VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2020 Jan Petrovský

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

TECHNIKY MAPOVÁNÍ GENOMU K REFERENČNÍ SEKVENCI
MAPPING TECHNIQUES FOR REFERENCE-BASED GENOME ASSEMBLY

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Jan Petrovský**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Vojtěch Bartoň**
SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Student: Jan Petrovský ***ID:*** 200211
Ročník: 3 ***Akademický rok:*** 2019/20

NÁZEV TÉMATU:

Techniky mapování genomu k referenční sekvenci

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma sestavování genomu z NGS dat. Zaměřte se především na mapování genomu k referenční sekvenci. 2) Vypracujte přehled metod mapování genomu a k tomu používaných nástrojů. 3) Vytvořte testovací dataset NGS dat s přesně definovanými pozicemi čtení. 4) Navrhněte metriku pro vyhodnocení kvality mapování a otestujte ji. 5) Namapujte dataset pomocí několika vybraných nástrojů. 6) Výsledky kvalitativně vyhodnoťte. 7) Diskutujte vhodné nastavení parametrů jednotlivých nástrojů pro různé typy organismů.

DOPORUČENÁ LITERATURA:

1) Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28: 3169-3177

2) Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, et al. (2014) Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed Res Int 2014: 309650

Termín zadání: 3.2.2020 ***Termín odevzdání:*** 5.6.2020

Vedoucí práce: Ing. Vojtěch Bartoň

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

ABSTRAKT

Bakalářská práce se zabývá nástroji používanými pro mapování genomu k referenční sekvenci. V teoretické části jsou popsány aktuálně využívané sekvenační technologie, ať už první generace nebo nejnovější technologie třetí generace, přehled nástrojů používaných pro sestavování genomu de novo, používané přístupy a nástroje pro mapování k referenci, program ART a s ním související datové formáty. Praktická část sestává z vytvoření testovacího datasetu NGS dat, vytvoření metrik vhodných pro vyhodnocení kvality mapování a jejich následné použití na vybrané mapovací nástroje.

KLÍČOVÁ SLOVA

sekvenační technologie, NGS, mapování genomu, Illumina, BWA, Bowtie2, Novoalign

ABSTRACT

The bachelor thesis deals with the tools used to map the genome to the reference sequence. The theoretical part describes the currently used sequencing technologies, whether the first generation or the latest technologies of the third generation, an overview of the tools used for de novo genome construction, approaches and tools used for mapping to the reference, the ART program and related data formats. The practical part consists of creating a test dataset of NGS data, creating appropriate metrics to evaluate the quality of the mapping and their subsequent usage on the chosen mapping tools.

KEYWORDS

sequencing technology, NGS, genome mapping, Illumina, BWA, Bowtie2, Novoalign

PETROVSKÝ, Jan. *Techniky mapování genomu k referenční sekvenci*. Brno, 2020, 72 s. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Vojtěch Bartoň,

PROHLÁŠENÍ

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

Brno ....

podpis autora

PODĚKOVÁNÍ

Rád bych poděkoval svému vedoucímu bakalářské práce panu Ing. Vojtěchu Bartoňovi za ochotu, konzultace, trpělivost a podnětné návrhy k práci. Dále bych rád poděkoval své rodině, která mě při práci podporovala.

Obsah

Úvod

Sekvenování DNA má množství aplikací ve vědě, medicíně nebo například v zemědělství. Mohou se sekvenovat celé genomy, případně jenom některé části genomů. Ve fylogenetice představuje DNA sekvenování možnost studovat fylogenezi (evoluční vývoj) organizmů. Sekvenování je běžně využíváno v oblasti studia genetické variability nebo transkriptomové analýzy kódujících i nekódujících úseků DNA.

Vývoj sekvenačních technologií nové generace vedl k vytváření milionů krátkých čtení v jediném běhu. Se zvyšující se délkou čtení a zdokonalováním technologií, jako jsou například *Ion Torrent* nebo *Illumina*, je potřeba vyvíjet nové, efektivnější mapovací nástroje. Bez znalosti principů a parametrů jednotlivých nástrojů nelze vybrat optimální mapovací program pro zkoumaná data. V osekvenovaném genomu se obvykle vyskytují sekvenační chyby a variace (např. opakující se úseky nebo polymorfismy). Pokud je v takovém případě vybrán nevhodný software, mohou výsledky způsobit mylné interpretace. Mapovacích softwarů je na trhu velké množství a ne všechny jsou vhodné pro namapování určitého typu organismu. Proto je pro biology důležité zvážit vhodnost jednotlivých nástrojů podle jejich přesnosti, výkonnosti a vlastností.

Proces mapování čtení k referenčnímu genomu stále zůstává časově náročným a vyžaduje vývoj rychlejších a přesnějších mapovacích nástrojů. Stávající nástroje vytvářejí různé kompromisy mezi přesností a rychlostí mapování. Navíc mnoho důležitých aspektů je během porovnávání nově vyvinutých nástrojů přehlíženo. Proto je důležité objektivní zhodnocení metod, které pokrývají všechny aspekty.

První část práce tvoří teoretický úvod, kde se zabýváme sekvenačními technologiemi. Jsou popsány jejich základní principy a současné využití ve vědě a jsou uvedeny jejich hlavní výhody a nevýhody. V teoretické části je také představeno a popsáno několik nástrojů pro mapování genomu k referenční sekvenci.

V praktické části práce se snažíme vyhodnotit testované mapovací nástroje s cílem určit jejich vhodnost pro různé typy organismů. Použili jsme 3 nástroje, které patří k nejznámějším a také často používaným: Bowtie2 a BWA využívající Burrowsova–Wheelerovu transformaci a Novolalign založený na hashovacích algoritmech. Vybrané nástroje jsme testovali pomocí navržených a naprogramovaných metrik. Tyto metriky sloužily ke komplexní analýze vybraných mapovacích nástrojů. Testování jsme prováděli na datasetu vytvořeném v simulačním programu ART, který simuloval platformu *Illumina MiSeq*. Dataset byl vytvořen ze sekvencí genomů 21 různých organismů.

1 Teoretický úvod

1.1 DNA

DNA, nebo-li deoxyribonukleová kyselina je jedním z typů nukleových kyselin. Je velice důležitá pro uchování genetické informace. Deoxyribonukleové kyseliny jsou přítomné téměř ve všech buňkách, nejčastěji v jádře. Výjimkou jsou bezjaderné buňky. Jejich význam spočívá v přenosu a uchovávání genetické informace a určování průběhu biosyntézy bílkovin v buňkách. Základní stavební prvek je tvořen bází, cukrem a fosfátem. Těmto třem látkám navázaným na sebe se říká nukleotid. Nukleotidy se skládají z cukru, fosfátu a jedné ze 4 nukleových bazí (A – adenin, G – guanin, C – cytosin, T – thymin). Báze rozlišujeme na purinové (A, G) a pyrimidinové (C, T).

1.1.1 Primární struktura

DNA je polymer v podobě řetězce nukleotidů. Pořadí těchto nukleotidů v polynukleotidovém řetězci nám udává primární strukturu.

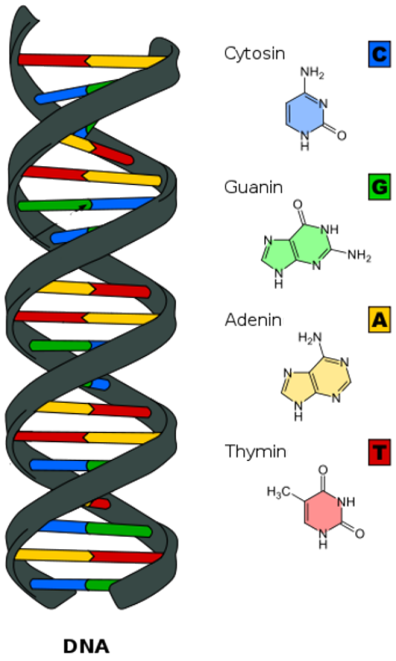
1.1.2 Sekundární struktura

Základní sekundární strukturou deoxyribonukleové kyseliny je alfa – Helix. Je to šroubovice dvou řetězců nebo dvou úseků téhož řetězce, který je stočen do tvaru připomínajícího spirálu. Oba řetězce dvoušroubovice jsou pravotočivé viz obrázek 1.1. To znamená, že cukry s navázaným fosfátem směřují k okraji, zatímco báze směřují k ose dvoušroubovice. Protilehlé báze jsou komplementární a jsou spojeny vodíkovými můstky. Adenin se páruje dvěma vodíkovými můstky s thyminem, zatímco guanin třemi vodíkovými můstky s cytosinem. Důležitou vlastností je opačná polarita fosfodiesterových vazeb obou řetězců. Nazýváme je antiparlelní. V jednom směru je směr vazeb 5'–3' a ve druhém 3'–5'. Druhou nejběžnější sekundární strukturou je beta – skládaný list. Zde jsou řetězce také dva, avšak jsou uspořádané rovnoběžně a antiparalelně ve dvou rovinách. Vzájemně jsou stabilizovány H-můstky. Kromě těchto dvou sekundárních struktur se vyskytuje i řada dalších, například Zn-prst nebo Leu-zip.

1.2 Genom

1.2.1 Základní charakteristika

Genom je veškerá genetická informace konkrétního organismu uložená v jeho DNA (u některých virů se nachází v RNA). Tímto pojmem se mohou zahrnout kódující



Obr. 1.1: Dvoušroubovice DNA převzato z [1]

i nekódující sekvence. Kódující úsek obsahuje důležitou informaci o stavbě a funkci buňky. Nekódující úseky jsou ty části genomu, které nejsou přepisovány v proteiny. Velmi často tvoří podstatnou část genomu. V případě eukaryotních organismů lze dále rozlišovat jaderný genom (genomická DNA) a mimojaderný genom (tvořený DNA sekvencemi v semiautonomních organelách - mitochondriích či plastidech). U člověka tedy obecně rozlišujeme mitochondriální a jaderný genom.

1.2.2 Projekt HUGO

Genom řady organismů byl v současné době celý osekvenován. *Human genome project* byl mezinárodní vědecký výzkum s cílem předpovědět sekvenci párů nukleotidových bazí, které vytváří lidskou DNA. Druhým cílem tohoto projektu bylo vyvinout nové technologie, pro mapování a sekvenování genomu. Je znám jako největší společný projekt v historii biologie. Původní myšlenka vznikla už v roce 1984 a formálně projekt započal v roce 1990, přičemž za hotový byl deklarován v roce 2003. Původně byl Human Genom Project určen ke zmapování nukleotidů v lidském haploidním referenčním genomu (více než 3 miliardy neukleotidů).[2, 3]

Genom jakéhokoli jedince je jedinečný. Základem výzkumu bylo sekvenování různých částí genomů u několika dobrovolníků, kteří patřili do etnicky různorodých skupin. Takto získané sekvence byly následně složeny dohromady, aby daly vzniknout kompletní sekvenci pro každý chromosom. Můžeme říci, že hotový lidský genom byl

jakousi mozaikou, která nereprezentovala jednoho jedince. Základní princip tedy byl, že získané sekvence dobrovolníků byly jednotlivě pomnoženy do několika milionů kopií. Tím se získal vzorek sekvencí, ty na sebe ovšem nenavazovaly v odpovídajícím pořadí. Toto pořadí bylo určováno následně pomocí algoritmů za využití velmi výkonných počítačů. Tyto algoritmy zpracovaly mnoho milionů dat a porovnáváním jednotlivých částí lidských genomů určily pořadí jednotlivých párů bazí. Ty odpovídaly konečnému počtu u všech 23 párů chromozomů.[4]

2 Sekvenační technologie

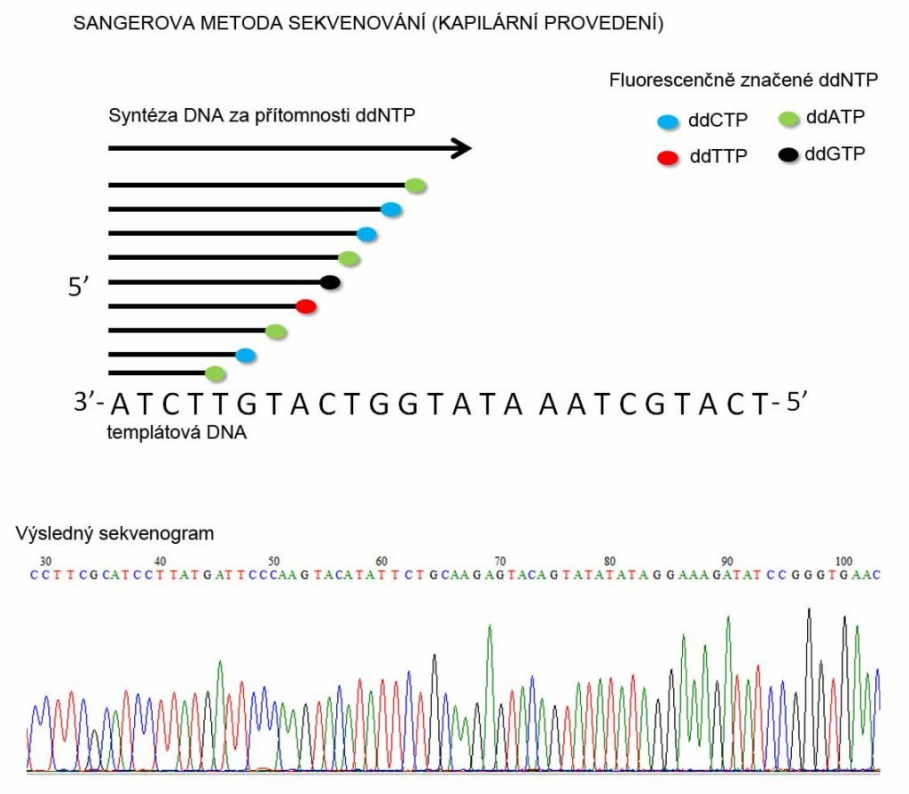
Hlavním úkolem sekvenačních technologií je určení pořadí nukleotidů v řetězci DNA, tzv. primární struktury. Ze znalosti primární struktury se vychází, při molekulárně-genetických analýzách biologického materiálu. V lékařství nachází široké uplatnění, především v oblasti diagnostiky dědičných chorob a nádorových onemocnění.

2.1 První generace

Po objevení sekundární struktury DNA se vědci začali zajímat o pořadí jednotlivých nukleotidů. To se povedlo objevit dvěma vědeckým týmům z Anglie a USA. V druhé polovině sedmdesátých let přišly na svět dvě různé, na sobě nezávislé metody. Byly pojmenovány podle svých tvůrců: Sangerova a Maxam–Gilbertova metoda [5]. Ačkoli obě metody přinesly svým autorům Nobelovu cenu, více používanou se stala Sangerova metoda, která nyní bude podrobněji vysvětlena.

2.1.1 Sangerova metoda

Pro samotné sekvenování nám nestačí pouze zastoupení jednotlivých nukleotidů, ale musíme vědět i jejich přesnou pozici. Toho jsme schopni dosáhnout pomocí Sangerovy metody. Tuto metodu publikoval v roce 1977 Frederick Sanger, po kterém je také pojmenována. Později mu byla právě za tento objev udělena Nobelova cena. Je jednou z nejjednodušších a zároveň nejzákladnějších myšlenek sekvenace genomu. Metoda je založena na principu terminace rostoucího řetězce. [7] Sangerova metoda může být také nazývána dideoxy sekvenační metoda. [8] V DNA jsou nukleotidy běžně označovány jako deoxynukleotidy. Obsahují 2–deoxyribózu (tzn. na druhém uhlíku ribózy chybí –OH skupina). Dideoxynukleotid obsahuje 2,3–dideoxyribózu (tzn. má deficit –OH skupiny na druhém i třetím uhlíku). Reakce probíhá přidáváním nukleotidů ve směru 5' - 3'. K syntéze je potřeba přítomnost jednovláknové DNA (teplátu), DNA – polymerázy, DNA primeru, normální deoxynukleosidtrifosfáty (dNTP) a jejich modifikované verze dideoxynukleosidtrifosfát (ddNTP), které jsou radioaktivně označeny. Dideoxyukleotidtrifosfáty tedy nemají volnou 3' -OH skupinu, která je nezbytná pro vznik fosfodiesterové vazby s následujícím nukleotidem. Jakmile se ddNTP začlení do rostoucí sekvence, proces prodlužování se ukončí. Princip metody si můžeme představit takto: Do čtyř různých zkumavek vložíme vzorek DNA. V jedné zkumavce se budou nacházet všechny čtyři typy „normálních“ nukleotidů (dATP, dGTP, dTTP, dCTP) a jeden dideoxynukleotid (ddCTP). Syntéza nového vlákna DNA se zastaví v případě, když se na templátovém řetězci objeví báze guaninu (řídíme se zde komplementaritou bází, tzn. A – T, G – C).



Obr. 2.1: Analýza signálu u Sangerovi metody, převzato z [7]

Koncentrace ddNTP je asi 100× nižší než dNTP. Tedy v našem případě na jeden ddCTP připadá 100 dCTP. Toho se využívá pro zjištění pozic dalších guaninů. Při syntéze je přítomno velké množství templátových vláken. Když se DNA–polymeráza dostane do místa, kde je na templátu přítomný guanin je pravděpodobnost 1:100, že se připojí ddCTP a dojde k ukončení elongace. Pokud se připojí dCTP, syntéza pokračuje, dokud se nepřipojí ddCTP. Z každé zkumavky dostaneme několik různě dlouhých úseků, které denaturujeme, aby se oddělila templátová vlákna od syntetizovaných. Vzorky následně vyhodnocujeme pomocí gelové elektroforézy. V agarózovém gelu se rozdělí jednotlivá syntetizovaná vlákna podle velikosti. Stejně postupujeme u Cytosinu, Adeninu a Thyminu. Nyní se tato metoda provádí pomocí kapilární elektroforézy. Používá se fluorescenční barvivo, tudíž použité reagence jsou netoxické. [7,9] Přístroje, které se používají, vyhodnocují výslednou sekvenci pomocí sekvenogramu. Analýzu signálu můžeme vidět na obrázku 2.1. [6, 5]

Sangerova metoda je využívána hlavně z důvodu, že poskytuje velmi dlouhé sekvence, které obsahují minimální množství chyb. Dodnes je široce využívaná, při sekvenování jednotlivých částí DNA. Například úseky DNA používané při klonování nebo pro PCR.

2.1.2 Maxam - Gilbertova metoda

Metoda, kterou vyvinuli Allan Maxam a Walter Gilbert, se někdy označuje jako „chemické sekvenování“. Vzorek obsahuje krátkou sekvenci DNA (dvouvláknovou tak i jednovláknovou), která je na svém 5’ konci radioaktivně označena fosforem 32P. Tento vzorek je rozdělen na čtyři oddělené části a každá [4] z částí je vystavena reagenciím, které specificky rozštěpí sekvenci DNA v místě, kde rozpoznají určitý nukleotid. Všechny sekvence DNA se umístí do polyakrylamidového gelu a je spuštěna elektroforéza. V gelu se všechny sekvence seřadí podle velikosti (nejdále v gelu doputují ty nejkratší sekvence). Dalším krokem je autoradiografie. Tedy přiložení tohoto gelu k filmu citlivému na rentgenové záření. Na vyvolaném filmu jsou pak všechny sekvence z gelu s 5’ koncem označeným radioaktivním fosforem patrné jako svítící proužky a na 3’ konci končí nukleotidem, na kterém proběhlo štěpení. V poslední fázi se podle pozice těchto proužků ve srovnání s ostatními proužky vyhodnocuje, jaké bylo původní řazení nukleotidů ve vzorku DNA.

2.2 Druhá generace

Pojmem Next Generation Sequencing (NGS) bývá označována druhá generace sekvenačních metod. Sangerova metoda byla revolučním objevem, avšak v moderní době nedostačuje požadavkům vědecké komunity. V roce 2005 nastal přelom s příchodem nové technologie sekvenace pomocí syntézy (sequencing by synthesis), vytvořenou firmou 454 Life Sciences.[8]

Pro NGS je charakteristické velké množství templátových řetězců, které se syntetizují paralelně. To vede k produkci většího množství sekvencí za velmi krátký čas. Všechny tyto metody mají stejné kroky. Příprava templátů, samotná sekvenace a interpretace, a vyhodnocení dat. Jedinečná kombinace postupů odlišuje jednu metodu od druhé a určuje typ dat vyprodukovaných z každé platformy. [9]

2.2.1 Pyrosekvenace

Pyrosekvenace je jednou z prvních technologií sekvenování druhé generace. Sekvenátory založené na pyrosekvenování jsou známy pod názvem Roche 454.

Na začátku máme dlouhou úsek DNA, který fragmentujeme. Dostaneme několik kratších dvouvláknových úseků o délce několika set párů bazí. Na konce těchto úseků připojíme adaptery. Poté je DNA denaturována na jednovláknovou. Adapter na konci je určen k pevnému spojení úseku s řádným povrchem. Tímto povrchem je kulička pokrytá streptavidinem. Na kuličkách je následně provedena PCR, která namnoží přichycený úsek. Na jedné kuličce se tedy nachází obrovské množství stejných fragmentů. Jednotlivé kuličky jsou následně přeneseny na pikotitrační destičku

tak, aby v jedné jamce byla právě jedna kulička. Do každé jamky jsou následně přidány enzymy. Při pyrosekvenaci jsou použity 4 enzymy a to DNA Polymeráza, ATP sulfyráza, Luciferáza a Apyráza. Reakční směs také obsahuje templátový řetězec s připojeným primerem, který je použit jako odrazový můstek pro DNA polymerázu. Čtyři druhy dNTP se přidávají v jeden moment, iterativně a cyklicky. To je umožněno díky minimální době trvání reakcí. Enzymatické reakce probíhající ve směsi jsou následující. První reakce DNA polymerizace nastane, pokud přidaný nukleotid vytvoří pár s nukleotidem na templátu a takto je začleněn do rostoucího vlákna DNA. V tomto případě dojde vlivem vznikající fosfodiesterové vazby k uvolnění pyrofosfátu. Ten okamžitě vstupuje do další reakce katalyzované sulfurylázou a vzniká ATP. Ve třetí a čtvrté fázi se za přítomnosti ATP přemění luciferin na oxyluciferin a to pomocí Luciferázy. Následně dochází k emisi viditelného záření. Světlo, které je vyprodukováno, se zaznamenává na velmi citlivou kameru s CCD čipem. Reakcí s aspyrázou se vymyje stávající dNTP a destička je promývána jiným dNTP. Požadované pořadí nukleotidů v sekvenci získáme analýzou flowgramů. S vyšším počtem začleněných dNTP dochází k vyšší emisi světla.[10, 11]

2.2.2 SOLiD

Technologie SOLiD (sequencing by oligonucleotide ligation and detecion) je založena na ligaci nukleotidových úseků, tudíž není zapotřebí používat DNA polymerázu.

Příprava vzorků je velmi podobná jako u pyrosekvenace. Fragmety DNA jsou ligovány na adaptery a pomocí emulzní PCR amplifikovány na kuličkách. DNA je denaturována a kuličky jsou umístěny na skleněnou destičku.

Vytvářejí se tzv. oktomery. Ty se skládají ze 2 specifických nukleotidů, 3 nukleotidů, které chceme zjistit, a 3 koncových nukleotidů označených barvivem. Pro každou dvojici nukleotidů je určena určitá barva. Využívají se 4 fluorescenční barviva, která jsou připevněna na 5‘ konec.

Do systému se přidá primer a poté se na referenční vlákno přidávají (na základě komplementarity bazí) oktomery podle toho, jaká dvojice nukleotidů je na začátku.[12] Ty, co se nehodí, jsou vymyty. Poslední 3 báze s barvivem se vymyjí pryč; toto uvolněné barvivo vydá světelný signál, který se detekuje. Po elongaci úseku požadované délky dojde k denaturaci úseku a na templátové vlákno nasedne primer, kratší o jeden nukleotid oproti předchozímu. Celkově je použito pět sekvenačních primerů, což zajistí, že každý nukleotid obsažený v řetězci bude přečten právě dvakrát. Na základě analýzy vlnové délky uvolněného světla, se sestavuje referenční sekvence po dvojicích nukleotidů. Ilustrovaný postup vidíme na obrázku 1.4.

2.2.3 Illumina

Tento sekvenátor byl poprvé představen firmou Solexa v roce 2006 a později dále vyvíjen firmou Illumina. Ze začátku byly výsledky srovnatelné se Sangerovou metodou. Délka výsledných čtení byla velmi krátká (asi 36 bp – párů bazí). Dnes se však díky technickému pokroku můžeme dostat až na délku 100 bp jednoho čtení. Tato metoda je v dnešní době jednou z nejvíce známých a používaných metod pro sekvenování. Je založena na syntéze řetězce DNA.

Příprava vzorků probíhá rozbitím velkého úseku DNA na menší fragmenty. Na konce těchto fragmentů se ligují adaptery. Všechny dále uvedené kroky se odehrávají na sekvenační destičce (flow cell). Flow cell je skleněné destička obsahující řádky. V každém řádku je položen „trávník“, který se skládá ze dvou typů adapterů. Závisejíce na specifické Illumina platformě může být destička rozdělena až na 8 oddělených kanálů.

Illumina používá „můstkovou“ amplifikaci. Sekvence pro amplifikaci je připevněna, komplementárně k sekvenci adapteru, k celému vnitřnímu povrchu řádků destičky. Jako první krok připevnění fragmentů na destičku je denaturace dvouvláknových fragmentů na jednovláknové. Jakmile je jeden adapter na destičce, pomocí hybridizace se k jejímu povrchu připojí i druhý a vytvoří tzv. most. Ten slouží jako předloha a DNA polymeráza vytvoří komplementární vlákno vytvářející dvouvláknový most. Tento most je denaturován a vznikají 2 vlákna připevněná jedním koncem k destičce. Tato vlákna se opět jedním koncem hybridizují k destičce. Tento proces se několikrát opakuje. Vznikají tak shluky stejných fragmentů tzv. klastry. Tomuto procesu se říká shlukování (clustering). Na destičce tak vzniká několik milionů klastrů, přičemž každý z nich tvoří identické molekuly DNA. Reverzní vlákna jsou rozštěpena a vymyta ze systému. Nyní je do systému potřeba přidat primery komplementární k adapterům, všechny 4 typy dNTP s fluorescenčním barvením a DNA polymerázu. Sekvenování začíná rozšířením prvního primeru. Každý ze 4 druhů dNTP (A, C, T, G) má svoje vlastní značení, které slouží k identifikaci báze a funguje jako reversibilní terminátor, který umožňuje přidání právě jednoho nukleotidu. Po zařazení jednoho dNTP jsou ostatní vymyta. Zároveň je fluorescenční barvivo odstraněno a reversibilní terminátor je deaktivován. Nyní je templátové vlákno připraveno pro přiřazení dalšího nukleotidu. Takto se proces cyklicky opakuje, dokud nedosáhneme chtěné délky sekvence. Tento proces se odehrává v každém kanálu na destičce. Díky této paralelizaci dostaneme velké množství čtení (až miliony) v krátkém čase.

Po ozáření klastrů laserem dojde k emisi světelného signálů. Pomocí analýzy těchto signálů, citlivými CCD čipy, dostaneme chtěnou sekvenci. Kamera v každém jednotlivém kole syntézy řetězce DNA snímá signál z celé destičky a podle rozdílné fluorescence pozná, jaké písmeno bylo přidáno u každého z milionů clusterů. Počítač

pak opět analyzuje záznam krok po kroku a podle toho, jak se mění fluorescenční signál v rámci každé skupiny (skládající se z identických molekul DNA), zrekonstruuje přesnou sekvenci molekul DNA v příslušné skupině. Ilustrovaný postup vidíme na obrázku 2.2.

Chybovost této metody je u různých jejích platforem různá (viz tabulka 1). Chyby mohou být způsobeny špatnou identifikací určitého nukleotidu. Systém také může vynechávat zkrácené sekvence.[13] Pro celogenomové de novo sekvenování malých genomů se doporučuje například platforma MiSeq, která umožňuje oproti ostatním sekvenátorům této společnosti číst delší úseky DNA, a pro sekvenování velkých genomů platformy NextSeq, HiSeq nebo NovaSeq.[14, 15]

2.2.4 Ion Torrent

Metoda Ion Torrent využívá emulzní PCR a svým principem spadá do kategorie sekvenování syntézou. Nejde však již o detekci světelného signálu. Namísto něj se detekuje změna pH vyvolaná uvolněním vodíkového iontu při inkorporaci dNTP DNA polymerázou.

Využívá se destičky s jamkami. Pod každou jamkou je čip ISFET (Ion-sensitive field effect transistor), který detekuje změny pH, které pozorujeme po vyloučení vodíkového iontu. Se změnou pH pozorujeme změnu napětí. [14]

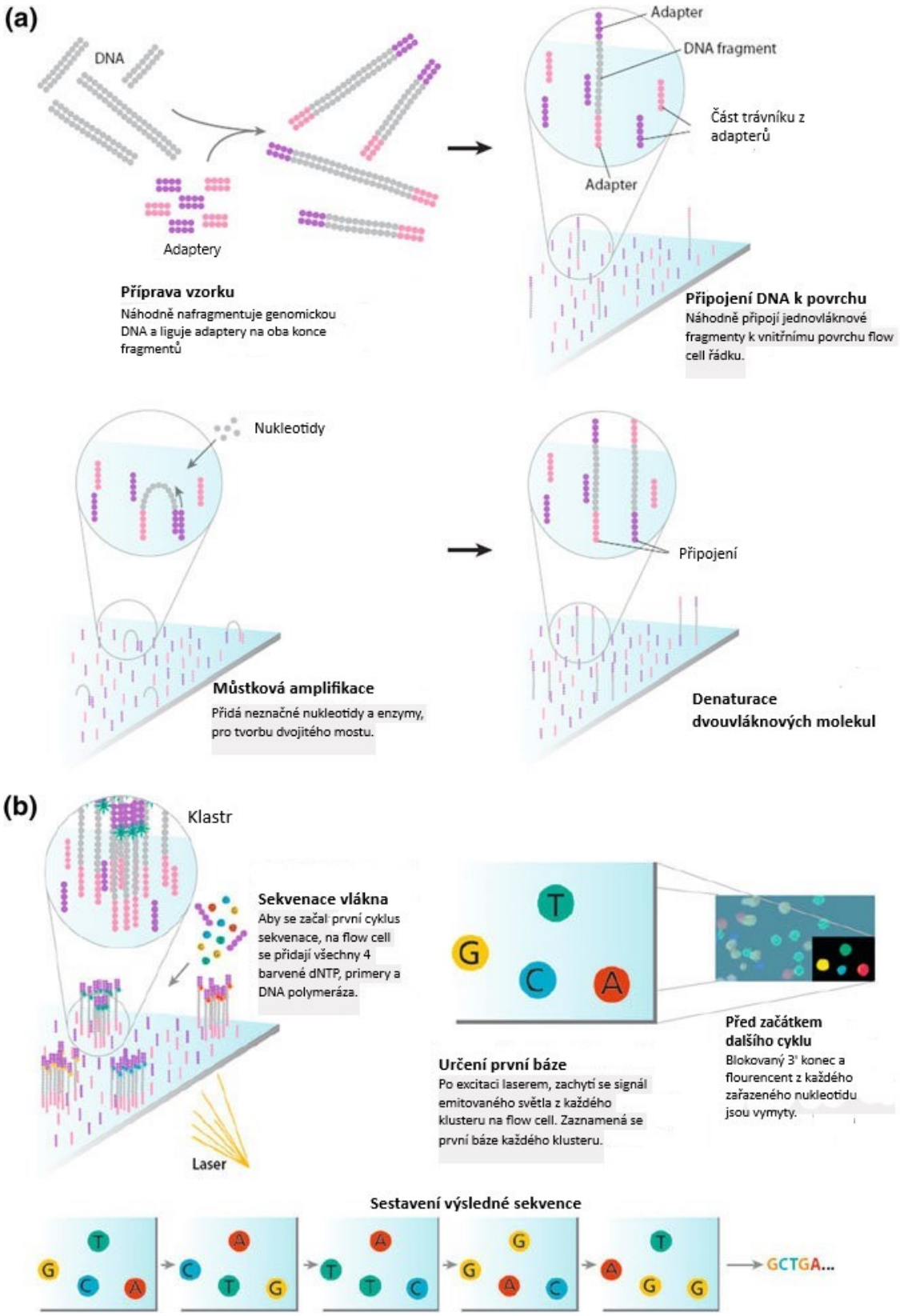
Příprava DNA je obdobná jako u pyrosekvenace 454. Všechny jamky jsou zaplaveny jedním ze čtyř typů dNTP. Pokud je dNTP uvnitř jamky komplementární k bázi templátu, je DNA polymerázou zařazen do syntetizovaného vlákna a je uvolněn vodíkový iont, který změní pH a to je zaznamenáno pomocí senzorů. Zbylé nukleotidy jsou vymyty a nahrazeny novými. Podle intenzity změny pH můžeme poznat, kolik nukleotidů bylo přiřazeno. K chybám může docházet u homopolymerních řetězců složených z jednoho druhu nukleotidů. V tuto chvíli nebude systém schopný rozpoznat přesný počet začleněních nukleotidů.

2.3 Třetí generace

Pro třetí generace je typická práce pouze s jednou molekulou DNA. Jsou rozděleny do 3 základních kategorií, které se navzájem liší. Každou kategorii bude reprezentovat jedna metoda.

2.3.1 SMRT

U této metody se pozorují jednotlivé molekuly DNA polymerázy, jak syntetizují jednu molekulu DNA.



Obr. 2.2: Postup můstkové PCR Illumina, převzato z [14]

SMRT (Single-molecule real-time sequencing) byla vyvinuta firmou Pacic Biosciences v roce 2009. Základem metody je zero-mode waveguide (ZMW). ZMW je díra o průměru desítek nanometrů ve 100 nm tlustém kovovém filmu položeném na skleněné destičce. Na skleněné dno každé dírky je přikotvena molekula DNA polymerázy a jedno vlákno templátové DNA. Jsou zde přítomny fluorescenčně barvené dNTP, kterými je celý systém zaplaven. Barevná značka je umístěna na konci fosfátové skupiny. Nukleotidy sestupují k DNA polymeráze, a pokud je detekován správný nukleotid, je zařazen do řetězce a uvolňuje se značka. Ve stejném čase detektor, umístěný u dna jamky, zaznamená barevný záblesk, jehož barva odpovídá určité bázi. Následuje připojení dalšího nukleotidu a cyklus se opakuje.

Tato metoda je schopna vytvářet velmi dlouhá čtení (až 10 000 bp), průměrně se však využívá délka 1 000 bp. Je velmi rychlá a navíc zde není potřeba vymývání a amplifikace, což celý proces ještě urychluje. Jako jediná zvládá detekovat změny na nukleotidech.[16]

2.3.2 Oxford nanopore

Nanopore – sekvenační technologie, ve které je jedna molekula DNA protažena nanopórem nebo umístěna v blízkosti nanopóru a jednotlivé báze jsou detekovány, jak ovlivňují elektrický proud nebo optický signál, když prochází nanopórem.

U této metody se využívají buď umělé, nebo biologické póry. Součástí každého systému je vždy nanopór a elektrolytický roztok. Systém se umístí do elektrického pole. Jednotlivé nukleotidy se rozpoznávají na základě velikosti proudu, který v systému vzniká.[16]

2.3.3 Halcyon Molecular

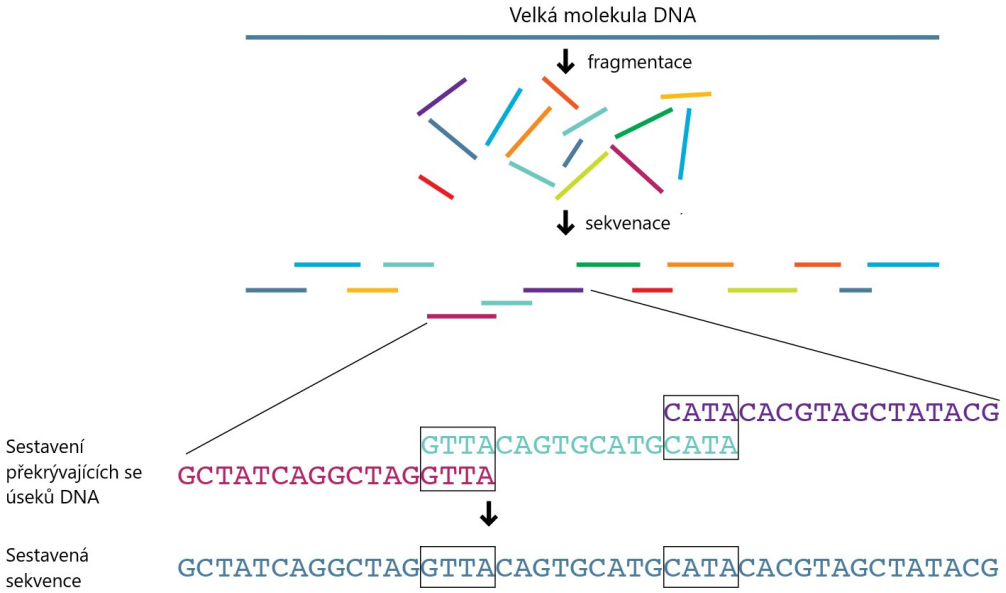
Tyto metody jsou založené na přímém zobrazení molekuly DNA pomocí specializovaných mikroskopických technik.

Pomocí transmisní elektronové mikroskopie (TEM) se snažíme přímo zobrazit a chemicky detekovat atomy, které by byly unikátní pro každý nukleotid v templátovém řetězci. Atomy se spolehlivě detekují sledováním odchylek na tmavém prstencovém poli, za využití neperiodického materiálu na plochém povrchu. [16]

2.4 Sekvenační techniky

2.4.1 Shotgun sekvenování

Je to technika založená na myšlence „rozbít“ velice dlouhou sekvenci na několik krátkých úseků. Tyto krátké úseky nasekvenovat a následně je, za pomoci výkonných



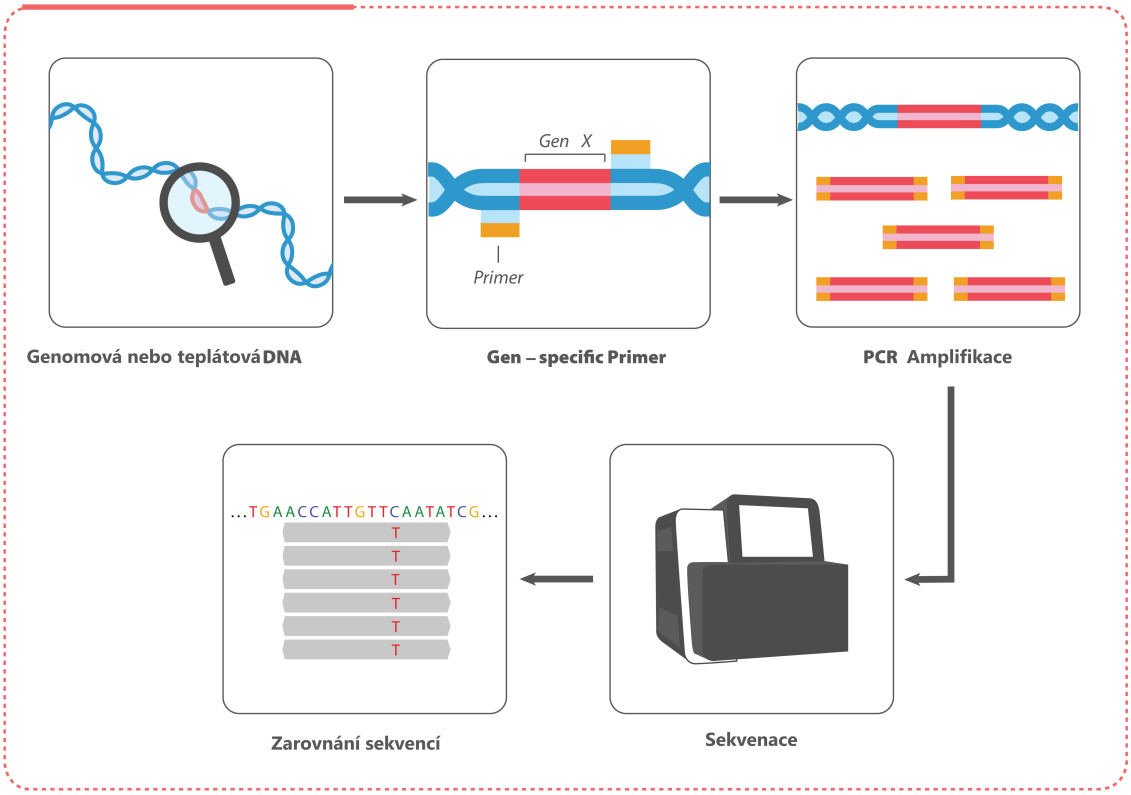
Obr. 2.3: Vizualizace shotgun sekvenace, převzato z [18]

počítačů a algoritmů, spojit do kompletní hledané sekvence.

Na začátku máme dlouhou sekvenci (např. lidský genom), který chceme sekvenovat. Tuto sekvenci rozdělíme na několik dostatečně malých úseků, které se mohou dále sekvenovat samostatně, nějakou z metod (Sangerova, NGS). Narušení původní sekvence se provádí náhodně. Vzniklé fragmenty DNA jsou v počítači analyzovány algoritmem, který hledá vzájemně identické úseky sekvence. Jakmile jsou tyto části identifikované jako stejné, vzájemně se překryjí a to umožní spojení dvou úseků sekvence dohromady. Jakmile jsou překrývající se úseky jednou seřazeny, algoritmus je schopen vytvořit celou sekvenci. Jelikož je fragmentování náhodné, nemusí se tato metoda řídit podle striktních pravidel, což se považuje za velkou výhodu. Některé nukleotidy mohou být ztraceny, tudíž chybí data v sekvenci. Pokud nenajdeme vzájemně se překrývající fragmenty, tato část sekvence zůstává neznámá. Přesně toto se stalo, když Shotgun sekvenovací metodu použili na Human Genome project. Velké množství regionů lidského genomu zůstalo neznámých, protože jsme nebyli schopni najít tyto sekvence. Grafické znázornění můžeme vidět na obrázku 2.3. [17]

2.4.2 Amplicon sekvenovaní

Jako amplikon nazýváme úsek DNA nebo RNA, který vznikl jako produkt PCR (polymerázová řetězová reakce) nebo LCR (ligázová řetězová reakce), popřípadě bě-



Obr. 2.4: Vizualizace amplicon sekvenování, převzato z [21]

hem replikace přímo v organismu. Amplicon sekvenace je vysoce přesný postup, umožňující vědcům analyzovat genetické změny ve specifických oblastech genomu.

Tato metoda je založena na hlubokém sekvenování (deep sequencing) amplikonů, ve kterých dokáže zachytit genetické změny. Tato technika využívá oligonukleotidové sondy, které zaměří a zachytí úsek zájmu. Tento úsek si vymezíme pomocí dvou primerů. Chtěný úsek můžeme sekvenovat jednotlivě pomocí nějaké z NGS metod. Hluboké sekvenování umožňuje zaměřit se na jediný gen, popřípadě větší úsek genů. Grafické znázornění je vidět na obrázku 2.4.

Amplicon sekvenace má velké uplatnění - ať už v ověřování nebo objevování změn v určitém úseku genomu či ve sledování klonů. Velice užitečná je pro hledání vzácných tělesných mutací v komplexním vzorku (například nádory). Další zcela běžná aplikace této metody je metagenomika rRNA bakteriálního genu 16S různých druhů. Využití se najde také v taxonomických a fylogenetických výzkumech.[19, 20]

3 Sestavování genomů

Výsledkem sekvenování genomové DNA je soubor čtení zkoumaného organismu. Cílem sestavení genomu je z těchto čtení složit původní sekvenci DNA. Pokud tuto sekvenci neznáme, hovoříme o sestavení genomu „de novo“. Každou pozici genomu máme ideálně pokrytou více čteními, která se vzájemně překrývají (koncová část jednoho čtení je shodná s počáteční částí čtení druhého). O mapování hovoříme, jestliže existuje nějaká referenční sekvence a sestavovací algoritmus pouze staví sekvenci podobnou již známe sekvenci – v podstatě jde o zarovnání. Toto je výpočetně spíše jednodušší problém než sestavení genomu de novo a zvládne to obyčejný počítač. I při mapování se však musí počítat s chybami sekvenace.[22]

Úkolem de novo celogenomového sestavení je rekonstruovat z krátkých úseků kompletní sekvenci DNA, rozdělenou do jednotlivých chromozomů daného organismu. Většina sekvenátorů poskytuje čtení o délce desítek až tisíc bází. Pokud sekvenujeme celý genom, je potřeba zpracovat mnohem větší množství párů bazí – například lidský genom má přibližně 2,9 milionů párů bází. Pokud použijeme vysoké číslo pokrytí (napřílad 30×) výsledný počet čtení je velmi velký a dostáváme náročnou výpočetní úlohu. Aby se sestavení genomu stalo proveditelným, provádí se sekvenování s velkým pokrytím. Našim cílem je získat dlouhé překryvy na všech místech sekvence. Mohou se objevit i chyby v překryvech. Pravděpodobnost špatného zarovnání překryvů roste s kratší délkou čtení. Někdy mohou být chyby čtení opraveny jejich zarovnáním a porovnáním, avšak je těžké detekovat chyby polymorfických změn a někdy jsou tyto chyby započítány do celkové chybovosti čtení. [22]

3.1 Používané nástroje pro de novo

Algoritmů pro sestavení sekvence de novo je velké množství. V této sekci bude uvedeno několik vybraných.

3.1.1 SPAdes

SPAdes (St. Petersburg genome assembler) je sestavovací algoritmus navržený pro datové soubory jednobuněčných a vícebuněčných bakterií. Není tedy vhodný pro dlouhé genomové sekvence. Je založen na pricipu de Bruijn grafu.[23]

3.1.2 Velvet

Velvet je algoritmus, používaný (podobně jako SPAdes) pro sestavení krátkých sekvencí. Toho je dosaženo pohybováním de Bruijn grafů podle odstraňování chyb a díky zjednodušení opakovaných úseků.[24, 25]

3.1.3 SSAKE a ABySS

SSAKE je algoritmus pro sestavení genomu s využitím nepárových krátkých čtení stejné délky.[26, 25] Využívají se ABySS je de novo, paralelní algoritmus pro sestavení velkého genomu krátkých čtení. Oba softwary pracují v operačním systému Linux. [27]

3.2 Mapování k referenci

Hledání pozice na referenční sekvenci, která přísluší dané sekvenci čtení, je obtížné. Nejjednodušší způsob, jak toho docílit, je postupně porovnat čtení se všemi pozicemi v genomu. Takovýto způsob by pro genom délky A a čtení délky B vyžadoval přibližně A $\times$ B porovnání bází. Ve skutečnosti to je ale ještě složitější, protože při porovnání všech možných pozic se nebere v úvahu možnost existence mezer. Avšak existují algoritmy, které dokážou najít optimální zarovnání sekvencí včetně mezer také s použitím A $\times$ B kroků. Problém mapování k referenčnímu genomu je ovšem tak rozsáhlý, že ani tyto algoritmy nelze jednoduše využít. Pokud bychom chtěli provést resekvenování lidského genomu s úrovní pokrytí 20×, získali bychom zhruba 600 milionů čtení o délce 100 bp. Pro namapování každého z nich by bylo potřeba asi 3×10^{11} kroků – dohromady tedy $1{,}8 \times 10^{12}$ porovnání nukleotidů. Dnešní procesory pracují na frekvenci okolo 4 GHz – provedou tedy 4×10^9 operací za sekundu. Pokud by procesor zvládl v jedné operaci porovnat jeden nukleotid, trvalo by mapování genomu $4{,}5 \times 10^{10}$ sekund, tedy 1 427 let.[28]

Z těchto důvodů se u mapování používá několik technik, které značně zkracují potřebný čas. Hlavní technikou je indexování. Nad sekvencí genomu je vytvořen index, což je jakýsi rejstřík. Ten se potom využívá na prohledávání genomu. Dalo by se to přirovnat ke způsobu, jakým člověk využívá rejstřík v knihách. Tvorba indexu je náročná, avšak pro každý genom ho stačí vytvořit jen jednou a poté je možné ho využít pro všechna budoucí mapování. V současnosti nejpoužívanější technikou tvorby rejstříku (indexu) je tzv. Burrowsova- Wheelerova transformace. Dalším způsobem, jak urychlit mapování, je jeho přesnost. Nehledá se vždy optimální zarovnání. Je zde možnost výskytu chybně namapovaných čtení, což celý postup velmi urychlí. Porovnání vybraných sekvenátorů je uvedeno v tabulce 3.1 [28]

3.2.1 Používané přístupy

Pro většinu mapovacích nástrojů začíná mapování sestavením indexu pro referenční genom nebo čtení. Poté je tento index použit, aby našel korespondující genomickou pozici každého čtení. Je mnoho technik používaných pro tvorbu indexu, v této sekci dvě nejpoužívanější.

Tab. 3.1: Porovnání sekvenátorů [14]

Sekvenátor	**Délka čtení**	**Počet čtení**	**Čas**
Roche 454 GS Junior+	700	1×10^5	18 h.
Roche 454 GS FLX+	1000	1×10^6	23 h.
SOLiD LifeTech5500xl	75	15×10^9	24 h.
Illumina MiSeq	2×300	15×10^9	56 h.
Illumina HiSeq2500	2×250	300×10^9	60 h.
Ion Torrent PGM 316	>100	1×10^6	2 h.
Oxford Nanopore	>100 000	-	-

Nástroje využívající přístup hašovací tabulky jsou pomalé, avšak jsou dobře optimalizované pro různorodé genomy. Typickými nástroji zde jsou například MAQ, SOAP, Novoalign, SSAHA, GSNAP.

Druhým přístupem jsou algoritmy, využívající Burrows-Wheelerovu Transformaci (BWT). Typickými zástupci těchto nástrojů jsou BWA, SOAP2 a Bowtie2. BWT je rychlá, ale vhodná jen pro krátká čtení a genomy s malým stupněm polymorfismu.

Hash table algoritmy

Metody založené na rozdrcení (hash) se dělí na dva typy: rozdrcení čtení nebo rozdrcení genomu. Obecně je hlavní myšlenka pro oba typy stejná – a to sestavit hašovací tabulku pro podsekvence nebo genom. Klíč každého vstupu je podsekvence, zatímco výstupní hodnoty představují seznam pozic, kde přesně můžeme danou podsekvenci najít. [29]

V základu tento program přesně zarovnává obrovské množství dat, vytvořené současnými sekvenačními přístroji, pomocí vícekrokové strategie semínkových algoritmů. Tato strategie je založena na principu „zasaď a prodluž“. Srovnávaná sekvence je roztříštěna na k menších podvláken. Prvním krokem je pokus o lokalizování těchto k-tých podvláken skrze hašovací tabulku. Tento proces nazýváme semínková detekce. Tento krok je primárně určen pro zrychlení propustnosti krátkých čtení. Abychom předpověděli přesnou pozici čtení v referenční sekvenci, musíme čtení zarovnat. Druhým krokem je tedy provést rozsáhlé zarovnání semínek pomalejším, avšak přesnějším algoritmem dynamického programování, jako jsou například Smith–Watermanův nebo Needleman–Wunschův algoritmus. Hašovací tabulka je typ tabulky s vyšší strukturou indexování, do které můžeme nahlédnout.[30]

Burrows-Wheelerova Transformace (BWT)

Je to zpětná permutace znaků v řetězci, původně určená pro kompresi. BWT na rozdíl od předchozího přístupu zarovnává celá čtení, namísto semínek čtení proti podvláknům navzorkovaným z referenčního genomu. Je to účinná, data indexující metoda, která si při prohledávání datasetu udržuje poměrně malé paměťové využití paměti. BWT byla rozšířena na novou datovou strukturu podporující přesné párování, pojmenovanou FM-index. Zasloužili se za to P. Ferragina a G. Manzini.[31] Transformací genomu do indexu FM se vylepší vyhledávací výkonnost algoritmu v případech, kdy se jedno čtení páruje s několika oblastmi genomu. Vylepšená výkonnost však přichází s výrazně delší dobou vytváření indexu (ve srovnání s hašovacími tabulkami) [30, 28].

Filtrace pomocí q-gramů

Máme referenční genom a velké množství sekvencí. Jednotlivá čtení jsou nasekána na opakující se kousky. Hlavním úkolem je najít všechny shody mezi referencí a těmito kousky čtení. Reference a čtení od sebe mají vzdálenost k. Ke spočítání těchto k-vzdáleností se využívá Levenshteinova vzdálenost, která na rozdíl od Hammingovy vzdálenosti zohledňuje inserce a delece. Nástrojem, využívajícím tuto metodu je například ShRiMP2.[30]

3.2.2 Používané nástroje

GSNAP

GSNAP je nástroj pro indexování genomu. Hašovací tabulka je sestavena rozdělením referenčního genomu na opakující se oligomery (molekula složená z podjednotek) o délce 12, vzorkované každé 3 nukleotidy. Mapovací fáze začíná rodělením čtení na více podvláken. Poté se snaží najít vhodnou oblast pro každé z těchto podvláken a konečně nakombinuje všechny oblasti každého z podvláken, aby vytvořil finální sekvenci. Využívá se především pro zarovnání RNA-Seq a DNA-Seq datasetů s genomem. Zvládne zpracovat delší čtení a větší objemy dat. Aplikuje se na Illumina data a data získaná Sangerovou metodou. [32, 29]

Novoalign

Tento nástroj je založen na přístupu hašovací tabulky. Tabulka je podobně jako u GSNAP sestavena rozdělením čtení na překrývající se oligomery. Při mapování se používá Needleman-Wunschův algoritmus s penalizací afinních mezer, který nám poskytne ideální globální zarovnání. Novoalign akceptuje spoustu formátů dat. Nejčastěji to jsou FASTA a FASTQ. Využívá se především s platformou Illumina.[33]

MAQ

MAQ je nástroj využívaný na indexaci čtení. Algoritmus pracuje tak, že nejprve vytvoří několik hašovacích tabulek pro jednotlivá čtení. Následně je referenční genom skenován proti tabulkám, aby se zjistilo umístění mapovaných oblastí. Výhodou zde je, že využívá velmi málo paměti RAM počítače. Využívá binární kodování pro kompresi dat. Referenční sekvence je ve formátu FASTA, popřípadě GenBank. Na výstupu jsou data ve formátu map. MAQ je specializovaný na zpracování velmi krátkých sekvencí. Původně byl vyvinut pro zarovnání dat z platformy Illumina-Solexa. [29]

mrFAST a mrsFAST

Jsou to nástroje pro indexaci genomu. Vytvářejí hašovací tabulku pro indexování k- genů genomu. MrFAST (microread Fast Alignment Search Tool) a mrsFAST (microread Fast Alignment and Search Tool) [30] jsou založeny na stejné metodě, ale první z nich podporuje mezery a nespárování, zatímco druhý podporuje pouze nespárování a díky tomu má rychlejší běh. Je využíván pro platformu Illumina. [29]

Bowtie2

Bowtie je velmi rychlý, paměťově efektivní program pro zarovnávání krátkých sekvencí DNA do velkých genomů. Je založený na BWT. Začíná vytvořením FM indexu pro referenční genom a poté použije upravený porovnávací algoritmus Ferragina a Manzini, aby našel mapovanou oblast. Existují dvě hlavní verze Bowtie, a to Bowtie a Bowtie2. Bowtie2 je navržen tak, aby zvládal především čtení delší než 50 bps.

Co se týče lidského genomu, indexace Burrows-Wheelerovou transformací umožňuje zarovnat více než 25 milionů čtení za hodinu s využitou pamětí přibližně 1,3 GB. Bowtie2 rozšiřuje předchozí techniky Burrows-Wheelera o nový algoritmus zpětného sledování s vědomím kvality, který pracuje i při nespárovaní nukleotidů.[34]

Často pracuje s daty z Illumina sekvenátorů. Bowtie2 je open source [1] [29]

SHRiMP

SHRiMP (SHort Read Mapping Package) je software, který zarovnává čtení proti refenčenímu genomu. Primárně byl určen k zarovnávání krátkých čtení, které nemůžou mít velký stupeň polymorfismu. Také se využívá pro barevně reprezentované zarovnání. Je hojně využívaný s platformou AB SOLiD.

SHRiMP2, podobně jako jeho předchůdce, je schopen zarovnat čtení s velkou mírou polymorfismu a chybovosti. Vykazuje značné zrychlení oproti předchozí verzi.

[1] software s otevřeným zdrojovým kódem [35]

Oba systémy podporují FASTa a FASTAq vstupy. Na výstupu potom najdeme data ve formátu SAM, který bude popsán později. SHRiMP2 podporuje známé platformy jako jsou Illumina-Solexa, Roche-454 a AB SOLiD. [36, 37]

BWA

BWA (Burrows-Wheeler Alignment tool) je zarovnávací balíček založený na zpětném vyhledávání pomocí BWT. Využívá přídavné pole (založené na BWT) pro rychlé vyhledávání podsekvencí. Efektivně zarovnává krátké sekvence čtení oproti rozsáhlému genomu, jako je například lidský. Zohledňuje nespárování a mezery. Podporuje data pocházejí z Illumina sekvenačních přístrojů a také barevně reprezentovaná čtení z AB SOLiD. [28, 30]

SOAP2

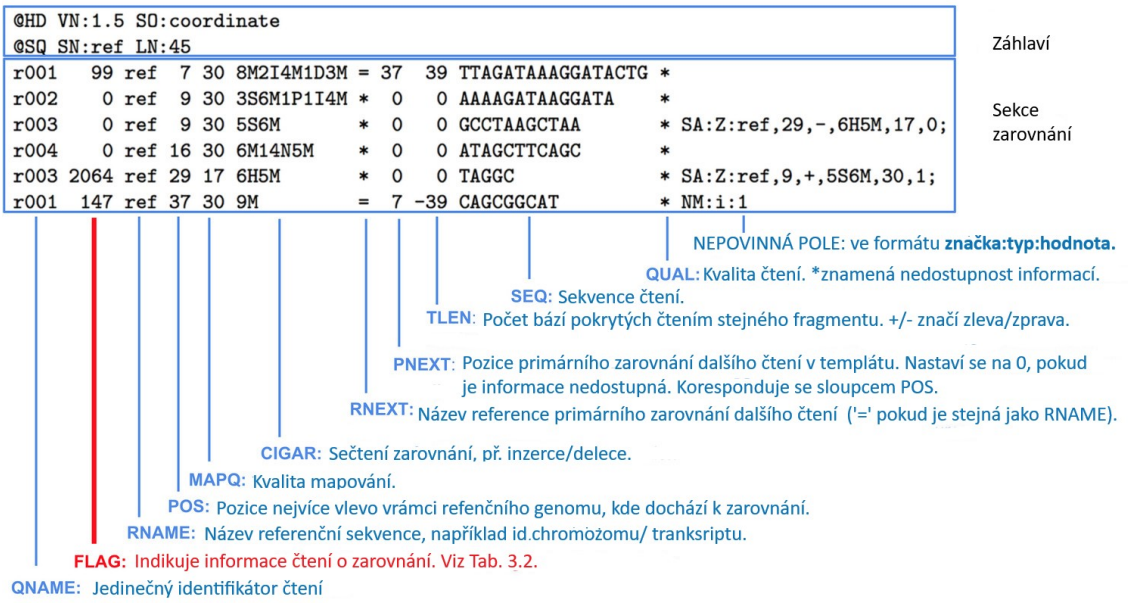
SOAP2 (Short Oligonucleotide Alignment Program) je určen pro superrychlé zarovnání krátkých čtení oproti dlouhé referenční sekvenci. Byl vyvinut ze svého předchůdce SOAP, který pro indexaci využíval semínkových algortimů. SOAP2 funguje jinak než ostatní nástroje založené na BWT. Pro indexování referenčního genomu používá techniky BWT a hašovací tabulky, což je užitečné k urychlení procesu přesného porovnávání. Používá „split-read strategii“, tj. rozdělí čtení na fragmenty na základě počtu neshod, aby nalezl nepřesné shody.

SOAP2 byl navržen pro Illumina GA sekvenování s délkou čtení přes 50bp. Referenční sekvence může být načtena jako textový dokument nebo FASTA formát. Chtěná čtení se načítají ve formátech FASTA a FASTAQ. Výstupní data obsahují SOAP textovou tabulku oddělenou tabulátory a SAM, popřípadě BAM data. SOAP2 je open source program. [38]

3.2.3 SAM/BAM formát

S rozvojem NGS technologií se vyvíjí i velké množství zarovnávacích nástrojů. Tyto nástroje generují zarovnání v různých formátech, což však komplikuje následné zpracování. Společný formát zarovnání, který podporuje všechny typy sekvencí a nástrojů pro zarovnání by vytvářel dobře definované rozhraní mezi zarovnáním a následnými analýzami. SAM (Sequence Alignment Map) byl navržen tak, aby dosáhl tohoto cíle. Je to TAB-delimited [2] textový formát pro ukládání zarovnaných sekvencí k referenční sekvenci. Dokáže škálovat zarovnaná data o délce 100 mld. párů bází. [39, 40]

[2]TAB-delimited složka je textový formát pro ukládání dat v tabulkové struktuře.



Obr. 3.1: Ukázka a popis SAM formátu, převzato z [41]

Pro vylepšení výkonu se navrhl doprovodný formát BAM (Binary Alignment Map). Je to binární reprezentace SAM a obsahuje totožné informace jako SAM.[39]

Skládá se ze záhlaví, které je volitelné, a ze sekcí zarovnání. Pokud je záhlaví přítomné, musí být před před zarovnáním. Každé zarovnání má 11 povinných polí se základními informacemi viz obrázek 3.1.

3.2.4 FLAG hodnoty

FLAG je přirozené číslo, které je součtem několika čísel, která reprezentují jednotlivé vlastnosti čtení. Významy jednotlivých hodnot můžeme vidět ve sloupečku Popis čtení v tabulce 3.2.4 . Zvolená vlajka (flag) je decimální reprezentace součtu vybraných bitů.

3.2.5 Phred skóre

Phred skóre (nebo Q skóre) je využíváno k posouzení kvality výstupu sekvenátoru. Indikuje pravděpodobnost, s jakou by konkrétní báze mohla být nesprávně přiřazena sekvenátorem. Parametry sekvenačního postupu se porovnají s empiricky získanými daty o známé přesnosti. Rovnice definující Phred skóre je: $Q = 10log10P$,kde P je pravděpodobnost nesprávného přiřazení báze. Phred skóre je logaritmicky přiřazeno k pravděpodobnostem chyby. Pokud je bázi přiřazeno skóre 30 (Q30), tak je to ekvivalentní výskytu jedné chyby z 1000. Tedy přesnost přiřazení báze je 99,9 procent. Skóre 20 (Q20) by znamenalo pravděpodobnost chyby 1:100. [42]

Tab. 3.2: Popis jednotlivých FLAG hodnot

#	**Decimální**	**Popis čtení**
1	1	Čtení spárováno
2	2	Čtení namapováno v řádném páru
3	4	Čtení nenamapováno
4	8	Sekvence nenamapována
5	16	Vlákno reverzního čtení
6	32	Vlákno reverzní sekvence
7	64	První v páru
8	128	Druhý v páru
9	256	Není primární zarovnání
10	512	Kontrola čtení na selhání platformy
11	1024	Čtení je PCR nebo optická kopie
12	2048	Doplňkové zarovnání

3.3 ART

Program ART je set simulačních nástrojů vytvářející umělá čtení NGS metod. Tato vlastnost je základem pro testování výkonnosti nástrojů pro analýzu NGS dat, včetně zarovnání a de novo sestavování. ART vytváří simulovaná sekvenační čtení emulací sekvanačního procesu s vestavěnými, pro každou technologii specifickými, chybovými modely a profily hodnot kvality bází empiricky parametrizovaných ve velkých datasetech.

ART aktuálně podporuje tři hlavní sekvenační platfomy nové generace: Roche 454, Illumina Solexa a Applied Biosystems SOLiD. Výstupní data, která dostaneme, jsou totožná s daty, která bychom dostali jako výstup některé z výše uvedených sekvenačních platforem. Také umožňuje flexibilitu při použití přizpůsobených parametrů modelu chyby čtení a profilů kvality. Právě pomocí programu ART jsme vytvořili dataset artificiálních dat, na kterém se následně testovalo. Program simuloval sekvenaci pomocí sekvenátoru Illumina MiSeq. Nastavení parametrů programu jsme ponechali v základním nastavení. Délka čtení byla 150 bp, chybové rozložení jsme zvolili na 10. Průměrné pokrytí bylo nastaveno na 10X.

Program načte soubor DNA sekvencí (reprezentující například referenční genom) a vytváří "umělá"sekvenační čtení způsobem, který kopíruje proces typický pro danou NGS technologii. ART má k dispozici set technologicky-specifických profilů chyb čtení, ale také je možné vytvářet uživatelem chtěné profily pro generování dat s přizpůsobenou délkou čtení a vlastnostmi chyb. ART podporuje všechny tři typy běžných sekvanačních chyb: substituce báze, inserce a delece. Standardně má tento

program na výstupu data ve formátu SAM a BAM. [43]

4 Metodika

4.1 Zvolená množina dat

Abychom prozkoumali kvalitu sestavení sekvenčních čtení poskytovaných platformami NGS, analyzovali jsme data sekvenováním celých genomů různých organizmů. Vybrali jsme si 21 zástupců různých organismů. Největší zastoupení budou mít bakterie, poté několik druhů ptáků a savců (viz tabulka 4.1). Sekvence všech bakterií mají podobnou délku s jedinou výjimkou a to Enterococcus aquimarinus, který je ve srovnání s ostatními sekvencemi kratší. Naopak genomy savců a ptáků jsou výrazně obsáhlejší. Nejdelšími sekvencemi jsou tedy Muss musculus a Macaca mulatta.

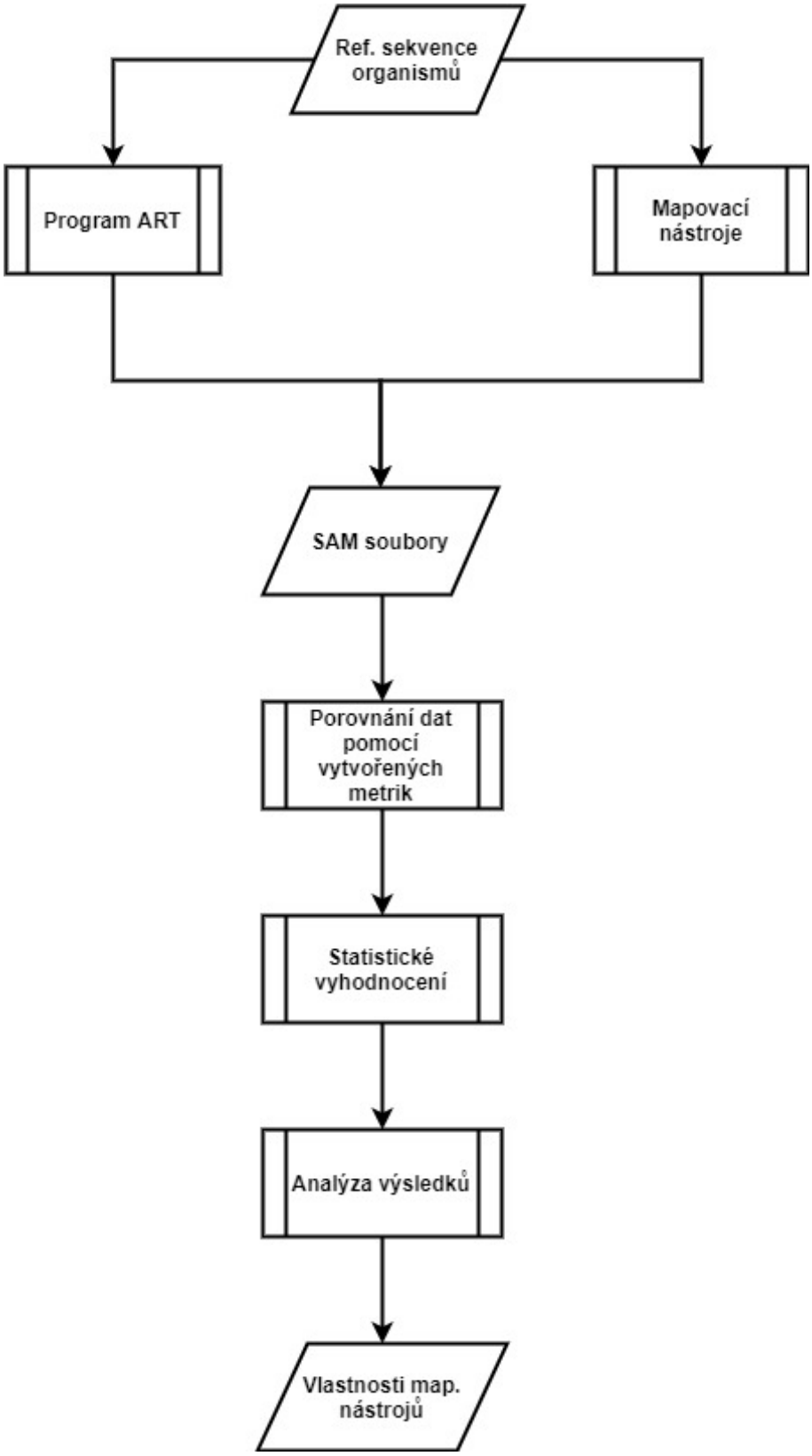
Následně se pomocí mapperu BWA namapovala sekvence a vznikl další SAM soubor, který se opět převedl do přehledné struktury. V našem případě jsme použili BWA-MEM. Takto se postupovalo také při mapování pomocí Bowtie2 a Novoalign. Jednotlivé vzniklé struktury, které byly na výstupech mapperů se porovnaly s daty vzniklými v programu ART. Data se porovnávala na základě vytvořených metrik uvedených v kapitole 4.2. Pro určení nejvhodnějšího nástroje pro určitý organismus, při zvoleném nastavení parametrů je potřeba testování provést pro všechny kombinace mapperů a testovacích metrik. Výše popsaný postup sloužil k získání výsledků pro každý z jednotlivých organismů uvedených v tabulce Tab. 4.1. Přesnou posloupnost metodiky ukazuje schéma viz Obr. 4.1

Pomocí parseru jsme vytvořený SAM soubor vložili do jednoduché a přehledné struktury (dataframu), ve které se nacházejí všechny informace o jednotlivých čteních (název, FLAG, CIGAR a další). Tímto způsobem se do do dataframu převedly všechny SAM soubory.

4.2 Metrika porovnání zvolených mapperů

Abychom zhodnotili kvalitu sestavení, tedy míru shody mezi referencí a sestavením, jednotlivých mapperů, uvažovali jsme několik hodnocených kritérií. V simulátoru ART byla délka čtení nastavena na 150 bp a standardní odchylka na 10. U všech tří mapperů jsme použili defaultní nastavení parametrů.

V prvé řadě jsme se zaměřili na úspěšnost namapování jednotlivých čtení. Přesněji tedy procentuální vyjádření namapovaných čtení z celkového počtu všech identifikovaných čtení. S tímto parametrem souvisí i další parametr, a to přesnost mapovaní. Zde se hodnotí čtení, která se namapovala správně. Tento parametr se vyhodnocoval na základě výpočtu Hammingovi vzdálenosti jednotlivých identifikovaných čtení získaných ze zvoleného mapovacího nástroje oproti referenčním čtením vygenerovaným v programu ART. Rovněž jsme analyzovali CIGAR-string a získali z něj



Obr. 4.1: Postup zpracování a vyhodnocení dat

Tab. 4.1: Organismy použité pro vytvoření datasetu

Název	**Doména/říše**	**Třída**	**Velikost**
Staphylococcus aureus	Bakterie	Bacilii	2,8 Mbp
Enterococcus faecalis	Bakterie	Bacilii	3,2 Mbp
Enterococcus hirae	Bakterie	Bacilii	2,8 Mbp
Enterococcus casseliflavus	Bakterie	Bacilii	3,4 Mbp
Enterococcus cecorum	Bakterie	Bacilii	2,3 Mbp
Enterococcus aquimarinus	Bakterie	Bacilii	335 kbp
Streptococcus pneumoniae	Bakterie	Bacilii	2,4 Mbp
Treponema brennaborense	Bakterie	Spirochaetes	3,1 Mbp
Treponema succinifaciens	Bakterie	Spirochaetes	2,7 Mbp
Treponema denticola	Bakterie	Spirochaetes	2,8 Mbp
Treponema pallidum	Bakterie	Spirochaetes	1,1 Mbp
Treponema pedis	Bakterie	Spirochaetes	2,8 Mbp
Mycobacterium tuberculosis	Bakterie	Actinomycetales	4,4 Mbp
Salmonella enterica	Bakterie	Proteobacteria	4,8 Mbp
Camarhynchus parvulus	Eukaryota	Ptáci	152 Mbp
Aquila chrysaetos	Eukaryota	Ptáci	854 Mbp
Limosa lapponica baueri	Eukaryota	Ptáci	5,4 Mbp
Hirundo rustica	Eukaryota	Ptáci	98 Mbp
Mus musculus	Eukaryota	Savci	195 Mbp
Olobus angolensis	Eukaryota	Savci	35 Mbp
Macaca mulatta	Eukaryota	Savci	223 Mbp

informace o počtech shod(M), neshod(X), inzercí(I) a delecí(D) jednotlivých bazí identifikovaného čtení, které jsou vhodným parametrem pro porovnání vybraných mapperů. Dnešním zarovnávacím nástrojům jsou kladeny požadavky na jejich rychlost a výpočetní náročnost. Proto je pro porovnání jednotlivých nástrojů jedním z klíčových parametrů doba běhu mapování, kterou jsme rovněž zahrnuli do testovaných parametrů. Byl použit PC s procesorem Intel Pentium G2130 (dvě jádra, takt 3,2 GHz) a 4 GB RAM.

Hammingova vzdálenost dvou řetězců udává, na kolika pozicích se vzájemně liší. Pro názorný příklad je na obrázku 4.2 uvedena dvojice řetězců, které se liší ve dvou znacích. Hammingova vzdálenost je tedy 2. Jinými slovy můžeme říci, že Hammingova vzdálenost měří minimální počet substitucí potřebných ke změně jednoho řetězce na druhý. Tato vzdálenost je také někdy nazývána Manhattanská vzdálenost.[44]

$$d_{ham}(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{4.1}$$

Obr. 4.2: Řetězce s Hammingovou vzdáleností 2.

5 Výsledky metrik

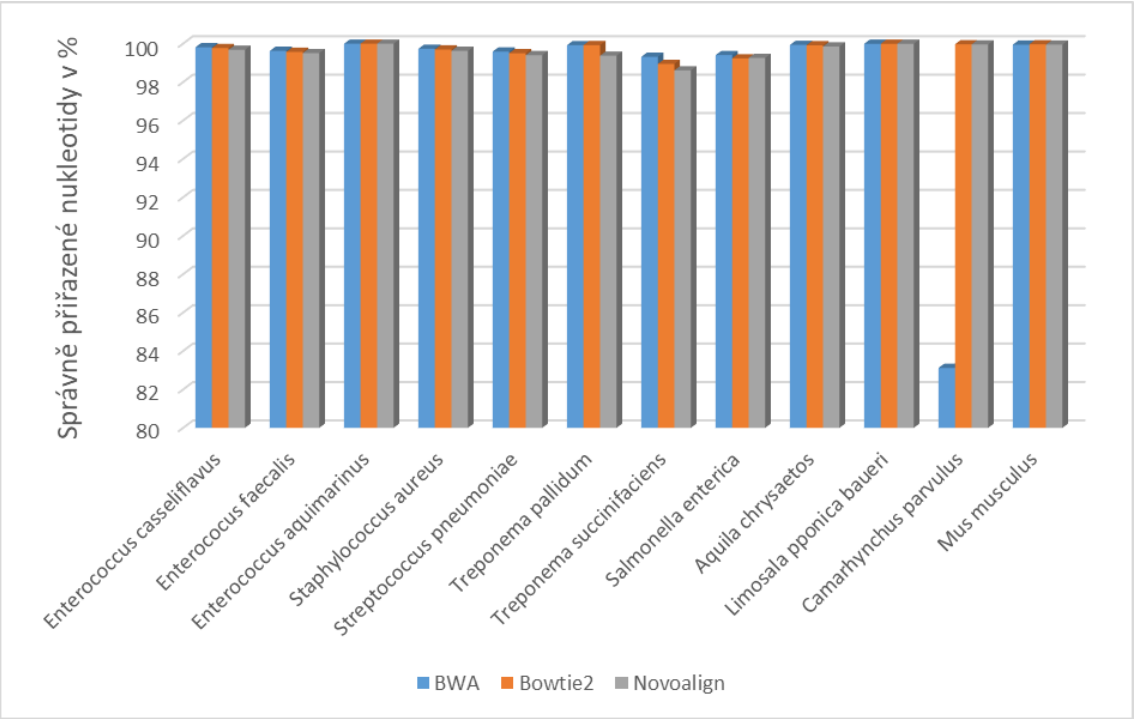
Výsledky získané pomocí metrik pro vyhodnocení kvality mapování jsem kvalitativně vyhodnotil pro tři mapovací nástroje, přičemž bylo použito všech 21 různých organismů. Výsledné skóre jednotlivých testů přikládám v příloze. Z důvodu, přehlednosti jsou v následujících podkapitolách vytvořeny grafy pouze pro několik organismů. Tyto organismy jsem vybíral na základě podobných výsledků v rámci stejné domény a třídy.

Pro snížení výpočetních nároků na zpracování testů jsem v případě velkých genomů testoval vždy pouze jeden mapper, a ne jako v případě menších genomů, kdy se daly testovat všechny tři mappery v jednom cyklu.

Jako nejvhodnější mapovací nástroj bakteriálních genomů se potvrdil BWA. Téměř u všech bakteriálních genomů dosahoval nejlepších kvality mapování kromě Treponemy Pedis, kde byla zaznamenána výrazně nízká hodnota p-distance. Časy mapovaní měl BWA u většiny organismů nejnižší, jedinou výjimkou byl organismus Mus musculus, kde byl Bowtie2 rychlejší. Co se týče Bowtie2, ten projevil nejlepší úspěšnost mapování téměř u všech organismů, včetně bakteriálních genomů. V oblasti živočichů poté dosahoval vyšších hodnot p-distancí než BWA i Novoalign, u bakteriálních genomů byl v tomto směru vhodnější BWA. Prakticky ve všech testech Novoalign lehce zaostával za jedním ze dvou dalších použitých mapperů. Za jeho hlavní nevýhodu by se dala označit velmi dlouhá doba mapování, která se ovšem zkracovala s rostoucí velikostí referenčního genomu. V oblasti živočišných genomů byly výsledky Novoalign srovnatelné s výsledky BWA, avšak jako nejvhodnější se v této oblasti potvrdil nástroj Bowtie2.

5.1 p-distance

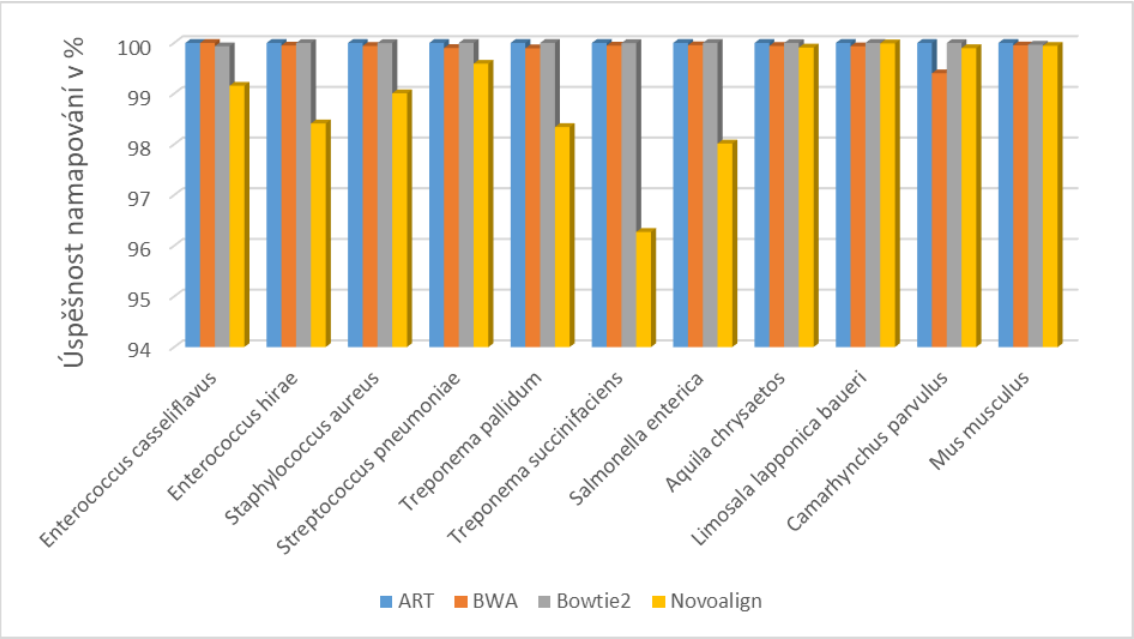
V grafu (viz Obr. 5.1) můžeme vidět srovnání p-distancí, tedy procentuální úspěšnost namapování správného nukleotidu oproti referenci, vybraných organismů. Je patrné, že bakteriální genomy byly nejpřesněji namapovány pomocí BWA. V doméně eukaryot měly všechny tři mappery podobnou přesnost, avšak nad ostatními lehce vyčníval Bowtie2. U organismů Enterococcus aquimarinus a Enterococcus cecorum se téměř ve všech případech namapovalo 100 % čtení. Tyto organismy měly také nejkratší sekvence a tudíž nejmenší počet čtení. Naproti tomu o mnoho delší sekvence, jako například Mus musculus nebo Aquila chrysaetos, měly úspěšnost namapování ve většině případů jen málo odlišnou. U 20 organismů se povedlo správně namapovat přes 98 % nukleotidů. Jedinou výjimkou byla bakterie Enterococcus hirae, u které p-distance nepřekročila 50 %.



Obr. 5.1: Srovnání p-distancí vybraných organismů.

5.2 Úspěšnost mapování

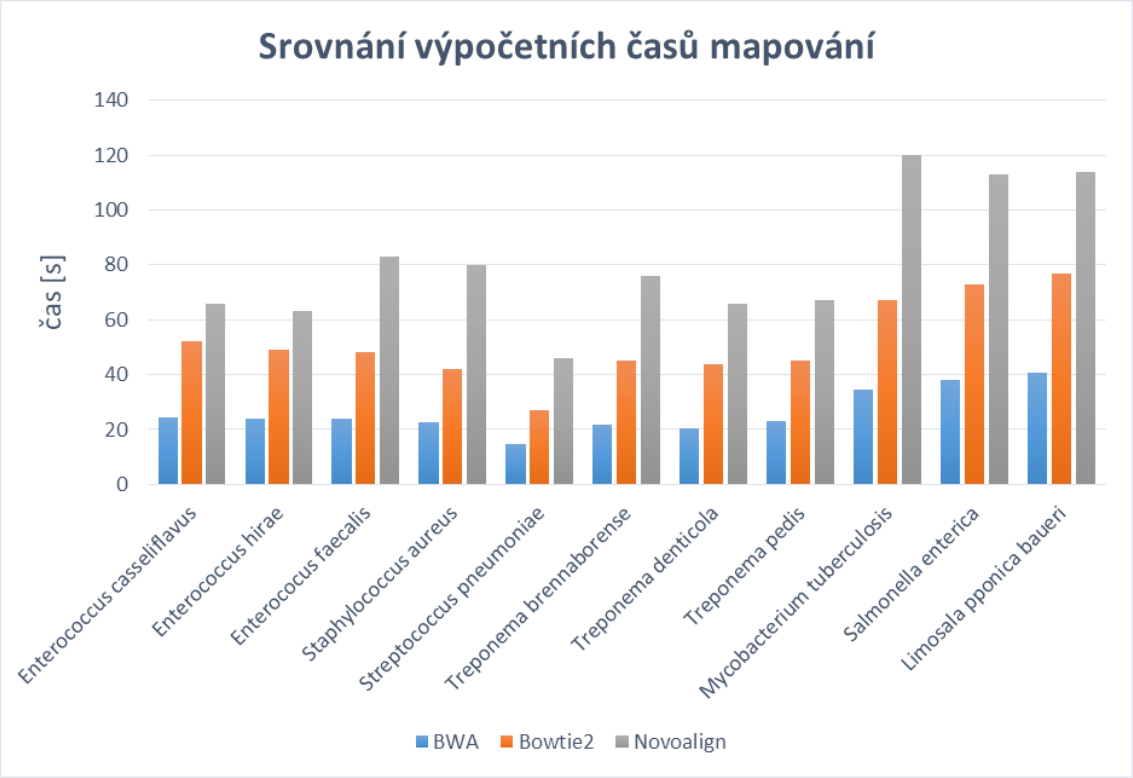
To, jestli se čtení namapuje, nebo nenamapuje je velmi důležitým aspektem a jde ruku v ruce i s dalšími parametry. V SAM souboru jsme zkontrolovali, jaké flag-hodnoty obsahuje a následně, pomocí webového nástroje pro dekódování flag-hodnot ze SAM formátu a naprogramovaného algoritmu, spočítali úspěšnost namapovaní. [45] Výsledky pro vybrané organismy lze vyčíst z grafu Obr. 5.2. Modrý sloupeček udává hodnotu pro referenční sekvenci vytvořenou pomocí programu ART. Jakožto referenční sekvence má také stoprocentní úspěšnost namapování. Jen o málo horší výsledky má Bowtie2, který dosahuje o několik procenta lepší úspěšnosti namapování než BWA. V případě nástroje Novoalign byl ve srovnání s dalšími dvěma mappery (ve většině případů) počet namapovaných čtení nízký . Pokud však vezmeme v úvahu celkovou dobu mapování jednotlivými mapovacími nástroji (viz Obr. 5.3 a Obr. 5.4) a u každého organismu porovnáme úspěšnost namapování právě s časem potřebným pro namapovaní daným mapovacím nástrojem, získáme lepší představu o efektivitě těchto nástrojů. Z tohoto hlediska se jako nelepší jeví BWA.



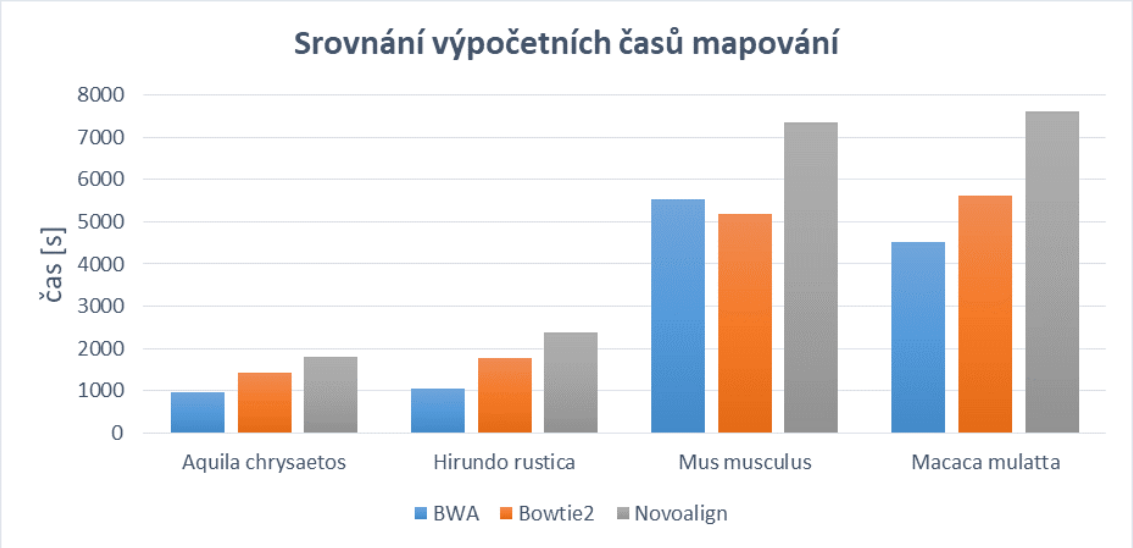
Obr. 5.2: Srovnání úspěšnosti namapování vybraných organismů.

5.3 Čas mapování

Časy mapování jsou zobrazeny pro všechny tři mappery (viz Obr. 5.3 a Obr. 5.4). Chtěli jsme zjistit dopad velikosti referenčního genomu na dobu mapování. Můžeme říci, že u všech mappery ukázaly lineární závislost mezi dobou mapování a velikostí referenčního genomu. Z průměrných výsledků pro dataset vytvořený pomocí sekvenační platformy Illumina shledáváme jako značně nejrychlejší BWA. Bowtie2 není nejrychlejší, avšak mapování pomocí tohoto nástroje bylo výrazně rychlejší než u Novoalign. Tyto výsledky by mohly být vysvětleny pomocí Burrows–Wheelerovi transformace u BWA a Bowtie2, což má za následek redukci doby mapování. Rozdíly jsou výrazné především u krátkých genomů, tedy převážně u bakterií. Se zvětšující se velikostí referenčního genomu se poměrově rozdíly v časech mapování zmenšují a hodnoty časů mapování Novoalign se alespoň lehce přibližují zbylým dvěma mapperům.



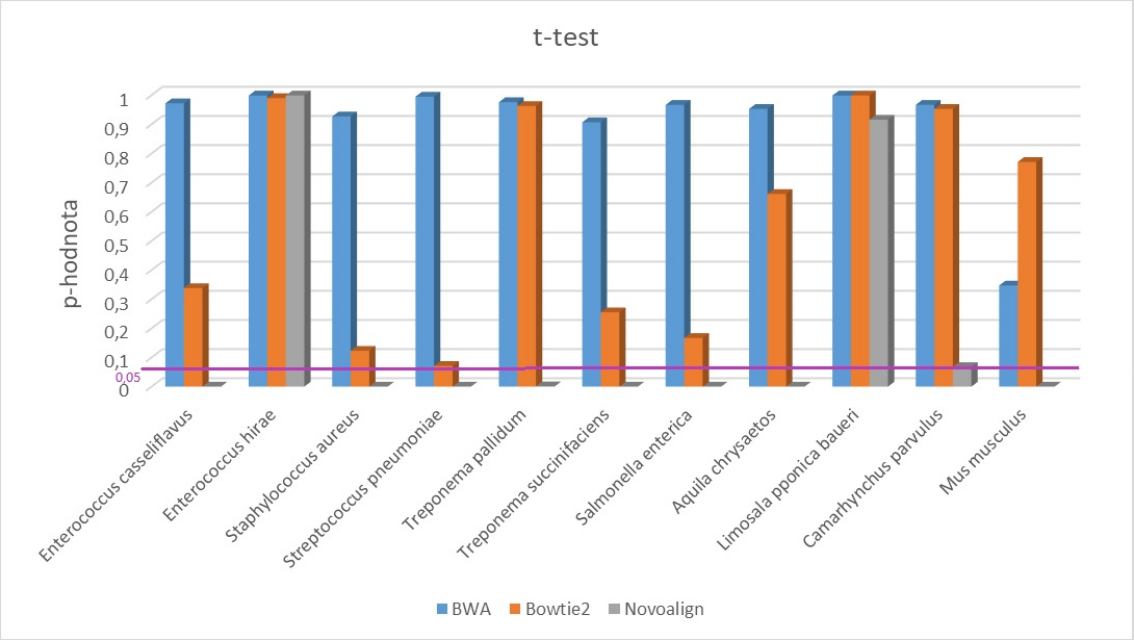
Obr. 5.3: Srovnání doby běhu mapperů pro vybrané organismy s krátkým genomem.



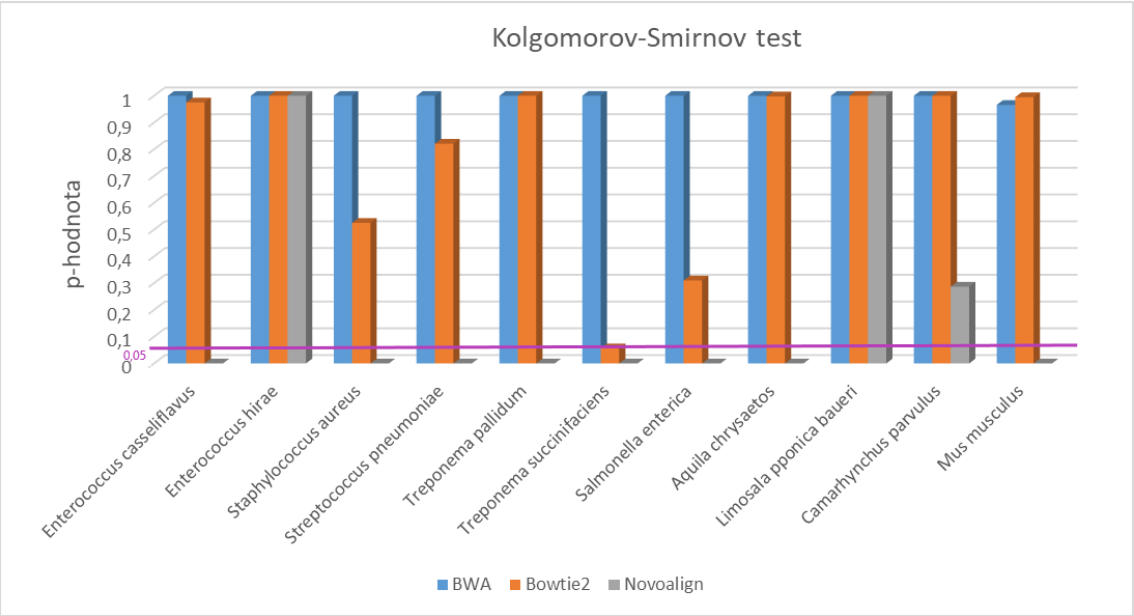
Obr. 5.4: Srovnání doby běhu mapperů pro vybrané organismy s dlouhým genomem.

5.4 Coverage

Při analýze sekvencí je velmi důležité jejich celkové pokrytí. V našem případě stačí vektor počátečních pozic jednotlivých namapovaných čtení. Vzhledem k tomu, že jednotlivá čtení se příliš neliší délkou a zároveň nepředpokládáme velký výskyt indelů, jak se nám potvrdilo při hodnocení CIGAR stringů, které by se projevily na celkové statistice, a tedy vektor počátečních pozic dobře aproximuje celkovou coverage celého genomu. Pro statistické vyhodnocení jsme použili t-test s testem párových diferencí. Zde testujeme nulovou hypotézu, jestli coverage z reference a z namapované sekvence pochází ze stejného normálního rozložení. Dále byl použit Kolgomorov-Smirnov test. Datasety byly seřazeny podle názvu čtení, aby si jednotlivé vzorky z obou distribucí odpovídaly, a následně byly testovány. Pro tyto účely jsme použili knihovnu *scipy*, která je dostupná pro Python. V případě t-testu je nutné si uvědomit, že t-test sleduje průměry obou distribucí, a tento parametr může být značně ovlivněn repeticemi a jejich pozicí v sekvenovaných genomech (obecně můžeme říci, čím více repetic, tím horší výsledky bude t-test mít). Výsledky testů jsou uvedeny v grafech Obr. 5.5 a Obr. 5.6. V obou grafech je naznačena hranice, při které ještě potvrzujeme nulovou hypotézu. Výsledky t-testu mají větší sílu než výsledky K-S testu. V případě BWA a Bowtie2 se u všech organismů potvrdilo normální rozložení vektorů počátečních pozic.



Obr. 5.5: Srovnání p-hodnot t-testu.

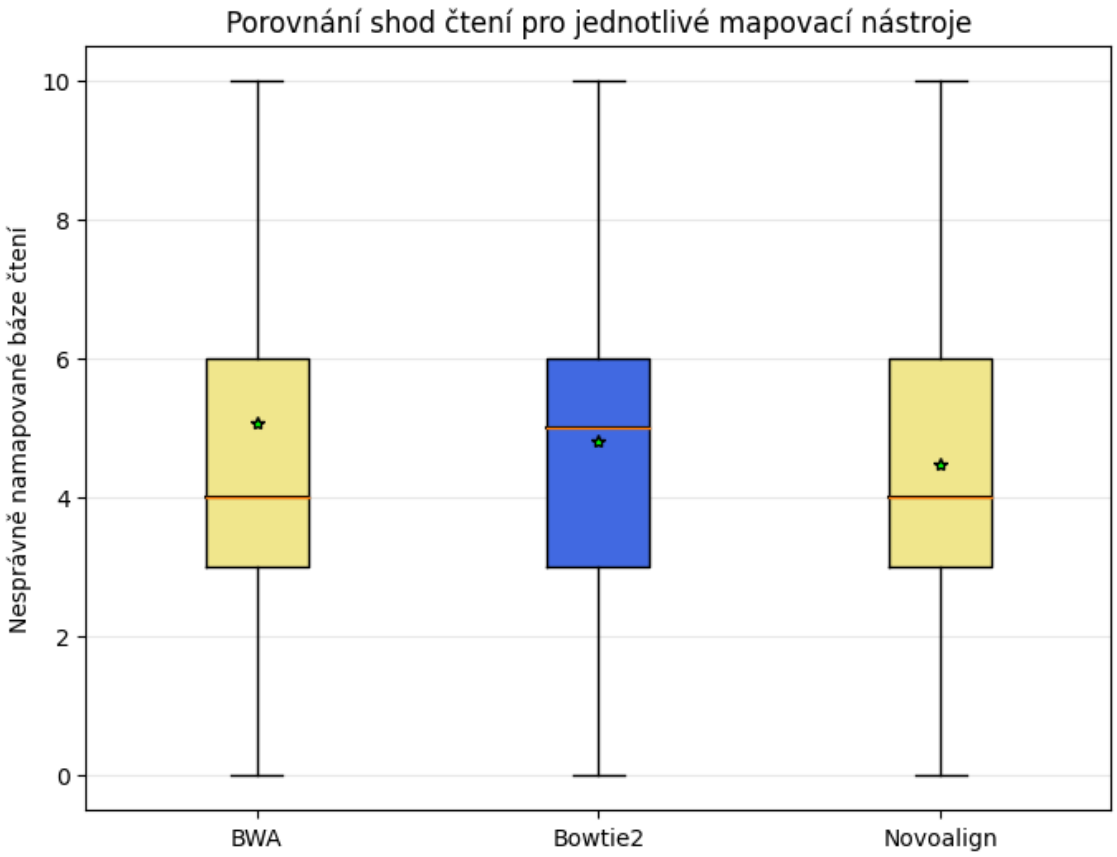


Obr. 5.6: Srovnání p-hodnot Kolgomorov-Smirnov testu.

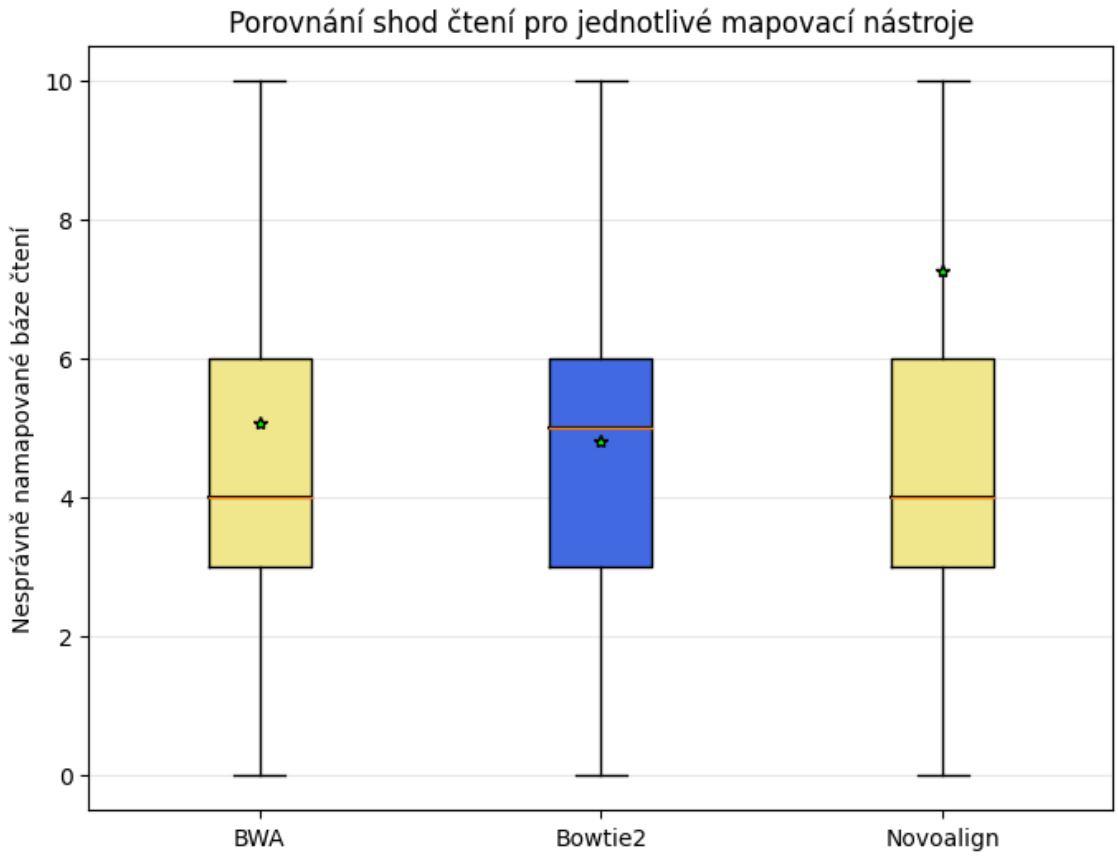
5.5 Analýza CIGAR-stringu

Ze SAM souborů jsme si vyextrahovali šestý sloupeček, ve kterém je uložena komprimovaná reprezentace zarovnání. Sledovali jsme počty shod a indelů pro každé identifikované čtení. Tento počet jsme porovnávali mezi čtením, které jsme dostali na výstupu mapperu a referenčním čtením vytvořeným v programu ART. Rozložení počtů shod ve čteních můžeme vidět v grafu Obr. 5.7 a Obr. 5.8. U většiny genomů vypadaly grafy pro Bowtie2 a BWA stejně. Bowtie2 a BWA měly konstantní výsledky prakticky pro všech 21 organismů. Tedy průměrný rozdíl shod čtení (mapper x reference) byl ve všech případech pro Bowtie2 4,8 a v případě BWA byla průměrná hodnota 5,1. U Novoalign byly výsledky poměrně nekonzistentní. V některých případech dosahovaly dobrých hodnot (například pro Enterococcus cecorum byl průměrný rozdíl shod 4,5), avšak Ve většině případů dosahoval mapper nejhorších výsledků - jak lze vidět např. na obrázku Obr. 5.8. BWA a Novoalign měly stejnou mediánovou hodnotu rozdílu shod mezi čteními, která se rovnala 4. Medián u Bowtie2 byl 5,2.

Vzhledem k tomu, že frekvence výskytu indelů je v podstatě nulová, výsledky se nevykreslovaly do žádného grafu.



Obr. 5.7: Boxplot rozdílů počtu shod čtení (mapperu) oproti referenci Aquila chrysaetos.



Obr. 5.8: Boxplot rozdílů počtu shod čtení (mapperu) oproti referenci Salmonella enterica.

6 Diskuze

V této kapitole jsou zhodnoceny dosažené výsledky a okomentována nastavení některých parametrů u všech tří mapovacích nástrojů. Výsledky jsou stručně shrnuty a případně interpretovány, dále je poukázáno na problémy, které bylo v práci nutné překonat a jsou uvedeny návrhy na rozšíření práce.

V každém případě lze konstatovat, že výsledky byly ovlivněny velikostí referenčního genomu. U genomů s malou velikostí dosahovaly Bowtie2 a BWA velice podobných výsledků, zatímco Novoalign za nimi lehce zaostával. Se zvětšující se velikostí genomu se zvětšovala přesnost i citlivost Novoalign a získali jsme výsledky srovnatelné s dalšími dvěma mappery. Obecně můžeme říci, že BWA nejlépe pracovalo s bakteriálními genomy, jako jsou např. Enterococus faecalis nebo Treponema brennaborense. Jedinou bakterií, kterou se podařilo namapovat pomocí Bowtie2 lépe byla Treponema denticola. Bakteriální genomy mají oproti živočišným genomům několikanásobně menší velikost. Právě s delšími genomy živočichů nejlépe pracoval nástroj Bowtie2, pro který všechny testované metriky dosahovaly lepších výsledků, než metriky u zbylých dvou mapovacích nástrojů v oblasti živočišných genomů. V této oblasti dosahovaly všechny mappery velmi podobných výsledků. Pokud bychom chtěli s Novoalign dosáhnout lepších výsledků i u bakteriálních genomů, nejspíše by se toho dalo dosáhnout správným nastavením parametrů. Při správném nastavení na určitý organismus by mohl Novoalign podávat stejné výsledky, jako další dva mappery.

Od každého mapovacího nástroje se očekává, že namapuje sadu čtení podle svých mapovacích kriterií. Avšak některá čtení se nemusejí namapovat (tj. falešněnegativní výsledky). Důvodem může být limitace defaultních nastavení těchto nástrojů.

V této práci jsme používali základní nastavení mapovacích nástrojů. Avšak toto nastavení není u všech mapperů stejné. Zvolení vhodného nástroje je obtížný úkol a abychom dosáhli co nejlepších výsledků, je nutné vzít v úvahu typ genomu a parametry nástrojů. Pro jeden genom je vhodnější jeden mapovací nástroj a pro druhý zase jiný. Nedá se říci, že je jeden nejlepší a vhodný na všechny typy dat. Každý z nástrojů je výborný za určitých podmínek. Mapování krátkých sekvencí je stále problém a je potřeba vyvinout nový, ideální mapovací nástroj. Prozatím se toto dá nahradit právě specifickým použitím nástrojů.

U každého z námi použitých mapovacích nástrojů lze nastavit počet povolených neshod pro identifikaci čtení jako namapovaného. Například u Novoalign je tento parametr (v základním nastavení) lehce odlišný (3), než u BWA a Bowtie2, kde je přípustný počet neshod nastaven na 0. Zvýšením tohoto parametru získáme lepší úspěšnost namapování, avšak za cenu menší relevance dosažených výsledků. Pro

Tab. 6.1: Srovnání úspěšnosti namapování při změně povolených neshod u Novoalign.

Organismus	**BWA**	**Bowtie2**	**Novoalign zákl.**	**Novoalign změn.**
Treponema succinifaciens	99,946	99,997	96,271	99,994
Salmonella enterica	99,953	99,999	98,015	99,251
Enterococcus hirae	99,948	99,998	98,411	99,421

ukázku jsme tento parametr změnili u Novoalign a to na 6 povolených neshod ve čtení. Úspěšnost namapovaní se se zvýšením tohoto parametru dle očekávání zlepšila. V případě p-distance došlo jen k nepatrně horším výsledkům v řádu setin procenta. Změnu nastavení jsme provedli u 3 zástupců bakteriálních genomů, u kterých byla nízká úspěšnost namapování při základním nastavení. Srovnání výsledků můžeme vidět v tabulce Tab. 6.1.

Je zde také možnost nastavit počet vláken (*threads*), na kterých bude software pracovat. Jejich maximální počet se odvíjí od procesoru počítače, na kterém je software spouštěn. S větším počtem vláken se výrazně zkracuje doba mapování až do určitého počtu vláken – počet je často individuální pro každý typ genomu. V našem případě jsme však nastavení tohoto parametru ponechali bez jakýchkoli změn oproti základnímu nastavení. Se správným počtem vláken při mapování by se daly časy mapování výrazně snížit a jejich rozdíly by nebyly tak znatelné. Mappery BWA a Bowtie2 mají také možnost nastavit délku podvlákna, které se má zarovnávat. Nižší hodnoty výrazně zpomalují proces mapování, avšak vedou k přesnějším výsledkům. Bowtie2 je defaultně přednastaveno na mód *sensitive* (je to jeden ze 4 módů, které jsou přednastavené se specifickými parametry – pro úsporu času nastavováním parametrů uživatelem). V tomto módu je délka podvlákna nastavena na 20. V případě BWA-MEM je délka podvlákna nastavena na 19. U každého mapovacího nástroje se dá nastavit minimální kvalita čtení, při které se báze zařadí do procesu mapování. U každého z námi využitých nástrojů je funkce tohoto parametru využita jinak a má jiné počáteční nastavení. U BWA-MEM se báze s kvalitou menší než zvolená hodnota nedostane do výstupního signálu. Tato kvalita je přednastavena na 30. Bowtie2 má pouze 2 možné nastavení minimální kvality báze, kdy se báze vůbec nebude účastnit procesu, a to 33 nebo 64. U Novoalign se však nestanovuje minimální přípustná kvalita, ale možný počet bazí s dostatečně dobrou kvalitou v jednom čtení. Proces nebude vůbec spuštěn pro čtení s méně bázemi, než je nastaveno. Tento počet bazí je v základu nastaven jako $\log_4 Ng + 5$, kde Ng je délka referenčního genomu. [46, 47, 48]

Závěr

Cílem bakalářské práce bylo zjistit vhodnost použití různých mapovacích nástrojů pro různé typy organismů. V rámci práce byla vypracována literární rešerše na téma sekvenační technologie a sestavování genomu z NGS dat. První část práce je věnována sekvenačním technologiím a sekvenačním technikám. Důkladně popisuje princip sekvenace pomocí sekvenátoru Illumina. Jsou zde také vysvětleny základní principy a mechanismy sestavování genomu se zaměřením na mapování genomu k referenci. Dále jsou uvedeny současné mapovací softwarové nástroje a jejich základní principy. Součástí práce je také popis formátu SAM, který slouží pro ukládání a manipulaci zarovnaných biologických sekvencí.

Důležitou částí práce bylo vytvoření testovacího datasetu NGS dat s přesně definovanými pozicemi čtení, na kterém následně probíhalo testování. Pro vytvoření datasetu jsem zvolil simulační program ART, simulující sekvenátor *Illumina MiSeq*. ART jsem popsal a definoval jeho nastavení. Dataset obsahuje 21 různých genomů organismů se zástupci z domény bakterií i říše živočichů. V rámci této práce jsem vytvořil jednotlivé metriky pro vyhodnocení kvality mapování, které byly základem pro komplexní analýzu kvality mapování na navrženém datasetu. Následně jsem na základě navržených metrik vytvořil programové kódy pro porovnání mapovacích nástrojů.

V druhé části práce jsem dataset namapoval pomocí několika vybraných nástrojů (konkrétně BWA, Bowtie2 a Novoalign) a výsledky vyhodnotil. Testování jsem provedl pro všechny naprogramované metriky. Veškeré výsledky testů jsou uvedeny v tabulkách v Příloze A této práce. Srovnání výsledků všech tří mapovacích nástrojů u jednotlivých sledovaných parametrů prezentuji slovně a ve formě grafického znázornění v kapitole 5 Výsledky metrik.

Z výsledků je možno určit, že pro mapování bakteriálních genomů je nejvhodnější BWA. Tento mapovací nástroj je vhodný zejména pro mapování krátkých genomů o velkém objemu, kde klademe důraz na rychlost tohoto procesu. Pokud chceme dosáhnout vysoké úspěšnosti namapování za cenu delší doby mapování, pak je vhodný nástroj Bowtie2. Z testovaných mapovacích nástrojů se Novoalign ukázal jako nástroj s nejnižší kvalitou mapování. V poslední části jsem diskutoval nastavení parametrů jednotlivých mapovacích nástrojů a zjištění jejich vlivu na výsledek. U několika vybraných organismů jsem zjistil, že změnou parametru u Novoalign lze docílit lepších výsledků, které se blíží výsledkům ostatních dvou mapovacích nástrojů. Ověřil jsem tedy skutečnost, že vhodným nastavením, byť jen jednoho parametru mapovacího nástroje, se dá docílit lepších výsledků.

Seznam obrázků

Seznam tabulek

Literatura

[1] Sekundární struktura dna, 2019. URL: http://web2.mendelu.cz/af_291_projekty2/vseo/print.php?page=1707&typ=html.

[2] J D Watson and R M Cook-Deegan. Origins of the human genome project. *The FASEB Journal*, 5(1):8–11, 1991. URL: http://www.fasebj.org/doi/10.1096/fasebj.5.1.1991595, doi:10.1096/fasebj.5.1.1991595.

[3] Historie projektu hugo, 2019. URL: http://www.hugo-international.org/history.

[4] J. Craig Venter, Mark D. Adams, and Eugene W. Myers et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001-02-16. URL: http://www.sciencemag.org/lookup/doi/10.1126/science.1058040, doi:10.1126/science.1058040.

[5] Sarah E. Walker and Jon Lorsch. Sanger dideoxy sequencing of dna. In *Laboratory Methods in Enzymology: DNA*, pages 171–184. Elsevier, 2013. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780124186873000148, doi:10.1016/B978-0-12-418687-3.00014-8.

[6] Sangerova metoda. URL: https://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2003/Obenrader/sanger_method_page.htm.

[7] Sangerova metoda sekvenování, 2009. URL: https://labguide.cz/wp-content/uploads/2015/02/Sangerova-metoda.jpg.

[8] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–18, 2008. URL: http://www.nature.com/articles/nmeth1156, doi:10.1038/nmeth1156.

[9] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010. URL: http://www.nature.com/articles/nrg2626, doi:10.1038/nrg2626.

[10] Afshin Ahmadian, Maria Ehn, and Sophia Hober. Pyrosequencing. *Clinica Chimica Acta*, 363(1-2):83–94, 2006. URL: https://linkinghub.elsevier.com/retrieve/pii/S0009898105004274, doi:10.1016/j.cccn.2005.04.038.

[11] J. F. Petrosino, S. Highlander, R. A. Luna, R. A. Gibbs, and J. Versalovic. Metagenomic pyrosequencing and microbial identification. *Clinical Chemistry*, 55(5):856–866, 2009-04-24. URL: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2008.107565, doi:10.1373/clinchem.2008.107565.

[12] Nicole Rusk and Veronique Kiermer. Primer. *Nature Methods*, 5(1):15–15, 2008. URL: http://www.nature.com/articles/nmeth1155, doi:10.1038/nmeth1155.

[13] Franziska Pfeiffer, Carsten Gröber, Michael Blank, Kristian Händler, Marc Beyer, Joachim L. Schultze, and Günter Mayer. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, 8(1), 2018. URL: http://www.nature.com/articles/s41598-018-29325-6, doi:10.1038/s41598-018-29325-6.

[14] Ali Masoudi-Nejad, Zahra Narimani, and Nazanin Hosseinkhan. De novo assembly algorithms. In *Next Generation Sequencing and Sequence Assembly*, pages 55–83. Springer New York, New York, NY, 2013. URL: http://link.springer.com/10.1007/978-1-4614-7726-6_4, doi:10.1007/978-1-4614-7726-6_4.

[15] Illumina, 2019. URL: https://www.illumina.com/science/technology/next-generation-sequencing.html.

[16] E. E. Schadt, S. Turner, and A. Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):R227–R240, 2010-10-12. URL: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddq416, doi:10.1093/hmg/ddq416.

[17] J. C. Venter. Genomics. *Science*, 280(5369):1540–1542. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.280.5369.1540, doi:10.1126/science.280.5369.1540.

[18] Shotgun sekvenace, 2014. URL: https://www.genome.gov/sites/default/files/tg/en/illustration/shotgun_sequencing.jpg.

[19] Amlikon sekvenace, 2015. URL: https://www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/amplicon-sequencing.html.

[20] M.Steven Oberste, William A Nix, Kaija Maher, and Mark A Pallansch. Improved molecular identification of enteroviruses by rt-pcr and amplicon sequencing. *Journal of Clinical Virology*, 26(3):375–377, 2003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1386653203000040, doi:10.1016/S1386-6532(03)00004-0.

[21] Amplicon sekvenace. URL: https://www.coderegenesis.com/wp-content/uploads/2019/03/Amplicon-Seq-v4.png.

[22] Jason R. Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0888754310000492, doi:10.1016/j.ygeno.2010.03.001.

[23] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Spades. *Journal of Computational Biology*, 19(5):455–477, 2012. URL: http://www.liebertpub.com/doi/10.1089/cmb.2012.0021, doi:10.1089/cmb.2012.0021.

[24] D. R. Zerbino and E. Birney. Velvet. *Genome Research*, 18(5):821–829, 2008-02-21. URL: http://www.genome.org/cgi/doi/10.1101/gr.074492.107, doi:10.1101/gr.074492.107.

[25] Abdul Rafay Khan, Muhammad Tariq Pervez, Masroor Ellahi Babar, Nasir Naveed, and Muhammad Shoaib. A comprehensive study of de novo genome assemblers. *Evolutionary Bioinformatics*, 14, 2018-02-20. URL: http://journals.sagepub.com/doi/10.1177/1176934318758650, doi:10.1177/1176934318758650.

[26] R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt. Assembling millions of short dna sequences using ssake. *Bioinformatics*, 23(4):500–501, 2007-02-14. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl629, doi:10.1093/bioinformatics/btl629.

[27] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J.M. Jones, and I. Birol. Abyss. *Genome Research*, 19(6):1117–1123, 2009-06-01. URL: http://genome.cshlp.org/cgi/doi/10.1101/gr.089532.108, doi:10.1101/gr.089532.108.

[28] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009-07-02. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324, doi:10.1093/bioinformatics/btp324.

[29] Ayat Hatem, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1), 2013. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-184, doi:10.1186/1471-2105-14-184.

[30] Jing Shang, Fei Zhu, Wanwipa Vongsangnak, Yifei Tang, Wenyu Zhang, and Bairong Shen. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *BioMed Research International*, 2014:1–16, 2014. URL: http://www.hindawi.com/journals/bmri/2014/309650/, doi:10.1155/2014/309650.

[31] P. Ferragina and G. Manzini. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398. IEEE Comput. Soc, 2000. URL: http://ieeexplore.ieee.org/document/892127/, doi:10.1109/SFCS.2000.892127.

[32] Thomas D. Wu, Jens Reeder, Michael Lawrence, Gabe Becker, and Matthew J. Brauer. Gmap and gsnap for genomic sequence alignment. In *Statistical Genomics*, pages 283–334. Springer New York, New York, NY, 2016. URL: http://link.springer.com/10.1007/978-1-4939-3578-9_15, doi:10.1007/978-1-4939-3578-9_15.

[33] Novoalign, 2014. URL: http://www.novocraft.com/products/novoalign/.

[34] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3), 2009. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25, doi:10.1186/gb-2009-10-3-r25.

[35] Open source, 2019. URL: https://opensource.com/resources/what-open-source?fbclid=IwAR1kE9Td2-YvikNYUiYKgw5w7OnyGFP2ApNhb0JC-c3MMsxyFsVldkI

[36] Stephen M. Rumble, Phil Lacroute, Adrian V. Dalca, Marc Fiume, Arend Sidow, Michael Brudno, and Wyeth W. Wasserman. Shrimp. *PLoS Computational Biology*, 5(5), 2009-5-22. URL: https://dx.plos.org/10.1371/journal.pcbi.1000386, doi:10.1371/journal.pcbi.1000386.

[37] Matei David, Misko Dzamba, Dan Lister, Lucian Ilie, and Michael Brudno. Shrimp2. *Bioinformatics*, 27(7):1011–1012, 2011-4-1. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr046, doi:10.1093/bioinformatics/btr046.

[38] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang. Soap2. *Bioinformatics*, 25(15):1966–1967, 2009-07-17. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp336, doi:10.1093/bioinformatics/btp336.

[39] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/-map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009-08-07. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352, doi:10.1093/bioinformatics/btp352.

[40] Formát sam/bam, 2019. URL: http://samtools.github.io/hts-specs/SAMv1.pdf.

[41] Sam formát. URL: https://www.samformat.info/images/sam_format_annotated_example.5108a0cd.jpg.

[42] Phred skóre, 2011. URL: https://www.illumina.com/Documents/products/technotes/technote_Q-Scores.pdf.

[43] Weichun Huang, Leping Li, Jason R. Myers, and Gabor T. Marth. Art. *Bioinformatics*, 28(4):593–594, 2012-2-15. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr708, doi:10.1093/bioinformatics/btr708.

[44] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1061–1069. Curran Associates, Inc., 2012. URL: http://papers.nips.cc/paper/4808-hamming-distance-metric-learning.pdf.

[45] Broadinstitute.github.io. URL: https://broadinstitute.github.io/picard/explain-flags.html.

[46] Manual reference pages - bwa (1), 2013. URL: http://bio-bwa.sourceforge.net/bwa.shtml.

[47] Bowtie-bio.sourceforge.net/bowtie2/manual.shtml#options, 2020. URL: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#options.

[48] Www.novocraft.com, 2016. URL: http://www.novocraft.com/documentation/novoalign-2/novoalign-user-guide/novoalign-command-options/.

A Výsledky metrik

Tab. A.1: Výsledky t-testu.

Organismus	**BWA**	**Bowtie2**	**Novoalign**
Enterococcus casseliflavus	0,972	0,338	4,862E-11
Enterococcus hirae	0,999	0,990	0,999
Enterococus faecalis	0,792	0,632	9,33E-16
Enterococcus cecorum	0,845	0,323	9,65E-14
Enterococcus aquimarinus	0,999	0,999	0,999
Staphylococcus aureus	0,927	0,123	2,33E-08
Streptococcus pneumoniae	0,995	0,071	5,81E-18
Treponema brennaborense	0,893	0,652	4,74E-11
Treponema denticola	0,998	0,972	2,06E-09
Treponema pallidum	0,976	0,963	1,20E-03
Treponema succinifaciens	0,907	0,255	3,23E-56
Treponema pedis	0,997	0,962	2,00E-04
Mycobacterium tuberculosis	0,983	0,931	7,80E-24
Salmonella enterica	0,967	0,167	3,21E-54
Aquila chrysaetos	0,953	0,961	7,20E-26
Limosala lapponica baueri	0,999	0,999	9,16E-01
Olobus angolensis	0,999	0,961	1,18E-07
Hirundo rustica	0,972	0,856	6,82E-11
Camarhynchus parvulus	0,967	0,953	0,067
Mus musculus	0,347	0,771	0,00E+00
Macaca mulatta	0,397	0,963	0,00E+00

Tab. A.2: Výsledky Kolgomorov-Smirnov testu.

Organismus	**BWA**	**Bowtie2**	**Novoalign**
Enterococcus casseliflavus	1,00	0,98	1,66E-07
Enterococcus hirae	1,00	1,00	1,00
Enterococus faecalis	1,00	1,00	1,87E-16
Enterococcus cecorum	1,00	0,35	4,00E-21
Enterococcus aquimarinus	1,00	1,00	1,00E+00
Staphylococcus aureus	1,00	0,53	1,70E-08
Streptococcus pneumoniae	1,00	0,82	2,62E-15
Treponema brennaborense	1,00	0,99	2,88E-14
Treponema denticola	1,00	1,00	1,26E-10
Treponema pallidum	1,00	1,00	1,67E-09
Treponema succinifaciens	1,00	0,06	2,28E-101
Treponema pedis	1,00	1,00	2,00E-03
Mycobacterium tuberculosis	1,00	1,00	1,40E-22
Salmonella enterica	1,00	0,31	3,21E-55
Aquila chrysaetos	1,00	1,00	2,50E-38
Limosala lapponica baueri	1,00	1,00	1,00E+00
Olobus angolensis	1,00	1,00	3,79E-11
Hirundo rustica	1,00	1,00	3,26E-111
Camarhynchus parvulus	1,00	1,00	0,29
Mus musculus	0,97	1,00	0,00E+00
Macaca mulatta	1,00	1,00	0,00E+00

Tab. A.3: Výsledky hammingovi vzdálenosti.

Organismus	**BWA**	**Bowtie2**	**Novoalign**
Enterococcus casseliflavus	99,816	99,774	99,683
Enterococcus hirae	48,028	47,963	47,955
Enterococus faecalis	99,627	99,573	99,509
Enterococcus cecorum	100,00	99,999	100,00
Enterococcus aquimarinus	100,00	100,00	100,00
Staphylococcus aureus	99,733	99,698	99,629
Streptococcus pneumoniae	99,593	99,504	99,401
Treponema brennaborense	99,607	99,565	99,526
Treponema denticola	99,921	99,923	99,594
Treponema pallidum	99,922	99,923	99,374
Treponema succinifaciens	99,319	98,947	98,614
Treponema pedis	65,029	98,937	98,823
Mycobacterium tuberculosis	99,722	99,683	99,521
Salmonella enterica	99,405	99,236	99,257
Aquila chrysaetos	99,936	99,917	99,853
Limosala lapponica baueri	99,921	99,997	99,997
Olobus angolensis	99,942	99,963	99,934
Hirundo rustica	99,935	99,955	99,921
Camarhynchus parvulus	83,121	99,969	99,959
Mus musculus	99,948	99,968	99,952
Macaca mulatta	99,932	99,945	99,933

Tab. A.4: Výsledky procentuální úspěšnosti namapování.

Organismus	**BWA**	**Bowtie2**	**Novoalign**	ART
Enterococcus casseliflavus	99,999	99,933	99,155	100
Enterococcus hirae	99,948	99,998	98,411	100
Enterococus faecalis	99,951	99,999	98,688	100
Enterococcus cecorum	99,053	99,998	99,999	100
Enterococcus aquimarinus	99,969	99,996	99,996	100
Staphylococcus aureus	99,941	99,997	99,006	100
Streptococcus pneumoniae	99,898	99,999	99,594	100
Treponema brennaborense	99,964	99,999	98,749	100
Treponema denticola	99,937	99,999	98,886	100
Treponema pallidum	99,893	99,999	98,343	100
Treponema succinifaciens	99,946	99,997	96,271	100
Treponema pedis	99,943	99,999	99,389	100
Mycobacterium tuberculosis	99,949	99,999	98,681	100
Salmonella enterica	99,953	99,999	98,015	100
Aquila chrysaetos	99,942	99,998	99,908	100
Limosala lapponica baueri	99,933	99,999	99,991	100
Olobus angolensis	99,943	99,991	99,665	100
Hirundo rustica	99,423	99,996	99,373	100
Camarhynchus parvulus	99,401	99,998	99,896	100
Mus musculus	99,951	99,963	99,942	100
Macaca mulatta	99,943	99,953	99,946	100

Tab. A.5: Výsledky pro průměrný rozdíl počtu indelů mezi vygenerovaným souborem a referencí.

Organismus	**BWA**	**Bowtie2**	**Novoalign**
Enterococcus casseliflavus	0	0	0
Enterococcus hirae	0	0	0
Enterococus faecalis	0	0	0
Enterococcus cecorum	0	0	0
Enterococcus aquimarinus	0	0	0
Staphylococcus aureus	0	0	0
Streptococcus pneumoniae	0	0	0
Treponema brennaborense	0	0	0
Treponema denticola	0	0	0
Treponema pallidum	0	0	0
Treponema succinifaciens	0	0	0
Treponema pedis	0	0	0
Mycobacterium tuberculosis	0	0	0
Salmonella enterica	0	0	0
Aquila chrysaetos	0	0	0
Limosala lapponica baueri	0	0	0
Olobus angolensis	0	0	0
Hirundo rustica	0	0	0
Camarhynchus parvulus	0	0	0
Mus musculus	0	0	0
Macaca mulatta	0	0	0

Tab. A.6: Výsledky pro průměrný rozdíl počtu shod mezi vygenerovaným souborem a referencí.

Organismus	**BWA**	**Bowtie2**	**Novoalign**
Enterococcus casseliflavus	5,1	4,8	5,7
Enterococcus hirae	5,7	4,8	6,8
Enterococus faecalis	5,1	4,8	6,3
Enterococcus cecorum	5,1	4,8	4,5
Enterococcus aquimarinus	5,1	4,8	4,5
Staphylococcus aureus	5,1	4,8	5,9
Streptococcus pneumoniae	5,1	4,8	6,7
Treponema brennaborense	5,1	4,8	6,2
Treponema denticola	5,1	4,8	6,0
Treponema pallidum	5,1	4,8	6,8
Treponema succinifaciens	5,1	4,8	9,7
Treponema pedis	5,5	4,8	5,3
Mycobacterium tuberculosis	5,1	4,8	6,3
Salmonella enterica	5,1	4,8	7,3
Aquila chrysaetos	5,1	4,8	4,5
Limosala lapponica baueri	5,1	4,8	4,7
Olobus angolensis	5,1	4,8	4,5
Hirundo rustica	5,1	4,8	5,9
Camarhynchus parvulus	5,1	4,8	6,2
Mus musculus	5,1	4,8	5,7
Macaca mulatta	5,1	4,8	5,6

Tab. A.7: Časy mapování všech organismů

Organismus	**BWA** [s]	**Bowtie2** [s]	**Novoalign** [s]
Enterococcus casseliflavus	24	52	66
Enterococcus hirae	24	49	63
Enterococus faecalis	24	48	83
Enterococcus cecorum	17	31	51
Enterococcus aquimarinus	2	6	6
Staphylococcus aureus	23	42	80
Streptococcus pneumoniae	15	27	46
Treponema brennaborense	22	45	76
Treponema denticola	21	44	66
Treponema pallidum	8	16	22
Treponema succinifaciens	23	50	77
Treponema pedis	23	45	67
Mycobacterium tuberculosis	35	67	120
Salmonella enterica	38	73	113
Aquila chrysaetos	954	1431	1808
Limosala pponica baueri	41	77	114
olobus angolensis	371	602	740
Hirundo rustica	1051	1760	2367
Camarhynchus parvulus	288	268	631
Mus musculus	5527	5183	7348
Macaca mulatta	4515	5604	7613