ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Katedra biomedicínské informatiky

Detekce CNV v datech z celoexomového sekvenování pacientů s neurogenetickým onemocněním

The detection of CNV in Whole-exome sequencing data of patients with neurogenetic disease

Bakalářská práce

Studijní program:	Biomedicínská a klinická technika
Studijní obor:	Biomedicínská informatika

Autor bakalářské práce:	Jaroslav Iha
Vedoucí bakalářské práce:	Ing. et Ing. David Staněk, Ph.D.

Kladno 2020

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení:	**Iha**	Jméno:	**Jaroslav**	Osobní číslo: **474300**
Fakulta:	**Fakulta biomedicínského inženýrství**			
Garantující katedra:	**Katedra biomedicínské informatiky**			
Studijní program:	**Biomedicínská a klinická technika**			
Studijní obor:	**Biomedicínská informatika**			

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Detekce CNV v datech z celoexomového sekvenování pacientů s neurogenetickým onemocněním

Název bakalářské práce anglicky:

The detection of CNV in Whole-exome sequencing data of patients with neurogenetic disease

Pokyny pro vypracování:

Cílem práce je otestovat a zavést do praxe metodiku pro detekci variability počtu kopií segmentů DNA (CNV, Copy number variation). Při řešení spolupracujte s pracovištěm DNA laboratoře KDN 2.LF a FN Motol. Po vzájemné konzultaci definujte kohortu pacientů a postup zpracování řešení. V rámci bakalářské práce otestujte současnou metodiku pro detekci germinálních variant u WES dat a aplikujte jí na data pacinetů shromážděná v DNA laboratoři.

Seznam doporučené literatury:

[1] V. Buffalo, Bioinformatics data skills: Reproducible and robust research with open source tools, ed. 1, O'Reilly Media, Inc., 2015, ISBN 1449367372
[2] J. Pevsner, Bioinformatics and functional genomics, ed. 3, Wiley-Blackwell, 2015, ISBN 1118581784

Jméno a příjmení vedoucí(ho) bakalářské práce:

Ing. David Staněk

Jméno a příjmení konzultanta(ky) bakalářské práce:

Mgr. Radim Krupička, Ph.D.

Datum zadání bakalářské práce: **17.02.2020**
Platnost zadání bakalářské práce: **20.09.2021**

doc. Ing. Zoltán Szabó Ph.D.	prof. MUDr. Ivan Dylevský, DrSc.
podpis vedoucí(ho) katedry	podpis děkana(ky)

PROHLÁŠENÍ

Prohlašuji, že jsem bakalářskou práci s názvem *„Detekce CNV v datech z celoexomového sekvenování pacientů s neurogenetickým onemocněním“* vypracoval samostatně a použil k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k diplomové práci.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu § 60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů.

V Kladně dne ……. …..……..……...

Jaroslav Iha

PODĚKOVÁNÍ

Rád bych tímto poděkoval Ing. et Ing. Davidu Staňkovi, Ph.D., za cenné rady, výstižné připomínky a ochotu po celou dobu vedení mé bakalářské práce. Dále bych rád poděkoval mé rodině a přítelkyni za nepřetržitou podporu během celého mého studia.

ABSTRAKT

Detekce CNV v datech z celoexomového sekvenování pacientů s neurogenetickým onemocněním

Přestože je variabilita počtu kopií segmentů DNA (*Copy number variation, CNV*) genetickou příčinou vzniku mnoha neurogenetických onemocnění, její následné detekování z dat masivně paralelního sekvenování zůstává složitou výzvou. Cílem této práce bylo implementovat a otestovat metodiku pro detekci CNV v datech z celoexomového sekvenování. Pro řešení práce byl zvolen přístup dle doporučení GATK v kombinaci s cloudovou aplikací Terra App. Otestování bylo provedeno na skupině pacientů s dědičnou hluchotou. Analýzou dat byly nalezeny CNV varianty, které by mohly pomoci vysvětlit příčinu onemocnění vyšetřovaných jedinců.

Klíčová slova

NGS, CNV, WES, neurogenetická onemocnění

ABSTRACT

The detection of CNV in Whole-exome sequencing data of patients with neurogenetic disease

Although copy number variation (CNV) is a genetic cause of many neurogenetic diseases, its subsequent detection from massively parallel sequencing data remains a complex challenge. The aim of this work was to implement and test a methodology for the detection of CNV in data from whole exome sequencing. For the solution was chosen the approach according to the GATK recommendations in combination with the Terra App cloud application. Testing was performed on a group of patients with hereditary deafness. Analysis of the data revealed CNV variants that could help explain the cause of the disease in the subjects.

Keywords

NGS, CNV, WES, neurogenetic disorders

Obsah

Seznam zkratek

Seznam zkratek

Zkratka	Význam
AD	Autosomálně dominantní
AR	Autosomálně recesivní
BAM	*Binary alignment map*
BED	*Browser extensible data*
BQSR	*Base quality scores recalibration*
CIGAR	*Compact idiosyncratic gapped alignment report*
CNV	Variabilita počtu kopií segmentů DNA (*Copy number variation*)
DNA	Deoxyribonukleotidová kyselina (*Deoxyribonucleic acid*)
DSB	Dvouřetězcový zlom (*Double-strand DNA break*)
FoSTeS	Blokování replikační vidlice a změna templátového vlákna (*Fork stalling and template switching*)
GATK	*Genome Analysis Toolkit*
GD	Gonosomálně dominantní
GR	Gonosomálně recesivní
NAHR	Nealelická homologní rekombinace (*Nonallelic homologous recombination*)
NGS	Sekvenování nové generace (Next generation sequencing)
NHEJ	Spojování nehomologických konců DNA řetězců (*Nonhomologous end joining*)
MPS	Masivně paralelní sekvenování (*Massive parallel sequencing*)
PCR	Polymerázová řetězová reakce (*Polymerase chain reaction*)
SAM	*Sequence alignment map*
SNP	Jednonukleotidový polymorfismus (*Single nucleotide polymorphism*)
SNV	Jednonukleotidové varianty (*Single nucleotide variants*)
VCF	*Variant calling file*
WES	Celoexomové sekvenování (*Whole exome sequencing*)
WGS	Celogenomové sekvenování (*Whole genome sequencing*)

1 Úvod

Změny v DNA se vyskytují v rámci celé populace, je tedy důležité zabývat se jejich sledováním. Díky těmto změnám lze objasnit příčiny či původ vzniku závažných onemocnění. S rozvojem technologií a nových metod molekulární genetiky využívaných pro analýzu nukleových kyselin, přišly i nové poznatky ohledně strukturní variability lidské DNA, konkrétněji jde o variabilitu počtu kopií segmentů DNA (*Copy number variation, CNV*).

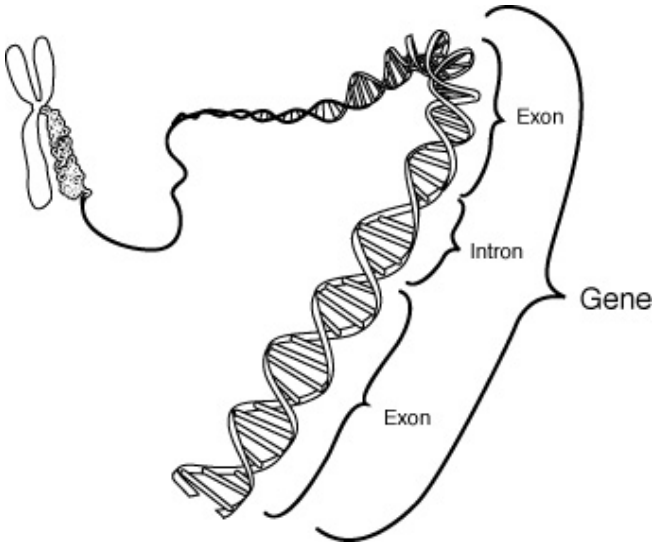
CNV je v rámci genomu častým a obvyklým jevem a udává se, že může mít na svědomí přibližně 12 % variability lidského genomu, což zdůrazňuje jeho význam v genetické rozmanitosti a evoluci. [1]

Tato práce se zabývá otestováním metodiky pro detekci CNV v datech z celoexomového sekvenování a jejím následným aplikováním na datech pacientů s neurogenetickým onemocněním.

1.1 Struktura genu

Gen je základní funkční a fyzickou jednotkou dědičnosti přecházející z rodičů na potomky. Geny jsou části DNA uspořádány jeden po druhém na strukturách nazývaných chromozomy. Většina genů obsahuje informace pro výrobu specifického proteinu. Lidé mají na chromozomech uspořádaných přibližně 20 000 genů. [2]

Exony jsou úseky DNA, které nesou kódující informaci genu. Introny jsou úseky nekódujících sekvencí oddělující jednotlivé exony (Obr. 1.1). Introny i exony mají velmi proměnlivou délku i počet, avšak zpravidla jsou introny mnohem delší než exony. [3]



Obrázek 1.1: Struktura genu. Zdroj: [4]

1.2 Genetický kód

Prostřednictvím čtyř bází (A – adenin, G – guanin, C – cytosin, U – uracil) obsažených v nukleových kyselinách se aminokyseliny kódují (Obr. 1.2) a řadí do polypeptidického řetězce. Trojice sousedních nukleotidů vždy rozhoduje o zařazení konkrétní aminokyseliny. Tato trojice se obecně nazývá triplet.

Pomocí 61 tripletů je možné kódovat 20 aminokyselin, z čehož plyne degenerace (redundance) genetického kódu. Znamená to, že jednu aminokyselinu lze kódovat více triplety. [5]

	U		**C**		**A**		**G**	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	stop	UGA	stop
	UUG	Leu	UCG	Ser	UAG	stop	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Obrázek 1.2: Genetický kód a kódování aminokyselin. Zdroj: [6]

1.3 Variabilita DNA

V celém genomu dochází ke změnám na úrovni sekvence DNA, lze je rozdělit do několika kategorií. Jednou jsou změny, při kterých se počet nukleotidů nezmění (tj. neovlivňují obsah DNA). Relativně často je zde například jeden nukleotid nahrazen jiným nukleotidem. Tyto typy změn mohou být bez fenotypového projevu, nemusí mít škodlivý vliv – synonymní varianty, nebo naopak mohou mít škodlivý vliv a způsobovat onemocnění – missense varianty.[7]

V další kategorii DNA změn dochází ke změně počtu kopií segmentů DNA, která se pohybuje v rozmezí od 100 nukleotidů až po velké záměny, pokrývající i celé geny. O tom, jestli jsou škodlivé, rozhoduje několik faktorů (umístění, rozsah atd.). V některých případech mohou vést k vývojovým syndromům a výjimečně ke spontánnímu potratu. [7]

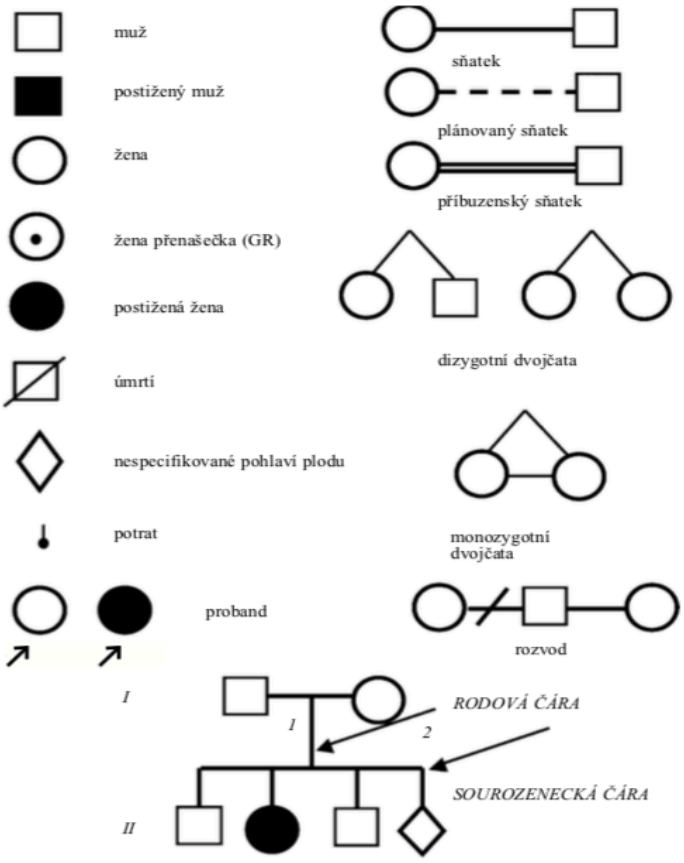
Celkově zahrnují nejčastější změny DNA pouze jediný nebo velmi malý počet nukleotidů. Takové změny, pokud leží mimo kódující oblast genu často nemají žádný efekt na fenotyp a v takovém případě jsou považovány za neutrální mutace. Stává se tak proto, že DNA toleruje malé změny v sekvenci bez zjevného účinku. Naopak změna i jediného nukleotidu uvnitř kódující oblasti genu může mít vážný dopad na organismus. [7] [8]

1.4 Genetická variabilita jako příčina onemocnění

Nové mutace se rozdělují do dvou variant podle místa vzniku – v zárodečné buňce (*germline*) nebo somatické (*somatic*). Zárodečné mutace mohou způsobit dědičné onemocnění, jelikož je u nich možné přenášení do dalších generací. Somatické mutace se na potomstvo nepřenášejí. Při vzniku v jedné buňce často nemají žádný negativní vliv na organismus díky kontinuálnímu nahrazování buněk. Občas ale mohou vést k nádorovému bujení, které vede k ohrožení života. [3]

1.5 Monogenně podmíněná onemocnění

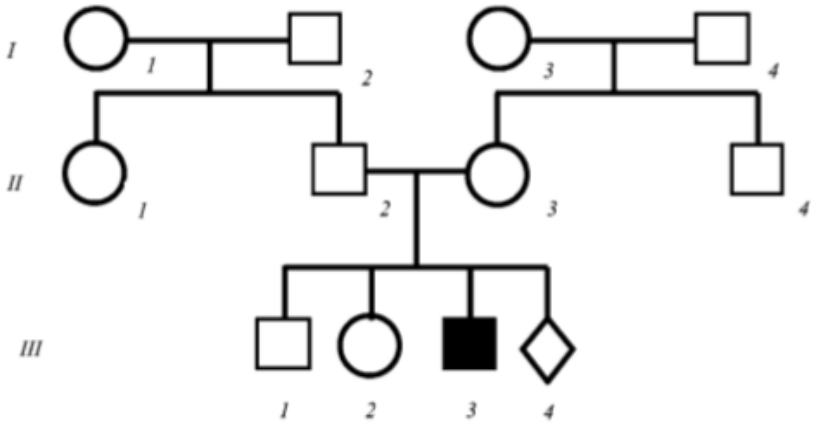
Monogenně děděná onemocnění vznikají mutací pouze v jediném genu a vykazují charakteristické vzorce dědičnosti. [9]



Obrázek 1.3: Vybrané symboly používané při sestavování rodokmenu. Zdroj: [5]

1.5.1 Autosomálně recesivní (AR) onemocnění

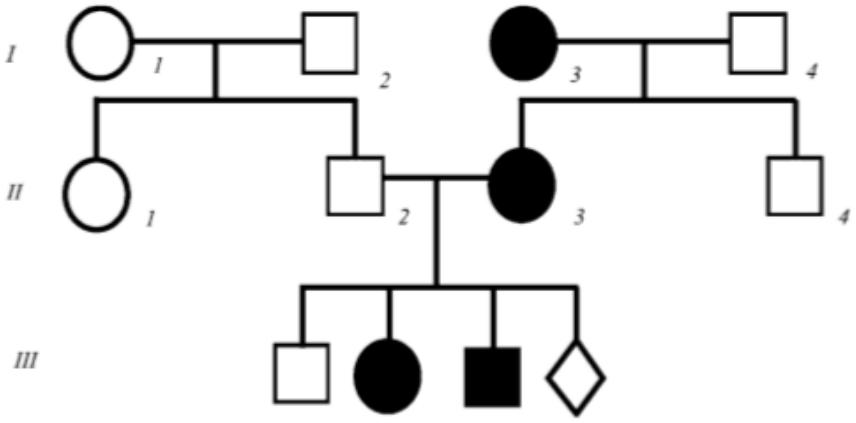
Vyskytuje se stejně často u mužů i žen. Postiženým bývá potomek zdravých rodičů. V případě vzácných AR onemocnění jde obvykle o příbuzenský sňatek rodičů postiženého. U jeho sourozenců je riziko onemocnění 25 %. U potomků postiženého probanda záleží riziko na četnosti choroby v populaci (malé riziko u vzácných onemocněních). Za typické AR onemocnění se považuje cystická fibróza (porucha činnosti žláz s vnější sekrecí), fenylketonurie (porucha metabolismu aminokyseliny fenylalaninu), srpkovitá anémie (porucha struktury hemoglobinu) a další. [5] [10]



Obrázek 1.4: Autosomálně recesivní dědičnost. Zdroj: [5]

1.5.2 Autosomálně dominantní (AD) onemocnění

Pravděpodobnost výskytu je stejná u obou pohlaví. Typické pro onemocnění je, že se projevuje většinou v každé generaci. Pro děti postiženého jedince platí riziko 50 % při sňatku se zdravým partnerem a pro děti zdravého jedince je riziko onemocnění nulové. Typickými AD chorobami jsou polydaktylie (typ víceprstosti), Huntingtonova chorea (neurodegenerativní onemocnění), polycystická choroba ledvin (nejčastější AD onemocnění). [5] [10]

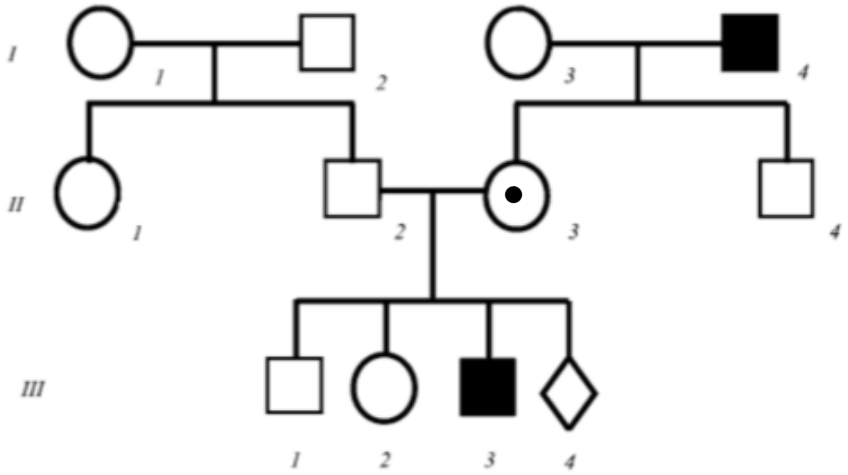


Obrázek 1.5: Autosomálně dominantní dědičnost. Zdroj: [5]

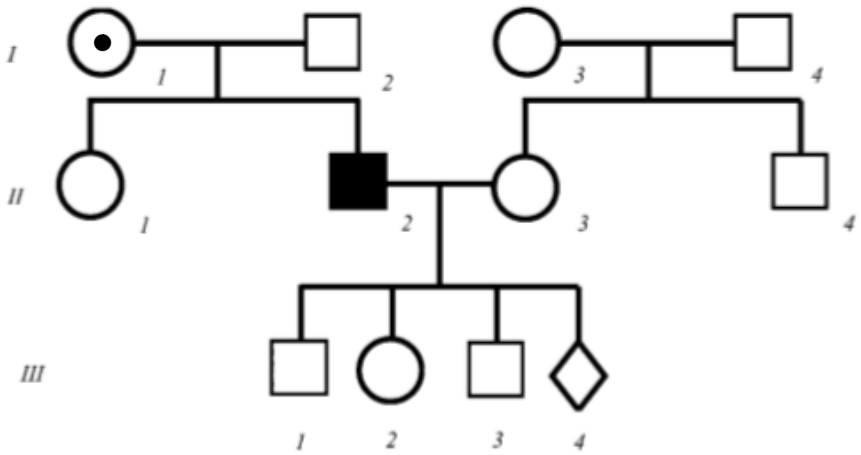
1.5.3 Gonosomálně recesivní (GR) onemocnění

Postiženi bývají téměř pouze muži. Ženy jsou zdravé přenašečky a onemocnění se u nich projevuje málokdy. GR choroby se často projevují „ob generaci“. To znamená, že všechny dcery postiženého muže jsou zdravé přenašečky a jeho synové jsou všichni zdraví. U zdravých přenašeček je potom riziko postižení u syna 50 %. Charakteristickými GR chorobami jsou daltonismus (barvoslepost), hemofilie A, hemofilie B (dědičná krvácivost), Duchennova muskulární dystrofie (porucha funkce svalů). [5] [10]

Na obrázcích (Obr. 1.6 a Obr. 1.7) jsou příklady dvou rozdílných výskytů gonosomálně recesivního onemocnění v rodokmenu.



Obrázek 1.6: Gonosomálně recesivní dědičnost. Žena II/3 - přenašečka.
Zdroj: [5]

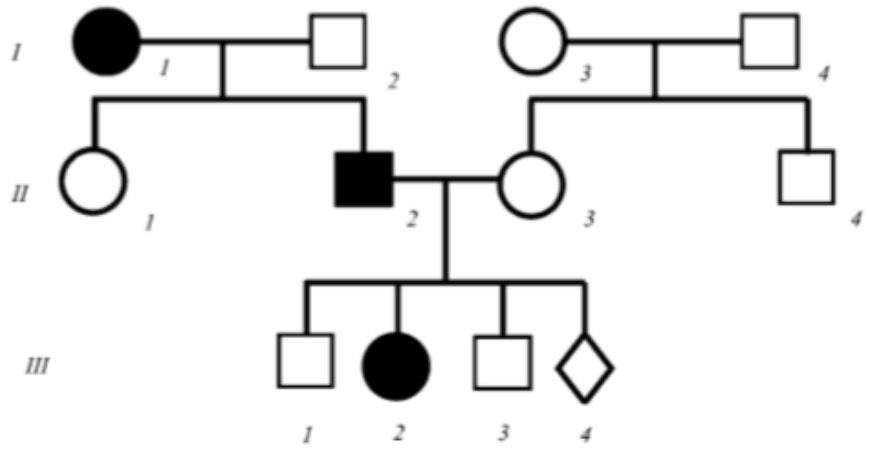


Obrázek 1.7: Gonosomálně recesivní dědičnost. Žena I/1 - přenašečka.
Zdroj: [5]

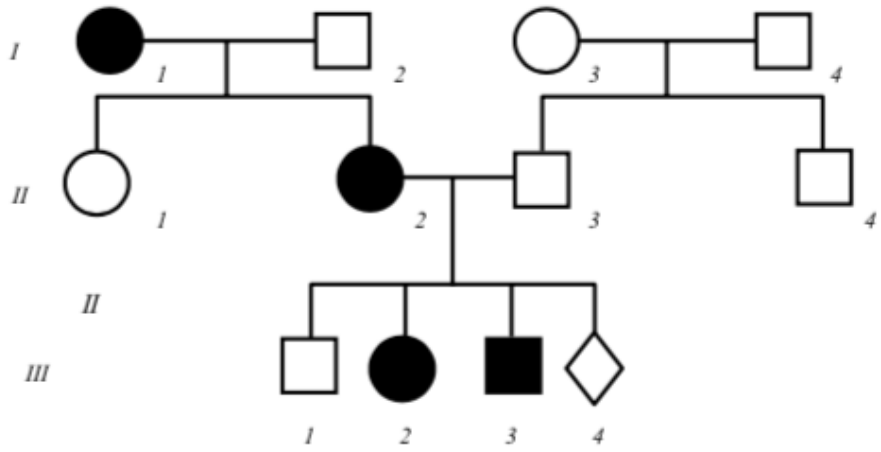
1.5.4 Gonosomálně dominantní (GD) onemocnění

Riziko postižení je pro ženy dvojnásobné než pro muže. Dcery postiženého muže jsou také všechny postiženy, naopak synové jsou všichni zdraví. V případě postižené ženy je riziko postižení pro její děti 50 % bez ohledu na pohlaví. Příklad typické GD choroby je D-vitamín rezistentní rachitis (křivice – postižení kostry). [5] [10]

Obrázky (Obr. 1.8 a Obr. 1.9) vyobrazují dva odlišné rodokmeny s přítomností gonosomálně dominantního onemocnění.



Obrázek 1.8: Gonosomálně dominantní dědičnost. Zdroj: [5]



Obrázek 1.9: Gonosomálně dominantní dědičnost. Zdroj: [5]

1.5.5 *De novo* varianta

De novo varianty se liší od zděděných tím, že ani jeden z rodičů postiženého jedince není nositelem varianty. *De novo* varianty vznikají v zárodečné buňce jednoho z rodičů nebo výjimečně ve vyvíjejícím se embryu. [11] Klinicky jsou *de novo* varianty zajímavé,

protože když se potvrdí *de novo* varianta u jedince, tak je velice malá pravděpodobnost, že se vyskytne totožná varianta u dalšího potomka stejných rodičů.

1.6 DNA varianty a polymorfismy

Při mutacích vznikají alternativní formy DNA, které jsou obecně známy jako DNA varianty (*DNA variants*). Avšak pokud je v populaci běžný výskyt větší než 1 %, je varianta popsána jako polymorfismus. Hodnota 1 % původně byla navržena proto, aby vylučovala opakující se mutaci. Zbylé varianty jsou často popisovány jako vzácné varianty (*rare variants*). [7]

1.6.1 Jednonukleotidové polymorfismy (SNPs)

Jako SNP (*Single nucleotide polymorphism*) se označuje nejmenší možná genetická změna, vztahující se k jednomu nukleotidu. Lépe řečeno, když je jeden nukleotid nahrazen jakýmkoliv jiným nukleotidem a zároveň se tato varianta vyskytuje ve významné části populace (četnost vyšší než 1 %), jedná se o SNP. Většina takových záměn je neutrální, jelikož leží mimo kódující sekvence. Díky degeneraci genetického kódu jsou zpravidla bez vlivu i ty ležící uvnitř exonů. Odhaduje se, že v lidském genomu se SNP průměrně nachází jednou za 1000 nukleotidů. [10]

Pokud je četnost výskytu těchto variant v populaci naopak menší než 1 %, jsou kvalifikovány jako SNV (*Single nucleotide variants*).

1.6.2 Inserce delece, variabilita počtu kopií segmentů DNA (CNV)

Variabilitu počtu kopií (CNV) lze formulovat jako typ strukturní variability zahrnující změny v počtu kopií specifických úseků genomu, ty mohou být buď deletovány nebo duplikovány. Tyto duplikace a delece zahrnují poměrně velké úseky DNA, které se ale mohou lišit co do prevalence, tak i velikostně. Variabilita v počtu kopií úseků DNA může být zděděná, ale výjimkou nejsou ani CNV vznikající *de novo*. [12] [13] Jako indel (zkr. pro inserce delece) se nazývají krátké úseky DNA (obvykle menší než 50 nukleotidů), které byly insertovány nebo deletovány. Tato práce se zaměřuje pouze na detekci variability počtu kopií (CNV), jelikož pro detekci indelů se používají zcela odlišné postupy.

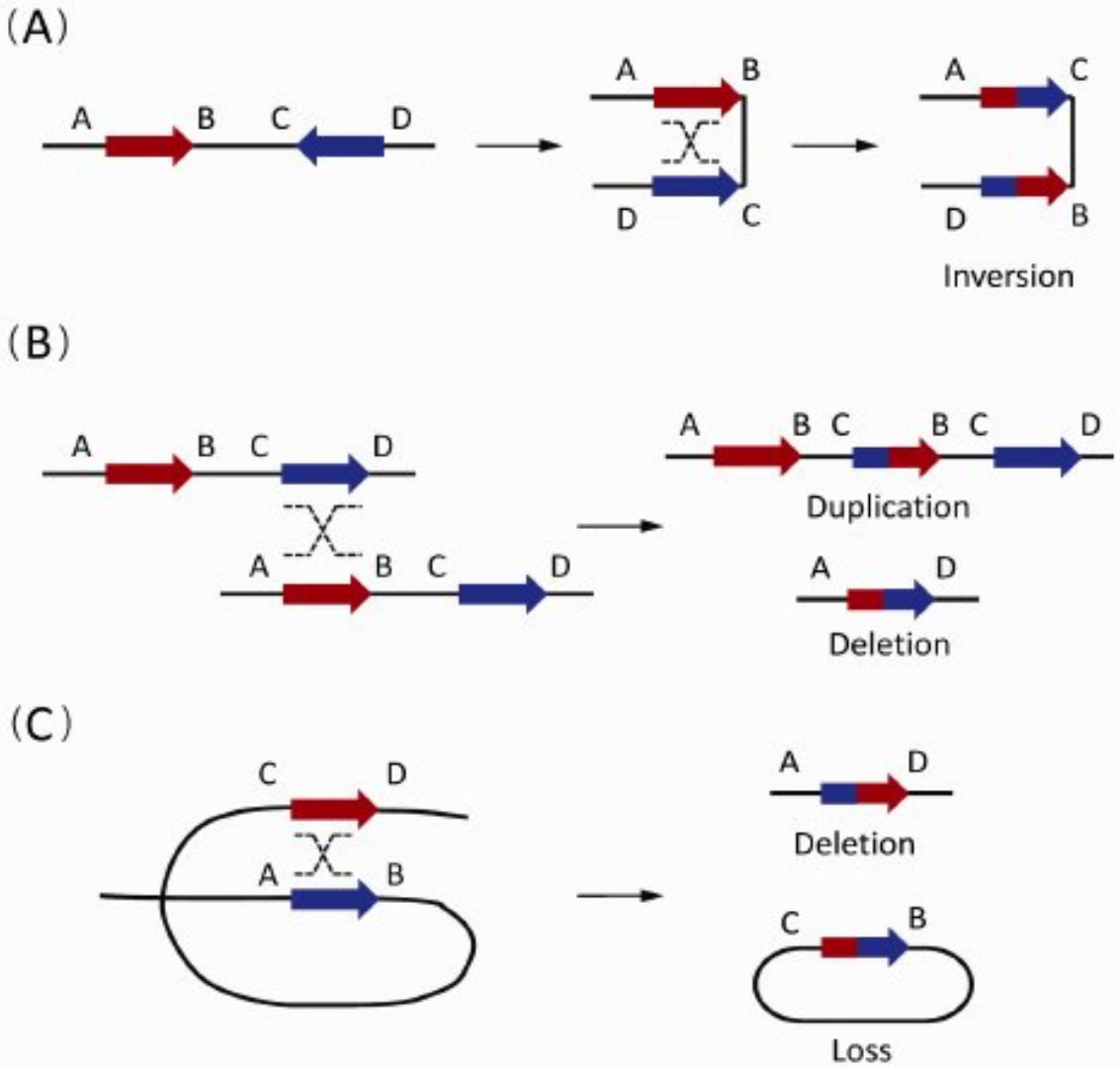
1.7 Mechanismy vzniku CNV

Dva hlavní mechanismy: NAHR a NHEJ, jsou příčinou změn v uspořádání lidského genomu a jsou v současnosti popisovány jako nejčastější příčina vzniku CNV. [14]

1.7.1 Nealelická homologní rekombinace (NAHR)

Nealelická homologní rekombinace (*Nonallelic homologous recombination, NAHR*) je způsobena zarovnáním a následným crossing-overem mezi dvěma nealelickými sekvencemi DNA (Obr. 1.10), které mají vysokou vzájemnou sekvenční podobnost. [14] Pomocí NAHR vznikají převážně chromozomové změny se shodnou délkou a zlomy na soustředěných místech chromozomu. Naopak jinými mechanismy vznikají častěji změny s proměnlivou délkou a náhodnými zlomy na chromozomu. [15] NAHR je považován za hlavní mechanismus vzniku patogenních CNV vyskytující se v meoticky i mitoticky dělících se buňkách. [16]

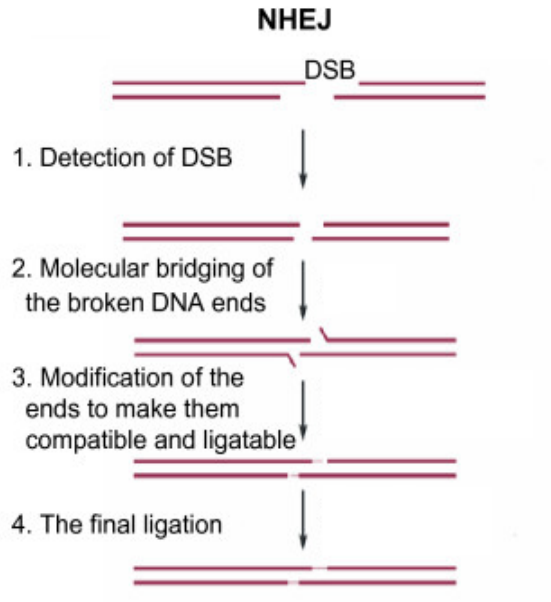
Mechanismem NAHR bývá nejčastěji postižen gen *PMP22*, u kterého dochází k duplikaci na jednom a deleci na druhém chromozomu. Projevem potom bývá dědičná neuropatie. [17]



Obrázek 1.10: Nealelická homologní rekombinace (NAHR). Výsledky této rekombinace mohou generovat: (A) Inverzi, která se považuje za neutrální strukturální variantu vzhledem k variabilitě počtu kopií segmentů DNA. (B) Duplikaci na jednom chromozomu, zatímco se kopie odstraní z druhého. (C) Deleci a prstencovitý úsek DNA, který bude ztracen v následujícím buněčném dělení. Zdroj: [18]

1.7.2 Spojování nehomologických konců řetězců DNA (NHEJ)

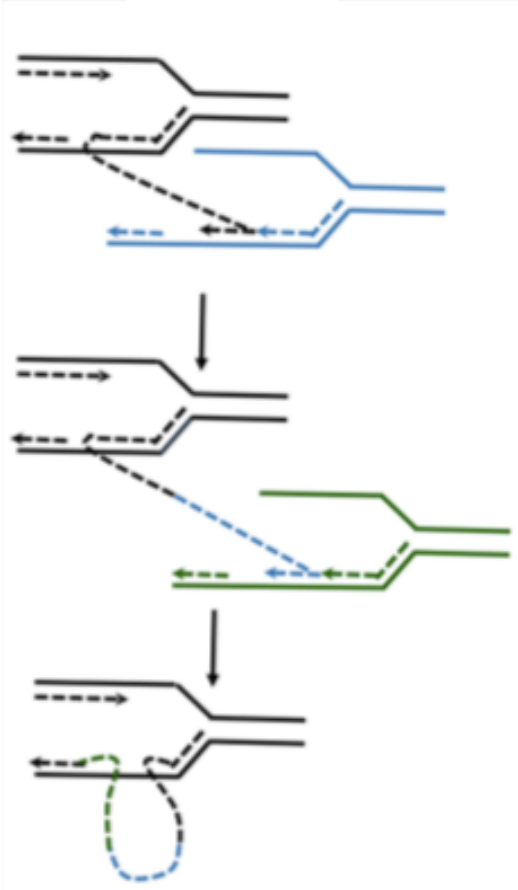
Mechanismus založený na spojování nehomologických konců DNA řetězců (*Nonhomologous end joining, NHEJ*) je jedním ze dvou hlavních mechanismů používaných eukaryotickými buňkami k opravě dvouřetězcových zlomů (*Double-strand DNA break, DSB)* a byl popsán v organismech od bakterií po savce. [19] [20] DSB v těle vznikají působením např. ionizujícího záření nebo reaktivních forem kyslíku. [21] NHEJ (Obr. 1.11) je rovněž momentálně považován za základní mechanismus spojující translokované chromozomy u rakoviny. [22]



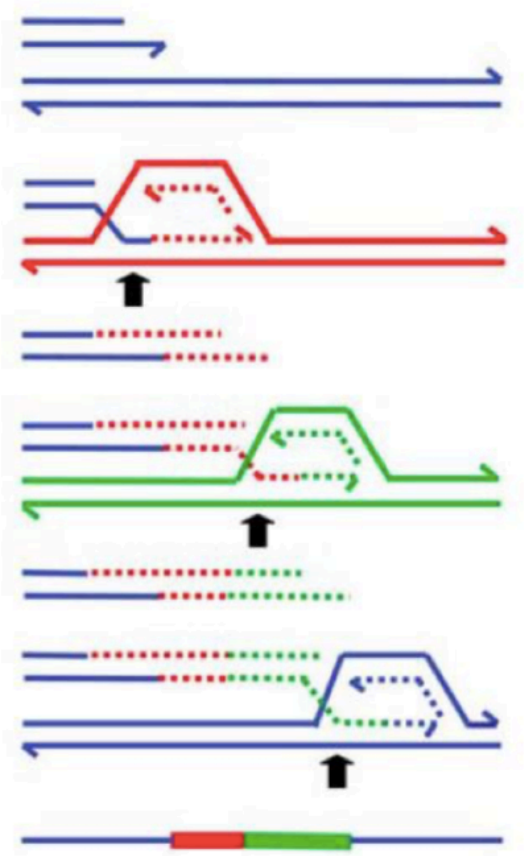
Obrázek 1.11: Spojování nehomologických konců řetězců DNA (NHEJ) se skládá ze čtyř kroků: 1. detekce DSB, 2. molekulární přemostění obou zlomených konců DNA, 3. modifikace konců tak, aby byly kompatibilní a ligovatelné, 4. konečná ligace. [20] Tento proces definuje dvě významné vlastnosti NHEJ. Zaprvé, pro NHEJ nejsou vyžadovány DNA segmenty s vysokou sekvenční podobností. Zadruhé, NHEJ zanechává „informační jizvu“ v místě opětovného spojení v podobě přidání nebo odebrání několika nukleotidů. [23] Zdroj: [24]

1.7.3 Další mechanismy vzniku CNV

Za další důležité, ale dosud nedostatečně prostudované principy vzniku lidských genomických změn jsou považovány mechanismy na bázi replikace – blokování replikační vidlice a změna templátového vlákna (*Fork stalling and template switching, FoSTeS*) a replikace vyvolaná zlomy mikrohomologických segmentů DNA (*microhomology-mediated break induced replication, MMBIR*). [25] Genomické změny způsobené FoSTeS (Obr. 1.12) nebo MMBIR (Obr. 1.13) mohou vést ke zdvojení či trojnásobení jednotlivého genu nebo dokonce ke změně v uspořádání jednotlivých exonů, což naznačuje důležitou roli při vývoji genu i celého genomu. [26]



Obrázek 1.12: Mechanismus FoSTeS. V replikační vidlici dojde ke zlomu řetězce a jeho uvolnění. Uvolněný řetězec se poté začlení do jiné replikační vidlice. Tato změna řetězce se může několikrát opakovat, dokud nedojde zase k replikaci na původní vidlici. Zdroj: [25]



Obrázek 1.13: Mechanismus MMBIR. Při replikaci se přechází do jiných pozic genomu, dokud nedojde k návratu na původní vlákno chromozomu. Výsledný produkt obsahuje úseky z jiných regionů genomu. Zdroj: [26]

1.8 Masivně paralelní sekvenování (MPS)

Sekvenovat lidský genom pomocí původních sekvenačních metod by bylo, kvůli jeho obrovské velikosti, velice nákladné a časově náročné. Aby bylo toto omezení překonáno byly vyvinuty sekvenační metody nové generace (*Next generation sequencing, NGS*), které umožňují rychlé sekvenování velkých segmentů nukleových kyselin i celých genomů. [27]

Masivně paralelní sekvenování (*Massive parallel sequencing, MPS*) využívá technologie, které jsou schopny paralelně zpracovávat více sekvencí DNA najednou, čímž získáváme veliké množství dat v relativně krátkém časovém úseku. [27] Nové platformy jsou schopny vygenerovat až 120 Gb dat za 27 hodin, což umožňuje osekvenovat celý genom během 24 hodin. [28]

Princip, kterým se takto velký objem dat generuje, je vesměs velice podobný. Fragmentovaná DNA se sekvenuje dvakrát (jednou v opačném směru) po krátkých úsecích o délce přibližně 150 bází. Jednotlivé přečtené fragmenty jsou poté řazeny za sebe a zarovnávány s referenčním genomem. Díky sekvenaci v obou směrech je metoda přesnější v detekování inzercí a delecí. Naopak nevýhodou je její větší chybovost. [28]

1.8.1 Platforma Illumina

Platforma Illumina [29] je momentálně nejrozšířenějším MPS systémem, který byl uveden na trh v roce 2006 firmou Solexa, později byl odkoupen společností Illumina. Princip metody sekvenování využívaný platformou Illumina je založen na sekvenaci syntézou. [30]

2 Cíle práce

- Implementovat bioinformatické postupy (*workflows*) pro detekci CNV dle doporučení GATK
- Vytvořit základní CNV modely v „*cohort*“ režimu
  - Testovací model z WES dat alespoň 10 pacientů
  - Finální model z WES dat alespoň 40 pacientů
- Otestovat oba modely a porovnat jejich výsledky na skupině alespoň 10 pacientů s dědičnou hluchotou

3 Metody

3.1 Datové formáty

Pro následné zpracování dat, získaných ze sekvenátorů, je důležité zvolit jejich správnou podobu. Měl by jí totiž rozumět jak počítač, tak i expert, který vyhodnocuje výsledek.

V každé části analýzy se používá jiný vhodný datový formát. Pro zpracování genomu počítačem a porovnávání s referenční sekvencí je využívána podoba dat, která je lépe zpracovatelná počítačem (formáty FASTQ a BAM). Naopak při reprezentaci výsledku je potřeba mít data ve formátu čitelnějším pro člověka (VCF).

3.1.1 FASTQ

Soubory ve formátu FASTQ jsou standardem pro výstup ze sekvenátorů. Obvykle se pracuje se dvěma soubory pro každý vzorek. První je pro dopřednou (*forward*) a druhý pro zpětnou (*revers*) sekvenci.

Formát FASTQ je v dnešní době velmi využívaný, jelikož dokáže k informaci o přečtené sekvenci ukládat číselné skóre kvality čtení spojené s každým nukleotidem. Tato přidaná informace nám pomáhá zjišťovat s jakou pravděpodobností je čtení na daném úseku chybné. [31]

V datovém formátu jsou definovány čtyři typy řádků (Obr. 3.1). Nejprve řádek začínající znakem „@“, jenž často obsahuje pouze identifikátor záznamu. Může ale i zahrnovat libovolné komentáře, anotace či délku sekvence. Druhým typem řádku je přečtená sekvence. Zatřetí, začíná sekce znakem „+“, který značí ukončení sekvenčních řádků. Většinou se sestává pouze z tohoto jednoho znaku, čímž se významně zmenšuje velikost souboru, ale může obsahovat alternativní identifikátory. Poslední typ musí být stejně dlouhý jako sekvence a vyjadřuje skóre kvality čtení každého nukleotidu. Číselné hodnoty jsou zde kvůli kompaktnosti a větší přehlednosti dále kódovány pouze jedním znakem. [31]

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Obrázek 3.1: Příklad datového formátu FASTQ. Zdroj: [31]

3.1.2 SAM/BAM – Sequence alignment map / Binary alignment map

Soubory formátu SAM obsahují přečtenou sekvenci zarovnanou s referenční sekvencí. BAM soubory jsou v praxi více využívané, protože jsou téměř totožné jako soubory SAM, ale na rozdíl od nich jsou komprimované a indexované. Díky tomu jsou soubory menší a nezabírají tolik místa v paměti počítače. Indexování v praxi znamená to, že je soubor rozčleněný do menších bloků. Poté je vytvořen indexový soubor, který odkazuje na každý blok ve velkém souboru. Kvůli tomu je ale zpracování BAM souborů pomalejší. [32]

Každý soubor se skládá ze dvou částí, a to hlavičky a zarovnání sekvence na referenci (Obr. 3.2). Řádky v hlavičce začínají znakem „@“ a obsahují informace o formátu/verzi souboru, referenční sekvenci (název, délka atd.), identifikaci vzorku (ID pacienta, sekvenační platforma) a použitém algoritmu (parametry). Sekce zarovnání sekvence, pak obsahuje nejméně 11 oddílů včetně přečtené sekvence, zarovnání na referenční sekvenci, kvality čtení a řetězce CIGAR (*Compact idiosyncratic gapped alignment report*). CIGAR popisuje operace, které algoritmus provedl a definuje tři označení: (M) kolik nukleotidů bylo stejných s referencí, (D) – kolik nukleotidů bylo oproti referenci deletováno nebo (I) insertováno. [32]

```
@HD VN:1.3  SO:coordinate
@SQ SN:ref  LN:45
@SQ SN:ref2 LN:40
r001    163 ref 7   30  8M4I4M1D3M  =   37  39  TTAGATAAAGAGGATACTG *   XX:B:S,12561,2,20,112
r002    0   ref 9   30  1D2I6M1D1I1D1I4M2I  *   0   0   AAAAGATAAGGGATAAA    *
r003    0   ref 9   30  5D6M    *   0   0   AGCTAA  *
r004    0   ref 16  30  6M14D1I5M   *   0   0   ATAGCTCTCAGC    *
r003    16  ref 29  30  6I5M    *   0   0   TAGGC   *
r001    83  ref 37  30  9M  =   7   -39 CAGCGCCAT   *
```

Obrázek 3.2: Příklad datového formátu SAM/BAM. Zdroj: [32]

3.1.3 VCF – Variant calling file

Soubor formátu VCF bývá strukturován do tabulky a používá se jako výstup bioinformatické analýzy. Obsahuje 3 části: meta-informace, hlavičku a řádky, z nichž každý obsahuje informace o dané variantě (Obr. 3.3). Jedná se tak o seznam všech nalezených variant/odchylek od referenční sekvence. Díky struktuře lze poměrně snadno filtrovat varianty podle údajů v tabulce. [33]

Každý řádek v první části (meta-informace) musí začínat znakem „##“. Následují informace o verzi VCF, předešlém filtrování, referenční sekvenci atd. Dále je doporučeno, aby meta-informace obsahovaly údaje o každém parametru vyskytujícím se ve sloupcích FILTER, INFO a FORMAT. Na každém řádku by měly být údaje pouze o jednom parametru (identifikátor, datový typ a popis). Hlavička se skládá z názvů sloupců, z nichž prvních osm musí být v každém souboru:

1. #CHROM – označení chromozomu z referenční sekvence
2. POS – pozice varianty na chromozomu
3. ID – jedinečný identifikátor, pokud je k dispozici
4. REF – báze podle zvolené referenční sekvence
5. ALT – alternativní báze z přečtené sekvence
6. QUAL – skóre kvality přečtení báze
7. FILTER – hodnota „PASS“ jestliže varianta prošla přes všechny filtry, pokud varianta neprošla nějakým filtrem, nastaví se hodnota na jeho název
8. INFO – dodatečné informace o variantě, všechny parametry zde použité musí být definovány v meta-informacích

V dalších sloupcích může formát také obsahovat genotypové informace pro každý vzorek. [33]

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Obrázek 3.3: Příklad datového formátu VCF. Zdroj: [33]

GVCF – Genomic VCF

Základní specifikace je totožná jako u VCF. Rozdíl je v tom, že obsahuje informace o všech sekvenovaných úsecích včetně těch, kde nebyla nalezena varianta. Formát se využívá při vyvolávání variant u více vyšetřovaných jedinců. [34]

3.1.4 BED – Browser Extensible Data

Formát BED je používaný ke stanovení určitých úseků chromozomu, na kterých má probíhat analýza. Formát je strukturován do sloupců (Obr. 3.4), z nichž jsou povinné minimálně 3:

1. chrom – označení chromozomu
2. chromStart – začáteční pozice/báze intervalu (báze jsou číslované od 0)

3. chromEnd – konečná pozice/báze intervalu (nepočítá se do intervalu)

Důležité je dávat si pozor na číslování bází. Například, prvních 100 bází chromozomu se definuje jako: *chromStart = 0*, *chromEnd = 100*. Interval bude zahrnovat báze 0–99. [35]

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
chr2  158365864  158367031
chr3  127477031  127478198
chr3  127478198  127479365
chr3  127479365  127480532
chr3  127480532  127481699
```

Obrázek 3.4: Příklad datového formátu BED.
Zdroj: [35]

3.2 Proces zpracování bioinformatických dat

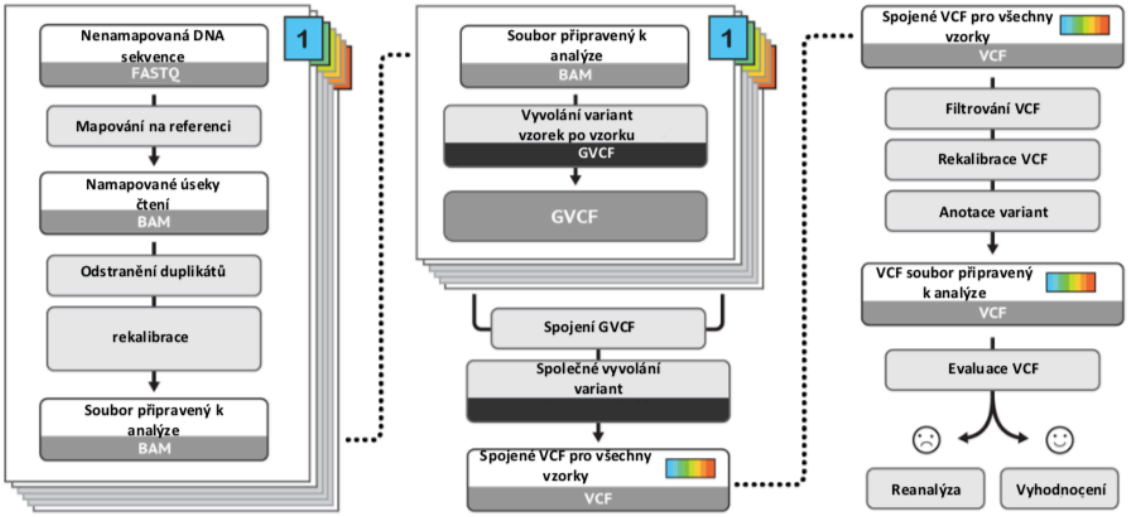
Většina nástrojů na bioinformatické zpracování dat dostupných v dnešní době používá metodiky založené na stejném principu, a tím je zarovnání přečtené sekvence na sekvenci referenční (*alignment*) a následné vyvolání variant (*variant calling*). Pracuje se i s identickými datovými formáty, takže každé zpracování dat začíná získáním dvou FASTQ souborů ze sekvenátorů a končí výstupním VCF souborem se seznamem variant. Celý proces zpracování bioinformatických dat je znázorněn na Obr. 3.5.

3.2.1 Zarovnání (*Alignment*)

1. Mapování přečtené sekvence na referenční genom. Reference lidského genomu má několik verzí. Nejčastěji používané verze mají označení hg19 a hg38 (novější). Vstupem je dvojice FASTQ souborů a výstupem je BAM soubor (.bam) a jeho index (.bai) pro každý vzorek.
2. Vyhledávání a následné odstranění PCR duplikátů, pokud se ve vzorku vyskytují. Tento krok je volitelný, záleží na zvolené knihovně. **PCR duplikáty** vznikají při sekvenování jedné nebo více kopií stejného úseku DNA.
3. Rekalibrace algoritmem – algoritmus na základě přepočítávání skóre kvality čtení detekuje systematické chyby způsobené sekvenčním přístrojem. Rekalibračním algoritmem je například BQSR (*Base Quality Scores Recalibration*), který je integrovaný například v GATK. [36]

3.2.2 Vyvolání variant (*Variant calling*)

1. Každý nástroj disponuje svým algoritmem pro vyvolání variant. Algoritmus se také liší podle variant, které chceme detekovat. Princip je ale většinou podobný. Výstupem je GVCF soubor pro každý vstupní BAM.
2. Společné vyvolání variant pro všechny spojené GVCF. Společným vyvoláním se dokáží detekovat i méně časté varianty ve vzorcích. Výsledkem je VCF soubor s variantami.
3. Filtrování variant – tvrdé filtrování nebo algoritmus (*Variant Quality Scores Recalibration*).
4. Výsledný VCF soubor připravený k anotaci dalšími nástroji.



Obrázek 3.5: Proces zpracování bioinformatických dat. Zdroj: [37]

3.3 Genome analysis toolkit (GATK)

S obrovským objemem dat generovaným pomocí MPS technologií se zvýšily nároky na nástroje používané k analýze. Neustálé vytváření účinných a robustních nástrojů by bylo velmi obtížné i pro profesionály v oboru. Proto byla ve spolupráci s 1000 Genomes Project vytvořena sada nástrojů GATK.

GATK je zkratkou pro Genome Analysis Toolkit. Jedná se o sadu nástrojů pro příkazový řádek používanou k analýze dat ze sekvenátorů. Primárně je sada zaměřena na zjišťování variant, ale obsahuje i množství dalších nástrojů (zpracování dat z MPS, filtrování, vyhodnocování variant atd.), které je možné využívat zřetězené do úplných pracovních postupů (*workflows*) nebo i jednotlivě. GATK poskytuje komplexní řešení postupů nazvané GATK Best Practices, kde lze nalézt konkrétní případy použití. [37]

Výhodou používání GATK je to, že lze poměrně snadno, rychle a efektivně vytvářet standardizované postupy zpracování dat s reprodukovatelnými výsledky. Přístup dle

doporučení GATK si zvolilo a aplikovalo již několik významných studií (The Cancer Genome Atlas či gnomAD). [38]

3.4 Detekce zárodečných CNV pomocí GATK

Princip detekce zárodečných CNV se z části liší od principů detekce jiných variant. Jednotlivé vzorky se porovnávají s modelem, který se nejprve vytvoří na základě všech vstupních vzorků.

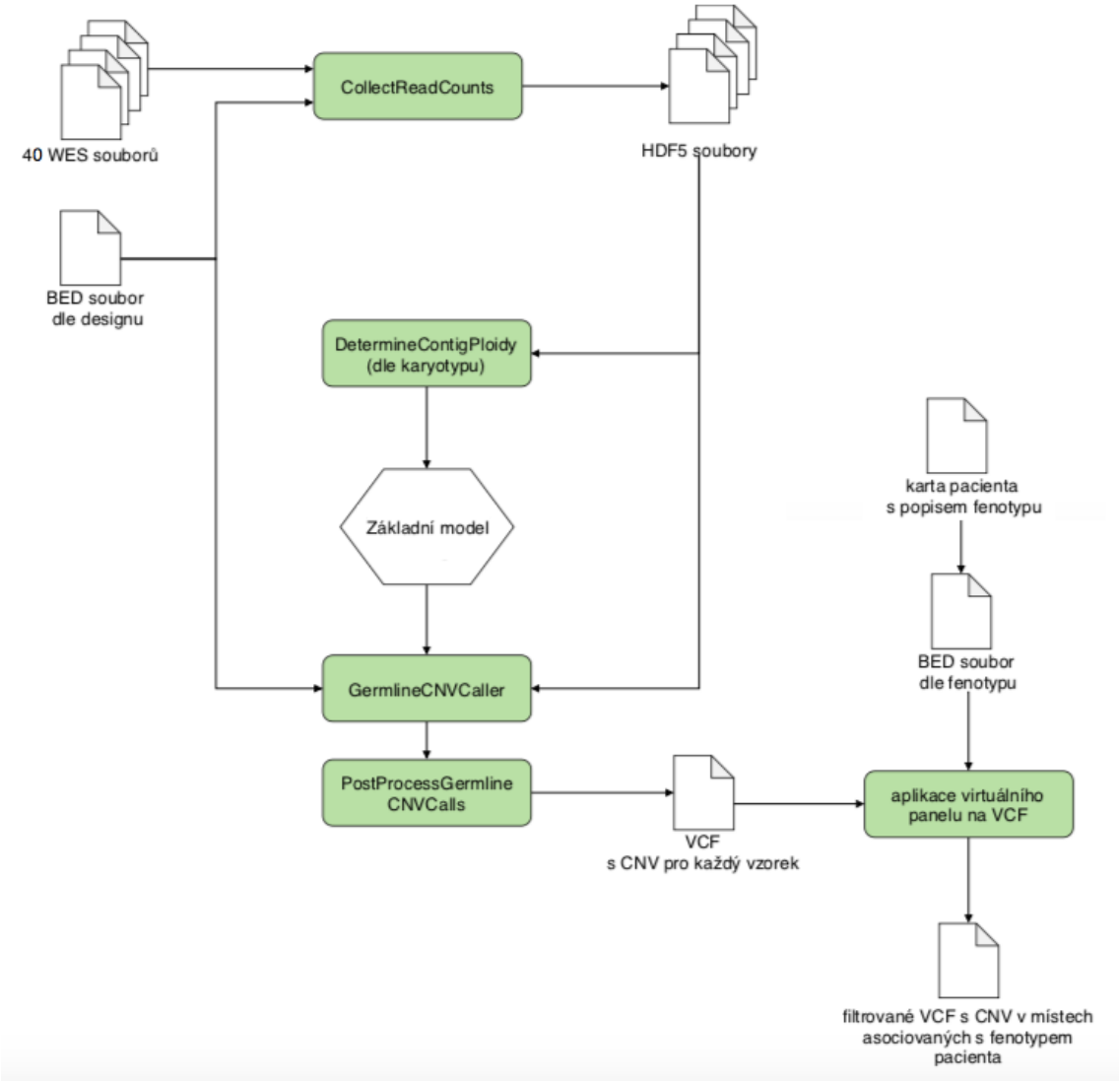
3.4.1 Postup analýzy zárodečných CNV

1. Ze vstupních BAM souborů se vytvoří model pro analýzu CNV.
2. V BED souborem definovaném intervalu se pomocí algoritmu *CollectReadCounts* určí počet úseků čtení. Pro každý vzorek vznikne výstup v podobě HDF5 souboru.
3. Algoritmus *DetermineGermlineContigPloidy* určí pro každý vzorek množství kopií každé informace. Vytvoří tak základní model potřebný pro vyvolání CNV.
4. Následně se porovnávají jednotlivé části sekvence dle intervalů a jejich pokrytí – algoritmus G*ermlineCNVCaller* pak vytvoří model pokrytí („coverage model“). Na základě odchylky od modelu lze detekovat rozdíl v počtu kopií.
5. Vyvolání jednotlivých CNV provádí algoritmus *PostprocessGermlineCNVCalls*. Výstupem pro každý vzorek je soubor ve formátu GVCF, ve kterém jsou CNV označeny číselnou hodnotou (0 – normální výsledek, 1 – delece, 2 – duplikace).

Kvůli velkému množství detekovaných CNV se doporučuje zaměřit se jen na ty, spojované s onemocněním vyšetřovaného jedince. Pomocí BED souboru, se hledání omezí jen na požadované oblasti. [39]

Celá analýza se dělí na dva režimy/módy. První *cohort mode*, připraví data do potřebné podoby a následně generuje základní model. Druhým režimem je *case mode*, který analyzuje každý vzorek samostatně na základě již vytvořeného (*cohort*) modelu. Stejný postup se vztahuje jak na data z celoexomového sekvenování (WES), tak i na data z celogenomového sekvenování (WGS).

Celý proces analýzy zárodečných CNV je znázorněn na Obr. 3.6.



Obrázek 3.6: Schéma analýzy zárodečných CNV dle doporučení GATK. Zdroj: [40]

3.5 Jiné nástroje pro detekci CNV v datech z MPS

Dalším nástrojem, vhodným k detekci CNV, je metaSV. Pomocí metaSV lze detekovat strukturální varianty (včetně CNV) na základě seskupování výstupů z dalších nástrojů. [41] Nástroje, které se používají k analýze dat před vstupem do metaSV, jsou:

- CNVkit – univerzální nástroj pro detekci CNV ve všech typech sekvenačních dat. Principem hledání CNV je zjištění průměrné hloubky čtení v celém BAM souboru a následné vyhledávání významných odchylek od průměrné hloubky v BED souborem definovaných oblastech. [42]
- CNVnator – hledá CNV využitím mean-shift algoritmu. Ve zvoleném okně se podle průměrné hloubky pokrytí vypočítá signál. Při velkých rozdílech

mezi hodnotami signálu se velikost skoku/rozdílu porovná s průměrnými hodnotami v genomu a na základě toho se detekují CNV. [43]

- Pindel, BreakSeq2 – nástroje fungující na principu hledání míst v sekvenci, kde došlo k inzerci/deleci genetické informace. Algoritmus zarovnává úseky přečtené sekvence na referenci, pokud část úseku neodpovídá referenční sekvenci, je tato část posunuta dál, dokud nedojde k zarovnání celého zbytku úseku na referenci. Vzniklá mezera je detekována jako potenciální varianta vzniklá delecí. [44] [45]

3.6 Analýza dat v cloudu

Zabývat se bioinformatickými analýzami na omezeném výpočetním výkonu je časově velice náročné. S příchodem technologií masivně paralelního sekvenování, se generování obrovského objemu dat ještě zrychlilo. Pro řešení této situace se začalo využívat cloudových služeb.

Uživatel nahraje data do cloudového úložiště a následná analýza probíhá na serverech (Google, Amazon) s mnohonásobně vyšším výpočetním výkonem, než mají lokální stroje. Velkou výhodou takového řešení je rychlost, s kterou analýzy probíhají. Nevýhodou je zpoplatnění za používání cloudových služeb a povinnost uživatele zajistit anonymizaci dat.

3.6.1 Terra App

V roce 2019 byla spuštěna cloudová aplikace Terra App běžící na platformě Google Cloud Platform [46], která umožňuje jednoduchou analýzu dat s již předpřipravenými pracovními postupy (*workflows*) podle doporučení GATK. Tzv. *workflows* jsou dostupné pro detekování několika typů variant a obsahují už předem nastavené parametry. Za službu se platí formou poplatků (hrazených přes Google Billing Accounts) za využité úložiště a výkon.

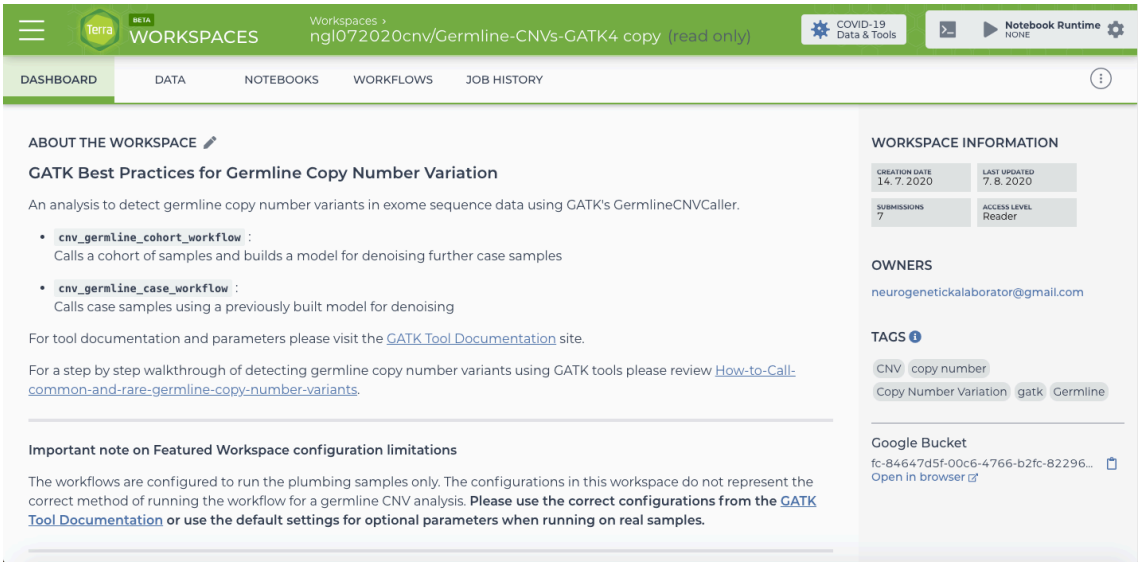
Pro řešení části mé práce jsem také zvolil aplikaci Terra App, jelikož vytvoření většího/finálního modelu (potřebného k detekování zárodečných CNV) je díky velkým výpočetním nárokům neproveditelný proces pro lokální pracovní stanici.

3.6.2 Práce s Terra App

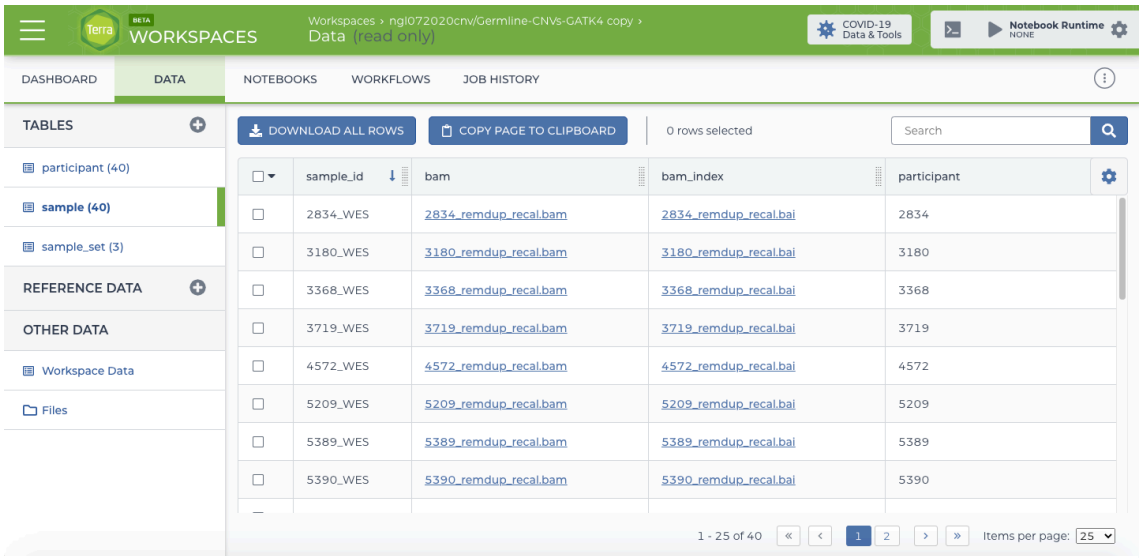
Pro přístup do cloudové aplikace Terra App je nutnost mít k dispozici Google účet. Po přihlášení má uživatel možnost vytvářet vlastní, klonovat či spravovat již existující pracovní prostory (*workspaces*). Dále může uživatel prohlížet, případně využívat veřejné analýzy (skripty, dokumentace) a data z velkých studií (např. 1000 Genomes, Human Cell Atlas, ENCODE Project atd.). V každém jednotlivém *workspace* je několik oddílů:

1. Dashboard (Obr. 3.7) – Zde by měli být obsaženy veškeré důležité informace potřebné k reprodukci analýzy jako například účel („Co se bude analyzovat“), očekávání (odhadovaná cena a čas, používané nástroje, očekávaný výstup) a postup popsaný krok po kroku.
2. Data (Obr. 3.8) – V tomto oddíle lze prohlížet a upravovat data, která jsou vizualizována v tabulkové struktuře.
3. Notebooks (Obr. 3.9) – Možnost vytváření a sdílení poznámek i s úryvky kódu. Využití *Notebooks* je velice účinné pro lepší pochopitelnost a reprodukovatelnost vytvářené analýzy.
4. Workflows (Obr. 3.10) – Zde si uživatel může vytvářet, spravovat a spouštět celé analýzy ale i jednotlivé kroky analýzy. V pododdíle S*cript* se nachází celý kód a lze s ním libovolně pracovat. Dalšími dvěma pododdíly jsou *Inputs* a *Outputs*, kde se dají upravovat vstupní, respektive výstupní parametry a jejich hodnoty.
5. Job History (Obr. 3.11) – Přehled všech úspěšně i neúspěšně proběhlých a současně spuštěných analýz. Uživatel může podrobně kontrolovat, co se zrovna děje a najít přímé odkazy na všechny zúčastněné vstupní a výstupní soubory.

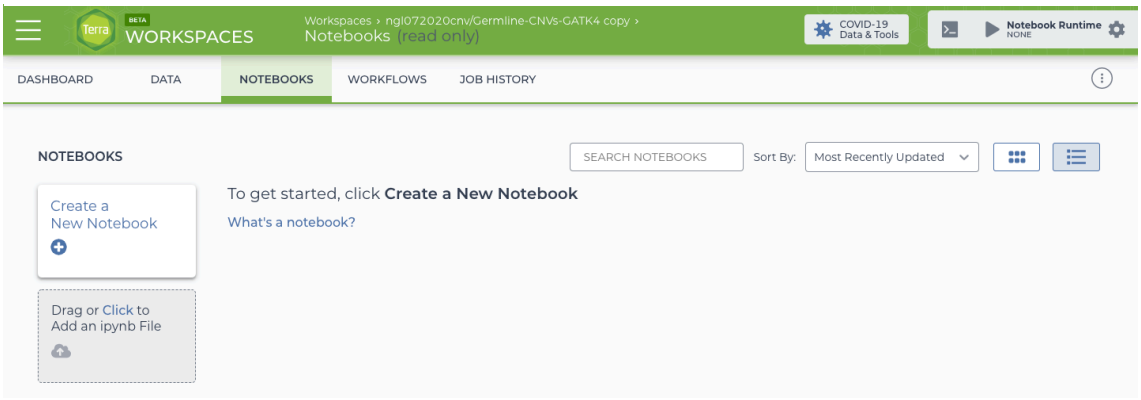
V mém případě nad daty poběží dvě části analýzy (*workflows*). První má za úkol vytvoření základního modelu (*cohort mode*) a druhá porovnávání jednotlivých sekvencí s vytvořeným modelem a následné detekování variant (*case mode*).



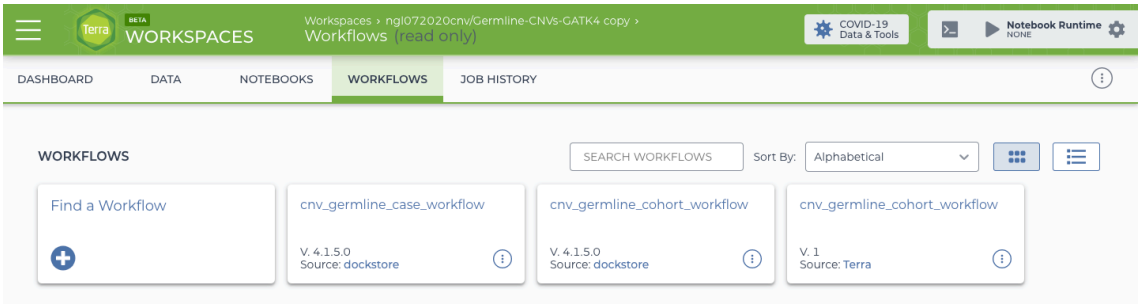
Obrázek 3.7: Terra App – Dashboard. Zdroj: vlastní



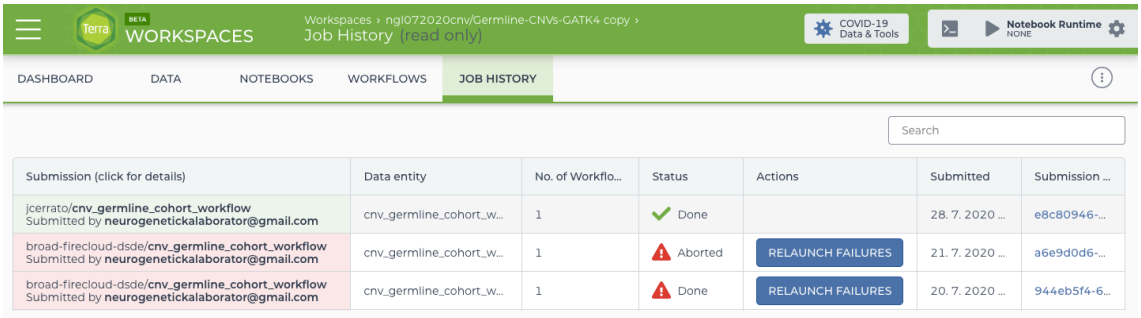
Obrázek 3.8: Terra App – Data. Zdroj: vlastní



Obrázek 3.9: Terra App – Notebooks. Zdroj: vlastní



Obrázek 3.10: Terra App – Workflows. Zdroj: vlastní



Obrázek 3.11: Terra App – Job History. Zdroj: vlastní

3.7 Pacienti a data

U vybraných pacientů se v první řadě začíná odběrem vzorku DNA a zjištěním co nejvíce souhrnných informací o pacientovi (celková anamnéza). Poté je zjištěna celková anamnéza obou rodičů včetně odebrání vzorku DNA.

V některých případech se u pacienta přistupuje pouze ke klasickému sekvenování vybraného genu. Při nenalezení patogenní varianty se využívá metod masivně paralelního sekvenování pomocí panelu genů, tzn. že se hledají patogenní varianty v předem definované množině (= panelu) genů. Jestliže není ani tentokrát nalezena patogenní varianta přistupuje se k celoexomovému (*Whole exome sequencing, WES*) případně k celogenomovému sekvenování (*Whole genome sequencing, WGS*). Prostřednictvím WES u pacienta a obou rodičů je možné hledat *de novo* varianty v genech, které nemají stále žádnou souvislost s lidskými onemocněními. WES se používá v případě, pokud nebyla při MPS panelem genů nalezena kauzální patogenní varianta. Výhodou WES oproti sekvenování panelem genů je pokrytí všech kódujících oblastí genů, naopak nevýhodou jsou vyšší náklady.

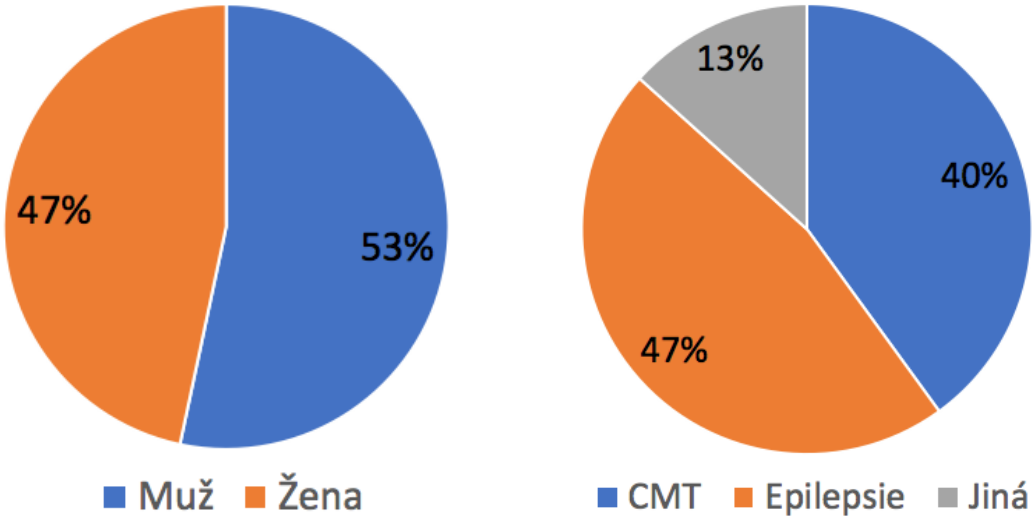
Níže v tabulkách Tab. 3.1 a Tab. 3.2 jsou uvedeny přehledy pacientů, jejichž data z WES byla použita v této práci pro vytvoření modelů pro analýzu zárodečných CNV. Na grafických přehledech Obr. 3.12 a Obr. 3.13 jsou znázorněny poměry mezi diagnózami a také poměry mezi počty mužů a žen.

Diagnóza	**Muži**	**Ženy**	**Celkem**
CMT	5	1	6
Epilepsie	2	5	7
Jiná	1	1	2
Celkem	8	7	**15**

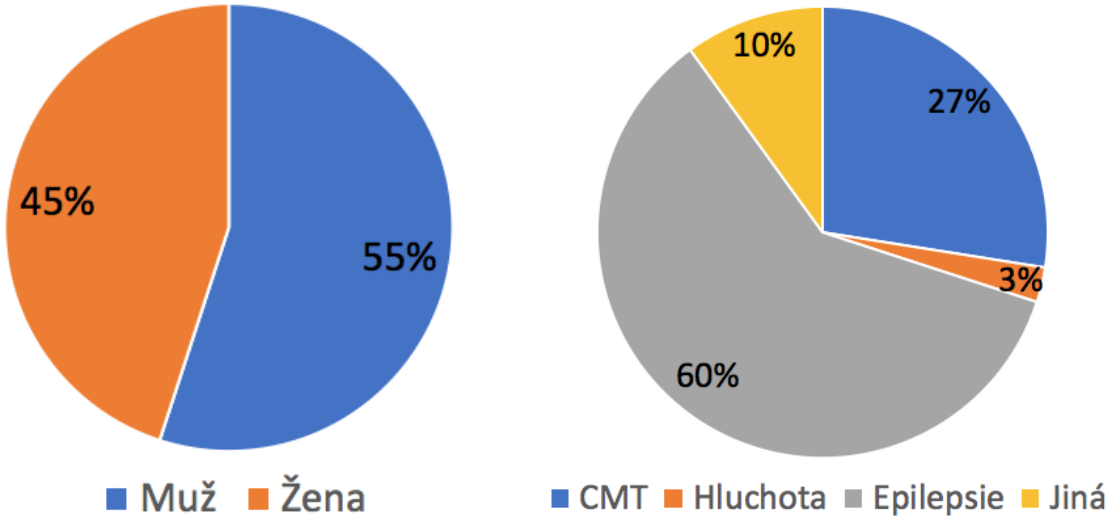
Tabulka 3.1: Přehled pacientů dle diagnóz a pohlaví. Jejich data byla použita k vytvoření testovacího modelu.

Diagnóza	**Muži**	**Ženy**	**Celkem**
CMT	9	2	11
Hluchota	0	1	1
Epilepsie	11	13	24
Jiná	2	2	4
Celkem	22	18	**40**

Tabulka 3.2: Přehled pacientů dle diagnóz a pohlaví. Jejich data byla použita k vytvoření finálního modelu.



Obrázek 3.12: Poměry pacientů dle pohlaví a diagnóz z testovacího modelu. Zdroj: vlastní



Obrázek 3.13: Poměry pacientů dle pohlaví a diagnóz z finálního modelu. Zdroj: vlastní

3.7.1 Skupina „monoalelických“ pacientů s dědičnou hluchotou

Pro otestování výsledných CNV modelů byla vybrána skupina pacientů s dědičnou hluchotou. Jedná se o 28 pacientů u kterých bylo provedeno celoexomové sekvenování. To odhalilo heterozygotní variantu na pozici c.35delG v genu *GJB2*. Další krok analýzy tedy spočívá v hledání další varianty v oblasti genu, popř. CNV způsobujícího deleci/duplikaci blízké oblasti genu *GJB2*. Rutinní analýzou variant nebyla nalezena patogenní varianta v souboru dat, proto je přistoupeno k CNV analýze tohoto vzorku.

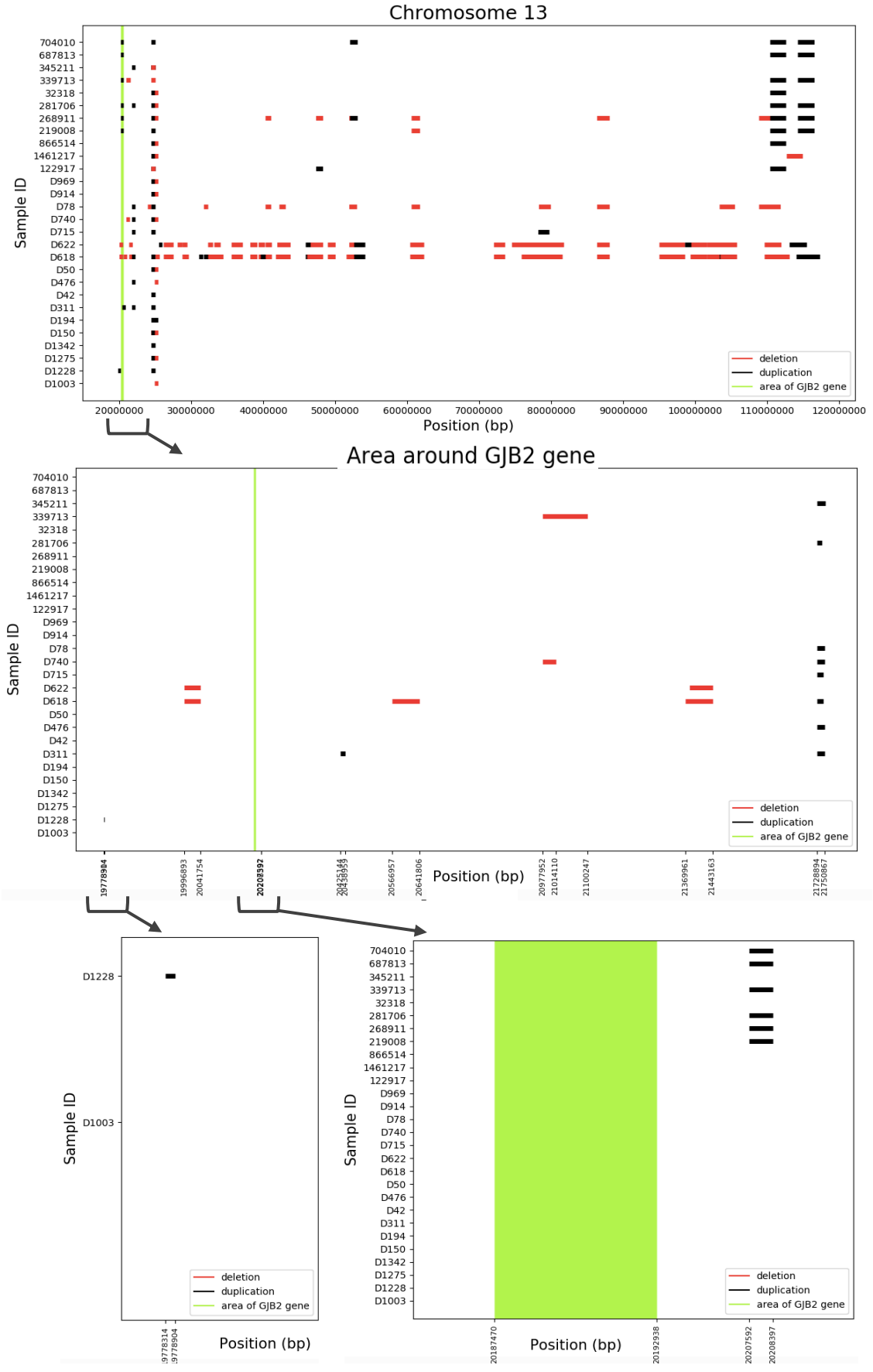
CNV analýza bude provedena porovnáváním dat vybraných pacientů s oběma modely, poté bude hledání omezeno na oblast chromozomu 13, na kterém se nachází gen *GJB2*. Cílem tohoto kroku je identifikovat možná CNV v definované oblasti.

4 Výsledky

V rámci bakalářské práce bylo analyzováno 28 pacientů pomocí testovacího modelu a 11 pacientů pomocí finálního modelu. Testovacím modelem bylo v průměru na jedince nalezeno 471 CNV vzniklých delecí s průměrným skóre kvality 258,07 a 261 CNV vzniklých duplikací s průměrným skóre kvality 25,66. U finálního modelu je průměrný počet CNV delecí 1910 (skóre kvality: 202,20) a duplikací 58 (skóre kvality: 33,56). V tabulkách Tab. 4.1 a Tab. 4.2 jsou celkové přehledy detekovaných CNV pro jednotlivé vzorky. Varianty nalezené na chromozomu 13 a v oblasti genu *GJB2* jsou graficky znázorněné na Obr. 4.1 a Obr. 4.2.

Sample_ID	Total number of DEL	Total number of DUP	Average quality score of DEL	Average quality score of DUP	Number of DEL on chr13	Number of DUP on chr13	Average quality score of DEL on chr13	Average quality score of DUP on chr13
D1003	269	163	204.80	12.10	1	0	29.82	0
D1228	290	158	272.80	25.90	0	2	0	15.02
D1275	317	150	398.46	24.26	1	1	49.69	11.9
D1342	295	144	274.78	24.88	0	1	0	5.32
D150	288	130	308.49	18.77	2	1	18.22	6.63
D194	310	168	289.79	9.73	0	2	0	4.59
D311	294	115	310.08	16.51	0	3	0	22.08
D42	316	132	344.23	14.47	0	1	0	3.75
D476	298	158	384.30	19.45	1	1	33.35	33.7
D50	250	176	282.86	12.88	1	1	34.39	8.24
D618	2067	1801	137.14	95.53	49	11	120.55	8.54
D622	2215	965	198.43	227.81	44	6	143.03	19.86
D715	332	156	318.15	13.72	0	3	0	15.24
D740	318	145	402.37	20.09	2	2	151.0	18.94
D78	606	132	160.90	16.72	11	2	77.76	9.29
D914	264	134	319.86	16.53	1	1	58.48	15.23
D969	276	148	274.37	20.57	1	1	54.0	3.09
122917	267	119	161.54	11.54	1	3	18.73	3.78
1461217	277	117	210.20	10.22	2	1	11.66	11.32
866514	293	128	328.93	21.11	1	2	12.44	14.29
219008	475	271	180.42	8.75	1	4	19.7	8.71
268911	551	278	111.27	8.94	7	5	30.10.	6.54
281706	469	293	206.27	9.75	1	5	14.92	7.75
32318	267	145	281.28	10.24	1	2	14.41	5.17
339713	432	291	219.08	10.68	2	3	30.26	11.77
345211	290	131	204.33	16.78	2	2	24.85	25.98
687813	434	282	229.28	10.61	0	3	0	9.03
704010	429	279	211.56	9.96	0	5	0	5.99

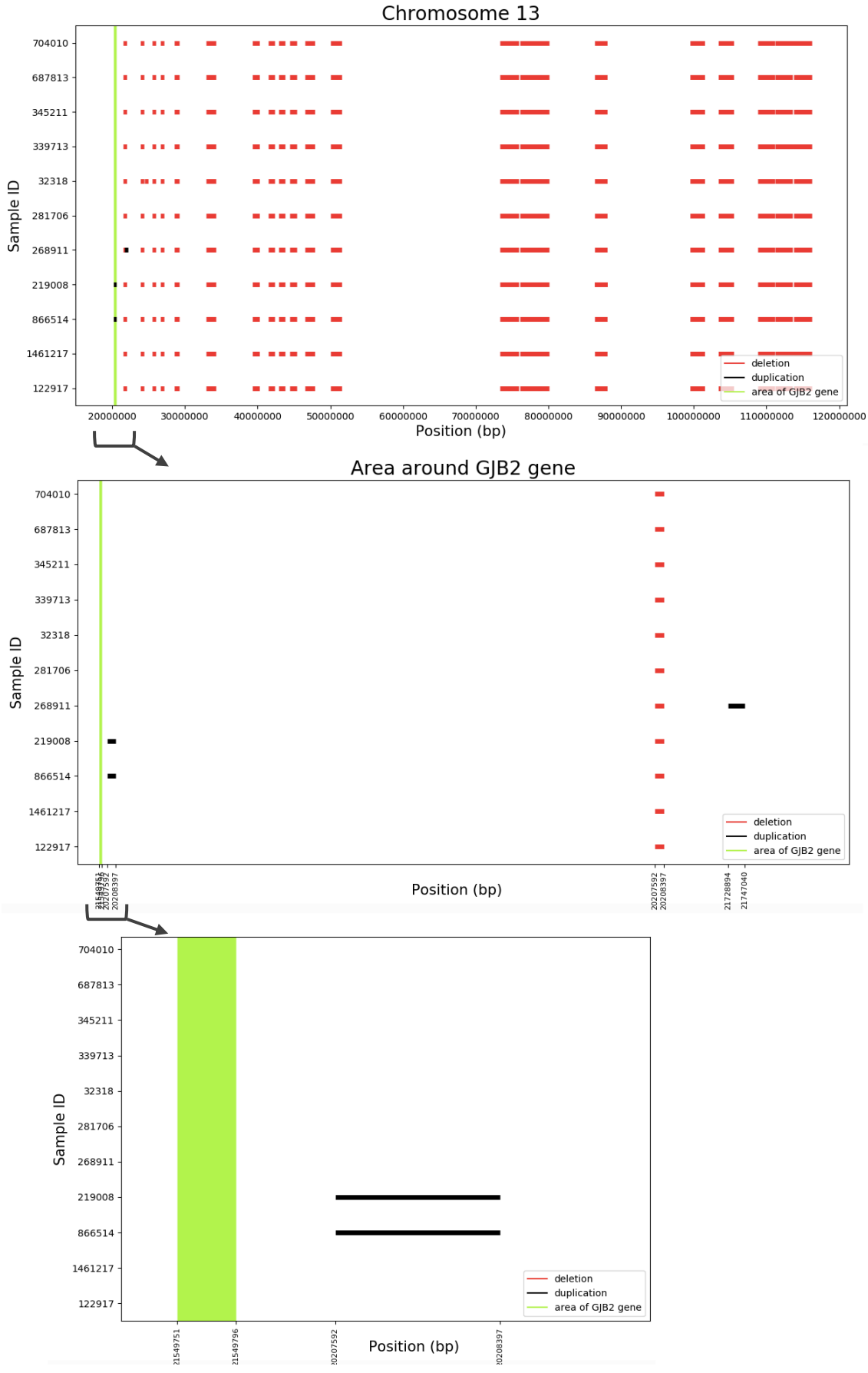
Tabulka 4.1: Výsledky analýzy na základě testovacího modelu. Celkové počty detekovaných CNV s průměrným skóre kvality i pro chromozom 13.



Obrázek 4.1: Grafické znázornění testovacím modelem nalezených CNV na chromozomu 13 pro každý ze 28 vzorků a přiblížení na oblast kolem genu *GJB2*. Zdroj: vlastní

Sample_ID	Total number of DEL	Total number of DUP	Average quality score of DEL	Average quality score of DUP	Number of DEL on chr13	Number of DUP on chr13	Average quality score of DEL on chr13	Average quality score of DUP on chr13
122917	1843	61	199.65	23.75	34	0	151.27	0
1461217	1843	61	244.38	16.49	35	0	178.04	0
219008	1959	61	168.42	40.06	34	1	116.28	4.18
268911	1950	64	159.10	34.29	35	1	116.41	5.72
281706	1970	57	180.05	39.08	34	1	125.10	52.32
32318	1868	61	221.78	21.06	34	0	146.78	0
339713	1950	49	176.64	47.34	35	0	114.10	0
345211	1857	62	246.67	40.87	35	0	183.49	0
687813	1958	54	177.84	41.27	34	0	119.93	0
704010	1929	51	180.14	41.78	34	0	122.32	0
866514	1883	62	268.72	23.16	35	0	180.43	0

Tabulka 4.2: Výsledky analýzy na základě finálního modelu. Celkové počty detekovaných CNV s průměrným skóre kvality i pro chromozom 13.



Obrázek 4.2: Grafické znázornění finálním modelem nalezených CNV na chromozomu 13 pro každý z 11 vzorků a přiblížení na oblast kolem genu *GJB2*. Zdroj: vlastní

5 Diskuse

Při porovnání výsledků z obou modelů, je vidět, že finální (větší model) je senzitivnější na detekování CNV vzniklých delecí, když na jeho základě bylo nalezeno 4krát více variant oproti testovacímu modelu. Rozdíl podle mě způsobuje větší citlivost finálního modelu na chyby sekvenačních metod. Na Obr. 4.2 je patrný výskyt CNV v každém vzorku podle určitého schématu. Je velice nepravděpodobné, aby 11 pacientů mělo podobný počet variant na téměř identických pozicích chromozomu 13. Data pacientů navíc pocházejí ze stejného cyklu sekvenování.

V článku [46] je zmíněno, že průměrný počet CNV vyskytující se v celém genomu každého jedince je roven 12. Ačkoli jsem vycházel pouze z WES dat, průměrný počet mnou detekovaných CNV na jedince je několikanásobně vyšší. Rozdíl je pravděpodobně způsoben nepřesnostmi sekvenačních metod, chybami algoritmů v průběhu analýzy nebo použitím WES dat. Data z celoexomového sekvenování mají nekontinuální pokrytí, skládají se totiž jen z kódujících oblastí genomu, a tím se zvyšuje počet detekovaných CNV. To je patrné i z Obr. 4.1 a Obr 4.2, kde varianty vyskytující se v každém vzorku na stejných pozicích mohou nejspíše být zmíněné chyby. Výsledky analýzy mohou být tím pádem snadno ovlivnitelné nekonzistencí vstupních dat.

Jelikož nebylo možné, kvůli vysokým technickým požadavkům, vytvořit na lokálním zařízení finální (velký) model, muselo se přistoupit k jinému řešení. Po konzultaci s vedoucím práce jsem se rozhodl pro využití cloudové aplikace Terra App. Výhodou používání cloudových služeb je obrovský výpočetní výkon, který je k dispozici. V úvahu jsem ale musel brát, že pracuji s daty té nejcitlivější povahy a při používání cloudových služeb uživatel nemá nikdy úplnou jistotu, kdo všechno má k jeho datům přístup. Proto je důležité se při práci s daty chovat obezřetně. Data zpracovávaná v rámci této práce byla po celou dobu uchovávána v zabezpečeném, šifrovaném úložišti a byla označena číselným kódem, který znemožňoval identifikaci pacienta.

Postup detekce zárodečných CNV použitý v této práci se dá považovat za relativně kvalitní, ale vzhledem k tomu, že se jedná stále o „*beta*“ verzi, je výpočetně, a tudíž i časově velice náročný. Velké množství času při vytváření modelů zabralo upravování parametrů a komunikace s podporou GATK, než se vůbec povedlo *cohort workflow* spustit. Problémem také byla v podstatě neexistující dokumentace k oběma *workflow*. Výsledkem prvního úspěšného spuštění byl model rozdělený na 11 000 částí, jehož vytvoření trvalo bezmála 8 dní a stálo přibližně 135 $. Postupným „laděním“ se časové i finanční nároky snižovaly až do stavu, kdy byl konečný finální model vytvořen během 2 hodin za necelých 5 $.

Do budoucna se jako perspektivní jeví vytvoření modelu z WES dat od přibližně 250 pacientů, který by byl dostatečně kvalitní pro rutinní běhy analýzy. Časový odhad

vytvoření modelu pomocí cloudové aplikace Terra App je zhruba 15 hodin a cenové náklady by se pohybovaly v rozmezí od 50 $ do 70 $ za předpokladu použití podobných parametrů jako při vytvoření finálního modelu. Další připadá v úvahu vytvoření většího modelu z dat celogenomového sekvenování (WGS) pacientů. Detekování CNV by bylo poté kvalitnější a přesnější, jelikož by k analýze byl k dispozici celý genom pacienta, a ne jenom jeho kódující oblasti, jako je tomu u WES. Nevýhodou jsou mnohonásobně vyšší náklady jak finanční, tak i časové.

V oblasti genu *GJB2* bylo oběma modely detekováno několik CNV, jejichž klinickou interpretaci provede zkušený expert, který se na problematiku zaměřuje.

6 Závěr

Cílem bakalářské práce bylo otestovat současnou metodiku pro detekci variability počtu kopií segmentů DNA a následně ji aplikovat na data z celoexomového sekvenování pacientů s neurogenetickým onemocněním. Splnění jednotlivých cílů bylo dosaženo pomocí nejmodernějších nástrojů a technologií, které jsou v dnešní době k dispozici.

Hlavním záměrem této práce bylo zavést do praxe bioinformatické postupy pro detekci zárodečných CNV, které budou jednoduchým způsobem opakovatelné a budou schopné generovat reprodukovatelné výsledky. Pro splnění tohoto cíle byl vytvořen skript v programovacím jazyce Python, který po spuštění s jediným argumentem, cestou ke složce s BAM soubory, zkontroluje správnost a úplnost vstupních dat a následně spustí analýzu zárodečných CNV v obou režimech (*cohort* i *case*). Používání skriptu není pro uživatele nijak složité, stačí základní znalost příkazové řádky. Skript je volně k dispozici na přiloženém CD a může být bez omezení využíván a upravován.

Pro další postup v práci bylo nutné v *cohort* režimu vytvořit CNV modely. Testovací model byl vytvořen z dat celoexomového sekvenování 15 pacientů. Finální model byl nakonec kvůli náročnosti vytvořen z WES dat od minimálního požadovaného počtu 40 pacientů.

Otestování proběhlo na skupině pacientů s dědičnou hluchotou, pro testovací model jich bylo 28 a pro model finální 11. Výběr skupiny byl vhodný vzhledem k tomu, že bylo možné případný výskyt CNV předpokládat v předem definované oblasti.

Seznam použité literatury

[1] REDON, Richard, Shumpei ISHIKAWA, Karen R. FITCH, et al. Global variation in copy number in the human genome. Nature [online]. 2006, 444(7118), 444-454 [cit. 2020-03-27]. DOI: 10.1038/nature05329. ISSN 0028-0836. Dostupné z: http://www.nature.com/articles/nature05329

[2] Gene | Talking Glossary of Genetic Terms | NHGRI. National Human Genome Research Institute Home | NHGRI [online]. 2019 [cit. 2020-03-28]. Dostupné z: https://www.genome.gov/genetics-glossary/Gene

[3] OTOVÁ, Berta, Milada KOHOUTOVÁ a Aleš PANCZAK. Lékařská biologie a genetika. 2., nezměněné vydání. Praha: Karolinum, 2019. ISBN 9788024637907.

[4] National Center for Biotechnology Information. MLA CE Course Manual: Molecular Biology Information Resources (Genetics Review: Gene) [online]. 2001 [cit. 2020-03-28]. Dostupné z: https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/gene.html

[5] OTOVÁ, Berta a Romana MIHALOVÁ. Základy biologie a genetiky člověka. V Praze: Karolinum, 2012. ISBN 9788024621098.

[6] BARTOŇ, Vojtěch a Denisa MADĚRÁNKOVÁ. Numerické reprezentace proteinových sekvencí pro klasifikaci. 2016.

[7] STRACHAN, T., J. GOODSHIP a Patrick F. CHINNERY. Genetics and genomics in medicine. New York: Garland Science/Taylor & Francis Group, [2015]. ISBN 978-0-8153-4480-3.

[8] STANĚK, David, Petra LAŠŠUTHOVÁ, Katalin ŠTĚRBOVÁ, Markéta VLČKOVÁ, Jana NEUPAUEROVÁ, Marcela KRŮTOVÁ a Pavel SEEMAN. Detection rate of causal variants in severe childhood epilepsy is highest in patients with seizure onset within the first four weeks of life. Orphanet Journal of Rare Diseases [online]. 2018, 13(1) [cit. 2020-08-08]. DOI: 10.1186/s13023-018-0812-8. ISSN 1750-1172. Dostupné z: https://ojrd.biomedcentral.com/articles/10.1186/s13023-018-0812-8

[9] FOX, Amanda A., Sonal SHARMA, J. Paul MOUNSEY, Marcel E. DURIEUX, Richard WHITLOCK a Elliott BENNETT-GUERRERO. Molecular and Genetic Cardiovascular Medicine and Systemic Inflammation. Kaplan's Essentials of Cardiac Anesthesia [online]. Elsevier, 2018, 2018 [cit. 2020-04-02]. DOI: 10.1016/B978-0-323-49798-5.00006-1. ISBN 9780323497985. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/B9780323497985000061

[10] OTOVÁ, Berta, Milada KOHOUTOVÁ a Aleš PANCZAK. Lékařská biologie a genetika. Praha: Karolinum, 2013. ISBN 978-80-246-1594-3.

[11] OLIVEIRA, Sofia, David COOPER a Luisa AZEVEDO. De Novo Mutations in Human Inherited Disease. [online]. 17 September 2018 [cit. 2020-04-04]. DOI: 10.1002/9780470015902.a0027866. Dostupné z: https://onlinelibrary.wiley.com/doi/full/10.1002/9780470015902.a0027866

[12] THAPAR, Anita a Miriam COOPER. Copy Number Variation: What Is It and What Has It Told Us About Child Psychiatric Disorders? Journal of the American Academy of Child & Adolescent Psychiatry [online]. 2013, 52(8), 772-774 [cit. 2020-03-20]. DOI: 10.1016/j.jaac.2013.05.013. ISSN 08908567. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0890856713003407

[13] LEE, Charles a Stephen W. SCHERER. The clinical context of copy number variation in the human genome. Expert Reviews in Molecular Medicine [online]. 2010, 12 [cit. 2020-03-21]. DOI: 10.1017/S1462399410001390. ISSN 1462-3994. Dostupné z: https://www.cambridge.org/core/product/identifier/S1462399410001390/type/journal_article

[14] STANKIEWICZ, Pawel a James R. LUPSKI. Genome architecture, rearrangements and genomic disorders. Trends in Genetics [online]. 2002, 18(2), 74-82 [cit. 2020-04-18]. DOI: 10.1016/S0168-9525(02)02592-1. ISSN 01689525. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0168952502025921

[15] LUPSKI, James R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends in Genetics [online]. 1998, 14(10), 417-422 [cit. 2020-04-19]. DOI: 10.1016/S0168-9525(98)01555-8. ISSN 01689525. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0168952598015558

[16] CHEN, Jian-Min, David N. COOPER, Claude FÉREC, Hildegard KEHRER-SAWATZKI a George P. PATRINOS. Genomic rearrangements in inherited disease and cancer. Seminars in Cancer Biology [online]. 2010, 20(4), 222-233 [cit. 2020-04-19]. DOI: 10.1016/j.semcancer.2010.05.007. ISSN 1044579X. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S1044579X10000428

[17] ZHANG, Feng, Pavel SEEMAN, Pengfei LIU, et al. Mechanisms for Nonrecurrent Genomic Rearrangements Associated with CMT1A or HNPP: Rare CNVs as a Cause for Missing Heritability. The American Journal of Human Genetics [online]. 2010, 86(6), 892-903 [cit. 2020-06-23]. DOI: 10.1016/j.ajhg.2010.05.001. ISSN 00029297. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0002929710002181

[18] CHEN, Lu, Weichen ZHOU, Ling ZHANG a Feng ZHANG. Genome Architecture and Its Roles in Human Copy Number Variation. Genomics & Informatics [online]. 2014, 12(4) [cit. 2020-08-08]. DOI: 10.5808/GI.2014.12.4.136. ISSN 2234-0742. Dostupné z: http://genominfo.org/journal/view.php?doi=10.5808/GI.2014.12.4.136

[19] LIEBER, Michael R., Yunmei MA, Ulrich PANNICKE a Klaus SCHWARZ. Mechanism and regulation of human non-homologous DNA end-joining. Nature Reviews Molecular Cell Biology [online]. 2003, 4(9), 712-720 [cit. 2020-04-19]. DOI: 10.1038/nrm1202. ISSN 1471-0072. Dostupné z: http://www.nature.com/articles/nrm1202

[20] WETERINGS, Eric a Dik C. VAN GENT. The mechanism of non-homologous end-joining: a synopsis of synapsis. DNA Repair [online]. 2004, 3(11), 1425-1435 [cit. 2020-04-19]. DOI: 10.1016/j.dnarep.2004.06.003. ISSN 15687864. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S156878640400179X

[21] SCHWARZ, Klaus, Yunmei MA, Ulrich PANNICKE a Michael R. LIEBER. Human severe combined immune deficiency and DNA repair. BioEssays [online]. 2003, 25(11), 1061-1070 [cit. 2020-04-19]. DOI: 10.1002/bies.10344. ISSN 0265-9247. Dostupné z: http://doi.wiley.com/10.1002/bies.10344

[22] LIEBER, Michael R, Haihui LU, Jiafeng GU a Klaus SCHWARZ. Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell Research [online]. 2008, 18(1), 125-133 [cit. 2020-04-19]. DOI: 10.1038/cr.2007.108. ISSN 1001-0602. Dostupné z: http://www.nature.com/articles/cr2007108

[23] LIEBER, Michael R. The Mechanism of Human Nonhomologous DNA End Joining. Journal of Biological Chemistry [online]. 2007, 283(1), 1-5 [cit. 2020-04-19]. DOI: 10.1074/jbc.R700039200. ISSN 0021-9258. Dostupné z: http://www.jbc.org/lookup/doi/10.1074/jbc.R700039200

[24] GU, Wenli, Feng ZHANG a James R LUPSKI. Mechanisms for human genomic rearrangements. PathoGenetics [online]. 2008, 1(1) [cit. 2020-04-19]. DOI: 10.1186/1755-8417-1-4. ISSN 1755-8417. Dostupné z: http://pathogeneticsjournal.biomedcentral.com/articles/10.1186/1755-8417-1-4

[25] LEE, Jennifer A., Claudia M.B. CARVALHO a James R. LUPSKI. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. Cell [online]. 2007, 131(7), 1235-1247 [cit. 2020-04-19]. DOI: 10.1016/j.cell.2007.11.037. ISSN 00928674. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0092867407015413

[26] HASTINGS, P. J., James R. LUPSKI, Susan M. ROSENBERG a Grzegorz IRA. Mechanisms of change in gene copy number. Nature Reviews Genetics [online].

2009, 10(8), 551-564 [cit. 2020-08-08]. DOI: 10.1038/nrg2593. ISSN 1471-0056. Dostupné z: http://www.nature.com/articles/nrg2593WHEELER, David A., Maithreyan SRINIVASAN, Michael EGHOLM, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature [online]. 2008, 452(7189), 872-876 [cit. 2020-04-21]. DOI: 10.1038/nature06884. ISSN 0028-0836. Dostupné z: http://www.nature.com/articles/nature06884

[27] SHOKRALLA, SHADI, JENNIFER L. SPALL, JOEL F. GIBSON a MEHRDAD HAJIBABAEI. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology [online]. 2012, 21(8), 1794-1805 [cit. 2020-07-02]. DOI: 10.1111/j.1365-294X.2012.05538.x. ISSN 09621083. Dostupné z: http://doi.wiley.com/10.1111/j.1365-294X.2012.05538.x

[28] Illumina | Sequencing and array-based solutions for genetic research [online]. 2020 [cit. 2020-04-21]. Dostupné z: https://www.illumina.com

[29] ZHOU, Xiaoguang, Lufeng REN, Qingshu MENG, Yuntao LI, Yude YU a Jun YU. The next-generation sequencing technology and application. Protein & Cell [online]. 2010, 1(6), 520-536 [cit. 2020-04-21]. DOI: 10.1007/s13238-010-0065-3. ISSN 1674-800X. Dostupné z: http://link.springer.com/10.1007/s13238-010-0065-3

[30] COCK, Peter J. A., Christopher J. FIELDS, Naohisa GOTO, Michael L. HEUER a Peter M. RICE. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research [online]. 2010, 38(6), 1767-1771 [cit. 2020-07-20]. DOI: 10.1093/nar/gkp1137. ISSN 0305-1048. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp1137

[31] LI, H., B. HANDSAKER, A. WYSOKER, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [online]. 2009, 25(16), 2078-2079 [cit. 2020-07-20]. DOI: 10.1093/bioinformatics/btp352. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352

[32] The Variant Call Format (VCF) Version 4.2 Specification [online]. 2020 [cit. 2020-07-21]. Dostupné z: https://samtools.github.io/hts-specs/VCFv4.2.pdf

[33] GVCF - Genomic Variant Call Format - GATK. GATK [online]. c2020 [cit. 2020-07-23]. Dostupné z: https://gatk.broadinstitute.org/hc/en-us/articles/360035531812-GVCF-Genomic-Variant-Call-Format

[34] Genome Browser FAQ. UCSC Genome Browser Home [online]. c2000-2020 [cit. 2020-07-23]. Dostupné z: https://m.ensembl.org/info/website/upload/bed.html

[35] Base Quality Score Recalibration (BQSR) - GATK. *GATK* [online]. c2020 [cit. 2020-07-23]. Dostupné z: https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-

[36] GATK [online]. c2020 [cit. 2020-07-29]. Dostupné z: https://gatk.broadinstitute.org/hc/en-us

[37] MCKENNA, A., M. HANNA, E. BANKS, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research [online]. 2010, 20(9), 1297-1303 [cit. 2020-07-22]. DOI: 10.1101/gr.107524.110. ISSN 1088-9051. Dostupné z: http://genome.cshlp.org/cgi/doi/10.1101/gr.107524.110

[38] (How to) Call common and rare germline copy number variants - GATK. GATK [online]. c2020 [cit. 2020-07-23]. Dostupné z: https://gatk.broadinstitute.org/hc/en-us/articles/360035531152?id=11684

[39] STANĚK, David. Objasňování příčin neurogenetických onemocnění analýzou dat z MPS pomocí moderních algoritmů. Praha, 2019. Disertační práce. 2. lékařská fakulta Univerzity Karlovy.

[40] MOHIYUDDIN, Marghoob, John C. MU, Jian LI, Narges BANI ASADI, Mark B. GERSTEIN, Alexej ABYZOV, Wing H. WONG a Hugo Y.K. LAM. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics [online]. 2015, 31(16), 2741-2744 [cit. 2020-07-31]. DOI: 10.1093/bioinformatics/btv204. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv204

[41] TALEVICH, Eric, A. Hunter SHAIN, Thomas BOTTON a Boris C. BASTIAN. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Computational Biology [online]. 2016, 12(4) [cit. 2020-07-31]. DOI: 10.1371/journal.pcbi.1004873. ISSN 1553-7358. Dostupné z: https://dx.plos.org/10.1371/journal.pcbi.1004873

[42] ABYZOV, A., A. E. URBAN, M. SNYDER a M. GERSTEIN. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research [online]. 2011, 21(6), 974-984 [cit. 2020-07-31]. DOI: 10.1101/gr.114876.110. ISSN 1088-9051. Dostupné z: http://genome.cshlp.org/cgi/doi/10.1101/gr.114876.110

[43] YE, K., M. H. SCHULZ, Q. LONG, R. APWEILER a Z. NING. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics [online]. 2009, 25(21), 2865-2871 [cit. 2020-07-31]. DOI: 10.1093/bioinformatics/btp394. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp394

[44] ABYZOV, Alexej, Shantao LI, Daniel Rhee KIM, et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. Nature

Communications [online]. 2015, 6(1) [cit. 2020-07-31]. DOI: 10.1038/ncomms8256. ISSN 2041-1723. Dostupné z: http://www.nature.com/articles/ncomms8256

[45] Terra.bio [online]. [cit. 2020-07-23]. Dostupné z: https://terra.bio

[46] FEUK, Lars, Andrew R. CARSON a Stephen W. SCHERER. Structural variation in the human genome. Nature Reviews Genetics [online]. 2006, 7(2), 85-97 [cit. 2020-08-11]. DOI: 10.1038/nrg1767. ISSN 1471-0056. Dostupné z: http://www.nature.com/articles/nrg1767

Seznam obrázků

Obsah přiloženého CD

17PBIBP_ 474300_Jaroslav_Iha.pdf..... Kompletní bakalářská práce

Abstrakt_CZ.pdf ..... Abstrakt v českém jazyce

Abstract_EN.pdf .....Abstrakt v anglickém jazyce

Klicova_slova.pdf..... Klíčová slova

Skript_analyza.py .....Skript pro spuštění analýzy

Zadani_BP.pdf..... Zadání bakalářské práce