(G) Carbohydrate transport and metabolism
(H) Coenzyme transport and metabolism
(I) Lipid transport and metabolism
(P) Inorganic ion transport and metabolism
(Q) Secondary metabolites biosynthesis, transport, and catabolism
(D) Cell cycle control, cell division, chromosome partitioning
(M) Cell wall/membrane/envelope biogenesis
(N) Cell motility
(O) Post-translational modification, protein turnover, and chaperones
(S) Function unknown
(T) Signal transduction mechanisms
(U) Intracellular trafficking, secretion, and vesicular transport
(V) Defence mechanisms
(V) COG unknown

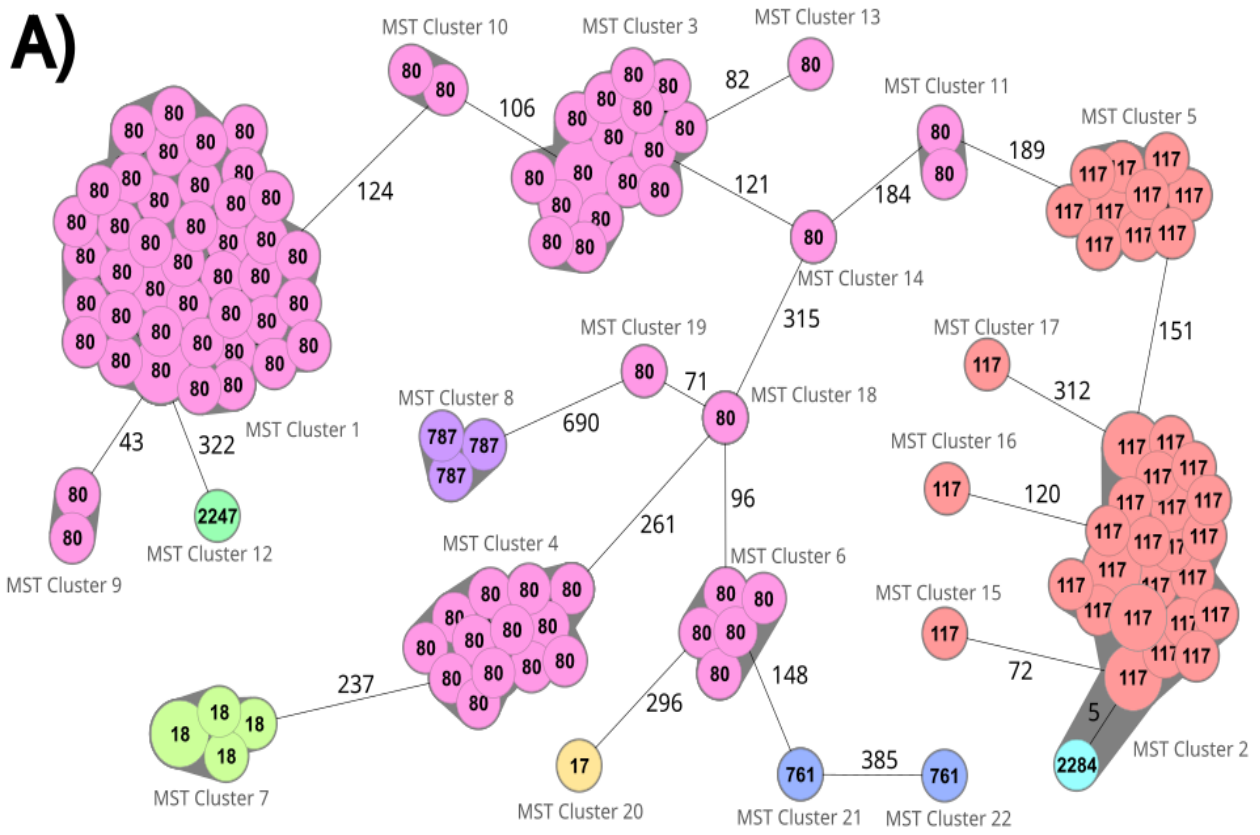
Králová *et al.*, 2021

Příklad analýzy 2 (epidemiologie)

Stanovení příbuznosti izolátů

VRE na základě WGS

- Vstup
  - Složené genomy
  - Metadata
- Výstupy
  - Rozdělení do clusterů na základě počtu rozdílů v alelách jednotlivých genů
  - Srovnání izolátů v rámci clusteru s dostupnými metadaty
  - Na základě výsledků případně zahájení epidemiologického šetření a zavedení příslušných opatření



Příklad analýzy 3 (biosyntetika)

- **Vstup: filtrovaná sekvenační data, po kvalitě kontroly**
- Vstup: fotografie
- Vstup: chromatografická data

MUNI

Příklad analýzy 3 (biosyntetika)

Overview 5.1 11.1 11.2 11.3 11.4 11.5 11.6 15.1 15.2 15.3 16.1 16.2 17.1 17.2 18.1 18.2 19.1 19.2 21.1

Identified secondary metabolite regions using strictness 'relaxed'

c00005_NZ_JABJ.. (original name was: NZ_JABJRC010000014.1)

Region	Type	From	To	Most similar known cluster	Similarity
Region 5.1	siderophore	53,440	65,008		

c00011_NZ_JABJ.. (original name was: NZ_JABJRC010000001.1)

Region	Type	From	To	Most similar known cluster		Similarity
Region 11.1	transAT-PKS, T1PKS, NRPS	586,329	666,426	9-methylstreptimidone	Polyketide:Modular type I	9%
Region 11.2	T3PKS	671,842	712,894	alkylresorcinol	Polyketide	100%
Region 11.3	melanin	1,268,489	1,278,860			
Region 11.4	siderophore	1,349,488	1,364,055			
Region 11.5	terpene	1,366,866	1,389,043	geosmin	Terpene	100%
Region 11.6	RRE-containing	1,499,997	1,521,088	CDA1b / CDA2a / CDA2b / CDA3a / CDA3b / CDA4a / CDA4b	NRP:Ca+-dependent lipopeptide	17%

c00015_NZ_JABJ.. (original name was: NZ_JABJRC010000002.1)

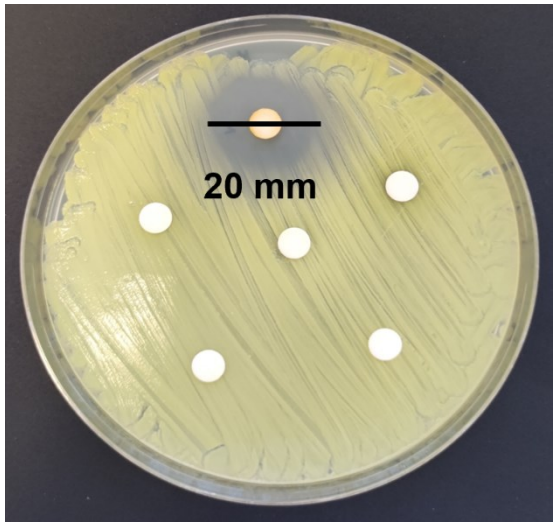
Region	Type	From	To	Most similar known cluster		Similarity
Region 15.1	NRPS-like, NRPS	380,365	485,540	thiocoraline	NRP:Cyclic depsipeptide	34%
Region 15.2	lanthipeptide-class-i	496,456	520,654			
Region 15.3	redox-cofactor	895,339	917,379	lankacidin C	NRP + Polyketide	20%

c00016_NZ_JABJ.. (original name was: NZ_JABJRC010000003.1)

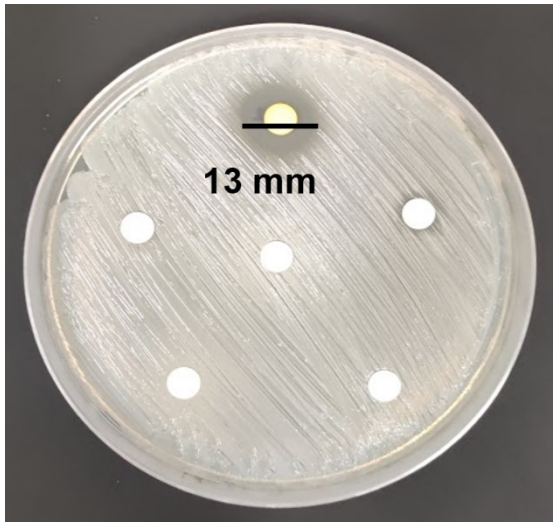
Region	Type	From	To	Most similar known cluster		Similarity
Region 16.1	lanthipeptide-class-iii	3,920	26,505	catenulipeptin	RiPP:Lanthipeptide	60%
Region 16.2	lanthipeptide-class-ii	31,433	54,363			

Příklad analýzy 3 (biosyntetika)

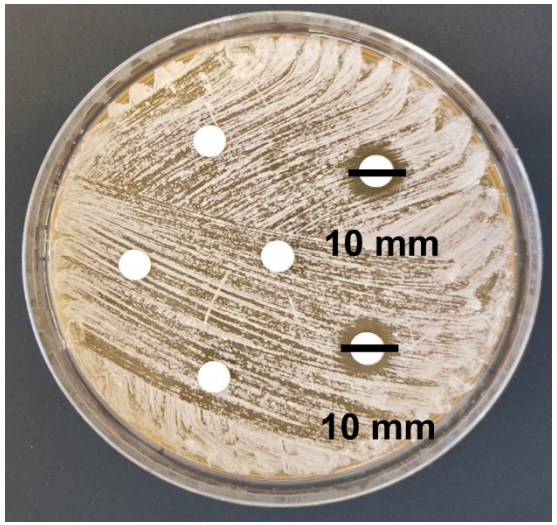
- Vstup: filtrovaná sekvenační data, po kontrole kvality
- **Vstup: fotografie**
- Vstup: chromatografická data



P12377, inhibition of *Micrococcus luteus*



P12377, inhibition of *Staphylococcus aureus*

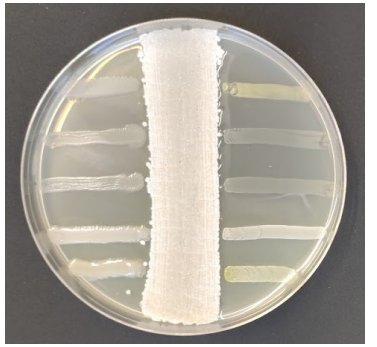


UV100, inhibition of *Saccharomyces cerevisiae*

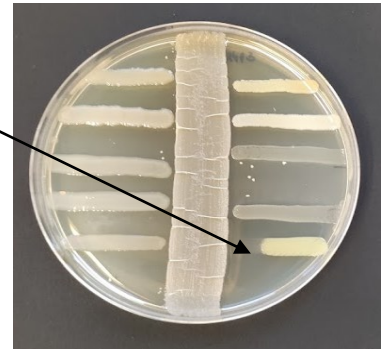
Příklad analýzy 3 (biosyntetika)

- Vstup: filtrovaná sekvenační data, po kvalitě kontroly
- **Vstup: fotografie**
- Vstup: chromatografická data

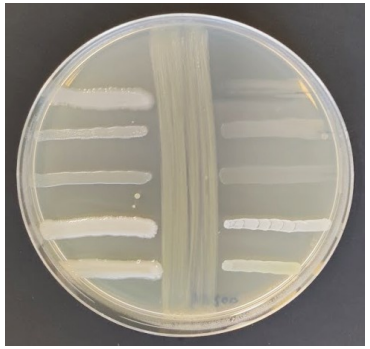
Streptomyces sp. UV100



Streptomyces sp. P12413



Arthrobacter sp. P12200

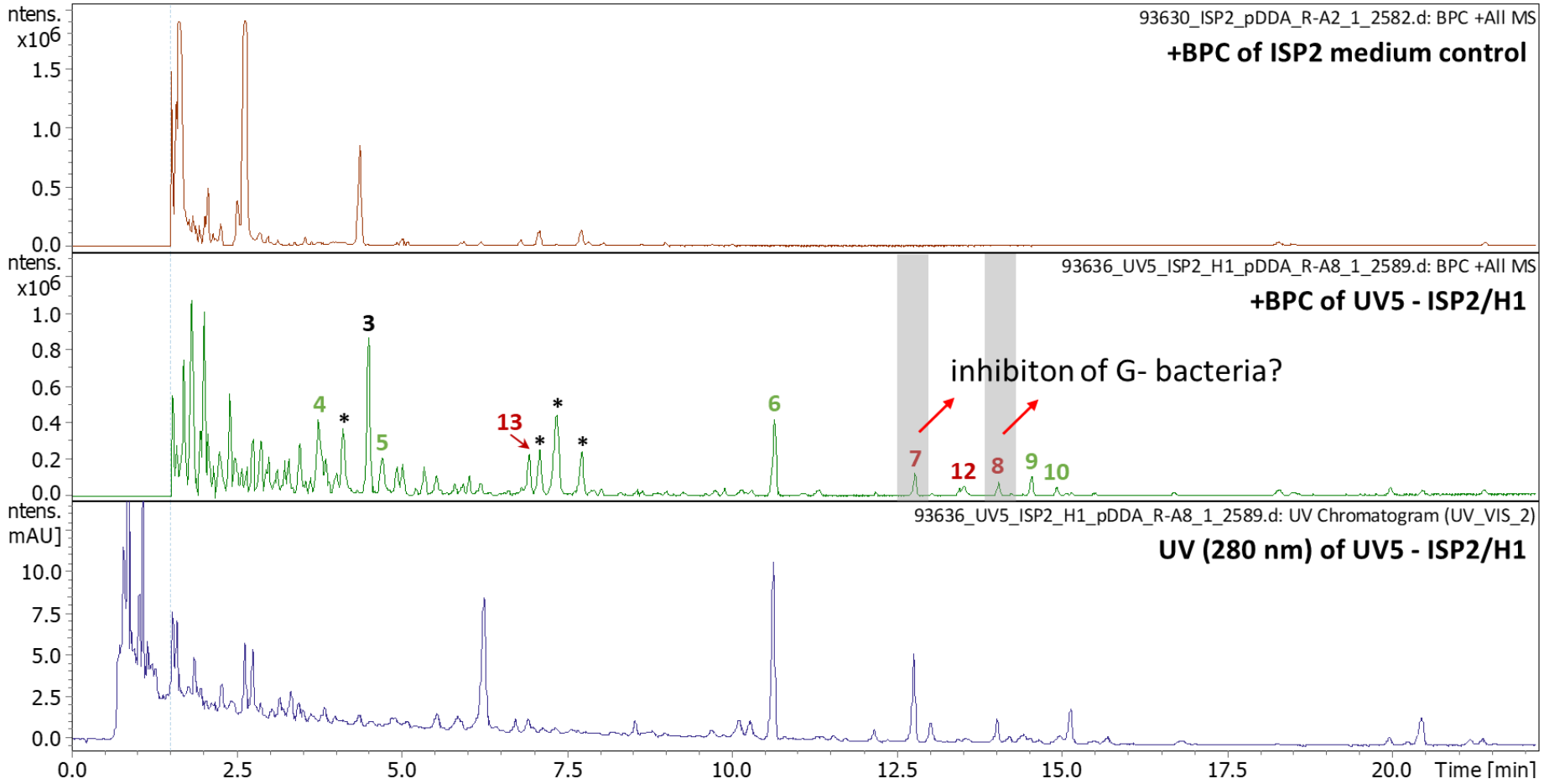


Příklad analýzy 3 (biosyntetika)

- Vstup: filtrovaná sekvenační data, po kvalitě kontroly
- Vstup: fotografie
- **Vstup: chromatografická data**

MUNI

Příklad analýzy 3 (biosyntetika)



MUNI

Archivace dat

- data v tištěné podobě – doba uchování, katalogizace
- data v elektronické podobě
- lokální úložiště
- cloudové úložiště
- uzavřené databáze
- open access databáze

- **ovlivnění grantovými agenturami – podmínky + příklady**

MUNI

Archivace dat – příklad: srovnání univerzit

Česká sbírka mikroorganismů, MU

- katalogizace v tištěné podobě
- katalogizace v digitální podobě, software MINE

- **výhody:**
  - dva systémy (záloha v tištěné podobě)
- **nevýhody:**
  - digitální systém zastaralý, neexistuje IT podpora
  - neexistuje záloha přístupná vzdáleně
  - neexistuje záloha přístupná všem

CMESS, UniVie

- katalogizace/návody primárně digitální
- papírové formy pouze výjimečně (návody k strojům)

- **výhody:**
  - univerzálně dostupném všem
  - pokud aktualizace, okamžitě u všech
  - vzdálený přístup samozřejmostí, pro všechny
  - denní zálohování na 3 serverech
- **nevýhody:**
  - nutný tisk pokud chcete vlastní kopii

MUNI

Archivace dat – online databáze a repozitáře

Ukládaní dat online má v biologii pravidla:

- **NUTNOST** před každou publikací
- minimální doba – 5 let po publikaci
- perzistentní identifikátor (DataCite DOI)
- umožněno review ukládaných datasetů!
- ideálně specifické repozitáře dle povahy dat nebo dle standardu v daném oboru
- open access
  - (výjimka: klinická data mohou vyžadovat „*Data Usage Agreements*“)

MUNI

Archivace dat – příklady

Data – nukleové kyseliny

Data types	Repository options	Data and metadata standards
Raw sequencing data (reads or traces) Genome assemblies Annotated sequences Sample metadata	INSDC repositories Genome Sequence Archive (GSA)	Browse data and metadata standards endorsed by the Genome Standards Consortium
Genetic variation data	dbSNP (human variations less than 50bp) dbVar (human variations greater than 50bp) ClinVar (human genotype & phenotype) European Variation Archive (EVA) (all species) Genome Sequence Archive for Human (GSA-Human)	

Data – proteinové sekvence

UniProtKB	view FAIRsharing entry

Data – funkční genomika

ArrayExpress	view FAIRsharing entry
Gene Expression Omnibus (GEO)	view FAIRsharing entry
GenomeRNAi	view FAIRsharing entry
dbGAP	view FAIRsharing entry
The European Genome-phenome Archive (EGA)	view FAIRsharing entry
Database of Interacting Proteins (DIP)	view FAIRsharing entry
IntAct	view FAIRsharing entry
Japanese Genotype-phenotype Archive (JGA)	view FAIRsharing entry
NCBI PubChem BioAssay	view FAIRsharing entry
Genomic Expression Archive (GEA)	view FAIRsharing entry
GWAS Catalog	view FAIRsharing entry

Data – obrázky

Image Data Resource	view FAIRsharing entry
The Cancer Imaging Archive	view FAIRsharing entry
SICAS Medical Image Repository	view FAIRsharing entry
Coherent X-ray Imaging Data Bank (CXIDB)	view FAIRsharing entry
Cell Image Library	view FAIRsharing entry

nature/repositories

MUNI

Archivace vzorků – mikrobiologie

TAXONOMIE: živé kultury – reprezentující nové druhy → musí byt dostupné

- nejde ale o open access!
- uchovávaní zabezpečují sbírky mikroorganismů → poskytování je placené

BIODIVERZITA/EKOLOGIE: živé kultury

- nejde o nutnost, ale často jsou uložené volně dostupné soubory izolátů
- nejde/jde o open access
- uchovávaní zabezpečují sbírky mikroorganismů / university → placené/neplacené

DSMZ Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH

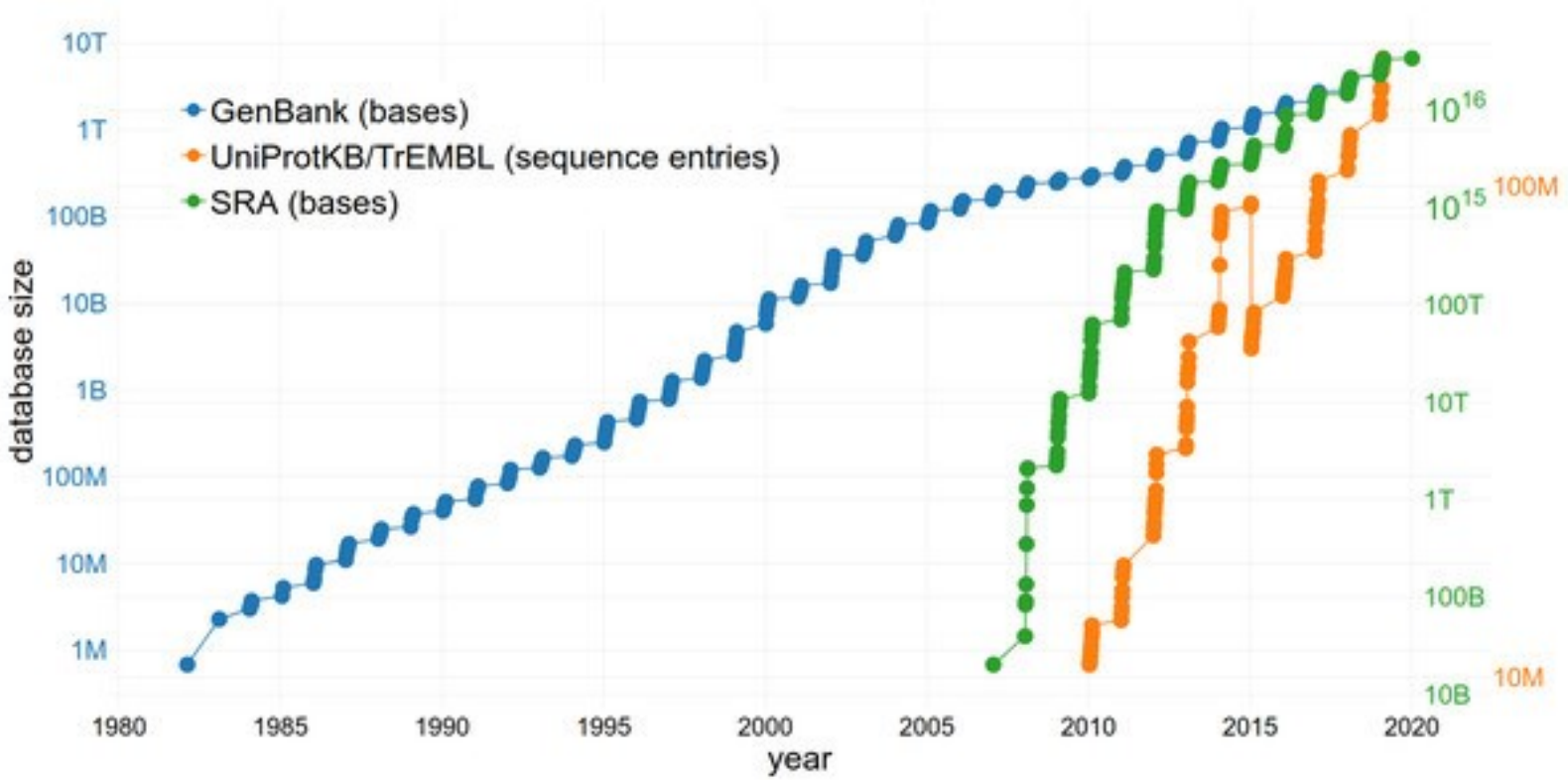
ATCC

MUNI SCI Ústav experimentální biologie Česká sbírka mikroorganismů

DSMZ Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures GmbH

MUNI

Data sharing



Sielemann *et al.*, 2020

MUNI

Data sharing – význam

V biologii nutnost:

- poskytnout **všechna data**
- nebo poskytnou **„minimální dataset“** → replikace výsledků (všechny data z článku, metadata, metody)

  „Data Availability Statement“ při submitování článku
- bez restrikce přístupu – open access
- pokud restrikce → etické/legální → jiný přístup **nutný** (agreement, official request, official permit)

MUNI

Data sharing

V biologii nutnost:

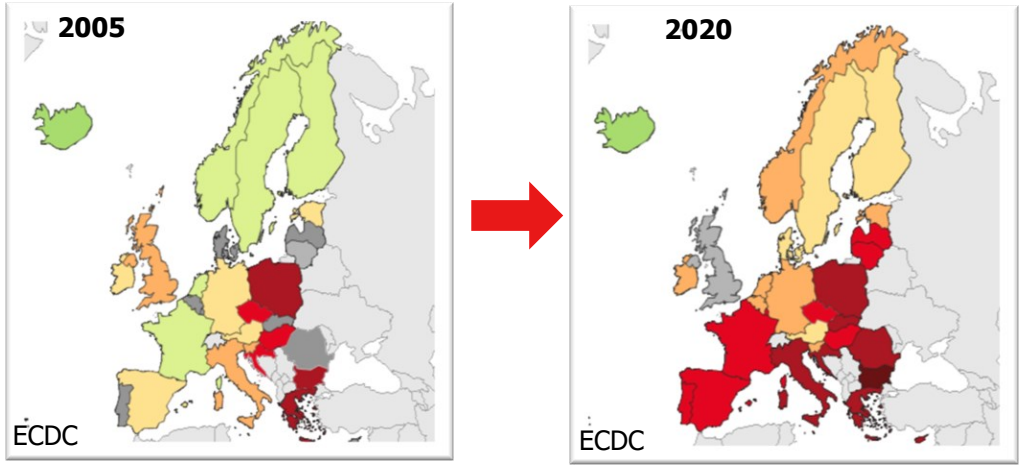
- validace výsledků
- replikace, re-analýza, re-interpretace
- nové analýzy
- reprodukovatelnost
- archivace dat → investice do výzkumu má dlouhodobé výsledky
- citace dat → viditelnost a rozeznatelnost autorů, tvůrců dat, kurátorů

Data sharing – význam i v praxi!

Sledování, reakce, predikce:

- infekční onemocnění (SARS-Cov-19, HAV v Brně, opičí neštovice...)
- šíření rezistentních patogenů (nemocnice, čističky vod...)
- šíření rezistence

Šíření rezistence k cefalosporinům v Evropě – *Klebsiella pneumoniae*



Zrychlení výzkumu:

- infekční onemocnění
- mikrobiom a související onemocnění
- Human Genome Project
  - rakovina
  - vzácná genetická onemocnění
  - farmakogenomika

MUNI

MUNI

Data reusability

- Použití **dříve nasbíraných/zpracovaných** dat za jiným účelem oproti původnímu plánu
- **Podmínka:**
  - vysoká kvalita
  - eticky v pořádku
  - dostatečně popsané – metadata
  - aktuální verze dat
  - uvedené podmínky opakovaného použití
  - **NUTNO OCITOVAT, UVÉST REFERENCI!!!**

Data reusability – význam

- nezávislé studium dat různými skupinami/vědci
- šetření finančních zdrojů
- urychlení výzkumu
- násobné možnosti zpracování a vyhodnocení (jedny data = mnoho výsledků)

- **Nevýhody:**
  - neúplné/nesprávné metadata
  - nízká kvalita dat
  - chybná data

Sequences: text strings describing sequential bases, including gaps in data and their length
(single gene/whole genome sequences, amino acid sequences)

```
>seq1
ATCGTTTAGCTAGACCTGATG
ATCCGATCGATTACGTG
>seq2
GACACGATCGTCAGAAATGCA
GTC
>seq3
ACGACAAATCATCTCC
```

```
>seq1
MYVRANQEFFK
>seq2
WTSMADCHLKV
>seq3
MGPKLHIGRQEFKLIHYWN
NNG
```

Metadata: information about the acquisition, processing and presentation of data
(methodology, sample size and origin)

Graphs: graphical representation indicating relationship
(pathway data, genetic maps, plasmid maps)

Annotations (gene annotation, gene model)

Algorithms and Software (code, bioinformatics pipelines)

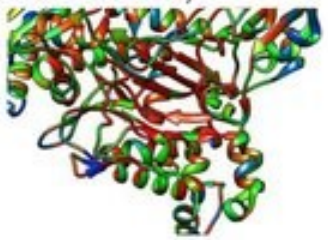
```
for ID in transcripts_per_genes[ gene ]:
    counter = 0
    for element in transcripts[ ID ]:
        if element['type'] == "CDS":
            counter += element['end']-element['start']
    CDS_len_per_transcript.append( { 'id': ID, 'len': counter } )
repr_trans = sorted( CDS_len_per_transcript, key=itemgetter('len') )[-1]
repr_transcripts.update( { repr_trans['id']: transcripts[ repr_trans['id'] ] } )
```

Hypotheses, evidence and prose (data interpretation)

Primary data

Plots and Images: natural, artificial and stylised imagery
(micrographs, radiographs, diagrams, microscopy of biological material, images of geographical regions)

Measurement parameters
(enzyme affinity/speed, chromatography results, mass spectra, geo data/coordinates)



Geometric information/Molecular structure data
(secondary and tertiary protein structures)

Patterns
(sequence motifs, regulatory sequences, expression profiles)

Primary databases: direct submission of experimentally-derived data from researchers; archival database
(DNA, RNA, protein, expression, disease, organism-specific databases)

PDBe
Protein Data Bank in Europe

Publications (text mining)

Meta-analyses

Derived data

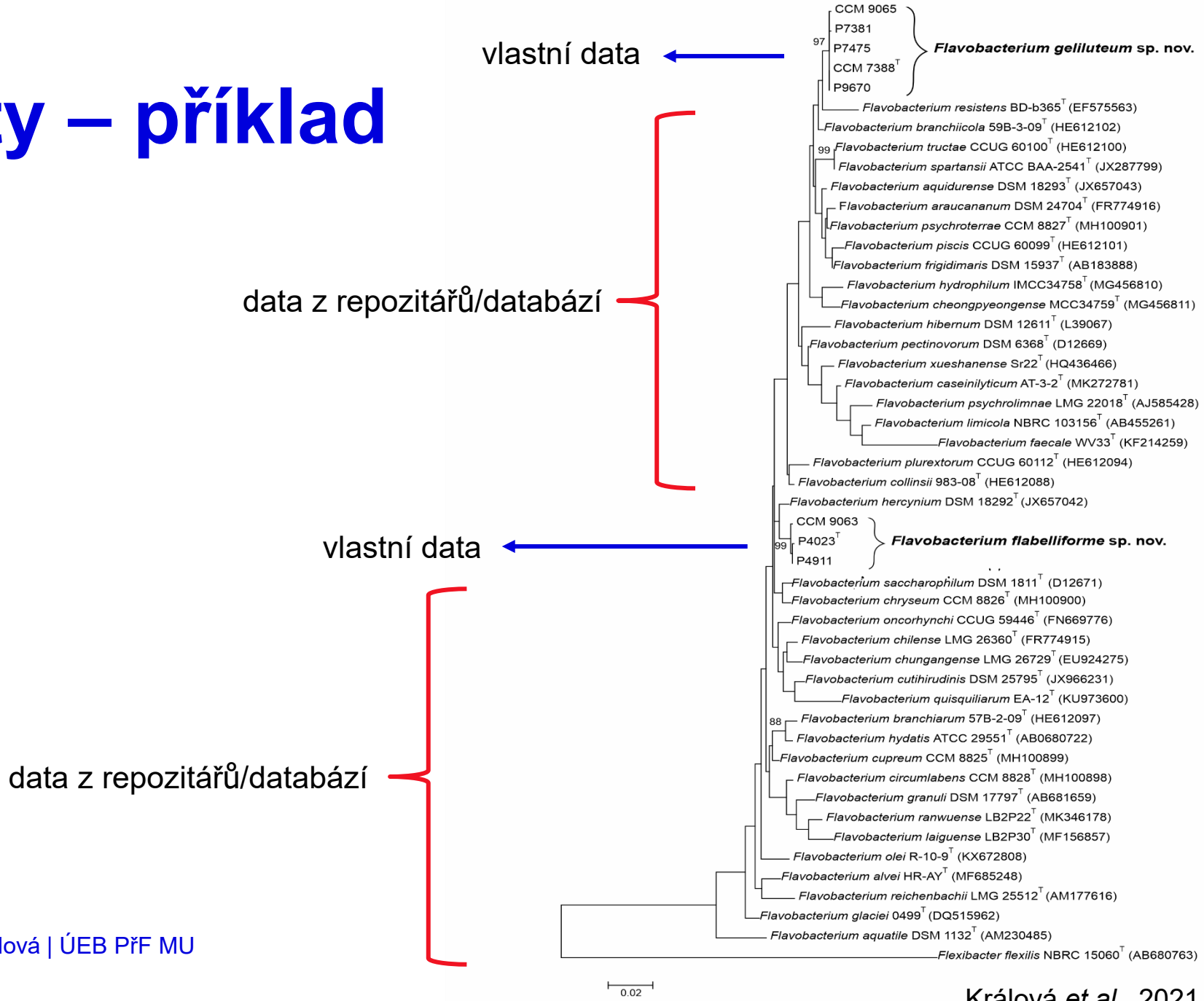
Derived databases: results of analysis, literature search and interpretation, curated database
(DNA, RNA, protein, expression, disease, organism-specific, phenotypic databases)

InterPro

Sielemann *et al*., 2020

Data reusability – příklad

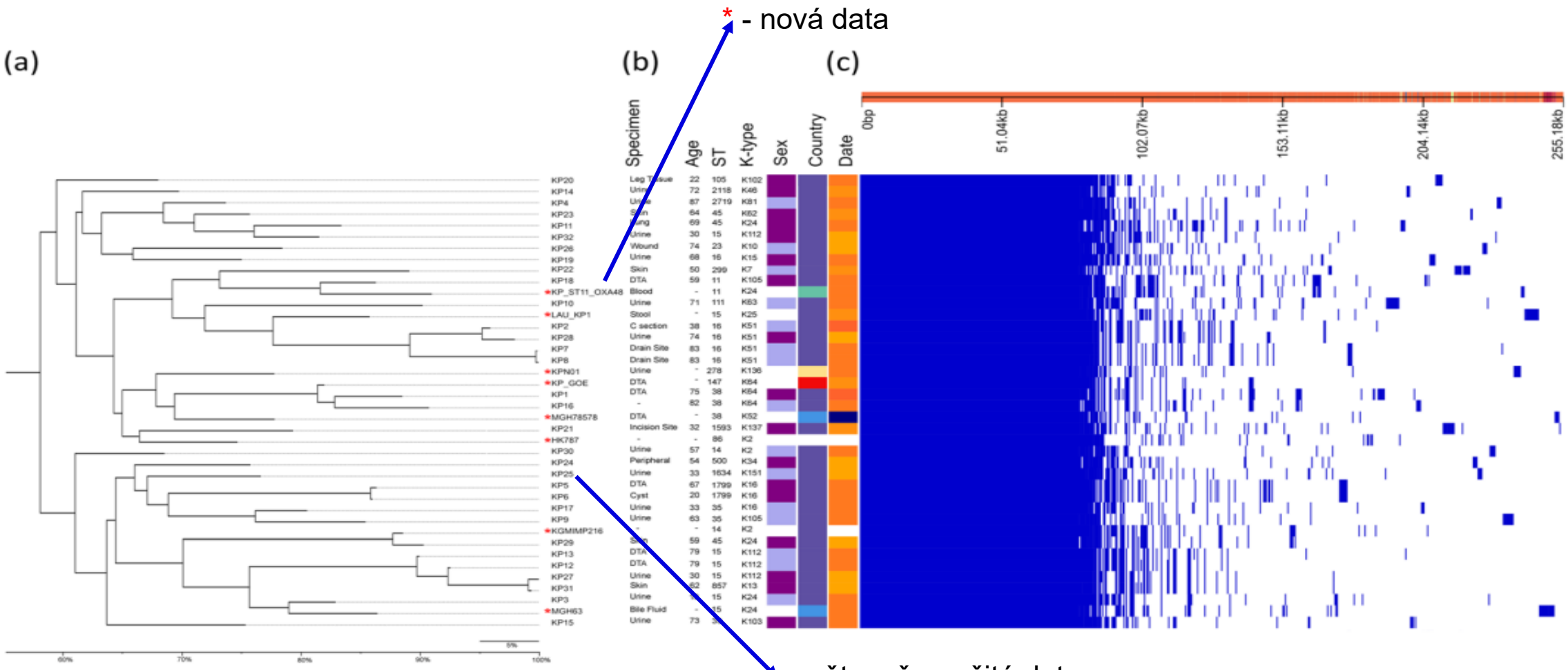
- každá fylogenetická analýza



Králová *et al.*, 2021

Data reusability – příklad

Pangenomická analýza – *Klebsiella pneumoniae*

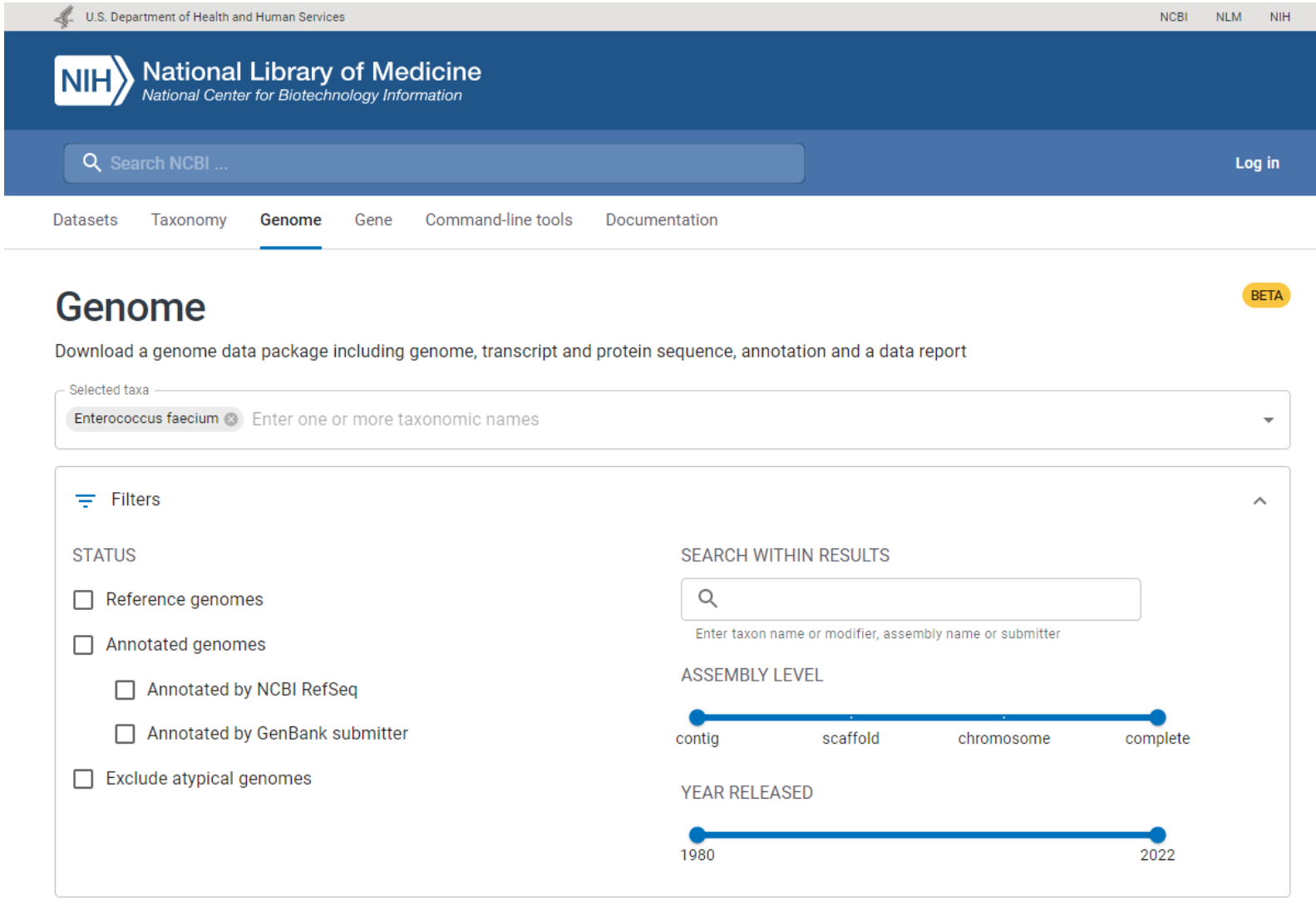


Arabaghian *et al.*, 2019

Data reusability – příklad – klinický výzkum

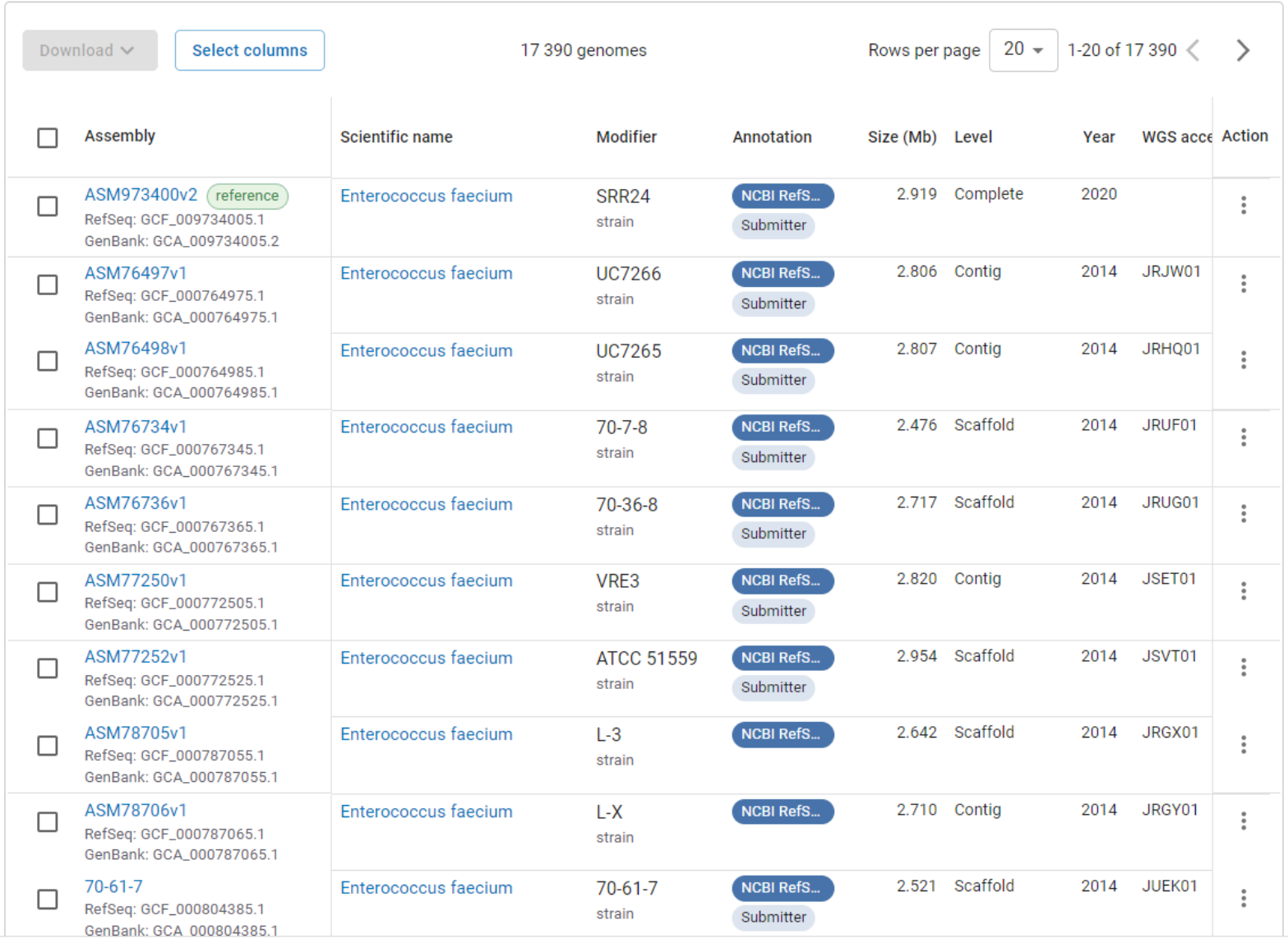
Tvorba epidemiologickych schémat

- využití NCBI databází
- využití genómové subdatabáze
- možnost filtrování výsledků
- možnost stáhnutí celého souboru



Data reusability – příklad – klinický výzkum

Tvorba epidemiologickych schémat



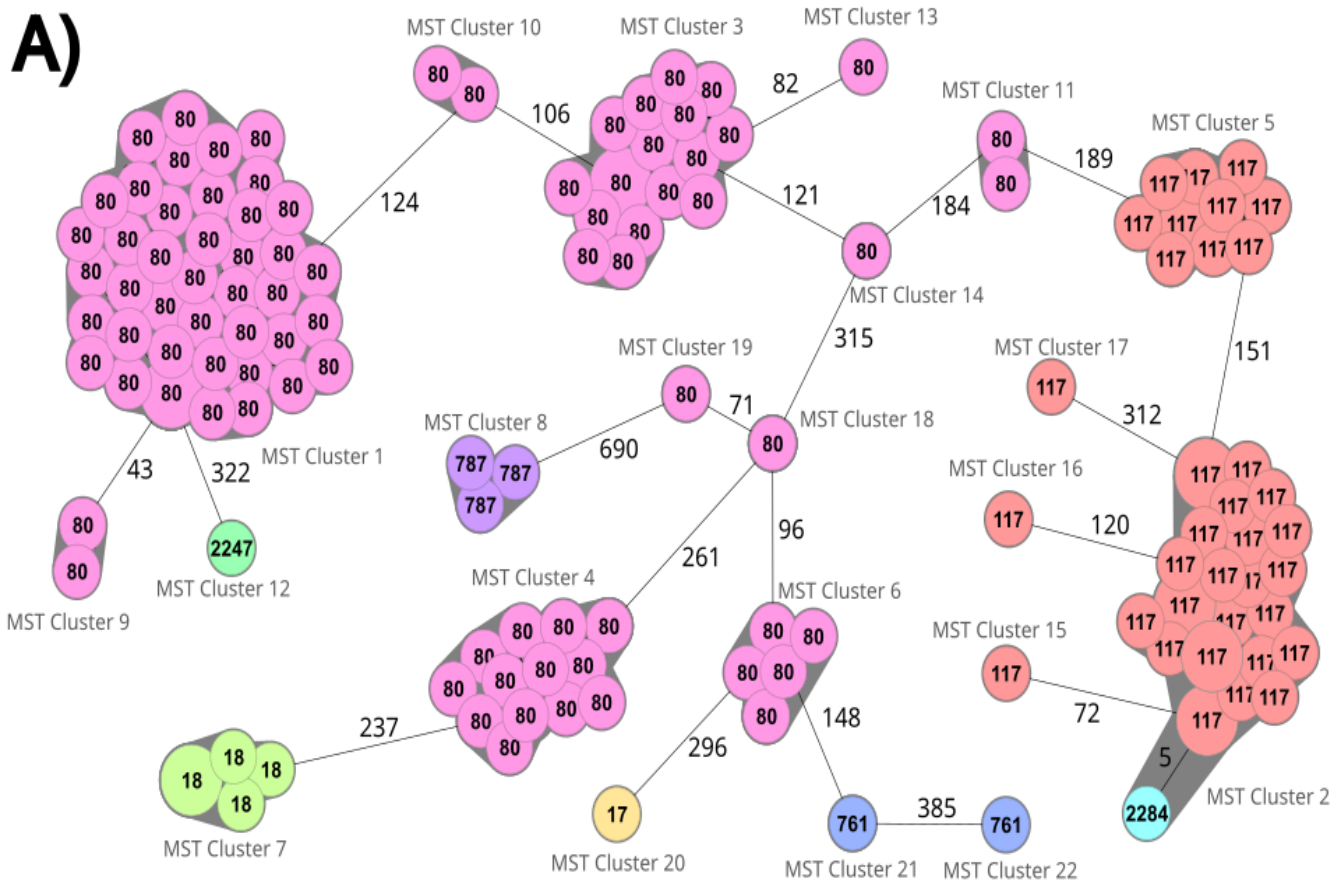
Download | Select columns | 17 390 genomes | Rows per page 20 | 1-20 of 17 390

Assembly	Scientific name	Modifier	Annotation	Size (Mb)	Level	Year	WGS acce	Action
ASM973400v2 reference RefSeq: GCF_009734005.1 GenBank: GCA_009734005.2	Enterococcus faecium	SRR24 strain	NCBI RefS... Submitter	2.919	Complete	2020		
ASM76497v1 RefSeq: GCF_000764975.1 GenBank: GCA_000764975.1	Enterococcus faecium	UC7266 strain	NCBI RefS... Submitter	2.806	Contig	2014	JRJW01	
ASM76498v1 RefSeq: GCF_000764985.1 GenBank: GCA_000764985.1	Enterococcus faecium	UC7265 strain	NCBI RefS... Submitter	2.807	Contig	2014	JRHQ01	
ASM76734v1 RefSeq: GCF_000767345.1 GenBank: GCA_000767345.1	Enterococcus faecium	70-7-8 strain	NCBI RefS... Submitter	2.476	Scaffold	2014	JRUF01	
ASM76736v1 RefSeq: GCF_000767365.1 GenBank: GCA_000767365.1	Enterococcus faecium	70-36-8 strain	NCBI RefS... Submitter	2.717	Scaffold	2014	JRUG01	
ASM77250v1 RefSeq: GCF_000772505.1 GenBank: GCA_000772505.1	Enterococcus faecium	VRE3 strain	NCBI RefS... Submitter	2.820	Contig	2014	JSET01	
ASM77252v1 RefSeq: GCF_000772525.1 GenBank: GCA_000772525.1	Enterococcus faecium	ATCC 51559 strain	NCBI RefS... Submitter	2.954	Scaffold	2014	JSVT01	
ASM78705v1 RefSeq: GCF_000787055.1 GenBank: GCA_000787055.1	Enterococcus faecium	L-3 strain	NCBI RefS...	2.642	Scaffold	2014	JRGX01	
ASM78706v1 RefSeq: GCF_000787065.1 GenBank: GCA_000787065.1	Enterococcus faecium	L-X strain	NCBI RefS...	2.710	Contig	2014	JRGY01	
70-61-7 RefSeq: GCF_000804385.1 GenBank: GCA_000804385.1	Enterococcus faecium	70-61-7 strain	NCBI RefS... Submitter	2.521	Scaffold	2014	JUEK01	

Data reusability – příklad – klinický výzkum

Tvorba epidemiologickych schémat

- příklad využití dat uložených za účelem publikace
- nový účel → epidemiologie



MUNI

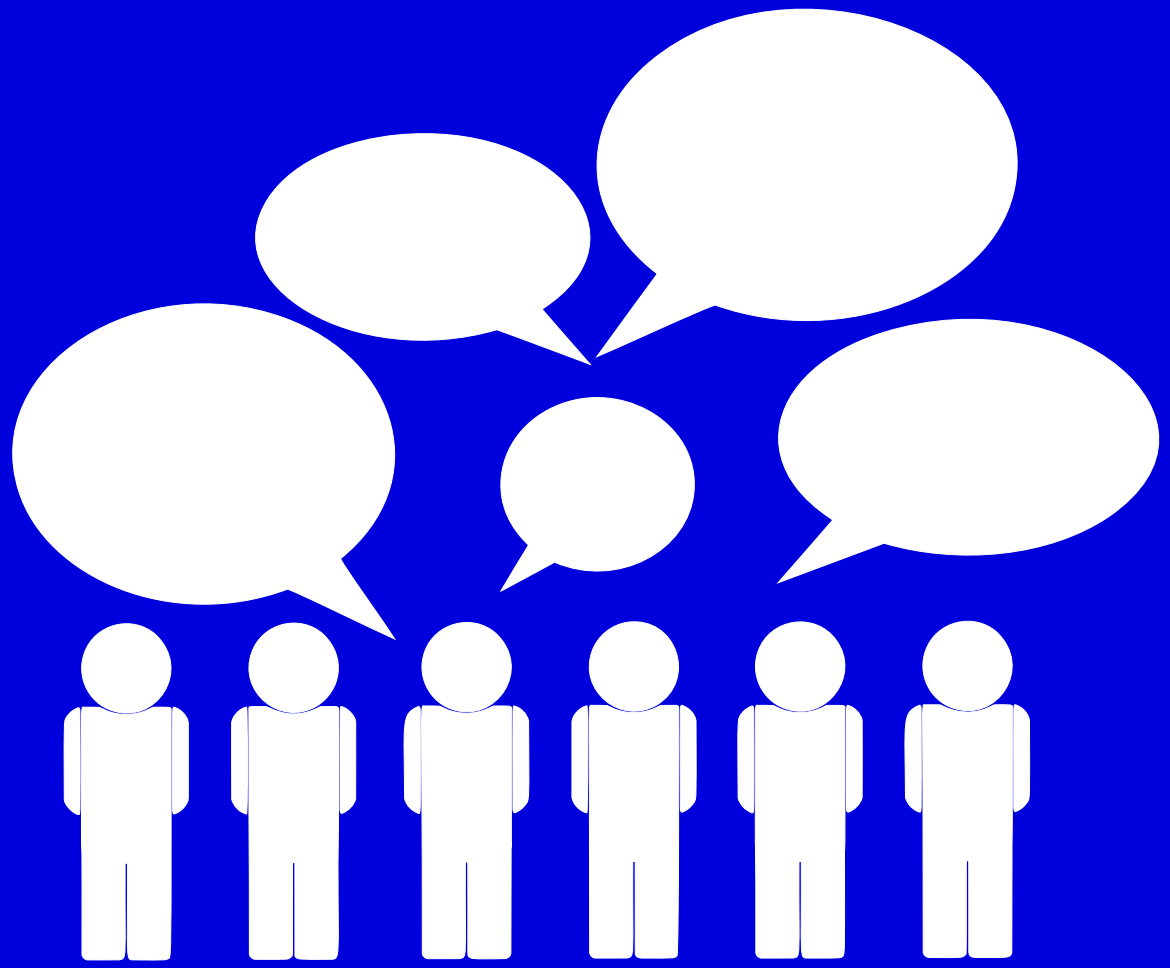
Shrnutí

6. přednáška CORE042

Shrnutí

- Cyklus dat v mikrobiologii/biologii → využíváme všechny kroky
- Generujeme velké množství variabilních dat (fotografie, HPLC, sekvence, měření, popisky, …)
- Generujeme enormní množství metadat (GPS souřadnice, typy materiálů, množství materiálů, časy/doby odběrů, klinická data, asociovaná data, …)
- Vysoká náročnost na **plan** → **collect** → **process** → **analyse**
- Výrazný podíl **sharing + reusability**
- Výrazný tlak na FAIR data (zejména **preserve + share** → **to be reused!**)
  - velké množství databází
  - specifické databáze
  - nevýhoda – databáze bez review procesu a kontroly → chyby

Diskuse



Zdroj: Communicate_communication_conference_2028004 od OpenClipart-Vectors z Pixabay