VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIPLOMOVÁ PRÁCE

Brno, 2018 Bc. Vojtěch Bartoň

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

ANALÝZA GENETICKÉ VARIABILITY V SEKVENAČNÍCH DATECH TREPONEMÁLNÍCH KMENŮ
ANALYSIS OF GENETIC VARIABILITY IN SEQUENCING DATA OF TREPONEMA STRAINS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Vojtěch Bartoň**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Denisa Maděránková, Ph.D.**
SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské inženýrství a bioinformatika**
Ústav biomedicínského inženýrství

Student: Bc. Vojtěch Bartoň ***ID:*** 164962
Ročník: 2 ***Akademický rok:*** 2017/18

NÁZEV TÉMATU:

Analýza genetické variability v sekvenačních datech treponemálních kmenů

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte rešerši sekvenačních přístupů/technik. 2) Navrhněte postup identifikace heterogenních míst v resekvenovaných genomech. 3) Navržený postup implementujte ve vhodném programovém prostředí (R, Python). 4) Identifikujte heterogenní místa v souboru resekvenovaných treponemálních kmenů. 5) Vyhodnoťte heterogenní oblasti v jednotlivých genomech a proveďte vzájemné srovnání v rámci celého souboru.

DOPORUČENÁ LITERATURA:

[1] ČEJKOVÁ, D.; STROUHAL, M.; NORRIS, S.J.; WEINSTOCK, G.M.; ŠMAJS, D.; PICARDEAU, M. A Retrospective Study on Genetic Heterogeneity within Treponema Strains: Subpopulations Are Genetically Distinct in a Limited Number of Positions. PLOS Neglected Tropical Diseases. 2015, 9(10), e0004110.

[2] PINTO, M.; BORGES, V.; ANTELO, M.; et al. Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation. Nature Microbiology. 2016, 2, 16190.

Termín zadání: 5.2.2018 ***Termín odevzdání:*** 18.5.2018

Vedoucí práce: Ing. Denisa Maděránková, Ph.D.
Konzultant: prof. MUDr. David Šmajs, Ph.D.

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

ABSTRAKT

Tato diplomová práce se zabývá metodami určení genetické variability v sekvenačních datech. Sekvenovaným organismem je několik kmenů bakterie *Treponema pallidum*. Bakterie byly osekvenovány na platformě Illumina. Navrhujeme postup určení variabilních míst v resekvenovaných genomech a jejich následné zkoumání v rámci jednoho genomu, tak i porovnání napříč všemi zpracovávanými genomy.

KLÍČOVÁ SLOVA

heterogenita, bioinformatika, *Treponema pallidum*, genetická variabilita, sekvenace

ABSTRACT

This diploma thesis is dealing with methods of identification genetic variability in sequencing data. The resarch is targeted to bacterial strains of *Treponema pallidum*. The sequencing was performed by Illumina platform. There is a proposition of method to identificate variable spots in resequenced genomes and their analysis and comparation across all processed genomes.

KEYWORDS

heterogeneity, bioinformatics, *Treponema pallidum*, genetic variability, sequencing

BARTOŇ, Vojtěch. *Analýza genetické variability v sekvenačních datech treponemálních kmenů*. Brno, 2018, 61 s. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Denisa Maděránková, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Analýza genetické variability v sekvenačních datech treponemálních kmenů“ jsem vypracoval samostatně pod vedením vedoucí diplomové 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 diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové 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 vedoucí diplomové práce Ing. Denise Maděránkové, Ph.D., za odborné vedení, trpělivost a podnětné návrhy k práci.
Dále bych rád poděkoval prof. MUDr. Davidu Šmajsovi, Ph.D., a jeho výzkumnému týmu z Biologického ústavu Lékařské fakulty Masarykovy univerzity za poskytnutí sekvenačních dat a četné konzultace, připomínky a rady během tvorby diplomové práce.
V neposlední řadě patří poděkování také rodině, která je mi oporou po celou dobu mého studia.

Tato práce vznikla za podpory projektů CERIT Scientific Cloud (LM2015085) a CESNET (LM2015042) financovaných z programu MŠMT Projekty velkých infrastruktur pro VaVaI.

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

S nástupem masivního rozšíření Next-Generation sekvenačních technologií jsme schopni produkovat obrovské množství dat, které nám umožňují mnohem lépe popsat danou sekvenovanou molekulu. Jednou z informací, která je pomocí sekvenování odhalitelná, je informace o vnitrogenomové variabilitě.

Analýza variability přináší nové informace o chování daného organismu, jeho evoluci nebo o jeho adaptačních schopnostech. V datech jsme schopni identifikovat místa, jež se oproti majoritní populaci liší a tato změna není způsobena náhodnou mutací, nýbrž jistým adaptačním tlakem daného organismu. Pro pochopení fungování takových organismů je analýza variability v sekvenačních datech nezbytná.

Cílem teoretické části této práce je podat přehled o používaných sekvenačních přístupech a metodách. Dále pak čtenáře stručně seznámit s kmenem bakterie *Treponema pallidum*, pro který máme k dispozici několik souborů osekvenovaných poddruhů. Tato bakterie a její poddruhy jsou známé především jako původce nomocí syfylis, endemická syfilis, či frambézie. Způsobená choroba se liší v závislosti na infekčním poddruhu. V práci se zabýváme poddruhy *T. pallidum*, *T. endemicum* a *T. pertenue*. U těchto chorob je léčba ztížená faktem, že bakterie neobsahuje příliš mnoho membránových proteinů, tedy farmaka není na co účinně vázat. Pochopení mechanismů fungování takovýchto bakterií jako celku, může přispět k efektivnější léčbě vyvolaných onemocnění.

V praktické části se zaměřuji na návrh postupu pro určení variabilních míst v osekvenovaném genomu. Budu diskutovat možnosti filtrace sekvenačních dat za účelem zlepšení jejich kvality a zvýšení informační výtěžnosti testovaného souboru pro analýzu variability. Zabývám se také způsobem identifikace variabilních míst v genomech a jejich srovnáním mezi jednotlivými testovanými bakteriemi.

V práci využíváme data poskytnutá Biologickám ústavem Lékařské fakulty Masarykovy univerzity. Při akvizici těchto sekvenačních dat, bylo využito specifického unikátního rozdělení každého sekvenovaného genomu do několika částí, čímž se zamezilo následnému falešnému mapování čteních pocházejících z sekvenčně si podobných oblastí. Tento způsob zpracování umožnil akvizici unikátních sekvenačních dat, umožňující hloubkovou analýzu variabilních pozic ve zkoumaných genomech.

Při hodnocení variabilit využívám také celogenomého zarovnaní celého zpracovávaného souboru. Díky tomu mohu hodnotit vzájemnou koincidenci výskytu variabilit na stejných pozicích v různých genomech, tak jako výskyt varibilit v rámci celých genů, a to i u ještě neanotovaných genomů.

1 SEKVENAČNÍ TECHNOLOGIE

Sekvenační technologie se zabývají metodami určení primární struktury DNA řetězce, tzn. pořadí nukleotidů v sekvenci. Ze znalosti primární struktury můžeme vycházet při určování kódujících částí, odhalování mutací a odvozování informací o tvorbě proteinů. V následující kapitole přináším přehled používaných sekvenačních metod rozčleněný podle generací.

1.1 První generace

Za objevem 3D struktury DNA šroubovice v roce 1953 stojí pánové Watson a Crick, kteří pracovali s krystalografickými snímky poskytnutými Rosalindou Franklin a Maricem Wilkinsem. Použitá metoda ale neumožňovala odhalení primární struktury DNA. To se povedlo až v druhé polovině sedmdesátých let dvacátého století. Nezávisle na sobě vznikají dvě první sekvenační metody pojmenované podle svých tvůrců - Maxam-Gilbertova a Sangerova metoda. [1]

1.1.1 Maxam-Gilbertovo sekvenování

Tato metoda vznikla v roce 1977. Za objevem stojí americký profesor Walter Gilbert, působící na Harvardské univerzitě, a jeho doktorand Allan Maxam. Spolu s Paulem Bergem a Frederickem Sangerem se v roce 1980 stali nositeli Nobelovy ceny za chemii za přispění k určení sekvence nukleotidů. [2]

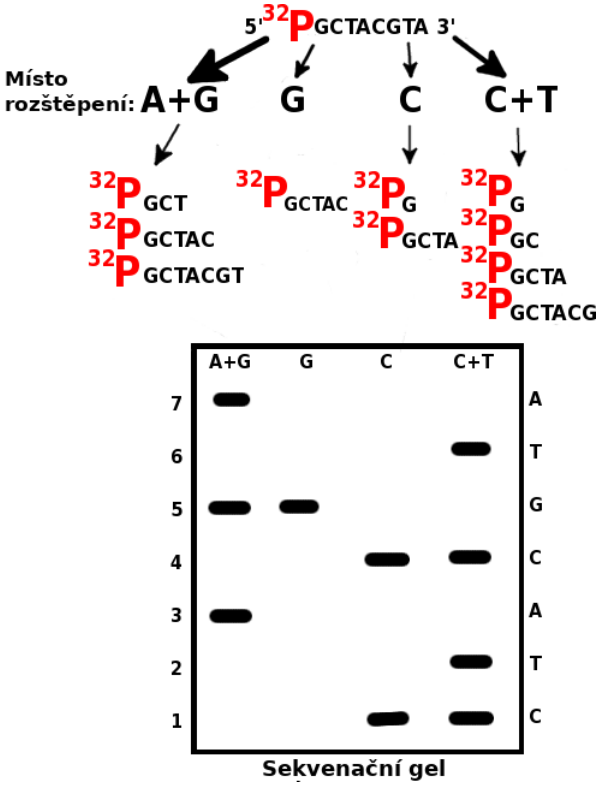
Metoda je založena na štěpení molekuly DNA v místech výskytu určitého nukleotidu a detekci radioaktivního záření. Vše začíná označením DNA fragmentu na 5' konci sekvenovaného řetězce nukleotidů radioaktivní značkou a purifikací molekuly. Takto připravený vzorek rozdělíme do čtyř oddělených systémů. Na každém vzorku bude probíhat jiná chemická reakce, která zaručí rozštěpení molekuly na místě výskytu daného nukleotidu. Do každého ze čtyř systémů je přidán jiný reagent, který zaručí odštěpení na dané bázi od deoxyribózy. Jednotlivé reagenty ukazuje tabulka 1.1.

Celý fragment je pak na dané modifikované pozici rozštěpen za použití horkého piperidinu. Celková koncentrace všech chemikálií je vypočtena tak, aby došlo jen k jedné modifikaci na molekulu DNA. Těmito reakcemi nám vznikají fragmenty DNA různé délky, které jsou na 5' konci radioaktivně označeny a na 3' konci končí nukleotidem u kterého proběhlo štěpení. Vyhodnocení sekvenace a určení posloupnosti nukleotidů probíhá pomocí elektroforézy. Produkty ze všech čtyř systémů, kde proběhlo štěpení, jsou naneseny na polyakrylamidový gel a odděleny elektroforézou na

Tabulka 1.1: Činidla pro štěpení u Maxam-Gilbertovy metody.

Nukleotid	Reagent
G	dimethylsulfát
A+G	kyselina mravenčí
C	hydrazin + NaCl
C+T	hydrazin

základě jejich velikosti. Následně celý proces vizualizujeme pomocí autoradiografie. Jednotlivé nukleotidy následně odečítáme z polohy proužků na gelu a postupně získáváme kompletní sekvenci DNA. Postup odečítání vidíme na obrázku 1.1 . [3]



Obrázek 1.1: Analýza elektroforézou u Maxam-Gilbertovy metody. Převzato z [4].

1.1.2 Sangerovo sekvenování

Metoda je nazvána po svém objeviteli Fredericku Sangerovi, který ji prvně publikoval v roce 1977. Patří do kategorie sekvenačních metod pomocí syntézy řetězce. Svému objeviteli přinesla Nobelovu cenu za chemii v roce 1980. [2]

Metoda je založena na terminaci narůstajícího řetězce. K jejímu provedení je třeba jednovláknová DNA, DNA primer, který označuje začátek sekvenovaného úseku, DNA polymerázu, normální deoxynukleosidtrifosfáty (dNTP) a jejich modifikaci v podobě dideoxynukleosidtrifosfátu (ddNTP). Modifikované ddNTP se od dNTP liší nepřítomností volné 3' -OH skupiny, která je nezbytná pro vznik fosfodiesterové vazby s následujícím nukleotidem. Začleněním ddNTP do narůstajícího řetězce DNA se ukončí jeho elongace. Zároveň je každý ddNTP radioaktivně označen.

Vzorek DNA je rozdělen do čtyř oddělených systémů. Každý systém obsahuje vzorek DNA, navržené primery, DNA polymerázu, normální dNTP všech čtyř druhů a pouze jeden ze čtyř ddNTP. Koncentrace ddNTP je asi 100x nižší než odpovídající dNTP, aby nedocházelo k předčasné terminaci všech jednovláknových úseků. Využívá se PCR (polymerázová řetězová reakce) v každém ze čtyř systémů. Nasednutím primeru na templátovou DNA a následná syntéza řetězce ve směru 5' -> 3'. Řetězec DNA postupně narůstá, než dojde k navázaní ddNTP, čímž je proces elongace ukončen. Takto ukončený úsek je zároveň radioaktivně označen. V každém systému obdržíme několik různě dlouhých a označených úseků DNA. Následně vzorky tepelně denaturujeme, aby došlo k oddělení templátových a syntetizovaných vláken.

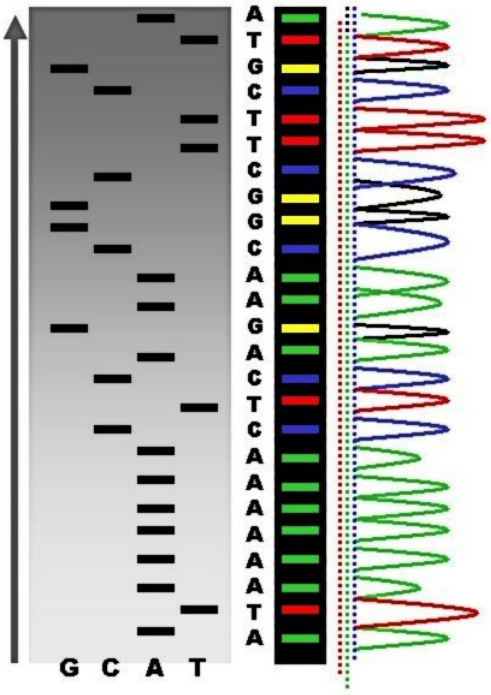
Vzorky následně hodnotíme podobně jako u Maxam-Gilbertovy metody za pomocí gelové elektroforézy. Pomocí ní oddělíme jednotlivá syntetizovaná vlákna na základě velikosti. Gel následně autoradiografií vizualizujeme a jeho čtením sestavujeme sekvenci templátové DNA. [5]

Tato metoda byla později modifikována za použití fluorescenčních barviv, místo radiačního značení, které umožnilo sloučení reakcí do jednoho systému. Proces byl dále upraven pro použití v automatických přístrojích pomocí kapilární elektroforézy. V těchto přístrojích se výsledná sekvence určuje na základě chromatogramu. Příklad můžeme vidět na obrázku 1.2 vpravo, zatímco vlevo vidíme analýzu pomoci gelové elektroforézy.

Sangerova metoda se stala masivně rozšířenou zejména díky snadné automatizaci a nepoužívání toxických reagencií. Výhodou je rovněž to, že poskytuje dlouhé sekvence, které obsahují velmi málo chyb. Dodnes se používá při sekvenování menších úseků DNA, či při validaci výsledků z modernějších sekvenátorů. [1]

1.2 Druhá generace - NGS

Druhá generace metod sekvenování bývá označována jako Next Generation Sequencing (NGS). Vývoj se přesunul z univerzitních pracoven do komerčních subjektů. Vyznačuje se zejména masivní paralelizací celého procesu a tím i k produkci mnohem většího počtu sekvencí za zlomek času. Rovněž se zde uplatňuje metoda *wash & scan*,



Obrázek 1.2: Analýza u Sangerovy metody. [7]

která spočívá v cyklickém přidávání reagentů, začlenění báze do řetězce, zastavení reakce, vymytí přebytku reagentů a detekce inkorporované báze. [6]

1.2.1 Pyrosekvenace

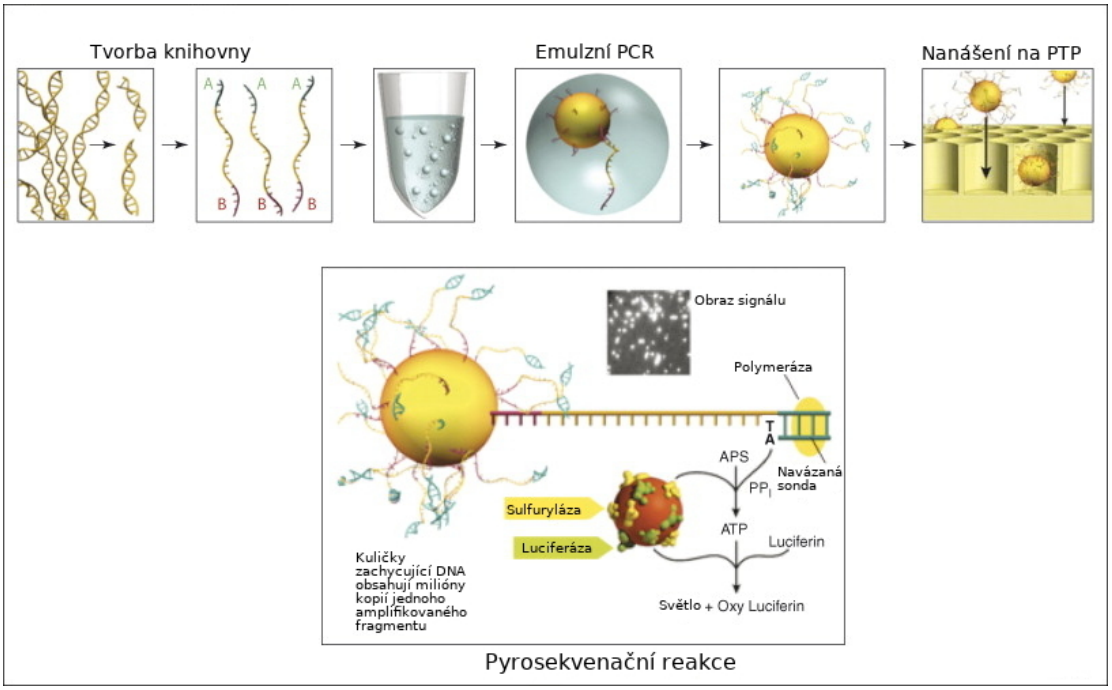
Jednou z prvních technologií druhé generace je využití pyrosekvenování vyvinuté v roce 2005 firmou 454 Life Sciences, která byla později koupena firmou Roche. Sekvenátory jsou známy pod jménem Roche 454. Metoda sekvenování využívá proces syntézy pyrosekvenací. Hlavním bodem procesu je uvolnění pyrofosfátu při začlenění nukleotidu polymerázou do narůstajícího vlákna DNA. [8]

Sekvenovaný úsek DNA je namnožen a fragmentován na kratší úseky o délce několika set párů bazí. Ke každému fragmentu jsou poté ligovány krátké adaptory na oba konce. DNA je poté denaturována na jednovláknovou. Adaptor na jednom konci je určen k navázaní na párující adaptor umístěný na povrchu malé streptavidinové kuličky. Na kuličkách je poté provedena emulzní PCR, která způsobí namnožení daného immobilizovaného fragmentu jednovláknové DNA. Po skončení emulzní PCR jsou na jedné kuličce navázány miliony stejných fragmentů jednovláknové DNA. Nenavázané produkty PCR jsou odmyty pryč.

Kuličky jsou poté přeneseny na pikotitrační destičku tak, aby do každé jamky zapadla právě jedna kulička. Ke každé jamce jsou přidány enzymy pro pyrosekve-

nování, ATP sulfuryláza a luciferáza a DNA polymeráza. Ke kuličkám jsou přidány primery, které nasednou na volný adaptor a umožní začít syntézu DNA. Následně je celá destička cyklicky promývána roztokem obsahujícím vždy jen jeden druh dNTP. V případě, že je daný dNTP inkorporován do narůstajícího řetězce, dojde vlivem vznikající fosfodiesterové vazby k uvolnění pyrofosfátu, jahož další přeměna na ATP je katalyzována sulfurylázou. Molekula ATP slouží k přeměně luciferinu na oxyluciferin za přítomnosti luciferázy a emise viditelného záření. [9]

Emitované světlo je snímáno velmi citlivým CCD čipem. Následně je roztok aktuálního dNTP odstraněn pomocí reakce s apyrázou a systém je promýván jiným dNTP. Analýzou flowgramů z CCD čipu poté sestavujeme požadovanou sekvenci nukleotidů v jednotlivých jamkách. Pokud je do řetězce začleněno v jednom cyklu více dNTP dojde k vyšší emisi světla. Tento jev ale není lineární a způsobuje chybovost metody při sekvenaci homopolymerních úseků. Celý proces je znázorněn na obrázku 1.3.



Obrázek 1.3: Postup pyrosekvenace. Převzato z [10].

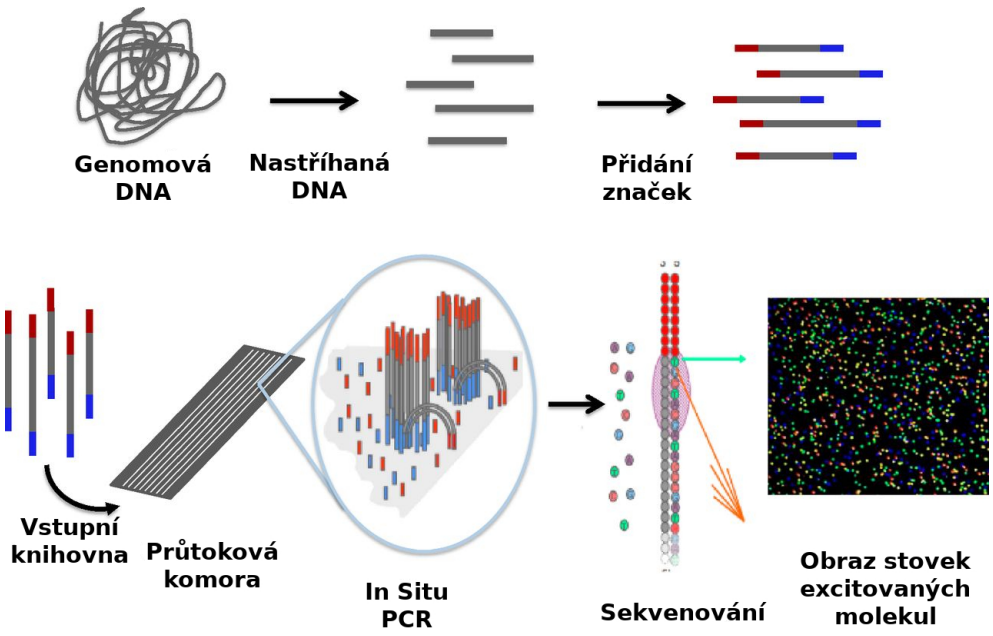
1.2.2 Illumina

Sekvenátor byl vyvinut v roce 2006 firmou Solexa, která jej ale prodala společnosti Illumina v roce 2007. Patří do kategorie sekvenace syntézou řetězce DNA. V současnosti jde asi o nejrozšířenější metodu sekvenování.

Příprava vzorku probíhá navázáním specifických adaptorů na oba konce fragmentovaných úseků DNA. Fragmenty jsou následně denaturovány a vzniknuvší jednovláknové úseky DNA jsou umístěny na sekvenační destičku (flow cell). Na destičce dojde k navázání fragmentů k destičce pomocí komplementarních adaptorů. K navázání zde dojde oběma konci fragmentu, vzniká tak jakýsi most, který dal této proceduře název "můstková PCR". K navázaným "mostům"je syntetizováno komplementární vlákno, následně je rozděleno a postup se opakuje. Na destičce tak vznikají shluky stejných fragmentů DNA. [8]

Reverzní vlákna jsou následně ze systému vymyta a přidávají se primery komplementární k volným adaptorům. Do systému se přidává DNA polymeráza a směs všech čtyř dNTP s barevnou sondou a navázaným reverzibilním terminátorem. Díky terminátoru dojde k navázání pouze jednoho nukleotidu k prolongovanému řetězci. Celý systém je excitován laserem a pomocí CCD čipu snímán barevný obraz. Před dalším cyklem dojde k odstranění terminátorů a opakujeme snímání obrazu.

Pomocí čtení barevných bodů v jednotlivých obrazech určujeme posloupnost nukleotidů vznikajícího vlákna v daném místě destičky. Celý postup je ilustrován na obrázku 1.4. [11]



Obrázek 1.4: Sekvenace pomocí můstkové PCR. Převzato z [12].

1.2.3 SOLiD

Technologie SOLiD (sequencing by oligonucleotide ligation and detecion) byla vyvinuta firmou Applied Biosystems v roce 2007. Jde o metodu, která již není založena na syntéze řetězce, ale na ligaci nukleotidových úseků. [13]

Příprava vzorku je obdobná jako u technologie Roche 454. Dochází k ligaci adaptorů na oba konce fragmentované DNA. Jednotlivé fragmenty jsou poté uchyceny díky kompatibilitě adaptorů na malé kuličky. Na každé kuličce se nachází pouze jeden fragment DNA. Nyní je provedena emulzní PCR na každé kuličce. Kuličky s amplifikovanou DNA jsou umístěny na destičku, kde jsou uchyceny ve svých pozicích pomocí kovalentní vazby k povrchu destičky.

Do systému je přidána DNA ligáza a speciální DNA sondy. Každá sonda se skládá ze dvou specifických nukleotidů následovaných třemi libovolnými nukleotidy a třemi terminačními s navázanou fluorescenční značkou. Fluorescenční značka je specifická podle prvních dvou nukleotidů sondy. Celkem jsou k dispozici čtyři různé značky. Sondy nasedají za primer a jsou na základě komplementarity bazí ligovány k takto se prodlužujícímu řetězci. Nenavázané sondy jsou vymyty a systém je excitován pro změření fluorescence. Poté dojde k odstřižení posledních tří terminačních nukleotidů s navázaným fluorescentem. Tím je umožněna ligace další sondy. Po elongaci celého úseku na požadovanou délku je nově vytvořený řetězec odstraněn a na původní templátové vlákno je navázen primer s délkou o jeden nukleotid kratší. [14]

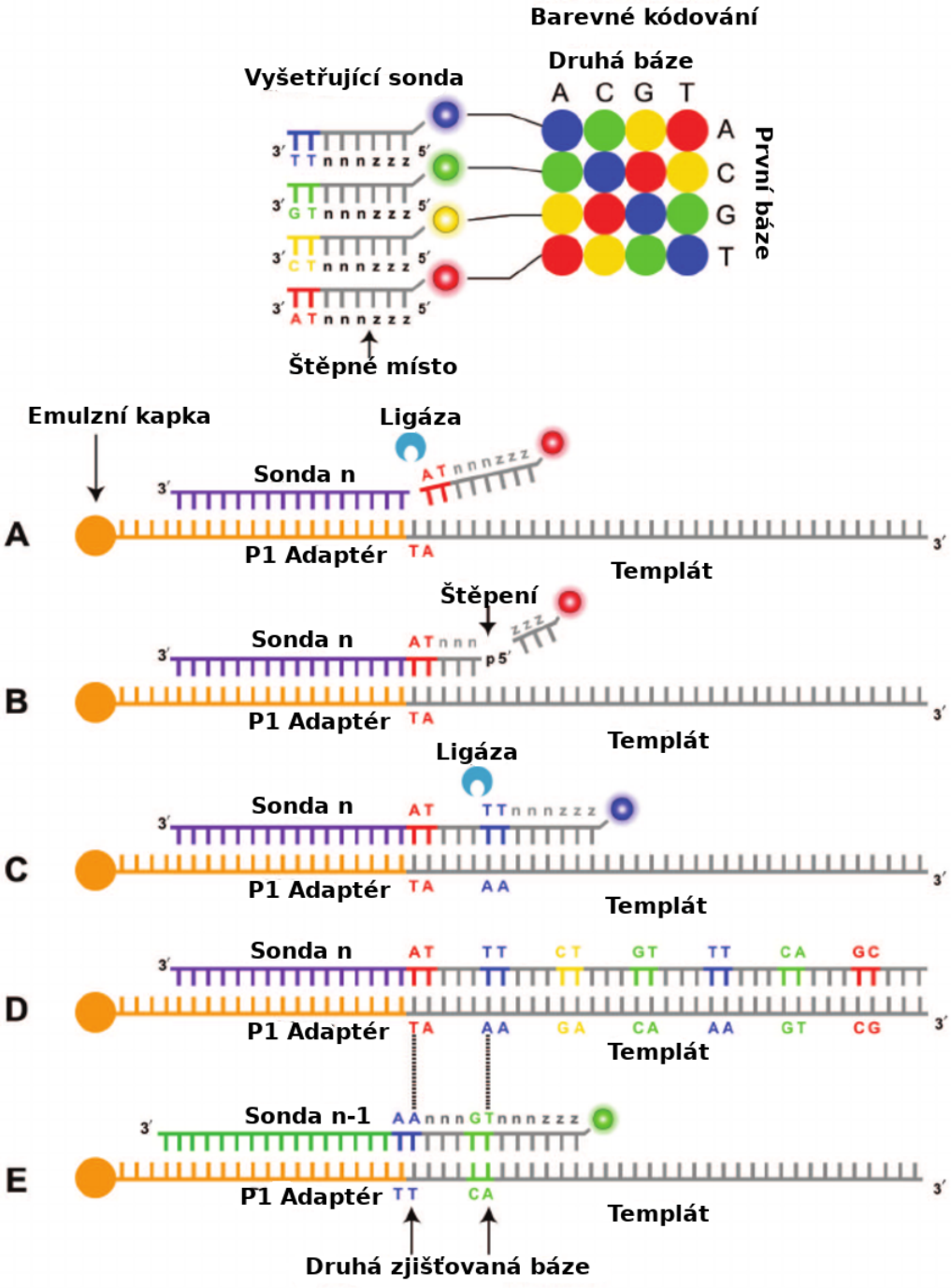
Nasednutí nového primeru je provedeno celkem pětkrát, což zajistí, že každý nukleotid v řetězci je přečten právě dvakrát. Na základě analýzy vlnové délky fluorescentů je poté sestavována sekvence templátového vlákna po dinukleotidech. Celý postup vidíme ilustrovaný na obrázku 1.5.

1.2.4 Ion Torrent

Metoda Ion Torrent byla představena společností Ion Torrent Systems Inc. v roce 2010. Spadá do kategorie sekvenačních přístupů pomocí syntézy řetězce. Ke svému chodu nevyužívá složité enzymatické reakce ani značené sondy či kamery. Sekvenace probíhá na citlivém polovodičovém čipu.

Základem metody je detekce vodíkového iontu, který se uvolní po začlenění nukleotidu do řetězce, tedy po vzniku fosfodiesterové vazby. [1]

Příprava vzorku je podobná Roche 454. Fragmenty DNA jsou immobilizovány na povrch mikročástice. Každá mikročástice má své místo v jamce na polovodičovém čipu a každá jamka je kontrolována pomocí ISFET (Ion-sensitive field effect transistor) iontového senzoru. Jamky jsou poté cyklicky zaplavovány vždy jen jedním druhem dNTP. Pokud je daný nukleotid komplementární, dojde k jeho začlenění do řetězce a tím k uvolnění vodíkového iontu. Uvolněný iont změní pH v jamce a tato



Obrázek 1.5: Postup sekvenace ligací SOLiD. Převzato z [16].

změna je detekována citlivým ISFET senzorem. Následně je aktuální směs dNTP vymyta a jamky jsou zaplaveny směsí jiného dNTP. [15]

1.3 Třetí generace

Třetí generace se vyznačuje především přístupem, kdy je čtena sekvence pouze z jedné molekuly a není tak nutné fragmenty DNA amplifikovat. Při amplifikaci mohou vznikat náhodné chyby, které pak proces sekvenace mohou ovlivnit.

1.3.1 SMRT

Metoda SMRT (Single Molecule Real-Time Sequencing) byla představena firmou Pacific Biosciences v roce 2009.

Metoda využívá zero-mode waveguide (ZMW) čipu. Jde o destičku s kruhovými jamkami. Do každé jamky je immobilizována právě jedna molekula DNA polymerázy a jedno vlákno templátové DNA. Celý systém obsahuje fluorescenčně značené dNTP. Značka je umístěna na konci fosfátové skupiny. Při navázání nukleotidu do vznikajícího řetězce je značka uvolněna a opouští detekční jamku. Detektor umístěný u dna jamky detekuje spektrum záření v jamce, a při opuštění značky zaznamená změnu spektra. Z této změny se poté určuje inkorporovaný nukleotid.

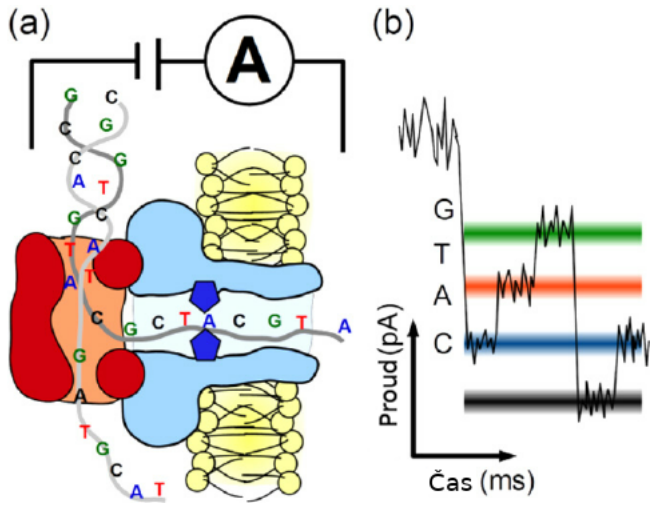
Tato metoda nepozastavuje proces prolongace komplementárního vlákna a začleněný nukloetid určuje v reálném čase. Tím je tato metoda extrémně rychlá a jako jediná zvládá detekovat změny na nukleotidech, například methylaci. [17]

1.3.2 Oxford Nanopore

Technologie byla představena už v roce 1995. Jde o představu molekuly DNA procházející malým pórem v membráně při aplikaci elektrického pole.

Využívá různé druhy pórů, buď biologické, například alfa hemolysin, nebo uměle vytvořené, například tunelování proudem elektronů do křemíkového čipu.

Systém vždy obsahuje nanopór a elektrolytický roztok. Při umístění systému do elektrického pole vzniká v systému elektrický proud. Velikost tohoto proudu v okolí nanopóru závisí na jeho velikosti a na typu nukleotidů procházejících pórem. Ze změny elektrického proudu tak určujeme procházející nukleotidy v póru. V případě příliš rychlého průchodu bazí pórem je třeba je brzdit navázáním objemnější molekuly. Postup je ilustrován na obrázku 1.6. [1]



Obrázek 1.6: Postup sekvenace pomocí nanopóru. Převzato z [18].

1.4 Sekvenační techniky

S nástupem Next-Generation metod je možné sekvenovat až milióny různých řetězců naráz. Tyto metody ale pracují s poměrně krátkými úseky DNA, které musí být navíc přítomny v milionech kopií. Dlouhé sekvence tedy musíme fragmentovat do kratších.

1.4.1 Shotgun sekvenování

Shotgun sekvenování je odvozeno od podobnosti s výstřelem z brokovnice. Broky po výstřelu mají náhodný rozptyl. Požadovanou sekvenci DNA náhodně fragmentujeme na kratší, často se překrývající úseky, a tyto úseky sekvenujeme zvolenou technologií. Tento přístup je vhodný pro celogenomové sekvenování.

1.4.2 Amplikonové sekvenování

Tato technika vyžaduje dopřednou informaci o námi sekvenované oblasti. Úsek který hodláme sekvenovat, si vymezíme pomocí dvou primerů. Následně amplifikujeme pouze úsek mezi těmito dvěma primery a ten je vstupem do sekvenátoru. Používá se při sekvenaci krátkých úseků, jednotlivých genů, nebo k vyhledávání polymmorfismů.

1.5 Sekvenační terminologie

V této kapitole je krátký přehled pojmů souvisejících s osekvenovaným genomem a se kterými se v rámci této práce ještě setkáme.

Čtení

Čtení je reprezentace fragmentu DNA, který byl sekvenován.

Délka čtení

Počet bazí v daném čtení. Dlouhá čtení (>400 bp) jsou vhodnější pro skládání dlouhých úseků. Krátká čtení (<150 bp) mají obvykle vyšší přesnost.

Výtěžek sekvenačního běhu

Celkový počet všech výstupních bazí ze všech čtení.

Hloubka pokrytí

Počet bazí ze čtení, které jsou mapovány na stejnou pozici. Jde o počet případů, kdy je konkrétní báze sekvenována nezávislými čteními.

Pokrytí

Udává kolikrát byl úsek osekvenován. Určujeme jako výtěžek sekvenačního běhu děleno velikostí sekvenovaného genomu.

Sestavení

Sloučení a seřazení jednotlivých čtení do delších úseků (kontigů), tak aby bylo možné určit původní sekvenci. Sestavení provádíme k referenční sekvenci (resekvenace) nebo de novo.

Skóre kvality

Udává kvalitu pro každou bázi v daném čtení. Tedy s jakou jistotou je báze určená správně. Nejčastější je Phred skóre, které lze vyjádřit rovnicí $Q = -10log_{10}P$, kde P je pravděpodobnost, že daná báze je určená špatně.

2 TREPONEMA PALLIDUM

Treponema pallidum je bakterie z kmene spirochet. Jde o Gramnegativní bakterii šroubovicovitého tvaru až 15 µm dlouhou s vnější i cytoplasmatickou membránou. [19]

Tento druh zahrnuje čtyři poddruhy:

- *Treponema pallidum pallidum*,
- *Treponema pallidum endemicum*,
- *Treponema pallidum careteum*,
- *Treponema pallidum pertenue*.

Treponema pallidum pallidum je původcem syfilidy, *T. endemicum* způsobuje endemickou syfilidu a bejel, poddruh *T. caretenum* způsobuje onemocnění zvané pinta a přítomnost poddruhu *T. pertenue* se projevuje jako yaws. Tyto poddruhy jsou morfologicky i sérologicky nerozeznatelné. *Treponema pallidum pallidum* je však pohlavně přenosná, její tvar jí umožňuje pohyb ve viskózním prostředí a přechází i do krve a lymfy. Ostatní poddruhy využívají jiné způsoby infekce.

Proti syfilidě není dostupná účinná vakcína, což je způsobeno především příliš malým počtem membránových proteinů na povrchu treponemální spirochety. Farmaka navržená proti této bakterii, tedy nemají dostatečný počet vazebných míst, na která by se mohla navázat a tím je značně snížen, či dokonce znemožněn jejich terapeutický účinek.

Laboratorní diagnostika této bakterie je umožněna mikroskopií v tmavém poli, či pomocí PCR, sloužící k identifikaci specifických genů, zejména genů z rodiny PolA. Rovněž je vyvinuto několik sérologických testů k identifikaci této bakterie v těle pacienta. [20]

Výzkum této bakterie ztěžuje i fakt, že je in-vitro nekultivovatelná a její kultivace je možná pouze v tkáňové kultuře. [21]

První kompletní sekvence *Treponemy pallidum* byla získána v roce 1998 [22]. Tato práce prokázala, že tato bakterie má jeden z nejmenších genomů. Celý genom obsahuje přibližně 1 140 000 párů bazí a asi 1041 ORF (open reading frame - úsek obsahující kódující sekvenci).

3 VARIABILITA GENOMU

Výskyt genomové heterogenity (zahrnující bodové mutace, indely a mobilní elementy jako plasmidy a fágy) je známý u kmenů mnoha patogenních bakterií. Heterogenní oblasti mohou přispívat k obraně vůči imunitní reakci hostitelského organismu, případně reprezentují adaptivní změny jako reakci na různorodé prostředí infikovaného organismu a jeho částí.

Identifikace heterogenních oblastí je důležitým krokem ke studiu infekčních mechanismů, šíření infekce a imunitních reakcí a identifikaci kmenových subpopulací.

Heterogenní oblasti v genomu se mohou vyskytovat pouze v malém zlomku celkové populace daného kmenu. Často se dostáváme na hodnoty výskytu v několik málo procentech a bývá tak obtížné určit, zda jde o heterogenní oblast, či pouze o náhodnou mutaci nebo chybu vyprodukovanou zvolenou sekvenační technologií. [23]

Variabilita v genomu je tedy důsledkem adaptace organismu na měnící se prostředí. Může být způsobena také náhodnou mutací jednotlivých pozic nukleotidů. Náhodné mutace, které se ukáží pro organismus jako nevýhodné jsou v rámci vývoje potlačeny, naopak mutace přinášející evoluční výhodu danému jedinci mají větší pravděpodobnost, že se rozšíří v rámci celé populace organismu.

Z chemické struktury nukleotidů rovněž vyplývá vyšší míra záměny jistých bazí za jiné. Obecně častější jsou záměny bazí s podobnou chemickou strukturou, tedy purinové báze za purinové a pyrimidinové za pyrimidinové. Takovéto bodové mutace nazýváme transice. Opakem transice je transverze.

V práci se budeme soustředit především na mutace genové, tedy takové, které probíhají na úrovní vlákna DNA a mění jeho primární strukturu. Kromě mutací genových rozlišujeme ještě mutace genomové a chromozomové. V případě genomových jde především o různé typy polyploidií, tedy znásobení celé chromozomální sady. V případě mutací chromozomových jde o strukturní změny na úrovni chromozomů, obecně označované jako aberace.

4 POUŽITÉ GENOMY A DATA

V rámci této práce budu pracovat se sekvenačními soubory dat poskytnutými Biologickým ústavem Lékařské fakulty Masarykovy univerzity.

Přehled všech sekvenovaných genomů ukazuje tabulka 4.1. Pokud jde o již publikovaný genom, uvádím zde i NCBI identifikátor kompletní nukleotidové sekvence daného genomu.

Několik genomů má již hotovou referenční sekvenci, na zveřejnění v databázi NCBI se ale stále čeká. U těchto genomů byla referenční sekvence pro zpracování genomu poskytnuta Biologickým ústavem Lékařské fakulty Masarykovy univerzity.

U několika dalších genomů bude jako reference využita sekvence jiného příbuzného genomu, vhodnost daného příbuzného organismu byla konzultována s odborníky z Biologického ústavu Lékařská fakulty Masarykovy univerzity.

Tabulka 4.1: Analyzované genomy

	Genom	**Reference***
TEN	IraqB	CP007548.1 (BosniaA)
TPA	Grady	CP004011.1 (SS14)
	HaitiB	LF BIO
	Madras	CP004010.2 (Nichols)
	Philadelpia-1	LF BIO
	UZ1974	LF BIO (Philadelphia-1)
TPE	CDC-1	LF BIO
	CDC-2575	CP020366.1
	Ghana-051	CP020365.1
	Kampung Dalan	LF BIO
	M540	LF BIO
	Sei Geringging	LF BIO

* GenBank identifikátor použité sekvence, v závorce je uveden referenční genom liší-li se od zpracovávaného. LF BIO = sekvence poskytnutá Biologickým ústavem Lékařské fakulty Masarykovy univerzity.

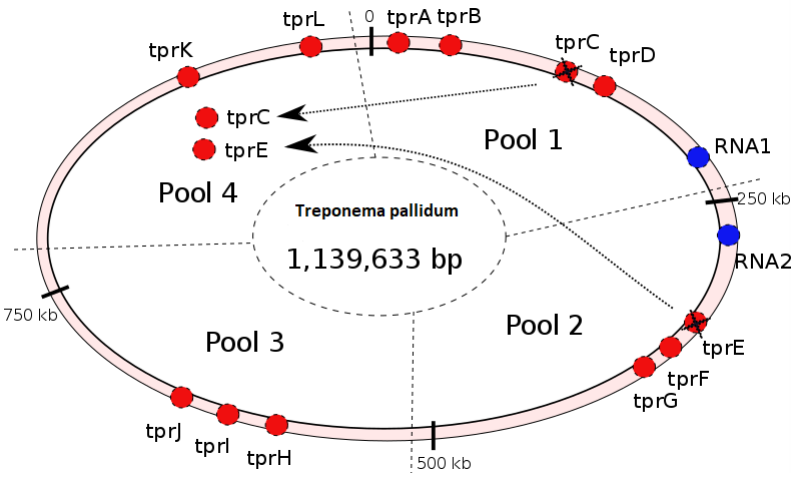
Označení kmene je následující. Bakterie označené jako TEN patří k *Treponema pallidum subsp. endemicum*, TPA jsou *Treponema pallidum subsp. pallidum* a TPE je *Treponema pallidum subsp. pertenue*.

5 PŘEDZPRACOVÁNÍ SEKVENAČNÍCH DAT

5.1 Sekvenování

Kruhový genom *Treponemy pallidum* obsahuje několik vzájemně si podobných genů, zejména z rodiny genů *tpr*. Při sekvenování je to třeba brát v potaz, aby výsledné čtení mohlo být jednoznačně přiřazeno ke svému původu. V případě nejednoznačnosti původu daného čtení je takovéto čtení při sestavování genomu umístěno na základě pravděpodobnosti, náhody, či zcela vyřazeno. Ve všech případech by to tedy příspívalo k nežádoucímu zkreslování sestavované sekvence a její nižší kvalitě, případně k falešně pozitivním identifikacím variabilních míst.

Z tohoto důvodu je celý genom rozdělen do čtyř částí, takzvaných poolů. Rozdělení je provedeno tak, aby příliš podobné geny nebyly sekvenovány v rámci stejného poolu. Rozdělení vidíme na obrázku 5.1. Vidíme, že mimo rozdělení na pooly dochází i k vystřižení genu tprC a genu tprE ze svých poolů a jejich sekvenování probíhá společně se sekvencí čtvrtého poolu.



Obrázek 5.1: Rozdělení genomu na pooly.

Jednotlivé pooly jsou rozděleny přibližně po čtvrtinách genomu a mají mezi sebou několik stovek nukleotidů přesah. Přesah poolů mezi sebou je volen záměrně, protože při sekvenování dochází ke snižování hloubky pokrytí na koncových částech sekvenovaných úseků. Při sestavení celého genomu se pak použijí čtení z obou překrývajících se poolů a dojde tak k navýšení počtu čtení v těchto překrývajících se oblastech. Parametry jednotlivých poolů jako je jejich délka, velikost překryvu a jejich ohraničující sekvence, které využijeme při identifikaci sekvencí jednotlivých poolů v sekvenci celého genomu, ukazuje tabulka 5.1.

Sekvenování genomů je provedeno na přístrojích Illumina MiSeq. Výstupem je soubor párových čtení pro každý pool zvlášť.

Tabulka 5.1: Parametry poolů v genomu.

POOL	**Délka***	**Překryv****		**Sekvence**
Pool 1	259 800 bp	800 bp	začátek	CGGCTGTATTTCGTTACTGTCTTGT
			konec	TTGGACACGGACCTACGTACGCCAT
Pool 2	253 300 bp	950 bp	začátek	GCGCAGTAAAAGAGGGACGACCTCT
			konec	CCGGTTCAAGTGGTGCATAGCAAGC
Pool 3	254 500 bp	2200 bp	začátek	CCTGCAGGGTACGTAAGTAGAGGAC
			konec	AAGCACGTAGCTCTTCCTCGCGTAT
Pool 4	376 200 bp	400 bp	začátek	TCGATAAAGGCTGCGATACCTCCAC
			konec	AACTTCTCCAAGGCAGGAGGAGTGT
(tprC)	4 700 bp	-	začátek	TTATCAGCCTGAATCGTATGTCCCT
			konec	TCACCGAGACTGCAACAATGGCTCT
(tprE)	5 700 bp	-	začátek	TTCCGAGCCATATCTGCGTACTGCG
			konec	AACACCGGCCGCAGAACATTAAACG
Genom	**1 154 200 bp**			

** Překryv je udáván s následujícím poolem. * Délka je přibližná, v konkrétních genomech se může lišit.

5.2 Sestavení

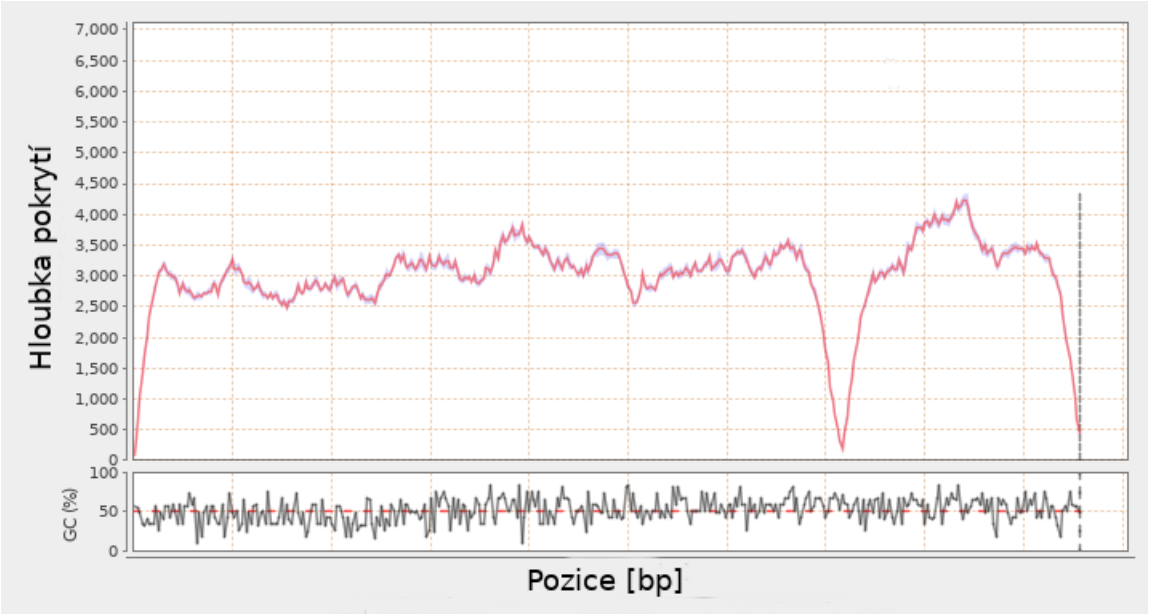
Při sestavování nesestavuji celý genom, ale opět postupuji po jednotlivých poolech a každý je sestaven zvlášť. Při sestavování využívám mapování vůči známé referenci.

Pokud není známá referenční sekvence daného organismu, lze využít i známou sekvenci nějakého příbuzného organismu, kde je předpoklad velké míry shody v DNA sekvencích obou organismů. U takovýchto sestavení, však musíme počítat s nižší kvalitou mapovaných čtení v neidentických úsecích a vyšším zastoupením indelů v sestaveném souboru. Dále takové sestavení samozřejmě zvyšuje hodnotu odhadu sekvenační chyby daného souboru.

Referenční genom si rozdělím na jednotlivé pooly. Při rozdělování využiji znalosti sekvencí ohraničujících jednotlivé pooly, neboť štěpení genomu probíhalo ve stejných místech u všech zpracovávaných genomů. Jednotlivá párová čtení nechám namapovat k referenci pomocí programu BWA a algoritmu mem, vhodným pro kratší čtení z platformy Illumina [24].

5.3 Kontrola kvality sestavení

Po sestavení jsou zkontrolovány parametry sestaveného souboru. Pro kontrolu parametrů využívám volně dostupného nástroje FastQC [25][26]. Zejména pak hloubku pokrytí, jelikož se v referenčních sekvencích nachází i hypervariabilní gen *tprK*, pro který často nebyla určena konsensuální sekvence a je tak v genomu značen 'N' bázemi. V okolí tohoto genu bude hloubka pokrytí značně klesat. Rovněž v místech vystřižených genů *tprC* a *tprE* dojde k poklesu hloubky pokrytí na nulovou hladinu. Příklad vidíme na obrázku 5.5.

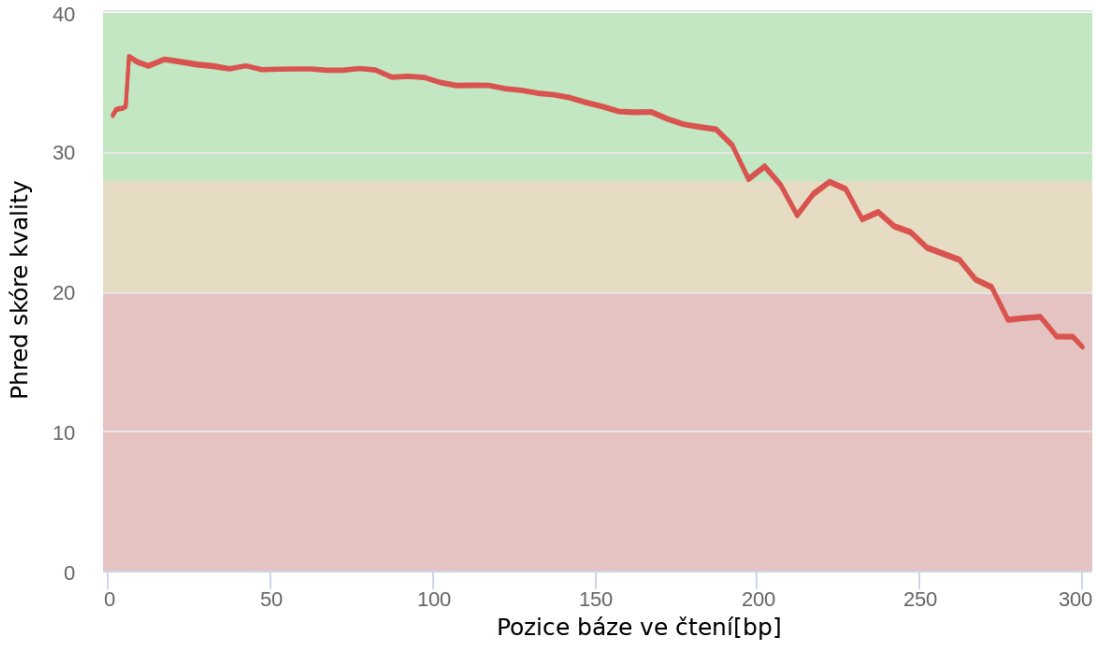


Obrázek 5.2: Hloubka pokrytí poolu 4 genomu Philadelphia-1.

Dále zkontroluji čtení na přítomnost adaptérů. Tedy připojených sekvencí, sloužících k odlišení sekvení v rámci sekvenování několika různých organismů zároveň. Před sestavováním souboru musí být tyto adaptéry ze čtení odstřiženy, neboť jejich sekvence namají původ v sekvenovaném souboru. Jejich přítomnost negativně ovlivní kvalitu sestavení, proto je třeba se jejich zahrnutí vyvarovat.

Dále sleduji míru zastoupení duplikovaných čtení. Tato hodnota by měla být co nejnižší. Vyšší zastoupení duplikátů čtení v sestavovaném souboru způsobuje jeho nižší kvalitu, protože duplikáty nepřinášejí žádnou novou informaci o úseku, na který jsou mapovány. V důsledku toho dochází k rozředění informace o zastoupené bázi.

Sleduji také skóre kvality jednotlivých čtení, respektive jejich průměrné hodnoty. Nižší kvalita čtení ukazuje na vyšší sekvenační chybu, či nepřesnosti v postupu přípravy vzorku. Průměrnou hodnoty kvalitativního skóre v rámci poolu 4 genomu Philadelphia-1 ukazuje obrázku 5.3.



Obrázek 5.3: Průměrná kvalita čtení báze v rámci celé délky čtení.

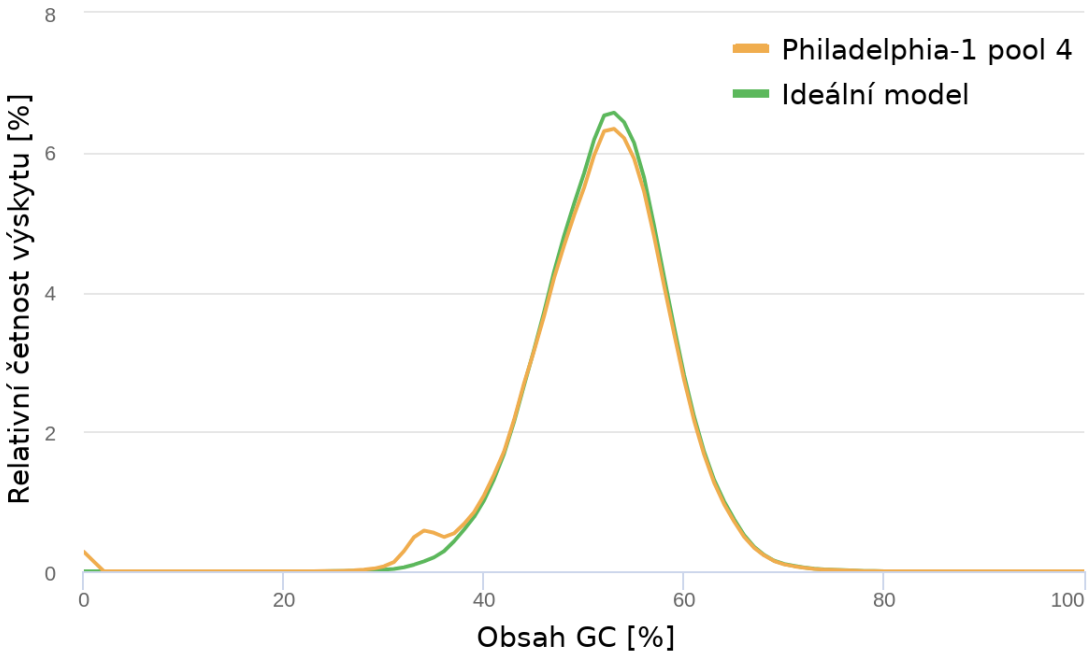
O kvalitě přípravy sekvenovaného vzorku vypovídá také poměr nenamapovaných čtení v sestaveném souboru. Pokud je tento parametr příliš vysoký, ukazuje to na kontaminaci vzorku cizí DNA, případně na použití špatné referenční sekvence.

O kvalitě sekvenace nám také vypovídá parametr udávající obsah GC v jednotlivých čteních. Distribuce tohoto parametru by měla vykazovat normální rozložení se střední hodnotou odpovídající zastoupení GC v sekvenovaném genomu. V našem případě by se střední hodnota měla pohybovat okolo 53 %. Pokud tento parametr nevykazuje normální rozložení, uvažujeme opět o kontaminaci sekvenovaného vzorku, či velmi nízké kvalitě sekvenace. Ukázku z genomu Philadelphia-1 pool 4 vidíme na obrázku 5.4.

5.4 Úprava sestavení

Po sestavení jednotlivých poolů je vhodné před další analýzou toto sestavení upravit s ohledem na další práci se souborem a na zjištěné kvalitativní parametry sestaveného souboru. Úpravou sestavení myslíme především filtraci a odstranění nekvalitních čtení ze souboru a tím zvýšit jeho kvalitu.

Zejména je vhodné ze souboru čtení odstranit ty, u kterých nebyl určen jejich původ a zůstaly nenamapované. Jejich sekvence neodpovídá žádné části v referenčním genomu, jde nejspíše o různá drobná znečištění sekvenovaného vzorku.



Obrázek 5.4: Zastoupení GC a ideální model.

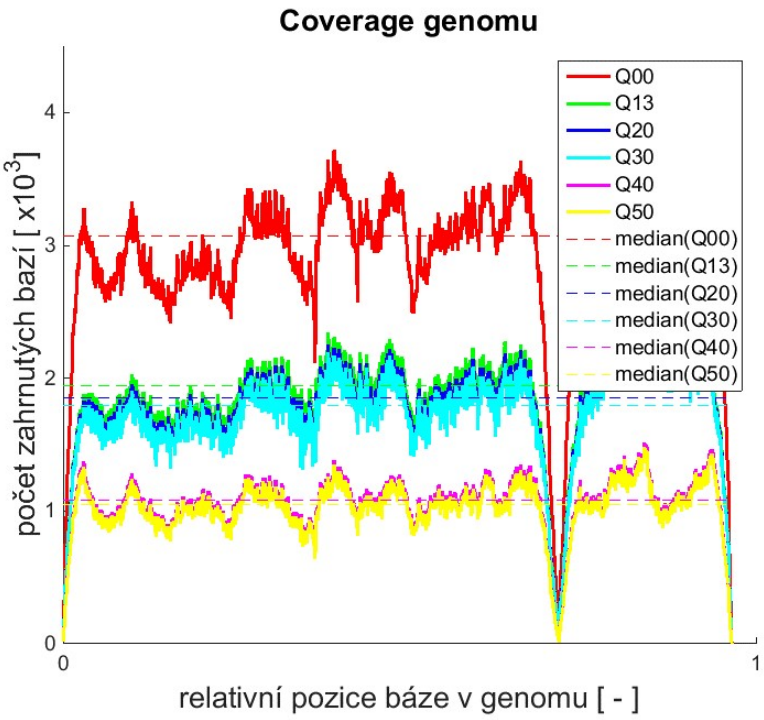
Dalším krokem je odstranění duplikovaných čtení. Tyto čtení vznikají špatným vymytím vzorku a proběhnutím PCR na vzniklých čteních. Takováto duplikovaná čtení by pak mohla vnášet chybu do následné analýzy variability.

Ze souboru rovněž odstraníme čtení s nízkou průměrnou kvalitou, která by mohla zkreslovat oblasti, na které se mapují.

Pro jednodušší a rychlejší práci je také vhodné si zpracovávaný soubor ve formátu SAM/BAM setřídit, tedy vypisovat čtení postupně tak, jak se daný úsek vyskytuje v referenčním genomu. Usnadní to následnou práci se souborem, vyhledávání v něm a také umožní efektivnější kompresi celého souboru, jelikož jde o poměrně velké textové soubory v řádu GB. Vhodné je rovněž vytvořit si index daného souboru pro rychlejší manipulaci a vyhledávání.

Pro úpravy sestavení využívám sady nástrojů SAMtools [27], zejména pak jeho příkazů pro filtraci na základě příznaků čtení, odstranění duplikovaných čtení a setřízení souboru.

Po těchto úpravách je vhodné opět zkontrolovat kvalitativní parametry sestaveného souboru. Mělo by dojít ke zlepšení jednotlivých hodnot, mimo hloubky pokrytí. Ta vlivem odstranění mnoha nekvalitních čtení může i výrazně poklesnout. V případě, že se nepodařilo úpravami sestavení kvalitativně vylepšit, je na místě uvažovat o chybě při zpracování celého vzorku a jaho sekvenaci provést opětovně, případně takový genom ze zpracovávaného souboru zcela vyřadit.



Obrázek 5.5: Hloubka pokrytí při různém skóre kvality.

Předchozí zmíněnné úpravy se týkaly především celých sekvencí čtení. Při identifikaci výsledné sekvence pracuji s jednotlivými pozicemi bází ve čtení. Celý soubor je převeden do pozičně orientovaného souboru *.vcf (variant call format). Z celkového sestavení vyexportujeme informaci o počtu a druhu čteních vztažených k jednotlivým nukleotidům referenční sekvence. Při převodu do tohoto formátu zavádím ještě jeden filtr kvality, kdy z jednotlivých sekvencí celých čtení, zachovávám pouze ty pozice, kde se nachází čtení dané pozice o vysoké kvalitě. Kvalita čtení jednotlivé pozice může totiž v rámci jedné sekvence čtení značně kolísat. Zachovávám tak pouze informace ze čtení pozice, u kterých je minimální šance na chybnou interpretaci sekvenační technologií.

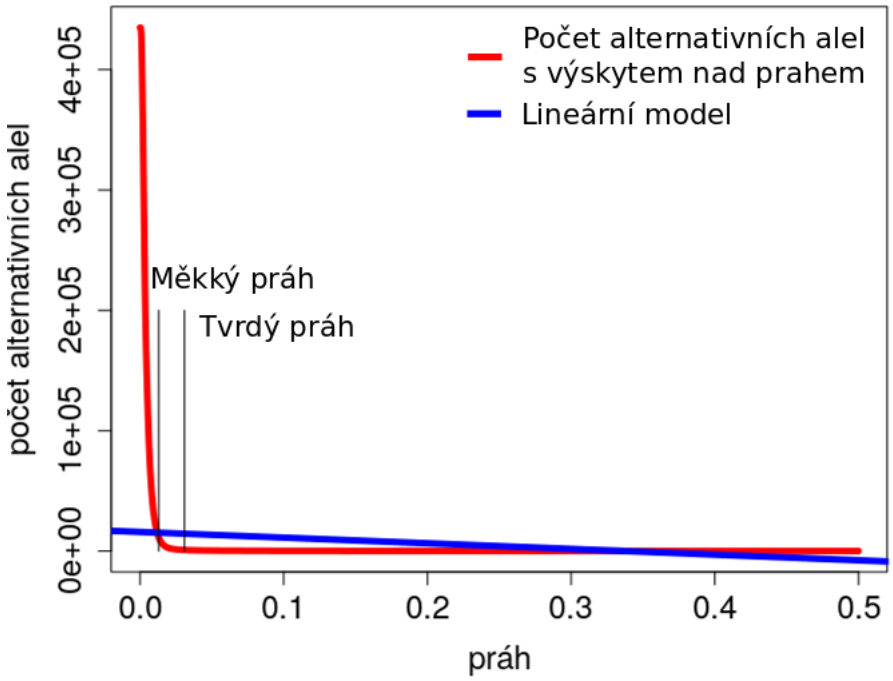
Tento filtr zvyšuje jistotu správně identifikované báze na dané pozici, ovšem opět za cenu značného snížení pokrytí dané báze. Vliv hloubky pokrytí na kvalitě zahrnutých čtení pozice je ukázán na obrázku 5.5.

6 IDENTIFIKACE VARIABILNÍCH MÍST

6.1 Odhad míry chybovosti

Na určování míry chybovosti v sekvenačních datech stále neexistuje jednoznačný ucelený postup. Nejčastěji metody počítají s odhadem na základě počítání rozdílných čtení pozice od pozice čtení v namapovaném souboru. Způsoby dalšího odhadu se liší od jednoduchých metod, které počítají poměr rozdílných čtení ku všem čtením, po sofistikovanější metody založené na stínové regresi a dalších matematických modelech. [28] [29]

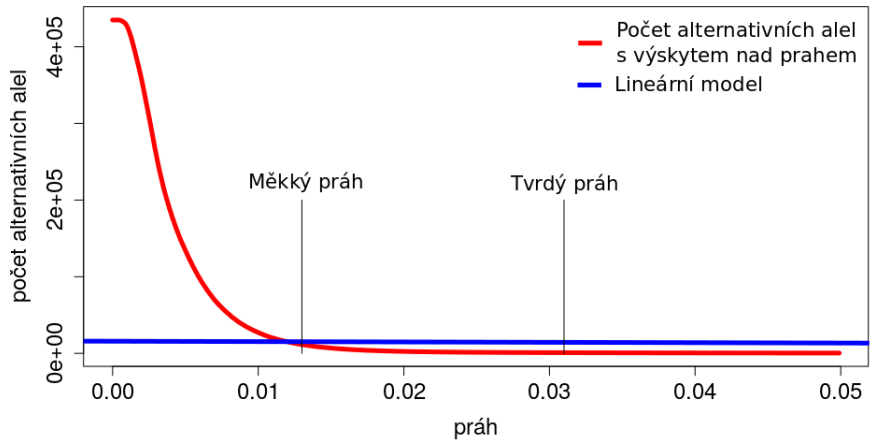
Pro účely této práce byla navržena vlastní metoda určování odhadu chybovosti. Pro její výpočet nejprve vytvořím datovou řadu závislosti počtu nalezených alternativnách alel, jejichž výskyt je vyšší jak stanovený práh, na velikosti tohoto prahu. Velikost prahu volím v rozmezí 0–0,5. Horní hodnota je nastavena na 0,5 záměrně, jelikož alely s výskytem vyšším jak tato mez, již nelze považovat za alely alternativní, ale za alely dominantní. Zobrazím-li si tuto závislost v grafu, pozoruji s klesajícím prahem nejprve zhruba lineární nárůst počtu nalezených alternativních alel s následným exponenciálním růstem.



Obrázek 6.1: Stanovení prahu.

Dále vycházím z poznatku, že heterogenní pozice a významná variabilní místa mají tendenci se v genomu konzervovat a propagovat se do dalších generací. Tedy jejich výskyt je v organismu častější. Tyto pozice nalezneme především v lineární části zmíněného grafu. V exponenciální části se proti tomu nachází i alely, které jsou způsobené čistě náhodnou mutací, sekvenační chybou či jiným šumem.

Tento navržený postup, ilustrován na obrázku 6.1 genomu TPA Madras, se tedy soustředí na identifikaci místa, kde lineární část přechází do exponenciálního průběhu. Celá datová řada je aproximována lineárním modelem. Následně je určen tvrdý práh jako velikost prahu, kde se nachází minimum residuí sestaveného modelu. Sklon linárního modelu je o něco vyšší než sklon lineární části průběhu. Minimum residuí je tedy taková pozice prahu, kde se lineární část přestává vzdalovat modelovanému průběhu. Měkký práh poté určím jako velikost prahu, kde se nachází minimum residuí navýšené o směrodatnou odchylku celého modelu. Celý postup vidíme na obrázku 6.2, kde je stanovení prahů ukázáno v detailu na genomu TPA Madras.



Obrázek 6.2: Stanovení prahu - detail.

6.2 Identifikace variabilních míst

V daném sestavení se snažím nalézt ty pozice, kde se co do identifikovaného nukleotidu jednotlivá čtení rozcházejí. Vzhledem k povaze heterogenity, která se vyskytuje i ve velmi malém procentu populace, musíme vytvořit postup, který bezpečně oddělí opravdu variabilní místa od míst, která jsou jako variabilní identifikována na základě náhodné mutace nebo sekvenační chyby či chyby sestavení.

Z analyzovaného souboru vyřadím všechny pozice, jež mají nízkou hloubku čtení nebo je referenční báze ve více jak 99 % čteních, takovýto úsek je považován za vysoce konzervovaný. Alternativní báze musí být podpořena alespoň 8 čteními, aby bylo zamezeno výskytu falešně pozitivní identifikace variability. Poměr čtení přímého a reverzního vlákna se musí pohybovat mezi 0,4–2,3 (maximální povolený poměr uvažujeme 30/70 (resp. 70/30)), v případě, že tomu tak není je úsek považován za vysoce chybově čtený. Ze souboru vyřadím také veškeré pozice nacházející se v homopolymerních úsecích (>5 bp) nebo v jejich okolí (± 5 bp), kde je chybovost technologie mnohem vyšší.

Ze všech čtení pozice navíc vyřadím ty, u kterých je zastoupení alternativní alely nižší jak identifikovaný práh variability. Práh variability určíme jako odhad sekvenační chyby. U alel, které mají zastoupení pod tímto prahem, není možno dostatečně bezpečně rozhodnout zda jde o reálně variabilní místo, či jde o chybné čtení dané pozice způsobené buď sekvenační technologii, nečistotammi, či jde o čistě náhodnou chybovou mutaci.

Jelikož zpracovávám i genomy, které jsou mapovány nikoli ke své vlastní referenci, ale k referenci příbuzných druhů, objeví se v souborech i pozice, které mají hloubku pokrytí referenční bází pod 50 %. U takovýchto pozic zaměním referenční a nejčastější alternativní alelu, neboť je pravděpodobné, že na tomto místě je rozdíl i v sekvencích zpracovávaného a referenčního genomu.

Shrňme si nyní přehledně veškeré aplikované podmínky, které byly určeny k identifikaci místa v genomu jako variabilního:

- vysoká kvalita čtení,
- dostatečná hloubka pokrytí,
- nejde o konzervovaný úsek,
- alternativní alela je podpořena minimálním počtem čtení,
- poměr čtení v obou směrech je maximálně 30/70,
- místo se nenachází v homopolymerním úseku ani jeho blízkém okolí,
- výskyt alternativní alely je vyšší jak hodnota odhadu míry chybovosti.

6.3 Parametrizace variabilních míst

Každé identifikované variabilní místo v genomu dále analyzuji. Zjistíme, zda místo spadá do kódujícího úseku genomu, a v případě, že ano, určím i zasažený gen. Uvádím rovněž i míru variability daného úseku. Nyní uvádím přehled nejdůležitějších zjišťovaných parametrů:

- pozice místa v genomu,
- pozice místa v alignmentu,
- pozice vztažená ke genomu Ghana-051,

- referenční alela,
- alternativní alela,
- relativní počet čtení pozice podporující alternativní alelu (frekvence výskytu alternativní alely),
- zasažený gen (vzhledem ke genomu Ghana-051).

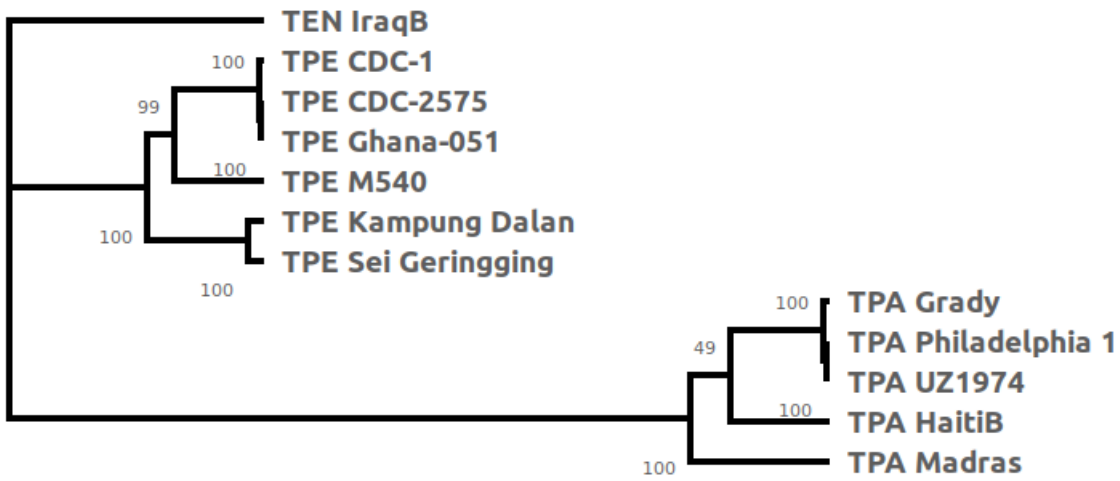
7 CELOGENOMOVÉ ZAROVNÁNÍ

Všechny použité sekvence treponemálních kmenů je třeba zarovnat. Jelikož jde o sekvence genomů stejného druhu (*Treponema pallidum*), jsou si všechny velmi podobné a nejsou od sebe fylogeneticky příliš vzdálené.

Pro zarovnání jsem využil algoritmu MUSCLE [30], především kvůli jeho vyšší rychlosti a nižší výpočetní náročnosti. Zarovnání probíhá v posuvném okně o délce 5 000 bp s překryvem 2 000 bp. Před posunem okna jsou odstraněny veškeré mezery, které jsou přidány na konec aktuálně zpracovávaného okna. Nastavené hodnoty jsou zvoleny tak, aby byl celý postup rychlý a zároveň dokázal zohlednit i poměrně dlouhé indely. Celý zarovnaný soubor je poté ještě manuálně zkontrolován.

V zarovnání je třeba zohlednit zejména různé sekvence alely stejného genu u různých poddruhů zpracovávaného organismu. Jde především o alely genu *tprD/D2* a repetiční úseky genu *arp*. Při dalším hodnocení variability v těchto oblastech je třeba brát v potaz, že ač může jít o stejnou pozici v zarovnání, nemusí jít o stejnou alelu v porovnávaných genomech.

Z celogenomového zarovnání následně sestrojím fylogenetický strom metodou UPGMA s využitím modelu Kimura-Nei, jelikož jde o značně příbuzné sekvence. K sestrojení využíváme softwarový nástroj MegaX [31]. Pro jeho otestování volíme bootstrapovou metodu na základě 500 stromů. Výslednou příbuznost zpracovávaných genomů vidíme na obrázku 7.1.



Obrázek 7.1: Fylogenetický strom zpracovávaných genomů.

8 VYHODNOCENÍ VARIABILNÍCH MÍST

Výše uvedený postup aplikuji na veškeré zpracovávané genomy, pro které jsou k dispozici soubory sekvenačních dat. U každého genomu určím hodnoty tvrdého i měkkého prahu a podle nich i počet identifikovaných variabilních pozic. Přehled těchto hodnot ukazuje tabulka 8.1.

Tabulka 8.1: Nalezené variabilní pozice a hodnoty prahů.

	Genom	**Průměrná hloubka pokrytí**	**Tvrdý práh**	**Variabilních pozic**	**Měkký práh**	**Variabilních pozic**
TEN	IraqB	850,75	2,70 %	77	0,80 %	234
TPA	Grady	446,35	2,80 %	25	1,10 %	32
	HaitiB	1200,26	1,70 %	10	0,60 %	19
	Madras	402,04	3,10 %	29	1,30 %	68
	Philadelphia-1	678,40	3,20 %	10	1,30 %	628
	UZ1974	282,02	3,20 %	53	1,20 %	89
TPE	CDC-1	148,59	4,10 %	7	1,80 %	7
	CDC-2575	148,58	4,10 %	7	1,80 %	7
	Ghana-051	446,39	4,40 %	40	1,30 %	148
	Kampung Dalan	374,53	2,30 %	0	1,00 %	2
	M540	62,92	9,10 %	4	3,50 %	11
	Sei Geringging	198,49	3,40 %	7	1,60 %	21

Nyní se soustřeďme na genomy TEN IraqB a TPA Grady, Madras a UZ1974. Pro tyto genomy bylo při sestavení k referenci využito referenčního genomu nějakého příbuzného organismu. Vzhledem k této skutečnosti se budou mezi variabilními alelami nacházet i takové, které jsou ve skutečnosti alelami dominantními, z důvodu rozdílné sekvence osekvenovaného genomu a použitého referenčního genomu. V těchto případech bude počet čtení alternativní alely vyšší jak 50 %, ve většině případů se bude dokonce blížit hodnotě 100 %. V takovýchto případech v datovém souboru zaměníme referenční a alternativní alelu a odpovídající hodnoty. Seznam takovýchto identifikovaných pozic nalezneme v příloze A.

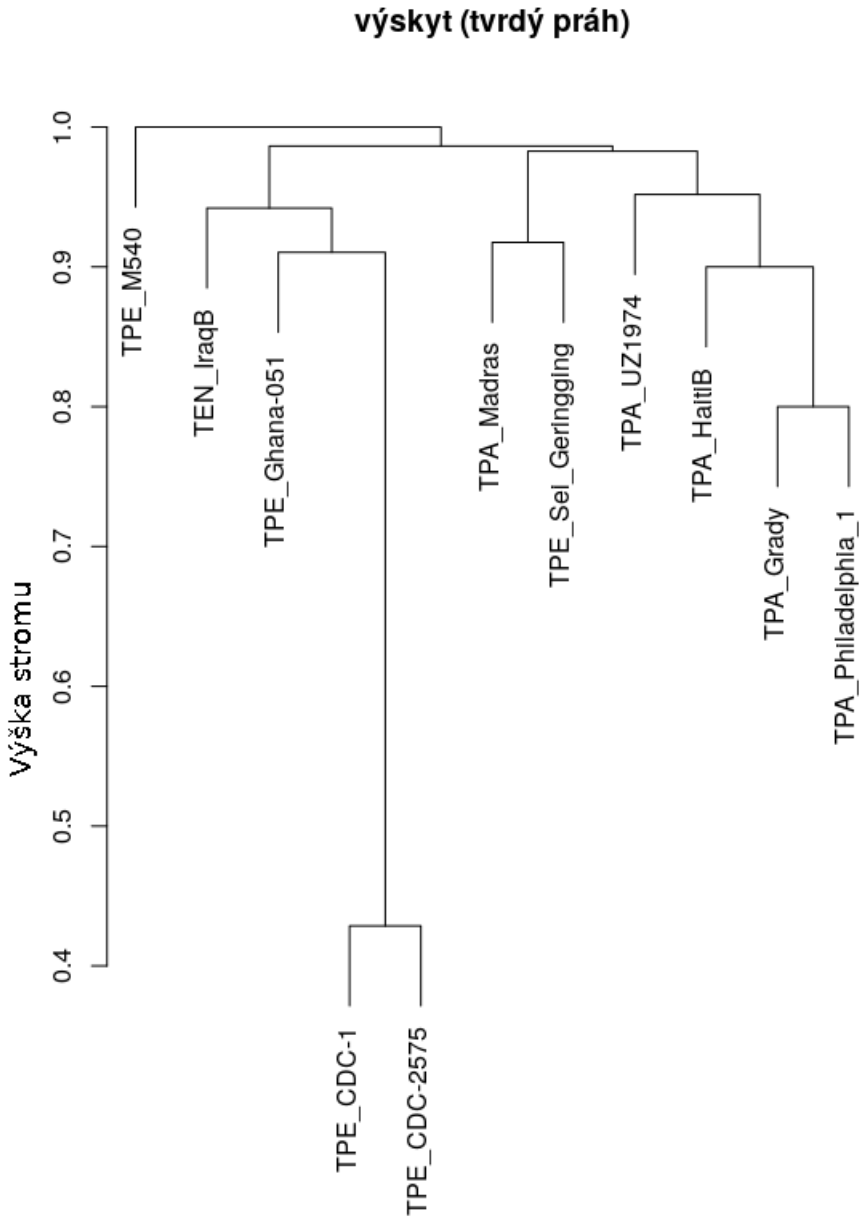
Po provedení záměny ověřím, zda tato pozice stále splňuje podmínky definované pro identifikaci variabilních úseků. Soubor podle toho upravím. Upravené počty variabilních míst ukazuje tabulka 8.2.

Nalezená variabilní místa jsou promítnuty do společného souboru zarovnání. Promítnutí provedu pro oba prahy a to jak čistě binárně na základě výskytu, tak i s nalezenými relativními četnostmi výskytu alternativních alel. K takovýmto datům

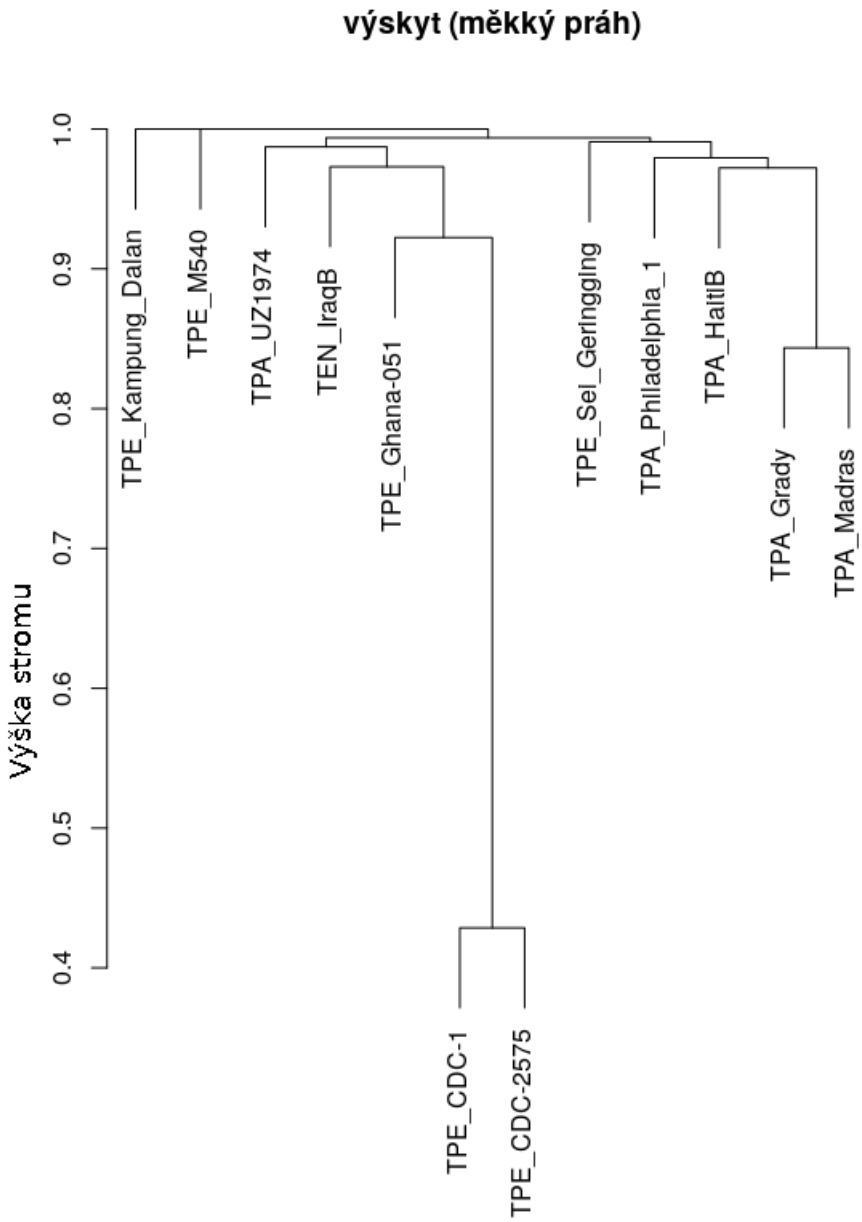
Tabulka 8.2: Nalezené nedominantní variabilní pozice genomů mapovaných k příbuzné referenci.

	Genom	**Průměrná hloubka pokrytí**	**Tvrdý práh**	**Variabilních pozic**	**Měkký práh**	**Variabilních pozic**
TEN	IraqB	850,75	2,70 %	50	0,80 %	207
TPA	Grady	446,35	2,80 %	10	1,10 %	17
	Madras	402,04	3,10 %	21	1,30 %	60
	UZ1974	282,02	3,20 %	43	1,20 %	79

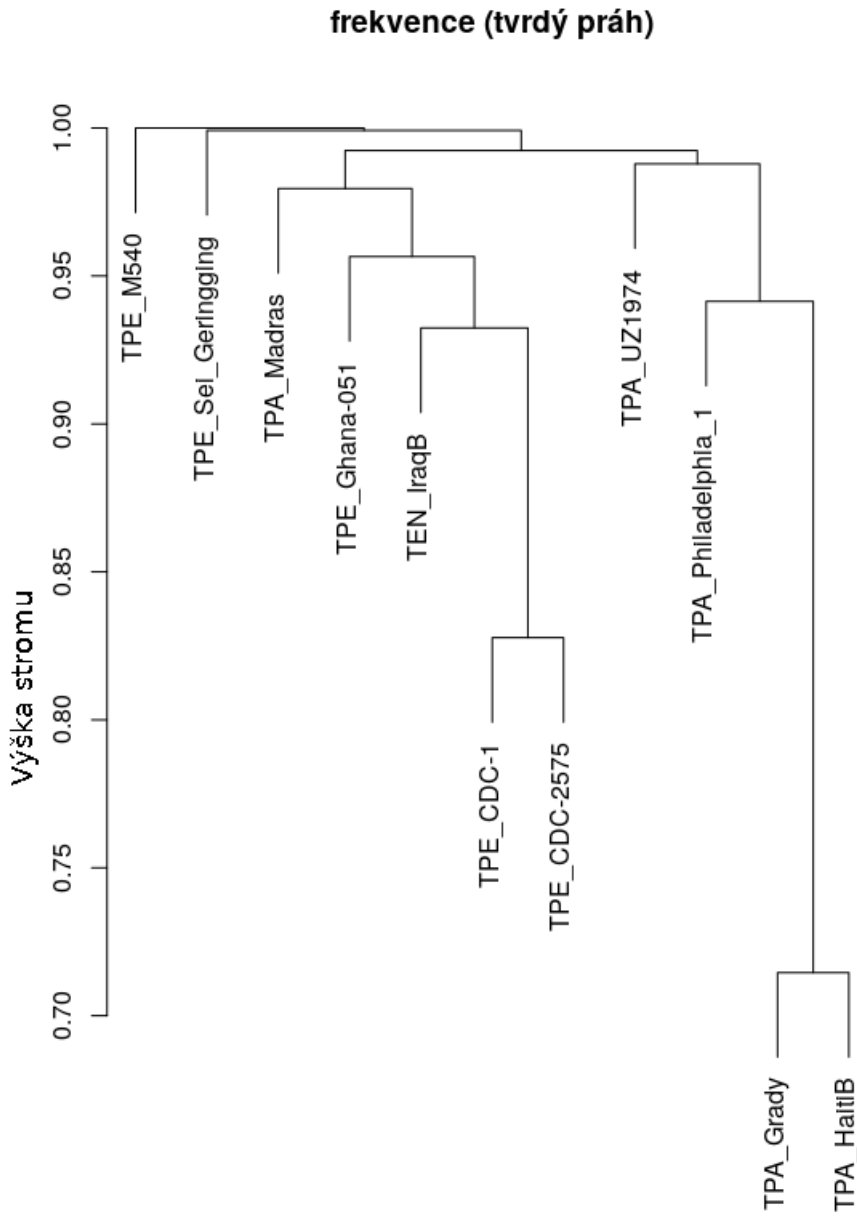
můžeme přistupovat také jako k signálu. Pro vizualizaci a představu o rozložení variabilních alel v genomech využiji hierarchickou shlukovou analýzu. Vstupem shlukové analýzy bude vzájemná korelace výskytu variabilních míst i relativní četnosti výskytu variability na dané pozici v genomech. Vzájemné vzdálenosti genomů vyjádřím pomocí Pearsonova korelačního koeficientu odečteného od jedné, aby jsme příbuznost převedli na vzdálenost. Vzdálenost následně normalizujeme do rozsahu 0–1. Z matice vzdáleností vytvořím stromy pomocí metody UPGMA (metoda párování pomocí nevážených aritmetických průměrů). Výsledný strom vytvořený čistě na základě výskytu variabilních alel s tvrdým prahováním vidíme na obrázku 8.1, při prahování pomocí měkkého prahu pak na obrázku 8.2. Stromy vytvořené i na základě zohlednění míry variability dané pozice pak ukazujeme na obrázku 8.3 a obrázku 8.4.



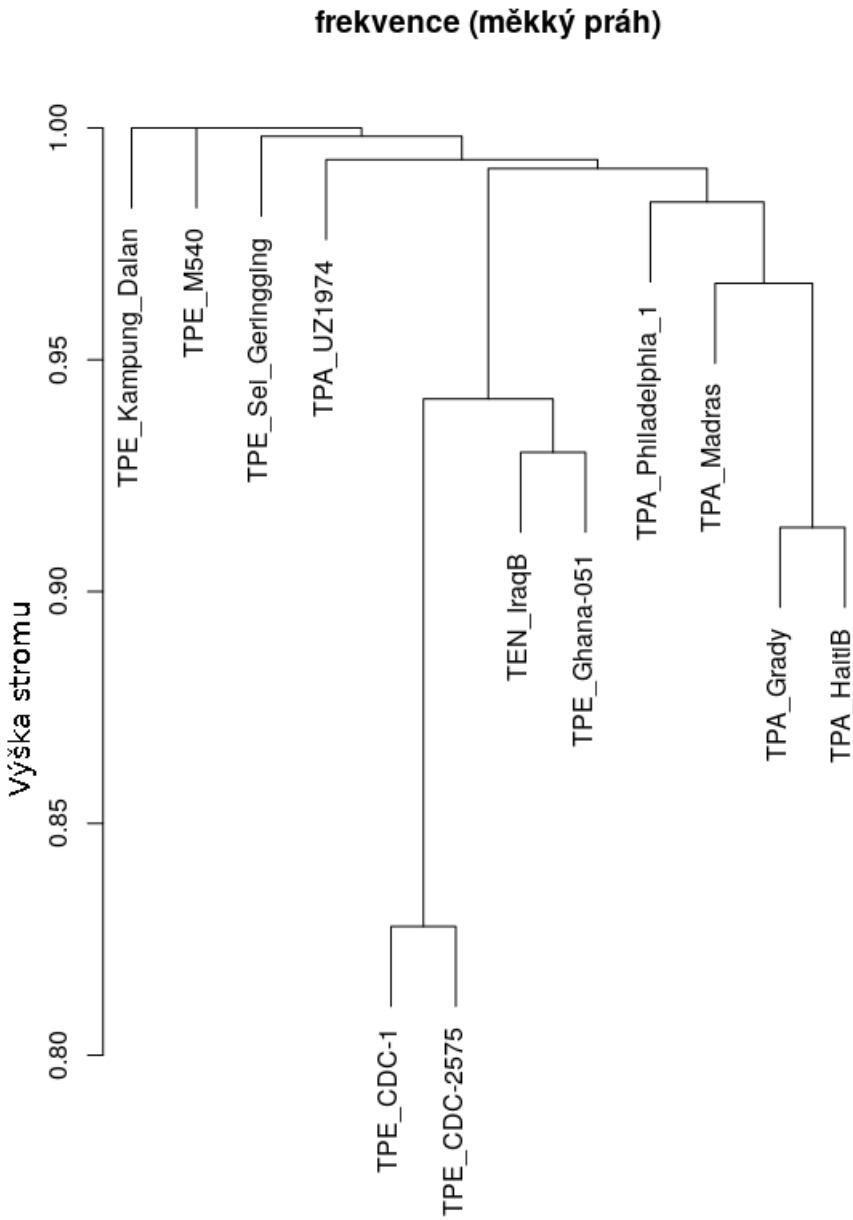
Obrázek 8.1: Strom na základě výskytu varibilních míst při tvrdém prahování.



Obrázek 8.2: Strom na základě výskytu varibilních míst při měkkém prahování.



Obrázek 8.3: Strom na základě relativní četnosti výskytu varibilních míst při tvrdém prahování.



Obrázek 8.4: Strom na základě relativní četnosti výskytu varibilních míst při měkkém prahování.

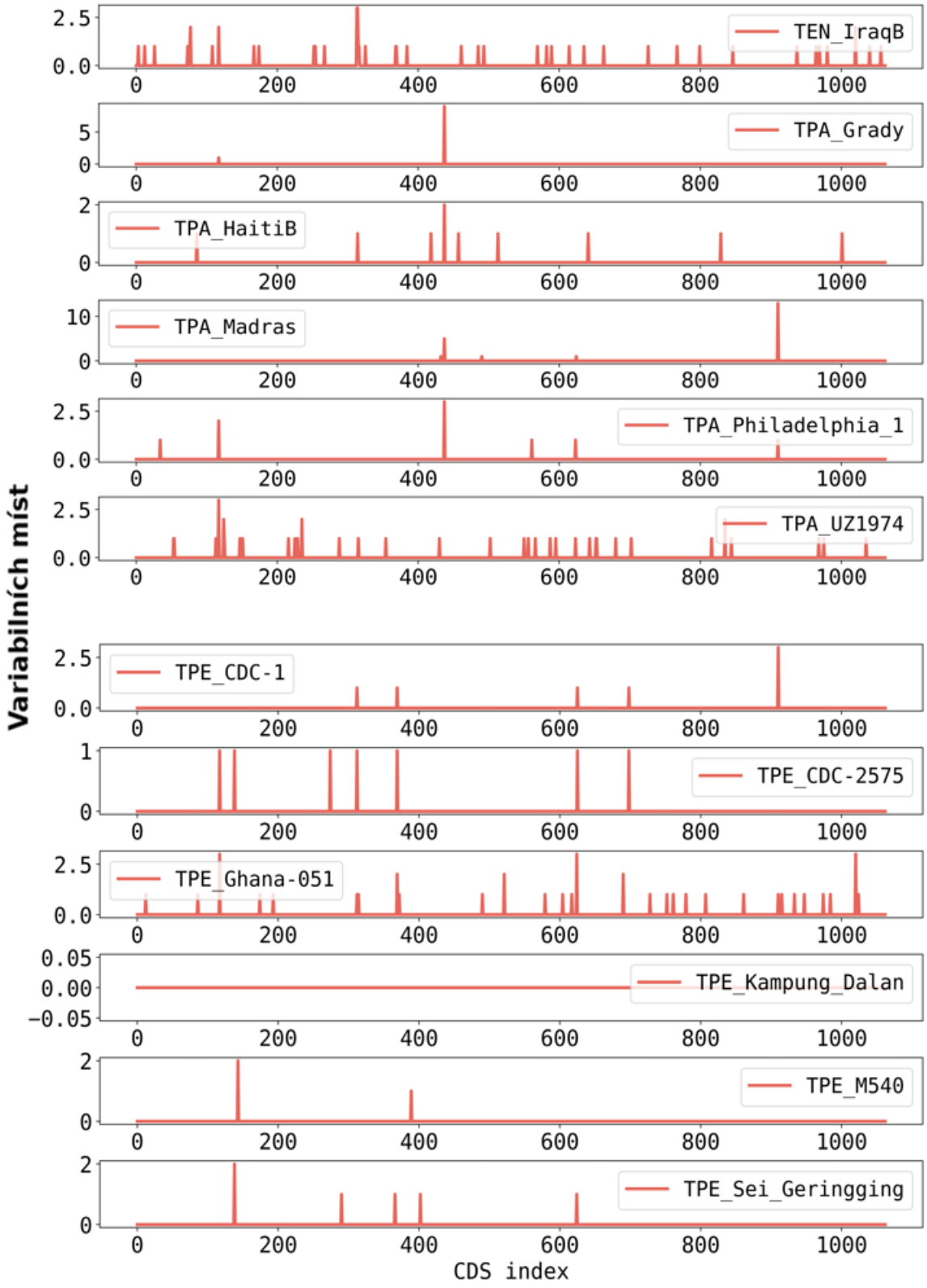
Díky zarovnání se nyní lze zaměřit na ty alely, které vykazují variabilitu na stejné pozici v zarovnání alespoň u dvou různých genomů. Výsledné počty ukazují na 14 takových míst při použití tvrdého prahování. Při použití měkkého prahu se tento počet zvýší na 29. Pravděpodobnost náhodné mutace u dvou a více genomů na shodné pozici je téměř nulová. Tyto alely jsou tedy v populaci zcela jistě variabilní. Výčet všech takovýchto nalezených pozic pomocí měkkého prahování nabízí tabulka 8.3.

Tabulka 8.3: Nalezené variabilní pozice shodné ve více genomech.

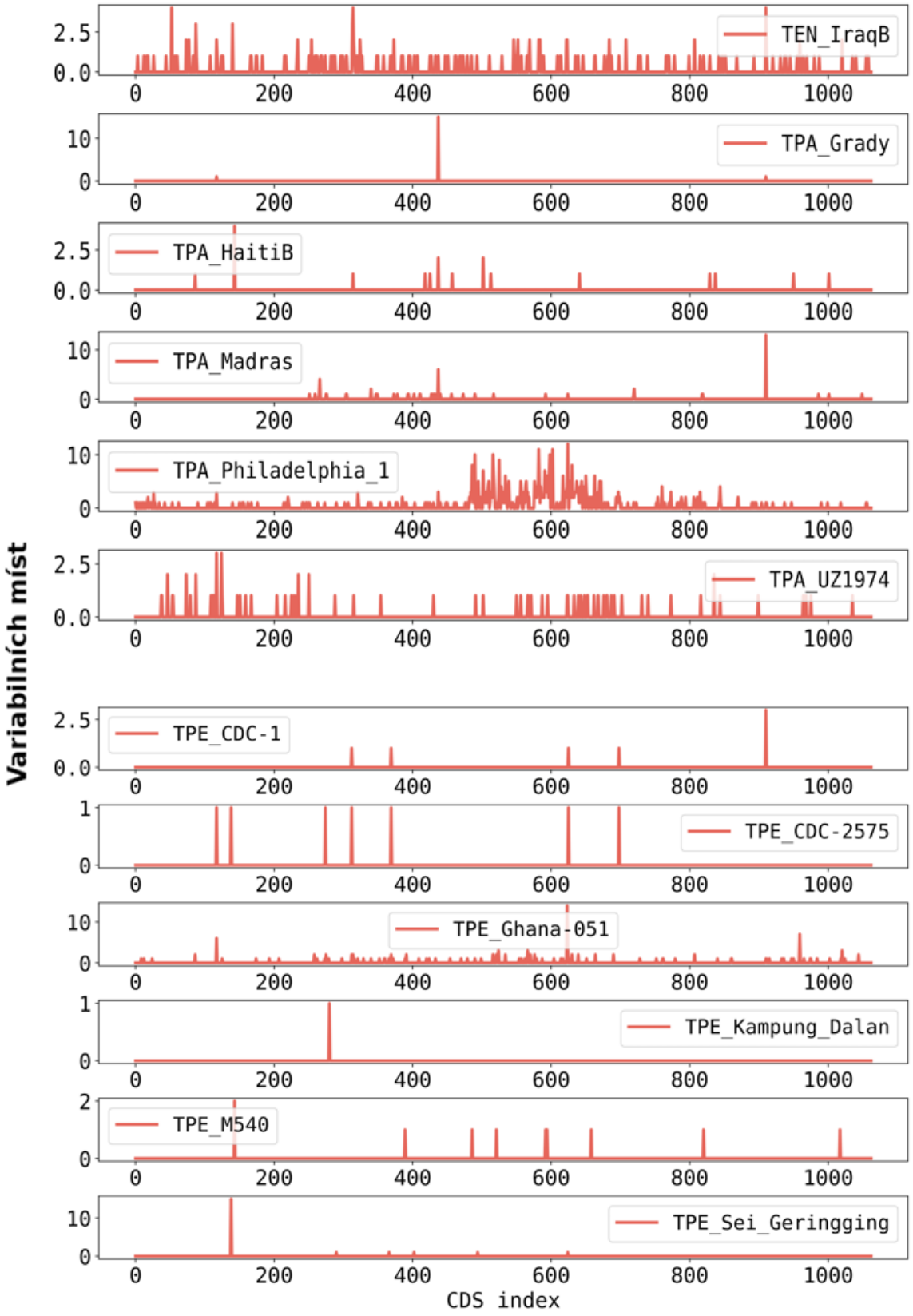
Pozice v alignmentu	Pozice v Ghana-051	gen	produkt	genomů s variabilitou	Záměna	TEN__IraqB	TPA__Grady	TPA__HaitiB	TPA__Madras	TPA__Philadelphia-1	TPA__UZ1974	TPE__CDC-1	TPE__CDC2575	TPE__Ghana-051	TPE__Sei__Geringging
96353	96331		hypothetical protein	2	C → A	2,45%					2,03%				
96360	96338		hypothetical protein	2	T → G	2,24%					2,17%				
135292	135269	tprC	Tpr protein C	2	T ↔ *C*	21,00%								2,33%	
135381	135358	tprC	Tpr protein C	3	T → C					7,11%	5,02%			13,50%	
135384	135361	tprC	Tpr protein C	4	G → A		4,49%			6,75%	4,82%			13,11%	
135954	135931	tprC	Tpr protein C	4	T → C	3,72%				3,14%	4,33%			6,62%	
156684	156504		putative outer membrane protein	2	A → G	2,25%				2,08%					
329818	329471	tprE	Tpr protein E	3	G → C							10,98%	10,98%	3,21%	
332853	332499	tprF	Tpr protein F	2	G → A	2,02%								3,08%	
333045	332691	tprF	Tpr protein F	2	A → G	6,01%								4,54%	
391053	390657	cheY	response regulator	4	G → A	11,19%						30,95%	30,95%	5,41%	
463212	462809	arp	arp	2	A → G		2,41%			3,60%					
463245	462842	arp	arp	3	C → T		9,79%	17,08%		4,91%					
464055	463052	arp	arp	2	A → G		11,20%		6,39%						
464059	463056	arp	arp	2	A → G		12,74%		6,83%						
464112	463109	arp	arp	2	G → A		1,72%		7,42%						
464115	463112	arp	arp	2	A → G		1,91%		7,47%						
464119	463116	arp	arp	2	A → G		1,68%		7,80%						
676552	675202	tprJ	Tpr protein J	2	A → G					2,36%	2,89%				
676563	675213	tprJ	Tpr protein J	3	C → T				3,83%	2,22%					5,21%
678591	677213		putative membrane protein	2	G → A							33,67%	33,67%		
700298	698920		putative membrane protein	2	G → A/T			5,94%		1,38%					
764485	763101		hypothetical protein	2	T → C							5,79%	5,79%		
978045	976199	tprK	Tpr protein K	2	C → A		1,94%			54,83%					
978549	976701	tprK	Tpr protein K	2	C → T	1,80%			34,71%						
978551	976703	tprK	Tpr protein K	2	C → T	1,54%			35,59%						
978557	976709	tprK	Tpr protein K	2	T → C	2,03%			29,33%						
978672	976818	tprK	Tpr protein K	2	T ↔ *C*				25,50%			30,00%			
978795	976929	tprK	Tpr protein K	2	T → C							46,58%		11,43%	

Variabilní místa vznikají často proto, aby ovlivnili výsledný exprimovaný protein. Z tohoto důvodu nemusí být významná pouze variabilita na dané pozici, ale také variabilita v rámci celého genu. Díky zarovnaným sekvencím je možné spočítat variabilní místa v jednotlivých CDS (Coding sequence - kódující sekvence) všech genomů. Vzhledem k tomu, že není k dispozici anotace všech zpracovávaných genomů, vycházím ze zarovnaných sekvencí. Při zarovnání předpokládám zarovnání také jednotlivých CDS na stejné souřadnice. Mohou nastat i individuální výjimky, které však pro účely této práce nyní zanedbám. Jako referenční anotaci pro celý zarovnaný soubor genomů je použita anotace genomu TPE Ghana-051 (CP020365.1) [32]. Tuto referenci volím po konzultaci s odborníky z Lékařské fakulty Masarykovy univerzity jako nejaktuálnější z anotovaných genomů a kvalitně zpracovanou. V referenci se nachází 1063 CDS oblastí.

Jednotlivé počty nalezených variabilních míst v jednotlivých CDS vynesu do grafu pro každý genom zvlášť. Jednotlivé grafy zastoupení alternativních alel v CDS regionech vidíme na obrázku 8.5 pro tvrdé prahování a na obrázku 8.6 pro prahování s měkkým prahem.



Obrázek 8.5: Zastoupení alternativních alel v jednotlivých CDS (tvrdé prahování).



Obrázek 8.6: Zastoupení alternativních alel v jednotlivých CDS (měkké prahování).

Z výše uvedených grafů pozoruji, že mezi genomy se výskyt některých variabilních alel v rámci oblastí CDS překrývá. Pro každý genom spočítám, kolik variabilních míst spadá do jednotlivých referenčních CDS. Každé CDS, které vykazuje alespoň jednu alternativní alelu označím za variabilní. Opět se podíváme kolik CDS je označeno jako variabilní alespoň ve dvou zpracovávaných genomech. Toto provedu pro oba typy prahů. Výsledek ukazuje tabulka 8.4.

Tabulka 8.4: Variabilní CDS (dle Ghana-051)

	Genom	**CDS s variabilitou (tvrdý práh)**	**CDS s variabilitou (měkký práh)**
TEN	IraqB	41	158
TPA	Grady	2	3
	HaitiB	9	14
	Madras	5	38
	Philadelphia-1	6	265
	UZ1974	34	65
TPE	CDC-1	5	5
	CDC-2575	7	7
	Ghana-051	30	89
	Kampung Dalan	0	1
	M540	2	9
	Sei Geringging	5	6
	celkem	116	471
	aspoň u dvou genomů	17	145

Z tabulky 8.4 je patrné, že při použití měkkého prahování bylo identifikováno 17 CDS, které se jako variabilní vyskytují alespoň u dvou genomů. Na tyto oblasti se podívám podrobněji, zejména pak o jaké oblasti jde a v kterých genomech se variabilní CDS nachází. Výsledek pro tvrdé prahování ukazuje tabulka 8.5.

Tabulka 8.5: Variabilní CDS (dle Ghana-051) alespoň u dvou genomů

gen	produkt	počet genomů	genom	Var. míst
pheT	phenylalanine–tRNA ligase, beta subunit	2	TEN_IraqB	1
			TPE_Ghana-051	1
	hypothetical protein	2	TPA_HaitiB	1
			TPE_Ghana-051	1
tprC	Tpr protein C	6	TEN_IraqB	2
			TPA_Grady	1
			TPA_Philadelphia-1	2
			TPA_UZ1974	3
			TPE_CDC-2575	1
			TPE_Ghana-051	3
tprD	Tpr protein D	2	TPE_CDC-2575	1
			TPE_Sei_Geringging	2
troD	troD iron (Fe2+)/zinc (Zn2+)/manganese (Mn2+) ABC	2	TEN_IraqB	1
			TPE_Ghana-051	1
tprE	Tpr protein E	3	TPE_CDC-1	1
			TPE_CDC-2575	1
			TPE_Ghana-051	1
tprF	Tpr protein F	3	TEN_IraqB	3
			TPA_HaitiB	1
			TPE_Ghana-051	1
tprG	Tpr protein G	2	TEN_IraqB	1
			TPA_UZ1974	1
cheY	response regulator	4	TEN_IraqB	1
			TPE_CDC-1	1
			TPE_CDC-2575	1
			TPE_Ghana-051	2
arp	arp	4	TPA_Grady	9
			TPA_HaitiB	2
			TPA_Madras	5
			TPA_Philadelphia-1	3
mcp2	methyl-accepting chemotaxis protein	2	TPA_Madras	1
			TPE_Ghana-051	1
tprI	Tpr protein I	2	TPA_Philadelphia-1	1
			TPA_UZ1974	1
tprJ	Tpr protein J	3	TPA_Madras	1
			TPE_Ghana-051	3
			TPE_Sei_Geringging	1
	putative membrane protein	2	TPE_CDC-1	1
			TPE_CDC-2575	1
	hypothetical protein	2	TPE_CDC-1	1
			TPE_CDC-2575	1
tprK	Tpr protein K	4	TPA_Madras	13
			TPA_Philadelphia-1	1
			TPE_CDC-1	3
			TPE_CDC-2575	1
ftsK	S-DNA-T family septal DNA translocator	2	TEN_IraqB	2
			TPE_Ghana-051	3

9 DISKUSE VÝSLEDKŮ

V rámci práce bylo zpracováno 12 genomů, které se svou kvalitou resekvenace lišily. Zejména průměrná hloubka pokrytí jednotlivých genomů kolísá od 62,92 čtení na pozici až po 1 200 čtení na pozici. Kvalita resekvenace ovlivňuje zejména výšku prahů, používaných k identifikaci variabilních míst. Pro každý genom bylo identifikováno několik variabilních pozic, s tím že pracuji se dvěmi prahovacími hodnotami. Při použití tvrdého prahu se počet nalezených variabilních míst v jednotlivých genomech pohybuje v řádu desítek, při použití měkkého prahování se však dostáváme do řádově vyšších hodnot, což je způsobeno nejspíš i zahrnutím některých nepříliš jasných chybových pozic.

Vedlejším produktem analýzy variability se stala identifikace rozdílných pozic mezi genomy a jejich příbuznými, které byly využity pro jejich namapování. Tyto identifikované pozice zpracuji a následně ze souboru odstraním, aby negativně neovlivňovaly následující analýzy.

Pomocí shlukové analýzy UPGMA srovnám zařazení genomů na základě jejich variability, se zařazením vůči fylogenetickému stromu vytvořenému stejnou metodou. Ve vytvořených stromech pozoruji tendenci shlukování stejných poddruhů. Zařazení je také ovlivněno kvalitou resekvenace, a v případě stromů na základě relativní četnosti zastoupení alternativních alel pak také výškou individuálních prahů. Kromě své vlastní individuální variability, usuzuji také nejspíše na variabilitu ovlivněnou typem poddruhu dané bakterie, neboť pozoruji shlukování stejných poddruhů.

Z výskytu variabilních míst nacházejících se na stejných místech v alespoň dvou genomech předpokládám jasný selekční tlak, nikoliv náhodnou mutaci těchto míst. Na těchto pozicích dochází zároveň k záměně stejných nukleotidů. Variabilní místa byla díky zvolenému způsobu sekvenace identifikována také v genech z rodiny *tpr*. Ve výčtu nechybí ani gen *tprK*, u kterého se ví o jeho extrémní hypervariabilitě. Tuto informaci podporují i značně vysoké relativní četnosti výskytu alternativních alel u zpracovávaných genomů.

Při analýze zastoupení variability v rámci jednotlivých kódujících úseků, he patrné poměrně rovnoměrné rozložení variabilních úseků v celé délce genomů. Srovnáním jednotlivých průběhů vidíme i opakující se stejná variabilní CDS u několika genomů. Z průběhů jasně vyčnívá genom *TPA Philadelphia-1* při použití měkkého prahování variabilních úseků. V tomto genomu pozoruji zvýšený počet identifikované variability v oblasti sekvenovaného poolu 3. To je nejspíš způsobeno sekvenováním v jiném sekvenačním běhu a v jiném čase. Tento úsek tak může vykazovat odlišné kvalitativní parametry od zbytku celého genomu složeného z ostatních poolů.

Variabilní geny, které se nacházejí alespoň u dvou genomů, srovnám napřed s tabulkou 8.3. Vidím, že tabulka 8.5 se seznam variabilních genů rozšiřuje o nově

nalezené. Opět je patrné silné zastoupení genů z rodiny *tpr*. Gen *tprK* je opět identifikován jako hypervariabilní. Kromě něj pozoruji i značnou variabilitu u genu *arp*, to je způsobeno nejspíše jeho sekvencí, která obsahuje repetitivní úseky.

Nalezené variabilní geny a kódující úseky porovnávám s přehledem identifikovaných variabilních úseků v treponemálních kmenech v publikovaných článcích [33]. Zde je patrná shoda v několika nalezených identifikovaných genech, pomocí představeného postupu bylo však identifikováno také několik genů nových.

10 PROGRAMOVÁ PODPORA

Pro účely této práce a zpracování veškerých dat vzniklo několik souborů. Jde především o posloupnost příkazů volajících programy použité při předzpracování dat, sestavení genomů a vytvoření **.vcf* souborů. Tato část je zpracována v Bashi. Pro správnou funkci je třeba mít naistalovaný nástroj pro mapování čtení *BWA* (verze 0.7.15) a balíček nástrojů *SamTools* (verze 1.4.1).

Následující zpracování je prováděno v jazyce R (verze 3.4.4). Kompletní vytvořený balíček je k dispozici na GitHubu (*https://github.com/VojtechBarton/TrepVar*). Analýza prováděná v jazyce R vychází z předzpracovaného souboru dat a pracuje s jeho **.vcf* verzí. Pro správnou funkci je třeba mít nainstalované balíčky *Biostrings* (verze 2.46.0), *tidyr* (verze 0.8.0) a *stringi* (verze 1.1.6).

Balíček obsahuje několik hlavních funkcí:

- **divRefToPools()**
  Funkce k rozdělení sekvence celého genomu na jednotlivé pooly, dále použitelné při tvorbě sestavení osekvenovaného poolu.
- **findHPolyRegions()**
  Funkce k nalezení homopolymerních úseků v sekvenci.
- **readVCF()**
  Funkce pro import dat z *vcf* formátu.
- **alignToRef()**
  Funkce pro sloučení přepočítání souřadnic poolů, a sloučení jednotlivých souborů poolů do jednoho souboru celého genomu.
- **exportIndels()**
  Funkce pro oddělení delších indelů ze souboru genomu.
- **joinRows()**
  Funkce pro sjednocení informace z překrývajících se úseků z jednotlivých poolů.
- **findFeatures()**
  Funkce pro výpočet příznaků.
- **findThresh()**
  Funkce pro určení odhadu sekvenační chyby.
- **findVariableSpots()**
  Funkce pro určení variabilních míst v genomu.

Součástí je i několik dalších souborů:

- **main.R**
  Skript obsahující navrhovaný postup analýzy.

- **praparation.sh**
  Bash soubor pro dávkové zpracování sekvenačních dat a jejich převod do formátu **.vcf*.
- **pools_coord.fa**
  Fasta soubor se sekvencemi ohraničujícími jednotlivé pooly.

11 ZÁVĚR

V práci je představen přehled nejdůležitějších sekvenačních metod a jejich rozdělení z pohledu generací. Věnujeme se představení principů fungování jednotlivých technologií. Důraz je kladen především na technologii Illumina, která byla použita při akvizici našich zpracovávaných dat. Krátce se věnujeme představení sekvenačních technologií a úvodu do odborné terminologie související se sekvenováním genomů.

Dále představuji bakterii *Treponema pallidum*, kterážto je využita jako analyzovaný organismus. Krátce se věnuji rozdílu v jednotlivých poddruzích této bakterie.

Další část práce je již věnována analýze variability v sekvenačních datech treponemálních kmenů. Představuji použitá sekvenační data jednotlivých do analýzy zařazených genomů. Veškeré genomy byly sekvenovány po částech, tak aby mohla být správně identifikována a interpretována i variabilita nacházející se v genech z rodiny *tpr*.

Zabývám se jednotlivými kroky metody předzpracování sekvenačních dat, s cílem zvýšit jejich kvalitu a tedy i informační výtěžnost zpracovávaného souboru. Popisuji jednotlivé kroky úprav a filtrací pomocí externích, volně dostupných programů a nástrojů.

V následující části představuji navržený a aplikovaný postup identifikace variabilních míst. Předpokladem pro úspěšnou analýzu je právě vhodně předzpracovaný a vysoce kvalitní soubor namapovaných čtení z předchozí části. Zavádím systém podmínek s cílem, co nejpřesnější identifikace variability a zamezení falešně pozitivně identifikované variability způsobené technologickou, či náhodnou chybou. Pro každý genom zvlášť je určena mez variability, tedy relativní četnost výskytu alternativní alely na dané pozici, kterou lze považovat za heterogenní.

Tento navržený postup je aplikován na soubor dvanácti treponemálních genomů, u kterých jsou dostupná sekvenační data. Nalezené variabilní pozice u každého genomu dále hodnotím a srovnávám napříč jednotlivými genomy. Při hodnocení koincidence výskytu alternativních pozic využívám vytvořený soubor celogenomových zarovnání. Dále přináším přehled identifikovaných variabilních míst, která se vyskytují na stejné pozici u více genomů. Variabilitu poté sleduji i z pohledu příslušnosti do jednotlivých kódujících úseků.

V rámci práce byly také identifikovány rozdíly mezi příbuznými genomy, pro které ještě nebyla stanovena referenční sekvence.

Kompletní postup předzpracování je připraven v BASH souboru. Následné analýzy provádím pomocí funkcí z sestaveného balíčku *TrepVar* vytvořeném pro jazyk R. Tyto soubory dávám volně k dispozici na platformě GitHub.

LITERATURA

[1] HEATHER, J. M. a B. CHAIN. The sequence of sequencers: The history of sequencing DNA. *Genomics* [online]. 2016, (107). DOI: 10.1016/j.ygeno.2015.11.003. ISBN 10.1016/j.ygeno.2015.11.003. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0888754315300410

[2] The Nobel Prize in Chemistry 1980. *Nobelprize.org* [online]. Nobel Media AB 2014, 2018. Dostupné z: https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/

[3] MAXAM, A. M. a W. GILBERT. *A new method for sequencing DNA* [online]. DOI: 10.1073/pnas.74.2.560. ISBN 10.1073/pnas.74.2.560. Dostupné z: http://www.pnas.org/cgi/doi/10.1073/pnas.74.2.560

[4] Wikimedia Commons contributors, File:Maxam gilbert sequencing.png, *Wikimedia Commons, the free media repository*. Dostupné z: http://commons.wikimedia.org/w/index.php?title=File:Maxam_gilbert_sequencing.png

[5] SANGER, F., S. NICKLEN a A. R. COULSON. *DNA sequencing with chain-terminating inhibitors* [online]. DOI: 10.1073/pnas.74.12.5463. ISBN 10.1073/pnas.74.12.5463. Dostupné z: http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463

[6] SCHADT, E. E., S. TURNER a A. KASARSKIS. *A window into third-generation sequencing* [online]. DOI: 10.1093/hmg/ddq416. ISBN 10.1093/hmg/ddq416. Dostupné z: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddq416

[7] Wikimedia Commons contributors, File:Radioactive Fluorescent Seq.jpg, Wikimedia Commons, the free media repository, http://commons.wikimedia.org/w/index.php?title=File:Radioactive_Fluorescent_Seq.jpg

[8] PETTERSSON, E., J. LUNDEBERG a A. AHMADIAN. *Generations of sequencing technologies* [online]. DOI: 10.1016/j.ygeno.2008.10.003. ISBN 10.1016/j.ygeno.2008.10.003. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0888754308002498

[9] MARGULIES, M., M. EGHOLM, W. E. ALTMAN, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* [online]. **437**, 376-380. DOI: 10.1038/nature03959.

[10] MARDIS, E. R. *The impact of next-generation sequencing technology on genetics* [online]. DOI: 10.1016/j.tig.2007.12.007. ISBN 10.1016/j.tig.2007.12.007. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0168952508000231

[11] *Illumina Sequencing Technology: Highest data accuracy, simple workflow, and a broad range of applications.* [online]. Illumina, 2010, 2010. Dostupné z: https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

[12] JOHNSEN, J. M., D. A. NICKERSON a A. P. REINER. *Massively parallel sequencing: the new frontier of hematologic genomics* [online]. DOI: 10.1182/blood-2013-07-460287. ISBN 10.1182/blood-2013-07-460287. Dostupné z: http://www.bloodjournal.org/cgi/doi/10.1182/blood-2013-07-460287

[13] NIEDRINGHAUS, T. P., D. MILANOVA, M. B. KERBY, M. P. SNYDER a A. E. BARRON. *Landscape of Next-Generation Sequencing Technologies* [online]. DOI: 10.1021/ac2010857. ISBN 10.1021/ac2010857. Dostupné z: http://pubs.acs.org/doi/abs/10.1021/ac2010857

[14] BREU, H. *A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction: Methods for Annotating 2 Base Color Encoded Reads in the SOLiD System* [online]. USA: Applied Biosystems, 2010. Dostupné z: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_058265.pdf

[15] PENNISI, E. *Semiconductors Inspire New Sequencing Technologies* [online]. DOI: 10.1126/science.327.5970.1190. ISBN 10.1126/science.327.5970.1190. Dostupné z: http://www.sciencemag.org/cgi/doi/10.1126/science.327.5970.1190

[16] VOELKERDING, K. V., S. A. DAMES a J. D. DURTSCHI. *Next-Generation Sequencing: From Basic Research to Diagnostics* [online]. DOI: 10.1373/clinchem.2008.112789. ISBN 10.1373/clinchem.2008.112789. Dostupné z: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2008.112789

[17] RHOADS, A. a K. F. AU. *PacBio Sequencing and Its Applications* [online]. DOI: 10.1016/j.gpb.2015.08.002. ISBN 10.1016/j.gpb.2015.08.002. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S1672022915001345

[18] STEINBOCK, L. J. a A. RADENOVIC. *The emergence of nanopores in next-generation sequencing* [online]. DOI: 10.1088/0957-4484/26/7/074003. ISBN 10.1088/0957-4484/26/7/074003. Dostupné z: http://stacks.iop.org/0957-4484/26/i=7/a=074003?key=crossref.e083fc814dc88fdb8d767b37bc7dd915

[19] MATĚJKOVÁ, P., M. STROUHAL, D. ŠMAJS, et al. *Complete genome sequence of Treponema pallidum ssp. pallidum strain SS14 determined with oligonucleotide arrays* [online]. DOI: 10.1186/1471-2180-8-76. ISBN 10.1186/1471-2180-8-76. Dostupné z: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-8-76

[20] BUFFET, M., P. A. GRANGE, P. GERHARDT, A. CARLOTTI, V. CALVEZ, A. BIANCHI a N. DUPIN. *Diagnosing Treponema pallidum in Secondary Syphilis by PCR and Immunohistochemistry* [online]. DOI: 10.1038/sj.jid.5700888. ISBN 10.1038/sj.jid.5700888. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0022202X15331444

[21] Fieldsteel, A. H., Cox, D. L., Moeckli, R. A. (1981). Cultivation of Virulent Treponema pallidum in Tissue Cultur*Infection and Immunity*, 32(2), 908–915.

[22] FRASER, C. M. Complete Genome Sequence of Treponema pallidum, the Syphilis Spirochete. *Science* [online]. 1998, **281**(5375), 375-388. DOI: 10.1126/science.281.5375.375. Dostupné z: http://www.sciencemag.org/cgi/doi/10.1126/science.281.5375.375

[23] ČEJKOVÁ, D., M. STROUHAL, S. J. NORRIS, G. M. WEINSTOCK, D. ŠMAJS a M. PICARDEAU. *A Retrospective Study on Genetic Heterogeneity within Treponema Strains: Subpopulations Are Genetically Distinct in a Limited Number of Positions* [online]. DOI: 10.1371/journal.pntd.0004110. ISBN 10.1371/journal.pntd.0004110. Dostupné z: http://dx.plos.org/10.1371/journal.pntd.0004110

[24] LI, H. a R. DURBIN. *Fast and accurate long-read alignment with Burrows—Wheeler transform* [online]. DOI: 10.1093/bioinformatics/btp698. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp698

[25] Andrews. A. FastQC A Quality Control tool for High Throughput Sequence Data. 2018. Dostupné z: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[26] EWELS, P., M. MAGNUSSON, S. LUNDIN a M. KÄLLER. MultiQC: summarize analysis results for multiple tools and samples in a single report [online]. DOI: 10.1093/bioinformatics/btw354. ISBN 10.1093/bioinformatics/btw354. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354

[27] LI, H., B. HANDSAKER, A. WYSOKER, et al. *The Sequence Alignment/Map format and SAMtools* [online]. DOI: 10.1093/bioinformatics/btp352. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352

[28] VICTORIA, X., N. BLADES, J. DING, R. SULTANA a G. PARMIGIANI. Estimation of sequencing error rates in short reads. *BMC Bioinformatics* [online]. 2012, **13**(185). DOI: 10.1186/1471-2105-13-185. ISBN 10.1186/1471-2105-13-185. Dostupné z: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-185

[29] ZHU, X., J. WANG, B. PENG a S. SHETE. Empirical estimation of sequencing error rates using smoothing splines. *BMC Bioinformatics* [online]. 2016, **17**(177). DOI: 10.1186/s12859-016-1052-3. ISBN 10.1186/s12859-016-1052-3. Dostupné z: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1052-3

[30] EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* [online]. 2004, **32**(5), 1792–1797. DOI: 10.1093/nar/gkh340. ISBN 10.1093/nar/gkh340. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh340

[31] KUMAR, Sudhir, Glen STECHER a Koichiro TAMURA. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* [online]. 2016, 2000, **33**(7), 1870-1874. DOI: 10.1093/molbev/msw054. ISBN 10.1093/molbev/msw054. Dostupné z: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw054

[32] STROUHAL, M., L. MIKALOVÁ, P. HAVLÍČKOVÁ, et al. Complete genome sequences of two strains of Treponema pallidum subsp. pertenue from Ghana, Africa: Identical genome sequences in samples isolated more than 7 years apart. *PLoS Negl Trop Dis.* [online]. 2017, **11**(9). DOI: 10.1371/journal.pntd.0005894. ISBN 10.1371/journal.pntd.0005894. Dostupné z: http://dx.plos.org/10.1371/journal.pntd.0005894

[33] ŠMAJS, David, Michal STROUHAL a Sascha KNAUF. Genetics of human and animal uncultivable treponemal pathogens. *Infection, Genetics and Evolution* [online]. 2018, (61), 92-107. DOI: 10.1016/J.MEEGID.2018.03.015. ISBN 10.1016/j.meegid.2018.03.015. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S1567134818301151

SEZNAM PŘÍLOH

A ROZDÍLNÉ POZICE V GENOMECH

A.1 IraqB

Genom TEN_IraqB byl namapován vůči referenčnímu genomu TEN_BosniaA. Na základě analýzy variability byly následující pozice identifikovány jako odlišné mezi oběma genomy.

Tabulka A.1: Rozdílné pozice mezi příbuznými genomy IraqB a BosniaA.

Pozice	**Báze v BosniaA (referenční)**	**Báze v IraqB (alternativní)**	**Hloubka pokrytí**	**Počet čtení alternativní alely**	**Poměr čtení forward/ reverse**
18409	A	G	698	100.00 %	0.442
18413	C	T	672	100.00 %	0.439
80362	C	T	632	100.00 %	0.549
135229	T	G	893	74.36 %	0.459
135230	C	G	887	74.75 %	0.441
135247	T	C	876	79.00 %	0.531
136529	T	C	1028	75.00 %	0.438
188041	T	G	740	100.00 %	0.709
203276	A	G	897	99.89 %	0.851
230531	A	C	2034	99.80 %	0.431
320545	G	A	927	100.00 %	0.739
333273	T	C	997	90.17 %	0.614
450054	A	G	737	99.86 %	0.704
468556	T	C	1357	99.93 %	0.719
497707	A	G	843	100.00 %	0.673
512068	C	G	611	99.51 %	0.652
521391	C	T	202	100.00 %	0.603
534767	A	G	1031	99.90 %	0.542
566815	G	T	906	100.00 %	0.483
592253	T	C	895	99.55 %	0.580
643060	A	G	937	100.00 %	0.524
690735	A	G	1030	99.81 %	0.406
702524	C	G	881	100.00 %	0.440
882335	G	A	1122	99.82 %	0.493
942850	A	G	520	100.00 %	0.516
993154	G	T	437	100.00 %	0.428
994941	C	T	797	100.00 %	0.518
1085949	C	T	463	99.57 %	0.547
1130225	G	A	277	100.00 %	0.428

A.2 Grady

Genom TPA_Grady byl namapován vůči referenčnímu genomu TPA_SS14. Na základě analýzy variability byly následující pozice identifikovány jako odlišné mezi oběma genomy.

Tabulka A.2: Rozdílné pozice mezi příbuznými genomy Grady a SS14.

Pozice	**Báze v SS14 (referenční)**	**Báze v Grady (alternativní)**	**Hloubka pokrytí**	**Počet čtení alternativní alely**	**Poměr čtení forward/ reverse**
94901	A	C	70	100.00 %	1.258
135108	G	C	533	100.00 %	0.563
226317	G	A	124	100.00 %	0.824
235246	G	A	60	100.00 %	1.400
283691	G	A	867	100.00 %	0.889
364888	T	C	204	100.00 %	1.429
495727	G	A	675	100.00 %	1.003
522907	A	G	218	100.00 %	1.295
674219	C	T	108	100.00 %	0.800
674227	A	C	92	100.00 %	0.769
674233	T	C	84	100.00 %	0.556
772846	C	T	623	100.00 %	0.865
773095	T	C	513	100.00 %	0.781
800482	A	G	508	100.00 %	0.841
861444	T	G	498	100.00 %	0.717

A.3 Madras

Genom TPA_Madras byl namapován vůči referenčnímu genomu TPA_Nichols. Na základě analýzy variability byly následující pozice identifikovány jako odlišné mezi oběma genomy.

Tabulka A.3: Rozdílné pozice mezi příbuznými genomy Madras a Nichols.

Pozice	**Báze v Nichols (referenční)**	**Báze v Madras (alternativní)**	**Hloubka pokrytí**	**Počet čtení alternativní alely**	**Poměr čtení forward/ reverse**
459976	T	C	518	99.81%	0.494
523540	A	G	299	71.57%	0.659
702170	C	T	299	100.00%	0.689
703137	T	A	310	100.00%	0.574
976383	T	G	94	54.26%	0.457
976411	C	T	118	100.00%	0.405
976413	G	A	128	85.16%	0.493
976597	C	G	122	98.36%	0.429
976601	C	G	150	75.33%	0.413

A.4 UZ1974

Genom TPA_UZ1974 byl namapován vůči referenčnímu genomu TPA_Philadelphia-1. Na základě analýzy variability byly následující pozice identifikovány jako odlišné mezi oběma genomy.

Tabulka A.4: Rozdílné pozice mezi příbuznými genomy UZ1974 a Philadelphia-1.

Pozice	**Báze v Philadelphia-1 (referenční)**	**Báze v UZ1974 (alternativní)**	**Hloubka pokrytí**	**Počet čtení alternativní alely**	**Poměr čtení forward/ reverse**
174177	C	T	444	100.00%	1.209
235246	A	G	272	100.00%	1.109
283691	A	G	138	100.00%	1.339
342703	G	A	97	100.00%	2.233
396035	C	T	174	100.00%	0.977
556154	C	T	562	100.00%	1.498
593294	G	A	341	100.00%	1.368
593298	G	A	342	100.00%	1.375
593912	A	G	449	100.00%	1.339
760092	T	C	369	100.00%	1.097