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

DETEKCE CNV V BAKTERIÁLNÍCH GENOMECH
CNV DETECTION IN BACTERIAL GENOMES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Michaela Lacinová

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. Helena Škutková, Ph.D.

BRNO 2019

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. Michaela Lacinová ***ID:*** 173118
Ročník: 2 ***Akademický rok:*** 2018/19

NÁZEV TÉMATU:

Detekce CNV v bakteriálních genomech

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se základními typy strukturní variability v bakteriálních genomech a zaměřte se na segmenty DNA vyskytující se ve variabilním počtu kopií (CNV - copy number variation). 2) Nastudujte výpočetní metody detekce CNV v eukaryotických i prokaryotických organismech. 3) Navrhněte algoritmus detekce CNV na základě nerovnoměrného pokrytí v genomovém sestavení, proměnlivého zastoupení GC obsahu a vzdálenosti sekvenčních čtení. Dílčí části návrhu otestujte. 4) Realizujte algoritmus detekce CNV ve zvoleném programovém prostředí. 5) Algoritmus otestujte na sekvenačních datech bakteriálních kmenů Clostridium difficile a Klebsiella pneumoniae získaných z Dětské nemocnice FN Brno. 6) Výsledky srovnejte s literaturou a diskutujte potenciální vliv detekovaných variabilních segmentů na šíření bakteriální antibiotické rezistence.

DOPORUČENÁ LITERATURA:

[1] ZHAO, Min, Qingguo WANG, Quan WANG, Peilin JIA a Zhongming ZHAO. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics. 2013, 14(Suppl 11), S1.

[2] PIROOZNIA, Mehdi, Fernando S. GOES a Peter P. ZANDI. Whole-genome CNV analysis: advances in computational approaches. Frontiers in Genetics. 2015, 6(138), 1-9. DOI: 10.3389/fgene.2015.00138. ISBN 1664-8021.

Termín zadání: 4.2.2019 ***Termín odevzdání:*** 17.5.2019

Vedoucí práce: Ing. Helena Škutková, Ph.D.
Konzultant:

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

Abstrakt

Tato diplomová práce se zabývá rozborem strukturní variability genomu a metodami jeho sekvenování napříč všemi generacemi. Součástí rozboru je popis variability počtu kopií a možnosti její detekce. Praktická část práce je zaměřena na návrh algoritmu pro detekci CNV na základě analýzy a testování simulovaných genomických dat podle nerovnoměrného pokrytí v genomovém sestavení, proměnlivého zastoupení GC obsahu a vzdálenosti sekvenčních čtení. Tento algoritmus je následně otestován na genomických datech bakterie *Klebsiella pneumoniae.*

Klíčová slova

Variabilita počtu kopií, CNV, sekvenování, bakteriální genom, *Klebsiella pneumoniae*

Abstract

This master thesis deals with analysis of structural variation of genome and with methods of its sequencing across all generations. Subsequently it contains a description of copy number variation and methods of its detection. The experimental part focuses on algorithm proposal for CNV detection according analysis and testing of uneven coverage in genome, variable representation of GC content and distance of sequence reads. Finally, the algorithm for detecting copy number variation is tested on genomic data of bacteria *Klebsiella pneumoniae.*

Keywords

Copy number variation, CNV, sequencing, bacterial genome, *Klebsiella pneumoniae*

LACINOVÁ, Michaela. Detekce CNV v bakteriálních genomech. Brno, 2019. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/118356. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce Ing. Helena Škutková, Ph.D.

Prohlášení

Prohlašuji, že svoji diplomovou práci na téma Detekce CNV v bakteriálních genomech jsem vypracovala samostatně pod vedením vedoucí diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako 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 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.

V Brně dne ....

(podpis autora)

Poděkování

Ráda bych poděkovala vedoucí diplomové práce Ing. Heleně Škutkové, Ph.D. za odborné vedení, trpělivost a podnětné návrhy při zpracování mé diplomové práce.

V Brně dne ....

(podpis autora)

Obsah

Seznam obrázků

Seznam tabulek

Úvod

S masivním rozvojem celogenomového sekvenování došlo k výraznému vzestupu analýzy variability počtu kopií (CNV) v sekvenci DNA. Převážná většina nástrojů pro detekci je vytvořena speciálně pro lidské genomy, kdy se výzkum soustředí na souvislost CNV a dědičných nebo onkologických onemocnění. Méně pozornosti bylo věnováno detekci CNV u prokaryotických organismů, přestože vykazují vysoké množství těchto úseků a uchovávají nadbytečné geny, které mohou následně tvořit určité jejich zvýhodnění za daných environmentálních podmínek. Tato práce se zabývá analýzou a detekcí variability počtu kopií u prokaryotických organismů se zaměřením na bakterii *Klebsiella pneumoniae* [1].

První objevy CNV lze vysledovat na počátek 20. století, kdy byl objeven lidský karyotyp s jedním chromozomem X bez chromozomu Y nebo trizomie 21 chromozomu. Tyto objevy se ovšem opíraly o poměrně nepřesné techniky, které umožňovaly detekovat variace pouze použitím optického mikroskopu a měly přímé klinické následky. S rozvojem techniky především v posledních letech lze pozorovat, že většina CNV představuje submikroskopické chromozomální strukturální změny, které je obtížné pozorovat i nejvýkonnějšími optickými mikroskopy. Významným přínosem pro porozumění těchto submikroskopických CNV byla první komplexní mapa lidských CNV, pomocí které bylo zjištěno, že CNV se objevují i u zdravých jedinců. Z toho lze odvodit, že CNV mohou hrát důležitou roli při adaptaci na různá prostředí, náchylnosti k běžným onemocněním nebo evoluci [2].

První část diplomové práce se věnuje rozboru strukturní variability DNA a metod jejího sekvenování. Dále je popsána variabilita počtu kopií a možnosti její detekce. Praktická část práce popisuje návrh algoritmu pro detekci CNV na simulovaných datech na základě nerovnoměrného pokrytí v genomovém sestavení, proměnlivého zastoupení GC obsahu a vzdálenosti sekvenčních čtení. Následně je tento algoritmus otestován na genomických datech bakterie *Klebsiella pneumoniae*.

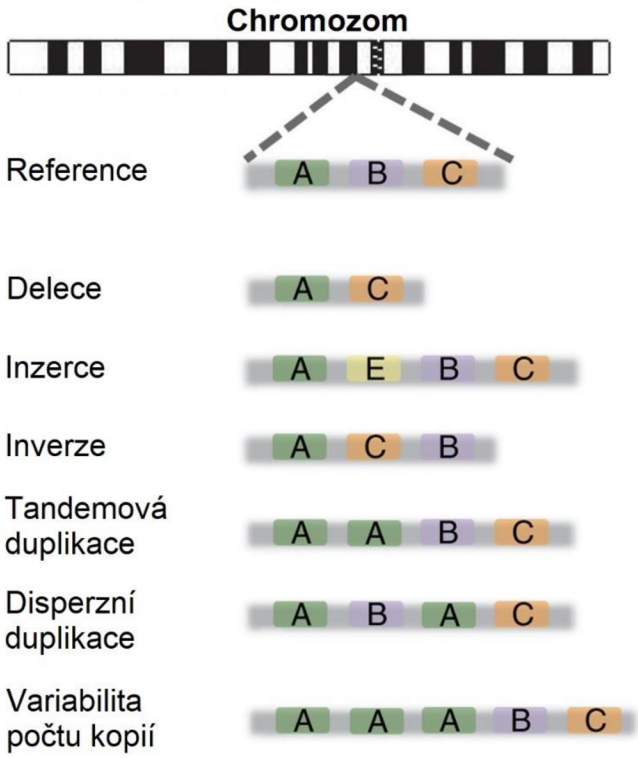
1 Genom

Genom daného organismu představuje kompletní sekvenci DNA, která obsahuje kódující a nekódující oblasti. Pro možnost výzkumu organismů a zdůvodnění případných abnormalit je důležité tuto sekvenci znát. V následujících podkapitolách budou popsány vybrané typy strukturní variability, které se projevují v sekvenci DNA. Dále budou rozebrány metody sekvenování genomů napříč jednotlivými generacemi.

1.1 Strukturní variabilita DNA

Zjištění strukturní variability mezi genomy daného druhu umožňuje zjistit fenotypovou variaci nebo náchylnost k různým nemocem. Před dostupností dnešních moderních sekvenačních technologií bylo obtížné pozorovat vzácné změny ve struktuře. Jednotlivé změny bylo možné sledovat pouze mikroskopicky, proto byly patrné především změny v množství a struktuře chromozomů. Tato mikroskopická strukturní variabilita byla definovaná pro velikost větší než tři megabáze. Mezi tyto odchylky se mohou zařadit aneuploidie nebo heteromorfismy. Při aneuploidiích dochází k nadbytku nebo k absenci chromozómu, vzniká tudíž nejčastěji monozomie nebo trizomie chromozómu. Heteromorfismy se projevují jako odchylky od fyziologické stavby chromozómu, přesto se ale nemusí vždy projevit jako patologická vlastnost [3].

S rozvojem sekvenátorů začaly být postupně sledovány menší modifikace. Variabilita o velikosti v rozmezí jedna kilobáze až tři megabáze byla definována jako submikroskopická strukturní variabilita. Do této skupiny se řadí např. jednonukleotidové polymorfismy (SNP), inzerce, delece, inverze, duplikace nebo translokace. Výběr těchto struktur je ukázán na obrázku č. 1, kde je nejprve ukázána reference, kterou následují ukázky variabilních struktur [3].



Obrázek 1 - Ukázka vybraných strukturních variabilit, upraveno [4]

SNP je změna jednoho nukleotidu na dané pozici v genomu na jiný nukleotid a nastává v populaci častěji než v 1 % případů. Inzerce značí začlenění určité části, která byla vyštěpena z chromozomu, do jiného chromozomu. Delece je naopak ztráta části chromozomu. Při inverzi dochází k přemístění nebo k otočení daného úseku v rámci jednoho chromozomu o 180°. S duplikacemi, které značí zdvojení dané části chromozomu, souvisí tandemové nebo disperzní repetice. Při tandemové repetici je kopírován sekvenční blok ve velkém množství za sebou na rozdíl od disperzních repetic, kdy jsou kopie sekvenčního bloku rozptýlené po celém genomu. Poslední zde uvedenou strukturní variabilitou jsou translokace, při nichž dochází k přesunu nebo k výměně určitého úseku mezi chromozomy [5].

Předpokládá se, že tyto menší strukturní varianty tvoří značnou část genetické variability, neboť se vyskytují s vyšší frekvencí než mikroskopické struktury. Ačkoli nemají v určitých úsecích genomu žádné zřejmé důsledky, v jiných mohou způsobit v kombinaci s genetickými nebo environmentálními faktory různá genetická onemocnění nebo predispozici k nim [3].

1.2 Metody sekvenování DNA

Pomocí metod sekvenování je zjišťováno pořadí nukleotidů v molekule DNA. Znalost těchto sekvencí tvoří podstatnou část informace pro následné zkoumání organismů a jejich funkcí. Metody sekvenování můžeme rozdělit na tři základní skupiny – první, druhá a třetí generace. Následující odstavce se budou postupně věnovat jednotlivým metodám sekvenování [6].

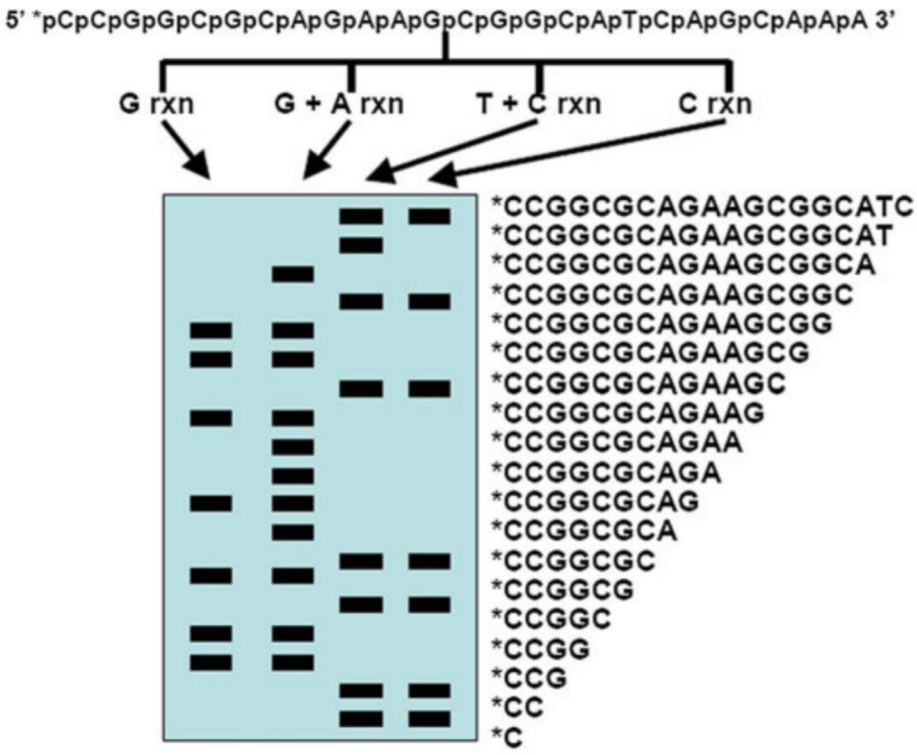
1.2.1 Sekvenátory první generace

Sekvenátory první generace byly vyvinuty v 80. letech minulého století a jsou založeny na metodách Maxam – Gilbert a Sanger [6].

Maxamova – Gilbertova metoda

Maxamova – Gilbertova metoda byla vyvinuta pro sekvenování jednovláknové DNA dvoukrokovým katalytickým procesem pomocí piperdinu a chemických látek, které selektivně napadají puriny a pyrimidiny. Na jeden konec daného vlákna DNA je nanesena radioaktivní značka a takto označené vlákno je vystavěno čtyřem odděleným reakcím, z nichž každá vytváří skupinu značených produktů končících známým nukleotidem. Dané skupiny se vyskytují jako samostatný nukleotid nebo jako dvojice nukleotidů a jedná se o kombinace G, G+A, T+C a C. Tyto čtyři reakce jsou naneseny na polyakrylamidový gel a fragmenty jsou separovány pomocí elektroforézy. Následně je gel přenesen na světlu odolné rentgenové kazety, na které je položen rentgenový film, a na několik dní umístěn do mrazničky. Poté je ze značených fragmentů v autoradiografu zjišťována výsledná sekvence [7].

Ukázka autoradiografu je na obrázku č. 2. Z tohoto grafu je možné vyčíst, že prvním nukleotidem je cytosin, protože je viditelný pruh ve skupině C i C+T. Pokud by byl pruh jen ve skupině C+T, jednalo by se o nukleotid thyminu. Stejný postup se uplatňuje pro skupiny G a G+A, kdy jsou identifikovány nukleotidy guanin a adenin [7].



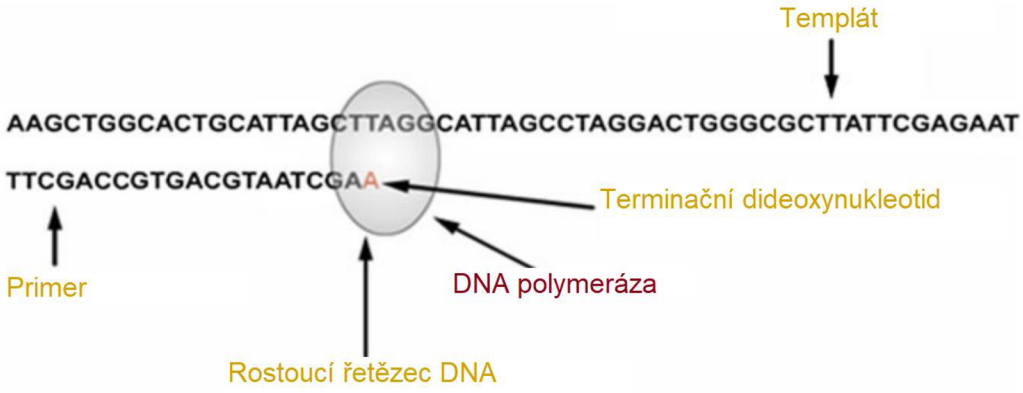
Obrázek 2 - Ukázka autoradiografu [7]

Sangerova metoda

Sangerova metoda sekvenování se i přes časovou náročnost a obtížnost stala velmi rozšířenou metodou. První automatizovaný sekvenátor byl uveden v roce 1987. Pracoval na principu separace jednotlivých úseků DNA pomocí kapilární elektroforézy a byl schopen za den osekvenovat 500 kilobází s délkou čtecích fragmentů 600 bází (bp). Automatické sekvenátory založené na této metodě a používané v současné době jsou schopny osekvenovat až 2,88 megabází s délkou čtecích fragmentů až 1 000 bp a stále jsou využívány v laboratořích díky své přesnosti. Ovšem limitace v podobě možnosti osekvenování jedné molekuly a tím i nevýhoda Sangerovy metody se nejvíce projevila při sekvenování lidského genomu a bylo zapotřebí vyvinout výkonnější a rychlejší metody [6], [8].

Při Sangerově metodě je podle jednoho daného templátu opětovně syntetizován nový řetězec DNA, který je pokaždé náhodně přerušen přiřazením modifikované nukleotidové báze, která tento proces syntetizace ukončí. Každá takováto báze nese charakteristické fluorescenční značení, čímž lze zjistit o jakou bázi se jedná. Opakováním procesu syntetizace jsou získány všechny možné délky řetězce, kdy každý je ukončen specifickou značkou označující danou bázi. Následným vzestupným seřazením těchto

nových řetězců DNA lze získat sekvenci DNA. Princip Sangerovy metody je ukázán na obrázku č. 3 [8].



Obrázek 3 - Princip Sangerovy metody, upraveno [7]

1.2.2 Sekvenátory druhé generace

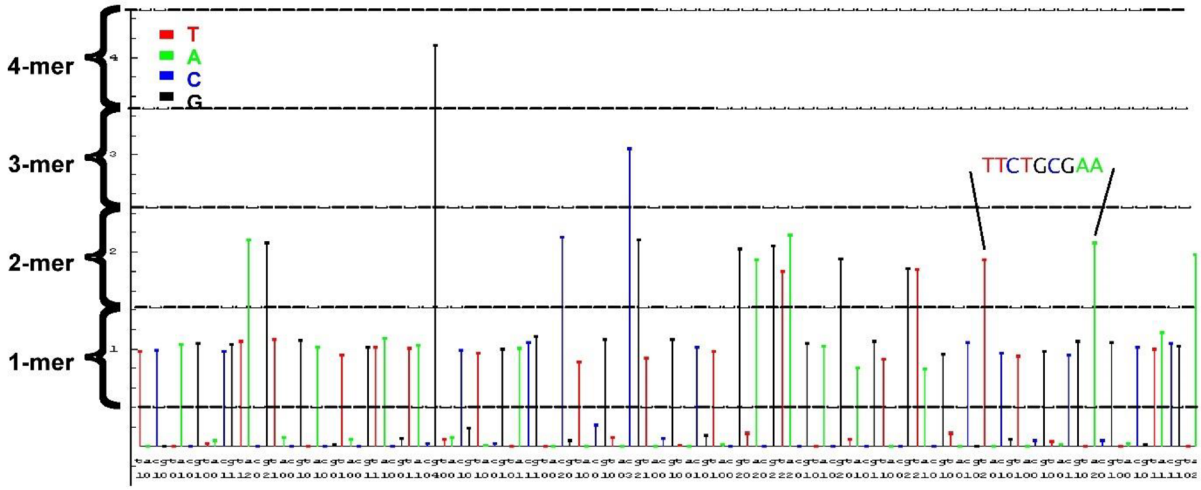
Sekvenátory druhé generace využívají podobného principu jako Sangerova metoda s rozdílem, že jsou schopny najednou osekvenovat až miliony odlišných molekul DNA. To představuje jejich značnou výhodu oproti předchozím metodám. Nevýhodou je krátká maximální délka čtecích sekvencí, která se pohybuje v řádu stovek bází, a zároveň i vyšší chybovost [8].

Roche 454

Sekvenování pomocí Roche 454 bylo vydáno v roce 2005 Jonathanem Rothbergem a bylo založeno na metodě pyrosekvenování. V roce 2016 byla ovšem podpora Roche 454 ukončena a nahrazena novějšími metodami sekvenování. Mezi výhody této metody patřila její vysoká rychlost a možnost velmi dlouhých čtení, čehož se využívalo při celogenomovém *de novo* sekvenování, resekvenování a při studiu bakteriální diverzity. Naopak největší nevýhoda spočívala ve vysoké chybovosti homopolymerních úseků, které byly detekovány pouze na základě množství vyzářeného světla [6], [8].

Při této metodě je nejprve DNA naštěpena na dvouvláknové fragmenty, k nimž jsou přiřazeny specifické adaptory (A a B), které jsou určeny k pozdější amplifikaci, purifikaci a sekvenování. Následně jsou tyto fragmenty navázány na magnetickou kuličku a denaturovány, kdy fragmenty s adaptorem A jsou uvolněny. Zbylé jednotlivé jednořetězcové fragmenty jsou navázány vždy na jednotlivé DNA kuličky, které nesou komplementární sekvenci DNA fungující jako primer. Dalším krokem je amplifikace každé

kuličky pomocí emulzní PCR, po které každá kulička nese kolem deseti milionů identických kopií původní DNA. Každá kulička je nanesena do jamky pikotitrační destičky společně se sekvenačním primerem a dalšími potřebnými enzymy. Do nově vznikajícího řetězce je vždy včleněn jeden daný komplementární nukleotid a je uvolněna molekula pyrofosfátu, který se následně účastní enzymatické reakce, při které dojde k emisi viditelného světla. Vysoce citlivá kamera snímá celou destičku a na základě rozsvícení a intenzity dokáže identifikovat, zda proběhlo přidání báze, případně kolik bází bylo přidáno najednou. Jednotlivé typy komplementárních nukleotidů jsou přidávány postupně, aby nedocházelo ke zkreslení výsledků. Celý proces je cyklicky opakován a sekvence je postupně načítána do pyrogramu, jehož ukázka je na obrázku č. 4 [6], [8].



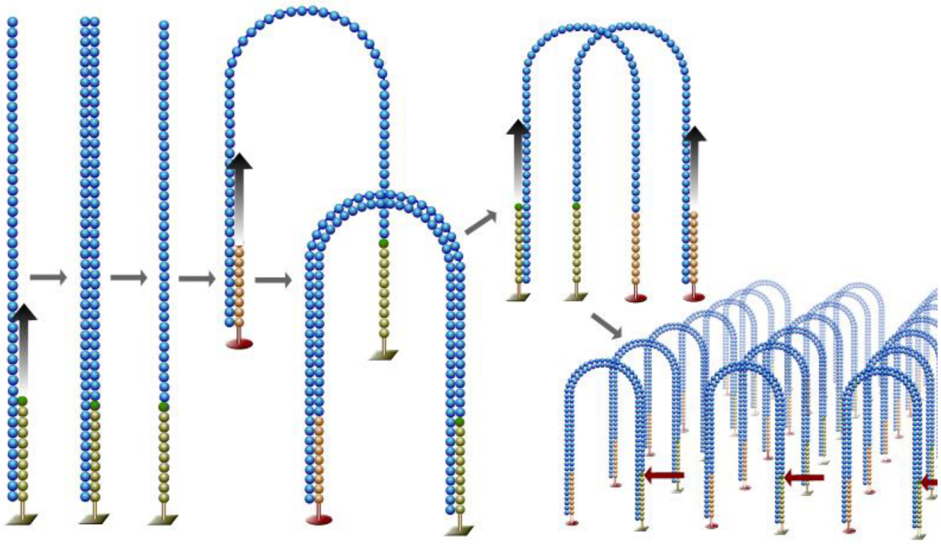
Obrázek 4 - Ukázka pyrogramu [9]

Illumina

Princip sekvenátorů Illumina byl představen roku 2007 a je založen na sekvenaci syntézou ve spojení s můstkovou amplifikací. V současné době jsou nabízeny sekvenátory různých výkonností vhodných jak pro sekvenování menších genomů nebo vybraných oblastí genů, tak i pro velké studie prováděné jen ve specializovaných centrech. Zároveň je to nejpoužívanější metoda sekvenace vzhledem k její výkonnosti. Další výhodou je cenová dostupnost oproti všem předchozím metodám. Mezi nevýhody se řadí nutnost využití jen krátkých sekvenačních čtení, které se průměrně pohybují o délce několik stovek bází, což může způsobit substituci nukleotidu. I přes tuto nevýhodu se úspěšnost udává 99 % [6], [8].

Při realizaci sekvenování metodou Illumina je příprava fragmentů velmi podobná předchozí metodě. DNA fragmenty jsou naštěpeny na velikost menší než 800 bp a na jejich

konce přidány adaptory. Po denaturaci jsou fragmenty navázány ke komplementárním adaptorům na reakční komůrce, čímž jsou jednotlivé fragmenty na jednom konci zcela imobilizovány ke komůrce, kde probíhá amplifikace a zároveň toto místo slouží jako primer pro syntézu dvouvláknové DNA při běhu PCR. Vzniklá dvouvláknová DNA je denaturována a původní templát je odmyt. Nové syntetizované vlákno zůstává jedním koncem navázáno k povrchu reakční komůrky a svým druhým volným koncem se napojí k adaptorům na povrch jiné reakční komůrky. Tímto způsobem dojde k ohybu vlákna a přemostění. V dalším cyklu PCR vznikne dvouvláknový most a celý tento proces se opakuje. Schéma vzniku dvouvláknových mostů znázorňuje obrázek č. 5. Vzniklé mosty jsou nakonec denaturovány a vzniknou klastry, které představují přibližně tisíc kopií DNA sloužící pro následnou sekvenaci. Při ní je do reakčních komůrek přilita směs odlišně značených nukleotidů fluorescenční barvou s inaktivovanou 3'- OH skupinou. Při spojení nukleotidu a řetězce je zaregistrována pozice a barva pomocí kamery a následně je inaktivovaná skupina, včetně barvy, odstraněna a cyklus přidání nové báze tak může být uskutečněn znovu [6].



Obrázek 5 - Znázornění můstkové amplifikace [10]

Sekvenování oligonukleotidovou ligací a detekcí

Třetím typem sekvenování druhé generace je metoda sekvenování oligonukleotidovou ligací a detekcí (SOLID). Oproti předchozím dvěma již zmíněným metodám SOLID využívá při sekvenování ligaci, pomocí níž je možné připojit úseky jednořetězových molekul DNA ke stávajícímu řetězci DNA. Technologie SOLID vyniká vysokou přesností detekce výsledné sekvence, případná chybovost se projevuje při čtení palindromatických úseků,

které mohou vytvářet smyčku v templátové DNA. Obdobně jako předchozí metody i metoda SOLID produkuje krátké sekvence o maximální délce 100 bp [6], [8].

Při samotném sekvenování se k templátu postupně připojují úseky DNA, které jsou označovány jako tzv. sondy. Sonda má na počátku vždy kombinaci dvou nukleotidů a obsahuje fluorescenční značení pro daný nukleotid. Celkově existuje šestnáct různých typů sond, kdy vždy čtyři jsou označeny stejnou značkou. Při tvoření nového řetězce se připojí sonda, která odpovídá templátové DNA, pomocí enzymu ligázy. Pomocí snímače je detekováno fluorescenční značení, následně je odstraněno a je přidána další sonda. K detekci celé sekvence je nutné opakované čtení templátové molekuly [8].

1.2.3 Sekvenátory třetí generace

Sekvenátory třetí generace se od předchozí generace liší především v tom, že templát DNA není namnožen, ale sekvenování probíhá na základě původní molekuly. Přestože tento typ sekvenace není prozatím výrazně rozšířen, umožňuje přečíst deset tisíc bází v rámci jedné molekuly DNA, čímž vytváří dlouhé sekvence. Tato vlastnost je výhodná při sekvenování celých genomů. Nevýhodu prozatím tvoří vysoká chybovost, která se pohybuje mezi 10-15 %. V současnosti jsou používané technologie PacBio a Oxford Nanopore [8].

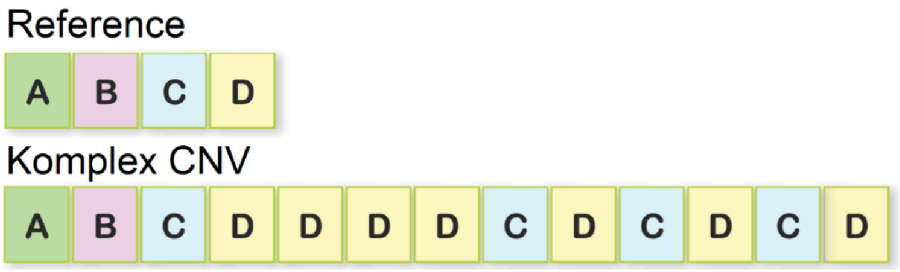
PacBio využívá k detekci výsledné sekvence fluorescenčně značené nukleotidy a vzhledem k vysoké citlivosti detektoru je možné zaznamenat přiřazení i jediného nukleotidu v reálném čase do řetězce DNA [8].

Při metodě Oxford Nanopore dochází k protahování jednořetězové molekuly DNA mikroskopickým pórem na membráně, která je syntetická. Vzhledem k tomu, že jednotlivé báze mají částečně různý tvar, nastává vždy odlišné zaplnění póru a snímače identifikují o jakou bázi se jedná. Nespornou výhodou tohoto sekvenátoru je jeho velikost, protože se jedná o kapesní přístroj, který lze lehce připojit k počítači [8].

2 Variabilita počtu kopií

Variabilita počtu kopií je jeden z typů strukturní variability, které odkazují na abnormality v sekvenci DNA a zároveň se podílí na fenotypových změnách. CNV mohou být zděděny z genomu rodičů nebo mohou vzniknout *de novo*. CNV, které jsou v populaci projeveny ve více než v 1 %, jsou definovány jako běžné CNV nebo polymorfismus počtu kopií. Naopak CNV, které se v populaci vyskytují v méně než 1 %, jsou označovány jako vzácné. Ve srovnání s jednonukleotidovými polymorfismy je četnost CNV nižší, ale délka sekvencí je výrazně vyšší, čímž stoupá významnost potenciálního vlivu na fenotyp a vývoj. CNV se tedy řadí mezi důležité zdroje genomického polymorfismu [2], [11].

CNV jsou zpravidla způsobeny strukturální změnou v chromozomech určitého segmentu DNA o délce od jedné kilobáze do několika megabazí v porovnání s referenční genomovou sekvencí. Nejčastěji se projevují delecí, tandemovou a disperzní repeticí. Na obrázku č. 6 je porovnána reference s komplexem CNV, který je tvořený jak tandemovou, tak i disperzní repeticí [11].



Obrázek 6 - Porovnání reference a komplexu CNV, upraveno [12]

2.1 Metody detekce CNV

Existují desítky dostupných nástrojů pro detekci CNV z dat získaných ze sekvenátorů druhé generace. Naprostá většina těchto nástrojů je ovšem vyvinuta pro detekci CNV v lidském genomu, tedy pro diploidie. To může představovat problém při aplikaci těchto nástrojů pro genomy prokaryotních organismů, které se mohou vyskytovat o různé ploidii. V této podkapitole bude uveden přehled statistických metod, které byly vyvinuty a testovány pro detekci CNV z dat obsahujících jednotlivá čtení. Tyto metody jsou založeny především na DNA čipech nebo byly zcela přizpůsobeny pro tento druh detekce [1], [13].

První uvedený algoritmus, který byl původně vytvořen pro genomová hybridizační data z DNA čipu, je založen na kruhové binární segmentaci (CBS). Prvním nástrojem pro zavedení CBS je ***SegSeq***, který porovnává zkoumaná data s kontrolním souborem k detekování CNV. Nejprve je vypočítán poměr počtu kopií dvou shodných vzorků v každém genomovém okně a následně je pomocí algoritmu identifikována hranice fragmentů DNA. Tento algoritmus je aplikován na data vygenerovaná ze sekvenátoru Illumina [13], [14].

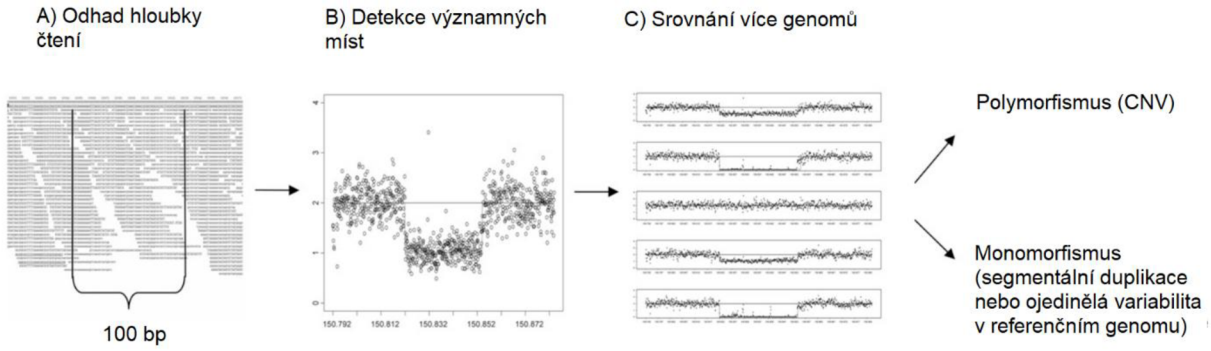
Metoda ***Mean Shift-Based*** (MSB) byla nejprve využívána pro analýzu dat z aCGH. Při MSB jsou sousední genomová okna s podobným pokrytím spojena podél chromozomu. Hraniční body jsou ohlášeny, pokud se pokrytí daného posuvného okna významně liší od pokrytí spojených oken. Na tomto principu byly vytvořeny dva nástroje, ***CNVnator*** a ***BIC - seq***. ***CNVnator*** pracuje na jednotlivých samostatných vzorcích a jeho výhodu tvoří schopnost detekovat CNV různých velikostí v rozsahu od stovek bází až po megabáze. Tento nástroj dosahuje vysoké senzitivity (86–96 %) a nízké míry falešné detekce (3–20 %). Naproti tomu ***BIC-seq*** využívá při implementaci odpovídající si páry dat. Tento algoritmus byl testován na datech ze sekvenovacích technologií Illumina a SOLID [13], [14], [15].

Následujícím uvedeným modelem je ***CNASeg***, který využívá skryté Markovy modely (HMM) v segmentačním kroku k identifikaci genomových oblastí s podobnou hodnotou pokrytí. Hlavním rysem této metody je možnost kontroly falešně pozitivních hodnot pomocí pokrytí mezi zkoumanými a kontrolními vzorky. Tato metoda byla původně aplikována pro sekvenci dat získaných při výzkumu rakoviny generovaných sekvenátorem Illumina [13], [14].

Rozšířením modelu ***Shifting level*** (SLM) je možné detekovat opakující se CNV z více vzorků. Zároveň tento model dosahuje kvalitních výsledků u detekce menších oblastí CNV o délce přibližně 500 bp. U tohoto modelu byla testována výkonnost u dat, ze sekvenátoru Illumina, o vysokém pokrytí z projektu 1000 genomů [13], [14].

Další metoda, která byla vytvořená přímo pro detekci CNV, se nazývá ***Event-wise testing*** (EWT). Tato metoda prohledává v genomu specifické třídy malých skupin, které splňují kritéria statistické významnosti na rozdíl od segmentačních metod, které zpravidla vyhodnocují pravděpodobnost pro každý bod genomu. Následně jsou tyto menší skupiny spojeny do větších. Zjištěné delece a duplikace v daném genomu se dále zkoumají v rámci více genomů k možné identifikaci polymorfismu mezi jednotlivci. Princip této metody je zobrazen na obrázku č. 7. Výsledky ukazují, že EWT dosahuje

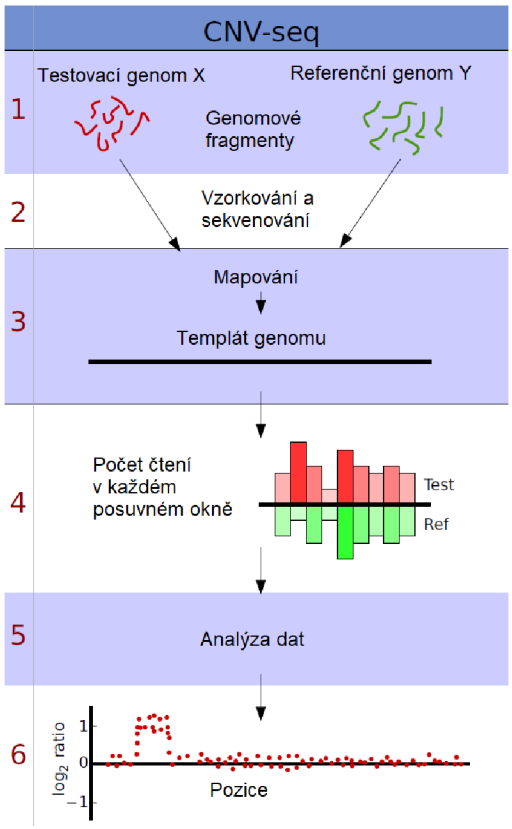
až 99,9 % úspěšnosti při detekci CNV o minimální velikosti jedné kilobáze v simulovaných datech. Tento algoritmus byl aplikován na data o vysokém pokrytí ze sekvenátoru Illumina a měřením na pěti reálných vzorcích genomů, které byly získány z projektu 1000 genomů. Bylo dosaženo úspěšnosti ověření 33-89 %. Nižší hodnota byla způsobena především vyšší mírou falešně pozitivních detekcí [13], [16].



Obrázek 7 - Princip metody EWT, upraveno [16]

Algoritmus ***CNOGpro*** byl vyvinut autory pro detekci CNV v prokaryotických genomech a využívá HMM. Tato metoda je schopna rychle odhadnout počet kopií jakéhokoli genu a zároveň je možná statistická verifikace metodou bootstrapping pro intervaly v oblastech, kde jsou očekávané CNV [1].

Model ***CNV-seq*** využívá posuvného okna k analýze poměrů mezi sekvenčním čtením dvou jedinců (např. zdravý a nemocný). Takto získané poměry jsou hodnoceny výpočtem pravděpodobnosti náhodného výskytu. Z výsledků je patrné, že klíčem k rozlišení detekce je počet čtení, a ne jejich délka, což podporuje sekvenování druhé generace. Dále bylo dosaženo specificity v rozmezí 91,7-99,9 % a senzitivity 72,2-96,5 %. Autoři testovali tento algoritmus na datech získaných Sangerovou metodou a ze sekvenátoru 454 Roche. Na obrázku č. 8 je ukázán základní princip postupu tohoto modelu [13], [17].



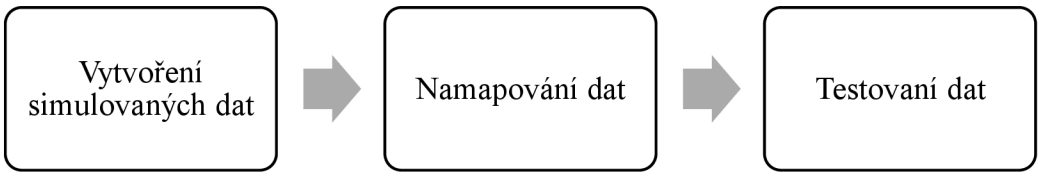
Obrázek 8 - Princip modelu CNV-seq, upraveno [17]

Uvedené algoritmy CBS, MSB, CNASeq a SLM je možné zařadit mezi segmentační metody, které umožňují rozdělit data sekvenčního čtení na segmenty, z nichž každý obsahuje stejný počet kopií DNA. Segmenty s pozměněným počtem kopií jsou detekovány pomocí prahování [13].

3 Návrh algoritmu detekce CNV

Tato kapitola se zabývá návrhem algoritmu pro detekci úseků CNV z bakteriálního genomu. Algoritmus bude navržen pro detekci na základě nerovnoměrného pokrytí v genomovém sestavení, proměnlivého zastoupení GC obsahu a vzdálenosti sekvenčních čtení. Vzhledem k tomu, že poskytnutá data bakteriálního genomu bakterie *Klebsiella pneumoniae* jsou neznámá a jsou určena pro další výzkum včetně detekce CNV, bylo výhodnější si nejprve vytvořit data simulovaná, u kterých je možnost následného zhodnocení úspěšnosti algoritmu.

Následující schéma shrnuje postup práce s daty v této kapitole. Ten bude dále podrobně vysvětlen.



Obrázek 9 - Schéma postupu práce s daty, vlastní zpracování

3.1 Vytvoření simulovaných dat

Simulovaná data byla vytvořena úpravou referenčního genomu bakterie *Klebsiella pneumoniae*. Jako reference byl vybrán patogenní kmen NTUH-K2044, který byl izolován z abscesu jater. Jeho délka činí 5 248 520 bp. Kmen NTUH-K2044 má mimo jiné vysokou virulenci a hypermukoviskozitu. Tyto faktory činí tento izolát vhodným modelovým kmenem pro genomické studie [18].

Pro simulaci dat byl do referenčního genomu vložen gen RbsR, který je represorem zodpovědným za regulaci exprese genu kódujícího ribózovou permeázu a nachází se v genomu právě jednou na pozici 13 492. Vzhledem k tomu, že gen RbsR má délku 1 398 bp a je výrazně delší než délka čtení, byl dále vybrán gen s označením KP10186, který má délku 108 bp. Pro oba tyto geny byla vytvořena simulovaná reference, kdy byl vždy daný gen vložen do sekvence navíc dvakrát, a to na pozici 500 000 a 3 500 000. Dále byly vytvořeny sekvence, ze kterých byla vytvořena simulovaná data. Přesná podoba jednotlivých sekvencí je popsána v kapitole 3.2.1 [19].

K vytvoření simulovaných dat byl využit volně dostupný simulátor *ART*, který byl spuštěn v terminálu operačního systému *UBUNTU*. Tento simulátor generuje umělá sekvenční čtení pomocí emulace sekvenčního procesu a v současné době podporuje simulaci ze sekvenátorů druhé generace Roche 454, Illumina a SOLID. Pro tuto práci byla vybrána simulace sekvenátoru Illumina pro párové čtení v závislosti na tom, že i reálná data byla sekvenována tímto způsobem. Vstupní hodnoty zadávané do tohoto simulátoru jsou zaneseny do tabulky č. 1. Hodnoty byly zvoleny dle [1] s mírnou úpravou, kdy autoři realizovali obdobnou simulaci dat. Změněna byla délka čtení a průměrné simulované pokrytí na hodnotu 150 bp z důvodu pokrytí celého genu KP10186 při mapování. Po proběhnutí simulace byly získány dva soubory typu *fastq*. Jedná se o soubory, které kromě sekvence nukleotidů v daném směru obsahují i ohodnocení její kvality [20], [21].

Tabulka 1 - Parametry vložené do simulátoru *ART*

Parametr	**Hodnota**
Délka čtení	150 bp
Průměrná odchylka DNA fragmentů	500 bp
Směrodatný odchylka	150 bp
Průměrné simulované pokrytí	150 bp
Sekvenovací systém	MiSeq v1

Tato simulovaná data ze sekvenátoru bylo nutné namapovat vůči referenčnímu genomu, kdy mapování proběhlo opět v terminálu operačního systému *UBUNTU* pomocí algoritmu *Burrows-Wheeler Alignment* (BWA). Jedná se o nástroj pro zarovnání sekvenčních čtení a je založen na zpětném hledání pomocí *Burrows–Wheeler* transformace k dosažení efektivního zarovnání krátkých sekvenčních čtení vůči velké referenční sekvenci, kterou tvoří zpravidla celý genom daného organismu. V rámci tohoto algoritmu jsou z výsledné sekvence automaticky odstraněny singletony, které vyjadřují čtení nesestavené do kontigů. Nástroj BWA podporuje data získaná ze sekvenátorů Illumina a SOLID a po jeho proběhnutí je vygenerován soubor typu *SAM*, který byl dále zpracován nástrojem *SAMtools*. Pomocí tohoto nástroje byla vygenerována konsensuální sekvence a určena míra pokrytí pro jednotlivé pozice [22], [23], [24].

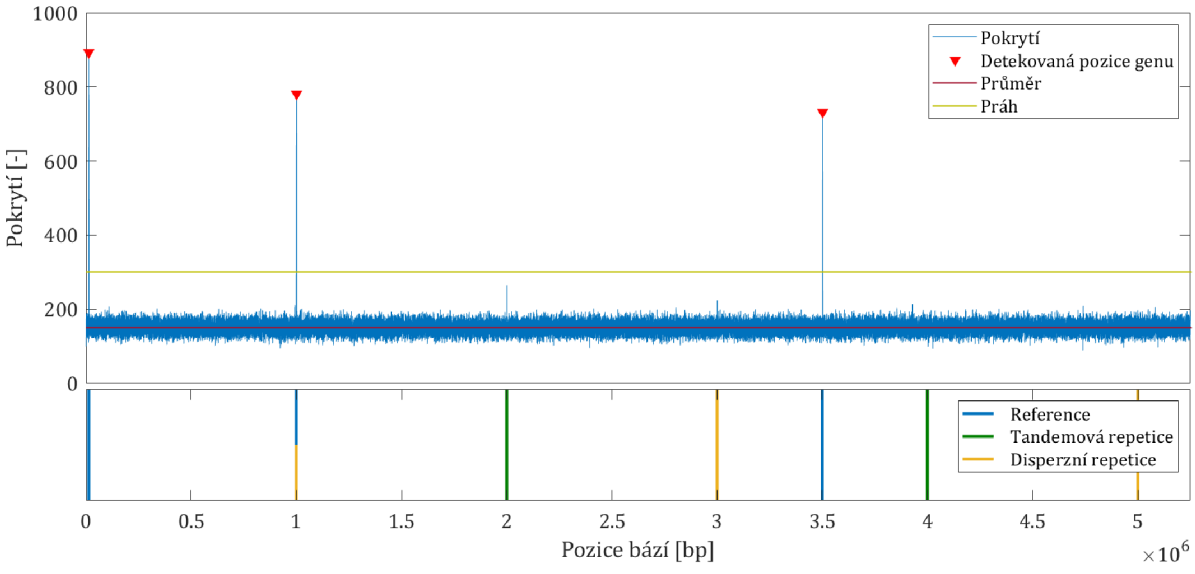
3.2 Testování dat

Pro navržení algoritmu detekce CNV byla simulovaná data nejprve otestována na základě nerovnoměrného pokrytí v genomovém sestavení, proměnlivého zastoupení GC obsahu a vzdálenosti sekvenčních čtení. Následující podkapitoly se zabývají jednotlivými částmi testování a shrnují získané poznatky.

3.2.1 Pokrytí v genomovém sestavení

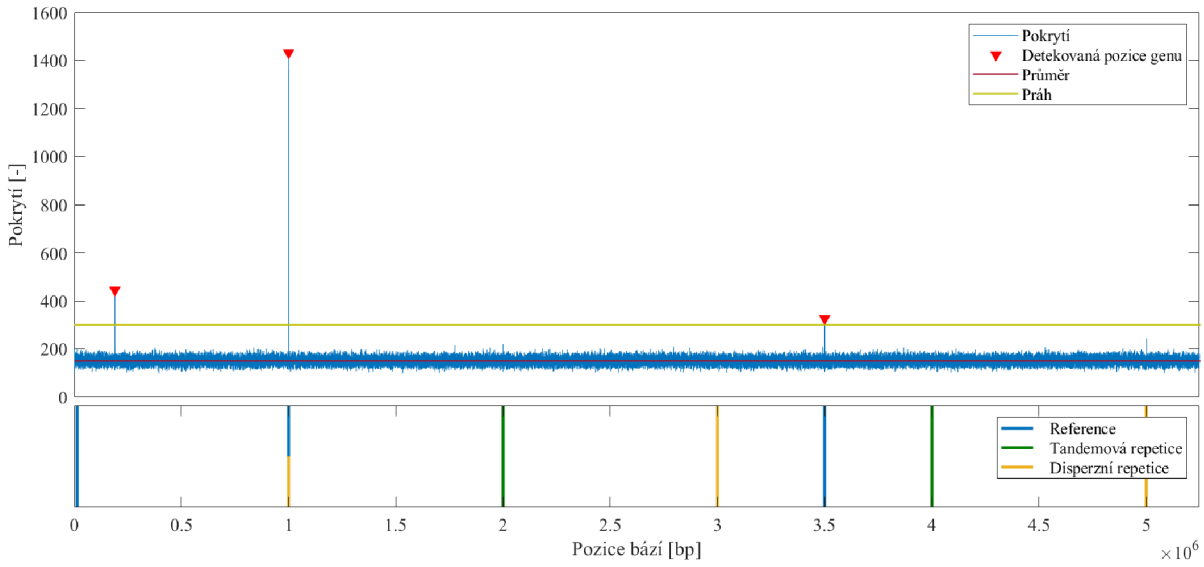
Pokrytí v genomu vyjadřuje počet jednotlivých čtení, které zahrnují daný nukleotid ve výsledné zarovnané sekvenci. Po simulaci a namapování upraveného genomu bylo zobrazeno výsledné pokrytí. Následně byly detekovány píky v oblasti potenciálního úseku CNV. Určení píku proběhlo při překročení stanoveného prahu, který byl určen jako dvojnásobek hodnoty průměrného pokrytí [25].

Obrázek č. 10 zobrazuje výsledné pokrytí při vložení genu RbsR jako kombinace tandemové a disperzní repetice. Gen byl vložen do sekvence vždy pětkrát za sebou na pozice 2 000 000 a 4 000 0000 a jedenkrát na pozice 1 000 000, 3 000 000 a 5 000 000. Z obrázku je patrné, že pokrytí bylo výrazně zvýšeno především na pozicích, kde se daný gen nachází v referenčním genomu, než kam byl gen vkládán v sekvenci. Důvod takto zvýšeného pokrytí je podrobněji vysvětlen s obrázkem č. 13. Pro přehlednost byla pod následující výsledné grafy přidána legenda, která zobrazuje přesná místa vložení jednotlivých genů.



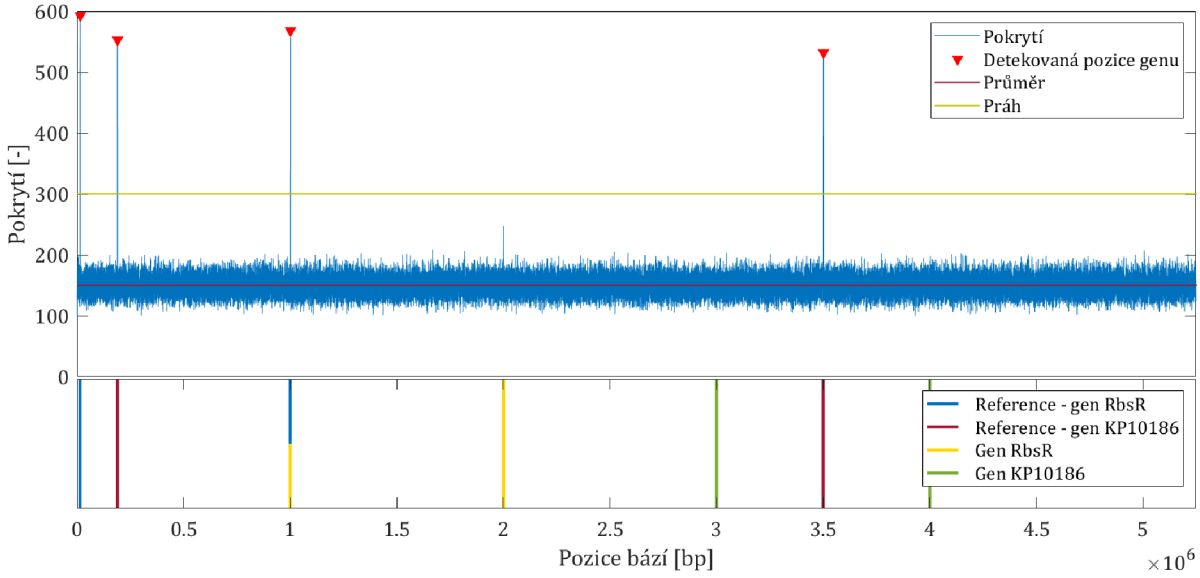
Obrázek 10 - Pokrytí genu RbsR o délce 1 398 bp při tandemové a disperzní repetici

Obrázek č. 11 zobrazuje výsledné pokrytí znovu pro kombinaci tandemové a disperzní repetici pro gen KP10186 o délce 108 bp. Tento gen byl vložen do sekvence na stejné pozice jako v předchozím případě. Z výsledného grafu je patrné, že byly opět detekovány pouze ty oblasti, kde se daný gen nachází v referenci. Hlavní rozdíl se projevil především v maximálním pokrytí, kdy gen vložený na stejnou pozici jako je v referenci dosahoval pokrytí téměř 1 400. Zbylé geny, které byly obsaženy pouze v referenci, dosahovaly hodnot až trojnásobně menších. To ve srovnání s předchozím případem nebylo sledováno. Z toho lze usoudit, že se snižující se délkou genu se snižuje i pokrytí genu, který se nachází v jiném místě než v referenci. To může mít vliv i na přesnost detekce.



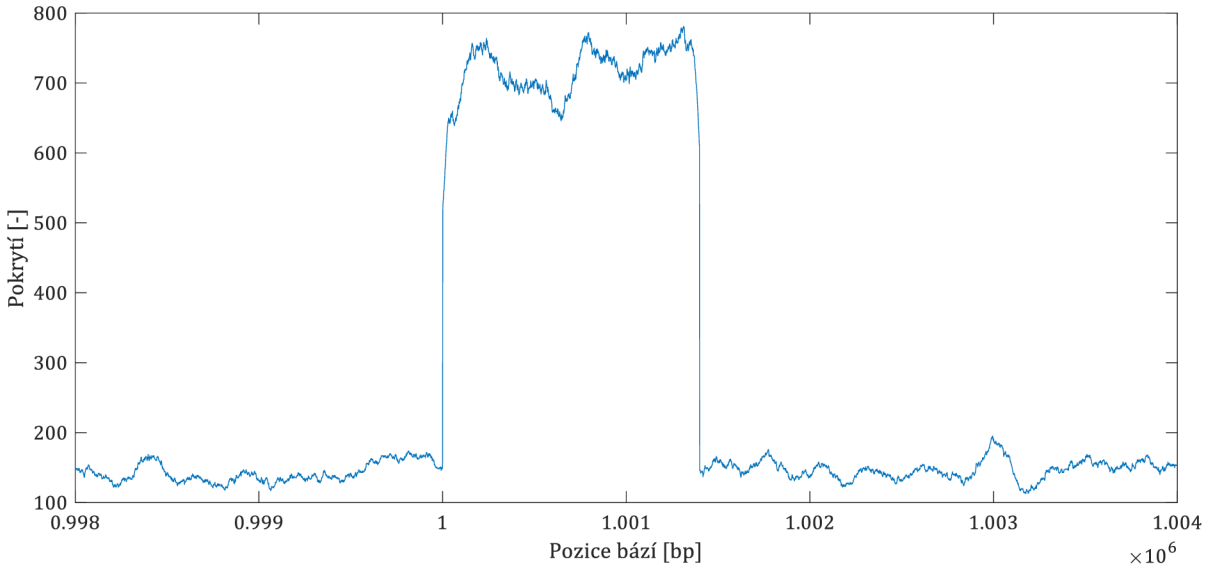
Obrázek 11 - Pokrytí genu KP10186 o délce 108 bp při tandemové a disperzní repetici

Následující obrázek č. 12 ukazuje kombinaci genů RbsR a KP10186 vložených do simulované sekvence. Gen RbsR byl vložen jednou na pozici 1 000 000 jako disperzní repetici a pětkrát na pozici 2 000 000 ve formě tandemové repetice. Druhý gen KP10186 byl vložen jako disperzní repetice jednou na pozici 3 000 000 a pětkrát na pozici 4 000 000 jako tandemová repetice. Ve výsledném grafu se projevilo zvýšení pokrytí pouze v oblastech, kde se nachází dané geny i v referenci. Rozdíl v maximálním pokrytí mezi danými geny není výrazně patrný.



Obrázek 12 - Pokrytí genů RbsR a KP10186

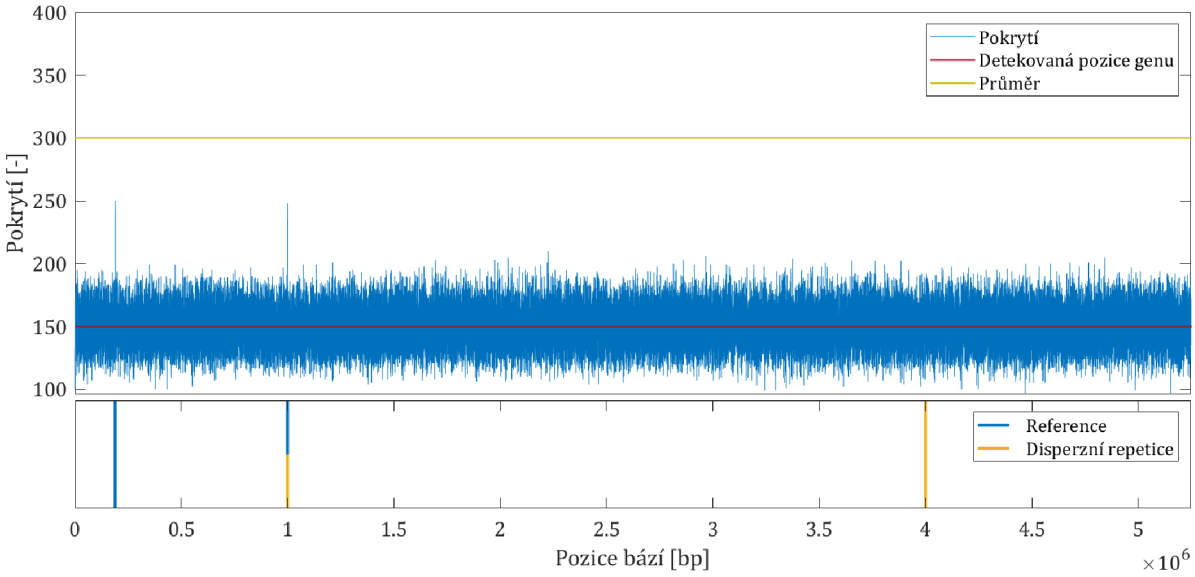
Obrázek č. 13 ukazuje detailní zobrazení pokrytí v okolí pozice 1 000 000 u genu RbsR. V grafu je nejprve patrný prudký vzestup. Výrazně zvýšená hodnota je způsobena vysokým počtem úseků čtení, které pochází z vložených genů jiných pozic. Tyto úseky byly do této oblasti namapovány, jelikož mají shodnou značnou část čtení s genem v referenci. Rozdíl mezi úseky čtení vložených genů a genů v referenci je tvořen na rozhraních, proto nedochází k takovému pokrytí a tím je způsoben pokles.



Obrázek 13 - Detailní ukázka pokrytí genu RbsR

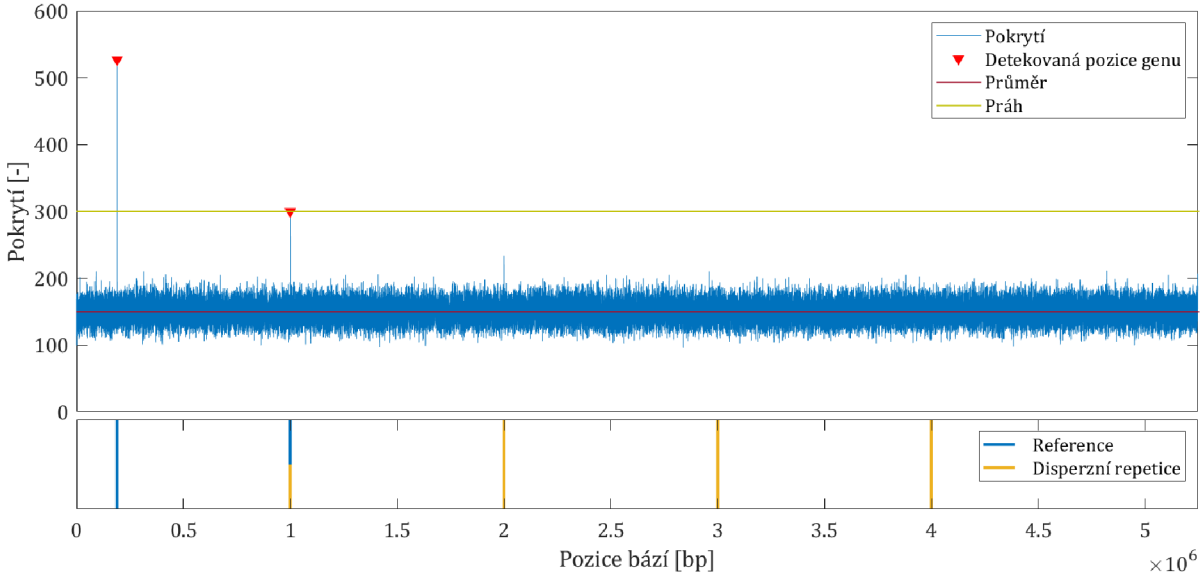
Dále byla zjišťována citlivost detekce potenciálních oblastí CNV. Gen KP10186 byl vložen pouze dvakrát do sekvence na pozice 1 000 000 a 4 000 000. V referenci se tento

gen nacházel také na dvou pozicích. Výsledný graf zobrazuje obrázek č. 14. Z něj je patrné, že tento gen nebyl vůbec detekován. V oblastech, kde je gen obsažen v referenci, lze sledovat mírně zvýšené pokrytí, to ale není dostatečné pro spolehlivou detekci, protože je pouze 1,63krát vyšší než průměrná hodnota pokrytí.



Obrázek 14 - Pokrytí genu KP10186 při analýze citlivosti

Vzhledem k předchozímu případu byla hledána citlivost detekce možných oblastí CNV. Při zanechání shodné reference a vložení genu KP10186 do sekvence na pozice 1 000 000, 2 000 000, 3 000 000 a 4 000 000 došlo již k detekci zvýšeného pokrytí, které je zobrazeno na obrázku č. 15. Z toho lze usoudit, že úsek CNV bude detekován, pokud se gen bude vyskytovat v sekvenci alespoň čtyřikrát.



Obrázek 15 - Pokrytí genu KP10186 při analýze citlivosti

Pro určení celého potenciálního úseku CNV bylo z předchozích výsledků zjištěno, že nárůst pokrytí v této oblasti má být vyšší než daný práh. Na základě těchto znalostí byl práh zvolen na hodnotu dvojnásobku průměrného pokrytí.

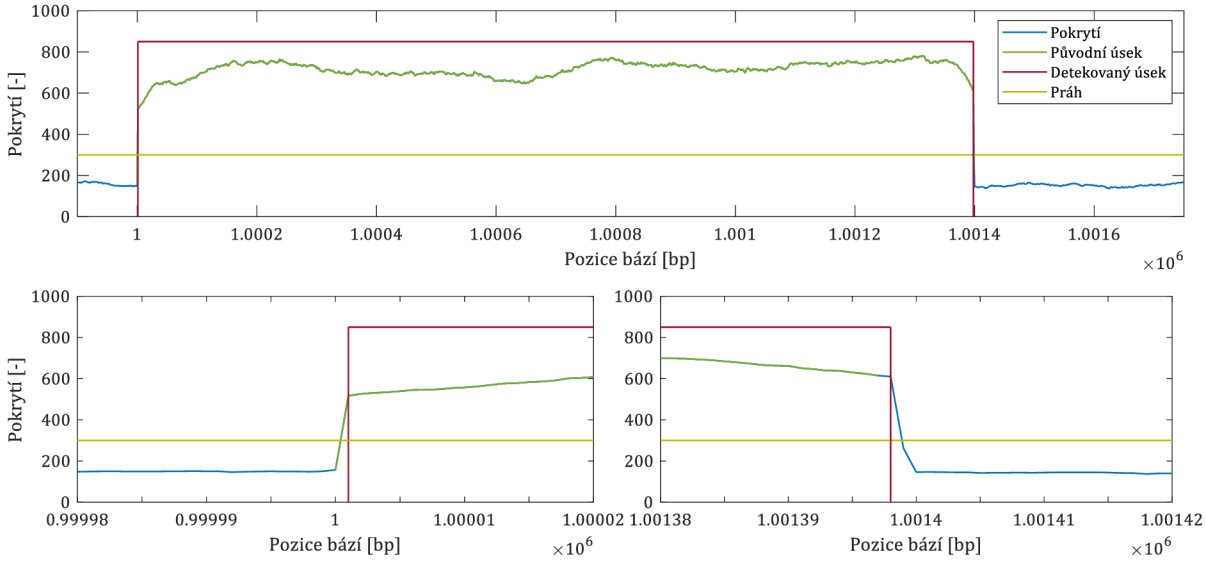
Při určování celého potenciálního úseku se vycházelo z detekovaného píku a byl použit algoritmus na principu seménkové detekce. Detekovaný pík byl označen za výchozí bod a následně se oběma směry analyzovalo pokrytí. Počáteční a konečný bod úseku byl označen v momentě, kdy pokrytí kleslo pod daný práh.

Následující tabulka č. 2 ukazuje příklad detekované pozice potenciální úseků CNV pro případ, kdy byly v sekvenci dva různé geny. Pokrytí tohoto úseku je zobrazeno na obrázku č. 12. Gen RbsR, který má originálně délku 1 398 bp, byl v jednom případě detekován s menší délkou. Jeho detekované délky byly 1 397 bp a 1 398 bp. To znamená, že došlo k chybě detekce v průměru o 0,05 % bp. Navýšení délky proběhlo u druhého genu KP10186, který má originálně délku 108 bp, kdy výsledné hodnoty byly 110 bp a 111 bp. Tento rozdíl představuje navýšení o 2,31 % bp.

Tabulka 2 - Detekovaná délka potenciálního úseku CNV

Gen	Detekovaný počáteční bod	Detekovaný konečný bod	Délka
RbsR	13 491	14 888	1 397
KP10186	189 761	189 871	110
RbsR	1 000 000	1 001 398	1 398
KP10186	3 501 397	3 501 508	111

Následující obrázek č. 16 srovnává třetí detekovaný úsek z předchozí tabulky se správnou pozicí tohoto genu. Nejprve je ukázán celý detekovaný gen (obrázek č. 16 nahoře), ale z důvodu jen malých rozdílů jsou následně v detailu zobrazeny místa začátku a konce tohoto genu (obrázek č. 16 dole). V grafu je zelenou barvou vyznačen původní úsek genu RbsR. Červené linie ohraničují detekovaný úsek. Levý spodní graf ukazuje detekci počátečního bodu, který byl detekován o 1 bp později oproti správnému původnímu úseku. Obdobný výsledek ukazuje pravý spodní graf konečného bodu. Původní úsek končí o 1 bp dříve než úsek detekovaný. Obě tyto odchylky jsou způsobeny nastavením prahu. Ovšem z obrázku lze pozorovat, že při jeho změně na vyšší ale i nižší hodnotu by došlo k většímu počtu falešně detekovaných míst.



Obrázek 16 - Porovnání původního a detekovaného úseku

3.2.2 Zastoupení GC obsahu

Druhou částí testování simulovaných dat byl výpočet zastoupení guanino-cytosinového (GC) komplementárního páru v konsensuální sekvenci. Oblasti bohaté na tyto nukleotidy vykazují zpravidla vyšší genovou hustotu nebo vyšší míru rekombinace. Naopak oblasti s podprůměrným GC zastoupením jsou projevem snížené kvality. GC obsah v sekvenci byl vypočítán dle rovnice

$$GC = \frac{G + C}{A + T + G + C} * 100\ [\%], \tag{1}$$

kde A, T, G, C značí počet jednotlivých nukleotidů v sekvenci [26], [27].

Výpočet GC obsahu byl proveden pro detekované potenciální úseky z předchozí podkapitoly ze simulovaných sekvencí pro geny RbsR, KP10186 a pro jejich kombinaci. Průměrná hodnota GC těchto úseků je 57,68 %. Směrodatná odchylka dosahuje hodnoty pouze 0,0003 %. Tato nízká hodnota je způsobena použitím této analýzy na simulovaných datech. Následující tabulky shrnují získané výsledky.

Tabulka č. 3 ukazuje výsledné hodnoty GC obsahu pro gen RbsR. Z tabulky lze vyčíst, že detekované potenciální úseky CNV mají nižší GC zastoupení oproti průměrné hodnotě. V případě prvního úseku je rozdíl nižší o 1,86 %, u druhého 2,12 % a u třetího 2,10 %.

Tabulka 3 - Hodnoty obsahu GC zastoupení pro gen RbsR

Úsek	**Hodnota GC obsahu [%]**
13487:14892	56,61
1000000:1001400	56,46
3501398:3502798	56,47
Průměr	**57,68**

Tabulka č. 4 zahrnuje GC obsah z detekovaných úseků genu KP10186. Výsledné hodnoty zastoupení GC obsahu nabývají hodnot nižších o oproti průměrné hodnotě o 6,41 %, 7,12 % a 12,33 %. Ve srovnání s předchozí tabulkou č. 3 jsou tyto hodnoty výrazně nižší.

Tabulka 4 - Hodnoty obsahu GC zastoupení pro gen KP10186

Úsek	**Hodnota GC obsahu [%]**
189761:189873	53,98
999999:1000110	53,57
3500107:3500193	50,57
Průměr	**57,68**

Tabulka č. 5 obsahuje hodnoty GC obsahu ze sekvence, která kombinuje geny RbsR a KP10186. Úseky, které odpovídají genu RbsR, dosahují snížení GC obsahu o 1,94 % a 2,17 %. GC zastoupení u oblastí odpovídající genu KP10186 se snižuje o 7,85 % a 8,67 %.

Tabulka 5 - Hodnoty obsahu GC zastoupení pro geny RbsR a KP10186

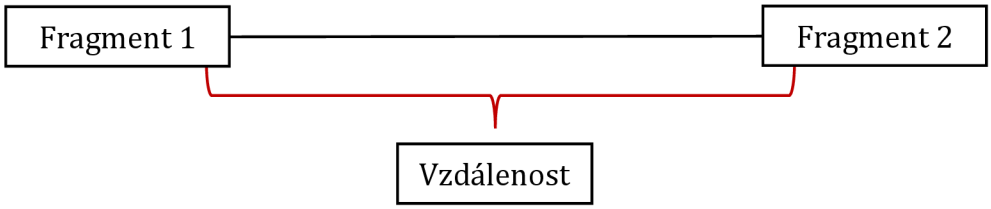
Úsek	**Hodnota GC obsahu [%]**
13491:14892	56,56
189761:189871	53,15
1000000:1001399	56,43
3501397:3501508	52,68
Průměr	**57,68**

Ze získaných hodnot lze konstatovat, že u všech detekovaných úseků docházelo ke snížení GC obsahu oproti průměru. Tyto hodnoty nesplňují teoretické předpoklady. To bylo způsobeno umělým vložení genů do simulovaných dat.

3.2.3 Vzdálenost sekvenčních čtení

Vzhledem k tomu, že při sekvenování dat byla zvolena varianta párového čtení dat, je možné na základě vzdálenosti těchto párů určit, zda se jedná o repetici. Obrázkem č. 13 bylo popsáno, že rozhraní poklesu a prudkého nárůstu pokrytí je způsobeno nedostatečným mapováním čtení z důvodu zcela odlišné původní pozice genu v testovaném genomu. Z tohoto důvodu může při mapování dojít k tomu, že tato párová čtení na rozhraní budou rozdělena do velké vzdálenosti od sebe. To může následně značit repetici [28].

Této znalosti bylo využito a pomocí volně dostupného bioinformatického systému *UGENE* byly získány pozice párových čtení v detekovaných potenciálních úsecích CNV. Z těchto pozic byla následně spočítána jejich absolutní hodnota vzdálenosti, která byla přehledně zobrazena pomocí histogramu. Obrázek č. 17 názorně ilustruje způsob určení vzdálenosti mezi pozicemi párových čtení, tedy mezi prvním a druhým párovým fragmentem [29].

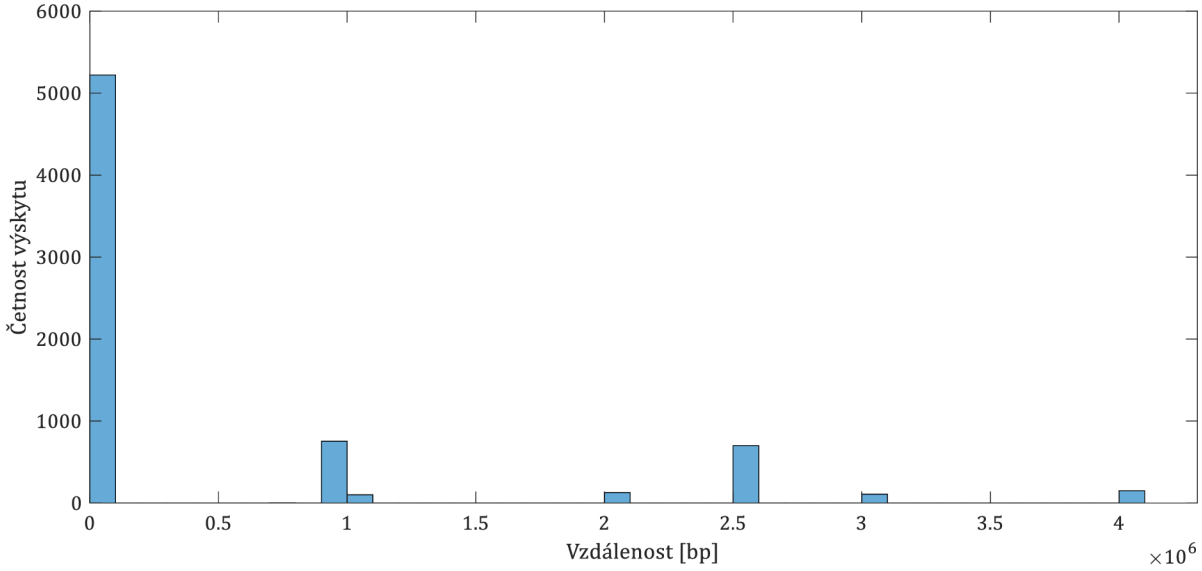


Obrázek 17 – Určení vzdálenosti u párových čtení, vlastní zpracování

Na obrázku č. 18 je zobrazen histogram četnosti výskytu vzdáleností z detekovaného úseku genu RbsR. Z něj je patrné, že v tomto úseku dochází k velkému rozptylu četnosti hodnot vzdáleností párových čtení. Nejvýrazněji jsou zastoupené hodnoty v místě s minimální vzdáleností. To je způsobeno správným namapováním jednotlivých párových čtení. Další detekované hodnoty vykazují již značnou vzdálenost, která je projevem namapování párových čtení z rozhraní detekovaných úseků. Silně zastoupené hodnoty jsou ve vzdálenosti 1 000 000 bp a 2 500 000 bp, které odpovídají místům v genomu, kde se daný gen nachází v referenci. Méně se projevila četnost hodnot ve vzdálenosti 2 000 000 bp, 3 000 000 bp a 4 000 000 bp, které odpovídají místům, kam byl gen umělé vkládán. Přehledná místa vložení jsou ukázána na obrázku č. 10.

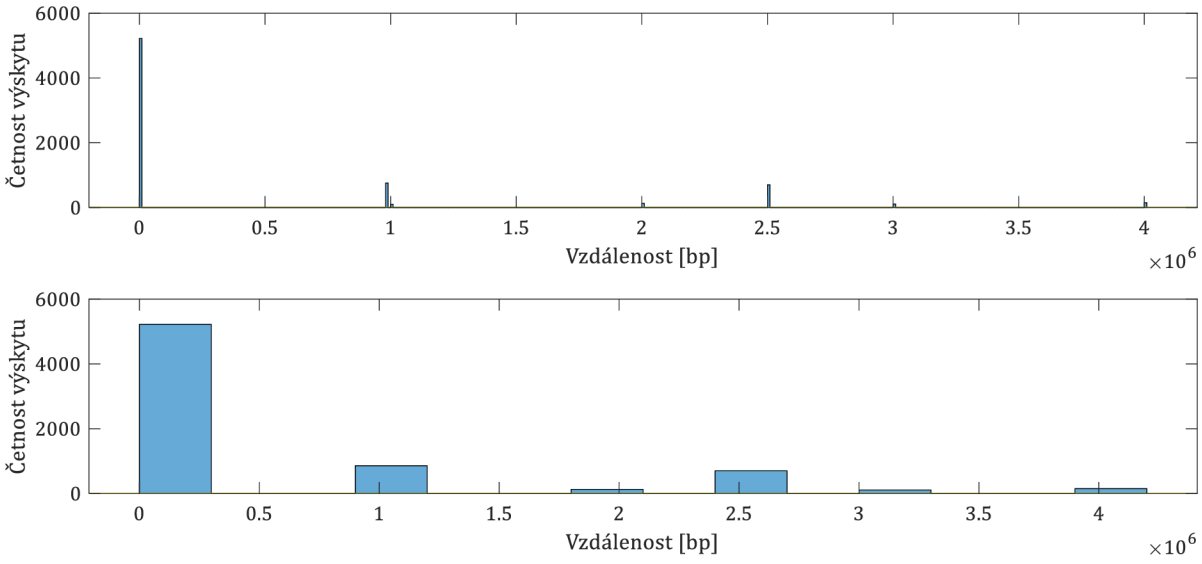
Místa zvýšené četnosti hodnot potvrzují popis obrázku č. 13, kdy dochází k namapování jednotlivých párových čtení do velké vzdálenosti. Na základě toho lze

konstatovat, že pro označení daného úseku jako CNV je nutné, aby došlo k projevení rozptylu četnosti hodnot vzdáleností.



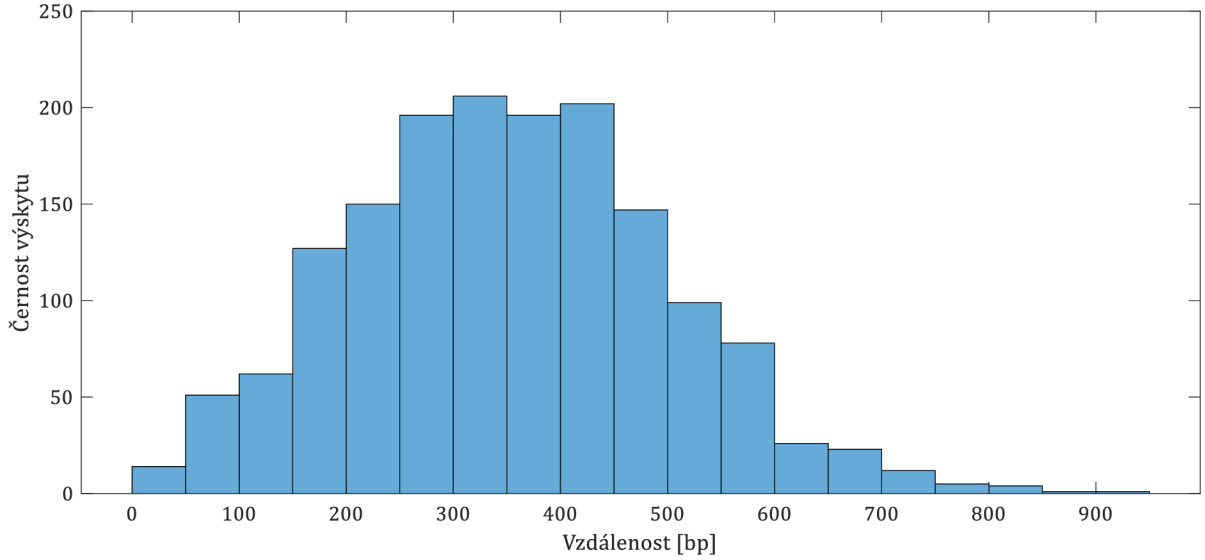
Obrázek 18 - Histogram četnosti vzdáleností genu RbsR pro detekovanou oblast

Z těchto dat také vyplývají požadavky pro správnou detekci, aby se zamezilo falešně pozitivním detekcím. Následující obrázek č. 19 ukazuje nastavení šířky sloupce u dvou histogramů na hodnoty 10 000 a 300 000. Při srovnání s obrázkem č. 18 je patrné, že u všech histogramů byly detekovány hodnoty potvrzující úsek CNV. Z tohoto důvodu byla nastavena šířka sloupce na výchozí hodnotu 100 000 a minimální počet výskytů v daném sloupci na hodnotu 10. Tím bylo docíleno toho, že byl detekován vždy dostatečný rozptyl četnosti hodnot.



Obrázek 19 - Srovnání histogramů četnosti vzdáleností párových čtení

Obrázek č. 20 ukazuje histogram výskytu vzdáleností pro oblast mimo detekovaný úsek CNV. Z histogramu je patrné, že dochází jen k minimálnímu rozptylu četnosti hodnot vzdáleností. Nejčastěji jsou jednotlivá párová čtení vzdálena od sebe v rozmezí pouze 250–450 bp. Tento histogram také odpovídá teoretickým předpokladům, že mimo úseky CNV nedochází k rozdělení párových čtení do větších vzdáleností. Z toho lze vyvodit, že tento úsek neodpovídá CNV.



Obrázek 20 - Histogram četnosti vzdáleností genu RbsR mimo oblast detekce

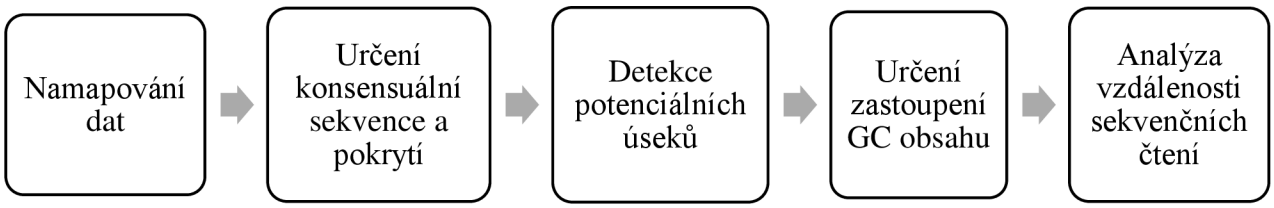
3.3 Návrh algoritmu pro detekci CNV

Na základě získaných výsledků testování v kapitole 3.2 byl vytvořen algoritmus pro detekci CNV v bakteriálním genomu. Následující část popisuje jednotlivé kroky, na obrázku č. 21 je ukázáno schéma navrhovaného algoritmu.

V navrženém algoritmu detekce úseků odpovídajících CNV bude nejprve provedeno genomové sestavení dle templátu pomocí algoritmu BWA a určení konsensuální sekvence a pokrytí. V pokrytí genomu budou detekovány nejprve jednotlivé píky na základě prahu. Ten bude určen pro všechna testovací data jako dvojnásobek průměrné hodnoty. Na základě těchto píků budou určeny celé potenciální úseky CNV pomocí algoritmu založeného na principu seménkové detekce.

U těchto určených úseků bude vypočítáno procentuální zastoupení GC obsahu, kde bude analyzováno, o kolik procent se daný výsledek odlišuje od průměrné

hodnoty celého genomu. Nakonec budou v potenciálních oblastech CNV analyzovány vzdálenosti sekvenčních párových čtení, pomocí nichž lze určit, zda se daný gen nachází v genomu vícekrát. Pokud bude detekována v této oblasti výrazný rozptyl hodnot, bude daný úsek určen jako CNV.



Obrázek 21 - Schéma navrhovaného algoritmu, vlastní zpracování

4 Testování algoritmu

V této kapitole bude otestován algoritmus vytvořený pro detekci CNV bakteriálního genomu vycházející ze znalostí simulovaných dat z předchozí kapitoly.

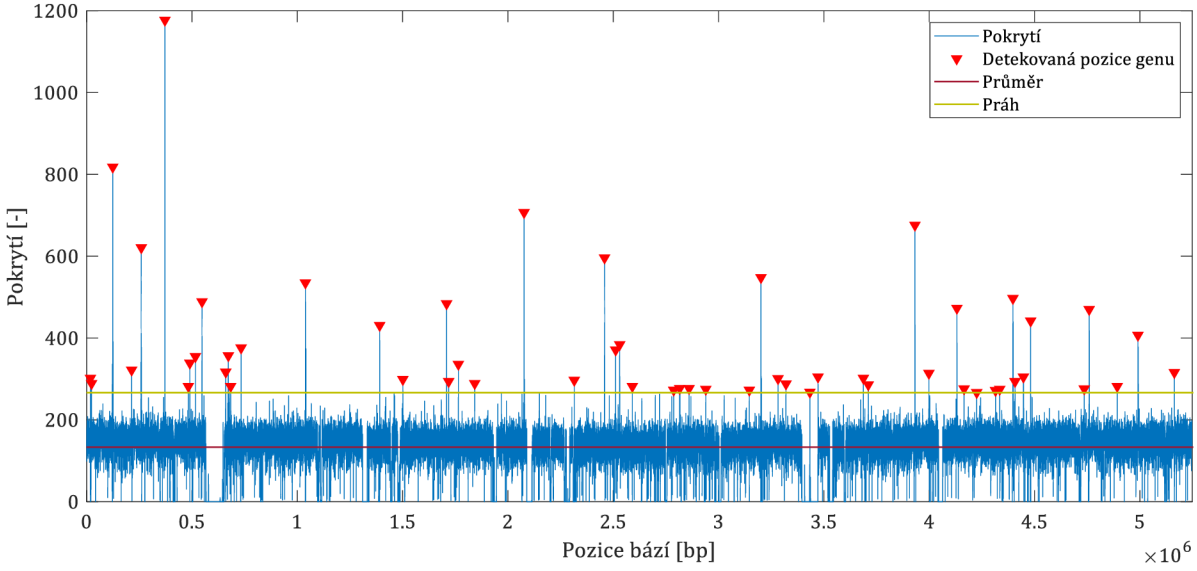
Genomická data, která jsou zpracovávána v této práci, pochází z oddělení mikrobiologie Fakultní nemocnice Brno a jedná se o bakterii *Klebsiella pneumoniae*. Sběr probíhal od roku 2014. Při zpracování těchto dat byla izolována bakteriální DNA, proběhla fragmentace a příprava knihovny pomocí kitu KAPA HyperPrep Kit. Tato připravená knihovna byla sekvenována pomocí sekvenátoru Illumina MiSeq. Po domluvě s vedoucí práce nebyla použita v této práci data bakterie *Clostridium difficile*. Tato bakterie nebyla v době zpracování dat považována za aktuální klinický problém a od dalšího sběru dat na pracovišti FN Brno bylo prozatím upuštěno.

Klebsiella pneumoniae je gram-negativní bakterie, která se řadí do čeledi *Enterobacteriaceae*. Vykazuje blízkou genetickou příbuznost s dalšími bakteriemi z této čeledi např. s *Escherichia*, *Salmonella*, *Shigella*, and *Yersinia*. Hlavním rozdílem je, že *Klebsiella pneumoniae* obsahuje silnou polysacharidovou kapsli. Předpokládá se, že tato kapsle je významným faktorem virulence. Infekce spojené s touto bakterií se objevují po celém světě, především v souvislosti s infekcí močových cest a nozokomiální nákazou. Zároveň se tato bakterie projevuje jako hlavní příčina bakterémie nebo infekcí rezistentních na léky [18].

Algoritmus detekce CNV byl aplikován na čtyřicet osm genomů bakterie *Klebsiella pneumoniae*. Následující část popisuje práci algoritmu na reálných datech a rozdíly oproti simulovaným, které byly rozebrány v kapitole 3.2.

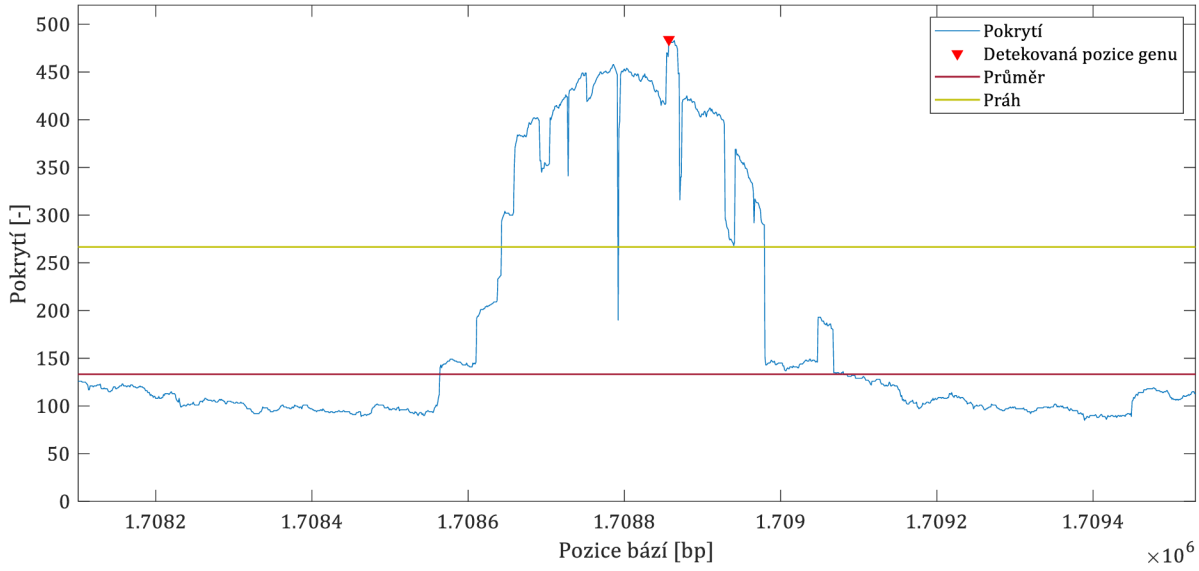
Nejprve bylo provedeno namapování pomocí algoritmu BWA a určení konsensuální sekvence a pokrytí. Tento krok byl shodný jako v případě simulovaných dat.

Obrázek č. 22 ukazuje pokrytí pro bakteriální genom s označením S05 včetně míst, kde byly detekovány potenciální úseky CNV. Z grafu je patrné, že bylo detekováno větší množství potenciálních úseků, v tomto případě padesát pět. V algoritmu bylo dále nastaveno, aby bylo počítáno pouze s úseky delšími než sto bází z důvodu eliminace falešných detekcí. Po této filtraci se počet detekovaných úseků snížil na devatenáct. Podobných hodnot bylo dosaženo u všech čtyřiceti osmi genomů.



Obrázek 22 - Pokrytí bakteriálního genomu S05 s detekovanými úseky

Dále bylo při podrobné analýze výsledků pokrytí zjištěno, že se v místech potenciálních úseků může objevit krátký lokální pokles, který nebyl u simulovaných dat pozorován. Tento pokles může předčasně ukončit detekci počátečních a koncových bodů. Ukázka takového místa je na obrázku č. 23. Z tohoto důvodu byl mírně upraven algoritmus detekce potenciálních úseků CNV.

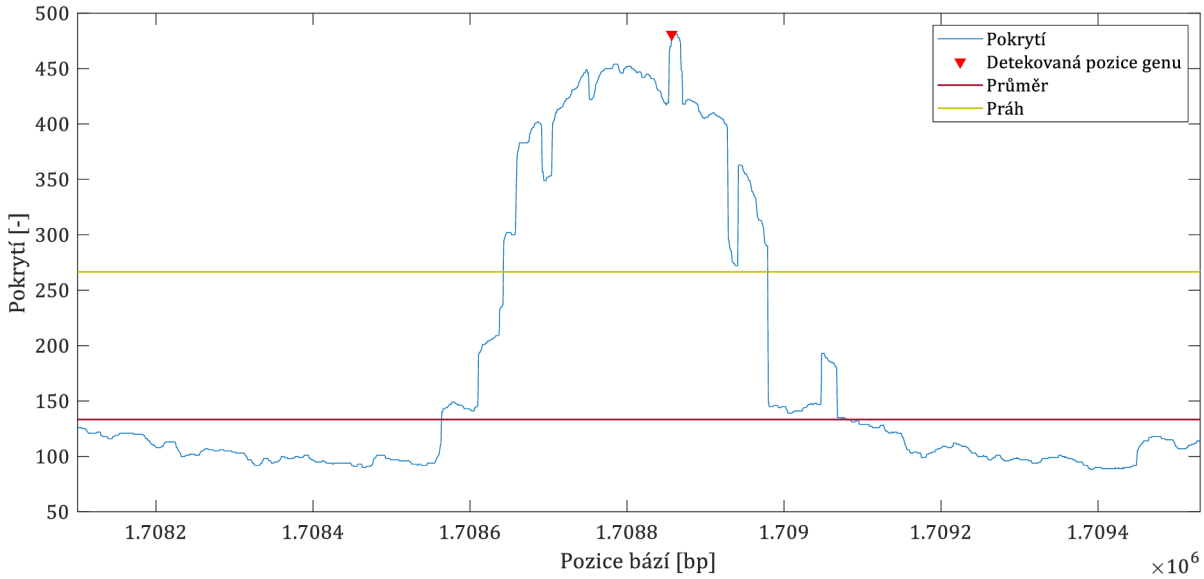


Obrázek 23 - Detail pokrytí u bakteriálního genomu S05

Pro eliminaci předčasného zastavení algoritmu vlivem možnosti lokálního poklesu zobrazeného výše, byly navrženy dvě možnosti. První možnost, založená na heuristickém přístupu, kontrolovala vždy po určení počátečního nebo koncového bodu i následující

tři hodnoty, zda nedochází opět k nárůstu nad stanovený práh. Pokud byl tento nárůst potvrzen, algoritmus pokračoval dále v detekci krajních bodů.

U druhé varianty, založené na signálovém zpracování, byla využita mediánová filtrace. Ta pracuje na principu posuvného okna, ve kterém jsou prvky vstupu seřazeny dle velikosti a na výstup je vybrána hodnota mediánu. Tento algoritmus je tedy schopen zcela odstranit lokální poklesy. Vzhledem k velmi krátkým lokálním poklesům byla zvolena délka okna na hodnotu sedm. Na obrázku č. 24 je ukázán výsledek po tomto typu filtrace, kde je patrné, že došlo k eliminaci poklesu ve srovnání s obrázkem č. 23.



Obrázek 24 - Detail pokrytí u bakteriálního genomu S05 po mediánové filtraci

Po srovnání obou metod bylo zjištěno, že počet výsledných detekovaných potenciálních úseků je shodný. Odlišnosti, které se projevily především v určení začátku a konce úseku, byly v řádu jednotek bází. V následující tabulce č. 6 je ukázán příklad detekovaných úseků pro genom S05, kde jsou zvýrazněny úseky, které byly detekovány odlišně. Pro přesné statistické zhodnocení byly určeny délky jednotlivých úseků, které se podrobily dvouvýběrovému t-testu. Ten porovnává, zda se tyto délky obou metod významně liší. Tento test potvrdil, že rozdíl délek u obou metod není statisticky významný na hladině významnosti 0,05. Co se týká časové náročnosti zmíněných metod k rychlejší detekci docházelo u první metody, a to v průměru o 7,46 %. Z tohoto důvodu byla využita v algoritmu právě tato metoda [30].

Tabulka 6 – Srovnání metod pro detekování potenciálních úseků CNV

Heuristický přístup	**Mediánová filtrace**
547427-547889	547427-547889
660423-660608	**660423-660609**
1390849-1391227	1390849-1391227
1501053-1501161	**1501053-1501158**
1708642-1708980	1708642-1708980
1842293-1842393	**1842293-1842395**
2076395-2076786	2076395-2076786
2458320-2458770	2458320-2458770
2511309-2511438	2511309-2511438
4446965-4447139	**4446964-4447139**
4759045-4759404	4759045-4759404
5163483-5163605	**5163483-5163603**

Následně bylo analyzováno GC zastoupení ve výše detekovaných úsecích. Vzhledem k tomu, že byl počítán GC obsah pro celý genom, tedy pro kódující i nekódující úseky, byly kromě úseků se zvýšeným GC zastoupením vybírány i úseky s průměrnými hodnotami. Filtrovány byly pouze úseky se sníženým GC obsahem, které mohou označovat místa s nižší kvalitou [27].

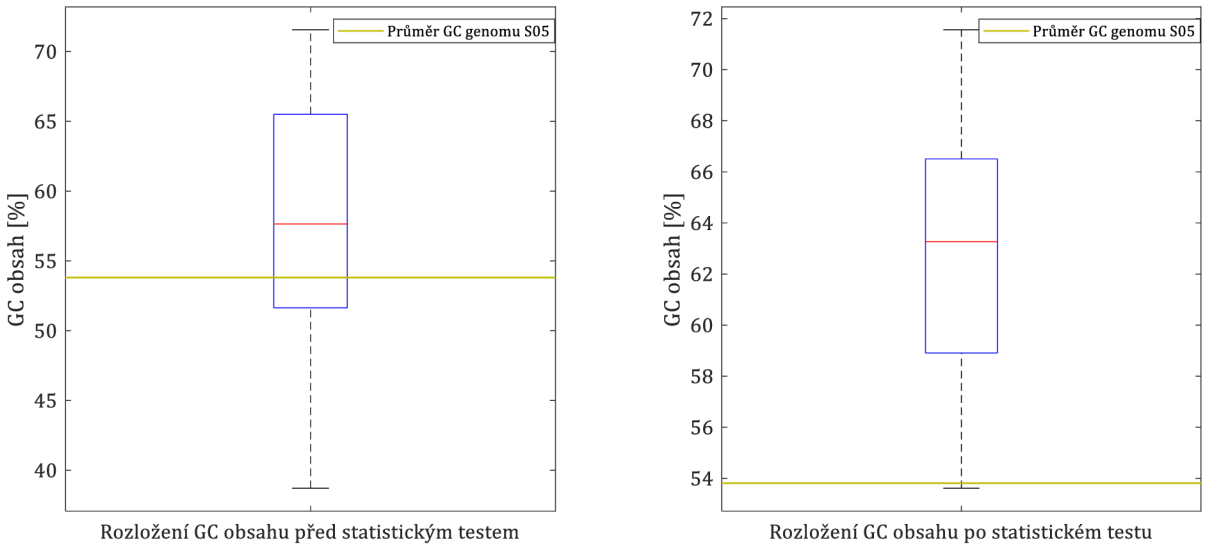
Jak již bylo zmíněno GC obsah každého úseku byl dále testován, kdy bylo nutné určit, které hodnoty lze vybrat kromě výrazně zvýšených jako průměrné a které již ne. Toto testování bylo provedeno pomocí jednovýběrového t-testu. Ten porovnával průměr GC obsahu ze všech genomů s hodnotou z daného úseku. Na základě zamítnutí nebo přijetí nulové hypotézy došlo k určení, zda daná hodnota spadá do průměrných hodnot [30].

Následující tabulka č. 7 ukazuje vybrané úseky dle GC zastoupení pro genom S05, kdy průměrný GC obsah je 53,81 %. Jak již bylo uvedeno dříve, pro tento genom bylo dle pokrytí detekováno celkem devatenáct úseků, z tabulky lze vyčíst, že po analýze GC obsahu jich bylo vybráno dvanáct.

Tabulka 7 - Vybrané úseky dle zastoupení GC obsahu pro bakteriální genom S05

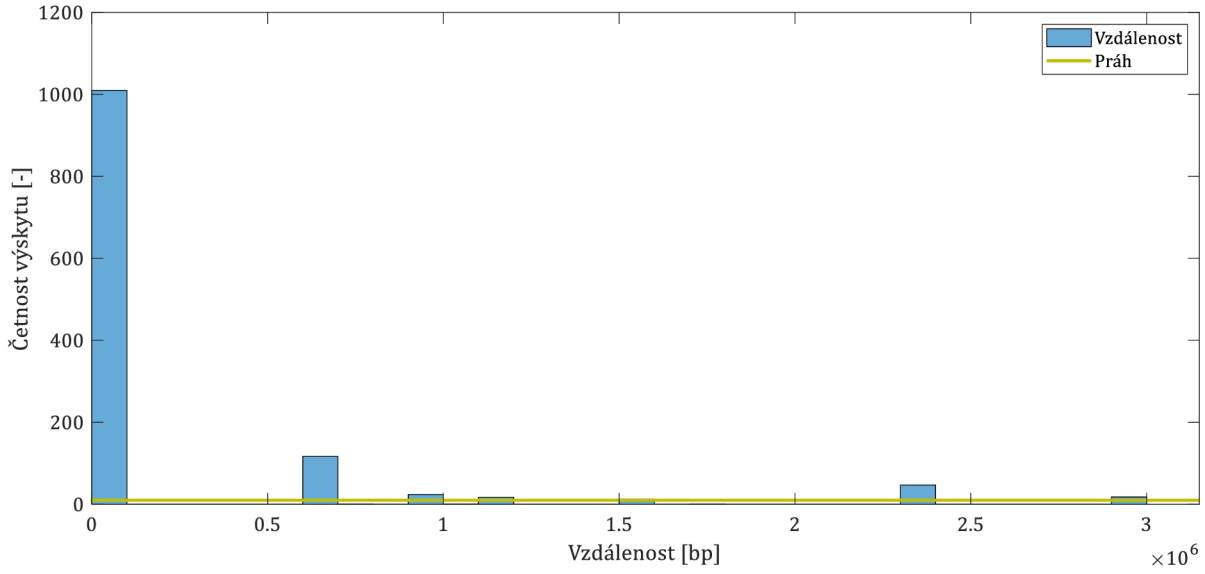
Úsek	**GC obsah [%]**
547427-547889	66,52
660423-660608	60,22
1390849-1391227	66,49
1501053-1501161	71,56
1708642-1708980	69,03
1842293-1842393	61,39
2076395-2076786	57,65
2458320-2458770	65,63
2511309-2511438	54,62
4446965-4447139	65,14
4759045-4759404	53,61
5163483-5163605	60,16

Pro srovnání výsledků z popsaného statistického testu byl vytvořen boxplot genomu S05 před a po testování, kde bylo zaznačeno i průměrné zastoupení GC obsahu v tomto genomu. Na obrázku č. 25 ukazuje levý boxplot GC obsah před výběrem hodnot pomocí statistického testu. Lze vyčíst, že rozmezí hodnot zastoupení se pohybuje od 40 % do 70 % bez výrazných odlehlých hodnot. Zároveň jsou patrné hodnoty pod průměrem. Z pravého boxplotu, který zobrazuje rozložení hodnot vybraných na základě statistického testu, je patrné, že došlo k odebrání hodnot, které jsou výrazně nižší, než je průměr. Tím došlo ke splnění teoretických předpokladů, kdy byly zachovány pouze průměrné a vyšší hodnoty GC obsahu.



Obrázek 25 - Srovnání boxplotů GC obsahu před a po statistické analýze

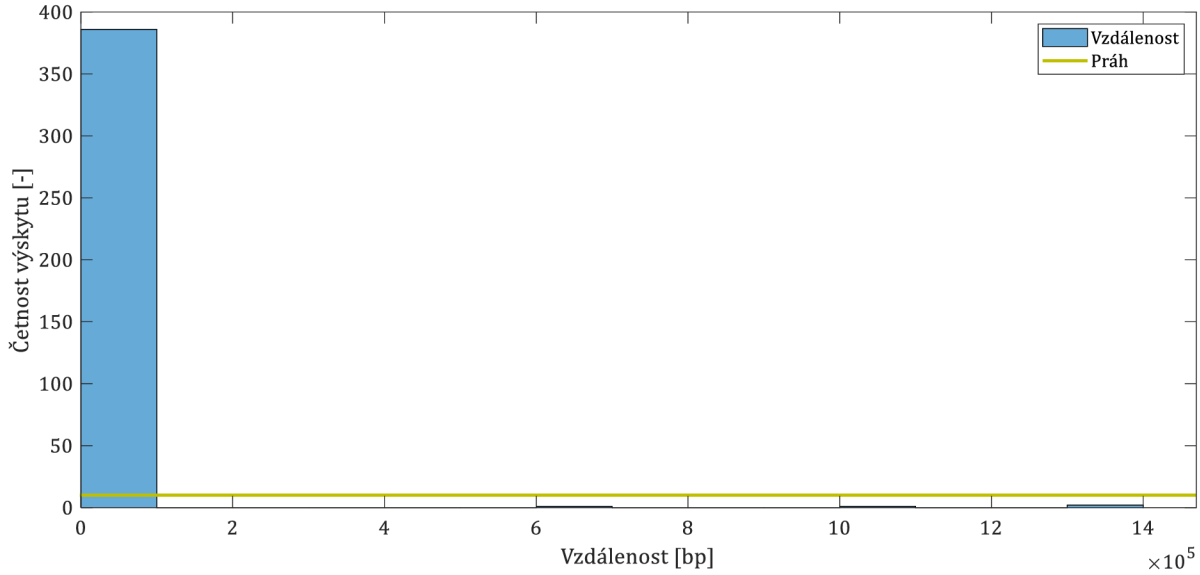
Vybrané úseky z předchozích dvou kroků byly nakonec analyzovány pomocí vzdálenosti jejich sekvenčních párových čtení popsané u simulovaných dat. Následující obrázek č. 26 ukazuje histogram četnosti výskytu vzdáleností párových čtení úseku o pozicích 2076395-2076786 u genomu S05. Z histogramu vyplývá jasně patrný vysoký rozptyl hodnot v počtu vyšším, než je práh, který je zde také zaznačen. Z tohoto důvodu byl tedy tento úsek označen jako CNV.



Obrázek 26 - Histogram četnosti jednotlivých vzdáleností úseku 2076395-2076786

Naopak na následujícím obrázku č. 27 je zaznačen histogram detekovaného úseku o pozicích 3472010-3472129 stejného genomu jako v předchozím případě. Lze sledovat,

že je zde minimální rozptyl hodnot, který nepřekračuje práh. Zároveň jsou tyto hodnoty i o řád nižší než u histogramu na obrázku č. 26. To odpovídá spíše falešným detekcím, a proto tento úsek nebyl označen jako CNV.



Obrázek 27 - Histogram četnosti výskytu vzdáleností úseku 3472010-3472129

5 Zhodnocení výsledků

Algoritmus byl aplikován na čtyřicet osm dostupných genomů bakterie *Klebsiella pneumoniae.* Výsledné detekované úseky CNV byly zaneseny do tabulky v příloze č. 1. Z tabulky je patrné, že bylo detekováno celkově dvacet jedna úseků jako CNV. Z nich pouze čtyři úseky byly určeny jen pro jeden daný genom. Všechny ostatní detekované úseky byly přiřazeny pro více testovaných genomů.

Výsledné detekované úseky CNV byly dále podrobeny analýze pomocí online nástroje BLAST. Jedná se o vyhledávací program sekvenčních podobností provozovaný institutem NCBI. Hledání těchto podobností probíhalo pro všechny kmeny bakterie *Klebsiella pneumoniae* dostupné v databázi NCBI. Byly získány informace o genových produktech daných úseků, které byly vybrány na základě e-hodnoty. Následující tabulka č. 8 obsahuje názvy jednotlivých genových produktů odpovídající detekovaným úsekům. Z tabulky je patrné, že pouze tři genové produkty jsou v detekovaných úsecích CNV zastoupené více než jednou.

Tabulka 8 - Označení jednotlivých úseků dle BLAST

Úsek	**Protein**
547077–547657	Chromosomal replication initiator protein DnaA
660496–661385	Carbohydrate porin
1109193–1109433	LysR family transcriptional regulator
1366216–1366459	Gfo/Idh/MocA family oxidoreductase
1390846–1391168	Class II fructose-bisphosphatase
1419824–1420105	IS3 family transposase
1459469–1459796	Yersiniabactin polyketide synthase HMWP1
1460964–1461247	Erythronate-4-phosphate dehydrogenase
1648031–1648271	Cell division protein CpoB
1708620–1708980	Nitrate ABC transporter substrate-binding protein
2076400–2076966	IS3 family transposase
2149373–2149664	Redox-regulated ATPase YchF
2458299–2458762	Hypothetical protein
2722863–2723156	Chromosomal replication initiator protein DnaA
2892904–2893852	Transposase
3145858–3146135	Chromosomal replication initiator protein DnaA
3435949–3436183	ISNCY family transposase
4447579–4448394	Transketolase
4869372–4869921	Glycogen-branching enzyme
4991933–4992237	IS3-like element ISKpn1 family transposase
5162934–5163594	Carbohydrate porin

Vzhledem k tomu, že pro některé genomy byly detekovány shodné úseky, které se lišily pouze v počátečních a koncových bodech v řádu jednotek bází, byly následně určeny jednotné počáteční a koncové body těchto úseků. Ty byly vypočítány pomocí průměru těchto bodů ze všech detekovaných úseků. Pro přehlednost ukazuje následující tabulka č. 9 detekované úseky pro bakteriální genom S05. Z tabulky lze vyčíst, že pro tento genom bylo celkem detekováno pět úseků.

Tabulka 9 - Úseky CNV pro genom S05

ID/Úsek	547077: 547657	660496: 661385	1708620:1708980	2076400:2076966	2458299:2458762
S05	+	+	+	+	+

Při podrobné analýze výsledných hodnot byly odhaleny podobnosti u vybraných genomů, kdy mezi nimi byly detekovány shodné úseky CNV. Bylo zjištěno, že tyto skupiny podobných genomů mají společnou vlastnost, kterou je melt typ. Ten slouží k rozlišení jednotlivých genomů na základě křivky teploty tání a byl získán laboratorní metodou použitím typizační metody mini-MLST. Tato metoda je založena na amplifikaci genů pomocí PCR a následné vysokorozlišovací analýze křivek tání. Výsledkem jsou rozdílné křivky tání, které představují jednotlivé tavící alely charakterizované převážně obsahem guanin-cytosinu. Následující tabulka č. 10 ukazuje souhrn tohoto rozdělení pro všechny genomy [31], [32].

Tabulka 10 - Rozdělení bakteriálních genomů dle melt typu

ID	Melt typ	ID	Melt typ	ID	Melt typ
S01	98	**S17**	23	**S33**	16
S02	29	**S18**	23	**S34**	16
S03	98	**S19**	61	**S35**	61
S04	98	**S20**	61	**S36**	61
S05	29	**S21**	61	**S37**	29
S06	98	**S22**	61	**S38**	29
S07	29	**S23**	61	**S39**	29
S08	98	**S24**	61	**S40**	29
S09	29	**S25**	61	**S41**	23
S10	98	**S26**	61	**S42**	29
S11	29	**S27**	14	**S43**	23
S12	29	**S28**	14	**S44**	61
S13	23	**S29**	98	**S45**	44
S14	23	**S30**	98	**S46**	44
S15	23	**S31**	14	**S47**	95
S16	23	**S32**	14	**S48**	29

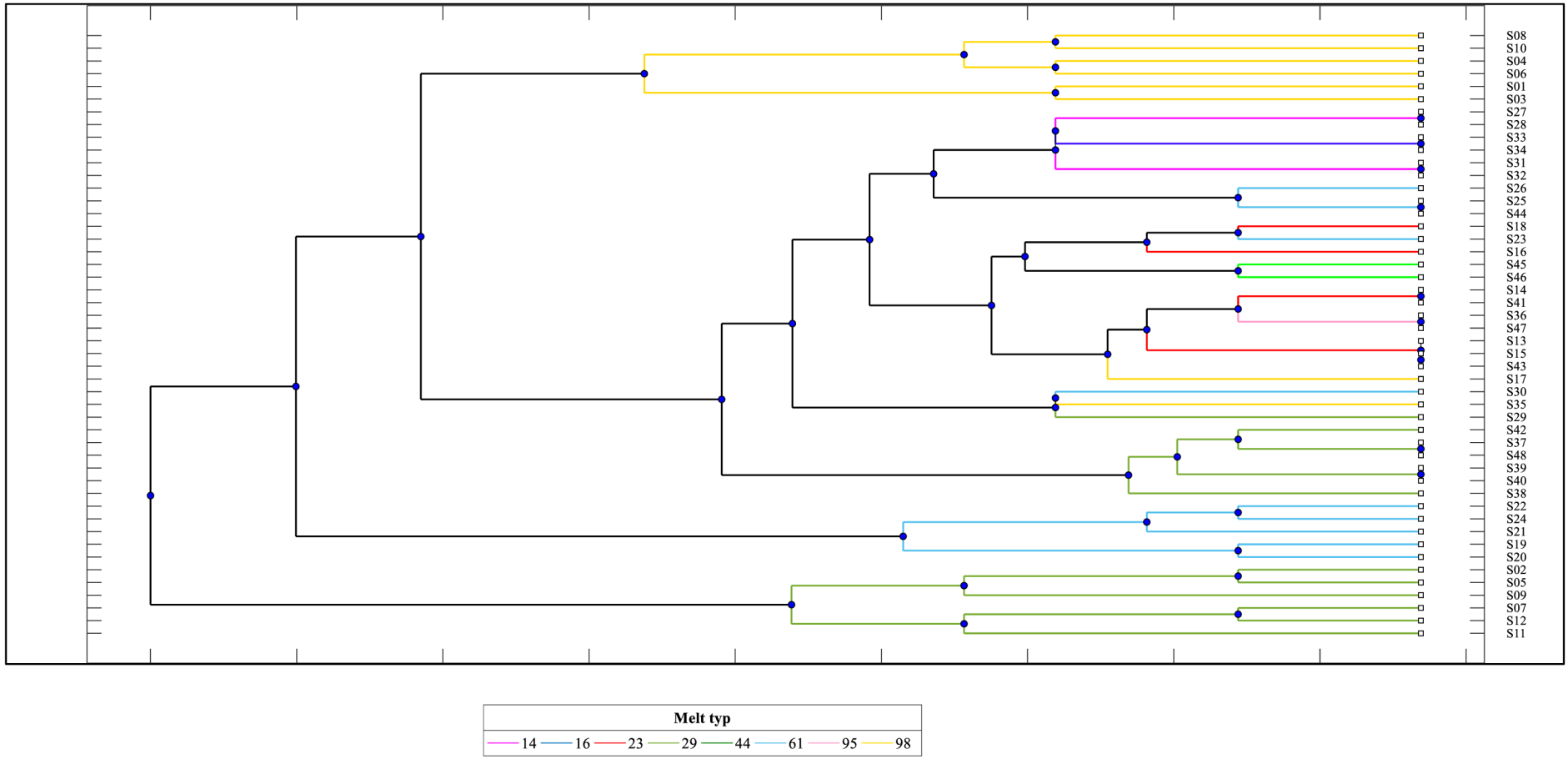
Tabulka č. 11 znázorňuje příklad pěti detekovaných úseků CNV u bakteriálních genomů S13-S18, S41 a S43, které mají melt typ 23. U všech těchto genomů byl detekován úsek CNV na pozicích 1459469-1459796. Dále byl u pěti genomů detekován jako CNV úsek 1109193-1109433. Naopak úsek 4869372-4869921 byl detekován pouze jednou nejen v rámci této skupiny, ale v rámci všech genomů.

Tabulka 11 - Genomy odpovídající melt typu 23

ID/Úsek	1109193:1109433	1459469:1459796	1648031:1648271	2149373:2149664	4869372:4869921
S13		+			
S14	+	+			
S15		+			
S16	+	+		+	+
S17	+	+	+		
S18	+	+		+	
S41	+	+			
S43		+			

Pro grafické znázornění výsledků detekce CNV byl vytvořen dendogram znázorňující vztahy mezi genomy pomocí metody UPGMA. Tato metoda vychází z distanční matice, kdy se spojí vždy dvojice s nejmenší vzdáleností. Místa v matici, které odpovídají spojeným genomům, jsou nahrazeny jejich aritmetickým průměrem. Jako metrika pro výpočet distanční matice mezi jednotlivými genomy byla zvolena Jaccardova vzdálenost. Ta popisuje podobnost každých dvou genomů jako podíl počtu společných atributů a celkovém počtu atributů [33], [34].

Výsledný dendogram je zobrazen na obrázku č. 28, kde jsou barevně odlišeny jednotlivé genomy dle melt typu. Na základě tohoto stromu lze usoudit, že většina testovaných genomů není zcela správně odlišena podle daných melt typů. Správně jsou v tomto stromě odděleny od ostatních pouze genomy patřící do melt typu 44, který ovšem obsahuje pouze dva genomy, a to S45 a S46. Ostatní skupiny genomů jsou rozděleny do větších vzdáleností od sebe. Melt typy 14, 29 a 98 se projevily rozdělením do dvou shluků. Nejhůře byly určeny melt typy 61 a 23, které vykazovali největší rozdíly vzdáleností mezi jednotlivými genomy.



Obrázek 28 - Dendogram rozlišující jednotlivé genomy dle melt typu

Dále bylo zjištěno, že testované genomy mezi sebou vykazují výraznou odlišnost v jejich průměrné hodnotě pokrytí. Následující tabulka č. 12 ukazuje průměrné pokrytí všech genomů. Z ní je patrné, že prvních dvacet čtyři genomů má výrazně vyšší míru pokrytí. Na základě této rozdílnosti byly vytvořeny další dva dendogramy, kdy každý byl vytvořen z poloviny genomů. Z prvního dendogramu, který je zobrazen v příloze č. 2, je patrné, že genomy byly seřazeny do shluků dle melt typu až na chybné zařazení genomu S23. Naopak druhý dendogram v příloze č. 3 zobrazuje rozdělení genomů s větším počtem odchylek. Správně byly seskupeny genomy mající melt typ 16, 29, 44 a 98. Ostatní data byla přiřazena do shluků, které kombinovaly genomy z více melt typů.

Tabulka 12 - Průměrné pokrytí genomů S1 – S48

ID	Pokrytí [-]	ID	Pokrytí [-]	ID	Pokrytí [-]
S1	102,63	**S17**	73,62	**S33**	31,33
S2	165,65	**S18**	98,08	**S34**	32,76
S3	88,83	**S19**	72,36	**S35**	25,61
S4	98,16	**S20**	65,20	**S36**	26,58
S5	133,30	**S21**	69,00	**S37**	30,03
S6	158,93	**S22**	60,82	**S38**	26,42
S7	136,31	**S23**	67,25	**S39**	26,33
S8	140,50	**S24**	55,92	**S40**	26,17
S9	76,80	**S25**	33,51	**S41**	36,69
S10	159,16	**S26**	31,36	**S42**	26,32
S11	195,90	**S27**	29,18	**S43**	27,01
S12	128,02	**S28**	25,08	**S44**	27,06
S13	96,78	**S29**	29,75	**S45**	24,91
S14	95,88	**S30**	54,42	**S46**	20,83
S15	86,21	**S31**	25,27	**S47**	24,57
S16	76,91	**S32**	28,42	**S48**	23,93

Z těchto výsledků lze vyvodit, že spolu souvisí nejen detekované CNV úseky genomů a jejich melt typ, ale také záleží i na míře pokrytí konkrétního genomu.

Jak již bylo zmíněno, bakterie *Klebsiella pneumoniae* je spojována s celou řadou onemocnění včetně rezistence vůči antibiotikům. Z výsledků prezentovaných v této práci a rozlišení jednotlivých bakteriálních genomů dle melt typů lze konstatovat, že antibiotická rezistence není jednostranná. Výsledky odpovídají tomu, že antibiotická rezistence závisí na celé řadě faktorů. Mezi ně je nutné zařadit i rozdělení genomů dle melt typů genomů a detekci úseků CNV. Ty mohou přímo určovat predispozice, ze kterých se může vyvinout rezistence na léčiva.

Závěr

Cílem této diplomové práce byla detekce CNV u genomů bakterie *Klebsiella pneumoniae*, kdy všechny testovací genomy byly shromážděny ve Fakultní nemocnici Brno. Práce přináší přehled teoretických znalostí, analýzu bakteriálních dat a samotnou detekci CNV úseků metodou vytvořenou přímo pro tuto problematiku.

Teoretická část se zaobírala literární rešerší genomu z pohledu strukturní variability, kde byly popsány nejčastější odchylky. Dále byly uvedeny metody sekvenování genomu napříč všemi generacemi. Poté se práce zabývala popisem variability počtu kopií a dostupnými nástroji pro jejich detekci. Bylo zjištěno, že velká část dostupných nástrojů využívaných pro detekci CNV, byla originálně vytvořena pro zcela odlišné účely.

Praktická část práce se věnovala nejprve přípravě a otestování simulovaných dat, která byla vytvořena z kmene NTUH-K2044 referenčního genomu *Klebsiella pneumoniae*. Ta byla analyzována na základě nerovnoměrného pokrytí v genomovém sestavení, zastoupení GC obsahu a vzdálenosti sekvenčních čtení. Zároveň byla tato simulovaná data porovnána u jednotlivých analýz s teoretickými předpoklady. Ze získaných znalostí byl navrhnut algoritmus detekce CNV.

Navržený algoritmus byl spuštěn na čtyřiceti osmi genomických datech. Po proběhnutí algoritmu byly detekovány úseky CNV pro jednotlivé genomy. Výsledky byly prezentovány mimo jiné graficky pomocí dendogramu. Bylo zjištěno, že vybrané výsledné detekované úseky vykazují podobnost mezi genomy na základě jejich melt typu. Přestože nebyly všechny genomy jednotlivých melt typů seskupeny s naprostou přesností, lze konstatovat, že detekované CNV úseky v rámci daného melt typu mohou spolu souviset a tvořit predispozice pro určité atributy u daného organismu.

Nejen vzrůstající antibiotická rezistence bakterie *Klebsiella pneumoniae*, ale i další onemocnění jí způsobenou jen potvrzují, že je nutné se ve výzkumu zaměřit více na prokaryotické organismy, které mohou způsobit až epidemie. Tato práce může být přínosem v oblasti detekce CNV a může sloužit jako podklad pro další zkoumání. Se stále větším rozvojem inovativních technologií, které se objevují v laboratořích, lze pokračovat v podrobném výzkumu této bakterie a dosáhnout pochopení veškerých jejich vlastností. Tyto závěry mohou vést až k potlačení jejích negativních atributů.

Seznam literatury

[1] BRYNILDSRUD, Ola, Lars-Gustav SNIPEN a Jon BOHLIN. CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data. *Bioinformatics* [online]. 2015, **31**(11), 1708-1715 [cit. 2018-10-26]. DOI: 10.1093/bioinformatics/btv070. ISSN 1460-2059. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv070

[2] YANG, Tie-Lin, Yan GUO, Christopher PAPASIAN a Hong-Wen DENG. Copy Number Variation. *Genetics of Bone Biology and Skeletal Disease* [online]. Elsevier, 2013, s. 123-132 [cit. 2018-10-26]. DOI: 10.1016/B978-0-12-387829-8.00009-3. ISBN 9780123878298. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/B9780123878298000093

[3] FEUK, Lars, Andrew CARSON a Stephen SCHERER. Structural variation in the human genome. *Nature Reviews Genetics* [online]. 2006, **7**(2), 85-97 [cit. 2018-11-10]. DOI: 10.1038/nrg1767. ISSN 1471-0056. Dostupné z: http://www.nature.com/doifinder/10.1038/nrg1767

[4] BAKER, Monya. Structural variation: the genome's hidden architecture. *Nature Methods* [online]. 2012, **9**(2), 133-137 [cit. 2018-11-24]. DOI: 10.1038/nmeth.1858. ISSN 1548-7091. Dostupné z: http://www.nature.com/articles/nmeth.1858

[5] BHAT, Tariq. *Chromosome structure and aberrations*. New York, NY: Springer Berlin Heidelberg, 2017. ISBN 978-81-322-3671-9.

[6] KOUBKOVÁ, L., B. VOJTĚŠEK a R. VYZULA. Sekvenování nové generace a možnosti jeho využití v onkologické praxi. *Klin Onkol*. 2014, (), 61-68. ISSN 0862-495X. Dostupné také z: https://www.linkos.cz/files/klinicka-onkologie/186/4483.pdf

[7] RAVI, Indu, Jyoti SAXENA a Mamta BAUNTHIYAL. *Advances in biotechnology*. New York: Springer, 2013. ISBN 978-81-322-1553-0.

[8] KOLÍSKO, Martin. Moderní metody sekvenování DNA. *Živa*. Nakladatelství Academia, 2017, (32017), 73-76. Dostupné také z: http://ziva.avcr.cz/files/ziva/pdf/moderni-metody-sekvenovani-dna.pdf

[9] Flowgram. In: *An assembly of reads, contigs and scaffolds* [online]. 2010 [cit. 2018-10-28]. Dostupné z: https://contig.wordpress.com/2010/10/28/newbler-input-i-the-sff-file/

[10] LIEKE, Thorsten. Vervielfältigung von NGS-Bibliotheken ohne PCR. *LaborJournal* [online]. b.r. [cit. 2018-12-13]. Dostupné z: https://www.laborjournal.de/rubric/methoden/methoden/v137.lasso

[11] GONG, Jiao, Tingcai CHENG, Yuqian WU, Xi YANG, Qili FENG a Kazuei MITA. Genome-wide patterns of copy number variations in Spodoptera litura. *Genomics* [online]. 2018 [cit. 2018-10-26]. DOI: 10.1016/j.ygeno.2018.08.002. ISSN 08887543. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0888754318302222

[12] ESTIVILL, Xavier a Lluís ARMENGOL. Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies. *PLoS Genetics* [online]. 2007, **3**(10), 07---0579 [cit. 2018-11-24]. DOI: 10.1371/journal.pgen.0030190. ISSN 1553-7390. Dostupné z: https://dx.plos.org/10.1371/journal.pgen.0030190

[13] MAGI, Alberto, Lorenzo TATTINI, Tommaso PIPPUCCI, Francesca TORRICELLI a Matteo BENELLI. Read count approach for DNA copy number variants detection. *Bioinformatics* [online]. 2012, **28**(4), 470-478 [cit. 2018-11-24]. DOI: 10.1093/bioinformatics/btr707. ISSN 1460-2059. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr707

[14] ZHAO, Min, Qingguo WANG, Quan WANG, Peilin JIA a Zhongming ZHAO. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* [online]. 2013, **14**(11) [cit. 2018-11-24]. DOI: 10.1186/1471-2105-14-S11-S1. ISSN 1471-2105. Dostupné z: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-S11-S1

[15] ABYZOV, A., A. URBAN, M. SNYDER a M. GERSTEIN. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research* [online]. 2011, **21**(6), 974-984 [cit. 2018-12-15]. DOI: 10.1101/gr.114876.110. ISSN 1088-9051. Dostupné z: http://genome.cshlp.org/cgi/doi/10.1101/gr.114876.110

[16] YOON, S., Z. XUAN, V. MAKAROV, K. YE a J. SEBAT. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* [online]. 2009, **19**(9), 1586-1592 [cit. 2018-12-15]. DOI: 10.1101/gr.092981.109. ISSN 1088-9051. Dostupné z: http://genome.cshlp.org/cgi/doi/10.1101/gr.092981.109

[17] XIE, Chao a Martti TAMMI. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* [online]. 2009, **10**(1) [cit. 2018-12-15]. DOI: 10.1186/1471-2105-10-80. ISSN 1471-2105. Dostupné z: http://www.biomedcentral.com/1471-2105/10/80

[18] WU, K.-M., L.-H. LI, J.-J. YAN et al. Genome Sequencing and Comparative Analysis of Klebsiella pneumoniae NTUH-K2044, a Strain Causing Liver Abscess and Meningitis. *Journal of Bacteriology* [online]. 2009, **191**(14), 4492-4501 [cit. 2019-04-26]. DOI: 10.1128/JB.00315-09. ISSN 0021-9193. Dostupné z: http://jb.asm.org/cgi/doi/10.1128/JB.00315-09

[19] LAI, Y.-C., H.-L. PENG a H.-Y. CHANG. Identification of Genes Induced In Vivo during Klebsiella pneumoniae CG43 Infection. *Infection and Immunity* [online]. 2001, **69**(11), 7140-7145 [cit. 2018-11-19]. DOI: 10.1128/IAI.69.11.7140-7145.2001. ISSN 0019-9567. Dostupné z: http://iai.asm.org/cgi/doi/10.1128/IAI.69.11.7140-7145.2001

[20] HUANG, Weichun, Leping LI, Jason MYERS a Gabor MARTH. ART: a next-generation sequencing read simulator. *Bioinformatics* [online]. 2012, **28**(4), 593-594 [cit. 2018-11-09]. DOI: 10.1093/bioinformatics/btr708. ISSN 1460-2059. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr708

[21] COCK, Peter, Christopher FIELDS, Naohisa GOTO, Michael HEUER a Peter RICE. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* [online]. 2010, **38**(6), 1767-1771 [cit. 2018-

11-19]. DOI: 10.1093/nar/gkp1137. ISSN 0305-1048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp1137

[22] LI, H. a R. DURBIN. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* [online]. 2009, **25**(14), 1754-1760 [cit. 2018-11-19]. DOI: 10.1093/bioinformatics/btp324. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324

[23] LI, H., B. HANDSAKER, A. WYSOKER et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [online]. 2009, **25**(16), 2078-2079 [cit. 2018-11-19]. DOI: 10.1093/bioinformatics/btp352. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352

[24] BIAGINI, Tommaso, Barbara BARTOLINI, Emanuela GIOMBINI, Maria CAPOBIANCHI, Fabrizio FERRè, Giovanni CHILLEMI a Alessandro DESIDERI. Performances of Bioinformatics Pipelines for the Identification of Pathogens in Clinical Samples with the De Novo Assembly Approaches: Focus on 2009 Pandemic Influenza A (H1N1). *The Open Bioinformatics Journal* [online]. 2014, **8**(1), 1-5 [cit. 2019-05-09]. DOI: 10.2174/1875036201408010001. ISSN 1875-0362. Dostupné z: https://openbioinformaticsjournal.com/VOLUME/8/PAGE/1/

[25] SIMS, David, Ian SUDBERY, Nicholas ILOTT, Andreas HEGER a Chris PONTING. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* [online]. 2014, **15**(2), 121-132 [cit. 2018-11-19]. DOI: 10.1038/nrg3642. ISSN 1471-0056. Dostupné z: http://www.nature.com/articles/nrg3642

[26] GLÉMIN, Sylvain, Yves CLÉMENT, Jacques DAVID a Adrienne RESSAYRE. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends in Genetics* [online]. 2014, **30**(7), 263-270 [cit. 2018-11-23]. DOI: 10.1016/j.tig.2014.05.002. ISSN 01689525. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0168952514000808

[27] SANTANI, Avni, Jill MURRELL, Birgit FUNKE et al. Development and Validation of Targeted Next-Generation Sequencing Panels for Detection of Germline Variants in Inherited Diseases. *Archives of Pathology & Laboratory Medicine* [online]. 2017,

141(6), 787-797 [cit. 2019-05-10]. DOI: 10.5858/ arpa.2016-0517-RA. Dostupné z: http://doi.org/10.5858/arpa.2016-0517-RA

[28] ZARE, Fatima, Michelle DOW, Nicholas MONTELEONE, Abdelrahman HOSNY a Sheida NABAVI. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics* [online]. 2017, **18**(1) [cit. 2019-05-09]. DOI: 10.1186/s12859-017-1705-x. ISSN 1471-2105. Dostupné z: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1705-x

[29] OKONECHNIKOV, Konstantin, Olga GOLOSOVA a Mikhail FURSOV. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* [online]. 2012, **28**(8), 1166-1167 [cit. 2019-05-09]. DOI: 10.1093/bioinformatics/bts091. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article/28/8/1166/195474

[30] ANDĚL, Jiří. *Matematická statistika*. 1. Praha: SNTL - Nakladatelství technické literatury, 1985.

[31] NYKRYNOVA, Marketa, Denisa MADERANKOVA, Helena SKUTKOVA, Matej BEZDICEK a Martina LENGEROVA. Bioinformatic tools for genotyping of Klebsiella pneumoniae isolates. *Proceedings of the Advances in Intelligent Systems and Computing*. 2019, , 419-428.

[32] 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. *Diagnostic Microbiology and Infectious Disease* [online]. 2016, **86**(1), 44-49 [cit. 2019-05-09]. DOI: 10.1016/j.diagmicrobio.2016.03.010. ISSN 07328893. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0732889316300505

[33] IRANI, Jasmine, Nitin PISE a Madhura PHATAK. Clustering Techniques and the Similarity Measures used in Clustering: A Survey. *International Journal of Computer Applications* [online]. 2016, **134**(7), 9-14 [cit. 2019-05-09]. DOI: 10.5120/ijca2016907841. ISSN 09758887. Dostupné z: http://www.ijcaonline.org/research/volume134/number7/irani-2016-ijca-907841.pdf

[34] MOULTON, Vincent, Andreas SPILLNER a Taoyang WU. UPGMA and the normalized equidistant minimum evolution problem. *Theoretical Computer Science* [online]. 2018, **721**, 1-15 [cit. 2019-05-09]. DOI: 10.1016/j.tcs.2018.01.022. ISSN 03043975. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0304397518300690

Seznam symbolů, veličin a zkratek

A	Adenin
aCGH	Array Comparative Genomic Hybridization
BLAST	Basic Local Alignment Search Tool
bp	Báze
BWA	Burrows-Wheeler Alignment Tool
C	Cytosin
CBS	Circular Binary Segmentation
CNV	Variabilita počtu kopií (Copy number variation)
DNA	Deoxyribonukleová kyselina
EWT	Event-Wise Testing
G	Guanin
GC	Guanino-cytosinový komplementární pár
HMM	Skrytý Markovův model
MSB	Mean Shift-Based
NCBI	National Center for Biotechnology Information
PCR	Polymerázová řetězová reakce (Polymerase Chain Reaction)
RbsR	Ribose operon repressor
SAM	The Sequence Alignment/Map
SLM	Shifting Level Model
SNP	Jednonukleotidové polymorfismy
SOLID	Sekvenování oligonukleotidovou ligací a detekcí
T	Thymin
UPGMA	Metoda párování pomocí nevážených aritmetických průměrů

Seznam příloh

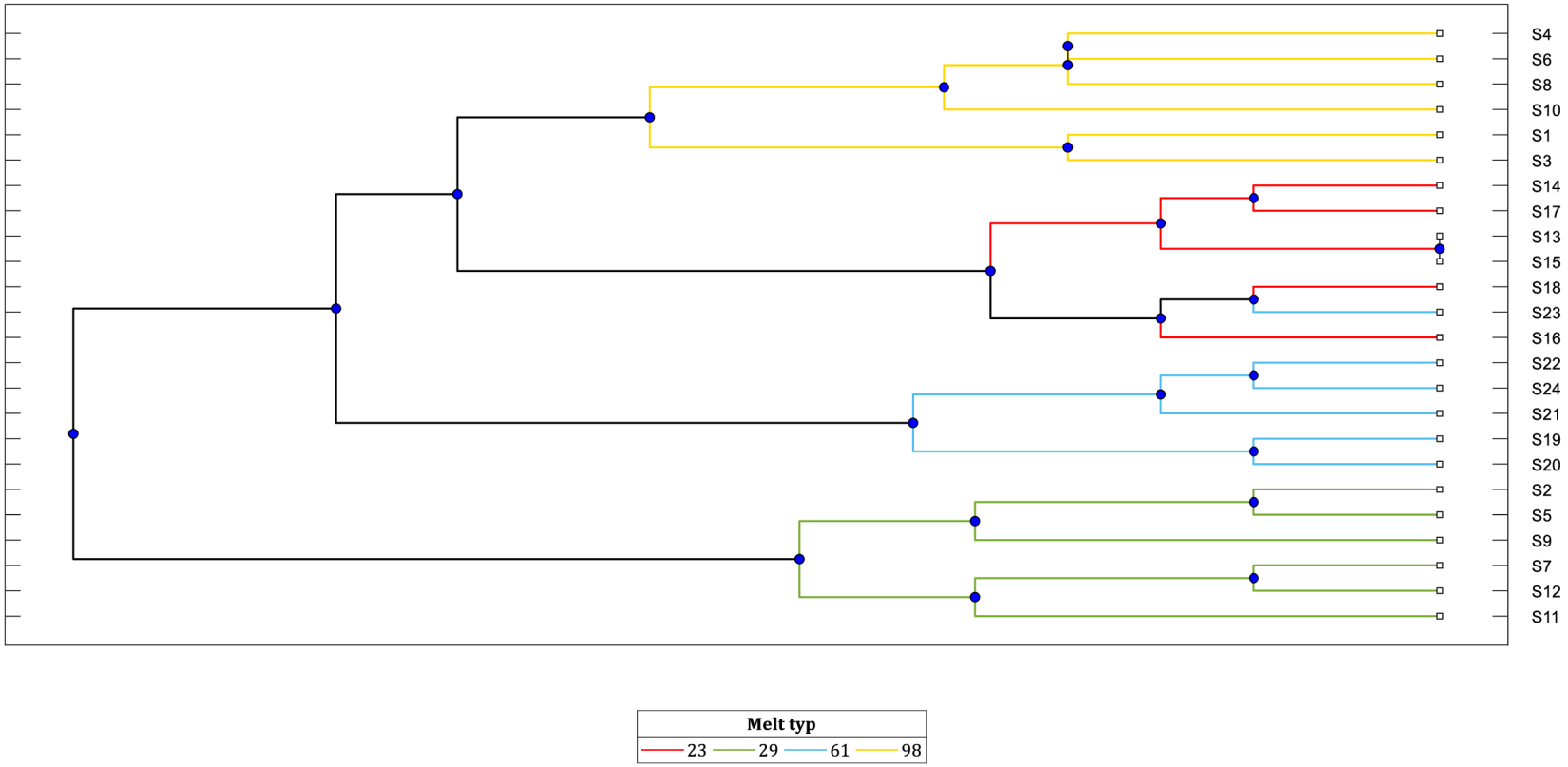
Příloha 1 - Detekované úseky CNV všech genomů

ID/Úsek	547077: 547657	660496: 661385	1109193: 1109433	1366216: 1366459	1390846: 1391168	1419824: 1420105	1459469: 1459796	1460964: 1461247	1648031: 1648271	1708620: 1708980	2076400: 2076966	2149373: 2149664	2458299: 2458762	2722863: 2723156	2892904: 2893852	3145858: 3146135	3435949: 3436183	4447579: 4448394	4869372: 4869921	4991933: 4992237	5162934: 5163594
S01					+	+	+					+	+	+							
S02	+	+				+				+	+		+								
S03						+	+					+	+								
S04	+						+						+								
S05	+	+								+	+		+								
S06	+						+						+	+		+					
S07	+				+	+				+	+		+								
S08	+					+	+						+			+					
S09	+									+	+	+	+								
S10	+				+	+	+						+	+		+					
S11	+				+	+				+	+	+	+					+			
S12	+				+	+				+	+		+							+	
S13							+														
S14			+				+														
S15							+														
S16			+				+					+							+		
S17			+				+		+												
S18			+				+					+									
S19			+	+		+						+	+				+				+
S20			+			+						+	+				+				
S21			+	+		+		+				+	+								
S22			+			+		+				+	+								
S23			+									+									
S24	+		+			+		+				+	+								
S25								+				+									
S26								+													
S27		+																			
S28		+																			
S29													+			+					
S30					+								+								
S31	+																				
S32	+																				
S33															+						

Příloha 1 - pokračování z předchozí strany

ID/Úsek	547077: 547657	660496: 661385	1109193: 1109433	1366216: 1366459	1390846: 1391168	1419824: 1420105	1459469: 1459796	1460964: 1461247	1648031: 1648271	1708620: 1708980	2076400: 2076966	2149373: 2149664	2458299: 2458762	2722863: 2723156	2892904: 2893852	3145858: 3146135	3435949: 3436183	4447579: 4448394	4869372: 4869921	4991933: 4992237	5162934: 5163594
S34															+						
S35			+										+								
S36			+																		
S37							+				+									+	
S38	+						+				+									+	
S39											+									+	
S40											+									+	
S41			+				+														
S42							+				+										
S43							+														
S44								+				+									
S45							+					+									
S46		+					+					+									
S47			+																		
S48							+				+									+	

Příloha 2 - Dendogram rozlišující genomy S1-S24 dle melt typu



Příloha 3 - Dendogram rozlišující genomy S25-S48 dle melt typu