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 SEKVENAČNÍCH DATECH

CNV DETECTION IN THE SEQUENCING DATA

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. Barbora Pleskačová

AUTHOR

VEDOUCÍ PRÁCE — Ing. Robin Jugas

SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

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

Ústav biomedicínského inženýrství

Studentka: Bc. Barbora Pleskačová ***ID:*** 185955

Ročník: 2 ***Akademický rok:*** 2019/20

NÁZEV TÉMATU:

Detekce CNV v sekvenačních datech

POKYNY PRO VYPRACOVÁNÍ:

1) Proveďte literární rešerši na téma detekce CNV v genomických datech. 2) Otestujte možnosti využití různých metod pro odhalování CNV. 3) Část algoritmu pro detekci CNV implementujte ve vhodném programovacím jazyce. 4) Kompletní softwarové řešení implementujte ve formě balíčku funkcí. 5) Spolehlivost algoritmu otestujte a statisticky vyhodnoťte na umělých i reálných datech, srovnejte s dalšími metodami. 6) Proveďte diskusi získaných výsledků a zhodnoťte účinnost a využitelnost řešení.
Pro splnění semestrálního projektu je nutné vypracování alespoň bodů 1) až 3).

DOPORUČENÁ LITERATURA:

[1] MEDVEDEV, Paul, STANCIU, Monica and BRUDNO, Michael. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods. 2009. Vol. 6, no. 11S, p. S13. DOI 10.1038/nmeth.1374.

[2] ALKAN, Can, COE, Bradley P. and EICHLER, Evan E. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011. Vol. 12, no. 5, p. 363–376. DOI 10.1038/nrg2958.

Termín zadání: 3.2.2020 ***Termín odevzdání:*** 29.5.2020

Vedoucí práce: Ing. Robin Jugas

prof. Ing. Stanislav Hanus, CSc.
předseda rady studijního programu

Abstrakt

Detekci variability počtu kopií v prokaryotických organismech je v současné době věnováno čím dál více pozornosti, a to zejména díky souvislosti CNV s patogenitou a antibiotickou rezistencí bakterií. Algoritmus navržený v této práci využívá k odhalování CNV segmentů detekci extrémů v signálu s hloubkou pokrytí. Pokrytí čtení je běžně získáno mapováním osekvenovaných čtení jednoho jedince, k již známé referenční sekvenci jiného jedince stejného druhu. Dva jedinci se však vždy budou v určitém množství genů lišit, vznikají tak nenamapovaná čtení, která jsou zbytečně zahozena. Tato práce proto předpokládá, že biologická přesnost detekce CNV se dá zvýšit použitím nové reference, která je vytvořena ze stejného setu čtení jako čtení k této referenci mapovaná. Pro ověření tohoto tvrzení je využito sekvenačních čtení jedinců bakterie *Klebsiella pneumoniae*.

Klíčová slova

Variabilita počtu kopií, CNV, strukturní variabilita, prokaryotní genom, sekvenační čtení, pokrytí

Abstract

Copy number variation detection in prokaryotic organisms is currently receiving more and more attention, mainly due to the association of CNV with pathogenicity and antibiotic resistance in bacteria. The algorithm designed in this thesis uses peak detection in sequencing coverage to detect CNV segments. Read coverage is commonly obtained by mapping sequencing reads of one individual to an already known reference of another individual of the same species. However, two individuals will always differ in a certain number of genes, resulting in unmapped reads that are unnecessarily discarded. Therefore, this work assumes that the biological accuracy of CNV detection can be increased by using a new reference that is created from the same set of reads as the reads mapped to this reference. Sequencing reads of Klebsiella pneumoniae individuals are used to verify this assertion.

Keywords

Copy number variation, CNV, structural variation, prokaryotic genome, sequencing reads, coverage

Bibliografická citace:

PLESKAČOVÁ, Barbora. *Detekce CNV v sekvenačních datech.* Brno, 2020. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/126845. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce Robin Jugas.

Prohlášení

„Prohlašuji, že svou závěrečnou práci na téma Detekce CNV v sekvenačních datech 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 autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne 29. května 2020 ……

podpis autora

Poděkování

Děkuji vedoucímu diplomové práce Ing. Robinovi Jugasovi za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

V Brně dne 29. května 2020 .....

podpis autorky

Obsah

Seznam obrázků

Seznam tabulek

ÚVOD

V posledním desetiletí se výrazně změnil pohled na genomickou strukturální variabilitu (SV). Strukturní variabilita, která byla dříve považována spíše za vzácnou událost, je nyní uznávána jako největší zdroj genetické variability mezi jedinci. Ukázalo se, že kromě jednonukleotidového polymorfismu (SNP) má velký vliv na genotypovou a fenotypovou variabilitu organismu také variabilita počtů kopií – copy number variants (CNV). CNV je jev, kdy se určitá část genomu opakuje a počet opakování se liší mezi jedinci v populaci.

Vliv CNV je nejvíce prozkoumán v souvislosti s lidským genomem. Bylo prokázáno mnoho onemocnění, se kterými je variabilita počtu kopií spojována, nebo se na jejich vzniku přímo podílí. Změny počtu kopií se však vyskytují také v řadě dalších organismů, včetně bakterií. Dříve se považovalo přeskupení genu pro bakteriální životaschopnost za škodlivé, v současné době je však poukazováno také na výhody takových přestaveb. Příkladem může být antibiotická rezistence či patogenita bakterií. Schopnost bakterií přizpůsobit se stresovým podmínkám prostředí spojená s obrovským počtem bakterií na planetě podstatně omezuje schopnost kontrolovat nemoci bakteriemi způsobené. Objevení a zkoumání variability počtu kopií v bakteriálním genomu může pomoci právě v boji s bakteriálním onemocněním. V této práci je proto navržený algoritmus detekce CNV aplikován na osekvenované genomy bakterie *Klebsiella pneumoniae*.

Současné počítačové metody detekce CNV, využívající sekvenační data, pracují zejména s namapovanými čteními k referenčnímu genomu, který je běžně získán z veřejné databáze (NCBI). Ze získaného pokrytí genomu se pak dále odvozuje detekcí extrémů informace o výskytu CNV. Genomy jedinců stejného bakteriálního druhu však neobsahují nikdy naprosto totožnou genetickou informaci, proto může při mapování vzniknout mnoho nenamapovaných čtení. Výzvou je proto zjistit, zda čtení, která nebyla namapována k referenci, v sobě neobsahují informaci usnadňující detekci delecí, duplikačních inzercí či duplikací. Za tímto účelem je navržen algoritmus, založený na metrice z – skóre, využívající novou referenční sekvenci sestavenou ze stejného souboru čtení jako čtení, která jsou k této referenci následně mapována. Většina čtení je pak využita a nedochází k jejich zbytečnému zahození. Předpokládá se, že tento přístup může zvýšit biologickou přesnost detekce CNV a poskytnout tak daleko robustnější výsledky.

Praktické části však předchází část teoretická. V úvodu práce je připomenuta definice genomu. Práce se zaměřuje zejména na genom prokaryot, který je v mnoha

ohledech specifický ve srovnání s tím eukaryotickým. Pro genom prokaryot je typická vysoká organizovanost a částečná genetická nestabilita, která umožňuje bakteriím přizpůsobit se extrémnímu prostředí.

Další část práce se zabývá již zmiňovanými strukturálními variabilitami a jejich charakteristikou. Důraz je kladen na variabilitu počtu kopií, která se v posledních letech dostala do oblasti zájmu řady studií.

Následující kapitola pojednává o sekvenaci DNA, tedy stanovení primární struktury DNA. Uvedeny jsou však pouze dvě dnes nevýznamnější metody – Illumina a Oxford Nanopore. I přes vyspělost současných sekvenačních technik, stále není možné sekvenovat celé genomy najednou. Z toho důvodu jsou sekvenována pouze kratší čtení, která jsou následně sestavena na základě jejich překryvů do jedné dlouhé sekvence metodami sestavení genomu, kterými jsem se věnovala ve své bakalářské práci.

V poslední teoretické kapitole jsou rozebrány metody detekce CNV, které jsou rozděleny na metody laboratorní a počítačové.

1 GENOM

Genom je kompletní sada DNA konkrétního organismu, která zahrnuje všechny jeho kódující i nekódující sekvence. Díky genetickému kódu, který každému tripletu bází adenin (A), cytosin (C), thymin (T), guanin (G) přiřazuje příslušnou aminokyselinu, obsahuje DNA veškeré informace potřebné k vybudování celého organismu. [9]
Díky rapidnímu vývoji metod sekvenace DNA, je v současnosti možné sekvenovat celé genomy organismů za poměrně krátký čas (několik hodin až několik dní) a nízkou cenu (od 500 $/ genom). [9]

V roce 2003 byla díky Projektu lidského genomu (HGP) vytvořena velmi kvalitní verze sekvence lidského genomu, která je dostupná ve veřejných databázích. Bylo zjištěno, že lidský genom obsahuje více než 3 miliardy párů bází, které jsou jako kompletní sada sekvencí nukleových kyselin kódované uvnitř 23 párů chromozomů. Odhaduje se, že v lidském genomu je uschováno 20 000 – 25 000 genů. [9]

Sekvenování DNA přispělo mimo jiné k rozmachu molekulární biologie. Genomické studie zdůraznily velmi zvláštní způsob evoluce genomu prokaryot ve srovnání s modelovými eukaryoty. Zajímavé je odhalení evoluční dynamiky prokaryotických genomů, která daným organismům umožňuje být vysoce organizovanými, ale také extrémně plastickými v evoluční časové stupnici. Chromozomy jsou totiž organizovány podle jejich interakcí s buněčnými procesy. To vysvětluje přizpůsobivost prokaryot pro různá, někdy až extrémní prostředí, a mimořádně rychlou replikaci. [8]

1.1 Genom prokaryot

Genetická informace prokaryot je uložena v bezmembránovém nukleoproteinovém komplexu zvaný nukleoid. Nukleoid obsahuje jednu dlouhou molekulu DNA, oba dva její konce jsou typicky kovalentně spojeny a tvoří tak kruhovitý chromozom. Chromozom je několiksetkrát přeložen a stlačen tak, aby se vešel do zhruba jedné pětiny objemu buňky. K vyšší hutnosti a celistvosti nukleoidu přispívá cytosol s vysokým obsahem makromolekul. Další dva klíčové faktory přispívající ke „zhutnění“ nukleotidu jsou negativní supercoiling (nadšroubovicové vinutí) topoizomerázami a kondenzace navázáním proteinů k nukleoidové struktuře. Supercoiling je nezbytný pro přežití buňky, protože podporuje rozvinutí DNA a tím mnoho buněčných mechanismů interagujících s DNA, zejména transkripci. Exprese mnoha genů je tedy ovlivněna změnami úrovně nadšroubovicového vinutí a dále také nukleoidovými proteiny. [14][30]

Nukleoid je umístěn ve středu prokaryotické buňky, RNA polymeráza leží na jejím obvodu a ribozomy se nacházejí v okrajích a interagují s vnitřní membránou. Toto uspořádání má pozitivní vliv na genovou expresi. Bakteriální genomy mají značnou výhodu, že obvykle postrádají introny, což podstatně usnadňuje proces identifikace hranic genu. Také úroveň repetitivní DNA je u prokaryot mnohem nižší než u eukaryot.[4] [30]

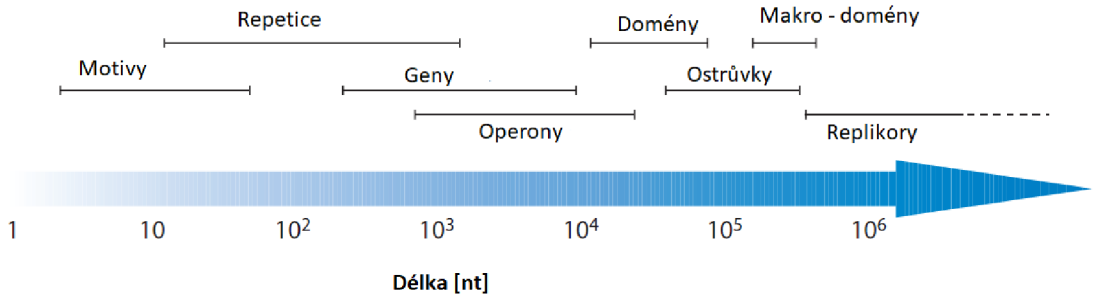
Geny jsou obvykle transkribovány v jednotkách zvaných operony. Příbuzné geny kódující enzymy bývají navíc v bakteriálním chromozomu kolokalizovány. Kromě toho, pořadí těchto genů se řídí řádem odpovídajících enzymových aktivit v metabolických drahách a jsou často kódovány ve stejném operonu, tudíž exprimovány najednou. Protože prokaryotická DNA může interagovat s cytoplazmou, dochází současně s transkripcí také ke translaci, která stejně jako u eukaryot probíhá na ribozomech. [30]

Také replikace má svá specifická pravidla. Například modelový organismus, bakterie Escherichia coli, má jediný replikační počátek (oriC), ve kterém začínají dvě replikační vidlice. Chromozom je replikován v opačných směrech do té doby, dokud vidlice nenarazí na terminační oblast a replikace je tak ukončena. Chromozom je tak rozdělen do dvou polovin, replikorů, které jsou replikované různými vidlicemi. V populaci rostoucích buněk to má za následek gradient počtu kopií, takže sekvence vyskytující se blíže k replikačnímu počátku jsou frekventovanější [34]. Zatímco transkripce formuje chromozomovou strukturu v lokálním měřítku, replikace, díky své vlastní symetrii, tak činí na úrovni replikorů, tj. celého genomu. [14][30]

Kromě chromozomální DNA bakterie často obsahují také plazmidy – malé cirkulární molekuly DNA. Bakterie mohou získat nové plazmidy z jiných bakteriálních buněk (během konjugace) nebo z prostředí. Mohou je také snadno ztratit – například při dělení, kdy plazmid zdědí pouze jedna dceřiná buňka. Každý plazmid má svůj vlastní „počátek replikace“ - úsek DNA, který umožňuje replikaci hostitelskou bakterií. Z tohoto důvodu se plazmidy mohou kopírovat samy o sobě nezávisle na bakteriálním chromozomu, takže v jedné bakteriální buňce může být mnoho kopií plazmidu – i stovky. Plazmidy obsahují jen několik genů, ale na hostitelskou bakterii mají velký vliv. Tyto geny obvykle nejsou nezbytné pro každodenní přežití bakterie, avšak místo toho pomáhají bakterii překonat stresové situace. Například mnoho plazmidů obsahuje geny, po jejichž expresi se stává hostitelská bakterie rezistentní na antibiotikum. Jiné plazmidy obsahují geny, které pomáhají hostiteli strávit neobvyklé látky nebo zneškodnit jiné typy bakterií. Plazmidy nacházejí své využití v molekulární biotechnologii. Slouží jako doručovací nástroj či vektor k zavedení cizí DNA do bakterií. [4]

Většina prokaryot se rozmnožuje nepohlavně a je haploidní, což znamená, že je přítomna pouze jedna kopie každého genu. Díky tomu je relativně snadné generovat mutace v laboratoři a studovat výsledné fenotypy. Nedávné studie však naznačily, že některá prokaryota obsahují až čtyři lineární nebo kruhové chromozomy. [14]

Uspořádání genomu můžeme chápat ve více měřítkách, od malých motivů až po velké chromozomální oblasti (Obr. 1.1).



Obr. 1.1 Různá měřítka organizace bakteriálního genomu [30]

Bakteriální genomy jsou vysoce proměnlivé z hlediska repertoáru genů, ale pozoruhodně stabilní, co se týče organizace chromozomu. Míra přeuspořádání v E. Coli se blíží rychlosti genomové mutace $\sim 10^{-3}$–10^{-4} změn / (generace genomu). U prokaryot je jedním z hlavních mechanismů přispívajících k diverzifikaci genomů horizontální přenos genů (HGT). Genomy mutují, mění se co do velikosti a přeskupují se. Většina spontánních přeuspořádání vede k nižší kondici celé buňky, např. zpomalením růstu. Některé změny však mohou přispívat k evoluci, divergenci a patogenitě bakterií. Velká přeuspořádání jsou však v rozporu s přirozeným výběrem, protože jsou zvláště škodlivá pro organizaci genomu. Konflikt mezi dynamikou genomu a organizací chromozomu je přirozeným výběrem formován a závisí na ekologických a buněčných procesech, které mohou být díky komparativní genomice odhaleny. Komparativní genomika spolu s experimentálními metodami patří ke dvěma hlavním přístupům analýzy stability genomu. Oba přístupy poskytují zásadně odlišné informace. Experimentální práce umožňuje stanovit míru přeuspořádání, zatímco komparativní genomika zjišťuje, jak se přeuspořádání nahromadilo v průběhu evoluční historie. Kvantifikace stability genomu umožňuje testování patogenity buňky nebo hustoty repetic. [30]

Variabilita genetické informace se objevuje ve více úrovních, od velkých alternací v karyotypu až po změny v jediném nukleotidu. V následující kapitole se zaměřím na strukturální variability, do kterých spadá také variabilita počtu kopií (CNV).

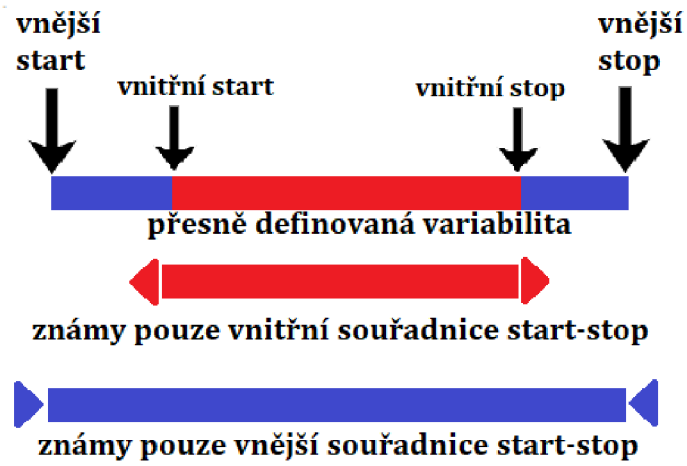
2 STRUKTURNÍ VARIABILITA

Strukturní variabilita (SV) je obecně definována jako oblast DNA, která vykazuje změnu v počtu kopií (delece, inzerce a duplikace), změnu orientace (inverze) nebo změnu chromozomálního umístění mezi jednotlivci (translokace) [13].

Strukturní variabilita může být balancovaná, bez ztráty genetické informace (inverze, translokace), nebo nebalancovaná, pokud je část genetické informace získána, ztracena či duplikována (inzerce, delece, duplikace). [13]

SV se tradičně týká genomických změn, které jsou větší než 1 kbp na délku. S metodami sekvenování lidských genomů, které se nyní stávají rutinní, se operační spektrum strukturních variabilit a variabilit počtu kopií rozšířilo o mnohem menší události (> 50 bp). [1]

Reprezentace strukturních variací může být složitá. Současné technologie zřídka poskytují přesné rozlišení bodů zlomů (počátek a konec SV) na úrovni párů bází. Existuje však základní soubor dat, který zachycuje všechny potřebné informace o variabilitě, včetně míry nejistoty v místě přerušení. Souřadnice start-stop se používají k definování událostí, u nichž jsou body zlomu známy na úrovni páru bází. Vnitřní souřadnice start-stop slouží k definování oblastí, o nichž je známo, že jsou ovlivněny variabilitou, ale nedefinují skutečné body zlomu. Body zlomu leží mimo definovanou oblast. Vnější souřadnice start-stop slouží k definování absolutní vnější hranice dané variability, ale nedefinují skutečné body zlomu. Body zlomu leží uvnitř definované oblasti (Obr. 2.1). [26]



Obr. 2.1 Příklad zobrazení strukturální variability

V roce 1991 bylo onemocnění Charcot-Marie Tooth prvním autosomálně dominantním onemocněním spojeným s účinkem dávky genů v důsledku zděděného

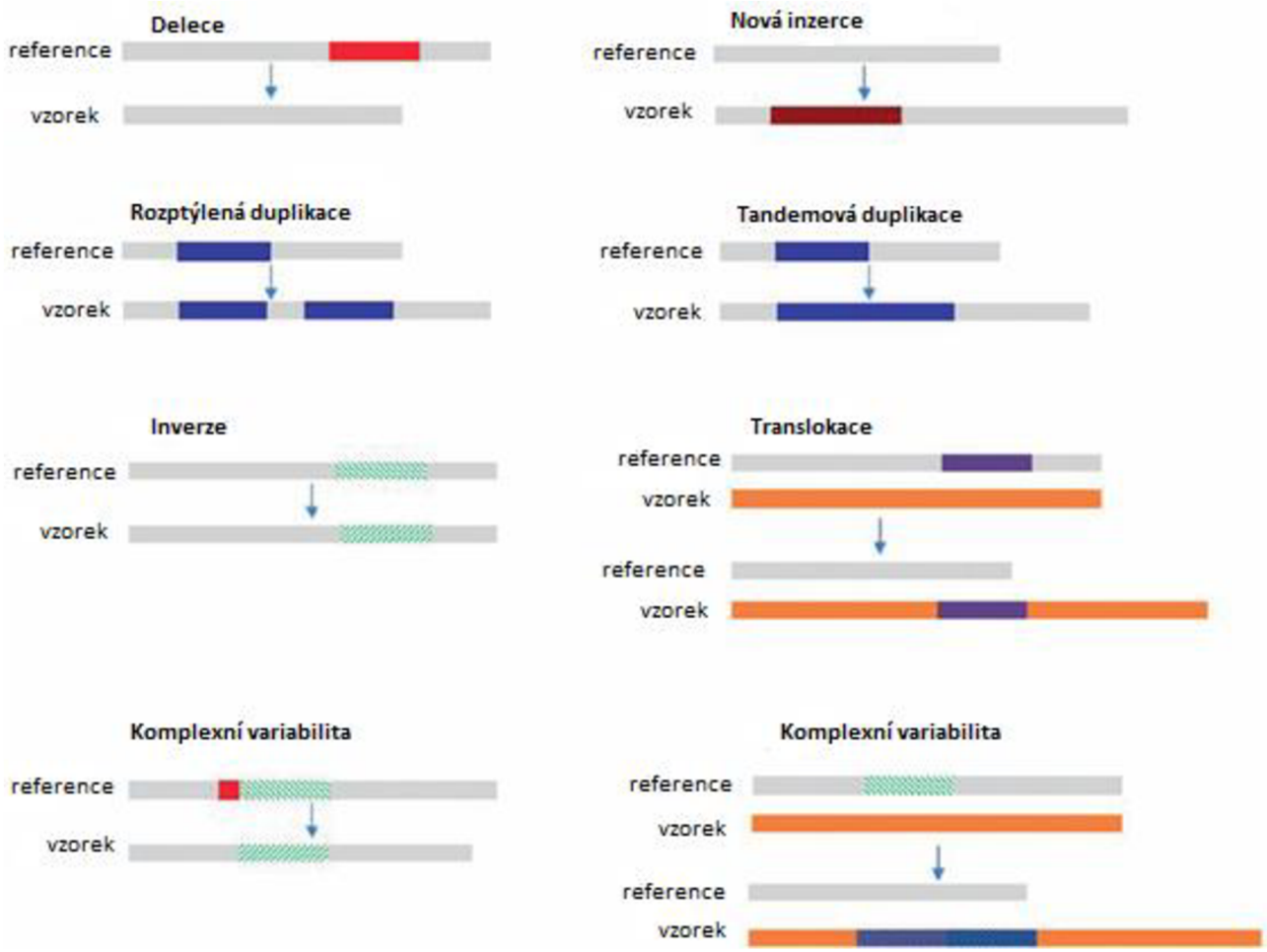
přeuspořádání DNA. Většina případů tohoto onemocnění je spojena s tandemovou duplikací postihující gen PMP22. Fenotyp onemocnění je výsledkem tří kopií normálního genu. Jediná kopie genu PMP22, má za následek klinicky odlišnou dědičnou neuropatii s tendencí k tlakovým obrnám (HNPP). [26]

Nedávné studie však ukazují, že přeuspořádání genomu může mít hluboký dopad také na prokaryotické genomy, což vede ke změně buněčného fenotypu. Sekvenování velkého počtu prokaryotických genomů a porovnání sekvencí blízce příbuzných druhů odhalilo, že mikrobiální genomy jsou často přeskupovány, což vede k diskontinuitě DNA kmenů stejného druhu. Množství studií zaměřených na variabilitu prokaryotických genomů se soustředilo převážně na jednonukleotidové variace a malé inzerce a delece. Nyní se stále více připouští, že kromě jednonukleotidových variací nejsou u bakteriálních genomů vzácné ani jiné typy variabilit, které představují mnohem větší genomické přestavby. Mezi různými typy genetických variabilit zůstaly strukturální variace nejobtížněji identifikovatelné a interpretovatelné. Studie strukturních variabilit však poskytují mnohem hlubší vhled do organizace bakteriálních genomů. Strukturální variace se mohou objevit v důsledku různých buněčných mechanismů, jako je DNA rekombinace, replikace a oprava DNA. Některé nestability jsou řízené, zatímco jiné jsou náhodné. Následkem těchto mechanismů mohou být změněny počty kopií genů, vytvořeny nové geny, změněna exprese genů a mnoho dalších funkčních důsledků. Ukázalo se například, že rozsáhlé přestavby v úzce příbuzných kmenech významně přispívají k bakteriálnímu vývoji či divergenci. Nestabilita genomu je také používána patogenními bakteriemi ke snadnější infekci hostitele, aniž by byla napadena jeho imunitním systémem. [10][27] V ideálním případě objevení a genotypizace strukturních variabilit vyžaduje přesnou předpověď tří vlastností: kopií, obsahu a struktury DNA. V praxi tento cíl zůstal nedosažitelný, protože SV mají tendenci vyskytovat se v repetitivních úsecích DNA, což ztěžuje jejich charakterizaci. [10][27]

Výzvou je nyní odhalit plný rozsah strukturních variabilit, aby bylo možné je běžně genotypizovat a porozumět jejich účinkům na lidská onemocnění, zjistit komplexní rysy i jejich vývoj.[1]

V případě prokaryot můžou být získané informace o SV využity v oborech jako je syntetická biologie a genomové inženýrství, kde by cílené SV mohly sloužit k přesné úpravě bakteriálního genomu. S rostoucím zájmem o inženýrské bakteriální genomy roste i potřeba vyvinout účinné nástroje pro jejich úspěšnou manipulaci. Příkladem takových nástrojů může být systém Cre-lox, který byl použit pro vytvoření delecí v E.coli a velkých inverzí v Lactococcus lactis. Pro genomové inženýrství prakticky všech bakterií byla nedávno představena technologie GETR (Genome Editing via

Targetrons and Recombinases). Technologie byla účinně použita k zavedení inzercí, delecí, inverzí a translokací v E.coli, S.aureus a Bacillus subtilis. Pro účinné genomové inženýrství je dále nutné indukovat DSB (dvouvláknové zlomy) v DNA a zahájit tak rekombinační proces. To bylo dlouhou dobu nemožné kvůli nedostatku prostředků pro cílení DSB na konkrétní místa. Umožnilo to až zavedení technologií pro syntézu a sestavení velkých fragmentů DNA a nástrojů, které jsou schopné přesně upravovat genomické oblasti. [27]



Obr. 2.2 Strukturní variability [13]

2.1 Variabilita počtu kopií

Variabilita počtu kopií, copy number variants (CNV), je převládající formou genetické variability, která vede k různému počtu kopií segmentů DNA v buňce. CNV zahrnují delece a duplikace genomických oblastí. [14]

Nejlépe jsou CNV a jejich fenotypový projev prostudovány u lidského genomu. CNV mohou být buď zděděny z rodičovského genomu, nebo mohou být způsobeny mutací, poté je označujeme jako *de novo* CNV. CNV, které dosahují populační frekvence větší než 1 %, jsou definovány jako běžné CNV nebo jako polymorfismus počtu kopií (CNP). Naopak CNV, které se vyskytují u méně než 1 % populace, se označují jako vzácné CNV. [37]

Odhaduje se, že přibližně 12 % (přibližně 360 Mbp) genomu v lidských populacích podléhá změně počtu kopií. Předpokládá se, že tyto všudypřítomné CNV v genomu mají stejný příspěvek ke genetické i fenotypové variabilitě u lidí jako jednonukleotidové polymorfismy (SNP), které byly dlouho považovány za nejhojnější genetické variability u lidí. Přibližně polovina zjištěných CNV se překrývá s oblastmi kódujícími proteiny. Tyto zisky a ztráty genových kopií můžou přímo ovlivnit množství genů v oblastech CNV, což také může vést ke změně úrovně genové exprese. [39]

První objev CNV u lidí lze vysledovat až do počátku 20. století, kdy byl u člověka rozpoznán pouze chromozom X, bez přítomnosti druhého pohlavního chromozomu. Trisomie 21. chromosomu představuje první objev lidského autosomálního CNV. Identifikace těchto CNV se opírala o relativně necitlivé techniky, zahrnující například optickou mikroskopii, která umožňovala detekci variabilit jako jsou aneuploidie, nadbytečné chromozomy, přeskupení a změny velikosti chromozomů. [37]

Podobně jako SNP nemusí mít CNV nezbytně negativní dopad na lidské zdraví. Mnoho genů, u kterých bylo zjištěno CNV, jsou zapojeny do enviromentálních reakcí, jako je smyslové snímání (čichové receptory) a imunita. [26]

U velkého počtu CNV však může mít variabilita v počtu kopií souvislost s nemocemi a fenotypy, jako jsou rakovina a neuropsychiatrické poruchy, či se jich mohou přímo účastnit [39]. Nalezení a genotypizace strukturálních variabilit byly pro pochopení těchto asociací onemocnění klíčové [26]. S CNV je spojováno spoustu nemocí, jako je Williams – Beuernův syndrom, DiGeorgeův syndrom, spinální svalová atrofie, Duchennova svalová dystrofie, ale i běžné lidské choroby, včetně obezity, schizofrenie a autismu. [37] Počet nových genomických poruch navíc roste. Vliv CNV byl dokázán například také při poruchách učení [26].

CNV jsou zdrojem významné genetické variace také u bakterií. Ačkoliv zkoumání role CNV u prakaryot nebylo věnováno tolik prostoru jako u eukaryot, bakterie rovněž vykazují značné rozdíly v počtu kopií. Mimořádné náklady na udržení nadbytečného genu přicházejí s odměnou selektivní výhody za určitých podmínek daných prostředím [5]. V současné době se vkládá velké úsilí zejména do zkoumání

variability počtu kopií a její role v antibiotické rezistenci v bakteriálních genomech. [17]

2.1.1 Delece

Delece znamenají ztrátu genomického segmentu a mohly by vést k inaktivaci genu, nebo ke ztrátě jedné nebo více funkčních domén, či ke změně funkce celého genu. Intergenové delece by mohly potenciálně ovlivnit regulační oblasti, a tím ovlivnit expresi sousedních genů. Fenotypové účinky delecí závisí na velikosti a umístění deletovaných chromozomálních segmentů v genomu. Větší delece pravděpodobně zahrnují více genů, což může vést k drasticky pozměněným fenotypům. Delece zahrnující ztrátu esenciálních genů mohou významně omezit životaschopnost buňky. [27]

2.1.2 Duplikace

Duplikace se vyznačují přítomností dvou nebo více kopií genomické oblasti, nebo genomického segmentu. Duplikované oblasti ležící vedle sebe jsou označovány jako tandemové duplikace, duplikované segmenty nacházející se na jiném genomickém místě jsou označovány jako inzerční duplikace (rozptýlené). Duplikace obecně vede k získání kopie segmentu DNA nesoucího nějaké informace. Funkční důsledek duplikace se může lišit v závislosti na informačním obsahu duplikovaného genomického segmentu a na kontextu duplikované oblasti. [27]

3 SEKVENACE DNA

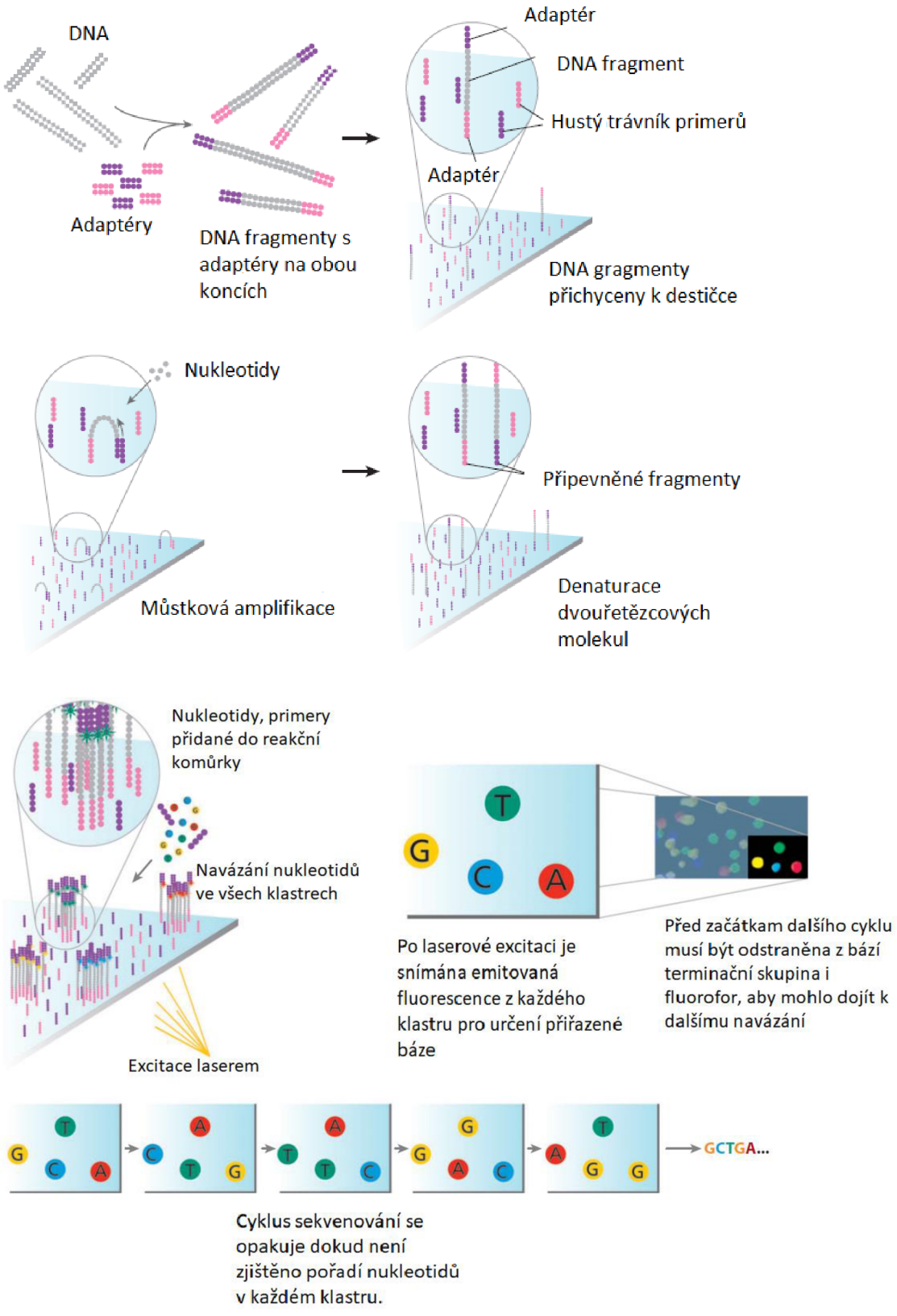
Metody sekvenování DNA patří v současné době k nejrozšířenějším způsobům zkoumání biologického materiálu. Bez sekvenování bychom neznali samotný obsah genetické informace v genomu či sekvenovaném úseku. Díky sekvenování máme přístup ke genetické informaci daného jedince, kterou můžeme dále studovat a analyzovat. Bylo vyvinuto několik klíčových metod. Mezi prvotní metody patří dodnes příležitostně používaná Sangerova sekvenace. V práci však popíšu jen dvě dnes nejvýznamnější – sekvenace platformy Illumina a Oxford Nanopore.

3.1 Illumina

Komerčně velice úspěšná technologie od společnosti Illumina patří do skupiny metod druhé generace (Next Generation Sequencing, NGS). Illumina, běžně označovaná také jako „Solexa“, je schopná najednou osekvenovat až 900 miliard párů bází, a hlavně díky tomu se řadí mezi nejpoužívanější metody sekvenace. Nedostatek této technologie tkví v kratších sekvencích, jejichž velikost se pohybuje kolem stovky bází. Sekvenace probíhá v reakční komůrce, která má osm nezávislých pruhů, takže během jednoho běhu přístroje může být paralelně sekvenováno až osm nezávislých knihoven DNA fragmentů. [18][32]

Knihovna může být konstruována jakýmkoliv způsobem, který vede ke směsi fragmentů, na jejichž oba konce jsou ligovány adaptéry (krátké oligonukleotidy). [18] Na začátku procesu jsou nastříhané úseky DNA uchyceny s využitím adaptérů na malou amplifikační destičku v reakční komůrce (flow cell). Každá molekula je pak opakovaně replikována pomocí můstkové PCR a na destičce vznikají dvouvláknové můstky. Následná denaturace vytvoří mozaiku milionů klastrů. Každý klastr obsahuje přibližně milion navzájem stejných, jednovláknových kopií původního fragmentu. To je dostatečné pro získání požadované intenzity signálu, který vypovídá o začleněné bázi. [18][22] [32]

V následující fázi sekvenace jsou pak do rostoucího řetězce pomocí DNA polymerázy řazeny nukleotidy označeny fluorescenční barvou, která je pro daný nukleotid specifická. Nukleotidy jsou také modifikovány do podoby reverzibilních terminátorů, takže v každém cyklu může být přiřazena pouze jedna báze s příslušným značením. Narozdíl od Sangerovy sekvenace je však tato blokace vratná. [18]



Obr. 3.1 Schéma metody Illumina [22]

Nově přidaná báze je po laserové excitaci přečtena vysoce citlivou kamerou, a následně dochází k enzymatickému odstranění fluorescenčního značení i blokující části molekuly. Další kolo reakce muže nastat přidáním nové báze. V každém kole syntézy řetězce DNA kamera snímá signál z celé destičky a na základě rozdílné fluorescence rozpoznává, jaký nukleotid byl navázán v každé skupině. Záznam je pak počítačově analyzován a dle měnící se fluorescence v každém klastru je sestavena příslušná sekvence DNA. [18][22]

Illumina se chlubí velmi vysokou přesností čtení DNA (99 až 99,9 %). Nejčastěji dochází ke špatnému přečtení báze, tedy záměně jednoho nukleotidu. V současné době je běžná délka čtení 36 bp. Delší čtení jsou možná, ale mohou způsobit vyšší chybovost. Samotná délka čtení je omezena několika faktory, které způsobují rozpad signálu, například neúplné štěpení fluorescenčních značek nebo terminačních skupin. [18][32]

3.2 Oxford Nanopore

V posledních několika letech se na scéně objevily také metody třetí generace. Od metod druhé generace se liší tím, že DNA templát není před samotným procesem sekvenace replikován a sekvenace se provádí z původní molekuly. Mezi komerčně nejvýznamnější zástupce metod třetí generace patří technologie od společností Pacific Biosciences a Oxford Nanopore. [18]

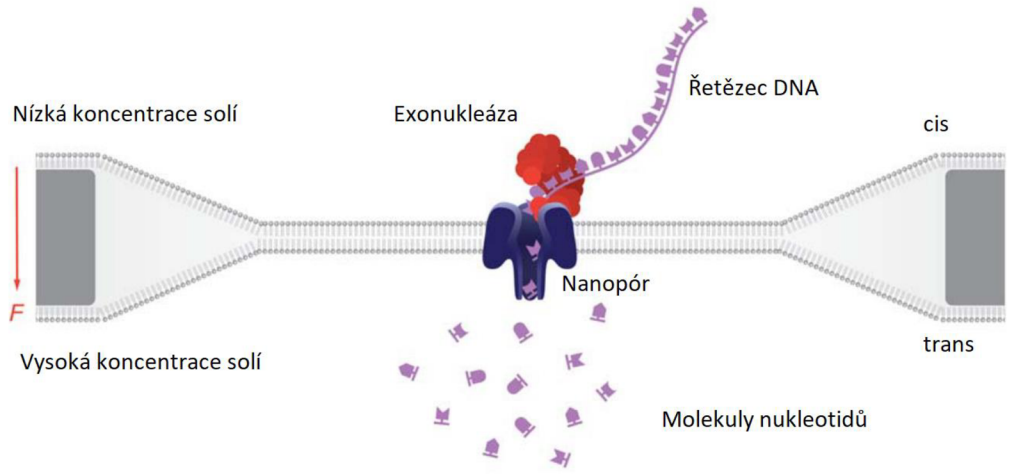
Při použití metody Oxford Nanopore je jednořetězcová molekula DNA elektroforeticky translokována mikroskopickým pórem (> 1,2 nm) na syntetické membráně. Nanopóry můžeme rozdělit na biologické a syntetické nanopóry solid-state. Biologické nanopóry mají velmi dobře prozkoumanou strukturu a můžou být cílenou mutagenezí modifikovány [19]. Biologický nanopór je vytvořený z upraveného α-hemolysinového póru, který má připojenou exonukleázu na extracelulární ploše póru. K vnitřnímu povrchu póru je kovalentně připojen syntetický cyklodextrinový senzor. Tento systém je umístěn do syntetické lipidové dvojvrstvy. [33]

Oproti tomu solid-state nanopóry jsou vyráběné litografickými metodami v pevných substrátech. Nejvíce jsou používány membrány na bázi křemíku (SiN, Si_2N, Si_3N_4, SiO_2), skla či grafenu. Tyto nanopóry jsou na rozdíl od proteinových biopórů mnohem více mechanicky, chemicky a teplotně stabilní. Lze si libovolně vybrat velikost i tvar póru, nevýhodu představuje velmi složitá reprodukovatelnost. Póry navíc disponují větší kapacitancí a šumem. [19]

Změnou koncentrace solí na obou stranách membrány (vnější-cis, vnitřní-trans) je na membránu přivedeno napětí. V případě, že je na cis stranu nanopóru nanesen DNA řetězec, exonukleáza štěpí jednotlivé nukleotidy, které prochází díky vzniklému iontovému proudu přes nanopór. [33]

Iontová vodivost je v nejužší oblasti nanopóru zvláště citlivá na přítomnost hmoty nukleotidové báze i na její přidružené elektrické pole, proto je právě toto místo použito jako snímací oblast [11]. Každá DNA báze při průchodu specificky narušuje iontový proud, což je vyhodnocováno citlivými snímači přístroje. V daný okamžik snímače zjistí míru narušení iontového proudu a na základě toho určí, jaká báze právě membránou prochází. [18][33]

Avšak i to nejmenší napětí posouvá každou bázi přes snímací oblast za méně než ~ 1–10 ms. Identifikace bází však vyžaduje vyšší poměr signál-šum, což není možné dosáhnout za více než 10 ms. Dosažení vyššího poměru signálu k šumu vyžaduje, aby každá báze spočívala v oblasti snímání nanopóru po dobu více jak 100 ms. Tento problém se řeší enzymaticky s využitím specifických DNA polymeráz, které kontrolují translokaci DNA přes nanopór. Použitý enzym má tak vliv na přesnost celého sekvenování, a právě v této fázi sekvenace nastává i v současné době nejvíce chyb. Aktivita enzymu je totiž stochastická, proto jsou intervaly posunu DNA variabilní. Z toho důvodu může dojít k velmi krátkému pozdržení báze v oblasti snímání a daná báze tak není identifikována a dochází k deleci ve vznikajícím řetězci. [11]



Obr. 3.2 Oxford Nanopore – membrána s nanopórem [33]

Metoda Oxford Nanopore však dokáže sekvenovat až 150 tisíc bází v rámci jedné analyzované molekuly DNA a vytváří tak velmi dlouhé úseky osekvenované DNA, což má své výhody hlavně při sekvenování celých genomů [11]. Sekvenování s využitím nanopóru nepotřebuje značení drahými fluorescenčními reagenty ani

nákladné CCD kamery pro snímání z optických čipů, což dělá tuto metodu cenově velmi přívětivou [31]. Určitou nevýhodou je poměrně vysoká chybovost, která se pohybuje okolo 10 až 15 % [18]. V současné době lze však sekvenovat taky dvouvláknovou DNA s využitím enzymu helikázy, která dvouvlákno rozplétá – 2D čtení, přesnost sekvenace je tak zvýšena. [11]

Další výhodu sekvenátorů Oxford Nanopore představuje rozměr přístroje. Společnost poskytuje malá kapesní zařízení (MinION), která se mohou přes USB připojit k počítači po celém světě a být využita k sekvenaci. V roce 2015, v západní Africe byl tento přístroj použit například pro sekvenování viru Ebola přímo v terénu po vypuknutí epidemie krvácivé horečky způsobené tímto virem. [18]

4 METODY DETEKCE CNV

Problematika detekce variability počtu kopií (CNV) segmentů DNA je pro tuto práci stěžejní, proto jí věnuji celou kapitolu.

V roce 2003 bylo dosaženo detekce CNV v celém genomu pomocí přesnějších přístupů založených na srovnávací genomové hybridizaci (array-CGH) či jednonukleotidovém polymorfismu (SNP). Tyto přístupy však trpěly několika nevýhodami, včetně hybridizačního šumu, omezeného pokrytí genomu, nízkého rozlišení a obtížnosti při detekci nových a vzácných mutací. [39]

V posledních letech se však rozvinuly techniky sekvenování nové generace (NGS), díky kterým lze analyzovat sekvenační data s vysokým rozlišením [3]. Silná poptávka po analýzách CNV na bázi NGS podpořila vývoj četných výpočetních metod a nástrojů pro detekci CNV. Od roku 2005 několik komerčních platforem, včetně platforem 454 Life Sciences, Illumina, Inc. a Life Technologies, bezprecedentně zvýšilo výkonnost generování čtení a přesnost volání bází, což umožnilo sekvenaci genomů za mnohem nižší cenu a rychlejší dobu. [39]

Většina současných algoritmů byla testována na eukaryotických (konkrétně lidských) genomech, avšak mechanismus tvorby strukturních variabilit v eukaryotech i prokaryotech se zdá obdobný, proto mohou být nástroje a algoritmy v podstatě aplikovány i na studium SV v prokaryotech s určitými omezeními [27]. Je totiž potřeba brát v úvahu, že prokaryota jsou organismy neurčité ploidie ve srovnání s lidskými diploidními genomy. Výklad v počtu kopií se proto může různit. U lidských genomů by počet kopií 1,5 znamenal, že daný segment měl dvě kopie na jednom chromozomu, ale jen jednu kopii na druhém. U bakterií by se interpretace lišila druh od druhu v závislosti na počtu sad chromozomů. Bakterie se totiž často nachází v různé fázi dělícího cyklu, a jsou navíc schopny replikovat chromozomy daleko rychleji, než je samotný proces dělení, což se odráží v sekvenačních datech DNA. Výsledný počet kopií je ovlivněn také růstovými mechanismy bakterie, například při procesu replikace vznikajícím gradientem počtu kopií v závislosti na vzdálenost od replikačního počátku. Nejen z těchto důvodů se v posledních letech začaly vyvíjet nástroje přímo pro detekci CNV v prokaryotických organismech. [5]

V následující podkapitole je uveden obecný přehled metod pro odhalování CNV.

4.1 Laboratorní metody

4.1.1 Mikročipy

Mikročipy byly hnacím motorem experimentálního objevování a genotypizace CNV. Tyto metody jsou reprezentovány komparativní genomovou hybridizací na čipu (array-CGH) a SNP mikročipy. Obě technologie založené na hybridizaci odvozují zisky nebo ztráty počtu kopií ve srovnání s referenčním vzorkem nebo populací, ale liší se v detailech a aplikaci molekulárních testů. Mikročipy nabízejí výraznou výhodu, pokud jde o výkon a náklady. Rozsáhlá aplikace přístupů založených na hybridizaci u tisíců normálních i nemocných vzorků poskytla prvotní pohled na oblast větších CNV. [2]

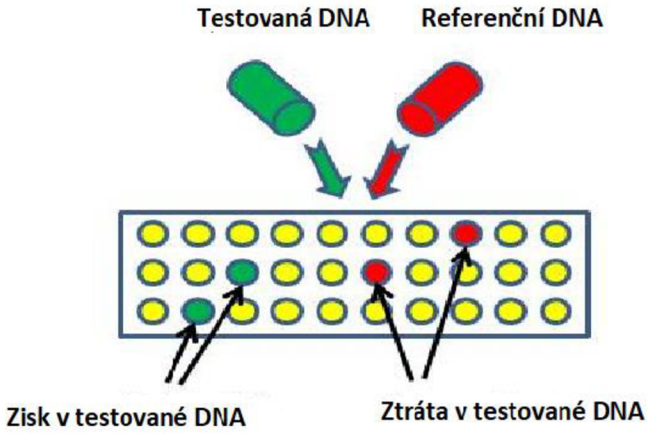
4.2 Komparativní genomová hybridizace na čipu

CGH platformy na čipu jsou založeny na principu komparativní hybridizace dvou fluorochromem značených vzorků (testovací a referenční) se souborem hybridizačních standardů (typicky dlouhé oligonukleotidy nebo bakteriální umělé chromozomové klony BAC). Každý oliogonukleotid má své přesné umístění v rámci mřížky mikročipu a může představovat určitý gen či jeho úsek. Na základě hodnot intenzit testovací a referenční DNA je sestaven CGH profil v podobě křivky vyjadřující poměr intenzit obou fluorochromů v průběhu každého chromozomu. Poměr signálů se pak použije jako zastoupení pro počet kopií. Důležitým hlediskem je vliv referenčního vzorku na profil počtu kopií. Například když je vyšetřen pouze jeden vzorek, ztráta v referenčním vzorku je nerozeznatelná od zisku v testovaném vzorku. Z tohoto důvodu je dobře určená reference klíčem ke správné interpretaci array CGH dat. [2] [24]

V současné době jsou Roche NimbleGen a Agilent Technologies hlavními dodavateli genomových CGH platforem na čipu a rutinně produkují pole s až 2,1 a 1 milionem oligonukleotidů (50–75 merů) na mikročip. Detekce CNV typicky vyžaduje signál z alespoň 3 až 10 po sobě jdoucích průzkumů, v důsledku toho mohou SNP a CGH mikročipy rutinně detekovat kdekoli od desítek do několika stovek událostí na genomu v závislosti na použité platformě. Dvě studie nedávno použily pole s ultra-vysokým rozlišením (sondy 24M až 42M) pro detekci SV založené na array CGH ve vzorcích jednotlivců z HapMapu (projekt, jehož cílem je identifikovat a katalogizovat odlišnosti v jednotlivých pozicích v lidské populaci). Ačkoli takováto pole s vysokou hustotou nejsou pro velký počet vzorků praktická (bylo použito 30 a 40 vzorků v těchto studiích), tyto přístupy umožnily objevení CNV až na úrovni 500 bp, s místy

zlomů dostatečně přesnými, aby umožnily identifikaci sekvenčních motivů v podmnožině variabilit. [1]

Přístup komparativní genomové hybridizace na čipu má omezenou schopnost rozlišovat rozdíly v počtu kopií, zejména proto, že se zvyšuje počet duplikovaných genů a rozdíl v počtu kopií vzhledem k referenčnímu genomu se stává mizivě malý. [1]



Obr. 4.1 Schéma aCHG [24]

4.2.1 SNP mikročipy

Čipy na bázi jednonukleotidového polymorfismu (single-nucleotide polymorphism) jsou také založeny na hybridizaci, s několika klíčovými rozdíly od technologií CGH. Za prvé, hybridizace se provádí s jedinou testovací DNA na mikročip a log-transformované poměry se vytvoří shlukováním intenzit měřených z každé sondy napříč mnoha vzorky. Za druhé, platformy SNP využívají sondy, které jsou specificky navržené tak, aby byly schopné rozlišit jednonukleotidové polymorfismy mezi sekvencemi DNA, a to buď metodami prodloužení o jednu bázi – Single-base-extension (Illumina) nebo diferenciální hybridizací (Affymetrix)[1]. Ačkoli tato pole byla původně navržena pro jednoduché genotypizace SNP, stále více se „těží“ pro analýzu CNV opětovnou analýzou informací o intenzitě sondy [37]. Jednou z klíčových nevýhod je to, že na jednu sondu mají SNP mikročipy tendenci nabízet nižší poměr signál-šum než nejlepší platformy CGH na čipu. Klíčovou výhodou mikročipů SNP je však použití SNP alel-specifických sond pro zvýšení citlivosti na CNV, rozlišení alel a identifikaci oblastí uniparentální disomie. [1]

4.2.2 Single-molecule analýza

Přístupy microarray nemohou identifikovat balancované strukturální variability, nebo v případě duplikace určit umístění duplikované sekvence. Pochopení struktury a umístění tradičně vyžaduje vizualizaci na úrovni jedné molekuly. První pohledy na běžné a vzácné strukturní variability genomu poskytly přístupy, jako je fluorescenční in situ hybridizace (FISH), FISH provedená na vlákně a spektrální karyotypizace. Nízká výkonnost a nízké rozlišení však omezují jejich použití na více jednotlivců a na zvláště velké strukturální rozdíly (~ 500 kb až 5 Mb). Pro zlepšení rozlišení a škálovatelnosti vyvíjí mnoho firem metody pro přímou vizualizaci struktury v "natažených" fragmentech DNA ve velkém měřítku. Například optické mapování, technika původně vyvinutá pro analýzu kvasinkových genomů, byla nedávno použita na analýzu SV lidského genomu. Tato výkonná technologie umožňuje jemnou strukturální analýzu genomů, detekci inverzí a translokací, jakož i změny počtu kopií a jejich umístění. Přestože je technika schopna detekovat nové inzerce, je omezena svou závislostí na referenčním genomu a v současné době má velmi omezenou výrobní kapacitu. [1]

Metoda DNA barcodingu je slibnou alternativou, která jednoho dne umožní vysoce výkonnou detekci balancovaných strukturálních variabilit. Tyto metody zahrnují skenování fluorescenčně značené molekuly DNA a použití SNP-specifického značení natažené DNA pro rozlišení haplotypů.[1]

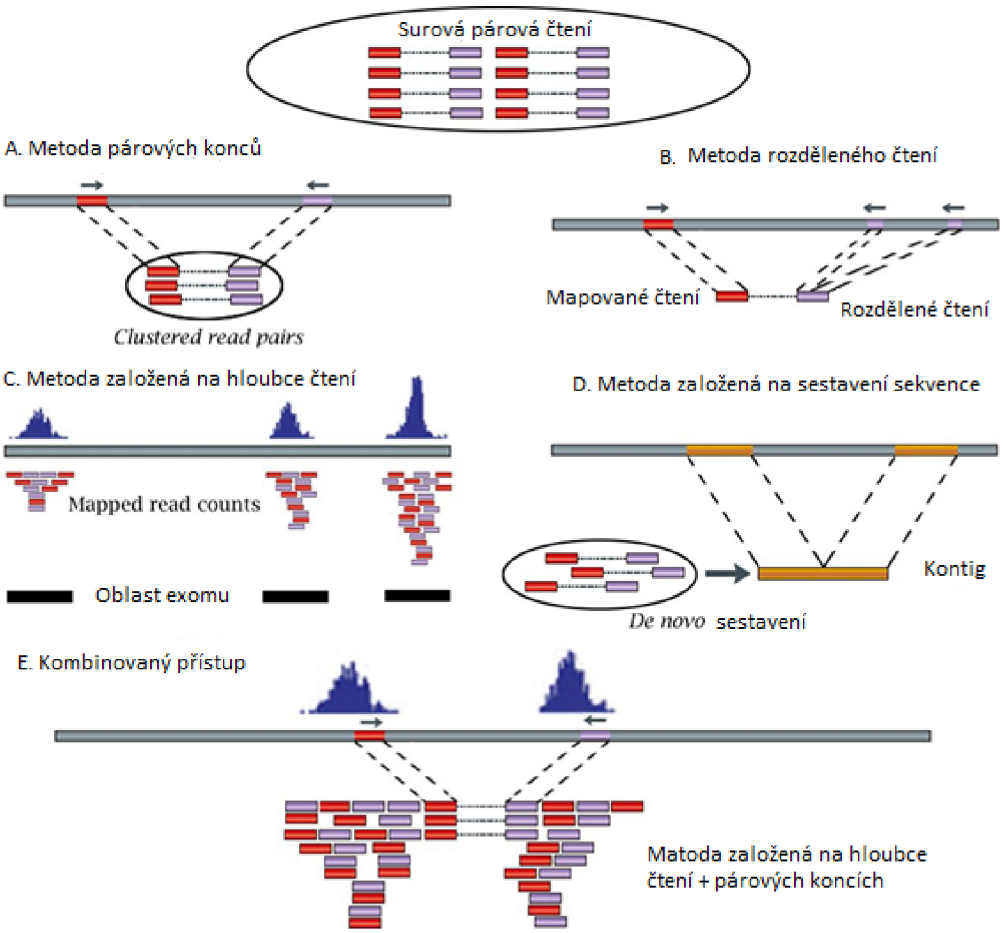
Podobně i absolutní odhady počtu kopií provedené amplifikací jednotlivých molekul pomocí emulzní PCR nebo single – molecule sekvenace lidských genomů nabízejí obrovský potenciál pochopit strukturální změny na buněčné úrovni.[1]

4.3 Počítačové metody

Příchod technologií sekvenace nové generace (NGS) slibuje revoluci ve studiích strukturálních variabilit, a nakonec nahradí mikročipy jako platformu pro jejich objevování a genotypizaci. Přístupy NGS však představují značné výpočetní a bioinformatické výzvy.[1]

Takovou výzvou může být například nejednoznačnost v mapování čtení. Pro zlepšení tohoto zkreslení je třeba delších čtení a insertů ke zvýšení specificity v mapování čtení. Odhaduje se však, že zhruba 1,5 % lidského genomu nelze jedinečně pokrýt ani při délce čtení 1 kbp. Dalším problémem je sekvenační pokrytí, definované jako průměrný počet toho, kolikrát je každý pár bází v genomu reprezentován v zarovnaném čtení. Sekvenční pokrytí je důležitým faktorem pro dosažení vysoké citlivosti a specificity detekce SV. [1]

Mezi největší výhodu NGS přístupu patří to, že je možné objevit celou řadu tříd variabilit pomocí jediného sekvenačního experimentu. Mezi další plusy patří také vyšší pokrytí a rozlišení, přesnější odhad počtu kopií, přesnější detekce míst zlomů oproti hybridizačním metodám a vyšší schopnost identifikovat nové CNV [39]. Kromě toho jsou sekvenační data do značné míry objektivní a představují potenciál pro pochopení kompletního spektra genetické variability. Analýza celého genomu je možná bez apriorních informací, specificita a lineární dynamický rozsah odezvy dat NGS nabízí mnoho výhod pro odhad počtu kopií. Metoda založená na hloubce čtení může přesně odhadnut absolutní počet kopií duplikovaných oblastí lidského genomu. Popis absolutních počtů kopií a schopnost rozlišovat mezi paralogními kopiemi duplikovaných genových rodin jsou nezbytné k lepší genotypizaci těchto dynamických oblastí genomu, což je zase nezbytné pro pochopení fenotypového účinku duplikací a jejich evoluční význam. [1]



Obr. 4.2 Základní výpočetní metody pro detekci CNV [1]

Existují čtyři obecné typy strategií (Obr. 2.2), z nichž všechny se zaměřují na mapování sekvenačních čtení k referenčnímu genomu a následně identifikují

nesouhlasné charakteristické rysy nebo vzory, které jsou příznačné pro různé třídy strukturních variabilit. [20]

4.3.1 Metoda využívající párová čtení

Jako pair-end čtení nazýváme dvě čtení osekvenovaná od začátku a konce stejné molekuly, oblast mezi nimi nazýváme inzert. Metody párových čtení hodnotí rozsah a orientaci čtení se spárovanými konci (Paired-end mapping, PEM) a shlukují ta pair-end čtení, ve kterých vzdálenost (velikost inzertu) a/nebo orientace namapovaných párů čtení je v rozporu s referenčním genomem. V zásadě lze detekovat většinu tříd variací. Párová čtení, která jsou mapována příliš daleko od sebe, určují delece, a ty, která jsou nalezena příliš blízko u sebe, poukazují na inzerce. Neshody v orientaci mohou znamenat inverze a specifickou třídu tandemových duplikací. Párová čtení, ve kterých se pouze jeden konec shlukuje a druhý konce se nemapuje k referenci, byly použity k označení variabilních sekvencí, které nejsou zahrnuty v referenčním genomu (nové inzerce). [1][23]

Přístupy typu read-pair jsou výkonné, ale řešení nejednoznačných přiřazení při mapování v opakujících se regionech je náročné. Ta čtení, která jsou mapována na více míst jsou často ignorována, což ztěžuje detekci variabilit v repetitivních oblastech genomu [23]. Přesná predikce míst zlomů SV závisí na pevném rozložení velikosti fragmentů, což má vliv na obtížnost a náklady tvorby knihovny. Na základě typických velikostí fragmentů NGS je více než 90 % objevených událostí menších než 1 kb a většina z nich jsou spíše delece než inzerce. Problémem jsou však menší delece, které nepřekračují velikost samotných čtení nebo naopak inzerce větší než průměrná velikost inzertu genomové knihovny. [1][23]

V dnešní době existuje mnoho výpočetních nástrojů založených na přístupu typu read-pair, včetně PEMer, VariationHunter, BreakDancer, MoDIL, MoGUL, HYDRA, Corona a SPANNER. [1][23]

Skovgaard (Skovgaard et. al, 2011) představil nový způsob využití technologie Variation Hunter, kdy ji kombinoval s analýzou počtu kopií templátové DNA v rychle rostoucích bakteriálních kulturách E.coli. Namísto mapování čtení získaných ze stacionární kultury, kde je počet kopií konstantní, získali kontrastní chování čtení z DNA exponenciálně rostoucích kultur, což potvrdilo přítomnost velkých inverzí. [27]

4.3.2 Metoda založená na hloubce čtení

Hlavně díky velkému množství čtení s vysokým pokrytím, které poskytují sekvenátory nové generace, se metody založené na hloubce čtení (read depth - RD) v poslední době staly hlavním přístupem k odhadu počtu kopií. Základní hypotéza

metod založených na RD spočívá v tom, že hloubka pokrytí v genomické oblasti je v korelaci s počtem kopií v dané oblasti. [39]

Přístupy postavené na hloubce čtení předpokládají náhodnou (typicky Poissonovou nebo modifikovanou Poissonovu) distribuci v hloubce mapování a zkoumají odchylky od této distribuce k objevení duplikace a delece v sekvenovaném vzorku. Duplicitní oblasti budou vykazovat výrazně vyšší hloubku čtení a delece budou vykazovat sníženou hloubku čtení ve srovnání s duplikovanými oblastmi. [23]

Přístupy read-depth využívající data NGS byly poprvé použity pro definování přestavby genetické informace u rakoviny, segmentové duplikace a sestavení mapy absolutního počtu kopií v lidských genomech. Přistup založený na hloubce čtení dokáže poměrně přesně předpovědět absolutní počet kopií, což postrádají nástroje založené na rozdělených či párových čtení, které pracují naopak s informací o poloze. Rozlišení míst zlomu je však touto metodou často nedokonalé. Metody založené na RD se v zásadě řídí postupem ve čtyř krocích k nalezení CNV: mapování, normalizace, odhad počtu kopií a segmentace. V kroku mapování jsou krátké čtení zarovnány s referenčním genomem a hloubka čtení se vypočítá podle počtu mapovaných čtení v předdefinovaných oknech. Hloubka pokrytí musí být v daném okně shodná, ale v porovnání s hloubkami pokrytí sousedních oken musí být značný rozdíl. Druhý krok se zaměřuje mimo jiné na normalizaci a korekci potenciálních zkreslení v hloubce čtení způsobených zejména obsahem GC a repetitivními oblastmi genomu. S normalizovanými hloubkami čtení je úkolem třetího kroku odhadnout přesný počet kopií podél chromozomu, aby se stanovil zisk nebo ztráta. Nakonec se sloučí genomické oblasti s podobným počtem kopií[39]

Metody, které se pokouší objevit menší delece a duplikace při lepším rozlišení bodu zlomu, zahrnují algoritmus EWT (event-wise-testing) a CNVnator. [39]
Mezi RD metody patří SegSeq, PEMer, VariationHunter, MoDIL, Pindel, BreakDancer a ABI SOLiD softwarové nástroje.[23]

4.3.3 Metoda založená na rozdělených čtení

Metody založené na rozdělených čtení – split-reads (SR) jsou schopny detekovat delece a malé inzerce až na úrovni jednoho páru bází a byly nejprve aplikovány na delší Sangerova sekvenační čtení. Metody SR používají pár čtení, ve kterém je jedno čtení zarovnáno s referenčním genomem, zatímco druhé čtení není mapováno vůbec, nebo jen částečně. Nenamapovaná čtení jsou rozdělena na více fragmentů. První a poslední fragmenty každého rozděleného čtení jsou pak samostatně zarovnány s referenčním genomem. Nepřetržitý úsek mezer ve čtení indikuje deleci, úsek mezer v referenci indikuje inzerci. Tento krok přemapování proto poskytuje přesné počáteční

a koncové polohy inzercí či delecí. Rozšíření tohoto přístupu může také detekovat inzerce mobilních elementů (MEI), pokud jsou čtení dostatečně dlouhá, aby překlenula mobilní element (například > 400 bp pro Alu prvky), a charakterizovala tak obsah celé sekvence. [39]

Algoritmy split-read slouží pro detekci širokého rozsahu tříd SV s přesným rozlišením míst zlomu. Nicméně je tato metoda v současné době spolehlivá pouze v unikátních oblastech genomu. [39]

Algoritmus Pindel byl první metodou založenou na SR k identifikaci míst zlomů velkých delecí (1 bp - 10 Kb) a středně velkých inzercí (1 - 20 bp). Mezi další nástroje založené na metodě rozděleného čtení patří Alignment with Gap Excision (ACE), split-read identification, calibrated (SRiC) a SLOPE. [39]

V roce 2012 Sun a kol. (2012) použil 454 pyro-sekvenování v kombinaci s výpočetní metodou „split mapping" k detekci spontánně se vyskytujících přeskupení genomu (SGR) z rychle rostoucí kultury Salmonella sp. Body zlomu byly stanoveny s rozlišením jednotlivých párů bází, experimentální ověření zlomů SGR bylo provedeno hybridizací s využitím zámkové sondy. [27]

4.3.4 Metoda založená na sestavení sekvence

Na rozdíl od přístupů RD, PEM a SR, které nejprve zarovnávají čtení NGS se známým referenčním genomem před samotnou detekcí, metody založené na sestavení sekvence (sequence assembly, AS) nejprve rekonstruují fragmenty DNA, tj. kontigy, z krátkých čtení spojením překrývajících se čtení. Porovnáním sestavených kontigů s referencí se identifikují genomické oblasti s různými počty kopií. Přímé sestavení krátkých čtení bez použití reference se nazývá de novo assembly. Při sestavení sekvencí se může také použít referenční genom jako pomocník ke zlepšení výpočetní účinnosti a kvality kontigu.[39]

Metody AS potenciálně poskytují nestranný přístup k objevování nových genetických variabilit sahajících od úrovně jednoho páru bází po velké změny struktury. Metody AS se však v detekci CNV u eukaryotických genomů používají jen zřídka, a to z důvodu nízké kvality sestavených kontigů a velkých nároků na výpočetní zdroje.[39]

Sestavení sekvencí však slibuje, že bude nejvšestrannější metodou díky porovnání dvou genomů. Je potřeba ovšem vyřešit problém oblastí repetic a duplikací, ve kterých sestavení často selhává. Jeho aplikace na detekci SV není rutinní a bude vyžadovat podstatný vývoj. [1]

Známé algoritmy de novo sestavení pro data pocházející z nové generace celogenomového shotgun sekvenování (NG-WGS) zahrnují EULER-USR, ABySS,

SOAPdenovo a ALLPATHS-LG. Za použití assembleru Cortex lze sestavení genomu provést zcela de novo nebo s různým stupněm informací z reference. Cortex má také tu schopnost sestavovat více genomů zároveň a přiřazovat SV mezi vzorky bez potřeby reference. NovelSeq slučuje metody de novo a lokálního sestavení k charakterizaci nových sekvenčních inzercí, cílem TIGRA je zlepšit odhady míst zlomů při objevování SV. [2]

4.3.5 Kombinované přístupy

Žádný ze čtyř hlavních přístupů k objevování strukturálních variabilit pomocí sekvenčních dat není komplexní. Když je na stejné vzorky DNA aplikováno mnoho algoritmů a experimentálních metod, zůstává významná část uznaných variabilit jedinečná pro konkrétní přístup. Každá metoda má různé silné a slabé stránky v detekci, v závislosti na typu variability nebo vlastnostech sekvence v místě SV. Toto uvědomění vedlo nedávnému vývoji nástrojů, které začleňují více metodik najednou pro zlepšení citlivosti a specifičnosti. [1]

Příkladem jsou nástroje SVDetect, CNVer, genom STRiP, GASVPro a inGAP-sv, které kombinují přístupy read-pair a read-depth různými způsoby, aby spolehlivěji detekovaly CNV [39]. Například algoritmus mrFAST byl navrhnut k efektivnímu mapování velkého množství krátkých čteních k referenčnímu genomu a výpočtu přesné hloubky čtení. Zaměřuje se zejména na to, aby byla čtení mapována na všechny možné pozice v referenčním genomu, aby mohl předpovědět absolutní počet kopií i v duplikovaných sekvencích. [2][1]

5 ALGORITMUS DETEKCE CNV

V praktické části diplomové práce jsem se zaměřila na návrh algoritmu pro detekci CNV. Algoritmus je založen na hloubce čtení a pracuje tedy se soubory obsahujícími pokrytí genomu (anglicky coverage). Coverage soubor poskytuje počet seřazených čtení na každé pozici referenční sekvence. Jak už bylo popsáno v teoretické části (kap.4.3.2), úseky s větší hloubkou čtení mohou naznačovat výskyt duplikací, naopak úseky s menší hloubkou čtení jsou spojeny s výskytem delecí.

Pokrytí čtení je však ovlivněno řadou zkreslení – neurčitost v mapování čtení, zkreslení počtu kopií z důvodu různé fáze buněk v replikačním cyklu, a zejména pak obsah guaninu a cytosinu mající hlavní vliv na pokrytí ve všech současných platformách sekvenování [5]. Detekci CNV úseků proto předchází GC normalizace. Samotný algoritmus detekce CNV je založen na metrice zvané z – skóre. Z – skóre každé hodnotě pokrytí přiřazuje počet směrodatných odchylek od průměru celého pokrytí. Jako CNV úsek je detekována každá pozice, jejíž z – skóre splňuje stanovený práh. Dále je na signál aplikováno posuvné okno, které zaručí, že detekované úseky mají minimální nastavenou délku. Aby bylo možné zhodnotit úspěšnost detekce, byl vytvořen umělý dataset, na kterém je navržený algoritmus otestován.

5.1 GC normalizace

Zkreslení obsahem GC popisuje závislost mezi počtem GC v oblasti a počtem fragmentů (pokrytí, hloubka čtení) do této oblasti mapovaných. Vztah mezi GC obsahem a počtem fragmentů byl popsán jako unimodální – zastoupení fragmentů bohaté či chudé na obsah GC je nižší od průměru. Zkreslení je považováno za artefakt spojený s procesem PCR, vznikající zejména během kroku přípravy knihovny [5]. Toto zkreslení však není mezi jednotlivými genomy jednotné a neexistuje shoda ohledně nejlepší metody pro jeho odstranění.[5]

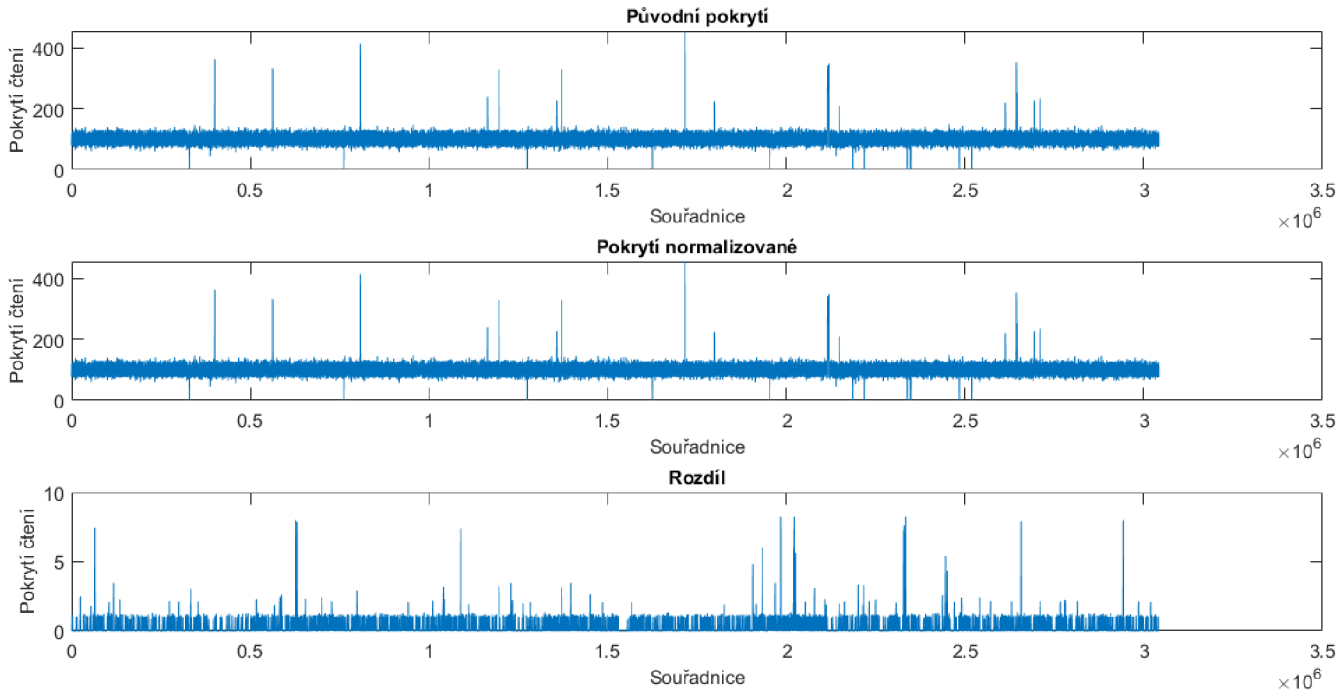
Jelikož detekční algoritmus, navržený v této práci, je založený na hloubce čtení, je GC obsahem ovlivněna i tato analýza CNV. Z toho důvodu je před detekováním CNV úseků na signál nesoucí pokrytí čtení aplikována GC normalizace.

Použitá normalizace je odvozená z již nevržené metody (Jugas et al., 2019). Nejprve je v klouzavém, překrývajícím okně o dané velikosti (70 bp) spočítán obsah guaninu a cytosinu. Dále je pro každou hodnotu GC obsahu spočítán medián hloubky čtení. To znamená, že například pro hodnotu GC obsahu 5 jsou nalezeny všechny pozice se stejným GC obsahem, na těchto pozicích je získán medián hloubky čtení a zapsán do předdefinované tabulky. Tímto postupem je vypočítán medián pokrytí pro

všechny hodnoty GC obsahu. Následně je korigováno pokrytí na všech pozicích signálu s hloubkou čtení větší jako 5 dle následujícího vzorce [6]:

$$RC_i^{corr} = RC_i \times \frac{mRC}{mRC_{GCi}}. \qquad 5.1$$

Hodnota pokrytí je korigována váhováním původní hodnoty pokrytí čtení podílem celkového mediánu pokrytí čtení a mediánu pokrytí čtení se stejným obsahem GC. RC_i ve vzorci představuje i – tou hodnotu pokrytí čtení (RC – read coverage), mRC je medián pokrytí čtení celého signálu, mRC_{GCi} je medián hodnot pokrytí čtení s odpovídajícím obsahem GC a RC_i^{corr} je korigovaná hodnota pokrytí čtení na pozici i. Díky přepočtu hloubky čtení na jednotlivých pozicích je tato metoda sice náročnější na výpočet, ale je také přesnější. [17]



Obr. 5.1 GC normalizace

5.2 Detekce CNV

Detekce CNV založená na z – skóre představuje jednoduchou a poměrně přesnou metodu, která se již osvědčila v několika studiích (Nord et al., 2011; Tolhuis, Karten, 2018; Derouault et al.,2020). Z – skóre bývá často používáno v kombinaci s dalšími

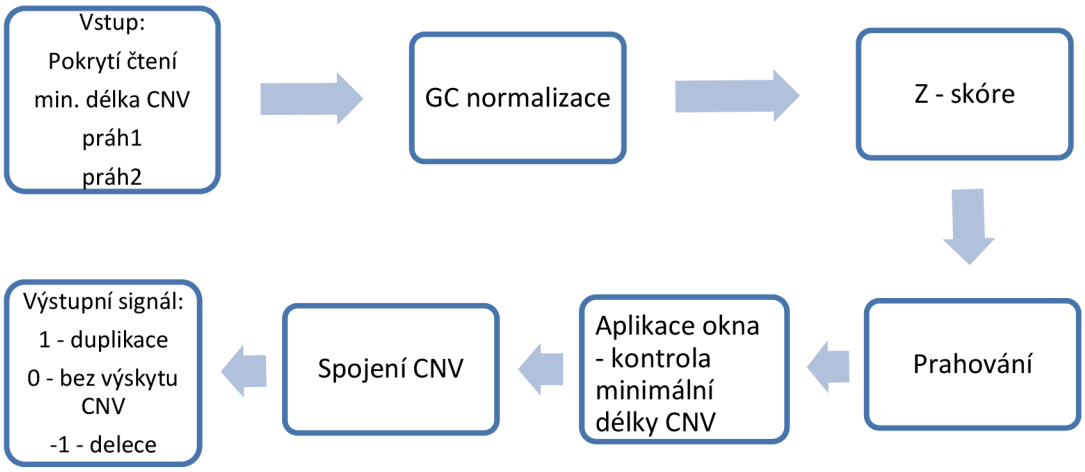
metrikami, jako je CNV podíl či variantní alelová frekvence (VAF) [7]. Tyto dvě metriky však vyžadují pro výpočet kontrolní vzorky. Z toho důvodu se v této práci spokojíme pouze s výpočtem z – skóre.

Postup detekce CNV je shrnut ve schématu (Obr. 5.2). Detekce je aplikována na signál, který byl normalizací částečně zbaven GC zkreslení. Z – skóre každou pozici normalizovaného pokrytí převádí na počet směrodatných odchylek vzdálených od celkového průměru signálu dle následujícího vzorce:

$$Z - score_i = \frac{nRC_i - \mu}{\sigma}. \qquad 5.2$$

$Z - score_i$ představuje vypočítané z-skóre na pozici i, nRC_i je normalizované pokrytí čtení na pozici i, μ představuje průměr celého signálu a σ směrodatnou odchylku.

Jako odpovídající počet směrodatných odchylek, poukazující na výskyt CNV úseku, se uvádí hodnota v rozmezí 4–5 (Xu et al.,2015; Nord et al., 2011; Tolhuis, Karten, 2018). Kladná hodnota tohoto prahu (práh1) značí duplikaci, naopak záporná hodnota ukazuje na deleci. Při aplikaci detekce na reálná data však byla záporná hodnota v rozmezí 4-5 pro detekci delecí příliš nízká. Jelikož však vždy známe minimální hodnotu z-skóre (pokrytí nulové), byl práh pro detekci delecí nastaven na 85 % z této minimální hodnoty.



Obr. 5.2 Schéma algoritmu detekce CNV

Výstupní signál pak obsahuje +1 na pozicích duplikace, -1 na pozicích delece a nuly na ostatních pozicích. Na tento signál je dále aplikováno posuvné okno o nastavitelné velikosti. Okno klouže po signálu pozici po pozici a jeho velikost udává minimální délku CNV úseků. Pouze pozice signálu, které naplní okno v požadovaném

rozsahu, jsou propuštěny na konečný výstup. Volitelným prahem (práh2) můžeme určit, zda detekovaným CNV úsekem musí být naplněný celý rozsah okna, nebo jen jeho procentuální část. V tom případě je v okně povolen výskyt i nulové hodnoty, ne však na počáteční a konečné pozici segmentu, aby nedocházelo k jeho falešnému prodlužování. Tento krok může zabránit předčasnému ukončení CNV úseku z důvodu lokálního poklesu hodnot pokrytí.

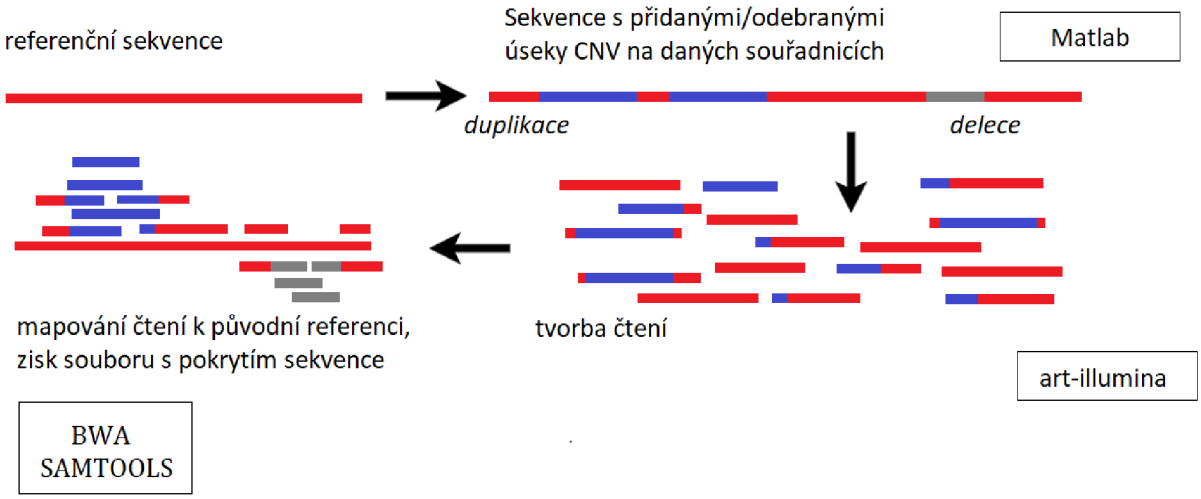
Dalším krokem, zvyšujícím robustnost proti lokálním poklesům, je možnost volby spojení CNV úseků. CNV úseky, které jsou od sebe odděleny mezerou kratší, než je povolená délka, a navíc průměrná hodnota z – skóre této mezery dosahuje zvoleného prahu, jsou tyto úseky spojeny do jednoho CNV segmentu.

Algoritmus dále zahrnuje funkci pro vystřihnutí CNV úseků z referenční sekvence na základě zjištěných souřadnic a uložení do fasta souboru.

5.3 Testování algoritmu detekce na umělých datech

Postup testování algoritmu je znázorněn na obrázku (

Obr. 5.3).



Obr. 5.3 Testování detekce

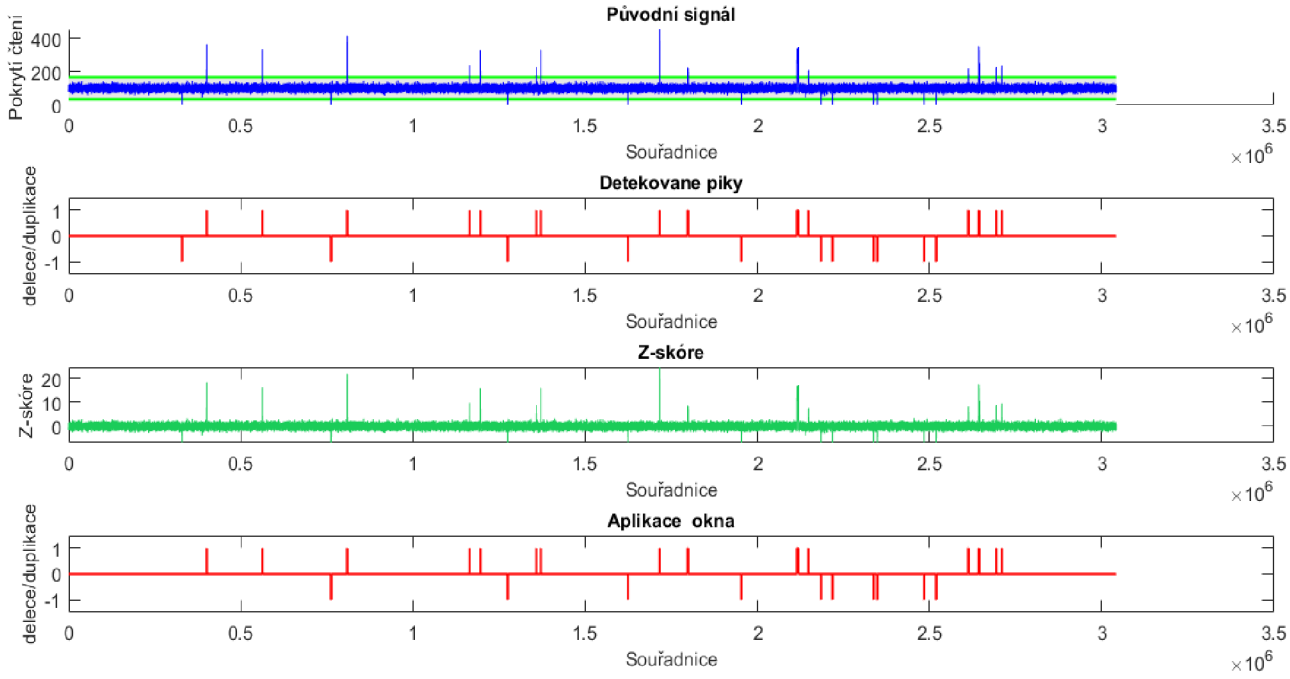
Pro otestování algoritmu detekce byl využit umělý dataset, který byl vytvořen na základě článku CNOG pro (Brynildsrud et al., 2015), a který pracuje s osekvenovaným genomem bakterie *Staphylococcus aureus TW2*.

Do referenční sekvence bylo v Matlabu vloženo či odebráno 30 úseků DNA na daných souřadnicích. Další zpracování probíhalo na virtuálním stroji s operačním systémem *Xubuntu*, použitá verze 18, s nastavením 4 jader CPU a 16 GB RAM. Z

obohacené sekvence o úseky CNV byla vytvořena sada čtení s využitím nástroje *art-illumina*. Jedná se o nástroj generující syntetická sekvenační čtení na principu Illumina sekvenátorů [16].

Tabulka 1.: Parametry nástroje ART-illumina

Čtení	**Párová, 76 bp**
Průměrné pokrytí	100 x
Délka insertu	500 bp
Směrodatná odchylka	100
Sekvenační systém	HiSeq 2500



Obr. 5.4 Detekce CNV v testovacím datasetu, minimální délka CNV 15 bp

Na čtení byly následně aplikovány nástroje *SAMtools* a *BWA*. *BWA* (Burrows-Wheeler Aligner) je softwarový balíček, který využívá zpětného vyhledávání s Burrowa-Wheeler transformací k mapování krátkých sekvenačních čtení k velkému referenčnímu genomu [21]. V tomto případě byla čtení mapována k původní referenci, tedy bez obsahu úseků CNV. Dále bylo využíváno nástroje *SAMtools* [20]. Namapovaný soubor byl převeden pomocí *samtools view* ze *SAM* formátu (Sequence

Alignment Map) do úspornější binární podoby – formátu *BAM* (Binary Alignment Map). Poté byla čtení seřazena podle výskytu v referenční sekvenci a také indexována. Po aplikaci nástroje *samtools depth* byl získán soubor s pokrytím sekvence, ve kterém mohly být CNV detekovány navrženým algoritmem. Pro určení pozitivně negativních detekcí byl stejným způsobem získán také soubor s pokrytím sekvence bez obsahu úseků CNV.

5.3.1 Výsledky

Souřadnice úseků CNV detekovaných navrženým algoritmem byly srovnány se známými souřadnicemi simulovaných CNV úseků a vyhodnoceny. Podrobné výsledky jsou uvedeny v přiložené tabulce (*Výsledky.xlsx*), zde jsou uvedeny pouze ty nejdůležitější. Pro stanovení úspěšnosti algoritmu byly zvoleny metriky senzitivita a specifita. Ke stanovení těchto hodnot je potřeba znát počty skutečně pozitivních (TP), falešně pozitivních (FP), skutečně negativních (TN) a falešně negativních (FN) nálezů. Senzitivita určuje pravděpodobnost, že detekce bude pozitivní v případě výskytu CNV (TP/(TP+FN)). Specifita naopak udává pravděpodobnost, že detekce bude negativní na pozicích, kde se žádné CNV nevyskytuje (TN/(TN+FP)). Detekce byla provedena s následujícími parametry (Tabulka 2).

Tabulka 2.: Parametry detekce

GC normalizace	70 bp
Práh	4.65
Minimální délka CNV	4 / 15 bp
Procento naplnění okna	100 %
Spojování CNV	Ne

Na základě znalosti umělého datasetu byly zvoleny dvě velikosti okna určující minimální délku CNV: 4 bp (Tabulka 3) a 15 bp (Tabulka 4). Procento naplnění okna bylo nastaveno na 100 %, tedy detekovaný úsek musel zaplnit okno v celém rozsahu. S nastaveným oknem 4 bp algoritmus detekoval 27 CNV úseků ze 30, s oknem 15 bp bylo detekováno o jeden CNV úsek méně. V obou případech byl CNV úsek na pozicích {2612863 - 2613755} detekován jako dva oddělené fragmenty, což bylo vyhodnoceno jako jeden skutečně pozitivní a jeden falešně pozitivní nález. Pro určení specifity byla detekce aplikována také na signál s pokrytím sekvence, která neobsahovala žádný CNV úsek. S oknem 4 bp byl detekován jeden úsek CNV, oknem 15 bp žádný CNV úsek

detekován nebyl. Menší okno tedy dosáhlo vyšší senzitivity, ale nižší specifity než okno dlouhé 15 bp.

Tabulka 3.: Výsledky testování detekce, okno 4 bp

Detekce CNV minimální délka 4 bp					
s CNV				bez CNV	
TP	**FP**	**FN**		**TN**	**FP**
27	1	3		29	1
Senzitivita	90 %	**Přesnost**	96,86 %	**Specifita**	96,67 %

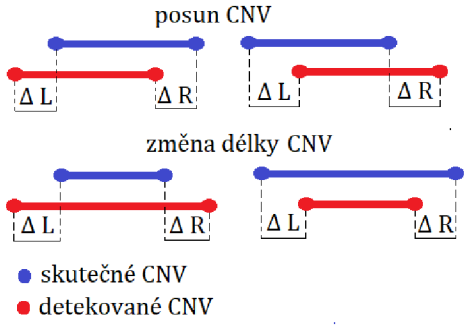
Tabulka 4.: Výsledky testování detekce, okno 15 bp

Detekce CNV minimální délka 15 bp					
s CNV				bez CNV	
TP	**FP**	**FN**		**TN**	**FP**
26	1	4		30	0
Senzitivita	86,67 %	**Přesnost**	96,3 %	**Specifita**	100 %

Dále byla detekce vyhodnocena z hlediska určení správné pozice úseku CNV a také jeho velikosti. Pro vyhodnocení byla využito následujícího vzorce [38]:

$$procento\ změny\ CNV = \frac{\Delta|L| - \Delta|R|}{2 * délka\ CNV} * 100. \qquad 5.3$$

Symboly $\Delta|L|$ a $\Delta|R|$ jsou objasněny na následujícím obrázku (Obr. 5.5).



Obr. 5.5 Schéma vyhodnocení změny CNV

Výsledky ukazují, že CNV úsek je průměrně změněn o 23,56 %, avšak bereme-li v potaz pouze úseky delší jako 100 bp, je tato hodnota snížena na 5,28 %. Při detekci docházelo spíše k posunu detekovaného CNV úseku a k chybnému určení bodů zlomů, které jsou průměrně změněny o 20 bp. Naopak délka CNV úseků byla určena poměrně

přesně. Skutečná délka CNV úseku se od té detekované liší průměrně o 4 bp (Tabulka 5).

Tabulka 5.: Vyhodnocení změny CNV

Procento posunu/změny délky	23,56 %
Procento posunu/změny délky pro CNV nad 100 bp	5,28 %
Průměrný posun začátku CNV	17,26 bp
Průměrný posun konce CNV	20,69 bp
Průměrná změna délky	4,34 bp

Nakonec bylo provedeno také srovnání s CNOGpro metodou, která byla testována na stejném datasetu. CNOGpro je nástroj navržený pro rychlou detekci variability počtu kopií v prokaryotickém genomu založený na skrytém Markovovu modelu [5]. Výsledky jsou uvedeny v tabulce (Tabulka 6).

Obě metody dosahují stoprocentní specifity, naopak se liší ve vypočítané senzitivitě. CNOGpro metoda nevykazuje příliš dobrých výsledků při detekci krátkých CNV úseků do 100 bp. Nástroj detekoval pouze 2 z 10 CNV úseků do 100 bp, metoda využívající z-skóre jich detekovala 7 z 10. Obě metody detekovaly všechny úseky delší než 100 bp. I přes nízkou senzitivitu, CNOGpro metoda dokáže alespoň částečně kvantifikovat počet kopií u devatenácti případů. Metoda navržená v této práci rozlišuje pouze delece a duplikace.

Tabulka 6.: Srovnání metod

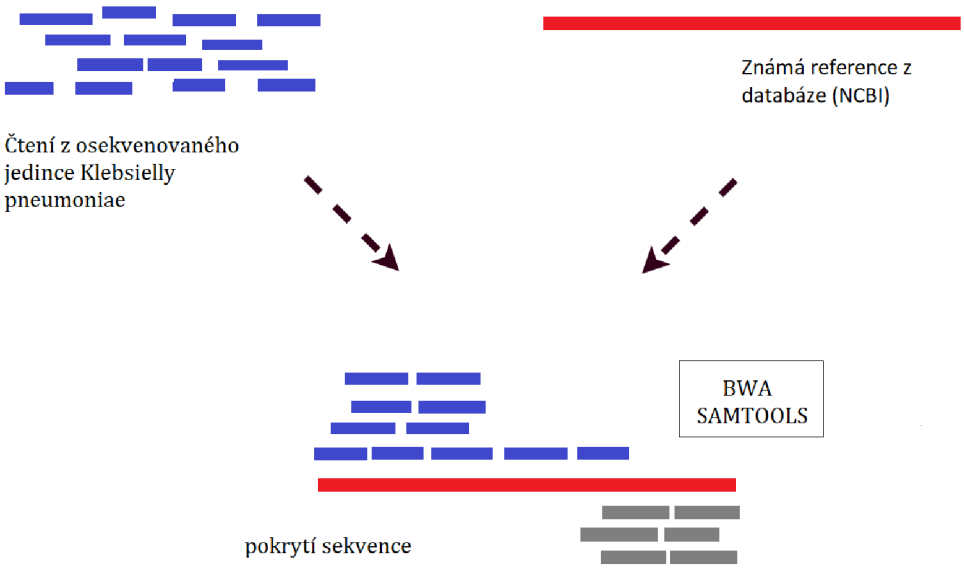
Srovnání metod	**CNOGpro**	**Z - skóre - 4**	**Z – skóre - 15**
specifita bez CNV	100 %	96,66 %	100 %
specifita s CNV	99,96	--	--
FP s CNV	0	1	1
Senzitivita	73 %	90 %	86,67 %
detekce CNV do 100 bp	2/10	7/10	6/10
detekce CNV nad 100 bp	20/20	20/20	20/20
správná/částečná kvantifikace	19/30	pouze delece/duplikace	

I přes poměrně pozitivní výsledky, je nutné poznačit, že simulovaná data nejsou zatížena stejným zkreslením jako data reálná, což se může při detekci na reálných datech negativně projevit.

6 DETEKCE CNV V DENOVO SEKVENAČNÍCH DATECH

Algoritmus byl aplikován na genomy bakterie *Klebsielly pneumoniae*. Osekvenovaná čtení z těchto bakterií pochází z Fakultní nemocnice u sv. Anny. Reálná data poskytnul vedoucí práce.

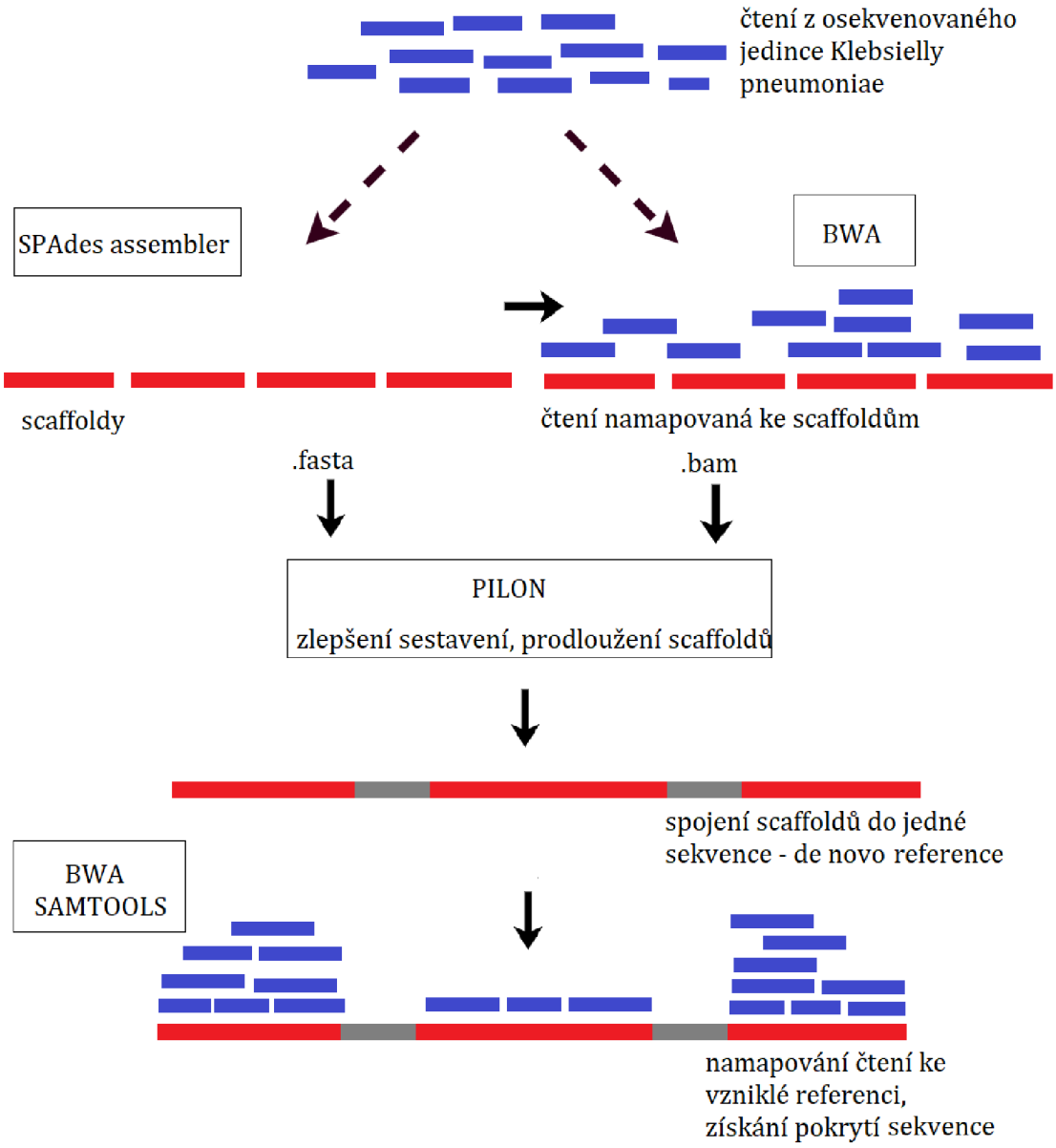
Algoritmus byl nejprve aplikován tradičním způsobem – osekvenovaná čtení jedince bakterie *Klebsielly pneumoniae* byla mapována ke známé referenční sekvenci jedince stejného druhu - *Klebsiella pneumoniae subsp. pneumoniae* NTUH-K2044 (5 248 520 bp), reference byla získána z databáze NCBI (Obr. 6.1). Genomy dvou jedinců však nebudou nikdy úplně totožné. Některé úseky obsažené v genomu zkoumaného jedince nemusí být obsaženy v genomu referenčním. Při mapování k referenci pak vzniká množství nenamapovaných čtení, které je zahozeno, a dochází tak ke ztrátě cenné informace. Z toho důvodu je v této práci navržený alternativní postup, ve kterém si vytváříme vlastní, de novo referenci z dostupných sekvenačních čtení daného jedince.



Obr. 6.1 Mapování ke známé referenci

6.1 Mapování *de novo*

Pro přehlednost je zde uveden postup sestavení *de novo* reference (Obr. 6.2). Zpracování opět probíhá v terminálu operačního systému *Xubuntu.* Ze sady osekvenovaných čtení jedince *Klebsielly pneumoniae* jsou nejdříve sestaveny kontigy pomocí assembleru SPAdes [3]. SPAdes je assembler založený na aplikaci více de Bruijnových grafů (každý vytvořený k-mery o různé velikosti k), díky tomu se pak lépe vypořádává s různým pokrytím napříč genomem. SPAdes na výstupu vydává sadu souvislých sekvencí DNA (kontigů).



Obr. 6.2 Mapování k de novo referenci

K těmto kontigům jsou následně mapovány čtení s využitím *BWA* nástroje. SAM soubor musí být převeden na BAM soubor a společně s vytvořenými kontigy (soubor.fasta) vstupuje do softwarového nástroje *PILON* [28]. *Pilon* se snaží vylepšit vytvořené kontigy tak, že usiluje o jejich prodloužení. Využívá k tomu namapovaná čtení k těmto kontigům obsažených v souboru BAM. Výstupní scaffoldy jsou poté spojeny do jedné dlouhé sekvence – vložením tisíce písmen N. Díky takto vytvořené referenci nejsme závislí na použití cizí reference jiného jedince stejného druhu, který se může lišit v obsahu genů. Nevzniká tak množství nenamapovaných čtení, které by bylo při mapování zahozeno, jelikož mapujeme stejná čtení, ze kterých byla nová reference vytvořena. Mapování je opět provedeno nástrojem *BWA* a pokrytí sekvence je získáno s využitím balíčku *SAMtools*.

Tabulka 7.: Použité příkazy pro sestavení sekvence denovo

SPADES	spades.py -1 $R1 -2 $R2 --threads 16 --careful --cov-cutoff auto -o $folder
--threads 16 – počet vláken --careful – snížení počtu záměn bází a krátkých indelů --cov-cutoff - auto -o – automatický výpočet prahu pro minimální pokrytí s využitím konzervativní strategie	
PILON	pilon-1.23.jar --threads 6 --fix "all" --genome $file --frags $bamfile --output $Styp --outdir pilon_output
-- threads 6 – stupeň paralelismu při zpracování --fix "all" – Pilon se pokusí opravit všechny možné chyby – chyby v jednotlivých pozicích, indely, vyplnění mezer, oprava chybných sestavení --dále jsou definovány vstupy a výstupy	
BWA	index -a mem -t 6
- indexace čtení, mapování čtení k referenci	
SAMtools	view -bt sort -o index depth -a
-převod na BAM soubor, seřazení namapovaných čtení dle souřadnic, indexace, výpočet hloubky čtení na každé pozici sekvence	

Uvedeným postupem bylo získáno 10 bakteriálních *denovo* referencí a jejich pokrytí. Pro vyhodnocení sestavených genomů byl použit nástroj QUAST, který umožňuje hodnotit kvalitu sestavení (assembly) i bez referenčního genomu [15]. Výsledná statistika je uvedena v tabulce (Tabulka 8). Statistika N50 je délka nejkratšího kontigu, který spolu s ostatními, stejně dlouhými nebo delšími kontigy, tvoří 50 % celkového assembly. L50 je pak nejmenší možný počet takovýchto kontigů. Kvalitní sestavení genomu je určeno vysokou hodnotou N50 a nízkou hodnotou L50

(např.: genom S18). Stejná definice platí také pro metriky N75 a L75, avšak zahrnující 75 % assembly. Sekvence dosahují celkově nižšího průměrného pokrytí, oproti tomu vyšší medián pokrytí je způsobený zejména vyšším pokrytím kratších kontigů na konci assembly.

Tabulka 8.: Statistika sestavení *de novo* sekvencí

Genom	Celková délka	Počet kontigů	Nejdelší kontig	GC %	N50	L50	N75	L75	medián pokrytí	průměrné pokrytí
S02	5427423	86	1890289	57,33	502827	3	340135	6	141,0	48,1
S04	5596424	164	1085651	57,14	436844	4	316799	8	79,0	25,7
S06	5634083	145	1085651	57,09	473116	4	324578	8	135,0	45,1
S11	5430923	92	1890289	57,32	502827	3	345116	6	169,0	57,6
S12	5428946	95	1038012	57,32	502827	4	345116	7	107,0	36,8
S13	5629214	154	653601	56,83	376804	6	188716	12	89,0	21,0
S18	5484924	105	2874596	57,3	2874596	1	377641	4	90,0	32,4
S19	5649330	80	1350651	57,08	511096	4	367924	7	67,0	25,6
S36	5732645	277	383776	57,06	229565	11	141306	19	18,0	6,2
S46	5766221	247	659171	56,65	291735	7	114342	14	20,0	6,9

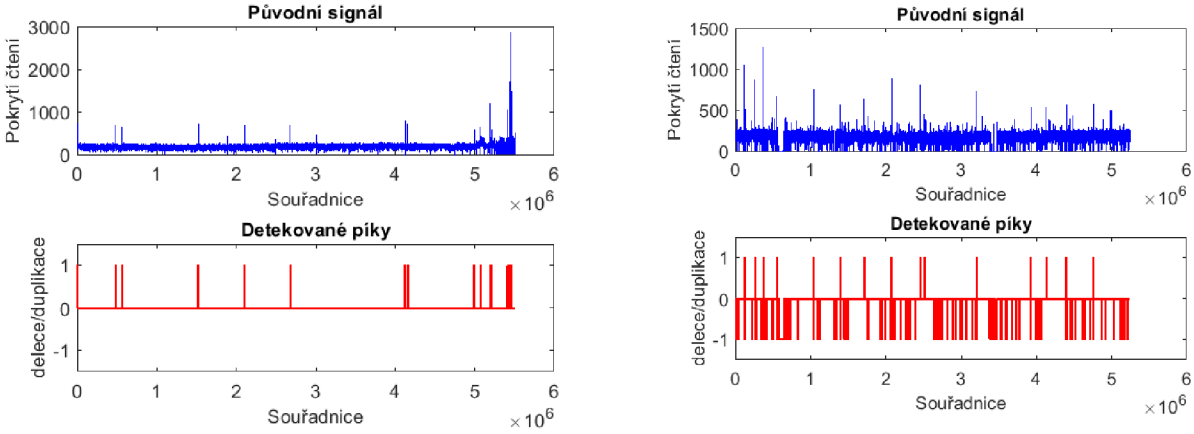
6.2 Srovnání přístupů

Pro srovnání obou přístupů detekce CNV – ze známé reference i sekvence *de novo*, byl algoritmus aplikován na deset jedinců *Klebsielly*. Parametry detekce jsou uvedeny v tabulce (Tabulka 9). Práh pro detekci duplikací byl nastaven na hodnotu 4 v obou případech. Práh pro detekci delecí v pokrytí známé reference byl nastavován automaticky, a to na 85 % z minimální hodnoty z – skóre, jelikož hodnota – 4 byla pro detekci příliš nízká. V pokrytí sekvencí sestavených *de novo* delece nejsou předpokládány. Povolené naplnění okna kontrolující minimální délku CNV bylo nastaveno na 80 %. Při zvolené délce okna 18 bp musí patnáct pozic v okně vykazovat deleci/duplikaci, zbylé tři pozice mohou být nulové. Spojení CNV bylo povoleno pro úseky s maximální mezerou 10 bp.

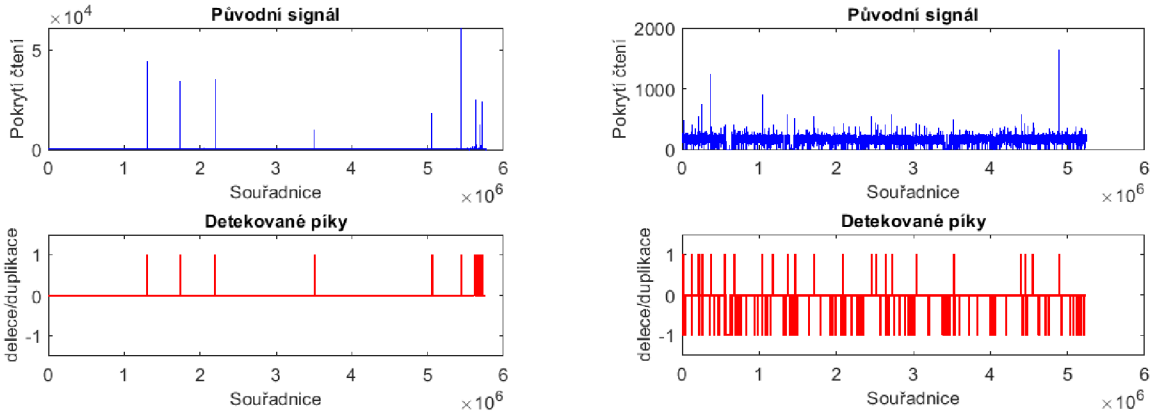
Tabulka 9.: Parametry pro detekci CNV

GC normalizace	70 bp
Práh	4 / auto
Minimální délka CNV	18 bp
Procento naplnění okna	80 %
Spojování CNV	10 bp

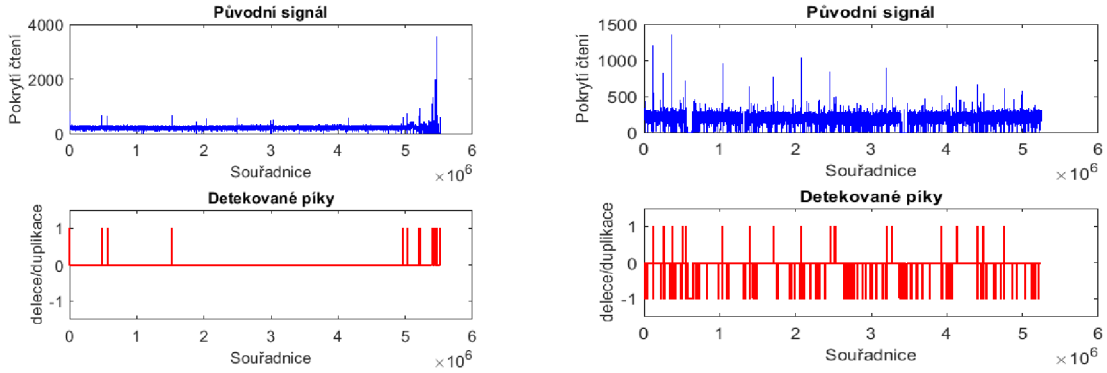
Vybrané výsledné detekce jsou znázorněny na následujících obrázcích.



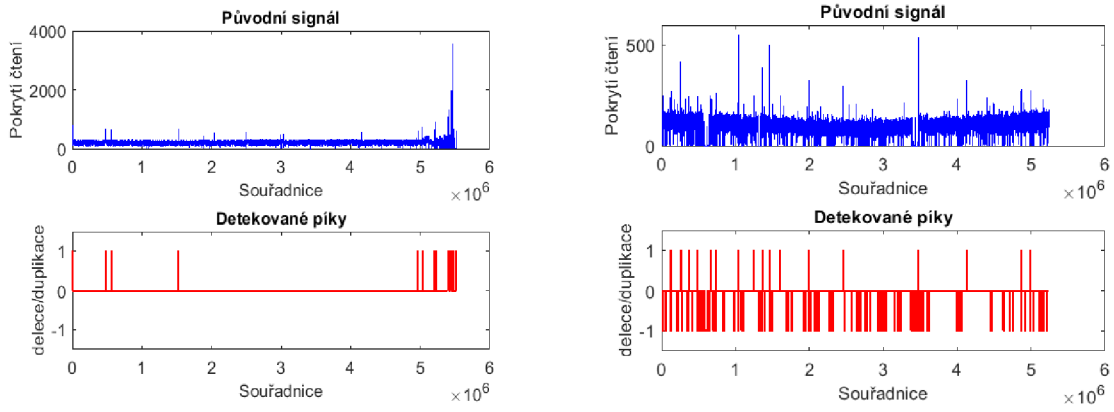
Obr. 6.3 Genom S02 – *de novo* reference (vlevo), známá reference (vpravo)



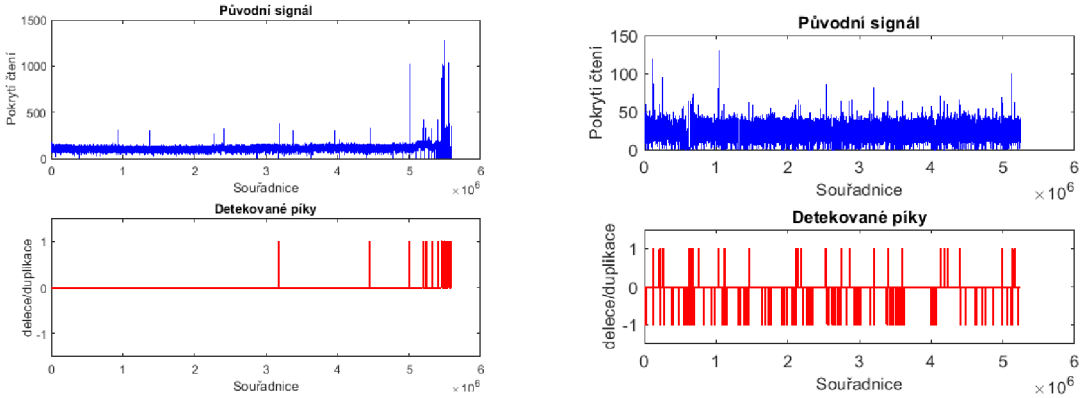
Obr. 6.4 Genom S06 – *de novo* reference (vlevo), známá reference (vpravo)



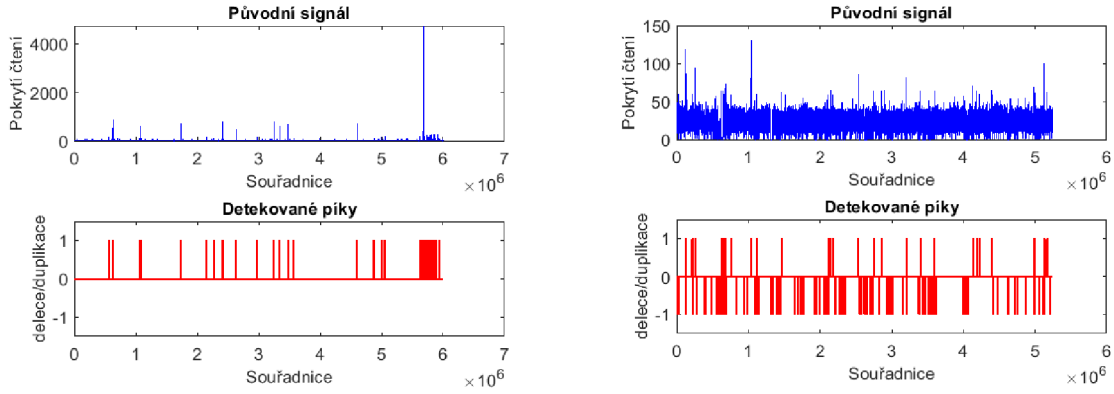
Obr. 6.5 Genom S11 - *de novo* reference (vlevo), známá reference (vpravo)



Obr. 6.6 Genom S13 - *de novo* reference (vlevo), známá reference (vpravo)



Obr. 6.7 Genom S18 - *de novo* reference (vlevo), známá reference (vpravo)



Obr. 6.8 Genom S46 - *de novo* reference (vlevo), známá reference (vpravo)

Na první pohled je patrné, že většina CNV segmentů v *de novo* referenci se nachází na konci sekvence. To si vysvětluji tím, že na konci assembly se nacházejí kratší kontigy, u kterých bylo pravděpodobně obtížnější určit polohu, jelikož se

v sekvenci vyskytují vícekrát. To dokazuje také vyšší pokrytí těchto segmentů oproti zbytku sekvence.

Aby bylo možné obě metody porovnat, údaje o detekovaných CNV segmentech byly zaneseny do tabulky (Tabulka 10). Pro účely této práce byly v úvahu brány pouze duplikované oblasti.

Tabulka 10.: Srovnání přístupů známá reference/ de novo

	Úseků (duplikace)		Počet bp		Průměrná délka		Maximální délka		Unikátní CNV	
Genom	reference	denovo	reference	denovo	reference	denovo	reference	denovo	Matlab	Blastn
S02	31	27	4987	16597	160,87	614,70	604	5273	20	15
S04	35	57	4769	21183	136,26	371,63	600	5350	49	49
S06	31	19	4148	3688	133,81	194,11	445	1287	9	12
S11	32	26	4655	20052	145,47	771,23	595	5334	17	16
S12	30	38	4174	19184	139,13	504,84	550	2395	29	29
S13	27	37	3359	21698	124,40	586,43	582	5371	27	28
S18	48	28	6167	17565	128,75	627,32	541	5403	22	21
S19	52	23	9408	14573	180,92	633,61	1403	5447	12	11
S36	78	38	6981	9682	89,50	254,79	842	3938	26	25
S46	45	47	4772	21103	94,93	449,00	528	5365	28	29

6.3 Vyhodnocení

Z uvedených výsledků lze usoudit, že CNV úseky detekované v nově sestavené referenci zahrnují téměř ve všech případech mnohem větší množství bází. Výjimku představuje genom S06, který vykazuje nejmenší počet CNV segmentů. Nebyla však nalezena žádná souvislost mezi detekovaným množstvím CNV úseků a statistikami sestavení referenčních sekvencí, či jejich pokrytím. Z výsledných hodnot je také zřejmé, že detekované segmenty v *de novo* sekvenci, dosahující maximální délky kolem 5 000 bp, jsou mnohem delší než segmenty detekované ve známé referenci. Důvodem takto dlouhých úseků může být spojení více CNV segmentů při sestavování sekvence.

Aby bylo dokázáno, že detekce CNV segmentů *de novo* poskytuje určité množství informace navíc, byly tyto segmenty z genomu vystřihnuty a následně porovnány se známou referencí. Porovnání proběhlo dvěma způsoby. Nejprve v Matlabu metodou komparace znakových řetězců, kdy byl každý úsek CNV srovnáván se známou referencí po celé její délce. Nejlepší shoda byla uložena. Dosahovala-li výsledná shoda 90 % celkové délky segmentu, byl nález vyhodnocen

jako pozitivní. V opačném případě byl segment vyhodnocen jako unikátní. Unikátním označujeme v této práci CNV úsek, který nemohl být ve známé referenci detekován. Představuje tak hodnotnou informaci, která je bežně zahozena. Druhý způsob představuje metodu *BLAST* (Basic Local Alignment Search Tool) [24]. Jedná se o algoritmus, který na základě heuristického přístupu vyhledává podobnost mezi dotazovanou a cílovou sekvencí a přiřazuje jí statistickou významnost. Využíváno bylo konkrétně nástroje *blastn*, který prohledává podobnost mezi CNV úseky a známou referencí. Dosáhne-li podobnost v některém místě statistické významnosti splňující nastavený práh, je úsek vyhodnocen s pozitivním nálezem. V opačném případě je CNV úsek opět vyhodnocen jako unikátní.

Počet jedinečných segmentů vyhodnocených oběma metodami je srovnatelný. U přístupu založeném na znakovém porovnávání řetězců se dá očekávat větší přesnost, avšak jedná se o výpočetně náročnou úlohu, která je s podobným efektem daleko rychleji provedena pomocí metody BLAST.

U všech genomů byly nalezeny úseky, které byly detekované *de novo*, ale ve známé referenci se vůbec nevyskytují. To poskytuje důkaz, že detekování CNV úseků touto cestou nabízí hlubší pohled do problematiky analýzy CNV v prokaryotických organismech. Ačkoliv tato detekce může být zatížena chybou při sestavení, je zřejmé, že se zde nachází množství informace, které při využití tradičního postupu ztrácíme. Zkombinování obou přístupů by mohlo vést k větší biologické přesnosti při detekci CNV segmentů, což by mohlo pomoci při zkoumání antibiotické rezistence a patogenity těchto kmenů.

7 ZÁVĚR

Stěžejním tématem této diplomové práce jsou metody detekce CNV. V úvodu práce je uvedena kapitola týkající se problematiky genomu. Jelikož detekci CNV v eukaryotních organismech bylo věnováno již spoustu pozornosti, tato práce se zaměřuje především na genom prokaryot.

Pro uvedení do kontextu následuje kapitola zabývající se tématem strukturální variability genomu, poněvadž jednou z převládajících forem variabilit je právě variabilita počtu kopií.

Dále práce představuje dnes nejvýznamnější metody sekvenace DNA – metodu Illumina a Oxford Nanopore. Technologie sekvenace nám poskytují čtení, ze kterých je skládána kompletní sekvence DNA. V této práci jsou sekvenační data používána také pro zjištění pokrytí genomu, na jehož základě jsou detekovány CNV úseky.

V následující kapitole jsou podrobně rozebrány metody detekce variabilit počtu kopií, které jsou rozděleny na metody laboratorní a počítačové. Počítačové metody mohou být postaveny na párových čtení, rozdělených čtení, sestavení sekvence či na hloubce čtení – pokrytí.

Variabilita počtu kopií má významný vliv na genetickou variabilitu organismu. V posledních letech se do ohniska zájmu dostává zejména objevení a zkoumání variability počtu kopií v bakteriálním genomu, a to hlavně ve spojitosti s antibiotickou rezistencí, která je zejména u některých druhů bakterií na vzestupu. Díky schopnosti bakteriálních genomů poměrně rychle se přizpůsobovat nepříznivým podmínkám, dochází u těchto organismů často k diverzifikaci genetické informace. To může negativně ovlivnit odhalování DNA segmentů s variabilním počtem kopií, kdy jsou osekvenovaná čtení jednoho jedince mapována k sestavené referenční sekvenci jedince jiného. Dva jedinci se však vždy budou lišit v obsahu genetické informace, dochází tak ke vzniku nenamapovaných čtení, která jsou zahozena.

Aby bylo zabráněno ztrátě této informace, byl navržen alternativní postup, který odhaluje CNV úseky v referenci sestavené ze stejných čtení, která jsou k ní následně mapována za účelem získání pokrytí této sekvence.

Navržený algoritmus detekce CNV pracuje s pokrytím genomu a je založený na metrice zvané z – skóre. Dále je obohacený o kontrolu minimální délky úseků CNV a také o možnost spojení CNV úseků v případě, že jsou od sebe vzdáleny povolený počet bází. Samotné detekci CNV předchází normalizace GC obsahu. Algoritmus byl otestován na umělém datasetu CNOGpro. Do sekvence bylo vloženo 30 úseků na daných souřadnicích, které byly následně porovnány se souřadnicemi detekovaných

CNV segmentů. Algoritmus dosahoval senzitivity 90 % a specifity 100 %. Při nastavených parametrech byl schopný detekovat i krátké úseky CNV o délce 4 bp. Algoritmus však nedokázal příliš dobře rozlišit body zlomu. CNV úseky byly změněny v délce či posunuty průměrně o 23,56 %. Pro úseky pouze delší než 100 bp je tato hodnota pouze 5,28 %. K další nevýhodě patří neschopnost kvantifikace počtu kopií.

Hlavním cílem této práce však bylo dokázat možnost využití nenamapovaných čtení k detekci CNV úseků. Za tímto účelem byly CNV úseky detekovány v deseti genomech bakterie *Klebsiella pneumoniae.* Detekce nejprve proběhla na souborech s pokrytím, které byly získány mapováním osekvenovaných čtení jedinců ke známé referenci. Poté byly CNV úseky odhalovány v souborech s pokrytím *de novo* sekvencí, které byly z dostupných sad čtení sestaveny. Údaje o detekovaných segmentech ze všech deseti jedinců, pro oba postupy byly zaneseny do tabulky a vyhodnoceny. Téměř ve všech případech bylo v segmentech detekovaných *de novo* zahrnuto mnohem větší množství nukletidů, úseky dosahovaly také mnohem větších hodnot délky. Metodami BLAST a prostým porovnáním řetězců bylo také ověřeno, že část těchto úseků představuje tzv. unikátní segmenty, které nejsou ve známé referenci vůbec obsaženy a představují tak cennou informaci, při běžném postupu zahozenou.

I když je tato detekce výrazně ovlivněna kvalitou sestavené reference, nabízí tento postup hlubší pohled do analýzy CNV v prokaryotických genomech. Segmenty s variabilním počtem kopií detekované v *de novo* sestaveném genomu mohou poskytnout větší biologickou přesnost detekce CNV a mohou také pomoci při pochopení jejich vlivu na vlastnosti a mechanismy probíhající v prokaryotech. Zásadní téma pro lidskou populaci je souvislost CNV s patogenitou a antibiotickou rezistencí bakterií. Tato práce může proto sloužit jako podklad pro další vědecké práce a přispět tak k boji s bakteriálními onemocněními.

Literatura

[1] ALKAN, Can, Bradley P. COE a Evan E. EICHLER. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* [online]. 2011, **12**(5), 363-376. DOI: 10.1038/nrg2958. ISSN 1471-0056. Dostupné z: http://www.nature.com/articles/nrg2958

[2] ALKAN, Can, Jeffrey M KIDD, Tomas MARQUES-BONET, Gozde AKSAY, Francesca ANTONACCI, Fereydoun HORMOZDIARI, Jacob O KITZMAN, Carl BAKER, Maika MALIG, Onur MUTLU, S Cenk SAHINALP, Richard A GIBBS a Evan E EICHLER. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*. 2009, **41**(10), 1061–7. ISSN 1546-1718. Dostupné z: DOI:10.1038/ng.437

[3] Assemble Reads with SPAdes - v3.13.0 | KBase App . *KBase* [online]. Dostupné z**:** https://kbase.us/applist/apps/kb_SPAdes/run_SPAdes/release?gclid=EAIaIQobChMI0LqYxubT5gIVF-h3Ch1f6g03EAAYASAAEgJgMvD_BwE

[4] Bacterial DNA – the role of plasmids — Science Learning Hub. *Science Learning Hub*. *25.3. 2014* [online]. Copyright © 2007. Dostupné z: https://www.sciencelearn.org.nz/resources/1900-bacterial-dna-the-role-of-plasmids?fbclid=IwAR1RElp5gulYl-ql-eYReI7NXLEQMatZApdhUTjLuyXw7iM4iwlKilHRXEU

[5] BENJAMINI, Yuval a Terence P. SPEED. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* [online]. 2012, **40**(10), e72-e72. DOI: 10.1093/nar/gks001. ISSN 1362-4962. Dostupné z: https://academic.oup.com/nar/article/40/10/e72/2411059

[6] BRYNILDSRUD, Ola, Lars-gustav SNIPEN a Jon BOHLIN. Genome analysis CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data [online]. 2015, 31(February), 1708–1715. Dostupné z: DOI:10.1093/bioinformatics/btv070

[7] CNV Calling Algorithm Overview — VarSeq CNV Caller Tutorial. Golden Helix Documentation [online]. Copyright © [cit. 11.05.2020]. Dostupné z: https://doc.goldenhelix.com/VarSeq/tutorials/cnv/overview.html

[8] CONSORTIUM, The 1000 Genomes Project. A global reference for human genetic variation. *Nature* [online]. 2015, 526, 68–74. Dostupné z: DOI:10.1038/nature15393

[9] Courtesy: National Human Genome Research Institute. A Brief Guide to Genomics [online]. Dostupné z: https://www.genome.gov/about-genomics/fact-sheets/A-

Brief-Guide-to-Genomicshttps://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics

[10] DARMON, Elise a David R F LEACH. Bacterial Genome Instability [online]. 2014, 78(1), 1–39. Dostupné z: DOI:10.1128/MMBR.00035-13

[11] DEAMER, David, Mark AKESON a Daniel BRANTON. Three decades of nanopore sequencing. *Nature Publishing Group* [online]. 2016, **34**(5), 518–524. ISSN 1087-0156. Dostupné z: DOI:10.1038/nbt.3423

[12] DEROUAULT, Paco, Jasmine CHAUZEIX, David RIZZO, Federica MIRESSI a Anne-sophie Lia ID. CovCopCan : An efficient tool to detect Copy Number Variation from amplicon sequencing data in inherited diseases and cancer. 2020, (Cll), 1–13.

[13] DOCAMPO, Elisa a Raquel RABIONET. A decade of structural variants: Description, history and methods to detect A decade of structural variants : description , history and methods to detect structural variation [online]. 2015, (April). Dostupné z: DOI:10.1093/bfgp/elv014

[14] GRISWOLD, Ann. (2008) Genome packaging in prokaryotes: the circular chromosome of E. coli. Nature Education 1(1):57. Dostupné z: https://www.nature.com/scitable/topicpage/genome-packaging-in-prokaryotes-the-circular-chromosome-9113/

[15] GUREVICH, Alexey, Vladislav SAVELIEV, Nikolay VYAHHI and Glenn TESLER, QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013, 29 (8): 1072-1075. DOI: 10.1093/bioinformatics/btt086.

[16] HUANG, Weichun, Leping LI, Jason R. MYERS a Gabor T. MARTH. ART: a next-generation sequencing read simulator. Bioinformatics [online]. 2012, 28(4), 593-594. DOI: 10.1093/bioinformatics/btr708. ISSN 1460-2059

[17] JUGAS,Robin, Martin VITEK, Denisa MADERANKOVA a Helena SKUTKOVA. Signal Processing Based CNV Detection in Bacterial Genomes [online]. B.m: Springer International Publishing, 2019. ISBN 9783030179380. Dostupné z: DOI:10.1007/978-3-030-17938-0

[18] KOLÍSKO, Martin. Moderní metody sekvenování DNA. Živa. 2017, 2017(3), 73–76.

[19] KUDR, Jiří, Branislav RUTTKAY-NEDECKÝ, Vojtěch ADAM a René KIZEK. Nanopóry jako moderní nástroj pro DNA sekvenování. nedatováno, 10–14.

[20] LI, Heng, B. HANDSAKER, A. WYSOKER, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [online]. 2009, 25(16), 2078-2079 [cit. 2020-05-12]. DOI: 10.1093/bioinformatics/btp352. ISSN 1367-4803.

[21] LI, Heng. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:1303.3997v1 [q-bio.GN].

[22] MARDIS, Elaine R. Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics [online]. 2008, 9(1), 387–402. ISSN 1527-8204. Dostupné z: DOI:10.1146/annurev.genom.9.081307.164359

[23] MEDVEDEV, Paul, STANCIU, Monica and BRUDNO, Michael. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*. 2009. Vol. 6, no. 11S, p. S13. DOI 10.1038/nmeth.1374.

[24] National Center for Biotechnology Information. *National Center for Biotechnology Information* [online]. Dostupné z: https://www.ncbi.nlm.nih.gov/

[25] NORD, Alex S, Ming LEE, Mary-claire KING a Tom WALSH. Accurate and exact CNV identification from targeted high-throughput sequence data [online]. 2011. Dostupné z: DOI:10.1186/1471-2164-12-184

[26] Overview of Structural Variation. *National Center for Biotechnology Information* [online]. Dostupné z: https://www.ncbi.nlm.nih.gov/dbvar/content/overview/https://www.ncbi.nlm.nih.gov/dbvar/content/overview/

[27] PERIWAL, Vinita a Vinod SCARIA. Insights into structural variations and genome rearrangements in prokaryotic genomes. Bioinformatics [online]. 2015, 31(1), 1–9. ISSN 14602059. Dostupné z: DOI:10.1093/bioinformatics/btu600

[28] PILON · GitHub. The world's leading software development platform · GitHub [online]. Copyright © 2020 GitHub, Inc. Dostupné z: https://github.com/broadinstitute/pilon/wiki

[29] PLESKAČOVÁ B. Vyhodnocení numerických reprezentací pro detekci překryvů. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2018. 57s. Vedoucí práce: Ing. Robin Jugas.

[30] ROCHA, Eduardo P C. The Organization of the Bacterial Genome [online]. 2008. Dostupné z: DOI:10.1146/annurev.genet.42.110807.091653

[31] RUSK, Nicole. Cheap third-generation sequencing. Nature Methods [online]. 2009, 6(4), 244–245. ISSN 15487091. Dostupné z: doi:10.1038/nmeth0409-244a

[32] SHENDURE, Jay a Hanlee JI. Next-generation DNA sequencing. Nature Biotechnology [online]. 2008, 26(10), 1135–1145. ISSN 10870156. Dostupné z: DOI:10.1038/nbt1486

[33] SCHADT, Eric E., Steve TURNER a Andrew KASARSKIS. A window into third-generation sequencing. Human Molecular Genetics [online]. 2010, 19(R2), 227–240. ISSN 09646906. Dostupné z: doi:10.1093/hmg/ddq416

[34] SKOVGAARD, Ole, et al. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome research*, 2011, 21.8: 1388-1393.

[35] TOLHUIS, Bas a Hans KARTEN. Validation of an ultra-fast CNV calling tool for Next Generation Sequencing data using MLPA-verified copy number alterations. 2018, 1–16. DOI: 10.1101/340505.

[36] XU, Hongtao, Xia ZHU, Zulong XU, Yue HU, Shiping BO, Tongjing XING a Kuichun ZHU. Non-invasive Analysis of Genomic Copy Number Variation in Patients with Hepatocellular Carcinoma by Next Generation DNA Sequencing. Journal of Cancer [online]. 2015, 6(3), 247-253. DOI: 10.7150/jca.10747. ISSN 1837-9664. Dostupné z: http://www.jcancer.org/v06p0247.htm

[37] YANG, Tie-lin, Yan GUO a Christopher J PAPASIAN. Copy Number Variation. 2013, 123–132.

[38] YAO, Ruen, Tingting YU, Yanrong QING, Jian WANG a Yiping SHEN. Evaluation of copy number variant detection from panel-based next-generation sequencing data. Molecular Genetics & Genomic Medicine [online]. 2019, 7(1) [cit. 2020-05-14]. DOI: 10.1002/mgg3.513. ISSN 23249269. Dostupné z: http://doi.wiley.com/10.1002/mgg3.513

[39] 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. 2013, 14(Suppl 11).

Seznam symbolů, veličin a zkratek

AS - Sequence assembly, sestavení sekvence

BAC - umělý bakteriální chromozom

BAM - Binary Alignment Map

bp - base pair – pár bází

BWA - Burrows-Wheeler Aligner

CGH - komparativní genomová hybridizace

CNV - copy number variants

DNA - deoxyribonukleová kyselina

DSB - double – strand breaks - dvouvláknové zlomy

HGP - Human genome project

HGT - horizontální přenos genů

MEI - inzerce mobilních elementů

NCBI - The National Center for Biotechnology Information

NGS - Next Generation Sequencing – Sekvenování druhé generace

PEM - paired-end mapping – párové mapování

RD - read depth – hloubka čtení

RNA - ribonukleová kyselina

SAM - Sequence Alignment Map

SNP - single-nucleotide polymorphism - jednonukleotidový polymorfismus

SR - split-read rozdělené čtení

SV - strukturní variabilita