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

GENOTYPIZACE KMENŮ BAKTERIE KLEBSIELLA PNEUMONIAE
GENOTYPING OF KLEBSIELLA PNEUMONIAE ISOLATES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Markéta Nykrýnová**
AUTHOR

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

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

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

Studentka: Bc. Markéta Nykrýnová ***ID:*** 164215
Ročník: 2 ***Akademický rok:*** 2017/18

NÁZEV TÉMATU:

Genotypizace kmenů bakterie Klebsiella pneumoniae

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši na metody typizace a genotypizace bakteriálních kmenů, především se zaměřte na metodu minim-typing MLST. 2) Ze sekvenačních dat poskládejte genomy bakteriálních kmenů Klebsiella pneumoniae a genomy anotujte. 3) Navrhněte algoritmus pro identifikaci genů vhodných pro genotypizaci, které mají vyšší míru variability napříč různými kmeny. 4) Algoritmus otestujte na kmenech bakterie Klebsiella pneumoniae. 5) Ze souboru genů vyberte co nejmenší počet genů, jejichž kombinace bude mít nejlepší diskriminační schopnost pro genotypizaci. 6) Výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] BRHELOVA, E.; KOCMANOVA, I.; RACIL, Z.; HANSLIANOVA, M.; ANTONOVA, M.; MAYER, J.; LENGEROVA, M. Validation of Minim typing for fast and accurate discrimination of extended-spectrum, beta-lactamase-producing Klebsiella pneumoniae isolates in tertiary care hospital. 2016, 86, 44-49.

[2] ANDERSSON, P.; TONG, S.Y.; BELL, J.M.; TURNIDGE, J.D.; GIFFARD, P.M. Minim typing - a rapid and low cost MLST based typing tool for Klebsiella pneumoniae. PLoS One. 2012, 7(3), e33530.

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

Vedoucí práce: Ing. Denisa Maděránková, Ph.D.
Konzultant:

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

ABSTRAKT

Tato diplomová práce se zabývá typizací kmenů bakterie *Klebsiella pneumoniae*. V první části jsou představeny metody typizace včetně jejich výhod a nevýhod. Následně je charakterizován bakteriální genom a popsána bakterie *Klebsiella pneumoniae*. V praktické části je uveden postup složení jednotlivých genomů včetně otestování jejich kvality a je představen navržený algoritmus pro nalezení variabilních úseků genů, které vykazují vyšší míru variability. Poté jsou uvedeny získané výsledky, které jsou následně otestovány na dalších genomech *Klebsiella pneumoniae*.

KLÍČOVÁ SLOVA

genotypizace, typizační metody, bakterie, *Klebsiella pneumoniae*

ABSTRACT

This master thesis deals with typing of *Klebsiella pneumoniae* isolates. The first part of the thesis introduces molecular typing methods. Then bacterial genomes and *Klebsiella pneumoniae* are characterized. Following part describes data validation, assembly of genomes and proposed algorithm for finding genes with high variability. In last part obtained results are presented and validated on other genomes of *Klebsiella pneumoniae*.

KEYWORDS

genotyping, typing methods, bacteria, *Klebsiella pneumoniae*

NYKRÝNOVÁ, Markéta. *Genotypizace kmenů bakterie Klebsiella pneumoniae*. Brno, 2018, 56 s. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Denisa Maděránková, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Genotypizace kmenů bakterie Klebsiella pneumoniae“ jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové 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

PODĚKOVÁNÍ

Děkuji vedoucí mé diplomové práce paní Ing. Denise Maděránkové, Ph.D. za trpělivost, konzultace, cenné a podnětné návrhy k práci.

Brno ....

podpis autorky

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Typizace bakterií představuje důležitý krok vedoucí k nalezení vztahů mezi jednotlivými bakteriálními liniemi, což je nezbytné pro porozumění šíření epidemií a vyhledání zdrojů nákazy. Typizace je zároveň nepostradatelná pro detekci přenosu nosokomiálních patogenů, včetně jejich diagnózy a léčby.

V minulosti byly typizační metody založené převážně na postupech zkoumající fenotyp a pouze omezená část se zabývala genotypem. Nicméně, v dnešní době se právě genotypizační metody dostávají do popředí, protože na rozdíl od fenotypových jsou schopny rozlišit velmi blízce příbuzné linie.

Jelikož některé typy bakterie *Klebsiella pneumoniae* jsou hypervirulentní a rezistentní vůči některým antibiotikům, a tudíž se velice rychle šíří napříč jednotlivými odděleními nemocnic, je potřeba mít k dispozici rychlou, přesnou a cenově nenáročnou typizační metodu, která bude schopna spolehlivě určit jednotlivé bakteriální linie, aby bylo možné v případě výskytu nebezpečného kmenu okamžitě zamezit dalšímu šíření a předejít potencionální epidemii.

K řešení výše uvedeného problému může přispět námi navržený postup, který slouží k nalezení variabilních genových úseků, díky nimž lze bezpečně rozlišit jednotlivé bakteriální linie. Nalezené úseky budou moci být v budoucnu použity pro metodu mini-MLST, kde bude pro každý genom určen jeho melt-typ, což umožní rozlišení linií.

Cílem teoretické části práce je představit čtenáři nejčastěji používané molekulární metody typizace, včetně jejich výhod a nevýhod. Následně je zde stručně charakterizován bakteriální genom a popsána bakterii *Klebsiella pneumoniae* včetně její rezistence vůči antibiotikům.

V praktické části je popsána kontrola kvality sekvenačních dat, jejich složení a následně navržen algoritmus pro nalezení genových úseků, které vykazují vyšší míru variability, a tudíž jsou vhodné pro bakteriální typizaci. V poslední části jsou uvedeny získané výsledky, které jsou následně otestovány a diskutovány.

1 TYPIZACE DNA

Typizační metody pro odlišení linií u bakterií stejného druhu jsou nezbytným epidemiologickým nástrojem pro kontrolu a prevenci infekcí. V posledních letech bylo vyvinuto několik nových metod, které slouží ke zkoumání molekulární epidemiologie mikrobiálních patogenů. Níže jsou uvedeny nejčastěji používané metody typizace, přičemž každá metoda má své výhody a nevýhody, tudíž její výběr záleží na analýze, kterou chceme provést.

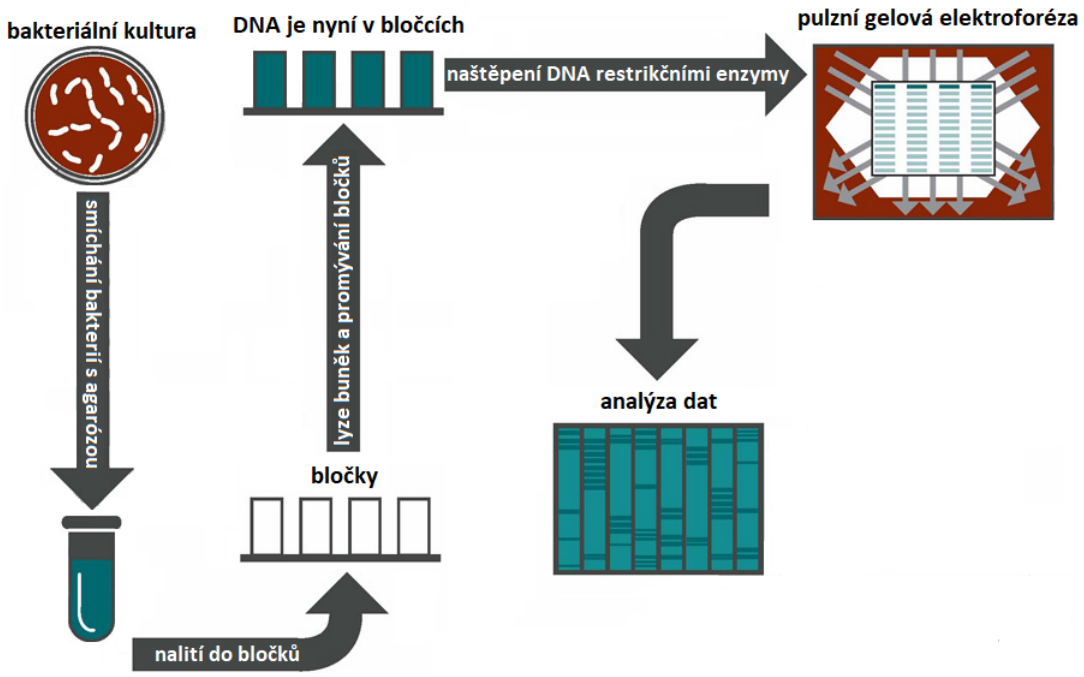
Obecně lze molekulární typizační metody rozdělit do tří hlavních skupin: metody založené na restrikční analýze bakteriální DNA, metody založené na amplifikaci vybraných genetických úseků pomocí polymerázové řetězové reakce (PCR) a metody založené na identifikaci polymorfismů v DNA sekvencích. [1], [2]

1.1 Pulzní gelová elektroforéza

Pulzní gelová elektroforéza (PFGE) je považována za zlatý standard mezi molekulárními typizačními metodami. Má velkou rozlišovací schopnost a velkou epidemiologickou shodu, je relativně levná a reprodukovatelná a umožňuje separaci velmi dlouhých fragmentů DNA, a to v rozsahu od 10 do 800 kb. Schéma metody je uvedeno na obr. 1.1.

Izoláty z bakterií rostoucí v médiu jsou smíchány s tekutou agarózou a poté vylity do bločků. Usazené bakterie v agaróze jsou následně lyzovány detergentním enzymem a naštěpeny méně častými restrikčními enzymy. V dalším kroku jsou naštěpené bakterie vloženy do agarózního gelu a umístěny do elektroforézy. V průběhu samotného procesu dochází v předem definovaných periodicky se opakujících intervalech ke změnám polarity proudu. Molekuly DNA se pohybují od katody k anodě v závislosti na své velikosti. Během periodických změn proudu se jednotlivé fragmenty přeorientovávají podle nového směru elektrického proudu, přičemž čas potřebný pro úpravu směru je nepřímo úměrný velikosti DNA fragmentu. Pro správné rozdělení dlouhých fragmentů DNA je požadováno použití pulzního elektrického pole, které vytváří 24 elektrod, přičemž tyto elektrody postupně mění směr pole o daný fixní úhel a to po dlouhou dobu. Gel je poté obarven fluorescenčním barvivem a zobrazen pod UV světlem. Následně může být vyfocen a podroben analýze.

Ačkoliv je pulzní gelová elektroforéza hojně využívaná, má několik nevýhod, mezi které patří například technická a časová náročnost a špatná rozlišovací schopnost pro proužky s téměř stejnou velikostí. [1], [3], [4]



Obr. 1.1: Schéma pulzní gelové elektroforézy, z [5].

1.2 Náhodná amplifikace polymorfní DNA

RAPD neboli náhodná amplifikace polymorfní DNA je metoda, která vznikla v roce 1990 a je založená na použití krátkých náhodných primerů (obvykle okolo 10 bází), které hybridizují s chromozomální DNA při nízké teplotě, takže mohou být použity k zahájení amplifikace oblastí z bakteriálního genomu. Namnožené fragmenty jsou následně separovány pomocí elektroforézy na agarovém gelu k vytvoření bakteriálního fingerprintingu. Elektroforeogramy jsou použity k porovnání příbuznosti bakteriálních linií.

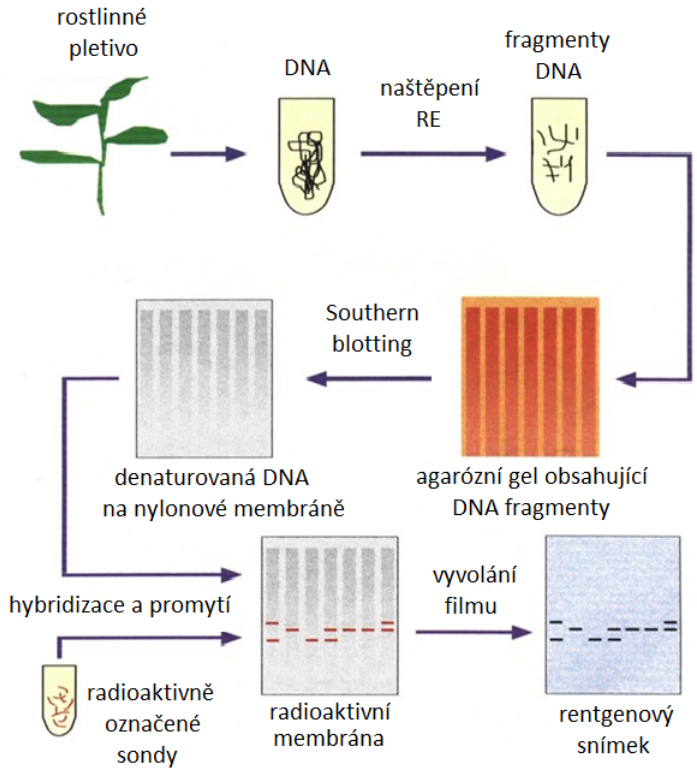
Ačkoliv má menší rozlišovací schopnost než pulzní gelová elektroforéza, je často používaná při vypuknutí epidemií, jelikož je jednoduchá, levná a rychlá. Mezi hlavní nevýhodu patří nízká reprodukční schopnost mezi jednotlivými laboratořemi, jelikož jsou použity velmi nízké teploty nasedání primerů. [1], [6]

1.3 RFLP analýza

RFLP znamená polymorfismus délky restrikčních fragmentů, přičemž právě analýza RFLP patří mezi jednu z prvních technik pro detekci variant v DNA sekvencích. Při této analýze měříme velikost restrikčních fragmentů, jež jsou separovány klasickou agarózní gelovou elektroforézou. Jelikož při použití restrikčních enzymů,

které se často využívají, může vzniknou velké množství krátkých fragmentů, dochází ke ztížení separace pomocí elektroforézy. Nicméně RFLP analýza může být zjednodušena, pokud použijeme Southern blotting. Separované molekuly DNA jsou přeneseny z agarózního gelu na nitrocelulózovou nebo nylonovou membránu. DNA je navázána na membránu a hybridizována označenými sondami homologními ke genu, který chceme určit. Následně již probíhá vizualizace výsledků. Schéma metody je zobrazeno na obr. 1.2.

Výhodou analýzy RFLP je, že se dá použít i pro neznámé vzorky, dále její efektivita, nízká cena a rychlost. Mezi nevýhody patří nízká reprodukovatelnost. [6], [7]



Obr. 1.2: Schéma RFLP analýzy, z [8].

1.4 Repetitivní PCR

Repetitivní polymerázová řetězová reakce byla poprvé popsána v roce 1991 a využívá primery, jež hybridizují s nekódujícími mezigenovými úseky, které se opakují v celém genomu. DNA mezi sousedícími repetitivními úseky je poté amplifikována pomocí PCR. Zároveň může být vytvořeno více amplikonů, přičemž záleží na rozložení repetitivních úseků v genomu. Získané fragmenty jsou analyzovány pomocí elektroforézy

za účelem zjištění jejich velikosti a v následujícím kroku jsou porovnány vzniklé elektroforeogramy. Následně lze určit genetickou příbuznost mezi jednotlivými analyzovanými liniemi.

Metoda je jednoduchá, rychlá, lze ji použít na velké i malé množství vzorků a pro velké množství bakteriálních genomů má velkou rozlišovací schopnost. Hlavní nevýhodou je menší reprodukovatelnost kvůli použití jiných reagencí při PCR a elektroforetických systémů. K vyhnutí se problému s reprodukovatelností vznikl poloautomatický rep-PCR komerční systém, který místo klasické gelové elektroforézy využívá mikrofluidní čipy pro separaci DNA fragmentů. [1], [6]

1.5 Polymorfismus délky amplifikovaných fragmentů

Polymorfismus délky amplifikovaných fragmentů neboli AFLP je metoda, jež je založená na amplifikaci vybraných fragmentů DNA, které vzniknou po naštěpení restrikčními enzymy. Jsou popsány dvě varianty AFLP. V případě první varianty se používají dva restrikční enzymy a dva primery, zatímco u druhé varianty se využívá pouze jeden restrikční enzym a jeden primer.

Metoda začíná naštěpením DNA pomocí restrikčních enzymů. Na naštěpené konce se poté navážou adaptory, což jsou známé úseky dvouvláknové DNA. Fragmenty s navázanými adaptory jsou amplifikovány pomocí PCR s použitím primerů, jež jsou komplementární k navázaným adaptorům. Pokud využijeme primery označené fluorescenční barvou, tak lze po tom, co byly fragmenty separovány na základě velikosti, použít k vyhodnocení automatický sekvenátor.

Výhodou metody je, že je reprodukovatelná, nejsou potřeba nasekvenovaná data pro konstrukci primerů a má velkou diskriminační schopnost. Mezi nevýhody patří vysoká cena, náročnost a potřeba mít vysoce kvalitní DNA. [1], [4], [9], [10]

1.6 Analýza variabilního množství tandemových repetic

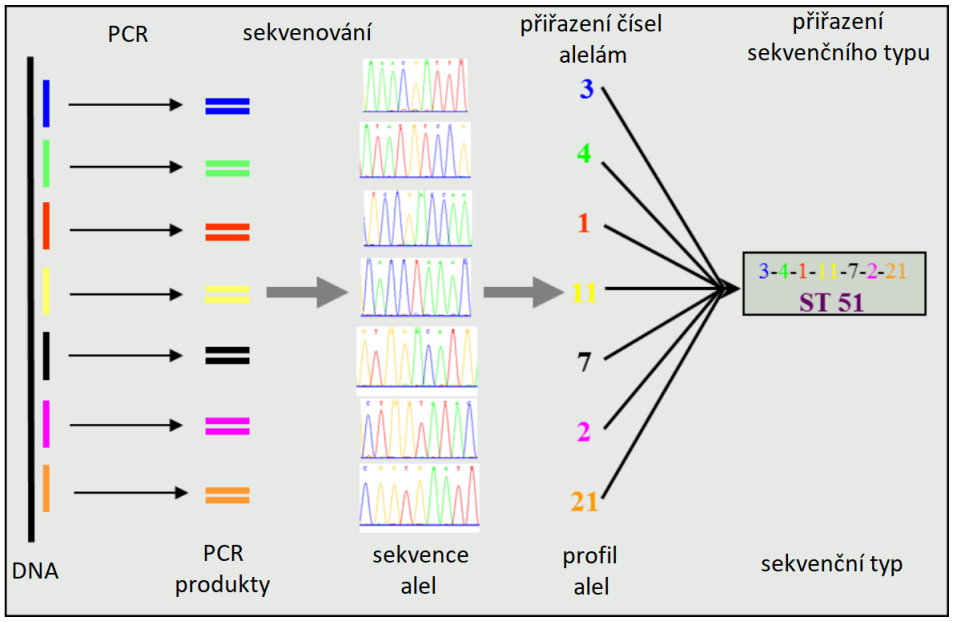
Velké množství bakteriálních genomů obsahuje regiony s nukleotidovými repeticemi, a to jak v kódujících, tak v nekódujících úsecích DNA o velikosti několik bází až po sto bp. Repetice jsou tandemové, což znamená, že několik kopií každého z repetitivních motivů je shromážděno u sebe a orientováno stejným směrem. Počet opakování v tandemu se může lehce lišit, a to i u linií stejného druhu.

VNTR (variabilní množství tandemových repetic) analýza je metoda založená na PCR, která spoléhá na amplifikaci DNA, jež zahrnuje krátké tandemové repetice DNA sekvence. Primery jsou navrhnuty tak, aby nasedly před opakující se úsek. Amplifikované produkty jsou poté separovány a změřeny k určení počtu opakování přítomných v amplikonu. Rozdíl v počtu repetitivních kopií na specifickém místě slouží k rozlišení linií.

Výhodou metody je její nízká cena, rychlost a to, že může být provedena i v laboratořích bez sofistikovaného elektroforetického systému. Mezi nevýhody patří, že získaná data nemohou být porovnána mezi laboratořemi. [1], [6]

1.7 Multilokusová sekvenční typizace

Multilokusová sekvenční typizace (MLST) je metoda, jež byla vyvinuta v roce 1998 za účelem předejití nízké reprodukovatelnosti mezi laboratořemi, která je typická pro starší metody typizace. MLST je založena na amplifikaci a následném osekvenování až 7 genů, které mají vysokou variabilitu. Jedná se převážně o provozní geny (z angl. housekeeping genes), jejichž sekvence jsou relativně stálé, jelikož kódují proteiny, jež jsou nezbytné pro funkci buňky. Pro vybrané geny je identifikována sekvence o délce 450 - 500 bp. Pro každý lokus je určen sekvenční typ a to tak, že unikátním sekvencím (alelám) jsou přiřazeny náhodná celá čísla a na základě kombinace identifikovaných alel je určen sekvenční typ. Schéma postupu je uvedeno na obr. 1.3.



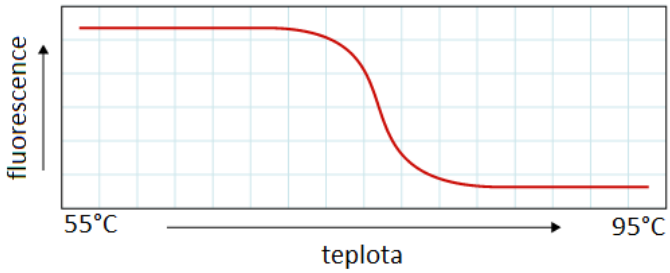
Obr. 1.3: Schéma metody MLST, z [11].

Výhodou MLST je, že získaná data jsou jednoznačná kvůli mezinárodně uznávané nomenklatuře, vysoce reprodukovatelná a mají velkou rozlišovací schopnost. Sekvence alel a profily sekvenčních typů jsou dostupné v centrálních databázích, které též obsahují software pro určení genetické příbuznosti mezi bakteriálními liniemi v rámci druhu. Mezi nevýhody metody patří vysoká cena, časová náročnost a pro některé patogeny nedostatečná diskriminační schopnost pro rutinní použití při vypuknutí epidemie a jejím sledování. [1], [9], [12], [13]

1.8 Mini-multilokusová sekvenční typizace

Mini-multilokusová sekvenční typizace (mini-MLST či minim-typing) je odvozená od multilokusové sekvenční typizace. V první části probíhá proces stejně jako u MLST, tedy v průběhu PCR dochází ke zmnožení genů jako u MLST, avšak v další části je sekvenování nahrazeno vysokorozlišovací analýzou křivek tání (HRM). Pro MLST jsou pomocí počítačové analýzy vybrány ty úseky DNA, které vykazují největší diverzitu v rámci genů a pro tyto úseky jsou následně navrhnuty primery.

Po provedení amplifikace genů probíhá HRM analýza, což je citlivá a rychlá metoda, která dokáže zachytit změny v jednom i více nukleotidech. DNA je obarvena pomocí barviva a následně během samotné analýzy dochází ke zvyšování teploty, což má za následek tání dvoušroubovice DNA a uvolňování barviva, čímž dochází k poklesu fluorescence, kterou sledujeme. Výsledkem je křivka tání, přičemž i jediná změna nukleotidu se v ní projeví. Příklad získané křivky tání je uveden na obr. 1.4. Zároveň během analýzy měříme i teplotu tání sekvence, kterou lze určit, kdy je 50 % DNA denaturováno. Následně je použit převodní klíč k přeložení naměřených dat do melt profilů a také k porovnání dat s MLST databází.



Obr. 1.4: Příklad křivky tání, z [14].

Výhodou metody je nízká cena, která se pohybuje v rozsahu zhruba 10-20 % MLST.

Z výše uvedeného plyne, že mini-MLST má potenciál k usnadnění nízkonákladových průzkumů, použití v případě podezření na vypuknutí nákaz a pro studie zahrnující velké množství linií.

Metoda mini-multilokusové sekvenční typizace byla již použita pro *Streptococcus pyogenes, Staphylococcus aureus, Enterococcus faecium* a *Klebsiella pneumoniae.* [15], [16], [17], [18]

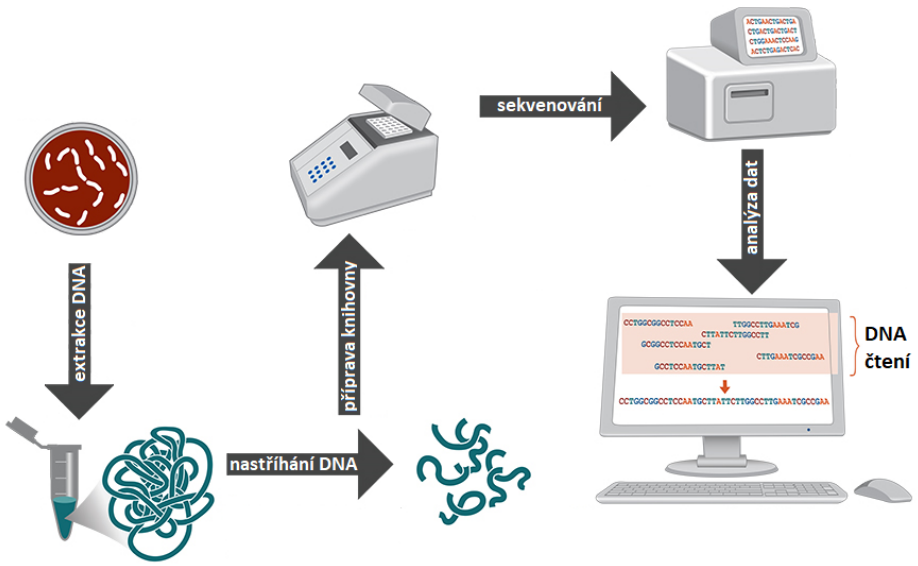
1.9 Celogenomové sekvenování

Sekvenování nové generace neboli NGS poskytlo cenově dostupnou cestu, jak prozkoumat varianty v celém genomu. Na rozdíl od Sangerova sekvenování, je během jednoho běhu schopno vyprodukovat miliony čtení (35 - 700 bp dlouhé) za relativně nízkou cenu.

Abychom dostali celou sekvenci, je nutné získaná čtení složit, a to buď *de novo*, tedy na základě překrývajících se úseků, nebo přiřazováním jednotlivých čtení k referenčnímu genomu.

Celogenomové sekvenování neboli WGS se stává velice atraktivním a mocným nástrojem pro epidemiologii. Je velmi pravděpodobné, že v blízké budoucnosti bude WGS běžně používané k přesné identifikaci a charakterizaci bakteriálních linií, přičemž hlavním cílem nebude produkovat sekvenační data, ale rychle je zpracovat a interpretovat důležité informace z objemného souboru, přičemž by získané informace měly být porovnatelné s výsledky získanými konvenčními metodami typizace. Postup celogenomového sekvenování je uveden na obr. 1.5.

Ovšem nalezneme zde i několik nevýhod. Produkovaná čtení jsou relativně krátká, občas není získán celý genom, ale pouze např. 90 % a je zde problém s VNTR. [1], [19]



Obr. 1.5: Schéma celogenomového sekvenování, z [20].

2 BAKTERIE *KLEBSIELLA PNEUMONIAE*

2.1 Bakteriální genom

Bakteriální genom se skládá z jedné kruhové molekuly DNA o průměrné molekulové hmotnosti 2,5 $\cdot 10^9$ a o délce téměř 1 mm, která se nazývá nukleoid. Chromozom není ohraničen membránou a v jednom místě je připevněn k plazmatické membráně buňky. Geny jsou zde uspořádány vedle sebe pouze s minimálním mezigenovým prostorem.

Kromě nukleoidu obsahuje bakteriální buňka též plazmidy, což jsou kruhové molekuly DNA, které jsou zhruba 100x menší než chromozom. Obvykle nesou pouze malé množství genů, které nejsou esenciální pro přežití buňky, ale slouží ke zvýhodnění bakterií. [21], [22]

2.2 *Klebsiella pneumoniae*

Klebsiella pneumoniae patří mezi Gram-negativní bakterie do čeledi Enterobacteriaceae, avšak na rozdíl od jiných zástupců z této čeledi nemá bičíky, a tudíž se jedná o nepohyblivou bakterii. Zároveň je obalena polysacharidovou kapsulou, jíž ji zajišťuje rezistenci proti mnohým obranným mechanismům hostitele. Je celosvětově rozšířena a můžeme ji nalézt v půdě, vodě i na povrchu rostlin. V nemocnicích ohrožuje pacienty se sníženou imunitou. Při infekci postihuje plíce, kde vyvolává pneumonii a též způsobuje infekce vylučovacího systému. Pokud pronikne do krevního oběhu, může u pacienta vyvolat sepsi. Mezi nejzávažnější onemocnění patří gastroenteritidy, meningitidy novorozenců a již uvedené sepse.

Průměrná velikost genomu je 5,5 Mbp a obsahuje přibližně 5500 genů, přičemž pomocí celogenomového sekvenování bylo zjištěno, že méně jak 2000 genů je společných pro všechny linie *Klebsiella pneumoniae*. Zbývajících 3500 genů, které lze nalézt jen u některých linií má následující funkce: 19 % slouží pro metabolismus sacharidů, 18 % pro ostatní metabolické dráhy, 13 % se podílí na membránovém transportu, 11 % je pro exopolysacharidovou kapsuli, 2 % pro rezistenci a metabolismus železa, 1 % pro rezistenci vůči antibiotikům, těžkým kovům a stresu a třetina genů nemá zatím objasněnou funkci. [23], [24], [25], [26]

Bakterie *Klebsiella pneumoniae* je zobrazena na obr. 2.1.

2.2.1 Rezistence vůči antibiotikům

Mezi jeden z druhů, u kterého se v současné době mluví o rostoucí antibiotické rezistenci patří právě *Klebsiella pneumoniae*. Z výzkumu vyplývá, že rezistence

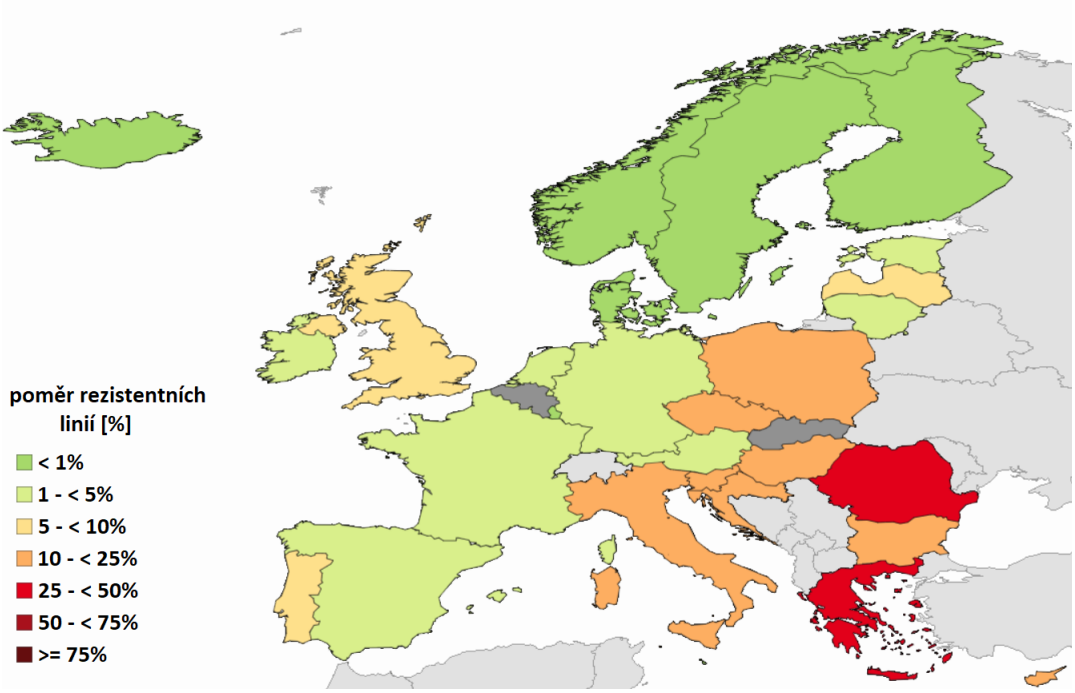
Obr. 2.1: Bakterie *Klebsiella pneumoniae* pod elektronovým mikroskopem, z [24].

vůči antibiotikům u ní vzniká mnohem snadněji a rychleji než u jiných bakterií a navíc je schopna stále vytvářet nové mechanismy rezistence, a tudíž v současné době zůstává již málo dostupných léčebných postupů.

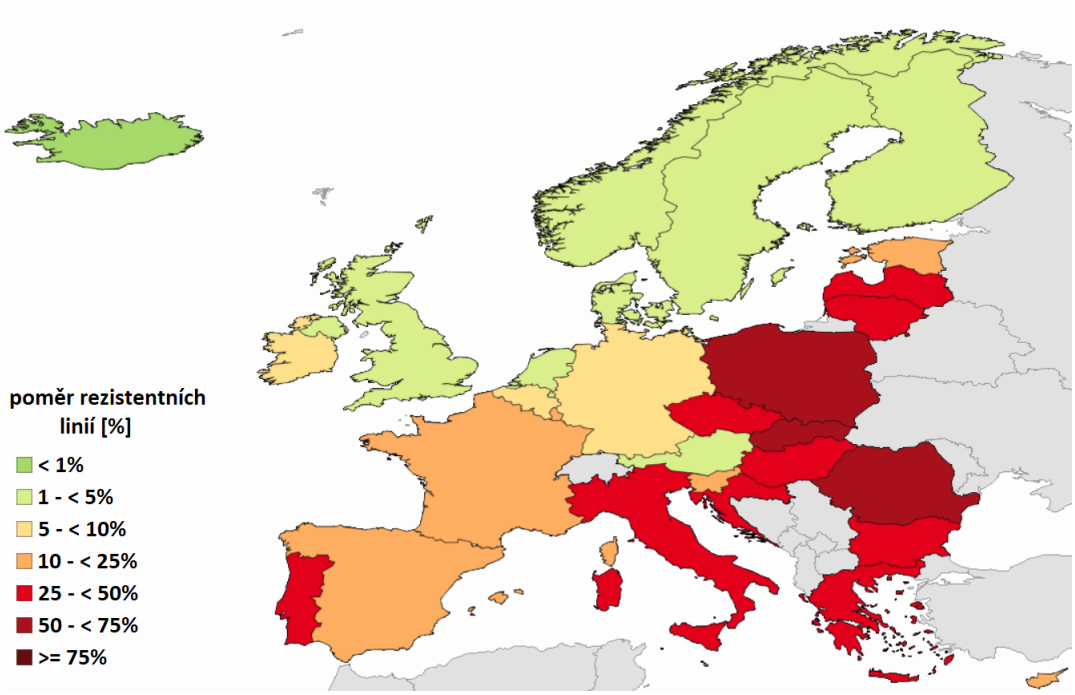
U bakterie *Klebsiella pneumoniae* byly pozorovány dva typy antibiotické rezistence. První mechanismus zahrnuje expresi beta-laktamáz širokého spektra, které poskytují bakterii odolnost proti cefalosporinům a monobaktamům. Druhý mechanismus je exprese karbapenemáz, které poskytují bakterii rezistenci k téměř všem dostupným beta-laktamovým antibiotikům (peniciliny, cefalosporiny, karbapenemy, ...). [27], [28]

2.2.2 Bakteriální rezistence v Evropě

Na obr. 2.2 a obr. 2.3 je zobrazen výskyt rezistentních linií bakterie *Klebsiella pneumoniae* s kombinovanou rezistencí (3. generace cefalosporinů, fluorochilony, aminoglykosidy) v roce 2006 a 2016. Jak vidíme, mezi uvedenými lety došlo k nárůstu výskytu rezistentních linií a to v některých krajinách dokonce o 50 %. Vyobrazené mapy rezistenci dokládají, jak již bylo uvedeno výše, že antibiotická rezistence u *Klebsiella pneumoniae* je stále rostoucí problém.



Obr. 2.2: Výskyt rezistentních kmenů bakterie *Klebsiella pneumoniae* s kombinovanou rezistencí v Evropě v roce 2006, z [29].



Obr. 2.3: Výskyt rezistentních kmenů bakterie *Klebsiella pneumoniae* s kombinovanou rezistencí v Evropě v roce 2016, z [29].

3 IDENTIFIKACE NOVÝCH GENŮ PRO TYPIZACI

V praktické části se zabýváme analýzou 24 genomů bakterie *Klebsiella pneumoniae*, která se skládá z otestování kvality sekvenačních dat, následném složení jednotlivých genomů a nalezením vhodných genů pro bakteriální typizaci. Poté jsou zde uvedeny získané výsledky, které jsou následně otestovány na dalších 12 genomech bakterie *Klebsiella pneumoniae*. Ze získaných výsledků jsou vybrány variabilní genové úseky, které jsou schopny odlišit jednotlivé bakteriální linie, a tudíž je lze využít pro genotypizaci. [30]

3.1 Vstupní data

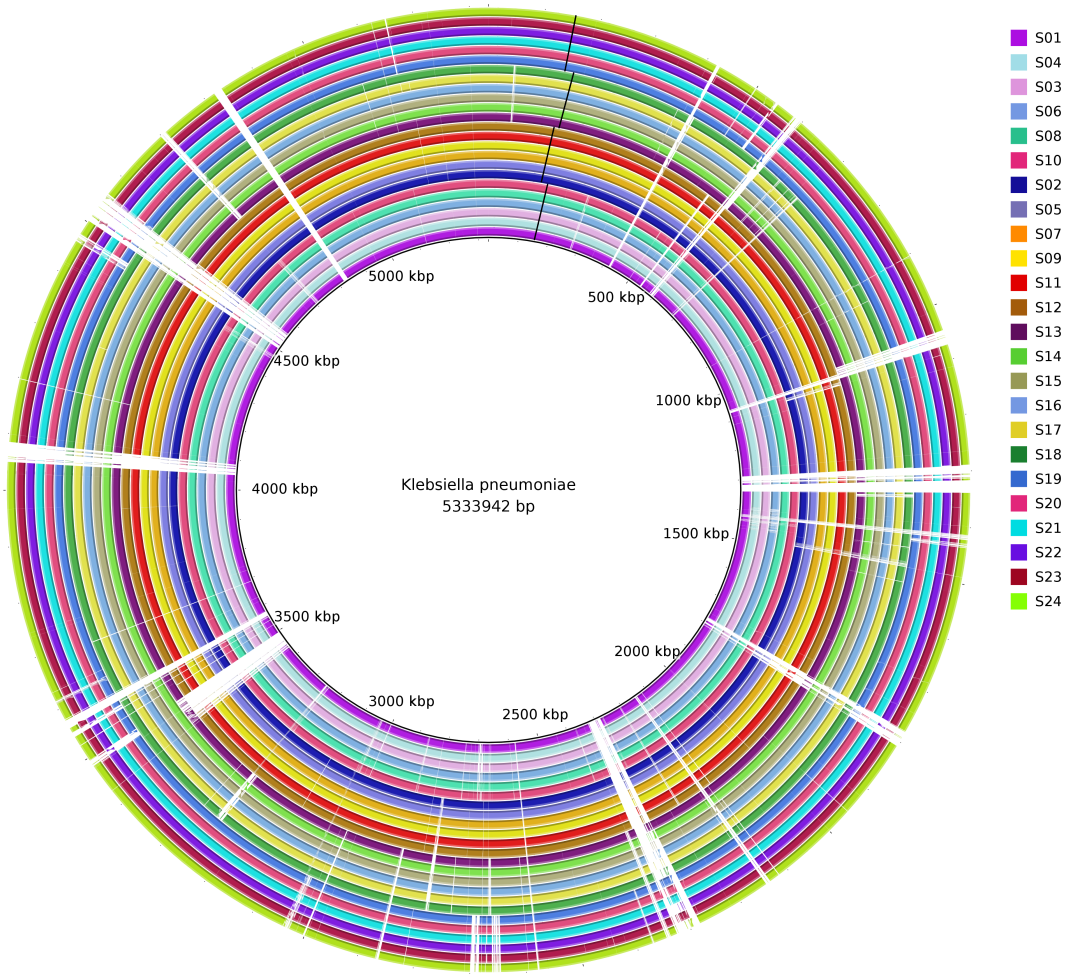
V práci se zabýváme 24 genomy bakterie *Klebsiella pneumoniae*, přičemž pro každý genom byl určen jeho melt-typ za použití typizační metody mini-MLST. Jednotlivé genomy označené pořadovým číslem S01 až S24 a k nim určené melt-typy jsou uvedeny v tab. 3.1.

Tab. 3.1: Analyzované genomy a jejich melt-typy.

genom	**melt-typ**	**genom**	**melt-typ**
S01	98	S13	23
S02	29	S14	23
S03	98	S15	23
S04	98	S16	23
S05	29	S17	23
S06	98	S18	23
S07	29	S19	61
S08	98	S20	61
S09	29	S21	61
S10	98	S22	61
S11	29	S23	61
S12	29	S24	61

Na obr. 3.1 je zobrazen kruhový graf s jednotlivými analyzovanými genomy vytvořený pomocí programu BRIG [31], kde každý kruh odpovídá jednomu genomu. Genomy, které patří do stejného melt-typu jsou označeny černou čárou. Jak vidíme v grafu, jednotlivé genomy se mezi sebou liší, což dokládají bílé úseky, ve kterých chybí části sekvence vzhledem k referenčnímu genomu (*Klebsiella pneumoniae subsp.*

pneumoniae HS11286, NC_016845). Z grafu vyplývá, že genomy nejsou zcela totožné, a proto je možné nalézt genové úseky, pomocí kterých jsme schopni odlišit jednotlivé linie, respektive melt-typy.



Obr. 3.1: Kruhový graf zobrazující srovnání analyzovaných genomů bakterie *Klebsiella pneumoniae*.

3.2 Otestování kvality vstupních dat

Před samotným složením genomů je potřeba otestovat sekvenační data a zjistit, zda jsou dostatečně kvalitní, neobsahují chyby či zda nedošlo ke kontaminaci vzorků.

Jednotlivé genomy byly sekvenovány pomocí Illumina sekvenátoru a bylo použito párové-koncové sekvenování (z angl. paired-end sequencing), což znamená, že pro jeden fragment máme sekvenované oba konce. Pro otestování kvality čtení jsme

použili program FastQC [32], který provádí samotnou kontrolu dat a MultiQC [33] pro zobrazení jednotlivých výsledků.

Jelikož bylo použito párové-koncové sekvenování, dostali jsme pro každý genom dva soubory (čtení v přímém a zpětném směru), a tudíž následná analýza probíhá pro oba dva.

Pro každý genom byla nejprve vyhodnocena obecná statistika, která zahrnuje název souborů s jednotlivými čteními, počet duplikovaných čtení, průměrný obsah GC, průměrnou délku čtení a celkový počet čtení. Příklad získaných údajů pro genom S10 je uveden v tab. 3.2.

Tab. 3.2: Obecné informace o testovaných čtení pro genom S10.

název	počet duplikovaných čtení [%]	průměrný obsah GC [%]	průměrná délka čtení [bp]	celkový počet čtení [miliony]
KP177_S10_L001_R1_001	4,1	56	247	0,8
KP177_S10_L001_R2_001	3,8	57	247	0,8

V další části se analyzuje průměrná kvalita jednotlivých pozic ve všech čtení, tzn. Phred skóre (obr. 3.2), která by neměla klesnou pod hodnotu 20, přičemž čím vyšší je hodnota kvality, tím lépe probíhá rozpoznání báze (z angl. base calling).

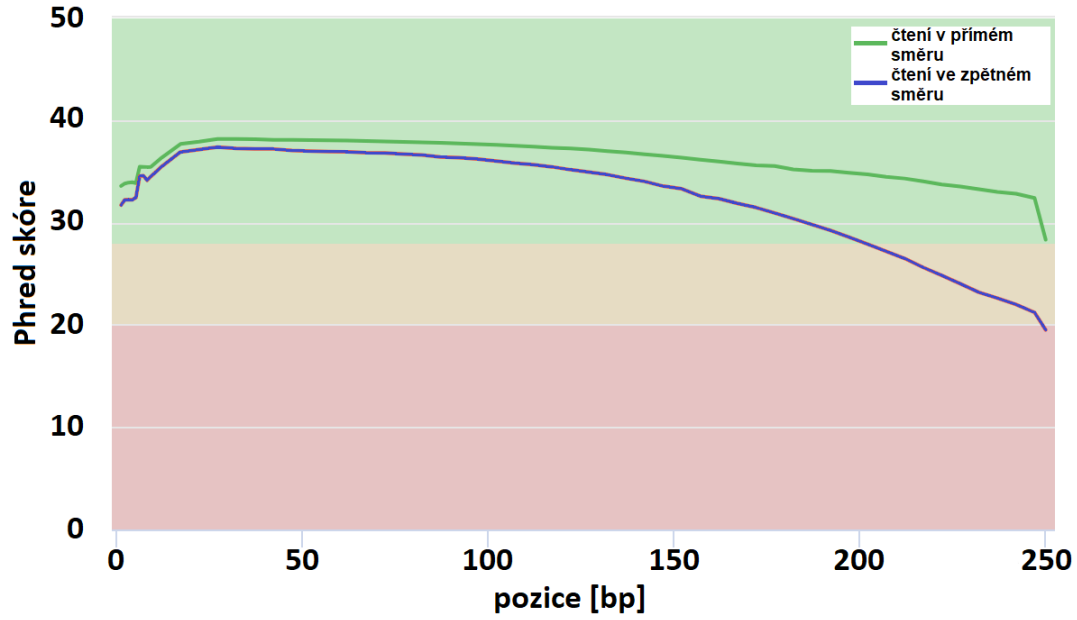
Phred skóre se používá pro změření přesnosti rozpoznání báze, přičemž se jedná o běžnou metriku využívanou pro ohodnocení přesnosti sekvenovací platformy. Označuje tedy pravděpodobnost, že daná báze je sekvenátorem rozpoznaná špatně.

Následně se vyhodnocuje průměrná kvalita pro čtení, tzn. kolik čtení o dané průměrné kvalitě sekvenační data obsahují (obr. 3.3) a poměr zastoupení jednotlivých bází (A, C, G, T) v rámci jedné pozice ve čtení (obr. 3.4), přičemž rozdíl mezi zastoupením A a T nebo G a C by neměl dosahovat hodnoty vyšší jak 20 %.

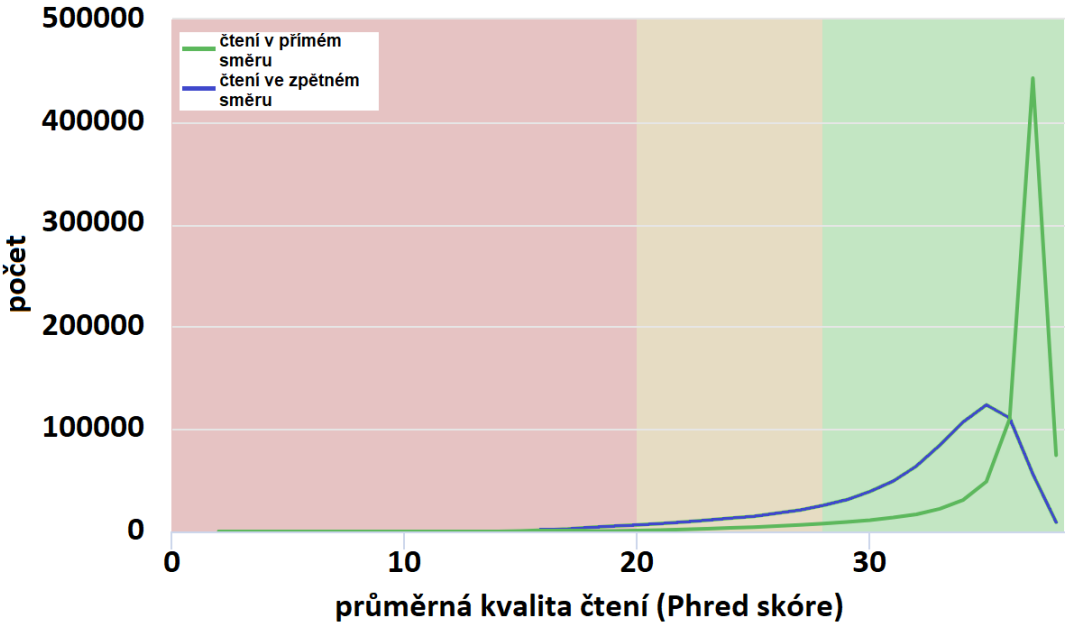
Dále testujeme průměrný obsah GC pro celou délku každého čtení (obr. 3.5), přičemž je zkoumáno, zda získaná data kopírují křivku normálního rozdělení obsahu GC. Pokud se tak nestalo a získaná křivka má neobvyklý tvar, mohlo dojít ke kontaminaci knihovny.

Jestliže sekvenátor nedokáže s dostatečnou přesností rozpoznat bázi, přiřadí na konkrétní pozici písmeno N. V analýze se kontroluje obsah N pro jednotlivé pozice ve všech čtení (obr. 3.6), který by neměl být větší než 20 %.

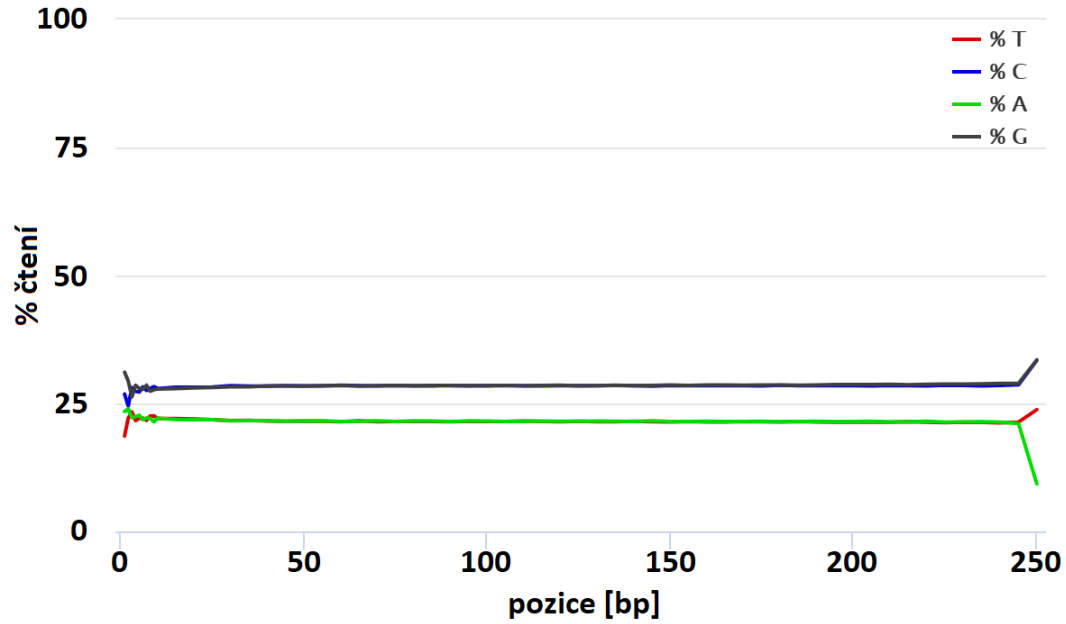
Následně se kontroluje počet čtení o jednotlivých délkách (obr. 3.7) a stupeň duplikace každého čtení v knihovně (obr. 3.8), přičemž by se nemělo stát, že počet neunikátních čtení bude větší než 50 % všech čtení.



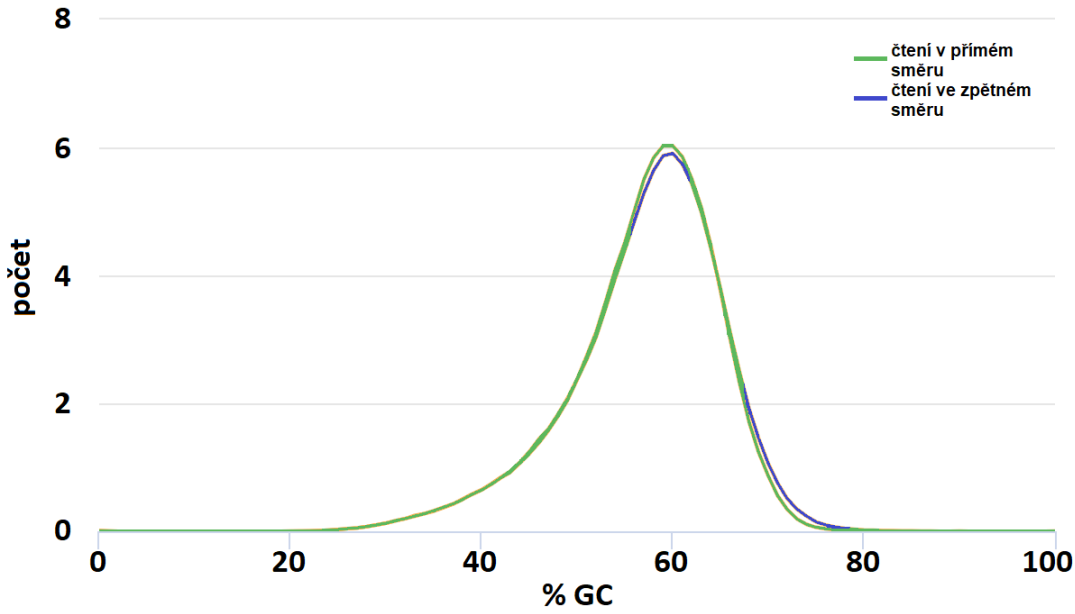
Obr. 3.2: Graf ukazující průměrnou kvalitu pozic ve všech čtení pro genom S10.



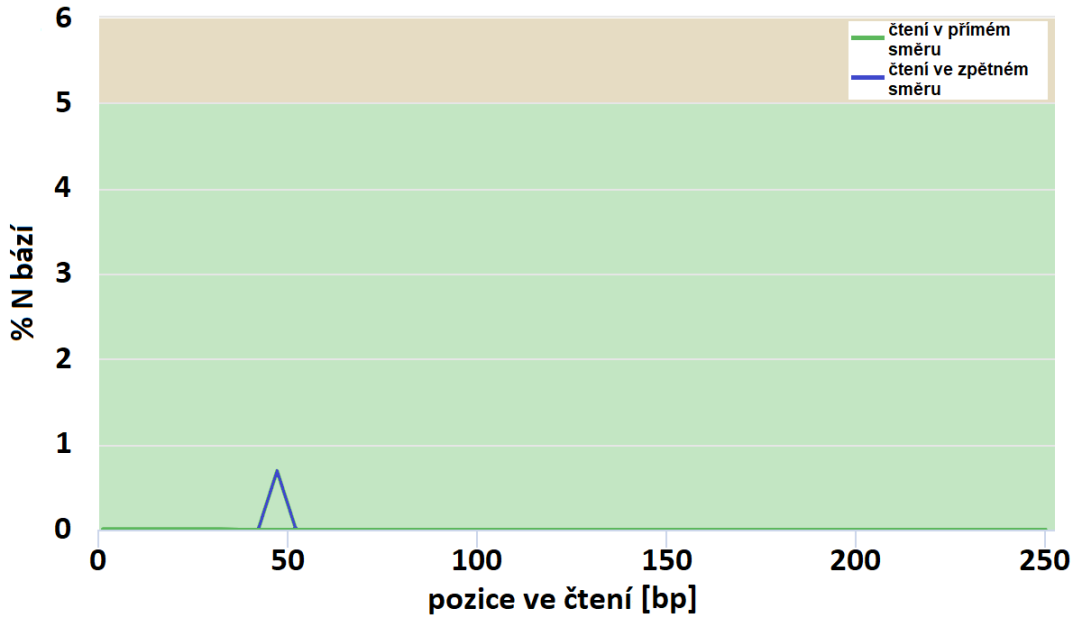
Obr. 3.3: Graf znázorňující počet čtení o dané kvalitě pro genom S10.



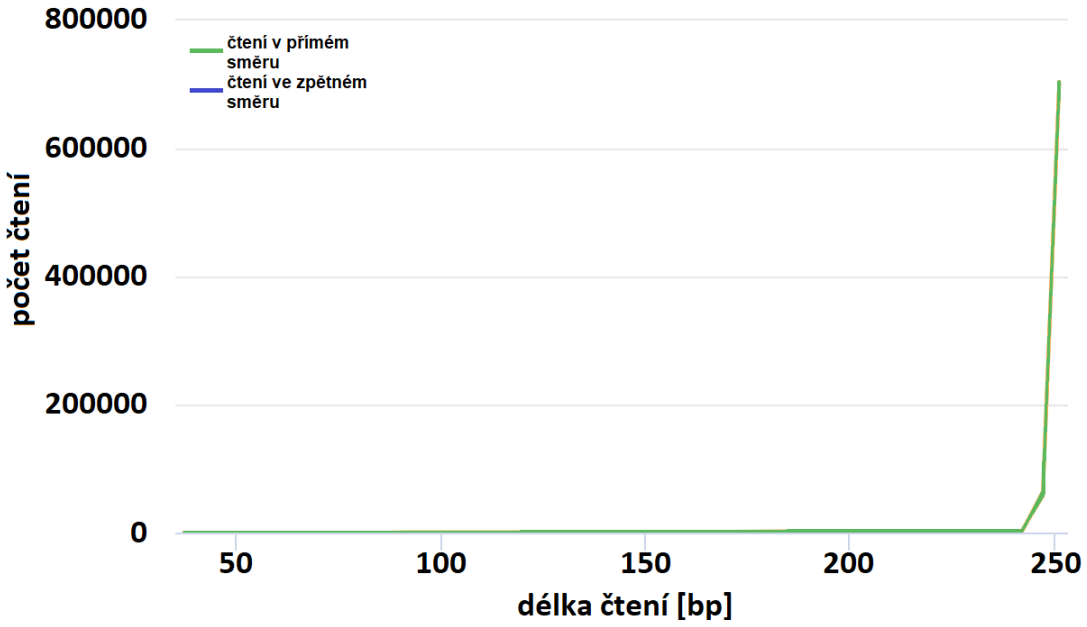
Obr. 3.4: Graf zobrazující poměr zastoupení jednotlivých nukleotidových bází ve čtení pro genom S10.



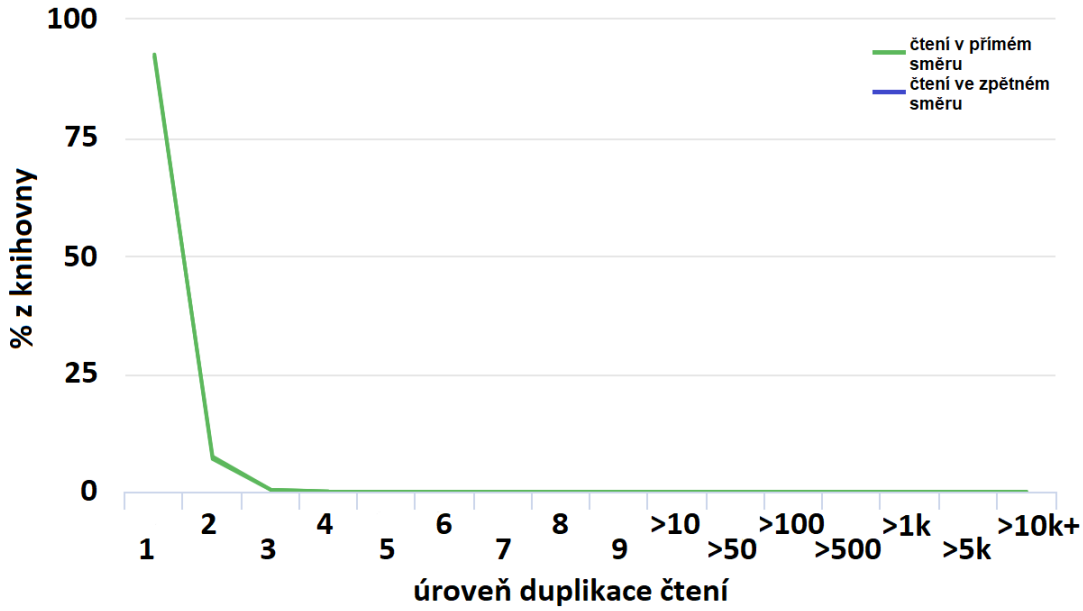
Obr. 3.5: Graf s průměrným obsah GC ve čtení pro genom S10.



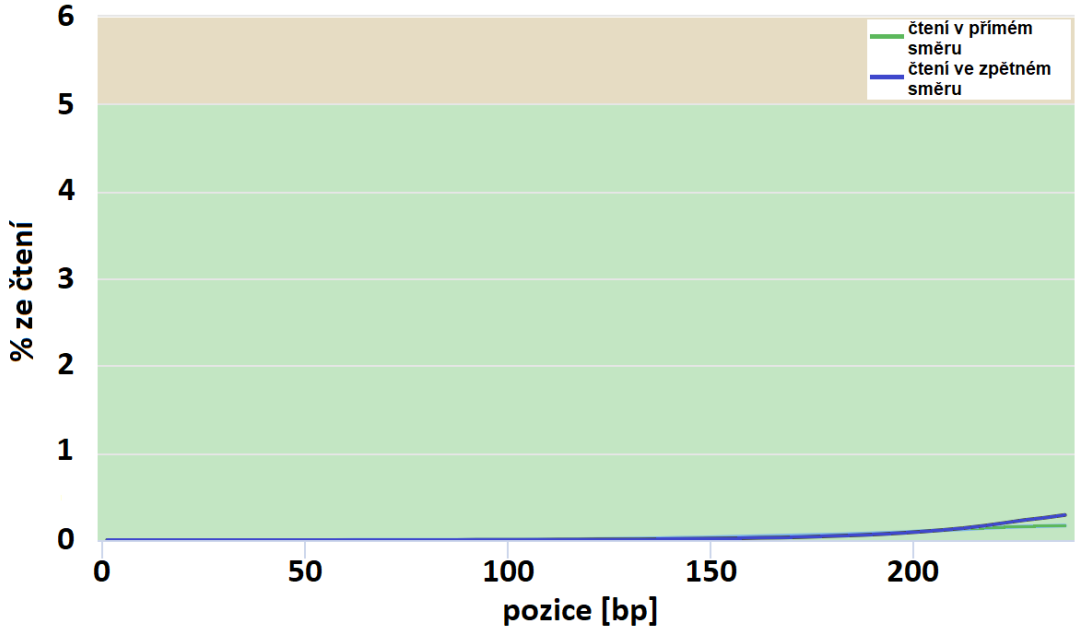
Obr. 3.6: Graf popisující obsah bází označených N pro genom S10.



Obr. 3.7: Graf vyobrazující počet čtení dané délce pro genom S10.



Obr. 3.8: Graf popisující počet duplikovaných čtení pro genom S10.



Obr. 3.9: Graf zobrazující obsah adaptorů pro genom S10.

Zároveň se sleduje, zda se některá sekvence nevyskytuje až příliš často, což by mohlo indikovat kontaminaci a též se testuje obsah adaptorů (3.9), což jsou krátké oligonukleotidové sekvence, od kterých probíhá sekvenování fragmentů.

Testování bylo provedeno pro všech 24 genomů *Klebsiella pneumoniae*. U všech genomů byla zjištěna snížená kvalita čtení ve zpětném směru, přičemž tato skutečnost bývá velmi často způsobena vysokou hustotou sekvenačních shluků na sekvenační destičce. Nicméně, pokles kvality není extrémní, a tudíž je kvalita pro naši práci dostačující.

3.3 Složení genomů

V dalším kroku bylo již provedeno samotné složení sekvenačních dat. Pro skládání genomu byl použit program Burrows-Wheeler Aligner (BWA) [34]. Skládání bylo provedeno oproti referenčnímu genomu *Klebsiella pneumoniae subsp. pneumoniae HS11286 (NC_016845)*, který byl získán z databáze NCBI. Tento genom je v databázi uváděn jako referenční genom pro zkoumanou bakterii, a proto byl použit. Během skládání byl nejprve nastaven referenční genom a následně byla načtena jednotlivá čtení, která byla poté přiřazena k referenčnímu genomu, čímž jsme získali složený genom ve formátu SAM (Sequence Alignment Map).

Poté byl použit program Samtools [35] k odstranění nenamapovaných čtení a k převedení sestaveného genomu do formátu BAM (Binary Alignment Map). Úsek získaného složeného genomu je zobrazen na obr. 3.10 pomocí programu Tablet [36], kde jsou vidět jednonukleotidové změny, které mohou sloužit pro odlišení jednotlivých bakteriálních linií.

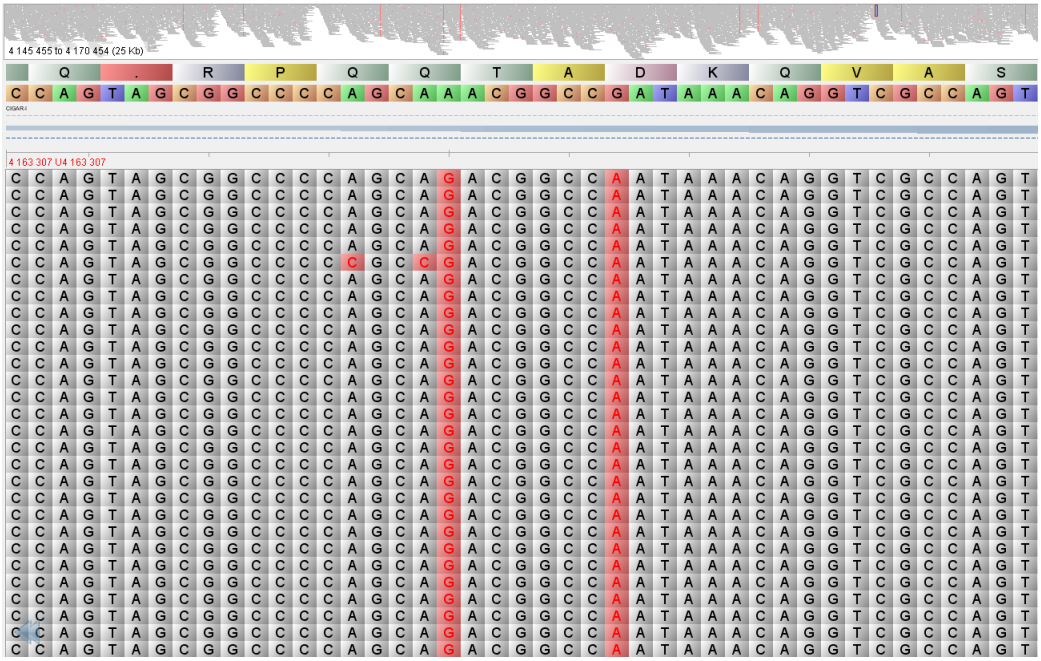
Počet celkových, namapovaných a nenamapovaných čtení je uveden v tab. B.1.

Z našich sestavených dat byl nakonec pro každý genom vytvořen konsenzus, čímž jsme získali jednotlivé sekvence pro všech 24 genomů ve formátu FASTA.

3.4 Hledání genů

Z referenčního genomu získaného z NCBI byly vyextrahovány jednotlivé geny, které byly následně vyhledávány v našich analyzovaných genomech. Vyhledávání bylo provedeno pomocí algoritmu BLAST (Basic Local Alignment Search Tool) [37], který slouží právě k nalezení podobných oblastí v biologických sekvencích.

Zde se však ukazuje nevýhoda skládání sekvenačních dat vůči referenci, jelikož pokud byly některé geny obsaženy pouze v analyzovaných genomech, a ne v referenci, nedošlo během skládání k jejich namapování a skončily v nenamapovaných čtení. Z tohoto důvodu by bylo lepší poskládat data *de novo*, nicméně zde bychom



Obr. 3.10: Ukázka zarovnání v programu Tablet.

neobdrželi jednu dlouhou sekvenci, ale velké množství někdy i krátkých kontigů, se kterými by se následně hůře pracovalo a celá analýza by byla výpočetně náročnější.

Nicméně abychom zjistili, kolik genů mohlo být obsaženo v analyzovaných genomech, a nikoliv v referenci, byla provedena následující analýza. Pro 4 vybrané genomy, kde každý náleží jinému melt-typu došlo nejprve k namapování jednotlivých čtení k referenčnímu genomu a k přesunutí nenamapovaných čtení do nového souboru, na němž bylo následně provedeno skládání *de novo* za pomocí programu CLC Genomics Workbench [38]. Výsledkem bylo získání v průměru 5000 kontigů o různém průměrném pokrytí (z angl. average coverage), z nichž byly vybrány ty, jejichž průměrné pokrytí dosahovalo hodnoty alespoň 100 a více, čímž bylo získáno průměrně 100 kontigů. Tyto kontigy byly poté vyhledány pomocí webové verze algoritmu BLAST, která je propojena s databází NCBI, a jako referenční organismus byla zvolena bakterie *Klebsiella pneumoniae*. Ze získaných výsledků jsme zjistili, že většina nenamapovaných čtení obsahuje úseky z různých plazmidů a pouze malá část (cca 10 %) obsahuje geny, jež se nacházejí na chromozomu a které nemá zvolená referenční sekvence. Jelikož tedy počet genů v nenamapovaných čtení není výrazně signifikantní, bylo vybráno mapování k referenci.

3.5 Navržený program

Program *find_variable_parts* byl vytvořen v prostředí MATLAB R2017a, přičemž při jeho tvorbě byl použit i Bioinformatics toolbox. Úkolem programu je projít všechny FASTA soubory, které obsahují geny z referenčního genomu, které byly nalezené pomocí algoritmu BLAST v analyzovaných genomech a z nich vybrat ty, které vykazují větší míru variability, a tudíž je možné je použít pro genotypizaci.

Samotný program se skládá z několika funkcí, jejichž princip je teoreticky popsán v následujících sekcích. Vývojový diagram navrženého algoritmu je uveden v příloze A.1.

3.5.1 Načtení sekvencí a kontrola obsahu

V prvním kroku dochází k načtení seznamu všech FASTA souborů pomocí funkce *fasta_from_current_dir*, čímž získáme strukturu obsahující především názvy jednotlivých souborů, jejich velikost a počet genů.

Následně je pomocí funkce *load_sequences* načten vždy jeden soubor, který obsahuje jeden gen v analyzovaných genomech, čímž získáme jednotlivé sekvence včetně hlaviček. Zároveň je před každým načtením ověřeno, zda soubor obsahuje sekvence. Pokud ne, je takovýto soubor zapsán do seznamu *empty_genes*, což znamená, že se gen z reference nevyskytuje v našich analyzovaných genomech. Seznam těchto genů a jejich genových produktů je uveden v tab. B.2. V opačném případě je sekvence s hlavičkami nahrána a zároveň je určen počet sekvencí v souboru.

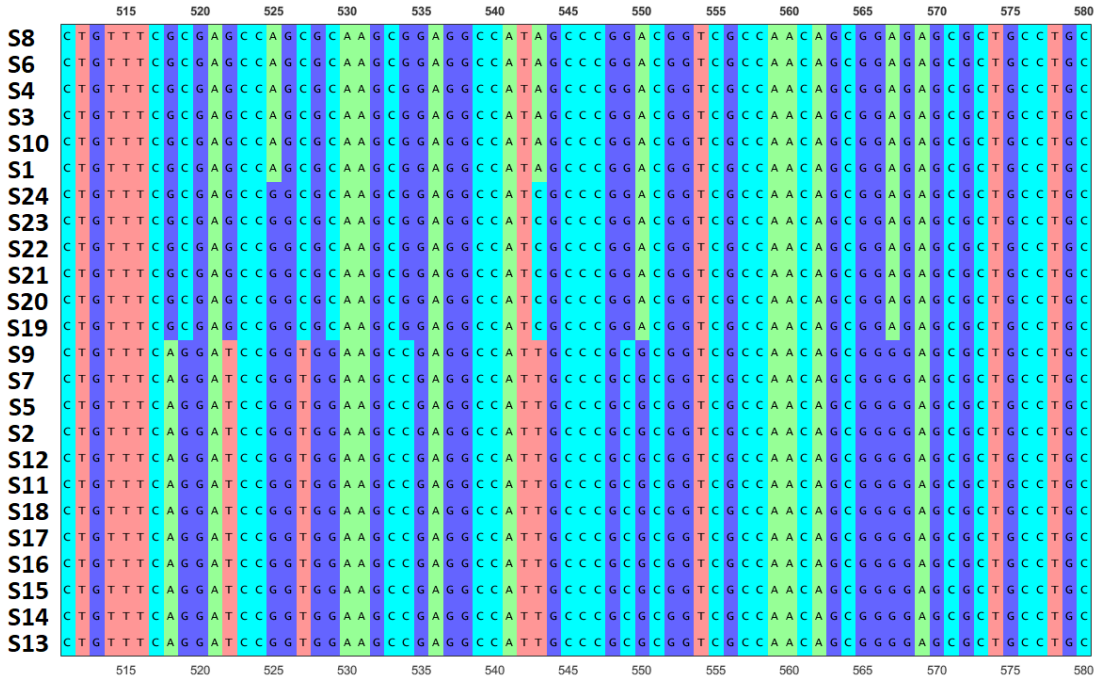
Nakonec je ověřeno, zda je v souboru přítomen gen u všech analyzovaných genomů. Jestliže ano, postupuje soubor v analýze dál, v opačném případě je soubor zapsán do seznamu *missing_gene_in_sequence* a již nevstupuje do dalšího kroku, protože genové úseky, které chceme použít pro genotypizaci, se musí vyskytovat u všech genomů.

Program *find_variable_part* probíhá postupně vždy pro jeden soubor, tzn. pro jeden gen.

3.5.2 Zarovnání sekvencí

Sekvence v rámci jednoho genu jsou vícenásobně zarovnány pomocí příkazu *multialign*. Jelikož se jedná o nukleotidové sekvence je jako skórovací matice použita 'NUC44'. Příklad úseku ze zarovnaného genu je uveden na obr. 3.11, kde vidíme i variabilní pozice, které mohou sloužit k rozlišení jednotlivých bakteriálních linií.

Následně je v rámci zarovnání jednoho genu určena délka každé sekvence bez zarovnávacích mezer. Pokud jsou získané délky stejně dlouhé, neprovádí se žádná akce



Obr. 3.11: Ukázka části zarovnaného genu.

a soubor s genem pokračuje dál. Avšak jestliže jsou délky různě dlouhé, dojde k ořezu zarovnání.

3.5.3 Ořezání zarovnaných sekvencí

K ořezání zarovnaných sekvencí dojde v případě, že jednotlivé sekvence daného genu nejsou stejně dlouhé, tzn. obsahují zarovnávací mezery. Nejprve se pomocí funkce *is_cut_effective* zjistí, zda existuje úsek, který je společný pro všechny sekvence. Pokud se žádný nenajde, nemá smysl ořezávat zarovnání a soubor s genem je vyloučen z analýzy.

Jestliže je nalezen společný úsek, dojde k ořezání pomocí funkce *neccessary_to_cut_sequences*. Pro zarovnané sekvence jsou zjištěny oblasti, ve kterých jsou přítomny všechny sekvence (bez zarovnávacích mezer) a následně jsou takovéto části zarovnání ořezány a postupují v analýze dále.

Výstupem funkce je zarovnání, které obsahuje sekvence bez zarovnávacích mezer, přičemž pro jeden soubor můžeme získat i větší počet zarovnání.

3.5.4 Variabilita genu

Abychom mohli z analyzovaných genů vybrat potencionálně vhodné geny pro genotypizaci je provedeno testování variability. Nejprve je pomocí funkce *count_procentual*

_changes_for_alignment, do které vstupuje zarovnání daného genu, vypočteno, zda je gen alespoň ze 3 % variabilní. Variabilita je spočtena jako počet pozic v zarovnání, kde se nachází alespoň jedna rozdílná báze vůči ostatním, dělená délkou zarovnání a poté převedena na procentuální hodnotu.

Následně je zjištěno, zda vybraná variabilní zarovnání mají délku alespoň 70 bp. Jestliže nesplňují uvedené kritérium, jsou takováto zarovnání vyřazena.

Pomocí funkce *count_alignment_index* jsou vybrány variabilní úseky ze zarovnaných genů. Pokud je délka zarovnání mezi 70 až 100 bp, je jednoduše spočten počet variabilních pozic (pozice, kde se nachází alespoň jeden jiný nukleotid oproti ostatním). Pokud je takovýchto pozic alespoň 10, je gen označen jako potencionálně vybraný a postupuje v algoritmu do dalšího kroku.

Pro zarovnání delší než 100 bp, je též proveden výpočet variabilních pozic, avšak s tím rozdílem, že výpočet probíhá v posuvném okně o délce 100 a s překryvem 1 nukleotid. S oknem začínáme na začátku zarovnání a vždy je v něm vypočítán počet variabilních pozic. Pokud je výsledný počet alespoň 10 bází, je pro první okno uložena hodnota 1, pokud zde není dosaženo požadované variability je uložena hodnota 0. Následně se okno posune o jeden nukleotid dále a výpočet je proveden znovu. Jako výsledek získáme vektor 0 a 1, přičemž 1 značí variabilní oblasti. Pro okna označené 1 je získán jejich začátek a konec v zarovnaných sekvencích. Pokud se nachází několik 1 za sebou, je získán index začátku pro začátek prvního okna a koncový index pro konec posledního okna. Jako výsledek dostaneme seznam indexů variabilních oblastí v zarovnání, pomocí kterého lze následně získat sekvence variabilních úseků.

Zvolení jednotlivých prahů byla vybráno na základě testování, které je uvedeno v kapitole 3.6.1.

3.5.5 Fylogenetická analýza

Pro analyzované sekvence je nejprve zjištěno, zda neobsahují na některé pozici bázi označenou N. Pokud ano, je takovýto gen, který obsahuje v některé sekvenci uvedenou bázi, vyloučen z analýzy.

Pro variabilní úseky sekvencí je následně spočítána evoluční vzdálenost pomocí modelu Jukes-Cantor dle následujícího vzorce:

$$d = -\frac{3}{4} \cdot ln(1 - p \cdot \frac{4}{3}), \tag{3.1}$$

kde d je hledaná evoluční vzdálenost a p je proporcionální vzdálenost mezi dvěma sekvencemi (počet bodových mutací vztažený na délku). Ze získaných evolučních vzdáleností je zkonstruován fylogenetický strom pomocí metody UPGMA.

Následně je na získaných datech zkontrolováno, zda byly jednotlivé sekvence rozděleny do správných skupin, které odpovídají jednotlivým melt-typům. K tomu je použit příkaz *cluster*, který označí sekvence hodnotou odpovídajícího shluku (1-4). Pro každý shluk je poté zjištěno, zda obsahuje všechny sekvence daného melt-typu. Pokud ano, postupuje variabilní úsek v analýze dále.

3.5.6 Teplota tání

Pro všechny sekvence v rámci jednoho genu je vypočtena teplota tání metodou nejbližšího souseda [39], která je označena za nejvíc přesnou metodu pro výpočet.

Teplota je stanovena dle následujícího vzorce:

$$T_m = \frac{\Delta H}{A + \Delta S + R \cdot ln(\frac{C}{4}))} - 273,15 + 16,6 \cdot log\left[Na^+\right], \tag{3.2}$$

kde T_m je teplota tání, ΔH označuje změnu entalpie, A je konstanta (A = -0,0108 $kcal\ K^{-1} \cdot mol^{-1}$), ΔS značí změnu entropie, R je plynová konstanta (R = 0,00199 $kcal\ K^{-1} \cdot mol^{-1}$), C popisuje koncentraci oligonukleotidů (byla použita hodnota 0,0000005 $mol\ L^{-1}$), -273,15 je konverzní faktor k přeměně očekávané teploty v K do $°C$ a Na^+ je koncentrace sodíkových iontů (byla použita hodnota 0,05 $mol \cdot L^{-1}$).

Hodnoty entalpie a entropie pro nejbližší sousedy byly použity podle tab. 3.3.

Jestliže se v sekvencích vyskytla báze, která není uvedena v tabulce (např. R - značí A nebo G), je vypočtena hodnota entropie a entalpie jako průměr hodnot možných sousedů.

Poté je pro každý pár genomů určena vzdálenost d_{st} pomocí Euklidovské metriky jako:

$$d_{st}^2 = (x_s - x_t)(x_s - x_t)', \tag{3.3}$$

kde x_s a x_t jsou dvě různé teploty tání.

Následně je spočítána průměrná vzdálenost d mezi všemi dvojicemi objektů v každém shluku dle vzorce:

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}), \tag{3.4}$$

kde r a s jsou shluky, n_r a n_s je počet objektů v shluku r a s a x_{ri} a x_{sj} je i-tý nebo j-tý objekt ve shluku r nebo ve shluku s. Pro vytvoření dendrogramu byla použita metoda UPGMA. [40]

Po shlukové analýze bylo stejně jako u fylogenetické analýzy zjištěno, zda byla analyzovaná data správně rozdělena do čtyř shluků dle jednotlivých melt-typů. Pokud ano, jsou úseky označené jako variabilní.

Tab. 3.3: Hodnoty entalpie a entropie pro dvojice sousedních bází.

interakce	ΔH $[kcal \cdot mol^{-1}]$	ΔS $[kcal \cdot K^{-1} \cdot mol^{-1}]$
AA	-9,1000	-0,0240
AT	-8,6000	-0,0239
TA	-6,0000	-0,0169
CA	-5,8000	-0,0129
GT	-6,5000	-0,0173
CT	-7,8000	-0,0208
GA	-5,6000	-0,0135
CG	-11,9000	-0,0278
GC	-11,1000	-0,0267
GG	-11,0000	-0,0266
AC	-6,5000	-0,0173
AG	-7,8000	-0,0208
TC	-5,6000	-0,0135
TG	-5,8000	-0,0129
CC	-11,0000	-0,0266
TT	-9,1000	-0,0240

3.6 Testování programu

V následujících kapitolách jsou uvedeny získané výsledky včetně nastavení jednotlivých parametrů. Následně je zde popsáno testování získaných výsledků na nových genomech.

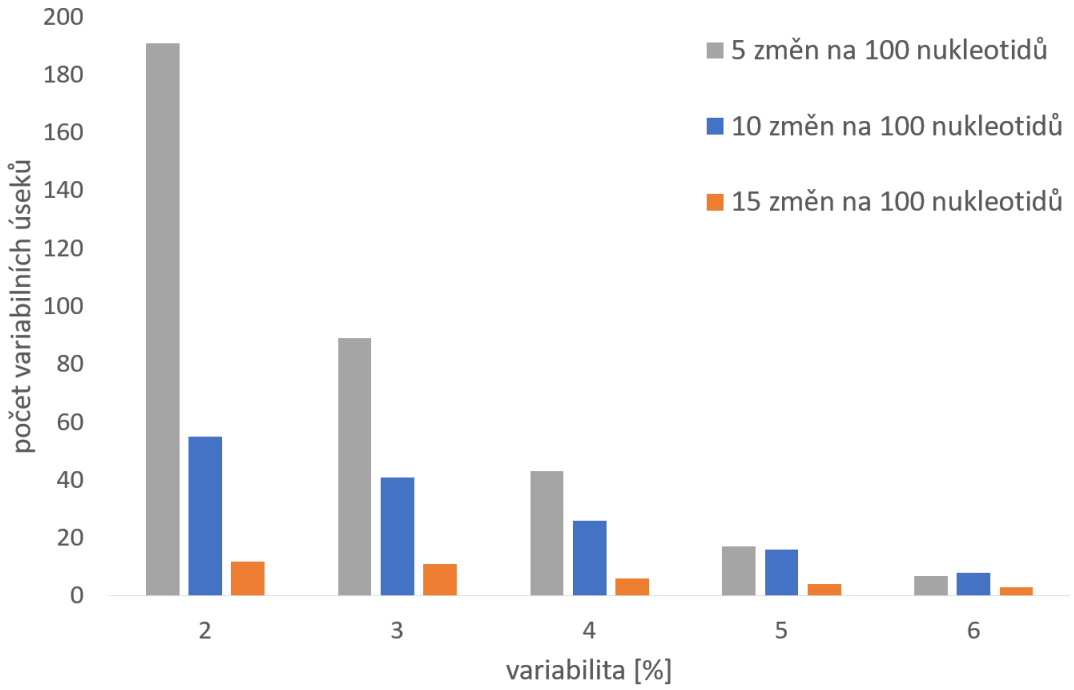
3.6.1 Nastavení parametrů

V rámci programu *find_variable_part* lze nastavit dva prahy, které nám umožňují vyfiltrovat méně variabilní geny. Prvním prahem, který se dá zvolit je variabilita genu, která je určena jako celkový počet pozic v zarovnání jednoho genu, kde se nachází alespoň jedna odlišná báze, podělený celkovou délkou zarovnání a převedený na procenta. V rámci testování bylo vyzkoušeno rozmezí hodnot od 2 % do 6 %.

Dalším parametrem, který lze zvolit je počet variabilních pozic v okně o délce 100 nukleotidů, pro který bude daný úsek vyhodnocen jako variabilní. Byly vyzkoušeny hodnoty 5, 10 a 15 pozic na 100 nukleotidů, přičemž pro úseky kratší než 100 nukleotidů, byly testovány hodnoty 5 a 10.

Počet nalezených genových úseků pro různá nastavení parametrů je zobrazen

na obr. 3.12 a kompletní výsledky jsou uvedeny v tab. B.3, tab. B.4, tab. B.5. Jako finální hodnoty prahů byly vybrány: 3 % variabilita a 10 variabilních pozic na 100 nukleotidů, protože zde vycházel optimální počet variabilních úseků. Pokud by byly hodnoty prahů nastaveny na příliš vysokou hodnotu, došlo by k nalezení menšího počtu variabilních genových úseků, které by nemusely být schopny rozlišit jednotlivé bakteriální linie ve všech případech. Naopak pokud by pro hodnoty prahů byly zvoleny nízké hodnoty, vedlo by to k nalezení velkého počtu variabilních úseků, a tudíž by se zvýšila časová náročnost provedení mini-MLST v laboratořích.



Obr. 3.12: Počet nalezených variabilních úseků pro různé hodnoty variability a počtu změn na 100 nukleotidů.

3.6.2 Výběr evolučního modelu

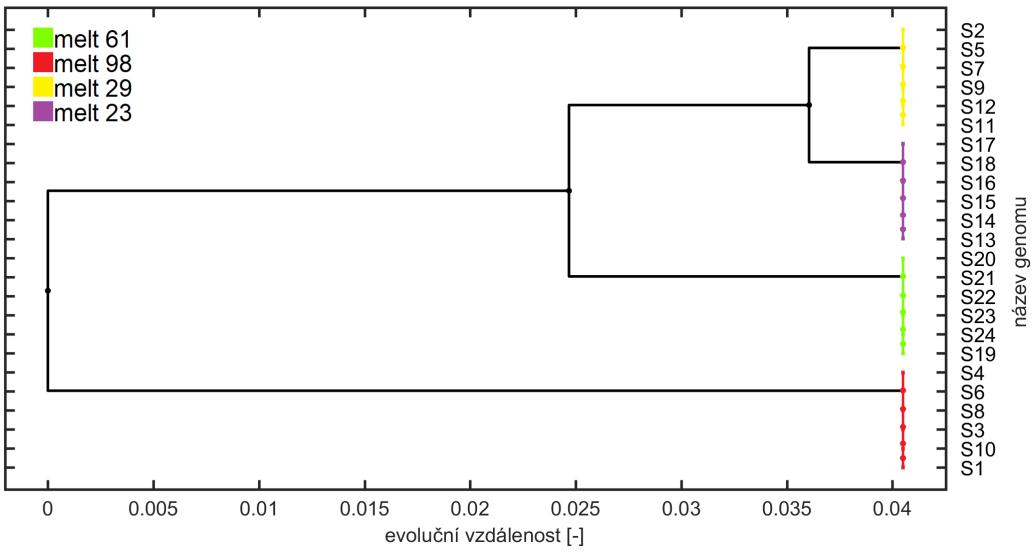
Pro výpočet evoluční vzdálenosti mezi sekvencemi, kterou používáme k vytvoření fylogenetických stromů, byl použit model Jukes-Cantor, který předpokládá, že se všechny nukleotidy vyskytují se stejnou frekvencí. Při použití tohoto modelu bylo získáno 41 variabilních genových úseků. Též byl otestován evoluční model Kimura, který je založen na předpokladu, že tranzice se vyskytují častěji než transverze. Nicméně z testovacích dat bylo potřeba odstranit sekvence, které obsahují jiné nukleotidy než základní čtyři (A, C, G, T), jelikož s nimi použitý model neumí pracovat.

Bylo zjištěno, že výběr evolučního modelu signifikantně nemění počet nalezených variabilních genových úseků, a proto byl pro jednoduchost zvolen model Jukes-Cantor.

3.6.3 Fylogenetická analýza a analýza teplot tání

Fylogenetická analýza je schopna rozlišit jednotlivé linie na základě evoluční vzdálenosti. Tedy jako výsledek analýzy obdržíme fylogenetický strom, kde příbuzné sekvence jsou blíže u sebe než vzdálené a vycházejí ze stejného uzlu.

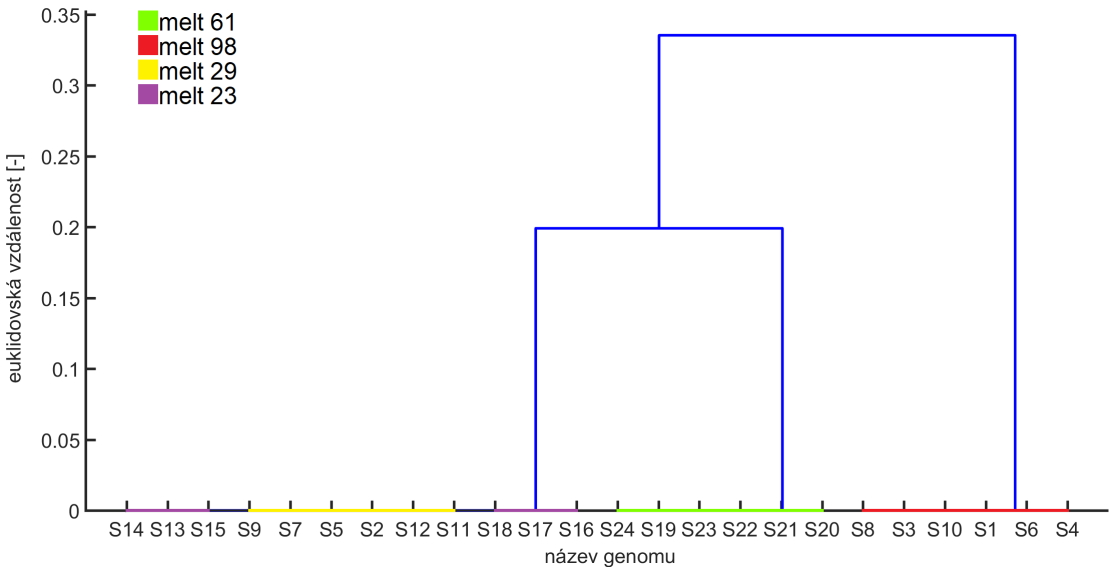
V našem případě byla za fylogenetickou část zařazena analýza, která je založena na teplotách tání sekvencí, jelikož během testování bylo zjištěno, že ačkoliv fylogeneticky lze sekvence správně roztřídit do skupin odpovídající jednotlivým melt-typům, mohou mít podobnou teplotu tání, a tudíž by během typizaci za použití mini-MLST, která analyzuje právě křivky tání, nedošlo ke správnému odlišení bakteriálních linií. Příklad takového genového úseku je uveden na obr. 3.13, kde ve fylogenetickém stromě jasně vidíme čtyři vytvořené shluky a na obr. 3.14, kde po analýze teplot tání byly identifikovány pouze tři.



Obr. 3.13: Fylogenetický strom.

3.6.4 Výsledky

Získané výsledky jsou uvedeny v tab. 3.4. Délka nalezených úseků se pohybuje v rozmezí 100 - 346 bp a počet variabilních pozic je od 10 do 43. Různorodé oblasti byly



Obr. 3.14: Shluková analýza teplot tání.

nalezeny ve 38 genech, přičemž pro 3 geny byly nalezeny 2 variabilní oblasti. V tab. B.6 je uveden seznam genů, které obsahují variabilní úseky a jejich genové produkty.

3.6.5 Testování variabilních úseků

Variabilní úseky byly testovány na dalších 12 genomech bakterie *Klebsiella pneumoniae*. Jednotlivé názvy genomů a jejich melt-typy jsou uvedeny v tab. 3.5.

Nejprve byly pomocí nástroje BLAST vyhledány variabilní geny, které obsahují variabilní genové úseky, v analyzovaných genomech. Poté byly v nalezených genech vyhledány variabilní genové úseky, a to opět pomocí uvedeného nástroje. Jako variabilní část, které se hledala, byl vždy použit variabilní úsek z genomu S1. V dalším kroku bylo zkontrolováno, zda se hledaná část nalézá u všech genomů a zda je všude stejně dlouhá. Bylo zjištěno, že úseky se nalézaly vždy u všech genomů, avšak u některých genových úseků bylo potřeba provést ořez, přičemž se postupovalo stejně jako je uvedeno v kapitole 3.5.3. Následně byly variabilní úseky analyzovány.

V prvním kroku byla opět provedena fylogenetické analýza. Pro získané výsledky bylo zkontrolováno, zda byly jednotlivé genomy správně zařazeny do shluku, který odpovídá danému melt-typu. Ze 41 analyzovaných variabilních úseků byl jeden úsek vyloučen, jelikož obsahoval báze označené písmenem N. Celkové výsledky pro všechny variabilní úseky jsou uvedeny v tab. 3.6, přičemž melt 1 obsahuje genomy S25, S26, S35, S36, do meltu 2 patří S27, S28, S31, S32, melt 3 se skládá z S29 a S30 a do posledního meltu 4 patří S33 a S34. Úhlopříčně jsou vyznačena procenta správně klasifikovaných genomů dle jednotlivých melt-typů. Jak vidíme, hodnoty se

Tab. 3.4: Seznam nalezených variabilních úseků schopných správně rozlišit jednotlivé melt-typy u genomů S01 - S24.

číslo genu	**délka variabilního úseku [bp]**	**počet variabilních pozic**	**číslo genu**	**délka variabilního úseku [bp]**	**počet variabilních pozic**
224	307	37	2738	222	23
250	172	14	2745	177	17
448	175	15	2752	193	17
449	114	11	2840	203	20
536	177	16	2883	100	10
569	103	10	3010	161	13
582	356	33	3019	167	15
582_2	217	22	3030	106	11
584	157	13	3052	195	17
584_2	298	28	3065	106	10
632	164	16	3534	225	25
789	218	23	3554	145	12
863	121	10	3554_2	100	10
2067	174	15	3715	346	41
2296	109	10	4372	286	27
2315	196	18	4641	144	12
2376	321	36	5230	199	19
2380	343	43	5235	130	11
2702	178	21	5236	106	11
2720	157	14	5239	154	14
2728	121	11			

Tab. 3.5: Seznam testovaných genomů a jejich melt-typy.

genom	**melt-typ**	**genom**	**melt-typ**
S25	61	S31	14
S26	61	S32	14
S27	14	S33	16
S28	14	S34	16
S29	98	S35	61
S30	98	S36	61

pohybují průměrně okolo 84,50 %. Nejlepší hodnoty byly získány pro melt 1, kde bylo správně zařazeno 93,13 % a nejhorší hodnoty byly obdrženy pro melt 4, kde bylo správně klasifikováno pouhých 76,25 %.

Ve druhém kroku byly pro všechny variabilní úseky kromě jednoho, který byl vyloučen v předchozím kroku, vypočteny teploty tání a následně byla provedena shluková analýza. Opět bylo zkontrolováno, zda jsou genomy stejného melt-typu ve stejném shluku. Získané výsledky jsou uvedeny v tab. 3.7. Zařazení genomů do melt-typů je stejné jako v případě fylogenetické analýzy a úhlopříčně jsou opět vyznačena procenta, která označují počet správně klasifikovaných genomů do meltů, přičemž průměr hodnot je zhruba 83,4 %. Nejlépe vyšel melt 2, kde bylo správně zařazeno 96,25 % a nejhůře opět dopadl melt 4, kde bylo správně rozpoznáno pouhých 61,25 %.

Tab. 3.6: Matice zmatení pro výsledky získané z fylogenetické analýzy.

		předpokládané			
		melt 1	**melt 2**	**melt 3**	**melt 4**
skutečné	**melt 1**	**93,13 %**	2,50 %	1,88 %	2,50 %
	melt 2	10,00 %	**88,75 %**	1,25 %	0,00 %
	melt 3	0,00 %	20,00 %	**80,00 %**	0,00 %
	melt 4	5,00 %	16,25 %	2,50 %	**76,25 %**

Tab. 3.7: Matice zmatení pro výsledky získané z analýzy teplot tání.

		předpokládané			
		melt 1	**melt 2**	**melt 3**	**melt 4**
skutečné	**melt 1**	**91,25 %**	4,38 %	0,63 %	3,75 %
	melt 2	2,50 %	**96,25 %**	1,25 %	0,00 %
	melt 3	5,00 %	10,00 %	**85,00 %**	0,00 %
	melt 4	8,75 %	22,50 %	7,50 %	**61,25 %**

Zároveň bylo zjištěno, že 16 variabilní úseků dokáže na základě fylogenetické analýzy správně odlišit jednotlivé melt-typy a 14 variabilní úseků rozliší melt-typ na základě analýzy teplot tání (viz tab. B.7). Porovnáním úseků získaných z jednotlivých analýz bylo zjištěno, že 14 úseků se shoduje v obou případech, a proto byly vybrány jako finální úseky, které jsou schopny správně provést genotypizaci a rozlišit jednotlivé linie bakterie *Klebsiella pneumoniae*. Čísla genů, které obsahují variabilní úseky, délka úseků a počet variabilních pozic je uveden v tab. 3.8.

Tab. 3.8: Seznam variabilních genových úseků schopných správně rozlišit jednotlivé melt-typy u genomů S25 - S36.

číslo genu	**délka variabilního úseku [bp]**	**počet variabilních pozic**	**číslo genu**	**délka variabilního úseku [bp]**	**počet variabilních pozic**
448	175	14	3010	161	13
584_2	296	27	3019	167	7
2296	109	10	3030	105	9
2315	168	12	3554	145	9
2376	321	35	3554_2	99	9
2840	203	18	4372	276	23
2883	90	5	5239	154	14

4 ZÁVĚR

V první polovině teoretické části jsou popsány typizační metody, včetně jejich výhod a nevýhod. Následně je čtenář seznámen s bakteriálním genomem, s organismem *Klebsiella pneumoniae* včetně jeho genomu a je mu objasněna antibiotická rezistence výše uvedené bakterie.

V praktické části práce je popsáno otestování vstupních dat a jejich následné složení, poté je představen navržený program pro hledání variabilních genových úseků. V další sekci je popsáno nastavení jednotlivých parametrů a zdůvodněn jejich výběr. Následně jsou uvedeny získané výsledky, jež zahrnují vybrané variabilní genové úseky včetně jejich délky a počtu variabilních pozic. Tyto úseky jsou poté otestovány na dalších 12 genomech. V poslední části jsou uvedeny genové úseky, které byly vybrány jako finální a jsou tedy schopny rozlišit jednotlivé bakteriální linie, respektive melt-typy.

Program *find_variable_parts* byl spuštěn na 24 genomech bakterie *Klebsiella pneumoniae*, které patřily do 4 melt-typů. Jako výsledek bylo obdrženo 41 variabilních genových úseků, přičemž byly nalezeny u celkem 38 genů (u 3 genů byly identifikovány 2 variabilní oblasti).

Výsledky byly otestovány na dalších 12 genomech, které patřily opět do 4 melt-typů. Pro každý variabilní úsek byl zkonstruován fylogenetický strom a strom vzniklý shlukovou analýzou teplot tání. Ze získaných dat bylo zjištěno, že 16 variabilních úseků správně rozliší jednotlivé melt-typy na základě fylogenetické analýzy a 14 úseků je schopno klasifikovat bakteriální linie pomocí analýzy teplot tání. Porovnáním obou skupin bylo tedy zjištěno, že 14 genových úseků je schopno správně rozlišit bakteriální linie u testovacích dat, a proto byly právě tyto úseky vybrány jako finální.

Zvolené variabilní genové úseky by měly být v budoucnu použity pro typizaci pomocí mini-multilokusové sekvenční typizace, kde by se měly zařadit vedle provozních genů a tím zvýšit rozlišovací schopnost uvedené metody.

Rezistence vůči antibiotikům je u bakterie *Klebsiella pneumoniae* stále na vzestupu a též roste počet multirezistentních linií, které mohou v budoucnu způsobit těžko zvladatelné epidemie. Z uvedeného plyne, že je nezbytné stále pracovat na nových způsobech typizace, popřípadě zkvalitňovat již existující metody, abychom byli schopni v případě potřeby včas identifikovat a izolovat právě rezistentní linie, přičemž kritickými parametry je nadále přesnost, rychlost a dostupnost.

LITERATURA

[1] SABAT, A.J., A. BUDIMIR, D. NASHEV, et al. *Overview of molecular typing methods for outbreak detection and epidemiological surveillance.* Euro Surveill [online]. 2013, 18(4) [cit. 2017-10-13]. Dostupné z URL: <http://www.eurosurveillance.org/images/dynamic/EE/V18N04/art20380.pdf>.

[2] FOLEY, Steven L., Aaron M. LYNNE a Rajesh NAYAK. *Molecular typing methodologies for microbial source tracking and epidemiological investigations of Gram-negative bacterial foodborne pathogens.* Infection, Genetics and Evolution [online]. 2009, 9(4), 430-440 [cit. 2017-12-08]. DOI: 10.1016/j.meegid.2009.03.004. ISBN 10.1016/j.meegid.2009.03.004. Dostupné z URL: <http://linkinghub.elsevier.com/retrieve/pii/S1567134809000495>.

[3] HE, Yiping, Yanping XIE a Sue REED. *Pulsed-Field Gel Electrophoresis Typing of Staphylococcus aureus Isolates* [online]. 2014, 103-111 [cit. 2017-10-27]. DOI: 10.1007/978-1-62703-664-1_6. ISBN 10.1007/978-1-62703-664-1_6. Dostupné z URL: <http://link.springer.com/10.1007/978-1-62703-664-1_6>.

[4] OLIVE, D.Michael a Pamela BEAN. *Principles and Applications of Methods for DNA-Based Typing of Microbial Organisms.* Journal of Clinical Microbiology [online]. 1999, 37(6), 1661-1669 [cit. 2017-10-27]. Dostupné z URL: <http://jcm.asm.org/content/37/6/1661>.

[5] ResearchGate. *Schematic overview of the Pulse Field Gel Electrophoresis Process* [online]. 2016 [cit. 2017-12-07]. Dostupné z URL: <https://www.researchgate.net/profile/Christian_Vinueza/publication/313905045/figure/fig1/AS:465057122394112@1487889666861/Figure-1-Schematic-overview-of-the-Pulse-Field-Gel-\Electrophoresis-Process-Source-CDC.jpg>.

[6] RANJBAR, Reza, Ali KARAMI, Shohreh FARSHAD, Giovanni M. GIAMMANCO a Caterina MAMMINA. *Typing methods used in the molecular epidemiology of microbial pathogens: a how-to guide.* The new microbiologica [online]. 2014, 37(1-15) [cit. 2017-11-03]. Dostupné z URL: <http://www.newmicrobiologica.org/pub/allegati_pdf/2014/1/1.pdf>.

[7] TABIT, Frederick Tawi. *Advantages and limitations of potential methods for the analysis of bacteria in milk: a review.* Journal of Food Science and Technology [online]. 2016, 53(1), 42-49 [cit. 2017-12-06]. DOI: 10.1007/s13197-015-1993-y.

ISSN 0022-1155. Dostupné z URL: <http://link.springer.com/10.1007/s13197-015-1993-y>.

[8] SlidePlayer *Procedures in RFLP.* [online]. [cit. 2017-12-05]. Dostupné z URL: <http://player.slideplayer.com/17/5310705/data/images/img1.jpg>.

[9] CASTRO-ESCARPULLI, Graciela, Nayelli Maribel ALONSO-AGUILAR, Gildardo Rivera SÁNCHEZ, et al. *Identification and Typing Methods for the Study of Bacterial Infections: a Brief Review and Mycobacterial as Case of Study. Archives of Clinical Microbiology.* Euro Surveill [online]. 2015, 7(1:3) [cit. 2017-11-03]. ISSN 1989-8436. Dostupné z URL: <http://www.acmicrob.com/microbiology/identification-and-typing-methods-for-thestudy-of-bacterial-\infections-a-brief-reviewand-mycobacterial-as-case-of-study.pdf>.

[10] Amplified Fragment Length Polymorphism (AFLP). *Wageningen University & Research.* Euro Surveill [online]. [cit. 2017-11-03]. Dostupné z URL: <https://www.wur.nl/en/show/Amplified-Fragment-Length-Polymorphism-AFLP.htm>.

[11] RUPPITSCH, Werner. *Molecular typing of bacteria for epidemiological surveillance and outbreak investigation.* Die Bodenkultur: Journal of Land Management, Food and Environment [online]. 2016, 67(4) [cit. 2017-12-01]. DOI: 10.1515/boku-2016-0017. ISBN 10.1515/boku-2016-0017. ISSN 0006-5471. Dostupné z URL: <https://www.degruyter.com/view/j/boku.2016.67.issue-4/boku-2016-0017/boku-2016-0017.xml>.

[12] MAIDEN, Martin C.J., Jane A. BYGRAVES, Edward FEIL, et al. *Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.* Proc Natl Acad Sci U S A [online]. 1998, 95, 3140–3145 [cit. 2017-12-01]. Dostupné z URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC19708/pdf/pq003140.pdf>.

[13] LI, Wenjun, Didier RAOULT a Pierre-Edouard FOURNIER. *Bacterial strain typing in the genomic era.* FEMS Microbiol Reviews [online]. 2009, 33(5), 892-916 [cit. 2017-12-01]. DOI: 10.1111/j.1574-6976.2009.00182.x. ISBN 10.1111/j.1574-6976.2009.00182.x. Dostupné z URL: <https://academic.oup.com/femsre/article-lookup/doi/10.1111/j.1574-6976.2009.00182.x>.

[14] PREMIER Biosoft. *High Resolution Melting Analysis* [online]. [cit. 2017-12-06]. Dostupné z URL: <http://www.premierbiosoft.com/tech_notes/high_resolution_melting_analysis.html>.

[15] BRHELOVA, Eva, Iva KOCMANOVA, Zdenek RACIL, Marketa HANSLIANOVA, Mariya ANTONOVA, Jiri MAYER a Martina LENGEROVA. *Validation of Minim typing for fast and accurate discrimination of extended-spectrum, beta-lactamase-producing Klebsiella pneumoniae isolates in tertiary care hospital* [online]. 2016 [cit. 2017-10-13]. DOI: 10.1016/j.diagmicrobio.2016.03.010. ISBN 10.1016/j.diagmicrobio.2016.03.010. Dostupné z URL: <http://linkinghub.elsevier.com/retrieve/pii/S0732889316300505>.

[16] ANDERSSON, Patiyan, Steven Y. C. TONG, Jan M. BELL, John D. TURNIDGE, Philip M. GIFFARD a Igor MOKROUSOV. *Minim Typing – A Rapid and Low Cost MLST Based Typing Tool for Klebsiella pneumoniae* [online]. 2012 [cit. 2017-10-13]. DOI: 10.1371/journal.pone.0033530. ISBN 10.1371/journal.pone.0033530. Dostupné z URL: <http://dx.plos.org/10.1371/journal.pone.0033530>.

[17] RICHARDSON, L.J., S.Y.C. TONG, R.J. TOWERS, et al. *Preliminary validation of a novel high-resolution melt-based typing method based on the multilocus sequence typing scheme of Streptococcus pyogenes.* Clinical Microbiology and Infection [online]. 2011, 17(9), 1426-1434 [cit. 2017-10-13]. DOI: 10.1111/j.1469-0691.2010.03433.x. ISSN 1198743x. Dostupné z URL: <http://linkinghub.elsevier.com/retrieve/pii/S1198743X14612295>.

[18] TONG, Steven Y. C., Shirley XIE, Leisha J. RICHARDSON, et al. *High-Resolution Melting Genotyping of Enterococcus faecium Based on Multilocus Sequence Typing Derived Single Nucleotide Polymorphisms* PLoS ONE [online]. 2011, 6(12), 1-8 [cit. 2018-01-02]. DOI: 10.1371/journal.pone.0029189. ISBN 10.1371/journal.pone.0029189. Dostupné z URL: <http://dx.plos.org/10.1371/journal.pone.0029189>.

[19] JOENSEN, K. G., F. SCHEUTZ, O. LUND, H. HASMAN, R. S. KAAS, E. M. NIELSEN a F. M. AARESTRUP. *Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli* Journal of Clinical Microbiology [online]. 2014, 52(5), 1501-1510 [cit. 2017-12-02]. DOI: 10.1128/JCM.03617-13. ISSN 0095-1137. Dostupné z URL: <http://jcm.asm.org/cgi/doi/10.1128/JCM.03617-13>.

[20] Centers for Disease Control and Prevention. *Whole Genome Sequencing (WGS)* [online]. 2016 [cit. 2017-12-07]. Dostupné z URL: <https://www.cdc.gov/pulsenet/images/genome-sequencing-508c-900px.jpg>.

[21] NEČAS, Oldřich. *Obecná biologie pro lékařské fakulty.* 3. přeprac. vyd., V nakl. H. Jinočany: H, 2000, 554 s. ISBN 80-860-2246-3.

[22] CAMPBELL, Neil A a Jane B REECE. *Biologie.* Vyd. 1. Brno: Computer Press, c2006, xxxiv, 1332 s. ISBN 80-251-1178-4.

[23] National Center for Biotechnology Information *Klebsiella pneumoniae* [online]. [cit. 2017-10-13]. Dostupné z URL: <https://www.ncbi.nlm.nih.gov/genome/?term=klebsiella+pneumoniae>.

[24] QURESHI, Shahab. *Klebsiella Infections* [online]. [cit. 2017-10-13]. Dostupné z URL: <https://emedicine.medscape.com/article/219907-overview>.

[25] PODSCHUN, R. a U. ULLMANN. *Klebsiella spp. as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors.* Clinical Microbiology Reviews [online]. 1998, 11(4), 589-603 [cit. 2017-10-13]. Dostupné z URL: <http://cmr.asm.org/content/11/4/589.full.pdf+html>.

[26] WYRES, Kelly L. a Kathryn E. HOLT. *Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones.* Trends in Microbiology [online]. 2016, 24(12), 944-956 [cit. 2018-04-04]. DOI: 10.1016/j.tim.2016.09.007. ISSN 0966842X. Dostupné z URL: <http://linkinghub.elsevier.com/retrieve/pii/S0966842X16301391>.

[27] PACZOSA, Michelle K. a Joan MECSAS. *Klebsiella pneumoniae: Going on the Offense with a Strong Defense.* Microbiology and Molecular Biology Reviews [online]. 2016, 80(3), 629-661 [cit. 2018-04-11]. DOI: 10.1128/MMBR.00078-15. ISSN 1092-2172. Dostupné z URL: <http://mmbr.asm.org/lookup/doi/10.1128/MMBR.00078-15>.

[28] *Antimicrobial resistance: global report on surveillance* [online]. Geneva, Switzerland: World Health Organization, 2014 [cit. 2018-04-11]. ISBN 978-924-1564-748. Dostupné z URL: <http://apps.who.int/iris/bitstream/handle/10665/112642/9789241564748_eng.pdf?sequence=1>.

[29] Surveillance Atlas of Infectious Diseases. *European Centre for Disease Prevention and Control* [online]. Solna (Sweden), c2018 [cit. 2018-04-11]. Dostupné z URL: <https://ecdc.europa.eu/en/surveillance-atlas-infectious-diseases>.

[30] NYKRÝNOVÁ, Markéta a Denisa MADĚRÁNKOVÁ. *Genotyping of Klebsiella pneumoniae isolates.* In: Proceedings of the 24th Conference STUDENT EEICT 2018. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2018, s. 264-266. ISBN 978-80-214-5614-3.

[31] ALIKHAN, Nabil-Fareed, Nicola K PETTY, Nouri L BEN ZAKOUR a Scott A BEATSON. *BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons.* BMC Genomics [online]. 2011, 12(1), - [cit. 2017-12-10]. DOI: 10.1186/1471-2164-12-402. ISSN 1471-2164. Dostupné z URL: <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-402>.

[32] ANDREWS, Simon. *FastQC: A Quality Control tool for High Throughput Sequence Data* [online]. [cit. 2017-12-10]. Dostupné z URL: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

[33] EWELS, Philip, Måns MAGNUSSON, Sverker LUNDIN a Max KÄLLER. *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics [online]. 2016, 32(19), 3047-3048 [cit. 2018-01-03]. DOI: 10.1093/bioinformatics/btw354. ISSN 1367-4803. Dostupné z URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354>.

[34] LI, H. a R. DURBIN. *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics [online]. 2009, 25(14), 1754-1760 [cit. 2017-12-12]. DOI: 10.1093/bioinformatics/btp324. ISSN 1367-4803. Dostupné z URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324>.

[35] LI, H., B. HANDSAKER, A. WYSOKER, et al. *The Sequence Alignment/-Map format and SAMtools.* Bioinformatics [online]. 2009, 25(16), 2078-2079 [cit. 2018-01-03]. DOI: 10.1093/bioinformatics/btp352. ISSN 1367-4803. Dostupné z URL: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352>.

[36] MILNE, I., G. STEPHEN, M. BAYER, P. J. A. COCK, L. PRITCHARD, L. CARDLE, P. D. SHAW a D. MARSHALL. *Using Tablet for visual exploration of second-generation sequencing data: simple prokaryote genome comparisons.* Briefings in Bioinformatics [online]. 2013, 2000, 14(2), 193-202 [cit. 2018-01-03]. DOI: 10.1093/bib/bbs012. ISSN 1467-5463. Dostupné z URL: <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs012>.

[37] ALTSCHUL, Stephen F., Warren GISH, Webb MILLER, Eugene W. MYERS a David J. LIPMAN. *Basic local alignment search tool.* Journal of Molecular Biology [online]. 1990, 215(3), 403-410 [cit. 2018-04-18]. DOI: 10.1016/S0022-2836(05)80360-2. ISSN 00222836. Dostupné z URL: <http://linkinghub.elsevier.com/retrieve/pii/S0022283605803602>.

[38] QIAGEN Bioinformatics. *CLC Genomics Workbench 9.5.3.* [online]. c2017 [cit. 2018-05-10]. Dostupné z URL: <https://www.qiagenbioinformatics.com/products/clc-genomics-workbench>.

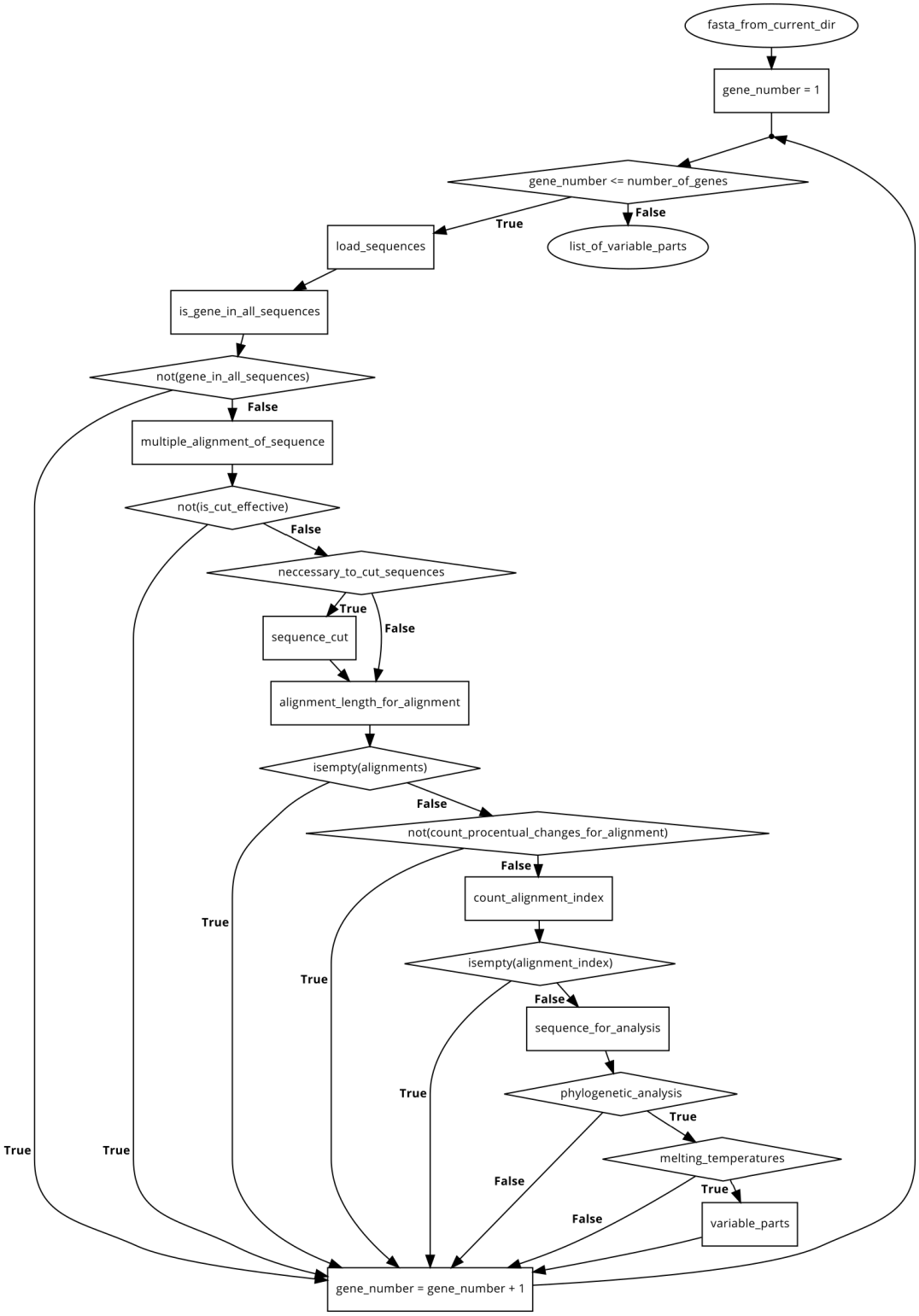
[39] Sigma-Aldrich. *Oligonucleotide Melting Temperature.* [online]. Darmstadt: Merck, c2018 [cit. 2018-03-11]. Dostupné z URL: <https://www.sigmaaldrich.com/technical-documents/articles/biology/oligos-melting-temp.html>.

[40] NYKRÝNOVÁ, Markéta, Denisa MADĚRÁNKOVÁ, Matěj BEZDÍČEK, Martina LENGEROVÁ a Helena ŠKUTKOVÁ. *Bioinformatic tools for genotyping of Klebsiella pneumoniae isolates.* In: Information Technologies in Medicine, 7th International Conference. ITIB 2018 Kamień Śląski, Polsko, 2018. (v tisku)

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

A	adenin
AFLP	polymorfismus délky amplifikovaných fragmentů
BAM	Binary Alignment Map
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
bp	páru bazí
C	cytosin
DNA	deoxyribonukleová kyselina
G	guanin
HRM	vysokorozlišovací analýza křivek tání
kb	kilobáze
mini-MLST	mini-multilokusová sekvenční typizace
MLST	multilokusová sekvenční typizace
NCBI	National Center for Biotechnology Information
NGS	sekvenování další generace
PCR	polymerázová řetězová reakce
PFGE	pulzní gelová elektroforéza
RAPD	náhodná amplifikace polymorfní DNA
RE	restrikční enzym
rep-PCR	repetitivní polymerázová řetězová reakce
RFLP	polymorfismus délky restrikčních fragmentů
SAM	Sequence Alignment Map
SNP	jednonukleotidový polymorfismus
T	thymin
UV	ultrafialové záření
VNTR	variabilní množství tandemových repetic
WGS	celogenomové sekvenování

A VÝVOJOVÝ DIAGRAM PROGRAMU



Obr. A.1: Vývojový diagram programu *find_variable_parts*.

B TABULKY

Tab. B.1: Počet celkových, namapovaných a nenamapovaných čtení pro analyzované genomy.

název genomu	**celkový počet čtení**	**počet namapovaných čtení**	**počet nenamapovaných čtení**
S01	3 305 724	2 274 686 / 68,81 %	1 031 038 / 31,19 %
S02	4 650 400	3 637 485 / 78,22 %	1 012 915 / 21,78 %
S03	2 662 038	1 928 886 / 72,46 %	733 152 / 27,54 %
S04	3 014 868	2 115 162 / 70,16 %	899 706 / 29,84 %
S05	3 780 800	2 922 275 / 77,29 %	858 525 / 22,71 %
S06	4 739 680	3 438 640 / 72,55 %	1 301 040 / 27,45 %
S07	3 905 676	3 028 505 / 77,54 %	877 171 / 22,46 %
S08	4 276 292	3 050 361 / 71,33 %	1 225 931 / 28,67 %
S09	2 296 492	1 738 282 / 75,69 %	558 210 / 24,31 %
S10	4 788 804	3 475 937 / 72,58 %	1 312 867 / 27,42 %
S11	5 474 258	4 304 504 / 78,63 %	1 169 754 / 21,37 %
S12	3 636 324	2 843 990 / 78,21 %	792 334 / 21,79 %
S13	2 753 496	2 129 558 / 77,34 %	623 938 / 22,66 %
S14	2 562 874	2 093 840 / 81,70 %	469 034 / 18,30 %
S15	2 299 112	1 855 290 / 80,70 %	443 822 / 19,30 %
S16	2 074 024	1 686 060 / 81,29 %	387 964 / 18,71 %
S17	1 939 256	1 587 903 / 81,88 %	351 353 / 18,12 %
S18	2 659 314	2 140 931 / 80,51 %	518 383 / 19,49 %
S19	1 971 630	1 567 503 / 79,50 %	404 127 / 20,50 %
S20	1 740 434	1 405 754 / 80,77 %	334 680 / 19,23 %
S21	1 817 226	1 472 310 / 81,02 %	344 916 / 18,98 %
S22	1 627 112	1 279 554 / 78,64 %	347 558 / 21,36 %
S23	2 060 586	1 499 025 / 72,75 %	561 561 / 27,25 %
S24	1 514 126	1 192 178 / 78,74 %	321 948 / 21,26 %

Tab. B.2: Seznam genů, které se nenacházejí v analyzovaných genomech a jejich genové produkty.

číslo genu	**genový produkt**
541	hypothetical protein
542	hypothetical protein
550	phage immunity repressor protein
1001	hypothetical protein
1005	phage immunity repressor protein
1007	hypothetical protein
1011	hypothetical protein
1012	cobyrinic acid a,c-diamide synthase
1238	putative phage-like protein
1240	hypothetical protein
1243	hypothetical protein
1246	bacteriophage CII family protein
1251	hypothetical protein
1252	hypothetical protein
1256	putative ninG protein
1261	hypothetical protein
1280	hypothetical protein
1283	hypothetical protein
1287	hypothetical protein
1363	LuxR family transcriptional regulator
1716	putative DinI-like damage-inducible protein
1717	hypothetical protein
1718	hypothetical protein
1720	terminase, ATPase subunit
1751	prophage P2 Ogr protein
2241	DNA-binding protein Roi
2250	hypothetical protein
2284	hypothetical protein
2291	hypothetical protein
2292	hypothetical protein
3561	hypothetical protein
3562	hypothetical protein
3563	family 2 glycosyl transferase
Pokračování na další stránce	

Tab. B.2 – pokračování z předcházející strany

číslo genu	**genový produkt**
3567	UDP-Gal::undecaprenolphosphate Gal-1-P transferase
4005	hypothetical protein
4007	hypothetical protein
4008	putative exonuclease CP81
4009	hypothetical protein
4015	hypothetical protein
4022	putative prophage phage head completion protein
4473	hypothetical protein
4480	hypothetical protein
4492	hypothetical protein
4495	hypothetical protein
4508	hypothetical protein
4510	hypothetical protein
4520	hypothetical protein
4522	hypothetical protein
4523	hypothetical protein
4526	hypothetical protein
4808	phage QLRG family, putative DNA packaging
4813	hypothetical protein
4815	putative prophage protein

Tab. B.3: Počet nalezených variabilních úseků pro různé hodnoty variabilit a pro práh 5 změn na 100 nukleotidů.

variabilita [%]	**počet variabilních úseků**
2	191
3	89
4	43
5	17
6	7

Tab. B.4: Počet nalezených variabilních úseků pro různé hodnoty variabilit a pro práh 10 změn na 100 nukleotidů.

variabilita [%]	**počet variabilních úseků**
2	55
3	41
4	26
5	16
6	8

Tab. B.5: Počet nalezených variabilních úseků pro různé hodnoty variabilit a pro práh 15 změn na 100 nukleotidů.

variabilita [%]	**počet variabilních úseků**
2	12
3	11
4	6
5	4
6	3

Tab. B.6: Seznam genů obsahujících variabilní úseky a jejich genové produkty.

číslo genu	**genový produkt**
224	DNA-binding response regulator in two-component regulatory system with ZraS
250	hypothetical protein
448	putative cationic amino acid transport protein
449	putative helix-turn-helix AraC-type transcriptional regulator
536	putative SN-glycerol-3-phosphate transport system permease
569	hypothetical protein
582	RND family efflux transporter MFP subunit
584	copper/silver efflux system outer membrane protein CusC
632	4-hydroxyphenylacetate catabolism
789	MFS family transporter
863	ABC transporter permease
2067	putative LysR-family transcriptional regulator
2296	hypothetical protein
2315	type VI secretion-associated protein
2376	3-hydroxybutyryl-CoA dehydrogenase
2380	transcriptional repressor for phenylacetic acid degradation
2702	hypothetical protein
2720	amidohydrolase, AtzE family
2728	putative beta-ketoacyl synthase
2738	hypothetical protein
2745	guanine deaminase
2752	TENA/THI-4 family protein
2840	NADH oxidoreductase for HCP
2883	hypothetical protein
3010	LysR family transcriptional regulator
3019	putative cytosine deaminase
3030	LysR family transcriptional regulator
3052	LysR family transcriptional regulator
3065	putative ABC transport system ATP-binding

Pokračování na další stránce

Tab. B.6 – pokračování z předcházející strany

číslo genu	**genový produkt**
3534	putative LysR-family transcriptional regulator
3554	UDP-glucose 6-dehydrogenase
3715	sn-glycerol-3-phosphate dehydrogenase subunit C
4372	Ars family arsenical pump
4641	putative SorC-family transcriptional regulator
5230	2-aminoethylphosphonate ABC transporter ATP-binding protein
5235	phosphonoacetaldehyde hydrolase
5236	putative inner membrane transport protein
5239	hypothetical protein

Tab. B.7: Nalezené genové úseky schopné odlišit jednotlivé melt-typy u genomů S25 - S36 získané z fylogenetické analýzy a analýzy teplot tání.

fylogenetická analýza		**analýza teplot tání**	
číslo genu		**číslo genu**	
448	3010	448	3010
584_2	3019	584_2	3019
632	3030	2296	3030
2296	3554	2315	3554
2315	3554_2	2376	3554_2
2376	4372	2840	4372
2840	5235	2883	5239
2883	5239		

C OBSAH PŘILOŽENÉHO CD

- **data/** - složka s vyhledanými geny v analyzovaných genomech, zarovnáním genů a s referenčním genomem
- **tex/** - složka se zdrojovým kódem práce pro systém LaTeX
- **matlab/** - zdrojový kód programu `find_variable_parts`
- **results/** - složka s nalezenými variabilními úseky ve formátu FASTA