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

ZPRACOVÁNÍ A ANALÝZA LIDSKÉHO STŘEVNÍHO MIKROBIOMU ZE SEKVENAČNÍCH DAT 16S RDNA

PROCESSING AND ANALYSIS OF THE HUMAN GUT MICROBIOME FROM 16S RDNA SEQUENCING DATA

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — Michaela Zbudilová

AUTHOR

VEDOUCÍ PRÁCE — Ing. Markéta Nykrýnová

SUPERVISOR

BRNO 2023

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Michaela Zbudilová ***ID:*** 230352

Ročník: 3 ***Akademický rok:*** 2022/23

NÁZEV TÉMATU:

Zpracování a analýza lidského střevního mikrobiomu ze sekvenačních dat 16S rDNA

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma střevní mikrobiom a jeho složení, zaměřte se na metody analýzy střevního mikrobiomu a jeho následného vyhodnocení. 2) Navrhněte postup pro zpracování sekvenačních dat střevního mikrobiomu z Illumina MiSeq a dílčí části realizujte. 3) Všechna získaná sekvenační data zpracujte na základě navrženého postupu. 4) Statisticky a graficky vyhodnoťte diversitu mikrobiomu, stanovte jeho taxonomické rozdělení. Získané výsledky diskutujte. 5) Vytvořte skript pro generování reportu, který se použije pro diagnostické účely ve FN Brno.

DOPORUČENÁ LITERATURA:

[1] QIAN, Xu-Bo, Tong CHEN, Yi-Ping XU, Lei CHEN, Fu-Xiang SUN, Mei-Ping LU a Yong-Xin LIU. A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. Chinese Medical Journal. 2020, 133(15), 1844-1855. ISSN 0366-6999. DOI:10.1097/CM9.0000000000000871

[2] BHARTI, Richa a Dominik G GRIMM. Current challenges and best-practice protocols for microbiome analysis. Briefings in Bioinformatics. 2021, 22(1), 178-193. ISSN 1477-4054. DOI:10.1093/bib/bbz155

Termín zadání: 6.2.2023 ***Termín odevzdání:*** 29.5.2023

Vedoucí práce: Ing. Markéta Nykrýnová
Konzultant: Mgr. Matěj Bezdíček, Ph.D.

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

ABSTRAKT

Tato bakalářská práce se zabývá analýzou lidského střevního mikrobiomu z dat ze 16S rRNA. V první části je teoreticky popsán střevní mikrobiom, způsoby jeho zpracování a vyhodnocení pomocí analýzy taxonomických jednotek a diverzity mikrobiomu. Další část je zaměřena na data, která jsou v práci zpracována a na formát, v jakém jsou tyto data poskytnuta. V třetí části je popsán navržený algoritmus sloužící ke zpracování dat a zároveň jsou i vyhodnoceny výsledky získané spuštěním právě tohoto algoritmu. V další části práce jsou vzorky z Fakultní nemocnice Brno zpracovány pomocí navrženého algoritmu. Poslední část práce se zabývá popisem skriptu sloužícím ke generování reportů, které mohou být využity k diagnostickým účelům ve Fakultní nemocnici Brno.

KLÍČOVÁ SLOVA

střevní mikrobiom, 16S rRNA, analýza, diverzita, taxonomie

ABSTRACT

This bachelor´s thesis deals with the analysis of the human intestinal microbiome from 16S rRNA data. In the first part, the intestinal microbiome is theoretically described, and then the methods of its processing and evaluation using analysis of taxonomic categories and sample diversity are mentioned. The second part focuses on the data processed in the thesis and the format in which those data are provided. In the third part, the proposed algorithm used to process the data is described, and the results obtained by running this algorithm are evaluated. In the fourth part of the thesis, the samples from the University Hospital Brno are processed using the proposed algorithm. The last part of the thesis focuses on the script, which is used to generate the reports which can be used for diagnostic purposes in the University Hospital Brno.

KEYWORDS

intestinal microbiome, 16S rRNA, analysis, diversity, taxonomy

ZBUDILOVÁ, Michaela. *Zpracování a analýza lidského střevního mikrobiomu ze sekvenačních dat 16S rDNA*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2023, 60 s. Bakalářská práce. Vedoucí práce: Ing. Markéta Nykrýnová

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Michaela Zbudilová
VUT ID autora:	230352
Typ práce:	Bakalářská práce
Akademický rok:	2022/23
Téma závěrečné práce:	Zpracování a analýza lidského střevního mikrobiomu ze sekvenačních dat 16S rDNA

Prohlašuji, že svou závěrečnou práci jsem vypracovala 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 autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma 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 autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí mé bakalářské práce paní Ing. Markétě Nykrýnové za odborné vedení, konzultace, trpělivost a ochotu, kterou mi v průběhu zpracování této práce věnovala.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Střevní mikrobiom je nejobsáhlejší a pro lidský život nejdůležitější mikrobiom. Je důležitý pro zdravé fungování trávicí soustavy, kdy jeho činnost pomáhá při boji proti patogenům a tím výrazně ovlivňuje náš imunitní systém. Jeho rovnovážný stav lze však jednoduše narušit různými vnějšími jevy, mezi které kromě nevyvážené stravy a nezdravého životního stylu patří i například častá konzumace léků (převážně antibiotik). Kvůli těmto vnějším vlivům (například vlivem chemoterapie u onkologických pacientů) je dobré tyto změny co nejkvalitněji a co nejrychleji detekovat pomocí bioanalýzy.

Úvodní část bakalářské práce seznamuje čtenáře se základními pojmy a postupy, které jsou potřebné k porozumění toho, jak funguje analýza střevního mikrobiomu. V první části je teoreticky popsán střevní mikrobiom, jeho zpracování a způsoby vyhodnocení - konkrétně anatomie a fyziologie střev, složení střevního mikrobiomu, sekvenátory, postup sekvenace a nakonec vyhodnocení střevního mikrobiomu pomocí taxonomických jednotek a diverzity mikrobiomu. V druhé části práce jsou popsána data, která slouží jako testovací data pro návrh algoritmu a zároveň formát, ve kterém jsou tyto data poskytnuta. Třetí částí práce je kapitola nesoucí název Bioinformatické zpracování dat a je zaměřena na návrh algoritmu, popis jeho jednotlivých částí a aplikování na testovací data. V další části práce je popsáno aplikování navrženého algoritmu z předchozí kapitoly na data, která byla poskytnuta z Fakultní nemocnice Brno z Interní hematologické a onkologické kliniky, z Centra molekulární biologie a genové terapie. Na závěr práce je popsána tvorba a vzhled generovaných reportů, které slouží k diagnostickým účelům pro Fakultní nemocnici Brno.

1 Střevní mikrobiom, jeho zpracování a vyhodnocení

V této části práce je čtenář v prvních dvou kapitolách seznámen s anatomií a fyziologií střev a dále se samotným střevním mikrobiomem, který je pro práci klíčovým tématem. V další kapitole jsou podrobně rozebrána sekvenační data získaná ze střevního mikrobiomu (konkrétně 16S ribozomální ribonukleová kyselina), sekvenátory a samotné sekvenování sloužící k získání dat mikrobiomu. Nakonec je popsáno vyhodnocování střevního mikrobiomu na základě biodiverzit vzorku.

1.1 Anatomie a fyziologie střev

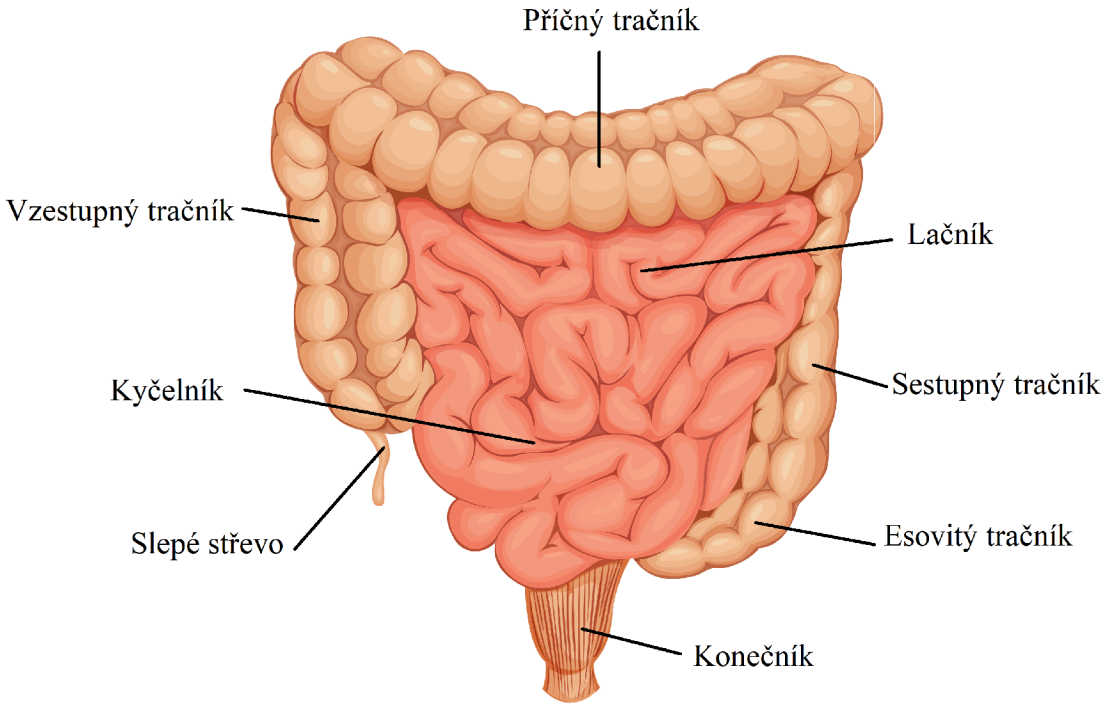
Střevo, jehož anatomie je k vidění na obrázku 1.1, je jednou z částí trávicí soustavy - a to konkrétněji částí trávicí trubice. Hlavními funkcemi trávicího systému je trávení, vstřebávání, přeměna živin, skladování živin a v neposlední řadě vylučování. Každá část trávicí trubice na základě její funkce má specifickou stavbu stěny, která se obecně skládá ze sliznice, podslizničního vaziva, zevní svalové vrstvy a serózy. [1], [2], [3]
Střevo dělíme na dva základní typy - na tenké střevo a tlusté střevo.

1.1.1 Tenké střevo

Tenké střevo zajišťuje trávení a vstřebávání živin. Svými rozměry je mnohem delší, než střevo tlusté - měří až sedm metrů. Z anatomického pohledu dělíme tenké střevo na tři části - dvanáctník, lačník a kyčelník.

Dvanáctník se svými 25 cm je nejkratším úsekem tenkého střeva. Jeho hlavní funkcí je mísení obsahu přicházejícího z žaludku se sekrety dvanáctníku, s žlučí či s enzymy slinivky břišní. Z důvodu velkého množství enzymů a trávicích šťáv (například žaludeční kyselina chlorovodíková) je vnitřní stěna dvanáctníku pokryta silnou vrstvou hlenového sekretu, který chrání střeva před chemickým i mechanickým poškozením.

Lačník a kyčelník tvoří druhou část tenkého střeva a jejich hlavní úlohou je především zpracování předpřipraveného obsahu z dvanáctníku. Aby docházelo co k nejefektivnějšímu způsobu štěpení a zároveň vstřebávání živin, je třeba, aby trávenina byla v co největším kontaktu se sliznicí střev. To je způsobeno specifickými kývavými a segmentačními pohyby, které tráveninu mísí, a zároveň pohyby peristaltickými, které ji posunují do tlustého střeva. [5], [6]



Obr. 1.1: Anatomie tlustého a tenkého střeva, upraveno a převzato z [4]

1.1.2 Tlusté střevo

Tlusté střevo dělíme na 6 hlavních částí - slepé střevo, vzestupný tračník, příčný tračník, sestupný tračník, esovitý tračník a konečník. Jeho hlavní funkcí je co největší možné zahuštění obsahu a vstřebání přebytečné vody - živiny se v tlustém střevě již nevstřebávají. Z důvodu častého hnití a kvašení zbytků potravy musí být vnitřní stěna tlustého střeva chráněna hlenem, který je produkován z hlenových žláz.

Slepé střevo je primární částí tlustého střeva, do které ústí kyčelník. Vybíhá z něj červovitý výběžek, který je tvořen mízní tkání. Často se u lidí projevují záněty právě červovitého výběžku, které mohou způsobit i záněty dalších částí trávicí trubice, a proto je slepé střevo často z důvodu zánětu chirurgicky odstraňováno. [6], [7]

1.2 Střevní mikrobiom

Lidský mikrobiom hraje v životě člověka velmi důležitou roli. Výrazně ovlivňuje boj proti patogenům či rozvoj imunitního systému - a to zejména v dětství, kdy se nám střevní mikrobiom tvoří například z pití mateřského mléka. Největší a nejpodstatnější mikrobiom je ve střevech a to především v oblasti tlustého střeva. Zde

organismy pomáhají nejen odbourávat potravu, ale i udržovat funkční imunitní systém.

Během života se však mikrobiom lidského těla neustále mění. Ve stáří nám ubývají prospěšné mikroorganismy, zatímco těch škodlivých narůstá (příkladem jsou organismy způsobující záněty). [8], [9]

1.2.1 Základní pojmy - mikrobiom, mikrobiota, metagenom

Pod pojmem střevní *mikrobiom* si můžeme představit společenství mikroorganismů žijících ve střevech, jejich genetický materiál a prostředí, ve kterém žijí. Čistě genetická stránka mikroorganismů je známa pod pojmem *metagenom*. Jedná se o veškerou genetickou informaci jedné komunity organismů. Skupinu mikroorganismů bez jejich genetického obsahu nazýváme *mikrobiota*. Spadají sem nejen bakterie, ale i viry, houby, prvoci, plísně a kvasinky. Můžeme tedy říct, že mikrobiom dělíme na dvě části - mikrobiotu a metagenom. [8]

Mikroflóra je pojem, který znamená v podstatě to samé, co mikrobiota. Jedná se o společenství organismů - z názvu však můžeme vyvodit, že se jedná o organismy rostlinného charakteru. Často se s tímto pojmem setkáváme i při popisu lidského biomu, označení je však chybné.

1.2.2 Složení střevního mikrobiomu

Lidský mikrobiom se na různých částech těla výrazně liší. Mikroorganismy mají různé funkce a jsou různě zastoupeny. Nejobsáhlejší a nejpodstatnější mikrobiom je však právě v trávicím traktu, konkrétně ve střevech.

V trávicím traktu nalezneme kolem 10^{14} bakterií společně s viry, plísněmi či houbami. Jejich hlavními funkcemi je převážně obrana proti různým typům patogenů, tvorba vitamínů (B12, K1 či K2), štěpení laktózy či vstřebávání minerálů. [10]

Bakterie

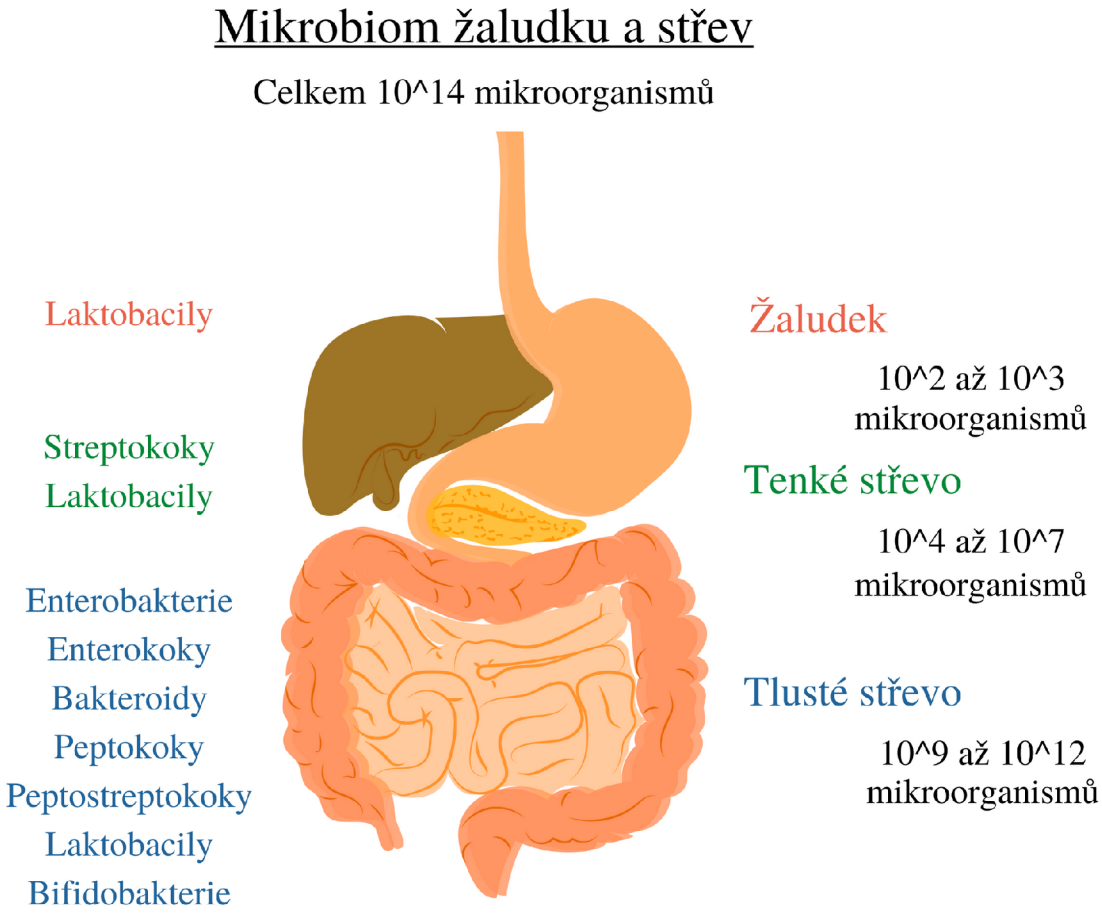
Nejrozšířenějším typem mikroorganismů střevního mikrobiomu jsou bezpochyby bakterie, které tvoří kolem 90 % z veškerého množství organismů. Tyto bakterie dělíme na dva základní typy - na anaeroby a aeroby. Anarobní bakterie k životu potřebují prostředí bez přítomnosti kyslíku. Opakem jsou aerobní bakterie, které pro svůj metabolismus kyslík potřebují. Poměr těchto dvou typů bakterií se ve střevech mění. V tenkém střevě je poměr 1:1, v tlustém střevě však vyskytuje mnohem více anaerobů. [11], [12]

Z pohledu kvalitativního zastoupení bylo zjištěno, že ve střevech máme dva dominantní mikrobiální kmeny: *Firmicutes* a *Bacteroidetes*. Tyto bakterie tvoří přibližně

90 % všech bakterií střev. Mezi další důležité bakteriální kmeny patří i *Actinobacteria*, *Proteobacteria* či *Verrucomicrobia*. Kvantitativní a kvalitativní zastoupení střevního mikrobiomu je k vidění na obrázku 1.2. [13], [14]

Bakterie typu *Firmicutes* jsou Gram-pozitivní bakterie, které jsou striktně anaerobní. Ve střevech jich nalezneme kolem 77,8 %. Příkladem mohou být například *Bacilli* a *Clostridia*. Stejně jako *Firmicutes* jsou i *Bacteroidetes* striktně anaerobní, jsou však Gram-negativní. Ve střevech jich je kolem 12,5 % z celkového množství bakterií. [15]

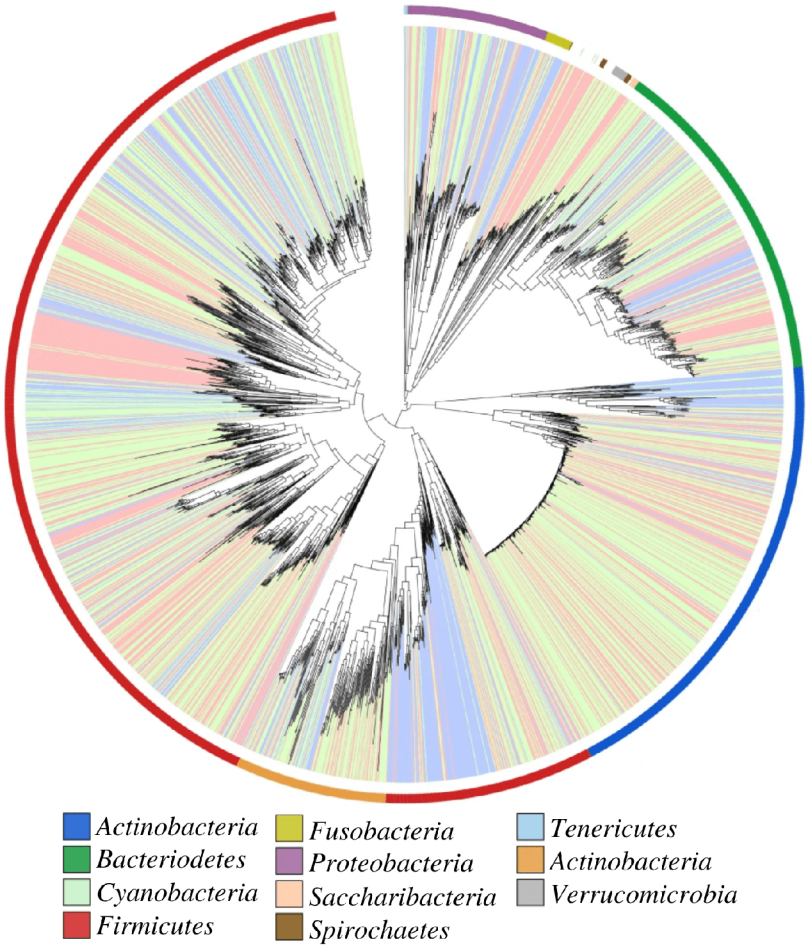
Na obrázku 1.3 je k vidění fylogenetický strom střevního mikrobiomu. V primární vrstvě u grafu jsou barevné škály, které podle legendy označují bakteriální kmeny.



Obr. 1.2: Kvantitativní a kvalitativní složení mikrobiomu žaludku a střev, upraveno a převzato z [13]

Viry

Množství virů v lidském mikrobiomu je oproti bakteriím zanedbatelné a kvůli tomu jsou viry i nedostatečně prostudovány. Nejčastěji se zkoumá vliv virů při různých



Obr. 1.3: Fylogenetický strom střevního mikrobiomu, upraveno a převzato z [16]

typech onemocnění a možnosti terapie. Z dostupných zdrojů víme, že kolem 90 % virů jsou bakteriofágové. Jedná se o viry, které díky své stavbě dokáží infikovat různé bakterie. Zbylých 10 % virů jsou viry eukaryotické. [14]

Houby

Stejně jako viry i houby jsou málo prozkoumanou oblastí lidského mikrobiomu - víme však, že nejrozšířenějším zástupcem hub ve střevech je kvasinka rodu *Candida*. Pro člověka a jeho zdraví životní styl je důležitá především *Candida albicans*. Jedná se o kvasinku, která při přemnožení způsobuje kandidózy - kvasinkové infekce. Nestojí zatím však jen *Candida albicans*, ale i jiné typy jako například *Candida tropicalis, Candida krusie* či *Candida parapsilosis*. Ty jsou přítomny nejčastěji ve střevech, ale i v dutině ústní. Jejich přemnožení nastává hlavně při snížené imunitě člověka a způsobují různé typy kandidóz - kožní, na sliznicích či vaginální. [14], [17]

1.2.3 Vývoj a ovlivňování střevního mikrobiomu během života

Vývoj lidského mikrobiomu je velmi proměnlivý během života každého jednotlivce. Rozvoj mikrobiomu jedince začíná již v děloze matky, kdy se do placenty přes krevní řečiště dostávají bakterie matky. Můžeme tedy říct, že mikrobiom plodu je vlastně mikrobiomem matky. Po porodu velkou roli pro rozvoj střevního mikrobiomu hraje pozření prvního mateřského mléka. Nejvíce se u novorozených dětí vyskytují *Stafylokoci* a *Enterobakterie*. U kojenců se jedná o *Bifidobakterie*. Během dospívání *Bifidobakterií* ubývá a začínají je nahrazovat kmeny bakterií *Bacteroidetes* a *Firmicutes*.

Složení se však mění nejen kvůli různým životním fázím, ale také kvůli našemu životnímu stylu. Mezi další faktory patří i genetika, strava, geografické prostředí, ze kterého pocházíme, ale i stres, hygienické návyky, prášky či infekce. [18], [19]

1.3 Data, sekvenátory a samotné sekvenování

1.3.1 16S rRNA

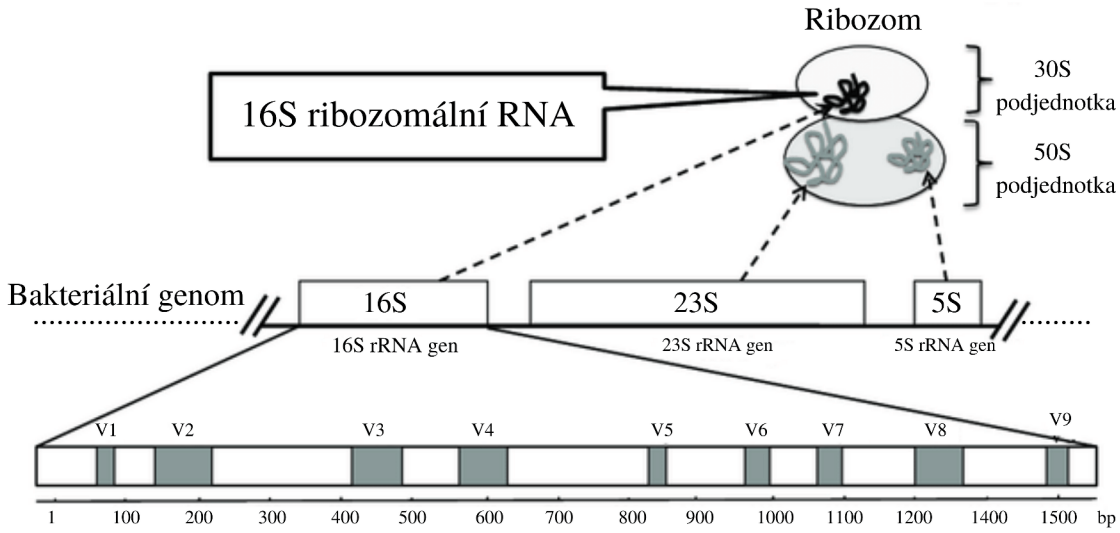
Data, na která je zaměřena tato práce, jsou krátké úseky ribonukleové kyseliny (RNA) nesoucí informace o složení lidského střevního mikrobiomu. Tyto geny se jmenují 16S ribozomální ribonukleové kyseliny (rRNA). Jedná se o úsek, který posunul vědu kupředu ve dvou hlavních kategoriích - evoluce a studium mikroorganismů. V případě evoluce jsme díky tomuto úseku rRNA zjistili, že pět říší, které jsme dříve považovali za nejvyšší klasifikační stupeň, můžeme přerozdělit do tří vyšších skupin - domén. V případě studia mikroorganismů jsme díky klonování a sekvenování těchto genů podrobněji charakterizovali mikrobiální diverzitu. [20]

Celý gen 16S rRNA je dlouhý kolem 1500 bp a je umístěn v malé podjednotce ribozomu 30S. Zbylé části genomu se nazývají 23S rRNA a 5S rRNA a jsou součástí velké podjednotky 50S. 16S rRNA obsahuje devět hypervariabilních oblastí (V1-V9), díky kterým jsme schopni rozlišit jednotlivé bakterie (a tím diverzitu vzorku) a také konzervativní oblasti, na které můžeme navrhovat konkrétní primery. Schématické zobrazení tohoto genu je k vidění na obrázku 1.4. [21]

Při sekvenování pomocí sekvenátoru Illumina MiSeq (kterému se budeme věnovat v další kapitole), jsou potřeba pouze oblasti V3 a V4 s přibližnou celkovou délkou 460 bp.

1.3.2 Sekvenátory nové generace

V roce 1953 byla definována Watsonem a Crickem struktura deoxyribonukleové kyseliny (DNA) dvojšroubovice. Na základě toho se začalo pracovat na možných



Obr. 1.4: Schéma ribozomálního komplexu a 16S rRNA genu, upraveno a převzato z [22]

způsobech sekvenování. Jedná se o metody, díky kterým jsme schopni získat pořadí nukleotidů z různých částí DNA či RNA. V sedmdesátých až osmdesátých letech dvacátého století paralelně vznikají dvě metody sekvenování - Maxam-Gilbert a Sanger, které společně patří do první generace sekvenátorů.

Další generace sekvenátorů přinesla mnoho kvalitních metod, které se denno-denně používají na celém světě, a to hlavně z důvodu jejich rychlosti a nízké ceny. Jedná se o takzvané sekvenátory druhé generace (NGS, z angl. Next Generation Sequencing), do které patří platformy jako *Illumina*, pomocí které byly získány vzorky pro tuto práci, dále *Roche 454 pyrosekvenování*, *Ion Torrent* či *SOLiD*. [23]

Illumina Solexa

Illumina je technologická firma, která vznikla v roce 1998 v San Diegu a v roce 2008 odkoupila firmu *Solexa*. Tohle spojení vedlo k vývoji nových sekvenátorů, přičemž mezi aktuálně používané patří *iSeq 100, MiniSeq, MiSeq Series, NextSeq 550 Series* či *NextSeq 1000 & 2000*.

Základním principem Illumina Solexa je sekvenování pomocí syntézy (SBS, z angl. sequencing-by-synthesis). Hlavními výhodami této platformy jsou rychlost a spolehlivost, nevýhodou je však vysoká cena oproti ostatním metodám. [24], [25], [26]

1.3.3 Příprava knihovny pro sekvenování pomocí Illumina sekvenátorem

Ze získaných vzorků bakteriální DNA je třeba vytvořit knihovnu, která je vhodná na samotné sekvenování. Příprava knihovny je prováděna nejčastěji na základě protokolu *16S Metagenomic Sequencing Library Preparation* [27]. Jednotlivé kroky jsou rozepsány v následujících kapitolách.

PCR amplifikace

Polymerázová řetězová reakce (PCR, Polymerase Chain Reaction) je metoda sloužící k amplifikaci DNA. Jedná se o základní princip využívaný k sekvenaci.

Prvním krokem při přípravě knihovny je amplifikace V3 a V4 oblastí 16S rRNA. K tomu jsou potřeba konkrétní primery - dopředný primer (5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG) a zpětný primer (5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC). Pomocí těchto primerů se vymezí oblast, kterou chceme namnožit.

U reakční směsi s templátovou mikrobiální DNA a konkrétními primery provedeme PCR - směs je vystavena cyklu, během kterého se na různou dobu mění teplota. [27], [28]

Přesný program je:

1. 95°C na 3 minuty
2. 25 opakujících se cyklů:
    - 95°C na 30 sekund
    - 55°C na 30 sekund
    - 72°C na 30 sekund
3. 72°C na 5 minut

První purifikace

Purifikace slouží k očištění DNA od všech nežádoucích částí vzorků. Používá se na ni mnoho komerčně dostupných purifikačních kitů. Nejčastěji však fungují na principu separace pomocí gravitačních kolonek. Při purifikaci pomocí fenol-chloroformu odstraňujeme z roztoku proteiny. Při čištění úseků 16S rRNA V3 a V4 se kromě jiného využívá i vymytí vzorku 80% ethanolem. [27], [28]

Indexace DNA

Po jednom týdnu uschování vzorku v -20°C následuje duálních indexace DNA amplikonu pomocí navržených indexů Nextera XT ve speciálních podmínkách, které jsou nastaveny následovně: [27], [28]

1. 95°C na 3 minuty
2. 8 opakujících se cyklů:
    - 95°C na 30 sekund
    - 55°C na 30 sekund
    - 72°C na 30 sekund
3. 72°C na 5 minut
4. Ponechat v klidu na 4 minuty

Druhá purifikace

Po indexaci je opět potřeba vzorek očistit a to stejným způsobem, jako při první purifikaci. Po dokončení čištění je vzorek taktéž na týden uschován v -20°C. [27], [28]

Ověření správnosti knihovny

Po druhém očištění je potřeba si ověřit, zda dosavadní kroky byly úspěšné. Jednou z metod může být zjištění délky amplikonu pomocí čipu *Bioanalyzer DNA*. Díky tomu, že víme, že předmětem naší analýzy jsou části V3 a V4 z 16S rRNA, můžeme očekávat výslednou délku kolem 630 bp. [27], [28]

Příprava na sekvenaci

Tato finální příprava obsahuje mnoho kroků, jako jsou: kvantifikace, normalizace, sdružování a denaturace. Kvantifikace lze provádět pomocí fluorometrické kvantifikace. Po normalizaci a sdružení knihoven je třeba vzorek denaturovat pomocí NaOH (hydroxid sodný) a zředit na požadovanou koncentraci. V tu chvíli je vzorek připraven na nanesení na kazetu MiSeq, kde se provede cílené nasekvenování. [27], [28]

1.3.4 Princip Illumina sekvenování

Po přípravě knihovny je možné zahájit sekvenování. Základním principem je SBS - na krátkou jednovláknovou sekvenci DNA se postupně navazují kompatibilní báze. Každá báze má na sobě navázanou specifickou fluorescenční látku, díky které jsem schopni identifikovat, jaká báze se na sekvenci navázala. Toto fluorescenční značení je detekováno vysoce citlivou kamerou a poté inhibováno, aby mohlo dojít k navázání a detekování další báze. Po přečtení celé sekvence dojde k zpětné rekonstrukci detekované sekvence DNA. [29]

Obecně (nejen u Illumina sekvenátorů) můžeme sekvenovat pomocí dvou základních technik - amplikon a shotgun. Tyto metody slouží k přípravě knihovny, což je soubor fragmentů, které prošly předzpracováním a jsou připraveny k sekvenování.

Amplikony jsou DNA produkty generované pomocí PCR. Tato technika sekvenování umožňuje rychlou metataxonomickou detekci. Základním principem je používání přesně navržených dopředných a zpětných primerů, které se naváží a vymezí tím cílenou oblast, kterou chceme následně osekvenovat. Metoda je tedy vhodná v případě, že přesně víme, jaká je cílená oblast našeho zájmu. Právě proto je vhodná pro analýzu 16S rRNA, protože přesně víme, které části chceme osekvenovat. [30]

Shotgun metoda na rozdíl od amplikonu nevyužívá navázání primerů, nýbrž celý genom rozdělí na malé fragmenty, které jsou sekvenovány individuálně. Následně jsou tyto osekvenované fragmenty vyhodnoceny počítačem, který na základě překryvů zrekonstruuje genom nazpět. Hlavním nevýhodou však může být kontaminace vzorku jinou DNA či RNA. V tom případě je důležité tyto nechtěné vzorky odstranit, protože by mohly narušit a negativně ovlivnit zpětnou rekonstrukci fragmentů. [30], [31]

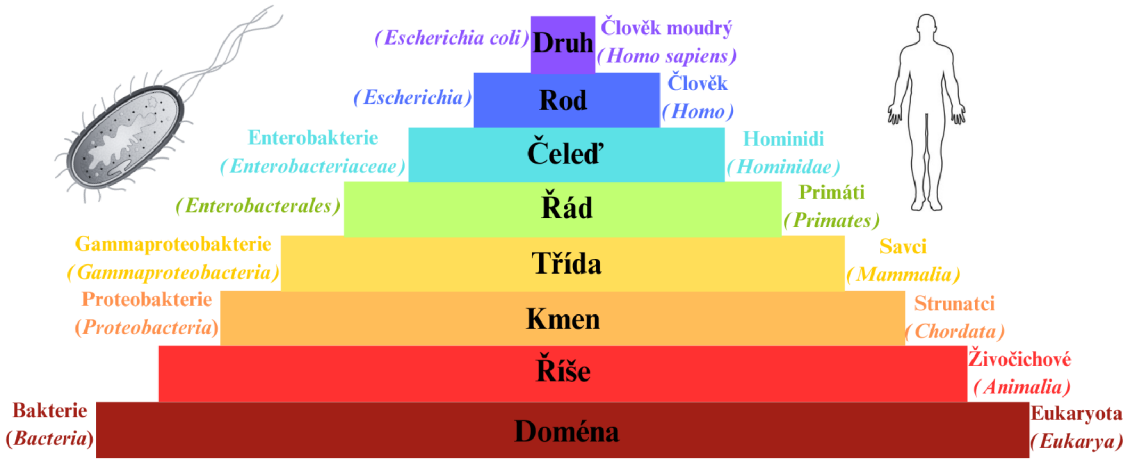
1.4 Vyhodnocení střevního mikrobiomu

Při analýze střevního mikrobiomu můžeme vyhodnocovat jeho složení podle taxonomických jednotek a jeho diverzitu. U složení můžeme bakterie definovat podle základní taxonomie bakterií, diverzitu definujeme na základě různých indexů a podobností.

1.4.1 Taxonomické jednotky

Taxonomie je vědní disciplína zabývající se klasifikací organismů do různých skupin - tyto skupiny se nazývají taxonomické kategorie. Máme osm hlavních kategorií - doména, říše, kmen, třída, řád, čeleď, rod a druh, přičemž druh je nejzákladnější jednotkou skládající se z jednotlivých konkrétních zástupců (v případě mikrobiomu se jedná o konkrétní bakterie). Existují tři hlavní domény - *Archea, Bacteria* a *Eucarya*, které jsou taxonomicky nejvýše. *Bacteria* se poté dělí na 29 větších kmenů, 90 % bakterií však spadá do pouhých 4 kmenů: *Proteobacteria, Actinobacteria, Firmicutes* a *Bacteriodetes*. Názorné taxonomické rozdělení jde vidět na obrázku 1.5, kde je pro ilustraci ukázána taxonomie člověka na pravé straně a taxonomie bakterie *Escherichia coli* na straně levé. [32], [33]

Taxonomii u bakterií dělíme na dva druhy - klasickou a molekulární. Klasická je založena na základě vizuálních podobností (fenotypu). Zde se bere v potaz morfologie bakterií (tedy velikost, tvar, počet či uspořádání bičíků), pohyblivost, fyziologie, výživa a další. Všechny tyto informace můžeme získat například mikroskopickým zkoumáním vzorků bez kultivace nebo pomocí barvení. U molekulární taxonomie se porovnávají informace z DNA (genotyp) a na základě podobnosti těchto informací



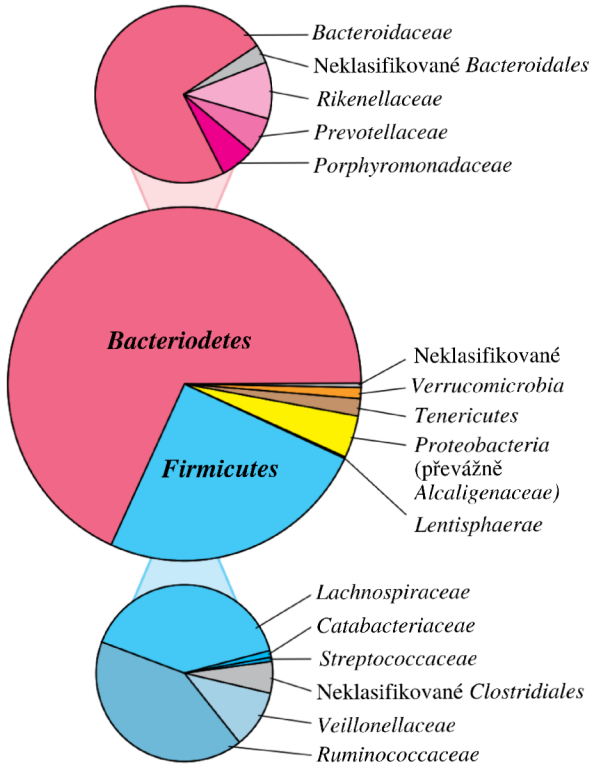
Obr. 1.5: Taxonomie člověka a bakterie *Escherichia coli*

opět probíhá klasifikace. Mezi molekulární techniky patří například analýza DNA, hybridizace nukleových kyselin, PCR či metoda fluorescenční *in situ* hybridizace. [33]

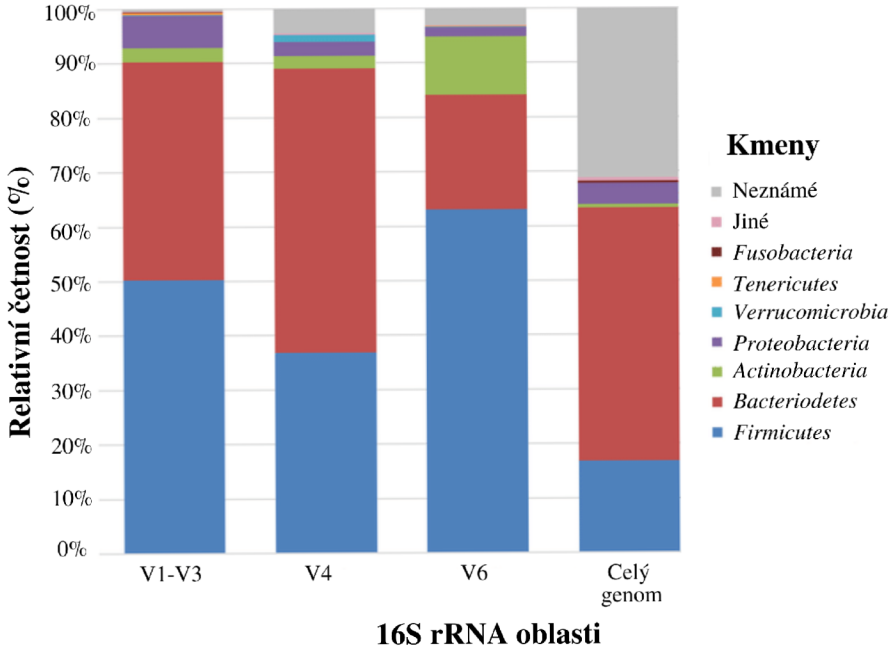
Díky moderním sekvenačním technologiím bylo v posledních letech možné konstruovat přesné fylogenetické stromy vedoucí k upřesňování bakteriální taxonomie. Moderní mikrobiální taxonomie byla převážně definována na základě sekvenace genu 16S rRNA. [34], [35]

Na základě výše zmíněných taxonomických kategorií můžeme při analýze střevního mikrobiomu tvořit přehledná grafická vyobrazení. Ukázka grafu je k vidění na obrázku 1.6, kde je zobrazena bakteriální diverzita stolice. Výsledný graf je sestrojen z výsledků analýzy přibližně milionu sekvencí ze 184 vzorků stolice, vše z variabilních oblastí V1 a V3 genu 16S rRNA. [33]

U střevního mikrobiomu taktéž ve spojitosti s taxonomickými kategoriemi můžeme vyhodnocovat jejich relativní četnost. Ta může být vztažená například na oblast výzkumu, na věk, pohlaví, část těla a podobně. Na následujícím obrázku 1.7 je vyobrazen sloupcový graf relativních četností, vyjádřený v procentech, bakteriálních kmenů v lidském střevním mikrobiomu vztažených na jednotlivé oblasti genu 16S rRNA. [36]



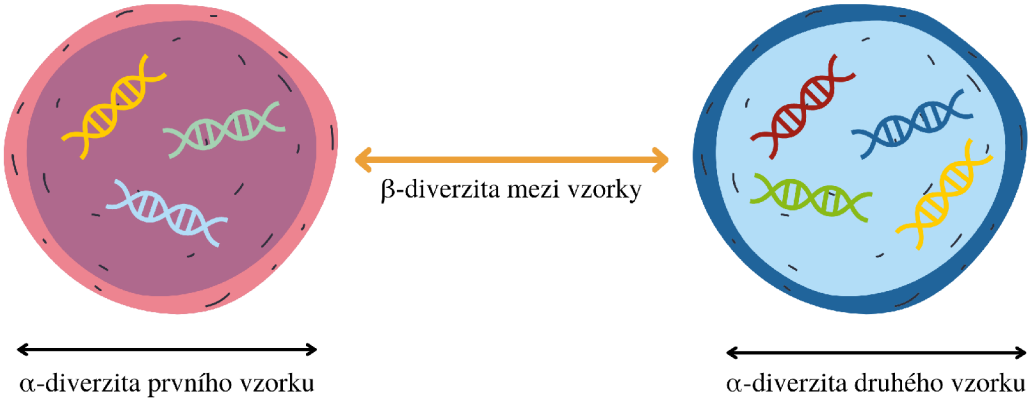
Obr. 1.6: Taxonomické zobrazení diverzity vzorků stolice, převzato a upraveno z [33]



Obr. 1.7: Relativní četnost [%] bakteriálních kmenů střevního mikrobiomu ve variabilních oblastech 16S rRNA, převzato a upraveno z [36]

1.4.2 Diverzita mikrobiomu

Pod pojmem diverzita (biodiverzita) si můžeme představit různorodost námi zkoumaného celku. Již v roce 1972 bylo R. H. Whittakerem popsány různé pohledy na biodiverzitu - konkrétně v jeho práci *Evoluce a měření druhové diverzity* [37]. Od té doby se biodiverzita stala velmi populární a zkoumanou problematikou. Podle výše zmíněného R. H. Whittekera existují 3 základní typy diverzit - alfa, beta a gama diverzita. V následujících kapitolách konkretizuji alfa a beta diverzitu, které jsou k vidění na obrázku 1.8, a jejich spojení se střevním mikrobiomem. Gama diverzita vyjadřuje celkovou druhovou diverzitu v jednom regionu získanou součinem alfa a beta diverzity. Ta však při analýze střevního mikrobiomu nehraje velkou roli, a proto není v této práci více definována. [38]



Obr. 1.8: Rozdíl mezi alfa a beta diverzitou

Alfa diverzita

Tento typ popisuje různorodost v rámci jedné oblasti (například ekosystém, mikrobiom či jiné lokální společenstvo). Konkrétně sledujeme počet druhů jednotlivých částí námi zkoumaného celku. Je tedy potřeba, abychom si přesně vymezili plochu nebo konkrétní společenstvo, které chceme zkoumat. [39]

Alfa diversitu můžeme popisovat dvěma typy indexů - Shannon-Wienerův indexem a Simpsonovým indexem.

Shannon-Wienerův index bere v potaz kromě počtu druhů i jejich početnost (kolik každého druhu je). Pomocí tohoto indexu lze vyjádřit vyrovnanost společenstva. Pokud by jeden druh ve společenstvu byl absolutně dominantní, index by nabýval hodnoty nula. Maximální hodnoty by index dosáhl ve chvíli, kdy by všechny druhy byly stejně početně zastoupeny.

Simpsonův index nabývá hodnot od nuly do jedné a na rozdíl od Shannon-Wienerova indexu nám říká, jak moc dominantní společenstvo může být. Pokud máme vysokou různorodost vzorku, tím je hodnota indexu menší. [40], [41]

Beta diverzita

Beta diverzita popisuje porovnání diverzit dvou systémů. Sledujeme tedy množství druhů jedné skupiny, které porovnáváme s množstvím druhé skupiny. Nejčastěji se k měření beta diverzity používá Jaccardova nepodobnost, která říká, jak moc nepodobné dvě zkoumané lokality jsou. [39], [40]

2 Datový formát a testovací data

V této části je nejdříve obecně popsán FASTQ formát, za kterým následuje rozbor konkrétních testovacích dat, která jsou v této práci bioinformaticky zpracována.

2.1 FASTQ formát

Existuje mnoho souborových formátů využívaných v bioinformatice, v této práci se však setkáváme pouze s jedním typem - FASTQ formát. V základním tvaru se jedná o čtyřřádkový textový formát sloužící k ukládání jednotlivých čtení ze sekvenátorů. První titulní řádek vždy začíná @ a nalezneme v něm ID sekvence. Druhý řádek představuje samotnou sekvenci nukleotidů. Na třetím řádku najdeme znaménko + signalizující konec sekvence a začátek čtvrtého řádku, kde je řetězec znamének kvality. [42]

Zde je k vidění příklad FASTQ formátu:

@M01938:189:000000000-B43KF:1:1101:17939:1792 1:N:0:1
CCTACGGGCGGCAGCAGTGGGGGATATTGC
+
AABBA33A@D?A2AEFGGFGGGA0EFGHBD

2.2 Testovací data

Data, která jsou zpracována v této práci, jsou součástí projektu PRJEB31801, jsou stažena z archivu ENA (European Nucleotide Archive, Evropský archiv nukleotidů) a slouží jako testovací data, na kterých byl navržen algoritmus, který je v další části práce aplikován na data z Fakultní nemocnice Brno. Jedná se o vzorky střevních mikrobiomů od osmi zdravých dobrovolníků. Sekvenace probíhala principem paired-end, což znamená, že fragment je čten nejprve z jedné strany a pak i z druhé. Proto od každého dobrovolníka máme dva FASTQ soubory, které ve svém názvu nesou buď R1 (read 1 = čtení 1) nebo R2 (read 2 = čtení 2). Délka jednoho čtení byla vždy 250 bp. Podrobnější informace o datech jsou k vidění níže v tabulce 2.1.

Tab. 2.1: Jednotlivé vzorky se stanoveným počtem čtení

Název vzorku	**Soubory FASTQ**	**Počet čtení**
ERR3237569	M4454_S1_L001_R1_001	150 832
	M4454_S1_L001_R2_001	150 832
ERR3237570	M4455_S2_L001_R1_001	200 365
	M4455_S2_L001_R2_001	200 365
ERR3237571	M4456_S3_L001_R1_001	247 692
	M4456_S3_L001_R2_001	247 692
ERR3237572	M4457_S4_L001_R1_001	92 357
	M4457_S4_L001_R2_001	92 357
ERR3237573	M4458_S5_L001_R1_001	186 625
	M4458_S5_L001_R2_001	186 625
ERR3237574	M4459_S6_L001_R1_001	191 739
	M4459_S6_L001_R2_001	191 739
ERR3237575	M4460_S7_L001_R1_001	212 378
	M4460_S7_L001_R2_001	212 378
ERR3237576	M4461_S8_L001_R1_001	270 582
	M4461_S8_L001_R2_001	270 582

3 Bioinformatické zpracování dat

V této kapitole je uveden návrh algoritmu na předzpracování a klasifikaci dat. Jednotlivé bloky diagramu jsou poté v podkapitolách podrobně popsány po teoretické stránce a taktéž obsahují jejich vyhodnocení pro testovací data.

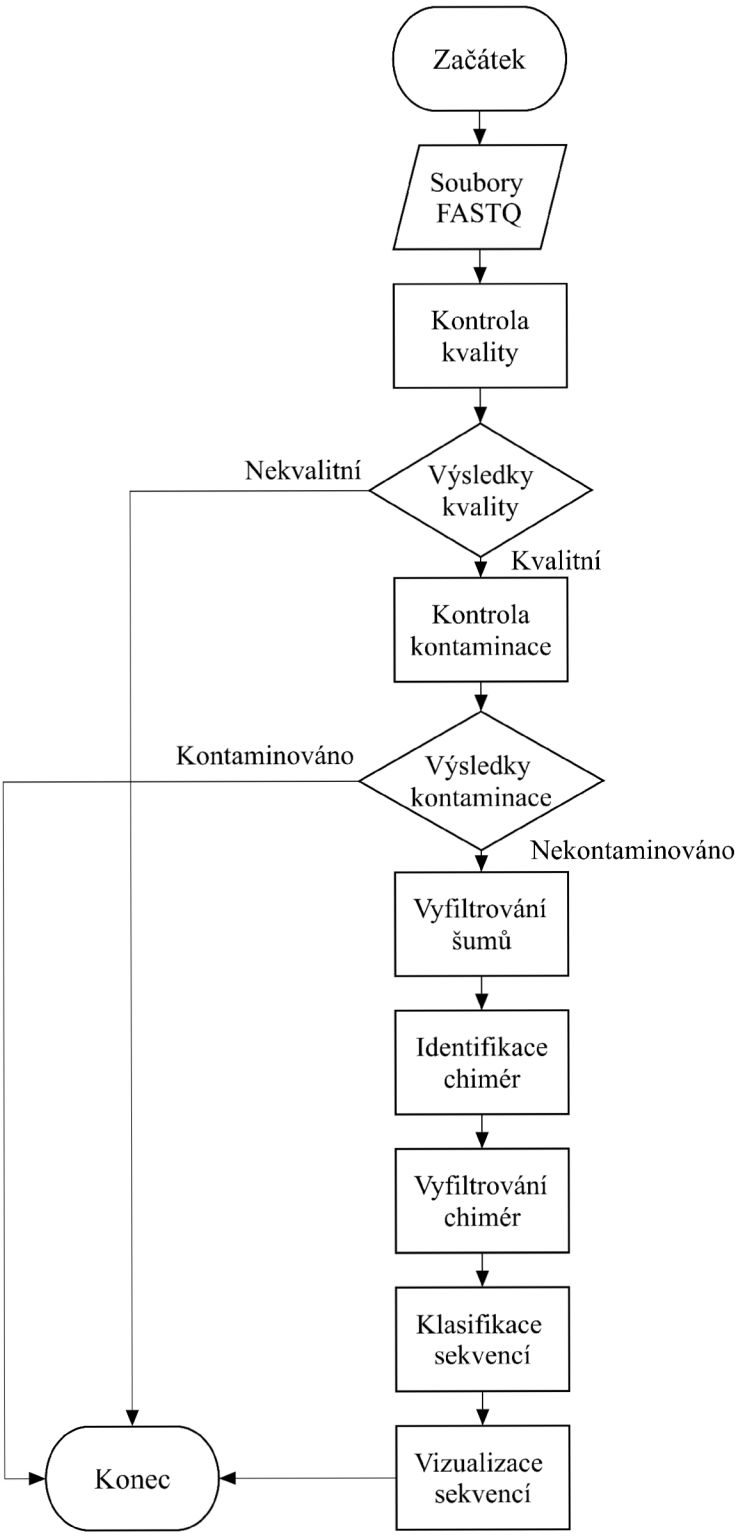
3.1 Návrh algoritmu

K bioinformatickému zpracování dat byl navržen algoritmus využívající nástroje QIIME2 [43]. Navržený postup je znázorněn blokovým diagramem, který je k vidění na obrázku 3.1, jehož jednotlivé části jsou podrobněji rozepsány v dalších kapitolách.

Vývojový diagram zobrazuje, jakým způsobem bude probíhat analýza dat. Na začátku diagramu nám vstupují FASTQ soubory s jednotlivými nasekvenovanými vzorky. Před cílenou klasifikací a vizualizací je potřeba provést kontrolu dat a ujistit se, zda jsou data dostatečně kvalitní, případně méně kvalitní části odfiltrovat.

Základní kontroly jsou dvě - kontrola kvality a kontrola kontaminace. Na základě výsledků těchto kontrol následuje v předzpracování další krok a tím je filtrace šumů. Pod pojmem šumy si můžeme představit části sekvencí s nízkou kvalitou či různé typy kontaminace, které by znehodnocovaly následnou analýzu dat. V případě, že by vzorky byly příliš kontaminovány či by měly velmi nízkou kvalitu, by bylo vhodné ukončit jejich předzpracování a odstranit je, aby negativně neovlivnily konečný výsledek analýzy.

Dalším krokem je identifikace chimér. Chiméry, jinak chimérické sekvence, jsou artefakty vzniklé nesprávným spojením dvou nebo více sekvencí, často během PCR. U 16S rRNA sekvencí se chimérických sekvencí vyskytuje řádově od 1 % do 5 %, i přes to je vhodné tyto části odfiltrovat. Odstraněním nekvalitních, kontaminovaných a chimérických části ze sekvencí je ukončeno předzpracování dat, a tím jsou data připravena na cílenou klasifikaci a vizualizaci. Výsledkem této analýzy je například taxonomická klasifikace či stanovení alfa diverzity vzorku. [44], [45]



Obr. 3.1: Vývojový diagram návrhu algoritmu

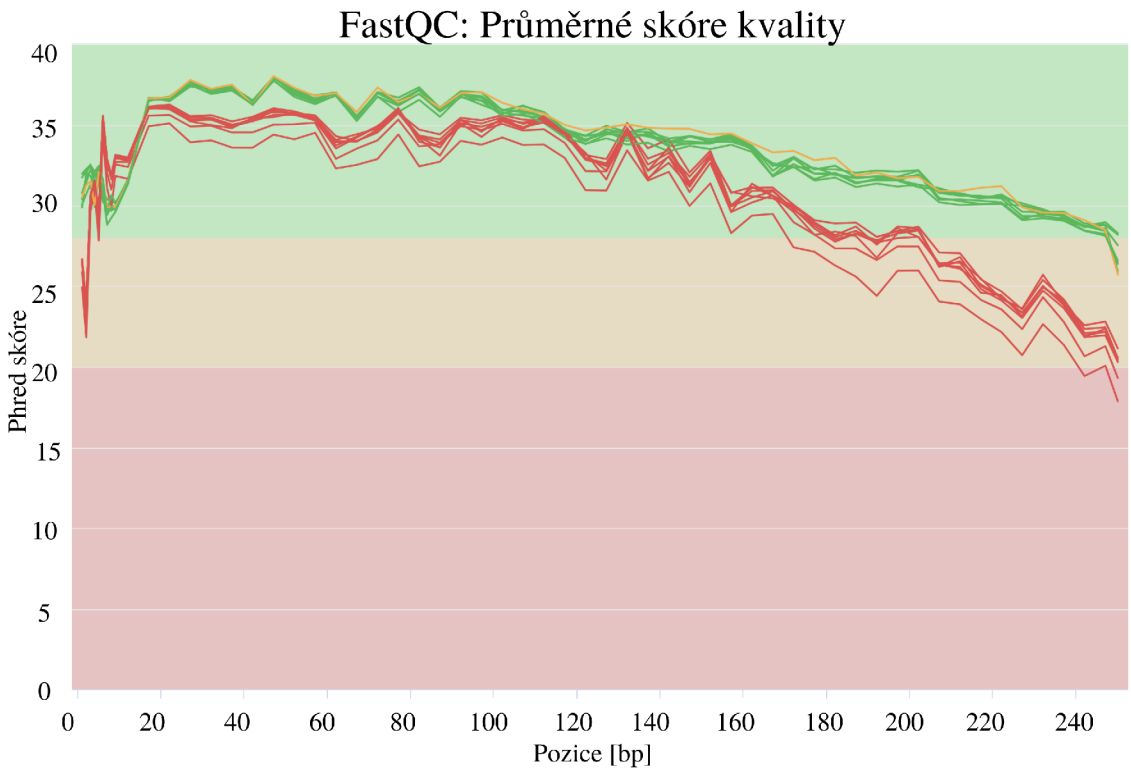
3.2 Kontrola kvality a její vyhodnocení

Prvním krokem při bioinformatické analýze by měla být kontrola kvality sekvenačních dat. Čtvrtý řádek FASTQ souborů, jak je nám již známo, představuje kvalitu jednotlivých bází sekvence. Tuto sekvenci znaků můžeme znát pod názvem *Phred skóre* nebo *Q skóre*. Skóre je zaznamenáno pomocí znaků ASCII, které jsou definovány na základě jednoduchých vzorců vycházejících z pravděpodobnosti chybného přiřazení báze. Vyšší skóre indikuje menší pravděpodobnost chyby, nižší skóre naopak značí chybu. [42]

Typickým znakem výstupů Illumina sekvenátorů je postupné klesání *Phred skóre* v závislosti na délce čtení. Jedná se o zcela normální stav způsobený samotným principem SBS, konkrétně fázováním. V každém cyklu je sekvenátor vymýván chemikáliemi, které blokují jednotlivé nukleotidy. Po dokončení cyklu je třeba odstranit tyto blokátory, aby cyklus mohl být znovu zahájen. Pokud blokátor není odstraněn, dochází k fázování a při dalším cyklu je starý nukleotid detekován podruhé. Od této chvíle bude tento fragment o jeden cyklus opožděn oproti ostatním fragmentům. S přibývajícími cykly jsou fragmenty čím dál tím více asynchronní, což se odrazí na klesající hodnotě *Phred skóre*. [46]

Kontrola kvality byla provedena pomocí programu FastQC (v0.11.5, [47]). Do ní vstupují jednotlivé FASTQ soubory, které pro každý vzorek vytvoří report v HTML prostředí. Díky programu MultiQC (v1.13, [48]) můžeme všechny výsledky dát do jednoho reportu, rovněž v HTML prostředí. Kromě základních statistických údajů (například délka jednotlivých sekvencí) nám program i graficky zobrazí průměrnou hodnotu kvality v závislosti na pozici ve čtení. Ukázka jednoho výstupu z MultiQC 3.2 je k vidění níže.

Podle grafu jsme schopni říct, které sekvence jsou kvalitní (zelené), nekvalitní (červené) nebo dostačující (oranžová). Taktéž si můžeme všimnout, že *Phred skóre* opravdu klesá tak, jak je zvykem u sekvencí z Illumina sekvenátorů. Podle grafu z programu MultiQC bychom mohli konstatovat, že kvalitních sekvencí je sedm a vždy se jedná o sekvence R1 - tedy sekvence, které byly čteny od začátku ke konci. Jedna dostačující je taktéž R1. Zbylých osm nekvalitních jsou na druhou stranu sekvence R2. To, že sekvence R2 jsou méně kvalitní je způsobeno výše popsaným znehodnocováním sekvencí vlivem fázování při SBS - nejdříve se sekvenovalo z jedné strany, poté z druhé, a proto je kvalita u R2 nižší než u R1. Nicméně i přes barevné rozlišení sekvencí na kvalitní a nekvalitní můžeme říct, že i nekvalitně zařazené sekvence jsou v závěru kvalitní - *Phred skóre* neklesá pod hodnotu 20, což je výsledek, se kterým by se šlo při bioinformatické analýze spokojit.



Obr. 3.2: Průměrné skóre kvality jednotlivých čtení

3.3 Kontrola kontaminace a její vyhodnocení

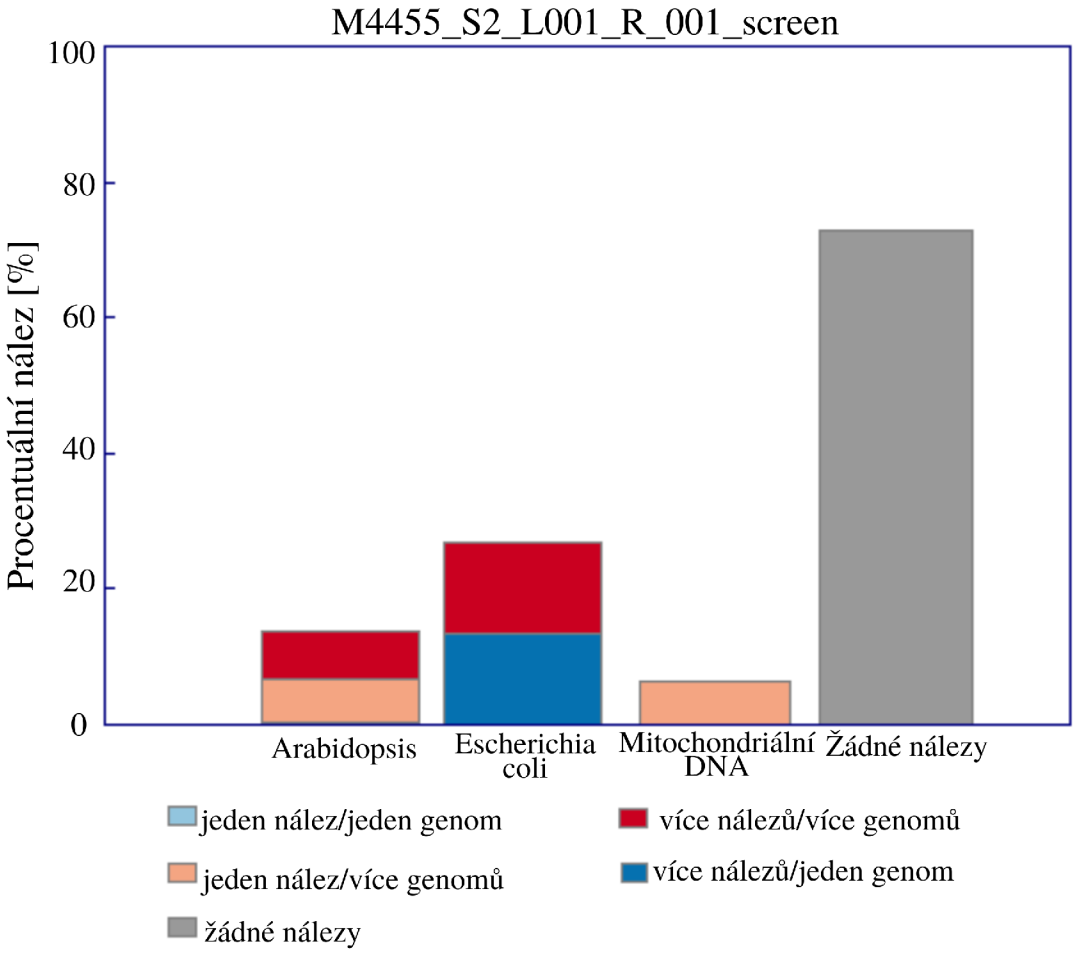
Dalším krokem by měla být kontrola, zda vzorky nebyly před či během sekvenování kontaminovány. Kontaminace může být způsobena všemi možnými způsoby, například kouskem kůže člověka, který vzorek připravoval. Kontrola probíhá tak, že sekvenační data jsou porovnávána s databází genomů, díky kterým můžeme zjistit, odkud data pocházejí.

Kontaminaci vzorku lze zjistit pomocí programu FastQ Screen (v0.15.1, [49]), který pro každou sekvenci vytvoří report v HTML prostředí a taktéž graf v PNG souboru, který je pro jeden vzorek k vidění na obrázku 3.3. Jedná se o sloupcový graf, který na ose x obsahuje jednotlivé genomové databáze a na ose y procentuální zastoupení. Níže ukázaný obrázek je upraven, na ose x byly kromě *Arabidopsis, Escherichia coli* a *mitochodnriální DNA* i další genomové databáze, jako třeba lidský, myší, potkaní či červí.

U zpracovávaných dat nebyla zjištěna markantní kontaminace (myšleno lidskými, myšími, červími či jinými genomy). Určitě procentuální zastoupení lze však pozorovat u skupin *Arabidopsis, Escherichia coli* a *mitochodnriální DNA*.

Arabidopsis označuje rod rostlin běžně se vyskytujících v mírném pásu - nejčastěji se jedná o roslinu *Arabidopsis thaliana*, česky známá jako huseníček rolní.

V jejich genomu jsou obsaženy geny, které se často shodují s geny bakterií, a proto se jedná o běžný nález u analýzy mikrobiomů. *Escherichia coli* je bakterie typická pro průjmová onemocnění. Přítomnost této bakterie ve vzorku ze střevního mikrobiomu je tedy běžná a analýzu nijak neovlivní. Mitochondriální DNA taktéž patří k obvyklým nálezům ve střevním mikrobiomu. [49], [50], [51]



Obr. 3.3: Ukázka výsledků kontroly kontaminace z nástroje FastQ Screen pro jeden vzorek

3.4 Filtrování šumů

K filtrování a samotné klasifikaci sekvencí byl použit tutoriál *Moving pictures* z platformy QIIME 2 [52]. Tento tutoriál obsahuje několik kroků, které vedou k získaní taxonomické klasifikace jednotlivých vzorků.

Prvním krokem v tomto tutoriálu je načtení dat, které se liší podle formátu vstupních dat. Formát dat, který používáme, se jmenuje Casava. Jedná se o typický výstupní formát FASTQ souborů z Illuminy sekvenátorů, kdy oproti klasickým FASTQ souborům je pozměněna hlavička následující za znakem "@". V hlavičce je navíc například uvedeno, jestli se jedná o R1 nebo R2. V případě, že vstupní data by byla multiplexována, by bylo potřeba provést ještě demultiplexaci. Naše vstupní data jsou však již demultiplexována, proto je tento krok vynechám.

Druhým krokem je filtrace dat. K tomu slouží dva nástroje - *DADA2* a *Deblur*. Při filtraci byl použit nástroj *Deblur*, který byl vytvořen právě na sekvenační data ze sekvenátorů Illumina MiSeq či Illumina HiSeq. Princip spočívá v tom, že sekvence se nejdříve seřadí podle abundance - to znamená, že se seřadí podle počtu čtení. Dále je vypočítána Hammingova vzdálenost (jedná se o vzdálenost určující, na kolika místech se dvě čtení od sebe liší), díky které jsou odstraněny sekvence, které nesplňují stanovené podmínky, a jsou tím pádem označeny za nevalidní. Po této části je dokončeno filtrování dat. [53]

3.5 Klasifikace sekvencí

Po filtraci následuje klasifikace, která je opět provedena pomocí tutoriálu *Moving pictures*. Aby bylo možné nálezy ve vzorcích taxonomicky přiřadit k správným bakteriím, je potřeba mít u vzorků klasifikátor. Jedná se o natrénovaný soubor, který v sobě obsahuje identifikace bakterií (obecně mikroorganismů) ve všech taxonomických úrovních. Díky tomu je možné v přiložených vzorcích porovnávat jednotlivé části sekvencí právě s tímto klasifikátorem a získat tak jejich taxonomickou klasifikaci. Klasifikátor je možné buď podle vlastní potřeby natrénovat a vytvořit si tak svůj vlastní z trénovacích dat nebo použít již vytvořené klasifikátory, které jsou volně stažitelné na internetu. V této práci byl použit již natrénovaný naivní Bayesovský klasifikátor z platformy QIIME2. Jedná se o klasifikátor získaný pomocí Bayesovské věty využívající techniku strojového učení. Konkrétně se jedná o klasifikátor z databáze Greengenes (verze 13.8, [54]) naformátovaný přímo pro použití s platformami QIIME1 a QIIME2. [55], [56]

3.6 Vyhodnocení výsledků

Hodnocení výsledků bylo provedeno dvěma způsoby - pomocí vizualizace taxonomických jednotek ve sloupcových grafech a vypočítáním alfa diverzity Shannon-Wienerovým indexem.

3.6.1 Vizualizace taxonomických jednotek

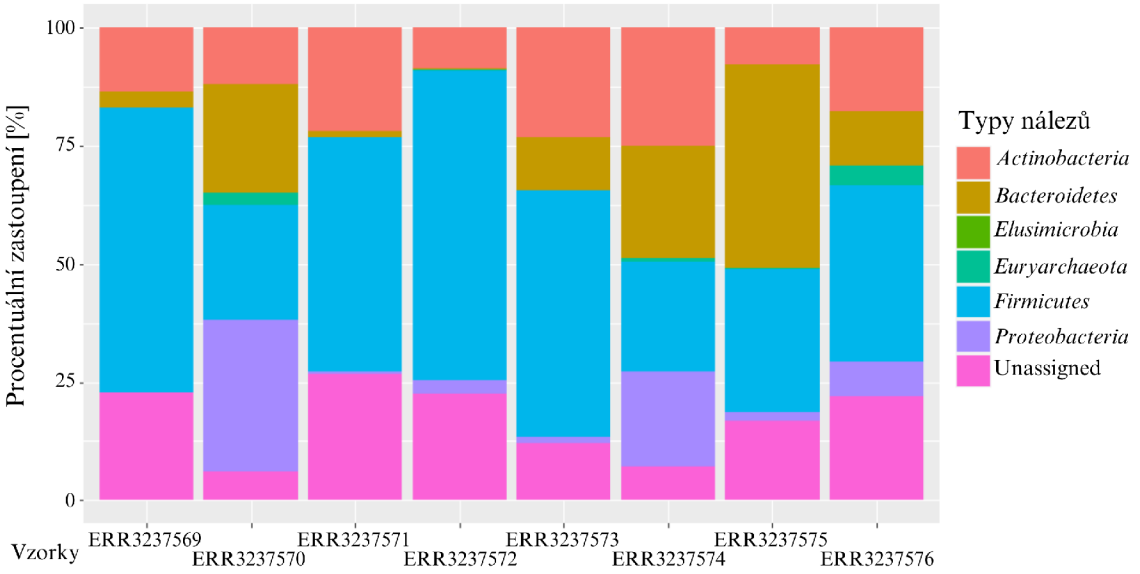
K vizualizaci klasifikace bylo třeba nechat si vygenerovat taxonomické výsledky. Pro každý vzorek tak byla vytvořena složka obsahující CSV soubory pro každou taxonomickou jednotku - celkově tedy sedm souborů, kdy v prvním souboru jsou uloženy výsledky na úrovni říše (nejvyšší taxonomická jednotka při analýze mikrobiomů, obecně je nejvyšší doména) a v sedmém souboru zase na úrovni druhu (nejnižší taxonomická jednotka).

K vykreslování grafů byly použity výsledky na úrovni kmenů (druhá nejvyšší jednotka) a na úrovni druhů. Při analýze mikrobiomů je drobná výjimka oproti jiným taxonomických analýzám - mikroorganismy nemají specifikovanou jednu celou kategorii - říši, která je v obecné hierarchické struktuře druhá nejvyšší (hned po doméně, viz obrázek 1.5). Pro usnadnění se tedy názvy domén (*Archea, Bacteria, Eucarya*) udávají jako názvy říší. Proto i v následujícím popisu dat tři základní domény budou popisovány jako říše.

Pro přehlednou vizualizaci výsledků byl vytvořen skript v jazyce R, který generuje sloupcové grafy zobrazující procentuální zastoupení nálezů ve vzorcích na různých úrovních. Do skriptu vstupuje pouze cesta do složky obsahující soubory s výsledky analýzy na druhé (kmenové) a sedmé (druhové) úrovni. Každý sloupec představuje jeden vzorek, jehož název je k vidění na ose x, na ose y je procentuálně vyjádřeno zastoupení nálezů, které jsou vypsány v legendě grafu. Mezi typy nálezů je vždy i kategorie Unassigned, kam spadají všechny nezařazené detekce v konkrétní úrovni (jinak řečeno, program dokázal rozpoznat, že daná část sekvence spadá do kmene *Firmicutes*, nedokázal však rozlišit, o jaký druh se jedná). Na kmenové úrovni jsou téměř všechny části zařazeny do konkrétních kmenů, na druhové úrovni je však mnohem těžší danou část sekvence zařadit ke specifickému druhu, a proto je v této úrovni mnohem větší podíl nezařazených detekcí oproti těm zařazeným.

První typ sloupcového grafu, k vidění na obrázku 3.4, vyjadřuje procentuální složení mikrobiomu na úrovni kmenů. S jistotou lze říct, že nejhojněji zastoupeným kmenem napříč téměř všemi vzorky je kmen *Firmicutes*. Hned po něj následují kmeny *Actinobacteria* a *Bacteroidetes*.

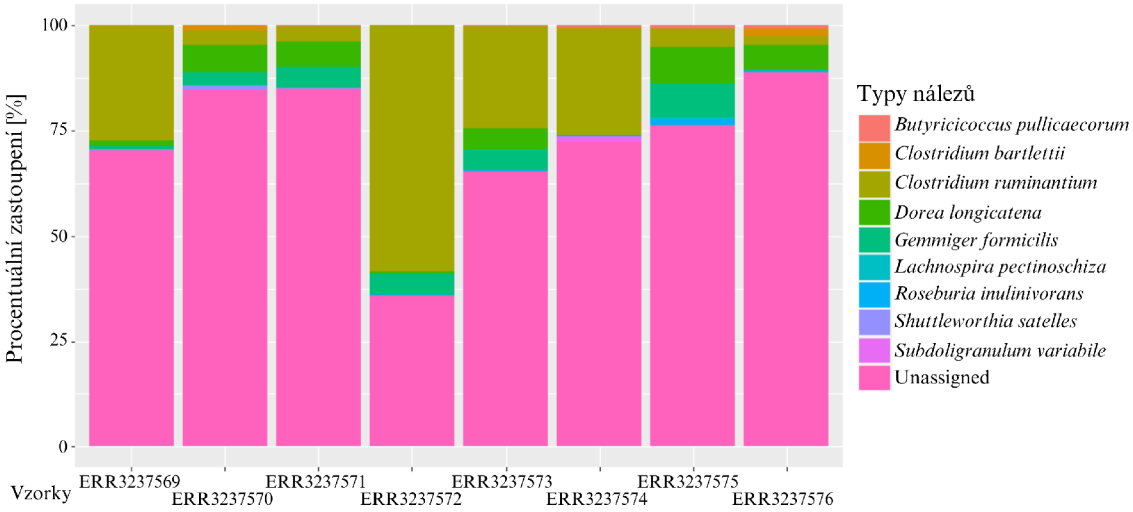
Dalším typem sloupcových grafů byly grafy zobrazující zastoupení druhů bakterií pro určitý kmen. Celkem tak byly vytvořeny čtyři grafy pro kmeny *Actinobacteria,*



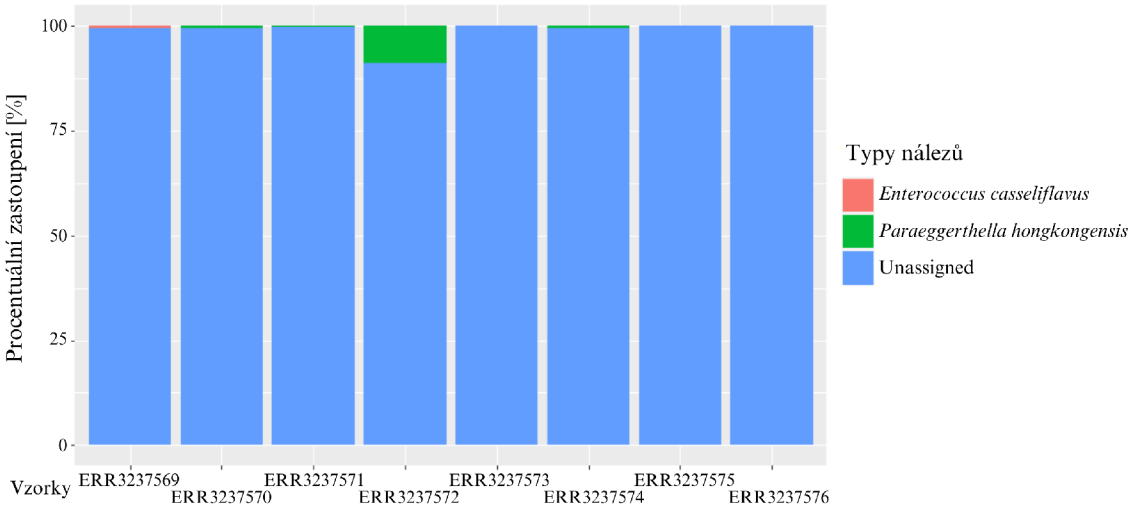
Obr. 3.4: Sloupcový graf zobrazující procentuální zastoupení [%] nálezů na úrovni kmene pro jednotlivé vzorky

Bacteroidetes, Firmicutes a *Proteobacteria*, a to z důvodu, že ve vzorcích právě tyto čtyři kmeny byly nejčastěji zastoupeny a měly nejrozmanitější druhovou kategorizaci. Na obrázku 3.5 je vidět kmen *Firmicutes* - u tohoto kmene bylo nejvíce detekovaných typů druhů bakterií, přičemž nevíce zastoupeným rodem bakterií byl rod *Clostridium*, speciálně druh *Clostridium ruminantium*. Celý tento rod se běžně vyskytuje ve střevních mikrobiomech obratlovců a bakterie jsou vůči hostitelům bráni jako komenzálové (organismy, které neškodně soužijí s hostitelem). Pokud ale dojde k přemnožení (zejména bakterie *Clostridium difficile*), může dojít ke vzniku klostridiové střevní infekce. [57], [58]

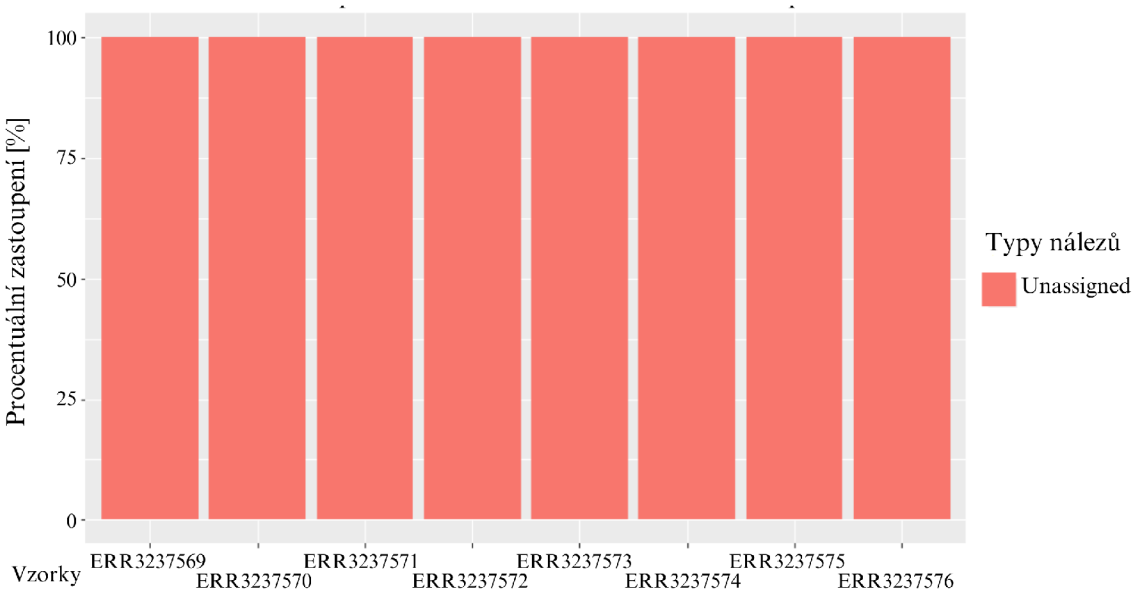
Na obrázku 3.6 je k vidění další druhová analýza, u které byly nalezeny konkrétní druhy bakterií, a to u kmene *Actinobacteria*, na rozdíl od kmenů *Bacteroidetes* (obrázek 3.7) a *Protebacteria* (obrázek 3.8), u nichž žádné druhy nebyly nalezeny.



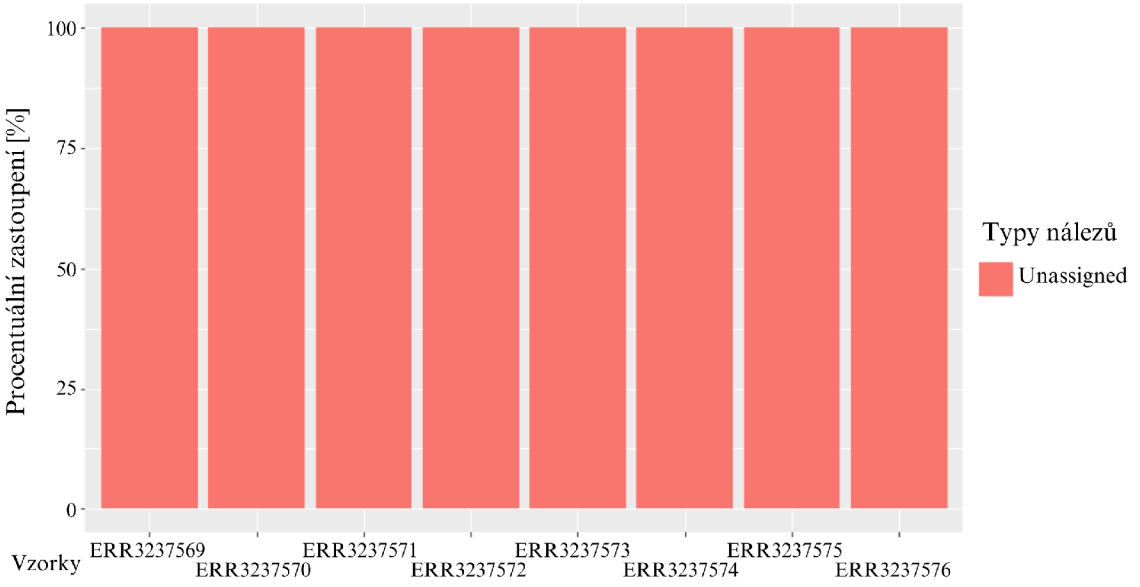
Obr. 3.5: Sloupcový graf zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Firmicutes* pro jednotlivé vzorky



Obr. 3.6: Sloupcový graf zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Actinobacteria* pro jednotlivé vzorky



Obr. 3.7: Sloupcový graf zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Bacteroidetes* pro jednotlivé vzorky



Obr. 3.8: Sloupcový graf zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Proteobacteria* pro jednotlivé vzorky

3.6.2 Analýza diverzity pomocí Shannon-Wienerova indexu

Pro určení diverzity (rozmanitosti) mikrobiomu byl pro každý vzorek vypočítán Shannon-Wiener index určující alfa diverzitu vzorku. Hodnota 0 znamená, že ve vzorku není žádná diverzita - jinak řečeno, v celém prostředí je pouze jeden jediný druh. Z toho vyplývá, že čím vyšší číslo, tím rozmanitější prostředí. Obvykle se hodnoty ale pohybují v rozmezí od 1.5 do 3.5. [59], [60]

K vypočítání byly použity výsledky z taxonomické analýzy na úrovni druhu nesoucí informace o počtu detekcí přiřazených ke konkrétnímu druhu bakterie. Všechny výsledky jsou k vidění v tabulce 3.1.

Tab. 3.1: Shannon-Wiener indexy značící alfa diverzitu pro jednotlivé vzorky

Název vzorku	**Shannon-Wiener index**
ERR3237569	2.27
ERR3237570	2.46
ERR3237571	2.38
ERR3237572	1.96
ERR3237573	2.47
ERR3237574	2.41
ERR3237575	2.14
ERR3237576	2.59

Z těchto výsledků můžeme usoudit, že nejrozmanitější diverzitu má vzorek ERR3237576 na rozdíl od vzorku ERR3237572, jehož diverzita je nejmenší. Stále jsou však všechny hodnoty v rozmezí 1.5 až 3.5, z čehož můžeme vyvodit, že ve všech případech byly vzorky dostatečně rozmanité.

4 Zpracování dat z Fakultní nemocnice Brno

Následující část bakalářské práce se zabývá zpracováním dat z Fakultní nemocnice (FN) Brno - konkrétně z Interní hematologické a onkologické kliniky, z Centra molekulární biologie a genové terapie. V pilotní studii bylo získáno od 22 hematologických pacientů (tedy od pacientů, u kterých bylo diagnostikované onemocnění krve) 59 vzorků stolice. Postup předzpracování i samotné analýzy byl stejný jako v navrženém algoritmu pro předchozí data, který je k vidění na obrázku 3.1.

4.1 Zpracovávaná data

Stejně jako v předchozí analýze dat jsou i data ze FN ve formátu FASTQ. K sekvenaci bylo zasláno 16 vzorků. U vzorků číslo 10, 11, 13 a 15 však došlo k chybě a úspěšně se osekvenovalo pouze 12 vzorků. Vzorky 1 až 8 byly připraveny pomocí kitu typu *PowerFecal* a vzorky 9 až 16 zase pomocí kitu typu *Mole P084B*. Sekvenační kit je sada obsahující všechny složky (chemikálie a potřebné nástroje) potřebné k sekvenování. [61]

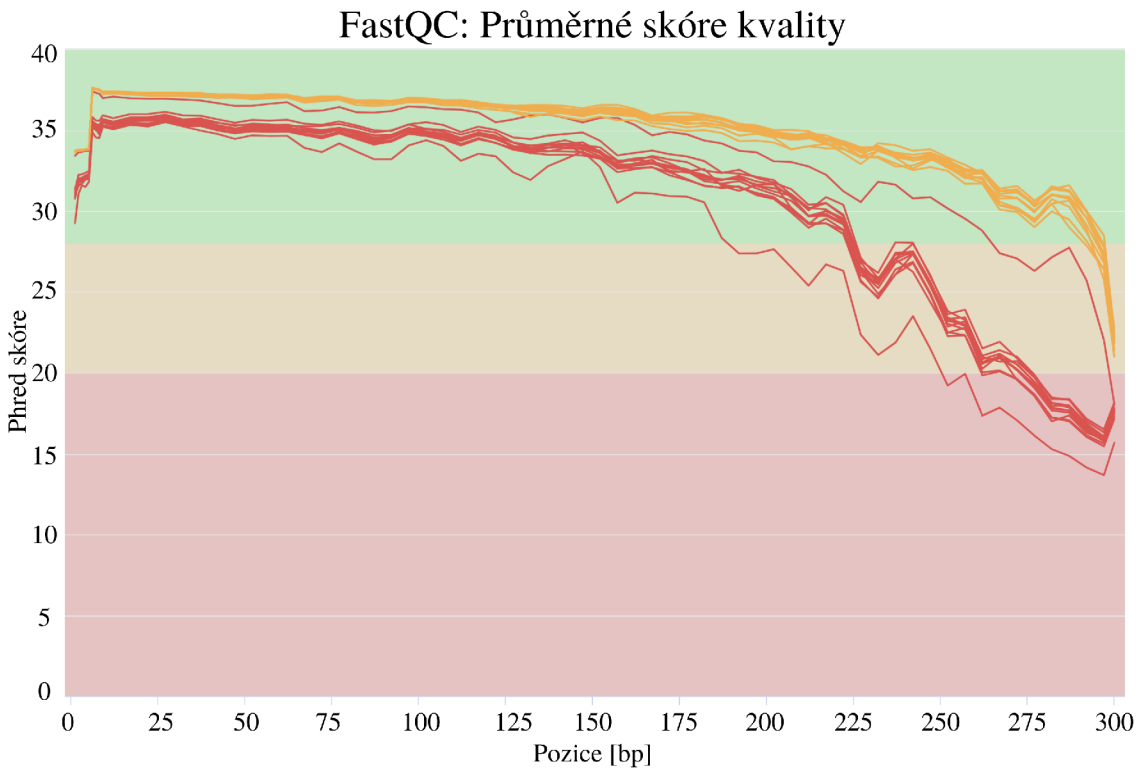
Podrobnější informace jsou uvedeny v tabulce 4.1. V tabulce vidíme, že sekvenace opět probíhala principem paired-end, takže od každého vzorku máme dva soubory a délka jednoho čtení je 300 bp.

4.2 Hodnocení kvality

Hodnocení kvality pro vzorky z FN Brno je k vidění na obrázku 4.1, který zobrazuje průměrné skóre kvality jednotlivých čtení na určité pozici. Křivky kvality jsou sice programem FastQC vybarveny oranžově a červeně (což značí, že jsou méně kvalitní až nekvalitní), hodnota Phred skóre se však z velké části pohybuje nad 20. V tom případě můžeme sekvence považovat za kvalitní a není proto třeba žádných dalších kroků vedoucích ke zlepšení jejich kvality. Taktéž můžeme vidět klesavou tendenci kvality, která je typická pro Illumina sekvenátory.

Tab. 4.1: Jednotlivé vzorky se stanoveným počtem čtení a s typem sekvenačního kitu

Název vzorku	**Soubory FASTQ**	**Počet čtení**	**Sekvenační kit**
GA-230303LADFMB01	S1_L001_R1_001 S1_L001_R2_001	225 634 225 634	PowerFecal
GA-230303LADFMB02	S2_L001_R1_001 S2_L001_R2_001	119 835 119 835	PowerFecal
GA-230303LADFMB03	S3_L001_R1_001 S3_L001_R2_001	216 106 216 106	PowerFecal
GA-230303LADFMB04	S4_L001_R1_001 S4_L001_R2_001	190 851 190 851	PowerFecal
GA-230303LADFMB05	S5_L001_R1_001 S5_L001_R2_001	183 068 183 068	PowerFecal
GA-230303LADFMB06	S6_L001_R1_001 S6_L001_R2_001	169 694 169 694	PowerFecal
GA-230303LADFMB07	S7_L001_R1_001 S7_L001_R2_001	191 792 191 792	PowerFecal
GA-230303LADFMB08	S8_L001_R1_001 S8_L001_R2_001	197 356 197 356	PowerFecal
GA-230303LADFMB09	S9_L001_R1_001 S9_L001_R2_001	184 064 184 064	Mole P084B
GA-230303LADFMB12	S10_L001_R1_001 S10_L001_R2_001	147 374 147 374	Mole P084B
GA-230303LADFMB14	S11_L001_R1_001 S11_L001_R2_001	201 583 201 583	Mole P084B
GA-230303LADFMB16	S12_L001_R1_001 S12_L001_R2_001	201 566 201 566	Mole P084B



Obr. 4.1: Průměrné skóre kvality jednotlivých čtení pro data z FN Brno

4.3 Hodnocení kontaminace

Po spuštění programu FastQ Screen (v0.15.1, [49]) byly patrné malé stopy kontaminace - konkrétně u myších, krysích, lidských a červích genomů byly nalezeny shody. I přes velmi malé hodnoty shod bylo potřeba kontaminaci ze vzorků vyfiltrovat, aby negativně neovlivňovala následující analýzu. Pro filtraci byly použity nástroje MiniMap2 (v2.26 r1175, [62]) a Samtools (v1.17, [63]). Pomocí Minimap2 byl namapován každý vzorek s referencí - s lidským genomem *GRCh38.p14* [64] staženým z databáze Národního centra pro biotechnologické informace. Pomocí nástroje Samtools2 poté dojde k vyfiltrování části, které nejsou namapované a nakonec výstup uloží ve formátu FASTQ.

4.4 Výsledky analýzy

Stejně jako u předchozích dat bylo po spuštění skriptů vygenerováno pět grafických reprezentací procentuálního zastoupení nálezů ve vzorcích. Na prvním vygenerovaných grafů je procentuální vyobrazení nálezů na úrovni kmene, který je na obrázku 4.2. Kmeny *Actinobacteria, Bacteroidetes, Cyanobacteria, Elusimicrobia, Firmicutes, Fusobacteria* a *Proteobacteria* jsou kmeny spadající do říše *Bacteria* na rozdíl

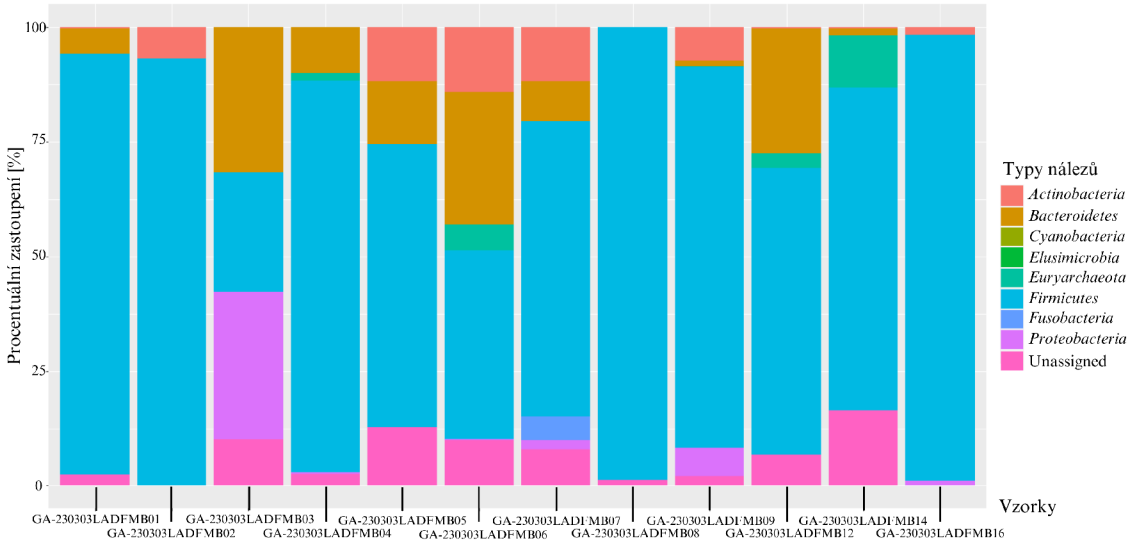
od kmene *Euryarchaeota*, který jako jediný z detekovaných kmenů patří do říše *Archaea*. U jedenácti vzorků lze jasně určit, že nejhojněji detekovaným kmenem je kmen *Firmicutes* - například u vzorku GA-230303LADFMB03 tvoří přes 80 % všech bakterií. Naopak u vzorku GA-230303LADFMB03 jsou nejvíce zastoupeny kmeny *Bacteroidetes* a *Proteobacteria*.

Na obrázku 4.3 je druhové zastoupení pro kmen *Actinobacteria* - můžeme vidět, že ve většině případů nebylo možné přesně zařadit bakterie až na úrovni druhu. U dvou vzorků však byly detekovány dvě konkrétní bakterie - u vzorku GA-230303LADFMB05 se jedná o bakterii *Paraeggerthella hongkongensis* a u vzorku GA-230303LADFMB07 o bakterii *Bifidobacterium breve*.

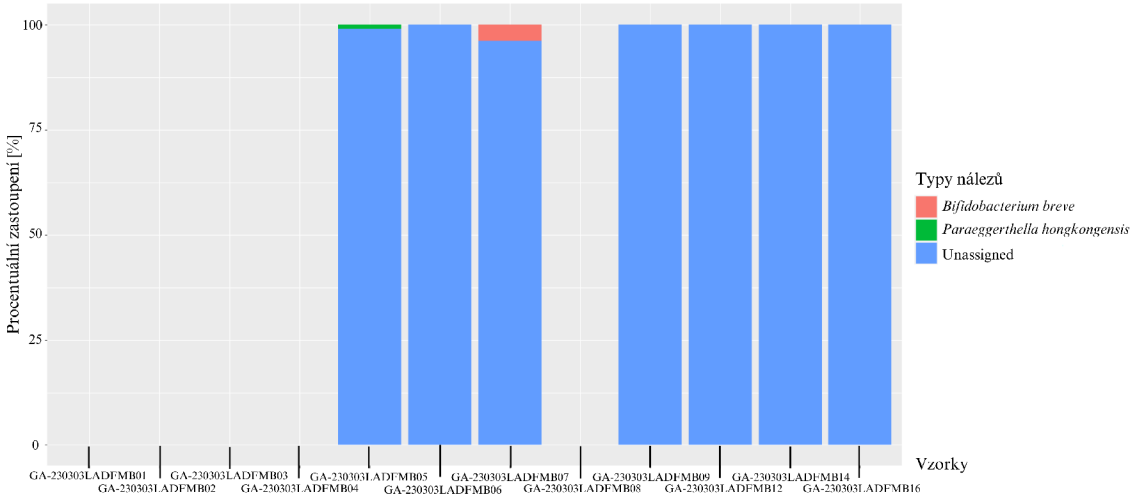
Na obrázku 4.4 je k vidění druhové zastoupení kmene *Bacteroidetes*. Zde jsou nejpočetnější nálezy *Alistipes onderdonkii* a *Human gut metagenome* (v tomto případě se nejedná o konkrétní druh bakterie, nýbrž o identifikovaný lidský střevní metagenom, který byl rozpoznán díky klasifikátoru při taxonomické analýze).

Obrázek 4.5 vyobrazuje druhovou analýzu pro kmen *Firmicutes*. Na první pohled lze usoudit, že oproti ostatním kmenům byl právě tento nejvíce dopodrobna detekován. Podle grafu souhrnně můžeme říct, že nejvíce se ve vzorcích vyskytuje rod bakterií *Clostridium* (konkrétně bylo detekováno šest druhů bakterií spadajících právě do tohoto rodu).

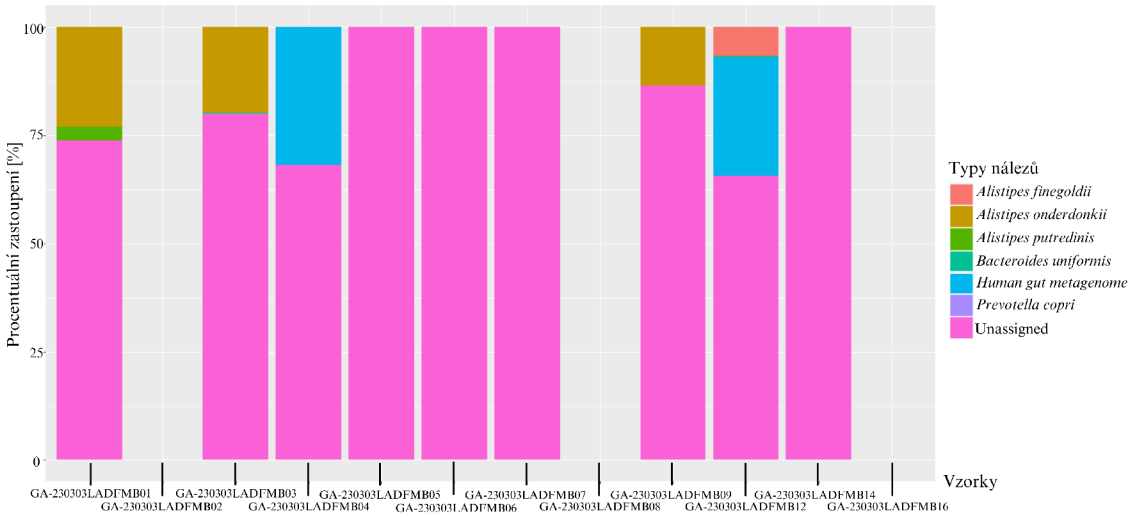
Poslední graf zobrazuje na obrázku 4.6 druhy kmene *Proteobacteria*. Zde byla detekován pouze jeden typ bakterie, a to u *Sphingobium estrogenivorans* u vzorku GA-230303LADFMB16.



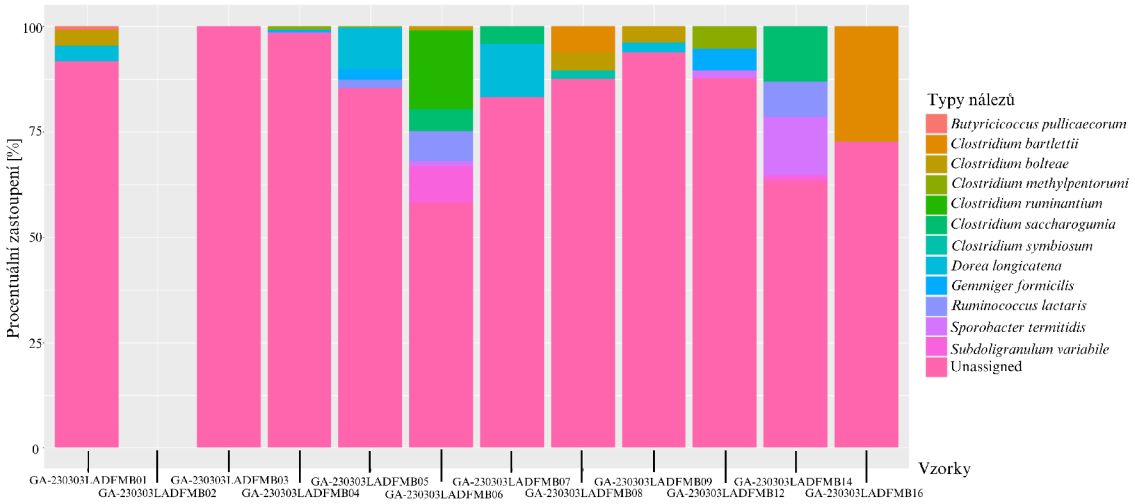
Obr. 4.2: Sloupcové grafy zobrazující procentuální zastoupení [%] nálezů na úrovni kmene pro jednotlivé vzorky



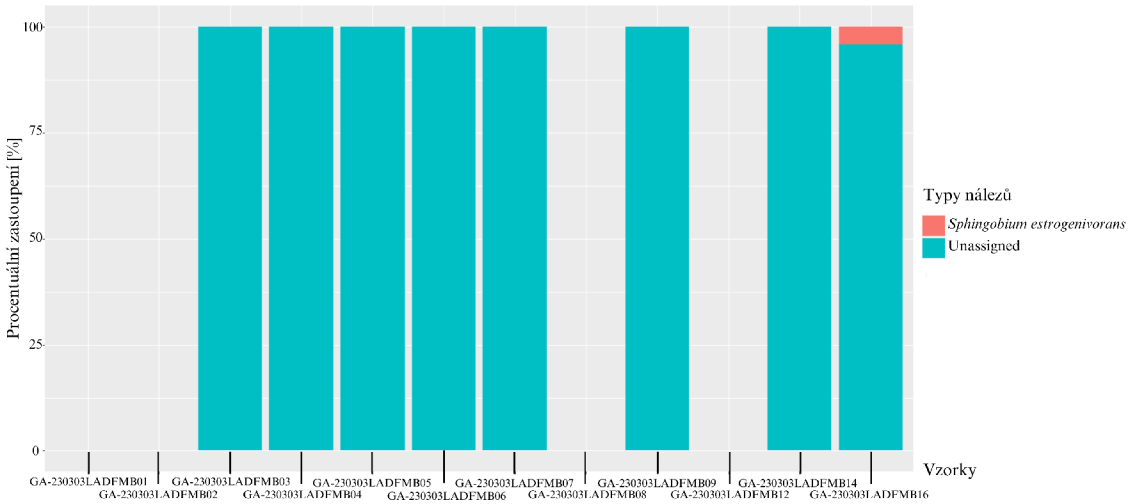
Obr. 4.3: Sloupcové grafy zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Actinobacteria* pro jednotlivé vzorky



Obr. 4.4: Sloupcové grafy zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Bacteroidetes* pro jednotlivé vzorky



Obr. 4.5: Sloupcové grafy zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Firmicutes* pro jednotlivé vzorky



Obr. 4.6: Sloupcové grafy zobrazující procentuální zastoupení [%] nálezů na úrovni druhu pro kmen *Proteobacteria* pro jednotlivé vzorky

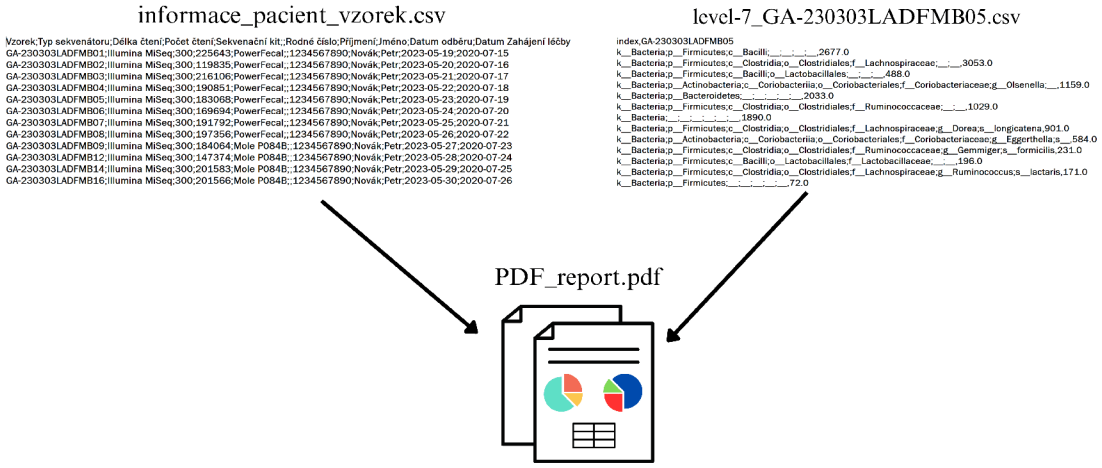
Taktéž byla vypočítána alfa diverzita pomocí Shannon-Wiener indexu, který je pro každý vzorek k vidění v tabulce 4.2. Podle výsledků můžeme říct, že nejrozmanitější byl vzorek GA-230303LADFMB12. U vzorku GA-230303LADFMB02 nebylo však možné diverzitu vypočítat a to z důvodu, že jako jediný nebyl schopen detekovat nálezy na nižší úrovni než je na úrovni kmene. Pokud bychom Shannon-Wiener index vypočítali z výsledků na úrovni kmene, dostali bychom hodnotu 0.25. Z ní můžeme usoudit, že vzorek měl velmi malou diverzitu, o čem se můžeme přesvědčit i z obrázku 4.2, kde vidíme, že přes 90 % vzorku je tvořeno bakteriemi kmene *Firmicutes*.

Tab. 4.2: Shannon-Wiener indexy značící alfa diverzitu pro jednotlivé vzorky z FN Brno

Název vzorku	**Shannon-Wiener index**
GA-230303LADFMB01	1.98
GA-230303LADFMB02	-
GA-230303LADFMB03	1.84
GA-230303LADFMB04	0.89
GA-230303LADFMB05	2.28
GA-230303LADFMB06	2.38
GA-230303LADFMB07	2.29
GA-230303LADFMB08	1.04
GA-230303LADFMB09	1.47
GA-230303LADFMB12	2.39
GA-230303LADFMB14	2.16
GA-230303LADFMB16	0.89

5 Generování reportu pro diagnostické účely ve Fakultní nemocnici Brno

Poslední části této práce bylo vytvořit skript, který automaticky bude generovat reporty sloužící k diagnostickým účelům ve FN Brno. Pro vygenerování jednoho reportu pro jeden vzorek musí do skriptu vstoupit dva CSV soubory. První v sobě obsahuje informace o taxonomické klasifikaci vzorku na úrovni druhu a druhý CSV soubor představuje souhrnnou tabulku s informacemi jak o pacientovi (jméno pacienta, rodné číslo, datum zahájení léčby a podobně), tak i informace týkající se vzorku i samotného průběhu sekvenování (jméno vzorku, délka jednoho čtení či typ sekvenátoru). Celý kód je zabalen do jedné velké funkce, která po spuštění vygeneruje do paměti počítače šest koláčových grafů zobrazující procentuální zastoupení nálezů na určitých úrovních a nakonec i samotný report ve formátu PDF, který se uloží na stejné místo jako zmíněné grafy. Schématické znázornění generování reportu je k vidění na obrázku 5.1.



Obr. 5.1: Schématické znázornění tvorby PDF reportu sloužícího k diagnostickým účelům ve Fakultní nemocnici Brno

První část reportu nese název *Informace o pacientovi*, kde jsou přehledně vypsány všechny informace z CSV souboru se souhrnnou tabulkou. Stejně vypadá i druhá část reportu s názvem *Technické informace*. Za tím následuje tabulka nesoucí hodnoty procentuálního zastoupení kmenů ve vzorku a zároveň i referenční rozsahy, ve kterých by se u zdravého jedince měly tyto hodnoty vyskytovat. Díky tomu je možné jednoduše interpretovat, jestli mikrobiom pacienta je v rovnovážném prospěšném stavu či je různými vnějšími i vnitřními jevy vyveden z rovnováhy. Poslední části úvodní strany je informace o diverzitě mikrobiomu vyjádřena Shannon-Wiener indexem definující alfa diverzitu vzorku.

Na druhé a třetí straně reportu je graficky znázorněna diverzita mikrobiomu pomocí šesti koláčových grafů. První dva v sobě nesou informace o složení na úrovni kmene a na úrovni druhu pro všechny kmeny dohromady. Čtyři další rozšiřují graf s druhovou analýzou - každý jeden zobrazuje druhy pro určitý kmen (opět byly použity čtyři nejčastěji vyskytující se kmeny ve střevním lidském mikrobiomu).

Ukázka jednoho vygenerovaného reportu je k vidění v příloze A.

Závěr

V této bakalářské práci je nejdříve vypracována ve čtyřech podkapitolách literární rešerše shrnující všechny teoretické informace potřebné k pochopení práce. První podkapitola se zabývá anatomií a fyziologií střev, druhá složením střevního mikrobiomu, kde byly popsány i základní pojmy jako mikrobiom či metagenom. Třetí podkapitola se věnuje datům z 16S rRNA, sekvenátorům, přípravou knihovny pro sekvenování Illumina sekvenátorem a taktéž samotným postupem Illumina sekvenování. Poslední čtvrtá podkapitola pojednává o možnostech vyhodnocení střevního mikrobiomu, a to pomocí taxonomických jednotek a vypočítáním diverzity mikrobiomu.

V druhé části práce jsou popsána testovací data, která byla v práci použita a taktéž FASTQ formát, který je pro takový typ dat typický.

Třetí část se věnuje navržení algoritmu sloužícímu k bioinformatickému zpracování dat. Algoritmus byl navržen a otestován na testovacích datech a poté byl použit na zpracování dat z Fakultní nemocnice Brno. Algoritmus obsahuje části, jako je kontrola kvality a kontaminace, filtrace, klasifikace sekvencí a nakonec vyhodnocení pomocí grafických reprezentací taxonomických jednotek a vypočítáním alfa diverzity vzorků Shannon-Wienerovým indexem.

Navržený algoritmus byl aplikován na dvanáct vzorků z Fakultní nemocnice Brno poskytnutých z Interní hematologické a onkologické kliniky, z Centra molekulární biologie a genové terapie, přičemž vzorky byly získány od hematologických pacientů. Bylo zjištěno, že nejvíce zastoupenými kmeny napříč vzorky jsou kmeny *Firmicutes* a *Bacteroidetes*. U běžného střevního mikrobiomu by 90 % bakterií střevního mikrobiomu mělo být právě z kmenů *Firmicutes* a *Bacteroidetes*, z čehož můžeme vyvodit, že žádný z pacientů neměl markantně poškozené složení střevního mikrobiomu vlivem léčby. Nejméně zastoupeným kmenem byl kmen *Euryarchaeota*, který byl detekován pouze u čtyř vzorků. Tento kmen však jako jediný ze zbylých detekovaných kmenů nepatří do říše *Bacteria*, nýbrž do říše *Archaea*.

Pro všechny vzorky byla taktéž pomocí Shannon-Wienerova indexu vypočítána alfa diverzita. Bylo zjištěno, že ne všechny vzorky mají diverzitu v referenčním rozsahu od 1.5 do 3.5. U vzorku číslo 2 nebylo možné získat výsledky taxonomie na úrovni druhu, a proto jako u jediného vzorku nebylo možné vypočítat druhovou diverzitu. Nejrozmanitější mikrobiom byl detekován u vzorku GA-230303LADFMB12, jehož hodnota Shannon-Wienerova indexu byla 2.39.

V poslední části práce je popsán proces tvorby reportů, které mohou sloužit k diagnostickým účelům ve FN Brno, kde právě probíhá pilotní studie analýzy střevního mikrobiomu u hematologických pacientů a bude díky nim možné sledovat jeho změny ve složení v závislosti na léčbě.

Literatura

[1] Trávicí soustava, 2016. URL: http://fblt.cz/skripta/ix-travici-soustava/.

[2] Ivan Dylevský and Petr Ježek. Trávicí systém. URL: https://vos.palestra.cz/skripta/anatomie/10.htm.

[3] Richard Rokyta. *Fyziologie*. Galén, Praha, třetí, přepracované vydání (první vydání v nakladatelství galén) edition, 2016.

[4] Hand drawn human organs large intestine small intestine vector hd png images. URL: https://pngtree.com/freepng/hand-drawn-human-organs-large-intestine-small-intestine_4554878.html.

[5] Peter H. Abrahams. *Jak pracuje lidské tělo*. Praha, 2014.

[6] Jitka Švíglerová and Jana Slavíková. *Fyziologie gastrointestinálního traktu*. Karolinum, Praha, 2., upr. vyd edition, 2013.

[7] Ivan Dylevský. *Základy funkční anatomie*. Poznání, Olomouc, 2011.

[8] B. Brett Finlay and Jessica M. Finlay. *Mikrobiom lidského těla*. Stanislav Juhaňák - Triton, Praha, 2020.

[9] Vladimír Zbořil. *Mikroflóra trávicího traktu*. Grada, Praha, 2005.

[10] Roman Stebel. Transplantace střevní mikrobioty – historie, současnost a budoucnost. *Gastroenterologie a hepatologie*, 2020(1):8, 2019.

[11] Přemysl Frič. Střevní mikroflóra, gastrointestinální ekosystém a probiotika. *Medicína pro praxi*, 7(11):408–413, 2010.

[12] Jan Lata and Jana Juránková. Střevní mikroflóra, slizniční bariéra a probiotika u některých interních chorob. *Medicína pro praxi*, 9(3):106–112, 2012.

[13] The gut microbiome and its impact on the brain. URL: https://med.libretexts.org/Bookshelves/Pharmacology_and_Neuroscience/Book%3A_Neuroscience_(Ju)/04%3A_Emergent_Topics_in_Neuroscience/4.01%3A_The_Gut_Microbiome_and_its_Impact_on_the_Brain.

[14] Ethan T. Hillman, Hang Lu, Tianming Yao, and Cindy H. Nakatsu. Microbial ecology along the gastrointestinal tract. *Microbes and environments*, 32(4):300–313, 2017. doi:10.1264/jsme2.ME17017.

[15] Cindy G. Boer, Djawad Radjabzadeh, Carolina Medina-Gomez, Sanzhima Garmaeva, Dieuwke Schiphof, Pascal Arp, Thomas Koet, Alexander Kurilshikov, Jingyuan Fu, M. Arfan Ikram, Sita Bierma-Zeinstra, André G. Uitterlinden, Robert Kraaij, Alexandra Zhernakova, and Joyce B. J. van Meurs. Intestinal microbiome composition and its relation to joint pain and inflammation. *Nature Communications*, 10(1), 2019. doi:10.1038/s41467-019-12873-4.

[16] Alexandre Almeida, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 2019-04-25. doi:10.1038/s41586-019-0965-1.

[17] Jana Fabryová. Candida albicans (kandidóza) a kvasinkové infekce, 2014. URL: https://www.doktor-zdravi.cz/candida-albicans-kandidoza-a-kvasinkove-infekce/.

[18] Kjersti Aagaard, Jun Ma, Kathleen M. Antony, Radhika Ganu, Joseph Petrosino, and James Versalovic. The placenta harbors a unique microbiome. *Science Translational Medicine*, 6(237), 2014-05-21. URL: https://www.science.org/doi/10.1126/scitranslmed.3008599, doi:10.1126/scitranslmed.3008599.

[19] Catherine A. Lozupone, Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, and Rob Knight. Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230, 2012. doi:10.1038/nature11550.

[20] Susannah G Tringe and Philip Hugenholtz. A renaissance for the pioneering 16s rrna gene. *Current Opinion in Microbiology*, 11(5):442–446, 2008. doi:10.1016/j.mib.2008.09.011.

[21] Jethro S. Johnson, Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, Blake M. Hanson, Hanako O. Agresta, Mark Gerstein, Erica Sodergren, and George M. Weinstock. Evaluation of 16s rrna gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1):442–446, 2019. doi:10.1038/s41467-019-13036-1.

[22] Kazumasa Fukuda, Midori Ogawa, Hatsumi Taniguchi, and Mitsumasa Saito. Molecular approaches to studying microbial communities. *Journal of UOEH*, 38(3):223–232, 2016. doi:10.7888/juoeh.38.223.

[23] James M. Heather and Benjamin Chain. The sequence of sequencers. *Genomics*, 107(1):1–8, 2016. doi:10.1016/j.ygeno.2015.11.003.

[24] Illumina sequencing platforms, 2023. URL: https://emea.illumina.com/systems/sequencing-platforms.html.

[25] Principle and workflow of illumina next-generation sequencing, 2018. URL: https://www.cd-genomics.com/blog/principle-and-workflow-of-illumina-next-generation-sequencing/.

[26] Sai Manasa Jandhyala. Role of the normal gut microbiota. *World Journal of Gastroenterology*, 21(29), 2015. doi:10.3748/wjg.v21.i29.8787.

[27] 16s metagenomic sequencing library preparation, 2013. URL: https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html.

[28] Moira Marizzoni, Thomas Gurry, Stefania Provasi, Gilbert Greub, Nicola Lopizzo, Federica Ribaldi, Cristina Festari, Monica Mazzelli, Elisa Mombelli, Marco Salvatore, Peppino Mirabelli, Monica Franzese, Andrea Soricelli, Giovanni B. Frisoni, and Annamaria Cattaneo. Comparison of bioinformatics pipelines and operating systems for the analyses of 16s rrna gene amplicon sequences in human fecal samples. *Frontiers in Microbiology*, 11, 2020-6-17. doi:10.3389/fmicb.2020.01262.

[29] Taishan Hu, Nilesh Chitnis, and Dimitri Monos. Next-generation sequencing technologies. *Human Immunology*, 82(11):801–811, 2021. doi:10.1016/j.humimm.2021.02.012.

[30] Amplicon-based next-generation sequencing vs. metagenomic shotgun sequencing. *CD Genomics*.

[31] Ravi Ranjan, Asha Rani, and Ahmed Metwally. Analysis of the microbiome. *Biochemical and Biophysical Research Communications*, 469(4):967–977, 2016. doi:10.1016/j.bbrc.2015.12.083.

[32] Taxonomic categories, 2020. URL: https://www.toppr.com/guides/biology/the-living-world/taxonomic-categories/.

[33] Michael T. Madigan, Kelly S. Bender, and Daniel H. Buckley. *Brock Biology of Microorganisms*. British Library Cataloguing-in-Publication Data, New York, fifteenth edition edition, 2019.

[34] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, 36(10):996–1004, 2018. doi:10.1038/nbt.4229.

[35] Ramon Rosselló-Móra and Rudolf Amann. Past and future species definitions for bacteria and archaea. *Systematic and Applied Microbiology*, 38(4):209–216, 2015. doi:10.1016/j.syapm.2015.02.001.

[36] Kane E. Deering, Amanda Devine, Therese A. O'Sullivan, Johnny Lo, Mary C. Boyce, and Claus T. Christophersen. Characterizing the composition of the pediatric gut microbiome. *Nutrients*, 12(1), 2020. doi:10.3390/nu12010016.

[37] R. H. Whittaker. Evolution and measurement of species diversity. *TAXON*, 21(2-3):213–251, 1972. doi:10.2307/1218190.

[38] Jiří Holčík and Martin Komenda. *Matematická biologie: e-learningová učebnice*. Brno: Masarykova univerzita, 1 edition, 2015.

[39] Jan Divíšek and Martin Culek. *Biogeografie*. Masarykova univerzita, Brno, 2., aktual. vydání edition, 2010.

[40] Nora Bynum. Alpha, beta, and gamma diversity. *Biodiversity*. URL: https://bio.libretexts.org/Bookshelves/Ecology/Biodiversity_(Bynum)/7:_Alpha_Beta_and_Gamma_Diversity.

[41] Xu-Bo Qian, Tong Chen, Yi-Ping Xu, Lei Chen, Fu-Xiang Sun, Mei-Ping Lu, and Yong-Xin Liu. A guide to human microbiome research. *Chinese Medical Journal*, 133(15):1844–1855, 2020. doi:10.1097/CM9.0000000000000871.

[42] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Research*, 38(6):1767–1771, 2010-04-01. doi:10.1093/nar/gkp1137.

[43] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, and 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.

[44] Adrian López-García, Carolina Pineda-Quiroga, Raquel Atxaerandio, Adrian Pérez, Itziar Hernández, Aser García-Rodríguez, and Oscar González-Recio. Comparison of mothur and qiime for the analysis of rumen microbiota composition based on 16s rrna amplicon sequences. *Frontiers in Microbiology*, 9, 2018-12-13. doi:10.3389/fmicb.2018.03010.

[45] Robert C. Edgar. Uchime2: improved chimera prediction for amplicon sequencing. page 37. doi:10.1101/074252.

[46] David Langenberger, Mario Fasold, Gero Doose, and Adele Feuerstein. Why does the per base sequence quality decrease over the read in illumina?, 2017. URL: https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina.

[47] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC. Babraham Institute, January 2012. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[48] P Ewels, M Magnusson, S Lundin, and M Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 10 2016. doi:10.1093/bioinformatics/btw354.

[49] Steven W. Wingett and Simon Andrews. Fastq screen. *F1000Research*, 7, 2018. doi:10.12688/f1000research.15931.2.

[50] Elliot M. Meyerowitz. Arabidopsis thaliana. *Annual Review of Genetics*, 21(1):93–111, 1987. doi:10.1146/annurev.ge.21.120187.000521.

[51] James P. Nataro and James B. Kaper. Diarrheagenic escherichia coli. *Clinical Microbiology Reviews*, 11(1):142–201, 1998. doi:10.1128/CMR.11.1.142.

[52] J Gregory Caporaso, Christian L Lauber, and Elizabeth K Costello. Moving pictures of the human microbiome. *Genome Biology*, 12(5), 2011. doi:10.1186/gb-2011-12-5-r50.

[53] Amnon Amir, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, Luke R. Thompson, Embriette R. Hyde, Antonio Gonzalez, Rob Knight, and Jack A. Gilbert. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2):e00191–16, 2017-04-21. doi:10.1128/mSystems.00191-16.

[54] Daniel McDonald, Morgan N Price, Julia Goodrich, Eric P Nawrocki, Todd Z DeSantis, Alexander Probst, Gary L Andersen, Rob Knight, and Philip Hugenholtz. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. In *The ISME Journal*, volume 6, pages 610–618, 2012. doi:10.1038/ismej.2011.139.

[55] Data resources. URL: https://docs.qiime2.org/2023.2/data-resources/#taxonomy-classifiers-for-use-with-q2-feature-classifier.

[56] Irene LG Newton and Guus Roeselers. The effect of training set on the classification of honey bee gut microbiota using the naïve bayesian classifier. *BMC Microbiology*, 12(1):1–9, 2012. doi:10.1186/1471-2180-12-221.

[57] Michal Maňas. Termín - komenzalismus. URL: https://www.biolib.cz/cz/glossaryterm/id2654/.

[58] Pavel Polák, Petr Husa, and Michaela Freibergerová. Kolitida způsobená clostridium difficile, její příčiny a aktuální možnosti léčby v širších souvislostech. *Interní medicína pro praxi*, 2014(6):241–243, 2014.

[59] Rita Rain. Shannon diversity index calculator, 2022. URL: https://www.omnicalculator.com/ecology/shannon-index#shannon-diversity-indexs-range-of-values.

[60] Suspense Averti Ifo, Jean-Marie Moutsambote, and Félix Koubouana. Tree species diversity, richness, and similarity in intact and degraded forest in the tropical rainforest of the congo basin. *International Journal of Forestry Research*, 2016:1–12, 2016. doi:10.1155/2016/7593681.

[61] Jan Chlumský. Sekvenování dna. URL: https://botanika.prf.jcu.cz/laboratory/sekvenovani.html.

[62] Heng Li and Inanc Birol. Minimap2. *Bioinformatics*, 34(18):3094–3100, 2018-09-15. doi:10.1093/bioinformatics/bty191.

[63] Heng Li, John Marshall, and Petr Danecek. Samtools documentation, 2023. URL: http://www.htslib.org/doc/samtools.html.

[64] Valerie A. Schneider, Tina Graves-Lindsay, and Kerstin Howe and. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017-05-01. doi:10.1101/gr.213611.116.

Seznam symbolů a zkratek

DNA	deoxyribonukleová kyselina
ENA	Evropský archiv nukleotidů
FN	Fakultní nemocnice
NGS	sekvenátory druhé generace
PCR	polymerázová řetězová reakce
RNA	ribonukleová kyselina
rRNA	ribozomální ribonukleová kyselina
SBS	sekvenování pomocí syntézy

A Vygenerovaný report sloužící k diagnostickým účelům do Fakultní nemocnice Brno

V příloze je k vidění vygenerovaný report pro vzorek GA-230303LADFMB05. Informace o pacientovi (jméno, příjemní, rodné číslo, datum odebrání vzorku a datum zahájení léčby) jsou smyšlené, technické informace jsou však již pravdivé.

Report analýzy lidského střevního mikrobiomu z 16S rRNA

Informace o pacientovi

Příjmení: Novák
Jméno: Petr
Rodné číslo: 1234567890
Datum odebrání vzorku: 2023-05-23
Datum zahájení léčby: 2020-07-19

Technické informace

ID vzorku: GA-230303LADFMB05
Sekvenátor: Illumina MiSeq
Délka jednoho čtení: 300 bp
Počet čtení: 183068
Typ sekvenačního kitu: PowerFecal

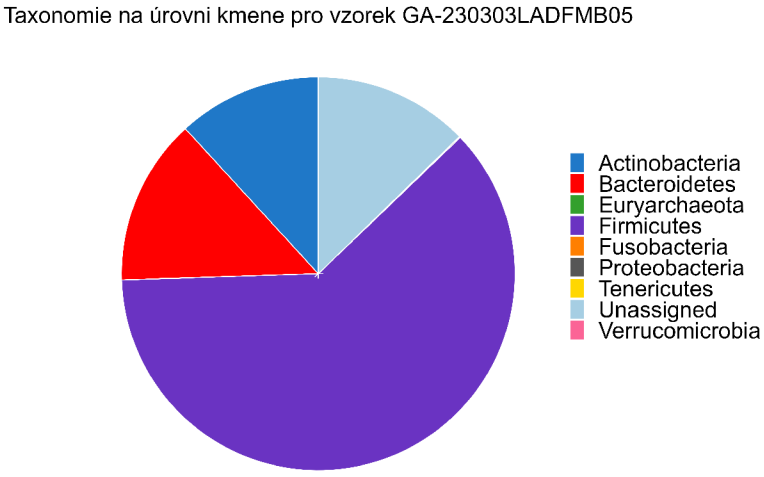
Zastoupení kmenů bakterií

Kmen	Procentualní zastoupení ve vzorku [%]	Referenční rozsah [%]
Actinobacteria	11.81	1.0-5.0
Bacteroidetes	13.69	30.0-60.0
Firmicutes	61.68	30.0-60.0
Fusobacteria	0.00	0.0-1.0
Proteobacteria	0.09	1.5-5.0
Verrucomicrobia	0.00	1.5-5.0
Tenericutes	0.00	0.02-0.4
Euryarchaeota	0.00	0.01-0.06
Unassigned	12.73	

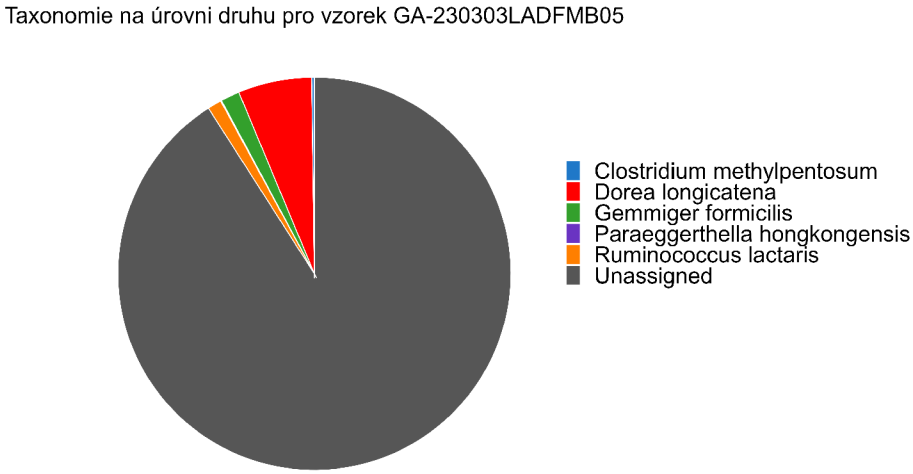
Diversita mikrobiomu

Shannon-Wiener index: 2.28 (referenční rozsah: 1.5-3.5)

Složení na úrovni kmene

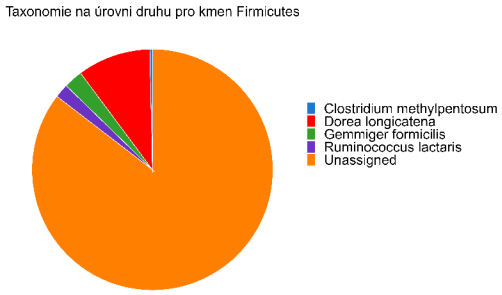


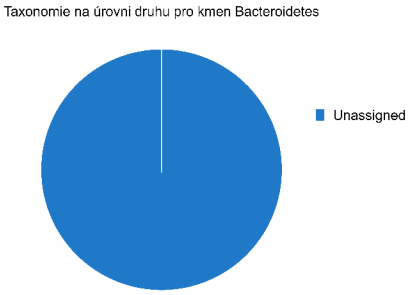
Složení na úrovni druhu



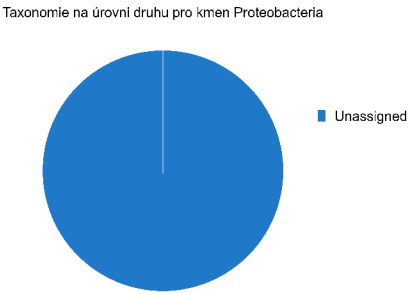
Složení na úrovni druhu pro jednotlivé kmeny

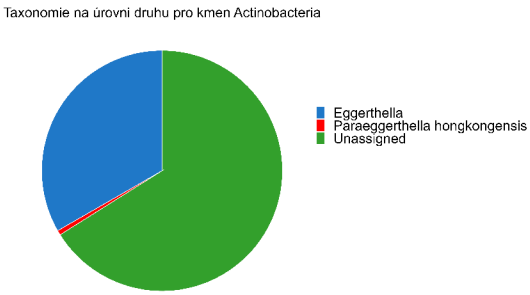
Kmen Firmicutes a kmen Bacteroidetes





Kmen Proteobacteria a kmen Actinobacteria





B Obsah elektronické přílohy

V elektronické příloze se nachází dvě složky obsahující skripty vytvořené pro potřeby bakalářské práce. Ve složce jsou uloženy dvě podsložky - v první podsložce nesoucí název *Předzpracování dat a analýza* jsou skripty naprogramované pomocí skriptovacího jazyku Bash, které byly spouštěné v prostředí Linuxu. Jedná se o soubory:

- *run_Fastqc.sh*
- *decontamination.sh*
- *Taxonomy_TestData.sh*
- *Taxonomy_FNData.sh*

První skript *run_Fastqc.sh* u dat vytvoří pomocí nástroje FastQC reporty hodnotící kvalitu sekvencí a zároveň i jeden souhrnný soubor vytvořený nástrojem MultiQC obsahující všechny reporty v jednom. Druhý skript *decontamination.sh* slouží k odstranění kontaminovaných částí sekvencí. Třetí a čtvrtý skript *Taxonomy_01data.sh* a *Taxonomy_FN.sh* slouží k samotné analýze dat.

V druhé podsložce s názvem *Vizualizace dat, generování reportu* jsou uloženy skripty naprogramované pomocí jazyku R, jeden skript naprogramovaný pomocí R Markdown sloužící k vykreslování souborů do PDF, jeden soubor ve formátu TEX definující záhlaví reportu a logo FN Brno ve formátu JPG. Výpis souborů je vidění zde:

- *BP_TestData_sloupce_kmeny.R*
- *BP_TestData_sloupce_druhy.R*
- *BP_FNData_sloupce_kmeny.R*
- *BP_FNData_sloupce_druhy.R*
- *BP_diverzita.R*
- *BP_generovani_reportu.R*
- *PDF_report.Rmd*
- *zahlavi.tex*
- *logo.jpg*

Skripty *BP_TestData_sloupce_kmeny.R*, *BP_TestData_sloupce_druhy.R*, *BP_FNData_sloupce_kmeny.R* a *BP_FNData_sloupce_druhy.R* slouží k vykreslování grafických reprezentací jak pro testovací data, tak pro data z FN Brno. Skript *BP_diverzita.R* vyhodnocuje diverzitu vzorků pomocí Shannon-Wiener indexu. Poslední dva skripty *BP_generovani_reportu.R* a *PDF_report.Rmd* slouží k vykreslování reportů - po spuštění prvního skriptu se zavolá skript druhý a vytvoří tak report ve formátu PDF.