

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2023

Martin Spanakis



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

**METAGENOMICKÁ ANALÝZA STŘEVNÍHO
MIKROBIOMU U ZVÍŘAT DLE TYPU DIETY**

METAGENOMIC ANALYSIS OF ANIMAL GUT MICROBIOME BASED ON DIET

BAKALÁRSKA PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

Martin Nikolaos Spanakis

VEDÚCI PRÁCE

ADVISOR

Ing. Vojtěch Bartoň

BRNO 2023

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Student: Martin Spanakis

ID: 229154

Ročník: 3

Akademický rok: 2022/23

NÁZEV TÉMATU:

Metagenomická analýza střevního mikrobiomu u zvířat dle typu diety

POKYNY PRO VYPRACOVÁNÍ:

1) Provedte literární rešerši na téma sekvenace mikrobiomu a jeho složení. 2) Sestavte z dostupných zdrojů vhodný dataset a vzorky popište. 3) Seznamte se s nástrojem QIIME2 pro analýzu mikrobiomu. 4) Data vhodně upravte a proveďte předzpracování. 5) Zpracujte vytvořený dataset a určete složení vzorků a diverzitu. 6) Srovnajte výsledky analýzy u druhů dle typu diety. 7) Výsledky vyhodnoťte a diskutujte závěry.

DOPORUČENÁ LITERATURA:

GUO, Xiurong, Hui LEI, Kailian ZHANG, Famin KE a Can SONG. Diversification of animal gut microbes and NRPS gene clusters in some carnivores, herbivores and omnivores. 2020, 34(1), 1280-1287. ISSN 1310-2818. doi:10.1080/13102818.2020.1835536

YOUNGBLUT, Nicholas D., Georg H. REISCHER, William WALTERS, Nathalie SCHUSTER, Chris WALZER, Gabrielle STALDER, Ruth E. LEY a Andreas H. FARNLEITNER. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. Nature Communications. 2019, 10(1). ISSN 2041-1723. doi:10.1038/s41467-019-10191-3

Termín zadání: 6.2.2023

Termín odevzdání: 29.5.2023

Vedoucí práce: Ing. Vojtěch Bartoň

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

UPOZORNĚNÍ:

Autor bakalářské práce nesmí při vytváření bakalářské práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

ABSTRAKT

Táto bakalárska práca sa zaoberá sekvenáciou mikrobiómu a jeho zložením. Zameriava sa na črevný mikrobióm zvierat, ktorý sa líši u zvierat s odlišnou stravou. Práca popisuje teoretické poznatky o metagenomickej analýze mikrobiómu, ako postupy pri odbere vzoriek, rôzne metódy sekvenácie a spracovanie získaných dát. V praktickej časti práce je obsiahnutá príprava datasetu, ktorá zahŕňa zhromaždenie dát a ich prípravu na nasledujúcu metagenomickú analýzu. Výsledkom práce je taxonomické zaradenie bakteriálnych druhov vo vzorkách a analýza ich diverzity podľa typu stravy jednotlivých živočíchov.

KLÚČOVÉ SLOVÁ

mikrobióm, metagenóm, sekvenácia, DNA, vzorka, 16S rRNA, FastQC, Fastp, QIIME2, taxonómia, diverzita

ABSTRACT

This bachelor's thesis deals with the sequencing of the microbiome and its composition. It focuses on the gut microbiome of animals, which differs among animals with different diets. The work describes theoretical knowledge about metagenomic analysis of the microbiome, such as sampling procedures, various sequencing methods and data processing. The practical part of the work includes the preparation of the dataset, which includes the collection of data and their preparation for the following metagenomic analysis. The result of the work is the taxonomic classification of bacterial species in the samples and the analysis of their diversity according to the type of diet of individual animals.

KEYWORDS

microbiome, metagenome, sequencing, DNA, sample, 16S rRNA, FastQC, Fastp, QIIME2, taxonomy, diversity

SPANAKIS, Martin Nikolaos. *Metagenomická analýza střevního mikrobiomu u zvířat dle typu diety*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2023, 66 s. Bakalárska práca. Vedúci práce: Ing. Vojtěch Bartoň

Vyhlásenie autora o pôvodnosti diela

Meno a priezvisko autora: Martin Nikolaos Spanakis
VUT ID autora: 229154
Typ práce: Bakalárska práca
Akademický rok: 2022/23
Téma závěrečné práce: Metagenomická analýza střevního mikrobiomu u zvířat dle typu diety

Vyhlasujem, že svoju záverečnú prácu som vypracoval samostatne pod vedením vedúcej/cého záverečnej práce, s využitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú všetky citované v práci a uvedené v zozname literatúry na konci práce.

Ako autor uvedenej záverečnej práce ďalej vyhlasujem, že v súvislosti s vytvorením tejto záverečnej práce som neporušil autorské práva tretích osôb, najmä som nezasiahol nedovoleným spôsobom do cudzích autorských práv osobnostných a/alebo majetkových a som si plne vedomý následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona Českej republiky č. 121/2000 Sb., o práve autorskom, o právach súvisiacich s právom autorským a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov, vrátane možných trestnoprávných dôsledkov vyplývajúcich z ustanovenia časti druhej, hlavy VI. diel 4 Trestného zákonníka Českej republiky č. 40/2009 Sb.

Brno

.....
podpis autora*

*Autor podpisuje iba v tlačenej verzii.

POĎAKOVANIE

Rád by som poďakoval vedúcemu bakalárskej práce Ing. Vojtěchovi Bartoňovi za odborné vedenie, konzultácie, trpezlivosť a podnetné návrhy k práci.

Obsah

Úvod	19
1 Genóm a mikrobióm	21
1.1 Genetika	21
1.2 Genomika	22
1.3 Metagenomika	23
1.4 Mikrobióm	23
1.4.1 Zloženie mikrobiómu	24
1.4.2 Mikrobióm tráviaceho traktu	24
1.4.3 Mikrobióm ako súčasť tiel živočíchov	24
2 Metódy analýzy mikrobiómu	25
2.1 Optická mikroskopia	25
2.2 Hmotnostná spektroskopia	25
2.3 Sekvenácia mikrobiómu	25
2.3.1 Úvod do sekvenácie mikrobiómu	25
2.3.2 Metódy sekvenácie	30
2.4 Taxonomická analýza 16S	35
2.4.1 OTU	35
2.4.2 ASV	35
2.4.3 Taxonomické úrovne	36
2.5 Analýza diverzity	36
2.5.1 Alfa diverzita	36
2.5.2 Beta diverzita	37
2.5.3 Gama diverzita	37
3 QIIME2	39
3.1 Využitie QIIME2	39
3.2 Dátové formáty v QIIME2	39
3.2.1 QIIME2 artifact	39
3.2.2 QIIME2 vizualizácia	40
4 Metagenomická analýza	41
4.1 Dataset	41
4.1.1 Pôvod datasetu	41
4.1.2 Štruktúra a obsah datasetu	41
4.2 Predspracovanie dát	43
4.2.1 Kontrola kvality pred úpravou	43

4.2.2	Zastrihnutie a filtrácia	48
4.2.3	Kontrola kvality po úprave - vyhodnotenie	48
4.3	Postup analýzy zloženia vzoriek	50
4.4	Výsledky analýzy diverzity	52
4.4.1	Alfa diverzita	52
4.4.2	Beta diverzita	55
Diskusia		57
Záver		59
Literatúra		61

Zoznam obrázkov

1.1	Príklad autozomálne recesívneho dedenia [1]	22
2.1	Schéma využitia primerov k zameraniu variabilných úsekov	28
2.2	Schéma ribozómu a 16S rRNA [2]	29
2.3	Štruktúra 16S rRNA [3]	30
2.4	Schéma prípravy knihovne metódou <i>shotgun</i>	31
2.5	Schéma prípravy knihovne metódou <i>amplicon</i>	31
2.6	Schéma PCR [4]	32
2.7	Skladanie DNA fragmentov	33
2.8	Porovnávanie nasekvenovaných úsekov so sekvenciami z databáz	34
3.1	Bloková schéma postupu spracovania dát pomocou QIIME2 [5]	40
4.1	Graf počtu unikátnych a duplikovaných readov v jednotlivých vzorkách pred úpravou	44
4.2	Závislosť kvality na pozícii báz (vľavo: ukážka kvality všetkých vzoriek, vpravo: ukážka kvality jednej vzorky)	45
4.3	Graf obsahu adaptérov v readoch pred úpravou	45
4.4	Graf výskytu báz na jednotlivých pozíciách vo vzorke SRR11773851_1	46
4.5	Porovnanie obsahu báz vo vzorkách SRR11773852_1, SRR11773853_1, SRR11773854_1	47
4.6	Graf počtu unikátnych a duplikovaných readov v jednotlivých vzorkách po úprave	49
4.7	Graf kvality vzoriek po úprave	50
4.8	Graf obsahu adaptérov v readoch po úprave	50
4.9	Výsledná kompozícia vzoriek [6]	52
4.10	Graf analýzy hlavných komponentov (vizualizácia podľa mena živočícha) [6]	53
4.11	Graf analýzy hlavných komponentov (vizualizácia podľa potravy živočícha) [6]	55

Zoznam tabuliek

2.1	Stavba prokaryotického ribozómu [7]	27
2.2	Stavba eukaryotického ribozómu [7]	27
2.3	Príklad zaradenia do taxonomických úrovní	36
4.1	Štruktúra datasetu	42

Úvod

Výskum črevnej mikroflóry zvierat sa stáva čoraz viac populárnym, pretože sa ukazuje, že mikroorganizmy žijúce v tráviacom trakte majú kľúčový vplyv na zdravie a výživu hostiteľa. Zloženie a funkcia mikrobiómov je úzko spojená s typom stravy, ktorú zvieratá konzumujú. Jedným z najpočetnejších mikrobiómov je črevný mikrobióm, ktorý sa u organizmov vytvára už v prenatálnom období, no najväčší rozmach nastáva prechodom na prirodzenú stravu daného jedinca (napr. prechod z výživy z materského mlieka na plnohodnotnú stravu). Črevný mikrobióm je kľúčovou súčasťou tiel organizmov, pretože sa významne podieľa na dôležitých metabolických, trofických či detoxikačných procesoch.

Štúdium rôznych druhov mikrobiómov sa stalo neoddeliteľnou súčasťou výskumov v oblasti medicíny, farmaceutického priemyslu či ekológie. Na zistenie zloženia jednotlivých komunít sa používa špecializovaná metóda sekvenácie, ktorá sa zameriava na gén kódujúci 16S rRNA reťazec. Vďaka hypervariabilným úsekom, ktoré tento gén obsahuje je možné efektívne a s dostatočnou presnosťou odlíšiť jednotlivé druhy mikroorganizmov obsiahnuté vo vzorke.

Cieľom bakalárskej práce je teoretické oboznámenie sa s tematikou a praktická časť, ktorá spočíva v príprave kvalitného datasetu a nasledujúcej analýze jeho zloženia. Samotná príprava datasetu zahŕňa zhromaždenie vhodných voľne dostupných dát, kontrolu ich kvality a prispôbenie na analýzu ich obsahu. V nadväzujúcej časti sa práca venuje metagenomickej analýze vzoriek pomocou nástroja QIIME2 a taxonomickému zaradeniu bakteriálnych druhov vo vzorkách. Z výsledných dát je napokon odvodená kompozičná diverzita vzoriek.

1 Genóm a mikrobióm

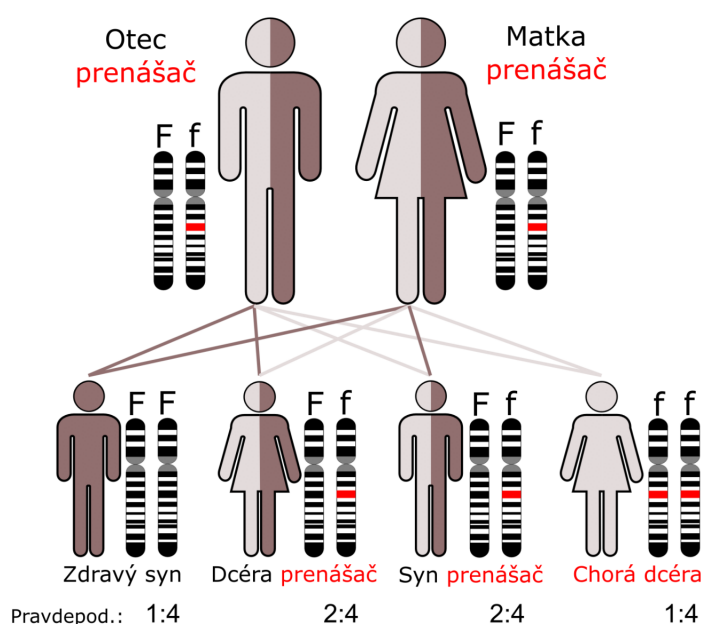
1.1 Genetika

Genetika je veda, ktorá skúma dedičnosť a gény, ktorými je podmienená. Gény sú predpokladom pre prenos genetickej informácie z generácie na generáciu. Genetická informácia je zakódovaná poradím nukleotidov v reťazci DNA, ktorý sa nachádza v každej bunke každého organizmu. Expresiou génu je tvorba proteínov, ktoré sa podieľajú na biochemických procesoch v každej bunke. Následný efekt sa prejaví v rámci celého organizmu.

Genetická charakteristika jedinca sa nazýva **genotyp**. Každý jedinec má svoju genetickú výbavu diploidnú (výnimkou sú gaméty, ktoré sú haploidné) - genotyp každého génu sa skladá z dvoch alel, z ktorých jedna pochádza od materského a druhá od otcovského jedinca. Následná kombinácia zdedených alel sa charakteristicky prejaví. Súbor týchto charakteristických znakov jedinca sa nazýva **fenotyp**, patria sem znaky morfologické (hmotnosť, výška, tvar končatín a i.), fyziologické (krvný tlak, hladina glukózy v krvi a i.) a biochemické (tvorba enzýmov a i.). Platí, že dvaja jedinci s rovnakým genotypom nemusia mať rovnaký fenotyp a naopak, dvaja jedinci s rovnakým fenotypom, nemusia mať rovnaký genotyp.

Napríklad u fenylketonúrie, ktorá patrí medzi autozomálne recesívne ochorenia (k mutácii dochádza na géne, ktorého lokus je na autozómoch a patologická je recesívna alela), existujú dva typy alel: fyziologická dominantná (F) a patologická recesívna (f). Môžu nastať tri situácie (obr. 1.1):

- dieťa zdedí od oboch rodičov zmutovanú recesívnu alelu (ff) - ochorenie sa prejaví,
- dieťa zdedí od jedného rodiča zmutovanú recesívnu alelu a od druhého fyziologickú dominantnú alelu (Ff) - ochorenie sa neprejaví, ale stáva sa prenášačom ochorenia, keďže jeho genotyp obsahuje patologickú alelu,
- dieťa zdedí od oboch rodičov fyziologickú dominantnú alelu (FF) - ochorenie sa neprejaví a nie je ani prenášačom.



Obr. 1.1: Príklad autozomálne recesívneho dedenia [1]

Keďže poradie nukleotidov v DNA je u každého jedinca unikátne, tvorba proteínov sa u každého o niečo líši. Preto zrejme nepoznáme dvoch ľudí, ktorí by vyzerali úplne rovnako, museli by mať totiž rovnaký fenotyp u všetkých génov, ktoré kódujú morfológické znaky (je to veľmi nepravdepodobné). V skutočnosti je ale zhoda u dvoch nezávislých ľudí asi 99,9% v rámci celého ľudského genómu, čo je asi 3,2 miliardy párov báz a zvyšná desatina percenta zapríčiňuje odlišnosť.[8]

1.2 Genomika

Jednou z podkategórií genetiky je genomika, novodobý obor, ktorý sa začal rozvíjať až v 21. storočí. Rozvoju genomiky predchádzali významné pokroky v sekvenčných metódach (Roche 454 a následne Illumina), bioinformatike a analýze genómu. Odhalenie štruktúry celého genómu pomocou týchto moderných technológií a postupov dnes už nepredstavuje žiadny problém. Genomika je veda zaoberajúca sa štúdiou genómu, teda kompletného genetického materiálu organizmu. Genóm obsahuje všetky gény, ktoré riadia dedičné vlastnosti organizmu, a tiež množstvo iných sekvencií DNA, ktoré môžu ovplyvniť jeho vývoj a funkciu. Genomika sa zaoberá nielen samotným sekvenčovaním DNA, ale aj analýzou genetických dát, aby sme mohli pochopiť, ako sa jednotlivé gény podieľajú na rôznych procesoch a funkciách organizmu. Genomika tiež zahŕňa výskum komplexných ochorení, na ktorých výskyt

podľa doterajších štúdií významne viac vplyva vzájomné pôsobenie prostredia a genetických faktorov ako samostatné gény jednotlivca. Medzi tieto ochorenia patria napríklad kardiovaskulárne ochorenia, astma, diabetes, rakovina a i. Výsledkom týchto štúdií sú nové spôsoby v liečbe a diagnostike komplexných ochorení či pokroky v umelej syntéze biologicky aktívnych látok (hormóny, neurotransmitery, protilátky, enzýmy). Je možné, že moderný rozvoj genomiky nás privedie k objavom, ktoré budú v budúcnosti predstavovať veľmi významný milník k pochopeniu jedného z najzložitejších biologických systémov, ako je mozog.[9]

1.3 Metagenomika

V porovnaní s genomikou, ktorá sa venuje výskumu genómu iba jedného organizmu, predmetom výskumu metagenomiky je súbor genómov organizmov z jedného prostredia. Tento postup umožňuje presnejšie analyzovať vplyv prostredia na mikrobiálnu diverzitu. Postupy tohoto odvetvia genomiky majú významné uplatnenie v medicíne (transplantácia mikrobiómu, analýza mikrobiálnej diverzity ľudskej pokožky [10]), poľnohospodárstve (analýza premnožených mikroorganizmov vo vode), genetike či vo vývoji biotechnológií (identifikácia bioaktívnych molekúl [11]).[12][13]

1.4 Mikrobióm

Mikrobióm je súbor všetkých mikroorganizmov (eukaryoty, baktérie, archeóny a vírusy), ktoré žijú v určitom prostredí.

Výskyt týchto komunít je prirodzený v každom živom organizme. V prípade rovnováhy tejto skupiny mikroorganizmov je jej vplyv na hostiteľa priaznivý, avšak v prípade premnoženia či výskytu patologických mikroorganizmov môže byť efekt patologický, čo môže v horších prípadoch viesť až k rozvinutiu ťažších ochorení.[14][15]

Počet mikroorganizmov obsiahnutých v celom mikrobióme tela živočíchov je približne rovnaký ako počet telových buniek hostiteľa, no u niektorých organizmov môže byť ich počet aj desaťnásobne vyšší. Absolútna väčšina je však súčasťou črevnej mikrobiómy. Ak by sme však porovnávali veľkosť genetickej informácie (počet rôznych génov) medzi hostiteľom a jeho mikrobiómom, pomer by sa mohol vyšplhať až na 1:100 v prospech mikrobiómu.

Rozmiestnenie mikroorganizmov po tele hostiteľa je výrazne heterogénne. V rôznych častiach tela sa vyskytujú mikroorganizmy rôznych druhov a v rôznom počte. Mikrobióm môžeme rozdeliť podľa toho, v ktorej časti tela sa nachádza, napríklad kožný, orálny, vaginálny, najbohatší črevný a i.[16]

1.4.1 Zloženie mikrobiómu

Tráviaci trakt organizmov obsahuje obrovské množstvo mikroorganizmov. Väčšina črevných mikroflór stavovcov obsahuje 500-1000 rôznych druhov baktérií, z ktorých viac ako 98% spadá pod kmeň *Bacteroidetes* alebo *Firmicutes*. Zreteľné rozdiely v zložení mikrobiómov jednotlivých stavovcov sú spôsobené rozdielmi pri narodení, v metódach kŕmenia a stravovacích návykoch, v životnom štýle, vo veku, v genetike hostiteľa, v antibiotikách či patogénoch, ktorým bol jedinec vystavený.[17]

1.4.2 Mikrobióm tráviaceho traktu

Tráviaci trakt je dokonalé prostredie pre vznik veľkej dynamickej komunity baktérií. Najbohatší súbor týchto komunít nazývame črevná mikroflóra. Aj keď značná časť z nich ešte nebola podrobne preskúmaná, s určitosťou môžeme povedať, že efekt týchto baktérií je pre správne fungovanie hostiteľského organizmu kľúčový. Medzi jej hlavné funkcie patrí podpora vstrebávania živín, úspora energie a absorbovateľných živín – **metabolické procesy**, pozitívne vplýva na funkčnosť tráviaceho epitelu – **trofický efekt** a celková ochrana kolonizovaného organizmu pred cudzími mikróbmami – **detoxikácia**. [18]

1.4.3 Mikrobióm ako súčasť tel živých organizmov

Telo živočícha s jeho črevnou mikroflórou a ostatnými mikroorganizmami tvorí komplexný ekosystém, akýsi „superorganizmus“, ktorý aktívne vykonáva mnohé biologické procesy [19]. Genóm samotného živočícha spolu s genómami všetkých mikroorganizmov v ňom by sme mohli považovať za **metagenóm** tohto živočícha [20].

2 Metódy analýzy mikrobiómu

Komunity mikroorganizmov sú neoddeliteľnou súčasťou každého ekosystému na Zemi. Potreba ich skúmania odštartovala rozvoj metód na detekciu a charakterizáciu baktérií, archeónov, vírusov a prokaryotov.

2.1 Optická mikroskopia

Jednou z prvých používaných metód bola optická mikroskopia, ktorá sa rozvíjala za účelom potreby „videnia“ týchto mikroorganizmov - obsah vzorky sa jednoducho pozoroval pod mikroskopom, ktorý však nebol dostačujúci k odhaleniu štruktúr na molekulárnej úrovni. To následne odštartovalo prudký záujem o vývoj ďalších technológií zameraných na klasifikáciu nepreskúmaných druhov, ich bunkového metabolizmu, vplyvu na prostredie, v ktorom žijú, a celkovo na biosféru.[21]

2.2 Hmotnostná spektroskopia

Ďalšou efektívnou metódou, využívanou na skúmanie zloženia mikrobiómu je hmotnostná spektroskopia. Je to metóda analytickej chémie, ktorej princíp spočíva v porovnávaní jednotlivých zložiek na základe pomeru m/Q , kde m je hmotnosť a Q je náboj zložky. Používa sa na určenie zloženia látok, ktoré sú typickými metabolitmi určitých mikroorganizmov, čím sa dokazuje ich prítomnosť vo vzorke.

2.3 Sekvenácia mikrobiómu

Sekvenácia je vrcholom analýzy genetických dát. Je to proces, ktorým sme schopní určiť primárnu štruktúru biopolymérov. Výsledkom sekvenácie je symbolický lineárny kód nazývaný **sekvencia**. Sekvenácia umožňuje presnú identifikáciu poradia nukleotidov v reťazcoch nukleových kyselín obsiahnutých vo vzorke, ktorá bola získaná zo skúmaného prostredia. Pomocou moderných sekvenačných metód dnes už dokážeme identifikovať konkrétne druhy mikroorganizmov, ktoré v tomto prostredí žijú.

2.3.1 Úvod do sekvenácie mikrobiómu

Jednou z hlavných funkcií každej bunky je expresia génu - **proteosyntéza**. Tento dej sa skladá z dvoch fáz, a to z transkripcie a translácie.

Transkripcia je súbor procesov, pri ktorých za prítomnosti potrebných enzýmov dochádza k prepisu DNA do mRNA podľa pravidiel komplementarity báz a

následným posttranskripčným úpravám vzniknutej mRNA (tie sa však vykonávajú len u prokaryotických buniek). Transkripčia prebieha u eukaryotických buniek v jadre a u prokaryotických buniek voľne v cytoplazme, keďže nemajú jadro.

Translácia je súbor procesov, pri ktorých dochádza k prekladu vzniknutej mRNA do poradia aminokyselín v polypeptidovom reťazci. Výsledný produkt je proteín. Translácia prebieha na ribozómoch endoplazmatického retikula u eukaryotických buniek a voľne v cytoplazme u prokaryotických buniek.

Ribozóm

Ribozóm je bunková štruktúra zložená zo špecializovanej rRNA a proteínov. Pomocou röntgenovej kryštalografie bolo zistené, že ribozomálne proteíny sa nevyskytujú v ribozómoch na mieste syntézy polypeptidového reťazca, čo znamená, že tieto proteíny sa nepodieľajú priamo na vzniku peptidových väzieb, ale fungujú skôr ako skelet ribozómov, ktorý zvyšuje schopnosť rRNA syntetizovať bielkoviny. Reťazce rRNA a proteíny sú usporiadané do dvoch celkov - malá a veľká ribozomálna podjednotka. Tieto dve podjednotky do seba zapadajú a pracujú ako „prekladač“ mRNA do polypeptidového reťazca.

Ribozomálne podjednotky u prokaryotických a eukaryotických buniek sú funkčne a štruktúrne veľmi podobné, líšia sa len vo veľkosti (počte nukleotidov v jednotlivých reťazcoch rRNA a počte proteínov medzi nimi). (tab. 2.1, 2.2)

Na charakterizáciu jednotlivých ribozomálnych podjednotiek a úsekov rRNA sa používa Svedbergov koeficient - miera sedimentácie, ktorá vyjadruje pomer medzi rýchlosťou pohybu biologického materiálu v centrifugačnej skúmavke a centrifugačným zrýchlením (veľkosť v nanometroch by v charakterizácii ribozomálnych podjednotiek nemala dostatočnú výpovednú hodnotu, pretože ribozómy majú výrazne nepravidelnú štruktúru). Jednotkou tejto veličiny je *S* (Svedberg), je to však ekvivalent jednotky času definovaný ako presne 10^{-13} sekundy. Hodnota sedimentačného koeficientu je ovplyvnená hmotnosťou, hustotou a tvarom skúmaného materiálu. Napríklad, ak bude mať ribozóm hodnotu sedimentačného koeficientu 26S (26×10^{-13} s), znamená to, že sa bude pohybovať rýchlosťou 26 mikrometrov za sekundu ($26 \times 10^{-6} \text{ m s}^{-2}$) pod vplyvom centrifugačného zrýchlenia miliónkrát väčšieho ako gravitačné zrýchlenie Zeme ($9,81 \times 10^6 \text{ m s}^{-2}$). [7]

Ribozóm	Podjednotka	rRNA	Počet proteínov
70S	50S	23S (2904 nt)	31
		5S (120 nt)	
	30S	16S (1542 nt)	21

Tab. 2.1: Stavba prokaryotického ribozómu [7]

Ribozóm	Podjednotka	rRNA	Počet proteínov
80S	60S	28S (4718 nt)	49
		5.8S (160 nt)	
		5S (120 nt)	
	40S	18S (1874 nt)	33

Tab. 2.2: Stavba eukaryotického ribozómu [7]

Ribozómy archeónov sú rozmerovo veľmi podobné prokaryotickým, avšak na sekvencijnej úrovni sa výrazne podobajú na ribozómy eukaryotické.

16S rRNA gén

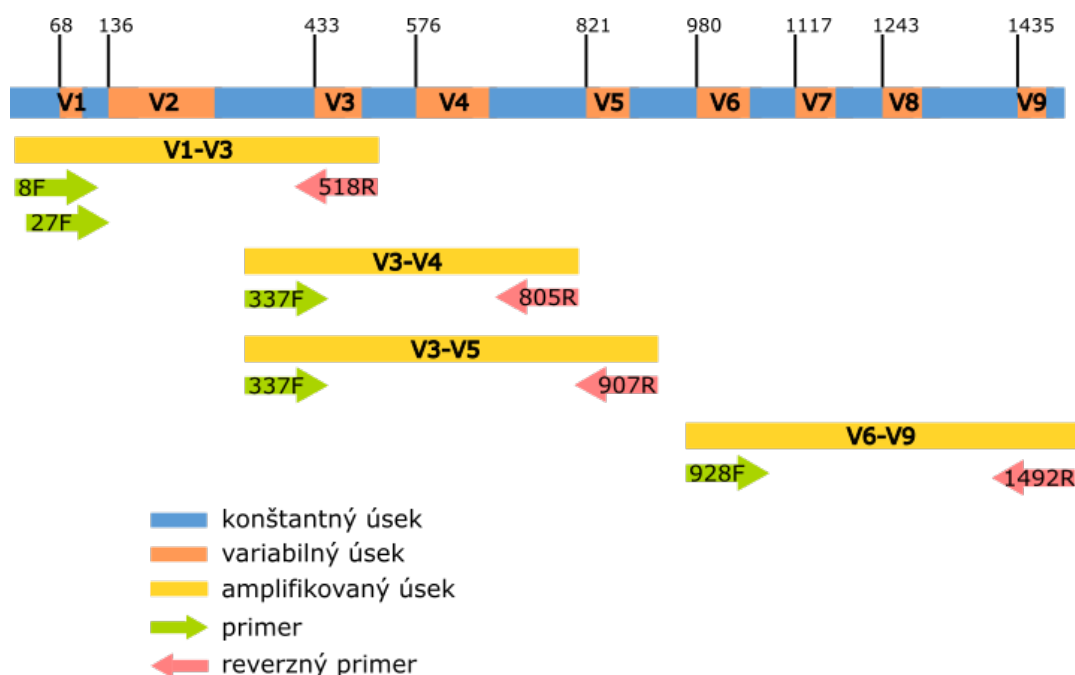
16S rRNA je nukleová kyselina, ktorá je stavebnou a funkčnou súčasťou malých podjednotiek prokaryotických ribozómov (30S). Jej hlavnými funkciami sú: určovanie polohy ribozomálnych proteínov v ribozóme, zabezpečenie väzby medzi malou a veľkou ribozomálnou podjednotkou počas translácie (úsek 16S rRNA interaguje s 23S rRNA) a zabezpečenie správnej iniciácie translácie - 3' koniec 16S rRNA obsahuje tzv. anti-Shine-Dalgarnovú sekvenciu, komplementárnu k tzv. Shine-Dalgarnovej sekvencii, ktorá sa nachádza v mRNA pred iniciačným kodónom AUG - pred samotnou iniciáciou translácie dochádza k vzniku vodíkových väzieb medzi týmito dvomi sekvenciami, čo je signálom pre ribozóm, že nasleduje kodón AUG, a teda translácia môže začať, (Shine-Dalgarnova sekvencia tiež zabezpečuje, aby bolo odlíšené, kedy kodón AUG kóduje aminokyselinu uprostred polypeptidového reťazca a kedy vystupuje ako start-kodón).[22][23]

Samotná sekvencia 16S rRNA je zakódovaná v DNA bakteriálnych buniek. Je prítomná takmer u všetkých druhov baktérií, práve preto je táto oblasť genetickej výbavy baktérií cieľom pri metagenomickej analýze mikrobiómu. 16S rRNA gén sa tiež nazýva ako „molekulárne hodiny“, pretože sa využíva na sledovanie miery mutácií v čase pre odhadnutie obdobia, kedy sa daný organizmus fylogeneticky odtrhol

od svojho predka. Tento gén má vysoké opodstatnenie v rekonštrukcii fylogenetických stromov, pretože počas evolúcie dochádza k zmenám v týchto úsekoch len zriedkavo (s rastúcim počtom zmien rastie aj zložitosť rekonštrukcie fylogenetických stromov - je náročnejšie odhadnúť „čo z čoho vzniklo“). Priekopníkom vo využití 16S rRNA génu vo fylogenetike bol Carl Woese, ktorý v roku 1977 prvýkrát definoval archeóny.[24][25]

16S rRNA gén sa skladá z 1542 nukleotidov usporiadaných do viacerých úsekov, ktoré sa delia na dva typy: konštantné a hypervariabilné (obr. 2.2).

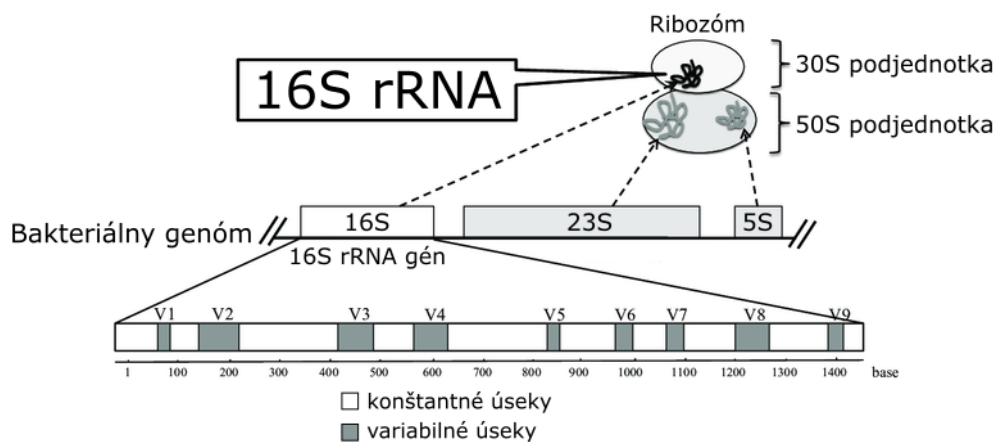
Konštantné úseky sú veľmi stále a podobné medzi rôznymi druhmi baktérií. Táto skutočnosť sa využíva v procese amplifikácie v PCR, kde je nutné zamerať amplifikované oblasti génu (hypervariabilné úseky). Keďže tieto konštantné úseky sa len vzácné menia u rôznych druhov baktérií, v PCR je možné využívať univerzálne primery (primery fungujú rovnako u takmer všetkých druhov baktérií), ktoré zaručia pomerne presnú lokalizáciu amplifikovaného úseku - určujú presné miesto na reťazci DNA, kde začne DNA polymeráza syntetizovať antiparalelný reťazec k obom reťazcom v procese PCR (obr. 2.1).[22]



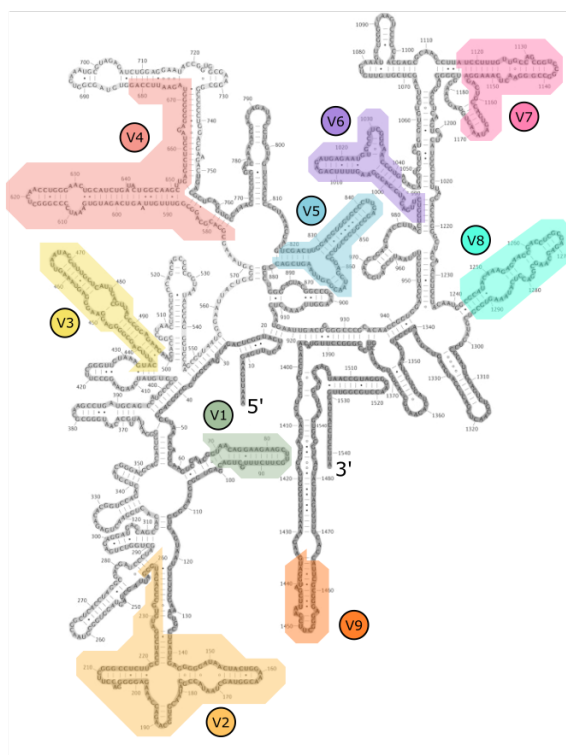
Obr. 2.1: Schéma využitia primerov k zameraniu variabilných úsekov

Hypervariabilné úseky sú v porovnaní s konštantnými veľmi špecifické pre konkrétny druh baktérie, preto sa tieto úseky využívajú k identifikácii konkrétneho druhu baktérie. 16S rRNA gén obsahuje 9 hypervariabilných úsekov (zn. V1 - V9), dlhých 30-100 párov nukleotidov. Miera variability medzi týmito úsekmi ale tiež

nie je konštantná, medzi druhmi na vyššej taxonomickej úrovni je variabilita menej výrazná ako na nižšej. Metagenomická analýza sa spravidla zameriava na určitú kombináciu hypervariabilných úsekov, väčšinou sú to dvojice alebo trojice hypervariabilných úsekov (V6-V8, V7-V8, V3-V4), no možná je aj analýza samostatného úseku (v prípade, že tento hypervariabilný úsek poskytuje dostatočnú informáciu k taxonomickej charakterizácii) či celogenómová sekvenácia 16S rRNA génu (je však zbytočne náročná v prípade, že je potrebné iba charakterizovať druhy baktérií vo vzorke). Úseky V1 a V9 sa v metagenomickej analýze nevyužívajú, pretože v početných výskumoch vykazovali nadmerne vysokú variabilitu aj u druhov, kde by sa takáto variabilita neočakávala.[22][26]



Obr. 2.2: Schéma ribozómu a 16S rRNA [2]



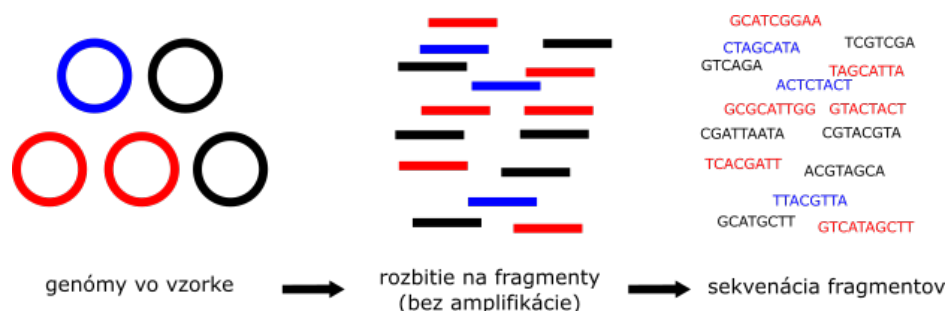
Obr. 2.3: Štruktúra 16S rRNA [3]

Aj keď môže variabilita hypervariabilných úsekov medzi rôznymi druhmi baktérií výrazne kolísať, v porovnaní s eukaryotickým ekvivalentom 18S rRNA je dĺžka sekvencie 16S rRNA génu pomerne stabilná u všetkých druhov. Umožňuje to výrazne jednoduchšie sekvenčné zarovnanie v analýze podobností v jednotlivých sekvenciách.[22]

2.3.2 Metódy sekvenácie

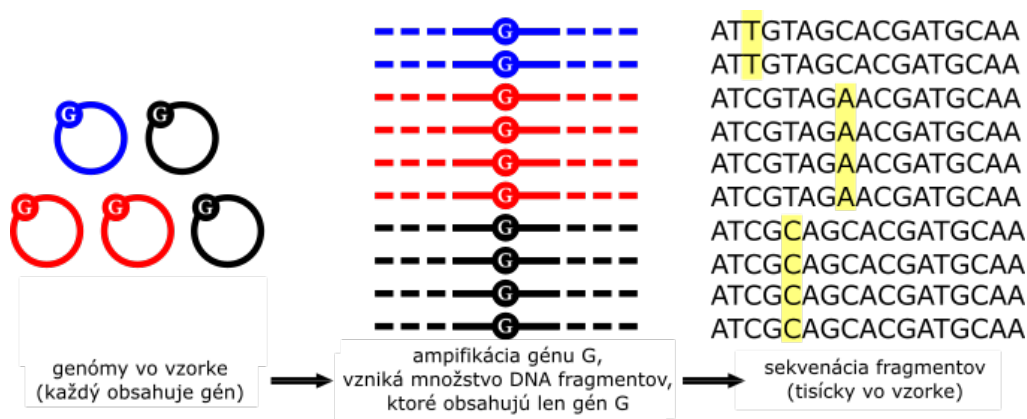
Samotnému procesu sekvenácie DNA predchádza príprava **knihovne**. Ide o proces, ktorého výsledkom sú fragmenty, ktoré sa budú sekvenovať. Existujú dva prístupy k príprave knihovne, a to *shotgun* a *amplicon*.

Shotgun metóda sa používa na sekvenáciu genómu alebo úsekov dlhších ako je dĺžka čítania sekvenátoru. Spočíva v rozbití DNA na fragmenty rôznych dĺžok, ktoré sa podľa dĺžky roztriedia pomocou elektroforézy a následnom sekvenovaní týchto fragmentov (obr. 2.4).[27][28]



Obr. 2.4: Schéma prípravy knihovne metódou *shotgun*

Amplicon je metóda, ktorá sa naopak používa pri sekvenácii kratších úsekov, ktoré pokryje čítanie sekvenátoru. Využíva sa možnosť výberu konkrétneho úseku DNA za pomoci primerov, ktoré v procese PCR vymedzia oblasť sekvencie DNA, ktorú bude DNA polymeráza syntetizovať (obr. 2.5).[29]

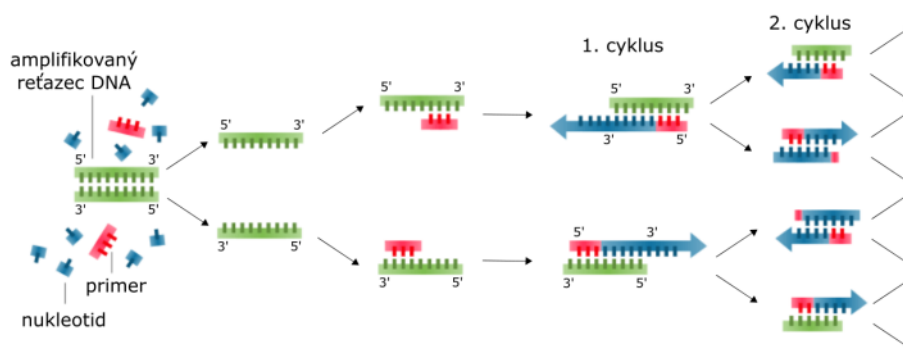


Obr. 2.5: Schéma prípravy knihovne metódou *amplicon*

PCR (polymerázová reťazová reakcia) je biochemická metóda využívaná v oblasti genetiky na namnoženie určitého úseku molekuly DNA. Je založená na princípe fungovania enzýmu DNA polymeráza, ktorá po denaturácii DNA syntetizuje komplementárne vlákno ku vláknu templátovému vždy v smere od 5' konca ku 3' koncu reťazca. DNA polymeráza však nedokáže nasyntetizovať komplementárne vlákno úplne od začiatku, potrebuje totiž primer¹, ktorý jej určí, kde má syntéza začať. Na zdvojenie jedného dvojvláknového reťazca DNA (jeden cyklus) je teda treba 2 primery (pre obe vlákna). V prípade, že je v reakčnej zmesi dostatočné množstvo DNA polymerázy, primerov a reakčného roztoku, tento proces prebieha na všetkých reťazcoch v reakčnej zmesi naraz. Tento cyklus sa opakuje dokým nie je dosiahnutý

¹Krátky úsek DNA (obvykle desiatky nukleových báz), ktorý sa pridáva do reakčnej zmesi PCR. Je nutné poznať okrajové oblasti amplifikovaného úseku, aby mohol byť primer správne nadizajnovaný.

potrebný počet namnožených reťazcov (v každom cykle sa množstvo reťazcov zdvojnásobí, teda s každým cyklom počet reťazcov exponenciálne stúpa) (obr. 2.6).[30]



Obr. 2.6: Schéma PCR [4]

Po príprave knihovne je potrebné nasekvenovať vzniknuté fragmenty. Najvyužívanějšími sekvenáčnymi platformami v analýze 16S rRNA génu sú *Illumina* (NGS²) a *Pacific Bioscience* (TGS³).

Illumina číta cieľové úseky DNA, ktoré boli naamplifikované pomocou mostíkovej PCR. Fyzykálno-chemický princíp detekcie, tzv. **sekvenácia syntézou**⁴, spočíva v tom, že syntéza komplementárneho vlákna k templátu prebieha zároveň so samotnou sekvenáciou. Nukleotidy, ktoré DNA polymeráza viaže k templátu sú fluorescenčne označené a v každom cykle, kedy sa práve jeden nukleotid naviaže na templát je emitované svetlo, ktoré prístroj zachytáva (pre každú bázu iná farba svetla). Následne je poradie zábleskov počítačovo spracované do poradia nukleotidov v DNA. Zároveň sa využíva metóda tzv. *single-end* (úsek DNA sa sekvenuje iba v jednom smere) alebo *paired-end* (úsek DNA sa sekvenuje v oboch smeroch, v priamom aj reverznom).[31][32]

Pacific Bioscience slúži na sekvenáciu dlhých reťazcov DNA. Využíva princíp **SMRT**⁵. Ide o *SMRT* čip, ktorý má na svojom povrchu tisíce jamiek hlbokých 100nm s priemerom 70nm. Na dne je zachytená DNA polymeráza, ktorá syntetizuje komplementárne vlákno k jednovláknovej DNA. Jednotlivé nukleotidy, z ktorých je syntetizované komplementárne vlákno sú fluorescenčne označené a pri naviazaní na templát je emitované svetlo, ktoré je detekované a spracované podobne ako v metóde Illumina.[33]

²Druhá generácia sekvenátorov (*Next generation sequencing*).

³Tretia generácia sekvenátorov (*Third generation sequencing*).

⁴Z anglického *sequencing by synthesis*.

⁵*Single-molecule real-time* (jednomolekulová sekvenácia v reálnom čase).

Dĺžka čítania sa u rôznych sekvenačných platforiem líši, spravidla je ale kratšia než skúmaná sekvencia. Táto výsledná sekvencia sa získava skladaním jednotlivých nasekvenovaných úsekov (obr. 2.7). V oblastiach, kde nasekvenované úseky obsahujú rovnakú časť celkovej DNA, sa jednotlivé úseky „prekrývajú“. Jednotlivé sekvencie a prekryvy sa analyzujú, čím sa získava celková sekvencia DNA. V prípade, že sa nasekvenované úseky porovnávajú s určitou referenčnou sekvenciou, ide o **mapovanie na referenciu**. Ak sa referenčná sekvencia nepoužíva, výsledná sekvencia sa vyskladá len z nasekvenovaných úsekov, ide o skladanie **de novo**.



Obr. 2.7: Skladanie DNA fragmentov

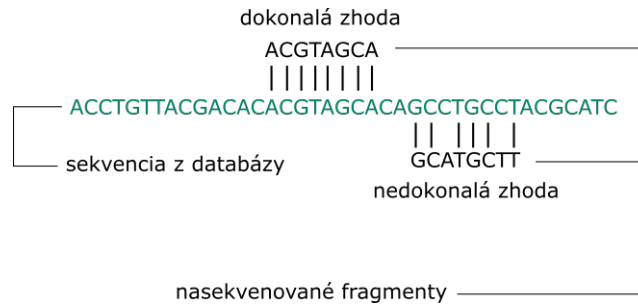
V samotnej analýze mikrobiómu sa využívajú dva prístupy, a to ampikonová alebo celogenómová analýza.

V **ampikonovej analýze** 16S rRNA génu sa vybraný hypervariabilný úsek amplifikuje v PCR použitím primerov, ktoré sú navrhnuté tak, aby došlo k amplifikácii len vybraných hypervariabilných úsekov. Najčastejšie sa používajú úseky V6-V8, V7-V8 alebo V3-V4, ich variabilita medzi rôznymi druhmi baktérií je totiž dostatočná na ich identifikáciu. Nasleduje sekvenácia a analýza. Výhody tejto metódy sú:

- je pomerne rýchla a lacná,
- presná a efektívna v prípade, ak ide o analýzu mikrobiómu na kmeňovej taxonomickej úrovni *phyla*,
- keďže je vo vedeckom prostredí vo veľkom používaná, existuje obrovské množstvo poznatkov, ktoré sú dostupné pre vedcov.

Nevýhody metódy:

- poskytuje len informácie na základe zvoleného hypervariabilného úseku,
- na druhej taxonomickej úrovni (*species*) presnosť analýzy klesá,
- špecifické gény sa neanalyzujú samostatne, ale sú aproximáciou s určitou toleranciou zaradené do taxonomickej úrovne (tab. 2.3),
- je vyžadovaná vysoká koncentrácia génu, aby bola analýza dostatočne presná.[34]



Obr. 2.8: Porovnávanie nasekvenovaných úsekov so sekvenciami z databáz

Riešením chybovosti amplikonovej analýzy v identifikácii baktérií na druhej taxonomickej úrovni je **celogenómová shotgun analýza**. Princíp spočíva v náhodnom rozbití všetkých genómov vo vzorke na fragmenty a ich následnom sekvenovaní. Výhody tejto metódy:

- poskytuje presné a kvalitné výsledky, pretože sekvenácia prebieha vrámci celého genómu mikroorganizmu, čo navyše umožňuje identifikáciu vzácnych či nových druhov baktérií,
- nevyžaduje amplifikáciu cielených úsekov.

Nevýhody metódy:

- je pomalá a drahá v porovnaní s amplikonovou analýzou,
- vysoké požiadavky na technické a programové vybavenie, pretože metóda poskytuje obrovské množstvo dát za krátky čas.[34][35]

Databázy sekvencií 16S rRNA

16S rRNA gén je používaný ako štandard v identifikácii a klasifikácii mikroorganizmov, pretože je vo väčšine z nich obsiahnutý vo veľkých množstvách a nesie mnohé dôležité informácie k tomuto spracovaniu. Mnohé sekvencie 16S rRNA génu sú verejne dostupné v databázach, ako napríklad NCBI (National Center for Biotechnology Information), avšak kvalita týchto sekvencií často nie je overená. Preto boli založené špeciálne databázy, ktoré sú zamerané na zhromažďovanie výhradne sekvencií 16S rRNA génov.

EzTaxon - databáza obsahujúca približne 63 tisíc celogenómových sekvencií baktérií a archeónov, ktoré boli taxonomicky identifikované na rodovej, druhej či poddruhej úrovni. Táto databáza je systematicky ošetrovaná a aktualizovaná.[36]

RDP (The Ribosomal Database Project) - udržiavaná databáza, ktorá ponúka fylogeneticky zarovnané sekvencie RNA (rRNA), odvodené fylogenetické stromy či rôzne programové vybavenie pre zjednodušenie analýzy a zobrazenia zarovnaní a stromov.[37]

SILVA - obsiahla databáza s kontrolou kvality, ktorá pravidelne aktualizuje svoje datasety. Obsahuje zarovnané sekvencie ribozomálnych rRNA malých i veľkých ribozomálnych jednotiek baktérií, archeónov a eukaryotov. Obsahuje tiež nástroje na prieskum, vyhľadávanie či sťahovanie materiálov.[38]

GreenGenes - databáza obsahuje fylogenetické stromy zostavené automaticky *de novo* zo 16S rRNA sekvencií, ktoré boli získané z verejných databáz a prešli testami kvality.[39]

2.4 Taxonomická analýza 16S

Po spracovaní nasekvenovaných dát nasleduje taxonomická analýza, ktorá využíva metódu tvorby OTU alebo ASV zhlukov.

2.4.1 OTU

Staršou metódou analýzy je vytváranie zhlukov **OTU**⁶. Ide o zhluky nasekvenovaných úsekov, ktoré sa porovnávajú s určitou toleranciou nezhody so sekvenciami z dostupných databáz 16S rRNA sekvencií. Táto metóda je výpočtovo jednoduchšia ako ASV metóda, pretože neanalyzuje unikátne sekvencie, ale len zaraďuje sekvencie podľa podobnosti s referenciou do skupín (tab. 2.3). Táto metóda sa tiež častejšie využíva v teréne. Nevýhodou je, že metóda funguje na vopred stanovenej tolerancii podobnosti, čo sa môže negatívne prejaviť na analýze vzdialenejších mikrobiálnych komunit. Kvalita analýzy takisto závisí na výbere databázy, s ktorou sa sekvencie budú porovnávať a samotnom zhlučovacom algoritme.[40]

2.4.2 ASV

Modernejšou metódou je vytváranie zhlukov **ASV**⁷. Sú to zhluky, ktoré vznikajú podobne ako zhluky OTU, ale v tomto prípade sa sekvencie porovnávajú s referenčnými na úrovni jednotlivých nukleotidov. Táto metóda je síce výpočtovo náročnejšia ako OTU metóda, avšak presnejšia, navyše umožňuje spoľahlivejšiu identifikáciu vzácných či nových druhov mikroorganizmov. Je takisto menej náchylná na výber referenčnej databázy či zhlučovací algoritmus, ale viac náchylná na chyby, vzniknuté v procese PCR.[41]

⁶Operational taxonomic unit

⁷Amplicon sequence variant

2.4.3 Taxonomické úrovne

Samotná taxonomická analýza však znamená, že jednotlivé mikroorganizmy vo vzorke dokážeme identifikovať, teda zaradiť ich do príslušných taxonomických úrovní. Ide o hierarchický systém, ktorý sa pravidelne aktualizuje. Kompletný hierarchický systém je zložitý, preto mikroorganizmy z použitej vzorky pre túto prácu sú taxonomicky zaradené iba do základných taxonomických úrovní (tab. 2.3 prvý stĺpec).

Taxonom. úroveň	Taxóny
Ríša	<i>Bacteria</i>
Kmeň	<i>Bacteroidetes</i>
Trieda	<i>Bacteroidia</i>
Rad	<i>Bacteroidales</i>
Čeľaď	<i>Bacteroidaceae</i>
Rod	<i>Bacteroides</i>
Druh	<i>Bacteroides dorei</i>

Tab. 2.3: Príklad zaradenia do taxonomických úrovní

2.5 Analýza diverzity

Analýza diverzity je proces kvantifikácie a porovnávania diverzity biologických spoločenstiev, ako sú mikrobiálne spoločenstvá v metagenomike, rastlinné spoločenstvá v ekológii alebo genetické variácie v populáciách. Cieľom analýzy diverzity je pochopiť procesy, ktoré diverzitu spôsobujú alebo ovplyvňujú jej početnosť v rámci komunity či medzi nimi.

Analýza diverzity v metagenomickej analýze zahŕňa kvantifikáciu a porovnanie diverzity mikrobiálnych spoločenstiev v rôznych vzorkách alebo prostrediach. Medzi základné postupy analýzy diverzity patria analýzy na rôznych úrovniach diverzity. Patrí sem alfa, beta či gama diverzita.

2.5.1 Alfa diverzita

Alfa diverzita popisuje štruktúru ekologického spoločenstva v rámci jedného biotopu, ekosystému alebo vzorky. V rámci alfa diverzity rozlišujeme dve základné vlastnosti biotopu, a to rozmanitosť (počet rôznych taxonomických skupín žijúcich v danom biotope) a rovnomernosť (distribúcia jednotlivých skupín v biotope). Keďže zmeny v spoločenstve výrazne ovplyvňujú alfa diverzitu spoločenstva, analýza alfa diverzity je jedným z prvých prístupov pri analýze spoločenstiev a ich vývoja.[42]

2.5.2 Beta diverzita

Podobne ako alfa diverzita, aj beta diverzita popisuje štruktúru ekologického spoločenstva, avšak v rámci rôznych biotopov, ekosystémov či vzoriek. Slúži na porovnanie obsahu druhov v odlišných podmienkach na rôznych miestach. Vysoká beta diverzita naznačuje, že druhové zloženie sa medzi lokalitami výrazne líši.[43]

2.5.3 Gama diverzita

Gamma diverzita predstavuje celkovú diverzitu druhov vo väčšej geografickej oblasti alebo krajine. Zohľadňuje lokálnu diverzitu (alfa diverzitu) i druhovú diverzitu medzi jednotlivými oblasťami (beta diverzitu) v rámci širšej priestorovej mierky.[44]

Samotná úroveň, na ktorej analýza diverzity prebieha však závisí na dopredu určených podmienkach. Hranica medzi alfa, beta či gama diverzitou nie je striktne stanovená.

3 QIIME2

3.1 Využitie QIIME2

QIIME2 je významný decentralizovaný verejne dostupný bioinformatický software, zameraný na analýzu dát, ktoré boli získané amplikonovou sekvenáciou mikrobiálnych komunit (najčastejšie sekvenácia génov 16S rRNA u prokaryotov a 18S rRNA u eukaryotov). Používateľovi ponúka celý proces spracovania od surových dát až po kvalitné grafické a štatistické závery. Proces zahŕňa demultiplexáciu a filtráciu kvality dát, výber OTU (pomocou UNCLUST alebo BLAST), taxonomické zaradenie, rekonštrukciu fylogenetických stromov, analýzu diverzity či grafické znázornenia analyzovaných dát a získaných štatistických záverov. Výhodou je, že software je výrazne zameraný na prehľadnosť dát, takže výstup softwaru je vo väčšine prípadov „user-friendly“. QIIME2 navyše ponúka veľké množstvo balíčkov, s ktorými je možné významne rozšíriť funkcionality softwaru.[45]

Verzia QIIME2 bola vydaná 1. januára 2018 ako kompletne redizajnovaný nástupca verzie QIIME1. Dizajn novej verzie bol vo väčšine prípadov inšpirovaný užívateľmi verzie QIIME1, ktorí v priebehu jej pôsobenia prispeli na seminároch či fórach svojimi nápadmi a dotazmi k jej vylepšeniu. Verzia QIIME1 už nie je podporovaná.[46]

3.2 Dátové formáty v QIIME2

Vstupom softwaru QIIME2 sú sekvencie, ktoré boli získané sekvenáciou (surové dáta). Pri definovaní vstupu je potrebné, aby užívateľ poskytol detaily o týchto dátach (čo je to za dáta, dátový formát, ako boli tieto dáta získané a i.). Najpoužívanejším formátom surových dát je FASTQ formát, ktorý obsahuje informácie o sekvencii, samotnú sekvenciu a korešpondujúce skóre kvality (ako surové dáta je možné použiť aj iné formáty, napr. FASTA, BIOM a i.).[47][45]

3.2.1 QIIME2 artifact

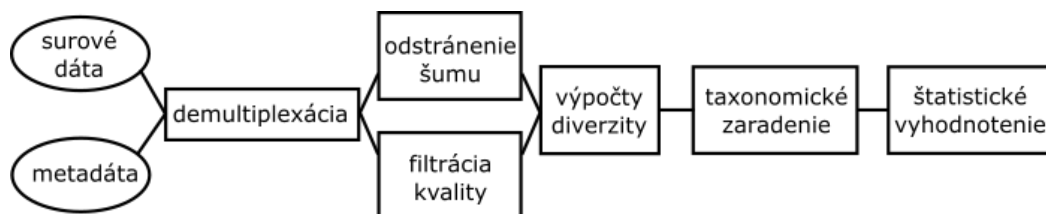
Jedným z výstupných dátových typov softwaru QIIME2 je tzv. **artifact** súbor. Je to *zip* súbor s koncovkou *.qza*, ktorý obsahuje dáta (samotné sekvencie, obvyčajne vo formáte FASTQ) a metadáta (špecifické metadáta pre QIIME2, ktoré poskytujú informácie o dátach, ako dátový formát, postup analýzy, ktorý bol zatiaľ vykonaný na dátach či postupy, ktoré QIIME2 využil na vytvorenie artifact-u). Tieto súbory je možné opakovane použiť ako vstup pre ďalšie analýzy.

Každý artifact súbor vytvorený pomocou QIIME2 má svoj príslušný sémantický typ, čo v prípade, že sa artifact použije ako vstup pre ďalšiu analýzu, umožňuje softwaru identifikovať ho ako kompatibilný vstup pre túto analýzu. Ide teda o istý spôsob prevencie proti vzniku chýb v analýzach.

Využitie artifact súborov umožňuje efektívnu prácu so softwarom QIIME2, keďže všetci užívatelia pracujú s tým istým dátovým typom. Môžu sa teda zamerať priamo na potrebné analýzy bez zbytočných konvertovaní z iných dátových typov.[47][45]

3.2.2 QIIME2 vizualizácia

Druhým výstupným dátovým typom softwaru QIIME2 je tzv. **vizualizácia**. Tieto súbory majú koncovku *.qzv*. Aj keď sú obsahovo veľmi podobné artifact súborom (obsahujú rovnaké metadáta), ich funkciou je poskytovať finálny výstup analýzy (štatistické výsledky, interaktívne vizualizácie, grafy a iné vizuálne reprezentácie dát), a teda nie je možné ich použiť ako vstup pre ďalšie analýzy.[47][45]



Obr. 3.1: Bloková schéma postupu spracovania dát pomocou QIIME2 [5]

4 Metagenomická analýza

4.1 Dataset

K pochopeniu vývoja črevnej mikrobioty podľa typu diéty som použil dataset obsahujúci nasekvenované úseky 16S rRNA génu, konkrétne ide o variabilné úseky V3-V4.

4.1.1 Pôvod datasetu

Vzorky stolice zvierat zo zoologickej záhrady v Chongqing boli odobrané hneď po prirodzenej defekácii živočíchov (vzorka neprišla do kontaktu s pôdou ani inými nečistotami). Následne boli vzorky zamrazené tekutým dusíkom a odoslané do laboratória v penových boxoch. Nukleové kyseliny, obsiahnuté vo vzorkách boli izolované pomocou špecializovanej súpravy (Qiagen QIAamp Stool Mini Kit) a ich koncentrácia bola určená UV absorbanciou pomocou spektrofotometra (Nano-Drop 2000). K PCR amplifikácii nukleových kyselín boli využité dostupné sekvencie primerov na vymedzenie oblastí s hypervariabilnými úsekmi V3-V4. Následne boli vzorky nasekvenované pomocou NGS (Next Generation Sequencing), konkrétne metódy *Illumina* (Illumina Miseq PE300). Sekvenácia prebiehala na fragmentoch v priamom aj reverznom smere (metóda *paired-end*, poskytujúca kvalitné dáta a zarovnateľnosť sekvencií).

Vzorky boli odobrané v roku 2018 v rámci projektu *Evolution of animal gut microbes and their NRPS gene clusters* (Evolúcia črevných mikróbov zvierat a ich NRPS génových zhlukov) a následne, 1.6.2021, bol zverejnený výsledný dataset v databáze GenBank. Informácie o projekte sú dostupné v databáze NCBI pod identifikátorom *PRJNA631813*.^[17]

4.1.2 Štruktúra a obsah datasetu

Podľa typu diéty sú zvieratá rozdelené do troch skupín, a to skupina A (bylinožravce), skupina B (mäsožravce) a skupina C (všežravé vtáky). Každá skupina obsahuje dvoch, resp. troch zástupcov (označení X, Y, resp. Z). Od každého zástupcu boli odobrané 3 vzorky stolice (označené 1, 2, 3). Dokopy bolo odobraných 24 vzoriek od ôsmich zástupcov.

Skupina **A** obsahuje troch zástupcov, a to *Elephas maximus Linnaeus* (slon ázijský, vzorky AXn¹), *Cervus nippon* (jelen sika, vzorky AYn) a *Tragelaphus oryx* (antilopa losia, vzorky AZn). Skupina **B** obsahuje rovnako troch zástupcov, a to

¹n - číslo odobranej vzorky, v tomto prípade 1, 2, 3

Panthera tigris Amoyensis (tiger juhočínsky, vzorky BXn), *Panthera leo* (lev púšťový, vzorky BYn) a *Vulpes vulpes fulva* (líška hrdzavá, vzorky BZn). V skupine C sú dvaja zástupci všežravých vtákov, a to *Pavo cristatus* (páv korunkatý, vzorky CXn) a *Struthio camelus* (pštros dvojprstý, vzorky CYn).[17]

Názorná ukážka štruktúry a obsahu datasetu je možné vidieť v tab. 4.1.

Skupiny podľa diéty	Zástupca	Vzorka	ID
Bylinožravce	<i>Elephas maximus Linnaeus</i>	AX1	SRR11773874
		AX2	SRR11773873
		AX3	SRR11773862
	<i>Cervus nippon</i>	AY1	SRR11773857
		AY2	SRR11773856
		AY3	SRR11773855
	<i>Tragelaphus oryx</i>	AZ1	SRR11773854
		AZ2	SRR11773853
		AZ3	SRR11773852
Mäsožravce	<i>Panthera tigris Amoyensis</i>	BX1	SRR11773851
		BX2	SRR11773872
		BX3	SRR11773871
	<i>Panthera leo</i>	BY1	SRR11773870
		BY2	SRR11773869
		BY3	SRR11773868
	<i>Vulpes vulpes fulva</i>	BZ1	SRR11773867
		BZ2	SRR11773866
		BZ3	SRR11773865
Všežravce	<i>Pavo cristatus</i>	CX1	SRR11773864
		CX2	SRR11773863
		CX3	SRR11773861
	<i>Struthio camelus</i>	CY1	SRR11773860
		CY2	SRR11773859
		CY3	SRR11773858

Tab. 4.1: Štruktúra datasetu

4.2 Predspracovanie dát

Pred samotnou analýzou je potrebné dáta predspracovať - skontrolovať kvalitu sekvenčných dát a následne podľa výsledkov nastaviť parametre zastrihnutia a filtrácie.

4.2.1 Kontrola kvality pred úpravou

Dataset obsahuje 48 súborov vo formáte *fastq*² - 24 párových vzoriek (napríklad SRR117738515_1.fastq a SRR11773851_2.fastq). Každý súbor obsahuje veľké množstvo readov o určitej dĺžke. Tieto vlastnosti datasetu som overil pomocou softwaru **FastQC** (verzia FastQC-0.11.5). *FastQC* je nástroj, ktorý poskytuje rôzne typy analýz veľkého množstva sekvenčných dát. Jeho najfrekvencovanejšie využitie je v kombinácii so sekvenčnými platformami generácie NGS³. Nástroj umožňuje používateľovi vykonať rýchlu kontrolu kvality sekvenčných dát. Výstupom nástroja je *html*⁴ súbor, ktorý obsahuje výsledky analýzy a štatistiky (väčšinou zobrazené a popísané v grafe). Následne som jednotlivé výstupy *FastQC* (pre každú vzorku) súhrnne zobrazil pomocou nástroja **MultiQC**. Výstupom *MultiQC* je rovnako *html* súbor, ktorý obsahuje graficky zobrazené agregované výsledky nástroja *FastQC*. Získaný *html* súbor je súčasťou elektronickej prílohy práce.[48]

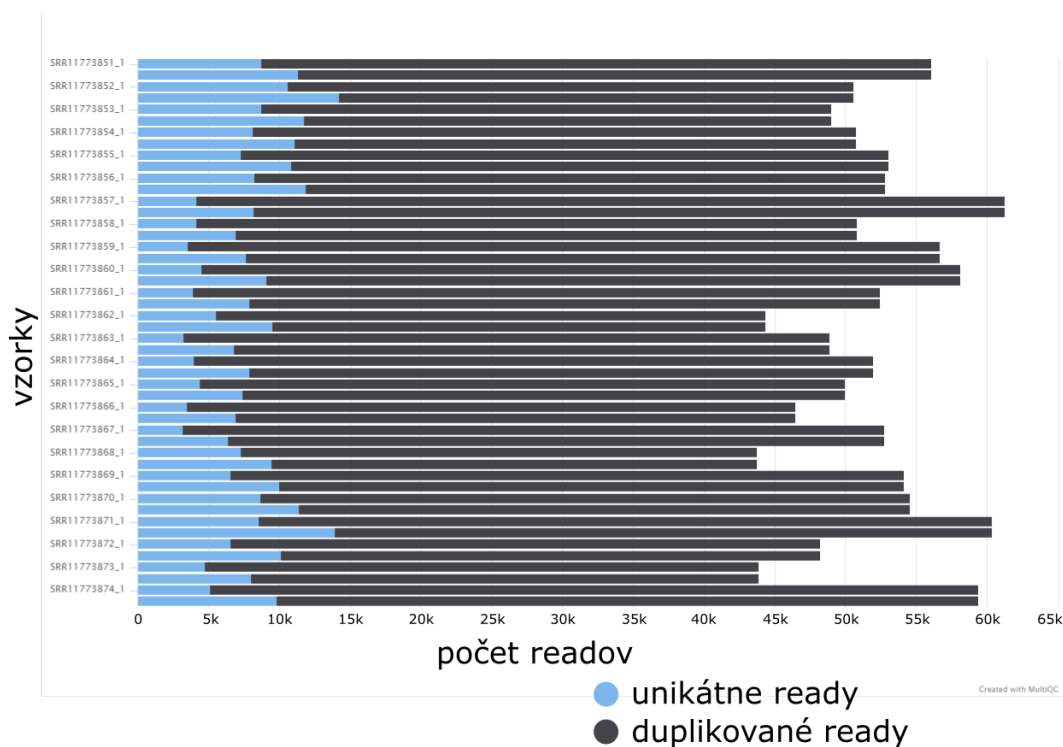
Počet a dĺžka readov

Z výstupu agregovanej kontroly kvality (MultiQC) som určil, že počet readov vo vzorkách sa pohyboval v rozmedzí od 43 751 do 61 263 (obr. 4.1), a že dĺžka jednotlivých readov je rovnaká, a to 301bp.

²Textový formát na uchovávanie biologických sekvencií a ich skóre kvality.

³Next generation sequencing.

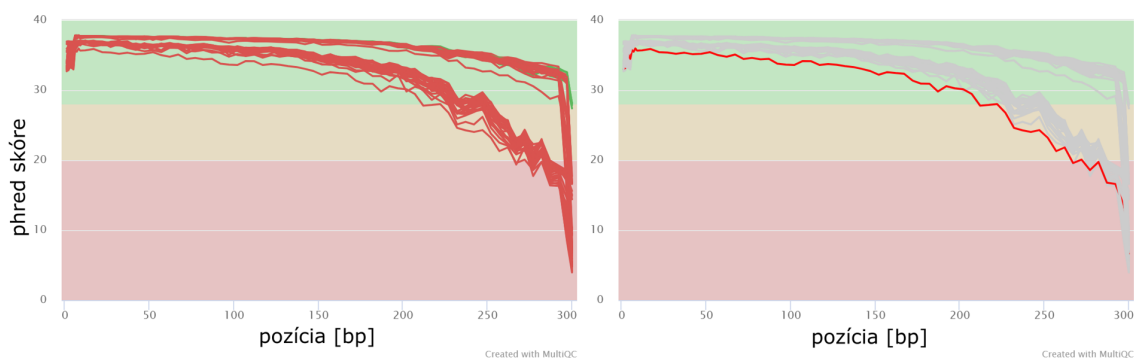
⁴Formát určený na zobrazenie v internetovom prehliadači.



Obr. 4.1: Graf počtu unikátnych a duplikovaných readov v jednotlivých vzorkách pred úpravou

Kvalita sekvencií

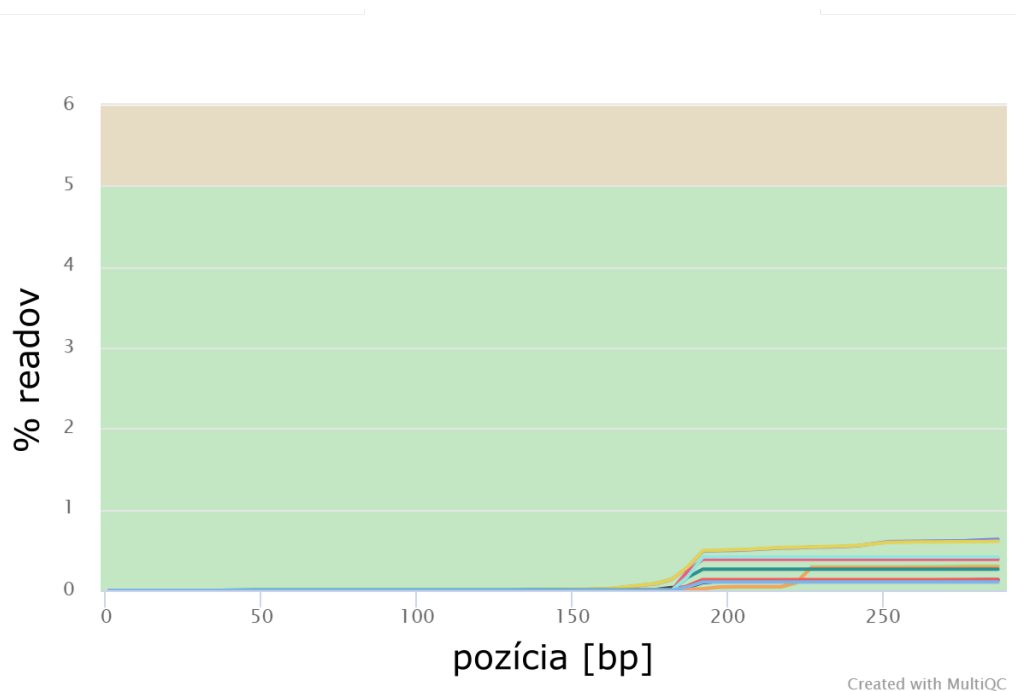
Na samotnú kontrolu kvality som použil nástroj FastQC, ktorý pracuje s hodnotami *phred* skóre pripadajúcim k jednotlivým bázam v readoch. Môj výsledný graf obsahuje 48 kriviek (48 vzoriek), pričom každý bod na krivke predstavuje priemernú kvalitu na danej pozícii v danej vzorke. Na zvislej ose grafu je *phred* skóre, ktoré vyjadruje kvalitu identifikácie bázy automatizovaným sekvenátorom. Výsledný graf som hodnotil podľa najnižšej kvality jednotlivých vzoriek, teda ak kvalita vzorky poklesla pod hodnotu *phred* skóre 20, hodnotil som ju ako nekvalitnú. V mojom prípade vyšli kvalitné vzorky len 4 zo 48, čo som hodnotil ako neuspokojivý výsledok (obr. 4.2 vľavo). Príklad jednej nekvalitnej vzorky je možné vidieť na obr. 4.2 vpravo. Na obrázkoch je vidieť, že kvalita sekvencií je najlepšia na začiatku sekvencií a postupne klesá ku koncu sekvencií. Ide o fenomén poklesu kvality na konci sekvenačného behu, ktorý vzniká z dôvodu postupujúceho procesu sekvenácie v prístroji alebo iných technických limitácií sekvenátorov. Tieto nekvalitné konce sekvencií sa spravidla odstraňujú zastrihnutím v procese spracovania sekvenačných dát, aby nedochádzalo k výrazným vplyvom na kvalitu analýzy.



Obr. 4.2: Závislosť kvality na pozícii báz (vľavo: ukážka kvality všetkých vzoriek, vpravo: ukážka kvality jednej vzorky)

Prítomnosť adaptérov

Adaptéry sa používajú v procese sekvenácie na imobilizáciu fragmentov v sekvenátore. Výskyt adaptérov v readoch je spôsobený tým, že sa sekvenátor „presekvenuje“ až do adaptéru. Prítomnosť týchto úsekov v readoch som overil na základe grafu, ktorý znázorňuje obsah adaptérovej sekvencie v readoch (obr. 4.3). V grafe je možné vidieť, že adaptéry sa vyskytovali v približne 0,5% readov.



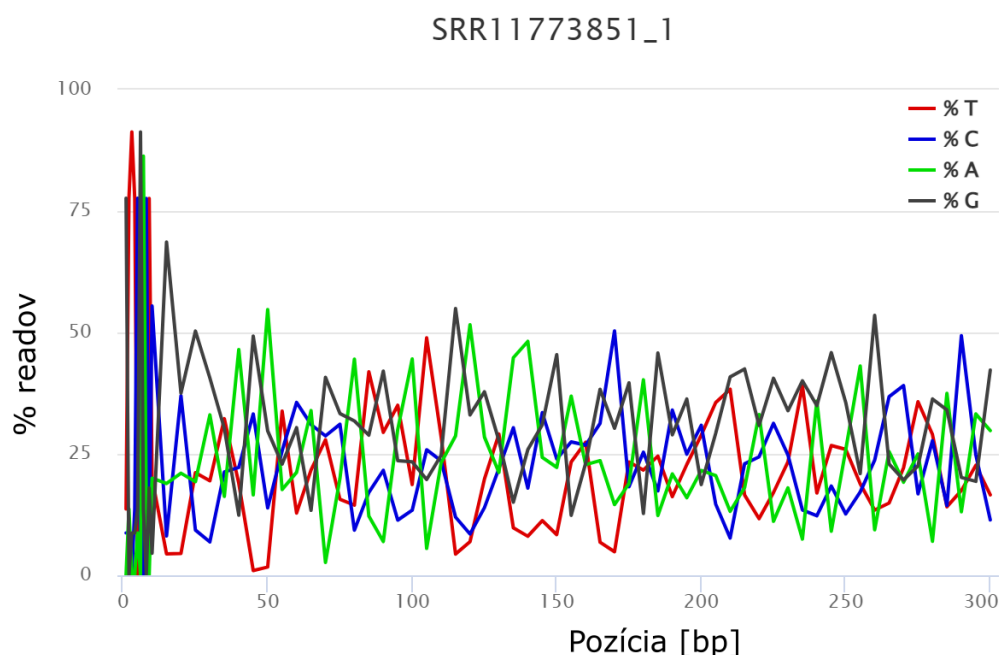
Obr. 4.3: Graf obsahu adaptérov v readoch pred úpravou

Výskyt N-báz

V prípade, že sekvenátor nie je schopný určiť bázu s určitou istotou, zapíše namiesto klasických báz (A, C, G, T) tzv. N-bázu. Prítomnosť týchto N-báz som overil z grafu, ktorý priamo zobrazoval percentuálny obsah N-báz na danú pozíciu pre jednotlivé vzorky. Obsah bol celkom nulový.

Distribúcia báz A, C, G, T

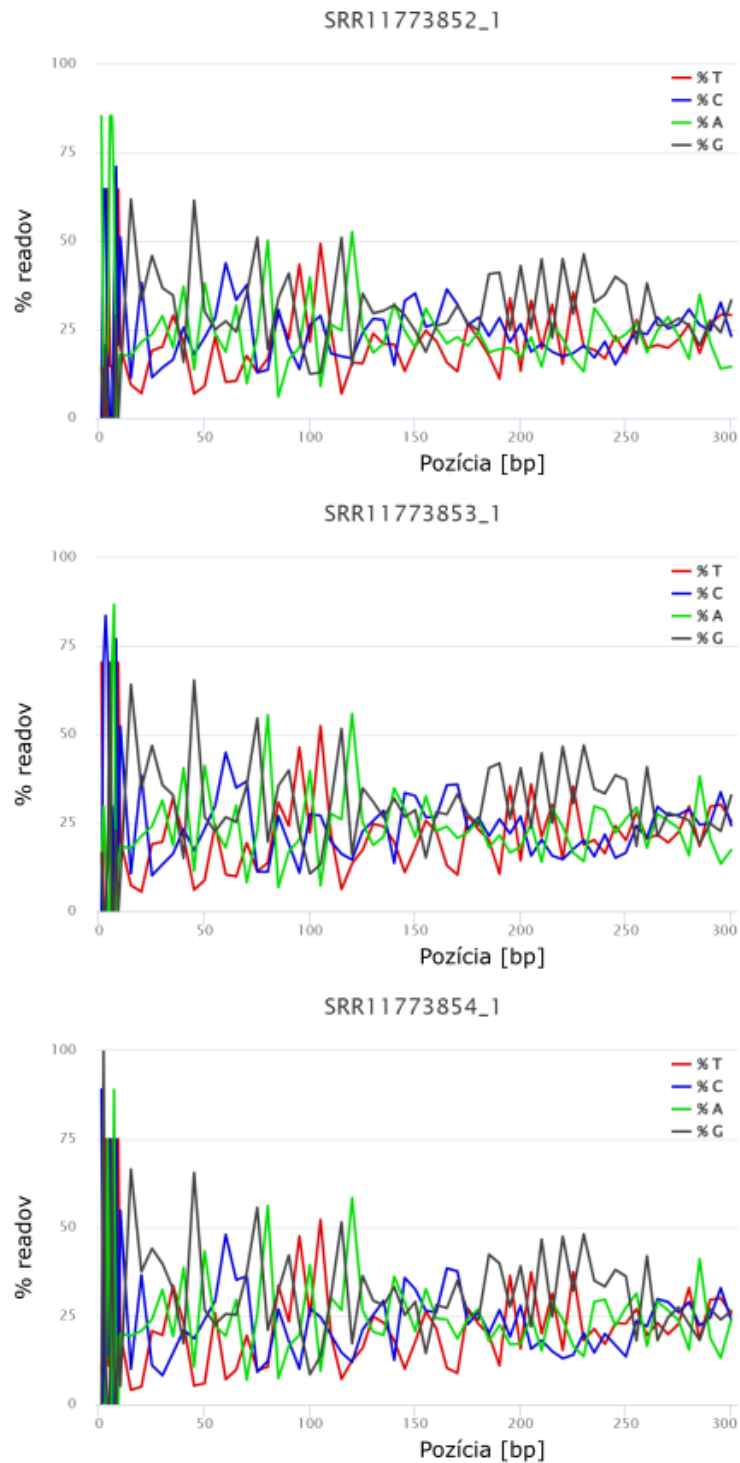
Obsah jednotlivých typov báz na jednotlivých pozíciách naprieč všetkými readmi vo vzorke som overil z grafov ako je na obrázku 4.4. Počiatočné vysoké hodnoty v grafe boli spôsobené prítomnosťou veľkého množstva rovnakých sekvencií na začiatkoch jednotlivých readov vo vzorke. Je to spôsobené výskytom adaptérových sekvencií na začiatkoch readov, ktoré sú naprieč vzorkou identické. V grafe je možné vidieť, že výskyt rôznych typov báz je celkom rozmanitý naprieč readmi vo vzorke. Dôvodom je, že vo vzorke sa nachádza veľké množstvo rôznych mikroorganizmov (rôzne hypervariabilné úseky), a dá sa teda očakávať, že aj sekvencie získané z tejto vzorky budú rozmanité.



Obr. 4.4: Graf výskytu báz na jednotlivých pozíciách vo vzorke SRR11773851_1

V nasledujúcich grafoch je možné vidieť, ako sa obsah báz v readoch trochu rôznych vzoriek čiastočne podobá (obr. 4.5). Dôvodom je, že tieto tri vzorky (SRR11773852_1, SRR11773853_1, SRR11773854_1) sú zo stolice jedného živočícha, teda metagenóm

v nich by mal byť podobný. Môžem teda vo vzorke očakávať prítomnosť podobných sekvencií.[49]



Obr. 4.5: Porovnanie obsahu báz vo vzorkách SRR11773852_1, SRR11773853_1, SRR11773854_1

4.2.2 Zastrihnutie a filtrácia

Na zastrihnutie a filtráciu sekvenčných dát som použil software **FastP** (v0.21.0). Je to nástroj určený na predspracovanie dát vo formáte *fastq*. Umožňuje kontrolu kvality, vyhľadanie a zastrihávanie adaptérov či filtráciu podľa kvality. Na základe predošlej kontroly kvality som nastavil parametre pre nástroj *FastP*.

Detekciu adaptérov je možné uskutočniť pomocou „analýzy prekrytia“. Ide o zarovnanie sekvencií podľa počiatočných báz v sekvenciách. Následne dochádza v miestach výskytu adaptérov k výraznej zhode medzi sekvenciami, pretože adaptéry, použité pri sekvenácii sú totožné naprieč sekvenciami. Následne zhoda výrazne klesá, čo je indikátorom, že na tomto mieste sa už nevyskytuje adaptérová sekvencia. Na detekciu adaptérov som použil príkaz `-detect_adapter_for_pe`, ktorý umožňuje automaticky detekovať a odstrániť adaptéry v jednotlivých readoch.

Pomocou príkazu `-3 -cut_tail_window_size 8 -cut_tail_mean_quality 30` som odstránil nekvalitné konce readov. Uvedený príkaz posúva okno (v mojom prípade dlhé 8 báz) od 3' konca smerom ku 5' koncu a zastriháva bázy v okne ak ich priemerná kvalita klesne pod určité *phred* skóre (v mojom prípade 30). Týmto postupom sa odstránia nekvalitné konce sekvencií, spôsobené fenoménom poklesu kvality na konci sekvenčného behu. Hodnoty prahového *phred* skóre, ktoré som testoval som si odvodil z grafu 4.2, kde od hodnoty 30 začína kvalita výraznejšie klesať. Hodnotu dĺžky okna som zvolil experimentálne podľa výsledkov. Dosiahnuť výbornú kvalitu výsledných dát je možné nastavením vysokej hodnoty prahového *phred* skóre, avšak spôsobuje to výraznejšiu stratu dát. Preto je nutné nájsť hodnotu, ktorá dáta dostatočne filtruje, ale zároveň neodstráni príliš veľké množstvo dát.

K odstráneniu readov, ktoré sú príliš nekvalitné som použil parameter `-e 30`, ktorý celkom odstráni ready, ktorých priemerná kvalita je nižšia ako zvolená hodnota *phred* skóre (v mojom prípade 30). Hodnotu parametru som zvolil prísnejšiu, pretože menej kvalitné ready spôsobujú veľkú chybu v analýze.

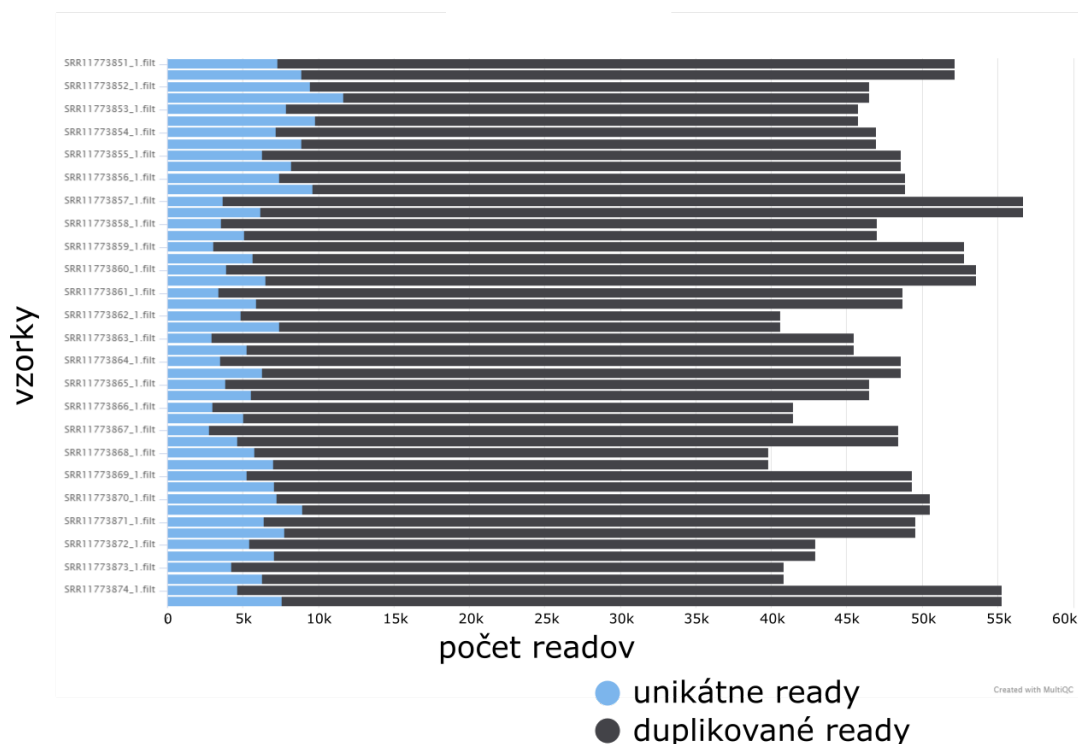
Príliš krátke ready som odstránil pomocou parametra `-l 30`, ktorý odstráni ready s obsahom báz nižším ako 30. Tento príkaz som umiestnil na koniec cyklu, pretože predošlými filtráciami dochádza k zmene dĺžky jednotlivých readov (nekvalitné ready sa môžu skrátiť až na dĺžku kratšiu ako 30).

4.2.3 Kontrola kvality po úprave - vyhodnotenie

Po filtrácii pomocou *FastP* som znova overil kvalitu sekvencií nástrojom *FastQC* a výsledky agregoval pomocou *MultiQC*. Získaný *html* súbor je súčasťou elektronickej prílohy práce.

Počet a dĺžka readov

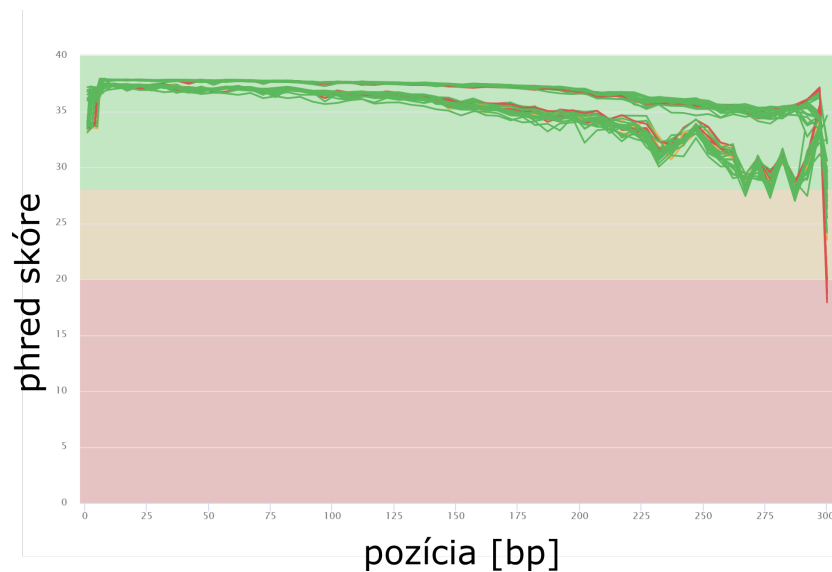
Počet readov vo vzorkách sa filtráciou znížil, a to na počet v rozmedzí od 39 858 do 56 716 (obr. 4.6). Zastrihnutím nekvalitných koncov sa dĺžky readov pozmenili na rôzne hodnoty. Priemerná dĺžka readov bola v rozmedzí od 239bp do 295bp.



Obr. 4.6: Graf počtu unikátnych a duplikovaných readov v jednotlivých vzorkách po úprave

Kvalita readov

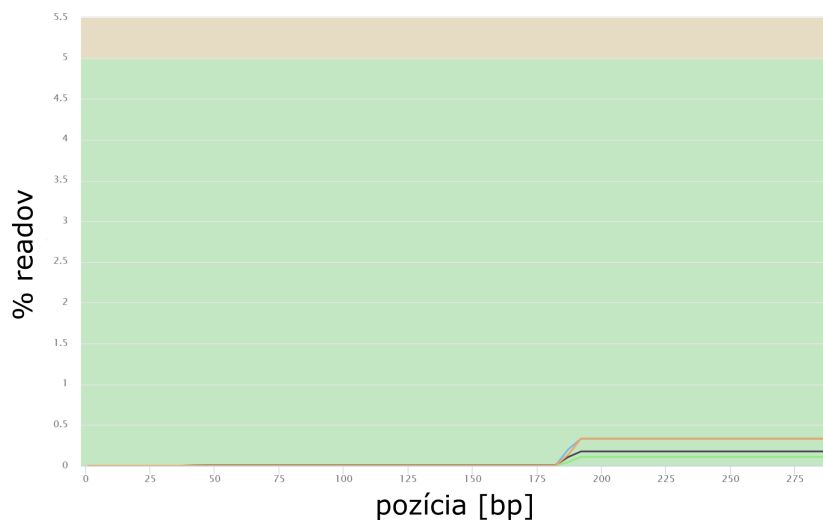
Zo 48 vzoriek prešlo testom kvality 38 vzoriek, čo bol dostačujúci výsledok (obr. 4.7). Na obrázku je vidieť, že kvalita sekvencií sa zlepšila. Ak zanedbám pokles kvality na konci readov, kvalita sekvencií sa nachádza nad hranicou *phred* skóre 27.



Obr. 4.7: Graf kvality vzoriek po úprave

Prítomnosť adaptérov

Výskyt adaptérových sekvencií bol po filtrácii nižší. V grafe je možné vidieť, že adaptéry sa vyskytovali v približne 0,34% readov, čo je pokles o 0,16% (obr. 4.8).[49]



Obr. 4.8: Graf obsahu adaptérov v readoch po úprave

4.3 Postup analýzy zloženia vzoriek

Na taxonomické zaradenie jednotlivých druhov baktérií vo vzorkách podľa nasekvenovaných dát som využil nástroj **QIIME2**. Táto metóda pozostáva z viacerých

krokov, z ktorých väčšinu je nutné vykonať v záväznom poradí (výstup jedného kroku je vhodným vstupom pre krok nasledujúci). Na začiatku som si pripravil súbor vo formáte *tsv* s metadátami o spracovávanom datasete (metóda vyžaduje metadáta v tomto formáte). Tento súbor som si nahral do nástroja QIIME2.

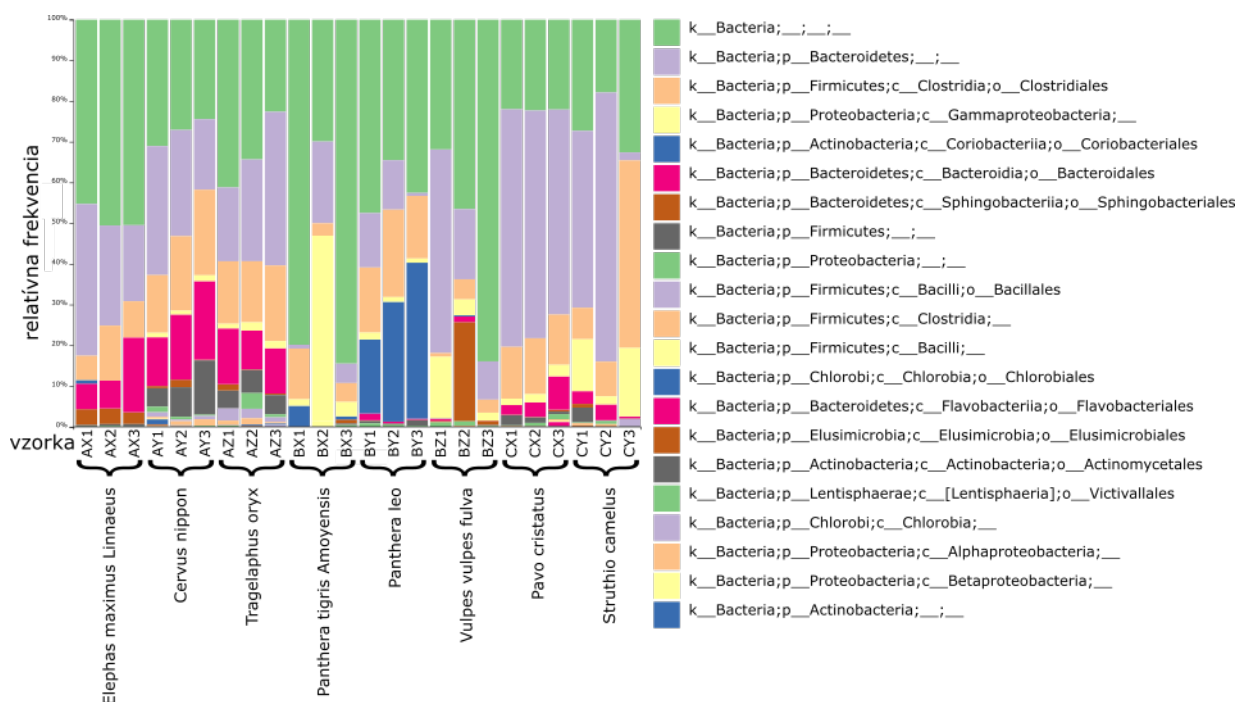
V úvode samotnej analýzy pomocou QIIME2 je možné využiť dve obdobné metódy predspracovania dát, a to **DADA2** alebo **Deblur**. Oba tieto algoritmy sú navrhnuté na spracovanie sekvencií získaných amplikonovou sekvenáciou. Zvolil som si metódu *Deblur*, pretože je všeobecne menej citlivá na vzácne druhy baktérií (na svoju analýzu som nepotreboval vysokú citlivosť metódy *DADA2*) a je menej výpočtovo náročná. Metóda *Deblur* vytvára ASV odstraňovaním chybných súčastí jednotlivých amplikonov. Rozlíšenie tejto metódy je na úrovni jednotlivých báz. Každé ASV teda predstavuje konkrétnu amplikonovú sekvenciu, ktorá je jedinečná v rámci dátovej sady. Z toho vyplýva, že ASV sú presnejšie ako OTU, ktoré vznikajú zhľukovaním sekvencií na základe ich podobnosti. Samotný proces úpravy sekvencií (filtrácia nekvalitných či chimérických sekvencií) však nemala významný dopad na dataset, pretože sekvencie prešli kontrolou kvality a filtráciou už v predspracovaní. *Deblur* bol však nutnou súčasťou analýzy aj z dôvodu nadväznosti vstupných a výstupných súborov nasledujúcich krokov analýzy.[50]

K nasledujúcej analýze bolo potrebné si vytvoriť fylogenetický strom z pripravených dát. Na zarovnanie sekvencií som využil nástroj **MAFFT**, ktorý je dostupný v balíku QIIME2. Po zarovnaní som odstránil pozície v sekvenciách, ktoré sa vyznačovali výraznou variabilitou. Tieto úseky totiž zvyšujú chybovosť výsledného fylogenetického stromu. Samotný strom som na záver zostavil pomocou nástroja **FastTree**.

Vytvorený fylogenetický strom som použil na následnú analýzu diverzity. Súčasťou analýzy diverzity bola aj analýza hlavných komponentov, ktorej výsledkom bol PCA graf 4.10, 4.11.

Na finálnu taxonomickú analýzu pomocou QIIME2 som využil klasifikátor *Naive Bayes*. Ide o model trénovaný na databázi GreenGenes [51], ktorý je určený na analýzu taxonomickej kompozície vzoriek. Výsledkom analýzy bola kompozícia jednotlivých vzoriek, ktorú som si vizualizoval vo forme barplotu. Z výsledného grafu 4.9, zobrazujúceho obsah jednotlivých bakteriálnych druhov vo vzorkách, som určil alfa a beta diverzitu vzoriek.

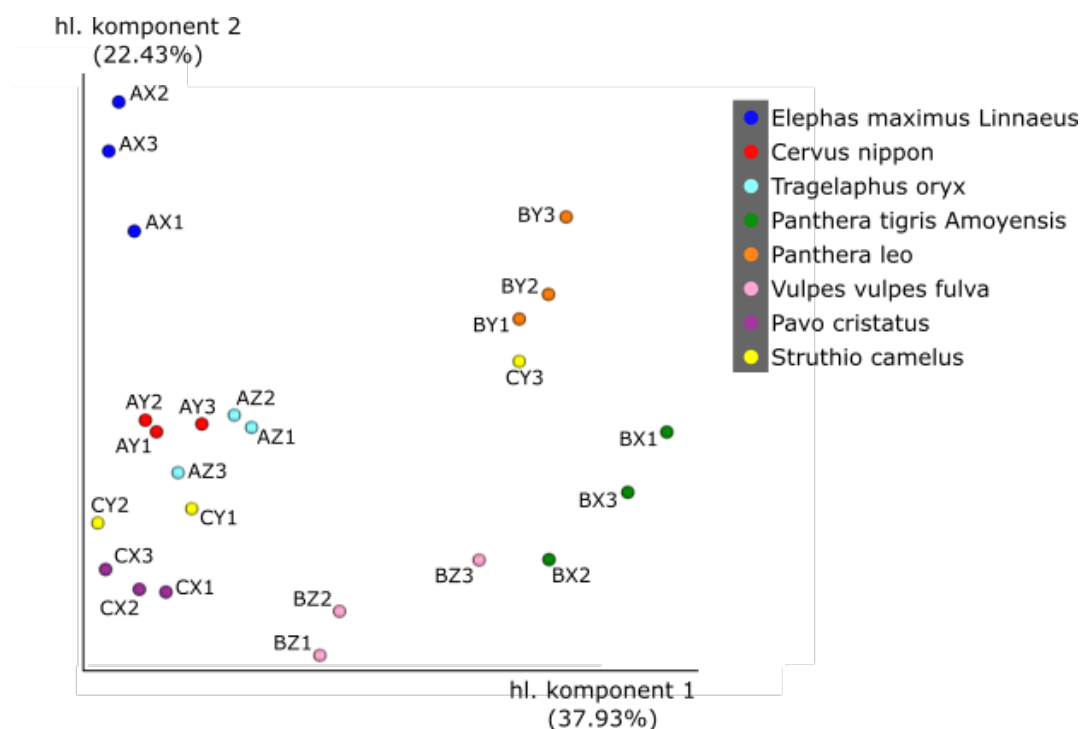
4.4 Výsledky analýzy diverzity



Obr. 4.9: Výsledná kompozícia vzoriek [6]

4.4.1 Alfa diverzita

Alfa diverzitu som analyzoval zvlášť na každej trojici vzoriek od jednotlivých zvierat a porovnal som ich kompozíciu v rámci skupín podľa typu potravy.



Obr. 4.10: Graf analýzy hlavných komponentov (vizualizácia podľa mena živočícha) [6]

V grafe 4.10 je možné vidieť, že diverzita medzi vzorkami od toho istého jedinca je takmer vždy pomerne nízka, čo súhlasí s očakávanými výsledkami. V grafe sa však nachádzajú aj odľahlé hodnoty (vzorky od *Struthio camelus*, *Panthera tigris Amoyensis* či *Vulpes vulpes fulva*), ktoré som v nasledujúcej časti analyzoval podrobnejšie.

Bylinožravce

Z grafu 4.10 som určil, že diverzita v skupine bylinožravcov bola u všetkých zástupcov nízka, s čím súhlasí aj graf 4.9. Obsah kmeňa *Bacteroidetes* je vo vzorkách podobný, rovnako ako rad *Clostridiales* či rad *Bacteroidales*. Mierny rozdiel je možné pozorovať u *Elephas maximus Linnaeus*, u ktorého sa v porovnaní s ostatnými vzorkami trieda *Gammaproteobacteria* a rad *Firmicutes* takmer vôbec nevyskytovali a naopak, rad *Sphingobacteriales* sa u neho vyskytoval vo väčšom množstve. Tento mierny rozdiel sa prejavil v grafe 4.10 tak, že vzorky od *Elephas maximus Linnaeus* sú čiastočne odstúpené od vzoriek od zvyšných dvoch zástupcov. Vzorky od *Cervus nippon* a *Tragelaphus oryx* sú kompozične veľmi podobné, preto spoločne vytvorili jeden zhhluk.

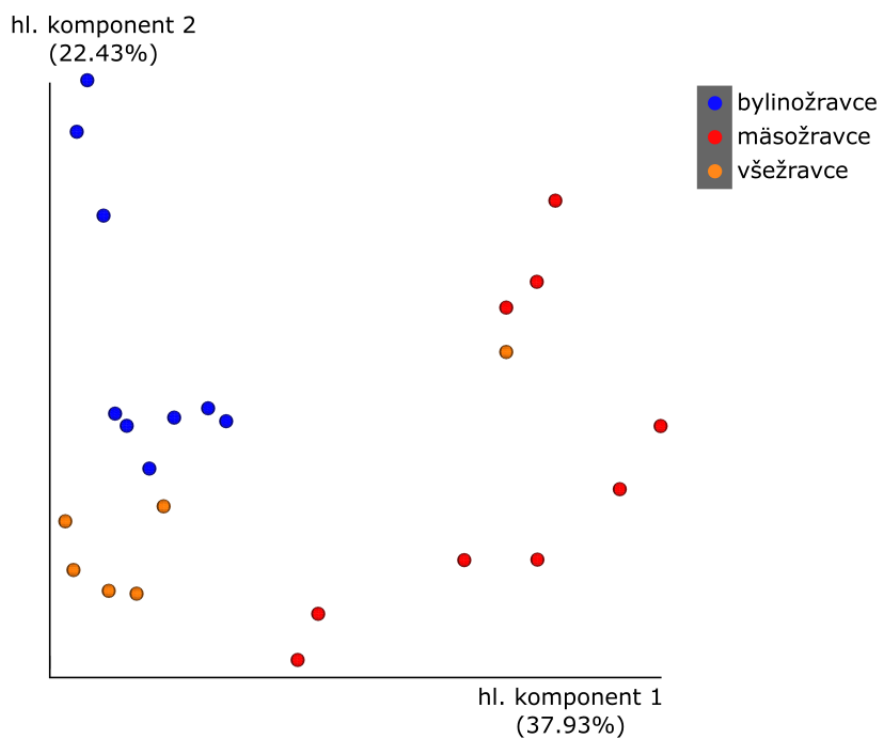
Mäsožravce

U mäsožravcov je variabilita rôznorodá. Podľa grafu 4.9 som určil, že vzorka BX2 u *Panthera tigris Amoyensis* sa výrazne odlišuje od BX1 a BX3 vysokým obsahom triedy *Gammaproteobacteria* a vzorka BZ2 u *Vulpes vulpes fulva* sa zase odlišuje od BZ1 a BZ3 výrazným zastúpením radu *Sphingobacteriales*. Dôvodom týchto výrazných odlišností bol zrejme jeden z faktorov ovplyvňujúcich kvalitu vzoriek (odber vzorky, kontaminácia vzorky a i.), pretože v oboch prípadoch sú zvyšné dve vzorky kompozične vcelku zhodné. U zástupcu *Panthera leo* je naopak diverzita nízka vo všetkých troch vzorkách, výrazný je najmä obsah rodov *Coriobacteriales* a *Clostridiales*. Tieto dva obsiahnuté rody však tohto zástupcu výrazne odlišujú od *Panthera tigris Amoyensis* i *Vulpes vulpes fulva*, a preto sú v grafe 4.10 vzorky týchto dvoch zástupcov mierne odstúpené od vzoriek zástupcu *Panthera leo*.

Všežravce

Vzorky v tejto skupine som podľa grafu 4.9 určil ako veľmi podobné. Výnimkou je vzorka CY3 od *Struthio camelus*, ktorá sa výrazne odkláňa od pomerne konštantnej kompozície ostatných vzoriek v skupine všežravcov. Má výrazne vyšší obsah rodu *Clostridiales* a zároveň výrazne nižší obsah kmeňa *Bacteroidetes*. Významná diverzita tejto vzorky bola takisto pravdepodobne spôsobená niektorým z faktorov ovplyvňujúcich kvalitu vzoriek. V grafe 4.10 je možné vidieť, ako sa vzorka CY3 zástupcu *Struthio camelus* vzdialila od zvyšných vzoriek všežravcov.

4.4.2 Beta diverzita



Obr. 4.11: Graf analýzy hlavných komponentov (vizualizácia podľa potravy živočíchov) [6]

Beta diverzitu som analyzoval medzi jednotlivými tromi skupinami, teda medzi bylinožravcami, mäsožravcami a všežravcami. V grafe 4.9 je možné vidieť, že u bylinožravcov je výrazný vyšší obsah radu *Bacteroidales* a u všežravcov zase kmeňa *Bacteroidetes*. Rod *Coriobacteriales* sa zase podstatne vyskytoval len u mäsožravcov, konkrétne významne len vo vzorkách od zástupcu *Panthera leo*. Vzorky od mäsožravcov sú však výrazne variabilné, teda presne poukázať na diverzitu v porovnaní s bylinožravcami a všežravcami nie je možné.

Diskusia

V rámci práce som vykonal zhukovaciu analýzu, v ktorej na základe diverzity mojich vzoriek došlo k vzniku zhukov podľa jednotlivých zástupcov živočíchov 4.10. Z toho konštatujem, že mikrobióm jednotlivých druhov je dostatočne diverzný. V rámci druhov je však vo všetkých prípadoch pomerne stabilný 4.9. Zhukovanie podľa typu potravy následne odhalilo podobné kompozičné vzory 4.11, z čoho som usúdil, že typ potravy má jednoznačne vplyv na zloženie črevného mikrobiómu zvierat.

Svoju analýzu považujem za úspešnú, pretože výsledky súhlasia s uskutočnenými odbornými štúdiami [17][52]. Aj napriek tomu však analýza má svoje limitácie, a to v počte vzoriek či v spôsobe ich odberu. Vzorky odobrané naraz od jednotlivých zástupcov nemusia striktne popisovať ich charakteristickú kompozíciu črevného mikrobiómu. Ten je významne ovplyvnený potravou, ktorú zviera konzumovalo za posledných niekoľko hodín. Analyzované boli len tri vzorky odobrané naraz od každého zástupcu, teda výsledky presne popisujú kompozíciu črevného mikrobiómu len v momente odberu, a nie jeho všeobecné zloženie u daného jedinca. Táto limitácia by sa dala zlepšiť odberom viacerých vzoriek v rôznom čase, čím by sa obmedzil vplyv konzumovanej potravy na kompozíciu črevného mikrobiómu. Tieto zmeny by však vyžadovali značné zmeny v spôsobe spracovania a dizajn štúdie by sa tým podstatne zmenil. Napokon by tieto zmeny musela zohľadňovať aj štatistická analýza.

Analyzoval som vzorky od troch bylinožravcov, troch mäsožravcov a dvoch všežravých vtákov. Tento dataset by sa mohol rozšíriť čisto početne a u všežravcov aj o iné živočíšne triedy. Navyše všetky vzorky boli odobierané od zvierat žijúcich v zoo, čo znamená, že konzumujú potravu sprostredkovanú človekom. Tento fakt by mohol predstavovať významný vplyv na kompozíciu črevného mikrobiómu, preto by bolo v rámci lepšej informačnej hodnoty výsledkov dobré odoberať a analyzovať vzorky aj od zástupcov žijúcich vo voľnej prírode, kde si potravu zabezpečujú sami. Následné porovnanie výsledkov medzi jedincami žijúcich v zoo a jedincami žijúcich vo voľnej prírode by mohlo poskytnúť informácie o kompozícii črevného mikrobiómu ovplyvneného potravou poskytnutou človekom.

Záver

V úvodnej časti bakalárskej práce som sa venoval teoretickým poznatkom, ktoré sú predpokladom k nasledujúcej metagenomickej analýzy. Popísal som základné pojmy ako genetika a genomika spolu s metagenomikou. Mikrobióm ako predmet analýzy v mojej bakalárskej práci spolu s jeho zložením, typmi a metódami, ktoré sa využívajú na jeho analýzu. Podrobnejšie som sa venoval sekvenácii mikrobiómu, ktorá je neoddeliteľnou súčasťou analýzy mikrobiómu. Popísal som štruktúru ribozómu spolu s jeho štruktúrnou podjednotkou 16S rRNA a génom, kódujúcim túto sekvenciu, ktorý je významným prvkom vo výskume mikrobiómov. Objasnil som postupy využívané v taxonomickej analýze vzoriek podľa génu 16S ako aj metódy, ktorými sa následne vyhodnocuje ich diverzita.

V rámci teoretickej časti som popísal aj súbor bioinformatických nástrojov združených pod názvom QIIME2, ktorý som sa rozhodol použiť v praktickej časti na komplexnú analýzu a prepojenie jednotlivých krokov. Popísal som základné dátové formáty využívané pri práci s QIIME2 spolu s postupom spracovania dát, ktorý QIIME2 ponúka.

Zostavil som si dataset z voľne dostupných surových sekvenačných dát. Popísal som jeho pôvod, štruktúru a jednotlivé vzorky. Využil som nástroj FastQC na kontrolu kvality a nástroj FastP na zastrihnutie a filtráciu dát. Týmto spôsobom som z nekvalitných dát vytvoril kvalitný dataset, ktorý mi poslúžil ako výborný materiál na nasledujúcu metagenomickú analýzu.

Na analýzu kompozície zložiek mikrobiómu som využil nástroj QIIME2, pomocou ktorého som zaznamenal jednotlivé bakteriálne druhy vo vzorkách a zaradil ich do prislúchajúcich taxónov. V procese analýzy pomocou QIIME2 som využil aj možnosť analýzy hlavných komponentov a vytvoril som si tak PCA vizualizácie, z ktorých som spolu s výsledným barplotom odvodil závery analýzy.

V závere praktickej časti som analyzoval diverzitu vzoriek na úrovni zvierat s rôznym typom potravy. Popísal som alfa diverzitu medzi vzorkami od jednotlivých zvierat a beta diverzitu celkovo medzi bylinožravcami, mäsožravcami a všežravcami. Podarilo sa mi dostatočne priblížiť k očakávaným výsledkom, ktoré boli zistené v rámci publikovaných odborných štúdií.

Nad rámec mojej bakalárskej práce by sa analýza mohla rozšíriť o zvieratá z voľnej prírody, čím by sa dal porovnať vplyv stravy od človeka a stravy z voľnej prírody na zvieratá. Takisto by sa mohla diverzita porovnávať na úrovni živočíšnych tried ako *Mammalia*, *Aves*, *Reptilia*, *Arachnida* a i. Ďalším cieľom, ktorý by bolo možné dosiahnuť zmenami v dizajne analýzy, by mohlo byť určenie funkčného potenciálu mikrobiálnych komunít u jednotlivých zvierat. Tieto modifikácie analýzy, spočívajúce v prepájaní viacerých dátových modalít, by mohli poskytnúť

komplexnejší náhľad na celú roblematiku a možnosť priblížiť sa k multi-ómickeému spracovaniu dát.

Z vyššie uvedeného konštatujem, že všetky body zadania mojej bakalárskej práce som úspešne splnil.

Literatúra

- [1] Autosomal recessive disorder, 2022. URL: https://commons.wikimedia.org/wiki/File:Autorecessive_en_01.png.
- [2] Kazumasa Fukuda, Midori OGAWA, Hatsumi TANIGUCHI, and Mitsu Saito. Molecular approaches to studying microbial communities: Targeting the 16s ribosomal rna gene. *Journal of UOEH*, 38:223–232, 09 2016. doi:10.7888/juoeh.38.223.
- [3] Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature Reviews Microbiology*, 12(9):635–645, 2014.
- [4] Microbiology info.com, 2022. URL: <https://microbiologyinfo.com/polymerase-chain-reaction-pcr-principle-procedure-types-applications-and-ani>.
- [5] Qiime2 workflow, 2020. URL: https://chmi-sops.github.io/mydoc_qiime2.html.
- [6] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva, Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin D. Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights, Irina Koester, Tomasz Kosciolk, Jorden Kreps, Morgan G. I. Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erika Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S. Roberson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi

- Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin J. J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H. D. Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, 37(8):852–857, 2019. URL: <https://www.nature.com/articles/s41587-019-0209-9>, doi:10.1038/s41587-019-0209-9.
- [7] Kundan Kishor Rajak. Structure of ribosome. 2017.
- [8] Wen-Hsiung Li and Lori A Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–523, 1991.
- [9] Genetics vs. genomics fact sheet, 2018. URL: <https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics>.
- [10] Elizabeth A Grice, Heidi H Kong, Sean Conlan, Clayton B Deming, Joie Davis, Alice C Young, NISC Comparative Sequencing Program, Gerard G Bouffard, Robert W Blakesley, Patrick R Murray, et al. Topographical and temporal diversity of the human skin microbiome. *science*, 324(5931):1190–1192, 2009.
- [11] Rajesh Gupta, Q Beg, and Patrick Lorenz. Bacterial alkaline proteases: molecular approaches and industrial applications. *Applied microbiology and biotechnology*, 59(1):15–32, 2002.
- [12] Jose Antonio Garrido-Cardenas and Francisco Manzano-Agugliaro. The metagenomics worldwide research. *Current genetics*, 63(5):819–829, 2017.
- [13] Metagenomics, 2022. URL: https://www.genome.gov/genetics-glossary/Metagenomics?fbclid=IwAR0--zsNMvYQGfiDceo62tXBRaIb7Hgcftbs857BAWcdy_RUXpTVar5Q_6E.
- [14] Emily B Hollister, Chunxu Gao, and James Versalovic. Compositional and functional features of the gastrointestinal microbiome and their effects on human health. *Gastroenterology*, 146(6):1449–1458, 2014.
- [15] Claudia S Plottel and Martin J Blaser. Microbiome and malignancy. *Cell host & microbe*, 10(4):324–335, 2011.

- [16] Jian Cao, Yongfei Hu, Fei Liu, Yanan Wang, Yuhai Bi, Na Lv, Jing Li, Baoli Zhu, and George F Gao. Metagenomic analysis reveals the microbiome and resistome in migratory birds. *Microbiome*, 8(1):1–18, 2020.
- [17] Xiurong Guo, Hui Lei, Kailian Zhang, Famin Ke, and Can Song. Diversification of animal gut microbes and nrps gene clusters in some carnivores, herbivores and omnivores. *Biotechnology & Biotechnological Equipment*, 34(1):1280–1287, 2020.
- [18] Kirsten Grond, Brett K Sandercock, Ari Jumpponen, and Lydia H Zeglin. The avian gut microbiota: community, physiology and function in wild birds. *Journal of Avian Biology*, 49(11):e01788, 2018.
- [19] Les Dethlefsen, Margaret McFall-Ngai, and David A Relman. An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, 449(7164):811–818, 2007.
- [20] Joshua Lederberg. Infectious history. *Science*, 288(5464):287–293, 2000.
- [21] Julie S Biteen, Paul C Blainey, Zoe G Cardon, Miyoung Chun, George M Church, Pieter C Dorrestein, Scott E Fraser, Jack A Gilbert, Janet K Jansson, Rob Knight, et al. Tools for the microbiome: nano and beyond. *ACS nano*, 10(1):6–37, 2016.
- [22] David J Lane, Bernadette Pace, Gary J Olsen, David A Stahl, Mitchell L Sogin, and Norman R Pace. Rapid determination of 16s ribosomal rna sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82(20):6955–6959, 1985.
- [23] Jin-Der Wen, Syue-Ting Kuo, and Hsin-Hung David Chou. The diversity of shine-dalgarno sequences sheds light on the evolution of translation initiation. *RNA biology*, 18(11):1489–1500, 2021.
- [24] AJ Martinez-Murcia, AI Anton, and F Rodriguez-Valera. Patterns of sequence variation in two regions of the 16s rrna multigene family of escherichia coli. *International Journal of systematic and evolutionary microbiology*, 49(2):601–610, 1999.
- [25] Soumitesh Chakravorty, Danica Helb, Michele Burday, Nancy Connell, and David Alland. A detailed analysis of 16s ribosomal rna gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, 69(2):330–339, 2007.

- [26] Bo Yang, Yong Wang, and Pei-Yuan Qian. Sensitivity and correlation of hyper-variable regions in 16s rrna genes in phylogenetic analysis. *BMC bioinformatics*, 17(1):1–8, 2016.
- [27] J Craig Venter, Mark D Adams, Granger G Sutton, Anthony R Kerlavage, Hamilton O Smith, and Michael Hunkapiller. Shotgun sequencing of the human genome, 1998.
- [28] Jonathan A Eisen. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS biology*, 5(3):e82, 2007.
- [29] Paul DN Hebert, Thomas WA Braukmann, Sean WJ Prosser, Sujeevan Ratnasingham, Jeremy R DeWaard, Natalia V Ivanova, Daniel H Janzen, Winnie Hallwachs, Suresh Naik, Jayme E Sones, et al. A sequel to sanger: amplicon sequencing that scales. *BMC genomics*, 19(1):1–14, 2018.
- [30] Gerald Schochetman, Chin-Yih Ou, and Wanda K. Jones. Polymerase chain reaction. *The Journal of Infectious Diseases*, 158(6):1154–1157, 1988. URL: <http://www.jstor.org/stable/30137034>.
- [31] Stephanie Hammond, Jeffrey C Swanberg, Mihailo Kaplarevic, and Kelvin H Lee. Genomic sequencing and analysis of a chinese hamster ovary cell line using illumina sequencing technology. *BMC genomics*, 12:1–8, 2011.
- [32] Erwin L Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.
- [33] Anthony Rhoads and Kin Fai Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.
- [34] Ravi Ranjan, Asha Rani, Ahmed Metwally, Halvor S McGee, and David L Perkins. Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochemical and biophysical research communications*, 469(4):967–977, 2016.
- [35] Elizabeth J Summer. Preparation of a phage dna fragment library for whole genome shotgun sequencing. In *Bacteriophages*, pages 27–46. Springer, 2009.
- [36] Seok-Hwan Yoon, Sung-Min Ha, Soonjae Kwon, Jeongmin Lim, Yeseul Kim, Hyungseok Seo, and Jongsik Chun. Introducing ezbiocloud: a taxonomically united database of 16s rrna gene sequences and whole-genome assemblies. *International journal of systematic and evolutionary microbiology*, 67(5):1613, 2017.

- [37] Bonnie L Maidak, Niels Larsen, Michael J McCaughey, Ross Overbeek, Gary J Olsen, Karl Fogel, James Blandy, and Carl R Woese. The ribosomal database project. *Nucleic acids research*, 22(17):3485–3487, 1994.
- [38] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(D1):D590–D596, 2012.
- [39] Monika Balvočiūtė and Daniel H Huson. Silva, rdp, greengenes, ncbi and ott—how do these taxonomies compare? *BMC genomics*, 18(2):1–8, 2017.
- [40] Mikhail Tikhonov, Robert W Leach, and Ned S Wingreen. Interpreting 16s metagenomic data without clustering to achieve sub-otu resolution. *The ISME journal*, 9(1):68–80, 2015.
- [41] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639–2643, 2017.
- [42] Amy D Willis. Rarefaction, alpha diversity, and statistics. *Frontiers in microbiology*, 10:2407, 2019.
- [43] M Luz Calle. Statistical analysis of metagenomics data. *Genomics & informatics*, 17(1), 2019.
- [44] Robert J Whittaker, Katherine J Willis, and Richard Field. Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of biogeography*, 28(4):453–470, 2001.
- [45] Qiime2, 2022. URL: <https://docs.qiime2.org/2022.11/concepts/>.
- [46] Qiime2, 2022. URL: <https://docs.qiime2.org/2022.11/about/>.
- [47] Mehrbod Estaki, Lingjing Jiang, Nicholas A Bokulich, Daniel McDonald, Antonio González, Tomasz Kosciółek, Cameron Martino, Qiyun Zhu, Amanda Birmingham, Yoshiki Vázquez-Baeza, et al. Qiime 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Current protocols in bioinformatics*, 70(1):e100, 2020.
- [48] Joseph Brown, Meg Pirrung, and Lee Ann McCue. Fqc dashboard: integrates fastqc results into a web-based, interactive, and extensible fastq quality control tool. *Bioinformatics*, 33(19):3137–3139, 2017.

- [49] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 06 2016. arXiv:<https://academic.oup.com/bioinformatics/article-pdf/32/19/3047/25072524/btw354.pdf>, doi:10.1093/bioinformatics/btw354.
- [50] Andrei Prodan, Valentina Tremaroli, Harald Brolin, Aeilko H Zwinderman, Max Nieuwdorp, and Evgeni Levin. Comparing bioinformatic pipelines for microbial 16s rna amplicon sequencing. *PLoS One*, 15(1):e0227434, 2020.
- [51] Todd Z DeSantis, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. Greengenes, a chimera-checked 16s rna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072, 2006.
- [52] Nicholas D Youngblut, Georg H Reischer, William Walters, Nathalie Schuster, Chris Walzer, Gabrielle Stalder, Ruth E Ley, and Andreas H Farnleitner. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nature communications*, 10(1):1–15, 2019.