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.