VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2021 Kristýna Heřmánková

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

DETEKCE A FILTRACE CHIMÉR V AMPLIKONOVÉ SEKVENACI
DETECTION OF CHIMERAS IN AMPLICON SEQUENCING

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Kristýna Heřmánková**
AUTHOR

VEDOUCÍ PRÁCE — **Mgr. Ing. Karel Sedlář, Ph.D.**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Kristýna Heřmánková ***ID:*** 211200

Ročník: 3 ***Akademický rok:*** 2020/21

NÁZEV TÉMATU:

Detekce a filtrace chimér v amplikonové sekvenaci

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši o výpočetním zpracování sekvenačních dat pro studie mikrobiomů, především o postupech pro detekci operačních taxonomických jednotek a jejich kvantifikaci. 2) Prostudujte princip laboratorních postupů používaných při získávání mikrobiomových dat. Zaměřte se na chyby, které mohou v datech vznikat, především na tzv. chimerické sekvence. 3) Navrhněte metodu pro detekci a filtraci chimerických sekvencí v mikrobiomových datech a její dílčí části otestujte. 4) Implementujte navrženou metodu ve zvoleném programovacím jazyce. 5) Proveďte ladění metody s využitím jak veřejně dostupných dat, tak unikátního datasetu poskytnutého spolupracujícím pracovištěm VÚVEL. 6) Srovnejte navrženou metodu s metodami existujícími, zhodnoťte rychlost i přesnost různých přístupů a výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] ALLALI, Imane, Jason W. ARNOLD, Jeffrey ROACH, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. BMC Microbiology. 2017, 17(1).

[2] EDGAR, Robert C., Brian J. HAAS, Jose C. CLEMENTE, Christopher QUINCE a Rob KNIGHT. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011, 27(16).

Termín zadání: 8.2.2021 ***Termín odevzdání:*** 28.5.2021

Vedoucí práce: Mgr. Ing. Karel Sedlář, Ph.D.

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

ABSTRAKT

Chimérické sekvence jsou častým artefaktem vyskytujícím se v sekvenačních datech po amplifikaci vzorků polymerázovou řetězovou reakcí. Výskyt těchto artefaktů může výrazně znehodnotit výsledky prováděné analýzy. Proto je detekce a následná filtrace chimérických sekvencí nezbytným krokem ve výpočetním zpracování sekvenačních dat. Součástí této práce je vysvětlení mechanismu vzniku těchto artefaktů a také možnosti omezení jejich výskytu. Cílem této práce je realizace algoritmu pro detekci a filtraci chimér v jazyce R a následné testování úspěšnosti algoritmu na vlastních datech poskytnutých Výzkumným ústavem veterinárního lekařství v Brně.

KLÍČOVÁ SLOVA

Operační taxonomická jednotka, metagenomika, polymerázová řetězová reakce, next-generation sekvenování

ABSTRACT

Chimeric sequences are the most common artifacts that can occur in sequencing data after the sample amplification using the polymerase chain reaction. The presence of these artifacts can negatively affect results of the analysis. Therefore, the detection and subsequent filtration of chimeric sequences is an important step in the computational processing of sequencing data. This work deals with the principle of chimera formation and possibility of reducing their occurrence. The aim of this work is to implement an algorithm for chimeras detection in R language and testing its accuracy on data provided by the Veterinary Research Institute in Brno.

KEYWORDS

Operational taxonomic unit, metagenomics, polymerase chain reaction, next-generation sequencing

HEŘMÁNKOVÁ, Kristýna. *Detekce a filtrace chimér v amplikonové sekvenaci*. Brno, 2021, 55 s. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Mgr. Ing. Karel Sedlář, PhD.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Detekce a filtrace chimér v amplikonové sekvenaci“ jsem vypracovala samostatně pod vedením vedoucího bakalářské 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 bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské 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.

V Brně dne 28. května 2021 .....

podpis autorky

PODĚKOVÁNÍ

Ráda bych poděkovala Mgr. Ing. Karlu Sedlářovi, Ph.D., vedoucímu bakalářské práce především za trpělivost, podnětné návrhy k práci a neustálou motivaci v průběhu celého semestru. Dále bych chtěla poděkovat své mamince za korekturu práce.

V Brně dne 28. května 2021 .....

podpis autorky

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Sekvenování genetické informace za účelem nahlédnutí na kompozici mikrobiální komunity prostředí se od vzniku Next-generation sekvenačních technologií (NGS) stalo nedílnou součástí mnoha oborů. Paralelní sekvenace, která je specifickou vlastností NGS technologií, zajišťuje nejen cenovou dostupnost sekvenování, ale šetří i čas strávený nad sekvenováním velkého množství dat.

Z oborů, které provádějí právě mikrobiální analýzy za účelem odhalení diverzity mikroorganismů, je to například medicína, která využívá sledování složení mikrobioty zdravých a nemocných pacientů pro odhalení možných souvislostí mezi onemocněním a změnou kompozice mikrobioty. Například v onkologii byla zaznamenána úzká souvislost mezi přítomností určitých mikroorganismů a jejich nemalým učinkem na karcinogenezi. Na druhou stranu je snaha určit kompozici normální mikrobioty, jejíž složení však není závislé jen na stavu pacienta (zdravý/nemocný), ale mnoha různých aspektech, které víceméně znemožňují její určení. [1]

Limitací mikrobiálních analýz zůstávají artefakty vzniklé především při amplifikaci vzorků. Výskyt těchto artefaktů v datech může vést ke špatné interpretaci dosažených výsledků, způsobené především výskytem falešných mikrobiálních komunit. Proto cílem mnoha bioinformatických algortimů zůstává odstranění či alespoň minimalizace těchto artefaktů, kterými můžou být například chimérické sekvence.

Cílem této práce je teoretické objasnění vzniku tzv. chimérických sekvencích, které tvoří významnou část artefaktů v mikrobiálních datech. Dále jsou v teoretické části uvedeny sekvenanční technologie, kterými daná mikrobiální data můžou být sekvenována.

V praktické části je uvedený popis navrženého algoritmu realizovaného v jazyce R. Ten je následně otestován na vlastních datech poskytnutých Výzkumným ústave veterinárního lékařství v Brně a jeho výsledky jsou srovnány s dostupnými algortimy pro detekci a filtraci chimér.

1 Metagenomika

Metagenomika se obdobně jako genomika zabývá studiem genomu organismu. Rozdílem však mezi těmito obory je předpona „meta“, která poukazuje na studium něčeho většího, než je samotný organismus. Z toho důvodu lze metagenomiku definovat jako studium genomů mikroorganismů vyskytujících se v určitém prostředí [2]. Jelikož metody založené na kultivaci bakterií nejsou schopné odhalit reálné zastoupení jednotlivých druhů ve vzorku, či některé bakterie jednoduše nelze kultivovat, metagenomika využívá metody, které kultivaci nevyžadují, a umožňuje tak objevovat nejen nové mikroorganismy, ale i význam jednotlivých mikroorganismů v určité komunitě [3].

S metagenomikou souvisí i další pojmy, které je vhodné uvést. Prvním z nich je mikrobiota, kterou lze popsat jako soubor všech mikroorganismů žijících v určitém prostředí [4]. Lze tak získat mikrobiotu z vody, půdy, ale i ze slin, kůže či střeva. Pojem lidská mikrobiota pak zahrnuje veškeré mikroorganismy žijící nejen v lidském těle, ale i na něm [5]. Pojmem metagenom pak můžeme chápat veškerou genetickou informaci mikrobioty. Nejčastěji zmiňovaným pojmem je však mikrobiom, který může být v různých případech různě interpretován. Podle Marchesi et al. [4], mikrobiom odpovídá nejen mikroorganismům a jejich genům, avšak i okolním podmínkám, ve kterých se mikroorganismy nacházejí.

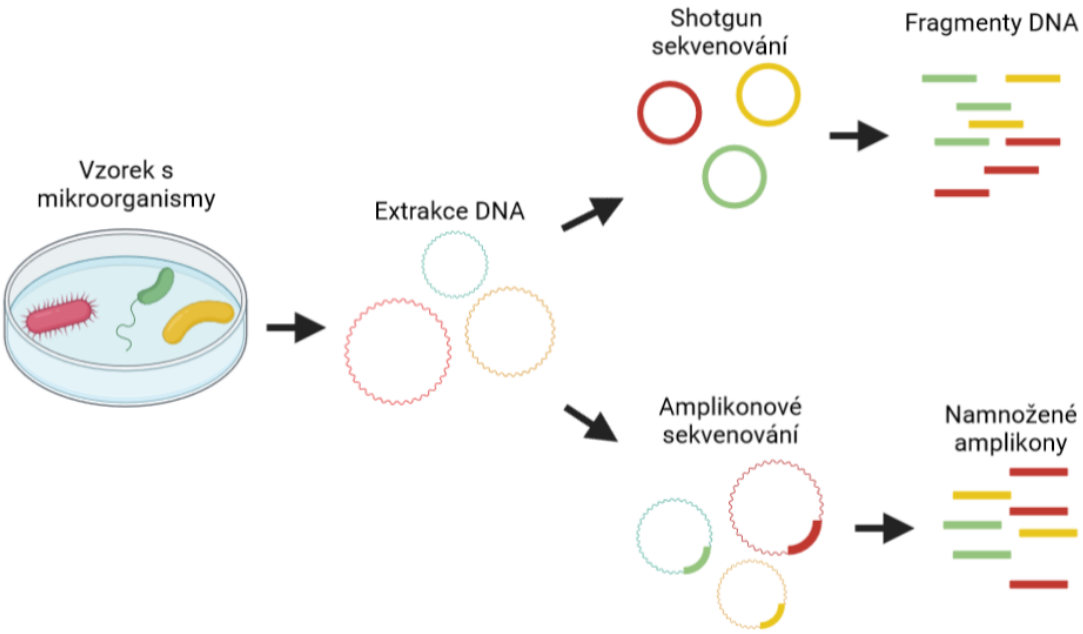
Pro komplexnost studie se k metagenomické analýze může přidat metatranskriptomika, která se zabývá funkčním profilem komunity, nebo metabonomika [6], studující metabolity produkované mikrobiotou. Kombinace těchto studií tak může vést k důkladnému pochopení mikrobiomů [7].

Jak již bylo zmíněno, metagenomika se zabývá analýzou veškerých genomů mikrobiální komunity. K přípravě těchto dat pro následnou sekvenaci je využívána metoda shotgun, která genetický materiál mechanicky rozruší na náhodně dlouhé úseky. Tyto úseky jsou poté sekvenovány zvolenou sekvenační technologií. Tento způsob přípravy a sekvenace je nazýván shotgun sekvenování a je tedy především využíván pro potřeby metagenomické analýzy. Metagenomickou analýzou lze tak získat přehled nejen o funkčním profilu komunity, ale i o tom, které mikroorganismy jsou ve vzorku přítomny.

Pokud je cílem analýzy pouze klasifikace mikroorganismů v rámci specifické komunity, je využíváno amplikonové sekvenování. Příprava dat pro sekvenování spočívá ve výběru pouze specifického genu z genetického materiálu, dle kterého lze stanovit diverzitu mikrobiální komunity.

Na obrázku č. 1.1 jsou pak znázorněné průběhy přípravy dat pro následující sekvenaci. Je zde zřejmé, že pro amplikonové sekvenování je potřeba obstarat pouze část genomu, resp. určitý gen, zatímco shotgun dále pracuje s veškerou genetickou

informací mikroorganismů ve vzorku.



Obr. 1.1: Princip amplikonového, resp. shotgun sekvenování.

1.1 Amplikonové sekvenování

Cílené amplikonové sekvenování se zaměřuje na sekvenaci specifických genů či jejich regionů. Analýza založená na této metodě umožňuje porovnávat různé varianty těchto specifických genů, kdy výsledkem pak může být např. sledování jejich vývoje či objevení mutací. Jak již bylo zmíněno, velmi často se tento přístup využívá k analýze mikrobiálních komunit, kdy jsou specifické geny studovány pro určení taxonomického profilu komunity.

Specificky navrženými primery, které vymezují sekvenci specifického genu, jsou následnou polymerázovou řetězovou reakcí (dále PCR) namnoženy početné kopie těchto genů. Výstupem každého sekvenačního procesu jsou úseky zvané čtení. Délka těchto čtení pak závisí na zvolené sekvenační technologii. Pro sekvenaci takto obsáhlých dat jsou dnes využívány Next-generation sekvenátory (NGS), jejichž objev umožnil paralelní sekvenaci velkého množství fragmentů DNA.

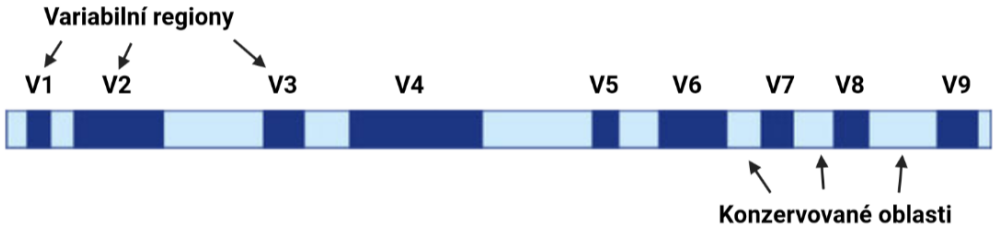
Specifický gen, který je využíván pro potřeby analýzy mikrobiálních komunit, musí splňovat několik kritérií. Tento gen by měl být přítomný ve všech mikroorganismech studovaného vzorku. Dále by měl obsahovat úseky, které jsou specifické

pro druh nebo alespoň rod a zároveň oblasti, které jsou ve všech mikroorganismech konstantní [8]. Pro doménu bakterií a archeí byl vybrán gen 16S rRNA.

16S rRNA

16S rRNA je část malé ribozomální podjednotky 30S prokaryotických organismů. Gen této části ribozomu je využíván k taxonomické analýze, jejímž výsledkem je zjištění evoluční vzdálenosti a příbuznosti organismů [9]. U eukaryotických organismů je využíván gen, který lze považovat za analogii ke genu 16S rRNA, kódující část ribozomu 18S rRNA. Naopak říši hub definuje region oddělující malou a velkou podjednotku ribozomální RNA a je znám pod zkratkou ITS (Internal transcribed spacer) [10].

Sekvence genu 16S rRNA je dlouhá zhruba 1550 bp a obsahuje jak variabilní, tak evolučně konzervované oblasti [9]. Rozložení těchto regionů je znázorněno na obrázku č. 1.2.



Obr. 1.2: Rozložení genu 16S rRNA [11]

Konzervované oblasti jsou využívány jako předlohy primerů a variabilní oblasti, V1-V9, jsou pak specifické pro jednotlivé druhy mikroorganismů, a tudíž slouží k jejich identifikaci. Limitací dostupných technologií je nemožnost sekvenace celé délky tohoto genu, který by poskytl nejpřesnější interpretaci mikrobiální diverzity. Pro analýzu se tudíž vybírá jeden či dva variabilní regiony, ze kterých i tak lze získat cenné informace.

1.2 Shotgun sekvenování

Tato metoda přípravy dat pro sekvenaci se využívá v případech, kdy je potřeba sekvenovat dlouhé geny, jejichž délka přesahuje možnosti sekvenačních technologií, nebo při celogenomové sekvenaci jako v případě metagenomických analýz. Po extrakci z buňky je DNA mechanicky rozrušena na malé fragmenty, které jsou poté sekvenovány zvolenou technologií. Následuje výpočetní zpracování výstupů sekvenátoru, tedy jednotlivých čtení, které je založeno na mapování čtení k referenční sekvenci či

de novo skládání, kdy referenční sekvence není k dispozici. Cílem skládání je uspořádat jednotlivé fragmenty tak, aby se překrývaly a následně vytvořit konsenzuální sekvenci. Kvalita sestavení konsenzuální sekvence je pak vyjádřena hodnotou pokrytí, která se počítá zvlášť pro každý nukleotid sestavené sekvence. Pokrytí je pak rovno podílu počtu nasekvenovaných bází, které byly namapovány k dané pozici, a délce sestavovaného genu či genomu.

Čím větší počet čtení je obstarán, tím je zde větší počet překrývajících se úseků, a tedy i větší možnost vytvořit přesnější konsensuální sekvenci [12].

1.3 Generace sekvenování

Cílem této podkapitoly je shrnout vývoj sekvenačních technologií a jejich základní principy. Přelomovým obdobím byla 90. léta, resp. začátek 21. století, kdy mnoho výzkumných pracovišť po celém světě pracovalo na projektu, jehož cílem bylo osekvenovat celý lidský genom o délce přibližně 3 miliardy bází. HGP (z anglického "Human Genome Project") [13] se zasloužil nejen o vývoj NGS technologií, ale i vývoj mnoha vědních oborů spojených s tímto projektem.

První generace

Sedmdesátá léta minulého století byla významným milníkem pro mnoho oborů zároveň. Nejen Sangerovo sekvenování syntézou [14], ale i Maxam-Gilbertova metoda chemického štěpení [15] umožnila vědcům rozklíčovat primární strukturu DNA - pořadí nukleotidů. Sangerova metoda, která s sebou přinesla základní principy následujícím generacím, byla častější volbou především díky její jednoduchosti a přesnosti [16]. Tato metoda je založená na vytvoření vhodného primeru, který je společně s volnými nukleotidy (dNTP), radioaktivně obarvenými dideoxy-nukleotidy (ddNTP), které postrádají volný konec 3', a DNA polymerázou obsažen ve směsi se sekvenovanou DNA [16]. Výsledné, různě dlouhé a obarvené sekvence byly vyhodnoceny gelovou elektroforézou. I dnes může Sangerovo sekvenování najít uplatnění, ale už není využíváno radioaktivní značení či gelová elektroforéza. Barví se pomocí fluorescenčních barviv a je používána kapilární elektroforéza, která zcela zautomatizovala tuto metodu [17].

Druhá generace

Generace taktéž známá pod názvem Next-generation sekvenování (NGS) se od té první liší především množstvím dat, které je možno sekvenovat najednou. Narozdíl od Sangerovy metody jsou platformy NGS schopny v rámci jednoho běhu obdržet

miliardy čtení [18]. Objev NGS tak udělal sekvenování cenově dostupným a časově méně náročným.

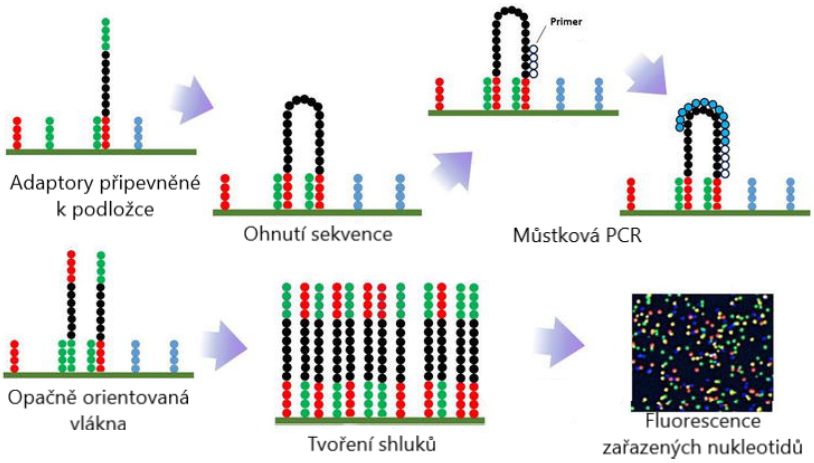
První dostupnou NGS technologii v roce 2005 uvedla společnost Roche [19]. Sekvenační technologie je známá pod názvem **454 pyrosekvenování**, které je podobně jako Sangerovo sekvenování založeno na sekvenaci syntézou. Pro namnožení vzorků se využívá emulzní PCR, kdy základem je emulze z vodného roztoku a oleje. Kapky vody pak ideálně obsahují jednotlivé templáty DNA s komponenty potřebnými pro amplifikaci [20]. Samotné sekvenování je prováděno postupným přidáváním volných dNTP, které se zařazují do nově se tvořícího řetězce. Při tomto zařazení je uvolněn pyrofosfát, jehož reakce s dalšími přitomnými látkami je doprovázena světelným zábleskem [20]. Světlo je detekováno výpočetním zařízením a ve výsledném pyrogramu [21] znázorněno jednotlivými píky grafu.

Další technologií umožňujicí paralelní sekvenaci je **SOLiD** od společnosti Life Technologies. Tentokrát je sekvenování prováděno ligací oligonukleotidů. Po navázaní oligonukleotidů k odpovídajícímu páru bází dochází k odstřižení barevné části, která je poté detekována [17]. Využívají se pouze 4 barvy, kdy každá z barev definuje čtyři určité páry bází. Pro přečtení určité sekvenece je proto provedeno několik cyklů, kdy jsou po detekci navázané oligonukleotidy i s primery odstraněny a v dalším kroce se navážou o jeden nukleotid dále [18].

Technologie **Ion Torrent** se od ostatních technologií NGS liší tím, že jako jediná nevyužívá barvení volných nukleotidů. Metoda je založená na změně pH roztoku, které se mění při každém zařazení nukleotidu do řetězce [22]. Měří se napětí, které odpovídá množství zařazených nukleotidů [18].

Hojně využívanou sekvenační technologií je **Illumina**, která opět využívá sekvenování syntézou. Množení vzorků probíhá pomocí můstkové PCR [23] na speciálně připravených podložkách s imobilizovanými adaptory. Komplementární adaptory jsou přidány na začátek i konec zvolené sekvence DNA. Po přichycení sekvencí na podložku probíhá amplifikace, kdy kolem vybraných sekvencí DNA vznikají shluky jejich kopií. Detekuje se fluorescence shluků, která je způsobena zařazením obarvených nukleotidů do nově vznikajícího řetězce. Jednotlivé obrazy pořízené při fluorescenci jsou dále zpracovávány a převáděny na sekvenci nukleotidů. Základní kroky této technologie jsou shrnuty na obrázku č. 1.3.

Co do počtu čtení během jednoho spuštění NGS mnohonásobně převažuje nad Sangerovou metodou, avšak v délce sekvenovaného řetězce se tyto technologie nemohou s první generací srovnávat. Zatímco jsou NGS v průměru schopny dosáhnout délky 500 bp, Sangerova metoda je schopná dosáhnout délky 1000 bp [24]. Na druhou stranu v ceně na milion bází Sangerova metoda mnohonásobně převyšuje nejlevnější sekvenaci Illuminy [22].



Obr. 1.3: Základní princip sekvenování technologií Illumina [17]

Třetí generace

Zatímco NGS produkuje relativně krátká, ale početná množství čtení, třetí generace zaznamenala i čtení dlouhé v jednotkách Mbp. Tato generace se především soustředí na sekvenování jednotlivých molekul DNA v reálném čase [22].

Mezi dostupné technologie třetí generace patří PacBio [25], který opět využívá sekvenaci syntézou, či Nanopore [26], který zaznamenává proud při průchodu sekvence skrz nanopór. Jednou z výhod třetí generace je, že nevyžaduje amplifikaci pomocí PCR, která zpomaluje proces získávání sekvencí a zároveň způsobuje jisté odchylky a chyby v takto získaných sekvencích.

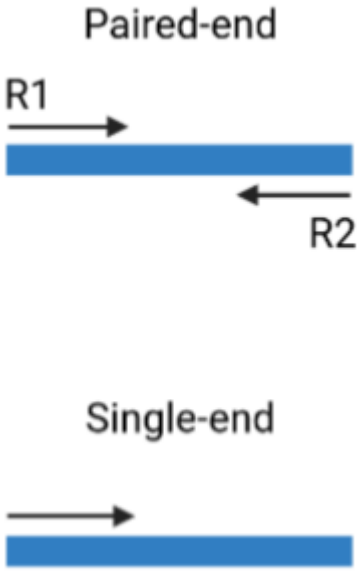
1.4 Generování čtení

Pro dosažení větší délky čtení, a tedy možnosti sekvenovat delší úseky DNA, některé NGS platformy umožňují generovat tzv. paired-end (dále PE) čtení. Zatímco klasické čtení, nazývané single-end (dále SE), je generováno od jednoho konce templátu dokud není přečteno určité množství bází dáno použitou sekvenančí technologií, PE čtení probíhá z obou konců templátu, který je sekvenován. Grafické znázornění způsobů čtení bází je pak znázorněno na obrázku č. 1.4.

První čtení PE metody začíná na jednom primeru, kdy je generováno čtení o určité délce dle použité technologie. Z druhého primeru, který nasedá na stejný templát, je poté generováno druhé čtení o stejné délce jako první. Na rozdíl od SE je při použití PE generováno dvakrát více čtení pro jeden templát, z čehož vyplývá dvojnásobně dražší provoz pro stejný počet nasekvenovaných úseků [27]. Naopak výhodou je již zmiňovaná možnost sekvenovat delší úseky sekvence či usnadnění

mapování při využití metody shotgun sekvenování.

Využívájí se různé způsoby PE čtení. Jedním z nich je částečný překryv. Po spojení obou čtení v místě překryvu tak získáváme sekvenci téměř dvojnásobné délky jednoho čtení. Dále je možný úplný překryv, pro zvýšení přesnosti sekvenování, či vzdálené čtení bez překryvu, které je velmi často využíváno při sekvenování genomu přístupem shotgun.



Obr. 1.4: Způsoby generování čtení

2 Extrakce a amplifikace

Obstarávání vzorků pro každou studii či analýzu hraje významnou roli v dosažení publikovatelných výsledků. Je tedy důležité dbát jak na kvalitu vzorku, tak i na dostatečnou reprezentativnost vzorku.

U některých studií, které sbírají vzorky s určitým časovým rozestupem, musí být tyto vzorky uskladněny. Podmínky, za kterých jsou vzorky krátkodobě uskladněny, nemají větší vliv na diverzitu či strukturu komunity mikroorganismů [28]. Naopak při dlouhodobém uchovávání je potřeba klást větší důraz na podmínky, za kterých jsou vzorky uskladněny. Nejčastěji se využívá zmražení na teplotu –80° C [29].

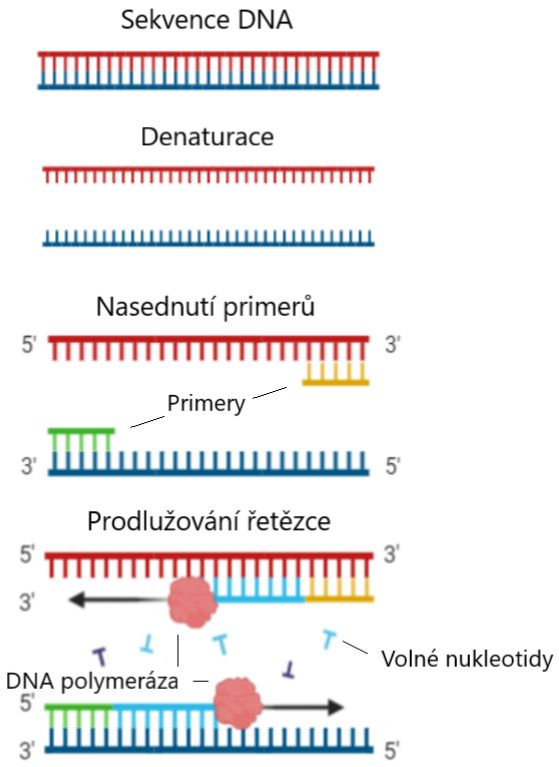
Jak již bylo zmíněno v první kapitole, při analýze mikrobiálních komunit je využíváno především amplikonové sekvenování, zatímco při metagenomické analýze je častěji využíván přistup shotgun. Obě tyto metody pracují s daty pocházejícími z genetické informace uložené v DNA. U eukaryotických buněk je DNA uložena v jádře, zatímco u buněk prokaryotických je genetická informace uložena v cytoplazmě. Pro obstarání těchto vzorků je využíván proces extrakce, který slouží k izolaci DNA z buňky. Nejprve je tedy potřeba se přes buněčnou membránu, případně i jadernou, dostat k DNA. To lze provést mechanickým rozrušením či chemickým rozpuštěním buňky [30]. V dalším kroku je nutné separovat genetickou informaci od zbylých molekul buňky. Mezi tyto metody patří např. organická extrakce či adsorpční metoda prováděna na silikagelu [30].

2.1 Polymerázová řetězová reakce

Polymerázová řetězová reakce je laboratorní metoda, která umožňuje replikaci genetické informace uložené v molekule DNA nebo RNA. Využívá se k několikanásobnému namnožení krátkých úseků těchto molekul.

Pro úspěšný průběh metody je potřeba několik komponent, které jsou smíchány ve vzorku s DNA, která má být namnožena (templátová DNA). Jednou z hlavních komponent je enzym DNA polymeráza, který katalyzuje tvorbu nového řetězce. Aby mohla být tvorba řetězce vůbec zahájena, je potřeba krátká jedno-řetězcová oligonukleotidová sekvence, na kterou tento enzym nasedá. Tato sekvence se nazývá primer a slouží pro vymezení oblasti, která má být amplifikována. Běžně je využívána dvojice primerů, které jsou vždy co nejpřesněji komplementární k templátu. Jeden z nich nasedá na vlákno v orientaci 5'-3', druhý na opačné vlákno téhož templátu. DNA polymeráza nasedá na primery a k tvorbě nového vlákna využívá volné nukleotidy, které jsou taktéž přidané do vzorku. Polymeráza tyto nukleotidy zařazuje do nově vznikající sekvence dle komplementarity bází k templátové DNA.

Metoda probíhá ve třech základních krocích, které jsou znázorněny na obrázku 2.1. První z nich je denaturace, kdy se zvýšením teploty naruší vodíkové můstky mezi vlákny DNA a původně dvouvláknová DNA je rozpletena na jednovláknové řetězce. V druhém kroku nasedají primery na 3‘ konce templátů. Nakonec DNA polymeráza nasedá na primery a tvoří komplementární vlákno k tepmlátu. Tyto kroky se opakují po libovolně zvolený počet cyklů, kdy v každém cyklu pak dochází ke zdvojnásobení počtu kopií.



Obr. 2.1: Základní kroky PCR

Taq polymeráza je nejčastěji využívanou DNA polymerázou pro potřeby PCR. Je taktéž známa jako termostabilní polymeráza, která byla izolována z termofilní bakterie *Termus aquaticus* [31]. Výhodou, jak vyplývá z jejího názvu, je možnost denaturace vysokými teplotami, aniž by došlo k jejímu zničení.

2.2 Artefakty

Jedním ze způsobů, jak minimalizovat veškeré chyby či odchylky v jakékoliv studii, je zachovávat konstantní postup pro veškeré vzorky a po celou dobu studie. Výhodou také může být zaznamenávání veškerých důležitých parametrů při práci se vzorky.

Takový záznam může poté sloužit pro zpětnou kontrolu či pro odhalení nesrovnalostí v datech.

I přes dodržování konstantních postupů a určených předpisů se ve výsledcích analýz mikrobiálních dat objevují chyby. Nejčastější příčinou těchto chyb je PCR, bez které nelze sekvenovat pomocí NGS technologií. Zabránit těmto chybám tedy úplně nelze, ale cílem je alespoň tyto chyby co nejvíce potlačit.

Podle Acinas et al. [32], artefakty způsobené PCR mohou být rozděleny do dvou skupin. PCR chyby, které se projevují ve výsledné sekvenci, a PCR odchylky, které zkreslují zastoupení jednotlivých populací v komunitě.

Výskyt odchylek je zaznamenáván v případech, kdy jsou použity univerzální primery pro sekvence různého původu. Proto je tento děj často pozorován u analýz mikrobiálních komunit, kdy je pomocí PCR amplifikováno velké množství různých mikroorganismů. Například při provádění taxonomické analýzy sekvencí 16S rRNA genu jsou voleny primery odpovídající konzervovaným regionům, které však nejsou stoprocentně shodné napříč celou mikrobiální komunitou. Podle Kanagawi a Takahira [33] pouze jedna neshoda se sekvencí primeru výrazně snižuje efektivitu amplifikace. Pro odstranění těchto odchylek se používají degenerované primery, které obsahují více možných variant bází na určité pozici [34].

Jednou z chyb vyskytujících se při použití PCR mohou být heteroduplexy. Jedná se o spojení jenodvláknových DNA, jejichž sekvence nejsou stoprocentně komplementární, do dvouvláknové DNA. Při replikaci pak může dojít ke špatnému přepisu nekomplementárních úseků a následnému vzniku umělé sekvence [33]. Pravděpodobnost formace těchto artefaktů roste s vyšším číslem PCR cyklu. V tento moment už je koncentrace primerů menší než koncentrace namnožených sekvencí, a tedy dochází ke spojení dvou téměř komplementárních sekvencí. Předejít těmto chybám lze například obnovením komponent potřebných pro PCR, především primerů a zvýšit tak jejich koncentraci [35].

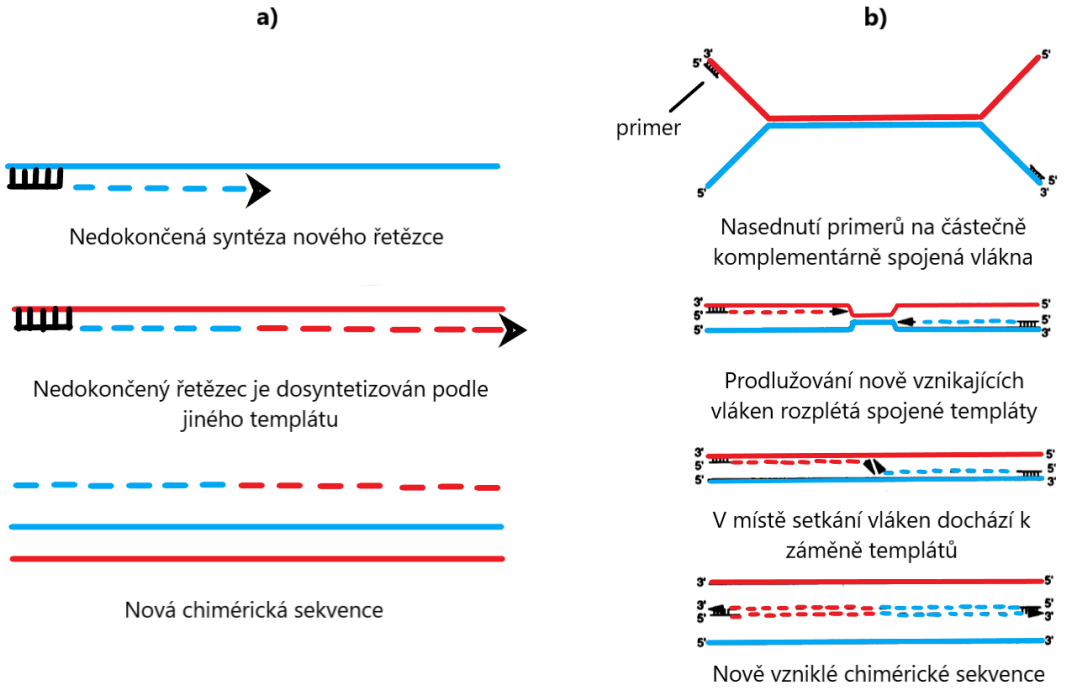
2.3 Chimérické sekvence

Převažující chybou produkovanou polymerázovou řetězovou reakcí je vznik chimérických sekvencí. Lze je popsat jako sekvence složené z několika částí, kdy každá patří jinému organismu. Chiméry jsou generovány, když se nedokončená extenze v jednom cyklu chová v dalším cyklu jako primer [36]. Tento proces generování chimérických sekvencí je znázorněn na obrázku 2.2 a). Jejich výskyt je častý při využití amplikonové sekvenace pro studium mikrobiální diverzity, kdy je amplifikace prováděna pro vzorek obsahující příbuzné mikroorganismy, jejichž sekvence vykazují vysokou míru podobnosti.

Problém může nastat, pokud chiméra vznikne v počátečních cyklech PCR. Takto jsou v dalších cyklech chimérické sekvence namnoženy a výsledkem je komunita s neexistujícími mikroorganismy.

Nedokončená extenze však nemusí být jediným procesem, který vede k vytvoření chimérických sekvencí. Podle Odelberga [37] je dalším procesem záměna templátu, která může být dále rozdělena do dvou skupin. V obou případech je záměna templátu způsobená komplementárním spojením dvou odlišných sekvencí. První z nich je zobrazena na obrázku č. 2.2 b). V tomto případě nově vznikající řetězec postupně rozplétá spojené templáty a v určité fázi dochází k záměně.

V druhém případě dochází k úplnému vysunutí nově vznikajícího řetězce mimo templát a následnému spojení nedokončeného řetězce s novou sekvencí či jiným nedokončeným řetězcem. Veškeré spojení je založeno na komplementaritě bází. Oba procesy záměny templátu jsou podobně jako heteroduplexy častější v posledních cyklech PCR, kdy ubývá koncentrace primerů [37].



Obr. 2.2: Tvoření chimérických sekvencí [37] a) Proces nedokončené extenze. b) Proces záměny templátů [37].

Tvorbu chimérických sekvencí lze snížit například prodloužením času pro nasednutí primerů [38]. Pro odstranění těchto artefaktů bylo taktéž navrženo několik algoritmů, které jsou diskutovány v následujících kapitolách.

2.4 Odstranění artefaktů

Již byly zmíněny možné artefakty, které se mohou v datech vyskytnout. Dále byly uvedeny příčiny a způsoby vzniku chimérických sekvencí, které jsou považovány za nejčastěji se vyskytující artefakt. Větší výskyt těchto arteficiálních sekvencí je zaznamenán u amplikonového sekvenování, kdy při amplifikaci jsou ve vzorku obsaženy pouze různé varianty stejných úseků genů, zatímco u shotgun sekvenování jsou ve vzorku přítomny i jiné geny či regiony genů [38].

Pro odstranění chimérických sekvencí již existuje několik metod založených na různých principech. Jedním z nich je porovnávání sekvencí s referenčními databázemi, které neobsahují chimérické sekvence. Takovými databázemi mohou být například GreenGenes [39], SILVA [40], RefSeq [41] či Ribososmal databse project (RDP) [42]. Potenciální chimérické sekvence jsou zarovnány s možnými předky těchto sekvencí a je hodnocena nejlepší shoda tohoto zarovnání. Jednotlivé metody se mohou lišit například ve způsobech, jakými jsou sekvence zarovnávány, či v hodnocení tohoto zarovnání.

Dalším principem, na kterém může být založena metoda pro detekci chimérických sekvencí, je *de novo* metoda. Podobně jako je tomu u *de novo* shlukování, ani v tomto případě není potřeba referenčních databází. Princip byl poprvé zmíněn v článku Quince et al. [36], kdy metoda Perseus předpokládá, že chimérická sekvence vznikla spojením více sekvencí, které se zúčastnily stejného či většího počtu cyklů PCR než chiméra. Využívá tedy zarovnání se všemi sekvencemi, které mají stejné či větší zastoupení.

Oba principy mohou být provedeny algoritmem UCHIME [43], který v módu srovnávání nejprve vyhledá potencionální předky chimérické sekvence a po provedeném zarovnání vyhodnotí přítomnost chiméry. V *de novo* módu se tvůrci algoritmu inspirovali algoritmem Perseus. Na rozdíl od předešlé práce se však předpokládá, že se předkové chiméry zúčastnili alespoň jednoho cyklu navíc.

Dále to je například ChimeraSlayer [38], který především využívá srovnávání s databázemi, aby odhalil chimérické sekvence.

3 Předzpracování dat

V první kapitole byly zmíněny různé sekvenační přístupy. Pro studium mikrobiálních komunit je to již zmiňované amplikonové sekvenování a pro metagenomickou analýzu pak shotgun sekvenování. Byly popsány i kroky pro obstarání mikrobiálních vzorků a uvedeny různé sekvenátory pro generování sekvenačních dat. Nedílnou součástí následné analýzy sekvenačních dat je jejich vhodné předzpracování. Cílem této kapitoly je tedy shrnout základní postup předzpracování a zároveň uvést kroky, které byly aplikovány na dostupná data.

3.1 Testovací data

Pro tuto práci byla poskytnuta data z Výzkumného ústavu veterinárního lékařství. Jedná se o datasety variabilních regionů V3-V4 genu 16S rRNA gram pozitivních bakterií, gram negativních, resp. jejich mix. Zároveň s daty byl poskytnut klíč, který obsahuje primery definující jednotlivé datasety.

Data byla sekvenována platformou MiSeq Illumina, která generuje PE čtení o délce 2x300 bp. Jednotlivé datasety jsou pak definovány dvěma soubory sekvencí ve formátu FASTQ, které odpovídají prvnímu, resp. druhému čtení PE metody.

Nutnost předzpracování obsáhlých dat generovaných NGS technologiemi vedla k vytvoření mnoha softwarů či programů, které ulehčují práci již hotovými algoritmy pro jednotlivé kroky předzpracování sekvenačních surových dat. Jedním z nejpoužívanějších softwarů, který se zabývá nejen předzpracováním rozsáhlých sekvenačních dat, se stal QIIME [44], veřejně dostupný software pro různorodé analýzy mikrobiálních dat. Jelikož byla poskytnutá data předzpracována právě v tomto softwaru, tato kapitola se také zmiňuje o základních parametrech a požadavcích sofwaru QIIME. Avšak než byla data vložena do QIIME, byla provedena tzv. demultiplexace těchto dat.

Reálné bakterie přítomné v jednotlivých datasetech jsou pak uvedeny v tabulce č. 3.1 a tabulka doplněná o bakterie v datasetu *p4* je dostupná v elektronické příloze v tabulce 3.1. Bakterie *Megasphaera elsdenii* je zde uvedena dvakrát, jelikož obsahuje více genů 16S rRNA, které jsou natolik odlišné, aby vytvořily různé shluky. Dataset *p4* je pak kombinací těchto dvou komunit bakterií.

3.2 Demultiplexace dat

Při využití NGS technologií, které umožňují sekvenování velkého množství vstupních dat, se nabízí sekvenovat více různých vzorků zároveň. Aby mohly být výstupní sekvence následně zpracovány dle vzorku, ze kterého pocházejí, je využívána metoda

Tab. 3.1: Tabulka bakterií v daných datasetech

p2 - gram-pozitivní bakterie	**p3 - gram-negativní bakterie**
Oscillibacter valericigenes	*Megamonas hypermegale*
Pseudoflavonifractor capillosus	*Bacteroides coprocola*
Butyricicoccus pullicaecorum	*Megamonas funiformis*
Anaerotruncus colihominis	*Bacteroides caecigallinarum*
Eubacterium contortum	*Bacteroides caecicola*
Lactobacillus aviarius	*Mediterranea massiliensis*
Lactobacillus ingluvie	*Bacteroides salanitronis*
Clostridium lactatifermentans	*Bacteroides plebeius*
Blautia producta	*Bifidobacterium saeculare*
	Megasphaera elsdenii
	*Megasphaera elsdenii**

multiplexace. Pro rychlejší a méně cenově náročné sekvenování jsou tak využívány specifické umělé sekvence, které jsou zařazeny mezi adaptér, nutný pro amplifikaci, a primer vymezující specifický sekvenovaný gen. Obvykle je tento krok prováděn ještě před amplifikací vzorků.

Poté, co proběhne sekvenování zvolenou technologií, je potřeba označené sekvence roztřídit podle zařazeného indetifikátoru. Tento proces je nazýván demultiplexace. Proces spočívá v nalezení specifických identifikátorů a následném zařazení sekvencí do odpovídajících vzorků, které jsou definovány právě tímto identifikátorem.

Pro získání jednotlivých datasetů, které obsahují již zmiňované bakteriální komunity, byla provedena demultiplexace poskytnutého sekvenačního běhu na základě poskytnutého klíče. Základním krokem bylo najít nejpočetnější skupinu sekvencí s primerem, který podle klíče definuje danou skupinu. Byly tedy prohledány různé pozice, na kterých se mohl primer nacházet. To bylo provedeno pro oba soubory sekvenančního běhu, tedy soubor s prvním, resp. druhým čtením. Jelikož tyto dva soubory obsahovaly stejný počet čtení, kdy odpovídající dvojice čtení byla na stejné pozici v souboru, nebyl zde zaznamenán výskyt tzv. singletonů, tedy čtení, kterému chybí párové čtení. Nalezené nejpočetnější skupiny byly extrahovány, a pokud dvojice čtení obsahovala daný primer, byla zařazena do odpovídajícího datasetu. Pokud byl primer nalezen pouze v jedné z dvojice čtení, byla tato dvojice odstraněna. Tabulka č. 3.2 pak obsahuje datasety s odpovídajícími primery a počtem sekvencí, které byly extrahovány z poskytnutého sekvenačního běhu.

Tab. 3.2: Počet sekvencí v datasetech s příslušnými primery

Dataset	**Primer**	**Počet sekvencí**
p2	ACAGTG	103 606
p3	ACACACT	71 610
p4	AGATGCT	64 029

3.3 Software QIIME

V roce 2010 Caporaso et al. [44] zveřejněnili software pro mikrobiální analýzu QIIME (Quantitative Insights Into Microbial Ecology). Cílem bylo vytvořit veřejně dostupný algoritmus zpracovávající surová data z různých NGS platforem.

V lednu roku 2018 byl QIIME nahrazen novou verzí QIIME 2, založenou na doplňkových modulech umožňujících spojení s ostatními softwary či doplňkovými algoritmy. Nová verze umožňuje ovládání přes příkazovou řádku a grafické uživatelské prostředí. Stejně jako u první verze je možné provést různé kroky mikrobiální analýzy a statisticky vyhodnotit získané informace. V rámci analýzy je možné provádět demultiplexaci surových dat, kotrolu kvality dat, OTU shlukování, filtraci chimérických sekvencí a další. Díky doplňkovým modulům je možné využít více možných algoritmů pro OTU shlukování či kontrolu kvality.

QIIME 2 umožňuje vkládání dat ve formátu FASTQ, který obsahuje kvalitu přečtení, či FASTA, u kterých informace o kvalitě přečtení chybí. Také je možné vkládat již upravená data, například fylogenetické stromy či tabulky příznaků, avšak tyto formáty musí splňovat určitá kritéria uvedená v dokumentaci softwaru.

Pro veškerou práci s daty jsou v rámci softwaru vyžadována data ve formátu QZA. Tento soubor obsahuje nejen vložená data, ale i informace o typu těchto dat, např. zda se jedná o tabulku frekvencí či samotné sekvence. Dále je v tomto formátu uchována informace o předešlých analýzách či úpravách v rámci QIIME 2 softwaru. To umožňuje uživatelům zpětnou kontrolu či objasnění provedených kroků.

Pro vizualizaci částečně či úplně zpracovaných dat, např. tabulek OTU či boxplotů kvality čtení, lze využít webového prostředí QIIME 2 View dostupného na view.qiime2.org. Je nutné, aby daný soubor s daty byl převeden z formátu QZA do vizualizovatelného formátu QZV odpovídajícími příkazy dostupnými v dokumentaci.

Mezi dostupné komplexní algoritmy dále patří například DADA (Divisive Amplicon Denoising Algorithm) [45] či UPARSE [46].

3.4 Odstřižení umělých sekvencí

Pokud sekvenační data obsahují umělé sekvence, např. identifikátory, které slouží pro demultiplexaci, je potřeba tyto části sekvencí odstranit z dat, aby neznehodnocovaly následnou analýzu. V tomto případě jsou za umělé sekvence považovány právě primery definující datsety bakeriálních komunit. Zatímco demultiplexace byla provedena v prostředí Matlab, pro další kroky předzpracování již byla data vložena do softwaru QIIME 2, což vyžadovalo převedení souborů FASTQ do již zmiňovaných souborů formátu QZA. Software QIIME 2 nabízí odstranění umělých sekvencí pomocí nástroje Cutadapt [47], kdy vstupem jsou jednotlivé sekvence a primery k odstranění. Při označení polohy primerů v sekvenci je zabráněno prohledávání možných pozic primeru a výsledný proces ořezávání je urychlen. V případě dostupných dat se primery nachází na začátku prvního i druhého čtení, a proto byla jejich pozice specificky označena.

3.5 Kontrola kvality a spojení čtení

Výstupní data sekvenačních platforem se mohou lišit formátem, ve kterém jsou zapsané. Nejčastěji využívaným formátem je FASTQ formát, který s přečtenou sekvencí uvádí i kvalitu přečtení jednotlivé báze. Toho lze využít pro prvotní filtraci, kdy jsou odstraněny sekvence, jejichž kvalita čtení nedosahuje požadované hodnoty, či odstřižení méně kvalitních konců sekvencí.

Kvalita přečtení bází je vyjádřena znaky kódu ASCII, kdy každý znak reprezentuje určitou hodnotu kvality Phred skóre. Nejčastěji využívaným spojením mezi znaky ASCII a kvalitou je Phred+33, kdy škála kvality začíná od 33. prvku tabulky ASCII. Vztah definující Phred skóre:

$$Q = -10 \cdot \log P, \tag{3.1}$$

kde Q je výsledné bezrozměrné skóre a P je pravděpodobnost nesprávného přečtení báze. V tabulce 3.3 jsou pak uvedeny konkrétní příklady skóre Q, jemu odpovídající pravděpodobnost P a přesnost přečtení. Reálně pak kvality přečtení dosahují skóre nad 40.

V případech, kdy je využita PE metoda čtení za účelem sekvenování delšího úseku, než je možné osekvenovat úsek SE metodou, musí být páry čtení pro další kroky analýzy spojeny. Stejně tomu je v QIIME 2, kdy čtení musí být spojena do konsenzuální sekvence přes překrývající se oblast, aby mohly být provedeny další kroky analýzy.

Tab. 3.3: Phred skóre

Q - skóre	P - pravděpodobnost nesprávného přečtení	Přesnost přečtení
1	0,7943	20,57%
10	0,1	90%
20	0,01	99%
30	0,001	99,9%
40	0,0001	99,99%

Trimmomatic

V případě párového čtení se nabízí použití vhodné metody pro ošetření kvality jednotlivých párů čtení před procesem spojování. Jak je patrné z boxplotů generovaných pomocí programu FastQC na obrázku č. 3.1, kde je znázorněna kvalita prvního, resp. druhého čtení, kvalita přečtení báze klesá ke konci 3'. Zároveň můžeme pozorovat modrou linii, která uvádí průměrnou kvalitu napříč čtením. Konce této linie u obou čtení sahají do červené oblasti grafu, která znázorňuje nízkou hodnotu Phred skóre, a tedy větší pravděpodobnost špatného přečtení bází.

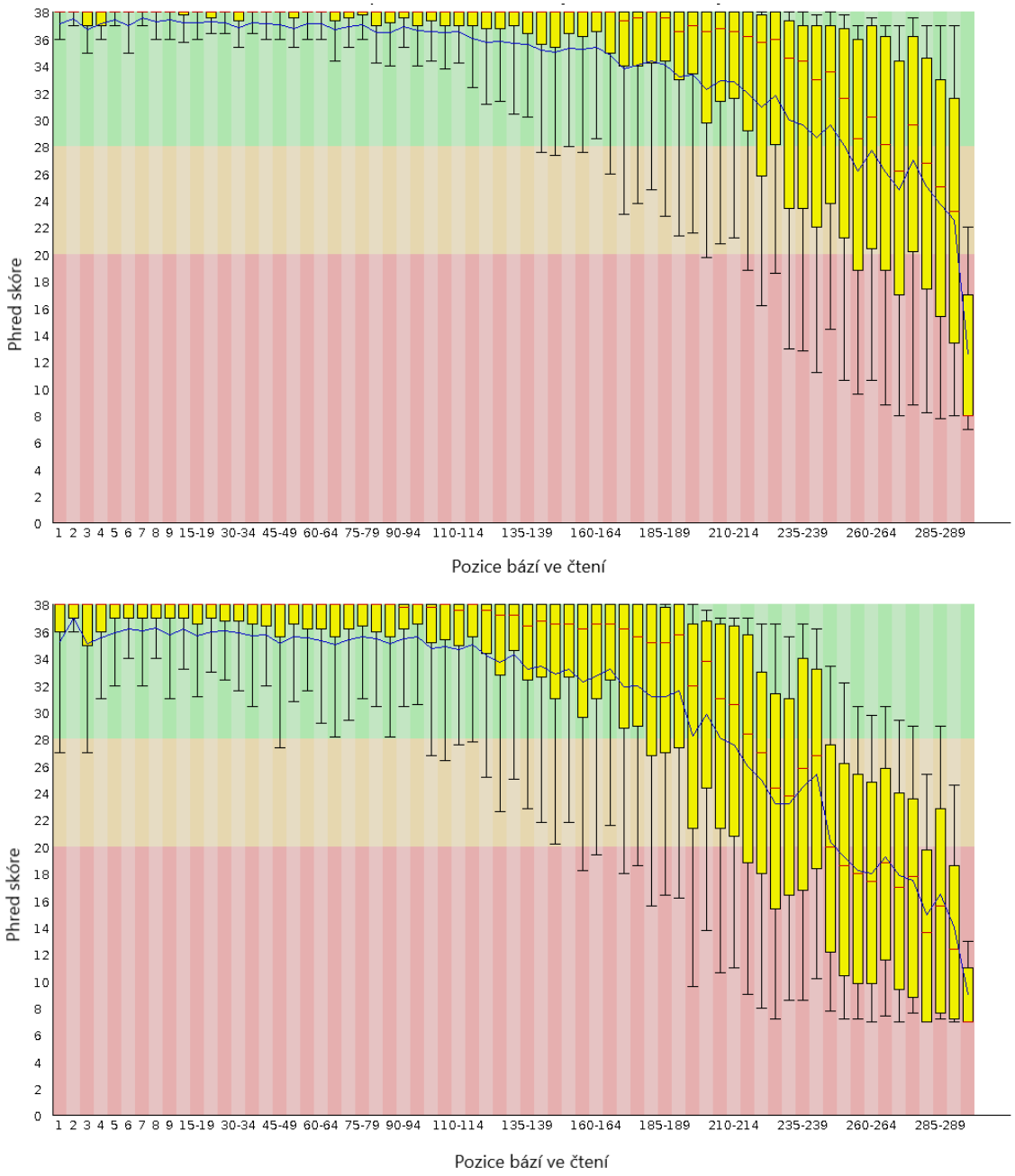
Pro odstranění těchto nekvalitních konců, které ovlivňují následný proces spojování čtení, byla data po odstřižení umělých sekvencí vložena do online verze nástroje Trimmomatic [48], [49]. Tento nástroj je využíván pro potřeby zpracování dat z platforem Illuminy, pro kterou jsou typické tyto méně kvalitní konce čtení.

Trimmomatic nabízí několik možností, jak ořezat sekvence na základě jejich kvality. Pro tyto účely byla vybrána metoda TRAILING, která ořezává jednotlivé báze na základě jejich kvality. Metoda začíná na 3' konci každého čtení a postupuje směrem k 5' konci. Veškeré báze, jejichž kvalita je nižší než předem stanovená prahová hodnota, jsou odstraněny. Metoda končí v momentě, kdy nalezne bázi o kvalitě alespoň shodné s prahovou hodnotou. Takto byla ošetřena veškerá PE čtení.

Grafy na obrázku č. 3.2 pak zobrazují kvalitu čtení po použití nástroje Trimmomatic na těch stejných datech. Můžeme vidět zlepšení kvality čtení na 3' koncích sekvencí. Taktéž je vidět, že u druhého čtení průměrná hodnota (znázorněna modrou linií) se již nenachází v červené oblasti grafu, tedy oblasti špatné kvality přečtení báze.

Spojení čtení

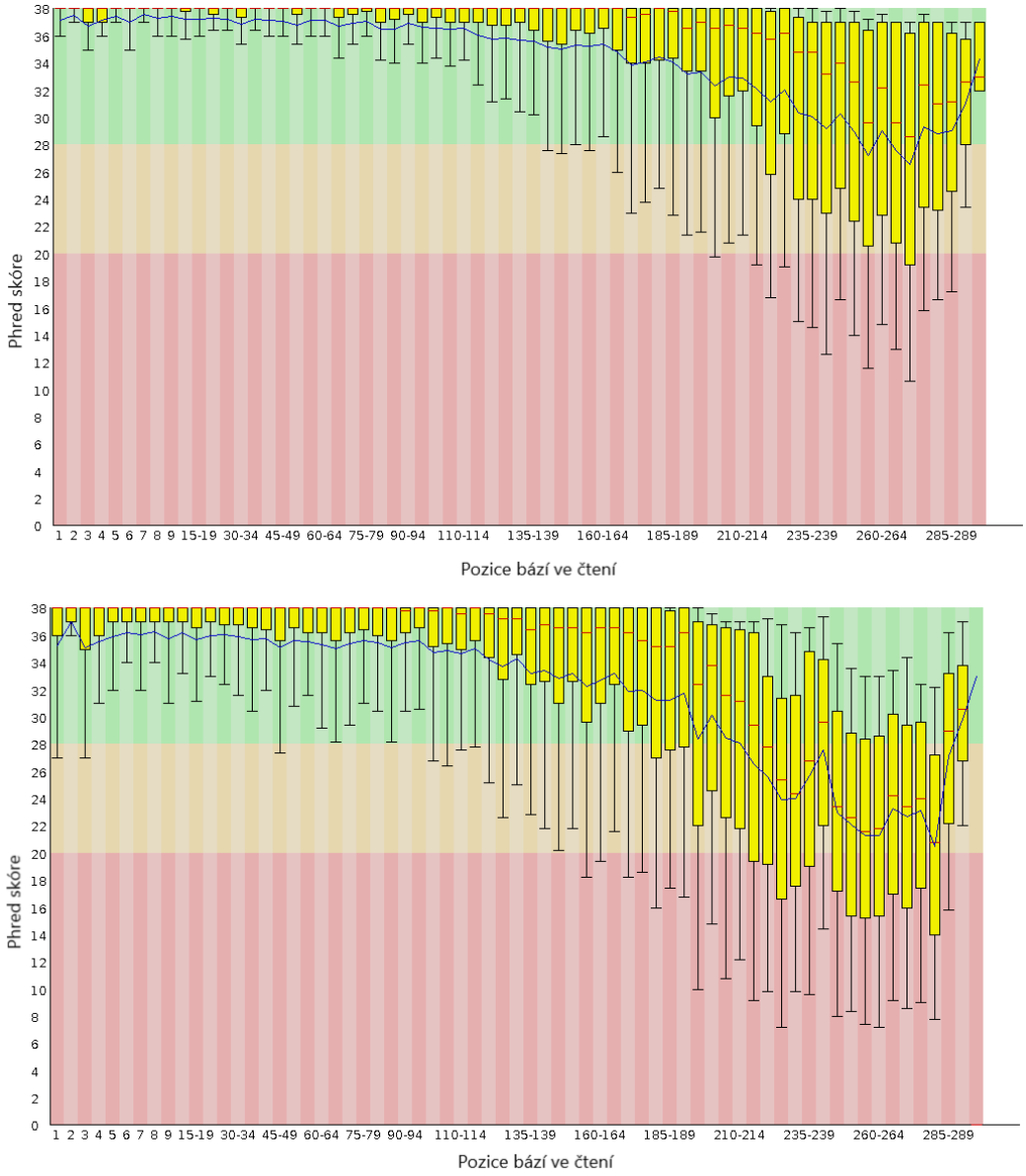
Jak již bylo zmíněno, v případech, kdy je potřeba sekvenovat delší úseky, než dovolují možnosti zvolené platformy, využívá se metod PE čtení, které se vzájemně



Obr. 3.1: FastQC boxploty před ořezáním

překrývají v daném úseku, jehož délka se může lišit. Pro další zacházení s PE čtením v QIIME 2 je potřeba jednotlivé páry čtení spojit přes překrývající se oblast do konsenzuálních sekvencí. Tyto sekvence pak představují jednotlivé úseky, které byly cílem sekvenování. Princip, na jakém je spojování čtení založeno, je znázorněn na obrázku č. 3.3.

Proces spojování tedy zahrnuje nejen vytvoření reverzního komplementu jednoho ze čtení, ale i následné hledání oblasti, která je dvojici čtení společná. Dalším kro-

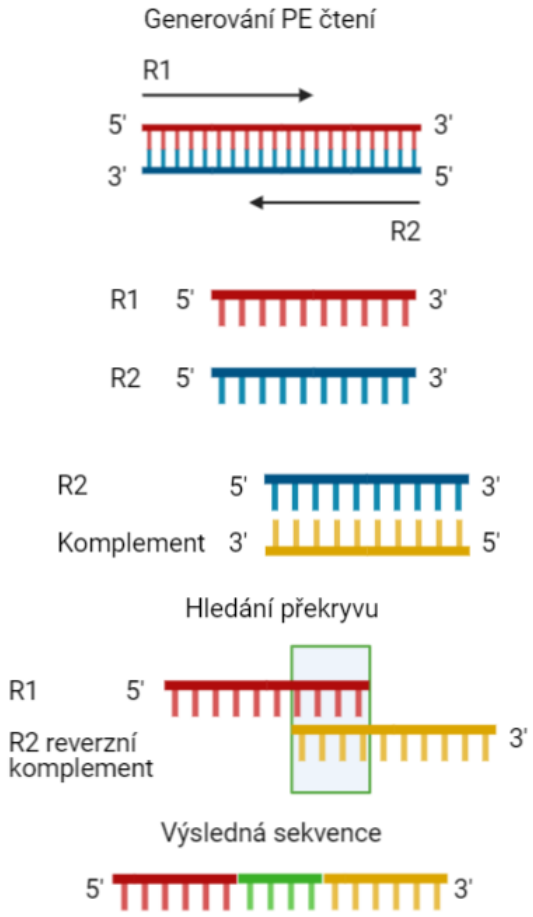


Obr. 3.2: FastQC boxploty po ořezání

kem je vytvoření konsenzuální sekvence. Ty části sekvencí, které se nepřekrývají, se přepisují do výsledné sekvence, avšak rozdíl je při vytváření výsledné sekvence z překrývající se části.

Jak vyplývá z článku Edgar et al. [50], báze jsou do konsenzuální sekvence v překrávající se oblasti zařazovány na základě jejich kvality. Při shodě v obou čteních je zařazena odpovídající báze a výsledná kvalita této báze se zvýší. Naopak je-li zaznamenána neshoda bází, do výsledné sekvence je zařazena ta báze, která má vyšší kvalitu, a výsledná kvalita této báze bude nižší.

Z principu spojování, které bylo i graficky znázorněno, je jasné, že se jednotlivé

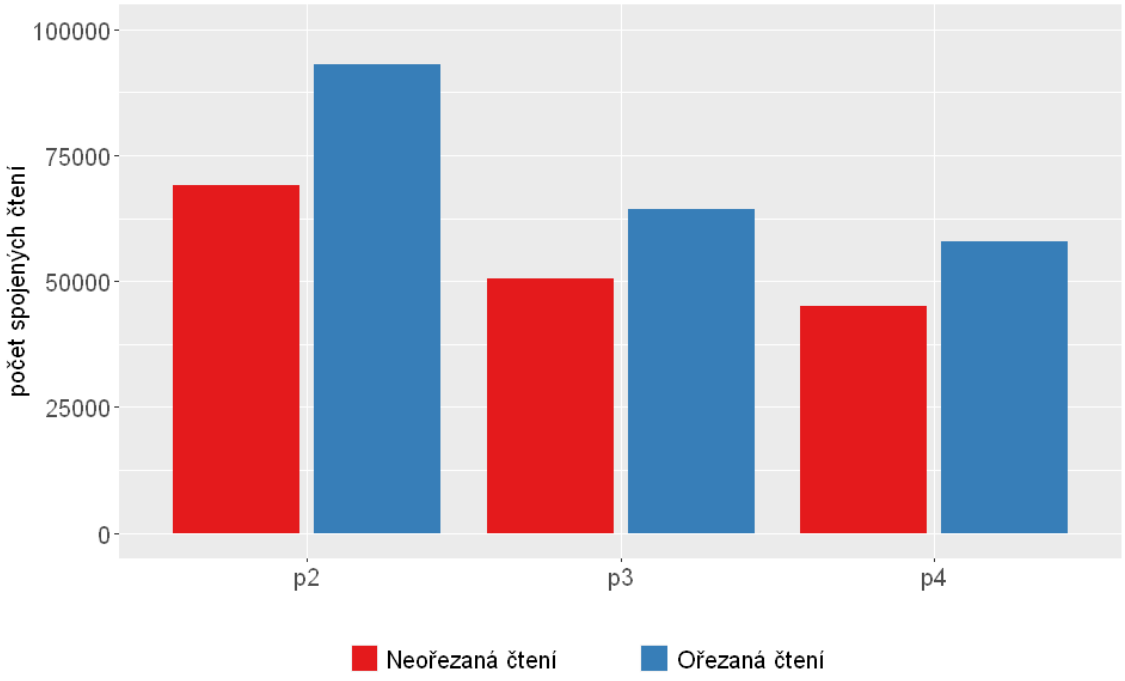


Obr. 3.3: Schéma principu metody PE čtení

páry čtení při procesu spojování překrývají méně kvalitními konci, jejichž přítomnost byla taktéž vysvětlena. Tento fakt může mít vliv na počet sekvencí, které jsou po procesu spojování zachovány. Protože výskyt méně kvalitních konců znamená větší pravděpodobnost špatně přečtených bází v této oblasti, roste i počet neshod v překrývající se oblasti.

Pro spojování čtení v softwaru QIIME 2 jsou požadovány pouze sekvence, které se mají spojit, a případně volitelné parametry pro průběh tohoto procesu. Jedním z nich je parametr *-p-maxdiffs*, který dovoluje určitý počet neshod v překrývající oblasti. Defaultní hodnota tohoto parametru je nastavena na 10, tedy pokud se vyskytne v překrývající se oblasti více jak 10 neshod, daná dvojice čtení je odstraněna. Z předešlého odstavce vyplývá, že by tento parametr měl mít velký vliv na úspěšnost procesu spojování. I přes rostoucí počty spojených čtení s vyšší hodnotou tohoto parametru byla využita defaultně nastavená hodnota. Důvodem je fakt, že větší počet neshod v překrývajícím se regionu by mohl vést k nekvalitním sekvencím,

které by mohly ovlivnit průběh analýzy. Jednotlivé hodnoty tohoto parametru a počtu úspěšně spojených sekvencí je znázorněn grafem na obrázku č. 3.4.



Obr. 3.4: Histogram spojování čtení před, resp. po kvalitním ořezaní

Problém ztráty dat při procesu spojování čtení nekvalitními konci byl tedy vyřešen ořezáním sekvencí nástrojem Trimmomatic, který byl zmíněn dříve. Graf, který znázorňuje vztah mezi sekvencemi spojenými bez použití nástroje Trimmomatic, resp. s použitím, je znázorněn na obrázku č. 3.4.

Pozorujeme tedy, že díky kvalitnímu ořezání byla téměř plně kompenzována 25% ztráta dat při procesu spojování párových čtení.

4 Operační taxonomická jednotka

Operační taxonomickou jednotku (OTU) lze chápat jako aproximaci taxonomické kategorie při klasifikaci organismů v určitém studovaném prostředí. Shlukováním organismů na základě jejich podobnosti vznikají OTU, které většinou reprezentují druh či rod. V případě mikrobiologických analýz organismů, resp. mikroorganismů, je OTU shlukování prováděno na základě podobnosti jejich genetické informace. Po OTU shlukování pak dochází ke kompresi dat, kterou je ulehčena následující práce s daty.

Shlukování organismů začíná výběrem vhodné prahové hodnoty, kterou je minimální identita dvou sekvencí, které jsou ještě zařazeny do stejného shluku. Pokud jsou si organismy podobné méně, jsou zařazeny do odlišných shluků, které jsou pak reprezentovány centroidem, tedy středovou sekvencí. Výběr prahové hodnoty záleží na použitém genu, dle kterého jsou organismy rozřazovány, či taxonomické kategorii, kterou mají jednotlivé OTU zastoupit [29]. Pro shlukování organismů ze stejného druhu se využívá 97% podobnost sekvencí a pro organismy z jednoho rodu je používáná podobnost sekvencí v 95%.

OTU shlukování je možné provádět třemi způsoby:

De novo

De novo shlukování je založeno na srovnávání veškerých dvojic sekvencí. Splňují-li sekvence dané kritérium podobnosti, jsou zařazeny do stejného shluku. Jako jediná metoda pracuje bez porovnávání s databází [51] . Nevýhodou je zde rostoucí výpočetní náročnost s rostoucím množstvím dat. Naopak výhodou je zařazení veškerých sekvencí do některého ze shluků. *De novo* shlukování se používá v případech, pracuje-li se s neobvyklým genem, jenž nemá referenční databází ke srovnávání.

Closed reference

Sekvence jsou srovnávány s předefinovanými centroidy shluků referenční databáze. Pokud sekvence není dostatečně identická s žádným centroidem, je tato sekvence považována za chybu a je následně odstraněna [29]. Tato metoda je vhodná pro velké množství dat. Naopak nevýhodou je, že nemůže dojít k objevení nového organismu. Zjednodušení pak spočívá ve tvorbě taxonomického stromu, který je možný vytvořit pomocí této databáze.

Open reference

Kombinací předchozích dvou způsobů vznikla open reference metoda. Nejprve jsou sekvence srovnávány s referenční databází a na ty sekvence, které nebyly přiřazeny do žádného shluku, je použita metoda *de novo*. Tímto způsobem tedy mohou být objevovány nové organismy a zároveň se snížila výpočetní náročnost díky prvotnímu srovnávání s referenční databází [51].

Referenční databáze

Mezi referenční databáze, které jsou využívány u druhého, resp. třetího způsobu OTU shlukování, patří již zmiňované GreenGenes, SILVA či RefSeq. Databáze Greengenes je určena pro říší bakterií a archeí, zatímco RefSeqs a SILVA slouží i pro porovnávání eukaryotických genů.

4.1 Dereplikace

Před tím, než jsou sekvence shlukovány do OTU, QIIME 2 vyžaduje tzv. dereplikaci, tedy shlukování shodných variant amplikonů (ASV) vyskytujících se v datasetu. Tímto jsou sekvence, které jsou si 100% podobné, zařazeny do stejného shluku. Pro tento krok je tedy typické, že se zde nachází mnoho sekvencí s velmi nízkou frekvencí výskytu. Důvodem je, že stačí změna jediného nukleotidu, a sekvence už není zařazena ke shluku, se kterým má zbytek nukleotidů shodných.

Výstupem dereplikace je pak tabulka jednotlivých ASV, které jsou zde reprezentovány identifikátorem, a frekvence jejich výskytu. Zároveň je k dispozici soubor, kde je k jednotlivým identifikátorům přiřazena odpovídající sekvence, která definuje jednotlivé ASV.

4.2 VSEARCH shlukování

Po dereplikaci testovacích dat algoritmem VSEARCH v rámci QIIME 2, bylo využito *de novo* shlukovací metody, která nepotřebuje přístup k referenční databázi a srovnává pouze sekvence v datasetu mezi sebou. Hranice pro zařazení sekvence do OTU byla nastavena na 97% podobnost, tedy pro zařazení do OTU musí být odlišnost sekvence a OTU reprezentující sekvence maximálně 3%.

Vstupní data jsou požadována ve výstupním formátu dereplikace, tedy tabulka frekvencí a soubor identifikátorů s odpovídajícími sekvencemi. Výstupem jsou pak stejné formáty, které však obsahují odlišné hodnoty. Tabulka frekvencí obsahuje

OTU reprezentované identifikátory, ke kterým jsou v druhém souboru přiřazeny sekvence, které odpovídají centroidům jednotlivých OTU.

Pro vysoký počet OTU byly vyfiltrovány ty shluky, které obsahovaly 10 a méně sekvencí. Výsledný počet OTU v jednotlivých datasetech je uveden v tabulce č. 4.1, kde je i mimo jiné uveden pro porovnání počet OTU získaných při shlukování na 99% podobnosti. I při tomto shlukování byly odstraněny OTU s nižší frekvencí výskytu než 10.

Tab. 4.1: Počet OTU v jednotlivých datasetech s počtem sekvencí po OTU shlukování

Dataset	**OTU 99%**	**Počet sekvencí**	**OTU 97%**	**Počet sekvencí**
p2	386	73 786	134	89 806
p3	278	47 124	122	61 053
p4	181	45 313	54	55 424

Dále lze z tabulky vyčíst i počet sekvencí po OTU shlukování, kde došlo ke snížení počtu oproti kroku spojení čtení. Tato ztráta byla způsobena již zmiňovanou filtrací OTU s nízkými frekvencemi výskytu. Je však patrné, že při shlukování na 99% došlo k vyšším ztrátám. Kritériem pro vybrání hranice podobnosti na 97% tedy byl nejen výrazně nižší počet OTU, ale i menší ztráta dat.

4.3 Kvantifikace shlukování

Obvyklým výstupem shlukování sekvencí do jednotlivých OTU jsou tzv. OTU tabulky, které obsahují nejen informaci o přítomných OTU, např. v podobě jejich identifikátoru (ID), ale i počet sekvencí, které byly do dané OTU zařazeny. Tabulka č 4.2 je zde názornou ukázkou takové OTU tabulky. Jedná se o tabulku po OTU shlukování na 97% a následné filtraci dle počtu zařazených sekvencí. Celá podoba OTU tabulky je dostupná v elektronické příloze v tabulce 4.1. Hodnoty frekvencí jednotlivých OTU pak mohou být nahrazeny relativními frekvencemi.

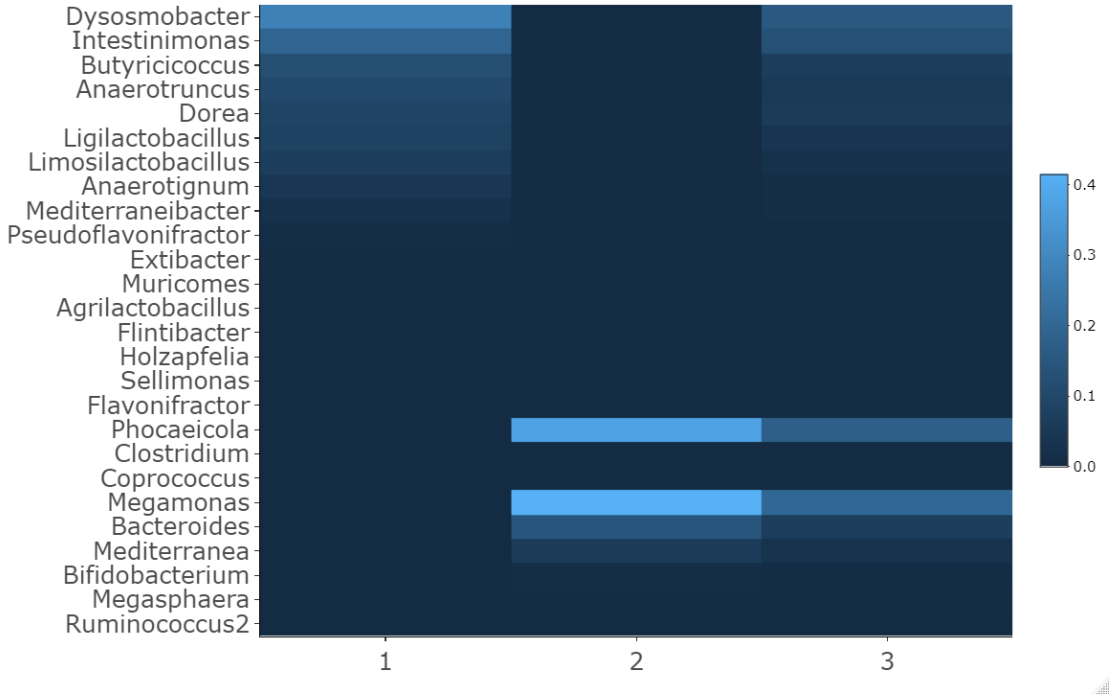
4.4 Klasifikace organismů

Pro zjištění, které bakterie se ve vzorcích nachází, byla provedena i klasifikace výsledných OTU. Pro tyto účely byl využit veřejně dostupný klasifikátor RDP [52], který umožňuje klasifikaci až po taxonomickou kategorii rod. Byl zvolen práh spolehlivosti na 95%. Pokud by spolehlivost klasifikace vybrané sekvence klesla pod tuto hodnotu, byla by sekvence na dané úrovni označena jako neklasifikována.

Tab. 4.2: Příklad OTU tabulky

OTU_ID	p2	p3	p4
866ae498cda3cae1066f783b342b001714070d9a	22861	0	0
56d31fdd8a262616bad30067096dddff997d5b9b	16588	0	0
07cf1af7b9058db9000c103b4707df039b198062	9606	0	0
b26b78dcb93822b8e56f17c0684c8b1bd70a736b	0	15338	7081
828fb71400fd6afb77e67f0eae33e0ef1ee17536	0	5956	1409
756b2d5bfe2ae53b4a6cf51e44131a9f4e501f6a	0	8910	0

Pro znázornění rodů vyskytujících se v datech byla vykreslena heatmapa (obrázek č. 4.1). Jsou zde uvedeny všechny rody, které se nachází v datech před filtrací chimér. Barevná škála pak znázorňuje relativní zastoupení rodů v jednotlivých datasetech.



Obr. 4.1: Heatmapa rodů bakterií před odstraněním chimér

V datasetu *p2*, jak lze z heatmapy vyčíst, je velké množství málo abundantních rodů, zatímco v datasetu *p3* pozorujeme, že zde převažuje rod *Megamonas* s rodem *Phocaeicola*. U datasetu *p4* je pak viditelná kombinace rodů z předchozích datasetů.

5 Detekce a filtrace chimér

Jak již bylo zmíněno, chimérické sekvence jsou nejčastěji se vyskytujícím artefaktem při amplikonové sekvenaci dat. Tudíž jejich odstranění je jedním z kritických kroků při předzpracování sekvenačních dat s cílem jejich následné analýzy. Existuje několik dostupných algoritmů, které se věnují této problematice. Jedním z nich je velmi spolehlivý a hojně využívaný algoritmus UCHIME. Volně dostupná verze tohoto algoritmu je součástí softwaru QIIME 2.

V rámci této práce byl realizován algoritmus pro detekci a filtraci chimér *ChimeraDetector* v jazyce R. Poskytnutá data se známým reálným složením bakteriálních komunit poté sloužila pro kontrolu a vyhodnocení správnosti algoritmu.

Tato kapitola obsahuje nejen popis samotného algoritmu *ChimeraDetector*, ale i dosažené výsledky v porovnání s již zmíněným algoritmem UCHIME.

5.1 ChimeraDetector

Metoda detekce chimér je založená na principu *de novo*, tedy bez potřeby přístupu k referenční databázi. Základem je pak předpoklad, že vzniklé chiméry jsou kombinací pouze sekvencí přítomných v daném datasetu, jejichž frekvence výskytu je alespoň dvakrát větší než výskyt chimérické frekvence. Dalším předpokladem algoritmu je vznik chimér pouze kombinací dvou sekvencí, což ovšem znemožňuje detekci tzv. multimér, tedy výskyt sekvencí, které vznikly kombinací více než dvou předků. Protože výskyt multimér není tak častý jako výskyt tzv. bimér, tedy chimér pouze z dvou předků, algoritmus není stavěný na detekci těchto sekvencí.

Algoritmus *ChimeraDerector* je realizován v IDE jazyka R - RStudio. Vstupem samotné funkce realizující detekci a filtraci chimér jsou sekvence ve formátu FASTA, tedy samotné sekvence a jejich odpovídající název či identifikátor (dále ID). Dalším vstupem je tabulka frekvencí jednotlivých sekvencí, které jsou zde reprezentovány pouze odpovídajícím názvem či ID. Výstupem funkce jsou pak 2 soubory FASTA obsahující chimérické sekvence a jejich ID, resp. nechimérické sekvence s příslušným ID. Třetí soubor na výstupu je také tabulka, která obsahuje ID chiméry, ID odpovídajících předků a podobnost chiméry s vytvořeným modelem.

V úvodu algoritmu je definována proměnná *refSeqs*, která slouží jako referenční databáze pouze těch sekvencí, které se vyskytují v daném datasetu. Do této proměnné jsou zařazeny ty sekvence, které nebyly označené jako chiméra a mohou být tedy potenciálními předky chiméry. Aby mohla začít detekce chimér, kdy je potřeba minimálně dvou předků pro jejich vznik, jsou do referenční databáze na úvod uloženy dvě sekvence, které jsou v daném datasetu nejvíce abundantní. Není totiž prakticky možné, aby tyto sekvence byly chiméry. Dále jsou tyto dvě sekvence a

jejich frekvence odstraněny z tabulky frekvencí, ze které jsou sekvence vybírány v pořadí od největší abundance po nejnižší.

Nalezení konzervovaného regionu

Prvním krokem algoritmu po definování proměnných je nalezení regionu, který je pro daný vzorek sekvencí nejvíce konstantní. Jak již bylo zmíněno, předpokládá se, že chiméry jsou kombinací pouze dvou předků, tedy tzv. biméry, jejichž vznik je více pravděpodobný, než vznik tzv. multimér, které jsou kombinaci 2 a více předků. Dále se předpokládá, že pokud chiméra vznikla, přechod mezi prvním a druhým předkem bude v místě, kde jsou si sekvence nejvíce podobné, ne-li totožné. Tento předpoklad, je založen na faktu, že navázání nedokončené sekvence na jinou sekvenci, a tedy častý princip vzniku chimér, je velmi pravděpodobný v místě největší shody.

U genu 16S rRNA, jak již bylo zmiňováno v úvodní kapitole, jsou tato místa nazývána konzervovanými regiony, které slouží mimo jiné k vytváření univerzálních primerů pro amplifikaci pomocí PCR. Pro nalezení nejvíce konzervované oblasti v rámci dostupných sekvencí slouží funkce *ConservedPart*, která má na vstupu právě všechny sekvence, a výstupem je část sekvence, která odpovídá nejstálejšímu regionu.

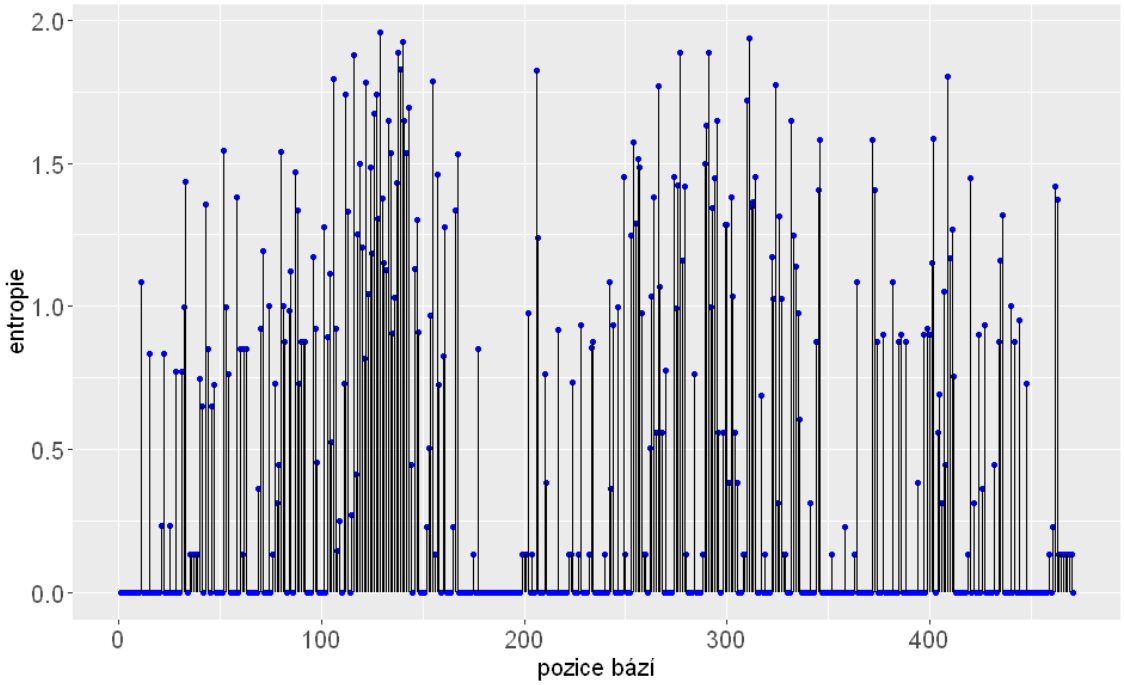
V rámci funkce pro nalezení tohoto regionu je nejprve provedeno zarovnání všech sekvencí z datasetu umožněného balíčkem *msa* [53], který mimo jiné umožňuje vícenásobné zarovnání algoritmem *ClustalOmega* v jazyce R. Z takto zarovnaných sekvencí je vypočítána Shannonova entropie pro každou pozici zarovnání podle vzorce [54]:

$$H = -\sum_{i} f_i + \log_2 f_i, \tag{5.1}$$

kdy H je entropie na dané pozici zarovnání a f zde odpovídá relativní frekvenci dané báze z množiny i = [A, C, G, T]. Jednotkou entropie je bit a v tomto případě může entropie nabývat hodnot od 0 do 2. Vysoká hodnota entropie pak odpovídá nejasnému přiřazení nukleotidu na danou pozici při vytváření konsenzuální sekvence z vícenásobného zarovnání. Naopak nulová hodnota entropie odpovídá 100% shodě nukleotidů na určité pozici zarovnání. Takto vypočítanou entropii pro konsenzuální sekvenci vícenásobného zarovnání datasetu p4 pak můžeme pozorovat na grafu č. 5.1.

Pro výpočetní nalezení dostatečně dlouhého konzervovaného regionu, byla entropie konsenzuální sekvence skenována průměrovacím oknem o délce 21 bp. Výsledná entropie sekvence po skenování oknem je znázorněna na grafu č. 5.2.

V grafech lze pozorovat rozložení oblastí s vysokou a nízkou entropií. Obzvlášť zajímavý region je v okolí 200 bp, tedy místo, kde se nachází delší úsek nulové entropie. Jelikož se jedná o entropii vypočtenou z jednoho z dostupných datasetů,



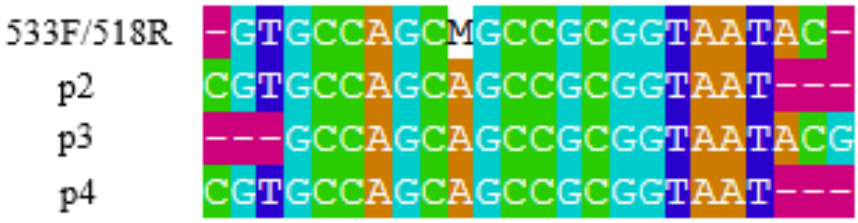
Obr. 5.1: Entropie konzervovaného regionu

Obr. 5.2: Průměrná entropie konzervovaného regionu

lze usoudit, že se jedná o konzervovaný region, který odděluje V3 a V4 variabilní

regiony genu 16S rRNA, ze kterého pochází poskytnutá data.

Jako pro většinu oblastí mezi variabilními regiony genu 16S rRNA, byl i pro tuto oblast vytvořen univerzální primer, který ji definuje a je využíván pro nalezení přilehlých variabilních regionů. Pro oblast mezi regiony V3 a V4 existují dva univerzální primery, které se z velké části překrývají. Jedná se o primery 518R a 533F, ze kterých byl pro pokrytí větší části konstantní oblasti vytvořen konsenzuální primer. Pro srovnání konsenzuálního primeru pro tuto oblast a pro oblast, která byla nalezena pomocí výpočtu Shannonovy entropie pro všechny dostupné datasety, je na obrázku č. 5.3 znázorněno jejich zarovnání.



Obr. 5.3: Zarovnání primerů pro konzervovanou oblast

Ze zarovnání lze tedy odvodit, že nalezení konzervovaného regionu pomocí Shannonovy entropie v tomto případě odpovídá reálné oblasti mezi V3 a V4 variabilními regiony genu 16S rRNA.

Nalezení potenciálních předků chiméry

Jelikož algoritmus detekce chimér je založen na principu metody *de novo*, je využito předpokladu, že chiméra vznikla až potom, co byla alespoň jednou zmnožena sekvence jejích předků. Protože v každém cyklu PCR vzniká dvojnásobné množství genetického materiálu, předek chiméry by měl být alespoň dvakrát více abundantní než chiméra. Tato podmínka je tedy využita k tomu, aby potenciální chiméra nebyla srovnávána se sekvencí, která nemůže být jejím předkem. K filtraci potenciálních předků slouží funkce *FilterPotentialParents*, jejímž vstupem je samotná *query*, referenční databáze možných předků *refSeqs* a tabulka frekvencí všech sekvencí v datasetu. Výstupem funkce jsou tedy jen ty sekvence, které splňují zmíněnou podmínku o frekvenci výskytu.

Po filtraci sekvencí založené na jejich abundanci, je *query* rozdělena na dvě části. Část před konzervovaným regionem, tedy v dostupných datech odpovídá V3 oblasti, a část za konzervovaným regionem, tedy V4 oblast. Pro každou z těchto oblastí je

hledána nejpodobnější sekvence, neboli předek pro danou oblast. K tomuto kroku slouží funkce *FindPotentialParents*.

Vstupem funkce *FindPotentialParents* jsou již filtrované sekvence na základě jejich abundance a odpovídající část *query*. Poté je oblast, V3 nebo V4, rozdělena na 4 podoblasti, ke kterým je pomocí lokálního zarovnání dvou sekvencí hledána nejlepší shoda. Výstupem funkce je tabulka, kde každý řádek odpovídá ID předka a sloupce odpovídají danému úseku. Hodnoty matice jsou jednotlivá skóre lokálního zarovnání query a předka. Jednotlivá skóre podoblastí každého předka jsou následně sečtena a potenciální předek chiméry je ten, jehož součet skóre je nejvyšší.

Pokud obě oblasti *query* mají shodného předka, algoritmus označí sekvenci jako nechimérickou a zařazuje ji do referenční databáze.

Model a určení chiméry

Pro zjištění, zda se doopravdy jedná o chiméru, je potřeba vytvořit její model. Modelem je složená sekvence z předka pro oblast před konzervovaným regionem, konzervovaným regionem a předkem pro oblast za tímto regionem.

Poté je provedno globální zarovnání *query* s modelem chiméry, sekvencí předka před konstantním regionem a sekvencí předka za touto oblastí. Jednotlivá skóre jsou srovnána, a pokud je *query* nejvíce podobná modelu chiméry, je vyhodnocena jako chimérická sekvence a zařazena do databáze chimér.

Kromě ID *query* a ID obou předků chiméry je vypočítána podobnost modelu a chiméry. K tomu slouží funkce *Identity*, která má na vstupu globálně zarovnanou právě *query* s modelem. Pro zjištění podobnosti je využit výpočet Hammingovy vzdálenosti, je odvozena od počtu substitucí v sekvencích. Jelikož je pro výpočet této vzdálenosti potřeba stejně dlouhých sekvencí a sekvence v datasetech nejsou stejné délky, v rámci funkce *Identity* musela být tato podmínka zajištěna. Veškeré mezery, které byly vloženy do sekvencí při globálním zarovnání, jsou vyhledány a odstraněny spolu s odpovídajícími bázemi na stejné pozici opačné sekvence. Tímto je zaručena stejná délka sekvencí a může být vypočítána Hammingova vzdálenost, která definuje podobnost sekvencí.

Výstupem funkce, jak již bylo zmíněno, jsou tedy chimérické sekvence, sekvence nechimérické a tabulka parametrů detekovaných chimér.

5.2 Program UCHIME

V rámci softwaru QIIME 2, jak již bylo zmíněno, je taktéž možné odstranit chimérické sekvence z dat. Nabízí se hned dvě metody programu UCHIME, a to referenční a *de novo* metoda. Soubory požadované na vstupu algoritmu jsou samotné sekvence,

které mají být prohledány a filtrovány, a dále tabulka frekvencí těchto sekvencí v datasetu. Pokud se jedná o referenční metodu, dalším vstupem je referenční databáze, která obsahuje prověřené sekvence reálných organismů. Předpokladem je pak, že výskyt artefaktů je v těchto databázích v ideálním případě nulový. Příkladem takových databází pak může být již zmiňovaná GreenGenes nebo SILVA.

Referenční metoda

Algoritmus považuje veškeré sekvence ve vloženém souboru za potenciální chiméry. Prvním krokem je tedy prohledávání refereční databáze s cílem nalezení nejvhodnějších předků, ze kterých mohla potenciální chiméra vzniknout.

Po nalezení vhodných předků je sestaven model z těchto sekvencí, který nejvíce odpovídá potenciální chiméře. Pokud je sestavený model chiméry podobnější samotné potenciální chiméře než jeden z předků sám, je dále počítáno skóre ze zarovnání potenciální chiméry a jejich předků. Pokud skóre dosáhne určité prahové hodnoty, analyzovaná sekvence je prohlášená za chimérickou a pokračuje analýza další sekvence v datasetu. Pokud je potenciální chiméra více podobná jedné konkrétní sekvenci v databázi než sestavenému modelu chiméry, je uložena do nechimérických sekvencí. To stejné platí v případě, nepřesáhne-li skóre při zarovnání zvoleného prahu. Takové sekvence pak podle algoritmu odpovídají reálným organismům v datasetu.

Pro hledání chimér referenční metodou algoritmu UCHIME byly využity referenční databáze GreenGenes. I přes poslední aktualizaci databáze v roce 2013 patří GreenGenes stále mezi hojně využívanou databázi. GreenGenes nabízí více databází, se kterými lze sekvence porovnávat. Tyto dílčí databáze se liší v hodnotě podobnosti, na základě které byly sekvence vzájemně shlukovány. Pro porovnání byly sekvence datasetu **p4** filtrovány dle databáze se shluky sekvencí na základě 99, 97 a 88% podobnosti. Tabulka č. 5.1 pak obsahuje právě úspěšnost detekce při aplikaci jednotlivých databází.

Tab. 5.1: Srovnání použitých referenčních databází

Referenční databáze	**Detkované chiméry**	**Se**	**Nedetekované chiméry**	**Počet FP**
GreenGenes-99	31	0.91	3	0
GreenGenes-97	30	0.88	4	0
GreenGenes-88	30	0.82	4	1

Referenční metoda byla aplikována na data již po OTU shlukování do jednotlivých druhů bakterií, tedy po shlukování sekvencí, které jsou si podobné na 97% a více. Nejlépe pak dopadla detekce chimér s referenční databází sekvencí s 99% podobností. I když detekce nebyla stoprocentní, pro další porovnání byla využívána právě tato databáze.

Zda má různé zařazení filtrace chimér do posloupnosti kroků předzpracování sekvenačních dat vliv na výsledné zastoupení bakterií ve vzorku, byla provedena filtrace chimér i před krokem OTU shlukování, tedy po dereplikaci sekvencí. Tabulka č. 5.2 pak obsahuje abundance vybraných bakterií po filtraci chimér refereční metodou před 97% OTU shlukováním a po něm. Pro obě metody byla zvolena referenční databáze GreenGenes se sekvencemi s 99% podobností. Referenční metoda použitá po dereplikaci je v tabulce označena symbolem *. Kompletní frekvence bakterií je dostupná v elektronické příloze v tabulce č. 5.1.

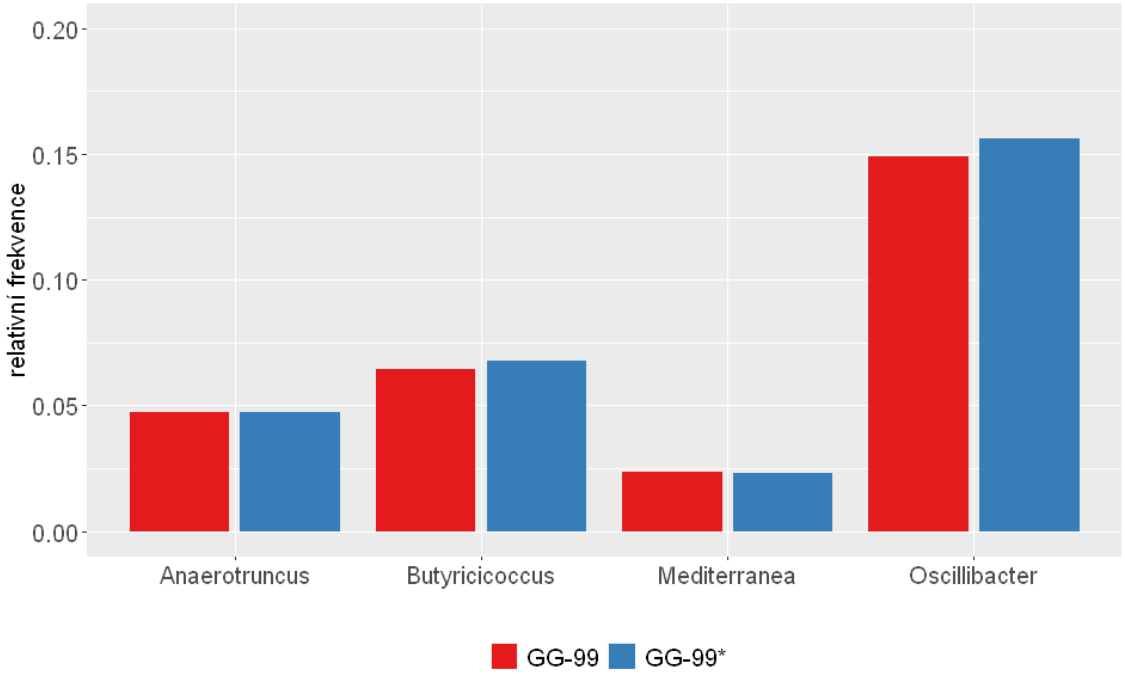
Tab. 5.2: Frekvence vybraných bakterií po filtraci chimér

	abundance	
Bakterie	**GG-99**	**GG-99***
Anaerotruncus colihominis	2605	2133
Butyricicoccus pullicaecorum	3553	3053
Mediterranea massilliensis	1294	1042
Oscillibacter valericigenes	8189	7049
celkový počet sekvencí	54957	45173

Při detekci a filtraci chimér po dereplikaci bylo filtrováno mnohem více chimér než při filtraci po shlukování bakterií do druhů. Je tedy patrné, že při filtraci po shlukování dochází k zařazení chimérických sekvencí do jednotlivých OTU, se kterými jsou velmi podobné. Také ale můžeme říct, že při filtraci po dereplikaci mohlo dojít k falešně pozitivní detekci chimér, tedy že sekvence, které nebyly chimérami, byly označeny jako chiméry.

Sloupcový graf č. 5.4 pak ukazuje, že i přes jednotlivé rozdíly v abundanci daných bakterií při odlišné filtraci jsou vypočtené relativní frekvence vybraných bakterií téměř shodné. Z toho lze usoudit, že nedochází k ovlivnění relativního zastoupení bakterií při různém zařazení kroku filtrace chimér referenční metodou v předzpracování mikrobiálních dat. Kompletní graf všech přítomných bakterií a jejich relativní

frekvence je dostupný v elektronické příloze pod číslem 5.1.



Obr. 5.4: Relativní frekvence bakterií

De novo metoda

Základ algoritmu pro *de novo* metodu je stejný. Hlavním rozdílem je, že *de novo* metoda začíná od sekvencí, které jsou v datech nejvíce abundantní, tudíž nemohou být chimérami. Předkem chiméry pak může být pouze taková sekvence, která je alespoň dvakrát více abundantní než samotná chiméra.

Jak Edgar et al. [43] uvádí, metoda *de novo* by měla být v postupu předzpracování dat zařazena za proces dereplikace a před OTU shlukování. Důvodem je, že v tomto místě předzpracování data obsahují ideálně všechny ASV v odpovídajících shlucích s jejich reálnou frekvencí výskytu, a tedy podmínka, že chiméra podstoupila alespoň o jeden cyklus PCR méně než její předek, je aplikovatelná.

Dalším důvodem, proč by metoda založená na principu *de novo* měla být použita před shlukováním OTU, je ten, že v případě shlukování mohou být chiméry dostatečně podobné sekvenci reálného organismu, aby byly zařazeny do stejného shluku. Tímto může být zkreslena analýza diverzit, která definuje daná společenstva organismů. Pro porovnání vlivu detekce chimér metodou *de novo* před a po OTU shlukování, byly použity obě dvě varianty.

Jelikož sekvence v datasetu jsou odlišných délek, krokem dereplikace jsou získány ASV, které obsahují pouze sekvence podobné na 100%. Aby tedy do odpovídajících ASV mohly být zařazeny i ty sekvence, které se liší například o jeden nukleotid, bylo provedeno OTU shlukování sekvencí, které jsou si podobné na 99%. Tímto se tedy předpokládá, že i nepatrné odchylky v délkách sekvencí jsou pro zařazení sekvence do odpovídající ASV ignorovány.

Aby mohl být aplikován algoritmus *ChimeraDetector* (CHD_f99), jehož výpočetní náročnost je vyšší než metody algoritmu UCHIME, byly tyto rozšířené shluky ASV filtrovány na základě jejich abundance. Veškeré shluky s abundancí nižší než 10 byly odstraněny. Tímto byl zajištěn poměrně nízký počet shluků a algoritmus *ChimeraDetector* s i tak vysokou výpočetní náročností byl použit. Po stejném předzpracování dat byla aplikována detekce chimér metodou *de novo* algoritmu UCHIME (UD_f99).

Dalším prvkem pro srovnání je metoda *de novo* algoritmu UCHIME (UD_f97 aplikovaná na sekvence po OTU shlukování na 97% a následné filtraci málo abundantních OTU ve vzorku. Ve stejném bodě předzpracování dat byl aplikován poslední prvek srovnání, a to algoritmus *ChimeraDetector* (CHD_f97).

Veškeré srovnání těchto metod je uvedeno v tabulce č. 5.3.

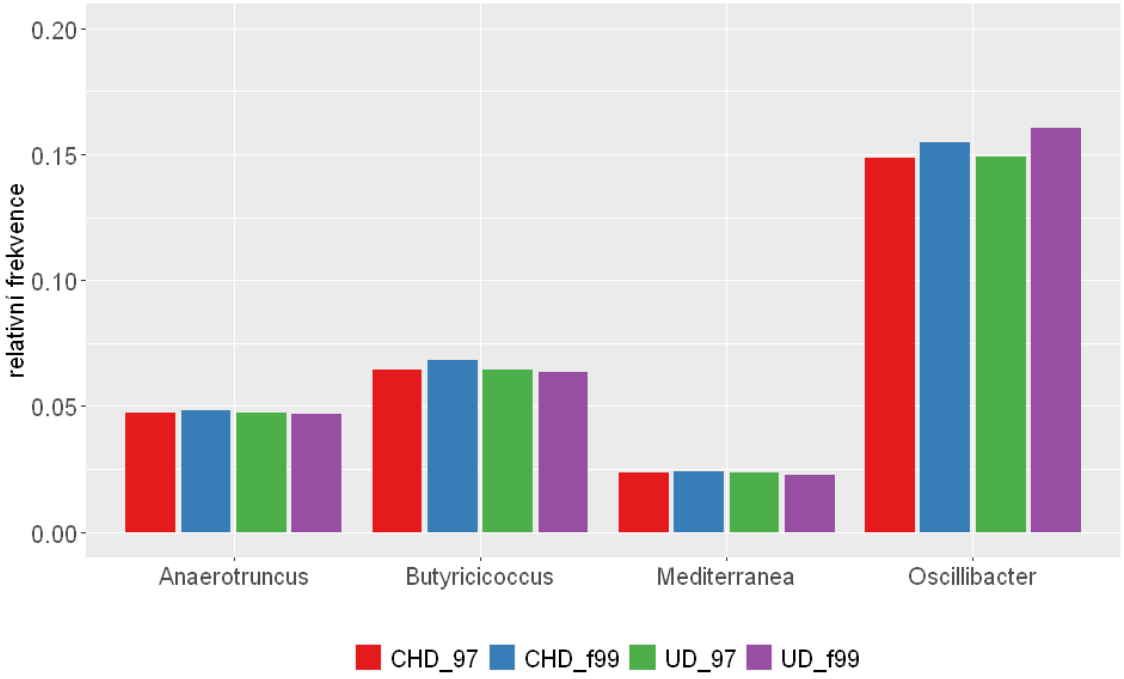
Tab. 5.3: Srovnání použitých *de novo* metod

	TN	**FP**	**Počet odstraněných chimérických sekvencí**	**Výsledný počet sekvencí**
CHD_f99	19	1	1937	43376
UD_f99	20	0	1714	43599
UD_f97	20	0	2049	54465
CHD_f97	20	0	396	55028

U metody CHDf99 pozorujeme vysoký počet odstraněných sekvencí, dokonce i falešně pozitivní detekci, tedy kdy nechimérická sekvence byla odstraněna spolu s chimérickými sekvencemi. Odstraněná sekvence poté odpovídala bakterii *Megasphera elsdenii*, která je v datatsetu zastoupena dvakrát s odlišnými formami genu 16S rRNA.

Znovu pro porovnání vlivu zařazení filtrace do procesu předzpracování dat byl vykreslen sloupcový graf uvedený na obrázku č. 5.5. Relativní frekvence jednotlivých bakterií odpovídají příslušným použitým metodám. Je patrný menší rozdíl v relativních frekvencích metod aplikovaných před OTU shlukováním do jednotlivých druhů. Především lze tento jev pozorovat u více abundantních sekvencí, zatímco

abundance méně zastoupených sekevencí jsou téměř shodné. I tak lze říci, že rozdíl mezi relativním zastoupením bakterií při různém zařazení filtrace chimér není jednoznačný.



Obr. 5.5: Sloupcový graf relativní frekvence bakterií po filtraci chimér

5.3 Výsledky použitých metod

I přes pádné důvody, proč se *de novo* metoda zařazuje před OTU shlukování, které byly v předešlé podkapitole uvedeny, tedy že *de novo* metoda detekce by měla být aplikována po dereplikaci sekvencí o stejných délkách, aby byly odhaleny veškeré chimérické sekvence, byl daný algoritmus *ChimeraDetector* především aplikován a otestován na datech, u kterých již bylo provedeno OTU shlukování na 97% podobnosti. Jednalo se tedy o shlukování sekvencí dle druhů. Následně byla provedena filtrace na základě frekvence jednotlivých OTU, kdy byly odstraněny OTU s frekvencí nižší než 10. Důvodem této úpravy dat bylo snížení výpočetního času algoritmu, který byl i tak vysoký. Je zřejmé, že nebyly odhaleny chiméry, které byly v kroku shlukování zařazeny do OTU, ale pouze ty chiméry, které byly natolik odlišné a abundantní, aby vytvořily samostatné OTU.

Zároveň byl stejný formát dat filtrován algoritmem UCHIME jak *de novo* metodou, tak referenční metodou za použití databáze GreenGenes 99.

Základní statistiky používané při hodnocení klasifikátorů, mezi které můžeme řadit i detekci chimér, jsou uvedeny v tabulce č. 5.4. Algoritmus *ChimeraDetector* (CHD), *de novo* metoda algoritmu UCHIME (UD) a referenční metoda (UR) byly použity na zbylé datasety, tedy *p2*, resp. *p3*.

Tab. 5.4: Základní statistiky při vyhodnocování klasifikátorů

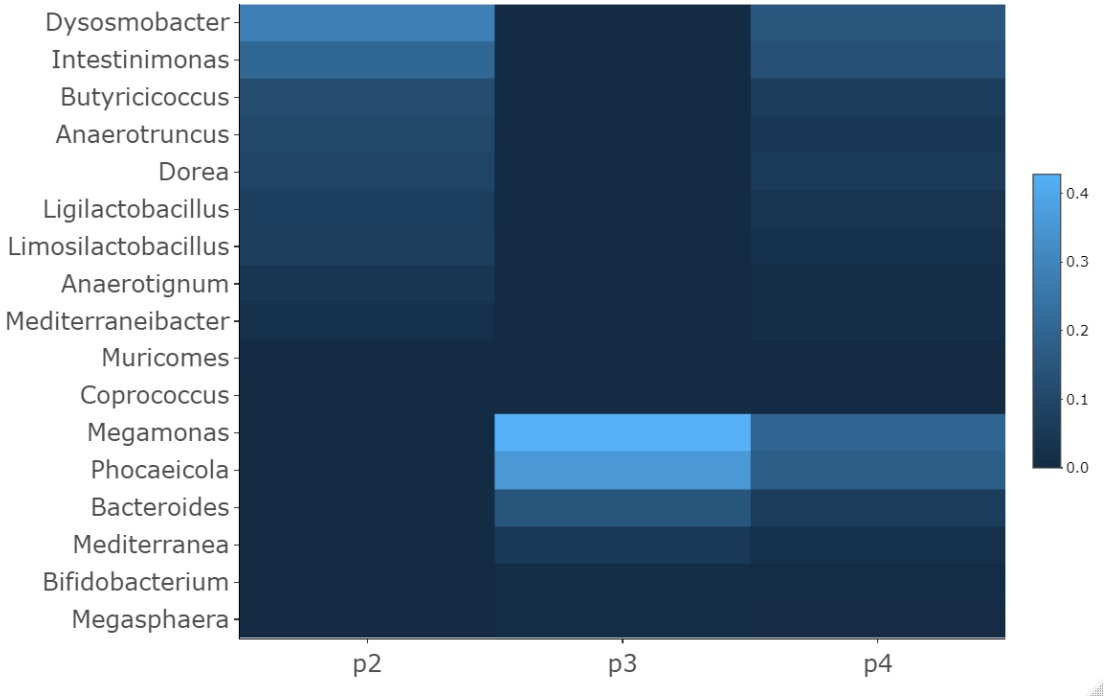
Metoda	**Se**	**Sp**	**Acc**	**Err**	**PPV**	**NPV**	**F_score**
p2_CHD	0.744	1	0.761	0.239	1	0.220	0.853
p3_CHD	0.748	1	0.770	0.230	1	0.282	0.856
p2_UD	0.976	1	0.978	0.022	1	0.750	0.988
p3_UD	0.982	1	0.984	0.016	1	0.846	0.991
p2_UR	0.952	1	0.955	0.045	1	0.600	0.975
p3_UR	0.802	1	0.820	0.180	1	0.333	0.890

Lze vyčíst nejvyšší úspěšnost metody *de novo* algoritmu UCHIME, která v obou datasetech odhalila největší počet chimér. Naopak detekce referenční metody se u obou datasetů výrazně liší. Je zřejmé, že z nějakého důvodu měla tato metoda problém s datasetem *p3*, obsahující gram-negativní bakterie. Důvodem pak může být například špatné zastoupení těchto bakterií v databázi. U obou metod je však zaznamenána stoprocentní pozitivní prediktivita, tedy že nebyl zaznamenán falešně pozitivní případ, který by nechimérickou sekvenci označil za chimérickou. Nejvíce nedetekovaných chimér pak zaznamenal algoritmus *ChimeraDetector*, jehož detekce není stoprocentní z několika důvodů. Především je algoritmus schopný detekce těch chimér, které vznikly spojením pouze dvou předků, a to v oblasti nalezeného konzervovaného regionu. Dalším omezením je výpočetní náročnost algoritmu, kdy výpočet uvedených datasetů trval podstatně déle než metody algoritmu UCHIME.

Za kladné vlastnosti algoritmu *ChimeraDetector* pak lze považovat jeho specificitu, kdy nebyla zaznamenána falešná pozitivita. Tím pádem je nejen specificita, ale i pozitivní prediktivita algoritmu rovna jedné. Za zmínku poté stojí i fakt, že ve většině případů určil algoritmus předky detekovaných chimér stejně jako *de novo* metoda algoritmu UCHIME. Statistiky proběhlé metody *de novo* algoritmu UCHIME a algoritmu *ChimeraDetector*, které obsahují mimo jiné i informaci o předcích chimér, jsou dostupné v elektronické příloze v tabulce 5.2.

Pro porovnání zastoupení jednotlivých rodů bakterií po filtraci chimér algoritmem *ChimeraDetector* se zastoupením bakterií před filtrací, které bylo znázorněno v předešlé kapitole, byla vykreslena heatmapa (obrázek č. 5.6) dostupných vyfiltrovaných datasetů. Můžeme tak pozorovat především snížení počtu rodů bakterií, které

byly tvořeny především chimérickými sekvencemi. Relativní zastoupení jednotlivých nechimérický druhů je poté srovnatelné s relativním zastoupením před filtrací.



Obr. 5.6: Heatmapa rodů bakterií po odstranění chimér

Závěr

Cílem bakalářské práce bylo vypracování literární rešerše týkající se problematiky chimérických sekvencí vyskytujících se v sekvenačních datech a následná realizace algoritmu v jazyce R pro jejich detekci a následnou filtraci. Na úvod byly vysvětleny pojmy týkající se práce s mikrobiálními daty a objasnění rozdílů mezi sekvenací dat pro metagenomickou, resp. mikrobiální analýzu. Dále byly uvedeny generace sekvenačních technologií a způsoby generování čtení.

Následující kapitola se zaměřila především na objasnění postupu amplifikace vzorku a její vliv na analýzu dat. Byly zde popsány různé artefakty vyskytující se v datech a důraz se kladl především na pochopení vzniku chimérických sekvencí. Dále zde byly uvedeny metody pro odstranění těchto artefaktů.

V třetí kapitole byly na úvod zmíněny základní informace o testovacích datasetech poskytnutých Výzkumným ústavem veterinárního lékařství v Brně. Dále byly uvedeny základní kroky, které se používají při předzpracování rozsáhlých sekvenačních dat, a byla popsána aplikace těchto kroků na vlastních datasetech.

Následně byla uvedena problematika operačních taxonomických jednotek, metody používané při shlukování OTU a následná kvantifikace tohoto kroku. Také zde byla uvedena klasifikace organismů klasifikátorem RDP a bylo zde znázorněno zastoupení jednotlivých bakterií v datasetech.

Poslední kapitola se pak zabývala samotnou realizací detektoru chimér v jazyce R. Kromě popisu fungování algoritmu zde byly uvedeny výsledky detekce na vlastních datasetech ve srovnání s dostupnými metodami pro detekci a filtraci chimérických sekvencí.

Jak výsledky napovídají, využití algoritmu *ChimeraDetector* v aktuální verzi by v konkurenci dostupných algoritmů pro detekci chimér nemělo velký úspěch. Vylepšení aktuální verze algoritmu by však mohlo vést k jeho uplatnění při zpracování mikrobiálních dat Výzkumným ústavem veterinárního lékařství v Brně. Příkladem takového vylepšení pak může být rozšíření o detekci tzv. multimér či snížení výpočetní náročnosti zakomponováním algoritmu BLAST pro nalezení předků chimér.

Literatura

[1] *GODOY-VITORINO, Filipa. Human microbial ecology and the rising new medicine. Annals of Translational Medicine. AME Publishing Company, 2019, roč. 7, č. 14, s. 342–342. ISSN 23055839. DOI: 10.21037/atm.2019.06.56*

[2] *THOMAS, Torsten, Jack GILBERT a Folker MEYER. Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation. Springer Science and Business Media LLC, 2012, roč. 2, č. 1, s. 3. ISSN 2042-5783. DOI: 10.1186/2042-5783-2-3*

[3] *WEINSTOCK, George M. Genomic approaches to studying the human microbiota. Nature. 2012, 489(7415), 250-256. ISSN 0028-0836. DOI:10.1038/nature11553*

[4] *MARCHESI, Julian R. a Jacques RAVEL. The vocabulary of microbiome research: a proposal. Microbiome. Springer Nature, 2015, roč. 3, č. 1. ISSN 2049-2618. DOI: 10.1186/s40168-015-0094-5*

[5] *PETERSON, Daniel A., Daniel N. FRANK, Norman R. PACE a Jeffrey I. GORDON. Metagenomic Approaches for Defining the Pathogenesis of Inflammatory Bowel Diseases. Cell Host Microbe. 2008, 3(6), 417-427. ISSN 19313128. DOI:10.1016/j.chom.2008.05.001*

[6] *NICHOLSON, Jeremy K., John C. LINDON a Elaine HOLMES. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 2008, 29(11), 1181-1189. ISSN 0049-8254. DOI:10.1080/004982599238047*

[7] *AGUIAR-PULIDO, Vanessa et al. Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis. Evolutionary Bioinformatics. Libertas Academica Ltd., 2016, roč. 12s1, s. EBO.S36436. ISSN 1176-9343. DOI: 10.4137/EBO.S36436*

[8] *SRINIVASAN, Ramya et al. Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. PLoS ONE. Public Library of Science, 2015, roč. 10, č. 2, s. 117617. ISSN 19326203. DOI: 10.1371/journal.pone.0117617*

[9] *CLARRIDGE, Jill E. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews. 2004, 17(4), 840-862. ISSN 0893-8512. DOI:10.1128/CMR.17.4.840-862.2004*

[10] *SNYDER, Anna K., Kenneth Z. ADKINS a Rita V.M. RIO. Use of the internal transcribed spacer (ITS) regions to examine symbiont divergence and as a diagnostic tool for Sodalis-related bacteria. Insects. Multidisciplinary Digital Publishing Institute (MDPI), 2011, roč. 2, č. 4, s. 515–531. ISSN 20754450. DOI: 10.3390/insects2040515*

[11] *Microbiome Analysis – Are You Aware Of All The Applications? - Eurofins Genomics BLOG. Dostupné z: https://www.genomics-expert.com/2019/05/29/microbiome-analysis-are-you-aware-of-all-the-applications/*

[12] *MYERS JR, Eugene W. A history of DNA sequence assembly. it - Information Technology. Walter de Gruyter GmbH, 2016, roč. 58, č. 3. ISSN 1611-2776. DOI: 10.1515/itit-2015-0047*

[13] *COLLINS, Francis S. The Human Genome Project: Lessons from Large-Scale Biology. Science. 300(5617), 286-290. ISSN 00368075. DOI:10.1126/science.1084564*

[14] *SANGER, Frederick a Alan R. COULSON. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology. Academic Press, 1975, roč. 94, č. 3, s. 441–448. ISSN 00222836. DOI: 10.1016/0022-2836(75)90213-2*

[15] *MAXAM, Allan M. a Walter GILBERT. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. National Academy of Sciences, 1977, roč. 74, č. 2, s. 560–564. ISSN 00278424. DOI: 10.1073/pnas.74.2.560*

[16] *HEATHER, James M. a Benjamin CHAIN. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016, 107(1), 1-8. ISSN 08887543. DOI:10.1016/j.ygeno.2015.11.003*

[17] *SHUIKAN, Ahmed, Sulaiman ALI ALHARBI, Dalal HUSSIEN M. ALKHALIFAH a Wael N. HOZZEIN. High-Throughput Sequencing and Metagenomic Data Analysis. N. HOZZEIN, Wael, ed. Metagenomics - Basics, Methods and Applications. IntechOpen, 2020, 2020-3-25. ISBN 978-1-83880-055-0. DOI:10.5772/intechopen.89944*

[18] *KCHOUK, Mehdi, Jean Francois GIBRAT a Mourad ELLOUMI. Generations of Sequencing Technologies: From First to Next Generation. Biology and Medicine. 2017, 09(03). ISSN 09748369. DOI:10.4172/0974-8369.1000395*

[19] *Roche Life Science | Welcome. Dostupné z: https://lifescience.roche.com/en$_c$z.html*

[20] *SIQUEIRA, José F., Ashraf F. FOUAD a Isabela N. RÔÇAS. Pyrosequencing as a tool for better understanding of human microbiomes. Journal of Oral Microbiology. 2012, 4(1). ISSN 2000-2297. DOI:10.3402/jom.v4i0.10743*

[21] *GHARIZADEH, Baback, Mehran GHADERI a Pål NYRÉN. Pyrosequencing Technology for Short DNA Sequencing and Whole Genome Sequencing. Seibutsu Butsuri. Biophysical Society of Japan, 2007, roč. 47, č. 2, s. 129–132. ISSN 0582-4052. DOI: 10.2142/biophys.47.129*

[22] *LIU, Lin, Yinhu LI, Siliang LI, Ni HU, Yimin HE, Ray PONG, Danni LIN, Lihua LU a Maggie LAW. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology. 2012, 2012, 1-11. ISSN 1110-7243. DOI:10.1155/2012/251364*

[23] *SLATKO, Barton E., Andrew F. GARDNER a Frederick M. AUSUBEL. Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology. Blackwell Publishing Inc., 2018, roč. 122, č. 1, s. e59. ISSN 19343647. DOI: 10.1002/cpmb.59*

[24] *SCHADT, Eric E., Steve TURNER a Andrew KASARSKIS. A window into third-generation sequencing. Human Molecular Genetics. Oxford Academic, 2010, roč. 19, č. R2, s. R227–R240. ISSN 09646906. DOI: 10.1093/hmg/ddq416*

[25] *EID, John et al. Real-time DNA sequencing from single polymerase molecules. Science. Harvard Univ. Press, 2009, roč. 323, č. 5910, s. 133–138. ISSN 00368075. DOI: 10.1126/science.1162986*

[26] *Oxford Nanopore Technologies. In: [cit. 04.01.2021]. Dostupné z: https://nanoporetech.com/*

[27] *FREEDMAN, Adam H., John M. GASPAR a Timothy B. SACKTON, 2020. Short paired-end reads trump long single-end reads for expression analysis. BMC Bioinformatics. 21(1), 1–11. ISSN 14712105. DOI:10.1186/s12859-020-3484-z*

[28] *WU, Gary D. et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiology. BioMed Central, 2010, roč. 10, s. 206. ISSN 14712180. DOI: 10.1186/1471-2180-10-206*

[29] *GOODRICH, Julia K., Sara C. DI RIENZI, Angela C. POOLE, Omry KOREN, William A. WALTERS, J. Gregory CAPORASO, Rob KNIGHT a Ruth E. LEY. Conducting a Microbiome Study. Cell. 2014, 158(2), 250-262. ISSN 00928674. DOI:10.1016/j.cell.2014.06.037*

[30] *GUPTA, Nalini. DNA extraction and polymerase chain reaction. Journal of Cytology. Wolters Kluwer Medknow Publications, 2019, roč. 36, č. 2, s. 116. ISSN 0970-9371. DOI: 10.4103/JOC.JOC11018*

[31] *AMILS, Ricardo. Taq Polymerase. GARGAUD, Muriel, Ricardo AMILS, José Cernicharo QUINTANILLA, Henderson James CLEAVES, William M. IRVINE, Daniele L. PINTI a Michel VISO, ed. Encyclopedia of Astrobiology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, 2011, s. 1648-1648. ISBN 978-3-642-11271-3. DOI:10.1007/978-3-642-11274-41561*

[32] *ACINAS, Silvia G. et al. PCR-induced sequence artifacts and bias: Insights from comparison of two 16s rRNA clone libraries constructed from the same sample. Applied and Environmental Microbiology. American Society for Microbiology (ASM), 2005, roč. 71, č. 12, s. 8966–8969. ISSN 00992240. DOI: 10.1128/AEM.71.12.8966-8969.2005*

[33] *KANAGAWA, Takahiro. Bias and artifacts in multitemplate polymerase chain reactions (PCR). Journal of Bioscience and Bioengineering. Elsevier BV, 2003, roč. 96, č. 4, s. 317–323. ISSN 13891723. DOI: 10.1016/s1389-1723(03)90130-7*

[34] *LINHART, Chaim a Ron SHAMIR. The Degenerate Primer Design Problem: Theory and Applications. Journal of Computational Biology. Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA , 2005, roč. 12, č. 4, s. 431–456. ISSN 1066-5277. DOI: 10.1089/cmb.2005.12.431*

[35] *THOMPSON, Janelle R., Luisa A. MARCELINO a Martin F. POLZ. Heteroduplexes in mixed-template amplifications: Formation, consequence and elimination by „reconditioning PCR". Nucleic Acids Research. Oxford University Press, 2002, roč. 30, č. 9, s. 2083–2088. ISSN 03051048. DOI: 10.1093/nar/30.9.2083*

[36] *QUINCE, Christopher et al. Removing Noise From Pyrosequenced Amplicons. BMC Bioinformatics. BioMed Central, 2011, roč. 12, č. 1, s. 38. ISSN 14712105. DOI: 10.1186/1471-2105-12-38*

[37] *ODELBERG, Shannon J. et al. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Research. Oxford University Press, 1995, roč. 23, č. 11, s. 2049–2057. ISSN 03051048. DOI: 10.1093/nar/23.11.2049*

[38] *HAAS, Brian J. et al. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research. Cold Spring Harbor Laboratory Press, 2011, roč. 21, č. 3, s. 494–504. ISSN 10889051. DOI: 10.1101/gr.112730.110*

[39] *YOKONO, Makio, Soichirou SATOH a Ayumi TANAKA. Comparative analyses of whole-genome protein sequences from multiple organisms. Scientific Reports. Nature Publishing Group, 2018, roč. 8, č. 1. ISSN 20452322. DOI: 10.1038/s41598-018-25090-8*

[40] *QUAST, Christian et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research. Oxford Academic, 2013, roč. 41, č. D1, s. D590–D596. ISSN 03051048. DOI: 10.1093/nar/gks1219*

[41] *O'LEARY, Nuala A. et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. Oxford University Press, 2016, roč. 44, č. D1, s. D733–D745. ISSN 13624962. DOI: 10.1093/nar/gkv1189*

[42] *COLE, James. R. et al. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. Nucleic Acids Research. Oxford University Press, 2009, roč. 37, č. SUPPL. 1, s. D141. ISSN 03051048. DOI: 10.1093/nar/gkn879*

[43] *EDGAR, Robert C. et al. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. Oxford University Press, 2011, roč. 27, č. 16, s. 2194–2200. ISSN 14602059. DOI: 10.1093/bioinformatics/btr381*

[44] *CAPORASO, J Gregory, Justin KUCZYNSKI, Jesse STOMBAUGH, Kyle BITTINGER, Frederic D BUSHMAN, Elizabeth K COSTELLO, Noah FIERER, Antonio Gonzalez PEÑA, Julia K GOODRICH, et al. QIIME allows analysis of high-throughput community sequencing data. Nature Methods. 2010, 7(5), 335-336. ISSN 1548-7091. DOI:10.1038/nmeth.f.303*

[45] *ROSEN, Michael J. et al. Denoising PCR-amplified metagenome data. BMC Bioinformatics. BioMed Central, 2012, roč. 13, č. 1, s. 283. ISSN 14712105. DOI: 10.1186/1471-2105-13-283*

[46] *EDGAR, Robert C. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nature Methods. Nat Methods, 2013, roč. 10, č. 10, s. 996–998. ISSN 15487091. DOI: 10.1038/nmeth.2604*

[47] *MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011, 17(1), 10-12. ISSN 2226-6089. DOI:10.14806/ej.17.1.200*

[48] *BOLGER, Anthony M., Marc LOHSE a Bjoern USADEL, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15), 2114-2120. ISSN 1460-2059. DOI:10.1093/bioinformatics/btu170*

[49] *AFGAN, Enis, Dannon BAKER, Marius VAN DEN BEEK, et al., 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research. 44(W1), W3-W10. ISSN 0305-1048. DOI:10.1093/nar/gkw343*

[50] *EDGAR, Robert C. a Henrik FLYVBJERG, 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics. 31(21), 3476-3482. ISSN 1367-4803. DOI:10.1093/bioinformatics/btv401*

[51] *RIDEOUT, Jai Ram et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. PeerJ Inc., 2014, roč. 2014, č. 1, s. e545. ISSN 21678359. DOI: 10.7717/peerj.545*

[52] *WANG, Qiong, George M. GARRITY, James M. TIEDJE a James R. COLE, 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology. 73(16), 5261–5267. ISSN 00992240. DOI:10.1128/AEM.00062-07*

[53] *BODENHOFER, Ulrich, Enrico BONATESTA, Christoph HOREJŠ-KAINRATH a Sepp HOCHREITER, 2015. Msa: An R package for multiple sequence alignment. Bioinformatics. 31(24), 3997–3999. ISSN 14602059. DOI:10.1093/bioinformatics/btv494*

[54] *GARCIA-MARTIN, Juan Antonio, Peter CLOTE a Emanuele PACI, 2015. RNA Thermodynamic Structural Entropy. PLOS ONE. 10(11). ISSN 1932-6203. DOI:10.1371/journal.pone.0137859*

Seznam zkratek

ASCII	americký standardní kód pro výměnu informací
ASV	varianty amplikonových sekvencí
bp	base pair (jednotka délky sekvence DNA či RNA)
ddNTP	dideoxyribonukleotidtrifosfát
dNTP	deoxyribonukleotidtrifosfát
HGP	Projekt lidského genomu
ITS	Internal transcribed spacer (část DNA)
NCBI	National Center of Biotechnology Information
NGS	Next-generation sekvenování
OTU	operační taxonomická jednotka
PacBio	Pacific Biosciences (společnost pro výrobu sekvenačních technologií)
PCR	polymerázová řetězová reakce
PE	paired-end (způsob generování čtení)
RDP	Ribosomal Database Project (databáze sekvencí ribozomálních RNA)
RefSeq	Reference Sequence (databáze nukleotidových i proteinových sekvencí)
rRNA	ribozomální RNA
SE	single-end (způsob generování čtení)

Seznam příloh

Příloha 1

Doplňující tabulky

Příloha 2

Rozšířené grafy

Příloha 3

Vzorové výsledky algoritmu *ChimeraDetector*

Příloha 4

Zdrojový kód algoritmu *ChimeraDetector* se vzorovými daty pro ověření funkčnosti