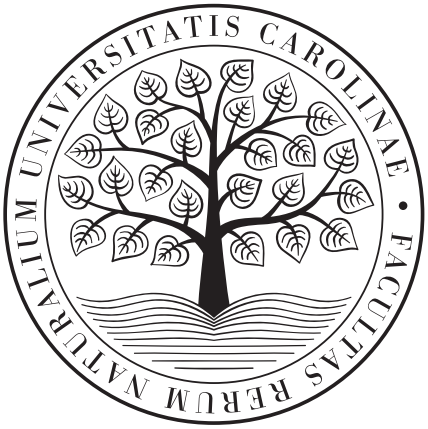
Univerzita Karlova
Přírodovědecká fakulta

BAKALÁŘSKÁ PRÁCE



Michal Vosyka

Využití sekvenačních metod pro studium mikrobiálních komunit: (meta)genomové assembly a populační genomika u bakterií

Katedra buněčné biologie (1510)

Vedoucí bakalářské práce: Mgr. Jakub Rídl, Ph.D.
Studijní program: Bioinformatika (B0688A140003)
Studijní obor: B-BINF (0688RA140003)

Praha 2024

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Poděkování. Mé díky patří především mému školiteli, doktoru Jakubu Rídlovi, za jeho odborný dohled a spolupráci. Dále děkuji své rodině za jejich podporu, bez níž by psaní práce bylo daleko obtížnější. Navíc děkuji všem ohleduplným lidem, kteří mě během společné cesty vlakem nerušili, a děkuji provozovatelům webových aplikací přehrávajících hudbu za jasné chvíle plné soustředění.

Název práce: Využití sekvenačních metod pro studium mikrobiálních komunit: (meta)genomové assembly a populační genomika u bakterií

Autor: Michal Vosyka

Katedra: Katedra buněčné biologie (1510)

Vedoucí bakalářské práce: Mgr. Jakub Rídl, Ph.D., Katedra zoologie (1700)

Abstrakt: Sekvenace DNA z mikrobiálních komunit umožňuje kromě taxonomické profilace přítomných mikrobiálních druhů také studium vnitropopulační genetické variability. Její popis je výsledkem informatické analýzy sekvenačních dat. Tato práce zkoumá, jaké bioinformatické nástroje jsou k dispozici pro identifikaci vnitropopulační variability z metagenomických dat *de novo* a jaké jsou algoritmické principy jejich fungování. Poskytuje perspektivu pro hodnocení správnosti výsledků, a začíná proto představením sekvenačních metod platforem Illumina, PacBio a Oxford Nanopore Technologies, včetně jejich limitací, a pokračuje popisem výpočetní rekonstrukce sekvencí genomů. Kromě představení nástrojů a benchmarků přináší pokus o konceptuální shrnutí různých přístupů studia variability z metagenomických dat.

Klíčová slova: Sekvenační metody Metagenomika Assembly Populační genomika Mikrobiální společenstva

Title: The new sequencing methods in metagenomics: (meta)genome assembly and population genomics in bacteria

Author: Michal Vosyka

Department: Department of Cell Biology (1510)

Supervisor: Mgr. Jakub Rídl, Ph.D., Department of Zoology (1700)

Abstract: Sequencing of DNA from microbial communities enables, besides taxonomic profilling, study of intrapopulation genetic variability. Its description is a result of a computational analysis of sequencing data. This thesis investigates what bioinformatics tools are available for *de novo* detection of intrapopulation variability in metagenomic data and how these tools function algorithmically. It provides a perspective for tool performance validation, and for that, it begins by discussing sequencing methods of platforms Illumina, PacBio and Oxford Nanopore Technologies, aiming at their limitations, and it continues with describing computational reconstruction of genomic sequences. Beyond the review of tools and benchmarks, it attempts to provide conceptual view on approaches to the study of variability based on metagenomic data.

Keywords: Sequencing methods Metagenomics Assembly Population genomics Microbial communities

Obsah

Úvod

Sekvenační metody přinesly revoluci do výzkumu mikroorganizmů. Nejprve byla revoluční možnost *amplikonového sekvenování*, které stanovuje (takzvaně čte) sekvenci specifického úseku genomu, tzv. markeru. Od začátku se jednalo zejména o sekvenci genu pro 16S rRNA, která je součástí malé podjednotky bakteriálního ribozomu. Gen pro 16S rRNA má silně evolučně konzervovanou sekvenci, což umožňuje srovnání i velmi evolučně vzdálených mikrobiálních genomů, fylogenetickou analýzu a přiřazení konkrétních sekvencí daným taxonům [1]. Navíc umožňuje získat taxonomický profil společenstva mikroorganizmů ve vzorku ze studovaného prostředí [2]. Analýza nukleových kyselin získaných z environmentálního vzorku a pocházejících z mikrobiální komunity (v kontrastu k analýze klonálních laboratorních kultur) dostala název metagenomika [3]. Význam metagenomiky si lépe uvědomíme díky tvrzení, že pouze zlomek mikrobiální rozmanitosti byl dosud kultivován v laboratoři a mnohé mikroby ani nejsme prakticky schopni kultivovat [4, 5]. Později se v souladu s rozvojem sekvenování nové generace začal uplatňovat přístup tzv. *shotgun* sekvenování [2], které se snaží do sekvenační knihovny nestranně zařadit fragmenty z veškeré DNA vzorku. Sekvenaci blíže popíšeme v kapitole 1. *Shotgun* sekvenování přináší např. možnost zkoumat teoreticky všechny přítomné geny nebo sestavit sekvence z celých genomů přítomných druhů *in-silico*, pomocí algoritmů pojednaných v kapitole 2. Sekvenci genomu přímo sestaveného z metagenomických čtení označíme kvůli stručnosti novým zkratkovým slovem MAG [1].

Mikrobiologie uznává vyčlenění bakteriálního druhu na základě kombinace fenotypových a genotypových rozlišujících znaků [4]. Přibývají důkazy, že genomy bakteriálních druhů jsou koherentní: Olm et al. [6] ukázali na základě MAGů z velkého množství environmentálních vzorků, že se sekvence genomů klastrují do skupin odpovídajících druhům, které mají vzájemně větší sekvenční identitu, než je tomu mezi skupinami. Nejlepší se jeví položit 95% jako hraniční hodnotu sekvenční identity mezi sekvencemi celých genomů a identičtější genomy zařadit do stejného druhu.

Zároveň se jiné studie zaměřují na těch několik procent divergence uvnitř druhu. Odlišnost sekvencí genomu se vyskytuje i v rámci jedné populace, jednoho vzorku bakterií [7]. Často se předpokládá, že tato vnitropopulační variabilita má

[1] Na základě anglické zkratky MAG zastupující *Metagenome Assembled Genome*.

podobu koexistence více klonálních kmenů [2], sestávájících z buněk téměř identických genomů. O vnitropopulační variabilitě můžeme uvažovat, aniž bychom předpokládali kmenové složení populace: Máme jednoduše množinu sekvencí genomů populace a zajímá nás, jakých konkrétních podob nabývá daný lokus genomu, jaké jeho varianty (jednonukleotidové či rozsáhlejší varianty, viz sekci 3.1) existují v množině genomů. Tam, kde budeme chtít asociovat varianty různých lokusů na základě toho, že se spolu vyskytují ve stejných genomech, budeme kromě výrazu kmen používat též *haplotyp*. Haplotyp, známý pojem z genetiky, se v kontextu eukaryotických (diploidních) organizmů používá pro označení alel fyzicky přítomných na tomtéž chromozomu. V dnešní době sekvenování genomů se ujal pojem *haplotypování*, nebo též *fázování*, pro proces rekonstrukce sekvence homologních chromozomů, jejichž heterozygotní sekvenční varianty mohly být detekovány v sekvenačních datech [8]. My zde přejmeme pojem haplotypování pro podobný úkol, který budeme řešit, pro rekonstrukci haplotypů zastupujících převládající kmeny v mikrobiální populaci (takto použito v [9]).

Mezi kroky pro zjištění vnitropopulační variability ze sekvenačních dat patří vedle zpracování vzorku a sekvenace ještě *in-silico* kroky předzpracování čtení, sestavení genomů, zarovnání sekvencí a detekce variant. Předkládaný text se snaží především popsat bioinformatické metody pro detekci variant zaměřené na data z mikrobiálních komunit. Vrátíme se k nim v kapitole 3. Jako technologicky vyzývavým a důležitým pro pochopení jiných kroků analýzy se budeme věnovat krokům sekvenace a sestavení genomů. Ačkoli ani mapování sekvenačních dat na referenční genomy komunit není triviální problém a je kritický pro následnou detekci variant, my jej zde budeme do jisté míry považovat za černou skříňku. V dalších kapitolách postupně popíšeme vytyčené kroky analýzy a u všech zdůrazníme hledisko přesnosti a správnosti a jeho kritický význam pro kroky následující.

[2] Kmen ve významu anglického *strain*. Kmen ve významu vyšší taxonomické jednotky *phylum* (z latiny) zde nebude nikde použit.

Kapitola 1

Sekvenační metody

Samotné sekvenaci předchází tvorba knihovny DNA molekul pečlivě připravených pro sekvenační reakci. Příprava DNA knihovny ze vzorku pro *shotgun* metagenomové sekvenování se liší podle vzorku a metody sekvenování. Prvním krokem typicky bývá získání DNA ze vzorku, s použitím lyze buněk (DNA se uvolní do roztoku) a extrakce DNA. Dále lze zajistit správnou délku DNA fragmentů, pomocí fragmentačních postupů a separace (např. gelovou elektroforézou) [10]. V této fragmentované podobě se objeví původní molekuly DNA v knihovně a nazýváme je inzerty. Inzerty jsou totiž typicky na koncích prodlouženy o specifické oligonukleotidy, tzv. adaptéry, které mohou plnit úlohu při nasedání primerů pro amplifikaci, úlohu identifikátorů, či asociovat s oligonukleotidy upevněnými na podkladu (metoda Illumina) [11, 12, 13]. Amplifikace, která bývá provedena v některých případech, se zakládá na technice PCR.

I příprava knihovny ovlivňuje validitu výsledků. Izolace DNA z obtížně lyzovatelných buněk může být snížená nebo její získání může být na úkor větší fragmentace [14, 15]. Chyby vznikají v případě amplifikace před sekvenací v důsledku nepřesnosti polymerázy a tyto chyby se během cyklů mohou někdy amplifikovat [14].

Sekvenační reakce používá chemické a fyzikální děje k tvorbě měřitelných signálů, z nichž lze odvodit (vyvolat [1]) sekvence bází v molekulách. Sekvenci vyvolaných bází nazýváme čtení [2]. Nejistota při určení báze může být kvantifikována do pozičně specifické pravděpodobnosti chyby, takže ideálně by velký počet bází s touž pravděpodobností chyby p obsahoval chybné báze s relativní četností p [16]. Místo p se udává transformovaná hodnota nazývaná *Phred* skóre, definované vztahem $q = -10 \cdot log_{10}(p)$, kde q je Phred skóre [16]. Pokrytí sekvence čteními můžeme definovat v kontextu studované sekvence (např. konsenzuální genom populace *Lactobacillus delbrueckii* subsp. *bulgaricus* ve vzorku jogurtu) jako zarovnání čtení na tuto sekvenci. O hloubce pokrytí dané pozice sekvence hovoříme ve významu počtu čtení s bází zarovnanou na tuto pozici. Průměrná

[1] Zavedeme zde vyvolávání bází jako překlad z angl. *base calling*.

[2] V angl. nazýváno *read*.

hloubka pokrytí sekvence (často jenom jako hloubka pokrytí) je průměr hloubek pokrytí přes její pozice. Šířka pokrytí znamená počet pozic sekvence s nenulovou hloubkou pokrytí.

Sekvenování cyklickou reverzibilní terminací

Sekvenování syntézou či přesněji řečeno sekvenování cyklickou reverzibilní terminací [11] je metoda uplatňovaná v produktech (sekvenátorech) společnosti Illumina, např. v zařízeních HiSeq [17], MiSeq, NextSeq a NovaSeq [18]. Sekvenační reakce se odehrává na povrchu destičky, tzv. *flow cell*, na níž jsou fixované oligonukleotidy komplementární k adaptérům [19, 13]. Destička je přelita roztokem jednovláknové DNA knihovny tak, aby fragmenty rozptýlené po destičce nasedly Watson-Crickovým párováním na fixní oligonukleotidy [19]. Po přidání reagencií DNA polymerázová reakce prodlouží fixní oligonukleotid do podoby molekuly komplementární k fragmentu [13]. Získáváme fragmenty s fixovaným koncem, jednotlivě rozeseté po destičce. Následuje amplifikace [19], s cílem zesílit signál sekvenační reakce. Volný konec svým adaptérem nasedá na komplementární fixní oligonukleotid poblíž a umožňuje jeho prodloužení [19, 13]. Opakování tohoto postupu vede k rozmnožení fixovaných kopií fragmentu kolem jeho původního umístění [19]. Kopie komplementární k původní jsou odštěpeny [13].

Samotná cyklická reverzibilní terminace používá nukleotid trifosfátů vázaných na skupinu zabraňující prodloužení DNA po tom, co je nukleotid inkorporován [11, 13]. Navíc je navázána na fluorescenční molekulu (fluorofor) s emisním spektrem specifickým pro každou bázi [13]. Cyklicky je opakován proces inkorporace jednoho dalšího modifikovaného nukleotidu komplementárního k templátu do rostoucího řetězce, odmytí reagencií, snímání signálu inkorporované báze, odstranění blokující skupiny a fluoroforu a jejich odmytí [11].

Takto je sekvenátor Illuminy schopen přečíst sekvence dlouhé nanejvýš několik set bází. Sekvenátor zvládá přečíst sekvenci z konce jednoho vlákna, následně podle něj na fixních oligonukleotidech nasyntetizovat komplementární vlákno, původní vlákna odštěpit a zopakovat sekvenační reakci pro přečtení druhého vlákna [13]. Takto je na obou vláknech přečteno až několik set bází z jejich $5'$ konce, což znamená přečtení sekvence na obou koncích inzertu [13]. Tato tzv. párová čtení umožňují kromě zdvojnásobení přečtených bází i získat informaci o molekule na větší velikostní škále [19]. Lze vytvořit knihovny s přibližně určenou délkou inzertu, která poskytuje informaci o vzdálenosti čtení stejného páru v sekvenci čtené molekuly. Pro delší inzerty je potřeba připravit *mate-pair* knihovnu, která v kratším inzertu drží konce z delšího fragmentu, čehož je docíleno cirkularizací delšího fragmentu a vystřižením úseku okolo místa ligace [13, 11].

Syntéza za přítomnosti všech druhů nukleotid trifosfátů minimalizuje šance špatně inkorporované báze [13, 20]. Reverzibilní znemožnění inkorporace více nukleotidů je považováno za výhodu minimalizující šanci kontextově specifických

chyb (např. chyba v počtu čtených bází v homopolymerním úseku [3]) [20]. Phred skóre bází dosahují nad 30 i 40, například v jednom testu zařízení MiSeq, resp. Hiseq, Phred skóre nad 30 mělo 89.7% bází, resp. 87.7% bází [20].

Single-molecule real-time sekvenování

Počátky *single-molecule real-time* (SMRT) metody sekvenování okolo roku 2000 jsou spjaté s pracemi dr. J. Korlacha a dr. S. Turnera, kteří se podílejí na firmě *Pacific Biosciences* (PacBio) poskytující sekvenátory pro SMRT sekvenování [21]. *Zero-mode waveguide* (ZMW) je reakční prohlubeň nanometrových rozměrů v metalické fólii pokrývající skleněnou destičku, která umožňuje efektivní excitaci fluoroforu a detekci emitovaných fotonů v malé oblasti u dna prohlubně [22]. DNA syntéza podle templátu katalyzovaná DNA polymerázou upevněnou na dně ZMW s použitím fluorescenčně značených nukleotid trifosfátů umožňuje pomocí detekce emisních signálů zachytit začlenění nukleotidu do rostoucího řetězce. Ve snímaném okolí polymerázy se volný značený nukleotid vyskytne typicky jen jednotlivě a krátce a lze jej odlišit od delšího charakteristického signálu začleňovaného nukleotidu [22]. Destičky k tomuto účelu jsou v produktech firmy PacBio dodávány jako tzv. *SMRT cells*. Jedna může obsahovat miliony ZMWs [23]. Použití vysoce procesivní polymerázy umožňuje sekvenovat jednotlivé molekuly v délce tisíců bází.

Tzv. HiFi sekvenace poskytovaná PacBio dosahuje prostřednictvím SMRT vysoce kvalitních dlouhých čtení [24]. Používá SMRTbell templátů [10], což jsou molekuly DNA vytvořené z dsDNA [4] inzertu napojením jejich konců na adaptérové vlásenky [12]. Vlásenky propojují komplementární vlákna inzertu do jedné kružnicové molekuly, po jejímž obvodu leží po řadě sekvence prvního vlákna, vlásenky, druhého vlákna a další vlásenky. S použitím primeru nasedajícího na vlásenku může polymeráza syntetizovat vlákno komplementární k molekule templátu a po obejití obvodu vytěsnit 5′ konec vznikající molekuly, a provést tak mnoho syntéz téže sekvence [12]. Zpracováním *in-silico* můžeme z tohoto opakovaného čtení téhož inzertu získat konsenzuální sekvenci s velmi velkou přesností, neboť pravděpodobnost stejné náhodné chyby ve více opakováních strmě klesá. Tento princip se nazývá *circular consensus sequencing* [12].

Nanopórové sekvenování

Nanopórové sekvenování, ve smyslu produktů společnosti *Oxford Nanopore Technologies* (ONT), není založeno na syntéze DNA. Sekvenace probíhá za použití molekulárního póru usazeného v membráně, kterým prochází sekvenovaná molekula DNA, tzn. jedno její vlákno [25]. Vlákno je protahováno (translokováno) skrze nanopór za pomoci enzymu a elektrického napětí [26] mezi opačnými povrchy membrány (DNA je záporně nabitá a migruje v elektrickém poli) [25]. Napětí

[3] Homopolymer je úsek DNA, v němž je na sebe navázáno několik stejných nukleotidů.
[4] Dvojvláknová DNA, z angl. *double-stranded DNA*.

vede k ustavení el. proudu iontů reakčního roztoku skrze pór spojující roztoky obou stran membrány, která je elektricky izolační. Translokace DNA interferuje s tokem iontů a citlivé snímání hodnot proudu je primárním signálem, který sekvenátor detekuje [25]. Konkrétní efekt DNA v póru je ovlivněn chemickou identitou bází v úseku DNA ve specifických místech póru [27]. Záznam o průběhu proudu v čase umožňuje vyvolat sekvenci bází, k čemuž je možné využít nejeden software. Používané softwary se zakládají na metodách strojového učení a použití neuronových sítí. Výpočty některých programů mohou být natolik rychlé, že umožňují průběžnou rekonstrukci sekvence během probíhající sekvenace [27].

Přednostmi nanopórového sekvenování jsou velká délka čtení, možnost čtení "v živém přenosu", čtení jednotlivých molekulových fragmentů původní DNA a schopnost číst modifikace bází těchto původních molekul [28]. Ačkoli nanopórové metodě byla v minulosti přisuzována větší chybovost (relativně k HiFi čtením), nejnovější hardware a software nabízí čtení s Phred skóre bází kolem 30 [28, 29]. Děje se tak pomocí sekvenování duplexu, což spočívá ve spojení komplementárních vláken oligonukleotidovou smyčkou, translokaci pórem, a tudíž přečtení obou vláken a *in-silico* vyvolání konsenzuální sekvence [28, 25]. Přítomnost homopolymerů zapříčiňuje vyšší chybovost čtení [25].

Kapitola 2

Rekonstrukce genomů ze sekvenačních dat

Základní problém řešený v procesu sestavování genomu (dále též nazývaném přejatým slovem assembly) je stejný pro tradiční genomové sestavování i pro sestavování skupiny genomů z metagenomických dat. Ze sekvenačních dat, která nám poskytují informaci o sekvenci různých fragmentů, je cílem získat sekvenci původních molekul reprezentujících genomy sekvenovaných organizmů či druhů. V případě sestavování eukaryotního genomu tak může jít o získání sekvencí chromozomů sekvenovaného jedince, přičemž se podle výzkumných cílů může jednat o pravdivé sestavení všech jednotlivých chromozomových sad, či o sestavení tzv. konsenzuálních sekvencí zastupujících navzájem homologní chromozomy. Dále také mohou být sestaveny extrachromozomální sekvence DNA (např. mitochondriální DNA).

V případě sestavování z metagenomických dat jsou žádaným výstupem sekvence genomů organizmů přítomných ve vzorku, typicky jde o cirkulární chromozomy bakterií a jejich plazmidy. Zdůrazňujeme význam toho, že se jedná v nějakém slova smyslu o konsenzuální genomy. Nebereme-li v úvahu metody amplifikovaného sekvenování jednotlivých buněk [30], každá sekvence na výstupu bude reprezentovat homologní a blízce si podobné sekvence molekul z více buněk přítomných ve vzorku. Ačkoli se tyto molekuly mohou lišit (jak bude probráno později) nebo mohou být odlišnosti vneseny chybami v procesu sekvenace, programy provádějící assembly musejí určit konsenzuální sekvenci, tedy zástupnou sekvenci, která bude co nejidentičtější se zastupovanými sekvencemi. Můžeme říci, že bude jakýmsi jejich průměrem.

Pokud k nějakému námi sestavovanému genomu známe sekvenci referenčního genomu, neboli takového, u něhož usuzujeme na velkou sekvenční podobnost s naším, můžeme přistoupit k *referencí řízenému sestavení*[1] [31]. Takový postup využívá možnost mapování sekvenačních čtení na referenční genom, což umožňuje zjistit odlišnosti obou genomů a upravit referenční genom na genom sestavo-

[1] Z angl. *reference-guided assembly.*

vaný. Dále v tomto textu se budeme plně soustředit na postupy uplatnitelné, i pokud žádnou referenci nemáme a musíme genomy sestavit pouze ze znalosti sekvenačních čtení, budeme se tedy zabývat *de novo sestavováním*[2].

V širší rovině můžeme pod procesem sestavování rozumět několik v praxi rozlišovaných podproblémů, což vychází ze skutečnosti, že bývá nerealizovatelná úplná rekonstrukce sekvencí molekul od začátku do konce. První podproblém tak pouze řeší spolehlivé sestavení čtení do větších fragmentů, tzv. contigů[3]. Následnými, blízce spjatými problémy jsou označení contigů patřících do stejného genomu, v angličtině *genome binning*, a případné uspořádání contigů a vymezení strukturálních závislostí mezi nimi, což se přirovnává k tvorbě genomové kostry či lešení, tzv. *scaffoldu*. Scaffoldy jsou výslednou aproximací genomů. Zpravidla bývá následně provedeno taxonomické přiřazení[4] a anotace prvků v sekvencích (např. promotory, kódující sekvence).

Pro sestavení genomů z metagenomu bylo vyvinuto množství algoritmických postupů a softwarových nástrojů, které se liší tím, které podproblémy postupu řeší, a mimo to se liší použitými algoritmy, zaměřením na různé druhy sekvenačních dat, dosahovanými výsledky a výpočetní náročností. Základem všech přístupů k sestavení contigů však je využití překryvů mezi přečtenými fragmenty pokrývajícími původní sekvence. Genomy se obvykle sekvenují s mnohanásobným pokrytím (viz kapitola 1) a sestavení contigu lze čekat pouze u části molekuly, která bude bez přerušení pokrytá překrývajícími se čteními. Pro základní pochopení práce těchto nástrojů je důležité se seznámit s používanými algoritmickými postupy.

2.1 Základy algoritmů pro assembly

Vstupem do programu provádějícího assembly (tzv. assembler) je sada sekvencí čtení, reprezentovaná např. souborem ve formátu FASTA či FASTQ. Existují tři přístupy implementovatelné v assembleru [31]. První, *hladový* přístup, se snaží o iterované připojování čtení s maximálním překryvem do vznikajících contigů [31]. Překryvem se nemyslí nic jiného než, že přípona jedné sekvence je předponou sekvence druhé (několik chybných bází bývá tolerováno). Další dva přístupy zachycují vztahy mezi čteními do grafové datové struktury, která usnadňuje extrakci sekvencí contigů.

V přístupu OLC, z angl. *overlap-layout-consensus*, pracujeme s grafem představujícím čtení a jejich překryvy, v této práci označovaném překryvový graf [5] [31]. Vrcholy v něm zastupují sekvence jednotlivých čtení. Pro konstrukci hran je potřeba prověřit dvojice čtení a hledat mezi nimi překryvy jejich konců [32, 33] Výklad v tomto místě zjednodušíme popisem, kde hrany grafu jsou orientované a

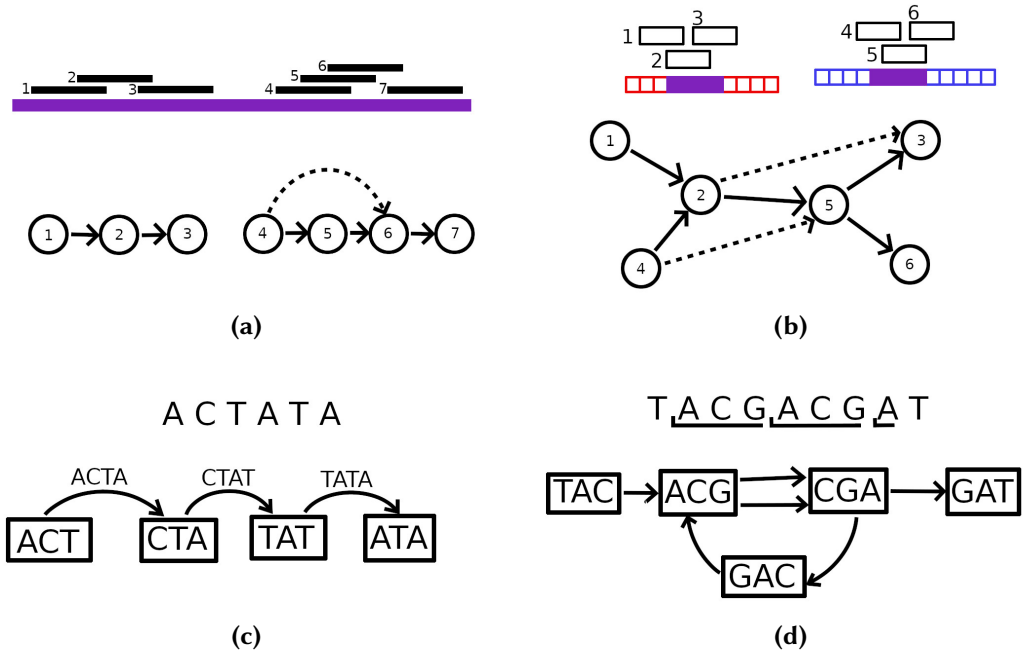
[2] Z angl. *de novo assembly*.
[3] Z angl. *contig*.
[4] V angl. *taxonomic binning*.
[5] Překlad angl. *overlap graph*.

hrana spojuje vrcholy, jejichž sekvence mají dostatečně dlouhý překryv [32] (viz ukázku na obrázku 2.1a). Hrana je intuitivně orientována z vrcholu sekvence překrývající se svou příponou do vrcholu sekvence překrývající se předponou [32]. Z více možných překryvů se vždy uvažuje ten nejdelší [32]. Graf může obsahovat tranzitivní hrany, které vznikají v posloupnosti čtení, kde se překrývají čtení se svými nejbližšími sousedy, ale kde se také překrývá čtení s jiným ob více sousedů. Dva vrcholy propojené hranou nasvědčují, že obě sekvence sdílejí stejný úsek původní sekvence a jejich spojení (ztotožněním překrývajících se částí) vede k rekonstrukci delšího úseku. Takovým spojením více vrcholů, které v grafu leží na cestě, rekonstruujeme možnou podobu contigu. V procesu assembly je jedním z prvních kroků po postavení překryvového grafu odstranění tranzitivních hran, které v grafu ruší jednoduché cesty [32, 33]. Zjednodušení spočívá v tom, že nevíme, jestli to které čtení pochází z vlákna nesoucí sekvenci, kterou nakonec zrekonstruujeme jako contig, nebo z vlákna komplementárního. Proto musíme brát v úvahu nejenom čtení, ale i jejich reverzní komplementy. Assemblery toto mohou řešit tak, že si vrcholy můžeme představit jako dvouvláknové fragmenty DNA (a rozlišení vláken na dopředné a reverzní) a hrany překryvového grafu jsou orientované na obou svých koncích zvlášť a tyto koncové orientace vypovídají o tom, v jaké orientaci se dva fragmenty překrývají [33].

Při pokusu o sestavení contigu z grafu vznikají nejednoznačná uspořádání, kde lze vrcholem vést více cest, odpovídajících alternativním podobám sekvence [32]. Zapříčinit to může sekvence, která se opakuje na více místech v genomu nebo která je přítomná ve více různých genomech (viz obrázek 2.1b). Čtení, která končí touto repetitivní sekvencí, se budou překrývat s čteními, která tak začínají, a to bez ohledu na to, zda pocházejí ze stejného opakování či z jiného. Potíž dělají repetitivní sekvence, které jsou delší než dosahovaná délka čtení, neboť není sekvenací možné získat čtení, které by se překrývalo po obou stranách repetice s unikátními úseky.

Druhou volbou grafu pro assembler je *de Bruijnův graf*. Pro dané číslo k se k-merem označuje sekvence délky k. Uvažujme k menší než průměrná délka čtení, v řádu desítek. De Bruijnův graf se skládá z orientovaných hran, které zastupují k-mery vyskytující se v sekvencích čtení [5, 34]. Vrcholy grafu reprezentují $(k-1)$-mery a do grafu umístíme takové vrcholy, aby každá hrana vytvořená na základě sekvence čtení vedla z vrcholu reprezentující $k-1$ dlouhou předponu k-meru hrany do vrcholu $k-1$ dlouhé přípony k-meru hrany. Spojení vrcholů hranou tak značí bezchybný překryv $(k-1)$-merů o délce $k-2$. Stejné k-mery (resp. $(k-1)$-mery) z různých čtení jsou ztotožněny v de Bruijnovu grafu. Můžeme učinit následující pozorování. Uvažme sadu bezchybných čtení pokrývající rekonstruovanou sekvenci s překryvy délky nejméně k. Jejich reprezentace de Bruijnovým grafem bude mít podobu cesty, pokud se v sekvenci neopakuje žádný $(k-1)$-mer (viz obrázek 2.1c). Počáteční vrchol cesty odpovídá počátečnímu $(k-1)$-meru a každý přechod po hraně přidá následující bázi.

Je užitečné si uvědomit, že různé hrany (resp. vrcholy) mohou zastupovat



Obrázek 2.1 Principy využití grafů pro assembly. *(a)* Čtení 1-7 jsou zarovnána na sekvenci (fialová), ze které pocházejí. Na základě překryvů čtení vznikl prezentovaný překryvový graf. Tranzitivní hrana je zobrazena čárkovaně. *(b)* Příklad rozvětveného překryvového grafu, který vznikl ze čtení pokrývajících dvě kopie repetitivní sekvenve (fialově vyplněné úseky). Reference a čtení jsou pro názornost rozdělena do bloků (např. 25 bp dlouhých). Tranzitivní hrany jsou čárkovaně. *(c)* Ilustrace de Bruijnova grafu reprezentujícího 1 čtení. *(d)* Příklad de Bruijnova grafu s kružnicí vzniklé v důsledku tandemové repetitivní sekvence v reprezentovaném čtení. Dvojice šipek mezi ACG a CGA naznačuje dva výskyty ACGA ve čtení.

sekvenci přítomnou v datech v různém počtu opakování. Hovoříme proto o pokrytí hran ve smyslu počtu čtení s daným *k*-merem [5] a na obrázku 2.1d je toto naznačeno zdvojenou hranou. Obrázek 2.1d také ilustruje nedostatek de Bruijnových grafů: ačkoli tu máme lineární sekvenci, její reprezentace není jednoduchou cestou, z které by se dalo jednoznačně rekonstruovat contig. Ve vyobrazeném příkladu tandemová repetice vedla k tvorbě kružnice. Dobré je si rozmyslet, že pro repetice delší než použité *k* platí: repetice uvnitř chromozomu povedou k uzavření kružnice, repetice mezi chromozomy povedou k tvorbě křížení. Přítomnost kopií chromozomu s různou bází na jedné pozici povedou k tvorbě tzv. bublin, tedy cesty větvící se do dvou a později splívající zpět do jedné cesty.

2.2 Nástroje pro metagenomické assembly

Ačkoli se dají úspěšně použít obecné nástroje sestavení genomu i na velmi jednoduché mikrobiální komunity, problémy objevující se u složitějších komunit motivovaly vývoj mnoha assemblerů dedikovaných pro metagenomická data [35]. Implementace se musí vyrovnat s podstatným objemem dat potřebných pro osekvenování komunity a zpracovat je s ohledem na omezenou primární paměť pro výpočet [36, 35]. Naším cílem zde není podat vyčerpávající průzkum publikovaných softwarů, přesto jich několik představíme a vybírat budeme podle jejich účasti a výsledků v soutěži CAMI II [37], kterou blíže představíme v sekci 2.3.

ABySS [38] je starší assembler, pracující s de Bruijnovým grafem na distribuovaných systémech. Distribuovaného uložení grafu je docíleno jednoznačným určením lokalizace *k*-meru na uzlu pomocí hašování a úspornou reprezentací sousedů *k*-meru díky pozorování, že lze vytvořit pouze 8 možných sousedů (4 odpovídají přechodu z *k*-meru na jiný rozšířený na konci jednou ze 4 bází, podobně 4 sousedi s různou počáteční bází mohou předcházet vrcholu). Před sestavením probíhá korekce grafu zahrnující iterované odstraňování krátkých terminálních větví a cest v malých bublinách tak, aby z každé bubliny zůstala pouze ta nejvíce pokrytá cesta. Pak jsou nerozbíhavé cesty spojeny do contigů. Pokud jsou čtení párová, je tato informace využita k dalšímu propojení contigů. Graf je prohledán pro každý contig s cílem nalézt cestu mezi daným contigem a dalšími contigy, které se na základě informace párových čtení nacházejí potenciálně někde v následující sekvenci. Tyto contigy najdeme pomocí zarovnání čtení na contigy a vybrání těch, co jsou spojeny párem čtení: jedno čtení páru je zarovnáno na první contig a další čtení na druhý contig. Navíc těchto spojujících párů musí být dostatečný počet. Takto identifikované cesty mezi contigy, které nasvědčují možné propojení contigů, jsou použity k propojení contigů v případech, které jsou jednoznačné [38].

GATB [39] je knihovna softwarů pro genomové assembly a zpracování, která je založená na reprezentaci dat de Bruijnovým grafem. Se zřetelem k nedostatku primární paměti mnoha počítačů byl v knihovně kladen důraz na adaptivní využití pevného disku a na úspornou reprezentaci de Bruijnova grafu, s využitím

Bloomova filtru. Součástí GATB je assembler Minia [40] a navíc nástroje pro detekci variant z grafu, discoSNP [41] pro hledání krátkých sekvenčních variant a TakeABreak [42] pro detekci konců invertovaných sekvencí (pro další nástroje viz kapitolu 3) [39].

MetaSPAdes [35] je adaptací assembleru SPAdes [5] cíleného na sestavování mikrobiálních genomů ze *single-cell* dat. Vstupní data jsou párová čtení a párovosti je důmyslně použito při sestavení. Program používá v kombinaci několik koncepcí rozvíjejících de Bruijnovy grafy. *Multisized* de Bruijnův graf využívá ke své konstrukci k-mery z čtení pro interval různých hodnot k, a obchází tím nevýhody fixní hodnoty k. Začlenění struktury *párového* de Bruijnova grafu je způsobem využití párovosti čtení přímo při sestavování contigů. Jeho efektivní použití je umožněno dopočítáním přesného odhadu délky inzertu [5]. MetaSPAdes se snaží zbavit minoritních variant komplikujících assembly. Vodítkem je relativní pokles pokrytí v některých cestách. Realizuje se odpojením alternativní cesty od konsenzuální cesty. Součástí implementace je modul exSPAnder, který tvoří contigy iterativním prodlužováním cest v grafu, u nichž je dostatečná podpora hodnotami lokálního pokrytí [35].

MetaHipMer [43] je rozšíření assembleru HipMer [44] pro metagenomická data. Návrh počítá se škálováním pro společné sestavení (*coassembly*) objemných sad dat ze složitých komunit s použitím paralelizace na distribuovaných systémech. Součástí algoritmického řešení je iterované sestavení contigů přes vzrůstající k použité v de Bruijnově grafu, přičemž contigy z minulé iterace přispívají ke stavbě grafu v další iteraci. Cílem je při iteracích s menším k sestavit contigy druhů s nízkým pokrytím, u nichž by větší k vedlo k fragmentaci, a následně použitím větších k vyřešit nejednoznačné úseky. Nakonec jsou contigy sestaveny do scaffoldů pomocí čtení propojujích konce contigů a na základě hloubek pokrytí podobně jako u řešení v exSPAnderu [43].

MEGAHIT [36] je assembler dedikovaný pro zpracování metagenomických dat. Založen je na de Bruijnových grafech. Do paměti je ukládá ve velmi úsporné reprezentaci s lineární paměťovou složitostí vzhledem k počtu hran a s lineární časovou složitostí konstrukce vzhledem k velikosti vstupních dat [45]. Také zde se používá iterativní skládání contigů přes vzrůstající hodnoty k z podobných důvodů jako výše [36].

OPERA-MS [46] je workflow pro sestavení genomů z metagenomických dat kombinujících dlouhá ONT čtení s krátkými čteními Illuminy. Základní myšlenkou je sestavit contigy z krátkých, ale kvalitních čtení a poté sestavit assembly graf, který spojuje hranami contigy propojené překryvem dlouhého čtení. Pro sestavení contigů se aplikuje jeden z vhodných assemblerů, v základním nastavení je MEGAHIT. Další práci programu shrnuli autoři takto: "Genomy jsou rozlišeny na úrovni druhů pomocí klastrování ... , dále přiřazeny do poddruhových klastrů předtím, než proběhne *scaffolding* a vyplnění mezer s OPERA-LG." Algoritmus klastrování contigů uplatňuje Bayesovský přístup [46].

Ray Meta [47] je určen pro distribuované sestavení metagenomických dat a

vychází z Ray [34]. Disponuje též nástrojem Ray Communities pro taxonomickou profilaci z *obarveného* de Bruijnova grafu. Barvení v tomto významu zastupuje zaznamenávání informací příslušejících prvkům grafu. Konkrétně zde je u každého vrcholu de Bruijnova grafu zaznamenána informace o výskytu odpovídajícího $(k-1)$-meru v referencích použité taxonomické databáze. Ray je založený na postavení (distribuovaného) de Bruijnova grafu a hladového prodlužování cest odpovídajících contigům. Prodlužování cest se řídí heuristikami, které uvažují u alternativních prodlužujících hran jejich podporu čteními a překryv těchto čtení s rekonstruovanou cestou (krátký překryv nemusí znamenat původ čtení v rekonstruovaném contigu). Podpora grafu čteními je uložena v grafu pro každé čtení u vybraného vrcholu $(k-1)$-meru z čtení [34].

2.3 Chybovost a dosahované výsledky

Nabízí se několik charakteristik metagenomických dat, jež jsou důležité pro návrh dobrého assembleru. Zaprvé může úkol ztěžovat neznámý počet rekonstruovaných genomů (kolik komunita obsahuje druhů nebo kolik haplotypů bude možné rekonstruovat) a jejich různá hloubka pokrytí [48, 35]. Přítomnost vnitropopulační variability pochopitelně vede ke složitější rekonstrukci [35]. Navíc kontaminace čtení sekvenačními chybami přidává další matoucí variabilitu a žádá si pozornost při zpracovávání [36]. Assemblery často volí přístup zpracování variabilních úseků do společné konsenzuální sekvence [48]. Ale i bez ohledu na faktory výše je nutné při algoritmizaci počítat s repetitivními úseky uvnitř genomu a s homologními úseky opakujícími se mezi různými genomy [35, 48].

Zhodnotit správnost získaných MAGů můžeme pouze s testovacími daty, u nichž známe původní genomy. Dobrou mírou pro hodnocení je NGA50, které lze definovat jako minimální délku sekvence takovou, že všechny contigy alespoň tak dlouhé pokrývají dohromady přes 50% pozic referenčního genomu [37]. I bez znalosti správných výsledků máme několik rozumných možností validace [48]. Pro skupinu contigů, např. přiřazených do jednoho genomu, je N50 hodnota minimální délky takové, že contigy alespoň takto dlouhé dávají dohromady přes 50% celkové délky contigů. Jde o robustní míru spojitosti sestavení. Pro validaci úplnosti a smysluplnosti MAGu můžeme hodnotit výskyt charakteristických genů a kódujících sekvencí. *Pravděpodobnostní přístup* poskytuje způsoby odhadu pravděpodobnosti pozorovaných dat v případě, že jejich zdrojem byly rekonstruované MAGy, na základě modelů pro proces sekvenování [48].

CAMI II [37] byla výzva bioinformatické komunitě v období mezi lety 2019 a 2021, která přijímala řešení pro zpracování připravených sad metagenomických dat. Přijímány byly výsledky v kategoriích: sestavení contigů, přiřazení contigů do genomů [6] a přiřazení čtení či contigů taxonomickým jednotkám [7]. CAMI II posky-

[6] Cizími slovy *genome binning*.

[7] Cizími slovy *taxonomic binning*.

tuje objektivní a ucelené testování postupů účastníků, které se lišily též použitými assemblery. Řešení měla být dodána s reprodukovatelným záznamem postupu. Datové sady představovaly případy mořského mikrobiomu, rhizobiomu a vysoce rozmanité komunity. Byly připraveny z množství popsaných genomů (včetně tehdy nezveřejněných) pomocí simulace jejich sekvenačních čtení napodobujících párová čtení Illuminy, resp. dlouhá čtení PacBio (pro jisté sady i ONT). Některá řešení pracovala pouze s krátkými čteními, některá s kombinací dlouhých a krátkých čtení (hybridní). Byla hodnocena řada metrik: rekonstruované procento genomu, relativní četnost špatně určených pozic, počet contigů chybné struktury, NGA50, zlomek kvalitních (haplotypům odpovídajících) genomů ze všech skutečně přítomných genomů, preciznost rekonstrukce samostatných haplotypů a doba běhu výpočtu a maximální využití paměti. Analýza nejlepších řešení různých assemblerů názorně demonstrovala rostoucí závislost rekonstruovaného procenta genomu s rostoucí hloubkou pokrytí. Zjednodušeně si můžeme představit sigmoidální křivku [8]. Navíc nároky na pokrytí se lišily, v popsaném případě v rozpětí 9.2× (SPAdes) až 19.5× (Ray meta). Některé assemblery (ABySS na krátkých čteních a OPERA-MS na hybridních) nezvládaly rekonstrukci plazmidů, navzdory jejich extrémní hloubce pokrytí (Ray na krátkých čteních a GATB na hybridních též některé plazmidy nezrekonstruovaly). Mezi vyvozenými závěry bylo shrnuto, že verze programu neměla příliš vliv, ale parametry a předzpracování (např. ořezávání nekvalitních bází nebo korekce čtení) ano. Dále hybridní assembly neukázaly signifikantně vyšší kvalitu, ale lépe se vypořádaly s podobnými kmeny a lépe rekonstruovaly komplikované úseky [37].

LMAS [49] je nedávno vydaná platforma pro *benchmarking* a hodnocení assemblerů na metagenomických datech. Pomocí LMAS její autoři porovnali různé assemblery. Kromě simulovaných sad čtení byly testovány i sady reálných čtení z připravených komunit známé skladby. Reálná data byla pro assemblery obtížnější než simulovaná. Navíc nebyla ve výsledcích znát nadřazenost assemblerů dedikovaných pro metagenomiku oproti ostatním [49].

[8] Konkrétně tato křivka má pro malé hodnoty pokrytí asymptoticky nula procent genomu rekonstruovaných, na jistém intervalu hodnot pokrytí rostoucí vývoj, poté jeho udržení blízkou 100% [37].

Kapitola 3

Analýza vnitropopulační variability mikrobiálních komunit

Minulá kapitola popisovala bioinformatické postupy, kterými lze dospět ke genomovému assembly z metagenomických dat. Výsledné genomy byly jakýmsi *průměrem* přes všechny genomy jedinců v populaci vzorku, zachycenými sekvenací. Ve smyslu tohoto přirovnání se tato kapitola zabývá *rozptylem* studovaných sekvencí, neboli popisem sekvenčních variant, které existují mezi genomy jednotlivých buněk. Budeme v tomto významu hovořit o *vnitropopulační variabilitě*, v literatuře se někdy používá také výrazu mikrodiverzita [50, 4]. Oba výrazy mohou vést k nejednoznačnému chápání (první se opírá o vymezení populace, druhý se odkazuje na příbuznost buněk v nejednoznačném mikro měřítku), my upřednostníme první. Studie ukazují, že tato vnitropopulační variabilita bývá spojená s funkčním rozrůzněním, hraje svou roli v ekologii dané populace a může přispět k porozumění evolučních procesů mikroorganizmů [30, 50, 7, 51]. Nejprve popíšeme a klasifikujeme, jakými způsoby se mohou lišit genomy buněk v populaci. Dále ukážeme, jak lze při sběru dat a následném zpracování postupovat různými cestami. Zaměříme se na informatická řešení detekce a popisu variant. Nakonec budeme vše demonstrovat na vybraných studiích.

3.1 Klasifikace variant

Variace v populaci se může objevit v důsledku různých procesů, které zahrnují imigraci z jiné populace, horizontální genový přenos (HGT [1]), aktivitu mobilních elementů a mutace. Při srovnávání sekvencí homologních genomů může mít mnoho podob. Pro naše účely analýzy variant je nejzásadnější rozlišení mezi strukturálními variantami a jednonukleotidovými variantami a k nim přiřazovaným krátkým indelům (max. 50 bp). Jako jednonukleotidovou variantu (SNV [2])

[1] Z angl. *Horizontal Gene Transfer.*
[2] Z angl. *Single Nucleotide Variation.*

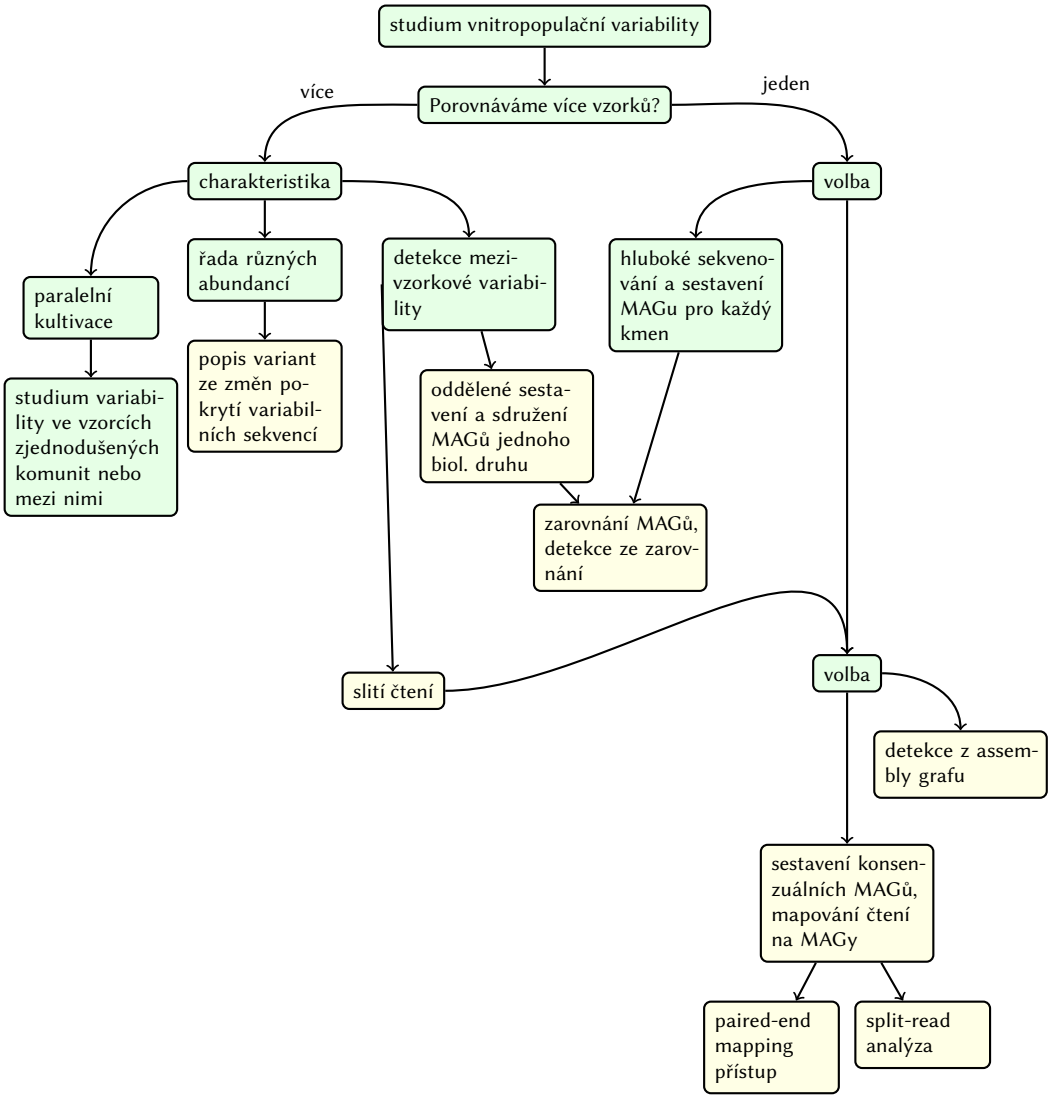
označíme takovou pozici v genomu, která má v různých verzích genomu přiřazenou různou bázi (A, C, G, T), případně v některých verzích tato pozice chybí (indel). Jako indel označíme i případ, kdy dotčená sekvence je delší něž jediný (komplementární) pár bází. Indel je v referenci na variantu obsahující sekvenci vnímán jako delece a v referenci na variantu postrádající sekvenci jako inzerce. Strukturními variantami (SVs [3]) se obecně rozumí změna delší souvislé sekvence nukleotidů a mezi nejčastěji uvažované strukturní varianty patří inzerce, delece, duplikace, inverze a translokace. Minimální délku nutnou k tomu, aby změna sekvence byla označena za SV, řeší různé práce různě, v některých případech jako 50 bp [52, 53], jindy např. jako 10 bp [54], a při podrobném rozboru vědeckých prací je třeba si na to dát pozor.

Jak bylo řečeno, SVs zahrnují bohatou skupinu jevů, zde poskytneme bližší popis. Delece jsou SVs, kde některým verzím genomu chybí úsek v porovnání s referencí (často se za referenci uvažuje nejčetnější alela, verze uvažované proměnlivé sekvence). Inzerce jsou naopak SVs, u kterých verze genomu obsahuje sekvenci navíc, často se myslí sekvenci, co nemá původ jinde v genomu [55]. Duplikace jsou vložené sekvence, které svou sekvencí kopírují jiný genomový úsek. Tandemová duplikace pak je speciálním případem, kdy je kopie vložená bezprostředně před nebo za svou vzorovou sekvenci. Inverze zachovává pozměněný úsek, ale obrací pořadí, v němž báze úseku následují. Translokace je delece dané sekvence v místě jejího původního položení a její inzerce na jiné místo genomu. Mnoho dalších, složitějších, případů SVs lze popsat jako kombinaci několika předchozích typů variant na překrývajících se úsecích sekvence [56].

3.2 Možné experimentální postupy

Úkol získat z metagenomických dat popis všech variant, které sekvenování bylo schopné vzhledem k omezenému pokrytí zachytit, přesně a bezchybně, je náročný. Navíc pouhý výčet detekovaných variant v různých místech genomu nemusí stačit pro zodpovězení některých otázek, pro které by bylo třeba znát, jaké varianty se spolu vyskytují v týchž buňkách, což je problém, o kterém lze smýšlet jako o rekonstrukci *haplotypů* přítomných ve vzorku. Cílem je zachytit genetickou informaci kmenů koexistujících v populaci. V literatuře je toto označováno také jako *fázování*. Jako v jiných oblastech bioinformatické analýzy biologických dat je i zde možná pouze určitá přesnost a vždy existuje šance na výskyt falešně negativních a pozitivních detekcí. Bylo zveřejněno značné množství nástrojů dedikovaných pro detekci variant v metagenomech a každý přispívá do diskuze vlastním pohledem a inovací a poskytuje software různých parametrů. Diverzita v problematice popisu variability se neomezuje pouze na existující nástroje, ale také na přístupy k návrhu experimentu, o čemž zde krátce pojednáme.

[3] Z angl. *Structural Variants*. Zkratky anglických pojmů budeme končit písmenem s, když budeme diskutovat o pojmu v množném čísle.



Obrázek 3.1 Rozbor možností, kterými lze přistoupit k identifikaci variací uvnitř mikrobiální populace. Blíže vysvětleno v sekci 3.2.

Obrázek 3.1 ukazuje podstatná specifika experimentu, která vedou k určitým algoritmickým východiskům (žluté rámečky). Základním přístupem pro nás bude cesta v diagramu na přímé větvi jdoucí vpravo seshora dolů. Jedná se o konfiguraci s jediným vzorkem a jediným z něj získaným datasetem sekvencí. Tento dataset potenciálně obsahuje příspěvěk z mnoha příbuzných haplotypů genomu a cílem je tuto skutečnost zachytit. Obecně toho dosáhneme ve dvou krocích, přičemž předpokládáme, že postupujeme *de novo* a nemáme referenci z databáze, což bude pravdivý případ velkého množství prokaryotických druhů [57]. Nejprve zkonstruujeme MAG pro každou skupinu blízkých haplotypů (reference pro druh) a na tyto MAGy namapujeme tento dataset čtení. Ať už pracujeme s jakýmikoli čteními, krátkými či dlouhými, můžeme z tohoto zarovnání získat informace o potenciálních variantách pomocí charakteristických odchylek, kdy sekvence čtení není identicky zarovnaná na souvislý úsek reference (detekce SVs fragmentuje mapované čtení a mluvíme o tzv. *split-read analysis*) [58]. Pokud jsou naše čtení párová, můžeme využít, že důsledkem SV je změněná vzdálenost čtení stejného páru při mapování na referenci oproti skutečnosti a detekce takových odchylek od očekávaného rozdělení vzdáleností poukazuje na variaci (tzv. *paired-end mapping* přístup) [58]. Pro detekci SVs založenou na mapování je žádoucí schovat veškerou vnitropopulační variabilitu do konsenzuálního MAGu, tvorba více genomů by komplikovala analýzu. Vicedomini et al. [59] uznávají tuto skutečnost a používají pro takovýto konsenzuální MAG výraz *strain-oblivious assembly* [4] a pro konkrétní genom haplotypu výraz *strain-aware assembly* [5].

Obejít se můžeme i bez mapování na referenční sekvenci, využitím grafových struktur (na obrázku 3.1 rámeček "detekce z assembly grafu"). Grafy používané assemblery, ať už překryvové či de Bruijnovy, zachycují větší variace rozvětvením [57]. De Bruijnův graf s $(k-1)$-mery asociovanými s vrcholy bude obsahovat bublinu i pro substituci jednoho nukleotidu, neboť by se na cestě změnilo $k-1$ vrcholů. Různá spletitější propojení grafu přispívají k tomu, že assembler nezvládne jednoznačné sestavení genomů a na výstupu dává pouze kratší contigy, tyto contigy bývají doplněny o graf, který má jako vrcholy contigy, a rekonstrukce cest mezi propojenými contigy je způsobem hledání SVs, např. užívaným nástrojem MaryGold [57].

Pro úplnost v diagramu uvádíme ještě další možnost detekce variability z jediného vzorku, v pravé polovině obrázku 3.1 jde o levou větev. Pokud bychom dokázali všechny zajímavé haplotypy sestavit jako samostatné MAGy, můžeme variaci zachytit ze zarovnání těchto MAGů a porovnání neidenticky zarovnaných míst [58, 60].

Analýza jednoho vzorku metagenomu může postačit pro prozkoumání variability v mikrobiální komunitě daného prostředí. Současné zpracování více různých vzorků však má také své místo. Zaprvé nás může zajímat, zda se populace stejného druhu z různých vzorků liší genetickými variantami a jejich frekvencemi

[4] Volně bychom přeložili jako sestavení skrývající kmeny.

[5] Volně bychom přeložili jako sestavení rozlišující kmeny.

(rámeček ”detekce mezivzorkové variability” na obrázku 3.1). Prvním řešením je spojení datových sad porovnávaných vzorků (”slití čtení”) a hledání variací způsoby popsanými pro jeden vzorek. Zřejmým důsledkem při přímočaré realizaci tohoto řešení je spojení efektů variability uvnitř populace s rozdíly mezi populacemi vzorků. Druhým řešením je oddělené sestavení genomů z různých vzorků a sdružení MAGů příslušejících stejnému biologickému druhu, načež by bylo možné tyto MAGy mezi sebou porovnat zarovnáním, stejně jako by to bylo možné pro více MAGů téhož druhu původem z jednoho vzorku (rámeček ”zarovnání MAGů” společný oběma přístupům).

Přístupem, který byl úspešně využit pro lepší popis variability, ale který přesahuje ryzí metagenomiku, je využití *paralelních konsorcií* kultivovaných z jedné komunity [50]. Kultivace podmnožiny přítomných druhů a haplotypů ubírá komplexitu během metagenomického zpracování a umožňuje citlivější detekci variability [50].

Při zkoumání vnitropopulační variability si rovněž můžeme pomoci specifickými informacemi získatelnými z více vzorků téže komunity za přiměřeně rozdílných podmínek, např. sebráním vzorků tvořících časovou řadu vývoje komunity nebo kultivací za trochu jiných podmínek. Základem těchto přístupů je představa, že rozdílné abundance haplotypů mezi vzorky se projeví změnami pokrytí v reprezentaci jejich sekvencí (assembly graf, MAG) a vzorce těchto změn jsou indikací struktury variant [54] či provázání variant v haplotypech [61].

3.3 Informatické nástroje detekce variant

Pro potřeby detekce genetických variant ze sekvenačních dat bylo publikováno nemalé množství softwarových implementací, přičemž mnoho z nich vzniklo za účelem analýzy jednotlivých genomů, např. pro výzkum biologie člověka. Takový software může o datech předpokládat původ z genomu určité ploidie, předpokládat frekvenci odpovídající heterozygotní alele, s rovnoměrným sekvenačním pokrytím apod. [9, 62]. Může být vcelku vhodný pro analýzu sekvenace klonálních vzorků a málo vhodný pro analýzu pravého metagenomu.

Mnohé nástroje jsou navrženy pouze na analýzu SVs, resp. SNVs. Nástroje z této katogorie jsme shrnuli do tabulky 3.1, resp. tabulky 3.3. Navíc několik nástrojů zvládá zpracování obou typů variant a ty jsme shrnuli do tabulky 3.2.

Co se týče analýzy SVs, raným softwarem pro zpracování metagenomických dat byl MaryGold [57]. Tento nástroj pracoval s ”contig grafem”, grafovou datovou strukturou zachycující možná propojení zrekonstruovaných úseků genomů, kterou je možné získat na základě výstupu z assembleru (autoři použili metAMOS [63]). Vrcholy grafu představují contigy a hrany jsou zvlášt orientované na obou svých koncích a propojení hranou znamená překryv daných contigů sekvencí čtení přes konce contigů vyznačené orientacemi hrany. MaryGold si kladl za cíl detekovat v tomto grafu oblasti vnášející nejednoznačnost při průchodu jinak jednoduché cesty, tzv. bubliny. Různé průchody těmito oblastmi značí

potenciální varianty sekvence. Algoritmus hledající bubliny nejprve určoval potenciální páry větvících se konců bublin. V neorientované verzi grafu bubliny tvoří 2-souvislé komponenty a konce bubliny tvoří pár, jehož odstranění vede ke ztrátě souvislosti grafu. Pro nalezení kandidátních párů pro konce bubliny byly hledány páry vrcholů rozdělující předtím identifikované 2-souvislé podgrafy do více komponent. Myšlenkou je provést rozklad 2-souvislých podgrafů do 3-souvislých komponent, k čemuž se využívá dříve vyvinutý postup zavádějící novou matematickou strukturu, SPQR stromy. Skutečné páry konců bublin jsou poté identifikovány prověřením cest v kandidátní bublině pomocí prohledávání do šířky. Alternativní cesty mezi konci validní bubliny jsou pomyslně nahrazeny jednou cestou se záznamem o podobě každé alternativní sekvence [57].

Zástupci přístupu detekce z mapování na referenci (viz sekce 3.2) jsou například DELLY [64], SVIM [55], Sniffles [56] či nedávno vydaný Sniffles2 [65], které jsou však zamýšlené pro analýzu izolovaných genomů a podporují zejména analýzu diploidního genomu. Máme-li metagenomický vzorek dobře reprezentovaný referencemi (třeba díky de novo sestavení), mohou být výsledky podávané těmito nededikovanými nástroji natolik dobré, že není velká potřeba specializovaného softwaru. Minimálně pro velmi jednoduché komunity má takový nástroj (Sniffles) své využití při hledání SVs [66]. Na druhou stranu, různý charakter sekvenačních dat s krátkými čteními a s dlouhými čteními si žádá různé přístupy: SVIM, Sniffles a Sniffles2 hledají SVs pomocí dlouhých čtení, zatímco starší DELLY používá krátká čtení (ideálně několik knihoven různých délek inzertů).

Podobným přístupem jako zarovnání čtení na genom, je zarovnání různých verzí genomu vůči sobě, v angličtině označováno *whole-genome alignment*. Pipeline metaSVs [67], která je uživatelsky přívětivým řešením pro zkoumání SVs z kombinace metagenomických dat sekvenovaných přístupem Illumina a ONT, využívá pro detekci variant program MUM&Co [68], který provádí detekce z celogenomových zarovnání, ale nebyl specificky vyvíjen v kontextu metagenomiky.

V duchu detekce variant z grafu vzniklo množství nástrojů dedikovaných metagenomice. Protože se používají grafy těsně související s procesem assembly, jsou i na úrovni nástrojů přesahy mezi assemblery a detektory variant. Např. metaFlye [69], assembler dedikovaný pro sestavování MAGů, umí pracovat v režimu $\text{metaFlye}_{\text{strain}}$, který se snaží sestavit genom každého haplotypu zvlášť.

Rhea [54] je zástupcem nástrojů detekujících SVs na základě sekvenačních dat z posloupnosti vzorků (např. časová řada) a vstupním grafem je právě assembly graf získaný pomocí $\text{metaFlye}_{\text{strain}}$, společný všem vzorkům. Čtení jsou namapovány na graf programem minigraph [70] a pokrytí grafu je transformováno do informace o proměnách pokrytí mezi vzorky. To slouží k jednoduché klasifikaci bublin jako strukturálních variant: např. delece nabývající abundance v časové řadě lze v grafu najít jako kružnice na 3 vrcholech, kde dva vrcholy se mezi vzorky mění velmi podobně, zatímco třetí vrchol (deletovaná sekvence) má klesající hloubku pokrytí [54].

Cortex [71] je nástroj, který sám sestaví de Bruijnův graf ze sady čtení a

název a rok publikace článku	principy práce	vstup
MaryGold 2013	Grafový algoritmus schopný zpracovat složitý graf více alel; tvorba konsenzu s uchováním variant.	*Contig graph* (výstup z assembleru); zamýšlen pro krátká čtení.
metaFlye 2020 (Flye 2019)	Assembler s módem zachovávajícím struktury rozlišující kmeny.	Dlouhá čtení.
metaSVs 2023	Uživatelsky přívětivá pipeline usnadňující analýzu čtení, včetně assembly a detekce SVs; detekce variant z celogenomového zarovnání programem MUM-&Co.	Dlouhá čtení (ONT) + krátká čtení (Illumina) + konfigurační složka.
rhea 2024	Více vzorků (časová řada); detekce ze změn pokrytí v assembly grafu.	Několik sad dlouhých čtení zachycujících metagenom v čase, nebo v jiných proměnách.

Tabulka 3.1 Souhrn vybraných nástrojů detekce strukturálních variant na základě sekvenačních dat z metagenomických vzorků.

detekuje variace z grafu kombinací dvou algoritmů. Cortex byl zamýšlen pro současné sestavení více vzorků z různých eukaryotních jedinců a práci s touto informací, kterou ukládá pomocí barvení grafu [71]. Podstata de Bruijnova grafu umožňuje Cortexu detekci SVs i SNVs.

Novější nástroj dedikovaný pro metagenomiku s názvem STRONG [61] detekuje obecné varianty a slouží k jejich haplotypování. Vyžaduje více vzorků překrývajícího se složení a analýzu variant omezuje namísto celých genomů na unikátní geny společné všem haplotypům, které jsou označovány jako SCGs z anglického *single-copy core genes*. Po konstrukci grafu společného všem vzorkům vhodným assemblerem (metaSPAdes [35]) je dalšími softwary provedeno přihrádkování do genomů, identifikace SCGs a extrakce jejich podgrafů. Na nich je provedena společná inference haplotypů v podgrafech pomocí Bayesovsky založeného programu BayesPaths [61].

Breseq [58] je pipeline disponující algoritmy pro detekci SNVs i SVs ze zarovnání krátkých čtení na referenční genom [72]. Reference by měla být kvalitní a co nejvíce souvislá. Detekce SVs probíhá cestou *split-read analysis*, konkrétně dojde k mapování pomocí bowtie2 [73] a všechna možná mapování s rozděleným čtením podstupují jako kandidáti zlomových pozic další prověření, zahrnující oříznutí nejednoznačných překryvů mapovaných částí, spojení evidencí více čtení a hodnocení na základě pokrytí [58].

Zojer et al. [62] s cílem vytvořit automatizované zpracování integrující dostupné nástroje pro studium evoluce mikrobiálních populací porovnali řadu nástrojů v kategoriích detekce SVs, indelů a SNVs. Například jedním ze zjištění byla malá senzitivita Cortexu při detekci SNVs vůči ostatním nástrojům. Výsledkem byl software VarCap používající informovaný výběr nástrojů lišících se přístupy a

název a rok publikace článku	principy práce	vstup
Cortex 2012	Grafový nástroj pracující s obarveným de Bruijnovým grafem a více vzorky.	Krátká čtení (teoreticky i dlouhá).
Breseq 2014	*Split-read* detekce SVs; detekce SNVs.	Krátká čtení + (anotované) referenční genomy vysoké kvality a souvislosti.
VarCap 2017	Workflow integrující různé nástroje.	Krátká párová čtení + reference.
STRONG 2021	Více vzorků (časová řada); grafová reprezentace; restrikce na SCGs fragmenty; extrakce obecných haplotypů Bayesovským algoritmem.	Krátká čtení z řady vzorků proměnlivých abundancí.

Tabulka 3.2 Souhrn vybraných nástrojů schopných identifikovat SNVs i SVs na základě sekvenačních dat z metagenomických vzorků.

schopnostmi. Varianty na výstupu jednotlivých nástrojů VarCap filtruje a slučuje různé detekce způsobené stejnou variantou. VarCap je používaný především na analýzu mikrobiálních izolátů, v kontextu studií experimentální evoluce [74]. Na vstupu přijímá sadu čtení a referenční genom, na který čtení mapuje a vytváří zarovnání pro detekci variant skupinou nástrojů [75]. Tento přístup může být vhodný pro taxonomicky jednoduché komunity v případech, kdy dokážeme zajistit zarovnání čtení na příslušné reference [76]. Nevýhodu by mohla představovat nutnost hlubšího pokrytí vzhledem k frekvenci varianty (400× pro variantu zastoupenou 2%) a horší zachycení inverzí [62].

Narozdíl od mnohotvárných SVs, které nádavkem mohou být výrazně delší než délky čtení, je detekce SNVs jednodušší, stačí například sekvenční průchod zarovnáním čtení na referenci. Na druhou stranu by detekce SNVs měla být citlivější k jednonukleotidovým chybám v sekvenování [77]. Vyvinuté programy se snaží detekovat SNVs s co největší citlivostí a přesností, často doprovázené snahou o kvantifikaci míry chyb detekce nebo o přenesení variant do podoby haplotypů.

LoFreq [78], jak název napovídá, se snaží o citlivou detekci vzácných variant ze zarovnání čtení na referenci, která se opírá o pravděpodobnostní model počítající p-hodnotu SNV z Phred skór čtení mapovaných na danou pozici, přičemž nulová hypotéza je původ změněných bází v chybách sekvenování. p-hodnoty jsou počítány rekurentním vzorečkem a výpočet je zrychlen vhodným ořezáváním a memoizací mezivýpočtů. LoFreq je určen pro práci na sekvencích s danými hodnotami chybovosti (viz strainFlye níže), při jejich absenci se pokusí odhadnout chybovost metodou *expectation-maximization* (EM), která se snaží sekvenační chyby zachytit hodnotami substituční tabulky mezi bázemi A, C, G a T [78].

Dalším nástrojem detekce SNVs je InStrain [79], který na vstupu bere jeden či

více vzorků krátkých párových čtení a pokouší se o přesnější detekci filtrováním čtení tak, že ponechá pouze páry namapované společně a s očekávanou délkou inzertu, s dobrým skóre mapování (MapQ) a dostatečnou sekvenční identitou. Detekci falešných variant je dále předcházeno volbou prahových hodnot pro absolutní i relativní četnost v pokrytí pozice. Navíc varianta musí svou evidencí převyšovat náhodu v nulovém modelu, který je narozdíl od LoFrequ poněkud zjednodušený: uvažuje uniformní míru chybovosti. Varianty zachované po těchto krocích půjdou na výstup. InStrain disponuje do jisté míry i schopností haplotypování: ty varianty, které mají dostatečné pokrytí sadou týchž párů čtení, budou předmětem výpočtu vazebné nerovnováhy[6]. InStrain se hodí pro posuzování divergence vzorků, protože ve variabilních pozicích hodnotí shodu všech variant (i minoritních) a za identitu 2 vzorků na dané pozici je považován každý neprázdný překryv variant mezi vzorky. S takto chápanou identitou pozice je spočítána průměrná sekvenční identita (ANI [7]) mezi vzorky, zde označená jako popANI [79].

Další představené nástroje, zařazené do tabulky 3.3, se vyznačují zaměřením na haplotypování, čili rekonstrukci koexistujících kmenů. Pohled na haplotypování a přehled dalších existujících nástrojů přináší Ghazi et al. [77], se zaměřením na nové nástroje vzhledem k roku 2022. Strainberry[59] je v té době vydaný software, který nebyl zmíněn. Strainberry je pipeline pracující s dlouhými čteními schopná separace několika jednotek haplotypů druhu ve velmi jednoduchých komunitách [59, 80]. Jako vstup je očekáváno assembly společné všem haplotypům (”*strain-oblivious*”) a složka mapovaných dlouhých čtení na toto assembly. Strainberry používá Longshot [81] s HapCUT2 [82] na detekci SNVs a přidělení čtení k pravděpodobně příslušným haplotypům. Jsou to nástroje vytvořené pro práci s diploidním genomem a zde jsou využívány iterovaně k odlišení jednoho haplotypu po druhém. Dále se provádí sestavení contigů haplotypů, trimmování a *scaffolding*, na výstupu jsou genomy haplotypů (”*strain-aware*”) [59].

Fedarko, Kolmogorov a Pevzner [83] poukazují na nevhodnost LoFrequ pro použití na dlouhá HiFi čtení, jelikož u nich ”Phred skóre není dostupné vždy, [84]” a EM odhad parametrů nebere v úvahu různý kontext. Jejich metoda strainFlye se zaměřuje na práci s HiFi čteními, provádí haplotypování a přichází se zajímavým přístupem k míře falešných pozitivů detekce. Uplatňují přístup zvaný *target-decoy approach*, který v podstatě vybírá z analyzovaných contigů ten, jenž má nejmenší variabilitu ve svých čteních, a je tedy nejlepším přiblížením nulovému modelu sekvence bez variant, a vůči němu poměřuje významnost variací u dalších contigů. Protože ani onen vybraný contig nemusí být prost variant, bude tendence hodnotit detekované varianty kriticky. StrainFlye haplotypování vypadá tak, že se provede oprava čtení, aby zůstaly pouze ty variace, které byly explicitně detekovány, přidají se ”virtuální” čtení v místech malého pokrytí a na základě této sady se postaví de Bruijnův graf, s kterým je dále pracováno [83].

[6] V angl. linkage disequilibrium.

[7] Z angl. average nucleotide identity. Obecně znamená ANI u zarovnání podíl počtu pozic se shodnými bázemi ku počtu všech pozic zarovnání.

StrainyMAG [9] je pipeline pro rekonstrukci haplotypů mikrobiálních genomů z dlouhých čtení z platformy PacBio i ONT. Počáteční assembly se uskutečňuje pomocí modifikovaného metaFlye$_{\text{strain}}$, který rozlišuje SVs. Vlastní program pro haplotypování, Strainy, zvládá pracovat na datech z několika málo jednotek druhů, a StrainyMAG proto rozděluje sestavené contigy do jednoduchých skupin. Strainy provede haplotypování ze zarovnání dlouhých čtení na tuto zjednodušenou skupinu contigů, využije nástroje detekující SNVs, postaví "*connection graph*" propojující čtení překrývající se *informativními variantami*, rozdělí graf do "hustě propojených" podgrafů (pomocí "*label propagation algorithm*") odpovídajících haplotypům. Informativní varianty jsou zde takové, které mají potenciál rozlišit dosud nerozlišené haplotypy, a Strainy iterativně rozděluje popsaným způsobem haplotypy a aktualizuje sadu informativních variant. Výsledkem je nahrazení haplotypům společných sekvencí v assembly grafu sekvencemi příslušejícími jednotlivým haplotypům (v tabulce značeno *linear phasing*) a jejich propojení do MAGů haplotypů [9].

Posledním popsaným softwarem je Hairsplitter [85], nedávno zveřejněná pipeline pro haplotypování. Vstupy jsou tvořeny sadou dlouhých čtení, která mohou být hodně chybová, a referencí. Hairsplitter provede mapování čtení na referenci nástrojem minimap2 [86] a pro zbytek analýzy rozštěpí analyzované pozice do oken podél reference. Na těchto oknech použije vlastní modul pro klastrování čtení podle haplotypu na základě podobnosti sekvencí. Dále rekonstruuje sekvence haplotypů s pomocí softwaru Racon [87] a nakonec propojí sestavení z jednotlivých oken do scaffoldu s pomocí GraphUnzip [88] [85]. Pro klastrování čtení však nově přibyla alternativa v podobě modulu strainMiner [80]. StrainMiner se v daném okně snaží mezi čteními (pokrývajícími alespoň 60% okna) najít čtení kovariující ve svých SNVs. Provádí to pomocí transformace zarovnání čtení do matice nul a jedniček, kde jednička znamená přítomnost majoritní varianty a nula druhé nejčetnější varianty. Matici se snaží postupně rozdělit do bloků převládajících jedniček a bloků převládajících nul. Čtení pokrytá stejnými bloky pak tvoří vlastní skupinu pro haplotypování. Svým způsobem náročný úkol hledání maximálního bloku požadovaných vlastností je zde řešen formulací problému *celočíselného lineárního programování* a jeho zpracováním softwarem Gurobi [89]. Oproti původnímu modulu Hairsplitteru se dosahuje zhruba řádového zmenšení maximálního využití paměti [80].

3.4 Hodnocení analýzy vnitropopulační variability bakterií

Ani se znalostí algoritmických řešení z minulých sekcí se nedokážeme dovtípit funkčnosti programů v praxi. K tomu slouží testy a tzv. *benchmarking*. *Benchmarking* má pro nás podobu testu, kterým můžeme hodnotit různé programy nebo celé experimentální protokoly. Základem je mít testovací sadu dat ze vzorku

název a rok publikace článku	principy práce	vstup
LoFreq 2012	Optimalizovaný výpočet p-hodnot rekurentním vzorečkem s použitím jednoduché teorie pravděpodobnosti; EM odhad při absenci Phred skór.	Mapovaná čtení (a Phred skóre).
InStrain 2021	Filtrování čtení a zachování párů; uniformní model pro chyby; výpočet popANI pro porovnání vzorků.	Reference + mapování krátkých párových čtení (1 a více vzorků).
Strainberry 2021	Iterované diploidní odlišení haplotypů (Longshot, HapCUT2); sestavení rozdělených readů; trimmování; *scaffolding*.	Assembly (“strain-oblivious”) +mapování dlouhých čtení.
strainFlye 2022	Odhad míry falešných pozitivů (FDR) *target-decoy* přístupem: vezmi nejméně variabilní contig a polož jeho míru mutací (hodnoceno jako počet nalezených variant ku max. možnému počtu variant) jako normalizaci pro cílové (*target*) contigy.	HiFi dlouhá čtení (nebo podobná) + sestavené contigy.
StrainyMAG pipeline (Strainy) 2023	Rozdělení contigů do jednoduchých skupin; na každou skupinu Strainy: operuje s assembly grafem, postaví v grafu haplotypům specifické sekvence (tzv. *linear phasing*); zjednodušení grafu a propojení do MAGů haplotypů.	Dlouhá čtení.
Hairsplitter pipeline 2023	1. mapování na referenci (minimap2), 2. klastrování čtení dle haplotypu, 3. lokální sestavení sekvence haplotypu (využívá Racon), 4. *scaffolding* (využívá GraphUnzip).	Dlouhá čtení + reference.
StrainMiner modul 2024	Modul Hairsplitteru vykonávající bod č. 2: Předzpracování a přenesení do celočíselného lineárního programování (řešeno pomocí Gurobi).	Čtení zarovnané na referenci.

Tabulka 3.3 Souhrn vybraných nástrojů pro analýzu SNVs na základě sekvenačních dat z metagenomických vzorků.

DNA, u něhož byla zjištěna variabilita jinými spolehlivými postupy. Takto určené varianty budou standardem pro posuzování správnosti a ideálních výsledků dosáhne nástroj, který určí právě a pouze varianty standardu. Pozice s variantou detekovanou během testu označujeme jako *pozitivní* nálezy. Mezi nimi ty, které se shodují se standardem, jsou *správně pozitivní* (SP), a ostatní jsou *falešně pozitivní* (FP). Pozice, na nichž nedošlo k detekci varianty, jsou negativní nálezy a ty, u nichž ani ve standardu není přítomná varianta, jsou *správně negativní* (SN), zbylé jsou *falešně negativní* (FN).

Olson et al. [14] podává přehled o testování nástrojů detekce variability (SNVs a indely) v kontextu mikrobiálních kultur. Zde zopakujeme potřebu stejných testovacích sad pro všechny nástroje i úskalí tohoto přístupu při porovnávání nástrojů uzpůsobených pro různé druhy sekvenačních dat (např. krátká a dlouhá čtení). Pak má smysl neporovnávat nástroje samotné, ale kombinace sekvenačních platforem a s nimi kompatibilních programů. Dalším doporučením je zvolení testovacích dat tak, aby dobře reprezentovala data z oblasti zamýšleného použití a rovnoměrně ji pokrývala [14].

Bush et al. [90] provedli *benchmarking* popisu variant u mikrobiálních izolátů, se simulovanými krátkými MiSeq čteními a paralelně s NextSeq čteními jako testovacími sadami. V dostupných sekvencích genomů mnoha různých kmenů 10 klinicky relevatních druhů bakterií změnili náhodně báze zlomku pozic a simulovali sady čtení. Každou sadu zvlášť zarovnali na původní genom nebo na jiný genom téhož druhu vysoké kvality. Testovali kombinace různých programů pro mapování a pro detekci SNVs. Studie mimojiné ukázala zásadní vliv sekvenční divergence mezi referencí a sekvenovaným genomem na validitu detekce variant. Detekce používající vzdálenou referenci vedly k horším výsledkům [90]. Použití MAGů jako referencí se tak jeví jako dobrá volba.

Andreu-Sánchez et al. [91] jako možná jediní provedli *benchmarking* na simulovaných metagenomických datech (simulovaná krátká čtení ze skupiny genomů bakterií střevního mikrobiomu). Testovali 7 nástrojů detekce SNVs za nastavení, kde nefigurovalo de novo sestavování a mapování na reference proběhlo pro všechny případy stejně. Překvapivým závěrem byly lepší výsledky pravděpodobnostních nástrojů nezaměřujících se na metagenomiku (Mutect2 a HaplotypeCaller) v porovnání s dedikovanými nástroji InStrain a metaSNV [92] [91]. Za zmínku stojí též benchmark [29] (v předtisku) zahrnující nástroje detekce variant, které využívají metody *hlubokého učení*, a testuje je na dlouhých čteních z mikrobiálních genomů. Ačkoli jsou k jejich trénování používána data z lidských genomů, výsledky studie jsou slibné pro použití na mikrobiálních datech [29].

Závěr

Příprava knihovny z celkové DNA mikrobiální komunity a následná sekvenace skrývá potenciál k popisu přítomných druhů i k rozlišení kmenů koexistujících v populaci. Rozvoj sekvenačních metod společně s vývojem bioinformatických nástrojů umožňuje studium variability mikrobiálních komunit až na úrovni celých genomů. Můžeme hovořit o populační genomice [66]. V této práci jsme popsali tři nejpoužívanější sekvenační metody, abychom pochopili charakter poskytovaných dat. Ukázali jsme proces sestavování genomů, který je zásadní pro nástroje popisu variability: Buď poskytuje MAGy použitelné programem jako reference nebo program využívá grafových struktur používaných při sestavení k rozlišení variant či haplotypů. Přístup založený na de Bruijnových grafech se vedle OLC nebo hladových algoritmů [31] jeví jako mocný teoretický nástroj. Oproti hladovým algoritmům poskytuje abstrakci do grafů, pro něž existuje řada popsaných grafových algoritmů. A umožňuje řadu modifikací a rozšíření (připomeňme např. barevné a párové de Bruijnovy grafy).

Nabídli jsme vlastní diagram pro orientaci mezi postupy uplatnitelnými pro výzkum variability v mikrobiálních komunitách. Přiblížili jsme vybraný software s různými schopnostmi: detekce SVs, SNVs, obou typů variant a takový software, který nad rámec SNVs řeší i jejich zařazení do haplotypů. Jak u assembly tak detekce variant jsme ukázali možnosti hodnocení správnosti, validace pomocí testů (benchmarking) a naznačili relevantní veličiny pro hodnocení.

Existuje značné množství nástrojů detekce variant a pokračuje publikování nových. Například nově vyvinutý assembler metaMDBG [93] rekonstruující MAGy s použitím tzv. *minimizer-space* de Bruijnových grafů slibuje paměťovou a časovou úsporu a autoři uvažují o využití konceptu *minimizer-space* grafů pro rozlišení haplotypů [8]. Vzhledem k tomuto dynamickému vývoji, vzhledem k paradoxu lepších výsledkům nástrojů vyvinutých mimo oblast metagenomiky v porovnání s nástroji dedikovanými metagenomice [91] a vzhledem k malému množství benchmarků zaměřujících se na detekci variant z metagenomických dat jsme přesvědčeni o potřebě více zevrubného benchmarkingu. Do budoucna bude důležité mít dostatek reprezentativních testovacích dat, možnosti unifikovaně a podle standardizovaných postupů hodnotit každý nový nástroj s ostatními. Simulované sady dat sice mají výhodu v přesné znalosti přítomných variant (sami

[8] Citována je starší verze článku (předtisk, bez recenzního řízení), který toto zmiňuje v diskuzní části.

totiž varianty generujeme), ale vzhledem k důkazům horších výsledků assemblerů na reálných sekvenačních datech [49] je důležité testovat proces v celé své komplexitě. Navíc specializace různých nástrojů na data různých sekvenačních metod dále klade přednost na testování na reálných vzorcích, pravděpodobně sekvenovaných různými postupy.

Tyto benchmarky by mohly odhalit slabiny jednotlivých metod a možná nalézt kombinaci doplňujících se nástrojů, které by vedly k superiorním výsledkům. Po vzoru VarCapu, který vznikal s motivací studia mikrobiální evoluce v laboratorních kulturách a který je 7 let starý, můžeme pomýšlet na software integrující více nástrojů pro maximální zužitkování potenciálu bioinformatické analýzy.

Seznam použité literatury

[1] Carl R Woese a George E Fox. "Phylogenetic structure of the prokaryotic domain: the primary kingdoms". In: *Proceedings of the National Academy of Sciences* 74.11 (1977), s. 5088–5090.

[2] Matthew B Scholz, Chien-Chi Lo a Patrick SG Chain. "Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis". In: *Current opinion in biotechnology* 23.1 (2012), s. 9–15.

[3] Carola Simon a Rolf Daniel. "Metagenomic analyses: past and future trends". In: *Applied and environmental microbiology* 77.4 (2011), s. 1153–1161.

[4] Michael Schloter et al. "Ecology and evolution of bacterial microdiversity". In: *FEMS microbiology reviews* 24.5 (2000), s. 647–660.

[5] Anton Bankevich et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of computational biology* 19.5 (2012), s. 455–477.

[6] Matthew R Olm et al. "Consistent metagenome-derived metrics verify and delineate bacterial species boundaries". In: *Msystems* 5.1 (2020), s. 10–1128.

[7] Gabriele Ghiotto et al. "Strain-resolved metagenomics approaches applied to biogas upgrading". In: *Environmental Research* 240 (2024), s. 117414.

[8] Ou Wang et al. "Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly". In: *Genome research* 29.5 (2019), s. 798–808.

[9] Ekaterina Kazantseva et al. "Strainy: phasing and assembly of strain haplotypes from long-read metagenome sequencing". In: *bioRxiv* (2023), s. 2023–01.

[10] *Preparing whole genome and metagenome libraries using SMRTbell prep kit 3.0*. URL: `https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-whole-genome-and-metagenome-libraries-using-SMRTbell-prep-kit-3.0.pdf`. (navštíveno 24. 4. 2024).

[11] Michael L Metzker. "Sequencing technologies - the next generation". In: *Nature reviews. Genetics* 11.1 (2010), s. 31–46. ISSN: 1471-0056.

[12] Kevin J Travers et al. “A flexible and efficient template format for circular consensus sequencing and SNP detection”. In: *Nucleic acids research* 38.15 (2010), e159–e159.

[13] Jennifer A Loch et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456.7218 (2008), s. 53–59. ISSN: 0028-0836.

[14] Nathan D Olson et al. “Best practices for evaluating single nucleotide variant calling methods for microbial genomics”. In: *Frontiers in genetics* 6 (2015), s. 235.

[15] Richa Bharti a Dominik G Grimm. “Current challenges and best-practice protocols for microbiome analysis”. In: *Briefings in bioinformatics* 22.1 (2021), s. 178–193.

[16] Brent Ewing a Phil Green. “Base-calling of automated sequencer traces using phred. II. Error probabilities”. In: *Genome research* 8.3 (1998), s. 186–194.

[17] *HiSeq Sequencing Systems*. URL: https://www.illumina.com/content/dam/illumina-support/documents/products/datasheets/datasheet_hiseq_systems.pdf. (navštíveno 29. 4. 2024).

[18] *Sequencing platforms*. URL: https://www.illumina.com/systems/sequencing-platforms.html. (navštíveno 29. 4. 2024).

[19] *An introduction to Next-Generation Sequencing Technology*. URL: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. (navštíveno 24. 4. 2024).

[20] *Quality Scores for Next-Generation Sequencing*. URL: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf. (navštíveno 28. 4. 2024).

[21] *Sequencing 101: long-read sequencing*. URL: https://www.pacb.com/blog/long-read-sequencing/. (navštíveno 24. 4. 2024).

[22] M. J Levene et al. “Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations”. In: *Science* 299.5607 (2003), s. 682–686. ISSN: 0036-8075.

[23] *Sequencing 101: The SMRT Cell in PacBio long-read sequencing*. URL: https://www.pacb.com/blog/smrt-cell/. (navštíveno 24. 4. 2024).

[24] *HiFi reads for highly accurate long-read sequencing*. URL: https://www.pacb.com/technology/hifi-sequencing/. (navštíveno 24. 4. 2024).

[25] Miten Jain et al. “The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community”. In: *Genome Biology* 17.1 (2016), s. 239–. ISSN: 1474-7596.

[26] Daniel Branton et al. “The potential and challenges of nanopore sequencing”. In: *Nature Biotechnology* 26.10 (2008), s. 1146–1153. ISSN: 1087-0156.

[27] *How basecalling works*. URL: https://nanoporetech.com/platform/technology/basecalling. (navštíveno 25. 4. 2024).

[28] *Our platform performance and accuracy*. URL: https://nanoporetech.com/platform/accuracy. (navštíveno 25. 4. 2024).

[29] Michael B Hall et al. “Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data”. In: *bioRxiv* (2024), s. 2024–03.

[30] Julliane D Medeiros et al. “Single-cell sequencing unveils the lifestyle and CRISPR-based population history of Hydrotalea sp. in acid mine drainage”. In: *Molecular ecology* 26.20 (2017), s. 5541–5551.

[31] Jay S Ghurye, Victoria Cepeda-Espinoza a Mihai Pop. “Focus: microbiome: metagenomic assembly: overview, challenges and applications”. In: *The Yale journal of biology and medicine* 89.3 (2016), s. 353.

[32] Ben Langmead. *Assembly in Practice: Part 1: OLC*. URL: https://www.cs.jhu.edu/~langmea/resources/lecture_notes/18_assembly_olc_v2.pdf. (navštíveno 29. 4. 2024).

[33] Jared T. Simpson a Richard Durbin. “Efficient de novo assembly of large genomes using compressed data structures”. In: *Genome research* 22.3 (2012), s. 549–556. ISSN: 1088-9051.

[34] Sébastien Boisvert, François Laviolette a Jacques Corbeil. “Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies”. In: *Journal of computational biology* 17.11 (2010), s. 1519–1533.

[35] Sergey Nurk et al. “metaSPAdes: a new versatile metagenomic assembler”. In: *Genome research* 27.5 (2017), s. 824–834.

[36] Dinghua Li et al. “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph”. In: *Bioinformatics* 31.10 (2015), s. 1674–1676.

[37] Fernando Meyer et al. “Critical assessment of metagenome interpretation: the second round of challenges”. In: *Nature methods* 19.4 (2022), s. 429–440.

[38] Jared T. Simpson et al. “ABySS: A parallel assembler for short read sequence data”. In: *Genome Research* 19.6 (2009), s. 1117–1123. ISSN: 1088-9051.

[39] Erwan Drezen et al. “GATB: genome assembly & analysis tool box”. In: *Bioinformatics* 30.20 (2014), s. 2959–2961.

[40] Rayan Chikhi a Guillaume Rizk. “Space-efficient and exact de Bruijn graph representation based on a Bloom filter”. In: *Algorithms for Molecular Biology* 8 (2013), s. 1–9.

[41] Raluca Uricaru et al. “Reference-free detection of isolated SNPs”. In: *Nucleic acids research* 43.2 (2015), e11–e11.

[42] Claire Lemaitre, Liviu Ciortuz a Pierre Peterlongo. “Mapping-free and assembly-free discovery of inversion breakpoints from raw NGS reads”. In: *Algorithms for Computational Biology: First International Conference, AlCoB 2014, Tarragona, Spain, July 1-3, 2014, Proceedigns 1*. Springer. 2014, s. 119–130.

[43] Steven Hofmeyr et al. “Terabase-scale metagenome coassembly with MetaHipMer”. In: *Scientific reports* 10.1 (2020), s. 10689.

[44] Evangelos Georganas et al. “HipMer: an extreme-scale de novo genome assembler”. In: *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*. Sv. 15-20-. ACM, 2015, s. 1–11. ISBN: 1450337236.

[45] Alexander Bowe et al. “Succinct de Bruijn Graphs”. In: *Algorithms in Bioinformatics*. Sv. 7534. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, s. 225–235. ISBN: 3642331211.

[46] Denis Bertrand et al. “Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes”. In: *Nature biotechnology* 37.8 (2019), s. 937–944.

[47] Sébastien Boisvert et al. “Ray Meta: scalable de novo metagenome assembly and profiling”. In: *Genome biology* 13 (2012), s. 1–13.

[48] Nathan D Olson et al. “Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes”. In: *Briefings in bioinformatics* 20.4 (2019), s. 1140–1150.

[49] Catarina Inês Mendes et al. “LMAS: evaluating metagenomic short de novo assembly methods through defined communities”. In: *GigaScience* 12 (2023), giac122.

[50] William C Nelson et al. “Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia”. In: *Applied and Environmental Microbiology* 82.1 (2016), s. 255–267.

[51] Astrid Lemire et al. “Populations, not clones, are the unit of vibrio pathogenesis in naturally infected oysters”. In: *The ISME Journal* 9.7 (2015), s. 1523–1531. ISSN: 1751-7362.

[52] Patrick T West, Rachael B Chanin a Ami S Bhatt. “From genome structure to function: insights into structural variation in microbiology”. In: *Current opinion in microbiology* 69 (2022), s. 102192.

[53] Moritz Smolka et al. “Detection of mosaic and population-level structural variants with Sniffles2”. In: *Nature biotechnology* (2024), s. 1–10.

[54] Kristen D Curry et al. “Reference-free Structural Variant Detection in Microbiomes via Long-read Coassembly Graphs”. In: *bioRxiv* (2024), s. 2024–01.

[55] David Heller a Martin Vingron. “SVIM: structural variant identification using mapped long reads”. In: *Bioinformatics* 35.17 (2019), s. 2907–2915.

[56] Fritz J Sedlazeck et al. “Accurate detection of complex structural variations using single-molecule sequencing”. In: *Nature methods* 15.6 (2018), s. 461–468.

[57] Jurgen F Nijkamp et al. “Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold”. In: *Bioinformatics* 29.22 (2013), s. 2826–2834.

[58] Jeffrey E Barrick et al. “Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq”. In: *BMC genomics* 15 (2014), s. 1–17.

[59] Riccardo Vicedomini et al. “Strainberry: automated strain separation in low-complexity metagenomes using long reads”. In: *Nature Communications* 12.1 (2021), s. 4485.

[60] Yingrui Li et al. “Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly”. In: *Nature biotechnology* 29.8 (2011), s. 723–730.

[61] Christopher Quince et al. “STRONG: metagenomics strain resolution on assembly graphs”. In: *Genome biology* 22 (2021), s. 1–34.

[62] Markus Zojer et al. “Variant profiling of evolving prokaryotic populations”. In: *PeerJ* 5 (2017), e2997.

[63] Todd J Treangen et al. “MetAMOS: a modular and open source metagenomic assembly and analysis pipeline”. In: *Genome biology* 14 (2013), s. 1–20.

[64] Tobias Rausch et al. “DELLY: structural variant discovery by integrated paired-end and split-read analysis”. In: *Bioinformatics* 28.18 (2012), s. i333–i339.

[65] Moritz Smolka et al. “Comprehensive structural variant detection: from mosaic to population-level”. In: *BioRxiv* (2022), s. 2022–04.

[66] Kateřina Burkartová et al. “Population Genomics of Microbial Biostalactites: Non-recombinogenic Genome Islands and Microdiversification by Transposons”. In: *Frontiers in Microbiology* 13 (2022), s. 828531.

[67] Yuejuan Li, Jiabao Cao a Jun Wang. “MetaSVs: A pipeline combining long and short reads for analysis and visualization of structural variants in metagenomes”. In: *iMeta* 2.4 (2023), e139.

[68] Samuel O'donnell a Gilles Fischer. "MUM&Co: accurate detection of all SV types through whole-genome alignment". In: *Bioinformatics* 36.10 (2020), s. 3242–3243.

[69] Mikhail Kolmogorov et al. "metaFlye: scalable long-read metagenome assembly using repeat graphs". In: *Nature methods* 17.11 (2020), s. 1103–1110.

[70] Heng Li, Xiaowen Feng a Chong Chu. "The design and construction of reference pangenome graphs with minigraph". In: *Genome biology* 21 (2020), s. 1–19.

[71] Zamin Iqbal et al. "De novo assembly and genotyping of variants using colored de Bruijn graphs". In: *Nature genetics* 44.2 (2012), s. 226–232.

[72] *breseq*. URL: `https://barricklab.org/twiki/bin/view/Lab/ToolsBacterialGenomeResequencing`. (navštíveno 6. 4. 2024).

[73] Ben Langmead a Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), s. 357–359.

[74] Paul Herrera et al. "Molecular causes of an evolutionary shift along the parasitism–mutualism continuum in a bacterial symbiont". In: *Proceedings of the National Academy of Sciences* 117.35 (2020), s. 21658–21666.

[75] *VarCap*. URL: `https://github.com/ma2o/VarCap/blob/master/README_varcap_3.0.txt`. (navštíveno 11. 4. 2024).

[76] Gavin M Douglas a Morgan GI Langille. "Current and promising approaches to identify horizontal gene transfer events in metagenomes". In: *Genome biology and evolution* 11.10 (2019), s. 2750–2766.

[77] Andrew R Ghazi et al. "Strain identification and quantitative analysis in microbial communities". In: *Journal of Molecular Biology* 434.15 (2022), s. 167582.

[78] Andreas Wilm et al. "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets". In: *Nucleic acids research* 40.22 (2012), s. 11189–11201.

[79] Matthew R Olm et al. "InStrain enables population genomic analysis from metagenomic data and sensitive detection of shared microbial strains". In: *Nature biotechnology* 39.6 (2021), s. 727.

[80] Tam Khac Minh Truong, Roland Faure a Rumen Andonov. "Assembling close strains in metagenome assemblies using discrete optimization". In: *BIOINFORMATICS 2024*. 2024.

[81] Peter Edge a Vikas Bansal. "Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing". In: *Nature communications* 10.1 (2019), s. 4660.

[82] Peter Edge, Vineet Bafna a Vikas Bansal. “HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies”. In: *Genome research* 27.5 (2017), s. 801–812.

[83] Marcus W Fedarko, Mikhail Kolmogorov a Pavel A Pevzner. “Analyzing rare mutations in metagenomes assembled using long and accurate reads”. In: *Genome Research* 32.11-12 (2022), s. 2119–2133.

[84] Yoshinori Fukasawa et al. “LongQC: a quality control tool for third generation sequencing long read data”. In: *G3: Genes, Genomes, Genetics* 10.4 (2020), s. 1193–1196.

[85] Roland Faure, Jean-François Flot a Dominique Lavenier. “HairSplitter: separating strains in metagenome assemblies with long reads”. In: *Proc. JOBIM*. 2023.

[86] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18 (2018), s. 3094–3100.

[87] Li Fang a Kai Wang. “Polishing high-quality genome assemblies”. In: *Nature Methods* 19.6 (2022), s. 649–650.

[88] Roland Faure, Nadège Guiglielmoni a Jean-François Flot. “GraphUnzip: unzipping assembly graphs with long reads and Hi-C”. In: *bioRxiv* (2021), s. 2021–01.

[89] *Gurobi*. URL: `https://www.gurobi.com/`. (navštíveno 19. 4. 2024).

[90] Stephen J Bush et al. “Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines”. In: *Gigascience* 9.2 (2020), giaa007.

[91] Sergio Andreu-Sánchez et al. “A benchmark of genetic variant calling pipelines using metagenomic short-read sequencing”. In: *Frontiers in genetics* 12 (2021), s. 648229.

[92] Paul Igor Costea et al. “metaSNV: a tool for metagenomic strain level analysis”. In: *PloS one* 12.7 (2017), e0182392.

[93] Gaëtan Benoit et al. “Efficient High-Quality Metagenome Assembly from Long Accurate Reads using Minimizer-space de Bruijn Graphs”. In: Cold Spring Harbor Laboratory Preprints, 2023.