VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

DASHBOARD PRO HODNOCENÍ VÝSLEDKŮ METATRANSKRIPTOMICKÉ ANALÝZY NÁSTROJI BIOBAKERY
METATRANSCRIPTOMIC EVALUATION DASHBOARD USING BIOBAKERY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — Jan Hýl
AUTHOR

VEDOUCÍ PRÁCE — Ing. Vojtěch Bartoň
ADVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Student: Jan Hýl ***ID:*** 240507

Ročník: 3 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Dashboard pro hodnocení výsledků metatranskriptomické analýzy nástroji bioBakery

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši na téma analýza metatranskriptomu. 2) Popište nástroje obsažené v balíčku bioBakery. 3) Sestavte vhodný testovací dataset a proveďte analýzu vhodným workflow. 4) Pro prezentaci výstupů programů MetaPhlAn a HumaNn vytvořte dashboard ve vhodném programovacím jazyce. 5) Na testovacím datasetu ověřte funkčnost dashboardu a diskutujte vhodnost použitých vizualizačních technik a interpretaci výsledků.
Práce bude vypracována ve spolupráci s pracovištěm RECETOX Masarykovy Univerzity.

DOPORUČENÁ LITERATURA:

[1] BEGHINI, Francesco, Lauren J MCIVER, Aitor BLANCO-MÍGUEZ, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. ELife [online]. 2021, 2021-05-04, 10. ISSN 2050-084X. doi:10.7554/eLife.65088
[2] MEHTA, Subina, Marie CRANE, Emma LEITH, et al. ASaiM-MT: a validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework. F1000Research [online]. 2021, 10 [cit. 2023-09-05]. ISSN 2046-1402. doi:10.12688/f1000research.28608.2

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 29.5.2024

Vedoucí práce: Ing. Vojtěch Bartoň

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

ABSTRAKT

Metatranskriptomická analýza poskytuje informaci o právě exprimovaných genech organismu v době odebrání analyzovaného vzorku. Rozmach zkoumání genů přišel s vývojem sekvenačních technologií a technologickým pokrokem. Tato práce pojednává o nukleových kyselinách a jejich významu v genetice. Zabývá se pojmy metatranskriptom, metagenom a mikrobiom. Popisuje metodu sekvenování od značky Illumina a představuje software vhodný pro metatranskriptomickou analýzu. Následně práce zahrnuje vytvoření testovacího datasetu, vytvoření dashboardu pro vizualizaci dat a vyzkoušení tohoto dashboardu na testovacím datasetu.

KLÍČOVÁ SLOVA

metatranskriptomická analýza, metatranskriptom, metagenom, nukleové kyseliny, HUMAnN, MetaPhlAn, sekvenování, mikrobiom

ABSTRACT

Metatranscriptomic analysis provides information about the currently expressed genes of an organism at the time the sample is collected. The boom in the study of genes came with the development of sequencing technologies and other technological advances. This thesis discusses nucleic acids and their importance in genetics. It explores the concepts of metatranscriptome, metagenome and microbiome. It describes the Illumina sequencing method and introduces software suitable for metatranscriptomic analysis. Subsequently, the work includes the creation of a test dataset, the creation of a dashboard for data visualization, and the testing of this dashboard using the test dataset.

KEYWORDS

metatranscriptomic analysis, metatranscriptome, metagenome, nucleic acids, HUMAnN, MetaPhlAn, sequencing, microbiome

Vysázeno pomocí balíčku `thesis` verze 4.07; `http://latex.feec.vutbr.cz`

HÝL, Jan. *Dashboard pro hodnocení výsledků metatranskriptomické analýzy nástroji BioBakery*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024, 42 s. Bakalářská práce. Vedoucí práce: Ing. Vojtěch Bartoň

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Jan Hýl
VUT ID autora:	240507
Typ práce:	Bakalářská práce
Akademický rok:	2023/24
Téma závěrečné práce:	Dashboard pro hodnocení výsledků meta-transkriptomické analýzy nástroji BioBakery

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu bakalářské práce panu Ing. Vojtěchu Bartoňovi za odborné vedení, časté a naučné konzultace, trpělivost, rychlou a pohotovou komunikaci, věcné a přínosné nápady a rozšíření rozhledu v oblasti bioinformatiky.
Autor/ři děkují výzkumné infrastruktuře RECETOX (č. LM2023069) financované Ministerstvem školství, mládeže a tělovýchovy za podpůrné zázemí. Výpočetní zdroje byly poskytnuty projektem e-INFRA CZ (ID:90254), podporovaným Ministerstvem školství, mládeže a tělovýchovy České republiky.

Obsah

Seznam obrázků

Úvod

Metatranskriptom přináší informace o exprimaci genů člověka, ale také informace o exprimaci genů mikrobiomů v našem těle a ve světě kolem nás. Nesmíme však zapomínat, že současné technologie umožňují zjišťovat metatranskriptom pouze v jednom časovém momentu a za specifických podmínek u konkrétního vzorku. Využívání metatranskriptomu je záležitostí počátku 21. století. Tento směr má před sebou ještě mnohé objevy a jistě se bude nemalou mírou podílet na zlepšení zdraví a kvality života nás všech.

Tato práce popisuje, co se pod slovem metatranskriptom skrývá, jak je možné jej zjistit a k čemu je možné jej využít. Zaměří se blíže na nástroje softwarového balíčku bioBakery 3 a na data analyzovaná pomocí programů z tohoto balíčku (MetaPhlAn 4 a HUMAnN 3). Pro tato data bude vytvořen dashboard s vizualizacemi, kde budou pokud možno, co nejlépe graficky vyobrazeny výsledky analýzy pomocí grafů, tabulek a diagramů a bude poukázáno na rozličné možnosti vizualizace.

Cílem práce je seznámit čtenáře se základními poznatky o genetické informaci a o tom, kde se nachází a jak je s ní nakládáno. Dále jsou čtenáři přiblíženy pojmy metagenom, mikrobiom a jejich souvislost s metatranskriptomem, následně je zaměřena pozornost na metody analýzy metatranskriptomu společně se souvisejícími metodami sekvenování. Jako nástroj pro analýzu je zde představen balíček bioBakery a přiblížena softwarová část analýzy. Na zmíněnou teorii navazuje návrh vizualizačního dashboardu, který má za cíl ukázat popis vizualizačního prostředí, nástroje pro jeho tvorbu a předat čtenáři informace o důležitosti správného vizualizování dat, ale také čtenáře varovat před možnými chybami v procesu vizualizace.

1 Nukleové kyseliny DNA a RNA

Nukleové kyseliny jsou biopolymery tvořené polynukleotidovými řetězci. Jejich funkce spočívá v uchování a přenosu genetické informace. Uplatňují se při proteosyntéze (transkripci a translaci).

Základem nukleotidových kyselin jsou tzv. dusíkaté báze: adenin, guanin (purinové báze), cytosin, thymin a uracil (pyrimidinové báze). Z dusíkatých bází a monosacharidů (ribosa a 2-deoxyribosa) poté vznikají nukleosidy adenosin, guanosin, cytidin, thymidin a uridin. U nukleosidů obsahujících 2-deoxyribosu přidáváme předponu deoxy-. Z nukleosidů jako jejich deriváty vznikají nukleotidy. Na ribosu nebo 2-deoxyribosu se esterovou vazbou váže kyselina fosforečná, nejčastěji v pozici 5' nebo 3'. Obecný název nukleotidů je poté nukleosid-fosfát. U konkrétních názvů se uvádí poloha 5' nebo 3', např. adenosin-5'-fosfát [1].

Báze	**Nukleosid**	**Nukleotid**
Adenin(A)	adenosin	adenosin-5'-monofosfát(AMP)
Guanin (G)	guanosin	guanosin-5'-monofosfát(GMP)
Cytosin (C)	cytidin	cytidin-5'-monofosfát(CMP)
Uracil (U)	uridin	uridin-5'-monofosfát(UMP)
Thymin (T)	thymidin	thymidin-5'-monofosfát(TMP)

Tab. 1.1: Přehled názvů bází nukleových kyselin. Převzato a přeloženo z [2].

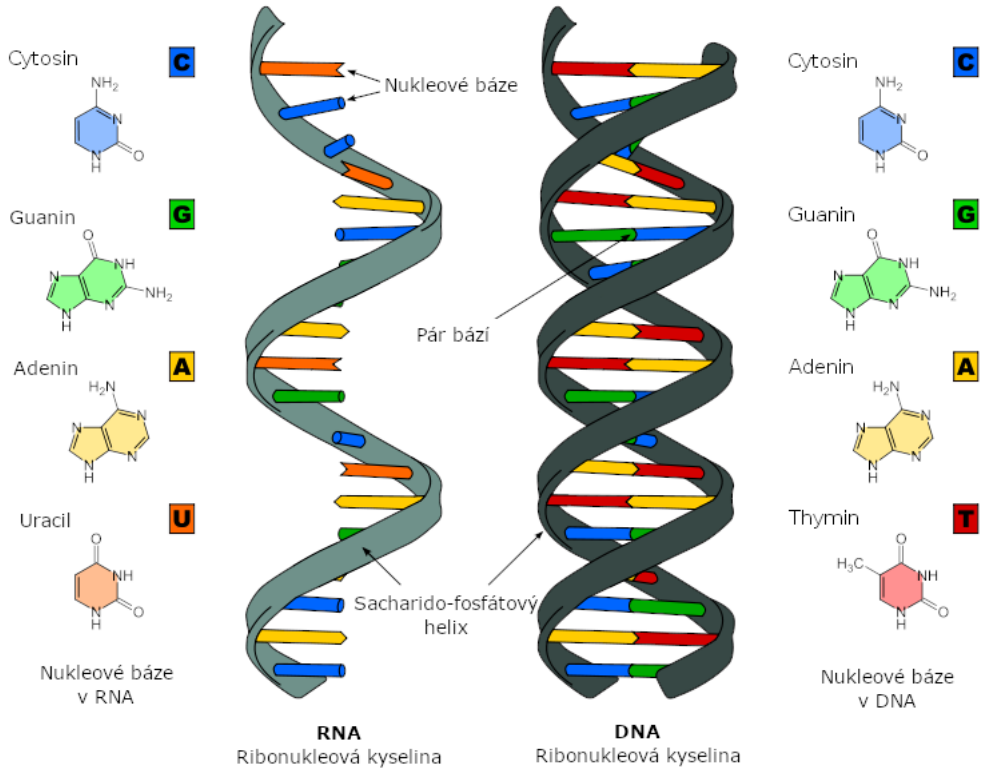
Do nukleotidových řetězců jsou jednotlivé nukleotidy spojeny prostřednictvím 3', 5'-fosfodiesterových vazeb (na 3' -OH skupinu nukleotidu se váže 5'-fosfátová skupina následujícího nukleotidu). Řetězce všech nukleových kyselin jsou uspořádány do pravidelně se opakující sekvence fosfát-pentosa, k nimž jsou navázány dusíkaté báze. Řetězce vždy končí na jedné straně 5'-koncem fosfátové skupiny a na druhé straně 3'-konce hydroxidové skupiny. To dává řetězcům polaritu a určuje, kterým směrem se bude číst genetická informace uložená v řetězci. 5'-konec má záporný náboj a 3' konec má kladný náboj. Jednotlivé nukleotidy jsou vždy čteny od 5'-konce ke 3'-konci.

Podle povahy sacharidové složky rozlišujeme deoxyribonukleové kyseliny (DNA), v nichž je obsažena 2-deoxyribosa, a ribonukleové kyseliny (RNA), v nichž je obsažena ribosa. V DNA se nacházejí purinové báze adenin, guanin a pyrimidinové báze thymin a cytosin. V RNA je thymin nahrazen uracilem [1].

1.1 Deoxyribonukleová kyselina - DNA

Deoxyribonukleovou kyselinu (DNA) si můžeme představit jako dvě vlákna nukleotidových řetězců, které se k sobě přes jednotlivé báze spojují vodíkovými můstky. Spojení však v prostoru nevypadá jako dva lineární řetězce spojené k sobě v jedné rovině, ale jako pravotočivá dvoušroubovice.

Spojování bází má v DNA svá pravidla, která nazýváme princip komplementárního párování nukleotidů. Pár mohou vždy utvořit pouze adenin s thyminem (A-T) a cytosin s guaninem (G-C). Vazba A-T je označována jako slabá vazba, protože je spojena pouze dvěma vodíkovými můstky a vazba G-C je označována jako silná vazba, protože je tvořena třemi vodíkovými můstky. Takto utvořená DNA je komplementární a můžeme z jednoho vlákna odvodit vlákno druhé, aniž bychom ho předem znali. DNA je dále antiparalelní, což znamená, že na jednom konci má jedno vlákno 3'-konec a druhé vlákno 5'-konec. Výše popsané je znázorněno v obrázku 1.1. DNA samotná slouží především pro uchování genetické informace [2].



Obr. 1.1: Sekundární struktura DNA a RNA. Převzato a přeloženo z [3].

1.2 Ribonukleová kyselina - RNA

Ribonukleová kyselina (RNA) je jednovláknový biopolymer, který je nejčastěji využívám na přenos genů z jádra buňky (transkripce) a jejich následné zpracování a vytvoření proteinů (translace). RNA je často kratší než DNA a nese informaci pouze o jednom či dvou genech. V RNA se místo dusíkaté báze thyminu nachází uracil. Princip komplementarity zůstává zachován, tudíž místo thyminu tvoří pár s adeninem uracil. RNA vždy vzniká transkripcí (viz Kapitola 1.3) a dělí se na několik druhů [1, 2]. Základními druhy jsou:

- **mRNA** - messengerová RNA, která dopravuje genetickou informaci z jádra do ribozomů. Tímto typem RNA se zabývá transkriptomická analýza.

- **rRNA** - ribozomální RNA tvoří dvě třetiny ribosomu a je zodpovědná za jeho funkci. Hraje důležitou roli při vytváření bílkovin. Je málo variabilní, a proto je využívána pro určování taxonomického původu organismu [2].

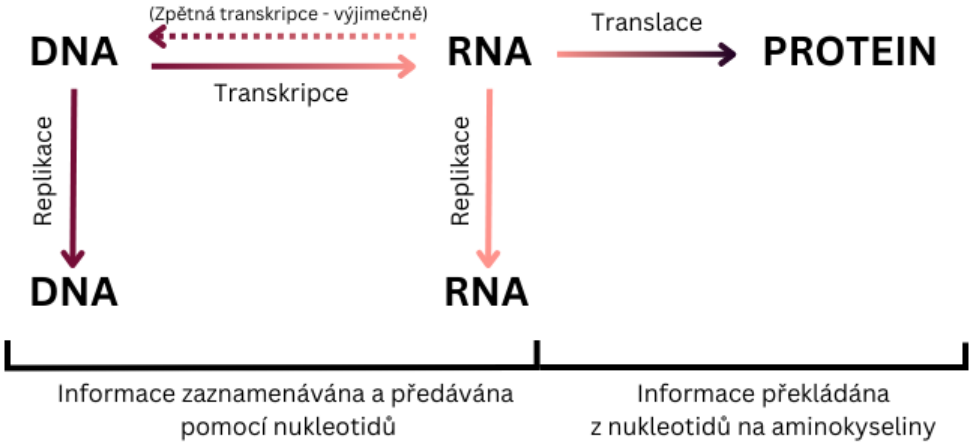
- **tRNA** - transferová RNA se podílí na syntéze bílkovin připojováním jednotlivých aminokyselin do polypeptidového řetězce [1].

1.3 Exprimování genetické informace

Správná exprese genetické informace je důležitá pro chod celého organismu. Fungování organismu zabezpečují bílkoviny. Aby mohla být vytvořena nová bílkovina, musí nejdříve proběhnout procesy zvané transkripce a translace. Při transkripci dochází ke čtení DNA a vytváření kopie genů, které jsou přenášeny v podobě RNA do ribozomů, kde jsou následně přeloženy (translace) na řetězec aminokyselin a upraveny na funkční bílkovinu. Dále bude podrobně popsán pouze proces transkripce, jelikož je stěžejní pro pochopení významu analýzy (meta)transkriptomu.

1.3.1 Centrální dogma molekulární biologie

Centrální dogma molekulární biologie udává pravidla a zastřešuje proces replikování a exprimování genetické informace. Je základním stavebním kamenem molekulární biologie. Genetická informace se může dle schématu 1.2 posouvat jen určitými směry [2]. Hrozbou pro fungování exprimace genetické informace a centrálního dogma je oxidace, konkrétně reaktivní kyslíkaté molekuly. Ty mohou poškozovat celý proces a narušovat strukturu nukleových kyselin. Kvůli tomu vznikají různá neurologická onemocnění jako je např. Alzheimerova choroba [4].



Obr. 1.2: Centrální dogma molekulární biologie.

1.3.2 Transkripce

Při transkripci genu z DNA je využíváno pouze jedno vlákno DNA, které se nazývá templátové vlákno. Proces zahajuje RNA-polymerasa a transkripční faktory, které se vážou na templátové vlákno DNA v místě před strukturním genem zvaném promotor. Následně dochází k syntéze nové RNA na základě principu komplementarity bází [1].

Nově vytvořená RNA se nazývá primární transkript. Primární transkript bývá často výsledná mRNA, která je připravena na translaci, ale není to vždy pravidlem. U člověka vzniká jako primární transkript tzv. pre-mRNA (prekurzorová mRNA). Pre-mRNA obsahuje úseky DNA, které kódují funkční geny (exony) a úseky DNA, které nenesou genovou informaci a nacházejí se mezi jednotlivými geny (introny). Pro vznik mRNA musí pre-mRNA projít úpravami, jednou z nich je například splicing (vystřižení intronů). Poté z pre-mRNA vzniká výsledná mRNA a putuje dále do ribozomů.

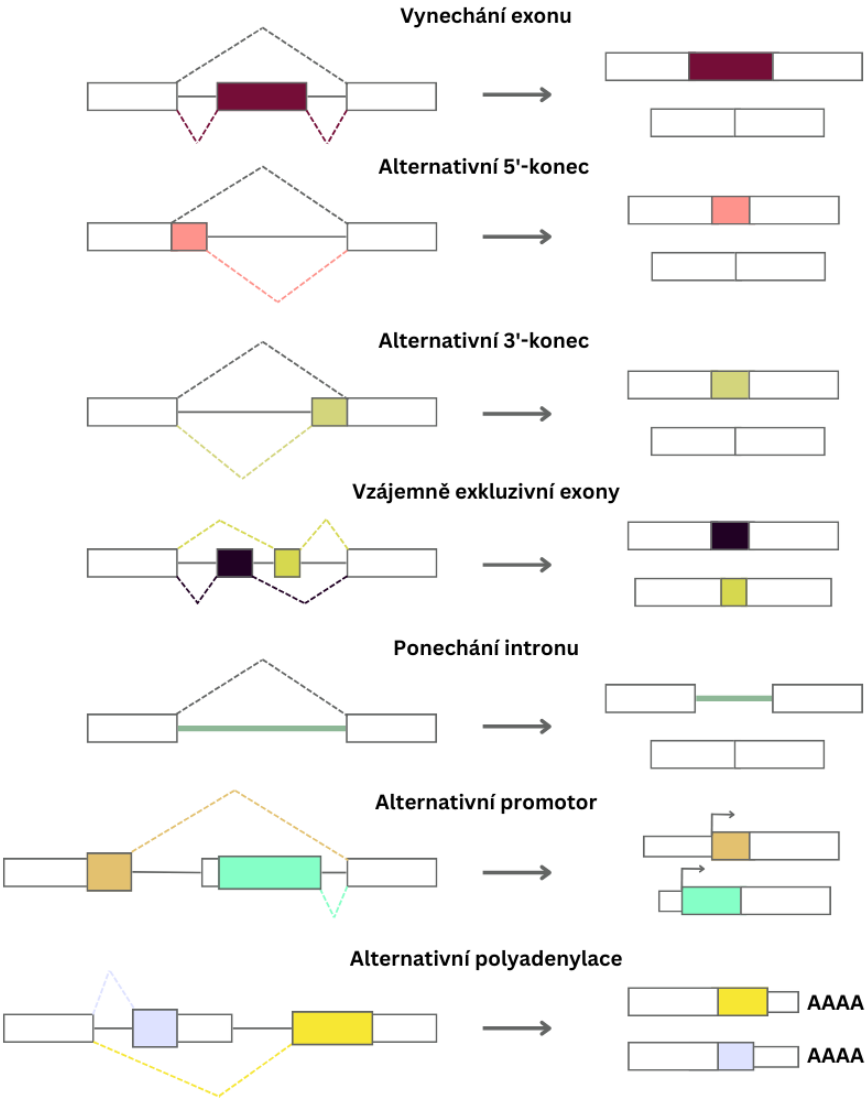
Analýza transkriptomu se zabývá právě zmíněnou pre-mRNA a mRNA a zkoumá jaké geny jsou exprimovány a v jaké míře jsou geny exprimovány. Bylo zjištěno, že významnou roli při exprimaci genů hrají právě introny, i když nenesou užitečnou genovou informaci [2].

1.3.3 Alternativní splicing

Konstitutivní (standardní) splicing je proces vystřihování intronů a spojování exonů v pořadí, v jakém se v genu vyskytují. U alternativního splicingu dochází k různým odchylkám od běžného splicingu. Různé faktory (např. spliceozomy) ovlivňují splicing, což vede k různým formám výsledné mRNA. Již tedy neplatí pravidlo jeden

gen – jedna RNA – jeden protein. Z jednoho genu může vzniknout více proteinů, které mohou mít podobné funkce. Dochází tím ke zvýšení diverzity exprimovaných proteinů.

Četné studie potvrdily stěžejní roli alternativního splicingu v biologických systémech. Alternativní splicing se objevuje především u eukaryot(Čím komplexnější eukaryota, tím vyšší podíl alternativního splicingu.), kde se podílí na rozmanitých biologických procesech, které nás provázejí od narození až po smrt. Příkladem může být vývoj jednoduchých ale i komplexních tkání jako jsou tkáně mozku, varlat či imunitního systému. Alternativní splicing můžeme tedy považovat za jeden z ústředních prvků genové exprese. Nejčastější varianty alternativního splicingu jsou graficky znázorněny v obrázku 1.3 [5].



Obr. 1.3: Varianty alternativního splicingu - převzato a přeloženo z [6].

2 Mikrobiom a metagenom lidského těla

Metagenomika (metagenomická analýza) se zabývá studiem celých biologických komunit v daném prostředí. Metagenom zachycuje kompletní genetickou informaci všech organismů ve vzorku, což znamená, že v sobě skrývá také informace o metatranskriptomu. Díky metagenomu víme, jakou mají organismy genetickou výbavu a při studiu metatranskriptomu zjišťujeme, do jaké míry tuto genetickou výbavu využívají [2].

2.1 Propojení metagenomu a metatranskriptomu

Metagenomická analýza je úzce spjatá s analýzou metatranskriptomu. Nejprve je nutné určit jaké organismy zkoumáme a až poté můžeme začít zjišťovat, jaké konkrétní geny jsou exprimovány. Z metagenomických dat lze za pomoci softwaru a referenčních databází určit metatranskriptom. Nelze se však spolehnout na jeho úplnou přesnost, protože pracujeme pouze s odhadem, který byl proveden na základě známých dat. Je tedy dobré mít na paměti, že se v tomto případě nedíváme na skutečný metatranskriptom. Pro zjištění přesnějšího metatranskriptomu je stále nutné provést metatranskriptomickou analýzu z pre-mRNA a mRNA [2].

2.2 Význam mikrobiomu nejen v lidském těle

Nejčastěji studujeme tzv. mikrobiom, což je soubor mikroorganismů v daném prostředí. Jedná se například o mikrobiom živočichů, ale také o mikrobiom na dně oceánů, či pod zemským povrchem.

Studium metagenomu pomáhá odhalit, co za organismy se nachází v mikrobiomu, jaká je jejich diverzita, ale také jaké nové proteiny, enzymy či biochemické cesty mohou existovat. Využívá se například již předem známých genů a genetické informace k určení přítomnosti daného organismu. Metagenomikou zkoumáme například bakterie žijící ve znečištěném prostředí a pozorujeme jejich přizpůsobivost. Ze zjištěných informacích lze poté vyčíst jaké enzymy bakterie produkují a jak se vypořádávají se znečištěním. Lidé pak mohou tyto znalosti využít ve svůj prospěch a zvrátit napáchané škody na životním prostředí nebo pomoci léčit nemocné pacienty.

Pro tuto práci bude podstatný mikrobiom lidského těla. Nejvýznamnějšími mikrobiomy jsou:

- orální (ústní) mikrobiom,
- kožní mikrobiom,
- vaginální mikrobiom,
- střevní mikrobiom - několikanásobně větší než předchozí jmenované.

Bylo zjištěno, že změna v mikrobiomu může být využita jako biomarker pro určení zdravotního stavu. Není však známo, jak přesně změna mikrobiomu ovlivňuje náš zdravotní stav. Poodhalit toto tajemství nám pomáhá právě meta-omická analýza (viz Kapitola 3) a můžeme tak využívat nové moderní techniky a nových poznatků k zlepšení kvality života lidstva [2].

3 Analýza metatranskriptomu

3.1 Základní poznatky

Metatranskriptomická analýza (též *metatranskriptomika*) je jednou ze základních metod zjišťování informací o společenstvu organismů. Mezi další metody patří:

- Metagenomická analýza - pro zjištění taxonomie a abundance jednotlivých mikroorganismů ve vzorku [7].
- Metaproteomická analýza - zjišťuje výskyt proteinů a dění kolem nich [8].
- Metabolomická analýza - zaměřuje se na metabolity produkované organismem [9].

Všechny tyto analýzy jsou navzájem provázány a dohromady vytvářejí komplexní obraz jednotlivých metagenomů. Souhrnně se dají označit jako meta-omická analýza [8]. Předpona meta- (řecky: *"nesrovnatelný"*) u jednotlivých analýz znamená, že se jedná o analýzu, která popisuje více (mikro)organismů v jednom vzorku. Zahrnuje tedy například více genomů či transkriptomů. Toto neplatí u metabolomické analýzy, kde je název odvozen od slova metabolit [7]. Metabolomická analýza je složitější na provedení než předchozí zmíněné analýzy. Pracuje totiž s nízkomolekulárními objekty s odlišnými fyzikálními vlastnostmi [9].

3.2 Využití metatranskriptomu

Zjišťování metatranskriptomu by nebylo možné bez vyvíjejících se technologií, zrychlování zpracování velkého množství dat a vývoje sekvenačních metod. Jednou z prvních analýz metatraskriptomu byly analýzy planktonu, mikrobiomů v moři a na mořském dně. Různé studie mapují změny v oceánech a odhalují informace o dosud nepoznaných organismech. Půda je dalším prostředím, na které vědci upřeli svůj zrak. V půdě se nachází nesčetně organismů a je jako taková velmi důležitá pro život na Zemi [10].

Hlavním zájmem je ale v tomto případě lidské tělo, jeho části, orgány, vznikající patologie a lidský mikrobiom. Pomocí metatranskriptomiky můžeme zkoumat vliv mikrobů na člověka a zjišťovat, jaké procesy probíhají při určitých chorobách. Mezi detailně zkoumané mikrobiomy patří střevní mikrobiom, ale i prostředí od úst až po žaludek [7, 11].

Velkou kategorii tvoří také nádorová onemocnění a rakovina obecně. Zvýšený výskyt rakoviny a karcinogenních látek trápí lidstvo čím dál více. Bezpečnou a univerzální léčbu je ale těžké najít a to zejména kvůli velké různorodosti nádorů a velké komplexnosti těchto onemocnění. Metatranskriptomika nám může pomoci odhalit, co se v daný okamžik s nádorem děje a jaké jeho složky jsou aktivní. To by mohlo

přinést další informace potřebné k efektivní léčbě. Analýzu právě funkčních genů lze také skloubit s různými zobrazovacími metodami a zacílit exprimaci konkrétního genu na konkrétní místo u nádoru. Jedním z možných využití je například získání informace, zda bude nádor metastázovat a do kterých míst to bude [6, 12].

3.3 Výzvy analýzy metatranskriptomu

Při výběru sekvenační technologie momentálně dominují sekvenátory pro krátká čtení a metoda shotgun sekvenování. Zejména proto, že pojmou velké množství dat za nízkou provozní cenu. Sekvenační technologie se ale stále zlepšují a zefektivňují a mohli by v budoucnu přinést větší přesnost při čtení a lepší výsledky. Obecně platí, že je potřeba zlepšovat a zrychlovat technologie, aby bylo možné získávat lepší a lepší výsledky.

Zlepšovat se musí nejen technologie, ale také sbírání a zpracování dat. Při analýzách chybí referenční a předem známá data, která by nám pomohla a usnadnila zpracování vzorků. Bude tedy nutné rozšiřovat databáze a sbírat další a další data. Není však důležité data jen sbírat, ale také sjednotit formát, ve kterém budou uložena. Protože pokud chceme data sdílet a chceme mít co nejvíce dostupných dat, musíme je zapisovat jedním stylem. V opačném případě dochází ke zpomalení analýz a získávání cenných informací [10].

4 Metody získání dat pro analýzu metatranskriptomu

Momentálně nejpřesnější metodou analýzy metatranskriptomu daného mikrobiomu v určitém časovém úseku a za určitých podmínek je RNA sekvenování (RNASeq). Pomocí této metody můžeme zachytit transkriptomy, geny s nízkou expresí, ale i nekódující části RNA.

Vzhledem ke složitosti mikrobiomu se pro metatranskriptomické studie nejčastěji používá sekvenování ve formě krátkých čtení obvykle generovaných sekvenační technologií Illumina, zejména pokud požadujeme vícevzorkovou analýzu a dobré pokrytí. Je však těžké správně určit parametry sekvenování a to kvůli nedostatku předem známých informací o vzorcích [10]. Metatranskriptomická analýza se tedy téměř vždy provádí společně s dalšími analýzami zmíněnými v kapitole 3.1, zejména společně s metagenomickou profilací.

4.1 Získání RNA-Seq ze vzorků

Pokud chceme zjistit jaká genetická informace je v daném vzorku exprimována, musíme izolovat pouze RNA molekuly, které chceme zkoumat. Pro izolování RNA sekvencí jsou dnes dostupné kity od různých společností. Každý kit má svůj standardní operační protokol (SOP), podle kterého se postupuje. Kity se mohou lišit v postupu a v použitých chemikáliích, základní idea nicméně zůstává stejná. Nejprve je potřeba ze vzorku odstranit veškerou DNA a ponechat pouze RNA. Dále musíme odstranit ribozomální RNA, která nepředstavuje právě exprimované geny. Konečnou úpravou před vytvořením knihovny a samotným sekvenováním je přetvoření RNA sekvencí na cDNA (*z anglického: complementary DNA (komplementární DNA)* sekvence. cDNA vzniká za použití reverzní transkriptázy, přičemž dochází k syntetizování komplementárního vlákna k vláknu RNA. Tento krok je nutné provést, abychom mohli využít technologie pro DNA sekvenování [13, 14].

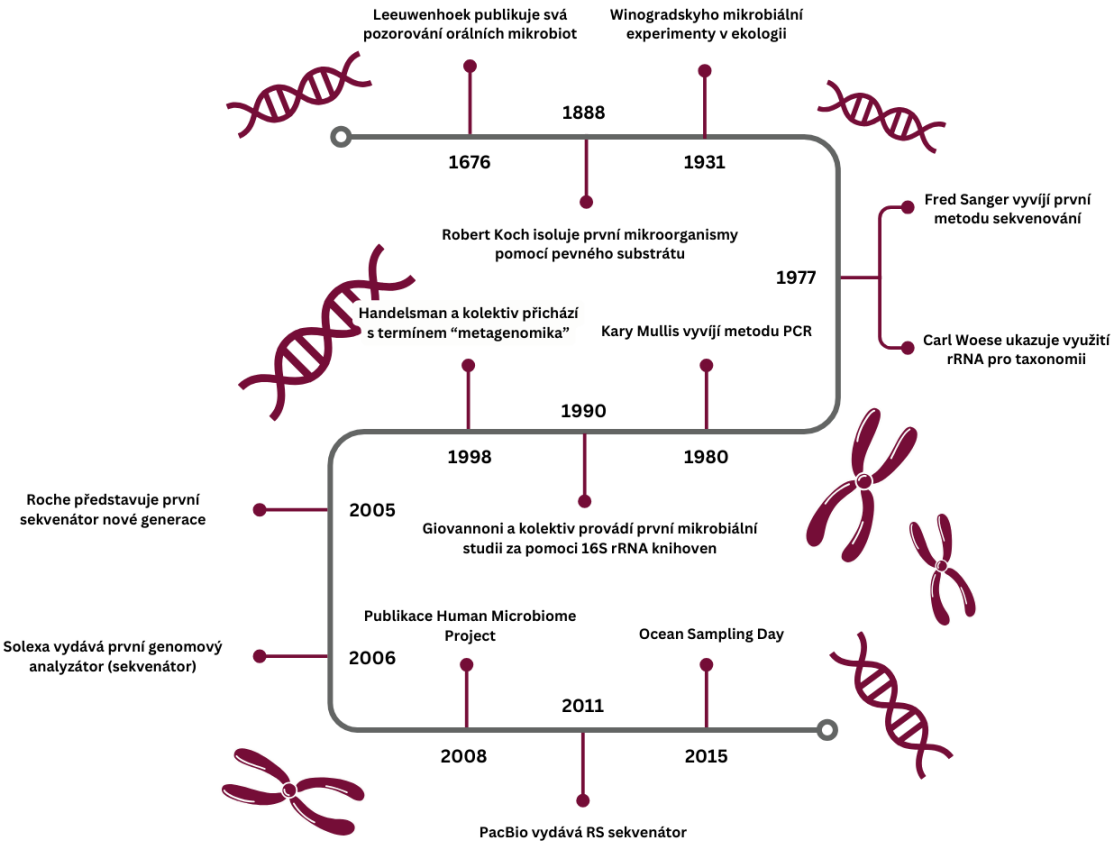
4.1.1 Technologie sekvenování DNA

Existuje několik technologií pro sekvenování DNA. Nejznámější technologie, z nichž se většina stále používá, jsou tyto:

- Maxam-Gilbert (již se nepoužívá)
- Sanger
- Roche 454 pyrosequencing (již se nepoužívá)
- Illumina

- SOLiD
- Ion Torrent
- Pacific Bioscience
- Oxford Nanopore

V těchto technologií se využívá buďto metoda čtení krátkých sekvencí (např. Illumina) nebo metoda čtení dlouhých sekvencí (např. Oxford Nanopore). První metoda je rychlá a velmi rozšířená. Druhá metoda je přesnější a dokáže lépe určit předem neznámé nukleotidové sekvence. V této práci bude zmíněna především sekvenační technologie Illumina, která je momentálně nejvyužívanější technologií pro transkriptomickou analýzu. Illumina se zaměřuje především na krátká čtení, ale vyvíjí také technologie pro dlouhá čtení. Podrobnější popis přípravy knihovny a procesu sekvenování je v kapitole 4.1.2 [15]. Přehled stručné historie s významnými body genomiky a sekvenování je možné vidět na časové mapě v obrázku 4.1.



Obr. 4.1: Stručný přehled historie sekvenování a genomiky. Důležité objevy pro vývoj sekvenování.

4.1.2 Sekvenování nové generace - Illumina

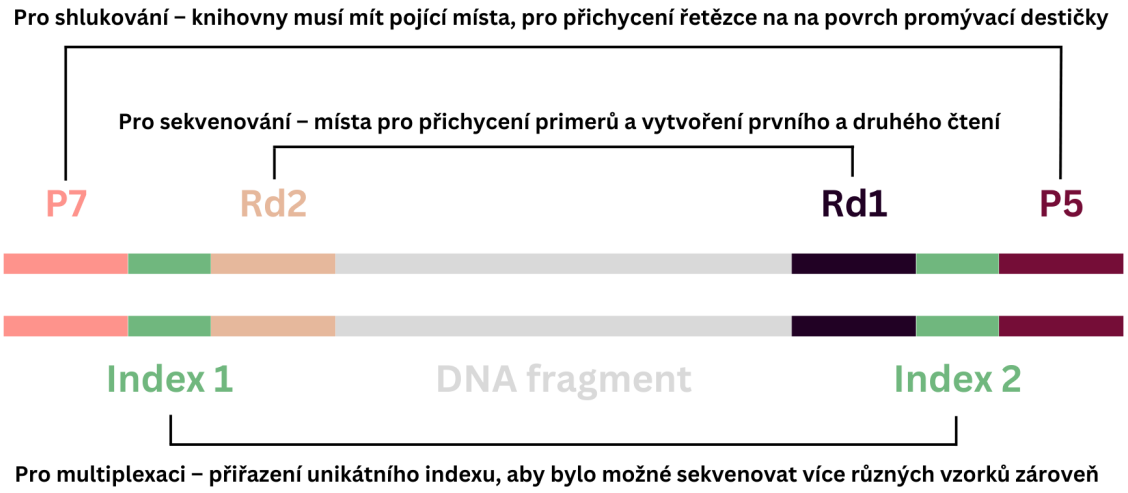
Abychom mohli zahájit sekvenování pomocí technologie Illumina, musíme nejprve připravit tzv. knihovnu z cDNA, které jsme získali ze vzorku, viz kapitola 4.1. Knihovnou myslíme DNA, která je upravena tak, aby ji bylo možno použít pro danou technologii (přidání adaptérů, označení DNA z jednotlivých vzorků). Pro přípravu knihovny je využíváno kitu, který Illumina vyrábí.

Příprava knihovny

Proces přípravy knihovny se dělí do těchto kroků:

1. Kvantifikace genetického materiálu - můžeme využít například metodu qPCR, abychom zjistili počet párů bází ve vzorku.
2. Fragmentace DNA na kratší části - délka sekvencí se bude v případě Illuminy pohybovat mezi 50-600 páry bází. Konkrétní délka závisí na druhu analýzy.
3. Přidání adaptérů k jednotlivým fragmentům - adaptéry jsou nutné pro uchycení fragmentů DNA na promývací destičku, pro rozpoznání, z kterého vzorku fragmenty pochází a pro označení místa začátku genetické informace.
4. Opětovné provedení kvantifikace vzniklé knihovny - zjištění, do jaké míry byla úspěšná fragmentace a přichycení adaptérů.
5. Vzniklá knihovna, která je připravena na vložení do sekvenátoru [14, 16].

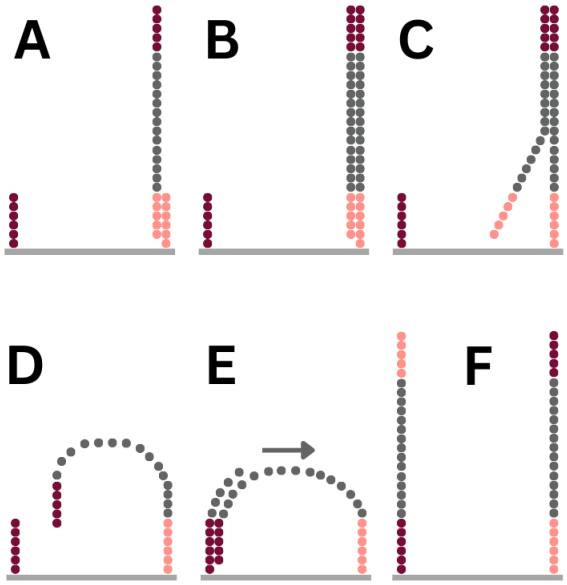
Správně upravený fragment DNA i s popisem důležitých adaptérů je možno vidět na obrázku 4.2. Připravenou knihovnu následně nanášíme na promývací destičku. Promývací destička má skleněný povrch rozčleněný do kanálů. Zde probíhá proces sekvenování.



Obr. 4.2: Správně vytvořená knihovna pro sekvenování – příklad jednoho správně upraveného fragmentu DNA – převzato a přeloženo z [16].

Shlukování

Fragmenty jednotlivých knihoven jsou denaturovány na ssDNA (single stranded DNA - jednořetězová DNA) a přichytávají se na destičku pomocí připojených adaptérů. Každý fragment v knihovně je poté klonován a dochází k vytváření shluků. Klonování probíhá pomocí můstkové PCR, která je názorně vysvětlena na schématu 4.3. Tento proces se neustále opakuje. Po dostatečném namnožení fragmentů dojde k linearizaci (narovnání všech fragmentů 4.3F) a probíhá kvantifikace knihovny. Je nutné zjistit, zda se na destičce nenachází příliš mnoho nebo příliš málo shluků. V obou případech by mohlo mít sekvenování větší chybovost. Při tvoření shluků se tedy snažíme docílit optimálního počtu shluků pro co nejlepší výsledek [14].

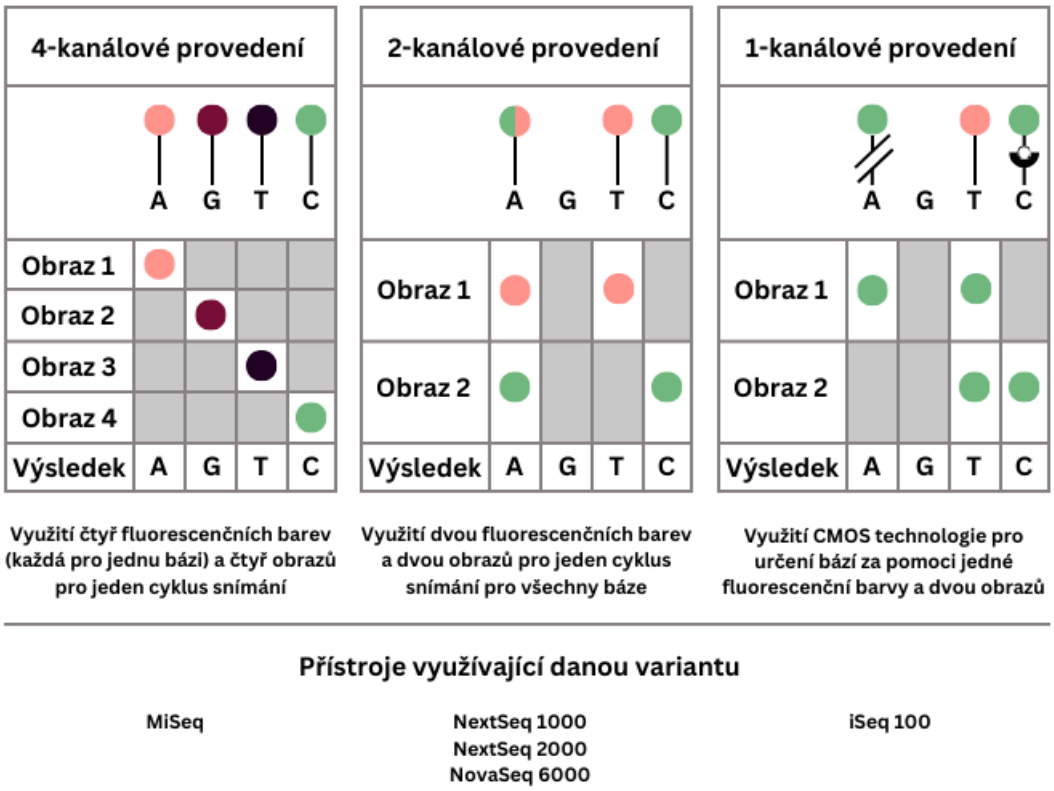


(A) Vlákno ssDNA se přichytí na promývací destičku pomocí adaptéru.
(B) Polymeráza replikuje vlákno, které bude upevněné k destičce.
(C) Vzniklá DNA denaturuje, původní přichycené vlákno se odděluje a je vymýváno pryč.
(D) Upevněné vlákno se ohýbá a spojuje se s nejbližším adaptérem, který má na svém druhém konci.
(E) Vzniká můstek a dochází k replikaci ssDNA.
(F) Dochází k denaturaci a vznikají dvě samostatná vlákna ssDNA upevněná k destičce.

Obr. 4.3: Schématické znázornění procesu můstkové PCR – převzato a přeloženo z [17].

Sekvenování

Při sekvenování vkládáme promývací destičku do sekvenátoru (Illumina vyvíjí různé typy sekvenátorů). K fragmentů DNA jsou přidávány reaktanty a chemikálie pro sekvenování včetně fluorescentně barvených nukleotidů. Při ozáření promývací destičky



Obr. 4.4: Možné varianty sekvenování metodou Illumina. Různé sekvenátory využívají různých technik snímání a barvení bází. Převzato a přeloženo z [16].

dojde k vyzáření barvy odpovídající pro jednotlivé nukleotidy a vzniká obraz, na kterém jsou vidět jednotlivé barvy. Následuje promytí destičky a ozáření další vrstvy nukleotidů. Celý tento proces se nazývá sekvenování syntézou (*anglicky: sequencing by synthesis (SBS)*.

Při sekvenování lze využít několika variant snímání obarvených bází. Tři používané varianty jsou zobrazeny na obrázku 4.4. U 4-kanálové varianty často dochází k překrytí barev a snižuje se tím kvalita čtení. Je nutné dobré rozeznávání barev a softwarové zpracování. Ve 2-kanálovém a 1-kanálovém zpracování nejde rozeznat zda je část přečtené sekvence se stejnými nukleotidy homopolymerem nebo špatně přečtenou sekvencí. Tyto dvě varianty totiž nezavádí možnost přiřazení písmena N místo přečteného nukleotidu v případě, že nejde rozeznat o jaký nukleotid se přesně jedná. Je to dáno tím, že nukleotid G nesvítí ani v jednom obrazu a je tedy přiřazen vždy, když není zaznamenána barva.

Po dokončení samotného sekvenování dochází ke sestavení konsenzu nukleotidových řetězců a stanovení kvality sekvenování pro jednotlivé báze. Vše je uloženo do souborů, kdy pro jeden vzorek dostáváme dva soubory: čtení 1 a čtení 2. Více o datovém formátu v kapitole 5.1 [14].

4.2 Využití softwaru pro predikci metatranskriptomu

Postupem času vzrůstá poptávka na zpětné zjišťování funkčního profilu u již odebraných vzorků. Bylo tedy vytvořeno několik nástrojů pro tento účel, mezi nimi i HUMAnN. Při zpětném zjišťování metatranskriptomu se nejčastěji využívá metod založených na anotování složených metagenomů či funkční profilaci dle reference. Při anotování je vzorek skenován a jsou hledány kódovací sekvence, nekódovací RNA a tRNA. Poté se provádí celková charakteristika a přiřazeních celých kódovacích sekvencí na základě ortologních vztahů, které jsou definovány v databázích.

Funkční profilace může být založena buď to na mapování k referenci pro genomová čtení nebo na prohledávání proteinových databází. U mapování jsou vzorky namapovány k referenci a poté znovu prohledány pro kódující místa. Naopak při prohledávání proteinových databází je vzorek nejprve translatován a potom je profilován na základě ortologních vztahů a četnosti metabolických drah. Poslední zmíněnou variantu využívá nástroj HUMAnN, který je využit v této práci [11, 18].

4.3 Ostatní metody

4.3.1 Hmotnostní spektrometrie

S rozvojem technologií se pro analýzu DNA a následné určení mikrobů začala na konci 20. století opět používat hmotnostní spektrometrie. Začaly se využívat zejména vylepšené a upravené metody hmotnostní spektrometrie, např. MALDI-TOF hmotnostní spektrometrie.

Matrix assisted laser desorption ionization-time of flight hmotnostní spektrometrie využívá roztoku tvz. matrice, do které je vzorek namočen. Při sušení začne roztok společně se vzorkem krystalizovat. Vzorek je poté v matrici ionizován. Při ionizace a desorpci laserem vznikají protonové ionty, které jsou urychlovány s jednotným potenciálem a jsou od sebe separovány na základě jejich poměru hmotnost ku náboji. Následně jsou tyto ionty zachytávány a měřeny pomocí zařízení určujících dobu letu.

Zmíněná metoda se pro svou rychlost, přesnost a malé náklady používá zejména pro mikrobiální identifikace a typizace kmenů, epidemiologické studie, detekce bojových biologických látek, detekce patogenů přenášených vodou a potravinami, detekce rezistence na antibiotika, detekce patogenů krve a močových cest atd [19].

4.3.2 Fluorescenční mikroskopie

Při využívání fluorescenční mikroskopie pro detekci úseků DNA je největším problémem barvení molekul jádra buňky a současné zachování jádra. Při barvení může dojít k poškození jádra či buňky a znemožnění další analýzy.

Mezi vhodné metody barvení patří např. využití pyrrol-imidazolových (Py-Im) polyamidů či různých dalších syntetických molekul. Samotná fluorescenční mikroskopie funguje na principu ozařování vzorku světlem o určitých vlnových délkách a pozorování preparátu pomocí speciální optiky [20].

4.3.3 Kvantitativní PCR

Kvatitativní PCR nebo také qPCR se využívá nejen pro kvantifikaci DNA či RNA ve vzorku, ale také pro zjištění relativní genové exprese, genotypizování, detekci mutací atd.

Hlavními používanými variantami qPCR jsou SYBR Green PCR a TaqMan PCR. SYBR Green PCR využívá barvu, která se váže na DNA a během PCR reakce vyzařuje světlo, podle kterého lze určit kvantitu DNA. TaqMan PCR využívá sond s fluorescenční barvou, které se přichytí pouze na požadovaná místa a požadované molekuly ve vzorku. Jakmile k nim dospěje proces polymerázové reakce, tak je vyzářena barva sondy, kterou je možné detekovat a zjistit tak nejen počet DNA, ale i jestli je přítomna konkrétní sekvence DNA [21, 22].

4.3.4 Microarrays - DNA čipy

DNA čipy byly vytvořeny za účelem automatizovaného paralelního sekvenování několika DNA sekvencí. V praxi je možné tímto způsobem sekvenovat tisíce DNA sekvencí najednou. Čipy se nejčastěji využívají pro sekvenování, detekci mutací a analýzu genové exprese.

Principem sekvenování u čipů je hybridizace fluorescenčně obarvená DNA v roztoku k jednovláknové DNA, která je trvale uchycena na čip. Takto upevněných DNA je na čipu více a tvoří pole bodů, které je skenováno a vyzářená barva je zaznamenávána. Podle vyzářené barvy se určuje, zda je daná sekvence přítomna. Využívá se k tomu speciálního softwaru.

5 Bioinformatické zpracování vzorků - balíček bioBakery

BioBakery poskytuje kompletní soubor nástrojů a analytického prostředí pro provádění meta-omických analýz. To zahrnuje metody pro jednotlivé kroky zpracování dat mikrobiálních komunit a dalších meta-omických analýz. Dále nabízí statistiky na nižší úrovni, integrované reprodukovatelné pracovní postupy a standardizované balení a dokumentaci prostřednictvím open-source repozitářů, jako jsou GitHub, Conda, PyPI a R/Bioconductor. bioBakery podporuje grid- a cloud-deployable obrazy pro platformy jako AWS, GCP a Docker.

K dispozici jsou online školení, demonstrační data a veřejné fórum pro komunitní podporu. Jeho jedinečnost spočívá v schopnosti generovat kontrolu kvality, taxonomický profil, funkční profil, profil kmene a výsledné datové produkty a zprávy v rámci jednoho pracovního postupu, přičemž udržuje verzování a záznamy o původu dat. bioBakery čerpá z databáze ChocoPhlAn, která využívá zdrojů jako je NCBI nebo UniProt [23].

Příkladem nástrojů je StrainPhlAn, pro detekování polymorfismů, PICRUSt, pro predikci funkčního obsahu z markrů, nebo KneadData, pro kontrolu kvality čtení. Tato práce se ale zaměřuje na dva důležité nástroje, kterými jsou MetaPhlAn a HUMAnN. Nejprve však budou krátce představeny datové formáty, s kterými je možné se při analýze setkat [24].

5.1 Datové formáty

Důležitým datovým formátem, pro výstup sekvenování a jako vstupní formát pro analýzu, je fotmát .fastq. Jedná se o textový člověkem čitelný formát, který slouží k uchování nukleotidových sekvencí a kvality, s jakou byly sekvenátorem přečteny. Pro každou sekvenci jsou v tomto formátu vyčleněny čtyři řádky:

1. @ a název sekvence
2. nukleotidová sekvence
3. znak +
4. kvalita přečtení jednotlivých nukleotidů

Velikost dat je často velká, proto se běžně můžeme setkat s komprimovanou variantou tohoto formátu .fastq.gz, kde .gz znamená gzip, což se nástroj pro komprimování a dekomprimování souborů.

Výstupy z provedené datové analýzy bývají často v běžně využívaném formátu tabulky .csv (comma separated values - čárkou oddělené hodnoty) nebo .tsv (tabulator separated values - tabulátorem oddělené hodnoty). Další výstupní formáty

jsou nejčastěji v textové či kódové podobě. Pro jejich přehledné zobrazení lze využít nástroj MultiQC [25], který shrne výsledky do přehledných tabulek a grafů a předá tak uživateli důležité informace o zpracovaných datech.

5.2 MetaPhlAn 4

MetaPhlAn 4 je nástroj navržený pro taxonomické profilování metagenomů, který přináší vylepšení a rozšíření svých stávajících možností. Přístup je založen na integraci rozsáhlého sestavování metagenomů s referenčními genomy bakterií a archeí uloženými v objemné databázi CHOCOPhlAnSGB, což umožňuje efektivní mapování metagenomů na miliony jedinečných markerových genů.

MetaPhlAn využívá genomových úseků na druhové úrovni (SGBs - species-level genome bins) jako primárních taxonomických jednotek, které sdružují mikrobiální genomy a metagenomicky sestavené genomy (MAGs - metagenome-assembled genomes) do konzistentních skupin na úrovni druhu. To usnadňuje přesné taxonomické analýzy a organizaci vzorků do operačních taxonomických jednotek (OTUs - operational taxonomic units).

Metoda integruje do struktury SGBs více než 1 milion MAGs a genomů, které vytvářejí jednu z největších databází spolehlivých referenčních sekvencí mikrobů. Tato integrace zajišťuje komplexní reprezentaci pro taxonomické profilování.

MetaPhlAn 4 zdokonaluje postup extrakce unikátních markerových genů z každého SGB a tím optimalizuje strategii referenčního mapování. Tento krok je klíčový pro přesnou taxonomickou identifikaci a kvantifikaci.

SGBs jsou klíčovým prvkem tohoto přístupu, který zahrnuje vymezení mikrobiálního druhu na základě genetických vzdáleností celých genomů s minimálně 5% shodou na úrovni genomické identity. Taxonomické označení může být přiřazeno SGB na základě přítomnosti nebo nepřítomnosti označených genomů z izolované sekvenace [26].

5.3 HUMAnN 3

HUMAnN je nástroj, který slouží k efektivnímu a přesnému profilování přítomnosti nebo absence a hojnosti mikrobiálních drah v mikrobiomu z metagenomických nebo metatranskriptomických sekvenačních dat (typicky miliony krátkých čtení DNA nebo RNA). Tento proces se nazývá funkční profilování a má za cíl popsat metabolický potenciál mikrobiální komunity a jejich členů. Jedná se tedy o určování metatranskriptomu. Obecně můžeme říci, že funkční profilování odpovídá na otázku: "Co mikroby v našem vzorku dělají nebo jsou schopny dělat?"[23].

Při používání HUMAnNu nejčastěji zjišťujeme metatranskriptom z metagenomických dat. HUMANnN využívá ortologních genomů, genových rodiny a četností metabolických drah ke stanovení odhadu metatranskriptomu. V procesu je nalezená genetická informace translatována a porovnávána s referenční databází DIAMOND, která je dostupná v různých verzích, popřípadě je možné zvolit databázi RAPSearch2. HUMAnN poté na základě různých parametrů určí predikci aktivních genových rodin, četnost aktivních metabolických drah a jejich pokrytí ve vzorku [18].

6 Návrh testovacího datasetu - využití NCBI

Pro návrh testovacího datasetu byla využita databáze NCBI (National Center for Biotechnology Information), konkrétně sekce SRA (Sequence Read Archive), kde jsou ukládány a zveřejňovány čtení z jednotlivých vzorků od daných experimentů. Všechny dostupné vzorky jsou veřejné a dostupné pro zpracování za akademickými účely. Pro testovací dataset bylo vybráno deset vzorků, jejich přehled je možné vidět v tabulce 6.1 níže. U každého vzorku je uveden odkaz na přehled informací ohledně vzorku, v jakém experimentu byl použit a kde je možné jej stáhnout.

Název vzorku	**Použitý sekvenátor**	**Odkaz na SRA**	**Místo odběru vzorku**	**Typ vzorku**
ERX10325893	Illumina HiSeq 3000	`www.ncbi.nlm.nih.gov/sra/ERX10325893`	střevní mikrobiom člověka	metagenomický
SRX20105627	Illumina NovaSeq 6000	`www.ncbi.nlm.nih.gov/sra/SRX20105627`	nádorová tkáň	metagenomický
SRX21412157	Illumina MiSeq	`www.ncbi.nlm.nih.gov/sra/SRX21412157`	horní cesty dýchací u člověka	metagenomický
SRX23032646	Illumina NovaSeq 6000	`www.ncbi.nlm.nih.gov/sra/SRX23032646`	dutina ústní (lidské sliny)	metagenomický
SRX23032643	Illumina NovaSeq 6000	`www.ncbi.nlm.nih.gov/sra/SRX23032643`	lidské oko	metagenomický
SRX3201554	Illumina HiSeq 2000	`www.ncbi.nlm.nih.gov/sra/SRX3201554`	střevní mikrobiom člověka (stolice)	metagenomický
SRX21188772	HiSeq X Ten	`www.ncbi.nlm.nih.gov/sra/SRX21188772`	střevní mikrobiom člověka (stolice)	metatraskriptomický
SRX20325651	Illumina NovaSeq 6000	`www.ncbi.nlm.nih.gov/sra/SRX20325651`	střevní mikrobiom člověka (stolice)	metatraskriptomický
ERX7017794	Illumina HiSeq 2500	`www.ncbi.nlm.nih.gov/sra/ERX7017794`	střevní mikrobiom člověka	metatraskriptomický
ERX5600663	HiSeq X Ten	`www.ncbi.nlm.nih.gov/sra/ERX5600663`	střevní mikrobiom člověka	metatraskriptomický

Tab. 6.1: Přehled vzorků použitých v testovacím datasetu

Vzorky nejčastěji pocházejí z částí lidského těla např. střev nebo z nádorů vznikajících v lidském těle. Záměrně byly vybrány vzorky z různých sekvenátorů, různých částí lidského mikrobiomu a různých experimentů, aby bylo možné výsledný dashboard otestovat na různých datech a zaručit jeho univerzálnost a použitelnost pro ši-

roký okruh dat. Testovací dataset slouží pouze pro testování dashboardu, nejedná se o dataset určený k vyhodnocení.

6.1 Postup zpracování

Vzorky z testovacího datasetu byly zpracovány pomocí vlastního postupu, jehož schematický návrh je možné vidět na obrázku 6.1. Po stažení vzorků z NCBI SRA byly jednotlivé vzorky podrobeny kontrole kvality pomocí softwaru FastQC [27] a FastQ Screen [28], pro odhalení nedostatečné kvality čtení, přebývajících adaptérů nebo případných kontaminací. Následně byla provedena filtrace vzorků nástrojem Fastp [29]. Parametry filtrace byly nastaveny na základě výsledků kontroly kvality, aby byla výsledná analýza relevantní. Při filtrování se odstraňují nedostatečně kvalitně osekvenovaná čtení a odstraňují se adaptéry.

Samotná analýza probíhá pomocí softwaru HUMAnN [23], který predikuje metatranskriptom a pomocí softwaru MetaPhlAn [26], který taxonomicky zařazuje jednotlivá čtení. Výsledné tabulky pro jednotlivé vzorky jsou poté spojovány prostřednictvím pomocným algoritmů softwaru HUMAnN a MetaPhlAn. Výstupy těchto algoritmů jsou následující:

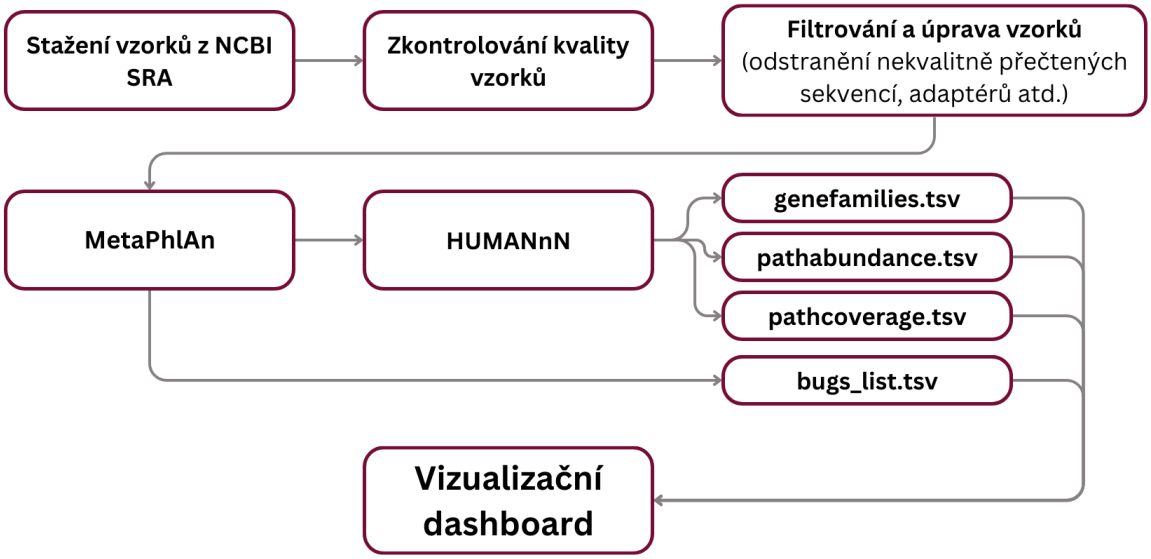
MetaPhlAn

- **bugs_list.tsv** - udává taxonomickou profilaci vzorku, tedy s jakou hojností jsou jednotlivé organismy zastoupené ve vzorku

HUMANnN

- **genefamilies.tsv** - udává hojnost zastoupení jednotlivých genových rodin ve vzorku
- **pathabundance.tsv** - udává hojnost zastoupení metabolických drah ve vzorku, ty představují funkční potenciál vzorku
- **pathcoverage.tsv** - udává jaké biochemické funkce mohou být prováděny mikrobiální komunitou ve vzorku, jaká je rozmanitost metabolických drah ve vzorku

Takto upravené výstupy jsou finální formou výsledků analýzy určenou pro vložení do vizualizačního dashboardu a následnou interpretaci. Všechny výpočetní úkony byly v rámci efektivního a pokud možno rychlého zpracování provedeny v gridovém výpočetním prostředí MetaCentrum.



Obr. 6.1: Blokové schéma zpracování vzorků pro testovací dataset.

7 Vizualizační dashboard

Vizualizační dashboard se skládá ze tří hlavních částí:

- **Overview** - umožňuje načtení, zobrazení a manipulaci s tabulkou,
- **Graphs** - zobrazuje základní grafy důležité pro metatranskriptomickou analýzu,
- **Statistics** - provádí základní statistickou analýzu a zobrazuje ji pomocí grafů.

Všechny části jsou níže podrobně popsány a okomentovány. Dashboard poskytuje základní důležité komponenty pro analýzu metatranskriptomu a pro hodnocení výsledků analýzy. Vše bylo naprogramováno pomocí programovacího jazyku Python, příslušných knihoven a frameworku Streamlit, který umožňuje vytvářet dynamické aplikace pro zobrazování dat. Vše je podrobněji popsáno v kapitole 7.1.

7.1 Programové řešení

V této práci byl využit programovací jazyk Python (dále pouze Python) a jeho knihovny. Dále byl použit Streamlit framework, který zabezpečuje snadné vyvíjení aplikačního prostředí a poskytuje nástroje pro vytvoření vizualizačního dashboardu.

7.1.1 Python a použité knihovny

Pro svou multifunkčnost, přehlednost, jednoduchost učení a velký počet dostupných balíčků pro analýzu byl pro implementaci dashboardu použit právě Python (verze 3.11.7). Tento programovací jazyk vyniká ve zpracování dat a datové analýze, ale také ve tvoření aplikací či jako serverová část pro webové stránky. Python je objektově orientovaný programovací jazyk na vysoké úrovní, což znamená, že řeší spoustu úkonů za uživatele (např. alokování paměti, kolik místa v paměti je potřeba, jaký typ proměnné se má nastavit) a usnadňuje proces samotného programování.

Je tedy vhodným programovacím jazykem pro začátečníky, kteří si chtějí vytvořit svou vlastní aplikaci, ale také pro zkušené programátory, kteří chtějí využít nespočetného množství možností, které Python nabízí. Dále budou popsány knihovny a balíčky použité při tvorbě dashboardu.

Knihovny pro zpracování dat

Knihovna Pandas (verze 2.2.2) je nástroj pro jednoduchou ale zároveň komplexní manipulaci s daty. Využívá objektů DataFrame a Series, do kterých jsou data nahrána a následně je možná úprava a jejich zpracování. Značná část knihovny je postavena na balíčku Numpy, který zjednodušuje a zefektivňuje práci s čísly v Pythonu.

Díky tomu je Pandas výkonnou a spolehlivou volbou pro zpracování dat. V práci zabezpečuje veškerou transformaci a úpravu dat.

Pro efektivní transformaci dat do Pandas DataFrame je využito balíčku IO (součást základní instalace Python), který umožní Pandas přečtení nahraného souboru ve formátu bytes.

Knihovny pro vizualizaci dat

Plotly (verze 5.22.0) je volně přístupná knihovna pro tvoření interaktivních a vysoce kvalitních grafů a vizualizací. Díky interaktivnosti může uživatel manipulovat s vytvořeným grafem, lépe mu porozumět a vyčíst z grafu všechny potřebné informace. Díky svému jednoduchému a propracovanému nastavení je ideální volbou pro zobrazení různorodých vizualizací. Další knihovnou, která úzce souvisí s Plotly, je Dash Bio (verze 1.0.2). Tento balíček zprostředkovává bioinformatické vizualizace, jako je například Clustergram využitý v dashboardu.

Knihovny pro statistické zpracování

Pro zpracování statistické analýzy byla využita knihovna Scikit-Bio (verze 0.6.0), která je vytvořena pro výpočty v bioinformatice. Knihovna obsahuje širokou škálu nástrojů pro práci s biologickými sekvencemi, fylogenetickými stromy nebo pro analýzu diverzity mikrobiomů. Z knihovny je zde využit balíček pro alfa a beta diverzitu a pro výpočet analýzy hlavních koordinát (PCoA). Podrobnější popis těchto balíčků se nachází v kapitole 7.4.1.

7.1.2 Django

První návrh vizualizačního dashboardu byl zpracováván pomocí frameworku Django a programovacího jazyku JavaScript. Django je robustní webový framework pro programovací jazyk Python. Jako takový poskytuje přehledné serverové zpracování dat propojené s navrhováním webové stránky a využití JavaScriptu, HTML a CSS pro zobrazování a interagování s daty na stránce. Django využívá přesměrovávání pomocí URL, lze v něm vytvořené projekty škálovat a znovu využívat. Dále nabízí přehledné zpracování formulářů, potřebnou ochranu dat při jejich posílání a zpracování a možné uživatelské admin rozhraní.

V základu pracuje Django dle architektury MTV (M - model, T - předloha, V - pohled). Pomocí modelů jsou ukládány data do relační databáze, vytvářeny formuláře a uložená data předávána pohledům pro další zpracování. Pohledy tedy zabezpečují základní logiku, úpravu dat a zasílání dat na webovou stránku pro

jejich zobrazení. Předlohy poté slouží pro udání struktury a interaktivnosti těmto zobrazeným datům.

Django je díky tomuto systému poměrně rigidní a nelze snadno nahrát do databáze například tabulku, která může mít pokaždé jiný počet sloupců. Z důvodu příliš složitého zpracování proměnlivých dat byl návrh dashboardu přenesen do vizualizačního prostředí Streamlit.

7.1.3 Streamlit

Streamlit je open-source Python knihovna určená pro zjednodušenou tvorbu různorodých webových aplikací týkajících se datového zpracování, bioinformatiky, strojového učení a vizualizací pro vědecké a akademické účely. Celá webová stránka je tvořena v Pythonu, za pomoci knihovny Streamlit a jejích komponentů, které jsou vkládány do kódu a zajišťují načítání dat, zobrazovaní a práci s tlačítky, zobrazování tabulek, textu a vykreslování grafů. Pomocí těchto jednoduchých komponentů je snadno dosaženo modifikování zobrazených dat uživatelem. Mezi těmito komponenty lze psát běžný kód pro logické operace, úpravu dat nebo tvorbu proměnných v Pythonu. Kompilování kódu a jeho načítání do aplikace funguje lineárně shora dolů. Streamlit tedy přečte kód pro jednotlivé stránky postupně po příkazech a provede logické operace či zobrazení dat v daném pořadí.

Pro zachování dat v proměnných mezi jednotlivými stránkami zde existuje tzv. session state (stav relace), do kterého je možné proměnné uložit. Při překliknutí na jinou stránku, zůstanou proměnné uloženy a připraveny pro použití na nové stránce. Další výhodou je ukládání dat do cache. Při opětovném načítání stejných dat již nemusí znovu proběhnout výpočet, ale jsou načteny data z mezipaměti, což urychluje chod celé aplikace.

Při spuštění aplikace a čtení kódu Streamlitem není nutné používat komponenty například pro jednoduché vypisování textu. Streamlit použije svou optimalizační knihovnu pro automatické určování typů dat a rozhodne, jakým způsobem data nejlépe vykreslí. Ve většině případů je však dobré stanovit přesně, jaký komponent použít. Důvodem je, aby nedocházelo k nechtěnému nebo špatnému vykreslování.

V jádru používá Streamlit pro vykreslování webové aplikace knihovnu React programovacího jazyku JavaScript. Po celou dobu běhu aplikace probíhá vyměňování dat mezi Pythonem (pro serverové zpracování dat) a JavaScriptem (pro zobrazení dat uživateli). Serverová databáze je zajištěna pomocí Tornado Framework [30].

Streamlit byl tedy ideální prostředím pro tvoření vizualizačních dashboardů a podobných webových aplikací. Nabízí jednoduše pochopitelný, ale komplexní systém pro tvorbu interaktivních grafů a vizualizaci statistických analýz.

7.2 Overview - zobrazení tabulky

Úvodní strana dashboradu s názvem Overview přináší uživateli možnost nahrát tabulku ve formátu .tsv v postranním panelu (panel je pro všechny následující strany stejný) a zobrazit ji uprostřed první strany. Je možné nahrávat více souborů najednou nebo soubory nahrávat postupně. Výběr konkrétního souboru se provede pomocí výběrového menu pod tlačítkem pro nahrání souboru. Při manuálním obnovení stránky (tedy při zmáčknutí tlačítka Obnovit v prohlížeči či zmáčknutí klávesy F5 na klávesnici) se všechny nahrané soubory odstraní a je možné nahrát soubory nové. Je tím zajištěno, aby v dashboardu nepřetrvávala data po zavření stránky či špatném obnovení prohlížeče.

Zobrazená tabulka je sama o sobě interaktivní a disponuje funkcemi jako zvětšení, řazení právě zobrazených hodnot nebo hledání v právě zobrazených hodnotách. Pro řazení celkového datasetu je nutno využít tlačítek nad tabulkou (konkrétně se jedná o tlačítka označená čísly 4 a 5). Tyto tlačítka seřadí celou tabulku podle zvoleného sloupce a zvoleného směru řazení. Tabulka je z důvodu velkého množství řádků vždy rozdělena do stránek o určitém počtu, který je nastavitelný. Toto nastavení společně s přepínáním a počtem stran lze nalézt naspodu tabulky.

Tabulka dále disponuje tlačítky pro normalizace tabulky (převede relativní abundanci na procentuální zastoupení daného řádku ve sloupci) a pro přidání sloupce s průměrem z jednotlivých řádků. U konkrétních tabulek se poté může objevit tlačítko pro volbu taxonomické úrovně, kterou chceme v tabulce zobrazit. Obrázek s designem celého Overview a legendou se nachází níže 7.1.

Obr. 7.1: Náhled stránky Overview z dashboardu, doplněný o číselné označení prvků. 1. Navigační menu pro přechod mezi jednotlivými stránkami dashboardu. 2. Tlačítko pro nahrání souborů z analýzy. 3. Tlačítko pro výběr nahraného souboru. 4. Tlačítko pro sloupce pro řazení sloupců. 5. Tlačítka pro přepínání mezi vzestupným a sestupním řazením. 6. Tlačítka pro přidání sloupce se průměrnými abundancemi a pro normalizování datasetu na procenta. 7. Tlačítko pro výběr taxonomické úrovně (pro vybrané soubory). 8. Zobrazená tabulka. 9. Počítadlo stránek. 10. Tlačítko pro přepínání stránek. 11. Tlačítko pro nastavení počtu řádků na stránku.

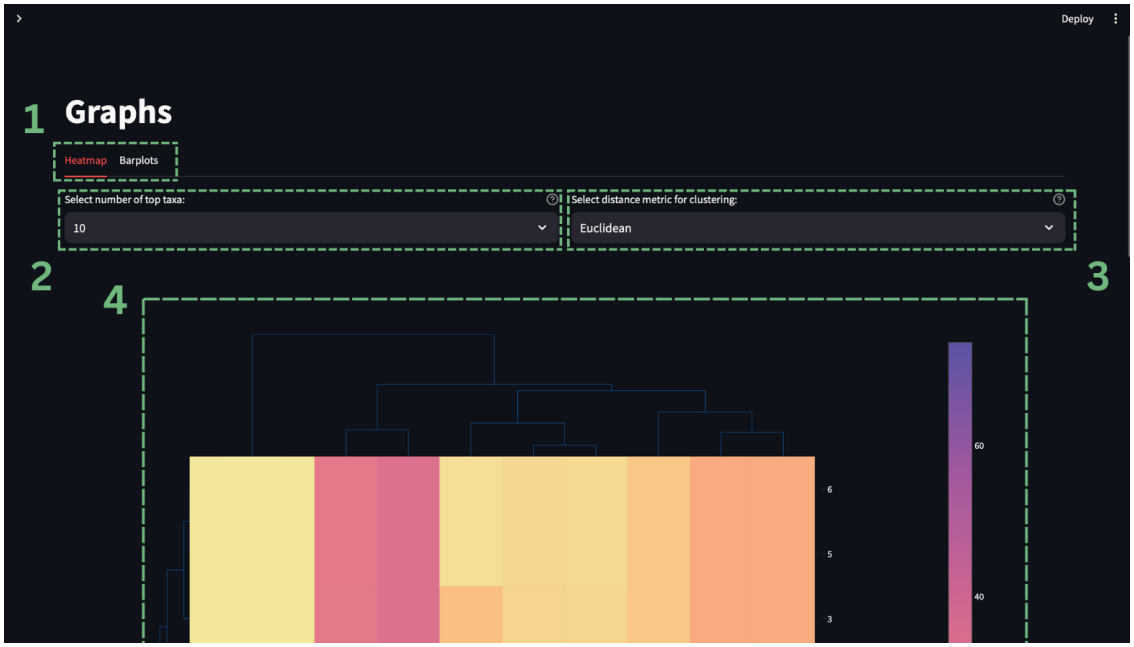
7.3 Graphs - grafické zobrazení dat

Stránka pro grafy vykresluje základní zobrazení průměrně nejabundantnějších řádků tabulek pomocí clustergramu (heatmapa s dendrogramy) a základní přehled rozložení hodnot v souboru pomocí barplotů. Grafické zobrazení poskytuje přehledné informace o zastoupení jednotlivých taxonů a vykresluje shluky pro celkově nejvíce zastoupené taxony. Předává tedy informaci potřebnou k prvotnímu zhodnocení analýzy.

7.3.1 Heatmapa pro vykreslení abundancí

Heatmapa je zde prezentována pomocí clustergramu, tedy heatmapy se shlukováním a vizualizací pomocí dendrogramů. Pro vykreslení je vždy vybrán zvolený počet průměrně nejabundantnějších taxonů. S použitím zvolené metriky je poté generován

clustergram který zobrazuje shluky vzdálenosti a shluky mezi jednotlivými taxony. Jednotlivé taxony jsou z důvodů dlouhých názvů označeny na vertikální ose čísly. Legenda k jednotlivým číslům je zobrazena pod clustergramem. Náhled je možno vidět na obrázku 7.2.



Obr. 7.2: Náhled stránky Graphs pro záložku Heatmap z dashboardu, doplněný o číselné označení prvků. 1. Záložky pro přepínání mezi heatmapou a barploty. 2. Tlačítko pro výběr počtu nejabundantnějších taxonů dle průměrné abundance napříč vzorky. 3. Tlačítko pro zvolení vzdálenostní metriky pro shlukování a vytvoření clustergramu 4. Zobrazení clustergramu s legendou pro řádky pod clustergramem.

7.3.2 Barploty

Pro relativní zastoupení jednotlivých taxonů ve vzorcích jsou vykresleny barploty, které v procentech udávají relativní abundanci taxonů na dané taxonomické úrovni. U tabulky genefamilies je pro velký počet řádků vykreslen vždy pouze jeden sloupec v absolutní abundanci. Sloupec se vždy seřazen sestupně podle abundance a je vykreslen zvolený počet nejabundantnějších řádků. Pro tabulku pathcoverage se barploty nevykreslují, jelikož pro ni není tato vizualizace vhodná. Ukázku vzhledu je možné vidět v obrázku 7.3.

Obr. 7.3: Náhled stránky Graphs pro záložku Barplots z dashboardu, doplněný o číselné označení prvků. 1. Záložky pro přepínání mezi heatmapou a barploty. 2. Tlačítko pro výběr taxonomických úrovní. 3. Zobrazení barplotu pro vybranou taxonomickou úroveň.

7.4 Statistics - grafické zobrazení statistické analýzy

V sekci pro statistickou analýzu najdeme alfa a beta diverzitu vzorku a odhad diferenciální exprese. Statistická analýza přidává další pohled na data a podrobuje je úpravám pro jejich lepší pochopení a zjištění významných porovnání dle různých výpočtů.

7.4.1 Alfa a Beta diverzita

Alfa diverzita se zabývá rozmanitostí taxonů v jednotlivých vzorcích. Dané vzorky sdružujeme dle nahraných metadat, která jsou potřeba pro smysluplné zobrazení alfa diverzity pomocí violin plotů. Pro výpočet alfa diverzity je možná zvolit Shannonův nebo Simpsonův index diverzity. Nechybí zde také volba jednotlivých taxonomických úrovní a volba, podle jakého příznaku sdružit vzorky do violin plotů. Náhled na vzhled stránky pro alfa diverzitu je možné vidět v obrázku 7.4. Níže jsou uvedeny vzorce pro Shannonův index,

$$H' = -\sum_{i=1}^{s}(p_i \log_2 p_i)$$

kde:

- H' je Shanonnův index,
- s je celkový počet taxonů,
- p_i je počet jednotlivců náležících i-tému taxonu.

A pro Simpsonův index diverzity,

$$1 - \sum_{i=1}^{s} p_i^2$$

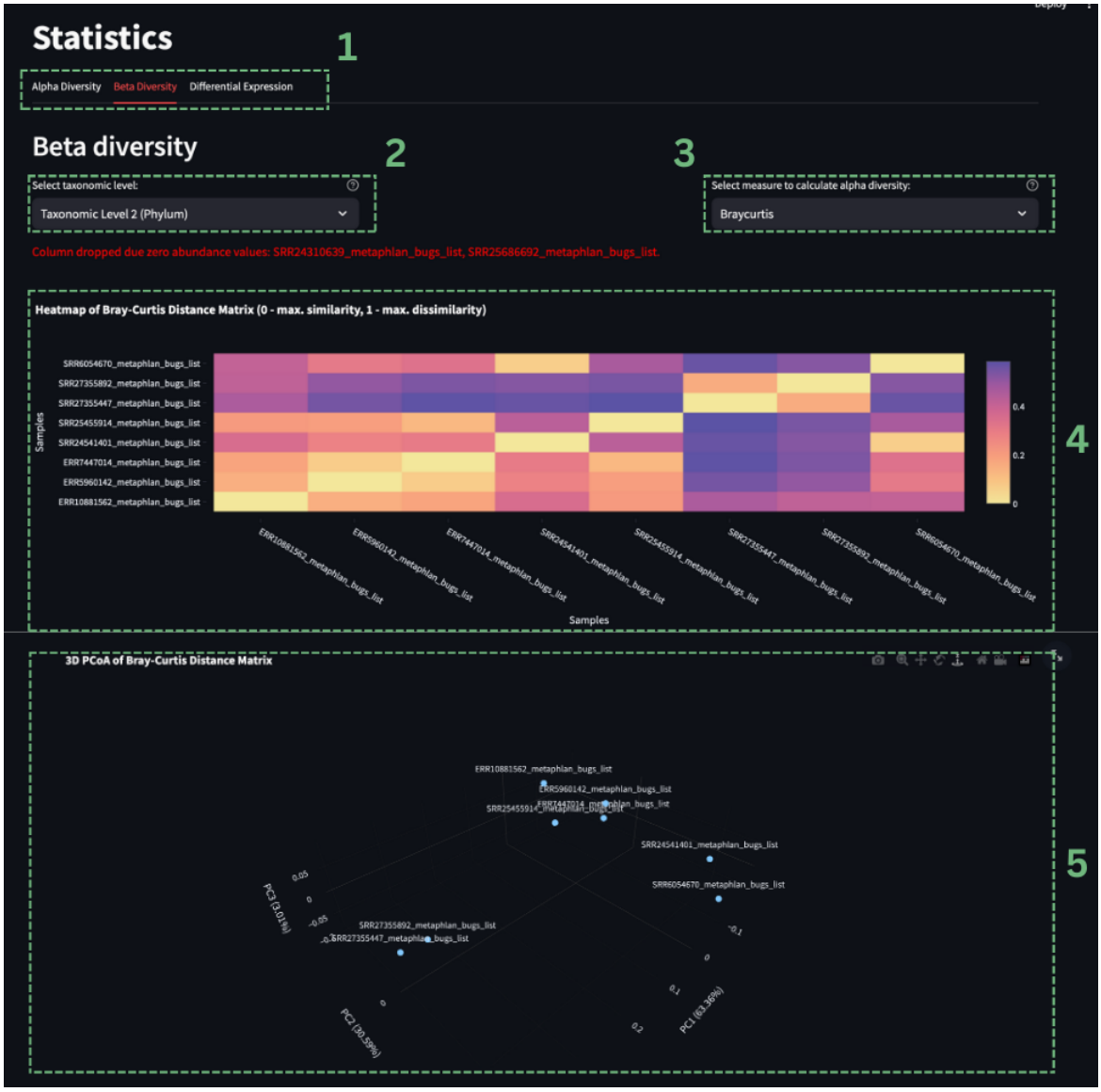
kde:

- s je celkový počet taxonů.
- p_i je počet jednotlivců náležících i-tému taxonu.

Obr. 7.4: Náhled stránky Statistics pro záložku Alpha Diversity z dashboardu, doplněný o číselné označení prvků. 1. Záložky pro přepínání mezi statistickými analýzami. 2. Tlačítko pro výběr taxonomické úrovně. 3. Tlačítko pro zvolení vzdálenostní metriky pro výpočet beta diverzity. 4. Vykreslení distanční matice. 5. Vykreslení 3D scatter plotu pro zobrazení prvních tří komponent z analýzy hlavních koordinát.

Beta diverzita zprostředkovává porovnání mezi jednotlivými vzorky, zabývá se tedy změnou taxonů v jednotlivých vzorcích. Pro výpočet beta diverzity je zde na výběr metrika Bray-Curtis nebo Jaccard. Pomocí těchto metrik vznikne distanční matice, která je i s popisem vykreslena pomocí heatmapy. Dále je pro lepší vyobrazení jednotlivých shluků dat provedena analýza hlavních koordinát (PCoA), která je následně

vynesena do 3D scatter plotu. PCoA je metoda využívající distančních neeuklidovských matic pro redukování dimenzionality dat, a je zde využita zejména pro účely vizualizace. Opět zde nechybí možnost vybrat jednotlivé taxonomické úrovně. Příklad vykreslení grafů a náhledu stránky je ukázán na obrázku 7.5.



Obr. 7.5: Náhled snímku stránky Statistics pro záložku Beta Diversity z dashboardu, doplněný o číselné označení prvků. 1. Záložky pro přepínání mezi statistickými analýzami. 2. Tlačítko pro výběr taxonomické úrovně. 3. Tlačítko pro výběr příznaku, podle kterého seskupit vzorky pro vykreslení alfa diverzity. 4. Tlačítko pro zvolení metriky pro výpočet alfa diverzity. 5. Vykreslení alfa diverzity pomocí violin plotu.

7.4.2 Diferenciální exprese (odhad)

Při výpočtu diferenciální exprese byl využit pouze její odhad pomocí odečtení dvou vybraných sloupců a vzniku sloupce s rozdílovou hodnotou, který byl převeden do absolutní hodnoty. Pomocí tohoto sloupce je vždy tabulka seřazena sestupně a je vykreslen zvolený počet nejrozdílnějších vzorků.

Pouhý odhad byl zvolen, protože tabulky jsou již různě normalizovány a bylo by obtížné a nepraktické je převádět na jiné jednotky. Pro vizualizaci rozdílnosti vzorků tedy postačí jejich samotný rozdíl a vykreslení do heatmapy, kde první dva sloupce jsou zvolené vzorky s jejich hodnotami pro dané taxony a ve třetím sloupci jsou hodnoty rozdílu. Lze zde tedy vyčíst, o kolik se jednotlivé taxony liší v jednotlivých sloupcích. Ve výpočtu rozdílu se vždy odečítá druhý sloupec od prvního. Heatmapa zobrazuje rozdíly pro zvolený počet nejrozdílnějších taxonů. Náhled této stránky je možné vidět na obrázku 7.6.

Obr. 7.6: Náhled stránky Statistics pro záložku Differencial Expression z dashboardu, doplněný o číselné označení prvků. 1. Záložky pro přepínání mezi statistickými analýzami. 2. Tlačítko pro zvolení prvního sloupce pro porovnání. 3. Tlačítko pro zvolení druhého sloupce pro porovnání. 4. Tlačítko pro zvolení počtu nejrozdílnějších taxonů. 5. Tlačítko pro zvolení taxonomické úrovně. 6. Vykreslení heatmapy pro hodnoty prvního sloupce, druhého sloupce a rozdílu sloupců.

8 Diskuze

Celometagenomové sekvenování a metagenomika obecně jsou využívány pro studium mikrobiálních společenstev a pomáhají nám pomocí sekvenování genetické informace nahlédnout blíže do prostředí, které je lidskému oku skryto. Takovým prostředím může být půda, do které zaséváme plodiny, vzduch, který dýcháme, voda, bez které by nevznikl život, nebo samotné lidské tělo, o kterém toho víme hodně, ale stále ne tak docela vše. Svět, ve kterém se odehrávají důležité procesy pro fungování života, můžeme spatřit pod okulárem moderních mikroskopů a zkoumat jeho koloběh. Přesné určování počtu mikroorganismů a jejich vlastností však šlo jinou, lépe uzpůsobenou cestou.

Za pomoci již zmíněného sekvenování přímo jednotlivých nukleotidů a následného čtení celých genomů, je možné přesně a efektivně určit počet a typ mikroorganismů v odebraném vzorku. Je potřeba zjistit, proč se nedaří rostlinám v konkrétním prostředí? Žádný problém. Odeberou se vzorky, v laboratoři se náležitě upraví a vloží se do sekvenátoru. Výsledná data dále analyzujeme pomocí počítačů a dostáváme přehledný výčet většiny mikroorganismů, které by mohly škodit půdě.

Správně a přesně určit viníka však nemusí být až tak jednoduché jak se zdá. Sekvenační metody a následná analýza pro poměrně rychlé sekvenování jsou do značné míry závislé na předem prozkoumaných mikroorganismech a uložených datech v databázích. Aby se mikroorganismus mohl do databáze dostat, musí být nejprve důkladně sekvenován přesnější, nákladnější a zdlouhavější metodou. I těchto metod využívá celometagenomové sekvenování pro objevování nového a neprobádaného života.

Další velkou výhodou celometagenomového sekvenování a metagenomiky je identifikace mikroorganismů, které není možno kultivovat v laboratoři a posvítit si na ně pod laboratorním sklem. Odpadá tedy nutnost konstruovat vhodné prostředí, jen abychom zjistili, o jaký mikroorganismus se jedná. Někdy totiž není možné takového prostředí docílit.

Pominout by se neměla aplikace sekvenování v oboru medicíny, biotechnologií, ekologie a dalších odvětvích lidské činnosti. Sekvenování zde nachází nesčetného využití, od zachraňování životů pro zlepšování lidského společenství.

Jak vyplývá z uvedených příkladů, znát, co se v prostředí kolem nás nachází, je důležité. Neméně důležité je však vědět, co se kolem nás děje. Tedy přesněji, co má který mikroorganismus právě na práci a jak to ovlivňuje jeho okolí.

V tomto ohledu nám pomůže metatranskriptom, který se přesně těmito otázkami zabývá. Při analýze se zaměřujeme na právě transkribovanou informaci, tedy na právě produkovanou mRNA. Dozvíme se tedy, co v daném časovém okamžiku mikrobiom dělá, můžeme ho podrobit experimentům a zkoumat jeho reakci. Abychom

nepochybili a zjistili co nejvíce, musíme zkoumané vzorky sekvenovat s dobrý pokrytím.

Mikroorganismy jsou schopné se adaptovat. Zmíněným experimentem můžeme vyvolat adaptační změnu a pomocí metatranskriptomiky zkoumat, jak změna probíhá a co se v mikrobiomu děje. Proces nám poskytne důležité odpovědi na řadu otázek, které mohou pomoci například v lékařském průmyslu k objevení nových a lépe cílených léčiv.

Při celometagenomové analýze a analýze metatranskriptomu narazíme na nemalá úskalí. Potíže mohou nastat již při odběru a zpracování vzorků. Sebemenší kontaminace či odchylka od postupu může zavést chybu, která drasticky ovlivní výsledek celé analýzy. Přítomné kontaminace lze v mnohých případech při počítačovém zpracování odhalit, v opačném případě by kontaminace mohla způsobit zcela klamné závěry a vyústit v další nesprávný postup.

Složitost zpracování celé analýzy se projevuje také v objemu dat, které je nutno počítačem zpracovat a uložit. V tomto případě musí být využito serverových gridových výpočetních systémů. Další překážkou je výběr vhodného softwaru pro takovou analýzu. Zprvu se to může zdát jako složitý případ pro leckterého detektiva a je tedy dobré, mít po ruce bioinformatika, který se umí správně navigovat velkým množstvím dostupného softwaru, popřípadě si sám navrhnout vlastní elegantní řešení.

U analýzy tohoto typu je možno nastavit nesčetně parametrů a využít celou řadu metrik a příslušných jednotek. Na metagenom či metatranskriptom se můžeme dívat z mnoha úhlů a každý potřebuje informace právě pro ten svůj. Problém může dále nastat ve standardizaci, kdy pro podobné věci používáme rozdílný zápis. Obecně platí, že pro zjednodušení celé analýzy je potřeba tvořit velké a standardizované databáze, které ulehčují získávání výsledků a urychlují celý proces.

Taktéž výsledky analýzy lze zobrazovat a interpretovat různě. Při využití balíčku bioBakery je zde několik možností vizualizace dat. Pro vytvoření kruhových taxonomických a fylogenetických stromů je zde GraPhlAn, samotné bioBakery workflows nabízí vytvoření reportu s vizualizacemi pro celometagenomová a 16S data. Pro vizualizaci výsledků HUMAnN a MetaPhlAn je zde program Hclust2, který tvoří heatmapy se shlukováním pomocí korelací. Všechny tyto možnosti vyžadují znalosti programování a používání kódu pro generování výstupu. Některé složitější než jiné. Postrádají interaktivní prvky a okamžité zobrazení výsledků.

Vizualizační dashboard představený v této práci poskytuje po nahrání dat možnost s daty ihned manipulovat a zobrazit si potřebné vizualizace. Například při otevírání velkého .tsv souboru v počítači, může dojít k přetížení editoru pro tabulková data. Dashboard zvládne velkou tabulku načíst a provádět různé operaci s daty. Tabulka je interaktivní a přizpůsobitelná. Dále dashboard poskytuje přehledné a interaktivní grafy pro vizualizaci dat a zobrazení dat pod různými úhly

pohledu a statistický náhled na zpracovaná data.

Všechny tyto vlastnosti dashboardu umožňují snadnější interpretaci dat. Dashboard má sloužit biologům, laboratorním pracovníkům, bioinformatikům a všem těm, kteří potřebují rychle nahlédnout na výsledky analýzy pomocí softwaru HUMAnN a MetaPhlAn.

Pro nastavování a správné fungování dashboradu byl použit testovací dataset (viz kapitola 6), který byl sestaven z různorodých vzorků, aby dashboard fungoval co nejuniverzálněji. Postup zpracování vzorků byl dle vlastního návrhu implementován za účelem získání relevantních výsledků. Bližší informace a schéma navrženého postupu lze nalézt v kapitole 6.1. Dashboard byl následně pomocí celého testovacího datasetu otestován a dále upraven pro lepší fungování.

Celé technické řešení je jednoduché, volně přístupné, škálovatelné a snadno modifikovatelné. Kód společně s informacemi o dashboardu je veřejně přístupný na platformě GitHub. Dashboard bude zajisté více než vhodným příspěvkem do bioinformatické komunity. Návrhy, změny či vlastní programové vylepšení dashbordu jsou vítány. Vhodných vizualizací pro lepší pochopení výsledků analýz není nikdy dost.

Závěr

Bakalářská práce shrnuje podstatné informace o exprimaci genetické informace člověka a zaměřuje se na geny exprimované v daném časovém momentu. Podrobně popisuje nukleové kyseliny, vysvětluje mikrobiom, metagenom a metatranskriptom. Ukazuje potenciál metatranskriptomiky, jejího využití a rozebírá složitost zpracování metatranskriptomických dat. Shrnuje a popisuje sekvenační technologii Illumina, která je jednou z nejvyužívanějších v oblasti sekvenování a metatranskriptomiky. Poskytuje ale i náhled ostatních sekvenačních metod.

Stručně je zde popsán využívaný software HUMAnN a MetaPhlAn (součást balíčku bioBakery), který se využívá při analyzování vzorků. Hlavní části analýzy jsou odhad metatranskriptomu a taxonomické zařazení jednotlivých čtení.

Podstatnou kapitolou pro práci je sestavení postupu zpracování pro testovací dataset. Pro zpracování datasetu byl sestaven vlastní návrh tak, aby byl testovací dataset zkontrolován, filtrován, analyzován a byl vhodný pro otestování dashboardu, který má být univerzální a má pojmout rozmanitá data. Dashboard má díky těmto vlastnostem nalézt co nejširší uplatnění v dané problematice. Samotný dataset musí projít zvolenými programy s vhodně upravenými parametry, aby mohl být brán jako relevantní a mohl být použit pro testování. Celý postup byl po návrhu implementován a kompletní analýza proběhla pro všechny vzorky v datasetu. Dataset byl dále využit pro nastavování a testování vizualizačního dashboardu.

Hlavním úkolem práce bylo vytvoření vizualizačního dashboardu pro výsledky z metatranskriptomické analýzy dat. Dashboard byl vytvořen s důrazem na jednoduchost, vhodné vizualizace, interaktivnost, ale také škálovatelnost, možnost úpravy dat a dalšího rozšíření. Pro správnou interpretaci analýzy je v tomto případě klíčová správná a užitečná vizualizace dat ale i jejich statistické zpracování. Do dashboradu byly zakomponovány všechny důležité prvky pro metatranskriptomickou analýzu.

Cílem práce také bylo vyzdvihnout důležitost vizualizace pro správné porozumění výsledkům analýz a ukázat, jaké možnosti vizualizace nabízí. Práce je proto doplněna o tématické obrázky, které byly vytvořeny za účelem názorného a přehledného přiblížení problematiky.

Kompletní bakalářský projekt se stručným návodem a popisem je možno nalézt pod tímto odkazem: https://github.com/HonzaHyl/VisualizationDashboard. Soubory jsou zprostředkovány tímto způsobem, aby mohl být dashborad dále rozvíjen a upravován, protože metatraskriptomická analýza je podstatnou součástí zkoumání mikroorganismů kolem nás a v budoucnu bude jistě rozšířena a zlepšována za účelem objevení nových poznatků a průlomů ve vědě.

Literatura

1. KRÁLÍKOVÁ, M.; PAULOVÁ, H.; KOL., a. *Biochemie pro biomedicínské techniky*. 1. vydání. Brno: MU Brno, 2021. ISBN 978-80-210-9857-2.
2. CLARK, David P.; PAZDERNIK, Nanette J.; MCGEHEE, Michelle R. *Molecular Biology*. 3rd edition. Elsevier, 2019. ISBN 978-0-12-813288-3.
3. *Rozdíl mezi strukturou DNA a RNA* [online]. [cit. 2023-10-18]. Dostupné z: https://www.wikiskripta.eu/w/Sekund%C3%A1rn%C3%AD_struktura_DNA.
4. FASNACHT, Michel; POLACEK, Norbert. Oxidative Stress in Bacteria and the Central Dogma of Molecular Biology. *Frontiers in Molecular Biosciences* [online]. 2021-5-10, roč. 8 [cit. 2023-11-12]. ISSN 2296-889X. Dostupné z DOI: 10.3389/fmolb.2021.671037.
5. WANG, YAN; LIU, JING; HUANG, BO; XU, YAN-MEI; LI, JING; HUANG, LIN-FENG; LIN, JIN; ZHANG, JING; MIN, QING-HUA; YANG, WEI-MING; WANG, XIAO-ZHONG. Mechanism of alternative splicing and its regulation. *Biomedical Reports* [online]. 2015-03-17, roč. 3, č. 2, s. 152–158 [cit. 2023-11-07]. ISSN 2049-9434. Dostupné z DOI: 10.3892/br.2014.407.
6. CHOI, Sunkyung; CHO, Namjoon; KIM, Eun-Mi; KIM, Kee K. The role of alternative pre-mRNA splicing in cancer progression. *Cancer Cell International* [online]. 2023, roč. 23, č. 1 [cit. 2023-11-07]. ISSN 1475-2867. Dostupné z DOI: 10.1186/s12935-023-03094-3.
7. ZHANG, Yancong; THOMPSON, Kelsey N.; BRANCK, Tobyn; YAN, Yan; NGUYEN, Long H.; FRANZOSA, Eric A.; HUTTENHOWER, Curtis. Metatranscriptomics for the Human Microbiome and Microbial Community Functional Profiling. *Annual Review of Biomedical Data Science*. 2021-07-20, roč. 4, č. 1, s. 279–311. ISSN 2574-3414. Dostupné z DOI: 10.1146/annurev-biodatasci-031121-103035.
8. ARMENGAUD, Jean. Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. *Environmental Microbiology*. 2023, roč. 25, č. 1, s. 115–125. ISSN 1462-2912. Dostupné z DOI: 10.1111/1462-2920.16238.
9. CLISH, Clary B. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies*. 2015-09-24, roč. 1, č. 1. ISSN 2373-2865. Dostupné z DOI: 10.1101/mcs.a000588.

10. SHAKYA, Migun; LO, Chien-Chi; CHAIN, Patrick S. G. Advances and Challenges in Metatranscriptomic Analysis. *Frontiers in Genetics* [online]. 2019-9-25, roč. 10 [cit. 2023-11-12]. ISSN 1664-8021. Dostupné z DOI: 10.3389/fgene.2019.00904.

11. SEGATA, Nicola; BOERNIGEN, Daniela; TICKLE, Timothy L; MORGAN, Xochitl C; GARRETT, Wendy S; HUTTENHOWER, Curtis. Computational meta'omics for microbial community studies. *Molecular Systems Biology*. 2013, roč. 9, č. 1. ISSN 1744-4292. Dostupné z DOI: 10.1038/msb.2013.22.

12. AITMANAITĖ, Lina; ŠIRMONAITIS, Karolis; RUSSO, Giancarlo. Microbiomes, Their Function, and Cancer: How Metatranscriptomics Can Close the Knowledge Gap. *International Journal of Molecular Sciences* [online]. 2023, roč. 24, č. 18 [cit. 2023-11-22]. ISSN 1422-0067. Dostupné z DOI: 10.3390/ijms241813786.

13. *RNA LEXICON Chapter #4 – RNA Extraction and Quality Control* [online]. 2021. [cit. 2023-11-22]. Dostupné z: https://www.lexogen.com/rna-lexicon-rna-extraction-and-quality-control/.

14. *An introduction to Next-Generation Sequencing Technology* [online]. 2017. [cit. 2023-11-22]. Dostupné z: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

15. HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. *Genomics* [online]. 2016, roč. 107, č. 1, s. 1–8 [cit. 2023-11-22]. ISSN 08887543. Dostupné z DOI: 10.1016/j.ygeno.2015.11.003.

16. *Illumina Sequencing Technology* [online]. [cit. 2023-12-29]. Dostupné z: https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.

17. NGUYEN, JULIE. *BRIDGE AMPLIFICATION SEQUENCING* [online]. 2021. [cit. 2023-11-30]. Dostupné z: https://apollo-institute.org/bridge-amplification-sequencing/.

18. ABUBUCKER, Sahar; SEGATA, Nicola; GOLL, Johannes; SCHUBERT, Alyxandria M.; IZARD, Jacques; CANTAREL, Brandi L.; RODRIGUEZ-MUELLER, Beltran; ZUCKER, Jeremy; THIAGARAJAN, Mathangi; HENRISSAT, Bernard; WHITE, Owen; KELLEY, Scott T.; METHÉ, Barbara; SCHLOSS, Patrick D.; GEVERS, Dirk; MITREVA, Makedonka; HUTTENHOWER, Curtis; EISEN, Jonathan A. Metabolic Reconstruction for Metagenomic Data and

Its Application to the Human Microbiome. *PLoS Computational Biology* [online]. 2012-6-13, roč. 8, č. 6 [cit. 2023-12-13]. ISSN 1553-7358. Dostupné z DOI: 10.1371/journal.pcbi.1002358.

19. SINGHAL, Neelja; KUMAR, Manish; KANAUJIA, Pawan K.; VIRDI, Jugsharan S. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology* [online]. 2015-08-05, roč. 6 [cit. 2023-12-19]. ISSN 1664-302X. Dostupné z DOI: 10.3389/fmicb.2015.00791.

20. PRADHAN, Shalini; APAYDIN, Sinem; BUCEVIČIUS, Jonas; GERASIMAITĖ, Rūta; KOSTIUK, Georgij; LUKINAVIČIUS, Gražvydas. Sequence-specific DNA labelling for fluorescence microscopy. *Biosensors and Bioelectronics* [online]. 2023, roč. 230 [cit. 2023-12-19]. ISSN 09565663. Dostupné z DOI: 10.1016/j.bios.2023.115256.

21. *How TaqMan Assays Work* [online]. [cit. 2023-12-19]. Dostupné z: https://www.thermofisher.com/cz/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learning-center/real-time-pcr-basics/how-taqman-assays-work.html.

22. MÉSZÁROS, Éva. *How does qPCR work: SYBR® Green vs TaqMan®* [online]. 2022. [cit. 2023-12-19]. Dostupné z: https://www.integra-biosciences.com/global/en/blog/article/how-does-qpcr-work-sybrr-green-vs-taqmanr.

23. BEGHINI, Francesco; MCIVER, Lauren J; BLANCO-MÍGUEZ, Aitor; DUBOIS, Leonard; ASNICAR, Francesco; MAHARJAN, Sagun; MAILYAN, Ana; MANGHI, Paolo; SCHOLZ, Matthias; THOMAS, Andrew Maltez; VALLES-COLOMER, Mireia; WEINGART, George; ZHANG, Yancong; ZOLFO, Moreno; HUTTENHOWER, Curtis; FRANZOSA, Eric A; SEGATA, Nicola. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *ELife*. 2021-05-04, roč. 10. ISSN 2050-084X. Dostupné z DOI: 10.7554/eLife.65088.

24. *BioBakery Workflows* [online]. [cit. 2023-12-13]. Dostupné z: https://huttenhower.sph.harvard.edu/biobakery_workflows/.

25. EWELS, Philip; MAGNUSSON, Måns; LUNDIN, Sverker; KÄLLER, Max. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* [online]. 2016-10-01, roč. 32, č. 19, s. 3047–3048 [cit. 2024-05-25]. ISSN 1367-4811. Dostupné z DOI: 10.1093/bioinformatics/btw354.

26. BLANCO-MÍGUEZ, Aitor; BEGHINI, Francesco; CUMBO, Fabio; MCIVER, Lauren J.; THOMPSON, Kelsey N.; ZOLFO, Moreno; MANGHI, Paolo; DUBOIS, Leonard; HUANG, Kun D.; THOMAS, Andrew Maltez; NICKOLS, William A.; PICCINNO, Gianmarco; PIPERNI, Elisa; PUNČOCHÁŘ, Michal; VALLES-COLOMER, Mireia; TETT, Adrian; GIORDANO, Francesca; DAVIES, Richard; WOLF, Jonathan; BERRY, Sarah E.; SPECTOR, Tim D.; FRANZOSA, Eric A.; PASOLLI, Edoardo; ASNICAR, Francesco; HUTTENHOWER, Curtis; SEGATA, Nicola. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology* [online]. 2023, roč. 41, č. 11, s. 1633–1644 [cit. 2023-12-13]. ISSN 1087-0156. Dostupné z DOI: 10.1038/s41587-023-01688-w.

27. *Babraham Bioinformatics* [online]. [cit. 2024-05-25]. Dostupné z: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

28. WINGETT, Steven W.; ANDREWS, Simon. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research* [online]. 2018, roč. 7 [cit. 2024-05-25]. ISSN 2046-1402. Dostupné z DOI: 10.12688/f1000research.15931.2.

29. CHEN, Shifu. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *IMeta* [online]. 2023, roč. 2, č. 2 [cit. 2024-05-25]. ISSN 2770-596X. Dostupné z DOI: 10.1002/imt2.107.

30. *Streamlit Forum* [online]. 2021. [cit. 2024-05-25]. Dostupné z: https://discuss.streamlit.io/t/backend-workings-of-streamlit/11834.

Elektronické přílohy

Veškeré zdrojové kódy a příklad naleznete na webové stránce GitHub pod tímto odkazem: https://github.com/HonzaHyl/VisualizationDashboard.