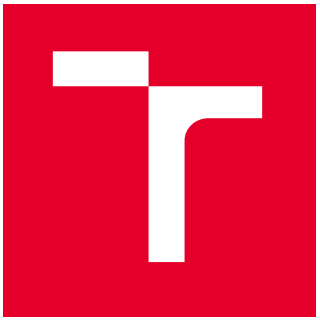


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

VÝVOJ POSTUPU PRO MLST TYPOVÁNÍ TREPONEMA PALLIDUM SUBSP. PALLIDUM

DEVELOPMENT OF THE WORKFLOW FOR MLST TYPING OF TREPONEMA PALLIDUM SUBSP. PALLIDUM

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Eliška Taliánová**
AUTHOR

VEDOUCÍ PRÁCE — **Mgr. Bc. Darina Čejková, Ph.D.**
SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní program **Biomedicínské inženýrství a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Bc. Eliška Taliánová ***ID:*** 211491

Ročník: 2 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Vývoj postupu pro MLST typování Treponema pallidum subsp. pallidum

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerší na téma typování Treponema pallidum. 2) Stáhněte sekvenační data z databáze SRA (short read archive). 3) Ověřte kvalitu kontroly dat. 4} Mapujte data na oblasti daných MLST lokusů – TP0136, TP0548, TP0705. 5) Proveďte post-alignment kontrolu a vyhodnoťte výsledky. 6) Zařaďte do postupu de novo assemblování pro identifikaci nových alel. 7) Diskutujte výsledky.

DOPORUČENÁ LITERATURA:

[1] GRILLOVÁ L., BAWA T., MIKALOVÁ L., et al. Molecular characterization of Treponema pallidum subsp. pallidum in Switzerland and France with a new multilocus sequence typing scheme. PLoS ONE. 2018,13(7):e0200773.

[2] CHEN W., ŠMAJS D., HU Y., et al. Analysis of Treponema pallidum strains from China using improved methods for whole-genome sequencing from primary syphilis chancres. Journal of Infection Diseases. 2021,223(5):848-853.

[3] 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(1):1-11.

[4] BEALE M.A., MARKS M., COLE M.J., et al. Global phylogeny of Treponema pallidum lineages reveals recent expansion and spread of contemporary syphilis. Nat Microbiol. 2021,6(12):1549-1560.

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 22.5.2024

Vedoucí práce: Mgr. Bc. Darina Čejková, Ph.D.

prof. Ing. Valentine Provazník, Ph.D.
předseda rady studijního programu

ABSTRAKT

Práce se zabývá problematikou MLST typování *Treponema pallidum* subsp. *pallidum*. Tato spirochetní bakterie způsobuje nemoc syfilis. MLST typování představuje významný zdroj informací, jak o sekvenčních změnách genů, tak o jejich dopadech na patogenitu této bakterie. Snaží se navrhnout postup pro určení alelického profilu s využitím SRA dat. Využívá BWA mapování na referenční genom.

KLÍČOVÁ SLOVA

MLST typování, *Treponema pallidum* subsp. *pallidum*, syfilis, SRA databáze, BWA, TP0136, TP0548, TP0705, kmen Nichols, kmen SS14

ABSTRACT

This thesis is focused on MLST typing *Treponema pallidum* subsp. *pallidum*. This spirochete bacterium causes the disease syphilis. MLST typing represents an important source of information, both on the sequence changes of genes and on their impact on the pathogenicity of this bacterium. It tries to design a metod for obtaining the allelic profile using SRA data. It uses BWA mapping to the reference genome.

KEYWORDS

MLST typing, *Treponema pallidum* subsp. *pallidum*, syphilis, SRA database, BWA, TP0136, TP0548, TP0705, Nichols strain, SS14 strain

TALIÁNOVÁ, Eliška. *Vývoj postupu pro in silico MLST typování Treponema pallidum subsp. pallidum*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024, 71 s. Diplomová práce. Vedoucí práce: Mgr.Bc. Darina Čejková, Ph.D.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Bc. Eliška Taliánová
VUT ID autora:	211491
Typ práce:	Diplomová práce
Akademický rok:	2023/24
Téma závěrečné práce:	Vývoj postupu pro in silico MLST typování *Treponema pallidum subsp. pallidum*

Prohlašuji, že svou závěrečnou práci jsem vypracovala samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí mé diplomové práce paní Mgr.Bc. Darině Čejkové, Ph.D. za konzultace, podnětné návrhy a připomínky k práci a v především za trpělivost, kterou se mnou měla.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Tato práce se zabývá problematikou typování *Treponema pallidum* subsp.*pallidum*.

Treponema pallidum subsp.*pallidum* spirální bakterie ze skupiny spirochét je původcem syfilis. Toto pohlavně přenášené onemocnění představuje poměrně závažný celosvětový zdravotní problém. World Health Organization (WHO) udává, že incidence nových případů se vyšplhá až na 11 miliónů ročně. Právě proto je lepší porozumění epidemiologii tohoto onemocnění klíčové.

Molekulární typování přináší řadu informací o kmenech těchto bakterií a o tom jakým způsobem se šíří v populaci. Díky molekulárnímu typování lze rozlišit jednotlivé poddruhy a do jisté míry předpovídat jejich chování na základě podobnosti s ostatními už známými příslušníky daného druhu.

Systémů pro typování této bakterie je hned několik a neustále se rozvíjejí. MLST typování představuje dosud poslední navržený přístup. K typizaci využívá tři geny - TP0136, TP0548 a TP0705.

Tato práce předkládá čtenáři přehledný výčet nemocí způsobených bakteriemi ze skupiny *Treponema pallidum*. Seznámí ho s morfologií organismu a předá základní informace k této problematice.

Dále nabízí vhled do problematiky DNA sekvenace organismů. Stručně popisuje historicky i v současnosti používané metody. Jejich základní metodiku a problémy. Zároveň zmiňuje metody používané k amplifikaci sekvenované DNA. Práce se zabývá problematikou MLST typování organizmů jako technikou, která ještě není zcela běžně známá široké veřejnosti, ale poskytuje cenné informace. Je zde obecně vysvětlena tato metoda a její přínosy v případě použití pro *Treponema pallidum* subsp.*pallidum*.

Dále se práce věnuje práci s raw sekvenačními daty. Databázím, které byly v rámci práce několikrát využity z mnoha důvodů, programům, které slouží ke zpracování takto náročných dat. Zmiňuje také problematiku hodnocení sekvenačních dat a programů, které k tomuto účelu slouží.

Práce se snaží navrhnout bioinformatický postup pro zjištění typu sekvence na základě alelického profilu surových sekvenačních dat. Jako podklad slouží data stažená z SRA databáze. Navržený postup poté aplikuje na několik vzorků z této databáze, hodnotí jejich kvalitu, jak co se týče přečtení jednotlivých bází, tak kvalitu mapování těchto čtení na referenční sekvenci pomocí BWA. Snaží se stanovit alelický profil dat a diskutuje výsledky.

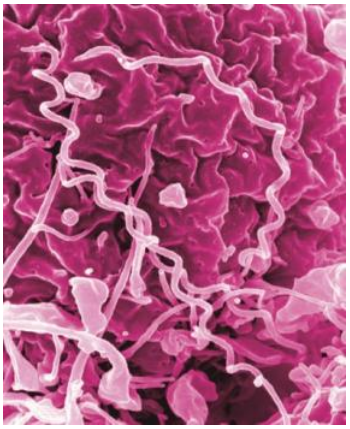
1 Treponema pallidum

Treponema pallidum je bakterie patřící mezi spirochety, která je původcem syfilis. Je náročná na zevní vlivy, vyznačuje se úzkým rozsahem pH (7,2 - 7,4) a teploty (30 - 37°C). Je inaktivována jakýmkoli výkyvem teplot mimo určený rozsah, vysycháním a většinou dezinfekčních prostředků, je mikroaerofilní, což znamená že pro svůj růst vyžaduje jen nízké koncentrace kyslíku, vyšší koncentrace vedou taktéž k inaktivaci. [1],[2],[3]

Klasifikace patogenních treponem bylo v minulosti založeno jen na rozlišení příznaků příslušných nemocí, které způsobují. *Treponema pallidum* subsp. *pallidum* způsobuje pohlavní syfilis; *Treponema pallidum* subsp. *pertenue* způsobuje frambézii; *Treponema pallidum* subsp. *pendemicum* způsobuje endemickou syfilis a *Treponema pallidum* subsp. *careteum* způsobuje nemoc pinta. V dnešní době s rozlišením pomáhá také genetika. [1]

1.1 Struktura

Jedná se o spirálovitě stočený organismus 6 - 15 μm dlouhý a 0,1 - 0,2 μm široký. Má vnější membránu s lemujícími periplazmickými bičíky (vmezeřeny mezi vnitřní a vnější membránu), které zajišťují pohyb.



Obr. 1.1: *Treponema pallidum* subsp. *pallidum* [4]

Ze 70 % je tvořena bílkovinami, 20 % lipidy a 5 % sacharidy. Ačkoliv má dvě membrány (vnější a cytoplazmatickou), svou strukturou se značně liší od gramnegativních bakterií. Její vnější membrána se vyznačuje extrémně nízkou hodnotou transmembránových proteinů. Zatímco cytoplazmatická obsahuje většinu integrálních membránových proteinů a je bohatá na lipoproteiny. Bakterie má typicky tři

bičíky, které ji obtáčejí. Důležitou složku představuje vrstva peptidoglykanů, která je důvodem pro náchylnost těchto bakterií na penicilin. Vrstva peptidoglykanů se nachází nad cytoplasmatickou membránou, nad ní jsou periplasmatické bičíky a vše uzavírá vnější membrána.[5]

Rozmnožuje se binárním dělením. Kultivace běžnými metodami *in vitro* je velmi náročná, k množení se obvykle používá pasážování v králičích varlatech. Některé zástupce lze *in vitro* kultivovat s využitím společné inkubace s epitelovými buňkami. Musí jim však být pravidelně měněno růstové médium. Životaschopné organismy mohou být v komplexních médiích udržovány až 21 dní. Ve tkáních lze vizualizovat metodami impregnace stříbra. Živé bakterie jsou příliš úzké pro konvenční světelnou mikroskopii, proto se pro jejich pozorování využívá temné pole. [1], [6]

1.2 Syfilis

Syfilis je léčitelná choroba, kterou způsobuje bakterie *Treponema pallidum* subsp. *pallidum*. Syfilis se přenáší pohlavním stykem, transfuzí krve (použitím nakažených jehel) a také z matky na plod. Přenos z matky na dítě je v případě neléčené syfilis téměř vždy letální pro plod. [7]

Infekce se do těla dostává kůží a sliznicemi, nejčastěji oblastí genitální, rektální a ústní sliznicí. Inkubační doba se pohybuje mezi 9 až 90 dny. Velká část lidí se syfilis netrpí žádnými příznaky, nebo má jen velmi slabé projevy. Z hlediska infekčnosti se onemocnění nyní dělí na dvě stádia - časné (infekční) a pozdní. [8]

- Primární syfilis:
  Po třech týdnech od infekce se začnou projevovat první příznaky. V místě vstupu infekce vzniká vřed. Většinou je jen jeden, není bolestivý a dost často se nachází na místech, kde si ho pacient nevšimne. Do šesti týdnů se obvykle zahojí. V pátém týdnu od infekce dochází k zduření místních uzlin.
- Sekundární syfilis:
  Začíná obvykle do 10 týdnů od začátku infekce. Bakterie se množí v krvi, objevují se kožní a slizniční projevy (exantém a enantém), postižení jiných orgánů a únava. Všechny slizniční projevy jsou velmi nebezpečným zdrojem infekce. Po určité době odezní a nastává období latence. Klinicky se nemoc neprojevuje a dá se určit pouze sérologicky. Období příznaků a latence se mohou střídat. Po dvou letech se nemoc většinou ustálí v latentním stádiu (pozdní latence), které může trvat i 20 a více let.
- Terciální syfilis:
  Orgánové postižení většinou jednoho orgánu, které nastupuje po 5 až 30 letech od nákazy. Charakteristickým projevem je specifický granulom - gumma. Jde

- o ostře ohraničené tuhé hrboly většinou růžové barvy. Jinak nemoc postihuje nervovou soustavu, kardiovaskulární systém, kosti a oči.

Přímý důkaz přítomnosti *Treponema pallidum* je možný pouze v prvním a druhém stádiu syfilis. Provádí se mikroskopickým pozorováním sekretu v temném poli. Negativní nález ale neznamená, že se bakterie v pacientově těle nenachází. Spolehlivějším důkazem syfilis je přímá imunofluorescence a PCR.

Raná stádia syfilis se léčí injekcemi benzathin benzylpenicilinu (PNC G), který patří do skupiny penicilinů s úzkým antibakteriálním spektrem. Mechanismus jeho účinku spočívá v inhibici syntézy buněčné stěny bakterií, což znamená že je ničí v růstové fázi. V druhé řadě mohou být nasazeny antibiotika - deoxycyclin (není vhodný pro těhotné), ceftriazon nebo azitromycin. Penicilin se používá i pro léčbu pozdějších stádií nemoci, jen ve vyšších dávkách. Použití penicilinu je ideální díky vysoké pronikavosti do tkání, včetně placenty. Dá se použít i při graviditě a může zabránit přenosu nemoci z matky na plod. [9, 10]

1.3 Frambézie (Yaws)

Jde o infekční dětské onemocnění způsobované bakterií *Treponema pallidum* subsp. *pertenue*. Vyskytuje se především v chudých částech Afriky, Asie a latinské Ameriky. Nízké sociálně-ekonomické standardy a špatné hygienické návyky vedou k šíření této nepohlavní nemoci. Nemoc se z počátku projevuje jako charakteristický papilom (kožní výrůstek), který je plný bakterií a umožňuje tedy snadné diagnostikování. Výrůstky jsou velmi infekční a pokud nejsou léčeny, snadno se šíří. Sekundární frambézie se objevuje týdny až měsíce po infekci a typicky se projevuje ve formě mnohočetných vystouplých žlutých lézí nebo bolesti a otoku dlouhých kostí a prstů.

Diagnóza se stanovuje na základě laboratorních sérologických testů jako jsou TPPA (Treponema pallidum particle agglutination) a RPR (rapid plasma reagin), které se využívají pro zjištění treponemových infekcí. Tyto testy nerozliší frambézii od syfilis, proto je u dospělých, žijících v endemických zónách, kde se frembézie vyskytuje, důležité pečlivé vyšetření a zhodnocení. Pro definitivní potvrzení nemoci se používá PCR. [11]

1.4 Pinta

Je velmi vzácné infekční onemocnění rozvíjející se nejčastěji u dětí, které způsobuje *Treponema pallidum* subsp. *careteum*. Přenáší se přímým dotykem s nakaženou částí kůže. Nemoc má opět tři stádia, která se projevují různými lézemi a zabarvením na kůži. Ostatní orgány nejsou ovlivněny. Inkubační doba je od 7 do 21 dní.

Ve většině případů jsou primární léze malé červené skvrny, které se objevují převážně na odhalených místech pokožky, jako jsou ruce a nohy. Mohou se vyskytovat i na obličeji nebo břiše. Občas svědí a mohou se rozšiřovat ve větší povlaky.

Měsíc až rok po rozvoji prvotní infekce se u infikovaných osob mohou na místě původních lézí objevit nové. Těmto lézím se říká pintids, infikují zpravidla už postižená místa a mohou být suchá a loupat se.

Pozdní fáze nemoci se objevuje dva až pět let od prvních příznaků. Jedná se o bílé až bezbarvé skvrny. Během této fáze se u pacientů může objevit nenormálně suchá a křehká kůže.[12]

2 Sekvenace

DNA sekvenování je laboratorní technika sloužící k přesnému určení sekvence nukleotidů (bazí) v molekule DNA. Sekvence bazí , většinou označovaných počátečními písmeny jejich názvů (A - adenosin, C - cytosin, G - guanin a T - thymin), kóduje biologickou informaci dané buňky. Zjištění sekvence fragmentu DNA má mnohostranné využití, jako je diagnostika nemocí, kontorla patogenů nebo provádění fylogenetických studií. Zároveň představuje klíč k porozumění funkci genů. V současnosti je k dispozici několik různých metod sekvenování, z nichž každá má jiné charakteristiky a je vhodná pro jiný typ práce. [13]

2.1 První generace

V rámci první generace sekvenování rozlišujeme dvě metody:

- Metoda terminace řetězce (Sanger et al., 1977), ve které je sekvence jednořetězcové DNA určována enzymatickou syntézou komplementárního řetězce, ukončeného na specifické nukleotidové pozici
- Metoda chemické degradace (Maxam a Gilbert, 1977), ve které je sekvence dvouřetězové DNA určena pomocí aplikace chemikálií, které ji naštěpí v určených nukleotidových pozicích

Obě metody byly ze začátku stejně populární, ale Sangerova metoda začala postupně převažovat. Ze začátku obě metody používaly radioaktivní značení, což představovalo riziko pro vědce, kteří práci prováděli. Důvodem pro rozmach Sangerovy metody byl nejen vynález fluorescenčních barviv, ale automechanizace metody, která byla nezbytně nutná z důvodu zpracování obrovského množství sekvenačních dat, které by bylo náročné zpracovávat ručně.

2.1.1 Sangerova metoda

DNA sekvenace metodou ukončování řetězce je založena na faktu, že jednořetězcové vlákna DNA, která se liší v délce jen o jeden nukleotid, lze separovat jedno od druhého pomocí gelové elektroforézy. To znamená, že po proběhnutí elektroforézy jsou na gelu viditelné bandy různě dlouhých řetězců DNA (v dnešní době až 1000 bp). Pro tuto metodu je nezbytně nutná příprava identických kopií jednořetězcových DNA molekul (templátů), existuje více variant přípravy:

- DNA klonování plazmidovým vektorem - Plazmidy jsou malé kruhové molekuly DNA, přirozeně se vyskytující v bakteriích. Tímto způsobem připravená DNA je dvouřetězcová, takže musí být před použitím denaturována. Jedná se o hojně používanou metodu, hlavně kvůli její nenáročnosti.

- DNA klonování pomocí fágového M13 vektoru - Bakterie hostitelského kmene je infikována bakteriofágem s vloženou cizí DNA. V průběhu lytického cyklu je namnožena celá fágová DNA včetně té cizí. Metoda je speciálně navržená k přípravě jedno-řetězcové DNA. Nevýhodou je, že při použití delšího fragmentu DNA může docházet k delecím, takže je vhodná jen pro přípravu krátkých úseků DNA (do 3 kb).
- DNA klonování fasmidem (=fagemidem) - Běžné bakteriální plazmidy, které nesou navíc část genomu některého bakteriofága. Obsahují jak počátek replikace ColE1, které zajišťují produkci dvouřetězcové plasmidové DNA, tak počátek replikace bakteriofága M13 (nebo jiného), který umožňuje vznik jednořetězcové DNA. Tento systém předchází chybám ke kterým dochází v předchozí možnosti a je proto vhodný pro fragmenty větších délek (10 kb a více).
- DNA klonování kosmidem - Modifikované plazmidové vektory, speciálně navržené pro přípravu velkých fragmentů DNA (až do 45 kb).
- Využití PCR - polymerázová řetězová reakce může být také použita k získání jednořetězcové DNA.

Prvním krokem Sangerovy metody je nasednutí krátkého oligonukleotidu na stejné místo na všech vláknech, tento oligonukleotid pak slouží jako primer pro syntézu druhého (komplementárního) řetězce DNA. Syntéza komplementu probíhá pomocí DNA polymerázy, za přítomnosti čtyř deoxyribonukleotidtrifosfátů (dNTPs - dATP, dCTP, dGTP a dTTP). Terminace řetězce je umožněna přítomností malého množství dideoxyribonukleotidtrifosfátů (ddNTPs), které se liší chybějící OH skupinou nezbytně nutnou pro pokračování replikace. DNA polymeráza nerozlišuje mezi dNTPs a ddNTPs, což znamená že dideoxyribonukleotid může být kdykoli začleněn do rostoucího řetězce, ale způsobí tím jeho ukončení. Tato reakce historicky probíhala ve čtyřech zkumavkách - v každé je zastoupen jiný dideoxynukleotid. V případě, že je přítomný ddATP je řetězec ukončen v místě výskytu thyminu v templátovém řetězci, ale protože je přítomný i dATP, syntéza není vždy ukončena hned při prvním výskytu T v templátu - může pokračovat dokud není templát dlouhý několik stovek nukleotidů, než dojde k začlenění ddATP. Výsledkem je tedy několik nových řetězců, lišících se jen v délce, ale končících vždy ddATP. To stejné se děje ve všech ostatních zkumavkách, vždy jen s tím rozdílem, že jsou řetězce ukončené jiným dideoxyribonukleotidem. Proto je použita gelová elektroforéza. Řetězce vzniklé syntézou v přítomnosti ddATP jsou naneseny do jednoho řádku gelu a stejně je naloženo i s řetězci vzniklými syntézou s dalšími třemi dideoxyribonukleotidtrifosfáty. Po proběhnutí elektroforézy může být DNA sekvence přečtena přímo z jednotlivých proužků přítomných v gelu. S tím, že proužek, který na gelu doputoval nejdále, je nejkratší. Jedná se tedy o řetězec, který byl ukončen připojením komplementárního ddNTP při prvním výskytu nukleotidu v DNA templátu. [14, 15, 16] Krátký oligo-

nukleotid (primer), který nasedá na templátové vlákno v prvním kroku metody je nezbytně nutný. Hlavně protože DNA polymeráza nemůže zahájit replikaci vlákna bez přítomnosti 3' konce na který pak navazuje další nukleotidy. Primer zároveň určuje od kterého místa bude replikace zahájena - která část molekuly bude replikována.

Postup popsaný výše už se v dnešní době nepoužívá, Sangerova metoda je používána v upravené formě. Už není potřeba probíhající reakce dělit do více zkumavek, probíhají dohromady, to je umožněno vynálezem fluorescenčních značek. Ty se používají ke značení ddNTPs - čtyři různé značky (s různými emisními spektry), každá pro jiný dideoxynukleotid. Po dostatečném množství reakcí je sekvenovaná DNA přítomna v různě dlouhých molekulách. Díky fluorescenčním barvivům je možné přejít z klasické elektroforézy na kapilární. Při ní se využívá kapilára naplněná polyacrylamidovým gelem. Molekuly DNA jsou seřazeny podle zvyšující se délky, každá z nich má na konci ddNTP s odpovídající fluorescenční značkou. Simultánně je prováděna detekce fluorescenčních barviv CCD detektorem (Charge-Coupled Device).

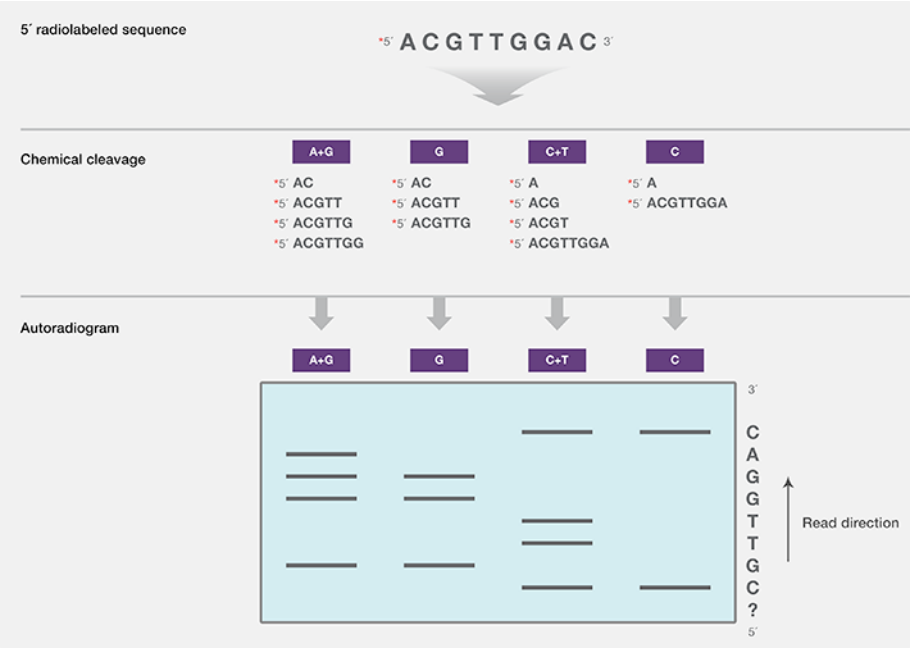
2.1.2 Maxam-Gilbertova metoda

Tato metoda je založena na poznatku, že lze na konci značenou DNA molekulu sekvenovat po jejím rozštěpení v určeném místě. Metoda pracuje s molekulou jednořetězcové DNA (nebo dvouřetězcové, která je denaturací nebo štěpením upravena na jednořetězcovou), která je na jednom konci (buď 3' nebo 5') radioaktivně značena pomocí ^{32}P. DNA sekvence je štěpena v místě adenosinu, cytosinu, guaninu nebo thyminu v závislosti na použité chemikálii. Reakce opět probíhá ve čtyřech zkumavkách, jedna pro puriny (A + G) jedna pro pyrimidiny (C + T) a po jedné pro G a C. Částečné štěpení v každé z určených bází vede ke vzniku sady radioaktivně značených sekvencí odlišující se v délce, které lze odlišit při použití gelové elektroforézy. [14, 17]

Na obrázku 2.1 je příklad sekvenace DNA označené na 5' konci. Gel po elektroforéze je o něco málo náročnější na přečtení než u Sangerovy metody (kde jsou bandy pro jednotlivé báze), ale i tak je snadno čitelný.

2.2 Druhá generace

I přesto, že byly předchozí dvě metody upraveny a zautomechanizovány, jsou omezeny délkou sekvence, kterou jsou schopny sekvenovat (do stovek párů bazí) v rámci jednoho experimentu. Ve srovnání s lidským genomem to znamená, že jsou v jednom běhu schopny sekvenovat jednu pětimiliontinu genomu. Proto nastupuje takzvaná next generation sekvenování, která umožňuje sekvenovat mnohonásobně vetší čtení.



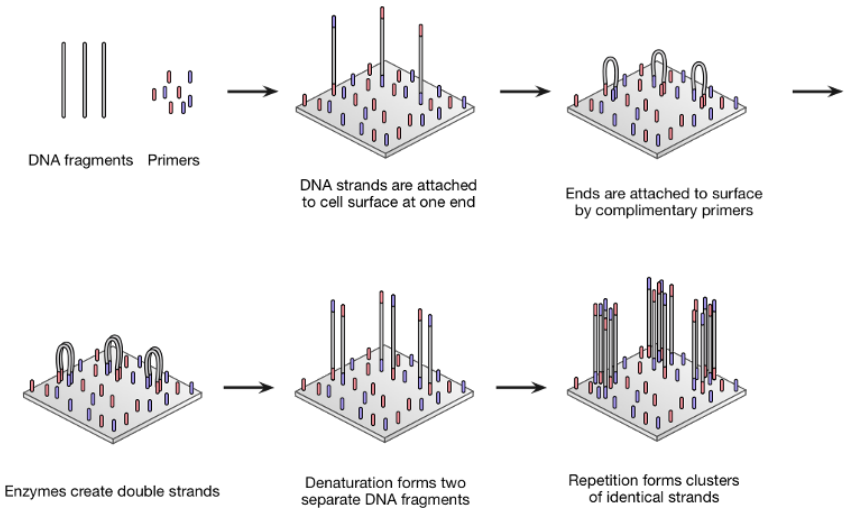
Obr. 2.1: Maxam-Gilbert sekvenování [17]

Metody druhé generace mohou být rozděleny do dvou kategorií:

- sekvenace hybridizací - Používá filtry se známými oligonukleotidovými sekvencemi, na které je hybridizován značený vzorek DNA. Po odplavení nežadoucí DNA, je možné určit zda sekvence zachycených značených fragmentů odpovídá sekvenci DNA oligonukleotidů na filtru.
- sekvenace syntézou (SBS) - Využívají polymerázu nebo ligázu k navázání (fluorescenčně značených) nukleotidů.

Metody druhé generace jsou bez výjimky závislé na amplifikaci DNA před samotným sekvenováním. Tento krok vede k tvorbě dostatečného množství kopií DNA a díky tomu je pak zajištěn dostatečný signál při stanovení jednotlivých bází. K aplifikaci DNA se konvenčně využívají metody:

- Emulzní PCR (polymerase chain reaction - polymerázová řetězová reakce): Funguje na podobném principu jako klasická PCR, jen s tím rozdílem, že celá reakce probíhá v kapénkách vody v emulzi s olejem. Amplifikovaná DNA je navázána na speciální adaptéry (syntetické kuličky) a použita jako templát pro PCR. Na každé kuličce se v průběhu reakce vytvoří až 10 miliónů kopií původní DNA. [18, 19]
- Můstková amplifikace: K amplifikaci slouží skleněná destička s navázanými primery dvou typů. Předpřipravené fragmenty DNA jsou jednořetězcové a mají na koncích navázaný adaptér komplementární k oligonukleotidovým primerům na destičce - dochází k jejich navázaní. Díky DNA polymeráze dochází k



Obr. 2.2: Můstková amplifikace - princip [19]

vytvoření komplementárního řeztězce. Dvouřetězcová DNA je denaturována, nově vytvořené vlákno se ohýbá a druhým koncem se navazuje na druhý typ primeru, dochází k syntéze komplementárního vlákna, to vede k vytvoření dvouřetězcového můstku, který je následnou denaturací zničen a celý proces se od začátku opakuje. [18, 19] Princip můstkové amplifikace je znázorněn na obrázku 2.2.

- Nanokuličky: Tato metoda nepoužívá k amplifikaci PCR, ale klonuje DNA fragmenty pomocí retrovirového vektoru. DNA spolu s adaptory je pomocí určitého oligonukleotidu zformována do kruhů, cirkulární DNA je replikována phi29 DNA polymerázou. Nově sekvenovaná vlákna společně s tamplátovou cirkulární DNA vytváří nanokuličku, která je dále sekvenována. [18]

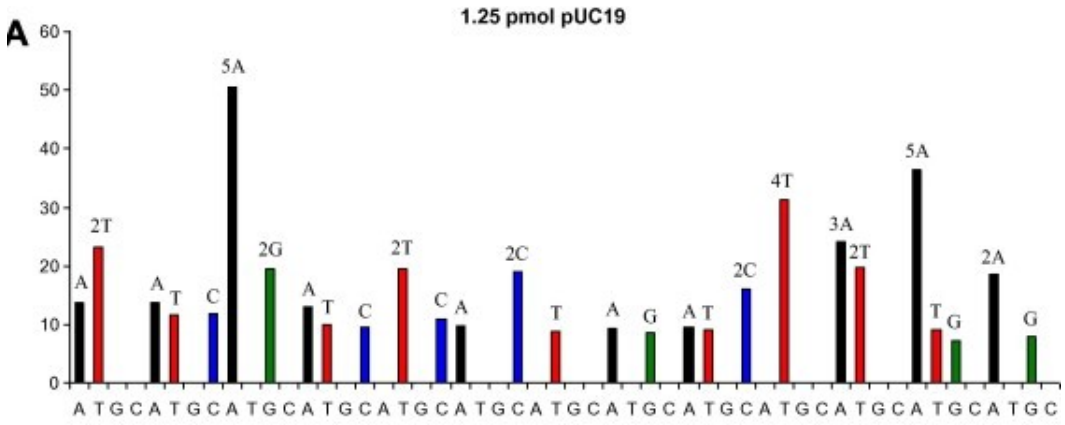
2.2.1 454 Pyrosekvenování

Pyrosekvenování je založeno na principu sekvenace syntézou, při které je syntetizováno komplementární vlákno. Narozdíl od předchozí metody založené na tomto principu (Sanger) zde nejsou využívány ddNTP. Skládá se ze série reakcí, které vedou k tomu, že jakmile je do nového komplementárního vlákna vložen nový nukleotid, je vyzářeno viditelné světlo.

DNA templát je v přítomnosti několika enzymů (DNA plymeráza, ATP sulfuryláza, luciferáza a apyráza) replikován. Při nasednutí deoxinukleotidtrifosfátu komplementárnímu k templátu, je uvolněn pyrofosfát. Ten slouží jako komponent pro tvorbu ATP, právě díky enzymu ATP sulfuryláze. ATP je využito enzymem luciferázou k přeměně luciferinu na oxyluciferin, tato reakce vede k produkci světla přímo

úměrného k počtu bazí, které byly navázány. Nenavázané nukleotidy jsou okamžitě degradovány nukleotidázou. [14, 20, 21]

Technologie 454 pyrosekvenování využívá této metodiky, ale posouvá ji dál. Sekvenovaná DNA je nejprve amplifikována pomocí emulzní PCR, poté je každá kapička uzavřena do samostatné komůrky, ve které jsou přítomné všechny látky potřebné pro pyrosekvenování. Tyto komůrky jsou součástí sekvenační destičky (PicoTiterPlate), která může obsahovat až jeden milion takových komůrek. Destička je opakovaně omývána deoxinukleotidtrifosfáty (T, A, C, G), když je daný nukleotid komplementární k templátu, je navázán DNA polymerázou a dříve popisovanou reakcí dojde k chemoluminiscenci. Světlo je zachycováno kamerou, intenzita je přepočítána na takzvanou flow-value a více méně odpovídá množství přidaných bazí. Informace o bázi a flow-value jsou ukládány do počitače. Ze sekvence bazí a hodnot intenzity lze zkonstruovat flowgram. Příklad je na obrázku 2.3. [21, 22]



Obr. 2.3: Flowgram - ukázka [22]

2.2.2 Ion torrent

Ion torrent sekvenování bylo uvedeno v roce 2010. Velký úspěch této metody byl způsoben použitím čipu (destičky), který obsahuje až milion CMOS (Complementary Metal-Oxide-Semiconductor) senzorů v matici a umožňuje tak zpracovávat data najednou jednoduše a poměrně levně. Další velkou inovací bylo představení elektrochemického ISFET (ion field sensitive transistor - iontově senzitivní tranzistor (tranzistor řízený polem)) senzoru, který je na dně každé komůrky. Tato metoda je, stejně jako předchozí, založena na konceptu sekvenace syntézou, ale na rozdíl od dalších metod druhé generace nevyužívá detekci světla nebo fluorescenčního značení. Namísto toho využívá detekci protonu (H^+), který je uvolněn při každém začlenění

nukletidu do nově sekvenovaného vlákna. Uvolnění vodíkových iontů vede ke změnám pH, které mohou být detekovány jako změny napětí pomocí CMOS-ISFET senzoru. Pokud do syntetizovaného řetězce není nukleotid připojen, neobjeví se žádný napěťový signál, díky tomu, na rozdíl od předchozích metod, není nutné nevyužité nukleotidy chemicky degradovat.

Sekvenování na této platformě probíhá podobně jako u předchozí metody. DNA sekvence je amplifikována využitím emulzní PCR a je nanesena na sekvenační polovodičovou destičku. Pravidelným omýváním komůrek dNTP (vždy jen jedním druhem) je stanovována DNA sekvence pomocí detekce napětí. Dojde-li k navázání dvou stejných nukleotidů hned po sobě je detekované napětí dvakrát větší. Problém ovšem nastává (stejně jako u předchozí metody) při stanovování repetetivních bazí (například 'AAAAAAAAAA'), kde je těžké rozlišit přesný počet po sobě jdoucích nukleotidů jen na základě napěťového údaje. [19, 20, 23]

2.2.3 SOLiD

Sekvenování ligací je metoda využívající DNA ligázu, enzym hojně využívaný v biotechnologiích, díky schopnosti vázat dvouřetězcové DNA vlákno. SOLiD (Sequencing by oligonucleotide ligation and detection) stejně jako fluorescenční úprava metody Sanger, je založen na detekci fluorescenčních signálů, narozdíl od Sangerova sekvenování, kde je fluorescenční barvivo vázáno na každý dideoxinukleotid, je u SOLiD sekvenování fluorofor použit k označení dvou nukleotidů. Na rozdíl od předchozích metod tedy není výstup této metody snadno čitelný, pro označení všech šestnácti možných kombinací nukleotidů, jsou používány jen čtyři fluorescenční sondy. [19, 14, 23]

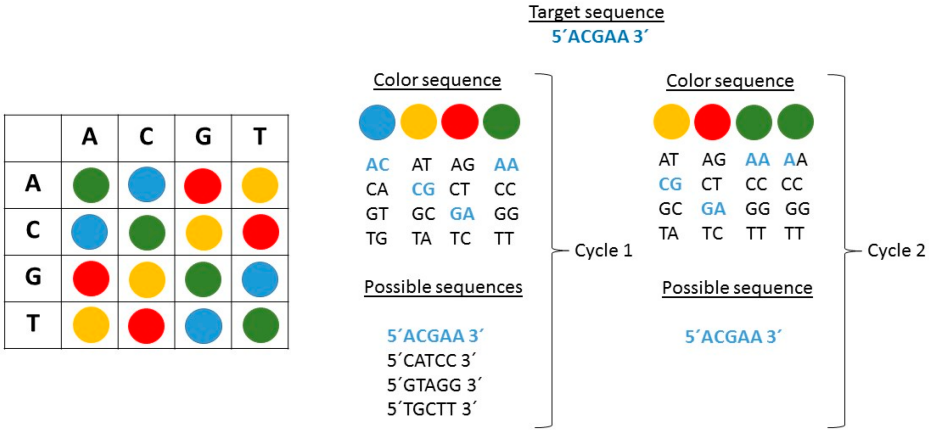
SOLiD sekvenování probíhá v několika krocích. V prvním kroku na sekvenovaný fragment DNA řetězce hybridizuje jedna ze 16 značených sond. Ta sestává ze dvou známých nukleotidů v polohách n a n+1, komplementárních k sekvenovanému fragmentu, nasledovaných sekvencí degenerovaných bází s fluorescenční značkou. Sonda je navázána DNA ligázou na použitý primer. Ve druhém kroku dochází k rozštěpení sondy, je uvolněn fluorescenčně značený konec (detekce fluorescence), na místě zůstává pět nukleotidů dlouhý řetězec s 5' fosfátovou skupinou na konci . Následuje série 10 cyklů hybridizace ligace a štěpení díky čemuž může celý proces proběhnout znovu. Po dokončení prvního cyklu sekvenace, probíhá několik dalších, pokaždé s kratším/delším primerem (o jeden nukleotid, o dva, atd.).

Na rozdíl od Sangerovy metrody, kde bylo čtení každé báze spojeno s fluorescenčním signálem, zde je k určení sekvence bází potřeba sada signálů. Technika colorspace byla novinkou představenou platformou SOLiD, která je jediná, která ji využívá. V této technice každý fluorescenční signál představuje dva nukleotidy. Každý

nukleotidový pár má svoji značku, ale jak je vidět v druhé části obrázku 2.4 barva určená pro různé páry není náhodná. Reverzní (například AC a CA), komplementární (AC a TG) a reverzně komplementární (AC a GT) dvojice sdílejí stejnou značku. [23, 24]

Read Position	n	n+1	n+2	n+3	n+4	n+5	n+6	n+7
Universal seq primer n	F	F	nF	nF	nF			
Universal seq primer n+1		F	F	nF	nF	nF		
Universal seq primer n+2			F	F	nF	nF	nF	
Universal seq primer n+3				F	F	nF	nF	nF

(a)



(b)

Obr. 2.4: SOLiD sekvenace - princip fluorescenčního značení [23]

Na obrázku 2.4 je znázorněn princip fluorescenčního značení a využití techniky color-space ke stanovení pořadí nukleotidů. F: dva fluorescentně značené nukleotidy (dohromady jedním fluoroforem), nF: neznačené báze.

2.2.4 Illumina

Opět se jedná o sekvenaci syntézou, na rozdíl od Sangerovy metody, používá Illumina fluorescentně značené reverzibilní terminátory, díky tomu je terminace řetězce u této metody vratná. Na rozdíl od jiných dříve zmiňovaných metod, se zde k amplifikaci DNA používá můstková PCR. Fragmenty sekvenované DNA jsou spojeny s primery imobilizovanými na povrchu. V každém cyklu jsou současně přidány čtyři fluorescentně značené nukleotidy, které mají na 3' konci místo OH skupiny 3'-o-azidomethylovou skupinu, tato chemická blokace vede k ukončení řetězce. Po začlenění nukleotidů je metodou TIRF detekována fluorescence. V dalším kroku jsou vymyty nenavázané nukleotidy a chemická blokáda 3' konce je odstraněna použitím

tris-(2-karboxyethyl)fosfinu (TCEP), díky čemuž lze pokračovat v syntéze řetězce. Tento postup se cyklicky opakuje dokud není sekvenován celý fragment DNA.[19, 23]

TIRF (Total Internal Reflection Fluorescence - Fluorescence s totálním vnitřním odrazem) je metoda využívající odrazu excitačního paprsku od rozhraní dvou prostředí. Díky tomu je fluorescence odkloněna od ohniskové roviny, což vede ke zvýšení SNR (Seignal to Noise Ratio - poměr signálu k šumu). Dochází ke vzniku evanescentní vlny, která excituje tenkou vrstvu vzorku (100 nm). Hlavní předností této metody je právě to, že je schopná detekovat fluorofory, které jsou velmi blízko pevného povrchu (sekvenační destičky). [23, 25]

V současnosti jsou Illumina sekvenátory nejrozšířenější. To je způsobeno nejen nízkými náklady na sekvenaci a jejich vysokou přesností, ale také tím že mají na trhu celou řadu možných variant, přizpůsobených různým projektům.[26]

Dostupné varianty:

- iSeq 100 (maximálně 1.2 Gb, délka čtení do 150 bp)
- MiniSeq (max 7.5 Gb, délka čtení do 150 bp)
- MiSeq (až 15 Gb, délka čtení do 300 bp)
- NextSeq 550 (až 120 Gb, délka čtení do 150 bp)
- NextSeq 1000 a NextSeq 2000 (do 540 Gb, délka čtení do 300 bp)
- NovaSeq 6000 (do 3 Tb, délka čtení do 250 bp)
- NovaSeq X (až 8 Tb, délka čtení do 150 bp)

2.3 Třetí generace

Ačkoli neexistuje žádná zřetelná hranice mezi druhou a třetí generací sekvenátorů, za specifikaci třetí generace by šlo považovat sekvenování v reálném čase a SMS (Single molecule sequencing - sekvenace jedné molekuly). Klíčovou vlastností metod třetí generace je schopnost přesně sekvenovat dlouhé řetězce DNA bez nutnosti ji předem nějak zpracovávat a amplifikovat. [27]

2.3.1 Pacific Biosciences

Sekvenátor od firmy Pacific Biosciences, využívá metodu SMRT (Single molecule real time - sekvenace jedné molekuly v reálném čase) a stále se jedná o nejvíce využívanou platformu pro tento typ práce. Na trh byl uveden roku 2011 pod názvem PacBio RS sequencer. Tento sekvenátor generoval poměrně krátká čtení (s průměrnou délkou 1.5 kb) s velkou chybovostí (13 %). V průběhu let došlo k rozvoji technologie a vyšel nový sekvenátor (PacBio Sequel System), který prodloužil délku čtení více než 10x.

Jako templát pro metodu SMRT slouží uzavřená (cirkulární) jednořetězcová DNA, která je vytvořena navázáním (hairpin - vlásenkových) adaptorů na oba konce molekuly (SMRTbells).

Narozdíl od předchozích metod, kde docházelo k sekvenaci DNA sekvencí navázaných na sekvenační destičku, nebo kuličku pomocí DNA polymerázy, která se pohybovala podél templátu, u této metody je DNA polymeráza fixována na dno speciální sekvenační komůrky (SMRTcell). Komůrky, kterých je v rámci sekvenátoru tisíce, mají na svém dně takzvané ZMW (zero mode waveguides), které slouží k usměrňování světla. V průběhu sekvenace je k templátovému řetězci, procházejícímu DNA polymerázou, navázán komplementární fluorescenčně značený nukleotid. Po osvícení laserem dojde k emitování světla, které je detekováno kamerou. Velikost ZMW je menší než vlnová délka excitačního světla, což vede k jeho exponenciálnímu útlumu. Ozářeno je tedy jen dno komůrky, což zamezuje detekci světla pocházejícího z ostatních fluorescenčně značených dNTPs, které jsou v roztoku. PacBio umožňuje sledovat detekci světla, emitovaného nově navázanými nukleotidy v reálném čase formou kontinuálního záznamu. Díky použití cirkulárních SMRTbells je umožněno temlátovou DNA sekvenovat v jednom běhu hned několikrát, pokud to životnost DNA polymerázy dovolí. Takovýto dlouhý kontinuální read lze pak rozdělit do více 'subreadů', díky odstranění adaptorů. To vede k vyšší přesnosti čtení. [28, 29, 30]

2.3.2 Oxford Nanopore

Technologie sekvenace pomocí nanopóru se začala rozvíjet už na konci 80. let. Nicméně kvůli technickým obtížím bylo prvního úspěšného výsledku dosaženo až v roce 2012. Dva roky na to vydala firma Oxford Nanopore Technologies (ONT) sekvenátor založený na této metodě.

Sekvenace nanopórem vychází z detekce změn elektického proudu, vyvolaných průchodem jednořetězcové molekuly DNA nebo RNA. Jednořetězcová sekvence prochází pórem a její průchod reguluje motor protein. Změna elektrického signálu je detekována a je charakteristická pro každý nukleotid.

Sekvenace probíhá v průtokové komůrce (flow cell), ta obsahuje elektricky odolnou membránu ze syntetického polymeru, obsahující nanopór. Membrána je ponořena do iontového roztoku a je na ni přivedeno napětí. To způsobí stálý průchod iontů skz kanál, který je narušován průchodem jednotlivých bazí sekvenované DNA. Elektrický signál z nanopóru je zaznamenáván a graficky reprezentován ve formě takzvaného squiggle plotu.

Na rozdíl od SMRT sekvenování není tato metoda limitována technologií, ale délkou sekvenované molekuly (DNA/RNA). Díky tomu je možné získat extrémně dlouhá čtení (v poslední době až 1 Mb). Nevýhodou je vysoká chybovost (15 %).

U sekvenování nanopórem není možné sekvenovat jeden řetězec několikrát po sobě jako u SMRT sekvenování, kde je tento postup používán k určování konsenzu s vyšší kvalitou. Pro větší přesnost byl vyvinut postup pro sekvenování obou vláken DNA. Na jeden konec dvouřetězcové DNA byl navázán vlásenkový adaptér, což vedlo k sekvenaci komplementárního vlákna hned po templátovém. Tento systém se nazýval dvousměrné sekvenování (2D), v poslední době byl ale nahrazen $1D^2$, který využívá normální adaptory s navázanou specifickou sekvencí, která podporuje vstup druhého řetězce.[28, 29, 30, 31]

3 Typování

Typování bakteriálních kmenů hraje významnou roli při určování rozmanitosti patogenů a epidemiologických infekcí. Může potvrdit epidemiologickou souvislost a poskytnout pohled na dynamiku populace. Existuje několik tradičních metod, které zůstávají první volbou při určování řetězce. Jednu z těchto metod představuje gelová elektroforéza s pulsním polem (PFGE). Tato metoda používá k separaci velkých DNA molekul elektrické pole různé intenzity, které v pravidelných intervalech mění směr, ze kterého působí na gel. Zkoumaná DNA je nejprve naštěpena pomocí speciálních restrikčních enzymů. Po proběhnutí elektroforézy jsou na gelu proužky tvořící specifické vzory. Tato metoda je známá jako zlatý standard, ale je náročná, jak co se týče standardizovaných protokolů pro zpracování patogenů, tak na laboratorní čas (příprava vzorků, gelu a následné zpracování). Další metodou využívající restrikční štěpení je REA (Restriction endonuclease analysis). Představuje jednu z prvních metod určování řetězce, využívá enzym (obvykle *Hind*III), který DNA sekvenci naštěpí a díky tomu je možné rozlišit po proběhnutí klasické gelové elektroforézy části řetězce ve formě proužků (bandů). Díky pokroku v sekvenačních technikách se přešlo k charakterizaci na základě sekvenování, jako je multilokusové sekvenční typování (MLST), sekvenční metoda stanovující typ sekvence na základě provozních genů. Tato metoda poskytuje jednoznačné výsledky a umožňuje jejich porovnání mezi laboratořemi pomocí centralizované databáze (PubMLST - kapitola 4.2). [32]

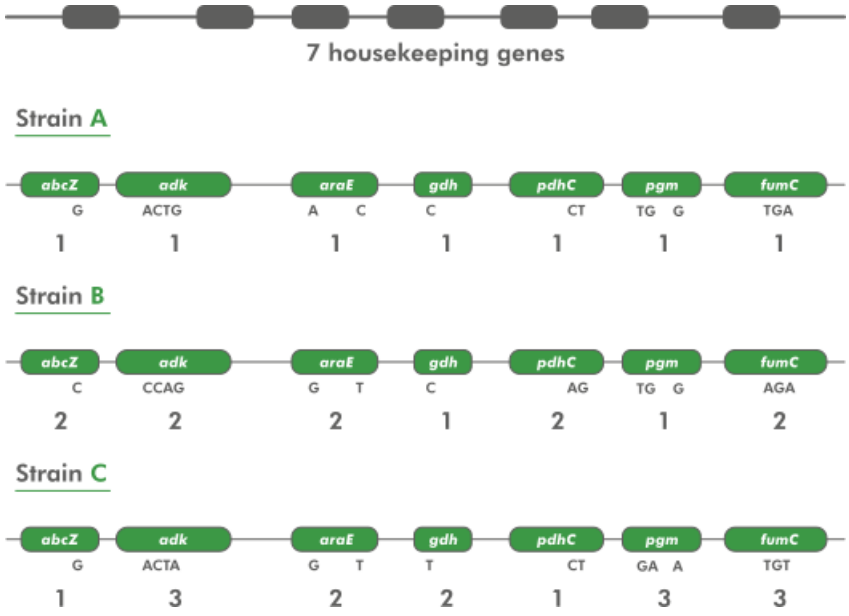
3.1 MLST typování

Multilokusové sekvenční typování (MLST typing) je metoda, která umožňuje charakterizaci bakteriálních izolátů na základě vnitřních sekvencí jejich provozních (housekeeping) genů.

To obvykle znamená systematické sekvenování šesti až sedmi dobře zachovávaných lokusů (genů) bakteriálního genomu. Variace v alelách jednotlivých lokosů jsou katalogizovány a typ sekvence je určen na základě porovnání s profily jiných izolátů v databázi.[33]

Dejme tomu, že organizmus typujeme na základě (obvykle) sedmi genů, které poskytují pro daný organismus největší diferenci v sekvencích. Z každého z těchto provozních genů je nasekvenován vnitřní fragment o délce 450 až 500 bp, odvíjející se od použitých primerů.Pro každý z nich je stanovena číslovaná tabulka možných alel, které mohou pro daný gen nastat. Alele, která se jakkoli liší od ostatních, už zaznamenaných, je přiděleno nové unikátní číslo.

Každý představitel druhu je tedy jednoznačně charakterizován sérií sedmi čísel alel jednotlivých house-keeping genů. Typ sekvence je určen na základě porovnání



Obr. 3.1: Princip MLST typování [35]

s alelovými profily jiných izolátů v databázi. Tento postup je jednoduše znázorněn na obrázku 3.1.

Velkou předností MLST typování je to, že sekvenční data a alelické profily jednotlivých bakterií mohou být snadno porovnány s těmi už uloženými v centrální databázi na internetu. Alelické profily lze navíc sestavit z klinických materiálů (z CSF (cerebrospinal fluid - mozkomíšní mok) nebo krve) pomocí PCR amplifikace předem určených sekvencí z house-keeping genů.[33],[34]

3.2 CDCT

Jde o typizační metodu navrženou pro *Treponema pallidum* subsp. *pallidum* vědci z CDC (Centers for Disease Control and Prevention) v roce 1998. Metoda je založena na stanovení počtu 60 bp dlouhých repetic v genu *arp* (acidic repeat protein) a na sekvenčních rozdílech v genech *tprE* (TP0313), *tprG* (TP0317) a *tprJ* (TP0621), patřících do skupiny *Treponema pallidum* repetetivních genů. Rozdíly jsou stanoveny na základě analýzy polymorfismu délky restrikčních fragmnetů (RFLP). Označení podtypu je stanoveno pomocí počtu repetic a vzoru RFLP - například 14a pro kmen Nichols. Přidáním genu TP0548 je zvýšena rozlišovací schopnost této metody - ECDCT (enhanced-CDCT). [36]

3.3 MLST typování *Treponema pallidum* subsp. *pallidum*

Syfilis představuje důležitý problém pro širokou veřejnost. Vzrůstající incidence této nemoci vede k potřebě charakterizovat změny v genotypu bakterie, pomocí čehož by bylo možné dopředu předpokládat její patogenitu a chování k farmakům. Navíc lze díky molekulárnímu typování získat strukturní data různých populací a na jejich základě určovat jejich geografické rozložení, podobnost mezi organismy s podobným typem sekvence. [37]

V dnešní době se k MLST typování *Treponema pallidum* subsp. *pallidum* využívají tři geny, jde o TP0136, TP0548 a TP0705. První dva z těchto house-keeping genů kódují proteiny vnější membrány, zatímco poslední kóduje protein vázající penicilin. Protein TP0136 hraje navíc důležitou roli při replikaci, protože váže fibronectin, důležitou složku extracelulární matrix hostitele a umožňuje tím adherenci. [37], [38], [39]

V databázi PubMLST lze k těmto housekeeping genům dohledat alely. Pro gen TP0136 je dostupných 38 možných alel, které se od sebe krom jiného liší i délkou - minimální 859 bp, maximální 1051 bp. Alely s číslem 8 a 38 v souboru chybí. Pro gen TP0548 je dostupno 81 možných alel opět s variabilní délkou (879 - 891 bp), v souboru referencí stažených z PubMLST databáze chybí alela číslo 21. Gen TP0750 má fixovanou délku 21 dostupných alel (732 bp). Jak již bylo dříve zmiňováno, postup se neustále rozvíjí, poslední změny v databázi proběhly 22. září minulého roku, kdy byl aktualizován dataset pro lokus TP0136.

Jako referenční byly použity sekvence kmene Nichols a kmene SS14 dostupné z [41] a [42]. Skupina SS14 zahrnuje kmeny SS14, Grady, Mexico A a Philadelphia 1. Skupina Nichols zahrnuje kmeny Nichols, Bal 73-1, DAL-1, MN-3, Philadelphia 2, Haiti B a Madras. Pro zástupce SS14 byla největší variabilita pozorována v genu TP0705, zatímco zástupci Nichols projevovali nejvyšší variabilitu v genu TP0548. [37],[43]

4 Práce s daty

Pro získání a při zpracování genomických dat *Treponema pallidum* subsp. *pallidum* bylo využito několik databází a programů, které k tomuto účelu slouží. V této kapitole je naznačeno jakým způsobem se dají využít a jak s nimi pracovat.

4.1 SRA databáze

The Sequence Read Archive (SRA) je databáze spadající pod NIH (National Institutes of Health) obsahující surová sekvenační data. Je součástí INSDC (International Nucleotide Sequence Database Collaboration), která propojuje tento archiv s EBI (European Bioinformatics Institute) a DDBJ (DNA Database of Japan). Jakákoli data nahraná do jedné z těchto databází jsou sdílena s ostatními.

V databázi bylo v době vydání této práce k dispozici 2242 vzorků pro dotaz *Treponema pallidum* subsp. *pallidum*. Sekvenační data pocházela převážně z přístrojů Illumina (MiSeq, HiSeq 2500, NextSeq 500, NovaSeq 6000), ale i ze sekvenátorů pacbio, 454 Roche a MinION.

Pro práci s touto databází je nutné si nainstalovat sra toolbox [45], který umožňuje data z databáze stahovat na základě ID kódu a následně je rozbalovat z archivu .sra na .fasta/.fastq/.sff/.sam soubory. Samozřejmostí je také zpětná konverce na .sra archiv. Celý toolkit je určen pro linux jako většina bioinformatických aplikací. Jednoduše se dá nainstalovat přes anacondu (distributor jazyků R a Python, který zjednodušuje správu a stahování balíčků) a pak rozjet přes příkazový řádek.

Tato práce byla navrhována na základě dat použitých v článku [46], která jsou volně dostupná právě v SRA databázi pod ID SRX1798946. [44] Jedná se o raw sekvenační data *Treponema pallidum*. Pomocí SRA toolboxu byla tato data převedena na fastq soubory. Tento textový formát obsahuje jak biologickou sekvenci, tak údaje o kvalitě čtení. Každý ze souborů obsahoval přes 3,5 miliónů čtení s průměrnou délkou 251 bp. Pro příjemnější práci s daty byly soubory pomocí programu MATLAB rozděleny na 1760 menších souborů - původní velikost 995 064 kB snížena na 556 kB.

4.2 PubMLST databáze

Jde o veřejnou databázi určenou pro molekulární typování - sbírku otevřených databází, které zahrnují data sekvencí populace s informacemi o jejich původu a fenotypovém profilu pro více než 100 různých mikrobiologických druhů. Celá databáze obsahuje více než 33 miliónů alel. Pro *Treponema pallidum* subsp. *pallidum* obsahuje skoro 5000 alelických sekvencí.

Pomocí této databáze byly zjištěny sekvence alel pro geny TP0136, TP0548 a TP0705. Tyto geny jsou momentálně používány pro MLST typování *Treponema pallidum* subsp. *pallidum*. Jedná se o postup navržený autory Grillová a kolektiv [37]. V následujících letech se počítá s rozšířením postupu o další geny.

4.3 Kvalita dat

Sekvenování generuje vysoko objemová data, která můžou být náročná na zpracování. Proto je důležité určit o jak kvalitní data se jedná. K tomuto účelu slouží takzvané Phred skóre, využívané převážně u druhé generace sekvenátárů. Každé bázi je při čtení přiřazeno hodnocení kvality správné identifikace. Tento přístup byl poprvé použit v rámci Human Genome projektu (autory Dr. P. Green a Dr. B. Erwing [40]) a je to doposud nejvyužívanější přístup jak v akademickém tak v komerciálním DNA sekvenování.

Existuje několik typů kódování Phred skóre v závislosti na použitém typu sekvenátoru. Použitá data z Illumina MiSeq používají typ kódování Phred+33. Na základě Phred skóre se určuje kvalita báze (správná identifikace báze přístrojem). K výpočtu kvality je použit vzorec:

$$Q(A) = -10 * log(P_A)$$

Kde P_a značí pravděpodobnost, že při určení báze došlo k chybě. Vyšší skóre tedy znamená nižší pravděpodobnost chyby. Například Q(A)= 1O odpovídá pravděpodobnosti, že jedna báze z desíti je určena špatně. Q(A) = 30 pak říká, že 1 z 1000 bazí je chybná. To znamená, že A (pravděpodobnost správného přečtení báze) je 99,9% [47], [48] Ve fastq souborech je Phred skóre uchováváno pomocí jednomístného ASCII kódu, který se liší v závislosti na typu kódování.

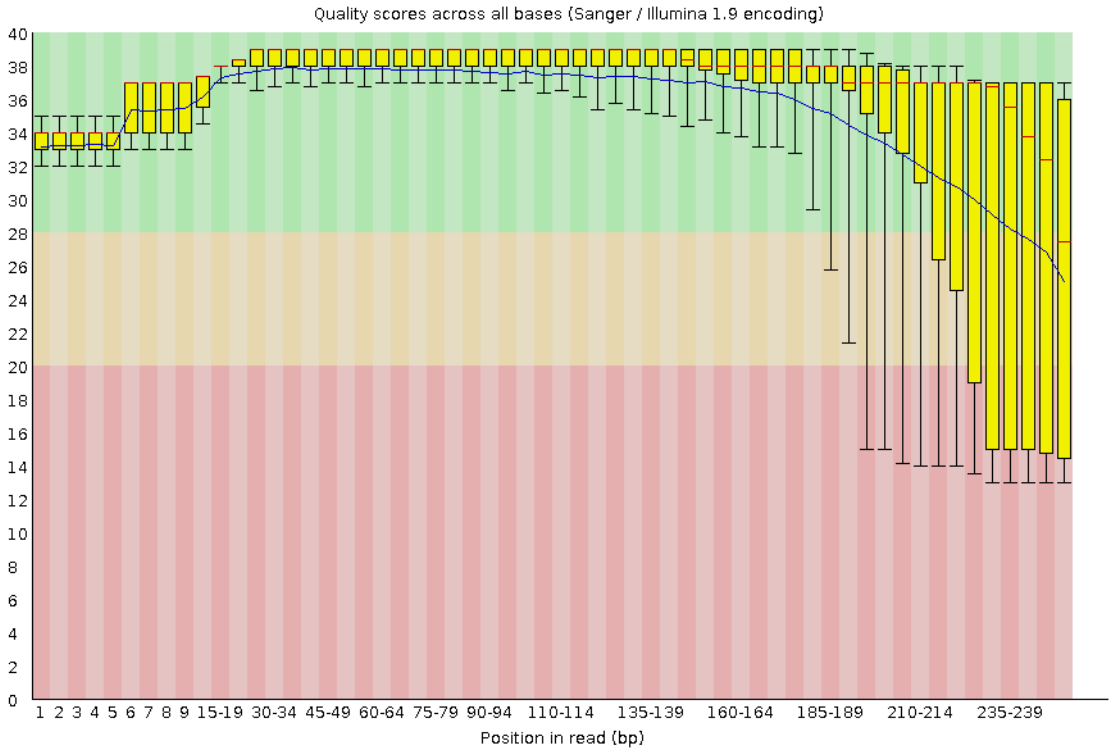
4.4 FastQC

Představuje jednoduchý způsob jak zkontrolovat kvalitu u raw sekvenačních dat. Nabízí širokou nabídku analýz, které poskytují přehled, o tom jaké problémy se mohou vyskytovat v datech. Program dokáže pracovat s daty v různých formátech (jako jsou .SAM, .BAM nebo .FastQ soubory), samozřejmostí jsou také grafy a tabulky hodnocených dat.[50]

Pro vizualizaci kvality bází napříč celou délkou čtení se velmi často používá krabicový diagram (takzvaný Whisker plot). Žluté boxy představují střední část dat, jsou ohraničeny 1. a 3. kvartilem (25 - 75 %), linie vycházející nahoru a dolu reprezentují variabilitu dat pod prvním a nad třetím kvartilem. Modrá linka znázorňuje průměrnou kvalitu čtení. Pozadí grafu je barevně rozlišeno podle kvality čtení na:

velmi dobrou (zelená), přiměřenou (oranžová) a špatnou (červená). Typicky se kvalita zhoršuje ke konci čtení.

Na obrázku 4.1 je příklad výstupu FastQC pro data SRX1798946 konkrétně pro čtení 1559001 až 1560000. Kvalita dat výrazně klesá ke konci čtení.

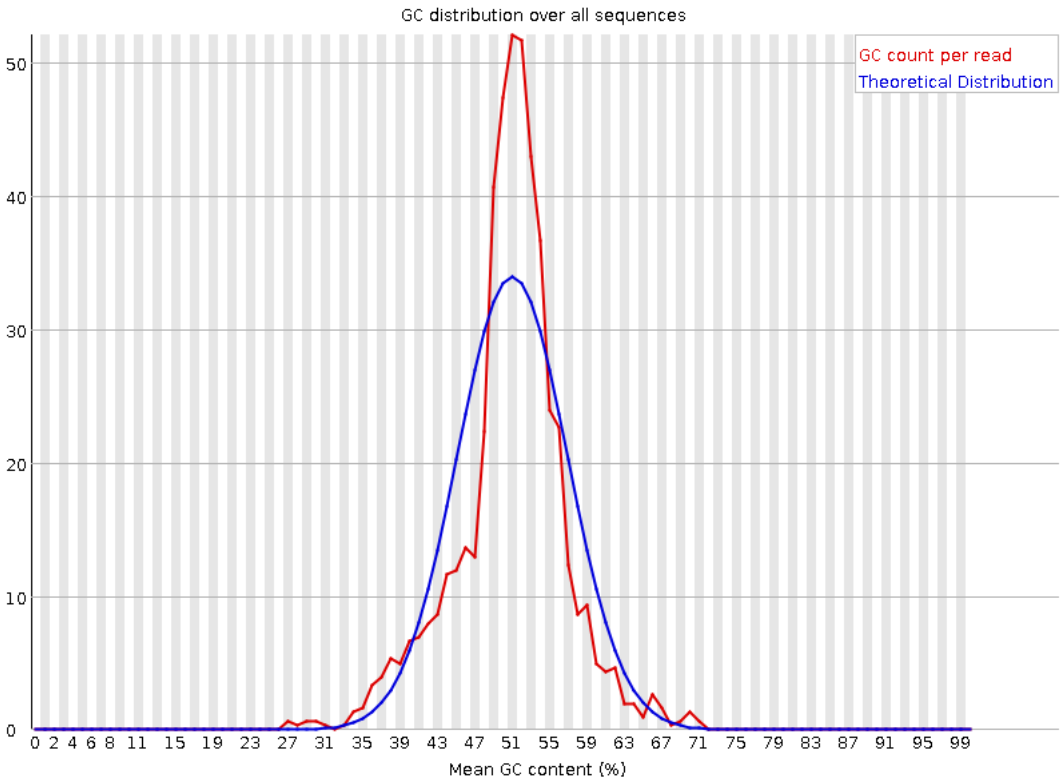


Obr. 4.1: FastQC - per base quality

Dalším výstupem FastQC je hodnocení zastoupení nukleotidů G (guanin) a C (cytosin) v DNA sekvenci, které ovlivňuje některé faktory, jako jsou například stabilita a teplota, při které DNA denaturuje. Obvykle se předpokládá, že bude nasekvenovaná DNA odpovídat normálnímu rozložení, je-li však distribuce neobvykle tvarovaná může to znamenat určité zkreslení nebo kontaminaci při přípravě knihovny. To je například vidět na obrázku 4.2, kde je zastoupení GC nukleotidů v sekvenci výrazně vyšší, než je předpokládaná hodnota.

Dále pak nástroj FastQC nabízí:

- graf zastoupení N v sekvenci (představuje procento pozic, u kterých nelze bázi s přesností určit)
- graf rozložení délek sekvencí (je potřeba vzít v úvahu v průběhu případného dalšího zpracování)
- graf úrovně duplikací (udává počet opakovaných sekvencí, pokud je duplikace vysoká může se jednat o kontaminaci vzorku)

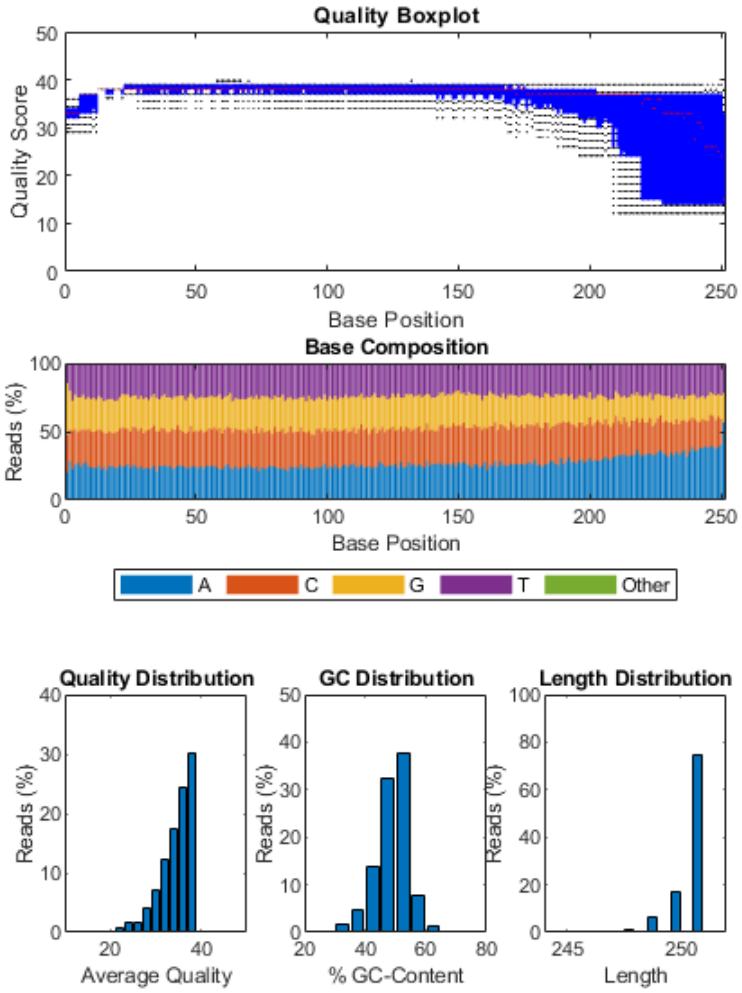


Obr. 4.2: FastQC - per sequence GC content

Jiný způsob, jak do jisté míry graficky zkontrolovat kvalitu dat, přináší i programové prostředí MATLAB, kde se dá využít funkce seqqcplot(), která generuje:

- graf průměrné kvality každé pozice sekvence
- graf zastoupení bazí na každé pozici sekvence
- histogram distribuce průměrné kvality sekvence
- histogram distribuce GC
- histogram délky sekvencí

Příklad výstupu této funkce je na obrázku 4.3, kde byla touto funkcí prověřena data SRX1798946 pro čtení 566001 až 567000. Tyto grafy nejsou tak moc přehledné jako ty z FastQC toolboxu, ale stále představují možnost, jak se na data podívat. Zhodnocení kvality dat je nezbytně nutné a slouží pro další zpracování, jako je například odstranění nekvalitních bazí, nebo jejich menší váha ve vytváření konsenzuální sekvence. Příklad možnosti zpracování dat je uveden v kapitole 5.3



Obr. 4.3: Hodnocení kvality - příklad výstupu funkce seqqcplot()

4.5 BLAST

Basic Local Alignment Search Tool zkrázeně BLAST slouží k hledání oblastí lokálních podobností mezi dvěma sekvencemi. Porovnává nukleotidové nebo proteinové sekvence se sekvencemi v databázi a počítá statistický význam nalezených shod. Nalezené podobnosti mohou mnohdy podat první informace o nově získané DNA sekvenci.

BLAST využívá heuristického přístupu, což umožnuje tak rychlé získání výsledků. Heuristický přístup ve zkratce znamená, že porovnávanou sekvenci rozbije na několik kratších úseků a pak až hledá shodu. Zároveň počítá takzvanou e-value (očekávaná hodnota), která určuje kolik shod by se vyskytlo při daném skóre. To umožňuje uživateli posoudit jak velkou důvěru k zarovnání mít.

4.6 BWA

Burrows-Wheeler Aligner (BWA) je program sloužící k zarovnávání poměrně krátkých nukleotidových sekvencí (například sekvenačních čtení) k dlouhým referenčním sekvencím (například lidskému genomu, nebo v tomto případě ke genomu *Treponema pallidum* subsp. *pallidum*). Skládá se ze tří programů:

- BWA-Backtrack;
- BWA-SW;
- BWA/MEM;

První algoritmus je určen pro zpracování sekvenačních dat, získaných pomocí Illumina sekvenátoru, pro ready kratší než 100 párů bazí. Zatímco druhé dva jsou určeny pro delší ready (od 70 bp po 1 Mbp). BWA-MEM a BWA-SW jsou podobné a oba zvládnou zpracovat i dlouhá čtení. BWA-MEM (nejnovější z algoritmů) je ale doporučován, protože je rychlejší a přesnější, používá se i pro Illumina ready s délkou kolem 70 bp -100 bp.[49] BWA vrací namapovaná data ve formě .sam souborů. Tento textový formát umožňuje ukládat zarovnané sekvence, včetně popisu jejich kvality a pozice v referenční sekvenci, na kterou se čtení mapovalo.

Tento program je opět určen pro linux, návod na jeho instalaci se dá najít nejen na githubu, ale je dostupný i formou videí na youtube.

5 Příprava algoritmů

Pro přípravu algoritmů posloužila data stažená z PubMLST databáze, jak celé sekvence, tak sekvence genů TP0136, TP0548 a TP0705. A také čtení SRX1798946 stažená z SRA databáze.

5.1 Určení alely

V programovém prostředí MATLAB byl navržen jednoduchý postup pro určení alely daného genu, využívající lokálního zarovnání pomocí funkce swalign(), do které vstupují sekvence, které chceme zarovnat a která vrací skóre nejlepšího lokálního zarovnání.

```
function [alela136] = ml136(sekvence)
alelyT136 = fastaread('alely/TP0136.fas');
al = [];
sc = [];
for i = 1:length(alelyT136)
    al = alelyT136(i).Sequence;
    Align = swalign(al,sekvence,'Alphabeth','NT');
    sc = [sc, Align];
end
[~,a] = max(sc);

% osetreni chyb vzniklych nepritomnosti dvou alel (8 a 38)
if a > 7
    a =a+1;
    if a>=38
        a = a+1;
    end
end

alela136 = a;
```

Sekvence stažené s PubMLST databáze mají známé alelické profily a byly tedy použity k ověření správnosti přístupu. Na základě známých lokusů daných genů v řetězci byly vybrány sekvence genu a jeho blízkého okolí, jak z testovaných sekvencí, tak ze vzorových sekvencí (Nichols a SS14). Tyto subsekvence byly lokálně zarovnány s alelickými sekvencemi a jako finální alela byla určena ta, která měla se vzorkem nejvyšší skóre. Tento jednoduchý přístup se ukázal být poměrně efektivním. Pro

lokus TP0136 pracovala funkce správně v 70% případů, problém představovaly alely 11, 12 a 8, 9 při jejichž určování docházelo k záměnám.

V případě záměny alely číslo 9 za alelu číslo 8 se jednalo o banální chybu. Alela číslo 8 v souboru alel pro gen TP0136 není, na jejím místě se nachází alela číslo 9. Tato chyba byla v kódu funkce ošetřena jednoduchým if() cyklem. O stejnou chybu se jednalo i v případě záměny alely 12 za 11. Alela číslo 12 je uložena v datasetu na místě 11.

Pro lokus TP0548 docházelo k záměnám v alelách 22, 23, 24 a 56, 57. Což bylo opět způsobeno nepřítomností alely číslo 21 v datasetu. Do funkce pro stanovení alely genu TP0548 byl proto přidán obdobný if() cyklus jako tomu bylo u funkce pro gen TP0136. Pro lokus genu TP0705, který se nachází na komplementním vlákně pracoval program správně. Ukázka zdrojového kódu obsahuje funkci sloužící k určení alely genu TP0136. Obdobně pak vypadají funkce pro ostatní geny.

5.2 Mapování dat

V semestrální práci na toto téma byla všechna čtení SRX1798946, stažená z SRA databáze, nejprve namapována na referenční sekvenci s využitím Burrows-Wheeler Aligner a až poté bylo určováno které z dostupných čtení se namapovalo na referenční genom. Tento přístup se ukázal být velmi výpočetně náročný – šlo o zpracovávání kapacitně velkých dat, která byla náročná, jak co se týče načtení, tak následného zpracování. Postupné procházení jednotlivých pozic mapování se ukázalo jako nemožné, z důvodu nedostatečné výpočetní kapacity použitého počítače.

Tento přístup byl v této práci přehodnocen. Data byla nejprve srovnána s referenční sekvencí daného genu za pomoci BLAST a teprve potom dále zpracovávána. Což se ukázalo být jednodušší a rychlejší co se programů týče. Samozřejmě tento přístup je nereálný v momentě, kdy je potřeba zpracovat vetší množství dat z SRA databáze.

5.3 Zhodnocení kvality dat

Jednotlivá čtení nalezená pomocí BLAST byla zhodnocena s využitím FastQC toolboxu. Čtení byla nejprve pomocí programu MATLAB uložena do společného fastq souboru a teprve potom použita k hodnocení.

```
talianova@gecko:~$ fastqc tp136_5.fastq
Started analysis of tp136_5.fastq
Analysis complete for tp136_5.fastq
```

Ukázka kódu 5.3 obsahuje příkaz fastqc a odpověď programu. Výsledek běhu programu je uložen v archivu zip do původní složky, není-li specifikováno jinak. Po rozbalení archivu lze získat obrázky grafů uváděných v kapitole 4.4.

Zároveň proběhla i kontrola mapovaných čtení získaných pomocí BWA. V získaném .sam souboru byla zkontrolována hodnota mapovací kvality, takzvané MAPQ skóre, které udává s jakou jistotou bylo čtení správně mapováno na danou pozici. Toto skóre může dosahovat hodnot 0-255 (255 znamená že MAPQ není dostupné), ale obvykle je většinou mapovacích nástrojů (včetně BWA) používáno skóre v rozsahu 0-60. Mapovaná čtení měla bez výjimek hodnotu MAPQ = 60, což indikovalo, že bylo čtení mapováno na správné místo (určené v souboru jako possition) referenční sekvence. Důležitou roli hrála také proměnná flag, která v .sam souboru reprezentuje různé informace o daném čtení. Jedná se o bitovou hodnotu, kde každý bit reprezentuje jinou vlastnost. V tomto případě měla čtení bez výjimek hodnotu flag = 0/16. Hodnota 0 znamená, že bylo čtení mapováno přímo na referenční sekvenci. Hodnota 16 znamená, že bylo čtení mapováno na reverzní vlákno (komplementární) a proto je nutné ho před dalším zpracováním převést na reverzní komplement.

5.4 Práce s geny TP0136 a TP0548

Z NCBI byly stažen genom kmene SS14, který sloužil jako reference. Z dostupných materiálů bylo vyhodnoceno, že icidence tohoto kmene napříč populací je vyšší. Na NCBI jsou uvedeny místa obou genů, díky nim byly geny indexovány v referenční sekvenci a uloženy jako reference daných genů. Tyto sekvence byly v BLASTu porovnány se všemi čteními z SRX1798946 v SRA databázi. Pro každý z genů bylo vybráno pět čtení s nejvyšší shodou.

SRA databáze umožňuje stahovat jednotlivá čtení přímo, pouze s využitím krátkých nukleotidových sekvencí, nebo pomocí ID které se v daném .sra souboru nachází. Bohužel velikost dat pravděpodobně znemožnila nalezení určeného čtení, takže bylo přistoupeno k použití SRA-toolboxu a stažení všech readů. Tento soubor byl nejprve rozdělen do více menších souborů, tak jak je popsáno v kapitole 4.1.

Nalezená čtení byla uložena do společného souboru. Byla provedena kontrola kvality čtení pomocí FastQC (obrázky uvedeny v kapitole 6). Případné nekvalitní konce sekvencí byly odstraněny. Následovalo mapování readů na referenční sekvenci SS14. Mapování probíhalo pro všechny geny na celou tuto referenční sekvenci, což sloužilo jako zpětná kontrola. Startovní pozice namapovaných čtení ležela v rozmezí daného genu. Teprve potom probíhalo určování konsenzuální sekvence (podle postupu uvedeném v kapitole 5.6).

5.5 Práce s genem TP0705

Práce s tímto genem byla poněkud náročnější. Nachází se totiž na komplementárním řetězci DNA a ten mnohdy není sekvenován. Ale jeho nalezení je i tak možné.

Získání reference genu TP0705 je možné dvěma způsoby. První možností je využití funkce seqrcomplement(), která v prostředí MATLAB slouží k získání reverzního komplementu. A následná indexace pozice genu. Druhá varianta je převedení indexů genu. Komplementární vlákno je rovněž číslováno ve směru od 5' ke 3' konci, což ale znamená, že se komplementární sekvence genů nachází ve forward sekvenci na pozici délka_sekvence - pozice_genu. Pro první variantu je na konci nutné ještě převést získanou sekvenci na reverzní komplement.

Referenční sekvence genu byla poté použita k nalezení čtení z SRX1798946 pomocí BLAST. Prvních pět čtení s nejvyšší shodou bylo použito k dalšímu zpracování za účelem nalezení konsenzu.

Tato čtení byla nejprve nalezena v souboru .fastq získaném pomocí SRA-toolboxu a následně uložena do společného .fastq souboru, který byl dále používán. Byla zkontrolována kvalita čtení a nekvalitní konce sekvencí byly ze souboru odstraněny. Mapovaná čtení byla použita k získání konsenzuální sekvence, která byla poté převedena na komplement funkcí seqcomplement(), která slouží k získání komplementu DNA (nebo i protinové) sekvence, a teprve poté došlo k určování alely.

5.6 Získání konsenzuální sekvence

Zpracovávání sekvencí probíhalo opět v prostředí MATLAB. Uložené .sam soubory (mapování) byly načteny a posloužily k získání sekvencí a pozice počátku zarovnání. Počátky se nacházely na pozici daného genu (TP0136: 157993-159471, TP0548: 593141-594457, TP0705: 772313-774967), bylo tedy nutné odečíst od nich počáteční místo genu. Zároveň bylo potřeba ověřit jestli se dané čtení mapovalo dopředně nebo na vedlejší řetězec, proto byla dále používána proměnná flags, získaná z .sam souboru mapování.

Upravené začátky zarovnání, sekvence čtení a flags byly použity jako vstup pro vytvořenou funkci consensus(), která sekvence uložené v buňkovém poli nejprve zarovná podle počáteční pozice a doplní je na stejnou délku (znakem: '-'), v případě flag 16 provede před doplněním převod na reverzní komplement a následně tvoří konsenzuální sekvenci podle následujícího popisu.

Prochází sekvence od počátku do maximální délky sekvence (tím, že jsou sekvence předpřipraveny podle počáteční pozice, délka stoupne z původních 250 bp - respektive z 230 bp po případném upravení podle kvality čtení). Znaky sekvencí na pozici i uloží do vektoru bases, který následně prochází dalším for cyklem. Určí

o jaký znak se jedná a přičte ho k hodnotě znaku (A_count, C_count, G_count, T_count), který je na začátku cyklu nastaven na hodnotu = 0. Tyto hodnoty nakonec uloží do vektoru se kterým se dále pracuje.

```
counts = [A_count,C_count,G_count,T_count];
    max_value = max(counts);

    if max_value==0
        consensus_base = '-';
    else
        possit = find(counts==max_value);
        if length(possit)==1
            consensus_base = out_base(possit);
        elseif length(possit)==2
            if (possit == [1 2])
                consensus_base='M';
            elseif (possit == [1 3]
                consensus_base='R';
            elseif (possit == [1 4]
                consensus_base='W';
            elseif (possit == [2 3]
                consensus_base='S';
            elseif (possit == [2 4]
                consensus_base='Y';
            elseif (possit == [3 4]
                consensus_base='K';
            end
        elseif length(possit)==3
            if (possit == [1 2 3])
                consensus_base='V';
            elseif (possit == [1 2 4]
                consensus_base='H';
            elseif (possit == [1 3 4]
                consensus_base='D';
            elseif (possit == [2 3 4]
                consensus_base='B';
        end
    end
```

Určuje se maximální hodnota ve vektoru, pokud je maximum = 0, znamená to

že je ve všech sekvencích na místě i znak '-' a do konsenzuální sekvence se zapisuje znak '-'. Pokud se maximum liší od 0, určí se pozice maxima. Pokud je vektor pozic (possis) dlouhý právě jeden prvek, znamená to že v sekvencích na pozici i převažuje jedna báze. Pozice uložená ve vektoru pozic possis určuje pozici báze ve vektoru možných bazí out_base. Ta je uložena nakonec konsenzuální sekvence (například je-li pozice maxima = 2, pak je jako cosensus_base uložena druhá báze z vektoru out_base = ['A','C','G','T'] → 'C') V případě, že je vektor pozic delší než jedna, znamená to že na pozici i v sekvencích převažuje více bází (například 2x 'A' a 2x 'G'). I tato varianta je ošetřena sérií cyklů if(). Pro tyto případy slouží takzvané UIPAC kódování, které zavádí společné znaky pro více bází viz obrázek 5.1.

```
W = A or T          ('Weak' base pairing)
S = C or G          ('Strong' base pairing)
R = A or G          (Purine)
Y = C or T          (Pyrimidine)
K = G or T          (Keto group on base)
M = A or C          (Amino group on base)
B = C, G, or T      (Not A)
D = A, G, or T      (Not C)
H = A, C, or T      (Not G)
V = A, C, or G      (Not T)
N = A, C, G, or T   (Any base)
```

Obr. 5.1: UIPAC kód [51]

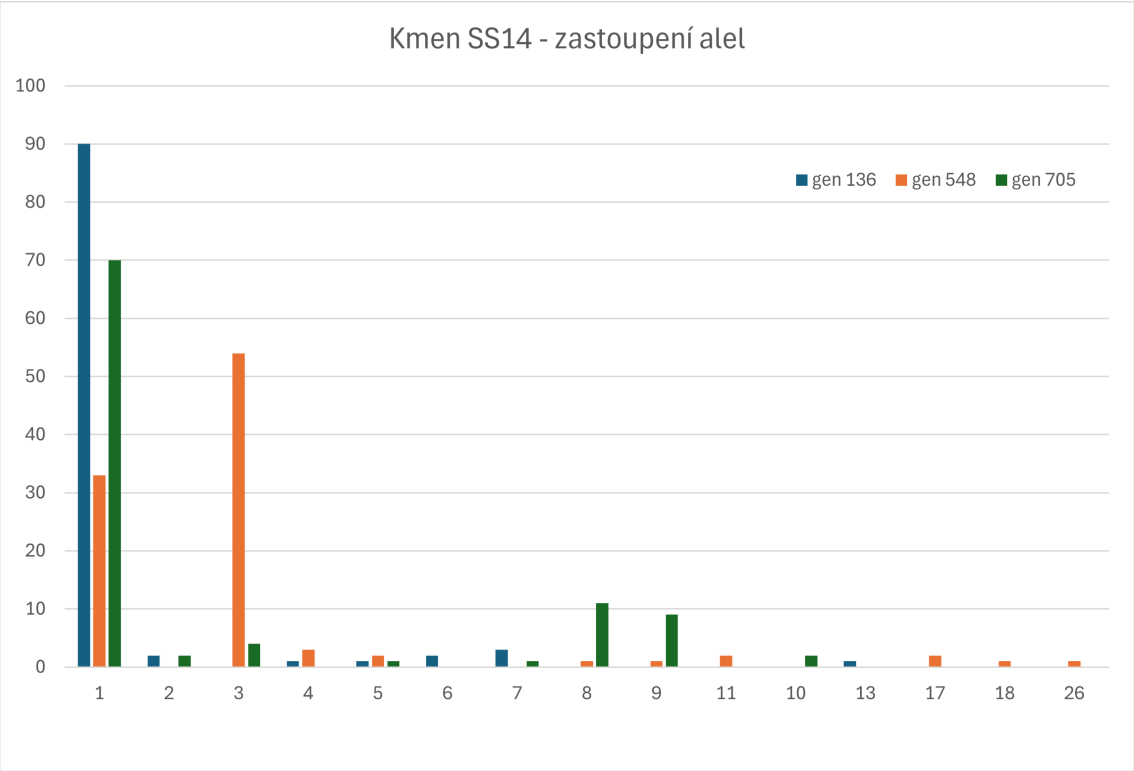
Pokud se tedy pozice maxima nacházely v prvním a druhém znaku byla jako konsenzuální báze určena báze 'M', pokud v prvním a třetím znaku báze 'R' a podobně pro všechny další možnosti.

Výstupem funkce je tedy konsenzus ve formě stringu dlouhý jako maximální délka upravených sekvencí. Ve funkci je přidána i možnost automatického uložení stringu do .fasta souboru s názvem Consensus.fasta, která je zakomentována a momentálně se nepoužívá, ale díky tomu lze funkci snadno modifikovat.

5.7 Určení kmene

V této práci bylo využito poznatku z článku [37], kde je uvedeno, že geny TP0136 a TP0705 rozlišují mezi oběma kmeny.

Bylo využito dat v pubMLST databázi k získání přehledů o alelickém rozložení typovaných genů. V grafu 5.2 je zobrazeno procentuální zastoupení alel genů pro kmen SS14. Je zde vidět, že pro gen TP0136 převládá alela číslo 1, která se ve sto

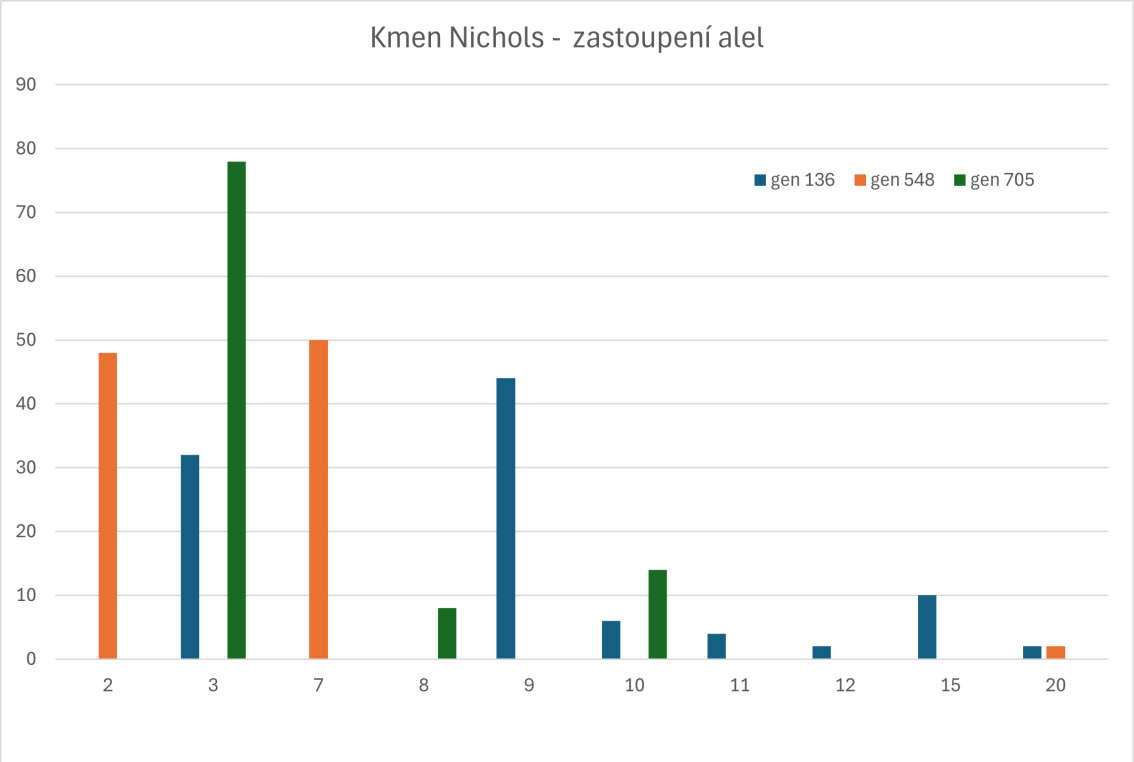


Obr. 5.2: Alely genů TP0136, TP0548 a TP0705 kmene SS14

vzorcích objevila v 90 případech. U genu TP0548 převládá alela číslo 3, která se objevila v 54, ale zde není převaha tak jasná. Pro gen TP0705 opět převládá alela číslo 1, která se vyskytla celkem 70x.

Graf 5.3 slouží k zobrazení alelického rozložení genů kmene Nichols. V genu TP0136 byla pozorována větší variabilita, z padesáti vzorků bylo 22 alely číslo 9 a 16 alely číslo 3.

Gen TP0548 měl pro padesát vzorků zastoupené pouze tři alely - 2, 7 a 20. Stejně tomu bylo i pro gen TP0705, který měl zastoupeny pouze alely číslo 3, 8 a 10.



Obr. 5.3: Alely genů TP0136, TP0548 a TP0705 kmene Nichols

6 Využití navržených algoritmů

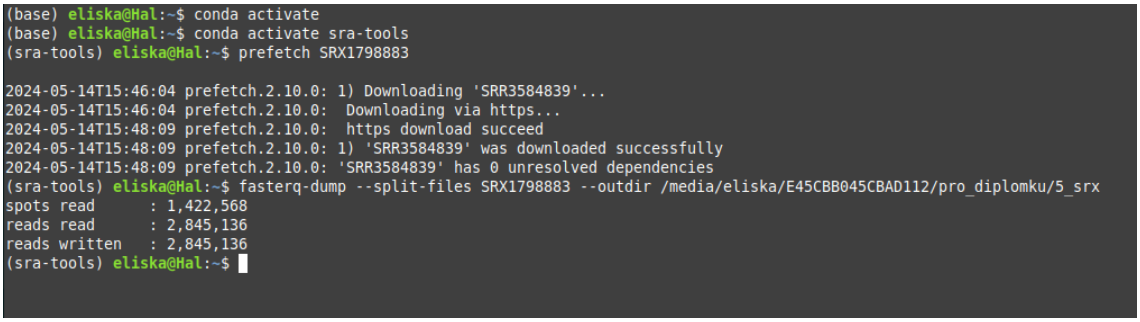
Předmětem této práce bylo vyvinutí postupu pro MLST typování *Treponema pallidum* subsp. *pallidum* významného organismu, jehož zkoumání umožňuje náhled do problematiky šíření a incidence nemoci sifilis.

K MLST typování se v dnešní době používá identifikace alelického profilu genů TP0136, TP548 a TP0705. V této práci byly využívány raw sekvenační data dostupná ve veřejné SRA databázi.

Navržené algoritmy byly použity pro MLST typování SRA dat:

- SRX1798946
- SRX1798900
- SRX1798896
- SRX1798886
- SRX1798883

Tato data byla nejprve stažena z SRA databáze pomocí sra-toolboxu a rozbalena z .sra archivu do fastq souborů.



Obr. 6.1: Příklad postupu stažení dat SRX1798883 z SRA databáze)

Postup stažení dat SRX1798883 z SRA databáze s využitím sra-tools je na obrázku 6.1, stažená data byla segmentována v programu MATLAB na menší části pomocí kódu uvedeného níže.

```
    %% načtení sra dat
%data 1
%data = fastqread('C:\pro_diplomku\5_sra\SRR3584965_1.fastq');
%index = 5;

    %% zpracování na menší soubory
delka = [1760, 1951, 1755, 1427, 1422];

for i = 1:delka(index)
```

```
    misto1 = i*1000 + 1;
    misto2 = misto1-1 + 1000;
    m = int2str(i+1);
    fastqwrite(strcat('C:\pro_diplomku\5_sra\3584_cut\',m,'.fastq'),
    data(misto1:misto2));

end
```

6.1 Data SRX1798946

Sloužila jako referenční data, na kterých byl testován postup, proto jsou v průběhu práce několikrát zmiňována. Zpracování dalších dat uváděných v následujících kapitolách probíhalo stejně.

Jedná se raw sekvenační data *Treponema pallidum* subsp. *pallidum* získaná ve studii [44]. Sekvenování proběhlo na přístroji Illumina MiSeq, celý soubor obsahuje něco přes 882 milionů bazí. Publikován byl 29.srpna 2016.

Data byla stažena z SRA databáze a uložena do menších souborů, jak už bylo zmiňováno. Byl využit BLAST a prvních 10 čtení bylo dále zpracováno. Naproti původní myšlence, kde bylo využíváno jen pět čtení, bylo využito deset z důvodu snahy o větší pokrytí lokusu a zároveň lepší kvalitu konsenzuální sekvence.

6.1.1 Kvalita dat

Čtení získaná po porovnání referenčních sekvencí s SRX1798946 pomocí BLAST, byla po úpravě zhodnocena pomocí nástroje FastQC.

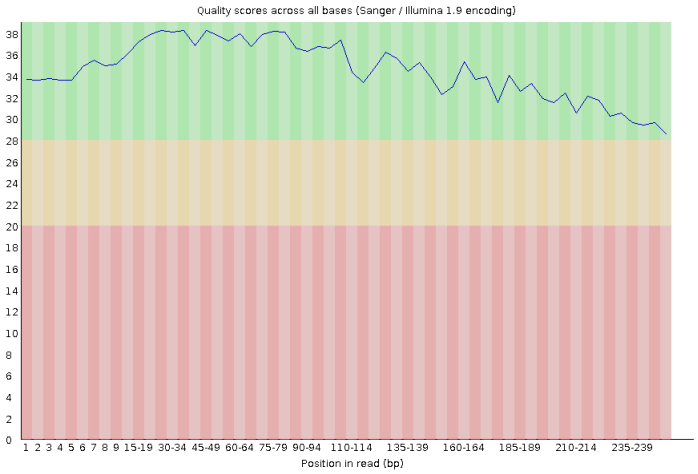
Pro gen TP0136 je výsledek analýzy kvality bazí na obrázku 6.2. Kvalita postupně klesá, ale stále zůstává dobrá.

Na obrázku 6.3 je kvalita bazí pro gen TP0548, zde kvalita čtení zůstává po celé délce dobrá.

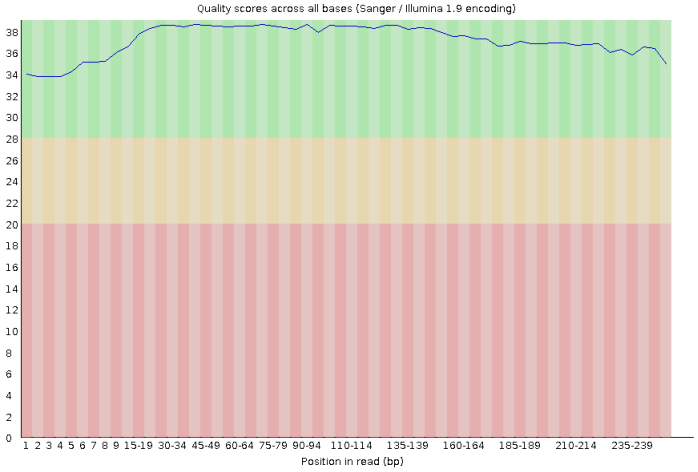
Podobně tomu je i u čtení genu TP0705, jak je vidět na obrázku 6.4. Data tedy nebylo nutné ořezávat a byla použita pro určení konsenzuální sekvence.

6.1.2 Určení alelického profilu

Mapovaná data byla zpracována v prostředí MATLAB tak jak je popisováno v kapitole 5.6. Byla provedena post-aligment kontrola, MAPQ skóre bylo pro všechna čtení všech tří genů = 60. Funkcí consensus() byly získány konsenzuální sekvence



Obr. 6.2: Kvalita čtení genu TP0136 (SRX1798946)

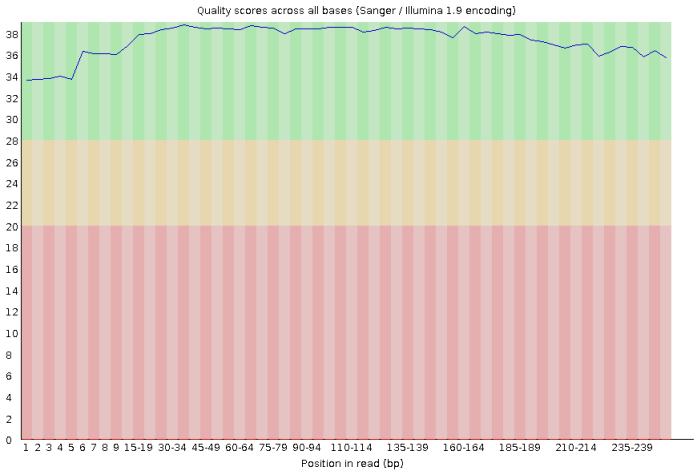


Obr. 6.3: Kvalita čtení genu TP0548 (SRX1798946)

pro všechny tři geny. Délka konsenzu se měnila na základě pozic mapování vybraných čtení. Pro gen TP0136 byla délka 1133 bp, pro gen TP0548 to bylo 1151 bp a pro gen TP0705 1890 bp.

Postupem uvedeným v kapitole 5.1 byla určena alela pro všechny tři geny z jejich konsenzuální sekvence. Pro gen TP0705 bylo potřeba nejprve získaný konsenzus převést na komplement funkcí seqcomplement(), která je součástí programu MATLAB.

Alelický profil dat SRX1798946 je uveden v tabulce 6.1. Na základě dostupných dat bylo určeno, že tento konkrétní vzorek patří do kmene SS14. Toto určení bylo konzultováno s pubMLST databází, která umožňuje typování na základě alelického



Obr. 6.4: Kvalita čtení genu TP0705 (SRX1798946)

profilu, po zadání čísel alel vypíše vzorky s nejvyšší podobností. V tomto případě pouze jeden, který patřil do kmene SS14.

6.2 Data SRX1798900

Vzorek byl odebrán v roce 2014 v Portugalsku, sekvenování proběhlo na přístoji Illumina MiSeq. Data byla zveřejněna v roce 2016 a obsahovala skoro 4 miliony čtení (979 Mb). Opět se jedná o raw sekvenační data ze studie [46].

Průměrná délka čtení byla opět 251 bp. Data byla stažena z SRA databáze za pomocí sra-toolboxu.

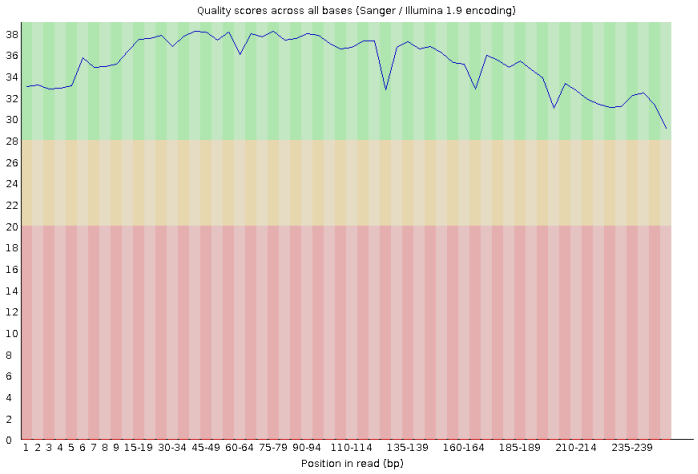
6.2.1 Kvalita dat

Čtení vyhodnocená za pomoci BLAST jako nejvíce podobná referenčním sekvencím byla uložena do souborů .fastq, pomocí programu MATLAB. Hodnocení kvality dat pak opět probíhalo pomocí nástroje FastQC.

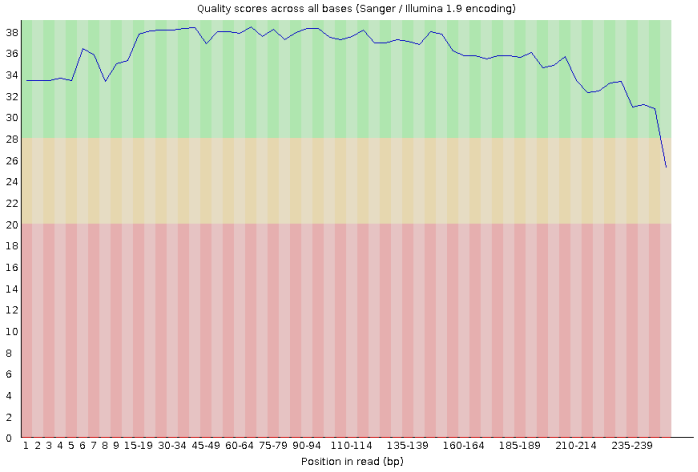
Na obrázku 6.5 je kvalita bazí napříč délkou čtení pro gen TP0136 tohoto vzorku. Kvalita kolísá, ale zůstává přijatelná. Obrázek 6.6 zobrazuje kvalitu čtení genu TP0548, zde kvalita klesá ke konci čtění, ale stále zůstává přijatelná. Dobrá kvalita dat je i u genu TP0705, ke kterému patří obrázek 6.7.

6.2.2 Určení alelického profilu

Díky přijatelné kvalitě dat bylo možné soubory dále zpracovávat bez úpravy. Čtení byla mapována pomocí BWA na referenční genom SS14 a uložena ve formě .sam



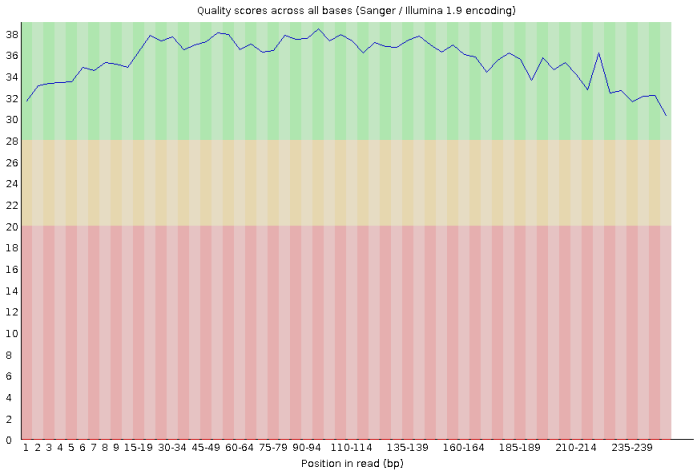
Obr. 6.5: Kvalita čtení genu TP0136 (SRX1798900)



Obr. 6.6: Kvalita čtení genu TP0548 (SRX1798900)

souborů. Ty byly načteny do programového prostředí MATLAB, kde bylo ověřeno, že byla čtení s dostatečnou kvalitou (MAPQ) mapována do pozice daného genu. Od počátku místa mapování byl odečten počátek genu (například pro gen TP0548: čtení mapováno na 593548 - počátek genu: 593141) a následně byla provedena tvorba konsenzuální sekvence. Délka konsenzu pro gen TP0136 byla 1195 bp, gen TP0548 měl 793 bp dlouhý konsenzus a gen TP0705 2064 bp.

Alelický profil dat SRX1798900 je opět dostupný v tabulce 6.1 na konci kapitoly. Na základě alelického profilu bylo určeno, že tento vzorek patří do kmene SS14. V pubMLST databázi byl opět nalezen jen jeden zástupce s tímto alelickým profilem, který náležel do téhož kmene.



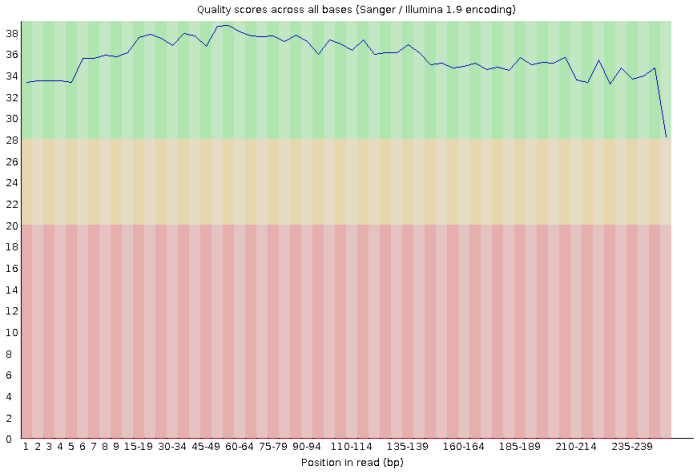
Obr. 6.7: Kvalita čtení genu TP0705 (SRX1798900)

6.3 Data SRX1798896

Data byla výsledkem celogenomového sekvenování pomocí nástroje Illumina MiSeq, zveřejněna v roce 2016 spolu s vydáním článku [46].

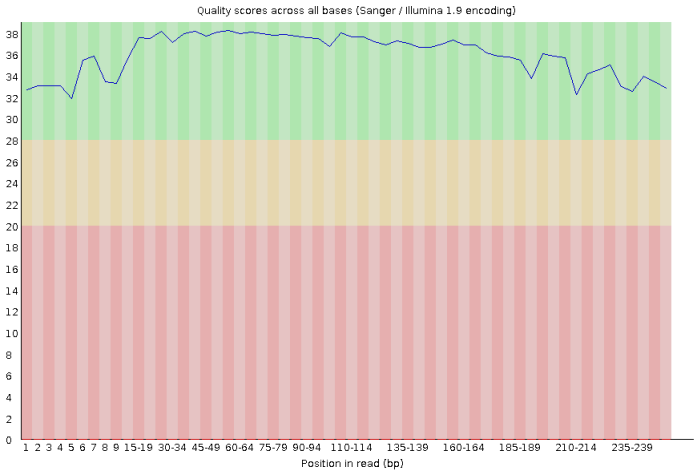
Jedná se o sekvenační data *Treponema pallidum* subsp. *pallidum* z klinických vzorků. Tato data obsahovala přes 3,5 milionů čtení. V SRA databázi jsou dostupná pod identifikátorem v názvu kapitoly, ale dají se najít i pod identifikátorem běhu - SRR3584879.

6.3.1 Kvalita dat

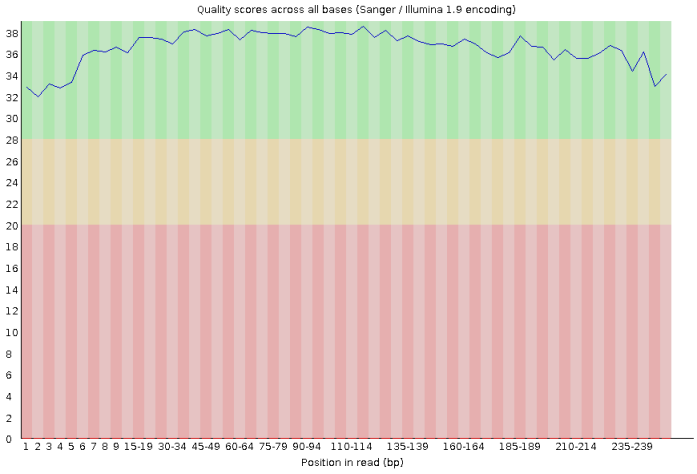


Obr. 6.8: Kvalita čtení genu TP0136 (SRX1798896)

Čtení nalezená za pomoci BLAST, byla vyhledána v předpřipravených (menších) souborech a uložena do formátu fastq, aby bylo možné hodnotit jejich kvalitu pomocí nástroje FastQC. Na obrázku 6.8 je kvalita bazí napříč délkou čtení pro gen TP0136 tohoto vzorku. Obrázek 6.6 zobrazuje kvalitu čtení genu TP0548. Stejně jako pro



Obr. 6.9: Kvalita čtení genu TP0548 (SRX1798896)



Obr. 6.10: Kvalita čtení genu TP0705 (SRX1798896)

předchozí dva geny je i pro gen TP0705 kvalita čtení dobrá po celé délce. Díky tomuto nebylo nutné daná čtení dále upravovat a bylo možné je použít pro mapování na referenční sekvenci a určení konsenzuální sekvence genů.

6.3.2 Alelický profil

Předpřipravená data byla mapována na referenční genom *Treponema pallidum* subsp. *pallidum* kmene SS14 pomocí BWA. V získaném souboru byly nalezeny indexy počátku mapování a bylo zkontrolováno, že leží v místě genu. Proběhla kontrola kvality mapování. Pomocí funkce consensus() byla získána konsenzuální sekvence všech genů z indexovaných čtení a byla použita pro určení alelického profilu dat, který je uveden v tabulce 6.1. Délky konsenzuální sekvence byly:

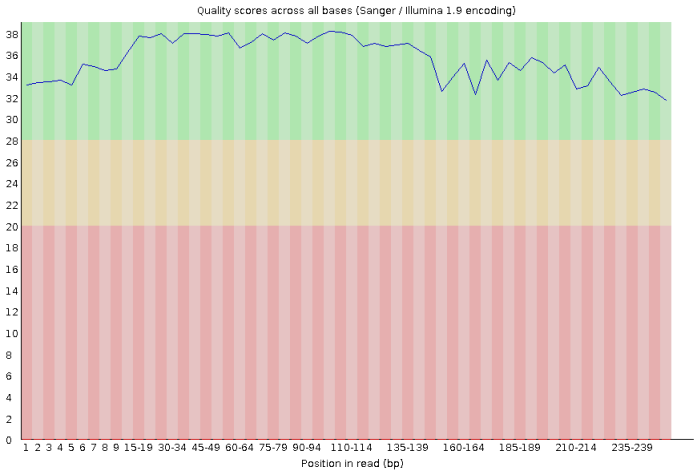
- pro gen TP0136: 1228 bp
- pro gen TP0548: 1128 bp
- pro gen TP0705: 1787 bp

6.4 Data SRX1798886

Jedná se o raw sekvenační data *Treponema pallidum* subsp. *pallidum* získaná sekvenováním klinicky získaného vzorku na platformě Illumina MiSeq. Data opět pochází z Portugalska a byla publikována společně se článkem [46].

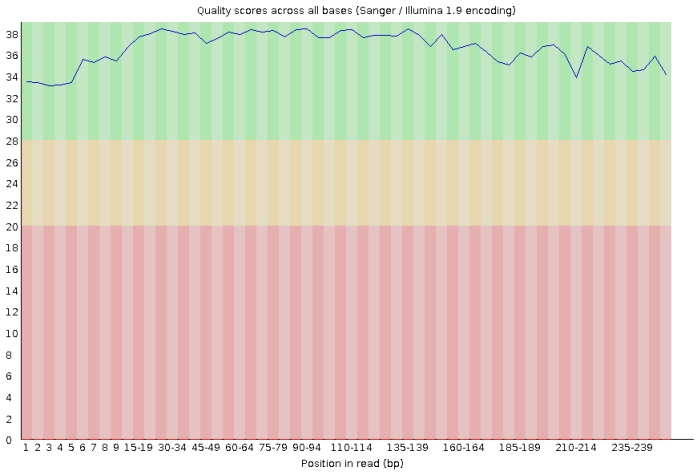
Data obsahovala přes 713 milionů bází, sekvenováno bylo 1,4 milionu míst. Celý dataset byl stažen z SRA databáze pomocí sra-tools.

6.4.1 Kvalita dat

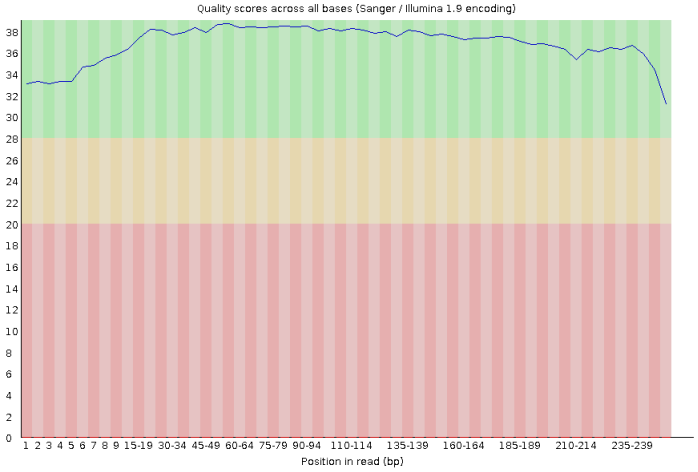


Obr. 6.11: Kvalita čtení genu TP0136 (SRX1798886)

Na obrázcích 6.11, 6.12 a 6.13 je znázorněna kvalita jednotlivých bází napříč délkou čtení genů TP0136, TP0548 a TP0705, získaná pomocí nástoje FastQC.



Obr. 6.12: Kvalita čtení genu TP0548 (SRX1798886)



Obr. 6.13: Kvalita čtení genu TP0705 (SRX1798886)

Kvalita čtení zůstává pro všechny tři případy v zelených hodnotách, takže nebylo potřeba čtení nějak ořezávat.

6.4.2 Alelický profil

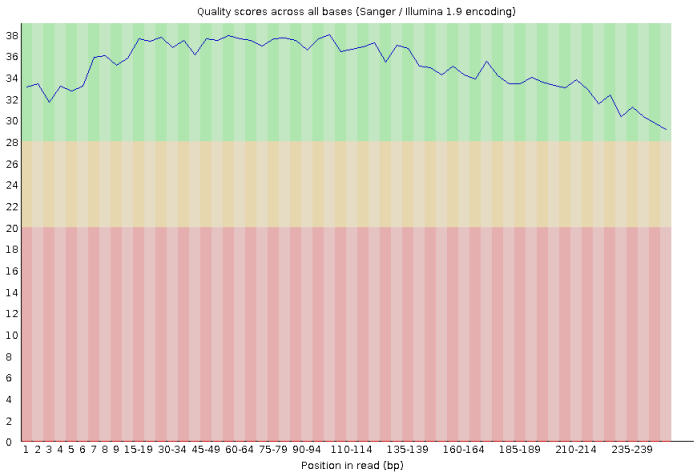
Po proběhnutí BWA mapování na referenční genom, byla stanovena konsenzuální sekvence všech tří genů. Konsenzus genu TP0136 byl dlouhý 1443 bp, genu TP0548 1061 bp a genu TP0705 1421 bp.

Alelický profil získaný na základě konsenzu jednotlivých genů je opět v tabulce 6.1.

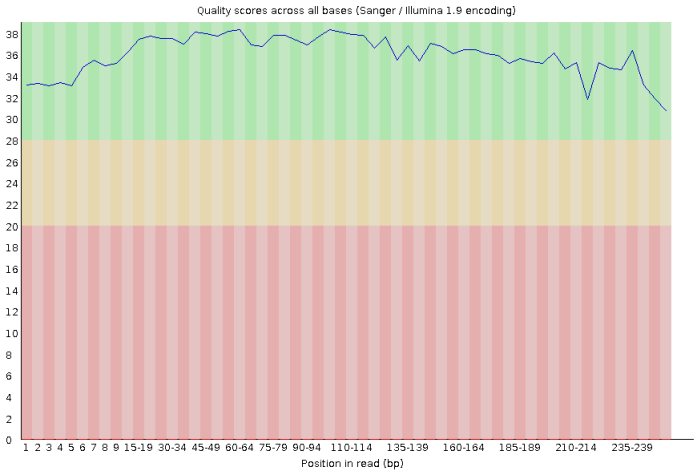
6.5 Data SRX1798883

Data pocházela ze stejné série dat z roku 2016, získané ve vědecké práci zmiňované ve článku [46]. Opět se jedná o raw data celogenomového sekvenování klinických vzorků *Treponema pallidum* subsp. *pallidum* nástrojem Illumina MiSeq. Dataset obsahoval přes 1,4 milionu sekvenovaných míst.

6.5.1 Kvalita dat

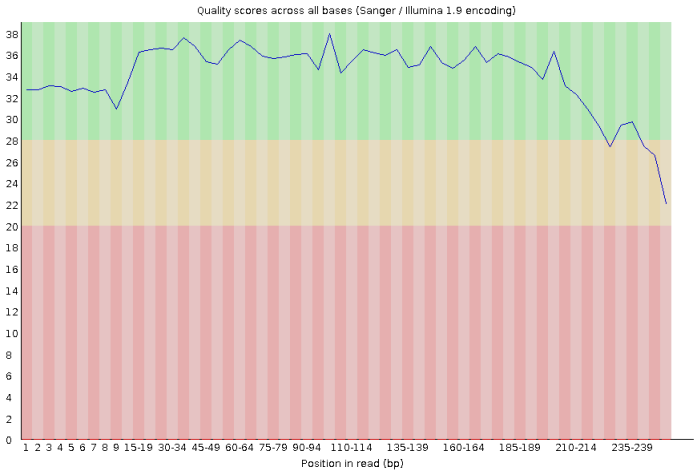


Obr. 6.14: Kvalita čtení genu TP0136 (SRX1798883)



Obr. 6.15: Kvalita čtení genu TP0548 (SRX1798883)

Na obrázcích 6.14 a 6.15 je hodnocení kvality přečtení báze napříč délkou čtení pro geny TP0136 a TP0548. Kvalita čtení těchto genů zůstává po celou délku dobrá.



Obr. 6.16: Kvalita čtení genu TP0705 (SRX1798883)

Pro gen TP0705, jehož kvalita čtení je na obrázku 6.16, klesá kvalita bází ke konci čtení k červeným hodnotám. Proto bylo přistoupeno k oříznutí nekvalitních konců čtení od 220 báze až do konce.

6.5.2 Alelický profil

Po úpravě čtení bylo přistoupeno ke tvorbě konsenzuálních sekvencí genů TP0136, TP0548 a TP0705. Na základě dat získaných pomocí BWA mapování na referenční sekvenci kmene SS14. Čtení byla bez výjimky mapována s dobrou mapovací kvalitou na oblasti daných lokusů. Konsenzuální sekvence měly délky 1443 bp, 1061 bp a 1821 bp. Alelický profil je opět v tabulce 6.1.

Tab. 6.1: Alelické profily dat

ID dat	gen TP0136	gen TP548	gen TP0705	Určený kmen
SRX1798946	1	1	13	SS14
SRX1798900	1	1	3	SS14
SRX1798896	1	1	1	SS14
SRX1798886	1	1	3	SS14
SRX1798883	1	1	1	SS14

6.6 Úprava algoritmu

V průběhu stanovování alelického profilu dat SRX1798883 bylo zjištěno, že předpřipravené funkce ml136(), ml548() a ml705() nejsou ošetřeny pro případ, že se více alel shoduje s konsenzuální sekvencí ve stejné míře. Problém nastával především v případě kdy byla čtení získaná pomocí BLAST mapována na vzájemně si blízké úseky genu. V takovém případě výsledná konsenzuální sekvence obsahovala poměrně dlouhé úseky mezer (znaku '-'), které mohly být zrovna v místě hledané alely, respektive v místě, kde se dvě různé alely od sebe lišily (například jednou substitucí). V případě, že bylo dosažené skóre lokálního zarovnání pro více alel shodné, vracela funkce na výstup první číslo alely s tímto skóre.

Proto byl daný postup trochu upraven a jako výstup sloužil vektor všech alel s nejvyšším skóre. Tabulka upravených alelických profilů 6.2 obsahuje všechny možnosti alel pro jednotlivé geny.

Tab. 6.2: Možné alely

ID dat	gen TP0136	gen TP548	gen TP0705
SRX1798946	1, 2, 5, 7, 14, 17, 21, 22, 25, 26, 27, 34, 35, 37, 39.	1	13
SRX1798900	1, 2, 5, 17, 25, 34, 37, 39, 40.	1, 3, 5, 11, 23, 29, 32, 36, 47, 48, 57, 65, 68, 72, 79, 82.	3, 6, 8, 9, 10, 12, 14, 15, 17, 18, 19.
SRX1798896	1, 4, 5, 13, 34, 40	1, 3, 5, 23, 29, 32, 36, 43, 47, 48, 68, 79, 82	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
SRX1798886	1, 4, 5, 13, 25, 37, 40	1, 3, 5, 11, 22, 23, 29, 32, 33, 36, 38, 43, 47, 48, 57, 65, 68, 72, 79, 82	3, 6, 8, 9, 10, 12, 14, 15, 17, 18, 19
SRX1798883	1, 2, 4, 5, 13, 17, 25, 37, 39, 40	1, 3, 5, 23, 32, 36, 47, 48, 68, 79, 82	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Zároveň byla provedena kontrola procentuálního pokrytí lokusů genů a podobnosti zarovnaných pozic pro globální zarovnání konsenzuální sekvence na referenční sekvenci genu. Hodnoty pro všechny vzorky jsou v tabulce 6.3. Nejlepší pokrytí bylo

pro gen TP0548. Nejhorší pokrytí bylo u vzorku SRX1798946 pro gen TP0136, kde kleslo pod 45 %. Podobnost konsenzuální sekvence genu s referenční sekvencí byla ve všech případech nad 80 %.

Tab. 6.3: Pokrytí a podobnost genů

	gen TP0136		gen TP548		gen TP0705	
ID dat	pokrytí	podobnost	pokrytí	podobnost	pokrytí	podobnost
SRX1798946	43,1 %	97,9 %	87,4 %	97.0 %	54,9 %	100 %
SRX1798900	65,2 %	85,1 %	60,2 %	97.7 %	63,3 %	89,2 %
SRX1798896	74,2 %	86,5 %	85,6 %	92.2 %	57,8 %	90,1 %
SRX1798886	69,1 %	99,2 %	67,4 %	84,7 %	56,6 %	94,2 %
SRX1798883	54,9 %	100 %	80,6 %	98,0 %	63,2 %	99,6 %

Vzorky SRX1798946, SRX1798900, SRX1798896, SRX1798886 a SRX1798883 jsou známé i pod označením kmene (PT_SIF...) a pod tímto identifikátorem jsou zaznamenané v PubMLST databázi. Jejich alelické profily jsou známé, byly do databáze přidány v srpnu roku 2018 paní Dr. L. Grillovou.

Tab. 6.4: Profil dat z PubMLST

ID dat	gen TP0136	gen TP548	gen TP0705	Kmen
PT_SIF0857	7	1	9	SS14
PT_SIF1299	1	3	1	SS14
PT_SIF1278	1	3	1	SS14
PT_SIF1183	1	1	1	SS14
PT_SIF1142	1	3	1	SS14

Alelický profil použitých dat, jak je dostupný v PubMLST databázi, je uveden v tabulce 6.4. Díky tomuto bylo možné porovnat alely určené použitím navrženého postupu s těmi, které byly stanoveny jako správné.

Pro vzorek SRX1798946 (PT_SIF0857) byla alela číslo 7 genu TP0136 mezi patnácti možnostmi stanovenými navrženým postupem. Gen TP0548 byl vyhodnocen správně. Pro gen TP0705 na druhou stranu program došel ke špatné alele. Pro zhodnocení podobnosti alel 9 a 13 byl využit nástroj Clustal Omega.

Na obrázku 6.17 je část výstupu z Clustal Omega pro porovnání alel číslo 9 a 13 genu TP0705. Na obrázku jsou patrné dvě odlišnosti - substituce A za G a G za A. Jinak jsou obě alely totožné. Celý výstup tohoto nástroje je k dispozici [52].

```
TP0705_9        GGTAGAACCGATTGCAGTGCGTTCAGTGGAGGATCGTTTAGGGCGGGTGATTTTGGATCC     360
TP0705_13       GGTAGAACCGATTGCAGTGCGTTCAGTGGAGGATCGTTTAGGGCGGGTGATTTTGGATCC     360
                ************************************************************

TP0705_9        AGAACGGGAAGTGCGGGCCCGCCTGCGCGCGCAGGGTGCGGCAACGCAACTGATCTCTGC     420
TP0705_13       AGAACGGGAAGTGCGGGCCCGCCTGCGCGCACAGGGTGCGGCAACGCAACTGATCTCTGC     420
                ***************************** ****************************

TP0705_9        GGAGAACGCGGCGCTCATGACGAATATGCTAGAGAAAACGGTAACGATGGGGACGTTGGC     480
TP0705_13       GGAGAACGCGGCGCTCATGACGAATGTGCTAGAGAAAACGGTAACGATGGGGACGTTGGC     480
                ************************ ***********************************

TP0705_9        GGTGGCCTCTGAGCGGGGGCGCGCATTTACATACCAAGACCCTGCAACGGGGCGATCGTT     540
TP0705_13       GGTGGCCTCTGAGCGGGGGCGCGCATTTACATACCAAGACCCTGCAACGGGGCGATCGTT     540
                ************************************************************
```

Obr. 6.17: Úsek porovnání alel 9 a 13 genu TP0705 nástrojem Clustal Omega

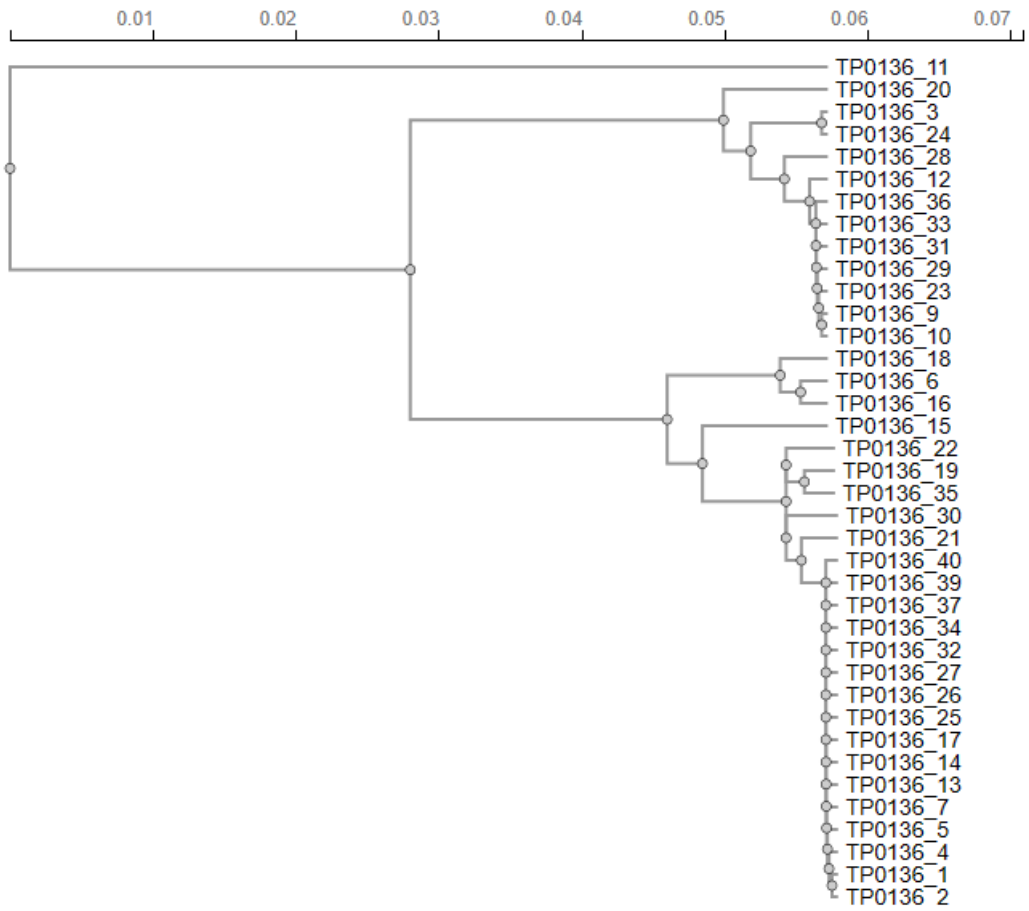
Pro vzorek SRX1798900 (PT_SIF1299) byly správné alely genu TP0136, TP0548 mezi devíti, respektive šestnácti možnostmi navrženými programem. U genu TP0705 nedošel program ke správnému řešení.

Pro vzorek SRX1798896 bylo správné řešení pro všechny tři geny mezi šesti, třinácti a dvaceti možnostmi. Zde je vidět, že algoritmus pro gen TP0705 měl poměrově největší problém, z 21 možných alel vrátil 20.

U vzorku SRX1798886 se program pro určení alely genu TP0705 opět nedostal ke správné variantě a místo toho navrhl jedenáct jiných možností. Správné alely genů TP0136 a TP0548 byly mezi sedmi a dvaceti možnostmi. U tohoto vzorku měl program pro určení alely genu TP0548 největší problém (nejnižší rozlišovací schopnost) určit z konsenzuální sekvence o jakou alelu se jedná. Procentuální podobnost konsenzuální sekvence tohoto genu s referenční sekvencí byla v tomto případě jen 84,7 %.

Vzorek SRX1798883 měl sice správné alely mezi výstupy programů, ale rozlišovací schopnost funkce ml705() pro určení alely genu 705 byla opět skoro nulová - program nabídl 20 z 21 možností.

Byla ověřena délka lokálního zarovnání alel s geny. Pro gen TP0136 byla délka skoro vždy větší než 830 bp. Samozřejmě s variabilní délkou mezer, která však nepřekročila 250 znaků. Naproti tomu u genu TP0705 byla délka lokálního zarovnání mnohonásobně nižší u vzorků 3 a 5, kde byla diskriminační schopnost navržené funkce ml705() velmi nízká, byla délka zarovnání jen 64 bp. Z toho samozřejmě vyplývá neschopnost programu určit o jakou alelu se jedná, neboť alely genu TP0705 se liší jen velmi málo. Což je patrné i na obrázku 6.20, kde je odlišnost jednotlivých alel genu naznačena formou fylogenetického stromu vygenerovaného nástrojem Clustal Omega.

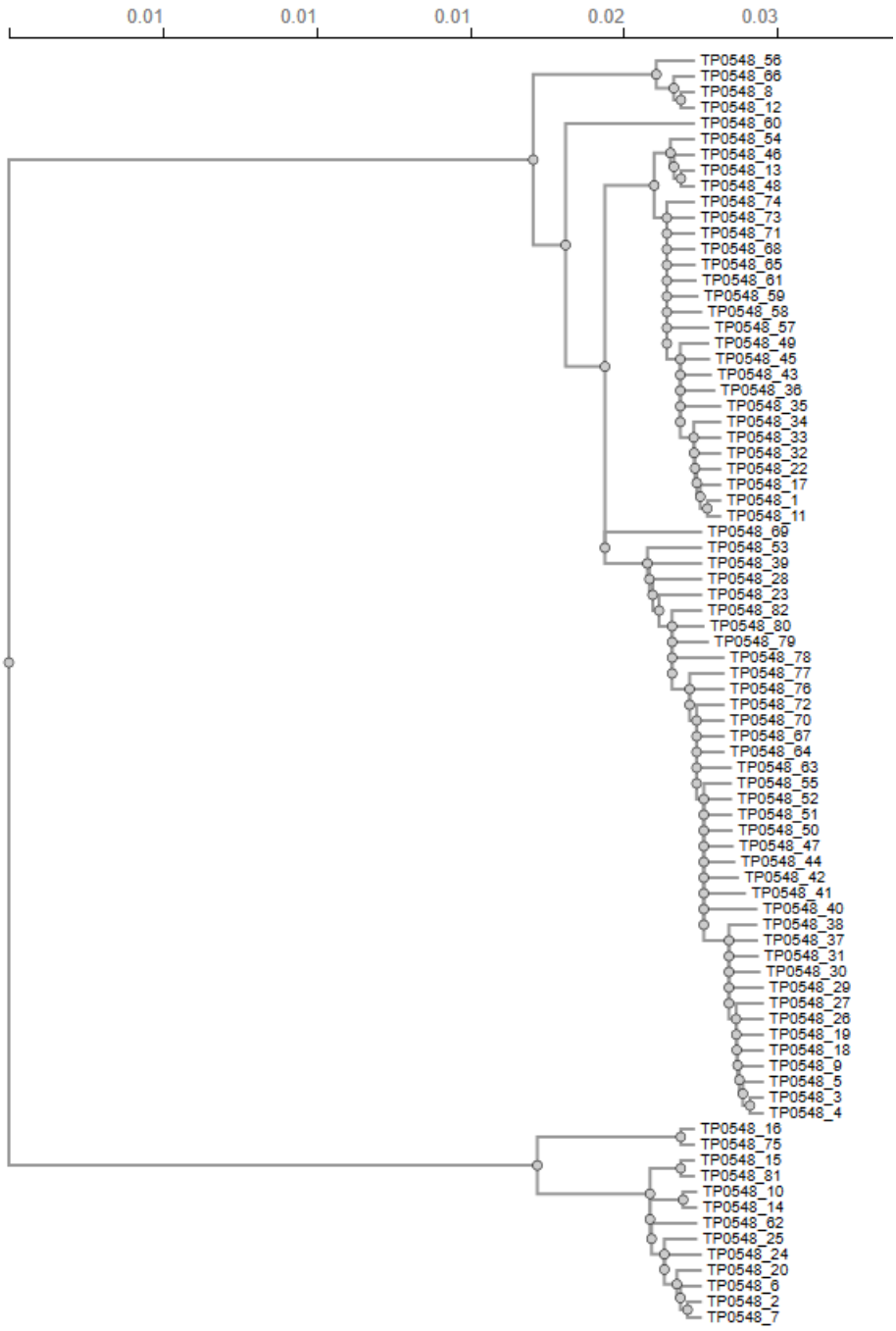


Obr. 6.18: Fylogenetický strom alel genu TP0136

Na obrázcích 6.18 a 6.19 jsou fylogenetické stromy alel dalších dvou genů. Jak je patrné, alely genu TP0705 jsou si velmi podobné, proto by bylo potřeba mnohem vyššího pokrytí k alespoň částečnému omezení možností. Alely genu TP0548 kterých je dohromady 81 by se daly rozdělit do dvou více odlišných skupin. V tomto případě je program schopný minimálně rozlišit o kterou z těchto dvou skupin se jedná. V případě alel genu TP0136, které jsou navzájem nejvíce odlišné, dochází program ke správnému zařazení do skupiny poměrně si podobných alel, které jsou na obrázku 6.18 dole.

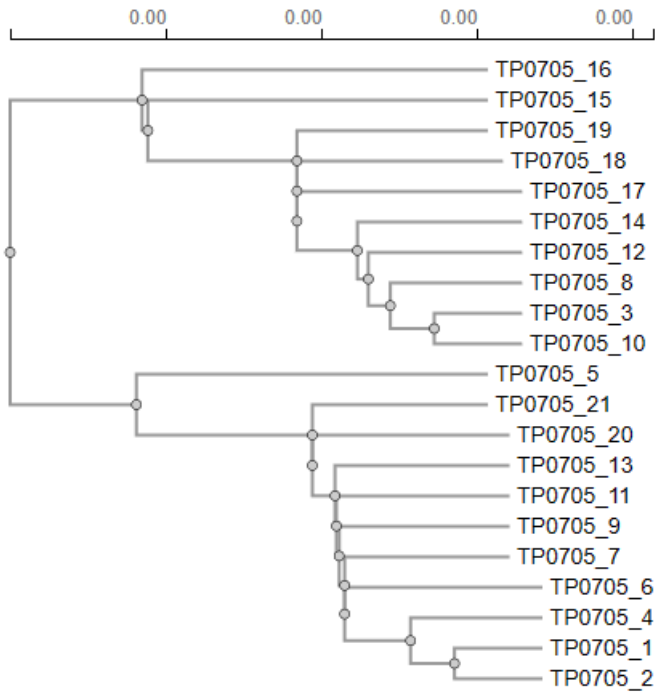
Celý výstup nástroje Clustal Omega, použitého pro analýzu těchto dat je dostupný: [53, 54, 55]

Na zakládě těchto informací bylo vyzkoušeno o kolik lepších výsledků by se dalo dosáhnout při použití 20 čtení pro každý gen. Jako testovací byla zvolena data z SRA databáze s ID SRX1793059. Jejich alelický profil byl vyhledán v PubMLST databázi - gen TP0136: alela číslo 7, gen TP0548: alela číslo 1 a gen TP0705: alela



Obr. 6.19: Fylogenetický strom alel genu TP0548

číslo 9. Zpracování dvaceti čtení probíhalo obdobným způsobem jako v předchozích případech. Mapovací kvalita byla opět = 60 pro všechna čtení, pro gen TP0136 mělo 12 čtení flag = 16 a bylo je proto nutné převést na reverzní komplement. Kvalita čtení bazí byla pro lokusy všech tří genů velmi dobrá - Q(A) nad 30. Pokrytí lokusu genu TP0136 bylo 66,9 % - žádné významné zvýšení, podobnost globálního zarovnání s



Obr. 6.20: Fylogenetický strom alel genu TP0705

referenční sekvencí byla 97,3 %. Algoritmus pro určení typu alely opět došel k deseti výsledkům - 1, 2, 4, 5, 7, 17, 25, 37, 39, 40. Konsenzuální sekvence o celkové délce 1311 bp obsahovala 321 bp dlouhou mezeru. Pokrytí genu TP0548 bylo na druhou stranu 99.9 %. Konsenzuální sekvence délky 1308 bp neobsahovala mezery a byla jen o 9 bp kratší než referenční sekvence tohoto lokusu. Podobnost byla 91,3 % a algoritmus pro určení alely genu TP0548 správně určil alelu číslo 1. Zde se díky vyššímu pokrytí lokusu podařilo dojít ke správné variantě. Pokud by se tedy pro ostatní dva geny podařilo vybrat vhodnější čtení, nebo jich bylo použito dostatek, vlivem vyššího pokrytí by rostla i diskriminační schopnost navrženého algoritmu. Pokrytí genu TP705 bylo 88 %, konsenzuální sekvence délky 2438 bp obsahovala dvě mezery s celkovou délkou 278 bp. Podobnost globálního zarovnání konsenzu s referencí byla 100 %. Algoritmus pro určení typu alely opět nefungoval, nabídl tři možnosti - 3, 5, 17.

Navržený přístup by bylo samozřejmě možné aplikovat i pro typování jiných genů. Stačilo by znát jejich polohu na referenční sekvenci. Referenční sekvence genu by byla použita pro vyhledání čtení, které spadají do oblasti zájmu. A dále by byl postup stejný. Stejným způsobem by se dala mapovat i sekvenační data kmene Nichols. Pokud by chyběla informace o tom o jaký kmen se jedná, bylo by možné

použít k zpracování referenční sekvence získané s využitím kmene SS14, jen by bylo třeba počítat se sníženou kvalitou takto získaných výsledků.

Závěr

Tato práce se zabývá problematikou typování *Treponema pallidum* subsp. *pallidum*. V první kapitole je čtenář seznámen s bakteriemi *Treponema pallidum* ze skupiny spirochet, s jejich strukturou a nemocemi, které tyto bakterie způsobují. *Treponema pallidum* subsp. *pallidum* jako zástupce této skupiny způsobuje nemoc sifilis. Incidence této nemoci stále roste, WHO udává, že v roce 2020 bylo nakažených 7,1 miliónů dospělých (mezi 15 a 49 lety). Proto začíná být nezbytně nutné bakterie typovat. Určovat do jaké skupiny patří, odkud pochází.

Pro tyto účely slouží typovací techniky. Těchto technik existuje řada. V minulosti se hojně používaly metody využívající různě modifikované verze gelové elektroforézy. Avšak díky pokroku v sekvenčních metodách, které jsou v této práci také rozebírány, bylo přistoupeno k technikám využívajících právě sekvenaci. Jednou z takovýchto technik je i MLST (Multilocus sequence typing - Multilokusové sekvenční typování), metoda stanovující typ sekvence, tedy druh kmene, na základě provozních genů. MLST typování se obvykle děje formou sekvenování vnitřních fragmentů dobře zachovávaných genů a následným určováním o jaký typ alely (z dostupného seznamu) se jedná. Určení typů alel pro všechny tyto geny vede k získání alelického profilu dané bakterie. Na jehož základě lze pak určit o jaký kmen bakterií se jedná. V případě MLST typování *Treponema pallidum* subsp. *pallidum* byl paní Dr. L. Grillovou a kolektivem autorů alelický profil sestaven ze tří genů - TP0136, TP0548 a TP0705. S jejich pomocí lze pak stanovit o který ze dvou hlavních kmenů (SS14 a Nichols) jde.

V této práci je proveden pokus o stanovení alelického profilu *Treponema pallidum* subsp. *pallidum* na základě veřejně dostupných raw sekvenačních dat. Proto je v této práci čtenář seznámen s databázemi a nástroji, které se v případě práce s těmito daty dají využít. Je zde zmíněna SRA databáze, ve které lze dohledat surová sekvenační data včetně informací k nim, možnosti ověření kvality dat a algoritmy využívané pro mapování čtení na referenční sekvenci.

Práce se dále snaží navrhnout postup pro zpracování takovýchto dat. Pro omezení výpočetní náročnosti využívá nástroj BLAST k určení čtení spadajících do oblasti zájmu, které jsou dále vyhledávány ve stažených souborech a zpracovávány. Pomocí nástroje FastQC je kontrolována kvalita přečtení báze napříč délkou čtení. Zároveň v programu zohledňuje variantu nízké kvality čtení (26 a méně) a takováto čtení dále zpracovává. K mapování na referenci využívá práce BWA (konkrétně algoritmus bwa-mem), mapovaná čtení slouží pro získání konsenzuální sekvence (s využitím UIPAC kódování), která je dále použitá pro určení typu alely daného genu (z datasetu získaného v PubMLST databázi).

Na začátku zpracování této práce byl vyvinut jednoduchý program sloužící pro

určení typu alely. Jako podklad posloužily data z PubMLST databáze - sekvence daných genů. V případě využití na tyto data pracoval program, využívající lokální zarovnání, velmi dobře. Bohužel v případě takových dat, jako jsou surová sekvenační data, začalo docházet k chybám a nepřesnostem. Konsenzuální sekvence skoro nikdy nezačíná na začátku lokusu daného genu, obsahuje mezery a nejednoznačné báze. Je možné, že hledaný úsek genu leží v místě, kde je v konsenzuální sekvenci vlivem nedostatečného pokrytí mezera. Jak se ukázalo dalším šetřením, deset čtení pro každý gen ani zdaleka nestačí pro dostatečné pokrytí lokusů. Dvojnásobné zvýšení počtu použitých čtení sice vede ke zvýšení pokrytí daných lokusů, ale konsenzuální sekvence stále obsahuje mezery, které správnému určení alely. Navržený postup pro určení typu alely sice více či méně omezuje varianty, ze kterých lze pro daný gen vybírat, ale není jednoznačný a občas je dokonce chybný. Nejhorší diskriminační schopnost má algoritmus pro určení alely genu TP0705. Což je samozřejmě dáno podobností referenčních alel, které se mnohdy liší jen velmi málo (například jen jednou substitucí).

Je možné, že pokud by byl navržený postup použit pro data získaná způsobem obvyklým pro tento přístup - sekvenace vnitřních fragmentů daných genů, byla by úspěšnost programu vyšší. Ale vše by opět záleželo hlavně na kvalitě získané konsenzuální sekvence (ať už jde o kvalitu čtení, či o dostatečné pokrytí lokusu). Bohužel pro takto zpracovávaná data je navržený postup určení alely velmi nevhodný. Pro určení typu alely by nejspíš mnohem lépe posloužila nějaká neuronová síť, natrénovaná na datech dostupných v PubMLST databázi.

Literatura

[1] University of Texas Medical Branch at Galveston: *Medical Microbiology. 4th edition.* [online]. Randolf D.J., Chapter 36, Treponema [cit. 15. 10. 2022]. Dostupné z:
<https://www.ncbi.nlm.nih.gov/books/NBK7716/#__NBK7716_dtls__>.

[2] Velký lékařský slovník: *Treponema pallidum* [online]. [cit. 15. 10. 2022]. Dostupné z:
<https://lekarske.slovniky.cz/pojem/treponema-pallidum>.

[3] Velký lékařský slovník: *mikroaerofilní* [online]. [cit. 15. 10. 2022]. Dostupné z:
<https://lekarske.slovniky.cz/pojem/mikroaerofilni>.

[4] National Institute of Allergy and Infectios Diseases *Syphilis* [online]. [cit. 16. 10. 2022]. Dostupné z:
<https://www.niaid.nih.gov/diseases-conditions/syphilis>.

[5] LIU J., HOWELL J.K., BRADLEY S.D., et al.: *Cellular Architecture of Treponema pallidum: Novel Flagellum, Periplasmic Cone, and Cell Envelope as Revealed by Cryo-Electron Tomography* [online]. Journal of molecular biology, 403(4), 546–561. Doi:
<https://doi.org/10.1016/j.jmb.2010.09.020>.

[6] EDMONDSON D.G., HU B., NORRIS S.J.: *Long-Term In Vitro Culture of the Syphilis Spirochete Treponema pallidum subsp. pallidum* [online]. Doi:
<https://doi.org/10.1128/mBio.01153-18>.

[7] Mayo clinic: *Syphilis* [online]. symptoms & causes [cit. 28. 10. 2022]. Dostupné z:
<https://www.mayoclinic.org/diseases-conditions/syphilis/symptoms-causes/syc-20351756>.

[8] POLÁČKOVÁ Z.: *Pohlavní choroby I. díl* [online]. Dermatologie pro praxi 2008, vol. 2, s. 74-76. [cit. 28. 10. 2022]. Dostupné z:
<http://www.solen.cz/pdfs/der/2008/02/06.pdf>.

[9] Centers for Disease Control and Prevention: *Sexually Transmitted Infections Treatment Guidelines, 2021* [online]. Primary and Secondary Syphilis [cit. 28. 12. 2022]. Dostupné z:
<https://www.cdc.gov/std/treatment-guidelines/p-and-s-syphilis.htm>.

[10] World Health Organisation: *Syphilis* [online]. [cit. 25. 05. 2024]. Dostupné z: <https://www.who.int/news-room/fact-sheets/detail/syphilis>.

[11] World Health Organisation: *Yaws*[online]. [cit. 29. 10. 2022]. Dostupné z: <https://www.who.int/news-room/fact-sheets/detail/yaws>.

[12] National Organization for Rare Disorders *Pinta*[online]. [cit. 2. 11. 2022]. Dostupné z: <https://rarediseases.org/rare-diseases/pinta/>.

[13] National Human Genome Research Institute *DNA SEQUENCING*[online]. [cit. 13. 3. 2024]. Dostupné z: <https://www.genome.gov/genetics-glossary>.

[14] BROWN TA. *Genomes. 2nd edition.*[online]. Oxford: Wiley-Liss; 2002. Chapter 6, Sequencing Genomes. [cit. 20. 3. 2024]. Dostupné z: <https://www.ncbi.nlm.nih.gov/books/NBK21117/>.

[15] LAB Guide průvodce laboratoří *Klonování*[online]. [cit. 21. 3. 2024]. Dostupné z: <https://labguide.cz/metody/klonovani/>.

[16] MOHAMMADI M.M., BAVI O. *DNA sequencing: an overview of solid-state and biological nanopore-based methods*[online]. Biophys Rev. 2021;14(1):99-110. Published 2021 Nov 23. [cit. 21. 3. 2024]. Doi: <https://doi.org/10.1007/s12551-021-00857-y>.

[17] ThermoFisher Scientific *DNA Sequencing Technologies–History and Overview*[online]. [cit. 21. 3. 2024]. Dostupné z: <https://www.thermofisher.com/cz/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/next-generation-sequencing/dna-sequencing-history.html>.

[18] MOBLEY I.: *DNA amplification techniques*[online]. Frontline Genomics. Dostupné z: <https://frontlinegenomics.com/dna-amplification-techniques>.

[19] ATDBio: *Sequencing, forensic analysis and genetic analysis*[online]. Nucleic Acids Book. Dostupné z: <https://atdbio.com/nucleic-acids-book/Next-generation-sequencing#Amplification>.

[20] SLATKO B.E., GARDNER A.F., AUSUBEL F.M.: *Overview of Next-Generation Sequencing Technologies*[online]. Current protocols in molecular biology, 122(1), e59. Doi: <https://doi.org/10.1002/cpmb.59>.

[21] PATRICK K.: *454 life sciences: Illiminating the future of genome sequencing and personalised medicine*[online]. The Yale journal of biology and medicine; 80(4): 191-194. Dostupné z: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2347365/#R1>.

[22] MIN G., SHI-HUI F.,LIFANG L., et al.: *Pyrosequencing enchancement for bette detection limit and sequencing homopolymers*[online]. Biochemical and Biophysical Research Communications, Volume 400, Issue 1, 2010, Pages 117-123. Doi: <https://doi.org/10.1016/j.bbrc.2010.09.022.>.

[23] GARRIDO-CARDENAS J.A, GARCIA-MAROTO F., et al.: *DNA Sequencing Sensors: An Overview*[online]. Sensors (Basel, Switzerland), 17(3), 588. Doi: <https://doi.org/10.3390/s17030588>.

[24] SALMELA L.: *Correction of sequencing errors in a mixed set of reads*[online]. Bioinformatics (Oxford, England), 26(10), 1284–1290. Doi: <https://doi.org/10.1093/bioinformatics/btq151>.

[25] ROSS S.T., SCHWARTZ S., et al.: *Total Internal Reflection Fluorescence (TIRF) Microscopy*[online]. MicroscopyU. Dostupné z: <https://www.microscopyu.com/techniques/fluorescence/total-internal-reflection-fluorescence-tirf-microscopy>.

[26] Illumina: *Sequencing platforms*[online].

Dostupné z: <https://emea.illumina.com/systems/sequencing-platforms.html>.

[27] MOHAMMADI M.M., BAVI O.: *DNA sequencing: an overview of solid-state and biological nanopore-based methods*[online]. Biophysical Reviews 14, 99–110 (2022) Doi: <https://doi.org/10.1007/s12551-021-00857-y>.

[28] LEE H., GURTOWSKI J., YOO S., et al.: *Third-generation sequencing and the future of genomics*[online]. Doi: <https://doi.org/10.1101/048603>.

[29] XIAO T., ZHOU W.: *The third generation sequencing: the advanced approach to genetic diseases*[online]. Translational pediatrics, 9(2), 163–173. Doi: <https://doi.org/10.21037/tp.2020.03.06>.

[30] ATHANASOPOULOU K., BOTI M.A., ADAMOPOULOS P.G., et al.: *Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics*[online]. Life (Basel, Switzerland), 12(1), 30. Doi: <https://doi.org/10.3390/life12010030>.

[31] Oxford Nanopore Technologies: *How nanopore sequencing works*[online]. Dostupné z: <https://nanoporetech.com/platform/technology>.

[32] SIMAR S.R., HANSON B.M., ARIAS C.A.: *Techniques in bacterial strain typing: past, present and future*[online]. Current opinion in infectious diseases, 34(4), 339–345. Doi: <https://doi.org/10.1097/QCO.0000000000000743>.

[33] DINGLE C.T., MACCANNELL R.D.: *Metods in Microbiology*[online]. Chapter 9 - Molecular Strain Typing and Characterisation of Toxigenic Clostridium defficile. Doi: <https://doi.org/10.1016/bs.mim.2015.07.001>.

[34] Public databases for molecular typing and microbial genome diversity: *Multi-Locus Sequence Typing*[online]. [cit. 8. 11. 2022]. Dostupné z: <https://pubmlst.org/multilocus-sequence-typing>.

[35] Biomérieux *Multilocus sequence typing (MLST) analysis*[online]. [cit. 12. 11. 2022]. Dostupné z: <https://www.applied-maths.com/applications/mlst>.

[36] MARRA C., SAHI S., TANTALO L., et al.: *Enhanced molecular typing of treponema pallidum: geographical distribution of strain types and association with neurosyphilis.*[online]. Doi: <https://doi.org/10.1086/656533>.

[37] GRILLOVÁ L., BAWA T., MIKALOVÁ L., et al.: *Molecular characterization of Treponema pallidum subsp. pallidum in Switzerland and France with a new multilocus sequence typing sheme*[online]. PLoS ONE. 2018 Jul 30;13(7):e0200773. Doi: <https://doi.org/10.1371/journal.pone.0200773>.

[38] KE W., MOLINI B.J., LUKEHART S.A., GIACANI L. *Treponema pallidum subsp. pallidum TP0136 protein is heterogeneous among isolates and binds cellular and plasma fibronectin via its NH2-terminal end*[online]. PLoS Negl Trop Dis. 2015; 9(3): e0003662. Dostupné z: <https://doi.org/10.1371/journal.pntd.0003662>.

[39] LUO X., LIN W.S., XU Q.Y, et al. *Tp0136 targets fibronectin (RGD) / β1 interactions promoting human microvascular endothelial cell migration*[online]. Experimental Cell Research, Volume 396, Issue 1, [cit. 18. 12. 2022]. Doi: <https://doi.org/10.1016/j.yexcr.2020.112289>.

[40] ERWING B., GREEN P.: *Base-calling of automated sequencer traces using phred. II. Error probabilities*[online]. Genome research, 8(3), 186-194 Dostupné z: <https://doi.org/10.1101/gr.8.3.186>.

[41] NCBI: *Treponema pallidum subsp. pallidum str. Nichols, complete genome*[online]. [cit. 18. 12. 2022]. Dostupné z: <https://www.ncbi.nlm.nih.gov/nuccore/CP004010.2/>.

[42] NCBI: *Treponema pallidum subsp. pallidum SS14, complete genome*[online]. [cit. 18. 12. 2022]. Dostupné z: <https://www.ncbi.nlm.nih.gov/nuccore/CP004011.1/>.

[43] NACHVÁTAL L.,PĚTROŠOVÁ H., GRILLOVÁ L., et al.: *Syphilis-causing strains belong to separate SS14-like or Nichols-like groups as defined by multilocus analysis of 19 Treponema pallidum strains*[online]. International Journal of Medical Microbiology, Volume 304, Issues 5–6, 2014, Pages 645-653, [cit. 19. 12. 2022]. Dostupné z: <https://doi.org/10.1016/j.ijmm.2014.04.007.>.

[44] National Institute of Health, Portugal: *SRX1798946: Treponema pallidum PT_SIF0857 - raw reads w/o human*[online]. [cit. 28. 12. 2022]. Dostupné z: <https://www.ncbi.nlm.nih.gov/sra/SRX1798946[accn]>.

[45] *SRA-toolkit*[online]. [cit. 1. 5. 2024]. Dostupné z: <https://github.com/ncbi/sra-tools>.

[46] PINTO M., BORGES V., ANTELO M., et al.: *Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation*[online]. Nat Microbiol 2, 16190 (2017). Doi: <https://doi.org/10.1038/nmicrobiol.2016.190>.

[47] Illumina: *Quality Scores*[online]. [cit. 29. 12. 2022]. Dostupné z: <https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScores_swBS.htm>.

[48] Illumina: *Quality Scores for Next-Generation Sequencing*[online]. [cit. 27. 2. 2024]. Dostupné z: <https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf>.

[49] Li H., Durbin R.: *Fast and accurate short read alignment with Burrows-Wheeler transform.*[online]. Bioinformatics. 2009;25(14):1754-1760. [cit. 29. 2. 2024]. Doi: <https://doi.org/10.1093/bioinformatics/btp324>.

[50] Babraham Bioinformatics: *FastQC* [online]. [cit. 30. 2. 2024]. Dostupné z: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

[51] GUHATHAKURTA D., STORMO D.G.: *Finding regulatory elements in DNa sequence*[online]. [cit. 19. 4. 2024]. Dostupné z: <https://www.researchgate.net/publication/265618126_Finding_regulatory_elements_in_DNa_sequence>.

[52] Clustal Omega - Multiple Sequence Alignment (MSA) Dostupné z: <https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240520-092102-0011-52708962-p1m&js=pass>.

[53] Clustal Omega - Multiple Sequence Alignment (MSA) Dostupné z: <https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240520-141056-0312-22134890-p1m>.

[54] Clustal Omega - Multiple Sequence Alignment (MSA) Dostupné z: <https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240520-141532-0606-61278330-p1m>.

[55] Clustal Omega - Multiple Sequence Alignment (MSA) Dostupné z: <https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240520-110638-0885-52994388-p1m>.

Seznam symbolů a zkratek

arp	acidic repeat protein
bp	base pair - pár bazí
BLAST	Basic Local Aligment Search Tool
BWA	Burrows-Wheeler Aligner
CCD	Charge-Coupled Device
CDC	Centers for Disease Control and Prevention
CMOS	Complementary Metal-Oxide-Semiconductor
CSF	cerebrospinal fluid - mozkomíšní mok
DDBJ	DNA Database of Japan
DNA	Deoxyribonukleová kyselina
ECDCT	enhanced-CDCT
EBI	European Bioinformatics Institute
HMEC-1	Human microvascular endothelial cells
INSDC	International Nucleotide Sequence Database Collaboratio
ISFET	ion field sensitive transistor - iontově senzitivní tranzistor (tranzistor řízený polem)
MLST	Multilocus sequence typing - Multilokusové sekvenční typování
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
PCR	polymerase chain reaction - polymerázová řetězová reakce
PNC	penicilin
RPR	rapid plasma reagin
REA	Restriction endonuclease analysis
SMS	Single molecule sequencing - sekvenace jedné molekuly

SMRT	Single molecule real time - sekvenace jedné molekuly v reálném čase
SNR	Seignal to Noise Ratio - poměr signálu k šumu
SRA	Sequence Read Archive
SOLiD	Sequencing by oligonucleotide ligation and detection
TCEP	tris(2-carboxyethyl)fosfin
TIRF	Total Internal Reflection Fluorescence - Fluorescence s totálním vnitřním odrazem
TPPA	Treponema pallidum particle agglutination
ZMW	zero mode waveguides