Univerzita Karlova

2. lékařská fakulta

Studijní program: Molekulární a buněčná biologie, genetika a virologie

Ing. et Ing. David Staněk

Objasňování příčin neurogenetických onemocnění analýzou dat z MPS pomocí moderních algoritmů

The elucidation of the causes of neurogenetic diseases by the MPS data analysis using advanced algorithms

Disertační práce

Školitelka: MUDr. Petra Laššuthová Ph.D.

Praha, 11. listopadu 2019

Identifikační záznam

STANĚK, David. Objasňování příčin neurogenetických onemocnění analýzou dat z MPS pomocí moderních algoritmů [The elucidation of the causes of the neurogenetic diseases by the MPS data using advanced algorithms]. Praha, 2019. 183 s., 10 příl. Disertační práce (Ph.D.). Univerzita Karlova, 2. lékařská fakulta, DNA laboratoř Kliniky dětské neurologie 2.LF a FN Motol. Vedoucí práce MUDr. Petra Laššuthová Ph.D.

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem řádně uvedl a citoval všechny použité prameny a literaturu. Současně prohlašuji, že práce nebyla využita k získání jiného nebo stejného titulu.

Souhlasím s trvalým uložením elektronické verze mé práce v databázi systému meziuniverzitního projektu Theses.cz za účelem soustavné kontroly podobnosti kvalifikačních prací.

V Praze dne 11. listopadu 2019

_____

David Staněk

Klíčová slova: Masivně paralelní sekvenování (MPS), Sekvenování nové generace (NGS), Celoexomové sekvenování (WES), Bioinformatika, Neurogenetická onemocnění

Key words: Massive parallel sequencing (MPS), Next-generation sequencing (NGS), Whole-exome sequencing (WES), Bioinformatics, Neurogenetic diseases

Poděkování Děkuji své školitelce, MUDr. Petře Laššuthové, Ph.D., za vedení mé doktorandské práce, za neustálou motivaci během mého studia, za trpělivost, konstruktivní kritiku a velkou podporu, bez které bych své studium nedokončil.

Děkuji též svým kolegům z DNA laboratoře KDN 2.LF UK a FN Motol, jmenovitě prof. Pavlu Seemanovi za konzultace mé práce, kolegyním RNDr. Anně Uhrové Mészárosové, Ph.D., RNDr. Janě Neupauerové, Ph.D., Ing. Lucii Sedláčkové Ph.D. a MUDr. Daně Šafce Brožkové Ph.D. za utvoření skvělého kolektivu a pomoc při vědecké činnosti.

Nakonec, svým rodičům a přítelkyni, kteří mi dodávali sílu a podporovali mě během celého mého studia.

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

S dokončením projektu lidského genomu (HGP) a příchodem metod masivně paralelního sekvenování DNA (MPS) se značně rozšířily možnosti celého odvětví molekulární biologie a genetiky a umožňují pokročilé hledaní nových i velmi vzácných příčin dědičných onemocnění. V současnosti MPS metody pokrývají široké spektrum aplikací, od určení malé změny jednoho nukleotidu v sekvenci, která je příčinou závažného onemocnění, po rozsáhlé studie populací čítajících stovky tisíc jedinců.

Aby bylo možné získanou informaci z MPS vytěžit, je nutné používat pokročilých počítačových metod.

Disertační práce byla zpracována v rámci DNA laboratoře KDN, která se zaměřuje dlouhodobě na diagnostiku dědičné periferní neuropatie (CMT) a závažné dětské epilepsie. Tyto klinické jednotky proto detailněji popisujeme níže.

1.1 Epilepsie a epileptické encefalopatie

Epilepsie je chronické onemocnění, jehož hlavním projevem jsou opakované záchvaty. V roce 2014 organizace International League Against Epilepsy (ILAE) definovala [Fisher et al. 2014] epilepsii jako stav, kdy:

- Byly přítomny alespoň dva spontánní epileptické záchvaty během 24 hodin

NEBO

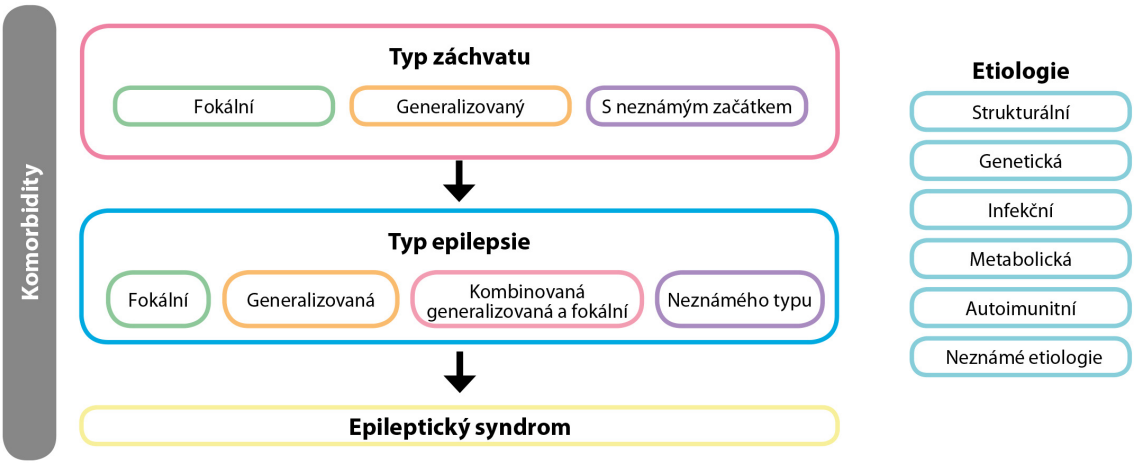
- Jeden spontánní epileptický záchvat a vysoké riziko opakování v nejbližších 10 letech

NEBO

- Byl diagnostikován epileptický syndrom

V roce 2017 [Fisher et al. 2017a], pak byla uvedená nová klasifikace epileptických záchvatů i epilepsií (klasifikace uvedená na Obr. 1.1), kdy dělíme záchvaty podle typu na fokální, generalizované a s neznámým začátkem. Z toho pak vychází typ epilepsie - fokální, generalizovaná, kombinovaná či neznámého typu.

Podle typu epileptických záchvatů a přidružených symptomů (komorbidit, jakou jsou mentální a motorické retardace, poruchy chování atd.) se určují epileptické syndromy. Posledním parametrem při klasifikaci epilepsií dle ILAE je etiologie. Prvním krokem při diagnostice epilepsie je obvykle MRI mozku, které je nezbytné pro určení případného strukturálního podkladu epilepsie. Mezi etiologie epilepsií patří infekční, genetické, metabolické či autoimunitní příčiny. Genetické příčiny zejména časných epilepsií a epileptických encefalopatií byly objeveny až v posledních několika letech a to díky zavedení metod MPS.[Fisher et al. 2017a]



Převzato z: [Marusic et al. 2018] přeloženo z původní: [Fisher et al. 2017a]

Obrázek 1.1: Klasifikace epilepsií dle ILAE 2017, česká verze

1.1.1 Epileptické encefalopatie (EE)

Epileptické encefalopatie jsou definovány jako stav, kdy epileptické záchvaty přispívají k závažným kognitivním a behaviorálním obtížím, dochází ke zpomalení a často i regresi vývoje. Nejčastěji se koncept EE uplatňuje u epilepsií s časným věkem začátku záchvatů a s genetickou etiologií. Koncept EE pracuje s hypotézou, že u pacientů dochází k postižení mozku kvůli záchvatům a změnám elektrické epileptogenní aktivity. EE jsou často farmakorezistivní. Příčina EE je často monogenně podmíněná (příkladem mohou být *CDKL5* encefalopatie nebo *CHD2* encefalopatie. [Khan a Al Baradie 2012; Capovilla et al. 2013; Scheffer et al. 2017]

1.1.1.1 Syndromy spojené s EE

Ohtahara syndrom (early infantile epileptic encephalopathy, EEIE)

Ohtahara syndrom má ze všech syndromů nejčasnější nástup již v prvních dnech života. Záchvaty jsou obvykle tonické, krátkého trvání, objevují se při spánku i bdění. U většiny případů dochází k vývoji směrem k Westovu či Lennox-Gastautovu syndromu. Prognóza je nepříznivá, časté jsou poruchy vývoje mozku, mortalita je 50%, léčba je velmi často neúspěšná. [Murakami, Ohtsuka a Ohtahara 1993]

Časná myoklonická encefalopatie (early myoclonic encephalopathy, EME)

EME je syndromem, který se projevuje již v novorozeneckém období (nástup do prvních měsíců věku) a prognóza je stejně jako u Ohtaharova syndromu velmi nepříznivá. Záchvaty nastupují často již v prvních hodinách po narození a mají myoklonický bloudivý charakter. Frekvence záchvatů je velmi variabilní, dochází ke zpomalení psychomotorického vývoje. [Dalla Bernardina et al. 1982]

Westův syndrom (Infantile spasms)

Westův syndrom patří mezi časné encefalopatie s nástupem obtíží během prvního roku věku. Charakteristické jsou infantilní spasmy a pokud nedochází k rychlému nasazení léčby dojde k regresi postmotorického vývoje. Příčinou vzniku mohou být cerebrální malformace, infekce, metabolické poruchy, či genetické etiologie (často spojené s Downovým syndromem) [Caraballo et al. 2011; Vigevano et al. 1993]. Úspěšnost léčby je často závislá na rychlosti diagnostiky a adekvátnosti terapie.

Dravetové syndrom (severe myoclonic epilepsy in infancy, SMEI)

Těžká myoklonická epilepsie je označována jako syndrom Dravetové. Obvykle začíná v prvním až druhém roce života, záchvaty jsou nejprve generalizované později se ale připojují myklonické záchvaty horních končetin a trupu. Záchvaty mohou být často vyvolány expozicí zdrojem tepla (teplá koupel) nebo světlem (fotosenzitivita). V batolecím období se objevuje ataxie, a regrese vývoje s retardací a autistickými rysy. Etiologicky jde nejčastěji o genetický původ variantou v genu *SCN1A*. [Khan a Al Baradie 2012; Komárek 2007]

Lennox-Gastautův syndrom (LGS)

LGS postihuje průměrně 5 % dětí s epileptickými záchvaty. Nástup onemocnění bývá nejčastěji kolem druhého roku života. Přítomné je široké spektrum záchvatů, obvyklé jsou noční tonické záchvaty nebo atypické absence během dne. LGS je v 90 % případů doprovázen mentální retardací s velmi špatnou prognózou. [Heiskala 1997; Arzimanoglou et al. 2009]

Landau-Kleffnerův syndrom (LKS)

LKS má pozdější nástup, ve čtvrtém až osmém roku života. Postižení jedinci postupně ztrácejí schopnost porozumění mluvenému slovu a později sami nejsou schopní mluvit, tento příznak je dominantní. Po nástupu onemocnění se zvyšuje agresivita a objevují se epileptické záchvaty. U LKS dochází v pozdějším věku k mírnému zlepšení klinických obtíží. [Komárek 2007; Pearl, Carrazana a Holmes 2001]

1.1.1.2 Geny asociované s EE

Při hledání patogenních variant se většinou zaměřujeme na geny dříve asociované s onemocněním. U epilepsií a epileptických syndromů bylo identifikováno mnoho genů se všemi typy dědičnosti. [Wang et al. 2017]

Vznik epileptických encefalopatií bývá většinou geneticky podmíněný a má velmi heterogenní charakter. Varianty způsobující EE bývají nejčastěji nalezeny v genech s AD dědičností vzniklé *de novo*. [Thomas a Berkovic 2014] V roce 2012 a 2013 byly publikovány výsledky analýzy konzorcia Epi4K [Consortium 2012; Consortium et al. 2013], identifikujících nejčastější *de novo* příčiny EE v trio WES vzorcích. V rámci této studie, pak byly publikovány asociace s EE u těchto genů: *SCN1A, STXBP1, CDKL5, GABRB3, SCN8A, SCN2A, CACNA1A, CHD2, FLNA, GABRA1, GRIN2B, IQSEC2, MTOR a NEDD4L*. Žádný z uvedených genů ale není prevalentní, proto je vyšetření pomocí MPS panelem genů (nebo WES) v současnosti nejefektivnějším postupem.

1.2 Hereditární motorické a senzitivní neuropatie (HMSN) / Charcot-Marie-Tooth choroba (CMT)

Hereditární motorické a senzitivní neuropatie jsou skupinou monogenně podmíněných onemocnění postihujících periferní nervovou soustavu. Choroba byla poprvé popsána již v roce 1886 třemi neurology – J.M. Charcotem, P.Mariem a H.H. Toothem a pojmenována jako „Charcot-Marie-Tooth disease“ [Charcot 1886]. Jedná se o nejčastější dědičné neuromuskulární onemocnění s prevalencí až 1:2500. [Skre 1974; Braathen 2012]

Mezi typické příznaky HMSN patří progresivní degenerace periferních nervů rezultující v oslabení distálních svalů nejprve dolních končetin a později i horních končetin. S dalším postupem nemoci vznikají deformity nohou tzv. pes cavus (zkrácení Achillovy šlachy, vysoký nárt a kladívkové prsty na nohou). HMSN podléhá všem typům dědičností – autozomálně dominantní, recesivní, X-vázané i mitochondriální. [Timmerman, Strickland a Züchner 2014; Rossor et al. 2013; Pareyson a Marchesi 2009]

Pro klasifikaci HMSN využíváme poznatků neurofyziologických i histopatologických. Na základě rychlosti vedení vzruchu (MCV) byly rozděleny na demyelinizační formy (skupina HMSN1/CMT1 a CMT4) s $MCV < 38ms^{-1}$, příčínou je primární postižení myelinu. Při axonální formách HMSN (HMSN2/CMT2) je rychlost vedení vzruchu snížená, ale stále nad $MCV > 38ms^{-1}$, jde o primární postižení axonu. [Harding a Thomas 1980]

1.2.1 Demyelinizační formy HMSN (HMSN I)

1.2.1.1 CMT1

CMT1 jsou primárně autosomálně dominantní demyelinizační neuropatie. Příčina onemocnění je silně heterogenní, až 80 % případů je spojováno s *CMT1A* (oblast 17.p11.2). Forma CMT1A tvoří skoro polovinu případů všech CMT (a 60 až 70 % případů ve skupině CMT1). Nejčastěji jde o postižení oblasti 17.p11.2 [Lupski et al. 1991], která obsahuje duplikovaný gen *PMP22*, způsobený mechanismem nealelické homologní rekombinace (NAHR, nonallelic homologous recombination) princip je uveden na Obr. 1.2. [Zhang et al. 2010]

Typický nástup potíží je v první dekádě života, pacient pozoruje poruchy chůze, zakopávání, vysoký nárt. Fenotyp CMT1A je variabilní, část pacientů trpí mírnou formou a je po celý život schopná chůze. Diagnostika tohoto typu onemocnění je nejčastěji indikována po zjištění snížené rychlosti vedení vzruchu (MCV). [Haberlová, Mazanec a Seeman 2006] Prvním krokem při genetické diagnostice je testování na přítomnost duplikace/delece lokusu 17p11.2, pokud je výsledek negativní, zaměřujeme se na SNV v genu *PMP22*. [Rossor et al. 2013]

1.2.1.2 Autozomálně recesivní HMSN I (AR-HMSN I / CMT4)

Autozomálně recesivní demyelinizační neuropatie, jsou v běžných populacích méně časté než AD CMT. Skupina CMT4 je typická závažným fenotypem s brzkým ná-

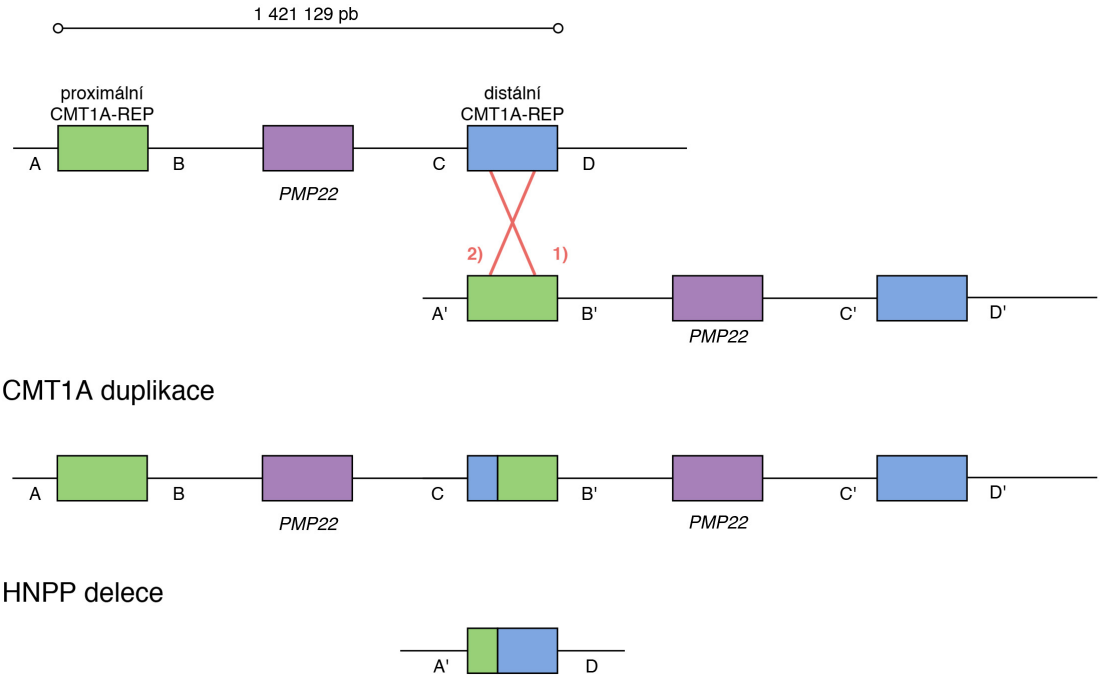


Schéma ukazuje crossing-over dvou nealelických segmentů DNA, které mají vysokou sekvenční homologii (LCR), proximální a distální oblasti CMT1A-REP se shodují v 99 % své sekvence. Dojde tedy k reciproké přestavbě, kdy sekvence se buď:
1) prodlouží - gen PMP22 mezi CMT1A-REP oblastmi je duplikován
2) nebo dojde k HNPP deleci, kdy je celá sekvence mezi LCR je deletována
Převzato z: [Lupski 2015]

Obrázek 1.2: Nealelická homologní rekombinace v oblasti 17.p11.2

stupem oproti AD CMT. Nemoc se projevuje u potomků zdravých rodičů a je pro ní typické, že sourozenci mají podobný fenotyp. Jsou často spojované se populacemi, u kterých je vyšší výskyt pokrevního příbuzenství. [Reilly, Murphy a Laurá 2011]

CMT4A je způsobená nejčastěji bodovými variantami v genu *GDAP1*. Nástup nemoci je velmi časný se závažným fenotypem – deformity nohou, progresivní slabosti a ztráta senzitivity nejdříve v dolních (v první dekádě) a poté i v horních končetinách. Do třicátého roku života jsou pak pacienti většinou upoutání na invalidní vozík. [Nelis et al. 2002]

CMT4B1 a **CMT4B2** jsou skupiny asociované s velmi brzkým nástupem proximální i distální slabostí a obličejovou slabostí. Příčinou CMT4B1 jsou nejčastěji SNV v genu *MTMR2*. Příčinou CMT4B2 jsou bodové varianty v genu *SBF2* (*MTMR13*). Demyelizační neuropatie začíná v dolních končetinách v první dekádě, má pomalý progres a je doprovázena deformitami nohou. Často se u tohoto typu onemocnění objevuje glaukom. [Azzedine et al. 2003] V rámci mezinárodní spolupráce byla na našem pracovišti publikována studie sedmi rodin, u kterých byly identifikovány dříve nepublikované varianty (další komentář v 4.2.3.1), jde o dosud největší popsanou kohortu pacientů, s detailním popisem fenotypu. [Laššuthová et al. 2018]

CMT4C je recesivně děděná forma CMT, způsobená variantami v genu *SH3TC2*. Typický fenotyp je charakterizován časným projevem onemocnění, demyelizační for-

mou neuropatie, s progresivní skoliózou. Varianty v genu *SH3TC2* se ukázaly jako velmi častá příčina HMSN I v české populaci, kdy v kohortě všech pacientů s CMT se varianta p.Arg954Stop našla v téměř dvou procentech případů (a u 95 % všech CMT4C). [Laššuthová et al. 2011]

CMT4D (HMSNL) je recesivně děděná demyelinizační neuropatie, která postihuje pacienty romského původu. CMT4D je způsobena variantou p.Arg148* v genu *NDRG1*.U romských pacientů v ČR jde o druhou nejčastější formu CMT (frekvence výskytu u romské populace byla 24 %, u celé kohorty pacientů s CMT byla frekvence 0,66%). [Brožková et al. 2017]

CMT4G (HSMNR) je další skupinou, která v rámci ČR postihuje romskou populaci. Je způsobená variantami v genu *HK1* (g.9712G>C) [Hantke et al. 2009b]. V populaci českých Romů se jedná o nejčastější formu CMT, dvakrát častější než CMT4D, v rámci našeho pracoviště jsme detekovali 20 pacientů s homozygotní g.9712G>C variantou v genu *HK1*. U celé kohorty pacientů se jednalo o 6. nejčastější příčinu CMT v naší populaci s výskytem 1,1 % (ale nejčastější pro populaci romskou). [Brožková et al. 2016]

1.2.1.3 CMT3/HMSN III Déjerine-Sottas syndrom

Jedná se o velmi časnou a těžkou formu CMT1, distální svalová slabost se projevuje v raném dětství. Klinicky jsou pacienti hypotoničtí, mají deformity nohou a také skoliózu páteře. Může se vyskytnout i porucha sluchu.[Mazanec et al. 2004] Příčinou jsou často *de novo* vzniklé dominantní varianty, nejčastěji se vyskytující v genech *PMP22* [Seeman et al. 2002], *MPZ* [Mazanec et al. 2004], *EGR2* [Mikesová et al. 2005].

1.2.1.4 X-vázané CMT (CMTX)

X-vázané CMT jsou druhé nejčastější, tvoří přibližně 10 % všech případů. Nejčastější příčinou onemocnění jsou varianty v genu *GJB1* (*Cx32*), kde bylo identifikováno více než 400 patogenních variant. [Panosyan et al. 2017] Gen *GJB1* je krátký gen, kódující protein connexin 32 dlouhý 238 AMK. Connexin 32 je protein tvořící intracelulární kanály pro přenos iontů a malých molekul mezi perinukleárním a periaxonálním kompartmentem Schwannových buněk. Změna funkce proteinu vede k poškození myelinové pochvy axonu a tím k axonopatii. [Seeman et al. 2000]

Jak vyplývá z gonozomální, X-vázané dědičnosti, muži jsou postiženi dříve a výrazněji než ženy. Některé ženy jsou dokonce bez fenotypu. Přenos nemoci v rámci rodokmenů odpovídá X-vázané dědičnosti – absentuje přenos z otce na syna, fenotyp žen je mírnější a nástup obtíží je pozdější než u mužů. Hodnota MCV je vyšší než u typu CMT1A (pohybuje se mezi $30 - 40 ms^{-1}$. [Seeman et al. 2000; Haberlová, Mazanec a Seeman 2006]

1.2.2 Axonální formy HMSN (HMSN II)

1.2.2.1 CMT2

CMT2 je skupina axonálních neuropatií, dědičnost CMT2 je autozomálně dominantní s nástupem v první až druhé dekádě. Oproti CMT1(A) bývá svalová slabost

výraznější, jsou přítomny atrofie. Deformity nohou jsou ale méně časté. Rychlost vedení vzruchu je vyšší než u CMT1, $MCV > 38ms^{-1}$, ale je redukována amplituda akčního potenciálu. [Harding a Thomas 1980]

Nejčastějším typem je CMT2A, tvořící téměř čtvrtinu všech případů CMT2. Jako příčina CMT2A byly identifkovány varianty v genu *MFN2* [Züchner et al. 2004; Lawson, Graham a Flanigan 2005; Cartoni a Martinou 2009]. Frekvence detekovaných variant byla v kohortě naších pacientů 7,2 % [Brožková et al. 2013]. V genu *MFN2* jsme také nalezli pomocí panelu genů variantu, která dříve nebyla Sangerovým sekvenováním detekována (primer mismatch). Tato varianta se nachází na pozici p.His361Tyr a byla dříve publikována [Verhoeven et al. 2006]. V této publikaci tak ukazujeme,že je potřeba myslet na možnost primer mismatch u Sangerova sekvenování. [Neupauerová et al. 2016]

Mezi autozomálně recesivní axonální formy CMT je řazen i typ CMT2B1 (gen *LMNA*), s velmi variabilním fenotypem i časem nástupu. [Benedetti et al. 2007] Gen *LMNA* byl označen jako hlavní příčína typu CMT2B1. V naší kohortě s axonálním typem CMT a brzkým nástupem jsme otestovali 98 pacientů, u kterých byla indikována AR CMT2. Z 98 pacientů byla nalezena pouze jedna potenciálně patogenní varianta v genu *LMNA*, na pozici c.1870C>T v heterozygotní formě, druhou variantu jsme ale v genu nanalezli. U všech pacientů byla provedena i MLPA analýza na přítomnost delecí a duplikací. Gen *LMNA* nebyl tedy prokazatelnou příčinou CMT2B1 u českých pacientů.[Lassuthová et al. 2009]

Dalším genem způsobujícím CMT2 je gen *MORC2*, kdy první varianty byly identifikovány pomocí WES u pacientů s významnou slabostí dolních, ale i horních končetin a hypotonie. Varianta p.Arg190Trp byla identifikována u pacientů různých národností [Sevilla et al. 2015]. Ve skupině našich pacientů byla tato varianta nalezena u šesti nepříbuzných rodin, kazuistika je publikována v [Laššuthová et al. 2016b].

1.3 Genetická variabilita na DNA úrovni

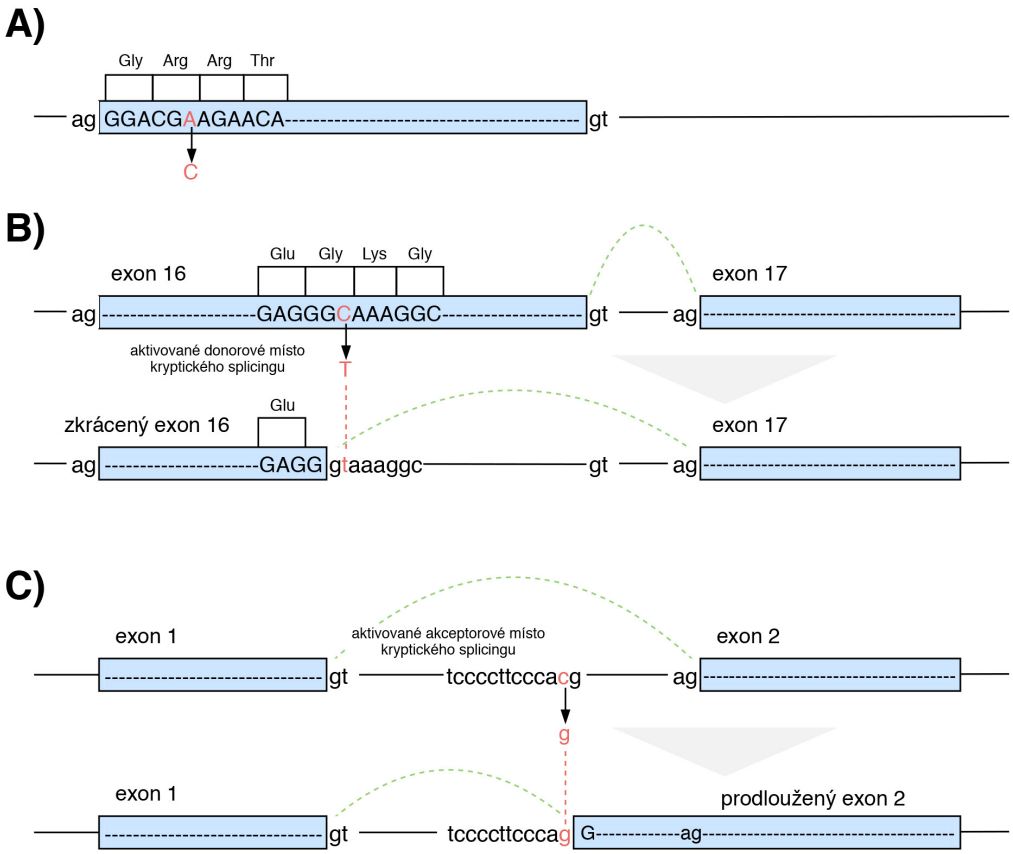
Změny v sekvenci DNA jsou přítomné v celém genomu, většina variant spadá do intergenové či intronové oblasti, která tvoří většinu genomu. Takové varianty jsou pak často dobře tolerované, bez fenotypového projevu. Navíc i některé geny jsou redundantní a tak změna v jedné kopii je dobře kompenzována jinou.

Záměny jednoho nukleotidu za jiný označujeme je jako SNV (Single nucleotide variants). Tyto jednoduché záměny se vyskytují v průměru jednou za 800 nukleotidů u každého jedince. Pokud takové varianty mají četnost v populaci vyšší než 1 % jsou označeny za SNP (Single nucleotide polymorphism) pro danou populaci. Zatímco SNP jsou ve většině případů benigní a nemají žádné známé asociace s onemocněním u mnoha SNV tato spojení nalézáme. Důležité je poté určit, jestli varianta vznikla v zárodečné fázi „germline varianty“ nebo až později během života organismu „somatické varianty“. U SNV rovněž rozlišujeme efekt, který záměna nukleotidu způsobí. [Strachan 2014]

1.3.1 Synonymní a nesynonymní typy variant

Synonymní SNV Nahrazením jednoho nukleotidu v kódující sekvenci za jiný dochází k změnám v tripletu – tedy kodonu. Díky redundanci v genetickém kódu to ale nemusí vždy znamenat, že dojde ke změně aminokyseliny (AMK) v polypeptidickém řetězci – pokud ke změně AMK nedojde jedná se o synonymní substituci. Pokud taková synonymní substituce neovlivňuje dále normální sestřih RNA, lze předpokládat, že nebude mít fenotypový projev. Pravděpodobnost nesynonymní substituce je závislá na pozici báze v rámci kodonu, na třetí pozici je pravděpodobnost nejnižší (pouze u jedné třetiny substitucí dochází ke změně aminokyseliny). Větší pravděpodobnost změny je pak u substitucí první a druhé báze kodonu – 100% na druhé pozici a 96% na první pozici (synonymní u Argininu a Leucinu). Synonymní substituce ale nemusí být vždy bez efektu na fenotyp. Synonymní substituce může v řetězci nahradit nukleotid tak, že nedojde k nahrazení AMK, ale dojde k aktivaci kryptického sestřihového místa (cryptic splice site) – vytvoří se sekvence rozhraní exon-intron a tím dojde ke zkrácení exonu se závažným fenotypem, tento jev ale může nastat i v opačném případě, kdy bodová varianta vznikne v intronu a tím vytvoří kryptické splicové místo – exon se pak naopak prodlouží (přehled efektů na Obr. 1.3). [Richard a Beckmann 1995]

Nesynonymní SNV –missense Nesynonymní substituce nemusí ale vždy znamenat závažný fenotyp, v tomto případě má velkou váhu k jaké změně aminokyseliny dojde. Pokud dojde k nahrazení AMK za chemicky podobnou nemá velký vliv na funkci proteinu (přibližně 30% pravděpodobnost). Dalším aspektem je důležitost AMK v řetězci polypeptidu např. pokud dojde ke změně aktivačního místa enzymu, může i podobná AMK měnit funkci, může docházet ke strukturálním změnám proteinů a výsledný efekt může být různý. Příkladem může být změna AMK za velmi odlišnou, kdy dojde ke změně hydrofobicity, tím dojde k přestavbě celého genového produktu a vyřazení jeho funkce. [Maquat 2001]



(A) Nahrazení A>C, nevede ke změně AMK, nemá efekt
(B) Synonymní substituce T>C i přes to, že jde o synonymní substituci vytváří nové kryptického sestřihové místo, dochází ke zkrácení exonu 16
(C) Aktivace kryptického místa v intronu mezi exony 1 a 2 provedla exonizaci intronu a tím prodloužení exonu;
Převzato z: [Strachan 2014]

Obrázek 1.3: Synonymní substituce a jejich efekt

Nesynonymní SNV – nonsense Nonsense varianty predikují záměnu původní AMK za terminační kodon a zkrácení proteinu. Terminační kodon je místo, kde se ribozom odpojuje od mRNA při translaci, jedná se o sekvenci tří nukleotidů (UAA, UAG, UGA), mnoho patogenních variant je ale spojováno s vytvořením nového terminačního kodonu, který má za následek tvorbu nekompletního polypeptidu. Výsledkem pak může být genový produkt kratší délky. Efekt této záměny závisí na tom, jak velká část produktu byla odstraněna a na stabilitě nově vzniklého produktu. Proti tomuto efektu probíhá v buňce pretranslační proces nonsense-mediated mRNA decay (NMD). Ten eliminuje defektní mRNA ještě před vytvořením proteinového produktu. V cytosolu probíhá první testovací kolo translace, při kterém ribozom překládá mRNA a „kontroluje“ přítomnost Exon junction complex (EJC), což jsou komplexy přítomné na exonech, mezi start a stop kodonem. Pokud ribozom narazí na stop kodon, kterým by ukončil translaci, ale zároveň je dále na sekvenci přítomný EJC, dojde k označení a degradaci mRNA. Tento proces tak umožňuje buňkám efek-

tivně bránit vzniku aberantních, krátkých proteinů, které by mohly mít za důsledek vážný fenotyp. [Kulkarni a Pfeifer 2014]

Varianty způsobené posunem čtecího rámce - frameshift Frameshift varianty mohou též vést ke vzniku terminačního kodonu, ale ne přímo. Jde o situaci, kdy do sekvence DNA je insertováno nebo deletováno několik nukleotidů. Pokud dojde ke změně počtu nukleotidů v řetězci, dochází k tvorbě zcela jiného polypeptidového řetězce a je velká pravděpodobnost narušení funkce výsledného produktu. Výjimkou jsou delece/inzerce tří nukleotidů, kdy čtecí rámec je zachován, ale dojde k odstranění/přidání jedné AMK. Pokud dojde ke vzniku SNV na rozhraní exon-intron, může dojít k narušení RNA splicingu – k tomu dochází nejčastěji při narušení sekvence „GT AG“. Výsledkem pak může být intronizace exonů (exon skipping) nebo exonizace intronů (intron retention), oba tyto stavy mohou mít velmi variabilní fenotyp. [Krawczak, Reiss a Cooper 1992]

1.3.2 Varianty ovlivňující aberantní RNA sestřih a varianty v regulatorních oblastech

Varianty měnící sestřih RNA tvoří přibližně 15 % všech variant. Jedná se o SNV, které postihují sestřihové 5' nebo 3' oblasti, čímž může dojít k zániku nebo vzniku nového sestřihového místa. [Mendell a Dietz 2001]

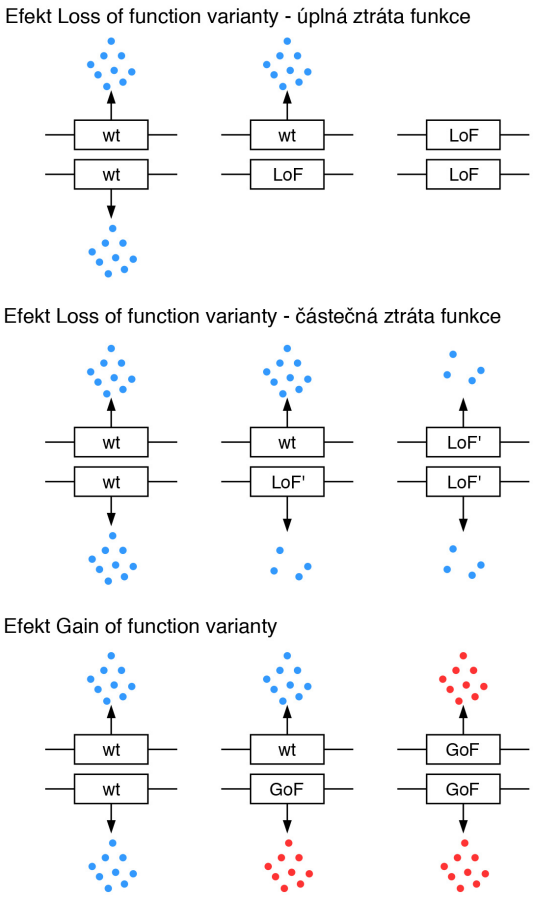
Většina bodových záměn se nalézá v kódujících oblastech genů, pokud ale substituce vznikne mimo tuto oblast, neznamená to, že nemůže mít žádný efekt. Záměny v regulatorních oblastechm jako jsou promotory, enhancerové oblasti, represory mohou ovlivňovat expresi genu a tím i způsobovat fenotyp. Příkladem přímého efektu jsou např. varianty v TATA boxu (promotor), způsobující beta-thalasémii. Kromě toho bývají SNV v těchto oblastech často označovány za faktory multifaktoriální dědičnosti. [Gonzalez-Redondo et al. 1989]

1.3.3 Efekt genetické variability na funkci proteinu

Výsledkem substituční změny jednoho nukleotidu může být změna sekvence genového produktu, čímž dojde i ke změně jeho funkce. Může dojít ke ztrátě funkce (loss of function), kdy produkt není schopen plnit svojí funkci nebo naopak k získání funkce nové (gain of function), která může způsobit i smrt buňky. Dalším možným efektem je změna množství produktu. Efekty jsou uvedené na obrázku Obr. 1.4.

Každá varianta může mít na výsledný produkt genu jiný efekt. Někdy dojde ke změnám v množství exprimovaného produktu, někdy může dojít k tomu, že produkt má zcela jinou funkci, nebo funkci ztratí. V případě Loss of function (LoF) varianty došlo ke změně nukleotidu na jedné alele, dochází k produkci menšího množství produktu. V tomto případě je často buňka schopná nedostatečné množství kompenzovat - buď jí polovina produktu stačí k funkčnostim nebo v některých případech dochází k vyšší produkci původního produktu wild-type alelou. Některé varianty ale mohou mít dominantní charakter, nastává případ, kdy buňka nedokáže kompenzovat dostatečné množství genového produktu, poté dochází k projevu fenotypu.

Gain of Function (GoF) je efekt genetické variability, kdy záměna nukleotidu vytváří zcela nový genový produkt, který je ale funkční. V případě heterozygotní varianty dochází k produkci nového genového produktu, který se projeví novým, dominantním fenotypem. [Griffiths et al. 2005]



A) V prvním případě nastává u homozygotní varianty úplná ztráta funkčnosti produktu, rezultující v odlišný fenotyp, u heterozygotní varianty je fenotyp často shodný s wt jelikož množství produktu stačí ke kompenzaci ztráty jedné alely (tento efekt je častý např. u metabolických poruch)
B) Ve druhém případě došlo ke částečné ztrátě funkčnosti produktu, i tak ale nedojde v případě homozygotní varianty k dostatečné kompenzaci a projeví se odlišný fenotyp
C) Případ GoF varianty, v případě heterozygotní varianty získává buňka nový produkt, často s dominantním efektem (tento efekt je častý např. u epilepsií). homozygotní formu provází často velmi vážná změnu fenotypu, není příliš častá
Převzato z: [Griffiths et al. 2005]

Obrázek 1.4: Efekt LoF a GoF variant

U AD onemocnění může být postižená alela variantou LoF i GoF, pokud se jedná o GoF, tak přestože jedna alela vytváří zdravý produkt, tak efekt poškozeného produktu z alterované alely převáží a fenotyp se projeví. U LoF variant je situace trochu

složitější, projev fenotypu je závislý na potřebné dávce produktu. Pokud je dávka od jedné alely dostatečná, onemocnění se neprojeví. Dalším příkladem, kdy se LoF projeví je při imprintingu Angelmanova syndromu – jedna LoF varianta v genu *UBE3A* na maternální alele v tomto případě způsobí fenotyp, neboť paternální alela je v tomto místě imprintovaná. Fenotypy způsobované LoF variantami jsou často spojované s heterogeními onemocněními. Výjimkou ale nemusí být kombinace LoF a GoF v jednom genu způsobující odlišný fenotyp než pouze jednotlivé varianty např. u duplikací a delecí oblasti 17p11:2 (*PMP22* gen asociovaný s HMSN). [Strachan 2014]

1.3.4 Inserce delece a změny v počtu kopií úseků (CNV)

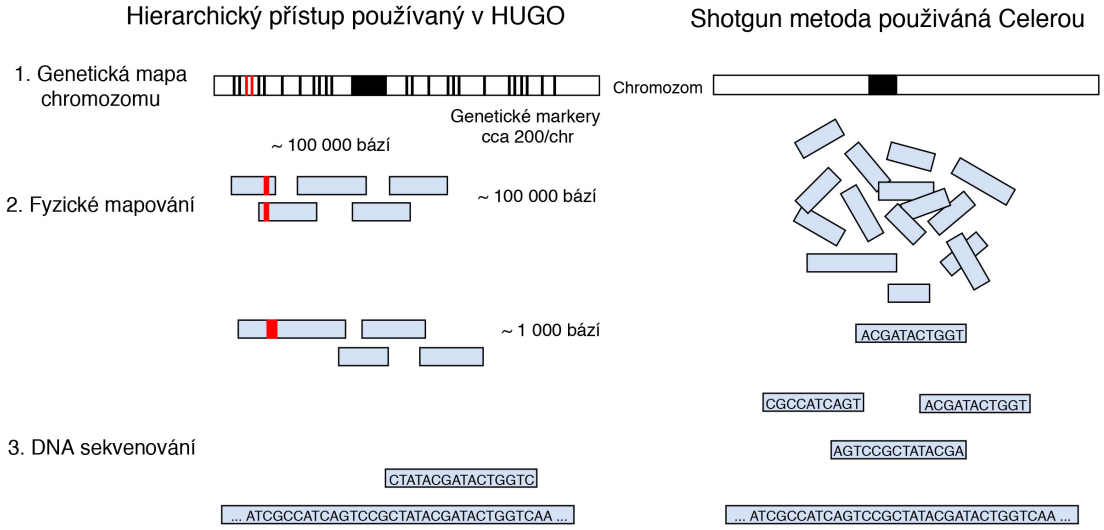
V některých případech může dojít k vytvoření variace DNA, kdy se finální počet nukleotidů liší - dochází k insercím nebo delecím nukleotidů. Pokud se jedná pouze o několik jednotek až desítek nukleotidů, označujeme tyto varianty jako indel (do 50 nukleotidů). U delších insertovaných / deletovaných sekvencí, délky většinou přesahující sto nukleotidů až po tisíce mluvíme o změnách v počtu kopií (Copy number variation, CNV). Frekvence delecí a insercí je cca desetinová oproti SNV, kdy 99 % tvoří krátké indely a jedno procento CNV.

V DNA laboratoři se nejčastěji setkáváme s CNV v případě duplikace / delece na chromozomu 17p.11.2-12, včetně genu *PMP22,* u CMT. K deleci nebo inzerci dochází mechanismem NAHR, kdy dojde při crossing-overu k překrytí blízkých homologních oblastí, a tím dojde k deleci nebo inzerci další kopie genu. Princip jsme popsali na obrázku Obr. 1.2.

1.4 Projekt lidského genomu

Se sekvenací prvních lidských genů přichází myšlenka osekvenovat celý lidský genom. Tyto myšlenky se začaly prosazovat na konci 80. let 20. století i přesto, že v tu dobu ještě nebyla známá technologie, která by tohoto byla schopná. V roce 1990 byl zahájen projekt lidského genomu HGP, kdy cílem bylo do 15 let získat celou DNA sekvenci člověka. Participovat na celém projektu měly laboratoře z celého světa a rovněž bylo ustanoveno, že všechny přečtené sekvence musí být do jednoho dne dostupné na internetu. [Mapping a Human Genome 1988]

Cílem projektu ale nebylo osekvenovat pouze lidský genom, ale i genom dalších pěti organismů – bakterie E.coli, kvasinky, háďátka, octomilky a myši. Z tohoto rozhodnutí těžíme dodnes, kdy znalost genomu těchto organismů nám pomáhá nejen porozumět funkcím jednotlivých genů, ale umožňuje i tyto organismy využít jako modelové pro testování efektu variant na organismus. Projekt lidského genomu zvolil postup, kdy prvním krokem bylo vytvoření map chromosomů – fyzické a později genetické na principu pravděpodobnosti crossing-overu. Jednotlivé úseky pak byly hierarchicky rozděleny na fragmenty pomocí restrikčních endonukleáz a uloženy do umělých bakteriálních chromozomů (BAC) o velikosti 150 000 až 200 000 pb. Při zajištění co nejmenšího překrytí fragmentů se tak jednalo o toho času nejlepší metodu pro přípravu a množení úseků k sekvenaci. Pro samotnou sekvenaci je pak nutné rozdělit genetickou informaci uloženou v BAC na další, menší fragmenty o délce řádově tisíců bází. Jednalo se o pomalou, ale velice přesnou a spolehlivou metodu, která vedla k výsledku. [Lander et al. 2001]



Převzato z: [Vácha 2016]

Obrázek 1.5: Srovnání hierarchické a shotgun metody sekvenování

Jako protiváha k tomuto projektu pak působila společnost Celera, v čele s C. Venterem, která lidský genom sekvenovala metodou „whole-genome shotgun“. Oproti hierarchickému dělení na „fragmenty fragmentů“ tato metoda genom dělí na mnoho

malých částí, které jsou osekvenovány a poté skládány pomocí počítačových algoritmů . [Venter et al. 2001] Srovnání obou metod je na obrázku Obr. 1.5 převzatého z URL[1].

Díky konkurenci obou skupin tak dochází ke značnému urychlení celého projektu, v červnu 2000 je oznámeno zdárné ukončení projektu, v únoru 2001 vychází v časopisech Nature [Lander et al. 2001] a Science [Venter et al. 2001] články s prvním konceptem sekvence. Konečná sekvence je pak oznámena v květnu 2003 a publikována v říjnu 2004 [Consortium 2004].

[1] http://www.zo.utexas.edu/faculty/sjasper/images/20.13.gif [online: 16.10.2019]

1.5 Masivně paralelní sekvenování (MPS)

Vývoj metod sekvenování DNA se datuje od 70. let 20. století, kdy sekvenování bylo velmi obtížné a provádělo se pouze po převodu molekuly do RNA, to se změnilo s příchodem metod první generace – Maxam-Gilbertovy [Maxam a Gilbert 1977] a metody Sangerovy [Sanger, Nicklen a Coulson 1977]. Sangerova enzymatická metoda s použitím terminačních dideoxynukleotidů umožnila první komerční aplikace sekvenačních metod. I přesto ale šlo o metody velmi náročné časově i finančně a to zejména proto, že se musí sekvenovat po jednotlivých bázích. Díky těmto metodám bylo možné osekvenovat lidský genom, ačkoliv celý projekt trval více než deset let.

Celé odvětví ale změnil příchod sekvenátorů druhé generace, nazývaných rovněž „Next generation sequencing“, který umožnil paralelní zpracování více molekul současně. Tím dochází k rapidnímu snížení časové náročnosti a nákladů. Metody se vyznačují perfektní škálovatelností - jsme schopní osekvenovat pár exonů, ale i celý genom najednou. Díky těmto metodám je možné dnes získat sekvenci genomu u probandů v řádu dní za částky kolem 1 000 $. S možností získat sekvenci celého genomu se ale pojí větší nároky na zpracování dat. Díky MPS dokážeme generovat obrovská množství dat, která je nutné „správně zpracovat“. Od získané sekvence k nalezení té pravé varianty (způsobující onemocnění) vede dlouhá cesta, která není pevně definována a skýtá mnohá úskalí. Pro vyhodnocení dat dnes existují stovky nástrojů a je vždy nutné přesně definovat postup a parametry celého postupu. [Zhou et al. 2010]

V současné době na trhu dominuje platforma Illumina, kterou využíváme i na našem pracovišti.

1.5.1 Platforma Illumina

Sekvenační platforma Illumina (Illumina, USA) byla vyvinuta v roce 2006 na univerzitě v Cambridge společností Solexa. Technika je založená na využití fluorescenčně značených bází, které jsou připojovány na sekvenovaný řetězec. Metodika má široké možnosti využití od sekvenování celého genomu až po RNA sekvenování.

Celý proces je založen na třech hlavní krocích, amplifikaci, sekvenování a analýze. Vstupem do procesu je purifikovaná DNA, která prochází fragmentací na úseky o velikosti kolem 200bp, úseky jsou pák dále zarovnávány.

Na takto připravené fragmenty jsou připojeny adaptory. Adaptory jsou sekvence, které obsahují specifické bloky pro sekvenaci – vazebnou sekvenci umožňující navázání komplementárního řetězce, tzv. index, který umožňuje identifikaci úseku čtení – readu, a sekvenci pro připojení k flowcell.

Každý fragment DNA opatřený adaptorem se připojuje ke komplementárnímu konci na reakční komoru, tzv. flowcell, kde pak probíhá samotná amplifikace. Amplifikace probíhá shlukově, to znamená, že sousední oligonukleotidy spolu sdílejí amplifikovaný řetězec (resp. adaptory, řetězce jsou komplementární vždy se dvěma sousedními konci). Tímto způsobem pak DNA polymeráza vytváří v jednotlivých clusterech kopie stejného řetězce. Po vygenerování clusterů jsou odmyty řetězce sekvenované reverzním primerem a dojde k jeho zablokování, aby nedošlo k dalšímu připojení k vedlejší sekvenci. Tímto vzniká „obnažený řetězec“, připojený na flow-

cell pouze jedním koncem, připravený k sekvenaci.

Sekvenace pak probíhá připojováním speciálních nukleotidů značených fluorochromem na templátový řetězec. Díky tomu dokáže sekvenátor rozlišit, jaká báze byla do řetězce připojená. Po přečtení celé sekvence dojde k odstranění blokované sekvence a tím k navázání na sousední imobilizovaný nukleotid – nastává pak druhé kolo sekvenace, tentokrát v opačném směru na stejném principu.

Výsledkem jsou pak přečtené jednotlivé úseky sekvence, které seřadíme dle indexů obsažených v adaptoru. Výhodou této metody je její vyšší přesnost a možnost detekce inzercí a delecí, což by v případě sekvenace pouze jednoho směru nebylo možné. Další výhodou je adaptabilita metody, kdy můžeme volit nejen délku sekvence, ale i hloubku čtení. Nevýhodou řešení je pak vyšší chybovost a nerovnoměrné pokrytí (zejména v AT a GC bohatých sekvencích) [Shokralla et al. 2012; Xuan et al. 2013].

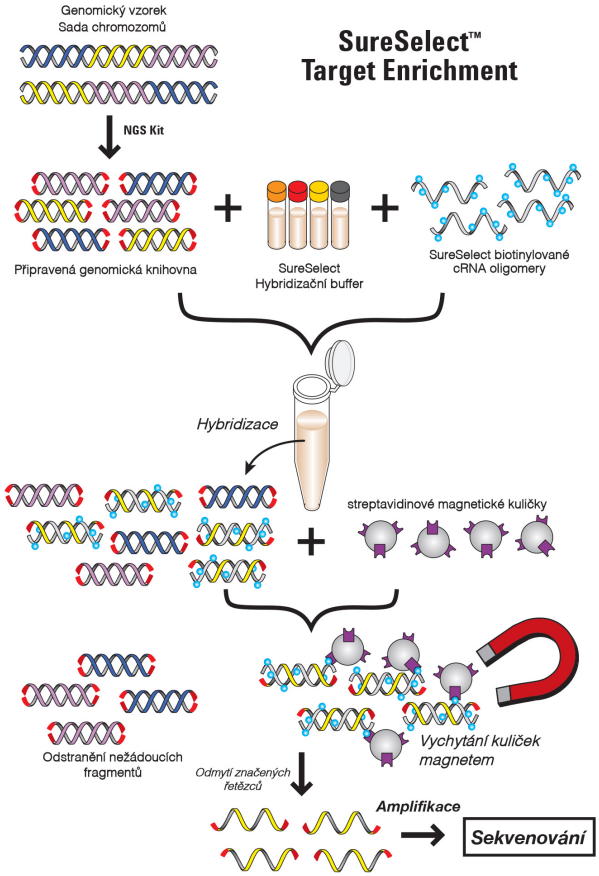
1.5.1.1 Cílené obohacování oblastí

Jak již bylo dříve zmíněno, tak masivně paralelní sekvenování umožňuje sekvenaci celého genomu. Někdy je ale vhodné se pomocí upravené knihovny zaměřit pouze na určité oblasti a to nejčastěji na všechny kódující oblasti všech genů (WES) nebo pouze na panel genů dle našeho zájmu (které předem vybereme).

V takovém případě je nutné nejprve vytvořit design knihovny – výrobci nejčastěji nabízí již předpřipravené knihovny pro WES a v některých případech i panely pro různá onemocnění. Z naší praxe se osvědčilo využití předpřipravené knihovny pro WES a vlastního návrhu panelu genů. To nám umožňuje značnou flexibilitu, co se týče přidávání a odebírání genů dle potřeby až do naplnění kapacity panelu. Po vytvoření knihovny získáme soubory s informací, které oblasti budou pokryty – nejčastěji se jedná o soubor formátu BED.

Po tomto kroku dochází k sekvenování s cíleným obohacováním oblastí, schéma na Obr. 1.6, převzato z URL[2]. Díky tomu můžeme na jednu flowcell vložit i více než 100 vzorků. Abychom ale jednotlivé vzorky mohli odlišit, je nutné přidat ke každému vzorku vlastní identifikátor, který se vloží mezi adaptér a fragment DNA (jedná se o specifický oligonukleotid označovaný jako barcode). Aby došlo k oddělení nežádoucí DNA je směs fragmentů smíchána s RNA knihovnou dle našich požadavků, a dojde k hybridizaci a zachycení pouze fragmentu komplementárních k RNA knihovně (na bázi magnetických korálků). Z této směsi se pak odstraní RNA sekvence a zbydou pouze cílové fragmenty DNA, u kterých přistupujeme k sekvenování [Mamanova et al. 2010; Shen et al. 2005].

[2]https://www.agilent.com/cs/library/usermanuals/Public/G7530-90000.pdf [online: 16.10.2019]



Převzato z URL: http://hpst.cz/molekularni-biologie/strand-specific-rna/princip-metody-sureselect-0

Obrázek 1.6: Schéma cíleného obohacování oblastí u knihovny SureSelect (Agilent, USA)

1.6 Bioinformatické zpracování dat v DNA laboratoři

1.6.1 Datové formáty

Abychom mohli data z MPS zpracovávat, musíme zvolit vhodnou formu, které bude rozumět jak počítač, tak i poté expert, který bude výsledek z počítače analyzovat.

Pojem data tedy můžeme chápat jako symbolickou reprezentaci informace tak, aby byla systematicky čitelná používanými algoritmy. V praxi pak využíváme konceptu, kdy na zdrojová data aplikujeme bioinformatické nástroje, které nám poskytnou odpověď na naší otázku.

Je ale nutné podoktnout, že vhodný datový formát je pro každý krok analýzy různý. Pro počítačové zpracování sekvence a párování s referenčním genomem je vhodnější zvolit formát, který nebude tolik srozumitelný pro běžného uživatele, ale bude lépe „pochopitelný" pro počítač. Pro alignment a variant calling jsou využívány formáty FASTQ a BAM, které by běžný uživatel bez dalšího zpracování nedokázal vyhodnotit. Oproti tomu výsledek celé analýzy, VCF, je tabulka variant, která je pro člověka snadno srozumitelná ale naopak náročnější pro další strojové zpracování.

1.6.1.1 FASTQ/FASTA formáty

Formální definice FASTQ a FASTA formátů byla čerpána dle: [Cock et al. 2010].

Běžná bioinformatická analýza začíná získáním FASTQ souborů ze sekvenátorů. Pro každý vzorek získáváme dva soubory, jeden pro forward sekvenci a druhý pro revers sekvenci. Pro zpracování jsou nutné oba soubory.

FASTQ formát se stal standardem, pro zprácování dat ze sekvenátorů, jedná se o datový formát, který nám umožňuje zkombinovat informaci o přečtené sekvenci s kvalitou čtení každé báze. Dozvídáme se tak, s jakou pravděpodobností je read chybný.

Datový formát je definován 4 sekcemi, pro každý záznam:

- Záznam vždy začíná znakem „@", který označuje hlavičku záznamu, identifikující daný read
- Další sekce obsahuje přečtenou sekvenci, začínající novým řádkem a končící znakem „+"
- Třetí sekce začíná od znaku „+" a může obsahovat další identifikátory, většinou ale bývá prázdná
- Poslední část je stejně dlouhá jako sekvence a obsahuje právě údaje o kvalitě čtení bází pomocí Phred skóre

Phred skóre Phred skóre je ukazatel popisující pravděpodobnost chybného přečtení dané báze. Hodnota se vypočítává podle vztahu:

$$Q = -10 \log_{10} P \tag{1.1}$$

Kde Q je výsledné skóre a P je pravděpodobnost chyby. Výsledné hodnoty pak odpovídají následující tabulce v Obr. 1.7.

Jelikož se ale snažíme o co nejkompaktnější zobrazení, tak se hodnoty dále kódují, aby se daly reprezentovat vždy pouze jedním znakem. Například pokud využíváme platformu Illumina, báze byla přečtená s přesností 99,9 %, je Phred skóre 30, abychom ale „ušetřili“ znak, kódujeme hodnotu podle tabulky na znak „?“. Jednoznakové kódování je rovněž důležité pro rozlišení sledu hodnot. Pokud bychom měli například za sebou tři báze, s hodnotami Phred 32 25 a 8, nelze tyto hodnoty uložit jako 32258 - z takové informace nelze zpětně odvodit jestli se jednalo o dvě nebo o tři hodnoty a jaké byly (kombinace 3, 22, 58 nebo 32, 25, 8, nebo 32, 2, 58). Proto je vhodné reprezentovat tyto hodnoty vždy jedním znakem, pravděpodobnostní a kódovací tabulky jsou uvedené na schématu Obr. 1.7.

Phred Skóre	Pravděpodobnost neprávného vyvolání	Přesnost vyvolání báze
10	1 z 10	90%
20	1 z 100	99%
30	1 z 1000	99,9%
40	1 z 10 000	99,99%
50	1 z 100 000	99,999%
60	1 z 1 000 000	99,9999%

Phred skóre	Kódovací znak	Phred skóre	Kódovací znak	Phred skóre	Kódovací znak	Phred skóre	Kódovací znak
0	!	**11**	,	**22**	7	**33**	B
1	"	**12**	-	**23**	8	**34**	C
2	#	**13**	.	**24**	9	**35**	D
3	$	**14**	/	**25**	:	**36**	E
4	%	**15**	0	**26**	;	**37**	F
5	&	**16**	1	**27**	<	**38**	G
6	'	**17**	2	**28**	=	**39**	H
7	(	**18**	3	**29**	>	**40**	I
8	)	**19**	4	**30**	?		
9	*	**20**	5	**31**	@		
10	+	**21**	6	**32**	A		

Obrázek 1.7: Tabulka hodnot Phred skóre a kódovací tabulka pro Illumina platformu

FASTA formát FASTA formát se od FASTQ formátu liší právě nepřítomností informace o kvalitě. Jde o formát, pomocí kterého nejčastěji reprezentujeme referenční sekvence. Záznam začíná znakem „>“ a je následován identifikátorem, na dalším řádku pak začíná samotná sekvence, která může být i přes několik řádků. Sekvence je popisována velkými písmeny, repetitivní sekvence pak písmeny malými.

1.6.1.2 SAM/BAM - Sequence alignment map / Binary alignment map

Formální definice SAM a BAM formátů byla čerpána dle: [Li et al. 2009].

Zkratka SAM znamená „Sequence alignment / map format“. Jedná se o soubor obsahující hlavičku a samotné záznamy obsahující sekvenci namapovanou na referenci. V praxi je častěji používaný BAM soubor, který je oproti SAM souboru komprimovaný a indexovaný. Indexování souboru zajišťuje jeho zarovnání do menších bloků a následnou kompresi. Abychom mohli snadno přistupovat k požadovanému bloku, je nutné vytvořit index, který spravuje práci se souborem (to znamená, že nás dokáže „nasměrovat do správného místa“ velkého souboru). Co se týče obsahu tak se soubory SAM a BAM neliší, liší se ve velikosti, jakou zabírají v paměti počítače, BAM je díky kompresi menší, ale zároveň díky tomu je práce s ním pomalejší.

Hlavička souboru je označena znakem „@“ a obsahuje informace o:

- formátu souboru, verze a jestli byl BAM seřazen dle koordinát – řádek HD
- použité referenční sekvenci, název sekvence, její délku, další informace o použitém genomu – řádek SQ
- identifikaci vzorku, tyto informace jsou předem definované během alignmentu, jde například o ID pacienta, sekvenační platformu, délku readů atd. - řádek RG
- alignmentu, jaký algoritmus byl použit a rovněž obsahuje všechny parametry příkazu, který byl vyvolán pro vytvoření souboru - řádek PG

Záznamy jsou pak blokově uspořádány a obsahují minimálně 11 částí, z nich nejdůležitější je blok přečtené sekvence, její umístění na referenční genom, kvalita přečteného readu a řetězec zvaný CIGAR (Compact idiosyncratic gapped alignment report). Pomocí tohoto identifikátoru dokážeme určit, jakou operaci algoritmus provedl a v jaké délce. Lze tak zjistit, kolik bází bylo shodných přesně s referencí (M), kolik bází v porovnání se sekvencí chybělo (D), nebo přebývalo (I) nebo pokud se celý read shoduje s referencí (=). Příklad readu ze souboru BAM je uvedený na obrázku Obr. 1.8.

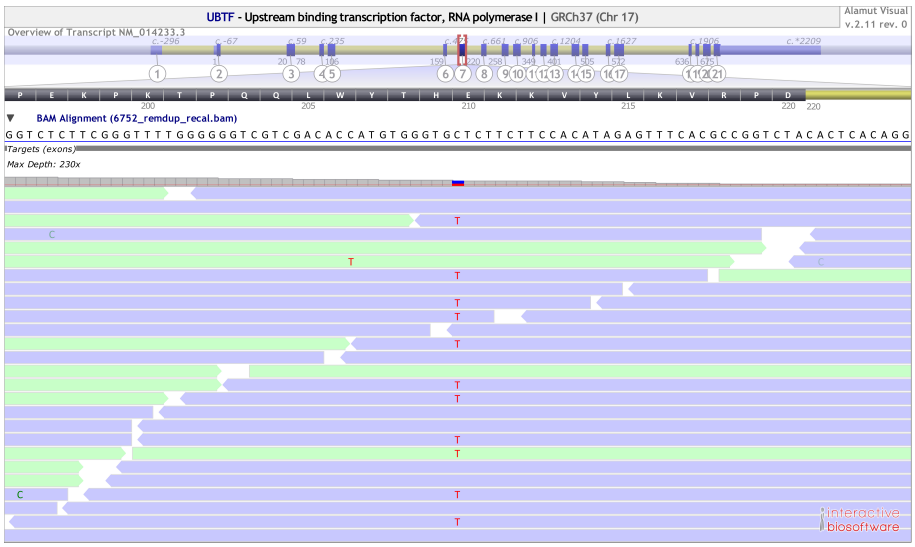
V praxi tento soubor využíváme při finálním vyhodnocení varianty, kdy prohlížíme variantu ve vizualizačním prohlížeči (IGV, Alamut Obr. 1.9) včetně jejího okolí, a prohlížíme a vyhodnocujeme informace o variantě [Li et al. 2009]. Podrobné specifikace formátu jsou dostupné online na URL[3].

[3]https://samtools.github.io/hts-specs/SAMv1.pdf [online: 16.10.2019]

```
@HD     VN:1.5     GO:none  SO:coordinate
@SQ     SN:chrM  LN:16571
@SQ     SN:chr1  LN:249250621
@SQ     SN:chr2  LN:243199373
@SQ     SN:chr3  LN:198022430
@RG     ID:7988    SM:GP_SS_EPI_042018_7988    PL:Illumina       PI:150
@PG     ID:bwa     PN:bwa    VN:0.7.16a-r1187-dirty  CL:/home/dnalab/
bioinformatics/bwa/bwa mem -M -t 14 -R @RG\tID:
7988\tSM:GP_SS_EPI_042018_7988\tPL:Illumina\tPI:150 /home/dnalab/
bioinformatics/hg19/ucsc.hg19.fasta /media/dnalab/6CA873B4A8737B80/
GAUK_data/GP_SS_EPI_042018/7988/fastq/7988_1.fastq.gz /media/dnalab/
6CA873B4A8737B80/GAUK_data/GP_SS_EPI_042018/7988/fastq/7988_2.fastq.gz
...
K00171:669:HT3FTBBXX:5:1107:11901:3829 99 chrM  1   60   18S83M =     196
296
AAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGA
GCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTATGCACGCGAT   @????
AC@CCACCB?AADBAACCCCDBACCAACCCBBCAA@AACBCCCCCC?
BBADDCCCCBCADDCA@@ECABBA@AD?BDDECCCCCBBBEEDC>B=@A
XA:Z:chr17,+22020709,101M,7;       MC:Z:101M
BD:Z:MMENQPPOMLQQPLOONOONNPOKJPONNMNMPOKNMONMLMMNNPK
OMOKNNLNNPPMMNOQQPQQQMEOPOONMEEOOOONRQNNNNPQPRSSORP
NPO       MD:Z:72G10    PG:Z:MarkDuplicates  RG:Z:7988
BI:Z:PPIQRRSQRORQONPPQQPPQPONNQQQQPQPPONQMPQPONPQQPNRM
OOQNMQPSSMPMOQQRNQRPJPQRRQPJJPQRSQSQOOOOSSSTQSQSRQQR
          NM:i:1     AS:i:78    XS:i:66
K00171:669:HT3FTBBXX:5:1116:3985:40649 353       chrM   1          50
65H36M   =     170  270  GATCACAGGTCTATCACCCTATTAACCACTCACGGG
A??@?CBCBACCA@CCCB@CAA@AAC@CCCCCC?BB  SA:Z:chrM,16507,+,
65M36S,50,0;  MC:Z:101M
BD:Z:MMOQSNMRQPONONQOKNMONMLMMNNPKOMOLOOM       MD:Z:36
PG:Z:MarkDuplicates       RG:Z:7988
BI:Z:PPRQQPPRRRRQQPPONQMPQPONPQQPNRMOOQON          NM:i:0
AS:i:36    XS:i:0
```

Hlavička označena červeně, nejdříve popis chromozomů z reference potom informace o vzorku a detaily analýzy. Modře pak jeden read vzorku

Obrázek 1.8: Hlavička souboru BAM s jedním readem



Obrázek 1.9: Náhled vizualizačního prohlížeče Alamut Visual

1.6.1.3 VCF - Variant calling file

Formální definice VCF a GVCF formátů byla čerpána dle: [Li et al. 2009].

Výsledkem bioinformatické analýzy nejčastěji bývá textový soubor VCF. Jde o tabulkové uspořádání, kde po hlavičce, ve které je definován každý sloupec, následuje tabulka s jednotlivými variantami – každý řádek je jedna varianta. Díky tomuto uspořádání je práce s těmito soubory uživatelsky přístupnější než jiná uspořádání – varianty lze snadno filtrovat dle parametrů v tabulce. Jde o výčet všech zjištěných variant ve vzorku, tedy všech odchylek od referenční sekvence a to i s dalšími charakteristikami každé varianty.

Tabulka je vždy doplněna o „preambuli“ s meta-informacemi, kde je popsaný každý parametr, který VCF obsahuje. Tyto řádky jsou označeny znakem „##“, měly by být přítomny pro každý parametr v polích INFO, FILTER a FORMAT a měly by obsahovat název parametru, jeho datový typ (jestli se jedná o číslo, textový řetězec) a popis. V této „preambuli“ rovněž nalezneme informace o použité referenci, o předchozím filtrování a o použitých nástrojích.

Následuje řádek s hlavičkou – názvy sloupců tabulky, z nichž prvních 8 musí být přítomno:

1. CHROM – označení chromozomu
2. POS – pozice varianty na chomozomu
3. ID – identifikátor dbSNP varianty, pokud je taková varianta v této databázi přítomná
4. REF – referenční báze dle zvolené reference
5. ALT – alternativní báze, jak byla přečtená ze sekvence probanda
6. QUAL – Phred skóre, vypočítané v předchozích krocích
7. FILTER – informace o tom, jestli varianta prošla všemi definovanými filtry, pokud ano, tak je uvedená hodnota PASS, pokud ne, tak je hodnota nastavená na název filtru, kterým neprošla (filtry jsou uvedeny v preambuli)
8. INFO – další informace o variantách, počet parametrů zde je libovolný, ale v INFO poli by se neměly nacházet parametry, které nebyly definovány v preambuli

V dalším sloupci (sloupcích) jsou uvedené genotypy pro každý vzorek i s informací o pokrytí.

Jak z definice vyplývá, tak do VCF souboru lze uložit mnoho informací, tento formát je vhodný zejména proto, že záznamy dokáže vhodně uspořádat i v případě, že je soubor anotován z mnoha různých zdrojů a sloupec INFO obsahuje i desítky různých parametrů. Pro ještě snadnější analýzu se pak často VCF soubor převádí na tabulku, kdy všechny sloupce oddělujeme tabulátorem pro ještě větší přehlednost. Příklad VCF souboru uvádíme na obrázku Obr. 1.10. Podrobné specifikace formátu jsou dostupné online na URL[4].

[4]https://samtools.github.io/hts-specs/VCFv4.2.pdf [online: 16.10.2019]

```
##fileformat=VCFv4.1
##GeneratedBy=SureCall2.0 Snppet
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward
bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Float,Description="Root-mean-square mapping quality of
covering reads">
##INFO=<ID=AF1,Number=1,Type=Float,Description="estimate of the first ALT allele
frequency">
##INFO=<ID=AN,Number=1,Type=Float,Description="Total number of alleles in called
genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=1,Type=Integer,Description="List of Phred-scaled genotype
likelihoods">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT
7089_1.fastq
chr1    762273  .       G       A       123.0   .
DP=47.0;DP4=0,0,4,43;STDP4=1,0,4,43;AF1=0.9791667;AN=3;MQ=41;RegionAverageDepth=0.47690016;I
(LOW|SILENT|tcC/tcT|S210|LINC00115|Non-coding_transcript|NON_CODING|NR_024321|
NR_024321.ex.1),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_015368|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047519|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047520|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047521|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047522|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047523|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047524|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047525|),UPSTREAM(MODIFIER||||LOC643837|Non-coding_transcript|NON_CODING|
NR_047526|
);SEL_PRIMARY_EFF=0;Gene_Description=762902;pValue=1.0E-123;Zygosity=HOM;Category=2
GT:PL:GQ        1/1:60,1,0:10
```

Povinné řádky jsou označeny červeně, první definuje verzi souboru, hlavička zeleně, popisuje jednotlivé sloupce a jejich přípustné hodnoty, druhý červený řádek označuje sloupce tabulky, černě první varianta

Obrázek 1.10: Příklad VCF hlavičky a první varianty

GVCF formát Tento formát je rozšířením VCF (Genomic VCF), základní definice je totožná s VCF, ale obsahuje další doplňující informace. Hlavním rozdílem je ale přítomnost informace o všech sekvenovaných oblastech, tedy i těch, kde se nenacházela varianta (odchylka od reference). Tyto oblasti jsou označeny řetězcem <NON_REF> ve sloupci pro alternativní bázi (bloky, které se shodují s referencí). Soubory tohoto typu jsou větší než VCF, protože obsahují informaci navíc. Využívají se pro vyvolávání VCF souboru u více probandů, kdy dojde ke spojení GVCF a poté vyvolání variant.

1.6.1.4 Další formáty používané v bioinformatické analýze

- BED dle definice z URL:[5] – sloupcově uspořádaný textový formát, slouží pro stanovení hranic, u kterých probíhá analýza. Tento soubor se využívá, pokud chceme provádět analýzu pouze v určitých intervalech daného chromosomu – např. u MPS panelu genů můžeme ohraničit každý exon genu. Výhodou tohoto postupu je daleko rychlejší analýza, nevýhodou je, že pokud by se hledaná varianta nacházela mimo interval, tak nedojde k jejímu vyvolání. Každý záznam vždy musí obsahovat tři sloupce – s označením chromosomu, začátkem a koncem intervalu.

- PED dle definice z URL:[6] – sloupcově uspořádaný textový formát, který stanovuje vztah mezi jednotlivými vzorky/probandy. V každém řádku identifi-

[5] https://software.broadinstitute.org/software/igv/BED [online: 16.10.2019]

[6] https://gatkforums.broadinstitute.org/gatk/discussion/7696/pedigree-ped-files [online: 16.10.2019]

kátorem označíme nejprve identifikátor rodiny, který je společný pro všechny členy, poté identifikátor probanda, identifikátor otce, matky, pohlaví probanda a jeho fenotyp (postižen / nepostižen / neuveden). Díky tomuto jednoduchému vyjádření lze popsat textově i komplexní rodokmeny. Pokud některý z parametrů není přítomen (např. nemáme vzorek otce), zapisuje se do souboru znak nuly.

1.6.2 Bioinformatická pipeline

Pojem bioinformatická pipeline označuje sled úkonů a postupů, které nám umožní získat informaci o genetických variantách probanda ze sekvence DNA přečtené sekvenátorem. Základním konceptem je přečtenou sekvenci DNA probanda nejprve namapovat na referenční genom – to znamená ke každé části sekvence – readu – získat informaci o její lokaci na chromozomu (alignment). S takto předzpracovaným souborem pak můžeme pracovat dále a hledat právě ta místa, kde se sekvence vzorku liší od reference (variant calling).

Pro zpracování dat v DNA laboratoři jsme zvolili přístup dle doporučení GATK (verze 3.8) [DePristo et al. 2011; Auwera et al. 2013], jedná se o sadu instrukcí / postupů, které byly aplikovány i na data ve významných velkých studiích (gnomAD či ExAC).

1.6.2.1 GATK pipeline pro detekci zárodečných (germinálních) variant

Tato pipeline [Auwera et al. 2013] popisuje postup pro identifikaci zárodečných variant. Vstupem je dvojice FASTQ souborů, nebo soubor s příponou uBAM (unmapped BAM), který má strukturu stejnou jako BAM, ale bez mapování na referenci. Celý proces je shrnut na schématu Obr. 1.11.

Jednotlivé kroky pipeline pak jsou:

Alignment

1. Mapování readů na referenční genom – dle volby uživatele – nejčastěji hg19 či hg38, čímž vzniká validní BAM soubor a jeho index (.bai)

2. V dalším kroku dochází k označení a odstranění PCR duplikátů, pokud jsou přítomny

3. Algoritmus BQSR (Base quality scores recalibration) přepočítává Phred skóre u každé báze na základě matice kovariancí. Tato matice obsahuje parametry specifické pro každou bázi (např. jestli byla báze přečtená správně, její Phred skóre), tak i parametry pro celý read – průměr Phred skóre celého readu, či větší okolní sekvence. Díky tomuto komplexnímu algoritmu lze přepočítat Phred skóre a eliminovat tak systematické chyby vzniklé sekvenátorem. Algoritmus tak může po přepočítání pohnout hodnotou Phred jak nahoru tak dolů.

Po tomto kroku je alignment vzorku hotový, jedná se o proces, který pro každý vzorek běží zvlášť, tzn. jeho výstupem pro n počet vzorků ($2n$ FASTQ souborů) je n validních BAM souborů.

Variant Calling

1. Vyvolání variant je provedeno algoritmem HaplotypeCaller, výstupem je GVCF soubor pro každý BAM, obsahující jak SNP, tak Indely.

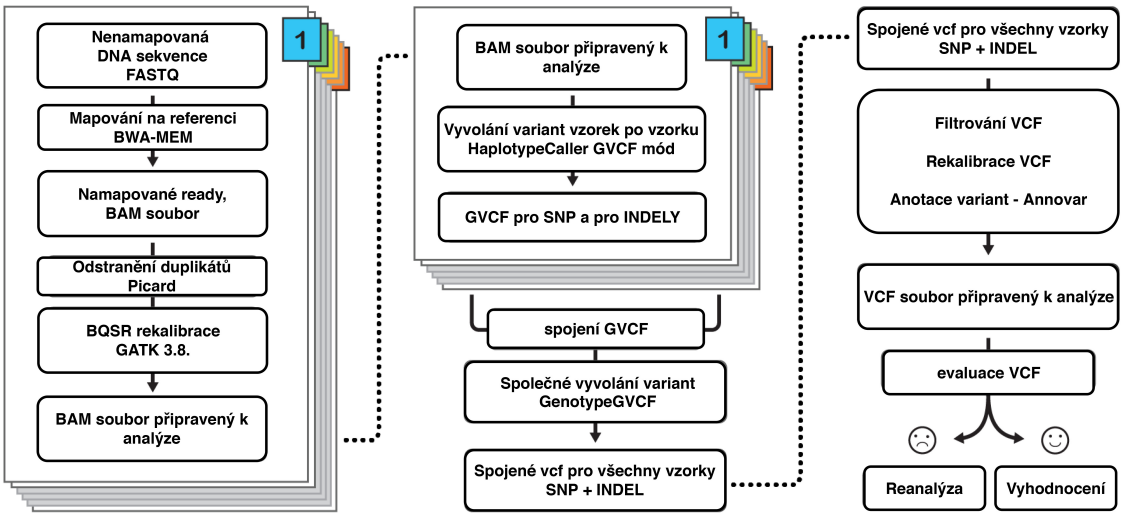
2. Spojení všech GVCF v knihovně a vyvolání variant pro společný soubor – společné vyvolání variant je výhodné v tom, že nepřijdeme o variantu, která by se necházela ve více vzorcích, ale byla by velmi málo pokrytá. Například pokud bude varianta pokrytá ve dvou readech u jednoho pacienta a budeme provádět variant calling jednotlivě, tak variantu nevyvoláme díky nízkému pokrytí. Pokud ale budeme mít knihovnu např. 20 vzorků, u nichž 5 z nich bude tuto variantu mít (s pokrytím 1 nebo 2), tak jí algoritmus vyvolá a lze s ní dále pracovat.

3. Výsledné VCF je pak rozdělené na SNP a Indel VCF.

4. Dochází k dalšímu filtrování variant, zde jsou dvě možnosti postupu u knihovny menší než je 30 vzorků WES se používá tvrdé filtrování, u knihovny větší se používá algoritmus VQSR (Variant quality scores recalibration), který vytváří pro každou knihovnu vlastní model, na základě kterého jsou varianty filtrovány.

5. Výsledkem jsou dvě vyfiltrovaná VCF, připravená k anotaci dalšími nástroji

6. Variant calling knihovny je hotový, pro knihovnu k s n vzorků získáváme $2k$ VCF (vždy jedno pro SNP a jedno pro Indely), v každém souboru jsou varianty pro n vzorků.



Převzato z URL: https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145, přeloženo

Obrázek 1.11: Schéma zpracování dle GATK best practices

1.6.2.2 Software pro zpracování dat

V současnosti je silný trend využívání vyhodnocovacího software vytvořeného jako tzv. all-in-one nástroj, který umožňuje uživatelům neznalým bioinformatiky všechny

parametry bioinformatické analýzy nastavit v uživatelsky přívětivém prostředí. Tato možnost přináší mnoho výhod, ale i nevýhody.

Hlavní výhodou je, že v tomto případě může analýzu provést i běžný uživatel počítače, kdy pomocí vhodně připraveného průvodce, lze získat kvalitní výsledek velmi snadno. Nevýhodou pak je, že v případě problémů je nesnadné nalézt příčinu nezdaru – program často vypíše pouze chybové hlášení typu „Analýza neproběhla“ a uživatel neví co dál, zároveň pak nemá plnou kontrolu nad všemi kroky, nemůže nastavit všechny parametry a jde tedy spíše o restriktivní přístup, kdy uživatel může udělat pouze to, co mu program dovolí.

Pro výpočetně náročné operace, jako je např. zpracování dat z WGS pak tyto programy často selhávají, neboť dojde k zahlcení paměti počítače a program si s tímto problémem často neporadí, navíc sám o sobě čerpá další cenné výpočetní zdroje. Výsledkem je pak pád aplikace a nutná reanalýza – což zejména u WGS, kdy zpracování jednoho vzorku trvá řádově desítky hodin, bývá nepříjemnou komplikací.

V DNA laboratoři jsou v současnosti využívány tři nástroje pro bioinformatickou analýzu. Dva ze tří nástrojů běží v klasickém prostředí OS Windows a slouží pro zpracování FASTQ souborů na VCF a BAM soubory. Jedná se o nástroje SureCall URL[7] (Agilent, USA) a NextGENe URL[8] (Softgenetics, USA).

Posledním nástrojem je Framework Galaxy [Afgan et al. 2018]. Ten vytváří virtuální obal kolem nástrojů ovládaných z příkazové řádky a díky tomu je možné všechny parametry nastavovat pomocí webového prostředí. Tento nástroj kombinuje výhody obou přístupů – kdy získáváme uživatelsky přívětivou aplikaci a zároveň máme možnost nastavovat všechny parametry a ovládat každý krok analýzy. Pro správné fungování je ale nutné (pokud není předpřipravené od vývojářů) si vytvořit vlastní workflow, s přesně definovanými vstupy a výstupy procesů, a rovněž definovat, jaké parametry, kterému nástroji můžeme předat. Tento krok většinou náleží bioinformatikovi, spouštění workflow pak už provádí běžný uživatel.

1.6.2.3 Analýza dat v cloudu, FireCloud

Díky zvyšujícím se požadavkům na vypočetní kapacitu se začínají prosazovat služby, které umožňují bioinformatickou analýzu provádět za nás. Takové služby nejčastěji nabízí komerční společnosti jako balíček k samotné sekvenaci. V praxi to pak znamená, že společnost neodesílá zákazníkovi pouze zpracované FASTQ soubory, ale již hotové VCF soubory.

Další možností je analýza dat přes cloudové služby, kdy zájemce nahraje data do cloudu a analýza probíhá za poplatek na serverech velkých korporací jako je Google nebo Amazon. Výhodou je pak rychlost analýzy, nevýhodou je nutnost data přenášet „neznámo kam“, protože, ve chvíli, kdy jsou data nahraná v cloudu, tak nelze plně kontrolovat, kdo k nim má přístup.

V druhé polovině roku 2017 byla uvolněna aplikace FireCloud [Birger et al. 2017], která pak umožňuje snadnou analýzu MPS dat, dle GATK doporučení. Tato aplikace nabízí předpřipravená workflow s aktuálními verzemi GATK pipeline pro všechny

[7] https://www.agilent.com/en/product/next-generation-sequencing/hybridization-based-next-generation-sequencing-ngs/ngs-software/surecall-232880 [online: 16.10.2019]

[8] https://softgenetics.com/NextGENe.php [online: 16.10.2019]

uživatele. Uživatel po registraci nahraje svá data do Google úložiště, které je privátní (s možností ho sdílet). Tato data jsou pak zpracovávána dle předpřipravených workflow.

Prozatímní nevýhodou je nutnost předzpracovat data do formátu uBAM, který spojí FASTQ soubory, ale nemapuje je na referenci. Soubor je pak nahrán do úložiště a dochází k jeho zpracování. Velkou výhodou je, že odpadává nutnost správného nastavení parametrů pro workflow a také, že čerpáme nesrovnatelně vyšší výkon ze serverů Google.

Služba funguje na principu plateb za využité úložiště a výkon, kdy uvedené workflow nabízejí zpracování WGS dat s cenou pod 5 dolarů + výdaj za uložení dat.

1.6.3 Prioritizace variant, populační databáze

Proces prioritizace variant nejčastěji začíná otevřením VCF obsahujícím všechny varianty nalezené ve vzorku DNA. Úkolem je pak shromáždit co nejvíce informací o každé variantě a pomocí těchto informací rozhodnout, která varianta může být pro nás suspektní (vysvětlující příčinu nemoci). Množství variant je variabilní dle typu sekvenační knihovny – pokud se jedná o sekvenování panelu genů získáváme maximálně stovky variant, v případě celoexomového sekvenování jde už o desítky tisíc variant a v případě genomu dokonce o více než milion variant na vzorek.

Pro uspořádání variant lze volit různé postupy, nejčastěji využíváme anotačních nástrojů, které prohledávají populační databáze, získávají informace od predikčních nástrojů a vše pak přiřazují k variantám. Výsledkem anotačního procesu je přidání několika desítek parametrů, které nám pomáhají rozhodovat o klasifikaci varianty. Pro prioritizaci variant jsou nejčastěji používány tyto zdroje:

1.6.3.1 Populační databáze pro filtrování variant

gnomAD Databáze gnomAD (The Genome Aggregation Database) [Karczewski et al. 2019] je v současnosti jeden z největších projektů sdružujících záznamy WES a WGS z mnoha studií. Data jsou veřejně a bezplatně přístupná celé komunitě a projekt navazuje na předchozí projekt ExAC [Karczewski et al. 2016].

V současnosti je k dispozici databáze obsahující 125 478 exomů a 15 708 genomů nepříbuzných jedinců z různých sekvenačních projektů zaměřených na různá onemocnění nebo na celé populace. Databáze obsahuje data celkem 141 456 jedinců. Ze skupiny byli vyloučeni probandi, kteří byli postiženi závažnými dětskými onemocněními včetně jejich nejbližších příbuzných.

Metodikou pro zpracování dat byla zvolená standardizovaná pipeline „BWA-Picard-GATK“, která je pokládána za standard ve zpracování MPS dat, a kterou rovněž používáme v DNA laboratoři. Zpracovaná data byla rozdělena do dalších 8 skupin dle populací (European – non-Finnish, African, Latino, Ashkenazi Jewish, East Asian, European (Finnish), South Asian, Other), umožňující získat frekvenci alely pro celou populaci i pro každou z podpopulací .

Frekvence uvedená v této databázi je nejčastěji používaným kritériem při filtrování variant. Základním předpokladem pro vyloučení varianty při filtrování je její frekvence větší než jedno procento v dané populaci. Vhodné je pak kombinovat frekvenci z obou databází (WES i WGS) a celé populace i subpopulace s větším důrazem kladeným na subpopulaci do které proband spadá.

1.6.3.2 Predikční programy a nástroje

PolyPhen-2 Predikční nástroj PolyPhen-2 [Adzhubei et al. 2010] slouží k určování efektu missense variant na proteinovou strukturu. Algoritmus srovnává sekvenci s wild-type alelou a s mutovanou alelou a detekuje strukturní a funkční změny v produktech sekvencí. Díky tomu dokážeme odhadnout, jakou bude mít změna báze (popř. aminokyseliny) efekt na výsledný produkt.

SIFT Sorting Intolerant from Tolerant (SIFT) [Sim et al. 2012] je dalším predikčním nástrojem sloužícím pro prioritizaci variant. Funguje na podobném principu jako Polyphen-2, ale zaměřuje se i na krátké indely. Sekvenci produktu porovnává se sekvencí uvedenou v databází UniProtkB. Výsledné skóre se pohybuje od 0 do 1, kdy hodnoty mezi 0 a 0.05 jsou pokládány za závažné a měnící funkci proteinu.

M-Cap Pro prioritizaci variant se využívá mnoho predikčních nástrojů a je těžké rozhodnout, který z nich poskytuje nejvalidnější výsledky. Nástroj M-Cap [Jagadeesh et al. 2016] slouží jako agregátor predikčních nástrojů (PolyPhen-2, CADD, MutationTaster, MutationAssesor, FATHMM, LRT, MetaLR, MetaSVM) a nástrojů k určení, jak vysoce konzervovaná je daná oblast(RVIS, PhyloP, PhastCons, PAM250, BLOSUM62, SIPHY, GERP). Pomocí metod strojového učení kombinuje jednotlivé parametry ve výsledné skóre. Díky výslednému parametru pak lze snadno filtrovat varianty ve WES datech s výslednou senzitivitou vyšší, než kdybychom používali jednotlivé nástroje samostatně.

OMIM Databáze OMIM[9] (Online Mendelian Inheritance in Man) [Hamosh et al. 2005] je databáze propojující lidské geny s klinickými fenotypy spravovaná univerzitou Johna Hopkinse. Záznamy jsou rozděleny do několika skupin (každá skupina má vlastní prefix kódu). Jednotlivé skupiny jsou pak dále děleny dle dědičnosti. V praxi využíváme databázi pro spárování genu s hledaným fenotypem u WES dat. Díky tomu si můžeme vyfiltrovat varianty pouze v genech, které mají spojitost s námi vyšetřovaným onemocněním.

V současné době databáze OMIM obsahuje násedující data (Obr. 1.12).

Popis	Prefix	Autozomální	X vázané	Y vázané	Mitochondriální	Celkem
Popis genu	*	15 239	732	49	37	**16 057**
Kombinace gen a fenotyp	+	46	0	0	0	**46**
Popis fenotypu se známou příčinou na molekulární úrovni	#	5 132	332	4	33	**5 501**
Popis fenotypu se neznámou příčinou na molekulární úrovni	%	1 442	122	4	0	**1 568**
Ostatní fenotypy s mendelovským typem dědičnosti		1 649	105	3	0	**1 757**
Celkem		**23 508**	**1 291**	**60**	**70**	**24 929**

Informace převzata z https://www.omim.org/statistics/entry

Obrázek 1.12: Přehled počtu záznamů v databázi OMIM (k červenci 2019)

ClinVar Veřejná databáze variant ClinVar [Landrum et al. 2016] je projektem NCBI (National Center for Biotechnology Information)[10] a poskytuje informace

[9] https://www.omim.org/ [online 16.10.2019]
[10] https://www.ncbi.nlm.nih.gov/clinvar/ [online 16.10.2019]

o zárodečných i somatických variantách a jejich příčinné souvislosti s onemocněním. Databáze je otevřená a umožňuje přidávání vlastních variant, přesto jednotlivé záznamy podléhají validaci a jejich klinická signifikance je pravidelně kriticky zhodnocována. V současnosti databáze obsahuje 782 636 klinicky významných variant v celkem 30 270 genech od 1 192 přispěvatelů (k červenci 2019). Každý záznam obsahuje lokaci varianty, odkud záznam pochází, případně se kterým onemocněním je varianta spojována a její klasifikaci.

UniProt UniProt [Consortium 2019] je v současnosti největší databáze anotovaných proteinů doplněných o jejich sekvenci. Obsahuje více než 120 miliónů proteinů mnoha organizmů (u člověka to je cca 80 000 proteinů), každý záznam obsahuje sekvenci proteinu, jeho domény a rovněž informaci o expresi v organizmu. Této informace nejčastěji využíváme při prioritizaci variant.

1.6.4 Klasifikace variant

Výsledkem procesu prioritizace variant by neměla být pouze jedna patogenní varianta ale soubor variant klasifikovaných do různých tříd. Pro třídění variant byl zaveden standard dle organizace ACMG [Li et al. 2017], který zavádí komplexní pravidla pro klasifikaci variant. Každá varianta je hodnocena s ohledem na níže uvedené podmínky a stanovuje se, zda splňuje dané kritérium.

1. Kritéria podporující patogenní charakter dané varianty:
   - Velmi silné
     - PVS1 Varianta typu nonsense, frameshift, splice site, v iniciačním kodonu, exonová delece v genu, kde je LOF známá jako přičina nemoci
     - Silné
       * PS1 Záměna AMK, která byla již dříve klasifikována jako patogenní
       * PS2 *De novo* varianta pokud byli testováni oba rodiče na přitomnost varianty
       * PS3 Varianta byla zkoumána funkční in vitro či in vivo studií s prokázaným dopadem na gen nebo genový produkt
       * PS4 Prevalence varianty je u postižených výrazně vyšší než u zdravých jedinců
     - Středně silné
       * PM1 Lokalizace v mutačním hotspotu nebo proteinové doméně bez benigních variant
       * PM2 Absence frekvence v populační databázi či velmi nízká frekvence v populační databázích u AR
       * PM3 V případě onemocnění s AR dědičnosti je v pozici trans s patogenní variantou

        * PM4 Dochází ke změně délky proteinu na základě efektu varianty
        * PM5 Missense varianta se záměnou AMK, v místě, kde již byla pozorována záměna za jinou AMK (př. Arg156His je patogenní, pozorujeme Arg156Cys)
        * PM6 *De novo* předpoklad, bez potvrzení varianty u rodičů
    - Podporující
        * PP1 Segregace varianty v rodině s onemocněním
        * PP2 Missense varianta v genu, který má velmi malé množství známých benigních variant, nebo vysoké množství známých patogenních variant
        * PP3 Označení varianty za patogenní dle in silico nástrojů
        * PP4 Velice specifický fenotyp pacienta
        * PP5 Potvrzení varianty ze spolehlivého zdroje

2. Kritéria podporující benigní charakter dané varianty
    - Postačující
        - BA1 Frekvence v populační databázi >5%
    - Silné
        - BS1 Frekvence v populačních databázích je vyšší než předpokládaná u onemocnění
        - BS2 Varianta byla pozorovaná u zdravého jedince
        - BS3 Varianta nemá poškozující efekt dle funkční studie
        - BS4 Variant nesegreguje v rodině
    - Podporující
        - BP1 Missense varianta v genu, kde se často vyskytují varianty zkracující protein (protein-truncating)
        - BP2 Varianta je v trans pozici s patogenní variantou při AD dědičnosti nebo v pozici cis u všech typů dědičností
        - BP3 In-frame indel v repetitivní oblasti
        - BP4 Označení varianty za benigní dle in silico nástrojů
        - BP5 Varianta byla nalezena u jiného probanda s jinou záměnou báze
        - BP6 Označení varianty za benigní spolehlivým zdrojem
        - BP7 Synonymní varianta bez vlivu na splicing

3. Po zhodnocení variant dle těchto pravidel dojde ke klasifikaci variant dle algoritmu:
    - Patogenní varianta je když splňuje tato kritéria:
        - PVS1 **A** min. 1× PS **NEBO**

        * min. 2× PM **NEBO**
        * 1× PM **A** 1× PS **NEBO**
      - min. 2× PS
      - 1 PS **A** min. 3× PM **NEBO**
        * 2× PM **A** 2× PP **NEBO**
        * 1× PM **A** min. 4× PP
  - Pravděpodobně patogenní musí splňovat tyto podmínky:
    - PVS1 **A** 1× PM
    - 1× PS **A** 1 až 2× PM
    - 1× PS **A** min. 2× PP
    - min. 3× PM
    - 2×PM **A** min. 2× PP
    - 1× PM **A** min. 4× PP
  - Benigní varianta je když splňuje podmínky:
    - 1 BA
    - min. 2× BS
  - Pravděpodobně benigní variant splňuje podmínky:
    - 1 BS **A** 1 BP
    - min. 2× BP
  - VUS (Bez jasného významu)
    - žádná z kritérií nejsou splněna
    - protichůdná kritéria pro patogenní variantu a benigní variantu jsou splněna

Na základě vyhodnocení těchto tabulek pak získáváme klasifikaci významu varianty. Jak z definice vyplývá, tak klasifikace je souborem několika kroků a fúze informací z více zdrojů, nelze tak spoléhat na klasifikaci, kterou nám poskytují nástroje pouze na základě VCF, neboť zde nezjistíme informace o segregaci v rodině, popř. o nálezu *de novo*. Takové klasifikace je tedy nutné brát s rezervou anebo nástrojům potřebné informace doplnit.

1.6.5 Sangerovo sekvenování

Sangerova metoda sekvenování [Sanger, Nicklen a Coulson 1977] využívá proces replikace DNA a funguje na podobném principu jako PCR.

V prvním kroku generujeme fragmenty DNA o různých délkách, které ale mají stejný počátek. Společným začátkem je 5' konec sekvenačního primeru. V reakční směsi jsou přítomny jak nukleotidy (dXTP), tak fluorescenčně značené dideoxynukleotidy (ddXTP). Jakmile se do syntetizovaného řetězce přidá fluorescenčně značený dideoxynukleotid, dojde k zastavení syntézy řetězce. Fragmenty jsou poté separovány

na základě délky elektroforézou, a tím seřazeny dle délky. Excitací fluorescenčních značek laserem dojde k přečtení báze (dle vlnové délky značeného ddXTP). Sekvence DNA se poté stanoví průběžným čtením sekvence těchto značek.[Snustad, Simmons a Relichová 2017]

Sangerovo sekvenování v DNA laboratoři využíváme v několika aplikacích. Metodou první volby je u vybraných pacientů, u kterých ještě před sekvenováním panelem genů předpokládáme diagnózu a chceme ji potvrdit. To je situace např. u pacientů se syndromem Dravetové, kdy cíleně sekvenujeme gen *SCN1A*. Další aplikace je při potvrzení přítomnosti varianty nalezené u probanda metodami MPS. Pomocí Sangerova sekvenování rovněž sekvenujeme vzorky rodičů při segregační analýze.

2 Cíle dizertační práce

1. Analýza MPS dat sekvenovaných panelem genů – cílem je optimální vyhodnocení minimálně 250 pacientů pomocí panelu genů
   a) Navržení vlastního workflow pro analýzu dat z MPS dat panelu genů
   b) Provést analýzu nalezených kauzálních variant podle postižených genů, jejich chromozomální lokalizace, původu varianty a typu dědičnosti genu
      i. Analýza dat více než 250 pacientů, kteří podstoupili sekvenování panelem genů asociovaných s epilepsií
         – Určíme patogenní a pravděpodobně patogenní varianty (P/PL) v souboru
         – Provedeme genovou analýzu P/PL variant, určíme dědičnost a původ vzniku variant
         – Provedeme analýzu vlivu věku pacientů při nástupu onemocnění na objasnitelnost případu
2. Analýza dat z celoexomového sekvenování (WES) – cílem je optimální vyhodnocení minimálně 150 WES
   a) Navržení vlastního workflow pro analýzu WES dat zejména u probandů, u kterých nebyla nalezena kauzální varianta sekvenováním pomocí panelu genů
   b) Zavedení nových metod anotace dat do workflow pro hledání kauzality variant pro vyšetřovaná onemocnění
   c) Zavést pokročilou metodiku pro detekci *de novo* variant pomocí Trio analýzy (pacient + oba rodiče) u pacientů s neurogenetickým onemocněním
   d) U pacientů, u kterých nebyla nalezena patogenní *de novo* varianta bude dále vytvořena metodika pro hledání patogenních variant (Singleton model) pro AR AD i X-vázanou dědičnost.
   e) Zavedení algoritmu pro hledání variability v počtu kopií segmentů DNA (copy number variation, CNV)
3. Analýza dat z celogenomového sekvenování (WGS) – cílem je vyhodnocení minimálně 20 WGS
   a) Navržení vlastního workflow pro analýzu WGS dat zejména u probandů, u kterých nebyla nalezena kauzální varianta v předchozí analýze pomocí MPS
   b) Navržení postupu organizace efektivní, spolehlivé a bezpečné správy WGS dat v rámci DNA laboratoře z kapacitních důvodů

   c) Otestování postupů aplikovaných na WES skupinu pro hledání patogenních variant u WGS vzorků

4. Databáze variant z MPS - vytvořit databázi pro optimalizaci postupu hledání P/PL variant
   a) Navržení a implementace In-house databáze variant z celoexomového sekvenování
      – Zavedení metodiky pro flexibilní analýzy pomocí in-house databáze variant
   b) Vytvořit databázi proteinových domén mapovaných na lidský genom
   c) Vytvořit databázi CMT variant pro sdílení v rámci mezinárodních týmů
5. Navrhnout a zavést systém pro správu dat a udržování databázových systémů

3 Pacienti a metody

3.1 Pacienti

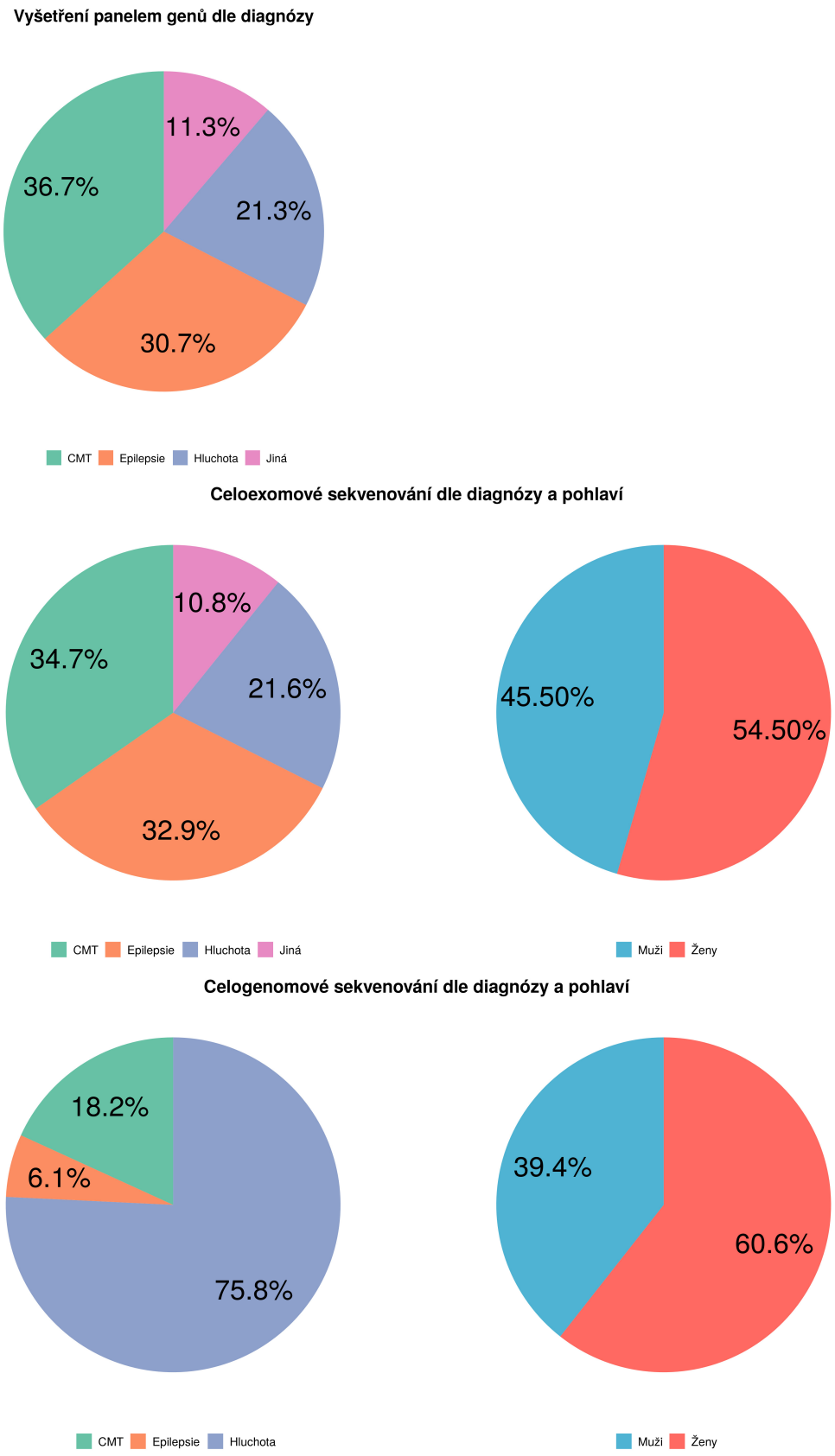
Výběr pacientů pro MPS probíhá v několika krocích. V DNA laboratoři zpracováváme vzorky pacientů z Kliniky dětské neurologie 2.LF UK a FN Motol. Prvním krokem při výběru pacienta je konzultace neurologa a genetika, kteří indikují vyšetření. Po indikaci přichází do DNA laboratoře vzorek DNA, se kterým dále pracujeme. Sekvenování probíhá v externí laboratoři. Při hledání kauzálních variant postupujeme v několika krocích, které jsou uvedené na obrázku Obr. 3.2. Níže uvádíme v tabulce Tab. 3.1 přehled všech pacientů, u kterých bylo provedeno MPS a rovněž v grafickém přehledu Obr. 3.1 poměry mezi jednotlivými diagnózami a u WES a WGS i poměr mezi počtem mužů a žen.

MPS panelem genů	
Diagnóza	**Celkem**
	905
CMT	332
Epilepsie	278
Hluchota	193
Jiná	102

WES			
Diagnóza	**Celkem**	**Muži**	**Ženy**
	222	121	101
CMT	77	49	28
Epilepsie	73	38	35
Hluchota	48	20	28
Jiná	24	14	10

WGS			
Diagnóza	**Celkem**	**Muži**	**Ženy**
	33	20	13
Epilepsie	6	4	2
CMT	2	0	2
Hluchota	25	16	9

Tabulka 3.1: Počty vyšetřených pacientů MPS panelu genů, WES a WGS dle diagnóz



Obrázek 3.1: Poměry vyšetřených pacientů pomocí MPS panelu genů, WES a WGS dle diagnóz

3.2 Metody

U vybraných pacientů je na Klinice dětské neurologie proveden odběr krve ke zpracování na našem pracovišti. Kromě samotného vzorku je nutné získat co nejkomplexnější informace o pacientovi, tedy jeho celkovou anamnézu. Pro segregační analýzu je pak odebrán vzorek krve pro izolaci DNA od obou rodičů, včetně informací o jejich komplexní anamnéze.

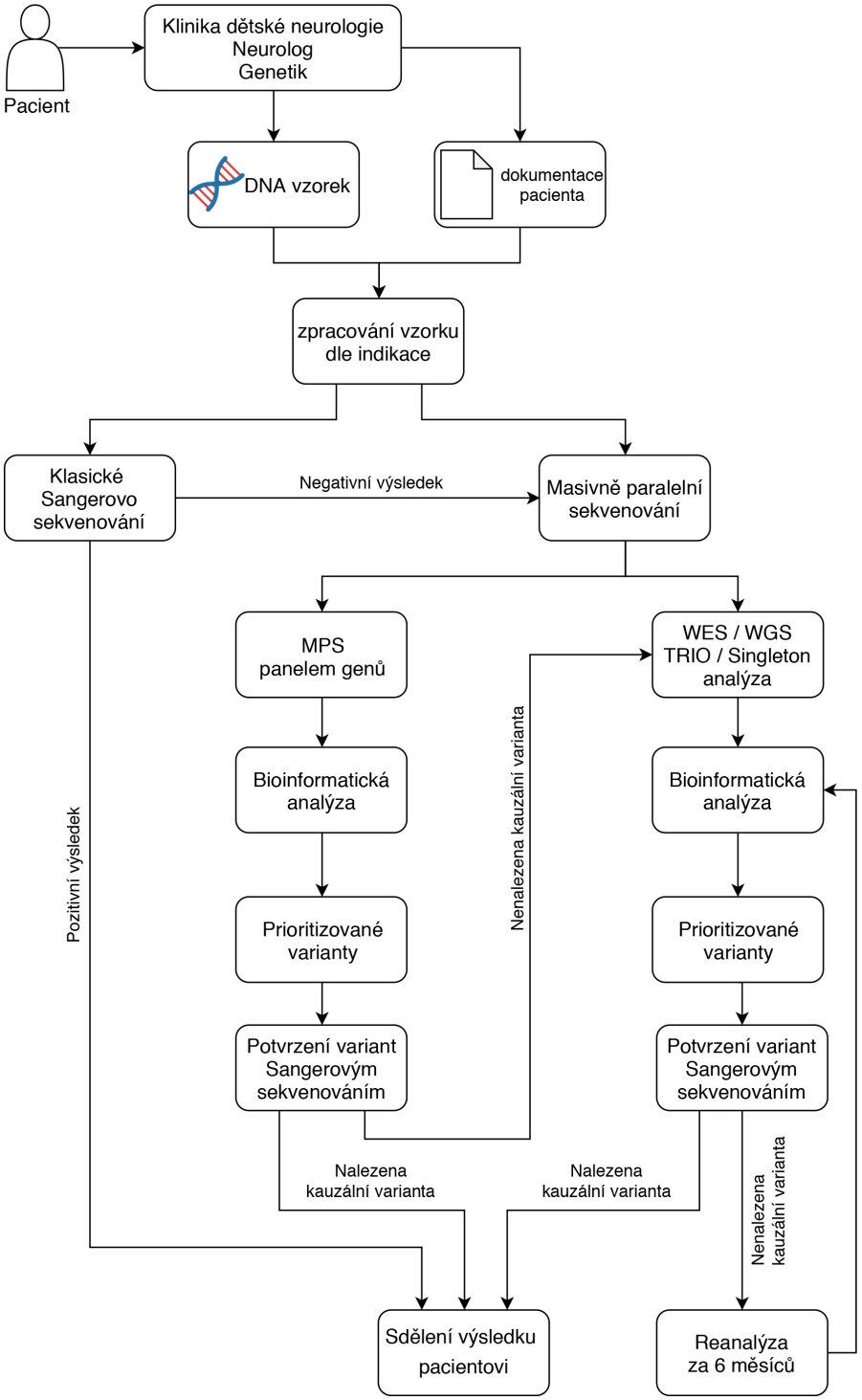
U některých pacientů se nepřistupuje k metodám MPS, ale využívá se pouze klasického sekvenování vybraného genu (př. *SCN1A* u syndromu Dravetové). Pokud ale nedojde k identifikaci patogenní varianty, přistupuje se k MPS a to v prvním kroku pomocí panelu genů.

Masivně paralelní sekvenování panelem genů hledá patogenní varianty v předem definované množině genů – panelu. V současné době na pracovišti využíváme několik knihoven pro jednotlivá onemocnění (např. HMSN/CMT – 103 genů, Epilepsie – 112 genů). Výhodou tohoto přístupu je vysoké pokrytí vyšetřovaných oblastí (pokrytí více než 500×). Limitací je nutná pravidelná aktualizace panelu – dle okolností přidávat nově popsané geny nebo v některých případech geny odebírat, při překročení kapacity panelu. Pokud nedojde k nalezení kauzální patogenní varianty v panelu genů, je po konzultaci provedeno WES nebo WGS. Pro tato vyšetření pomocí MPS je pro interpretaci variant vhodné mít k dispozici vzorky DNA od obou rodičů k dovyšetření segregace variant s onemocněním. Pomocí WES u pacienta a obou rodičů je možné hledat *de novo* varianty v genech dosud nespojovaných s lidskými onemocněními.

Celoexomové sekvenování se obvykle provádí u pacientů, u kterých nebyla nalezena kauzální varianta při MPS panelem genů. Výhodou WES je pokrytí téměř všech kódujících oblastí genů, nevýhodou je nižší pokrytí (jak z hlediska hloubky čtení, tak pokrytí oblastí) a vyšší cena.

Pokud analýza dat z WES či WGS nezachytí variantu, která s dostatečnou pravděpodobností vysvětluje příčinu onemocnění pacienta, dochází k opakování analýzy dat znovu nejpozději do 6 měsíců, pro případ, že by se objevily nové publikace, sdělující poznatky k danému případu.

Proces je znázorněný na schématu Obr. 3.2, které vykresluje zpracování vzorku v DNA laboratoři, samotné zpracování MPS dat je popsáno v další kapitole.



Obrázek 3.2: Průběh celého procesu od příjmu pacienta na oddělení po sdělení výsledku pacientovi

3.2.1 Sekvenování

3.2.1.1 Panel genů pro časné a těžké dětské epilepsie a epileptické encefalopatie

Sekvenování a příprava knihoven probíhá v laboratoři Macrogen Europe (Amsterdam, Nizozemsko) na sekvenátoru HiSeq 4000 (Illumina, USA) s 2×150bp sekvenačním kitem. Pro MPS byly do poloviny roku 2016 využívány kity HaloPlex Target enrichment (Agilent, USA) od této doby je využíván systém SureSelect Target Enrichment (Agilent, USA) s vyšším pokrytím (alespoň 1000×).

Design panelu genů Geny do panelu byly vybrány dle následujících kritérií:

1. Nejméně dvě nezávislé publikace, které spojují gen s epilepsií
2. Alespoň jedna publikace popisující kauzální varianty v genu u dvou nebo více nepříbuzných pacientů

Seznam genů v panelu je uvedený na obrázku Obr. 3.3.

3.2.1.2 Panel genů pro dědičné neuropatie

Sekvenování probíhá v laboratořích - dříve v EMBL genomics core facilities (Heidelberg, Německo), později v Macrogen Europe (Amsterdam, Nizozemsko). V současnosti je využíván stejný sekvenanční kit jako u panelu genů spojených s epilepsií na sekvenátoru HiSeq 4000 (Illumina, USA). Pro sekvenování byly do poloviny roku 2016 využívány kity HaloPlex Target enrichment (Agilent, USA), od této doby je využíván systém SureSelect Target Enrichment (Agilent, USA) s vyšším pokrytím (alespoň 500×).

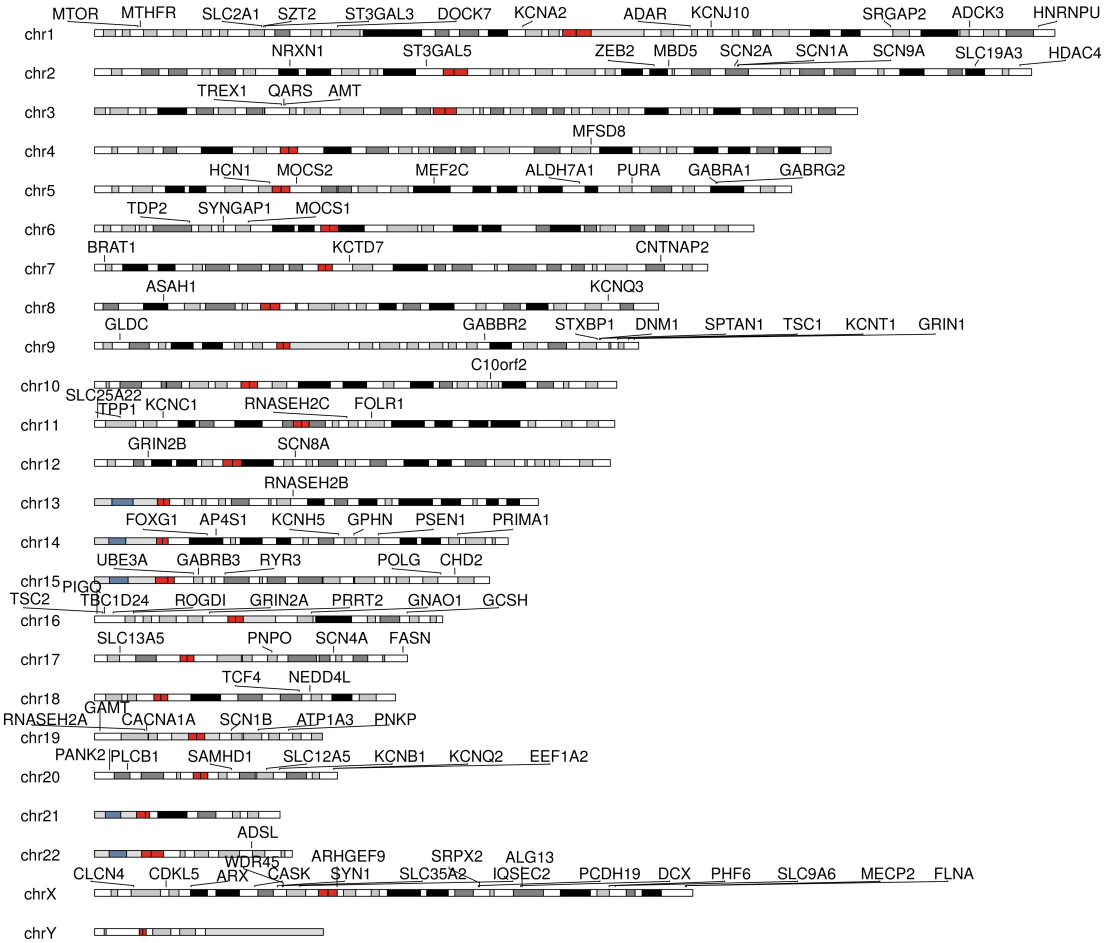
Design panelu genů Geny do panelu byly vybrány dle následujících kritérií:

1. Nejméně dvě nezávislé publikace, které spojují gen s dědičnou neuropatií
2. Geny, které byly publikované jako patogenní minimálně ve dvou rodinách nebo v jedné, rozvětvené rodině, kdy patogenicita byla ověřena funkční studií
3. Kandidátní geny, které byly detekovány u našich pacientů a hledáme druhou rodinu pro potvrzení (geny *SH3PB4*, *ITPR3*, *KLHL13* dle [Schabhüttl et al. 2014], a dále gen *SLC18A3)*

Seznam genů v panelu je uvedený na obrázku Obr. 3.4.

3.2.1.3 WES a WGS

Pro sekvenování WES vzorků využíváme laboratoř Macrogen Europe (Amsterdam, Nizozemsko), sekvenace probíhá pomocí kitů SureSelectXT verze 6 (Agilent, USA), s průměrnou hloubkou čtení alespoň 100×. Pro sekvenování WGS vzorků využíváme stejnou laboratoř, s využitím kitu TruSeq DNA PCR Free (Illumina, USA) s pokrytím alespoň 30×.



ADAR	*DOCK7*	*HNRNPU*	*NEDD4L*	*SAMHD1*	*SYNGAP1*
ADCK3	*EEF1A2*	*CHD2*	*NRXN1*	*SCN1A*	*SZT2*
ADSL	*FASN*	*IQSEC2*	*PANK2*	*SCN1B*	*TBC1D24*
ALDH7A1	*FLNA*	*KCNA2*	*PCDH19*	*SCN2A*	*TCF4*
ALG13	*FOLR1*	*KCNB1*	*PHF6*	*SCN4A*	*TDP2*
AMT	*FOXG1*	*KCNC1*	*PIGQ*	*SCN8A*	*TPP1*
AP4S1	*GABBR2*	*KCNH5*	*PLCB1*	*SLC12A5*	*TREX1*
ARHGEF9	*GABRA1*	*KCNJ10*	*PNKP*	*SLC13A5*	*TSC1*
ARX	*GABRB3*	*KCNQ2*	*PNPO*	*SLC19A3*	*TSC2*
ASAH1	*GABRG2*	*KCNQ3*	*POLG*	*SLC25A22*	*UBE3A*
ATP1A3	*GAMT*	*KCNT1*	*PRIMA1*	*SLC2A1*	*WDR45*
BRAT1	*GCSH*	*KCTD7*	*PRRT2*	*SLC35A2*	*ZEB2*
C10ORF2	*GLDC*	*MBD5*	*PSEN1*	*SLC9A6*	
CACNA1A	*GNAO1*	*MECP2*	*PURA*	*SPTAN1*	
CASK	*GPHN*	*MEF2C*	*QARS*	*SRGAP2*	
CDKL5	*GRIN1*	*MFSD8*	*RNASEH2A*	*SRPX2*	
CLCN4	*GRIN2A*	*MOCS1*	*RNASEH2B*	*ST3GAL3*	
CNTNAP2	*GRIN2B*	*MOCS2*	*RNASEH2C*	*ST3GAL5*	
DCX	*HCN1*	*MTHFR*	*ROGDI*	*STXBP1*	
DNM1	*HDAC4*	*MTOR*	*RYR3*	*SYN1*	

Obrázek 3.3: Panel vyšetřených genů, ve kterých jsou varianty asociované s epilepsií, s vyobrazením pozice genu na chromozomu a v tabulce

AARS	*EGR2*	*LRSAM1*	*PRPS1*	*SPTLC2*
AIFM1	*FAM134B*	*MARS*	*PRX*	*SURF*
ARHGEF10	*FBLN5*	*MED25*	*RAB7A*	*TFG*
ATL1	*FBXO38*	*MFN2*	*REEP1*	*TRIM2*
ATP1A1	*FGD4*	*MICAL1*	*SBF1*	*TRIM2*
ATP7A	*FIG4*	*MME*	*SBF2*	*TRPV4*
BAG3	*GAN*	*MORC2*	*SCN11A*	*TUBB3*
BICD2	*GARS*	*MPZ*	*SCN9A*	*TTR*
BSCL2	*GDAP1*	*MT-ATP6*	*SEPT9*	*VAPB*
CCT5	*GJB1*	*MTMR2*	*SETX*	*VCP*
CHCHD10	*IKBKA*	*MYH14*	*SH3BP4*	*WNK1*
COX6A1	*INF2*	*NDRG1*	*SH3TC2*	*YARS*
CTDP1	*ITPR3*	*NEFL*	*SIGMAR1*	
DCTN1	*KARS*	*NGF*	*SLC12A6*	
DHTKD1	*KIF1A*	*NIF3L1*	*SLC18A3*	
DNAJB2	*KIF1B*	*NTRK1*	*SLC5A7*	
DNM2	*KIF5A*	*PDK3*	*SOD1*	
DNMT1	*KLHL13*	*PLEKHG5*	*SOX10*	
DRP2,DST	*LITAF*	*PMP2*	*SPG11*	
DYNC1H1	*LMNA*	*PMP22*	*SPTLC1*	

Obrázek 3.4: Panel vyšetřených genů, ve kterých jsou varianty asociované s dědičnými neuropatiemi, s vyobrazením pozice genu na chromozomu a v tabulce

3.2.2 Bioinformatické zpracování

Bioinformatické zpracování dat začíná příjmem sekvenačních dat – jedná se o dva soubory FASTQ. Další zpracování se mírně liší dle použité knihovny. V této sekci popisujeme metody použité pro alignment a variant calling ze sekvenančních dat, anotace variant je popsána v sekci 3.2.5.

Na našem pracovišti v současnosti používáme 3 nástroje pro kompletní bioinformatickou analýzu – 2 komerční nástroje s grafickým uživatelským rozhraním (GUI) SureCall (Agilent, USA) a NextGENe (Softgenetics, USA) a „GATK best practices“ pipeline v příkazové řádce na pracovní stanici s linuxem. Všechny tři metodiky pracují na stejném principu - alignmentu a poté variant callingu, proto není vhodné využívat všechny tři zároveň. Pro vyhodnocování MPS dat se nám v DNA laboratoři osvědčilo využívat vždy kombinaci dvou nástrojů pro analýzu (pro každý jeden vzorek).

Celý proces bioinformatického zpracování je znázorněný na schématu Obr. 3.5.

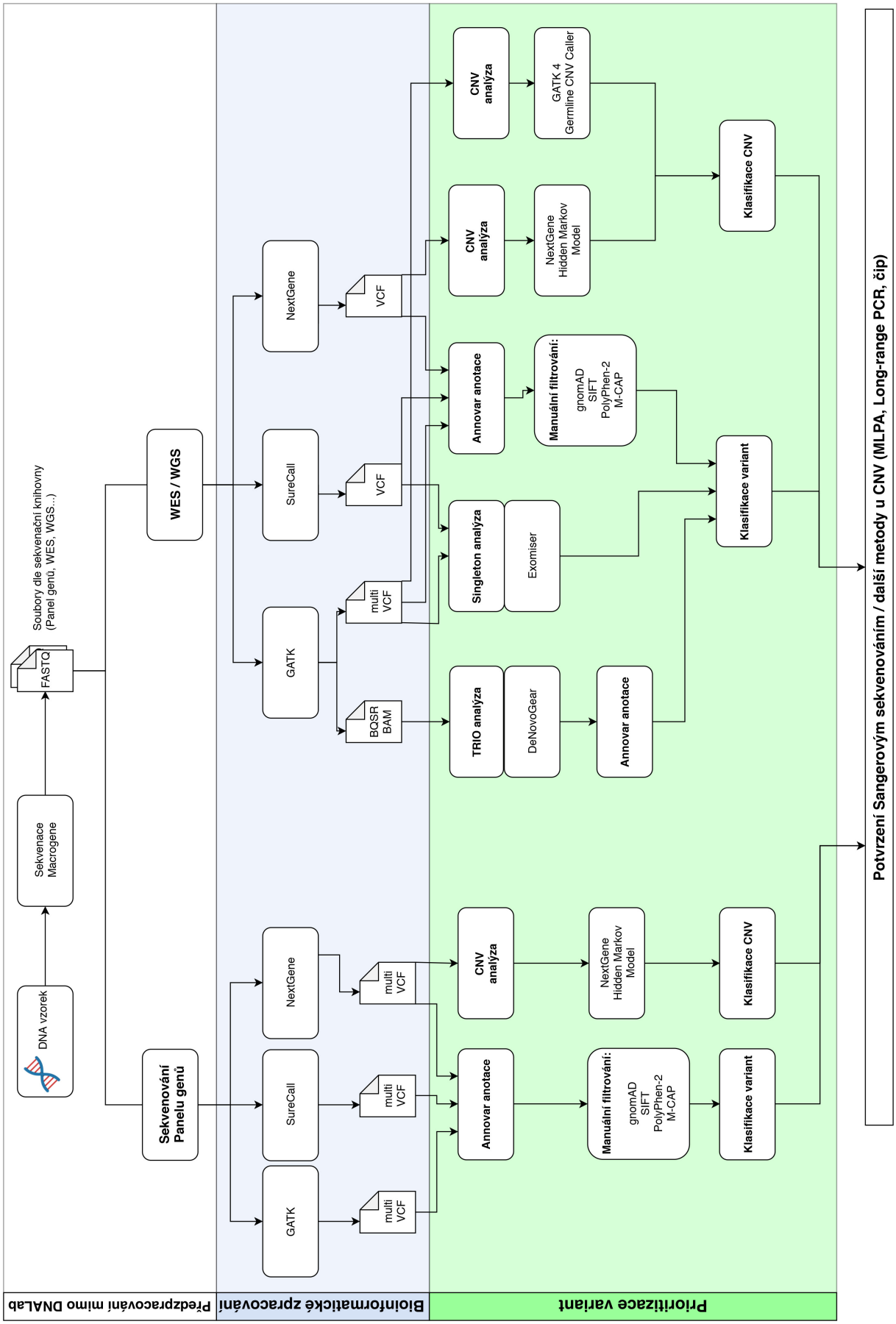
3.2.3 Nástroje NextGene a SureCall pro zpracování MPS dat

Implementace těchto nástrojů do vyhodnocovacího postupu DNA laboratoře umožňuje kvalitní analýzu dat i uživatelům bez bioinformatických znalostí. Celý proces analýzy se dá rozdělit do několika jednoduchých kroků:

1. Načtení dat – uživatel vybere dvojici FASTQ souborů (NextGene provádí konverzi FASTQ do FASTA)
2. Uživatel vybere referenci pro alignment – verze genomu – v současnosti využíváme hg19
3. Uživatel vybere design studie – panel genů, WES nebo WGS, podle toho je vybrán BED soubor obsahující koordináty cílové sekvence
4. Probíhá alignment FASTQ souborů s referenčním genomem
   a) V prvním kroku dojde k alignmentu na referenční sekvenci hg19
   b) Podle designu studie je (ne)provedeno odebrání PCR duplikátů ze sekvence
5. Dojde k uložení BAM souboru a probíhá variant calling
6. Výsledný soubor VCF se seznamem variant je dostupný pro uživatele pro další zpracování buď ve formě multiVCF – více vzorků z jedné knihovny sloučeno dohromady nebo VCF, které obsahuje vždy jeden unikátní vzorek.

Výstupem pro další analýzu jsou tak minimálně dva soubory, VCF s variantami a BAM, který nám ve vhodném prohlížeči (IGV, Alamut Visual) umožní prohlížet data z širší perspektivy – ne variantu po variantě, ale i po readech a větších úsecích.

Vhodná je kombinace těchto nástrojů, protože NextGene používá z velké části svoje vlastní algoritmy pro Alignment i Variant calling, oproti tomu SureCall lze chápat spíše jako grafické rozhraní obalující běžné, volně dostupné algoritmy (BWA-MEM a Picard pro variant calling).



Obrázek 3.5: Celý proces bioinformatického zpracování dat

3.2.4 GATK best practices

V této sekci jsou popsány bioinformatické postupy založené na doporučení GATK, dle této metodiky jsme vyvinuli celý systém pro správu dat v DNA laboratoři, popsanou v další sekci. Cílem je mít standardizovaný postup pro zpracování dat, který nám poskytne reprodukovatelné výsledky, které jsou porovnatelné s daty z velkých sekvenačních projektů (např. gnomAD [Karczewski et al. 2019]). Právě proto jsme zvolili GATK ve verzi 3.8. [DePristo et al. 2011; Auwera et al. 2013]

Proces analýzy se skládá z podobných kroků jako u předchozích nástrojů, ale přístup je odlišný, je využíváno několika nástrojů, kdy každý musí být vyvolán zvlášť. Schéma bioinformatického zpracování bylo již uvedeno v úvodu na Obr. 1.11.

Analýza probíhá v krocích:

1. Pro dvojici FASTQ souborů jsou definovány: název vzorku (číselný kód), typ knihovny (WES/WGS/TGP – pro panel), typ sekvenátoru (Illumina), délka readů (150), BED soubor a cesta k referenční sekvenci. Všechny tyto parametry jsou vstupem do programu BWA-MEM. Ten provádí alignment vzorku vůči referenci. Výsledkem je soubor SAM.

2. Soubor SAM je nástrojem samtools převeden do formátu BAM a dojde k jeho indexaci (vytvoření souboru .bam.bai).

3. Pokud uživatel zvolí, že chce odebrat PCR duplikáty je nástrojem Picard BAM dále zpracován – vzniká BAM označený jako remdup_BAM.

4. Nástroj GATK provádí rekalibraci kvality (BQSR) algoritmus v několika krocích – výstupem je remdup_BQSR_BAM (nebo BQSR_BAM pokud nedošlo k odstranění duplikátů) a je dokončený krok alignmentu.

5. Z BQSR_BAM souboru nástroj GATK tvoří g.vcf, obsahující všechny pokryté oblasti, nejen tedy oblasti s odlišností od reference.

6. Všechny g.vcf soubory z jednoho běhu jsou dále zpracovány algoritmem GATK HaplotypeCaller, který vyvolává varianty na základě výskytu ve všech vzorcích v daném běhu. Výsledkem pak je jeden multiVCF soubor obsahující varianty ze všech vzorků, kdy je samozřejmě možné určit, který vzorek variantu obsahuje a který ne (dle genotypu).

7. Výsledný VCF soubor může být dále filtrován a rekalibrován dle doporučení GATK, avšak rekalibrace je doporučována až od počtu 20 WES v rámci jednoho běhu.

8. Uživatel získává soubor multiVCF k prioritizaci.

3.2.5 Prioritizace variant – anotování

Proces prioritizace variant navazuje bezprostředně na bioinfomatickou analýzu. V DNA laboratoři se nám osvědčil postup alespoň dvou nezávislých analýz, kdy v rámci každé analýzy jsou zpracovány dva soubory VCF z různých zdrojů.

Prvním krokem je získání co největšího množství informací o každé variantě. K tomu napomáhají tzv. anotační nástroje, kdy využíváme dva – ANNOVAR [Wang, Li a Hakonarson 2010] a Alamut Batch (Interactive Biosoftware, Francie) dostupný z[1]. Tyto nástroje prohlíží vstupní soubor variantu po variantě a získávají o nich informace z různých zdrojů – z populačních databází, z predikčních nástrojů a dalších zdrojů. Tento krok je velmi časově náročný, ale probíhá automaticky, uživatel tedy není k anotaci potřeba.

Výstupem anotace je tabulka s variantami a všemi dostupnými informacemi. Nyní je na expertovi, který provádí prioritizaci variant, aby provedl filtrování – tedy rozhodl, které varianty jsou pro případ relevantní, a které nejsou. Kritéria mohou být značně variabilní a je vhodné pracovat s jednotlivými parametry jako s „měkkými filtry“ – to znamená, že pokud vidíme, že varianta je označená jedním predikčním nástrojem jako benigní, nelze jí zcela jistě vyloučit – jedná se tedy o vícekrokový proces, při kterém nám pomáhají i ACMG kritéria (viz výše 1.6.4).

Základní postup pro manuální prioritizaci je následující:

- Deprioritizace variant s frekvencí v dané populaci (European-non-Finnish) nebo s celkovou frekvencí v databázi větší než 1 %.
- Lokalizace varianty v rámci genu – preference exonových oblastí před ostatními (introny, UTR).
- Typ varianty – dle efektu na výsledný protein; preference missense, stop mutace preferovaná před synonymními.
- Konzervovanost nukleotidu – dle tohoto parametru vidíme, jestli se aminokyselina (nebo nukleotid) v sekvenci shoduje mezi různými organismy, čímž dokážeme posoudit „důležitost“ produktu – phyloP skóre.
- Predikční skóre na základě nástrojů – PolyPhen-2, SIFT, M-CAP, ClinVar.
- Informace o genu, ve kterém se varianta nachází – OMIM, UniProt.

Po vyfiltrování variant zůstává několik jednotek až nižších desítek variant, které jsou dále posuzovány – je možné si prohlédnout BAM soubor v prohlížeči (IGV dostupné z [2] nebo Alamut Visual dostupné z [3]), abychom viděli poměr přečtené referenční a alterované báze – tím dokážeme odhadnout jestli nejde o falešně pozitivní variantu. Posledním krokem je pak kontrola databáze HGMD [Stenson et al. 2003], která obsahuje informaci, jestli k dané variantě byla vydaná publikace.

Pokud i po těchto krocích existuje varianta, která se jeví jako zajímavá, přistupujeme k ověření varianty Sangerovým sekvenováním a poté k segregační analýze, která se provádí metodou Sangerova sekvenování jak u pacienta, tak u obou rodičů, pokud je k dispozici vzorek.

[1] https://www.interactive-biosoftware.com/alamut-batch/ [online: 16.10.2019]
[2] https://igv.org/ [online: 16.10.2019]
[3] https://www.interactive-biosoftware.com/alamut-visual/ [online: 16.10.2019]

3.2.6 Pokročilá prioritizace variant

Pro pokročilou prioritizaci variant nám nestačí pouze sekvence pacienta, ale potřebujeme zajistit další vstupy. K pokročilým metodám přistupujeme u pacientů s WES, kdy manuální metody již přestávají být efektivní. V praxi jsme testovali mnoho nástrojů, jak u singleton analýzy (analýza pouze probanda bez rodičů), tak u trio analýzy (sekvenace probanda i rodičů). V současnosti jsme po dlouhodobém testování do naší workflow zařadili nástroje DeNovoGear a Exomiser (s HPO termíny pro popis fenotypu).

DeNovoGear Nástroj DeNovoGear [Ramu et al. 2013] (DNG) slouží ke zpracování MPS dat pro detekci *de novo* variant z WES a WGS dat. U onemocnění, která jsou často způsobena *de novo* variantami přistupujeme k tzv. Trio analýze, tato analýza spočívá v porovnávání sekvence probanda a jeho rodičů. Pokud použijeme „naivní" přístup pro detekci *de novo*, tak nám stačí VCF soubor s variantami a srovnáváme jednotlivé genotypy probanda proti rodičům. Pokud je varianta u probanda přítomná a u rodičů ne, pokládáme jí za *de novo*. Takový přístup je ale zcela závislý na kvalitě sekvenačních dat a pokud dojde byť jen k jediné chybě ve variant callingu, tak už nikdy nedostaneme validní výsledek. Proto je nutné přistoupit k modelu, který vypočítává pravděpodobnosti výskytu *de novo* varianty z BAM souboru (tedy z readů přiřazených k referenční sekvenci).

Vstupem nástroje DNG jsou soubory BAM od probanda a rodičů, soubor PED s definovaným rodokmenem a referenční sekvence. Nástroj pak vypočítává pravděpodobnost jednotlivé *de novo* varianty dle zadané pravděpodobnosti variability (parametr definovaný uživatelem) a pravděpodobnosti přenosu od rodičů. Výsledkem je pak hodnota pravděpodobnosti, s jakou varianta vznikla *de novo*.

Exomiser Nástroj Exomiser byl publikovaný v roce 2014 [Robinson et al. 2014] a jeho zavedení do workflow pro analýzu WES dat pro nás znamenal velký pokrok. Nástroj Exomiser využívá algoritmu PHIVE (PHenotypic Interpretation of Variants in Exomes), který byl implementován pro výpočet shody mezi fenotypem a genotypem jedince. Míra shody je vypočítána na základě informací o fenotypu pacienta. Zde jsou dvě možnosti – pokud si jsme jistí správností diagnózy, je možné zadat jeho OMIM kód, pokud ne, tak je možné zadat HPO termíny popisující jednotlivé fenotypové znaky a algoritmus dle znaků hledá variantu, která je s takovým fenotypem spojována.

Vstupem je VCF, buď samotného probanda nebo VCF probanda a rodičů s definovaným vztahem pomocí PED souboru. Rovněž je nutné zadat HPO termíny nebo OMIM kódy, které jsou navrženy dle záznamů z karty pacienta (obsahující popis fenotypu). Analýza je časově nenáročná, do 15 minut je znám výsledek z WES VCF s jednotlivými variantami ohodnocenými parametrem PHIVE skóre.

HPO termíny Propojení mezi genotypem a fenotypem je velmi často komplikované, protože při vyhodnocování pracujeme s nestrukturalizovaným záznamem. Proto vznikl projekt The Human Phenotype Ontology (HPO) [Köhler et al. 2014], který se snaží o unifikaci popisu fenotypu a stanovení jejich vzájemných vztahů.

Výsledkem je databáze obsahující více než 10 000 záznamů a 13 tisíc vzájemných propojení. Díky tomu můžeme popsat i komplikované fenotypy několika málo HPO termíny. Záznamy jsou v databázi hierarchizované – to znamená, že pokud vyhledáme např. termín „Peripheral neuropathy“, tak můžeme dále pracovat s termíny nadřazenými (Abnormal peripheral nervous system morphology), tak i podřazenými (např. Sensory neuropathy, Polyneuropathy atd...). Výhodou takové relační databáze je možnost hledání vztahů mezi fenotypy – tzn. dokážeme fenotypy např. propojit s komplexním syndromem a nebo je využít pro spárování s genotypem – k dispozici je i informace o genech spojených s termínem. Každý termín má unikátní identifikátor, který se dále aplikuje při anotacích, zejména v programu Exomiser. [Köhler et al. 2014]

3.2.7 Metody pro detekci CNV v datech z MPS

Prvním krokem pro provádění CNV analýzy bylo nalezení vhodného nástroje pro detekci CNV. Jako první jsme se rozhodli otestovat nástroj metaSV [Mohiyuddin et al. 2015] , který je určený pro hledání strukturálních variant. Funguje na principu agregace výstupů z dalších nástrojů pro detekci strukturálních variant.

Mezi nástroje, jimiž jsou data analyzována před samotným vstupem do metaSV patří:

- CNVkit – [Talevich et al. 2016] – je univerzální nástroj pro hledání CNV aplikovatelný na všechny typy NGS dat – panely genů, WES i WGS data. Vyvolání CNV funguje na principu hloubky čtení. Nástroj nejprve získá informace o průměrné hloubce čtení v celém BAM souboru a poté hledá významné odchylky od průměrné hloubky v rámci prohledávaných oblastí (definovaných BED souborem). Při analýze WGS dat jsou ke zpřesnění analýzy dopočítány další parametry – obsah GC bází (GC content) a místa s repetitivními oblastmi, které by WGS CNV analýzu zkreslovaly.

- CNVnator – [Abyzov et al. 2011] – nástroj v základním nastavení určený primárně pro WGS data, využívá algoritmu mean-shift pro hledání CNV dle změn v signálu vypočítaného dle průměrné hloubky pokrytí zvoleného okna. Pro každé okno je vypočítaná průměrná hloubka čtení, kdy dochází k seskupování podobných hodnot a detekování co největších skoků mezi takovými shluky, podle toho se pak vyvolávají CNV při porovnání velikosti skoků vůči průměrným hodnotám v genomu.

- Nástroje pro detekci „break-pointů“ – Tyto nástroje fungují na principu hledání „break-pointů“, tedy míst v genomu, ve kterých došlo k deleci popř. inzerci genetické informace do sekvence. Algoritmus mapuje sekvenci pacienta k referenci, pokud dojde ke správnému mapování celého readu, tak pokračuje dál, pokud ale část readu neodpovídá referenci, dojde k posunutí této části readu dál po sekvenci, dokud nedojde k namapování zbytku readu na referenční sekvenci – vzniklá mezera mezi částmi readu v referenční sekvenci je tedy vyvolána jako potenciální delece – čím více readů vykazuje tento efekt, tak tím je skóre pro danou deleci vyšší. Jedná se o nástroje:

  - Pindel [Ye et al. 2009], Manta [Chen et al. 2015], BreakSeq2 [Abyzov et al. 2015], Lumpy [Layer et al. 2014], WHAM [Kronenberg et al. 2015]

- BreakDancer – [Fan et al. 2014] – slouží k detekci velkých strukturálních variant ve NGS datech, provádí analýzu tzv. ARP readů (anomalous read pair / anomální páry readů), tedy readů, u kterých došlo ke korektnímu mapování pouze jednoho z páru readů. Počet ARP readů je pak průměrován v rámci celé sekvence a hledají se sekvence, které se v tomto počtu liší od průměru. ARP ready mohou vznikat na rozhraní CNV, kdy díky změně sekvence (u probanda) nedojde ke korektnímu namapování na referenci.

Všechny nástroje byly předem otestovány na testovacích datech, která jsou standardně přiložená k nástrojům a testy proběhly bez chyb.

3.2.7.1 Detekce zárodečných CNV pomocí GATK 4 germline pipeline

V současné době (červen 2019) nebyla k dispozici metodika GATK Best Practices pro germinální CNV. Přestože jsou v aktuální verzi GATK 4.1.0.0 algoritmy pro detekci CNV připravené, jedná se o beta verzi, to znamená, že funkčnost nelze garantovat a není ověřená. Z tohoto důvodu jsme nevolili GATK germinální CNV workflow jako metodu první volby. Pro implementaci bylo využito dokumentace na URL [4].

Postup analýzy pro detekci zárodečných CNV Analýza je provedena v těchto krocích (přehled celého procesu na Obr. 3.6)

1. Prvním krokem analýzy je vytvoření modelu pro CNV analýzu, pro tento krok jsou vstupem BAM soubory z WES (v našem případě 45 WES BAM souborů).

2. Pomocí nástroje GATK algoritmu *CollectReadCounts* dojde ke určení počtu readů v prohledávaném intervalu. Interval je definovaný BED souborem. Výstupem je pro každý vzorek soubor HDF5.

3. Dochází ke karyotypování každého vzorku, tedy určení počtu kopií každé informace ve vzorku. Kdy u autozomů předpokládáme vždy dvě kopie genu, u chromozomu X buď jednu nebo dvě kopie a u chromozomu Y jednu nebo žádnou kopii genu. Tento krok je zpracován algoritmem *DetermineGermlineContigPloidy* - tím dojde k vytvoření základního modelu pro vyvolání CNV - nástroj vypočítává počet readů na daném contigu (část sekvence DNA dle intervalů) dle karyotypu.

4. Algoritmus *GermlineCNVCaller* porovnává jednotlivé contigy a jejich pokrytí, ze kterého pak modeluje model pokrytí „coverage model“. Na základě tohoto modelu pak dokáže určit odchylku a tím detekuje jak možné zvýšení, či snížení počtu kopií.

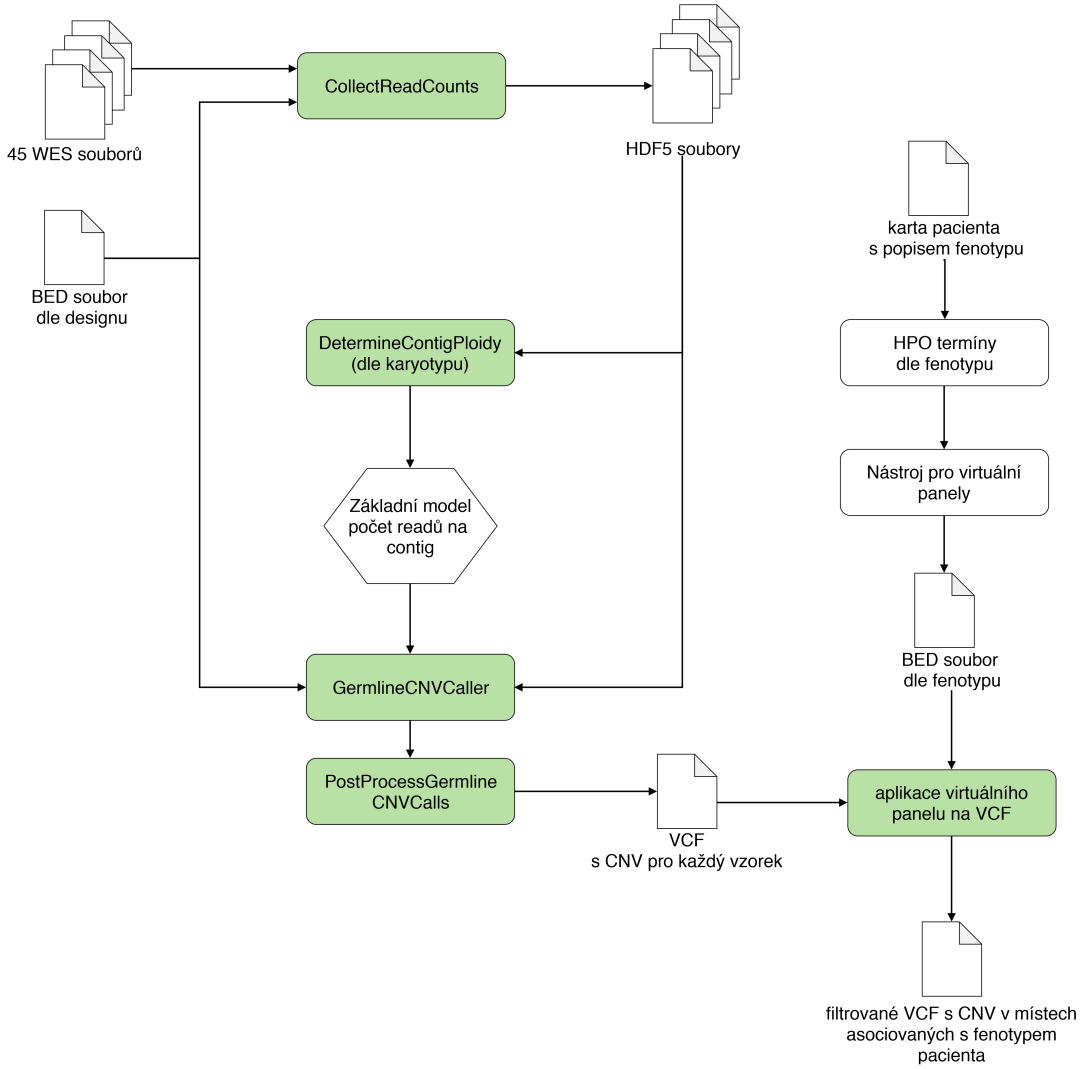
5. Posledním krokem je vyvolání algoritmu *PostprocessGermlineCNVCalls*, který provádí vyvolání jednotlivých CNV ve gVCF formátu pro každý vzorek. Interpretace je pak na základě hodnoty GT u každého CNV v souboru, kdy 0

[4] https://gatkforums.broadinstitute.org/gatk/discussion/11684 [online: 16.10.2019]

značí normální výsledek, 1 značí deleci a 2 duplikaci. Hodnota *QA* pak udává Phred-skóre popisující pravděpodobnost správného vyvolání.

6. Pro snadnější interpretaci jsou varianty vyfiltrovány pomocí virtuálního panelu (BED soubor dle fenotypu pacienta více v sekci 4.6)

Množství detekovaných CNV je velmi velké (kolem 250 tisíc na jeden WES vzorek), proto se snažíme toto číslo zmenšit pouze na oblasti, které jsou relevantní pro náš případ. Toho dosáhneme aplikací virtuálního panelu ve formě BED souboru, který hledání omezí na oblasti spojované s onemocněním.



Obrázek 3.6: Schéma analýzy CNV pomocí „GATK 4 germline pipeline“

3.2.8 In-house databáze WES variant u pacientů shromážděných v DNA laboratoři

Pro vybudování „in-house“ databáze variant jsme využili data všech pacientů s WES, kteří měli předchozí výsledek z testování panelem genů negativní. Jednalo se o 222 jednotlivců (121 mužů, 101 žen).

Data byla analyzována dle GATK best practices (BWA-MEM, Picard, BQSR rekalibrace) až po gVCF, poté došlo k vyvolání variant algoritmem JointGenotyping pro 222 WES vzorků dohromady, čímž vznikl výsledný soubor multiVCF obsahující všechny varianty u všech pacientů. Výsledný soubor byl anotován nástrojem ANNOVAR pro získání informací o výskytu variant v populačních databázích a predikcích na základě predikčních nástrojů.

Následně byla dopočítaná frekvence alel v naší subpopulaci dle vztahu:

$$dbAF = \frac{nHET + nHOM * 2}{nWES * 2} \tag{3.1}$$

, kde: *inhousedbAF* je frekvence alely v naší subpopulaci, *nHET* je počet pacientů s variantou v heterozygotní formě (GT pole ve VCF u vzorku udává hodnotu „0/1“) *nHOM* je počet pacientů s variantou v homozygotní formě (GT pole ve VCF u vzorku udává hodnotu „1/1“) a *nWES* je počet všech WES vzorků. Výsledkem je údaj o frekvenci kdy všechny uvedné varianty mají frekvenci větší než 0 (jinak by nebyly vyvolány).

Pro další analýzu jsme využili informace o délce genů z databáze RefSeq [O'Leary et al. 2015] dostupné na URL[5]. Z délky kódující informace genu jsme pak vypočítali parametr *varpb* dle vztahu:

$$varbp = \frac{nHETex + nHOMex * 2}{delka(cDNA)} \tag{3.2}$$

kde *nHETex* je počet pacientů s variantou v heterozygotní formě (GT pole ve VCF u vzorku udává hodnotu „0/1“), nacházející se v kódující oblasti genu, *nHOMex* je počet pacientů s variantou v homozygotní formě (GT pole ve VCF u vzorku udává hodnotu „1/1“), nacházející se v kódující oblasti genu, a *delka(cDNA)* je délka kódující sekvence daného genu, ve kterém byly varianty nalezeny v bp.

[5]https://www.ncbi.nlm.nih.gov/refseq/ [online: 16.10.2019]

3.2.9 Databáze proteinových domén prot2HG

Při prioritizaci variant se často řídíme funkční změnou, to znamená, jaký vliv má varianta na polypeptidový řetězec, jestli dochází ke změně jedné AMK nebo prodloužení/zkrácení celého řetězce. Při takovém procesu, ale často nevidíme o krok dál, tedy do proteinového produktu. Každý protein má několik funkčních částí, pokud tuto funkci známe a umíme jí popsat, tak tuto část nazýváme proteinovou doménou. Pokud tedy dokážeme determinovat funkci proteinové domény, lze předpokládat, že je její funkce signifikantní, a proto je důležité vědět, jestli nedošlo ke změně právě v doméně. Tuto informaci nám ale současné anotační nástroje nedávají v jednoduché, přístupné formě. Proto jsme ve spolupráci s pracovištěm Hussman Institute for Human Genomics v Miami USA, během mé stáže, vytvořili projekt, který si kladl za cíl namapovat všechny známé proteinové domény na genomickou sekvenci lidského genomu (verze hg19).

V pilotní fázi projektu bylo zpracováno jen kolem 100 genů, klinicky prokazatelně spojených s dědičnými neuropatiemi. V dřívější studii bylo ukázáno, že patogenní varianty v genu *SYT2* byly nalezeny právě v krátké proteinové doméně, čímž došlo k projevu fenotypu onemocnění [Whittaker et al. 2015]. Proto jsme přistoupili k pilotní fázi, kdy jsme zpracovávali geny asociované s dědičnými neuropatiemi, výsledky jsou shrnuté v 4.4.3. Jelikož se výsledek této pilotní fáze ukázal jako hodnotný pro vyhodnocování variant, pokročili jsme do další fáze, kdy jsme přistoupili ke zpracování všech publikovaných proteinových domén.

V tomto projektu jsme se zaměřili na dva typy proteinových domén, „Sites“ a „Regions“, dle definice RefSeq [O'Leary et al. 2015]. Doména typu „Site“ má velmi krátkou sekvenci s definovanou funkcí, tento typ je často funkční podjednotkou jiných, větších domén (jedna se často o transmembránové regiony, místa nitrosylace nebo fosforylace). Domény typu „Region“ jsou většinou delší, s délkou přes 100 AMK, jedná se o část proteinu formující 3D strukturu [Doolittle 1995]. Typickou doménou tohoto typu jsou zinkové prsty, nebo repetice bohaté na Leucin.

Pro namapování proteinových domén na referenční genomovou sekvenci jsme vytvořili 4 krokový postup, shrnutý na obrázku Obr. 3.7 a popsaný níže:

A) Získání dat z databáze NCBI Prvním krokem bylo sesbírání všech proteinových dat z databáze NCBI (dostupné z URL[6]). Nejdříve jsme vyexportovali přes webové rozhraní seznam všech známých proteinů t.č. 42 371 ve formě jejich RefSeq identifikátorů (tvar NP_ID). Pomocí tohoto seznamu a nástrojů EZ utilities [Sayers a Wheeler 2004] jsme získali kompletní informace o všech proteinech, jejich sekvenci, identifikátor, záznamy o všech doménách s koordináty v rámci proteinové sekvence, jejich délce a popisem. U každého proteinu byla v záznamu rovněž uvedená informace o jeho korespondujícím genu, jehož produktem je daný protein (identifikátor NM_ID). Díky tomu jsme mohli aplikovat stejný postup a získat informace o genech a jejich kódující sekvenci cDNA a pozici od které cDNA tvoří proteinový produkt.

B) Zpracování dat Zpracování dat probíhalo skriptem vytvořeným v Pythonu 3.5, kdy nejprve pomocí knihovny *BeautifulSoup* došlo k rozdělení dat do bloků dle typu

[6]https://www.ncbi.nlm.nih.gov/protein [online: 16.10.2019]

záznamu, a následnému uložení do vlastní datové struktury. Ta obsahovala všechny dříve získané informace pohromadě již správně identifikované (tento krok v praxi znamená, že skript ve staženém záznamu správně vyhledá blok sekvence a označí ho jako sekvenci, najde identifikátor genu atd.). Dalším krokem je pak předzpracování samotných domén. Nyní známe pro každou doménu její název, typ (Region nebo Site), její popis a původ a souřadnice, kde se nachází v proteinu. Díky tomu, že známe celou sekvenci, tak lze snadno zjistit sekvenci AMK pro danou doménu, takovou sekvenci si tedy uložíme pro všechny domény v proteinu.

C) Mapování domén k DNA sekvenci Získaná AMK sekvence nám nyní umožňuje zjistit, jak vypadá cDNA sekvence, tu získáváme algoritmem zpětné translace. Protože většina AMK je kódována více než jedním kodonem [Ycas 1969], využili jsme notaci dle IUPAC [Johnson 2010], která umožňuje nejednoznačné nukleotidy nahradit korespondujícím znakem (například: Glycin má kodony GGT, GGC, GGA, GGG, proto překládáme Glycin jako GGN, kdy N označuje jako možné nukleotidy A C G T). Tímto algoritmem tak získáváme přeloženou sekvenci (rtcDNA). Pro další zpracování ale potřebujeme určit, v jakém místě kódující DNA se doména nachází.

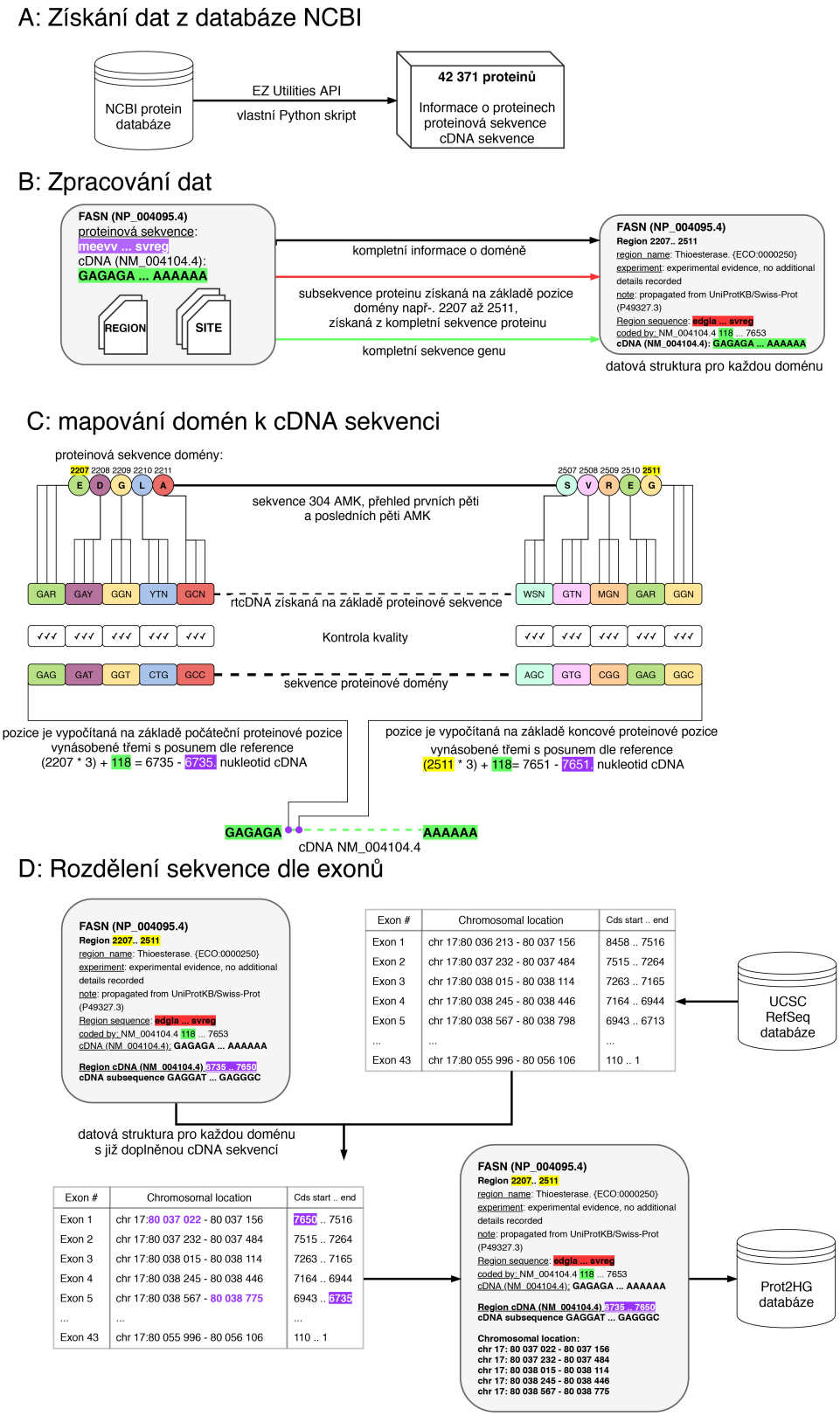
Zde použijeme kódující sekvenci genu, kterou jsme získali v kroku **A)**. Hledáme tedy v cDNA genu sekvenci, která je shodná se získanou zpětnou translací rtcDNA. Nejprve určíme souřadnice sekvence domény, souřadnice získáme jednoduchým výpočtem, kdy počáteční a koncovou souřadnici domény v rámci proteinu vynásobíme 3 (počet nukleotidů v kodonu) a následně tento interval posuneme, dle pozice od které cDNA tvoří produkt (získali jsme jí ze záznamu genu).

Pro další zpracování je ale potřeba kontrola správnosti, tedy jestli rtcDNA odpovídá nalezené cDNA. Máme souřadnice pro cDNA, získáme tedy požadovanou subsekvenci a porovnáme jí s námi vytvořenou rtcDNA. Z tohoto porovnávání pak vypočítáme mapping skóre (procento shodujících se nukleotidů), kdy předpokládáme hodnotu 1.0 (100 %).

D) Rozdělení sekvence dle exonů Posledním krokem je nalezení genomické pozice celé domény. Nyní známe lokaci domény v rámci cDNA, to nám ale pro anotování variant nestačí, chceme znát umístění sekvence, která danou doménu kóduje v rámci chromozomu.

Pro tento krok jsme nejprve stáhli genomické souřadnice všech kódujících exonů (vždy začátek a konec) z USCS RefSeq databáze s využitím MySQL API (dostupné z URL[7]). Nyní využijeme známé cDNA souřadnice, a najdeme jim odpovídající část pokrývající jeden nebo více exonů (nebo i část exonu), a tyto intervaly si uložíme do datové struktury, čímž jsou informace o proteinové doméně kompletní, známe pro každou doménu její genomické souřadnice. Tyto záznamy jsou pak uloženy do databáze typu SQL.

[7] https://genome.ucsc.edu/ [online: 16.10.2019]



A) Získání dat z databáze NCBI, B) Zpracování dat C) Mapování domén k DNA sekvenci, D) Rozdělení sekvence dle exonů

Obrázek 3.7: Proces mapování proteinových domén k referenční sekvenci hg19

3.2.10 Databáze variant spojených s dědičnou neuropatií

Tato kapitola shrnuje metody publikované ve článku [Saghira et al. 2018].

Při hledání příčin vzácných onemocnění se nelze spoléhat pouze na své vlastní poznatky. Přestože máme v rámci regionu pravděpodobně nejobsáhlejší databázi pacientů, je výhodné navzájem sdílet data, a tak společným úsilím objasnit další případy. Proto vznikla myšlenka databáze, která bude udržovat genetické varianty spojené s onemocněním CMT, ale zároveň umožní přidávání dat všem uživatelům, ne jen správcům databáze. Díky tomu je možné nejen zjistit, že variantu nalezl již jiný tým ve světě, ale se i případně spojit s týmem a navázat spolupráci. Dalším aspektem je pak vzájemná kontrola, kdy jednotlivé týmy fungují jako reference, to může mít jak pozitivní efekt – např. našli jsme variantu, kterou nemůžeme označit jako kauzální, ale již jí identifikovala jiná skupina, navážeme spolupráci a díky tomu budeme schopní variantu potvrdit jako příčinu, ale i efekt „negativní“, kdy odborná komunita „hlídá“ přeceňování nálezů.

Zpracování projektu Prvním krokem bylo shromáždění všech známých variant spojených s fenotypem dědičných neuropatií. Zdrojem těchto variant byla databáze konsorcia The Inherited Neuropathies Consortium, Athena and Quest diagnostics, HGMD a dalších vyzvaných skupin, zaměřujících se na onemocnění [Cooper, Ball a Krawczak 1998].

Úvodní dataset obsahoval pečlivě vybrané geny (82 genů), které mají známé propojení s CMT onemocněním, z principu projektu ale vyplývá, že uživatelé budou moci přidávat další geny. Všechny geny a varianty prošly validací pomocí HUGO Gene Nomenclature Committee's Multi-Symbol checker tool [Dunnen et al. 2016] a Mutalyzer Syntax Check. U všech uvedených variant je pak poskytnut údaj o původním zdroji varianty.

Celá databáze je postavená na systému Microsoft SQL Server s webovým rozhraním v PHP a AngularJS.

3.2.11 Nástroj pro virtuální panely

Manuální prioritizace variant z WES a WGS sekvenování je velice náročný proces, kdy začínáme se souborem obsahujícím více než 50 000 variant a my hledáme jednu nebo dvě, způsobující onemocnění. Primárním cílem jsou varianty v genech asociovaných s onemocněním, na které se zaměřujeme. Jelikož se množství asociací genů s fenotypem neustále zvyšuje (pro každé onemocnění se jedná o desítky až stovky genů), není možné si všechny pamatovat, proto přistupujeme k vytvoření tzv. virtuálního panelu. Sekvenování pomocí panelu genů se od virtuálního panelu liší v tom, že si oblast zájmu vybíráme před sekvenováním a tím můžeme více využít kapacitu sekvenačního kitu (vyšší pokrytí). Virtuální panel je ale tvořen až po sekvenaci, kdy vyvoláváme varianty pouze v oblastech, které jsou relevantní pro vyšetřované onemocnění. Databáze Human Phenotype Ontology [Köhler et al. 2014] nám umožňuje tyto asociace hledat – po zadání fenotypu, dokážeme získat názvy genů, díky tomu jsme pak schopní dohledat lokaci genu (jeho genomickou pozici). Cílem je tedy vytvoření nástroje, který tyto kroky spojí dohromady a po zadání HPO termínů nám vygeneruje validní BED file – tzn. takový, se kterým budou umět pracovat další nástroje, používané pro bioinformatické účely.

Zpracování projektu je znázorněno na schématu Obr. 3.8.

Před vytvořením samotného nástroje je prvním důležitým krokem shromáždění dat. Nejprve je potřeba získat aktuální databázi všech HPO termínů, ta je k dispozici na webové adrese[8], konkrétně se jednalo o soubor, který asociuje HPO termíny s geny, který má následující strukturu:

- HPO-ID – identifikátor HPO termínu
- HPO-Name – popis fenotypu
- Gene-ID – identifikátor genu
- Gene-Name – název genu (dle RefSeq)

Jako další je potřeba získat údaje o všech RefSeq genech [O'Leary et al. 2015] – genomické pozice celého genu i všech exonů. Tento soubor je dostupný přes USCS Genome browser na adrese: [9] a má následující strukturu:

- ID – identifikátor záznamu
- Gene_ID – RefSeq identifikátor genu
- CHR - chromozom
- txStart – počátek genu na chromozomu
- txEnd – konec genu na chromozomu
- cdsStart – začátek kódující sekvence genu
- cdsEnd - konec kódující sekvence genu

[8] https://hpo.jax.org/app/download/annotation [online: 16.10.2019]
[9] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ [online: 16.10.2019]

- exonCount – počet exonů genu
- exonStarts – začátky exonů, oddělené čárkou
- exonEnds – konce exonu, oddělené čárkou
- name2 – název genu (dle RefSeq)
- cdsStartStat – informace o kompletnosti sekvence
- cdsEndStat - informace o kompletnosti sekvence

Tyto dvě tabulky jsme navzájem propojili dle názvu genu (Gene-Name a name2 sloupce v tabulkách). Tím získáme kompletní záznam – každý HPO termín má nyní doplněné údaje o všech genech, které jsou s termínem asociované. Tím máme připravenou tabulku, kterou využíváme jako zdroj dat pro vybudování virtuálního panelu.

Pokud si chce uživatel vytvořit vlastní virtuální panel, vkládá do nástroje vstup ve formě textového souboru, kde na každém řádku je HPO termín ve správném formátu *HP:1234567*, pro každý takový termín jsou pak z tabulky vyvolány všechny řádky.

V dalším kroku probíhá transformace do požadovaného formátu – BED, který má první tři sloupce povinné a další nepovinné. Pro účely našeho projektu jsme formát zvolili následující:

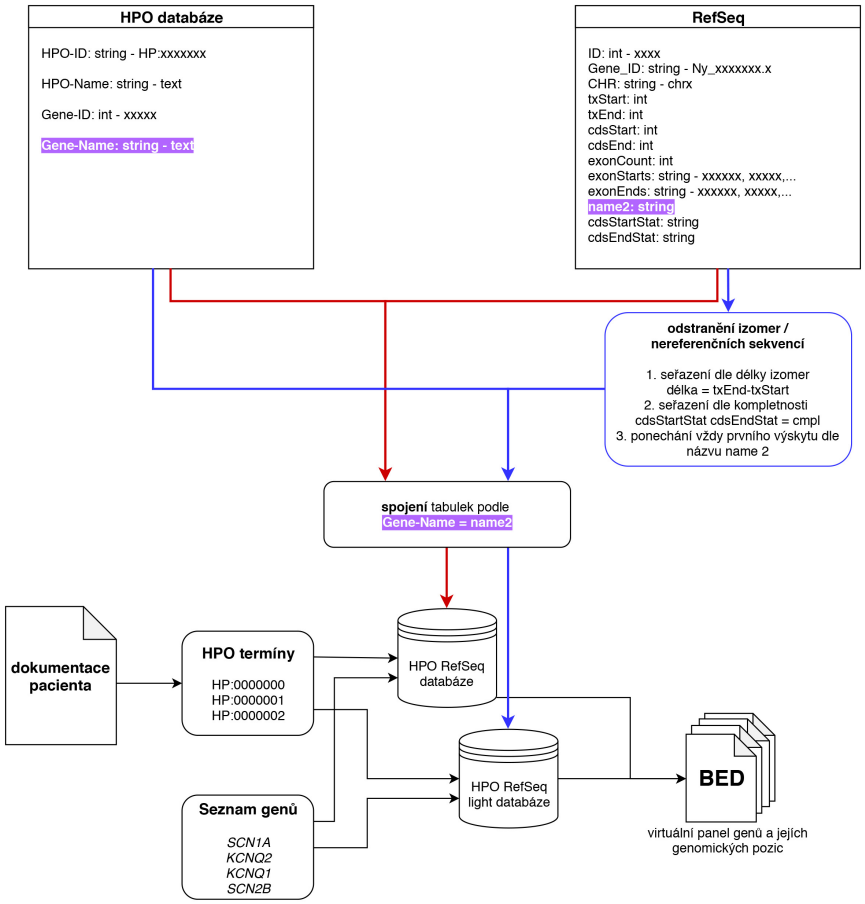
- chromozom
- začátek intervalu
- konec intervalu
- název genu (nepovinné)

Skript tedy nejprve tabulku seřadí dle chromozomu a počátku genu, poté prochází tabulku s řádky odpovídajícími uživatelskému vstupu a ty ukládá do dvou souborů – první obsahuje celé geny – to znamená, že interval obsahuje jak kódující, tak nekódující sekvenci, jedná se o souřadnice txStart, txEnd z tabulky RefSeq. V druhém případě, je generován soubor pokrývající pouze kódující sekvenci, kdy ze sloupců exonStarts a exonEnds vkládá začátky a konce intervalů. Výsledný BED soubor pak obsahuje více řádků, ale vyhneme se pak zahrnutí nekódujících oblastí.

V těchto základních výstupech ale narážíme na velkou redundanci dat, každý gen obsahuje několik různých záznamů, pro své vlastní izomery nebo nedokončené publikované sekvence (s RefSeq identifikátorem NR_xxxxxxx místo NM_xxxxxxx), proto byl nástroj doplněn o tzv. light verze BED souborů, které vycházejí s námi předfiltrované RefSeq tabulky. Filtrování proběhlo na základě toho, že u každého genu nás zajímá jeho nejdelší kompletní sekvence, pokud tedy má gen pouze jeden záznam, tak takový řádek ponecháme, pokud ale obsahuje řádek více různých sekvencí/izomer přistupujeme k filtrování a ponechání pouze jednoho řádku dle filtru, kdy je preferována kompletní sekvence (parametry cdsStartStat a cdsEndStat mají hodnotu „cmpl“) a délka celého genu (txEnd – txStart) je nejvyšší. Následný proces

je totožný s původní iterací nástroje – vytvoří se spojená tabulka s HPO termíny a provede se transformce do formátu BED.

Výsledkem tak jsou 4 BED soubory – s exonovými intervaly, s celými geny a verze „light“ obsahující exony a geny vždy pro jednu izomeru/sekvenci genu.



Pro vytvoření virtuálního panelu čerpá nástroj informace ze dvou databází HPO a RefSeq. V prvním případě (červeně) dochází k přímému propojení tabulek dle názvu genu (fialově), v druhém případě (modře) dochází k odstranění izomer a nereferenčních sekvencí, ponechána je ta nejdelší. Výsledná HPO RefSeq databáze a „light“ verze databáze jsou dotazovány na přítomnost HPO termínů (dle fenotypu pacienta) či alternativně na přítomnost genů dle názvu. Výsledkem jsou pak BED soubory z obou verzí databáze.

Obrázek 3.8: Schéma vytvoření nástroje pro virtuální panely

3.2.12 Správa a ukládání dat

Se zpracováním velkého množství NGS dat se pojí potřeba bezpečného a spolehlivého ukládání a zálohování dat a návrhu správy dat tak, aby nedocházelo ke ztrátám informací, data byla dostupná, zálohovaná alespoň ve dvou kopiích, a abychom se vyvarovali redundance dat. Cílem tedy bylo navrhnout takový systém, který zajistí všechny tyto požadavky a bude udržitelný v delším časovém horizontu. Data jsou zpracována na hlavní pracovní stanici s operačním systémem Linux, kde je pomocí GATK workflow provedeno bioinformatické zpracování od přijatých FASTQ souborů až do VCF souboru, který podléhá další analýze. Výstupy ze všech kroků jsou pak ukládány a zálohovány na diskové pole v rámci DNA laboratoře, kde jsou i dostupné všem uživatelům.

Jeden kompletně zpracovaný panel genů jednoho pacienta zabere na diskovém úložišti v počítači kolem 5 GB, u WES to pak je 25 GB, u WGS pak řádově stovky GB. Důležité je tak nastavit systém, u kterého budou data zabezpečená před ztrátou (zálohování) a dobře dohledatelná pro případnou reanalýzu. Pro zpracování dat v rámci projektu se nám osvědčila stromová struktura pro ukládání dat, schéma v kapitole 4.6.

Cílem projektu byl návrh samotného postupu zpracování dat, v rámci prostředí Linux bylo využito vestavěného skriptovacího jazyka Bash pro manipulaci se soubory.

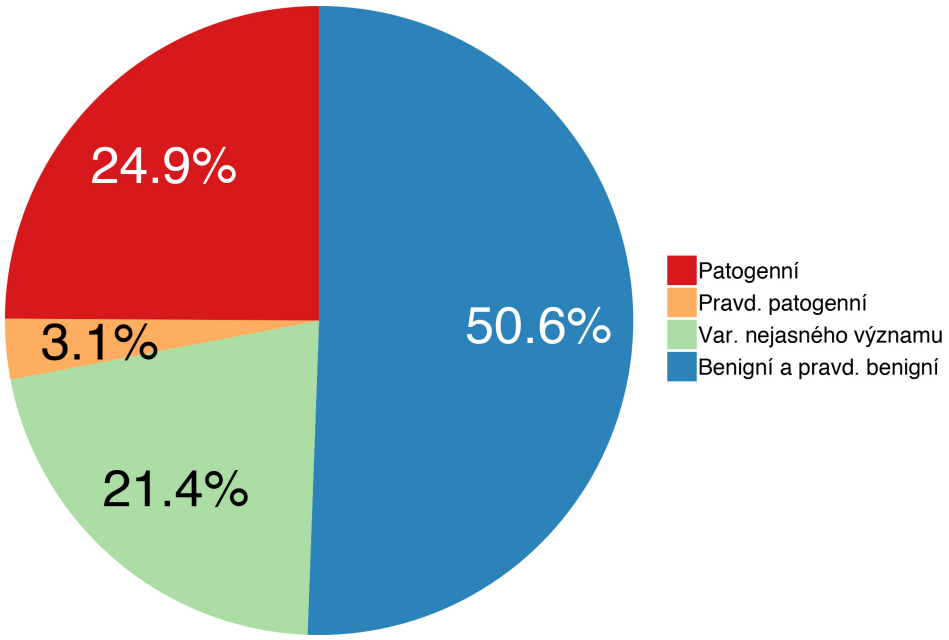
4 Výsledky

4.1 MPS panelu genů u pacientů s EE

K analýze pomocí námi připraveného panelu 112 genů bylo indikováno na Klinice dětské neurologie (KDN) celkem 257 pacientů s klinickým podezřením na epileptickou encefalopatii. Analýza proběhla vždy nejméně dvěma experty se zkušenostmi s vyhodnocováním NGS dat a nejméně dvěma nezávislými nástroji (modrá část v obr Obr. 3.5). Pro potvrzení variant byla prováděna segregační analýza Sangerovým sekvenováním jak u pacienta, tak u rodičů.

Varianty byly prioritizovány dle kritérií ACMG do čtyř tříd (poslední dvě sloučené):

- Patogenní (Pathogenic)
- Pravděpodobně patogenní (Likely pathogenic)
- Varianty nejasného významu (Variants of uncertaing significance, VUS)
- Benigní a pravděpodobně benigní (Benign, Likely bening)

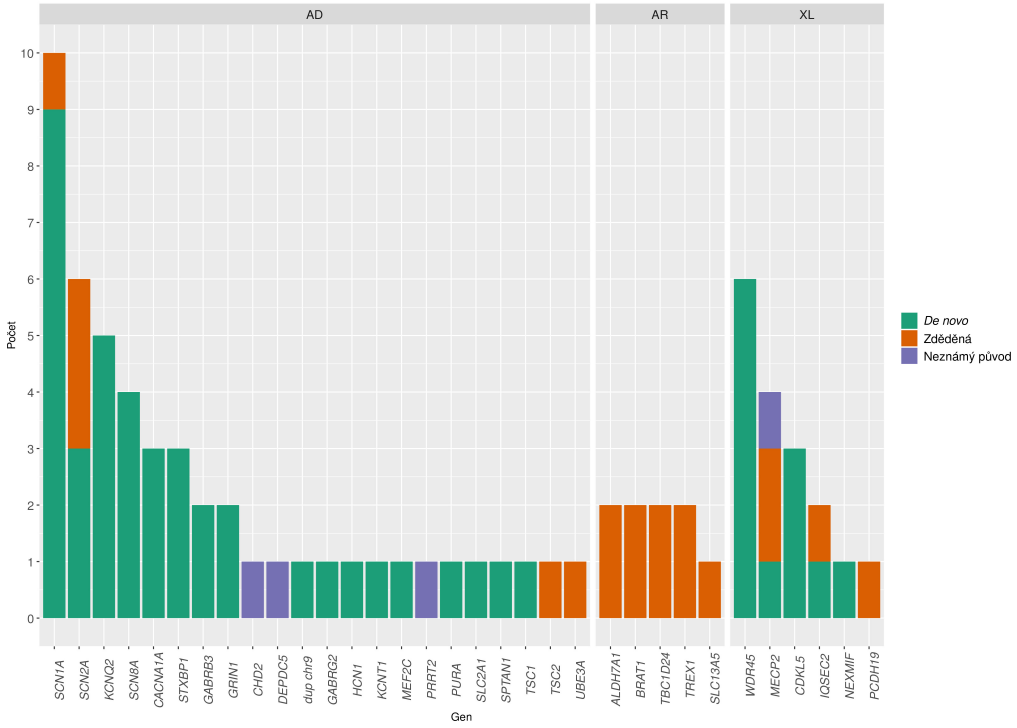


Obrázek 4.1: Výsledky pacientů dle nalezených variant

Pokud se na výsledky budeme dívat z hlediska pacientů, tak u 64 z 257 (24.9 %) byla nalezena přesvědčivě patogenní varianta, kterou pokládáme za příčinu EE. U osmi z 257 byly nalezeny varianty pravděpodobně patogenní (3.1 %), u 55 z 257 (21.4 %) byly nalezeny varianty s nejasným významem. U zbytku pacientů jsme nenašli žádnou variantu, kterou bychom mohli zařadit do vyšších tříd. Zařazení do skupin se řídí vždy nejvýše klasifikovanou variantou, to znamená, že pokud jsme u pacienta našli pravděpodobně patogenní variantu, je možné, že byly detekovány i varianty zařazené do nižších tříd (VUS, pravd. benigní).

4.1.1 Identifikované varianty

Pro analýzu typu mendelovské dědičnosti jsme použili patogenní a pravděpodobně patogenní varianty. Takových variant bylo celkem 75 (68 patogenních a 7 pravděpodobně patogenních). Všechny nalezené patogenní a pravděpodobně patogenní varianty jsou seskupeny dle genu a uvedeny v grafu níže Obr. 4.2, kde jsou rozděleny do skupin dle dědičnosti, seřazeny dle četnosti a barevně označeny dle vzniku.



Geny jsou rozdělené do skupin dle dědičnosti - AD, AR, XL (pro X-vázané), na ose X geny, seřazené dle četností v rámci skupiny, osa Y ukazuje počty variant v genu

Obrázek 4.2: Přehled genu u kterých byly identifikované patogenní a pravděpodobně patogenní varianty

Ve čtyřech případech byly u pacientů nalezeny dvě patogenní varianty – pacient s variantami v genu *ALDH7A1*, pacient s variantami v genu *TBC1D24*, pacient

s variantami v genu *TREX1* a pacient s variantami v genu *BRAT1*, jedná se o varianty v genech s AR dědičností.

Ze 112 genů v panelu byla patogenní nebo pravděpodobně patogenní varianta nalezena v 33 genech (29.5 % genů v panelu). Varianty byly nejčastěji nalezeny v genech s autosomálně dominantní dědičností – 49 variant ve 22 genech. Nejčastěji jsme nalezli patogenní nebo pravděpodobně patogenní variantu v genu *SCN1A* (10 výskytů), poté v genech *SCN2A* (6 výskytů) a *KCNQ2* (5 výskytů). U genů s autosomálně recesivní dědičností jsme nalezli 9 variant v celkem 5 genech a u X-vázaných genů jsme patogenní nebo pravděpodobně patogenní variantu našli v 17 případech a v 9 genech, kdy nejčastěji se jednalo o variantu v genu *WDR45*.

Z celkem 72 pacientů, u kterých jsme našli patogenní nebo pravděpodobně patogenní variantu, bylo možné provést segregační analýzu u 68 (71 variant). U zbytku nebylo možné získat vzorky DNA obou rodičů a proto nebylo možné přesně určit původ zkoumané varianty (4 varianty u 4 pacientů).

Díky segregační analýze bylo možné určit původ 71 variant, kdy 52 (73.2 %) bylo *de novo* vzniklých, 19 (26.8 %) variant bylo zděděných od rodičů.

Všechny varianty jsou pak uvedené v tabulce Tab. 4.1, kde uvádíme souřadnice každé varianty, výsledek predikčních nástrojů, dědičnost genu, ve kterém se nachází, původ varianty a její klasifikaci. V tabulce jsou uvedeny pouze varianty, které byly hodnocené jako patogenní či pravděpodobně patogenní.

Gen	Ref Seq	Genomické souřadnice	Proteinové souřadnice	Predikce (SIFT, PolyPhen2,ClinVar)			AD/AR	DN/INH	Klasifikace
CACNA1A	NM_001127221.1	c.13319826G>T					AD	DN	P
CACNA1A	NM_001127221.1	c.2663A>T	p.Gln888Leu				AD	DN	P
GABRB3	NM_000814.5	c.841A>G	p.Thr281Ala	D	PD		AD	DN	P
GABRG2	NM_000816.3	c.968G>A	p.Arg323Gln	D	PD		AD	DN	P
GRIN1	NM_007327.3	c.2443G>A	p.Gly815Arg	D	PD	P	AD	DN	P
GRIN1	NM_007327.3	c.1643G>A	p.Arg548Gln	T	PD		AD	DN	P
HCN1	NM_021072.3	c.1189A>G	p.Ile397Leu	T	B		AD	DN	P
KCNQ2	NM_172107.2	c.826A>C	p.Thr276Pro	D	B		AD	DN	P
KCNQ2	NM_172107.2	c.1004C>G	p.Pro335Arg	D	PD		AD	DN	P
KCNQ2	NM_172107.2	c.701C>T	p.Thr234Ile	D	PD		AD	DN	P
KCNQ2	NM_172107.2	c.913_915delTTC	p.Phe305del				AD	DN	P
KCNQ2	NM_172107.2	c.913_915delTTC	p.Phe305del				AD	DN	P
MEF2C	NM_002397.4	c.766C>T	p.Arg256*				AD	DN	P
PURA	NM_005859.4	c.812_814del	p.Phe271del				AD	DN	P
SCN1A	NM_001202435.1	c.1244T>A	p.Ile415Lys	D	PD		AD	DN	P
SCN1A	NM_001165963.1	c.5384A>G	p.Glu1795Gly	D	PD		AD	DN	P
SCN1A	NM_001165963.1	c.4384dup	p.Tyr1462Leufs*24				AD	DN	P
SCN1A	NM_001165963.1	c.1178G>A	p.Arg393His	D	PD	P	AD	DN	P
SCN1A	NM_001165963.1	c.1525C>T	p.Gln509*				AD	DN	P
SCN2A	NM_001040142.1	c.2774T>C	p.Met925Thr	D	PD		AD	DN	P
SCN2A	NM_021007.2	c.2291C>T	p.Ala764Val	T	PD		AD	INH	P
SCN2A	NM_001040142.1	c.5009C>T	p.Thr1862Ile	T	PD		AD	DN	P
SCN8A	NM_014191.3	c.4921C>G	p.Leu1641Val	D	PD		AD	DN	P
SCN8A	NM_014191.3	c.2549G>A	p.Arg850Gln	D	PD	LP	AD	DN	P
SCN8A	NM_014191.3	c.4850G>T	p.Arg1617Leu	D	PD		AD	DN	P
STXBP1	NM_003165.3	c.1654T>C	p.Cys552Arg	D	B		AD	DN	P
UBE3A	NM_130838.1	c.1149G>C	p.Glu383Asp				AD	INH	P
dup chr9							AD	DN	P
TSC1	NM_000368.4	c.1525C>T	p.Arg509*			P	AD	DN	P
SCN2A	NM_021007.2	c.4756C>T	p.Arg1586Cys	D	PD		AD	DN	P
GABRB3	NM_000814.5	c.863C>A	p.Thr288Asn	D	PD		AD	DN	P
SCN8A	NM_014191.3	c.4901C>T	p.Ala1634Val	D	PD		AD	DN	P
PRRT2	NM_145239.2	c.883C>T	p.Arg295Trp	D	PD		AD	UNK	P
SLC2A1	NM_006516.2	c.1296C>A	p.Tyr432*			P	AD	DN	P
STXBP1	NM_003165.3	c.1117_1118insGGG	p.Leu372_Ala373insGly				AD	DN	P
STXBP1	NM_003165.3	c.1060T>C	p.Cys354Arg	D	PD		AD	DN	P
KCNT1	NM_020822.2	c.1421G>A	p.Arg474His	D	PD	P	AD	DN	P
SCN1A	NM_001165963.1	c.1702C>T	p.Arg568*	D	PD	P	AD	DN	P
SCN1A	NM_001165963.1	c.602+1G>A				P	AD	DN	P
SCN1A	NM_001165963.2	c.1738C>T	p.Arg580*			P	AD	DN	P
SCN2A	NM_021007.2	c.668G>A	p.Arg223Gln	D	PD	P	AD	INH	P
SCN1A	NM_001165963.2	c.1336C>T	p.Gln446*				AD	DN	P
SPTAN1	NM_001130438.2	c.6886G>A	p.Ala2296Thr	D	PD		AD	DN	P
SCN2A	NM_021007.2	c.5624T>C	p.Leu1875Pro	D	PD		AD	INH	P
SCN1A	NM_001165963.2	c.3361G>A	p.Glu1121Lys	D	PD		AD	INH	P
ALDH7A1	NM_001182.4	c.1318-1G>C					AR	INH	P
ALDH7A1	NM_001182.4	c.518-14_518delinsCA					AR	INH	P
SLC13A5	NM_177550.3	c.425C>T	p.Thr142Met	D	PD	P	AR	INH	P
TREX1	NM_016381.3	c.10621072del	p.Leu354Phefs*22				AR	INH	P
TREX1	NM_016381.3	c.1072A>C	p.Thr358Pro	T		P	AR	INH	P
TBC1D24	NM_001199107.1	c.1505dupG	p.Ser503Glnfs*55				AR	INH	P
BRAT1	NM_152743.3	c.638dupA	p.Val214Glyfs*189			P	AR	INH	P
TBC1D24	NM_001199107.1	c.731C>T	p.Ala244Val	D	PD		AR	INH	P
BRAT1	NM_152743.3	c.224_226del	p.Phe75del				AR	INH	P
CDKL5	NM_003159.2	c.2578C>T	p.Gln860*				XL	DN	P
CDKL5	NM_003159.2	c.463+5G>A					XL	DN	P
CDKL5	NM_003159.2	c.1247_1248del	p.Glu416Valfs*2			P	XL	DN	P
IQSEC2	NM 001111125.2	c.3206G>C	p.Arg1069Pro	D	PD		XL	INH	P
MECP2	NM_004992.3	c.1219_1229del	p.Asp407Glnfs*25				XL	UNK	P
WDR45	NM_007075.3	c.654del	p.Arg219Alafs*69				XL	DN	P
WDR45	NM_007075.3	c.970_971del	p.Val324Hisfs*17				XL	DN	P
WDR45	NM_007075.3	c.511C>T	p.Gln171*				XL	DN	P
WDR45	NM_007075.3	c.344+4A>C					XL	DN	P
WDR45	NM_007075.3	c.343_344+2delAAGT	p.Lys115Aspfs*23			P	XL	DN	P
NEXMIF	NM_001008537.2	c.553C>T	p.Gln185*				XL	DN	P
WDR45	NM_007075.3	c.831-1G>A					XL	DN	P
MECP2	NM_004992.3	c.1357C>T	p.Arg453*			P	XL	DN	P
IQSEC2	NM_001111125.2	c.810_817dup	p.Gln273Argfs*2				XL	DN	P
DEPDC5	NM_001242896.1	c.752A>G	p.Tyr251Cys	D	B		AD	UNK	LP
CACNA1A	NM_001127221.1	c.184T>C	p.Tyr62His	T	PD		AD	DN	LP
CHD2	NM_001271.3	c.2672C>T	p.Pro891Leu	D	PD		AD	MF	LP
TSC2	NM_000548.3	c.2355 G>C	p.Gln785His	T	PD		AD	INH	LP
MECP2	NM_004992.3	c.925C>T	p.Arg309Trp	D	PD	VUS	XL	INH	LP
PCDH19	NM_001184880.1	c.698A>G	p.Asp233Gly	D	PD		XL	INH	LP
MECP2	NM_004992.3	c.397C>T	p.Arg133Cys	D	PD	P	XL	INH	LP

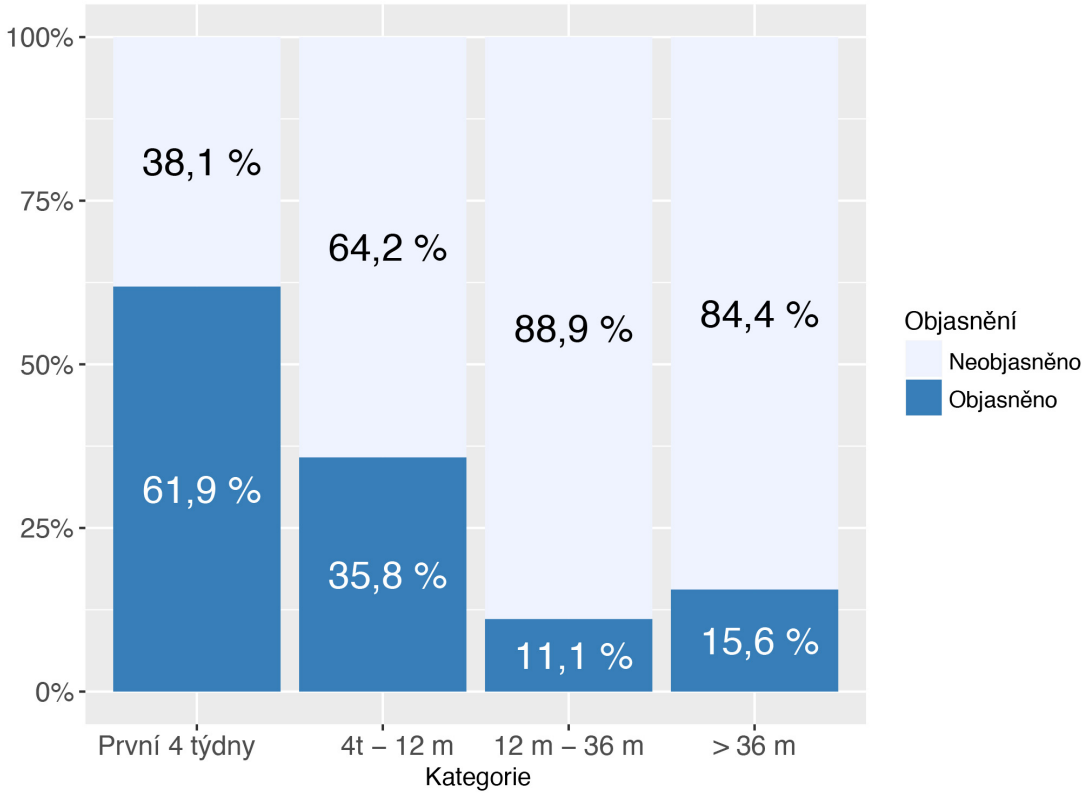
Varianty jsou seřazené dle klasifikace, dědičnosti a dle genů, predikce: SIFT: D - poškozující (damaging), T -tolerované ; PolyPhen2: PD - pravd. poškozující ,B - benigní ; ClinVar P - patogenní, LP - pravd. patogenní),dědičnost (AD, AR, a XL pro X vázanou, DN udává *de novo* vznik varianty, INH variantu zděděnou, UNK neznámý původ, klasifikace jsou P - patogenní a LP - pravd. patogenní

Tabulka 4.1: Tabulka všech patogenních a pravd. patogenních variant nalezených u pacientů s EE

4.1.2 Pravděpodobnost objasnění EE v závislosti na věku pacienta při prvním záchvatu

V publikaci [Staněk et al. 2018] jsme prokázali souvislost mezi věkem pacienta při prvním záchvatu a nalezení patogenní varianty (objasnitelnost). Analýza byla provedena u 151 pacientů, u kterých byl zaznamenán věk při prvním epileptickém záchvatu. Pacienti byli rozdělení do 4 skupin podle věku při objevení se prvních epileptických záchvatů – do 4 týdnů věku, mezi 4 týdny a 12 měsíci, po 12. měsíci a před 36. měsícem a starší než 36 měsíců při prvním záchvatu Obr. 4.3.

U první skupiny, s prvním záchvatem během prvních čtyř týdnů života, byla objasnitelnost téměř dvojnásobná než u ostatních skupin (62 % u první skupiny proti maximálně 35,8 %) a postupně s věkem nástupu epileptických záchvatů klesala. Tato hypotéza byla testována pomocí statistického Chí kvadrát testu, a výsledky byly signifikantní s $p < 0.05$.



Obrázek převzatý z prvoautorské publikace: [Staněk et al. 2018]

Obrázek 4.3: Přehled objasnitelnosti příčiny EE dle věku při prvním epileptickém záchvatu pacientů skupin

4.1.3 Vybrané publikované kazuistiky objasněné pomocí MPS panelem genů

4.1.3.1 Pacienti s novými variantami v genu *CDKL5* asociovanými s EE

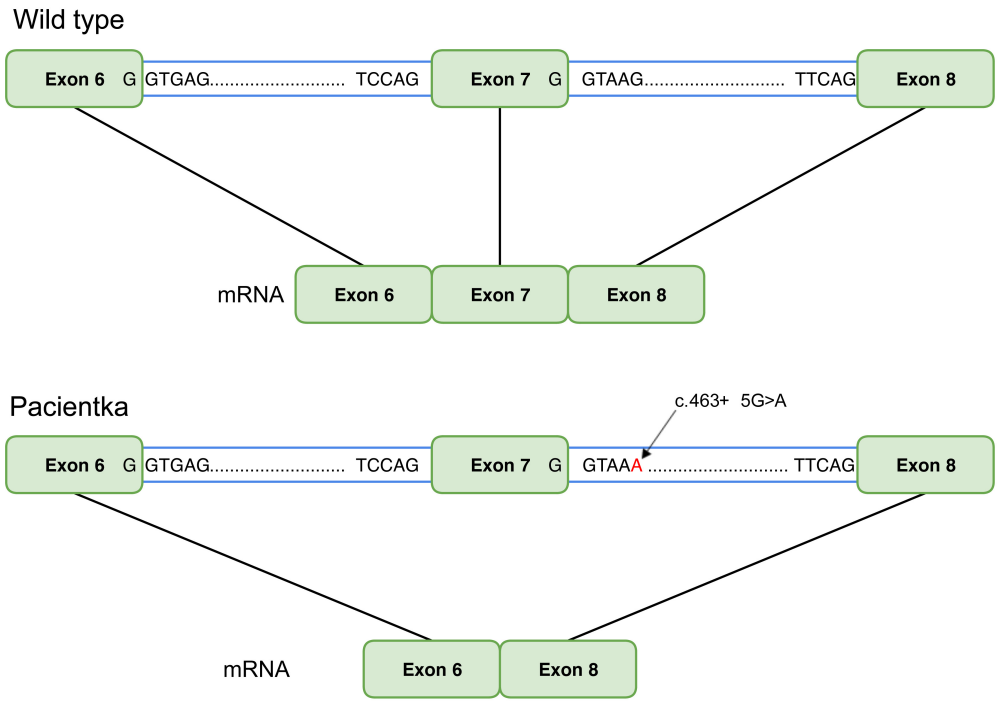
Tato kapitola shrnuje výsledky publikované ve článku [Neupauerová et al. 2017].

Prvním pacientem byl 3 letý chlapec s nástupem epileptických záchvatů ve čtyřech týdnech života (záškuby horních končetin a stáčení hlavy, EEG v normě) s charakterem typu infantilních spasmů. S postupným rozvojem třesu dolních končetin, stáčení očí a zíravým pohledem, poté vývoj myoklonických a tonických záchvatů. Od prvních měsíců života byl pozorován regres psychomotorického vývoje. V současnosti (věk tři a půl let) není pacient schopný kontrolovat pohyb hlavy, není schopen sedu.

Vyhodnocením dat z MPS panelem 112 genů byla u pacienta detekována varianta v genu *CDKL5* (NM_003159.2) na pozici c.2578C>T (p.Gln860*) v hemizygotním stavu, která dříve nebyla popsána a nemá frekvenci v populačních databázích. Jde o variantu predikující předčasné zařazení stop kodonu, má tedy závažný efekt, a neprokázali jsme ji o obou rodičů, nejspíše je tedy *de novo* vzniklá.

Druhou pacientkou byla dívka, které záchvaty začaly v šesti týdnech věku a byly bilaterální tonické. Pacientka má od počátku opožděný psychomotorický vývoj a výraznou hypotonii. Psychologickým vyšetřením byla diagnostikována středně těžká mentální retardace a poruchy autistického spektra. Záchvaty se postupně stávaly komplexnější asynchronější s přerušovanou reaktivitou. Stav byl farmakorezistivní s dobrou odezvou na ketogenní dietu.

U této pacientky byla detekována varianta v genu *CDKL5* (NM_003159.2) na pozici c.463+5G>A (g.18600075G>A) v nekódující části genu (v intronu 7) v heterozygotním stavu, která dosud nebyla popsána jako příčina epilepsie a nemá frekvenci v populačních databázích. Variantu jsme neprokázali u zdravých rodičů pacientky, je tedy nejspíše *de novo* vzniklá. Nalezená nepopsaná intronová *de novo* varianta, byla podezřelá z alterace normálního sestřihu mRNA. Proto bylo provedeno vyšetření na RNA úrovni (sekvenování cDNA, které prokázalo, aberantní sestřih mRNA s vynecháním exonu 7 (bez posunu čtecího rámce pro mRNA tzv. in frame), schématicky znázorněno na obrázku Obr. 4.4.



V prvním případě wild type, v druhém případě alteruje varianta na pozici c.463+5 normální sestřih, dochází k intronizaci exonu 7
obrázek převzatý ze spoluautorské publikace: [Neupauerová et al. 2017]

Obrázek 4.4: Schématické znázornění mechanismu aberantního sestřihu mRNA s vynecháním exonu 7

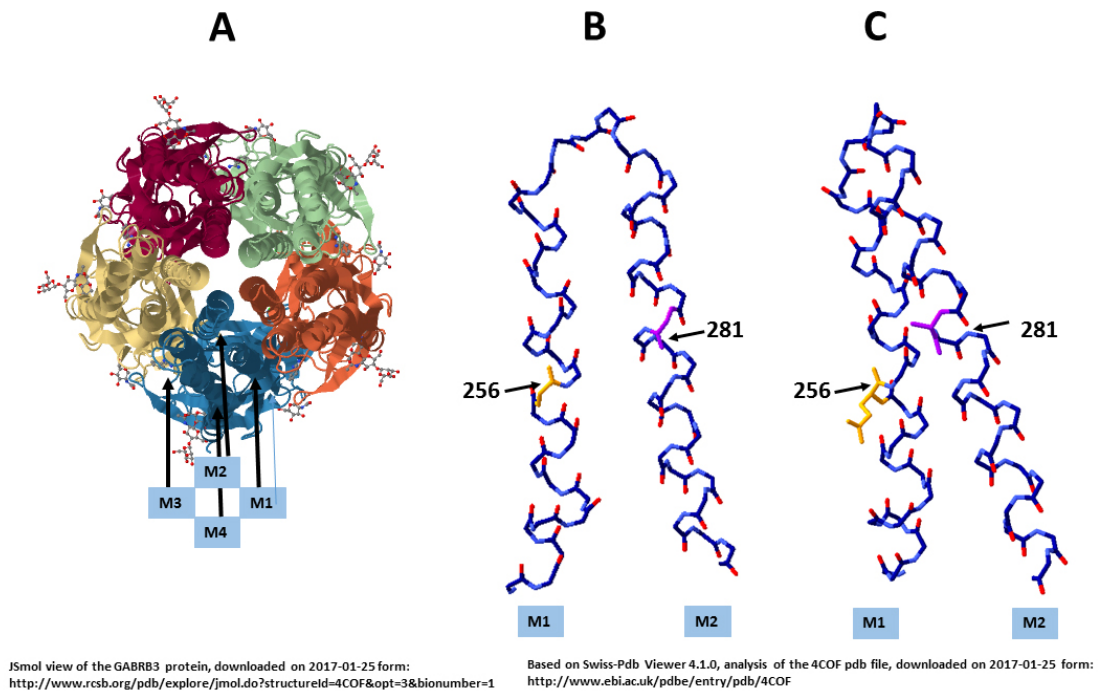
4.1.3.2 Jednovaječná dvojčata s novými variantami v genu *GABRB3* asociovanými s EIMFS

Tato kapitola shrnuje výsledky publikované ve článku [Štěrbová et al. 2018].

Jednovaječná dvojčata byla přijata na novorozenecké oddělení s poruchou poporodní adaptace, od porodu byla přítomna hypotonie, byl pozorován třes končetin a trupu. Epileptické záchvaty byly pozorovány již v prvních dvou týdnech života (s korelátem na EEG) a byly farmakorezistivní. S dalším rozvojem byla u obou dvojčat patrná mikrocefálie, faciální dysmorfismus, hypotelorismus, hypomimie, silné slinění.

U obou pacientek bylo provedeno vyšetření MPS panelu 112 genů asociovaných s EE a u obou pacientek-sester byla prokázána varianta v genu *GABRB3* (NM_000814.5) na pozici c.841A>G (p.Thr281Ala) v heterozygotním stavu, která dosud nebyla popsána, nemá frekvenci v populačních databázích a je hodnocena predikčními nástroji jako patogenní. Variantu jsme neprokázali u zdravých rodičů, je tedy nejspíše *de novo* vzniklá, což je typické pro příčiny těžkých dětský epilepsií a jiné pacienty s variantami v genu *GABRB3* [Møller et al. 2017]. Fenotyp pacientek je velmi podobný, jako u jiných dosud popsaných pacientů, u kterých byla detekována varianta na pozici c.767T>A (p. Leu256Gln) [Myers et al. 2016], jen začátek epileptických

záchvatů je časnější než u dosud popsaných pacientů. Obě tyto varianty se nachází v alfa-helixu a nejsou příliš vzdálené (Obr. 4.5).



(A) *GABRB3* je iontový kanál, homopentamer, každá transmembránová jednotka se skládá ze čtyř alfa-helixů (M1 až M4)
(B) Model alfa-helixů M1 a M2 wild-type. Pozice p.Leu256 je zvýrazněna (žlutě) a pozice p.Thr281 (fialově)
(C) Model alfa-helixů M1 a M2 s alternativními AMK na pozici p.Leu256Gln (žlutě) a p.Thr281Ala (fialově), obě záměny AMK způsobují formování nových silných vodíkových vazeb
Obrázek převzatý ze spoluautorské publikace: [Štěrbová et al. 2018]

Obrázek 4.5: Model *GABRB3* proteinu

4.2 Výsledky z celoexomového sekvenování

4.2.1 Porovnání bioinformatických postupů

V současné době jsou v DNA laboratoři pro analýzu dat z MPS využívány tři různé bioinformatické analýzy. Dva komerční nástroje – NextGENe (NG, SoftGenetics, USA) a Surecall (SC, Agilent, USA) a volně dostupná GATK workflow (GATK). Pro vzájemné porovnání jsme vybrali již dříve analyzovanou skupinu 24 pacientů s EE, u kterých byla po neobjasňujícím výsledku z MPS panelu genů provedena analýza WES dle sekvenačního kitu SureSelect All exons V6 (Agilent, USA). Data byla zanalyzována pomocí všech tří nástrojů, poté byla provedena anotace programem ANNOVAR a následovalo manuální filtrování variant. Analýza byla navržená tak, abychom dokázali porovnat všechny metody na srovnatelných datech, u kterých je již známý výsledek, proto jsme vybrali tuto skupinu, která obsahovala vzorky jak s nalezenou kauzální variantou, tak bez nalezených kauzálních variant.

Pro srovnání uvádíme nejprve počty variant v každém kroku manuálního filtrování a dále distribuci jednotlivých variant u predikčních nástrojů využívaných při filtrování. V nástrojích NG a SC bylo použito základní nastavení bez úprav parametrů, pro GATK workflow jsme použili nástroje BWA-MEM, Picard a samtools pro alignment a pro variant calling Haplotype Caller, který je součástí balíku GATK, data nepodléhala předfiltrování. Výstupy z NG a SC byly pak spojené do jednoho VCF souboru pro všechny vzorky (24) pomocí nástroje VCFBreakCreateMulti, který je součástí balíku VCFTools [Danecek et al. 2011], u GATK byly vyvolány společně z GVCF souborů (algoritmus JointGenotyping).

Postup při manuálním filtrování variant byl následující:

1. Byly odfiltrovány varianty, které měly alespoň jednu z frekvencí gnomAD exome ALL, gnomAD exome NFE, gnomAD genome ALL, gnomAD genome NFE větší než 0.01

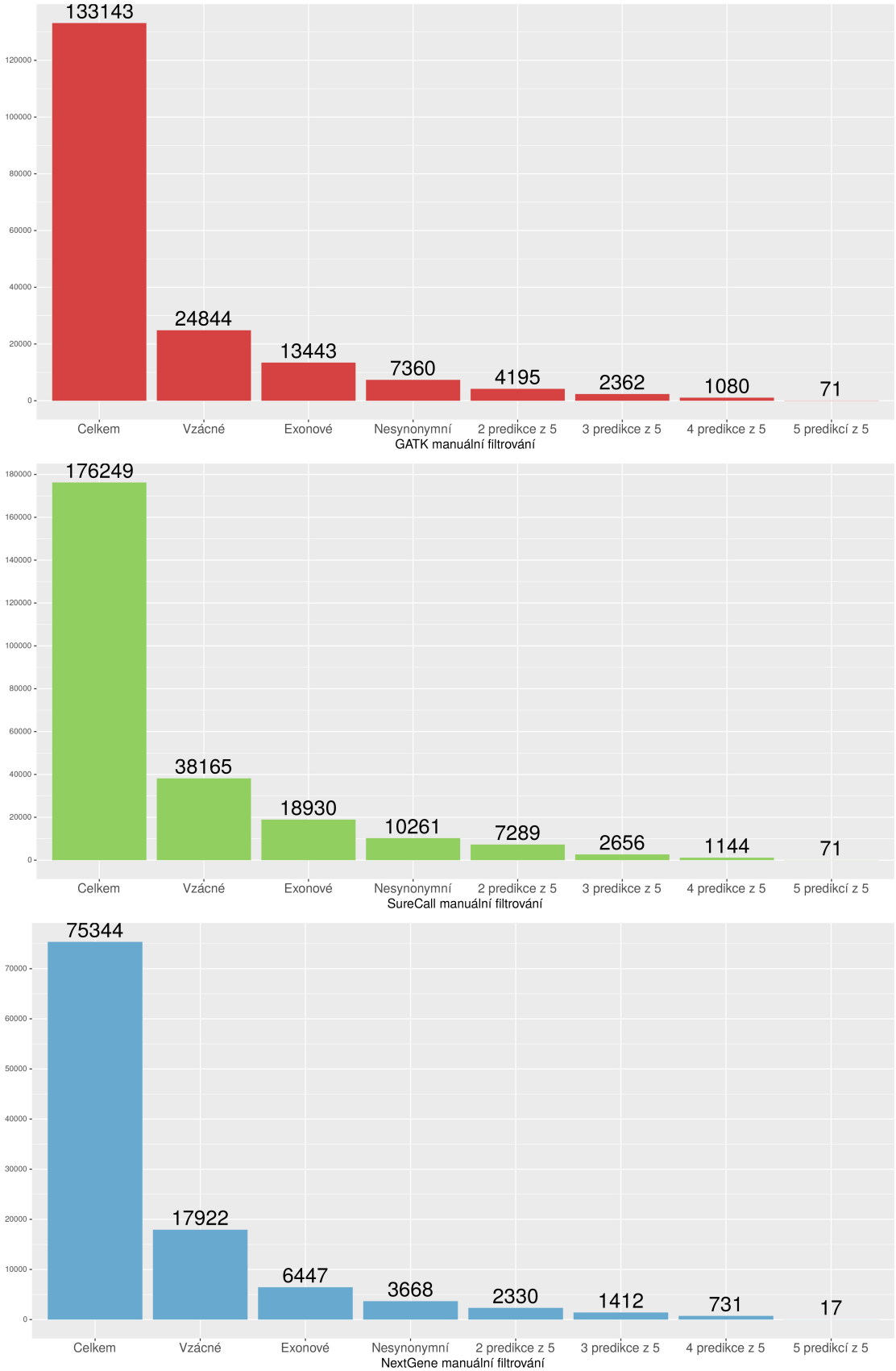
2. Byly odfiltrovány varianty, které se nachází mimo exonové oblasti

3. Byly odfiltrovány varianty, které byly označeny jako synonymní

4. Byly hodnoceny varianty dle predikčních nástrojů PolyPhen2-HVAR – hodnota: P (pathogenic), SIFT – hodnota D (damaging), MCAP – číselná hodnota větší než 0.025, ClinVar – Pathogenic nebo VUS, Intervar – predikce nebyla jako benigní nebo bez predikce

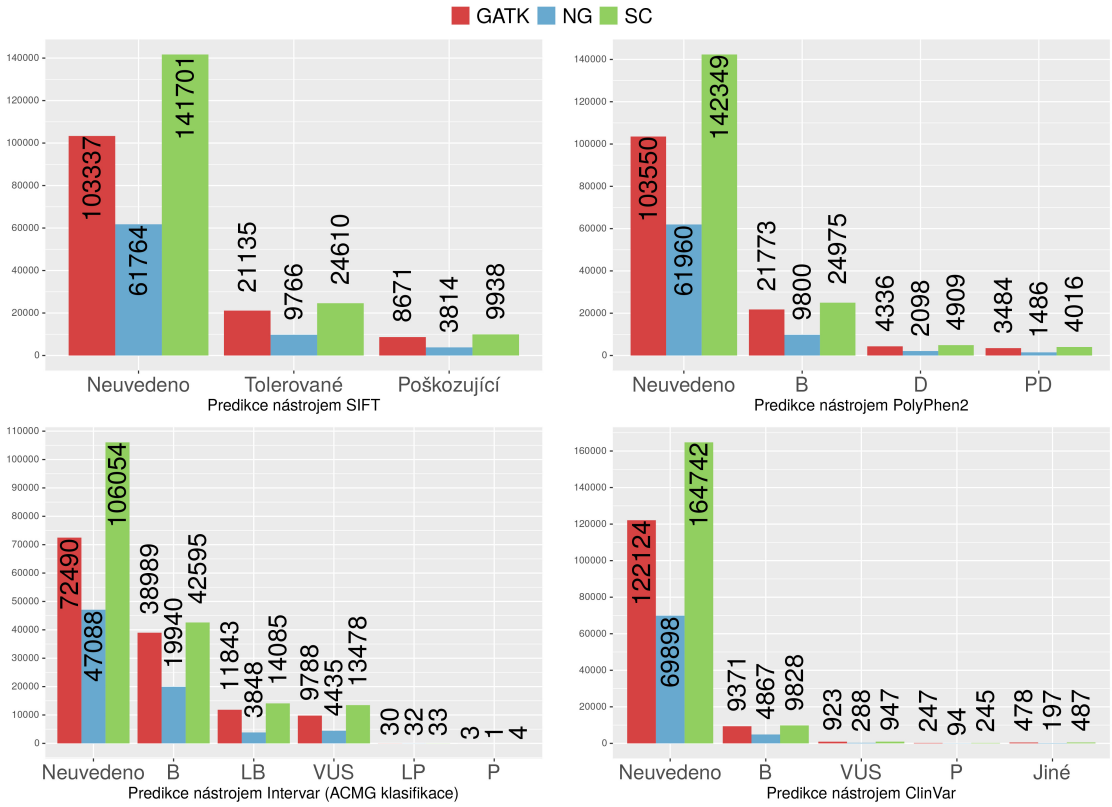
5. Byly spočítány počty variant splňující podmínky z bodu 4. a vždy alespoň 2 z 5, poté 3 z 5 atd.

Tento postup je mírně upravený oproti postupu, který aplikujeme při hledání patogenních variant. Zde byly použity tvrdé filtry na přítomnost varianty v exonové oblasti a na nesynonymní varianty pro porovnání počtu variant. A to s cílem porovnat efektivitu jednotlivých workflow. Průběh filtrování je znázorněn na obrázku Obr. 4.6, dále uvádíme přehled distribucí dat u vybraných predikčních nástrojů Obr. 4.7.



Vybrané predikční nástroje: PolyPhen2 (hodnota P, pathogenic), SIFT (hodnota D, damaging), M-CAP (větší než 0,025), ClinVar (Pathogenic, VUS), Intervar (s predikcí, ne benigní) - filtrování dle počtu splněných podmínek

Obrázek 4.6: Počty variant v jednotlivých krocích filtrování u všech tří metodik



PolyPhen 2 hodnocení: B benigní, D poškozující (damaging), PD pravděpodobně poškozující (probably damaging); Intervar a ClinVar hodnocení: B benigní, LB pravděpodobně benigní (likely benign), VUS nejasný význam (variant of uncertain significance), LP pravděpodobně patogenní (likely pathogenic), P patogenní (pathogenic)

Obrázek 4.7: Distribuce dat, dle počtu variant u každé metodiky, uvedeno u nástrojů SIFT, PolyPhen2, Intervar a ClinVar

Z výsledků je patrné, že GATK a SureCall vyvolávají přibližně stejný počet variant, to je způsobeno využitím stejných algoritmů pro alignment. Celý proces manuálního filtrování se liší v počtu variant jen velmi málo a končí s počtem 71 variant, které jsou predikovány jako závažné všemi sledovanými nástroji. I ze zpětné analýzy potvrzených variant je patrné, že obě tyto metody poskytují srovnatelné výsledky a díky tomu lze předpokládat, že námi implementovaná workflow poskytuje srovnatelné výsledky s komerčním nástrojem.

Oproti tomu nástroj NextGENe, vyvolal téměř o polovinu variant méně (75 tisíc proti 133 tisíc u GATK a 176 tisíc u SC), distribuce zkoumaných parametrů jsou poměrově shodné s porovnávanými metodami, pouze jsou nižší hodnoty. Menší počet vyvolaných variant je způsoben striktnějším základním nastavením, které je ale doporučené od výrobce programu. Jedná se tedy o vhodný doplňující nástroj k jednomu z dvojice GATK a SC s tím, že nám může díky jiným použitým algoritmům

pomoci odhalit jiné varianty. Výhodou je pak u NG snadná analýza CNV pomocí vestavěných nástrojů.

Z 24 pacientů se nám podařilo objasnit příčinu EE u 8 z nich, proto dalším krokem srovnání bylo určit, které postupy dokáží tyto námi dříve potvrzené varianty nalézt. Výsledkem je tabulka Tab. 4.2.

Varianta				GATK	SC	NG
DHDDS	(NM_024887.3)	c.110G>A	(p.Arg37His)	✓	✓	✓
GABRB2	(NM_021911.2)	c.895A>C	(p.Ile299Leu)	✓	✓	✓
HUWE1	(NM_031407.6)	c.12195 G>C	(p.Trp4065Cys)	✓	✓	✘
NARS2	(NM_024678.5)	c.83_84del	(p.Leu28Glnfs*17)	✓	✓	✘
NARS2	(NM_024678.5)	c.1339A>G	(p.Met447Val)	✓	✓	✓
PPP2R5D	(NM_0006245.3)	c.1267_1270del	(p.Leu423fs)	✓	✓	✘
SERPINI1	(NM_005025.4)	c.1174G>A	(p.Gly392Arg)	✓	✓	✓
SLC1A4	(NM_003038.4)	c.1370G>A	(p.Arg457Gln)	✓	✓	✓
UBTF	(NM_014233.3)	c.628G>A	(p.Glu210Lys)	✓	✓	✘

Tabulka 4.2: Přehled 9 kauzálních variant z 24 pacientů a jejich a jejich záchyt pomocí bioinformatických nástrojů

4.2.2 *De novo* model pro hledání kauzálních variant

Při hledání příčíny onemocnění pomocí WES využíváme dva hlavní přístupy, *de novo* model a singleton model. U de novo modelu je důležité mít k dispozici data z WES jak u pacienta, tak od jeho rodičů – tzv. Trio analýza. Díky tomu je pak možné vytvořit model, určující pravděpodobný vznik varianty. Nelze tedy vzít data pacienta a jeho rodičů a vyfiltrovat varianty, které jsou pouze u pacienta a ty pokládat za *de novo* – je nutné modelovat případy, kdy varianta např. nemusela být v některém ze vzorků vyvolána. Dále uvádíme několik publikovaných případů, u kterých bylo možné pomocí *de novo* modelu (nástrojem DeNovoGear) určit příčinu onemocnění.

4.2.2.1 Pacient s variantou v genu *UBTF*

Tato kapitola shrnuje výsledky publikované ve článku [Sedláčková et al. 2018].

Případ popisuje 13 letého pacienta (chlapce), který se narodil zdravým rodičům. Jeho motorický vývoj byl normální (chůze bez pomoci v 13 měsících věku), vývoj řeči byl mírně opožděný (první slova ve věku 18 měsíců). Před nástupem onemocnění byly u pacienta horečky následované apatickými stavy 3 týdny po očkování (v 15 měsících věku). Neurologické problémy začaly ve dvou letech, kdy rodina uvádí nestabilní chůzi a časté pády pacienta. Krátce po nástupu motorických potíží následovala regrese mentálního vývoje, pacient ztratil verbální schopnosti, izoloval se od jiných dětí, byl apatický.

Během dalších 6 měsíců došlo k progresi paleocerebelárního syndromu projevující se nestabilitou chůze, hypertonií, hyperreflexií a extrapyramidovým syndromem. Magnetická rezonance mozku ve věku tří let odalila nespecifické změny v bílé hmotě s rozvojem periventrikulární a korové atrofie.

Ve věku 6 let se u pacienta objevily a rozvíjí epileptické záchvaty (tonické, tonicko-klonické a záchvaty bez motorických projevů), od počátku projevu byly záchvaty farmakorezistentní. Vyšetření EEG bylo provedeno ve věku 3 let s normálním výsledkem, od 6 let abnormální záznam.

V současné době má pacient dva až tři záchvaty za hodinu. Od 11 let je vyživován gastrostomií, je imobilní, bez schopnosti mluvy, s pozitivní emoční reakcí na členy rodiny.

Po genetické konzultaci pacienta bylo indikováno sekvenování pomocí MPS panelu 112 genů asociovaných s EE bez nalezení kauzální varianty Proto bylo přistoupeno k dalšímu kroku, celoexomovému sekvenování vzorku pacienta a jeho obou rodičů (trio analýza). Analýza nástrojem DenovoGear odhalila heterozygotní variantu v genu *UBTF* (NM_014233.3) na pozici c.628G>A (p.Glu210Lys), která byla přítomna u pacienta a u rodičů nebyla detekována. Tento výsledek byl potvrzen Sangerovým sekvenováním. V době první analýzy WES dat nebyla ještě popsána asociace genu *UBTF* s onemocněním, proto nebylo možné případ uzavřít (varianta neměla žádné frekvence v populačních databázích ExAC, gnomAD).

Po reanalýze, která je u neuzavřených WES případů prováděna minimálně jednou za 6 měsíců byly nalezeny aktuální publikace [Toro et al. 2018; Edvardson et al. 2017] popisující stejnou variantu v genu *UBTF*, u sedmi pacientů s velmi podobným fenotypem, čímž došlo k potvrzení patogenicity varianty a uzavření případu.

Dále byla provedena sekvenace exonu 7 genu *UBTF* Sangerovým sekvenováním u dalších 112 pacientů s EE, bez dalších nálezů.

4.2.2.2 Pacienti s patogenní variantou v genu *SETBP1* se syndromem Schinzel-Giedion

Tato kapitola shrnuje výsledky publikované ve článku [Neupauerová et al. 2018].

Publikace popisuje dva případy pacientů s Schinzel-Giedion syndromem (SGS), autozomálně dominantním onemocněním, závažným zpožděním vývoje a neobvyklou faciální dysmorfií.

Prvnímu pacientovi je v současnosti 20 let, od narození byl pozorován opožděný psychomotorický vývoj. Ve třech letech věku byl schopný samostatné chůze, ve dvou letech používal slabiky, ale verbální schopnosti byly ztraceny. U pacienta byl diagnostikován Lennox-Gastautův syndrom.

Epileptické záchvaty začaly od 7 let, byly myoklonické a myoklonicko-astatické, později doplněné o tonicko-klonické a tonické (diagnóza Lennox-Gastautova syndromu). Vyšetření magnetickou rezonancí prokázalo difuzní mozkovou atrofii. V současnosti pacient nemluví (vydává jen neartikulované zvuky), bez schopnosti chůze, apatický. Pacient má klenuté čelo, výrazné nadočnicové oblouky, hypetrofický jazyk, který neustále vyplazuje, gotické patro, hypomimii.

Po genetické konzultaci pacienta bylo indikováno genetické testování pomocí panelu MPS panelu 97 genů asociovaných s EE bez nálezu kauzální varianty. Proto bylo přistoupeno k dalšímu kroku, celoexomovému testování pacienta a jeho obou rodičů (trio analýza). Analýza nástrojem DenovoGear odhalila heterozygotní variantu v genu *SETBP1* (NM_015559.2) na pozici c.2601C>G (p.Ser867Arg), která byla již publikována v článku [Carvalho et al. 2015]. Jelikož byl fenotyp pacienta

shodný s pacienty v publikaci, byl diagnostikován SGS.

Druhá pacientka je dívka ve věku jednoho roku, s faciální dysmorfií, mikrocefálií, nadměrnou vzdáleností očí, hirsutismem a exoftalmem. Po srovnání fenotypu s předchozím pacientem byl klinickým genetikem diagnostikován SGS, který byl následně potvrzen cíleným Sangerovým sekvenováním čtvrtého exonu genu *SETBP1*, výsledkem byla nalezená varianta c.2608G>A (p.Gly870Ser), kterou již byla dříve publikovaná [Hoischen et al. 2010] jako patogenní varianta způsobující SGS. Otestováním rodičů byl potvrzen *de novo* vznik varianty.

4.2.3 Singleton model

4.2.3.1 Pacienti s dědičnou periferní neuropatií způsobenou variantami v genu *SBF2*

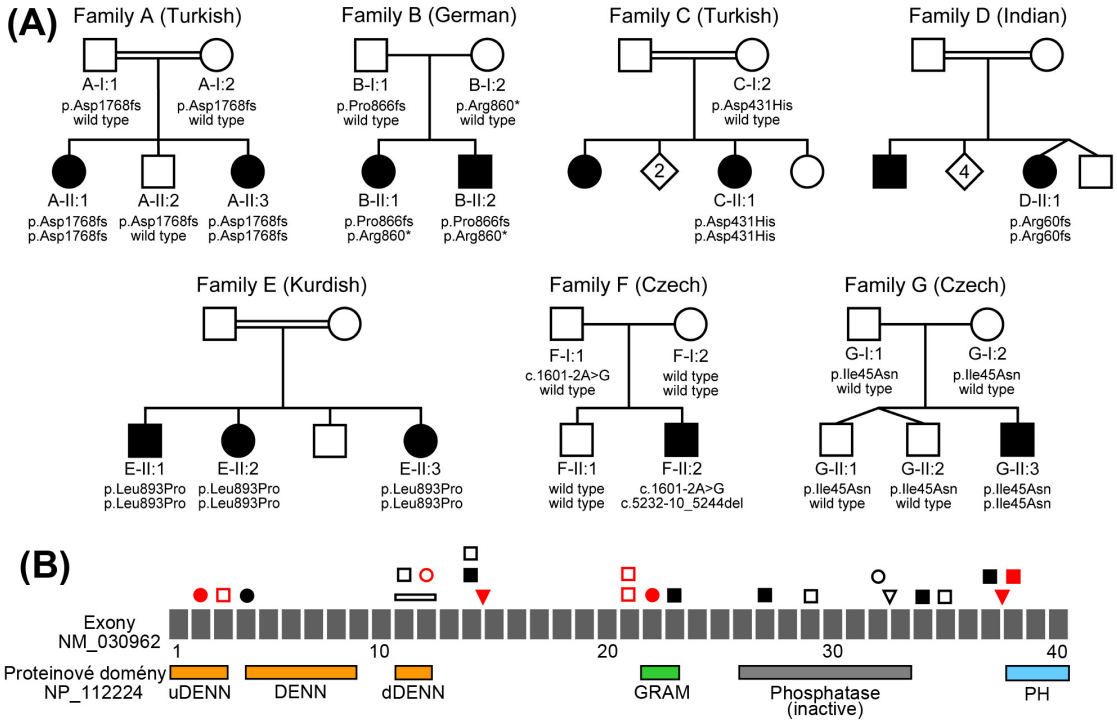
Tato kapitola shrnuje výsledky publikované ve článku [Laššuthová et al. 2018].

Publikace popisuje výsledky sedmi rodin z pěti států, s AR dědičnou periferní neuropatií typu CMT4B2. Celkem bylo detekováno 9 dříve nepopsaných patogenních variant v genu *SBF2* (tři missense varianty, tři frameshift, jedna nonsense a dvě splice-site varianty), všechny varianty nebyly dříve publikovány (s výjimkou varianty c.134T>A publikované [Laššuthová et al. 2016a]).

Ve všech uvedených případech byl potvrzen AR typ přenosu heterozygotních variant od obour rodičů s výjimkou varianty u probanda F-II:2, kdy byla detekována varianta c.1601-2A>G, zděděná od otce, a varianta c.5232-10_5244del na druhé alele genu *SBF2*, vzniklá *de novo*.

V níže uvedeném schématu uvádíme rodokmeny rodin s uvedenými variantami a lokalizaci variant v rámci exonů Obr. 4.8.

Co se týče klinických příznaků tak nejčastěji se CMT způsobené variantami v genu *SBF2* projevují nástupem na pomezí první a druhé dekády života, slabostí dolních končetin, snížením reflexů, rozpadem myelinu detekovatelným biopsií nervů, deformitami nohou. Ve více než 50 % případů došlo k rozvoji glaukomu, a v méně než 50 % případů k deformitám rukou, ztrátě sluchu či skolióze.



(A) Rodokmeny všech rodin s uvedenou variantou testovaných členů (B) Lokalizace variant v rámci exonů. Legenda: čtverec - nonsense a frameshift varianty, trojúhelník splice-site varianty, kroužek missense varianty, obdélník in-frame delece; červeně varianty uvedené v rámci studie, s výplní varianty způsobující neuropatii s glaukomem převzato ze spoluautorské publikace: [Laššuthová et al. 2018]

Obrázek 4.8: Přehled rodin s pacienty s variantou v genu *SBF2*

4.2.4 Další objasněné případy WES

Epileptické encefalopatie

– U probanda byla prokázána *de novo* varianta v genu *DHDDS* (NM_024887.3) na pozici c.110G>A (p.Arg37His) v heterozygotním stavu, která byla již dříve publikovaná [Hamdan et al. 2017] jako příčina epileptické encefalopatie. Varianta nemá frekvenci v populačních databázích a byla hodnocena jako patogenní predikčními nástroji. Tuto variantu jsme neprokázali u rodičů, je tedy nejspíše *de novo* vzniklá.

– U probanda byla prokázána *de novo* varianta v genu *SERPINI1* (NM_005025.4) na pozici c.1174G>A v heterozygotním stavu, která byla již dříve publikována [Ranza et al. 2017] jako příčina progresívní myoklonické epilepsie dětského věku nebo neurodegenerace s Collinsovými tělísky. Fenotypově odpovídá průběh onemocnění publikovaným případům. Pacientův vývoj probíhal normálně do 8 let věku, poté nastává regres s progresivní myoklonickou epilepsií. Varianta nemá frekvenci v populačních databázích a byla hodnocena jako pato-

genní predikčními nástroji. Tuto variantu jsme neprokázali u rodičů, je tedy nejspíše *de novo* vzniklá.

– U probanda byla prokázána *de novo* varianta v genu *SCN8A* (NM_014191.3) na pozici c.5630A>G v heterozygotním stavu, která byla již dříve publikována [Anand et al. 2016] jako příčina benigní familiární infantilní epilepsie. Je uvedená i v databázi ClinVar jako patogenní u časně infantilní epileptické encefalopatie, predikční nástroje označily variantu protichůdně, v populačních databázích se varianta nenachází. Tuto variantu jsme neprokázali u rodičů, je tedy nejspíše *de novo* vzniklá.

– U probandky byly prokázány dvě varianty v genu *SPATA5* (NM_145207.2) v heterozygotním stavu. Varianta c.2563C>T (p.Gln885*) nebyla dosud popsána jako patogenní, predikuje zařazení předčasného stop kodonu a nemá frekvenci v populačních databázích. Segregační analýzou byla prokázaná u zdravé matky v heterozygotním stavu. Druhá varianta na pozici c.394C>T (p.Gln132*) již byla popsána jako patogenní, v publikaci [Puusepp et al. 2018] byla označena jako přičina opoždění vývoje, mikrocefálie, epilepsie a poruchy sluchu s AR dědičností, varianta predikuje zařazení stop kodonu a byla prokázána u otce probandky v heterozygotním stavu. Varianty jsou v pozici trans.

– U probanda byla prokázána *de novo* varianta v genu *YWHAG* (NM_012479.3) na pozici c.170G>A (p.Arg57His) v heterozygotním stavu, která nebyla dosud publikována ani není přítomná v populačních databázích, predikční nástroje jí hodnotí jako patogenní. Varianta predikuje záměnu vysoce konzervované aminokyseliny za málo odlišnou a je predikčními programy hodnocena jako závažná - patogenní. Tuto variantu jsme neprokázali u rodičů, je tedy nejspíše *de novo* vzniklá. Pro patogenicitu této varianty rovněz hovoří publikace [Guella et al. 2017], která se popisuje missense varianty v genu *YWHAG* jako příčinu epileptické encefalopatie *s* fenotypem podobným jako je fenotyp u probanda.

– U probanda byla prokázána varianta v genu *SLC1A4* (NM_003038.4) na pozici c.1370G>A (p.Arg457Gln) v homozygotním stavu, která nebyla dosud popsána jako patogenní. Dříve ale byla popsána varianta postihující stejný kodon [Damseh et al. 2015]. Varianta je hodnocená predikčními nástroji jako patogenní, a v populačních databázích je uvedená pouze v heterozygotním stavu s frekvencí 0,002 %. Varianty v *SLC1A4* genu jsou příčinou AR těžkého neurovývojového onemocnění se záchvaty s různými stupni spasticity [Pironti et al. 2018]. Varianta byla segregační analýzou potvrzena u obou rodičů v heterozygotním stavu v pozici trans.

– U probandky byla prokázána varianta v genu *GABRB2* (NM_021911.2) na pozici c.895A>C (p.Ile299Leu) v heterozygotním stavu, která dosud nebyla popsána jako patogenní, nemá frekvenci v populačních databázích a predikčními nástroji je hodnocená jako patogenní. V genu *GABRB2* ale již byly popsány *de novo* missense varianty způsobující vývojové a epileptické encefalopatie [Ham-

dan et al. 2017]. Tuto variantu jsme neprokázali u rodičů, je tedy nejspíše *de novo* vzniklá.

– U probanda byly prokázány dvě varianty v genu *NARS2* (NM_024678.5) na pozici c.83_84del (p.Leu28Glnfs*17) a c.1339A>G (p.Met447Val), v heterozygotním stavu. Ani jedna z variant není uvedená v populačních databázích a nebyly dříve popsány jako patogenní. První varianta predikuje vznik předčasného stop kodonu a druhá byla hodnocena predikčními nástroji protichůdně. Segregační analýza prokázala výskyt varianty jednotlivě u každého z rodičů, jde tedy o varianty v pozici trans. Bialelické mutace inaktivující a missense jsou příčinou infantilní neurodegenerace CNS a poliodystrofie vč Alpersova syndromu [Mizuguchi et al. 2017].

Dědičné neuropatie

– U probanda byla prokázána varianta v heterozygotním stavu v genu *NEFL* (NM_006158.4) na pozici c.967_978 (p.Arg323_Asn326del). Varianta není přítomná v populačních databázích a způsobuje in-frame deleci, podobná delece již byla publikována jako příčina dědičné neuropatie. Varianta byla nalezena u podobně postižené matky a není přítomna u zdravého bratra probanda.

– U probanda byla prokázána varianta v genu *AIFM1* (NM_004208.3) na pozici c.134C>G (p.Pro45Arg) v hemizygotním stavu, která je uvedena v populačních databázích s nízkou frekvencí, predikční nástroje jí hodnotí protichůdně. V genu *AIFM1* byly publikovány další varianty způsobující axonální polyneuropatie [Sancho et al. 2017; Hu et al. 2017].

– U probanda byly prokázány dvě varianty v genu *SLC25A46* (NM_138773.2) na pozici c.1208T>G (p.Leu403*) a c.1075G>A (p.Val359Met) v heterozygotním stavu. Ani jedna z variant není uvedená v populačních databázích, ale jde o varianty závažné. První varianta způsobuje vznik předčasného stop kodonu a druhá byla hodnocena jako patogenní predikčními nástroji. Obě varianty, v heterozygotním stavu jsme prokázali u podobně postižené sestry a jedna z variant (p.Leu403*) byla prokázána v heterozygotním stavu u zdravé matky (vzorek od otce není k dostupný). Bialelické mutace v genu *SLC25A46* byly již dříve publikovány jako příčina autosomálně recesivní atrofie zrakových nervů s axonální polyneuropatií neuropatií [Abrams et al. 2015].

– U probandky byla prokázána varianta v genu *ATP1A1* (NM_001160233.1) na pozici c.620C>T (p.Ser207Phe) v heterozygotním stavu, která dosud nebyla popsána jako příčina onemocnění a nemá frekvenci v populačních databázích. Varianta je hodnocena jako patogenní predikčními nástroji. Heterozygotní patogenní mutace v genu *ATP1A1* jsou nově popsanou příčinou dominantní HMSN II [Lassuthova et al. 2018]. Varianta nebyla prokázána u údajně zdravé matky.

– U probandky byla prokázána varianta v genu *KCNK9* (NM_001282534.1) na pozici c.710C>A (p.Ala237Asp) v heterozygotním stavu, která dosud nebyla

popsána a nemá frekvenci v populačních databázích, predikční nástroje jí hodnotí jako patogenní. Varianty ve vedlejším kodonu byly dříve publikovány jako příčina „*Birk-Barel Mental Retardation Dysmorphism*“ [Graham et al. 2016]. Gen *KCNK9* podléhá imprintingu a je utlumen na paternální alele, varianta je zděděná od otce, který je bez fenotypu. Výsledek byl publikován v [Šedivá et al. 2019].

4.2.5 CNV analýza

Pro testování nástroje na detekci CNV jsme použili dva případy, u kterých jsme již dříve potvrdili CNV nález metodou *MLPA*. U prvního případu se jednalo o pacienta s hereditární spastickou paraparézou, se sporadickým výskytem v rodině, u kterého jsme nalezli heterozygotní variantu měnící sestřih mRNA. Pomocí *MLPA* poté byla detekována delece tří exonů 37 38 39 v genu *SPG11*.

Druhý pacient měl diagnostikovánu dědičnou, nesyndromovou hluchotu. U pacienta byla nalezena delece genu *STRC* (a sousedícího genu *CATSPER*), s potvrzením metodou *MLPA*. Gen *STRC* má pseudogen, jehož sekvence se shoduje ve více než 99,5 %, proto se tento případ ukázal jako vhodný pro otestování metod detekce CNV.

K celému testování jsme přistupovali ve dvou krocích, kdy první test byl proveden v základních nastaveních všech nástrojů tak, jak je uvedeno v dokumentaci, druhý test pak proběhl s nastavením řádově vyšší senzitivity všech nástrojů (pro detekci maximálního počtu variant).

Výsledkem analýzy nástrojem metaSV byl textový soubor s CNV variantami, kdy v základním nastavením bylo detekováno 240 303 CNV a s maximálně senzitivním 293 954 CNV, avšak bez detekce přítomné CNV v námi definovaných oblastech zájmu – genů *SPG11* a *STRC*.

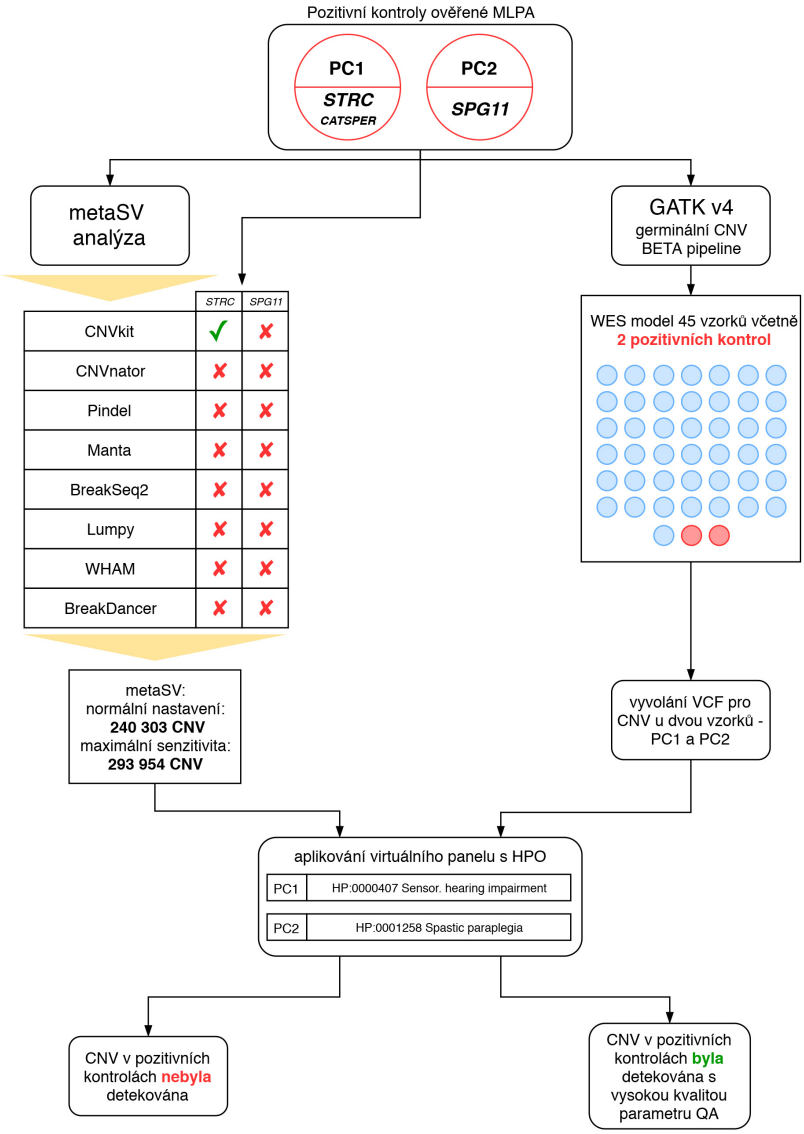
Tento výsledek byl pro nás signifikantní ze dvou důvodů. Algoritmy pro detekci CNV u WES dat mohou poskytovat velké množství výsledků a tudíž je nutné vědět, kde hledáme. Navíc v obou případech byly případy uzavřeny na základě předpokladu, že se jedná o deleci v genu spojovaném s onemocněním, proto jsme vytvořili nástroj na rychlou tvorbu virtuálních panelů, pomocí HPO termínů.

Druhým poznatkem bylo, že žádný z otestovaných nástrojů nedetekoval v našich případech přítomné a dříve prokázané CNV. Jedinou výjimkou byl nástroj CNVkit, který po mnohonásobném zvýšení senzitivity byl schopný detekovat CNV v *STRC* genu. Ostatní nástroje ale nedetekovaly CNV, důvodem mohlo být:

- neuzpůsobení nástroje na WES data
- nekontinuální pokrytí kitu, kdy nástroje si neumí poradit s nepokrytými místy a že se jednalo o příliš krátké, přerušované úseky
- některé nástroje nebyly rovněž určené pro detekci germinálních variant, ale pouze pro somatické varianty – očekávají jako vstup dva vzorky od stejného pacienta

Celá analýza byla poté opakována pomocí GATK v. 4, kdy jsme nejprve vytvořili model čítající 45 vzorků WES, sekvenovaných stejnou knihovnou (včetně dvou

probandů zmíněných výše). Poté dochází k vyvolání VCF pro každý vzorek z modelu. Na výsledné VCF byl aplikován virtuální panel dle fenotypu - v prvním případě „Spastic paraplegia HP:0001258“ a v druhém případě „Sensorineural hearing impairment HP:0000407“. Po vyfiltrování souboru byly detekovány obě delece v souladu s předchozími nálezy, kdy skóre (parametr *QA*) patřilo vždy mezi deset nejvyšších v souboru. Celý proces testování je zobrazen na schématu níže (Obr. 4.9).



Pro porovnání byly vybrány nástroje metaSV (průběh vlevo) a GATK 4 metodika pro detekci germinálních CNV v beta verzi, pozitivní kontroly, pro které byla analýza prováděna jsou označena textově jako PC1 a PC2, a graficky červeně.

Obrázek 4.9: Schéma testování CNV nástrojů

4.3 Výsledky z celogenomového sekvenování (WGS)

Zpracování dat z celogenomového sekvenování (WGS) znamená další krok ve vyhodnocování dat. K WGS přistupujeme nejčastěji u vybraných pacientů jejichž předchozí analýza pomocí WES neodhalila kauzální variantu. Při analýze dat volíme dva přístupy k vyhodnocování.

Prvním krokem je analýza s BED souborem z WES sekvenování, kdy získaný VCF soubor má shodné parametry jako předchozí WES, s tím, že dříve nepokryté oblasti (které jsou kódující, ale nejsou pokryté designem) jsou nyní zahrnuty do analýzy a je možné detekovat varianty.

Pomocí WGS jsme ale schopni nalézt nové varianty, které u předchozích metod nebyly detekovány. Během analýzy může dojít k nalezení varianty v nově popsaných genech, které dříve nebyly s onemocněním asociovány (nebo nebyly pokryté), nebo může být kauzalita způsobena zcela jiným mechanismem než jsou klasické SNV nebo CNV, příkladem můžou být strukturální varianty.

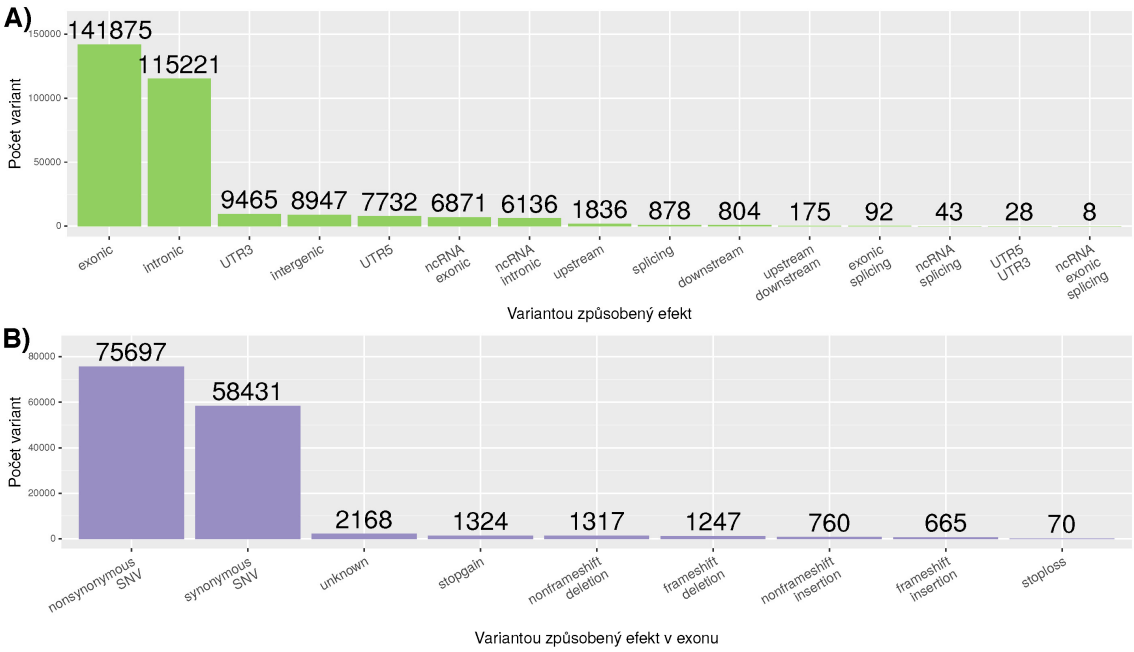
V DNA laboratoři jsme vyhodnotili celkem 33 WGS vzorků (přehled v kapitole Pacienti na obrázku Obr. 3.1), šest pacientů s CMT, dva s EE a 25 s hluchotou. U všech jsme provedli obě analýzy (jak virtuální WES, tak WGS). Dosud jsme však z WGS nenalezli žádnou vysvětlující kauzální variantu – to může být způsobeno i tím, že jde u těchto pacientů o dosud neobjevené typy onemocnění – v důsledku změn v genech dosud nespojovaných s onemocněním.

4.4 Bioinformatické databáze

4.4.1 Databáze WES variant DNA laboratoři

Celkem bylo vyvoláno 300 111 variant ve 17 512 genech, tyto varianty byly anotovány nástrojem ANNOVAR pro doplnění populačních frekvencí a predikčních nástrojů. Výsledná frekvence *dbAF* ukazuje relativní výskyt alely v celém datasetu. Medián alelické frekvence je 0,009 (což odpovídá vyvolání 4 alel), průměr 0,110 (odpovídá přibližně 49 alelám).

V souboru bylo nalezeno 141 875 kódujících variant. V této skupině bylo více než 54 % variant označeno jako nesynonymních (nejčastěji missense varianty) a 42 % variant jako synonymních. Přehled variant dle jejich efektu je uveden na Obr. 4.10, jak v rámci celé sekvence, tak i zvlášť pro exonové oblasti (dle RefSeq).



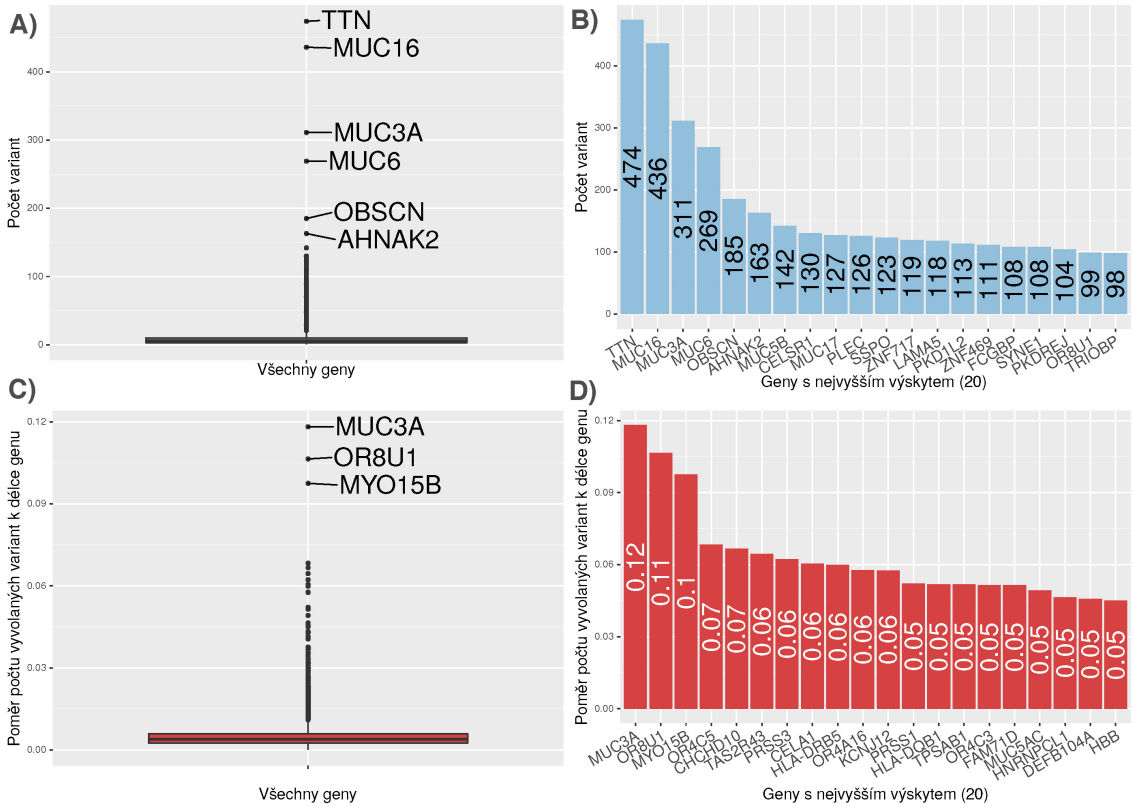
A) v rámci celé sekvence B) v rámci exonů

Obrázek 4.10: Přehledy typů variant podle lokalizace a predikovaného efektu

Dalším krokem byla analýza počtu variant dle genů.

Na obrázku Obr. 4.11 níže uvádíme distribuci variant dle jednotlivých genů (počtu exonových variant na gen) modře. Průměrně bylo v kódující sekvenci každém genu vyvoláno 8,1 variant (medián 6), nejvyšší počet vyvolaných exonových variant byl v genech *TTN* (474 variant), *MUC16* (436 variant), *MUC3A* (311 variant) *MUC6* (269 variant), další geny jsou uvedené v obrázku.

Stejnou analýzu jsme provedli pro parametr *varbp*. Tento parametr měl v průměru hodnotu 0,005 to znamená 5 variant na tisíc bází (medián 0,004). Vysoké hodnoty (nad 0,08) byly zaznamenány u genů *MUC3A* (nejvyšší hodnota 0,118), *OR8U1* (0,106) a *MYO15B* (0,098).



A) distribuce počtu variant dle genů v boxplotu B) výběr 20 genů s nejvyšším počtem vyvolaných variant C) distribuce počtu vyvolaných variant dle *varbp* D) výběr 20 genů s nejvyšší hodnotou počtu vyvolaných variant dle *varbp*

Obrázek 4.11: Vyvolané varianty dle genů

4.4.1.1 Prezentace variant

Pro zpřístupnění databáze jsme zvolili dvě cesty. První je privátní pro uživatele DNA laboratoře, druhá je veřejná. Hlavním rozdílem je různá dostupnost informací. Běžný uživatel získá přes webové rozhraní veřejný přístup k databázi, ve které může vyhledávat varianty dle genu popř. podle genomické pozice. Ve výsledné tabulce je pak vidět:

- název genu
- genomická pozice
- referenční a alterovaná báze
- frekvence v *dbAF*
- funkce varianty (v rámci celé sekvence i exonická)
- hodnocení dle databáze ClinVar
- frekvence v databázi gnomAD exome all

– frekvence v databázi gnomAD exome NFE (non-Finnish European)

V rámci DNA laboratoře pak můžeme využít anotační skript, který má přístup k „plné verzi“ databáze. Tím dokážeme získat informace o počtu pacientů, kteří danou variantu měli detekovanou v homozygotní či heterozygotní formě a pak identifikační číslo jejich vzorku. Tyto čtyři parametry nejsou dostupné pro všechny uživatele z důvodu anonymizace a ochrany dat.

Pro prezentaci variant jsme zvolili jednoduché webové rozhraní, které umožňuje připojení k relační databázi typu SQL. Uživatel má k dispozici jednoduché textové okno, kam zapisuje hledané údaje, vždy jeden na řádek. Validními vstupy jsou buď názvy genů (vždy jeden gen na řádek), popř. genomické pozice ve tvaru „chr10:546421654“ (vždy jedna na řádek). Po vyhledání pak uživatel v tabulce vidí varianty splňující hledané parametry a má možnost data dále prohledávat popř. exportovat do souboru. Databáze je přístupná pro prohlížení přes webové rozhraní na URL: **prot2hg.com/variantbrowser** .

4.4.2 Databáze proteinových domén prot2hg.com

Zpracováním prošlo celkem 42 371 proteinů, z nichž jsme vložili do databáze 808 886 proteinových domén (190 760 regionů a 616 126 sites). Regiony jsou průměrně dlouhé 315.6 nukleotidů a sites 7.3 nukleotidů, to znamená, že více než 75 % databáze tvoří velmi krátké proteinové domény. Pro každý záznam se v databázi nachází následující informace: název genu, RefSeq_ID (NP_ID a NM_ID), mapovaný řetězec (+ / −), typ domény (Site/Region), název domény, mapovací skóre, genomické pozice a další záznam, získaný v NCBI databázi.

Uživatelské rozhraní Dalším krokem bylo zpřístupnění dat. Vytvořili jsme jednoduchou webovou aplikaci na adrese **www.prot2hg.com**, vytvořenou na bázi PHP webového serveru s TwitterBootstrap HTML frameworkem (sadou nástrojů). Aplikace se připojuje k MySQL databázi, ve které jsou uložená data.

Hlavní stránka umožňuje anotaci jednotlivých variant po zadání jejich genomických souřadnic (je možné zadat až 100 variant pro jednu anotaci). Aplikace pak vypíše ty domény, do kterých daná varianta spadá. Druhou možností pro uživatele je stažení celé databáze a to ve více formátech – JSON, SQL a CSV. V tomto případě je soubor vhodný k zakomponování do vlastní bioinformatické pipeline, kdy je pak možné využívat anotaci i na větší datasety.

Anotační analýza Pro ověření správné funkce databáze a zjištění, jestli pozice v proteinové doméně je pro analýzu NGS dat zajímavá, jsme provedli anotační analýzu. Pro tuto analýzu jsme využili dva zdroje – s databází gnomAD exome [Karczewski et al. 2019], obsahující více než 15 milionů variant s určenou frekvencí v populaci a databázi ClinVar [Landrum et al. 2016], která obsahuje i informace o patogenicitě variant. Tím jsme získali vlastní databázi variant s frekvenci v populaci gnomAD_exome_ALL a s její patogenicitou dle ClinVar databáze.

Tuto databázi jsme pak anotovali pomocí prot2HG a po anotaci jsme vypočítali následující parametry:

- Počet unikátních anotovaných variant
- Počet unikátních variant anotovaných do regionu a do site
- Počet vzácných variant (s frekvencí v gnomAD_exome_ALL < 0.01)
- Počet polymorfismů (s frekvencí v gnomAD_exome_ALL > 0.01)
- Počet patogenních a pravděpodobně patogenních variant dle ClinVaru

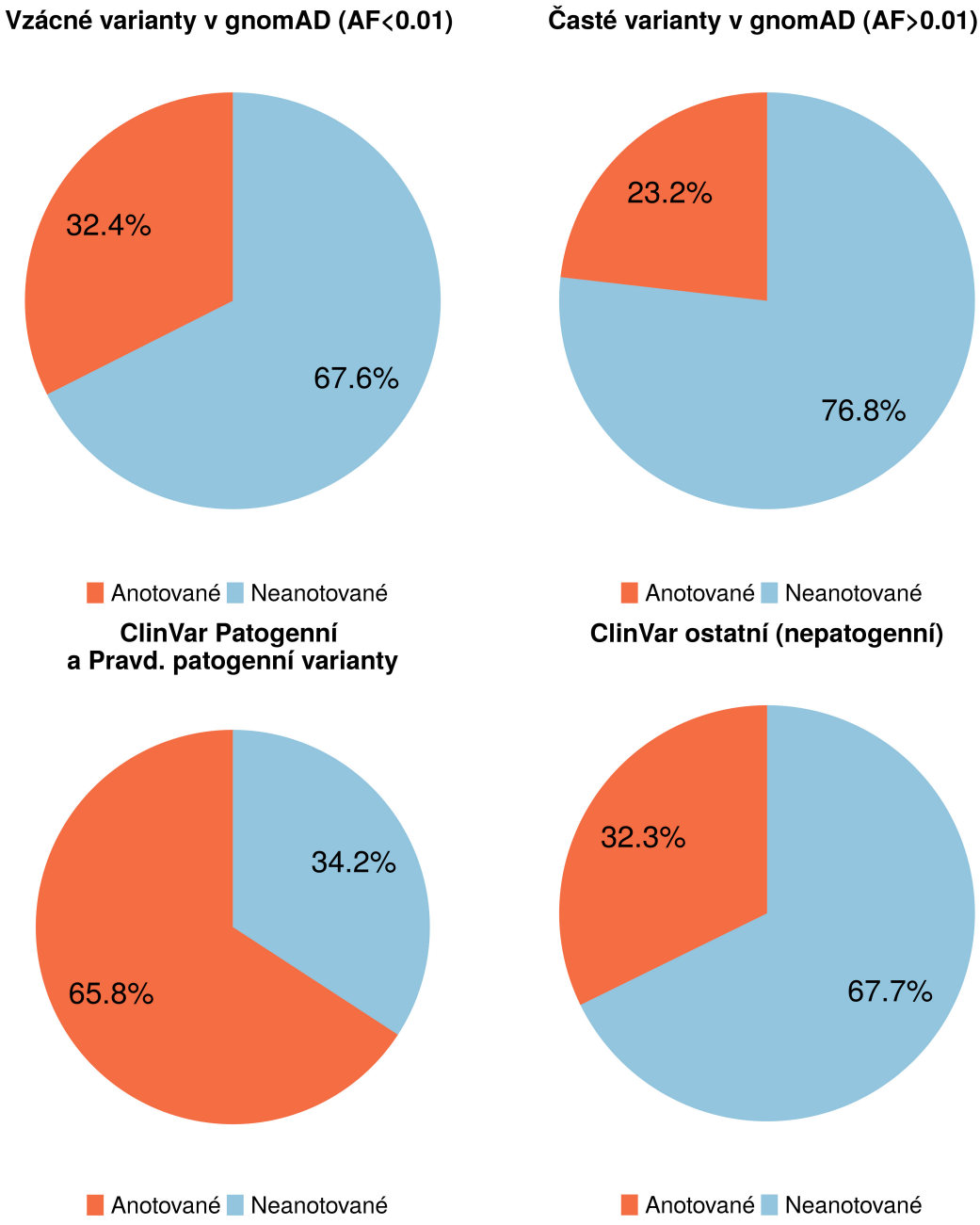
Cílem analýzy bylo zjistit, jestli se počty vzácných variant signifikantně liší ve skupině anotovaných a neanotovaných variant a rovněž počet patogenních variant by se měl lišit v obou skupinách. Pro statistické vyhodnocení jsme použili Chí kvadrát test nezávislosti.

Nulová H_{0a}hypotéza pro analýzu gnomAD dat předpokládá, že na základě populační frekvence není rozdíl mezi počtem vzácných variant z gnomAD, které budou spadat do oblasti domén (anotované) a které budou spadat mimo oblast domén (neanotované).

Po anotaci variant nacházejících se v gnomADu byl určen poměr anotovaných a neanotovaných variant. U vzácných (frekvence menší než 1 %) variant bylo anotováno do proteinové domény 32,4 % variant, u variant častých v populaci (frekvence větší než 1 %) bylo anotováno 23,2 % variant. Tento výsledek je signifikantní při p blížící se nule, proto můžeme hypotézu H_{0a} zamítnout, mezi frekvencí menší než 1 % a lokalizací uvnitř domény existuje vztah. Poměry jsou uvedeny na grafu Obr. 4.12 části A.

Nulová H_{0b}hypotéza pro analýzu ClinVar dat předpokládá, že není rozdíl mezi počtem patogenních a pravděpodobně patogenních variant dle hodnocení ClinVar, které budou spadat do oblasti domén (anotované) a které budou spadat mimo oblast domén (neanotované).

V tomto případě byly výsledky ještě přesvědčivější než při vyhodnocení dle gnomAD databáze. Z patogenních a pravděpodobně patogenních variant dle ClinVar databáze bylo 65,8 % variant lokalizováno v proteinové doméně anotací prot2HG. U variant, které nebyly hodnocené jako patogenní / pravděpodobně patogenní dle ClinVar databáze bylo anotovaných do proteinové domény 32,3 %. Tento výsledek byl signifikantní pro p blížící se 0, můžeme hypotézu H_{0b} zamítnout, vztah mezi patogenicitou dle ClinVar databáze a lokalizací v proteinové doméně byl determinován. Poměry jsou uvedeny na grafu Obr. 4.12 části B.



A) varianty v databázi gnomAD, u vzácných variant bylo anotováno 32,4 % a u častých variant pouze 23,2 %
B) rozdíl mezi patogenními a ostatními variantami v ClinVar databázi, u patogenních variant bylo anotováno 65,8 % variant a u ostatních pouze 32,3 % variant

Obrázek 4.12: Poměry anotovaných a neanotovaných variant

4.4.3 Databáze variant spojených s CMT

Tato kapitola shrnuje výsledky publikované ve článku [Saghira et al. 2018].

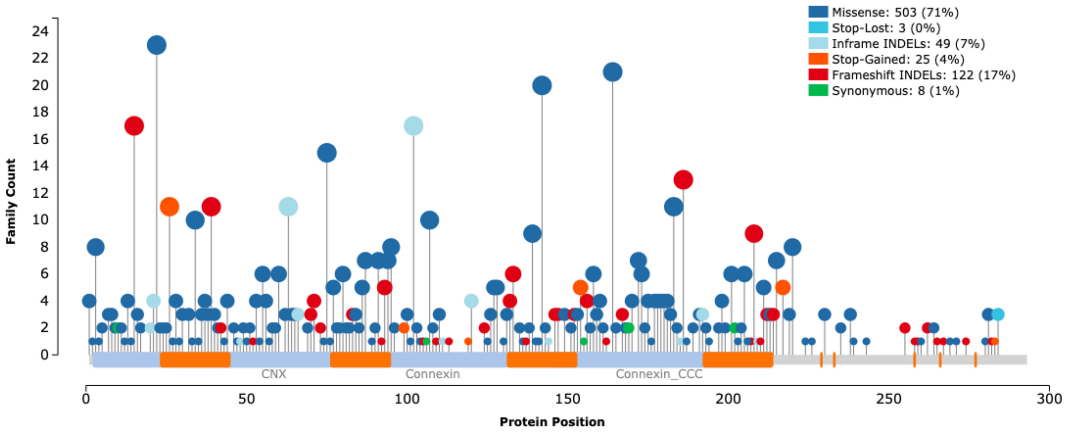
V prvotní fázi bylo v databází celkem 3 809 variant v 82 genech. Nejvíce variant bylo v genu *GJB1* (720 variant), medián byl 16 variant v genu. Celkem byla použita data 4 558 nepříbuzných rodin, 2 244 heterozygotních jedinců, 528 homozygotů a 301 složených heterozygotů.

Uživatelské rozhraní Webové rozhraní je přístupné z URL[1]zde můžeme hledat varianty dle genu (je možné vybrat i více genů najednou), typ varianty a zdroj odkud byla varianta získána.

Rozhraní zobrazuje rozprostření variant v rámci proteinu, kde je možné vidět nejen umístění v sekvenci (osa x), ale i jestli varianta spadá do některé proteinové domény, její četnost (osa y) a barevné rozdělení dle typu (missense, stop-gain, frameshift...), další možností je zobrazení variant seřazených dle frekvence v populaci (ExAC MAF)[Karczewski et al. 2019].

Každá varianta má vlastní hodnocení dle ACMG, vyjádřené počtem (1 až 5) hvězdiček, dle klasifikace – benigní, pravděpodobně benigní, varianta nejasného významu, pravděpodobně patogenní, patogenní. Registrovaný a přihlášený uživatel má možnost přidávat varianty vyplněním jednoduchého formuláře – vyplní se varianta, gen, typ varianty, její genotyp a identifikátor její sekvence s tím, že nakonec musí být ještě uvedený zdroj Náhled webové aplikace je na obrázku Obr. 4.13.

[1] hihg.med.miami.edu/neuropathybrowser [online 16.10.2019]



Variants

Page Size: 500

Filtered **431** of **3835** Total Variants

Gene	Variant	Protein Notation	Mutation Type	Rating	Links	Data Source
GJB1	c.1A>T	p.Met1Leu	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.2T>G	p.Met1Arg	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.3G>C	p.Met1Ile	Missense	★★★★★	E O C N	Clinical Report (1) Published Paper (1)
GJB1	c.3G>T	p.Met1Ile	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.-6G>A		5' UTR	☆☆☆☆☆	E O C N	Clinical Report (1)
GJB1	c.6C>G	p.Asn2Lys	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.7T>C	p.Trp3Arg	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.7T>G	p.Trp3Gly	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.8G>A	p.Trp3Ter	Stop-Gained	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.8G>C	p.Trp3Ser	Missense	☆☆☆☆☆	E O C N	Published Paper (3) Clinical Report (1)
GJB1	c.8G>T	p.Trp3Leu	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.11C>A	p.Thr4Lys	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.13G>A	p.Gly5Ser	Missense	☆☆☆☆☆	E O C N	Published Paper (1)
GJB1	c.14G>T	p.Gly5Val	Missense	★★★☆☆	E O C N	Research Finding (1)
GJB1	c.-16-3C>G		Splice Site	☆☆☆☆☆	E O C N	Published Paper (1)

Nahoře grafické vyjádření s počtem variant s pozicí v sekvenci, dole seznam variant včetně jejich typu, hodnocení a zdroje dat
Obrázek z webové adresy: hihg.med.miami.edu/neuropathybrowser

Obrázek 4.13: Přehled genu *GJB1* v Inherited Neuropathy Variant Browser

4.5 Nástroj pro virtuální panely

Nástroj pro virtuální panely genů je do velké míry univerzální a je možné ho aplikovat na filtrování variant v širokém spektru aplikací - při manuálním filtrování WES dat, WGS dat, při vyhodnocování CNV analýzy u WES/WGS dat (kdy nám pomáhá snížit počet suspektních CNV). Pro demonstraci „účinnosti“ zde uvádíme výsledky filtrování při využití datasetu in-house databáze WES variant ze sekce 4.4.1.

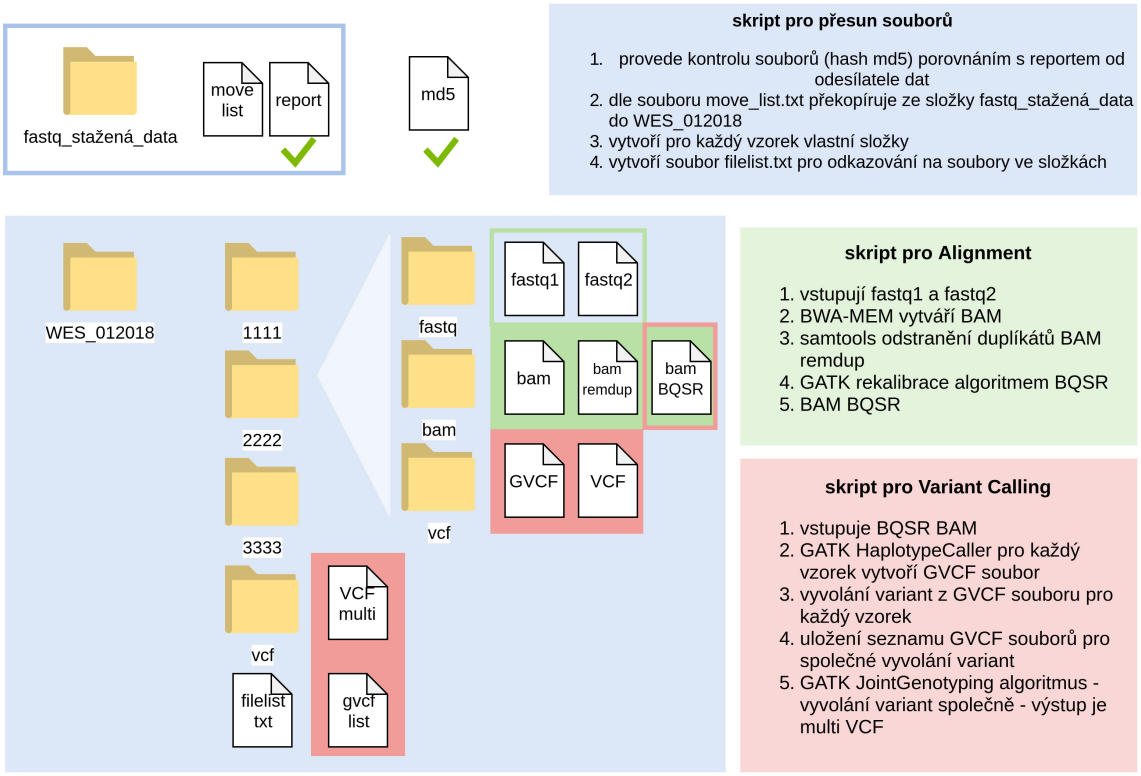
Pro demonstrační účely jsme vytvořili dva virtuální panely dle „méně specifických“ HPO termínů, pro EE to byl termín „Seizures HP:0001250“ a pro CMT to byl termín: „Peripheral neuropathy HP:0009830“.

V případě EE se jednalo v současnosti o 1 346 asociovaných genů s termínem „Seizures“, po vyfiltrování souboru zbylo v oblasti genů (exony i introny) 281 866 variant (3,1 % původního počtu) a po omezení pouze na kódující oblasti zbylo 17 076 variant (0,2 % původního počtu).

V případě CMT se jednalo o 493 asociovaných genů s termíny „Peripheral neuropathy“. Po vyfiltrování zbylo v oblasti genů (exony i introny) 82 423 variant (0,9 % původního počtu) a po omezení na exonovou oblast zbylo 6 862 variant (0,08 % původního počtu).

4.6 Správa dat v DNA laboratoři

V této kapitole je popsán námi navržený postup pro zpracování dat, konkrétně krok po kroku od přijetí dat po uzavření případu, schéma zpracování dat je uvedené na obrázku Obr. 4.14.



Obrázek 4.14: Schéma správy dat integrované v rámci DNA laboratoře

Prvním krokem je stažení a kontrola integrity dat pomocí md5 součtu. U každého přijatého FASTQ souboru je tato kontrola provedena. Kontrola probíhá tak, že u souboru se vypočítá tzv. md5 hash, což je řetězec vygenerovaných znaků unikátních pro každý soubor. Takto vygenerovaný řetězec se porovnává s řetězcem, který je zaslán odesílatelem souborů. Pokud se řetězce shodují, byl soubor doručen beze změn, pokud se liší, došlo k problému během doručení souboru (stahování, poškození disku na kterém se soubor nachází). Kontrola pomocí md5 je výhodná v tom, že i velmi malá změna souboru znamená vygenerování zcela jiného hash řetězce, čímž tuto změnu odhalíme.

Po kontrole dat se data ukládají do interního systému, který funguje na stromové bázi (složky). K tomu slouží přesouvací skript, jehož vstupem je uživatelem předpřipravený soubor *move_list.txt*, který slouží k importu informací a dat z NGS do cílového adresáře, soubor *move_list.txt* má následující formát:

- číslo vzorku [xxxxxxx]
- název složky [xxxxxxx] (bez lomítek)

- fastq1 [/fastq/xxxxx.fastq]
- fastq2 [/fastq/xxxxx.fastq]
- typ [GP, WES, WGS] - GP = panel genů, WES = celoexomové sekvenování, WGS = celogenomové sekvenování
- název knihovny [GP_MMRRRR / WES_MMRRRR] – MM značí měsíc a RRRR rok sekvenace

Data se pak přesouvají dle následujících pravidel:

- kořenový adresář _data, obsahující všechny knihovny NGS dat
- obsah kořenového adresáře:
  - skripty pro analýzu dat
  - jednotlivé složky knihoven NGS dat
    - každá složka je pojmenována dle následujícího klíče:
      - TypKnihovny_MěsícRokZpracování př. WES_082018 – pro WES data ze srpna 2018, nebo GP_012019 pro Panel EE z ledna 2019
      - obsahují soubor *filelist.txt* (popis níže)
      - obsahují soubor *gvcf.list* (popis níže)
      - obsahují složku vcf, ve které se nachází multi VCF soubor, obsahující varianty všech vzorků, z dané knihovny
      - obsahují složku označenou číselným kódem každého vzorku, ta obsahuje 3 podsložky, vcf, bam a fastq, kam se ukládají jednotlivé soubory během analýzy

Data jsou zpracována pomocí připravených skriptů v jazyce bash. Základní analýza probíhá ve dvou skriptech – první pro převod z FASTQ do BAM souboru a druhý pro převod BAM do VCF souboru. Analýza probíhá v rámci celé knihovny, kdy cesty k souborům jsou definovány v souboru filelist.txt, který obsahuje všechny potřebné informace o vzorcích v dané knihovně, má následující formát, který je vygenerovaný během přesouvacího skriptu. Soubor filelist slouží jako hlavní „mapa“, podle které probíhá další zpracování dat, má následující formát:

- číslo vzorku [xxxxxxx]
- celá cesta do složky vzorku [/../../xxxx] (bez lomítka na konci)
- fastq1 [/fastq/xxxxx.fastq]
- fastq2 [/fastq/xxxxx.fastq]
- typ [GP, WES, WGS] - GP = panel genů, WES = celoexomové sekvenování, WGS = celogenomové sekvenování
- název knihovny [GP/WES_MMRRRR] – MM značí měsíc a RRRR rok sekvenace

– cesta do nadřazeného adresáře s daty [/../../all_data] (zde jsou uloženy knihovny)

Druhým vstupem, nutným ke zpracování je parametr o odstraňování PCR duplikátů. Toto nastavení se mění dle sekvenačního kitu, nelze předem detekovat a proto ho musí uživatel definovat ručně, proto do skriptů pro zpracování NGS pomocí GATK worfklow musí zadávat jako vstupní parametr kromě souboru filelist.txt i hodnota „*true*“ pro odstranění duplikátů (sekvenační kit *SureSelect*) či hodnota „*false*“ (sekvenační kit *Haloplex*). Konkrétní kroky NGS analýzy pak jsou:

1. Vytvoření souboru *move_list.txt*, který odkazuje na FASTQ soubory připravené k analýze

2. Po přesunutí se v pracovním adresáři _data vytvoří složka knihovny, kde je soubor *filelist.txt*

3. Pokud se jedná o knihovnu s předpřipraveným designem, je nutné nahrát BED soubor do složky a pojmenovat ho shodným názvem jako je složka, s příponou BED

4. Spouští se skript fastq2bam.sh, se dvěma vstupy – souborem *filelist.txt* a hodnotou true/false pro odstranění/ponechání PCR duplikátů, v tomto kroku probíhá alignment, tedy přiřazení sekvence vzorku k referenční sekvenci hg19
   a) Prvním krokem je vytvoření souboru SAM algoritmem BWA-MEM, kdy vstupují FASTQ soubory, a údaje o vzorku – číslo, knihovna, typ, sekvenátor, velikost sekvenačního kitu
   b) Nástroj samtools převádí soubor SAM na typ BAM a vytvoří jeho index (BAI soubor)
   c) Pokud byla vstupní hodnota nastavena na „*true*“ dochází k odstranění duplikátů pomocí nástroje samtools a picard
   d) Dochází k rekalibraci readů algoritmem BQSR v rámci nástroje GATK (používaná verze 3.8)
   e) Výstupem jsou BAM soubory - původní, převedený ze SAM souboru, „_remdup.bam“ soubor, s odstraněnými duplikáty a „_recal.bam“ s provedenou rekalibrací nástrojem BQSR – ten se používá pro další analýzu

5. Probíhá variant calling skriptem bam2vcf.sh, se dvěma vstupy – souborem filelist.txt a hodnotou true/false pro informaci, pokud byly v předchozím kroku duplikáty odstraněny. Vyvolávání variant z BAM souboru probíhá v několika krocích:
   a) Prvním krokem je vyvolání variant algoritmem HaplotypeCaller v rámci nástroje GATK, na BAM soubor se aplikuje soubor BED s designem pro nastavení intervalu pro vyvolání variant, výstupem je GVCF soubor obsahující nejen varianty, odlišující se od reference ale i bloky shodující se s referencí, kvůli společnému vyvolání variant v rámci celé knihovny v dalším kroku, odkaz na umístění se uloží do souboru gvcf.list, který je v kořenové složce knihovny

b) Dalším krokem je vyvolání variant pro každý vzorek zvlášť, výstupem je VCF soubor obsahující pouze varianty jednoho vzorku

c) Po vygenerování jednotlivých GVCF a VCF (vzorek po vzorku) dochází k vyvolání společného multi VCF v rámci celé knihovny, kdy výsledkem je VCF soubor obsahující všechny vzorky, výhodou je, že při společném vyvolání (algoritmus JointGenotyping) dojde ke snížení počtu systematických sekvenačních chyb, přítomných ve více vzorcích knihovny. Výsledkem je multi VCF připravené k další analýze. U všech VCF souborů pak probíhá kontrola zarovnání doleva (Left Align) a rozdělení VCF na SNP a INDEL, většinou je ale pro další analýzu využíván nerozdělený multi VCF soubor

6. Dochází k dalšímu zpracování, které se liší dle typu knihovny a individuálních požadavků (filtrování, anotace, editace)

Tento standardní postup byl během využívání několikrát modifikován, došlo tak k vytvoření dalších skriptů:

- **fastq2bam_noBQSR.sh** – vynechává rekalibraci readů, z důvodu šetření kapacity diskového úložiště aplikováno na WGS data, rekalibrované soubory jsou totiž přibližně 2 až 3 krát větší než BAM soubory z předchozího kroku (úspora kapacity i času při generování)

- **gvcf_2vcf_multi.sh** – skript, který vytváří pouze multiVCF ze seznamu GVCF souborů (soubor *gvcf.list*), hlavní použití je, pokud chceme vyvolat společně varianty u vzorků z více knihoven, dojde k vytvoření vlastního seznamu GVCF souborů, z nichž je pak multi GVCF vygenerováno

- Sada skriptů **_nobed.sh** – pro účely analýzy v nástroji Ingenuity jsme se rozhodli generovat VCF soubory bez omezení BED souborem, proto jsme vytvořili sadu skriptů s příponou „*nobed*“, korektně pracujících bez přítomného BED souboru, další využití bylo i u WGS dat, která z principu předpřipravený design nemají

- Sada skriptů **_strict.sh** – tato sada funguje k provedení analýzy jednotlivých vzorků (nejčastěji WGS), kdy nechceme zpracovávat data z celé knihovny, uživatel pak musí pro každý běh skriptu provést úpravy a zadat absolutní cesty k souborům, využití je hlavně u WGS dat, kdy spouštíme celou analýzu vzorek po vzorku, aby nedošlo v případě vyčerpání volného místa k zastavení celé analýzy nebo v případě reanalýzy jednotlivých vzorků (či změně nastavení).

Díky systému je dosaženo snadné orientace i ve velkém množství dat a díky indexaci disků na pracovních stanicích je snadné dohledat dle čísla vzorku všechny potřebné soubory. Všechna data jsou pravidelně zálohována minimálně na dvou místech pro ochranu před ztrátou dat. K tomu slouží NAS úložiště obsahující 10 disků s kapacitou celkem 80 TB a menší NAS úložiště pro zálohu WGS se 4 disky a kapacitou celkem 24 TB.

Rovněž jsou ještě zajištěny další zálohy hrubých FASTQ dat, která mají velikost do 4TB a ze kterých je možné analýzu provézt znovu.

5 Diskuse

S rozvojem genomiky a zejména masivně paralelního sekvenování se možnosti diagnostiky dědičných onemocnění značně rozšířily a zlepšily. Je možné objasňovat i vzácnější a velmi vzácné příčiny dědičných onemocnění a objevovat stále nové příčiny a typy onemocnění. I přesto se nám podaří objasnit příčinu onemocnění sekvenováním panelem genů u přibližně čtvrtiny pacientů. [Staněk et al. 2018] Následným WES dokážeme objasnit příčinu onemocnění u přibližně třetiny zbylých pacientů u kterých sekvenování panelem genů příčinu nemoci neodhalilo. [Eldomery et al. 2017] Pokud do analýzy zařadíme WGS, získáváme „další vrstvu“ informací v podobě nekódujících oblastí a možnosti hledání dalších mechanismů vzniku onemocnění. Přesto je ale stále nemožné u řady pacient nalézt příčinu onemocnění, tedy kauzální variantu.

Při hledání kauzality využíváme populační databáze, naší interní „in-house“ databázi, predikční nástroje, které nám umožňují anotaci a evaluaci variant do té míry, že jsme schopní přesvědčivě určit původce onemocnění. Jako dalším efektivním nástrojem se ukazují analýzy na základě propojení fenotypu s onemocněním, využívané pro vytvoření virtuálních panelů (podrobněji v 4.5), při manuálním filtrování pomocí databáze OMIM [Hamosh et al. 2005] nebo při pokročilé prioritizaci variant nástrojem Exomiser (podrobněji v 3.2.6).

5.1 Klasifikace variant - kauzalita a negativní výsledek

Přestože MPS metody dokážou objasnit více než 50 % všech případů, stále zbývá velká část neobjasněných. Nalezení přesvědčivě kauzální varianty je několika krokový proces, který začíná objevením varianty ve VCF. S variantou, která se jeví jako patogenní, je ale nutné dále pracovat, abychom mohli výsledek potvrdit. Následuje tedy krok sekvenace pomocí Sangerova sekvenování jak pacienta, tak rodičů, kdy ověřujeme nepřítomnost varianty u zdravých rodičů (popř. přítomnost heterozygotní varianty). U MPS panelu genů nám takový výsledek, pokud se jedná o patogenní (dle ACMG), dříve popsanou variantu, vede často k uzavření případu - nalezli jsme variantu v genu, který má dříve prokázanou spojitost s onemocněním a jsme si jistí její přítomností u probanda.

U celoexomového sekvenování je situace jiná, protože často nalezneme vzácnou variantu v genu, který ještě nemá asociaci s onemocněním. Přestože vidíme, že varianta spadá do místa (např. proteinové domény), které by mohlo být kauzální pro fenotyp, nemůžeme případ uzavřít aniž bychom si byli jistí výsledkem. Obvykle nezbývá než vyčkat s interpretací dokud nejsou přesvědřivé důkazy o asociaci mezi variantou a fenotypem. Tento případ nastal v případě varianty v genu *UBTF* [Sed-

láčková et al. 2018], kdy jsme nalezli variantu, kterou jsme pokládali za kauzální, ale nemohli jsme naše podezření u jediného pacienta dále potvrdit. K potvrzení došlo až po publikování studie [Edvardson et al. 2017], která variantu v genu UBTF popisovala se shodným fenotypem.

Dalším případem, kdy od samotného sekvenování vedla dlouhá cesta k objevení nové příčiny onemocnění je případ české rodiny s CMT, kdy publikace nové příčiny CMT2 proběhla po 14 letech po první sekvenaci pacienta z rodiny. Celkem pak bylo vyšetřeno u české rodiny 18 pacientů (z toho 11 postižených). Díky mezinárodní spolupráci v rámci platformy Genesis [Gonzalez et al. 2015], která agreguje WES vzorky pacientů s CMT bylo možné nalézt dalších šest rodin s variantami v genu *ATP1A1*, který do té doby nebyl asociován s CMT. V rámci mezinárodní spolupráce, tedy bylo možné identifikovat gen *ATP1A1* jako příčinu CMT a tím uzavřít případ této rodiny. [Lassuthova et al. 2018]

Pokud ale analýza WES dat nenajde přesvědčivě kauzální variantu, nelze z toho usuzovat, že se taková varianta ve vzorku nenachází. Výsledek nelze označit za negativní. Takový výsledek neznamená, že pacient nemá žádnou variantu způsobující onemocnění, ale že jsme v současnosti takovou variantu neidentifikovali. Data z WES u kterých jsme nenalezli kauzální variantu jsou reanalyzována vždy jednou za půl roku, kdy je provedena nová anotace a u variant, které se zdály v předchozích analýzách „podezřelé“ je provedena rešerše na publikace, které by podobnou příčinou onemocnění popisovaly.

Další možností, je aktivní hledání jiného pacienta se shodným genotypem a fenotypem. Jak jsme ukázali výše, mohou být tyto iniciativy velmi užitečné. V praxi se využívá zejména portál GeneMatcher[1], který funguje jako síť výzkumných týmů. Pomocí webového rozhraní lze zadat gen či variantu, pro kterou hledáme dalšího pacienta (rodinu) a můžeme vidět, jestli už varianta byla identifikována jiným výzkumným týmem.

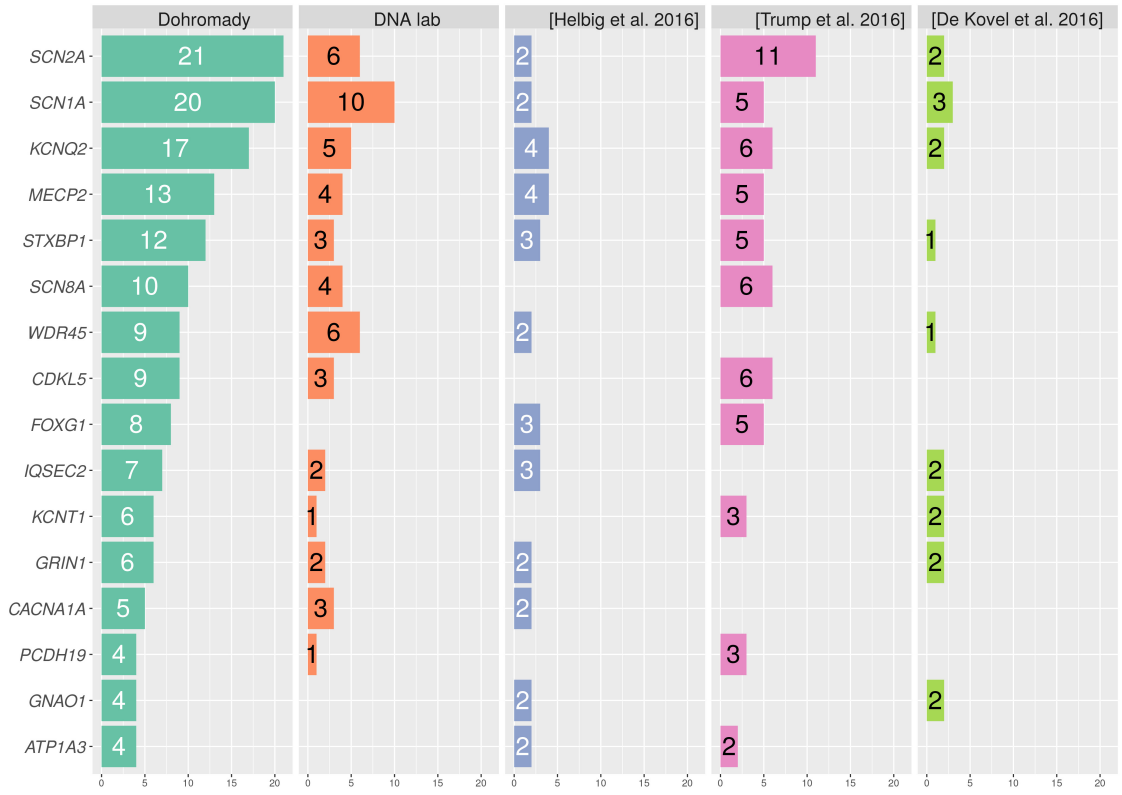
V případě, kdy nenalezneme žádný další tým, který by měl korespondující data, je další možností funkční studie na modelovém organismu. V současné době funguje v rámci Evropské unie iniciativa SolveRD (URL[2]), která slouží k propojování výzkumných týmů s týmy provádějící funkční studie. Díky tomuto programu je tedy možné navázat spolupráci a pomocí funkční studie ověřit kauzalitu varianty. [Boycott et al. 2017]

5.2 MPS panelem genů u pacientů s EE, srovnání

Pro srovnání výsledků našeho MPS panelu 112 genů asociovaných s EE jsme použili tři studie s podobným zaměřením: [Helbig et al. 2016],[Trump et al. 2016],[Kovel et al. 2016]. Výsledky v publikaci [Helbig et al. 2016] korespondují s našimi výsledky, objasnitelnost se pohybovala kolem jedné čtvrtiny až jedné třetiny případů, většina variant byla *de novo* vzniklých a v publikaci byla rovněž determinována souvislost mezi věkem nástupu onemocnění s objasnitelností - jak uvádíme ve výsledcích, objasnitelnost je až dvakrát vyšší, pokud se epileptické záchvaty objeví během prvních

[1] https://www.genematcher.org/ [online: 16.10.2019]
[2] http://solve-rd.eu/rdmm-europe/ [online: 16.10.2019]

čtyř týdnů života pacienta. Naše výsledky jsou též podobné s publikací [Trump et al. 2016], která ale měla nižší objasnitelnost (18 % oproti 28 % u našeho panelu). V poslední studii [Kovel et al. 2016], byl zvolen odlišný přístup, kdy do panelu genů bylo zařazeno 351 genů (tedy přibližně třikrát více). Takový návrh panelu se ale neukázal jako příliš efektivní, s 8% objasnitelností oproti očekávaným 30 %. Návrh většího panelu může pomoci objasnit případy, kdy onemocnění způsobí varianta v nepříliš známém genu (pro onemocnění). Takový panel je ale mnohem dražší, poskytuje nižší pokrytí, protože je více vyčerpána kapacita panelu. Z tohoto důvodu se nám osvědčil postup, kdy prvním krokem je sekvenování panelem genů a WES až ve druhém kroku.



	DNA lab	[Helbig et al. 2016]	[Trump et al. 2016]	[De Kovel et al. 2016]
Počet pacientů	257	314	400	360
Objasněných případů	28%	38%	18%	8%

Genů v panelu	112	Virtuální panel	46	351
Gen s nejvíce P/LP variantami	*SCN1A*	*KCNQ2, MECP2*	*SCN2A*	*SCN1A*
Počet genů s P/LP variantou	33	79	21	20
Počet P/LP variant	76	108	71	29

Obrázek 5.1: Srovnání výsledků MPS panelu genů u pacientů s EE v DNA laboratořemi s dalšími publikovanými studiemi[Helbig et al. 2016],[Trump et al. 2016],[Kovel et al. 2016]

5.3 MPS panelem genů vs WES

Sekvenování panelem genů je pro nás prvním krokem při hledání příčíny onemocnění pomocí MPS. Výhodou panelu je jeho snadná úprava dle našich požadavků. Toho využíváme jak u EE panelu genů, tak u panelu genů spojených s dědičnou neuropatií. Díky tomu je možné reagovat na nové poznatky a v případě identifikace nového genu takový gen do panelu přidat. Naopak geny, u kterých jsme žádné varianty nenalezli je možné nahradit. Například u panelu genů spojených s dědičnými neuropatiemi jsme navrhli 6 různých verzí, začínali jsme s 59 geny, a přes 64, 69, 78 jsme došli až k 103 genům. Udržovat panel aktuální je důležité, přídáním genů *ATP7A*, *COX6A1* nebo *DYNC1H1* se nám povedlo identifikovat kauzální varianty pro dědičné neuropatie, které bychom v předchozích případech neidentifikovali, pokud bychom nepřistoupili k WES.

Další otázkou, kterou je potřeba objasnit je, proč vlastně využívat panel genů a nepřistoupit rovnou k WES s aplikací virtuálního panelu na vybrané geny. Panel genů nám umožňuje získat informaci o sekvenci pacientovy DNA řádově vyšší hloubkou čtení proti WES a správně navržený design panelu poskytne téměř 100% pokrytí námi sledovaných oblastí. Pro příklad, u našeho panelu, který pokrývá 482 648 bází, zůstává průměrně 48 nukleotidů nepokryto. U celoexomového sekvenování dosahujeme u dobře nastaveného designu k pokrytí 98%. To znamená, že z 50 390 061 bází je nepokrytých 1 007 801, které mohou spadat do oblasti našeho zájmu. Naopak pokud bychom chtěli místo MPS panelu genů provádět Sangerovo sekvenování, tak by to znamenalo provést (v ideálním případě, kdy na každý exon provádíme jen jednu sekvenaci) tisíce sekvenací - i pokud bychom vybrali pouze geny, u kterých jsme nalezli P nebo LP variantu (33 genů) - jednalo by se o 8 481 Sangerových sekvenování.

Sekvenování panelem genů preferujeme jako první krok při MPS i z etického hlediska, kdy se řídíme doporučením ESHG [El et al. 2013]. Dle doporučení bychom se měli zaměřovat pouze na oblast, která byla dříve asociována s hledaným fenotypem. Nikoli zvolit „hunt for everything“ přístup, tedy za každou cenu nalézt zajímavou variantu asociovanou s naším zájmem.

Pokud vyhodnocujeme data z MPS, setkáváme se s desítkami tisíc variant, které jsou asociované s tisíci fenotypy. Pro příklad můžeme uvést hypervariabilní geny, se kterými jsme pracovali v rámci in-house databáze variant – u některých genů máme stovky asociovaných fenotypů. Jde o geny, ve kterých jsou varianty velmi časté, z takového výsledku pak nelze usuzovat žádné závěry.

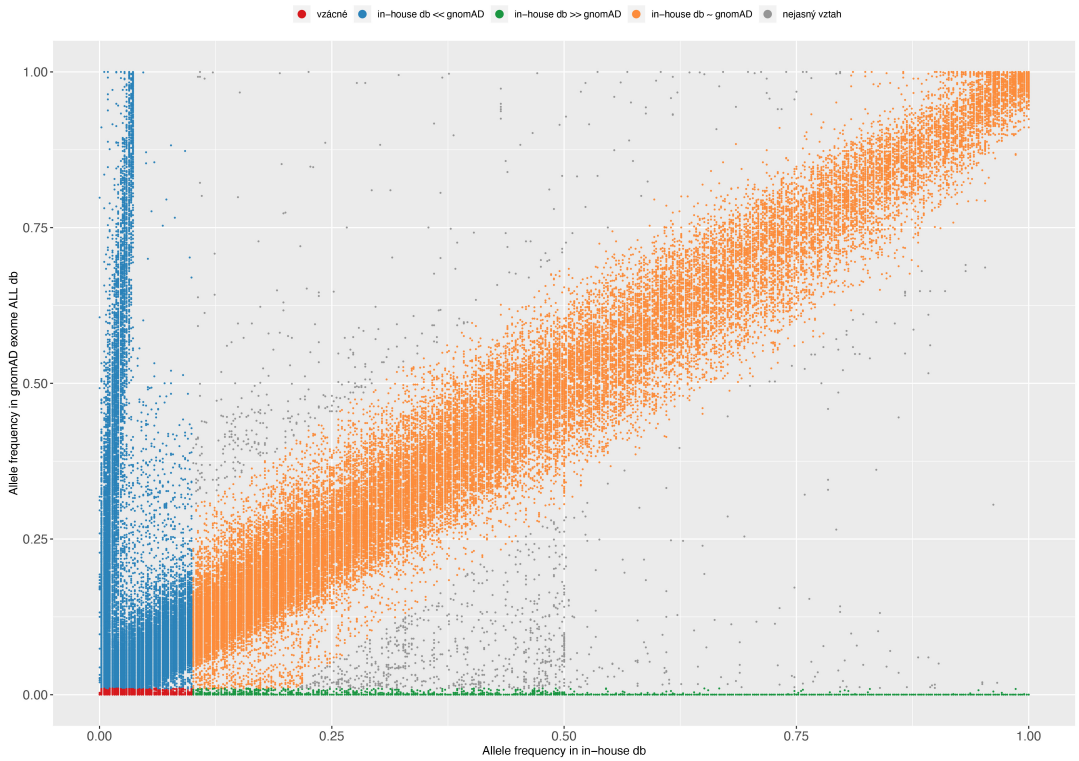
Naopak, pokud u WES dat nalezneme patogenní variantu způsobující závažné onemocnění v pozdějším věku, může být taková informace pro pacienta velmi důležitá. Je ale nutné vždy preferovat a respektovat volbu pacienta, jestli chce takový výsledek znát. S takovým přístupem dokážeme zajistit maximální bezpečnost pro pacienta a zabezpečení jeho dat.

5.4 In-house databáze DNA variant z WES

Pro další analýzu jsme varianty rozdělili do několika tříd, dle vztahu mezi frekvencí v in-house databázi a gnomAD exome ALL databází Obr. 5.2. Třídy jsme zvolili dle následujícího klíče:

Třída	Zkratka	In-house frekvence	GnomAD frekvence	Barva	Počet
Vzácné	**V**	<0,1	<0,01	červená	171 840
Vzácné lokálně	**VL**	<0,1	>0,01	modrá	56 697
Běžné lokálně	**BL**	>0,1	<0,01	zelená	14 782
Korespondující	**K**	$inhousedbAF = gnomAD \pm 0.2$		oranžová	55 453
Nejasný vztah	**NV**	ostatní případy, nelze odvodit vztah		šedá	1 339

Tabulka 5.1: Třídy variant dle vztahu mezi frekvencemi in-house databáze a gnomAD exome All databáze



Bodově vyznačené varianty: červeně skupina vzácných, potenciálně patogenních variant, zeleně varianty s vysokým výskytem v naší subpopulaci a nízkým v gnomAD, modře varianty s vysokým výskytem v gnomAD a nízkým v naší subpopulaci, oranžově varianty s frekvencí korelující mezi databázemi, a šedě spektrum variant u kterých nelze jednoznačně určit vztah

Obrázek 5.2: Srovnání frekvencí variant v in-house databázi s databází gnomAD exome ALL

Nejvíce variant jsme nalezli ve třídách V a BL, tedy s hodnotami frekvencí v databázi gnomAD menší než jedno procento. Počet variant zařazených do těchto skupin je ovlivněn počtem variant s nulovou nebo prázdnou gnomAD frekvencí. U interpretace těchto hodnot musíme být obezřetní, prázdná hodnota v gnomAD databázi v tomto případě totiž nemá shodný význam jako nulová hodnota. Ačkoli při vyhodnocování NGS dat nám nepřítomnost varianty v populačních databázích hovoří pro prioritizaci takové varianty, je v tomto případě nutné se zamyslet nad původem nulové hodnoty. Nepřítomnost hodnoty frekvence má nejčastěji dvě příčiny. První příčinou je, že pozice varianty nebyla pokrytá sekvenační knihovnou, ale v našem případě již pokrytá byla - v takovém případě dostáváme nepřesnou informaci, varianta vypadá dle frekvencí „vzácně“, ale i tak může být v populaci velmi častá, oproti tomu druhou možností je situace, kdy výskyt varianty byl opravdu negativní, v kohortě probandů gnomAD databáze byla přítomná pouze referenční alela, v tom případě se jedná o „false positive“ variantu.

5.4.1 Rozdělení variant do skupin

Varianty byly rozděleny do tříd dle významu. Cílem tohoto projektu nebylo hledání patogenních variant u jednotlivých vzorků, ale hledání souborů variant se společnými charakteristikami - proto došlo k rozdělení do tříd. Vytvoření tříd nám přinese další možnost filtrování variant, například varianty, které se vyskytují v naší populaci častěji dokážeme odfiltrovat a tím snížit počet variant pro manuální filtrování.

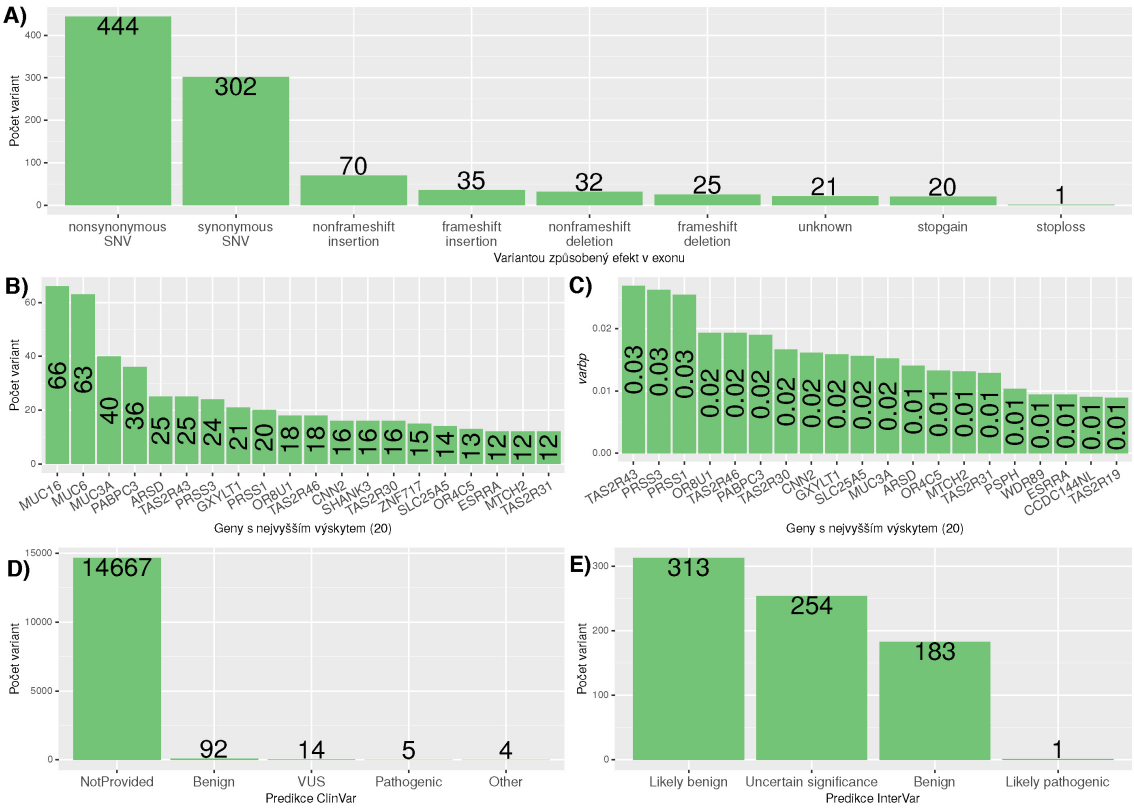
Třída V obsahuje varianty, které jsou vzácné dle frekvencí v obou populacích. Vyskytují se s frekvencí menší než 0,01 v gnomAD databázi a 0,1 v in-house databázi, takové varianty mohou mít patogenní potenciál, jsou to právě varianty, na které cílíme při analýze dat a hledání příčin onemocnění.

Ve **třídě VL** jsou varianty, které jsou vzácné v naší populaci, ale jsou časté v gnomAD databázi. Tato skupina variant nemá pro určování příčin onemocnění velký význam, protože jedním z prvních parametrů při prirotizaci variant je filtrování na frekvenci v gnomAD. Skupina bude rovněž nejvíce zatížena chybou malého počtu vzorků - nelze zcela přesně srovnat frekvenci mezi 222 pacienty v naší databázi a skupinou více než 120 tisíc probandů v gnomAD.

Ve **třídě K** jsou varianty, jejichž frekvence jsou přibližně stejné (rozdíl do 0,2 ve frekvenci) a zároveň nejsou ve třídě P. Tato skupina variant není pro další analýzu příliš signifikantní, všechny tyto varianty totiž mají obě frekvence vysoké, a tudíž by byly v rámci hledání variant deprioritizovány. Varianty v této skupině odpovídají evropské populaci. Podobný závěr lze udělat u skupiny NV, kdy opět jsou frekvence vysoké, ale navzájem nekorespondující, jedná se ale o varianty, které nejsou ve třídách BL ani VL, tudíž jejich vztah mezi populacemi nelze určit.

Třída BL je pro tento projekt nejzajímavější, protože obsahuje varianty, které jsou lokálně časté, ale přitom je jejich frekvence v gnomAD daleko nižší. Tyto varianty se v rámci naší analýzy jeví jako specifické pro naší populaci, odhalení takových variant nám napomáhá k identifikaci možných false positive variant - tyto varianty bychom prioritizovali při vyhodnocování, ale jelikož víme, že jejich frekvence je vysoká, lze předpokládat, že nebudou příčinou onemocnění.

Ve třídě BL se nachází 14 782 detekovaných variant, tyto varianty se nachází v 289 genech. Pro přehled uvádíme analýzu skupiny BL dle genů, predikce nástrojem ClinVar a Intervar (který hodnotí variant dle ACMG specifikací poskytnutých ve VCF) na obrázku Obr. 5.3.



A) počty variant dle exonových funkcí
B) geny s nejvíce vyvolanými variantami v rámci třídy
C) geny s nejvyšším poměrem variant na délku genu v rámci třídy
D) počty variant pro jednotlivá hodnocení ClinVarem
E) počty variant pro jednotlivá hodnocení InterVarem (ACMG automatické)

Obrázek 5.3: Přehled variant ve třídě BL

5.4.1.1 Asociace variant s fenotypem dle HPO

Pro asociační analýzu jsme vybrali geny, u kterých bylo nalezeno nejvíce variant, a které měly nejvyšší hodnotu *varbp*. U genů s nejvyšším počtem variant jsme pro analýzu odstranili geny, které měly velmi nízkou hodnotu *varbp*, to totiž značilo, že se jedná o geny, které mají vysoký počet variant, ale zároveň jsou velmi dlouhé. Kritérium pro odstranění takových genů byl 95. percentil (tedy hodnota 0,01) - pokud měl gen *varbp* nižší, nepokračovali jsme s takovým genem v další analýze. Pro asociační analýzu v tomto případě bylo využito 10 genů, které společně s dvaceti geny s nejvyšším *varbp* tvoří skupinu „hypervariabilních genů".

Stejným způsobem jsme předfiltrovali geny z třídy BL, zde byla hodnota 95. percentilu 0,013, pro asociační analýzu bylo využito 14 genů doplněných o dvacet genů s nejvyšším *varbp* v BL třídě, tvořící dohromady skupinu „BL hypervariabilních genů".

Postup výběru genů uvádíme v tabulce Tab. 5.2.

(A) Všechny varianty

Gen	Počet	*varbp* (95. percentil 0.01)
TTN	474	0,004389864
MUC16	436	0,010017462
MUC3A	311	0,118206005
MUC6	269	0,036748634
OBSCN	185	0,006910205
AHNAK2	163	0,009374281
MUC5B	142	0,008209042
CELSR1	130	0,014372582
MUC17	127	0,009419967
PLEC	126	0,008964781
ZNF717	119	0,043351548
LAMA5	118	0,010642136
ZNF469	111	0,009357613
FCGBP	108	0,006659267
SYNE1	108	0,004091839
PKDREJ	104	0,015380065
OR8U1	99	0,106451613
PIEZO1	98	0,012952683
TRIOBP	98	0,013806706
DNAH17	97	0,007207609

(B) BL třída

Gene	Počet	*varbp* (95. percentil 0.013)
MUC16	66	0,001516405
MUC6	63	0,008606557
MUC3A	40	0,015203345
PABPC3	36	0,018987342
ARSD	25	0,014029181
TAS2R43	25	0,02688172
PRSS3	24	0,026229508
GXYLT1	21	0,015873016
PRSS1	20	0,025445293
OR8U1	18	0,019354839
TAS2R46	18	0,019354839
CNN2	16	0,01611279
SHANK3	16	0,003051106
TAS2R30	16	0,016666667
ZNF717	15	0,005464481
SLC25A5	14	0,015607581
OR4C5	13	0,013251784
ESRRA	12	0,009433962
MTCH2	12	0,013157895
TAS2R31	12	0,012903226

A) dvacet genů s nejvyšším počtem variant v celé databázi, červeně geny, které byly vyloučeny pro další analýzu, protože vysoký počet variant je kompenzován délkou genu, pro další analýzu využito 10 genů
B) dvacet genů s nejvyšším počtem variant ve třídě BL, červeně geny, které byly vyloučeny pro další analýzu, protože vysoký počet variant je kompenzován délkou genu, pro další analýzu využito 14 genů

Tabulka 5.2: Tabulka genů s nejvíce variantami, výběr pro asociační analýzu

U dvou skupin genů – hypervariabilních genů a BL hypervariabilních genů jsme provedli asociační analýzu – seznam genů jsme anotovali pomocí HPO termínů. V tabulkách níže (Tab. 5.3) je možné vidět počty asociovaných fenotypů s geny.

Z tabulky je možné vyčíst, že většina uvedených genů je asociována s desítkami fenotypů nebo naopak s žádným fenotypem. Výjimkou je ale gen *TRIOBP*, který má asociaci pouze se třemi HPO termíny - Severe sensorineura hearing impairment, AR inheritance a Infantile onset - tyto fenotypy nelze ignorovat, protože v naší kohortě

se nachází pacienti s hluchotou. Proto jsme provedli analýzu všech variant (98), které se nachází v kódující oblasti. Tyto varianty jsme profiltrovali a vyhodnotili, jestli mohou být kauzální. Nakonec jsme zjistili, že většina variant se nenacházela u pacientů s diagnostikovanou hluchotou. Výjimkou byly tři varianty, které ale byly detekovány u pacientů, kteří již mají známou příčinu dědičné hluchoty. Jedna z variant v genu *TRIOBP* měla závažné predikce, jednalo se o stop variantu na pozici chr22:38131357. Tato varianta ale byla detekována u pacienta s CMT, který nemá se sluchem problémy.

Seznam všech asociovaných HPO termínů je uvedený v Příloze I (z důvodu velikosti).

(A) **Všechny varianty**

Geny dle počtu variant	
Gen	Počet nalezených HPO termínů
MUC16	0
MUC3A	0
MUC6	0
CELSR1	0
SSPO	0
ZNF717	0
LAMA5	0
PKD1L2	0
PKDREJ	0
OR8U1	0
TRIOBP	**3**

Geny dle *varbp*	
Gen	Počet nalezených HPO termínů
MUC3A	0
OR8U1	0
MYO15B	0
OR4C5	0
CHCHD10	**83**
TAS2R43	0
PRSS3	0
CELA1	0
HLA-DRB5	0
OR4A16	0
KCNJ12	0
PRSS1	**21**
HLA-DQB1	**22**
TPSAB1	0
OR4C3	0
FAM71D	0
MUC5AC	0
HNRNPCL1	0
DEFB104A	0
HBB	**89**

(B) **BL třída**

Geny dle počtu variant	
Gen	Počet nalezených HPO termínů
MUC3A	0
PABPC3	0
ARSD	0
TAS2R43	0
PRSS3	0
GXYLT1	0
PRSS1	**21**
OR8U1	0
TAS2R46	0
CNN2	0
TAS2R30	0
SLC25A5	0
OR4C5	0
MTCH2	0

Geny dle *varbp*	
Gen	Počet nalezených HPO termínů
TAS2R43	0
PRSS3	0
PRSS1	**21**
OR8U1	0
TAS2R46	0
PABPC3	0
TAS2R30	0
CNN2	0
GXYLT1	0
SLC25A5	0
MUC3A	0
ARSD	0
OR4C5	0
MTCH2	0
TAS2R31	0
PSPH	**24**
WDR89	0
ESRRA	0
CCDC144NL	0
TAS2R19	0

A) uvedené počty HPO termínů asociovaných s hypervariabiními geny, nenulové žlutě

B) uvedené počty HPO termínů asociovaných s hypervariabiními geny ze skupiny BL, nenulové žlutě

Tabulka 5.3: Tabulka asociací genů s nejvyšším počtem variant s HPO termíny

5.4.1.2 Použití „in-house“ databáze

Výslednou databázi můžeme v rámci vyhodnocování NGS využít několika způsoby. První možností je využití alelické frekvence pro anotaci všech variant. V tomto případě je důležité mít informace o skupině variant ze třídy BL - tedy o variantách, které mají velmi nízkou frekvenci v databázi gnomAD, ale naopak v naší databázi mají frekvenci vysokou. Takové varianty, ačkoli by na první pohled mohly vypadat podezřele, můžeme v rámci analýzy deprioritizovat. Během filtrování variant rovněž můžeme použít výsledky analýzy dle genů, kdy geny, ve kterých jsou varianty velmi časté (Obr. 4.11) můžeme opět deprioritizovat.

Kromě toho lze databázi využít na principu „exome-wide association study“, kdy dokážeme identifikovat skupiny variant, které jsou specifické pro naší kohortu. Tato informace může přinést cenná data nejen pro naše analýzy, ale i pro další vědecké skupiny v rámci regionu.

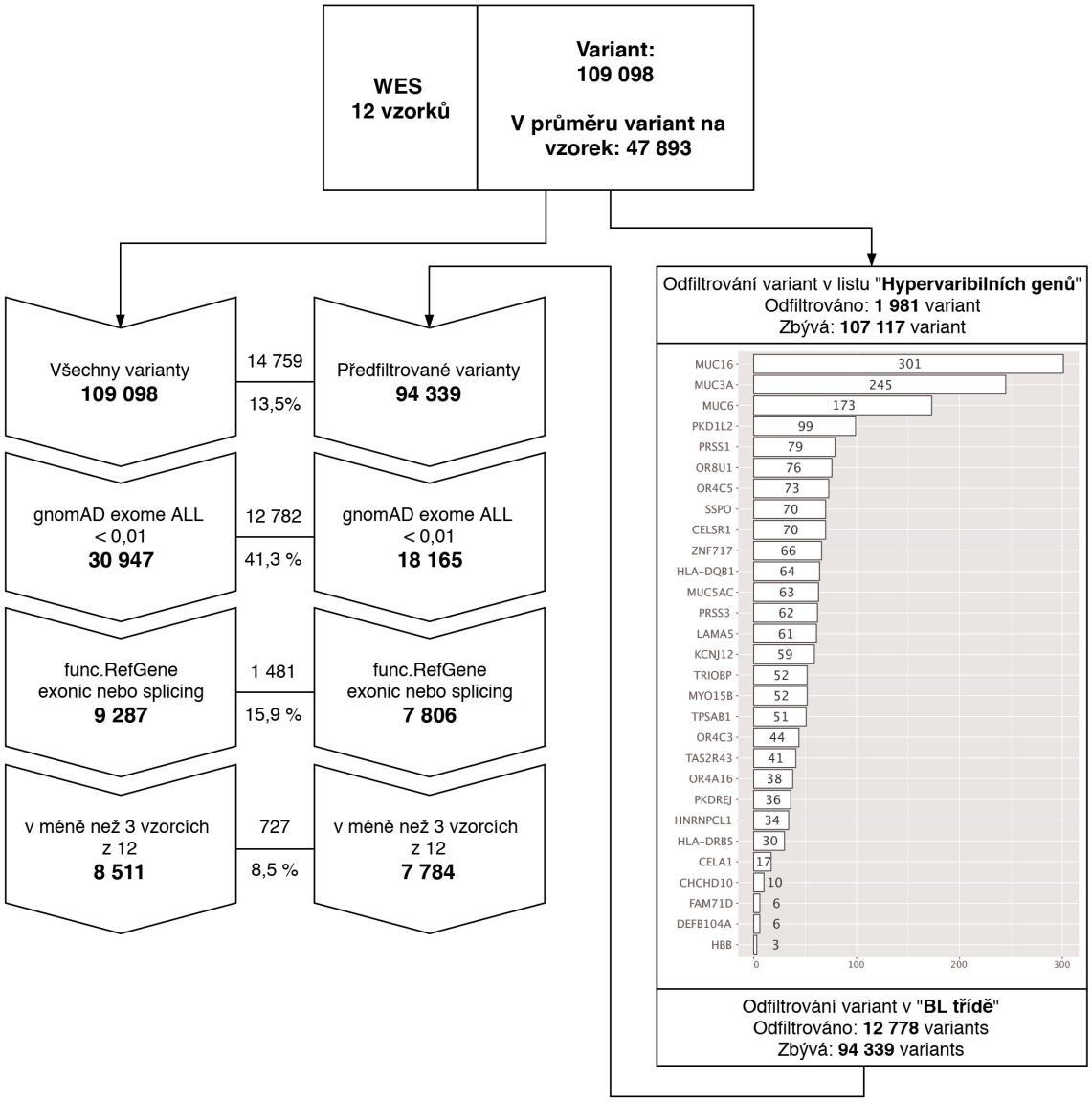
Pro demonstraci praktického využití databáze jsme provedli manuální filtrování dvanácti nových WES vzorků, které nebyly zahrnuty do in-house databáze. Celkem bylo vyvoláno 109 098 variant (průměrně 47 893 variant na vzorek).

Nejprve jsme provedli předfiltrování:

- byly odfiltrovány varianty, které se nacházejí ve skupině hypervariabilních genů - počet variant se snížil o 1 981 na 107 117 variant.
- byly odfiltrovány varianty, které byly v in-house databázi zařazeny do třídy BL - počet variant se snížil o 12 778 na 94 339
- bylo provedeno manuální filtrování pro porovnání

Předfiltrování souboru nám pomohlo deprioritizovat 14 759 variant (13,5 %). V níže uvedeném schématu (Obr. 5.4) uvádíme porovnání srovnání i v dalších krocích manuálního filtrování – deprioritizace variant s vysokou frekvencí v gnomAD, ponechání pouze exonových a splicing variant a v posledním kroku odstranění variant, které byly ve více než 3 vzorcích z 12 analyzovaných. Výsledkem bylo 7 784 variant u předfiltrovaného datasetu a 8 511 variant u datasetu bez předfiltru.

Druhou možnou aplikací je vypočítané alelické frekvence v in-house databázi použít jako další kritérium manuálního filtrování – pokud bychom z datasetu bez předfiltru odstranili varianty s alelickou frekvencí v in-house databázi vyšší než 0,025, snížili bychom počet variant na 7 325.



Obrázek 5.4: Porovnání procesu manuálního filtrování u předfiltrovaného datasetu a datasetu bez předfiltru

5.5 Databáze prot2HG

Z výsledků anotační analýzy je patrné, že byly nalezeny signifikantní vztahy mezi umístěním variant v doméně a jejich populační frekvencí. Vzácné varianty se častěji nachází v oblasti proteinových domén. To odpovídá předpokladu, že tyto oblasti jsou konzvervovanější než jiné části genomu. Můžeme tedy předpokládat menší variabilitu v oblastech proteinových domén.

Pro tuto analýzu jsme pokládali varianty jako vzácné, pokud jejich frekvence v gnomAD databázi byla nižší než 1 %. Během zpracování projektu bylo uvažováno o zvýšení prahu na ultra vzácné varianty („ultra rare“) s frekvencí nižší než 0,01 %, výsledky se ale lišily jen nepatrně (proto nebyly prezentovány).

V případě patogenních variant z ClinVar databáze je také vidět možná souvislost mezi patogenicitou a lokalizací uvnitř domény. Tyto výsledky můžou být ovlivněny systematickou chybou. Databáze ClinVar nemá žádnou centrální autoritu, která by kontrolovala správnost dat, může tedy docházet k „přeceňování“ hodnocení (uživatel vloží variantu a označí jí jako patogenní, přestože nemá patogenicitu ověřenou). Jako alternativní zdroj by bylo vhodné použít databázi HGMD Professional (URL[3]), kde autorita pro posuzování variant funguje. Tento zdroj ale není t.č. volně dostupný ke stažení a inkorporaci do analýzy.

5.5.1 Použití prot2HG databáze v praxi

Tento příklad byl prezentován v rámci konference ASHG, poster uveden příloze G.

Pro demonstraci jsme využili datasetu 60 dříve potvrzených kauzálních (patogenních a pravděpodobně patogenních) variant pacientů s EE. Tyto varianty jsme anotovali databází prot2HG k určení, jestli se nachází v oblasti proteinových domén.

Z 60 variant byla u 43 (71 %) potvrzena pozitivní anotace na proteinovou doménu. Z hlediska efektu domén se jednalo nejčastěji o regiony iontového transportu (ion transport regions) – 14 variant, u 6 variant o transmembránové oblasti (transmembrane region) a u 5 variant o iontové kanály (ion-channel transmembrane region). Většina variant byla propagována do více než jedné domény (43 variant do celkem 93 domén).

Na tomto příkladu je patrné, že anotace databází prot2HG nám může pomoci s hledáním kauzálních variant u NGS dat. Pomocí detekce proteinových domén a jejich funkcí, by mohlo dojít i k identifikování nových, dosud nepopsaných genů a určení jejich asociací s onemocněním (například u EE).

5.6 Databáze variant spojených s CMT - komentář k projektu

Tento projekt vznikl v rámci mezinárodní spolupráce s Hussman Institute for Human Genomics v Miami USA, kde jsem byl na několika měsíční stáži. Podílel jsem se na návrhu databáze a její implementaci. Kdy jsme začínali se shromažďováním

[3] http://www.hgmd.cf.ac.uk/ [online 16.9.2019]

dat a jejich přípravou před vložením do databáze, dalším krokem byl návrh a otestování uživatelského rozhraní, kdy jsem dostal na starost proces manuálního přidávání variant. Posledním krokem, byla již dříve zmíněná pilotní fáze databáze proteinových domén, kdy pro prvních 82 CMT genů, které byly přidány, jsme potřebovali získat informace o jejich doménách a prezentovat je v rámci uživatelského rozhraní. Data byla získána z databáze NCBI, námi navrženým skriptem a poté vložena do databáze. Výsledek projektu byl publikován v publikaci [Saghira et al. 2018].

5.7 Zavedení nových bioinformatických metod do DNA laboratoře

Jedním z cílů této práce bylo zavedení nových metod, umožňujících zvýšit objasnitelnost případů v DNA laboratoři. Pro vyhodnocování dat z MPS bylo v posledních letech publikováno mnoho nástrojů či databází, které nám pomáhají dosahovat lepších výsledků. V praxi je ale právě otestování nového nástroje a jeho případné zařazení do bioinformatické workflow klíčovou fází celého procesu.

Prvním krokem je výběr správného nástroje, kdy po přečtení publikací týkajících se tématu vybíráme dle vlastního uvážení vhodný nástroj. Při zavádění CNV analýzy jsme nejprve zvolili nástroj metaSV, který dle publikace měl být pro naše použití u WES dat vhodný. Ale jak praktické testování nástroje ukázalo, že aplikace není vhodná pro naše použití – určení bylo spíše pro detekování strukturálních variant v rámci WGS dat. Proto jsme se rozhodli nalézt jiný nástroj. Je vhodné po určité době neúspěšného testování nástroje zvolit jinou alternativu, v tomto případě CNV nástroj z knihovny GATK4.

V některých případech je nutné řešit problémy i se samotnými vývojáři. To byl případ nástroje pro detekci *de novo* variant ve WES - DenovoGear. Po aktualizaci nástroje na nejnovější verzi docházelo k problému s pamětí, kdy nástroj spotřeboval všechnu pamět počítače a analýza poté neproběhla. Po 6 měsících řešení s podporou byla chyba opravena a nástroj byl znovu funkční. Tento případ je tedy důkazem toho, že ne vždy musí být chyba v analýze u uživatele.

Základním postupem pro vyhodnocování variant z MPS dat je manuální filtrování. Anotované VCF je filtrováno dle informací z populačních databází (gnomAD, ExAC), predikčních nástrojů nebo dle efektu varianty. Tento postup nadále využíváme při filtrování dat z panelu genů, kdy není potřeba dalších, pokročilých metod, neboť se pohybujeme v rámci genů asociovaných s onemocněním.

Zcela odlišný postup jsme ale etablovali pro vyhodnocování WES dat. Zde se snažíme každý vzorek vyhodnocovat zvlášť a nevyužíváme pouze metod manuálního filtrování, naopak tento postup považujeme za sekundární. Prvním krokem při zpracování WES dat je zpracování pomocí bioinformatické pipeline, kdy získáme VCF soubor, který pak vstupuje do dalších analýz. Prioritní je tzv. Trio analýza, kdy je celoexomové sekvenování provedeno jak pacientovi, tak rodičům. Při takovém postupu pak hledáme *de novo* varianty - počet takových variant se pohybuje v řádu jednotek na pacienta. Vyhodnocování Trio modelu nelze zpracovat pouze porovnáním vyvolaných variant u pacienta a jeho rodičů, je nutné aplikovat pravděpodobnostní model, který zahrne i možnost nevyvolání varianty – z tohoto důvodu využíváme nástroj

DenovoGear, používající jako vstup BAM soubor - tedy celou sekvenci před variant callingem. Pokud nenalezneme v rámci Trio analýzy kauzální variantu, nebo nemáme vzorky rodičů k dispozici, volíme jiný postup pracující pouze s daty pacienta. Díky vhodné kooperaci vyšetřujících lékařů máme k dispozici kartu pacienta s dobře popsaným fenotypem, fenotyp pak můžeme zadat pomocí HPO termínů do nástroje Exomiser, který hledá varianty v genech asociovaných s námi popsaným fenotypem. Tyto postupy jsou dále doplněné manuálním filtrováním alespoň ze dvou zdrojů - GATK Pipeline a jeden z komerčních nástrojů (NextGene, SureCall).

K vyhodnocování WES a WGS dat ale využíváme i další zdroje informací, jedním z nich je in-house databáze variant, která nám poskytuje informaci o relativní frekvenci varianty v rámci naší kohorty. To je důležité pro odstranění nejen false positive varianty způsobených chybou sekvenování, ale i výskytem variant, které jsou časté v rámci naší populace.

V případě, že ani tato analýza neodhalí kandidátní variantu přistupujeme k detekci CNV ve vzorku pomocí GATK4 pipeline pro hledání germinálních variant, kde opět využíváme HPO termíny pomocí našeho nástroje pro vytváření virtuálních panelů.

V současnosti tedy máme zpracovaný postup, který nám umožňuje produkovat výsledky srovnatelné s celosvětovým měřítkem co se týče procenta objasnitelnosti pacientů. Je ale nutné podotknout, že jde o kontinuální proces, který je nutné neustále aktualizovat a přizpůsobovat se novým trendům. Jako příklad lze uvést integraci nového nástroje na detekci spliceových variant SpliceAI, který pomocí strojového učení dokáže detekovat varianty měnící sestřih mRNA sekvence.[Jaganathan et al. 2019]

5.8 Perspektivy ve vyhodnocování MPS / NGS dat

V současnosti se v celosvětovém měřítku daří určit původ vzácných genetických onemocnění pomocí WES přibližně u 30 % případů (ve vyjímečných případech monogenně podmíněných onemocnění až 50 %). Aplikováním poznatků dostupných v databázích jako je HGMD, ClinVar či OMIM a pravidelnou reanalýzou dříve negativních výsledků (bez nalezené kauzální varianty) se daří objasnit přibližně 10–12 % případů [Boycott et al. 2019].

Aby došlo ke zvýšení těchto čísel, je nutné kooperovat v rámci mezinárodních týmů. Každý týden je na světě sekvenováno tisíce exomů, ale většinou se jedná o izolované studie. Pro další vývoj je zásadní, aby byla data z WES sdílena mezi pracovišti, ať už pomocí „match-making“ platform jako je GeneMatcher pro hledání týmů se stejnými geny / variantami nebo seskupováním do velkých konsorcií, která budou všechna tato data agregovat. Stále je ale nutné zajistit a preferovat bezpečí a soukromí pacienta a jeho práva.

Dalším trendem je zpracování celého procesu v outsourcovaných laboratořích – v DNA laboratoři v současnosti využíváme outsourcingu při sekvenování, kdy po zaslání DNA vzorku pacienta získáváme jeho sekvenci, s garancí kvality provedení sekvenace a cenou, které bychom nedosáhli, ani pokud bychom sekvenovali u nás v laboratoři. Lze předpokládat, že tento trend bude následovat i další krok – bionformatickou analýzu, ať už to budou cloudové platformy TerraApp (dříve FireCloud), které

využívají obrovské výpočetní kapacity technologických společností jako jsou Google nebo Amazon a dokáží zpracovat WGS data za 5 dolarů (celá GATK analýza genomu). Kromě toho zpracování bioinformatických dat začínají nabízet i společnosti, které poskytují sekvenování (přesto, že jde pouze o pouhou aplikaci předpřipraveného prostředí v TerraApp). Úloha bioinformatika se tedy bude spíše posouvat od zpracování dat a hledání jednotlivých variant do hledání širších souvislostí jako jsou například GWAS (genome wide association) studie atd.

Přestože WES tvoří majoritní část všech MPS analýz, pro některé případy takový postup není vhodný. Prvním takovým příkladem může být mozaicismus. Detekovat patogenní variantu v mozaice je velmi složitý proces, distribuce varianty ve vzorku DNA se totiž liší jak lokálně (v periferní krvi může být jiná než ve svalové či nervové tkáni). Nalézt tak přesvědčivě patogenní efekt v mozaice vyžaduje velmi vysokou hloubku čtení.

Celoexomové sekvenování není příliš vhodné pro hledání indelů (menších delecí a insercí), CNV či chromozomálních přestaveb. Pro tyto genetické varianty je vhodnější využít WGS. Celogenomové sekvenování kromě těchto přestaveb dokáže identifikovat i varianty mimo kódující oblasti, které mají patogenní charakter. Jsou to varianty v oblastech promotorů, enhancerů, regulatorních sekvencí. U takových variant ale bývá velmi problematické potvrdit jejich kauzalitu. Předpokládá se, že WGS by oproti WES přineslo nalezení kauzální varianty v dalších cca 10 % případů. Interpretace nekódujících variant by mohla pomoci objasnit značné procento případů, u takových variant je ale proces určení patogenicity velmi komplexní. Efekt nekódující varianty lze poměrně snadno určit u sestřihové varianty, ale v intronových a intragenových oblastech se mohou vyskytovat další poruchy, vzniklé jinými mechanismy než jsou pouze SNV. Může jít o poruchy regulace transkripce, ale i během pozdějších úprav post-trasnkripčních i post-translačních.[Boycott et al. 2019]

Dalšími metodami, které by mohly zvýšit procento objasnitelnosti případů, jsou metody zaměřující se na funkční efekt variant. Například metody transkriptomového sekvenování umožnují zjistit efekt sestřihových nebo stop variant na výsledný produkt – jestli dojde ke zvýšení či snížení množství funkčního proteinu. Pro detekci poruch imprintingu jsou využívány metylační arraye, umožňující zjistit alterace v počtu imprintovaných genů nebo objasnit případy uniparentální disomie. Současné metody se zaměřují hlavně na monogenně podmíněná onemocnění, dalším krokem je ale zahrnutí komplexních etiologií do diangostiky, příkladem mohou být tzv. „common epilepsies“, což jsou epilepsie s mírnějším fenotypem, familiárním výskytem, u kterých je předpokládaná příčina polygenní s dalším vlivy [Allen et al. 2017]. Mnoho onemocnění může být ovlivněno či dokonce způsobeno enviromentálními vlivy, kdy může být postižen pre nebo post natální vývoj plodu. Proto je důležité se nezaměřit pouze na statická datam ale využít i dynamických parametrů např. lidského metabolomu a dalších multi-omických studií. [Aref-Eshghi et al. 2018; Soellner et al. 2017]

6 Závěr

Cílem disertační práce bylo objasňování příčin neurogenetických onemocnění analýzou dat z MPS pomocí moderních algoritmů. Celý projekt byl zpracováván v rámci pracoviště DNA laboratoře Kliniky dětské neurologie 2.LF a FN Motol. Zaměřovali jsme se na pacienty s neurogenetickým onemocněním, zejména na tři onemocnění - epileptické encefalopatie, CMT a dědičné hluchoty.

Souhrnem lze říci, že hlavní cíle byly splněny:

1. Analýza MPS dat sekvenovaných panelem genů
   - Navržená metodika poskytuje rychlou a spolehlivou možnost identifikování kauzálních variant pro onemocnění sledovaná v DNA laboratoři. Aplikace metodiky byla provedena u 257 pacientů s EE, kde jsme aplikovali panel 112 genů asociovaných s onemocněním. U těchto pacientů jsme identifikovali patogenní a pravděpodobně patogenní varianty v 28 % případů a provedli jsme další analýzu variant - dle jejich původu (*de novo* nebo děděné) a dle dědičnosti genu (AD, AR, X-vázané). Dále jsme provedli analýzu věku nástupu, kdy jsme prokázali vysokou objasnitelnost u pacientů s nástupem onemocnění v prvních čtyřech týdnech života. Výsledky byly publikované v[Staněk et al. 2018]. Kromě toho jsme publikovali další dvě kazuistiky, jejichž objasnění bylo dosaženo pomocí sekvenování panelem genů - [Neupauerová et al. 2017; Štěrbová et al. 2018].

2. Analýza celoexomových dat (WES)
   - Pro vyhodnocování WES bylo nutné nejprve metodiku navrhnout a otestovat. Nejprve jsme porovnali jednotlivé bioinformatické postupy využívané v DNA lab pro bioinformatickou analýzu a navrhli vhodnou kombinaci metod. Dále jsme zavedli metody pokročilého zpracování a anotace variant, které nám umožnily identifikovat varianty, které bychom metodou manuálního filtrování nenalezli. Jedná se o pokročilé filtrování pomocí HPO termínů, kdy dochází k asociaci fenotypu s geny. Rovněž bylo důležité zavést a otestovat nástroje pro detekci *de novo* variant pomocí Trio analýzy. Díky zavedení nástroje DenovoGear do naší workflow jsme identifikovali de novo varianty, mimo jiné i kauzální *UBTF* variantu, publikovanou v [Sedláčková et al. 2018].
   - Kromě základní workflow pro hledání kauzálních variant jsme rovněž otestovali mnoho nástrojů pro detekci CNV, kdy jsme nejlepších výsledků dosáhli využitím workflow GATK 4 pro hledání germinálních CNV.
   - Díky těmto metodikám se nám podařilo identifikovat kauzální varianty v mnoha dalších případech, které jsou uvedené ve výsledcích.

3. Analýza celogenomových dat (WGS)
   - Pro vyhodnocování WGS jsme navrhli a otestovali kromě klasické metodiky, využívané pro WES i možnost analýzy v rámci cloudových služeb. Nakonec jsme se ale rozhodli data zpracovávat shodnou metodikou jako u WES, kdy nejprve analyzujeme WGS data s omezením na oblasti shodné s WES a poté na oblast celého genomu. V současnosti po analýze 33 WGS vzorků jsme neidentifikovali žádnou kauzální variantu, ale nalezli jsme některé kandidátní varianty, které by mohly být v budoucnu potvrzeny.
4. Databáze variant z MPS
   - (In-house) Databáze variant z WES - vytvořili jsme databázi 222 WES vzorků, obsahující více než 300 000 variant. U variant jsme spočítali jejich frekvence v naší databázi a porovnali je s frekvencemi v databázi gnomAD. Tato databáze nám poskytuje přehled o alelické frekvenci variant v naší subpopulaci
   - Díky postupům zavedeným při implementaci in-house databáze máme nyní možnost vytvářet flexibilní analýzy skupin a podskupin pacientů, např. lze velmi rychle spočítat frekvenci variant v dané podskupině pacientů a porovnat s frekvencemi v celé in-house databázi.
   - Databáze proteinových domén prot2HG – vytvořili jsme databázi proteinových domén mapovaných na lidský genom. Tuto databázi můžeme využít při anotování variant, jak jsme ukázali, tak přítomnost varianty v proteinové doméně může být důležitou informací při prioritizaci variant
   - Databáze variant spojených s CMT – v rámci mezinárodní spolupráce jsme vytvořili databázi CMT variant na komunitní bázi, kdy uživatelé mohou sdílent varianty a informace u nich. Tato informace může pomoct při potvrzení kauzality variant.
5. Systém pro správu dat a udržování databázového systému v DNA laboratoři
   - V rámci reanalýzy všech MPS dat, které byly t.č. v DNA laboratoři jsme navrhli systém pro správu dat. Tento systém poskytuje přehlednou a účelnou formu reprezentace, tak aby byla všechna data snadno dohledatelná. V současnosti jsou všechna zpracovaná data dostupná, zálohovaná (alespoň na dvou nezávislých zařízeních) a jsou k dispozici všechny části bioinformatické analýzy.

6.1 Seznam publikací autora

Prvoautorské

David Staněk et al. (2018). "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* 13 (1), s. 71. ISSN: 1750-1172. DOI: 10.1186/s13023-018-0812-8 **(IF: 3,48) udělena cena Ervína Adama za rok 2019**

Spoluautorské

Cima Saghira et al. (2018). "Variant pathogenicity evaluation in the community-driven Inherited Neuropathy Variant Browser." *Human mutation* 39 (5), s. 635–642. ISSN: 1098-1004. DOI: 10.1002/humu.23412 **(IF: 5,35)**

Pe Laššuthová et al. (2018). "Novel SBF2 mutations and clinical spectrum of Charcot-Marie-Tooth neuropathy type 4B2." *Clinical genetics* 94 (5), s. 467–472. ISSN: 1399-0004. DOI: 10.1111/cge.13417 **(IF: 3,51)**

Lucie Sedláčková et al. (2018). "UBTF Mutation Causes Complex Phenotype of Neurodegeneration and Severe Epilepsy in Childhood." *Neuropediatrics*. ISSN: 1439-1899. DOI: 10.1055/s-0038-1676288 **(IF: 1,60)**

Katalin Štěrbová et al. (2018). "Neonatal Onset of Epilepsy of Infancy with Migrating Focal Seizures Associated with a Novel GABRB3 Variant in Monozygotic Twins." *Neuropediatrics* 49 (3), s. 204–208. ISSN: 1439-1899. DOI: 10.1055/s-0038-1626708 **(IF: 1,57)**

Jana Neupauerová et al. (2017). "Two Novel Variants Affecting CDKL5 Transcript Associated with Epileptic Encephalopathy." *Genetic testing and molecular biomarkers* 21 (10), s. 613–618. ISSN: 1945-0257. DOI: 10.1089/gtmb.2017.0110 **(IF: 1,26)**

Anna Uhrová Mészárosová et al. (2017). "Disease-Causing Variants in the ATL1 Gene Are a Rare Cause of Hereditary Spastic Paraplegia among Czech Patients". *Annals of human genetics* 81.6, s. 249–257 **(IF: 1,53)**

Simona Poisson Marková et al. (2018). "STRC Gene Mutations, Mainly Large Deletions, are a Very Important Cause of Early-Onset Hereditary Hearing Loss in the Czech Population." *Genetic testing and molecular biomarkers* 22 (2), s. 127–134. ISSN: 1945-0257. DOI: 10.1089/gtmb.2017.0155 **(IF: 1,18)**

7 Souhrn

V rámci disertační práce „Objasňování příčin neurogenetických onemocnění analýzou dat z MPS pomocí moderních algoritmů“ jsme zpracovávali MPS data sekvenovaná pomocí panelu genů, celoexomového (WES) a celogenomového (WGS) sekvenování.

Sekvenování panelem genů Při sekvenování pomocí panelu genů jsme využívali na našem pracovišti navržený panel genů, které jsou asociovány s onemocněním. Obecnou podmínkou pro zařazení genu do panelu jsou minimálně dvě nezávislé publikace asociující gen s onemocněním a nebo alespoň jedna publikace popisující kauzální varianty v genu ve dvou nebo více nepříbuzných pacientech. Těmto kritériím v případě panelu pro epileptickou encefalopatii (EE) vyhovovalo t.č. 112 genů. Sekvenování pomocí panelu genů bylo provedeno u 257 pacientů s epileptickou encefalopatií. Patogenní či pravděpodobně patogenní variantu jsme nalezli u 28 % případů (72 z 257 pacientů).

U patogenních a pravděpodobně patogenních 76 variant jsme provedli další analýzu variant – rozdělili jsme varianty dle genů do skupin dle dědičnosti a dle původu varianty na de novo, zděděné a s neznámým původem. Ze 112 genů v panelu jsme nalezli patogenní nebo pravděpodobně patogenní variantu ve 33 genech, z nich se nejčastěji jednalo o geny s autozomálně dominantní dědičností (50 variant ve 22 genech). Dle segregační analýzy bylo možné určit původ variant u 68 pacientů ze 72. De novo vznik jsme potvrdili u 70,3 % variant. Rovněž jsme prokázali spojitost mezi věkem nástupu onemocnění a objasnitelností – ta byla dvojnásobná v případě, že první záchvat se u pacienta objevil do 4 týdnů věku. Tato studie navazuje na publikaci [Staněk et al. 2018]. Kromě toho uvádíme další dvě spoluautorské publikace, které popisují kazuistiky pacientů s EE [Neupauerová et al. 2017] a [Štěrbová et al. 2018].

Celoexomové sekvenování (WES) V kapitole celoexomového sekvenování nejprve porovnáváme bioinformatické postupy využívané v DNA laboratoři. Pro analýzu bylo vybráno 24 WES vzorků pacientů s EE. Otestovali jsme bioinformatické zpracování třemi způsoby – GATK best practices workflow a dva komerční nástroje SureCall a NextGENe. Z výsledků vyplynulo, že GATK a SureCall poskytují oba kvalitní výsledky a proto budou v DNA laboratoři metodou první volby.

Při hledání příčiny onemocnění pomocí WES jsme definovali dva hlavní přístupy de novo model a singleton model. U de novo modelu je důležité mít k dispozici data z WES jak u pacienta, tak jeho rodičů. Další zpracování pak probíhá pomocí nástroje DeNovoGear, který se ukázal jako optimální pro hledání de novo variant u pacientů s WES. Díky zavedení této metodiky jsme identifikovali varianty, které byly následně publikovány ve spoluautorských publikacích [Sedláčková et al. 2018]

a [Neupauerová et al. 2018]. Druhou možností pro vyhodnocování je tzv. Singleton model, kdy hledáme příčinu onemocnění analýzou pouze probandova vzorku. V tomto případě jsme zavedli metodiku manuálního filtrování variant doplněnou o vyhodnocování pomocí asociací genotyp-fenotyp nástrojem Exomiser (za využití HPO termínů popisujících fenotyp).

U singleton modelu uvádíme spoluautorskou publikaci [Laššuthová et al. 2018], která popisuje variantu v AR genu *SBF2*, která byla uvedena jako příčina CMT v celkem sedmi rodinách. Dále jsou uvedené další objasněné případy WES. V poslední podkapitole se poté věnujeme CNV analýze, kdy jsme otestovali vhodné nástroje a povedlo se nám zavést metodiku vhodnou pro vyhodnocování CNV ve WES datech. Metodu jsme otestovali na dvou již dříve potvrzených případech CNV, k vyhodnocení dat jsme rovněž použili nástroj pro vytváření virtuálních panelů, implementovaný na našem pracovišti.

Bioinformatické databáze V další kapitole se věnujeme třem bioinformatickým databázím, které jsme na pracovišti implementovali a pomáhají při vyhodnocování MPS dat.

První databází je databáze variant všech 222 WES vzorků shromážděných v DNA laboratoři. Jedná se o celkem 300 111 variant ve 17 512 genech. Pro tyto varianty jsme vypočítali jejich alelickou frekvenci v naší populaci, uvedli jejich typ a provedli genovou analýzu. Dalším krokem pak bylo varianty rozdělit do tříd, porovnat alelické frekvence v naší databázi proti alelické frekvenci v databázi gnomAD. Tím jsme definovali varianty, které jsou v naší populaci častější, než by se dalo předpokládat dle alelické frekvence NFE (nefinské evropské) populace v gnomAD. Tuto databázi lze využít při manuálním filtrování variant jako další anotační zdroj, kdy první možností je vyfiltrování variant, které mají v naší databázi vysokou frekvenci. Další možností je pak předfiltrování variant, kdy dojde k odstranění hypervariabilních genů z naší kohorty a lokálně specificky častých variant.

Databáze proteinových domén poskytuje informace o genomických pozicích proteinových domén. Díky tomu dokážeme pomocí anotace touto databází určit, které varianty spadají do proteinových domén, což může predikovat patogenní riziko varianty. Databáze je dostupná na URL www.prot2hg.com. Pro ověření správnosti jsme provedli anotační analýzu, kdy jsme využili variant v gnomAD a ClinVar pro ověření hypotézy, že varianty spadající do proteinových domén mají vyšší patogenní potenciál. Další ověření jsme rovněž provedli anotací ověřených kauzálních variant z EE panelu – kdy více než 70 % variant spadalo do proteinových domén, toto srovnání jsme prezentovali na konferenci ASHG v roce 2018 (poster v příloze H).

Databáze variant spojených s CMT vznikla v rámci mezinárodní spolupráce s pracovištěm Hussman Insititute for Human Genomics v Miami (USA), jedná se o komunitně vedenou databázi variant spojených s CMT, kdy uživatelé mají možnost varianty přidávat, hodnotit a navzájem sdílet. Výsledkem tohoto projektu byla publikace [Saghira et al. 2018]. V rámci projektu jsem se podílel na návrhu databáze (technického řešení), doplňování dat a implementoval jsem důležitou komponentu pro zobrazování proteinových domén u CMT genů.

Správa dat Posledním cílem disertační práce bylo navržení udržitelné správy dat v DNA laboratoři. Se zvyšujícím se počtem MPS dat v laboratoři se zvyšují požadavky na správu. Proto jsme navrhli systém, který umožní dlouhodobě data uchovávat ve snadno dohledatelné formě (přesně definovaná struktura ukládání), data jsou zálohovaná na zařízeních NAS a dalších vzdálených serverech a rovněž díky definované metodice nedochází k redundanci uložených dat.

8 Summary

The thesis "The elucidation of the causes of neurogenetic diseases by the MPS data analysis using advanced algorithms" is focused on processing the massively parallel sequencing (MPS) data from a gene panel, whole-exome sequencing (WES) and whole-genome sequencing (WGS). The aim of the study was to develop a suitable pipeline to evaluate at least 250 MPS gene panel data, 150 WES data and 20 WGS data in order to improve molecular genetic testing of rare neurogenetic disorders. Associated data management and database implementation is also described.

Targeted gene panel sequencing A custom-designed gene panel consisting of genes previously associated with the disease was used. In the Epileptic Encephalopathy (EE) panel, two prerequisites need to be met for inclusion into the panel: the gene has to have been published in at least two independent publications OR at least in one publication but in multiple independent families. In the case of the EE panel, 112 genes were included. The targeted gene panel sequencing was then performed on 257 patients with EE. Pathogenic or likely pathogenic (according to ACMG criteria) variants have been found in 28% of patients (72 out of 257). Further analysis of the pathogenic or likely pathogenic variants was performed (76 in total); the variants were grouped by gene and then the genes were grouped by inheritance and origin of the variants. Out of 112 genes included in the custom gene panel, pathogenic or likely pathogenic variants were found in 33 genes. According to the segregation analysis, we were able to determine the origin of the variant in 68 patients out of 72. De novo origin was confirmed in 70.3% of variants. A relationship between the age of onset of the epileptic seizures and clarification rate was also demonstrated; the number of solved cases is almost two times higher in patients with onset of seizures in the first four weeks of life [Staněk et al. 2018]. In addition, two co-authored publications presenting case reports of patients with EE are included [Neupauerová et al. 2017] and [Štěrbová et al. 2018].

Whole exome sequencing (WES) For WES data analysis, three different bioinformatic pipelines were compared (GATK, SureCall, NextGene). Twenty-four WES samples of EE patients were used as a test dataset. It was shown that the optimal results were obtained with both GATK and SureCall. Therefore, these two algorithms were chosen to be the method of choice in the DNA lab.

Two main approaches (de novo and singleton model) were used for variant evaluation.

Based on the analysis of WES samples in Trio (proband and both parents), several different tools were tested for the de novo model. The DeNovoGear tool produced the best results for de novo variant detection and cases solved by using the de novo model have been reported, [Sedláčková et al. 2018] and [Neupauerová et al. 2018].

The second option for WES variant evaluation is a singleton model. This analysis is based on manual filtering followed by a genotype-phenotype association tool Exomiser (using HPO terms describing the phenotype). The singleton model was used to identify variants in the SBF2 gene – presented in the co-authored publication [Laššuthová et al. 2018]. We also describe some other cases solved by using proposed methods.

At the end of WES chapter, the CNV analysis workflow is presented. At first, we performed the analysis of available and suitable tools and finally we managed to introduce an optimized methodology for germline CNV in WES data by using GATK 4 beta pipeline. We tested the methodology on two previously confirmed CNV cases and with the GATK4 tool and the custom virtual panel tool (implemented in the DNA laboratory) we were able to detect both of the CNV in the samples.

Bioinformatic databases The first database is a variant database of 222 WES samples (all samples in our workplace). By using GATK, we detected 300,111 variants in 17,512 genes. For these variants we calculated their allele frequency in our subpopulation, listed their type and performed gene analysis. The variants were divided into five classes according to their allele frequency in our database and compared to gnomAD allele frequency. Further analysis of variants, that are more common in our subpopulation rather than in the gnomAD, was performed.

The protein domain database is available at url: *www.prot2hg.com*. Here we introduce a resource that addresses the question of whether a particular variant falls onto an annotated protein domain. When applied to patient genetic data, we found that rare (<1%) variants in the gnomAD were significantly more annotated onto a protein domain in comparison to common (>1%) variants. Similarly, variants described as pathogenic or likely pathogenic in ClinVar were more likely to be annotated onto a domain. In addition, we tested a dataset consisting of 60 causal variants in a cohort of patients with epileptic encephalopathy, and found that 71% of them (43 variants) were propagated onto protein domains (presented as a poster at the ASHG conference 2018; poster is shown in the supplementary section; manuscript has been submitted).

The Inherited Neuropathy Variant Browser is a database of CMT variants developed in an international collaboration with the Hussman Institute for Human Genomics in Miami (USA). It is a community-driven database of CMT variants where users can add, evaluate and share variants. The results of this project were published in a co-authored publication [Saghira et al. 2018]. My role in the project was to help to design the database structure, to load data into the database and to implement the component that projects protein domain location onto the CMT genes.

Data management The final aim of the thesis was to design a sustainable data management for our DNA laboratory. Therefore, we designed a simple and clear system allowing the long-term storage of data in an easily traceable form (precisely defined tree structure). Data is backed up on the NAS devices and other remote servers, and, with the defined methodology, we minimized redundancy of data and maximized data safety.

Bibliografie

Abrams, Alexander J et al. (2015). "Mutations in SLC25A46, encoding a UGO1-like protein, cause an optic atrophy spectrum disorder." *Nature genetics* 47 (8), s. 926–932. ISSN: 1546-1718. DOI: 10.1038/ng.3354.

Abyzov, Alexej et al. (2011). "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing". *Genome research* 21.6, s. 974–984.

Abyzov, Alexej et al. (2015). "Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms". *Nature communications* 6, s. 7256.

Adzhubei, Ivan A et al. (2010). "A method and server for predicting damaging missense mutations." *Nature methods* 7 (4), s. 248–249. ISSN: 1548-7105. DOI: 10.1038/nmeth0410-248.

Afgan, Enis et al. (2018). "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update." *Nucleic acids research* 46 (W1), W537–W544. ISSN: 1362-4962. DOI: 10.1093/nar/gky379.

Aicardi, Jean et al. (1998). *Diseases of the nervous system in childhood*. Sv. 559. Mac Keith Press London.

Allen, Andrew S et al. (2017). "Ultra-rare genetic variation in common epilepsies: a case-control sequencing study". *The Lancet Neurology* 16.2, s. 135–143.

Anand, G et al. (2016). "Autosomal dominant SCN8A mutation with an unusually mild phenotype." *European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society* 20 (5), s. 761–765. ISSN: 1532-2130. DOI: 10.1016/j.ejpn.2016.04.015.

Aref-Eshghi, Erfan et al. (2018). "Genomic DNA Methylation Signatures Enable Concurrent Diagnosis and Clinical Genetic Variant Classification in Neurodevelopmental Syndromes." *American journal of human genetics* 102 (1), s. 156–174. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.12.008.

Arzimanoglou, Alexis et al. (2009). "Lennox-Gastaut syndrome: a consensus approach on diagnosis, assessment, management, and trial methodology." *The Lancet. Neurology* 8 (1), s. 82–93. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(08)70292-8.

Auwera, Geraldine A Van der et al. (2013). "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline." *Current protocols in bioinformatics* 43, s. 11.10.1–11.1033. ISSN: 1934-340X. DOI: 10.1002/0471250953.bi1110s43.

Azzedine, H et al. (2003). "Mutations in MTMR13, a new pseudophosphatase homologue of MTMR2 and Sbf1, in two families with an autosomal recessive demyelinating form of Charcot-Marie-Tooth disease associated with early-onset glaucoma." *American journal of human genetics* 72 (5), s. 1141–1153. ISSN: 0002-9297. DOI: 10.1086/375034.

Benedetti, S et al. (2007). "Phenotypic clustering of lamin A/C mutations in neuromuscular patients." *Neurology* 69 (12), s. 1285–1292. ISSN: 1526-632X. DOI: 10.1212/01.wnl.0000261254.87181.80.

Bennett, Simon (2004). "Solexa Ltd." *Pharmacogenomics* 5 (4), s. 433–438. ISSN: 1462-2416. DOI: 10.1517/14622416.5.4.433.

Berg, Anne T et al. (2010). "Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005-2009." *Epilepsia* 51 (4), s. 676–685. ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2010.02522.x.

Bergoffen, J et al. (1993). "Connexin mutations in X-linked Charcot-Marie-Tooth disease." *Science (New York, N.Y.)* 262 (5142), s. 2039–2042. ISSN: 0036-8075.

Birger, Chet et al. (2017). "FireCloud, a scalable cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs". *bioRxiv*. DOI: 10.1101/209494. eprint: https://www.biorxiv.org/content/early/2017/11/03/209494.full.pdf. URL: https://www.biorxiv.org/content/early/2017/11/03/209494.

Birouk, Nazha et al. (2003). "Phenotypical features of a Moroccan family with autosomal recessive Charcot-Marie-Tooth disease associated with the S194X mutation in the GDAP1 gene". *Archives of neurology* 60.4, s. 598–604.

Boycott, Kym M et al. (2017). "International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases." *American journal of human genetics* 100 (5), s. 695–705. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.04.003.

Boycott, Kym M et al. (2019). "A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers." *Cell* 177 (1), s. 32–37. ISSN: 1097-4172. DOI: 10.1016/j.cell.2019.02.040.

Braathen, G J (2012). "Genetic epidemiology of Charcot-Marie-Tooth disease." *Acta neurologica Scandinavica. Supplementum* (193), s. iv–22. ISSN: 1600-5449. DOI: 10.1111/ane.12013.

Brozková, Dana et al. (2010). "Six new gap junction beta 1 gene mutations and their phenotypic expression in Czech patients with Charcot-Marie-Tooth disease." *Genetic testing and molecular biomarkers* 14 (1), s. 3–7. ISSN: 1945-0257. DOI: 10.1089/gtmb.2009.0093.

Brožková, D Šafka et al. (2016). "HSMNR belongs to the most frequent types of hereditary neuropathy in the Czech Republic and is twice more frequent than HMSNL." *Clinical genetics* 90 (2), s. 161–165. ISSN: 1399-0004. DOI: 10.1111/cge.12745.

Brožková, Dana Šafka et al. (2017). "HMSN Lom in 12 Czech patients, with one unusual case due to uniparental isodisomy of chromosome 8." *Journal of human genetics* 62 (3), s. 431–435. ISSN: 1435-232X. DOI: 10.1038/jhg.2016.148.

Brožková, Dana Šafka et al. (2013). "Spectrum and frequencies of mutations in the MFN2 gene and its phenotypical expression in Czech hereditary motor and sensory neuropathy type II patients." *Molecular medicine reports* 8 (6), s. 1779–1784. ISSN: 1791-3004. DOI: 10.3892/mmr.2013.1730.

Capovilla, Giuseppe et al. (2013). "The history of the concept of epileptic encephalopathy". *Epilepsia* 54, s. 2–5.

Caraballo, Roberto et al. (2011). "Long-term follow-up of the ketogenic diet for refractory epilepsy: multicenter Argentinean experience in 216 pediatric patients." *Seizure* 20 (8), s. 640–645. ISSN: 1532-2688. DOI: 10.1016/j.seizure.2011.06.009.

Cartoni, Romain a Jean-Claude Martinou (2009). "Role of mitofusin 2 mutations in the physiopathology of Charcot-Marie-Tooth disease type 2A." *Experimental neurology* 218 (2), s. 268–273. ISSN: 1090-2430. DOI: 10.1016/j.expneurol.2009.05.003.

Carvalho, Ellaine et al. (2015). "Schinzel-Giedion syndrome in two Brazilian patients: Report of a novel mutation in SETBP1 and literature review of the clinical features." *American journal of medical genetics. Part A* 167A (5), s. 1039–1046. ISSN: 1552-4833. DOI: 10.1002/ajmg.a.36789.

Charcot, Jean Martin (1886). "Sur une forme paticuliere d'atrophie musuculaire progressive souvent familial, debutante par les pieds et les jambes et atteignant plus tard les mains". *Rev. Med Fr* 6, s. 97–138.

Chen, Xiaoyu et al. (2015). "Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications". *Bioinformatics* 32.8, s. 1220–1222.

Cock, Peter J A et al. (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." *Nucleic acids research* 38 (6), s. 1767–1771. ISSN: 1362-4962. DOI: 10.1093/nar/gkp1137.

Consortium, 1000 Genomes Project et al. (2015). "A global reference for human genetic variation." *Nature* 526 (7571), s. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393.

Consortium, Epi4K (2012). "Epi4K: gene discovery in 4,000 genomes." *Epilepsia* 53 (8), s. 1457–1467. ISSN: 1528-1167. DOI: 10.1111/j.1528-1167.2012.03511.x.

Consortium, Epi4K et al. (2013). "De novo mutations in epileptic encephalopathies." *Nature* 501 (7466), s. 217–221. ISSN: 1476-4687. DOI: 10.1038/nature12439.

Consortium, International Human Genome Sequencing (2004). "Finishing the euchromatic sequence of the human genome." *Nature* 431 (7011), s. 931–945. ISSN: 1476-4687. DOI: 10.1038/nature03001.

Consortium, UniProt (2019). "UniProt: a worldwide hub of protein knowledge." *Nucleic acids research* 47 (D1), s. D506–D515. ISSN: 1362-4962. DOI: 10.1093/nar/gky1049.

Cooper, David N, Edward V Ball a Michael Krawczak (1998). "The human gene mutation database". *Nucleic acids research* 26.1, s. 285–287.

Dalla Bernardina, B et al. (1982). "[Early myoclonic epileptic encephalopathy (EMEE) (author's transl)]." *Revue d'electroencephalographie et de neurophysiologie clinique* 12 (1), s. 8–14. ISSN: 0370-4475.

Damseh, Nadirah et al. (2015). "Mutations in SLC1A4, encoding the brain serine transporter, are associated with developmental delay, microcephaly and hypomyelination." *Journal of medical genetics* 52 (8), s. 541–547. ISSN: 1468-6244. DOI: 10.1136/jmedgenet-2015-103104.

Danecek, Petr et al. (2011). "The variant call format and VCFtools." *Bioinformatics (Oxford, England)* 27 (15), s. 2156–2158. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr330.

DePristo, Mark A et al. (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nature genetics* 43 (5), s. 491–498. ISSN: 1546-1718. DOI: 10.1038/ng.806.

Doolittle, Russell F (1995). "The multiplicity of domains in proteins". *Annual review of biochemistry* 64.1, s. 287–314.

Dunnen, Johan T den et al. (2016). "HGVS Recommendations for the Description of Sequence Variants: 2016 Update." *Human mutation* 37 (6), s. 564–569. ISSN: 1098-1004. DOI: 10.1002/humu.22981.

Edvardson, Simon et al. (2017). "Heterozygous De Novo UBTF Gain-of-Function Variant Is Associated with Neurodegeneration in Childhood." *American journal of human genetics* 101 (2), s. 267–273. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.07.002.

El, Carla G van et al. (2013). "Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics." *European journal of human genetics : EJHG* 21 (6), s. 580–584. ISSN: 1476-5438. DOI: 10.1038/ejhg.2013.46.

Eldomery, Mohammad K et al. (2017). "Lessons learned from additional research analyses of unsolved clinical exome cases." *Genome medicine* 9 (1), s. 26. ISSN: 1756-994X. DOI: 10.1186/s13073-017-0412-6.

Engel, J a International League Against Epilepsy (ILAE) (2001). "A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ILAE Task Force on Classification and Terminology." *Epilepsia* 42 (6), s. 796–803. ISSN: 0013-9580.

Fan, Xian et al. (2014). "BreakDancer: Identification of genomic structural variation from paired-end read mapping". *Current protocols in bioinformatics* 45.1, s. 15–6.

Fisher, Robert S et al. (2014). "ILAE official report: a practical clinical definition of epilepsy." *Epilepsia* 55 (4), s. 475–482. ISSN: 1528-1167. DOI: 10.1111/epi.12550.

Fisher, Robert S et al. (2017a). "Instruction manual for the ILAE 2017 operational classification of seizure types." *Epilepsia* 58 (4), s. 531–542. ISSN: 1528-1167. DOI: 10.1111/epi.13671.

Fisher, Robert S et al. (2017b). "Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology." *Epilepsia* 58 (4), s. 522–530. ISSN: 1528-1167. DOI: 10.1111/epi.13670.

Gastaut, JL et al. (2000). "Charcot Marie Tooth disease: exacerbation in pregnancy". *Revue neurologique* 156.10, s. 890–891.

Gonzalez, Michael et al. (2015). "Innovative genomic collaboration using the GENESIS (GEM.app) platform." *Human mutation* 36 (10), s. 950–956. ISSN: 1098-1004. DOI: 10.1002/humu.22836.

Gonzalez-Redondo, J M et al. (1989). "A C—-T substitution at nt–101 in a conserved DNA sequence of the promotor region of the beta-globin gene is associated with "silent"beta-thalassemia." *Blood* 73 (6), s. 1705–1711. ISSN: 0006-4971.

Graham, John M et al. (2016). "KCNK9 imprinting syndrome-further delineation of a possible treatable disorder." *American journal of medical genetics. Part A* 170 (10), s. 2632–2637. ISSN: 1552-4833. DOI: 10.1002/ajmg.a.37740.

Griffiths, Anthony JF et al. (2005). *An introduction to genetic analysis*. Macmillan.

Guella, Ilaria et al. (2017). "De Novo Mutations in YWHAG Cause Early-Onset Epilepsy." *American journal of human genetics* 101 (2), s. 300–310. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.07.004.

Haberlová, Jana, Radim Mazanec a Pavel Seeman (2006). "Dědičné periferní neuropatie". *Neurologie pro praxi* 7.3, s. 147–152. ISSN: 1213-1814.

Hagberg, B a B Westerberg (1983). "HEREDITARY MOTOR AND SENSORY NEUROPATHIES IN SWEDISH CHILDREN I: Prevalence and Distribution by Disability Groups". *Acta Pædiatrica* 72.3, s. 379–383.

Hamdan, Fadi F et al. (2017). "High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies." *American journal of human genetics* 101 (5), s. 664–685. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2017.09.008.

Hamosh, Ada et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". *Nucleic acids research* 33.suppl_1, s. D514–D517.

Hantke, Janina et al. (2009a). "A mutation in an alternative untranslated exon of hexokinase 1 associated with hereditary motor and sensory neuropathy – Russe (HMSNR)." *European journal of human genetics : EJHG* 17 (12), s. 1606–1614. ISSN: 1476-5438. DOI: 10.1038/ejhg.2009.99.

Hantke, Janina et al. (2009b). "A mutation in an alternative untranslated exon of hexokinase 1 associated with hereditary motor and sensory neuropathy–Russe (HMSNR)". *European journal of human genetics* 17.12, s. 1606.

Harding, A E a P K Thomas (1980). "The clinical features of hereditary motor and sensory neuropathy types I and II." *Brain : a journal of neurology* 103 (2), s. 259–280. ISSN: 0006-8950. DOI: 10.1093/brain/103.2.259.

Heiskala, H (1997). "Community-based study of Lennox-Gastaut syndrome." *Epilepsia* 38 (5), s. 526–531. ISSN: 0013-9580.

Helbig, Katherine L et al. (2016). "Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy." *Genetics in medicine : official journal of the American College of Medical Genetics* 18 (9), s. 898–905. ISSN: 1530-0366. DOI: 10.1038/gim.2015.186.

Hildebrand, Michael S et al. (2013). "Recent advances in the molecular genetics of epilepsy." *Journal of medical genetics* 50 (5), s. 271–279. ISSN: 1468-6244. DOI: 10.1136/jmedgenet-2012-101448.

Hoischen, Alexander et al. (2010). "De novo mutations of SETBP1 cause Schinzel-Giedion syndrome." *Nature genetics* 42 (6), s. 483–485. ISSN: 1546-1718. DOI: 10.1038/ng.581.

Houlden, Henry et al. (2004). "A novel RAB7 mutation associated with ulcero-mutilating neuropathy." *Annals of neurology* 56 (4), s. 586–590. ISSN: 0364-5134. DOI: 10.1002/ana.20281.

Hu, B et al. (2017). "A novel missense mutation in AIFM1 results in axonal polyneuropathy and misassembly of OXPHOS complexes." *European journal of neurology* 24 (12), s. 1499–1506. ISSN: 1468-1331. DOI: 10.1111/ene.13452.

Igarashi, Masanori, Elizabeth I Thompson a Gaston K Rivera (1995). "Vincristine neuropathy in type I and type II Charcot-Marie-Tooth disease (hereditary motor sensory neuropathy)". *Medical and pediatric oncology* 25.2, s. 113–116.

ILAE. *Epileptic Encephalopathies in Infancy and Childhood | Epilepsy Foundation*. https://www.epilepsy.com/learn/professionals/about-epilepsy-seizures/epileptic-encephalopathies-infancy-and-childhood. (Accessed on 03/27/2019).

– (1981). "Proposal for revised clinical and electroencephalographic classification of epileptic seizures. From the Commission on Classification and Terminology of the International League Against Epilepsy." *Epilepsia* 22 (4), s. 489–501. ISSN: 0013-9580.

Jagadeesh, Karthik A et al. (2016). "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity". *Nature genetics* 48.12, s. 1581.

Jaganathan, Kishore et al. (2019). "Predicting Splicing from Primary Sequence with Deep Learning." *Cell* 176 (3), 535–548.e24. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.12.015.

Johnson, Andrew D (2010). "An extended IUPAC nomenclature code for polymorphic nucleic acids". *Bioinformatics* 26.10, s. 1386–1389.

Karczewski, Konrad J et al. (2016). "The ExAC browser: displaying reference data information from over 60 000 exomes". *Nucleic acids research* 45.D1, s. D840–D845.

Karczewski, Konrad J et al. (2019). "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes". *BioRxiv*, s. 531210.

Khan, Sonia a Raidah Al Baradie (2012). "Epileptic encephalopathies: an overview." *Epilepsy research and treatment* 2012, s. 403592. ISSN: 2090-1356. DOI: 10.1155/2012/403592.

King, Rosalind HM et al. (2011). "Ndrg1 in development and maintenance of the myelin sheath". *Neurobiology of disease* 42.3, s. 368–380.

Klein, C J et al. (2011). "TRPV4 mutations and cytotoxic hypercalcemia in axonal Charcot-Marie-Tooth neuropathies." *Neurology* 76 (10), s. 887–894. ISSN: 1526-632X. DOI: 10.1212/WNL.0b013e31820f2de3.

Komárek, Vladimír (2007). "Léčba epileptických syndromů u dětí". *Česká a slovenská neurologie a neurochirurgie* 70/103.5, s. 473–486. ISSN: 1210-7859.

Koubková, Lucie, Bořivoj Vojtěšek, Rostislav Vyzula et al. (2014). "Sekvenování nové generace a možnosti jeho využití v onkologické praxi". *Klinická onkologie* 27.3.

Kovel, Carolien G F de et al. (2016). "Targeted sequencing of 351 candidate genes for epileptic encephalopathy in a large cohort of patients." *Molecular genetics & genomic medicine* 4 (5), s. 568–580. ISSN: 2324-9269. DOI: 10.1002/mgg3.235.

Krawczak, M, J Reiss a D N Cooper (1992). "The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences." *Human genetics* 90 (1-2), s. 41–54. ISSN: 0340-6717.

Kronenberg, Zev N et al. (2015). "Wham: identifying structural variants of biological consequence". *PLoS computational biology* 11.12, e1004572.

Kulkarni, Shashikant a John Pfeifer (2014). *Clinical genomics*. Academic Press.

Köhler, Sebastian et al. (2014). "The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data." *Nucleic acids research* 42 (Database issue), s. D966–D974. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1026.

Landau, W M a F R Kleffner (1957). "Syndrome of acquired aphasia with convulsive disorder in children." *Neurology* 7 (8), s. 523–530. ISSN: 0028-3878. DOI: 10.1212/wnl.7.8.523.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* 409 (6822), s. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062.

Landrum, Melissa J et al. (2016). "ClinVar: public archive of interpretations of clinically relevant variants." *Nucleic acids research* 44 (D1), s. D862–D868. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1222.

Lassuthova, Petra et al. (2018). "Mutations in ATP1A1 Cause Dominant Charcot-Marie-Tooth Type 2." *American journal of human genetics* 102 (3), s. 505–514. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2018.01.023.

Lassuthová, Petra et al. (2009). "Mutations in the LMNA gene do not cause axonal CMT in Czech patients." *Journal of human genetics* 54 (6), s. 365–368. ISSN: 1435-232X. DOI: 10.1038/jhg.2009.43.

Lawson, Victoria H, Brad V Graham a Kevin M Flanigan (2005). "Clinical and electrophysiologic features of CMT2A with mutations in the mitofusin 2 gene." *Neurology* 65 (2), s. 197–204. ISSN: 1526-632X. DOI: 10.1212/01.wnl.0000168898.76071.70.

Layer, Ryan M et al. (2014). "LUMPY: a probabilistic framework for structural variant discovery". *Genome biology* 15.6, R84.

Laššuthová, P et al. (2011). "High frequency of SH3TC2 mutations in Czech HMSN I patients." *Clinical genetics* 80 (4), s. 334–345. ISSN: 1399-0004. DOI: 10.1111/j.1399-0004.2011.01640.x.

Laššuthová, Pe et al. (2018). "Novel SBF2 mutations and clinical spectrum of Charcot-Marie-Tooth neuropathy type 4B2." *Clinical genetics* 94 (5), s. 467–472. ISSN: 1399-0004. DOI: 10.1111/cge.13417.

Laššuthová, Petra et al. (2012). "Clinical, in silico, and experimental evidence for pathogenicity of two novel splice site mutations in the SH3TC2 gene." *Journal of neurogenetics* 26 (3-4), s. 413–420. ISSN: 1563-5260. DOI: 10.3109/01677063.2012.711398.

Laššuthová, Petra et al. (2016a). "Improving diagnosis of inherited peripheral neuropathies through gene panel analysis." *Orphanet journal of rare diseases* 11 (1), s. 118. ISSN: 1750-1172. DOI: 10.1186/s13023-016-0500-5.

Laššuthová, Petra et al. (2016b). "Severe axonal Charcot-Marie-Tooth disease with proximal weakness caused by de novo mutation in the MORC2 gene." *Brain : a journal of neurology* 139 (Pt 4), e26. ISSN: 1460-2156. DOI: 10.1093/brain/awv411.

Li, Heng et al. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics (Oxford, England)* 25 (16), s. 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Li, Marilyn M et al. (2017). "Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists." *The Journal of molecular diagnostics : JMD* 19 (1), s. 4–23. ISSN: 1943-7811. DOI: 10.1016/j.jmoldx.2016.10.002.

Lupski, J R et al. (1991). "DNA duplication associated with Charcot-Marie-Tooth disease type 1A." *Cell* 66 (2), s. 219–232. ISSN: 0092-8674. DOI: 10.1016/0092-8674(91)90613-4.

Lupski, James R (2015). "Structural variation mutagenesis of the human genome: Impact on disease and evolution." *Environmental and molecular mutagenesis* 56 (5), s. 419–436. ISSN: 1098-2280. DOI: 10.1002/em.21943.

Mamanova, Lira et al. (2010). "Target-enrichment strategies for next-generation sequencing." *Nature methods* 7 (2), s. 111–118. ISSN: 1548-7105. DOI: 10.1038/nmeth.1419.

Mapping, National Research Council (US) Committee on a Sequencing the Human Genome (1988). "Mapping and Sequencing the Human Genome".

Maquat, L E (2001). "The power of point mutations." *Nature genetics* 27 (1), s. 5–6. ISSN: 1061-4036. DOI: 10.1038/83759.

Marková, Simona Poisson et al. (2018). "STRC Gene Mutations, Mainly Large Deletions, are a Very Important Cause of Early-Onset Hereditary Hearing Loss in the Czech Population." *Genetic testing and molecular biomarkers* 22 (2), s. 127–134. ISSN: 1945-0257. DOI: 10.1089/gtmb.2017.0155.

Marusic, Petr et al. (2018). "Nove klasifikace epileptickych zachvatu a epilepsii ILAE 2017". *Neurology for practice* 19.1, s. 32–36. ISSN: 12131814. URL: https://www.neurologiepropraxi.cz/artkey/neu-201801-0007.php.

Maxam, A M a W Gilbert (1977). "A new method for sequencing DNA." *Proceedings of the National Academy of Sciences of the United States of America* 74 (2), s. 560–564. ISSN: 0027-8424. DOI: 10.1073/pnas.74.2.560.

Mazanec, Radim et al. (2004). "Divergentní fenotypy choroby Charcot-Marie-Tooth: demyelinizační s infantilním začátkem a axonální s pozdním začátkem a zpomalenou fotoreakcí následkem různých mutací myelin protein zero". *Česká a slovenská neurologie a neurochirurgie* 67/100.5, s. 321–329. ISSN: 1210-7859.

Mendell, J T a H C Dietz (2001). "When the message goes awry: disease-producing mutations that influence mRNA content and performance." *Cell* 107 (4), s. 411–414. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(01)00583-9.

Mészárosová, Anna Uhrová et al. (2017). "Disease-Causing Variants in the ATL1 Gene Are a Rare Cause of Hereditary Spastic Paraplegia among Czech Patients". *Annals of human genetics* 81.6, s. 249–257.

Mikesová, E et al. (2005). "Novel EGR2 mutation R359Q is associated with CMT type 1 and progressive scoliosis." *Neuromuscular disorders : NMD* 15 (11), s. 764–767. ISSN: 0960-8966. DOI: 10.1016/j.nmd.2005.08.001.

Mizuguchi, Takeshi et al. (2017). "PARS2 and NARS2 mutations in infantile-onset neurodegenerative disorder." *Journal of human genetics* 62 (5), s. 525–529. ISSN: 1435-232X. DOI: 10.1038/jhg.2016.163.

Mohiyuddin, Marghoob et al. (2015). "MetaSV: an accurate and integrative structural-variant caller for next generation sequencing." *Bioinformatics (Oxford, England)* 31 (16), s. 2741–2744. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btv204.

Murakami, N, Y Ohtsuka a S Ohtahara (1993). "Early infantile epileptic syndromes with suppression-bursts: early myoclonic encephalopathy vs. Ohtahara syndrome." *The Japanese journal of psychiatry and neurology* 47 (2), s. 197–200. ISSN: 0912-2036.

Myers, Candace T et al. (2016). "De novo mutations in SLC1A2 and CACNA1A are important causes of epileptic encephalopathies". *The American Journal of Human Genetics* 99.2, s. 287–298.

Møller, Rikke S et al. (2017). "Mutations in GABRB3: From febrile seizures to epileptic encephalopathies." *Neurology* 88 (5), s. 483–492. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000003565.

Nelis, E et al. (2002). "Mutations in GDAP1: autosomal recessive CMT with demyelination and axonopathy." *Neurology* 59 (12), s. 1865–1872. ISSN: 0028-3878. DOI: 10.1212/01.wnl.0000036272.36047.54.

Nelis, Eva a C van Broeckhoven (1996). "Estimation of the mutation frequencies in Charcot-Marie-Tooth disease type 1 and hereditary neuropathy with liability to pressure palsies: a European collaborative study".

Neupauerová, Jana et al. (2016). "Massively Parallel Sequencing Detected a Mutation in the MFN2 Gene Missed by Sanger Sequencing Due to a Primer Mismatch on an SNP Site." *Annals of human genetics* 80 (3), s. 182–186. ISSN: 1469-1809. DOI: 10.1111/ahg.12151.

Neupauerová, Jana et al. (2017). "Two Novel Variants Affecting CDKL5 Transcript Associated with Epileptic Encephalopathy." *Genetic testing and molecular biomarkers* 21 (10), s. 613–618. ISSN: 1945-0257. DOI: 10.1089/gtmb.2017.0110.

Neupauerová, Jana et al. (2018). "Schinzel-Giedion Syndrome: First Czech Patients Confirmed by Molecular Genetic Analysis". *Journal of Pediatric Neurology*.

Niemann, Axel et al. (2005). "Ganglioside-induced differentiation associated protein 1 is a regulator of the mitochondrial network: new implications for Charcot-Marie-Tooth disease". *J Cell Biol* 170.7, s. 1067–1078.

O'Leary, Nuala A et al. (2015). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". *Nucleic acids research* 44.D1, s. D733–D745.

Ouvrier, R A, J G McLeod a T E Conchin (1987). "The hypertrophic forms of hereditary motor and sensory neuropathy. A study of hypertrophic Charcot-Marie-Tooth disease (HMSN type I) and Dejerine-Sottas disease (HMSN type III) in childhood." *Brain : a journal of neurology* 110 (Pt 1), s. 121–148. ISSN: 0006-8950.

Ouvrier, Robert A, James Graham McLeod a John David Pollard (1999). *Peripheral neuropathy in childhood*. Cambridge University Press.

Panosyan, Francis B et al. (2017). "Cross-sectional analysis of a large cohort with X-linked Charcot-Marie-Tooth disease (CMTX1)." *Neurology* 89 (9), s. 927–935. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000004296.

Pareyson, Davide a Chiara Marchesi (2009). "Natural history and treatment of peripheral inherited neuropathies". *Inherited Neuromuscular Diseases*. Springer, s. 207–224.

Patel, P I et al. (1992). "The gene for the peripheral myelin protein PMP-22 is a candidate for Charcot-Marie-Tooth disease type 1A." *Nature genetics* 1 (3), s. 159–165. ISSN: 1061-4036. DOI: 10.1038/ng0692-159.

Pearl, Phillip L., Enrique J. Carrazana a Gregory L. Holmes (2001). "The Landau-Kleffner Syndrome." *Epilepsy currents* 1 (2), s. 39–45. ISSN: 1535-7597. DOI: 10.1046/j.1535-7597.2001.00012.x.

Pironti, Erica et al. (2018). "A novel SLC1A4 homozygous mutation causing congenital microcephaly, epileptic encephalopathy and spastic tetraparesis: a video-EEG and tractography - case study." *Journal of neurogenetics* 32 (4), s. 316–321. ISSN: 1563-5260. DOI: 10.1080/01677063.2018.1476510.

Previtali, Stefano C, Angelo Quattrini a Alessandra Bolino (2007). "Charcot–Marie–Tooth type 4B demyelinating neuropathy: deciphering the role of MTMR phosphatases". *Expert reviews in molecular medicine* 9.25, s. 1–16.

Puusepp, Sanna et al. (2018). "Compound heterozygous SPATA5 variants in four families and functional studies of SPATA5 deficiency." *European journal of human genetics : EJHG* 26 (3), s. 407–419. ISSN: 1476-5438. DOI: 10.1038/s41431-017-0001-6.

Raeymaekers, P et al. (1991). "Duplication in chromosome 17p11. 2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a)". *Neuromuscular disorders* 1.2, s. 93–97.

Ramu, Avinash et al. (2013). "DeNovoGear: de novo indel and point mutation discovery and phasing." *Nature methods* 10 (10), s. 985–987. ISSN: 1548-7105. DOI: 10.1038/nmeth.2611.

Ranza, Emmanuelle et al. (2017). "SERPINI1 pathogenic variants: An emerging cause of childhood-onset progressive myoclonic epilepsy." *American journal of medical genetics. Part A* 173 (9), s. 2456–2460. ISSN: 1552-4833. DOI: 10.1002/ajmg.a.38317.

Reilly, Mary M, Sinéad M Murphy a Matilde Laurá (2011). "Charcot-Marie-Tooth disease". *Journal of the peripheral nervous system* 16.1, s. 1–14.

Richard, I a J S Beckmann (1995). "How neutral are synonymous codon mutations?" *Nature genetics* 10 (3), s. 259. ISSN: 1061-4036. DOI: 10.1038/ng0795-259.

Robinson, Peter N et al. (2014). "Improved exome prioritization of disease genes through cross-species phenotype comparison". *Genome research* 24.2, s. 340–348.

Rogers, T et al. (2000). "A novel locus for autosomal recessive peripheral neuropathy in the EGR2 region on 10q23." *American journal of human genetics* 67 (3), s. 664–671. ISSN: 0002-9297. DOI: 10.1086/303053.

Rohkamm, Barbara et al. (2007). "Further evidence for genetic heterogeneity of distal HMN type V, CMT2 with predominant hand involvement and Silver syndrome." *Journal of the neurological sciences* 263 (1-2), s. 100–106. ISSN: 0022-510X. DOI: 10.1016/j.jns.2007.06.047.

Rossor, Alexander M et al. (2013). "Clinical implications of genetic advances in Charcot–Marie–Tooth disease". *Nature Reviews Neurology* 9.10, s. 562.

Saghira, Cima et al. (2018). "Variant pathogenicity evaluation in the community-driven Inherited Neuropathy Variant Browser." *Human mutation* 39 (5), s. 635–642. ISSN: 1098-1004. DOI: 10.1002/humu.23412.

Sancho, Paula et al. (2017). "A newly distal hereditary motor neuropathy caused by a rare AIFM1 mutation." *Neurogenetics* 18 (4), s. 245–250. ISSN: 1364-6753. DOI: 10.1007/s10048-017-0524-6.

Sanger, F, S Nicklen a A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12), s. 5463–5467. ISSN: 0027-8424.

Sayers, E a D Wheeler (2004). "Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils)". *NCBI Short Courses, Bethesda (MD): National Center for Biotechnology Information (US), Bookshelf, US National Library of Medicine, National Institutes of Health*.

Schabhüttl, Maria et al. (2014). "Whole-exome sequencing in patients with inherited neuropathies: outcome and challenges". *Journal of neurology* 261.5, s. 970–982.

Scheffer, Ingrid E et al. (2017). "ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology." *Epilepsia* 58 (4), s. 512–521. ISSN: 1528-1167. DOI: 10.1111/epi.13709.

Šedivá, Marie et al. (2019). "Novel variant in the KCNK9 gene in a girl with Birk Barel syndrome". *European journal of medical genetics*.

Sedláčková, Lucie et al. (2018). "UBTF Mutation Causes Complex Phenotype of Neurodegeneration and Severe Epilepsy in Childhood." *Neuropediatrics*. ISSN: 1439-1899. DOI: 10.1055/s-0038-1676288.

Seeman, P et al. (2004). "Hearing loss as the first feature of late-onset axonal CMT disease due to a novel P0 mutation." *Neurology* 63 (4), s. 733–735. ISSN: 1526-632X. DOI: 10.1212/01.wnl.0000134605.61307.de.

Seeman, Pavel et al. (2000). "Charcot-Marie-Tooth - gonosomálně dominantní typ (CMTX1) - první nálezy mutací v genu pro connexin 32 v České republice". *Česká a slovenská neurologie a neurochirurgie* 63/96.4, s. 219–225. ISSN: 1210-7859.

Seeman, Pavel et al. (2002). "Kongenitální hypomyelinizace v souvislosti s de-novo mutací v genu pro periferní myelin protein 22 - první prokázaný případ v České republice a přehled literatury". *Česká a slovenská neurologie a neurochirurgie* 65/98.3, s. 206–212. ISSN: 1210-7859.

Sevilla, Teresa et al. (2015). "Mutations in the MORC2 gene cause axonal Charcot–Marie–Tooth disease". *Brain* 139.1, s. 62–72.

Shen, Richard et al. (2005). "High-throughput SNP genotyping on universal bead arrays." *Mutation research* 573 (1-2), s. 70–82. ISSN: 0027-5107. DOI: 10.1016/j.mrfmmm.2004.07.022.

Shokralla, Shadi et al. (2012). "Next-generation sequencing technologies for environmental DNA research." *Molecular ecology* 21 (8), s. 1794–1805. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2012.05538.x.

Sim, Ngak-Leng et al. (2012). "SIFT web server: predicting effects of amino acid substitutions on proteins". *Nucleic acids research* 40.W1, W452–W457.

Skre, H (1974). "Genetic and clinical aspects of Charcot-Marie-Tooth's disease". *Clinical genetics* 6.2, s. 98–118.

Snustad, D. Peter, Michael J. Simmons a Jiřina Relichová (2017). *Genetika*. Druhé aktualizované vydání. Brno: Masarykova univerzita. ISBN: 978-80-210-8613-5.

Soellner, L et al. (2017). "Recent Advances in Imprinting Disorders." *Clinical genetics* 91 (1), s. 3–13. ISSN: 1399-0004. DOI: 10.1111/cge.12827.

Staněk, David et al. (2018). "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* 13 (1), s. 71. ISSN: 1750-1172. DOI: 10.1186/s13023-018-0812-8.

Stenson, Peter D et al. (2003). "Human Gene Mutation Database (HGMD): 2003 update." *Human mutation* 21 (6), s. 577–581. ISSN: 1098-1004. DOI: 10.1002/humu.10212.

Strachan, Tom (2014). *Genetics and genomics in medicine*. Taylor & Francis.

Street, V A et al. (2003). "Mutation of a putative protein degradation gene LITAF/SIMPLE in Charcot-Marie-Tooth disease 1C." *Neurology* 60 (1), s. 22–26. ISSN: 1526-632X.

Talevich, Eric et al. (2016). "CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing". *PLoS computational biology* 12.4, e1004873.

Thomas, PK a DB Calne (1974). "Motor nerve conduction velocity in peroneal muscular atrophy: evidence for genetic heterogeneity". *Journal of Neurology, Neurosurgery & Psychiatry* 37.1, s. 68–75.

Thomas, Rhys H a Samuel F Berkovic (2014). "The hidden genetics of epilepsy-a clinically important new paradigm." *Nature reviews. Neurology* 10 (5), s. 283–292. ISSN: 1759-4766. DOI: 10.1038/nrneurol.2014.62.

Timmerman, Vincent, Alleene Strickland a Stephan Züchner (2014). "Genetics of Charcot-Marie-Tooth (CMT) disease within the frame of the human genome project success". *Genes* 5.1, s. 13–32.

Toro, Camilo et al. (2018). "A recurrent de novo missense mutation in UBTF causes developmental neuroregression." *Human molecular genetics* 27 (7), s. 1310. ISSN: 1460-2083. DOI: 10.1093/hmg/ddy049.

Trump, Natalie et al. (2016). "Improving diagnosis and broadening the phenotypes in early-onset seizure and severe developmental delay disorders through gene panel analysis." *Journal of medical genetics* 53 (5), s. 310–317. ISSN: 1468-6244. DOI: 10.1136/jmedgenet-2015-103263.

Venter, J C et al. (2001). "The sequence of the human genome." *Science (New York, N.Y.)* 291 (5507), s. 1304–1351. ISSN: 0036-8075. DOI: 10.1126/science.1058040.

Verhoeven, Kristien et al. (2006). "MFN2 mutation distribution and genotype/phenotype correlation in Charcot-Marie-Tooth type 2." *Brain : a journal of neurology* 129 (Pt 8), s. 2093–2102. ISSN: 1460-2156. DOI: 10.1093/brain/awl126.

Vigevano, F et al. (1993). "The idiopathic form of West syndrome." *Epilepsia* 34 (4), s. 743–746. ISSN: 0013-9580.

Vácha, Marek (2016). "Příběh lidského genomu". *Živa* 64.5, s. 203–206. ISSN: 0044-4812. URL: http://ziva.avcr.cz/.

Wang, Jie et al. (2017). "Epilepsy-associated genes." *Seizure* 44, s. 11–20. ISSN: 1532-2688. DOI: 10.1016/j.seizure.2016.11.030.

Wang, Kai, Mingyao Li a Hakon Hakonarson (2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data". *Nucleic acids research* 38.16, e164–e164.

Warner, L E et al. (1996). "Clinical phenotypes of different MPZ (P0) mutations may include Charcot-Marie-Tooth type 1B, Dejerine-Sottas, and congenital hypomyelination." *Neuron* 17 (3), s. 451–460. ISSN: 0896-6273.

Warner, L E et al. (1999). "Functional consequences of mutations in the early growth response 2 gene (EGR2) correlate with severity of human myelinopathies." *Human molecular genetics* 8 (7), s. 1245–1251. ISSN: 0964-6906.

Whittaker, Roger G et al. (2015). "Electrophysiologic features of SYT2 mutations causing a treatable neuromuscular syndrome." *Neurology* 85 (22), s. 1964–1971. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000002185.

Xuan, Jiekun et al. (2013). "Next-generation sequencing in the clinic: promises and challenges." *Cancer letters* 340 (2), s. 284–295. ISSN: 1872-7980. DOI: 10.1016/j.canlet.2012.11.025.

Ycas, M (1969). *Biological Code (Frontiers of Biology S.)* North-Holland Publishing Company.

Ye, Kai et al. (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads". *Bioinformatics* 25.21, s. 2865–2871.

Yger, Marion et al. (2012). "Characteristics of clinical and electrophysiological pattern of Charcot-Marie-Tooth 4C". *Journal of the Peripheral Nervous System* 17.1, s. 112–122.

Zhang, Feng et al. (2010). "Mechanisms for nonrecurrent genomic rearrangements associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability." *American journal of human genetics* 86 (6), s. 892–903. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2010.05.001.

Zhou, Xiaoguang et al. (2010). "The next-generation sequencing technology and application." *Protein & cell* 1 (6), s. 520–536. ISSN: 1674-8018. DOI: 10.1007/s13238-010-0065-3.

Züchner, Stephan et al. (2004). "Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A." *Nature genetics* 36 (5), s. 449–451. ISSN: 1061-4036. DOI: 10.1038/ng1341.

Štěrbová, Katalin et al. (2018). "Neonatal Onset of Epilepsy of Infancy with Migrating Focal Seizures Associated with a Novel GABRB3 Variant in Monozygotic Twins." *Neuropediatrics* 49 (3), s. 204–208. ISSN: 1439-1899. DOI: 10.1055/s-0038-1626708.

Seznam zkratek

Zkratka	Význam anglicky	Význam česky
(g)VCF	Variant calling file	
ACMG	American Collegene of Medical Genetics	
AD		Autozomálně dominantní
AR		Autozomálně recesivní
ASHG	American Society of Human genetics	
BAM	Binary alignment map	
BQSR	Base quality score recalibration	
CMT	Charcot-Marie-Tooth disease	
CNV	Copy-number variation	
DN		*De novo*
DNG	DenovoGear	
EE		Epileptická encefalopatie
ESHG	European Society of Human genetics	
GATK	Genome analysis toolkit	
GoF	Gain of function	(Varianta způsobující) získání nové funkce
GUI		Grafické uživatelské rozhraní
GWAS	Genome wide association study	
HPO	Human Phenotype Ontology	
HPO	Human phenotype ontology	
LoF	Loss of function	(Varianta způsobující) ztrátu nové funkce
MPS	Massive parallel sequencing	Masivně paralelní sekvenování
NG	NextGene	
NGS	Next-generation sequencing	Sekvenování nové generace
OMIM	Online Mendelian inheritance in Man	
rtcDNA	Reverse translated coding DNA	Zpětně přeložená kódující DNA
SC	SureCall	
SNP	Single nucleotide polymorphism	
SNV	Single nucleotide variant	
WES	Whole exome sequencing	Celoexomové sekvenování
WGS	Whole genome sequencing	Celogenomové sekvenování
XL	X-linked	X-vázané

Přílohy

Seznam příloh:

Příloha A Prvoautorská publikace [Staněk et al. 2018], IF 3,48

Příloha B Spoluautorská publikace [Saghira et al. 2018], IF 5,35

Příloha C Spoluautorská publikace [Laššuthová et al. 2018], IF 3,51

Příloha D Spoluautorská publikace [Sedláčková et al. 2018], IF 1,62

Příloha E Spoluautorská publikace [Štěrbová et al. 2018], IF 1,57

Příloha F Spoluautorská publikace [Neupauerová et al. 2017], IF 1,26

Příloha G Poster z konference ESHG 2018

Příloha H Poster z konference ASHG 2018

Příloha I Geny s asociovanými fenotypy z 5.4.1.1

Příloha J Publikační profil autora