Univerzita Karlova

1. lékařská fakulta

Studijní program: Doktorské studijní programy v biomedicíně

Studijní obor: Molekulární a buněčná biologie, genetika a virologie

UNIVERZITA KARLOVA
1. lékařská fakulta

Mgr. et Mgr. Anna Přistoupilová

Využití nových metod analýzy genomu ve studiu molekulární podstaty vzácných geneticky podmíněných onemocnění

Genome analysis techniques and their applications in elucidation of molecular underpinnings of rare genetic diseases

Disertační práce

Vedoucí závěrečné práce/Školitel: **prof. Ing. Stanislav Kmoch, CSc.**

Praha, 2020

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem řádně uvedla a citovala 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, 10. 4. 2020

Jméno – Příjmení (hůlkovým písmem)

ANNA PŘISTOUPILOVÁ

Podpis

Identifikační záznam:

PŘISTOUPILOVÁ, Anna. *Využití nových metod analýzy genomu ve studiu molekulární podstaty vzácných geneticky podmíněných onemocnění. [Genome analysis techniques and their applications in elucidation of molecular underpinnings of rare genetic diseases]*. Praha, 2020. 75 s., 8 příl. Disertační práce. Univerzita Karlova, 1. lékařská fakulta, Klinika dětského a dorostového lékařství, Laboratoř pro studium vzácných nemocí. Vedoucí práce Kmoch, Stanislav

Abstrakt

Vzácná onemocnění představují heterogenní skupinu více než ~7000 různých onemocnění, která postihují 3,5–5,9 % celosvětové populace. Většina vzácných onemocnění je genetických, ale kauzální geny jsou známy jen u některých z nich. Řada pacientů se vzácným onemocněním zůstává bez diagnózy, která je klíčová pro genetické poradenství, prevenci a léčbu. S rozvojem nových metod analýzy genomu, klesající cenou sekvenování a rostoucími znalostmi o lidském genomu byl nastolen nový koncept identifikace onemocnění podmiňujících genů, založený na porovnávání genetické variability pacienta s genetickou variabilitou běžné populace. Tato disertační práce popisuje nové metody sekvenace genomu (NGS), bioinformatickou analýzu získaných dat a jejich využití ve studiu molekulární podstaty vzácných, geneticky podmíněných onemocnění. Tyto postupy vedly k určení a charakterizaci kauzálních genů a genových mutací u autosomálně dominantního tubulointersticiálního onemocnění ledvin (*SEC61A1, MUC1),* autosomálně dominantní neuronální ceroidní lipofuscinózy (*CLN6*, *DNAJC5),* neurodegenerativního onemocnění neznámé etiologie (*VPS15*), Akadské varianty Fanconiho syndromu (*NDUFAF6*) a spinální svalové atrofie (*SMN1*). Zavedením nových metod analýzy genomu se zvýšila diagnostická výtěžnost vzácných onemocnění z původního 1 % na 50 %.

Klíčová slova: vzácná onemocnění, nové metody sekvenace genomu, bioinformatická analýza

Abstract

Rare diseases represent a heterogeneous group of more than ~7000 different diseases, affecting 3,5-5,9% of the global population. Most rare diseases are genetic, but causal genes are known only in some of them. Many patients with rare diseases remain without a diagnosis, which is crucial for genetic counseling, prevention, and treatment. With the development of new methods of genome analysis, decreasing cost of sequencing, and increasing knowledge of the human genome, a new concept for identifying disease-causing genes was established. It is based on comparing the patient's genetic variability with the genetic variability of the general population. This dissertation describes next-generation sequencing technologies (NGS), bioinformatic analysis of acquired data and their applications in the elucidation of molecular underpinnings of rare genetic diseases. These procedures have led to the identification and characterization of causal genes and gene mutations in autosomal dominant tubulointerstitial kidney disease (*SEC61A1*, *MUC1*), autosomal dominant neuronal ceroid lipofuscinosis (*CLN6*, *DNAJC5*), neurodegenerative disease of unknown etiology (*VPS15*), Acadian variant of Fanconi syndrome (*NDUFAF6*) and spinal muscular atrophy (*SMN1*). The application of novel genome analysis techniques increased the diagnostic yield of rare diseases from the original 1 % to 50 %.

Key words: rare diseases, next-generation sequencing, bioinformatics analysis

Poděkování

Na tomto místě bych ráda poděkovala Standovi Kmochovi za vytvoření odborného a přátelského pracovního prostředí a za možnost být součástí jeho týmu. Dále děkuji všem svým kolegům za příjemnou atmosféru, podnětné rozhovory, cenné rady a odbornou expertízu, díky čemuž mohly vzniknout výsledky prezentované v této práci. Mé díky patří konkrétně Ivě Jedličkové, Aleně Čížkové, Hátě Hartmannové, Káče Hodaňové, Lence Noskové, Viktorovi Stráneckému, Petrovi Vyleťalovi, Lence Piherové, Martině Živné, Ditě Mušálkové, Heleně Myškové, Veronice Barešové a všem ostatním. Také děkuji za spolupráci a trpělivost všem rodinám se vzácnými onemocněními, díky kterým jsme mohli odhalit molekulární podstatu studovaných onemocnění a tím pomoci nejen jim, ale i dalším pacientům. Za podporu dále děkuji svým přátelům - Bartulce, Kátě, Petrovi, WTčkářům a mnohým dalším. Quiero agradecer a Ivo Gut, Marta Gut, Sergi Beltran, Sophia Derdak y Raúl Tonda del CNAG, por la oportunidad de poder crecer a nivel bioinformático y conocer otro ambiente de trabajo. También quiero dar las gracias a todos los otros CNAGers, por los momentos buenos que pasamos juntos. El mayor agradecimiento va a mi Pedazo, que estuvo siempre a mi lado, apoyandome, escúchandome y dispuesto a compartir mis fracasos y celebrar mis logros. A také děkuji svým rodičům za to, že mě vždy podporovali v mých aktivitách, studiu a cestování.

Finanční podporu pro projekty zmíněné v této práci poskytly následující grantové agentury a granty: institucionální programy Univerzity Karlovy: UNCE 204064, UNCE 204011, PRVOUK-P24/LF1/3, PROGRES-Q26/LF1, PROGRES-Q39/LF1, SVV 2016/260148, SVV 260367/2017; výzkumné projekty Ministerstva školství, mládeže a tělovýchovy: LL1204, LH12015, NT13116-4/2012, LO1304 NPU I, LQ1604 NPU II; programy a granty Ministerstva zdravotnictví ČR: NV15-28208A, NV17-29786A, NV19-08-137, RVO-VFN 64165; grantová agentura České Republiky: 14-36804G; Grantová agentura Univerzity Karlovy: 269615, 1402213; Akademie věd: RVO 67985823; Evropský fond pro regionální rozvoj: OPPK CZ.2.16/3.100/24022, OPPK CZ.2.16/3.1.00/24509, OP VaVpI CZ.1.05/2.1.00/19.0400, OP VaVpI CZ.1.05/1.1.00/02.0109; Nadace Carlose Slima: Slim Initiative for Genomic Medicine. Analýza souboru neselektovaných kontrol byla umožněna díky existence a podpoře vědecké infrastruktury Národního centra lékařské genomiky (LM2015091) a jeho projektu zaměřeného na vytvoření referenční databáze genetických variant České republiky (projekt CZ.02.1.01/0.0/0.0/16_013/0001634).

Obsah

Seznam zkratek

ACMG	Americká společnost lékařské genetiky a genomiky (American College of Medical Genetics)
ALOHOMORA	bioinformatický nástroj pro vazebnou analýzu
ANCL	autosomálně dominantní neuronální ceroidní lipofuscinózy (Autosomal Dominant Neuronal Ceroid Lipofuscinosis)
aNIID	adultní onemocnění s neuronálními intranukleárními inkluzemi (adult Neuronal Intranuclear Inclusion Disease)
ANNOVAR	bioinformatický nástroj (ANNOtate VARiation)
AVFS	Akadská varianta Fanconi syndromu (Acadian Variant of Fanconi syndrome)
BAM	binární forma formátu SAM (Binar Alignment/Map)
BCFtools	bioinformatický nástroj (Binary Calling Format tools)
bp	páru bází (base pair)
BWA	bioinformatický nástroj (Burrows–Wheeler Algorithm)
CADD	bioinformatický nástroj (Combined Annotation Dependent Depletion)
CCS	cirkulární konsenzuální sekvence, u technologie SMRT (Circular Consensus Sequence)
cDNA	komplementární deoxyribonukleová kyselina (complementary DNA)
CNAG	Národní centrum pro analýzu genomu (Centro Nacional de Análisis Genómico)
CNVs	změny genové dávky (Copy Number Variantions)
CRT	cyklická reverzibilní terminace (Cyclic Reversible Termination)
CSD	cystein-string doména (Cystein String Domain)
CSPα	cystein-string protein alfa (Cystein String Protein alpha)
ČAVO	Česká Asociace pro Vzácná Onemocnění
ddNTP	dideoxynukleotidtrifosfát (dideoxyNucleotide TriPhosphate)
DMP	Dědičné Metabolické Poruchy
DNA	deoxyribonukleová kyselina (DeoxyriboNucleic Acid)
dNTP	deoxynukleotidtrifosfát (deoxyNucleotide TriPhosphate)
dsDNA	dvouvláknová DNA (double stranded DNA)
exSTRa	bioinformatický nástroj (expanded STR algorithm)
FASTQ	datový formát pro ukládání sekvencí a jim odpovídajícím skóre kvality
FastQC	bioinformatický nástroj (Fastq Quality Control)
FGS	první generace sekvenování (First Generation Sequencing)
GB	jednotka udávající 10^9 bajtů
gDNA	genomová deoxyribonukleová kyselina
hg19	referenční sekvence lidského genomu verze 19 (Human Genome)
hg38	referenční sekvence lidského genomu verze 38 (Human Genome)
GEM	bioinformatický nástroj (The Genome Multitool)
GEMINI	bioinformatický nástroj (GEnome MINIng)
GERP	bioinformatický nástroj (Genomic Evolutionary Rate Profiling)
GERP++	bioinformatický nástroj (Genomic Evolutionary Rate Profiling ++)
gnomAD	databáze (GeNOMe Aggregation Database)
GTEx	databáze (Genotype-Tissue Expression)
HGMD	databáze (Human Gene Mutation Database)
HPO	ontologie (Human Phenotype Ontology)
HuGE	bioinformatický nástroj (Human Genome Epidemiology)
IBD	identické oblasti pocházejí od společného předka (Identity By Descent)

IGV	prohlížeč genomických variant (Integrative Genome Browser)
iNIID	infantilní onemocnění s neuronálními intranukleárními inkluzemi (infantile Neuronal Intranuclear Inclusion Disease)
IPD	interval mezi pulzy, u technologie SMRT (Inter Pulse Duration)
IRDiRC	Mezinárodní konsorcium pro výzkum vzácných onemocnění (International Rare Disease Research Consortium)
jNIID	juvenilní onemocnění s neuronálními intranukleárními inkluzemi (juvenile Neuronal Intranuclear Inclusion Disease)
kb	jednotka udávající 10^3 bází
KEGG	databáze (Kyoto Encyclopedia of Genes and Genomes)
KDDL	Klinika Dětského a Dorostového Lékařství
LAA	bioinformatický nástroj (Long Amplicon Analysis)
LOD	skóre využíváné ve vazebné analýze (Log Of the Odds)
LR-PCR	long-range PCR (Long-Range PCR)
LRT	bioinformatický nástroj (Likelihood Ratio Test)
MAF	frekvence minoritní alely, frekvence druhé nejčastější alely v populaci (Minor Allele Frequency)
Mb	jednotka udávající 10^6 bází
MeSH	lékařský tezaurus (Medical Subject Headings)
MLPA	MLPA esej (Multiplex Ligation dependent Probe Amplification)
mRNA	mediátorová RNA (messenger RNA)
MSA	mnohonásobné zarovnávání sekvencí (Multiple Sequence Alignment)
NCL	neuronální ceroidní lipofuscinóza (Neuronal Ceroid Lipofuscinosis)
NCLG	Národní Centrum Lékařské Genomiky
NGS	nové metody sekvenace (Next-Generation Sequecning)
NHGRI	Národní institut pro výzkum lidského genomu (National Human Genome Research Institute)
NIID	onemocnění s neuronálními intranukleárními inkluzemi (Neuronal Intranuclear Inclusion Disease)
NORD	Národní organizace pro vzácná onemocnění (National Organisation of Rare Disorders)
OMIM	databáze (Online Inheritance In Men)
ONT	firma zabývající se nanopórovým sekvenováním (Oxford Nanopore Technologies)
PacBio	firma zabývající se SMRT sekvenováním (Pacific Biosciences)
PCR	polymerázová řetězová reakce (Polymerase Chain Reaction)
Phevor	bioinformatický nástroj (Phenotype Driven Variant Ontological Re-ranking tool)
PHIVE	bioinformatický nástroj (PHenotypic Interpretation of Variants in Exomes)
RFLP	polymorfismus v délce restrikčních fragmentů (Restriction Fragment Length Polymorphism)
RNA	ribonukleová kyselina (RiboNucleic Acid)
ROH	bioinformatický nástroj (Runs Of Homozygosity)
RVIS	bioinformatický nástroj (Residual Variation Intolerance Score)
SAM	formát využívaný pro uchování alignmentu (Sequence Alignment/Map)
SAMtools	bioinformatický nástroj (Sequence Alignment/Map tool)
SBL	sekvenování ligací (Sequencing By Ligation)
SBS	sekvenování syntézou (Sequencing By Synthesis)
SD	směrodatná odchylka (Standard Deviation)
SGS	druhá generace sekvenování (Second Generation Sequencing)

SMA	spinální svalová atrofie (Spinal Muscular Atrophy)
SMRT	sekvenační technologie firmy Pacific Biosciences (Single Molecule Real Time Sequencing)
SNA	přidání jednoho nukleotidu (Single Nucleotide Addition)
SNPs	jednonukleotidové polymorfismy (Single Nucleotide Polymorphisms)
SOLiD	sekvenační technologie firmy Life Technologies/Thermo Fischer (Sequencing by Oligonucleotide Ligation and Detection)
SSIEM	organizace (Society for the Study of Inborn Errors of Metabolism)
STRetch	bioinformatický nástroj pro detekci tandemových repetic
STRING	bioinformatická databáze (Search Tool for the Retrieval of Interacting Genes/Proteins)
STRs	krátké tandemové repetice (STR, Short Tandem Repeat)
SWAN	syndromy, které nemají název (Syndrome Without a Name)
TB	jednotka udávající 10^{12} bajtů
TGS	třetí generace sekvenování (Third Generation Sequencing)
TREDPARSE	bioinformatický nástroj (Tri-nucleotide-REpeat Diseases PARSE)
UCSC	databáze (University of California, Santa Cruz)
UTRs	5′ a 3´ nepřekládané oblasti (UnTranslated Regions)
VCF	formát využívaný pro uchování variant (Variant Calling Format)
VNTR	variabilní počet tandemových repetic (Variable Number of Tandem Repeats)

1 Úvod

Tato disertační práce se zabývá studiem molekulární podstaty vzácných, geneticky podmíněných onemocnění. Identifikace kauzálních genů a mutací podmiňujících vzácná onemocnění byla dříve založena na postupech biochemické genetiky, pozičního klonování, funkčního klonování, genetického mapování a později i technologiích DNA čipů. Diagnostická výtěžnost těchto postupů byla pouhé 1 %. Většina pacientů se vzácným onemocněním tak zůstávala bez diagnózy, která je klíčová pro genetické poradenství, prevenci a léčbu.

V roce 2010 jsme začali v laboratoři pro studium vzácných nemocí na Klinice dětského a dorostového lékařství (KDDL; v té době ještě Ústav dědičných metabolických poruch) 1. lékařské fakulty UK a VFN využívat nových metod sekvenace genomu (NGS, Next-Generation Sequencing), čímž se kompletně změnil metodický přístup studia vzácných onemocnění. Již se nezaměřujeme na sekvenování jednotlivých genů a genomových oblastí tak, jako dříve. NGS metody nám umožňují rychlé a levné čtení celých genomů či jejich kódujících oblastí a následné porovnávání genetické variability pacienta nebo skupiny jedinců s genetickou variabilitou populace. Můžeme tak odhalit populačně vzácné či unikátní genetické varianty ve funkčně důležitých oblastech genomu, které by mohly vysvětlit studovaný fenotyp.

Ve své disertační práci prezentuji využitelnost a limity NGS na příkladech odhalování molekulární podstaty několika vybraných vzácných onemocnění.

1.1 Vzácná onemocnění a význam jejich studia

Vzácná onemocnění představují heterogenní skupinu více než ~7000 různých onemocnění, která postihují 3,5–5,9 % celosvětové populace, což dohromady představuje ~300 miliónů lidí (Nguengang Wakap et al., 2019). V Evropské unii je vzácné onemocnění definováno jako onemocnění s výskytem nižším než 50 na 100 000 obyvatel (European Union, 2000), zatímco v USA jako onemocnění postihující méně než 200 000 lidí v zemi (Federal Food, Drug, 1983).

Vzácná onemocnění jsou často chronická, závažná a život ohrožující či beroucí. Až 70 % vzácných onemocnění postihuje pacienty v dětském věku, z nichž 30 % umírá před pátým rokem života (The Lancet Diabetes & Endocrinology, 2019). Neznalost diagnózy a příčiny onemocnění snižuje nejen kvalitu života pacientů, ale ovlivňuje významně i kvalitu života jejich rodin. Určení kauzálních genů je základním předpokladem přesné klasifikace

studovaného onemocnění a východiskem pro cílenou DNA diagnostiku, kvalifikované genetické poradenství, prevenci a případně aplikaci či vývoj cílených terapeutických postupů.

Až 72 % vzácných onemocnění je genetického původu (Nguengang Wakap et al., 2019), nicméně kauzální geny nebyly u mnohých z nich stále určeny. V současné době je v databázi Online Inheritance In Men (OMIM, Hamosh et al., 2000), sdružující informace o lidských genech, genetických onemocněních a znacích, popsáno 9123 fenotypových jednotek. Molekulární podstata je známá u 5801 z nich (stav k 30. 3. 2020). Kauzální gen či geny tedy stále ještě neznáme u 36 % jednotek. Navíc některé fenotypové jednotky v databázi vůbec nejsou zaneseny, protože se jedná o syndromy, které nemají název (SWAN, Syndrome Without a Name). Jedinou možnou cestou ke zlepšení této situace je intenzivní základní výzkum dědičných onemocnění neznámé etiologie. Tento typ výzkumu je zároveň jednou z možností, jak vysvětlit základní biologické funkce kauzálních genů.

Základní výzkum vzácných onemocnění dále poskytuje unikátní biologické modely umožňující identifikaci a definici kandidátních genů a efektivní studium základních patofyziologických procesů v lidských buňkách, tkáních, metabolických a regulačních dráhách účastnících se rozvoje komplexních onemocnění.

Studium genetické podstaty vzácných onemocnění je mnohdy komplikováno tím, že na světě existuje jen několik pacientů s daným onemocněním. Je tedy velmi obtížné najít další rodinu se stejným onemocněním a prokázat tak kauzalitu kandidátního genu pomocí opakovaného výskytu mutace ve stejném genu při stejném onemocnění (rekurence). Nalezení i jen jednoho dalšího pacienta s mutací ve stejném genu a překrývajícím se fenotypem může být klíčové pro identifikaci kauzálního genu. Problémem je, že na případech vzácných onemocnění pracuje mnoho výzkumných týmů a klinických pracovišť a data pacientů jsou ukládána v izolovaných databázích bez možnosti přístupu ostatních pracovišť. S cílem vyřešit tento problém a usnadnit hledání genetické příčiny u pacientů s nevyřešeným vzácným onemocněním vznikla v roce 2013 iniciativa Matchmaker Exchange (Philippakis et al., 2015). Matchmaker Exchange propojuje jednotlivé uzly pomocí aplikačního programovacího rozhraní (Application Programming Interface) a umožňuje vyhledávání genů a fenotypických profilů napříč mnoha databázemi najednou. Jedním z uzlů této iniciativy je například GeneMatcher (Sobreira et al., 2015). Jedná se o veřejně dostupnou webovou stránku, která umožňuje propojení mezi lékaři, výzkumníky a pacienty z celého světa, kteří sdílejí zájem o stejný gen.

Studium genetické podstaty vzácných onemocnění je dále komplikováno jejich klinickou heterogenitou, fenotypickou variabilitou a nespecifickými symptomy. Pacienti ve snaze o

zjištění diagnózy podstupují náročná vyšetření u specialistů z různých lékařských oborů a od nástupu prvních příznaků onemocnění k určení správné diagnózy mnohdy uběhne i několik let. Není také výjimkou, že ke správnému určení diagnózy dojde až po úmrtí pacienta nebo také nikdy. Včasná diagnostika a případné včasné nasazení léčby je u některých vzácných onemocnění klíčové - léčba musí být zahájena dříve, než dojde k nenávratnému poškození tkání.

V současné době je jen málo vzácných onemocnění, pro která existuje specifická léčba. K roku 2018 schválil Americký Úřad pro kontrolu potravin a léčiv (Food and Drug Administration) 747 léků pro vzácná onemocnění a Evropská léková agentura (European Medicines Agency) 164 léků, přičemž tato čísla zahrnují i léky pro dětská onkologická onemocnění (Tambuyzer et al., 2019). Vývoj léčebných a terapeutických postupů pro vzácná onemocnění je pro farmaceutické společnosti finančně náročný kvůli malému počtu pacientů rozmístěných po celém světě, nedostatku validovaných biomarkerů a epidemiologických dat, limitované klinické expertíze a neznámé patofyziologii onemocnění.

V posledních desetiletích byla na celém světě věnována značná pozornost a úsilí stimulaci výzkumu vzácných onemocnění a vývoje léčivých přípravků. Byly zavedeny specifické regulace usnadňující vývoj léků a založeny mezinárodní organizace, programy, pacientské spolky a sítě pro vzácná onemocnění, jako jsou například americká Národní organizace pro vzácná onemocnění (NORD, National Organisation of Rare Disorders, https://rarediseases.org), Mezinárodní konsorcium pro výzkum vzácných onemocnění (IRDiRC, International Rare Disease Research Consortium, https://irdirc.org), Evropská referenční síť sdružující vysoce specializovaná centra, umožňující vzájemnou spolupráci v diagnostice a léčbě vzácných onemocnění a péči o pacienty (European Reference Networks, https://ec.europa.eu/health/ern_en), Evropská organizace pro vzácná onemocnění sdružující evropské pacientské organizace (EURORDIS, https://www.eurordis.org) nebo Česká asociace pro vzácná onemocnění sdružující české pacientské organizace (ČAVO, http://www.vzacna-onemocneni.cz). Tyto iniciativy usnadňují mezinárodní spolupráci mezi lékaři, výzkumníky a pacienty, která je klíčová pro zlepšení stávající situace na poli vzácných onemocnění.

2 Technologie a metody

Většina vzácných onemocnění je způsobená mutacemi jednotlivých genů či funkčně významných genetických elementů. Mezi dřívější metody identifikace kauzálních genů patřilo poziční klonování a funkční klonování. Metody pozičního klonování se používají v případech, kdy podstata onemocnění není známá. Pomocí polymorfních markerů vyhledáme úseky genomu, které segregují s daným onemocněním/fenotypem a v dané kandidátní oblasti poté sekvenujeme geny. Metody funkčního klonování jsou založené na vyhledání abnormálního metabolitu, určení aktivity příslušného enzymu, jeho izolaci, určení sekvence a následném vyhledání příslušného genu.

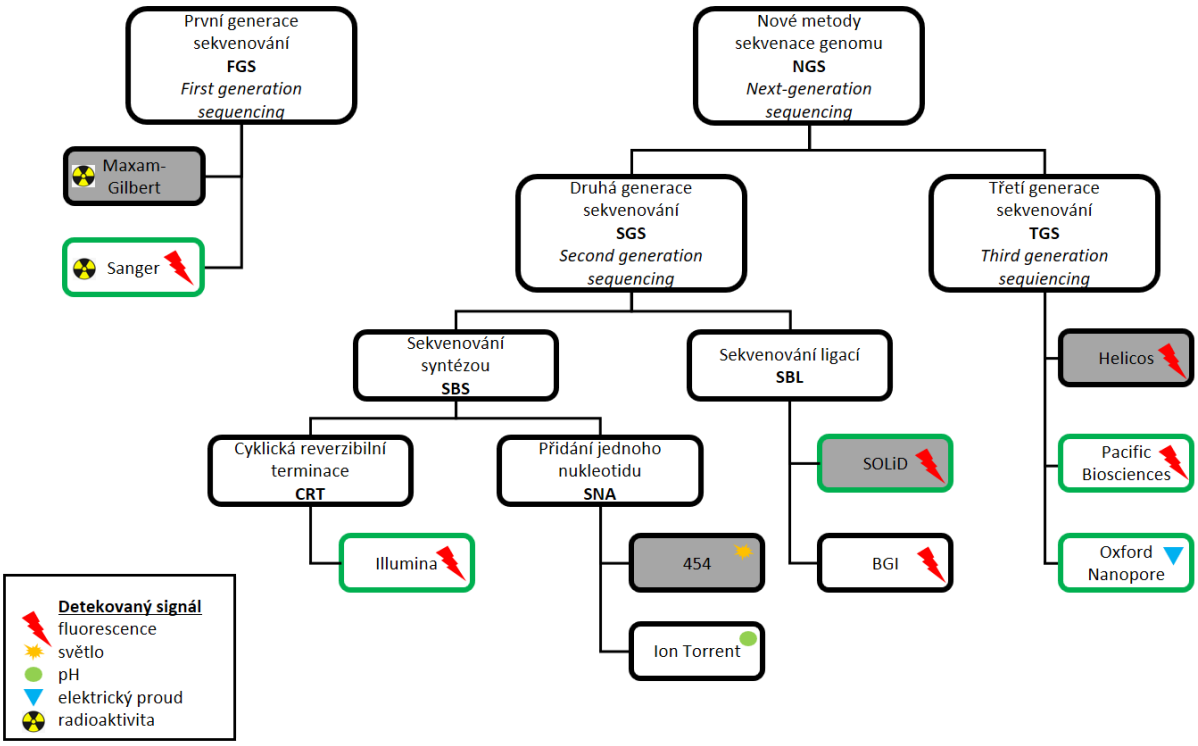
2.1 Sekvenační technologie

Zásadním milníkem ve studiu vzácných onemocnění byl rok 1977, kdy byly nezávisle na sobě vyvinuty dvě odlišné metody sekvenace DNA – Sangerova enzymatická metoda (Sanger et al., 1977) a Maxam-Gilbertova chemická metoda (Maxam and Gilbert, 1977). Jednalo se o sekvenační metody první generace (FGS, First Generation Sequencing). Tyto metody, poprvé v historii, umožnily zkoumání genetické informace všech živých organismů. Sekvenovaly se geny, genomové oblasti a nakonec i celé genomy. Limitacemi těchto technologii je vysoká cena sekvenování a možnost číst maximálně stovky vzorků najednou o délce ~1000 párů bází (bp).

Z těchto důvodů došlo k rozvoji NGS, které jsou díky masivní paralelizaci výrazně levnější a rychlejší. V jednom běhu jsou čteny až miliardy molekul. Tyto technologie se proto také někdy nazývají masivně paralelním sekvenováním (Massively Parallel Sequencing). Existují dva typy NGS metod. První z nich jsou metody vyžadující předchozí amplifikaci čtených fragmentů, jejichž délkový limit je v závislosti na použité platformě 35–500 bp. Protože je délka čtených fragmentů kratší než u Sangerova sekvenování, říká se těmto metodám také sekvenování krátkých čtení (Short Read Sequencing), nebo také sekvenační metody druhé generace (SGS, Second Generation Sequencing).

Třetí generací sekvenování (TGS, Third Generation Sequencing) jsou metody, které nevyžadují předchozí amplifikaci vstupního materiálu. Umožňují čtení jednotlivých molekul o délce až miliónů bází v reálném čase, díky čemuž se těmto metodám také říká sekvenování dlouhých čtení (Long Read Sequencing), nebo jednomolekulové sekvenování (Single Molecule Sequencing).

Názvosloví a dělení jednotlivých sekvenačních metod není v literatuře jednoznačné. Obrázek 1 shrnuje názvosloví použité v této disertační práci.



Obrázek 1: Schéma zobrazující dělení sekvenačních technologií. Členění podle generace sekvenování: FGS, NGS, SGS, TGS. Základní sekvenační principy: SBS (Sequencing By Synthesis), SBL (Sequencing By Ligation), CRT (Cyclic Reversible Termination), SNA (Single Nucleotide Addition) a konkrétní sekvenační platformy firem Illumina, 454, Ion Torrent, BGI, Helicos, Oxford Nanopore, Pacific Biosciences a platforma SOLiD. Zelené ohraničení – platformy, které jsem v průběhu svého doktorského studia využila. Šedá výplň – platformy, které se již nevyužívají a nejsou komerčně dostupné.

2.1.1 První generace sekvenování - FGS

Maxam-Gilbertova metoda sekvenování je založená na chemické modifikaci jednotlivých bází a jejich následném štěpení. DNA je nejprve denaturována a radioaktivně označena na 5′ konci. Reakce probíhá ve čtyřech odlišných zkumavkách, přičemž v každé z nich dochází ke specifické chemické modifikaci konkrétních bází a následnému náhodnému štěpení v místech modifikací. Vzniklá směs fragmentů je poté rozdělena v polyakrylamidovém gelu v závislosti na délce fragmentů. Sekvence je odečtena pomocí pozice jednotlivých bází ve všech čtyřech reakcích.

Sangerova metoda sekvenování využívá procesu replikace DNA. Reakce vyžaduje genomovou DNA, radioaktivně značený primer komplementární k začátku sekvenovaného místa, deoxynukleotidtrifosfáty (dNTP) a pro každou ze čtyř zkumavek vždy jeden

z dideoxynukleotidtrifosfátů (ddNTP). DNA je nejprve denaturována, primer nasedne na začátek sekvenovaného místa a dNTP a ddNTPs jsou pomocí DNA polymerázy postupně připojovány na 3′-OH skupinu ribózy. Ve chvíli, kdy dojde k začlenění ddNTP, je syntéza řetězce ukončena, protože ddNTP nemají OH skupinu a další elongace je tím blokována. Vzhledem k tomu, že k začleňování ddNTP dochází náhodně, vznikne v každé reakci směs různě dlouhých fragmentů, zakončených na různých pozicích ddNTP. Stejně jako u Maxam-Gilbertovy metody je vzniklá směs elektroforeticky separována, vizualizována a poté je odečtena výsledná sekvence. Pomocí této metody osekvenoval Frederick Sanger historicky první genom, genom viru phiX174 (Sanger et al., 1977). Za tuto metodu obdržel v roce 1980 Nobelovu cenu.

Sangerova enzymatická metoda nevyžadovala manipulaci s tolika nebezpečnými chemickými a radioaktivními látkami jako Maxam-Gilbertova chemická metoda, a proto se stala na dalších téměř 30 let metodou nejrozšířenější a stále má své místo v téměř všech molekulárně genetických laboratořích. Sangerova metoda byla dále vylepšována – radioaktivní značení primeru bylo nahrazeno fluorescenčním značením ddNTPs, byly vyvinuty automatické kapilární sekvenátory umožňující paralelní sekvenování stovek vzorků najednou a místo klonování DNA v bakteriálních vektorech se začalo využívat polymerázové řetězové reakce (PCR, Polymerase Chain Reaction; Saiki et al., 1985). Tato vylepšení Sangerovy metody umožnila sekvenaci kompletních genomů, včetně toho lidského, který byl dokončen v roce 2003. Jeho přečtení trvalo 13 let a stálo 2,7 biliónů dolarů (International Human Genome Sequencing Consortium, 2004; Lander et al., 2001).

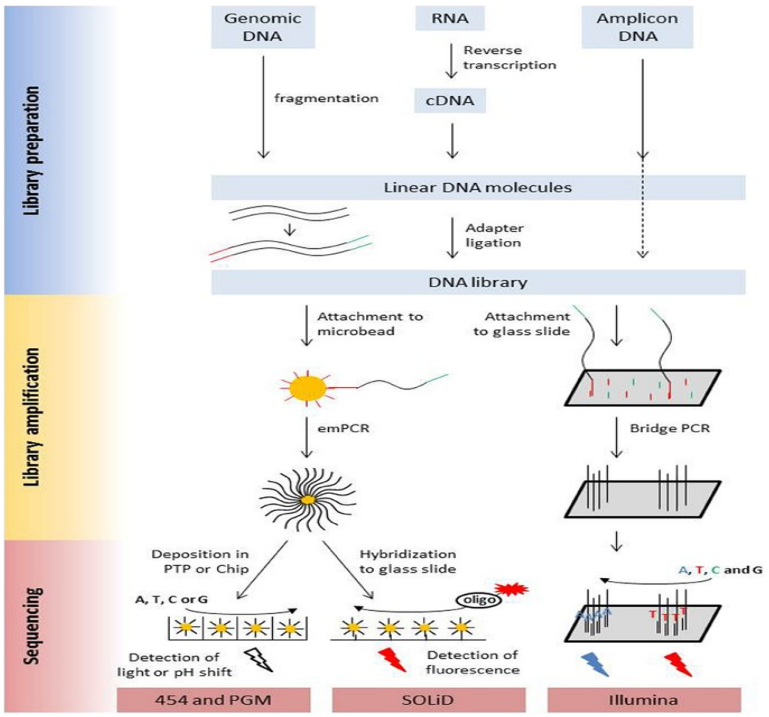
Tlak na vývoj výkonnějších, levnějších a rychlejších sekvenačních metod vyústil v roce 2004 ve vyhlášení grantového projektu Národním institutem pro výzkum lidského genomu (National Human Genome Research Institute, NHGRI). Jeho cílem bylo snížit do 10 let cenu sekvenování lidského genomu na 1000 $. Díky této výzvě došlo ke stimulaci vývoje a komercializaci nových metod sekvenace genomu.

2.1.2 Nové metody sekvenace genomu - NGS

Nové metody sekvenace genomu se liší od první generace sekvenování v několika aspektech: I) došlo k paralelizaci a v jednom běhu jsou najednou sekvenovány až miliardy jednotlivých molekul II) výsledky jsou odečítány přímo a již není potřeba následné elektroforetické dělení fragmentů pro zobrazení přečtené sekvence DNA.

2.1.2.1 Druhá generace sekvenování – SGS

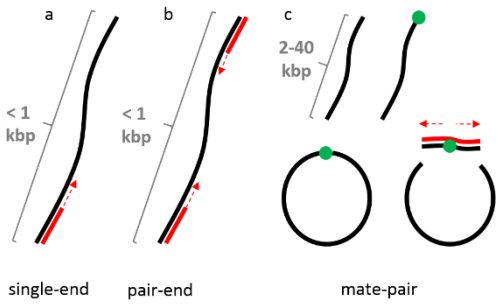
Základní princip všech v současnosti rozšířených SGS sekvenátorů je velmi podobný. Zahrnuje přípravu knihovny, amplifikaci a sekvenaci (Obrázek 2). Volitelným krokem je selektivní obohacení DNA knihovny o vybrané úseky genomu. Konkrétní detaily jednotlivých kroků se liší v závislosti na zvolené sekvenační platformě. V této práci popíši podrobněji pouze sekvenační platformy SOLiD a Illumina, které jsem ve své disertační práci využila.



Obrázek 2: Schéma SGS sekvenování. Příprava knihovny, amplifikace a sekvenace. Příprava knihovny je obdobná pro všechny sekvenační platformy. Dochází ke štěpení DNA a ligaci adaptorů. Amplifikace probíhá buď pomocí emulzního PCR (454, SOLiD, PGM = Ion Torrent), nebo pomocí můstkové amplifikace (Illumina). Sekvenace se pro jednotlivé platformy liší podle sekvenačního principu a detekovaného signálu. Převzato z Knief, 2014.

2.1.2.1.1 Příprava knihovny

V první fázi přípravy knihovny je potřeba vstupní materiál **naštěpit** na kratší fragmenty. Vstupním materiálem může být genomová DNA (gDNA), komplementární DNA (cDNA) nebo PCR produkt. Pro štěpení se využívají metody mechanické (akustická sonikace, hydrodynamické štěpení, nebulizace), enzymatické (restrikční endonukleázy, transponáza, fragmentáza, fragmentace polymerizací; Ignatov et al., 2019) a chemické (hydrolýza DNA teplem s bivalentními ionty kovů). Následuje **enzymatická úprava konců** a **ligace adaptorů** sloužících pro amplifikaci a rozlišení DNA fragmentů odpovídajících jednotlivým vzorků v případě, že jich sekvenujeme více najednou (Head et al., 2014). Existuje



Obrázek 3: Schéma typů sekvenačních knihoven: a) fragmentová – čtená z jedné strany, c) fragmentová – čtená z obou stran, d) mate-pair - čtení obou konců najednou.

několik různých typů knihoven (Obrázek 3). Knihovna fragmentová, kterou je možné sekvenovat buď jen z jedné strany (single-end), nebo z obou stran (pair-end) a knihovna, kdy fragment nejprve cirkularizujeme a poté čteme oba původní konce najednou (mate-pair).

Amplifikace

V závislosti na použité sekvenační platformě je získaná DNA knihovna namnožena pomocí PCR jednou z uvedených metod: můstkovou amplifikací na destičce (Illumina), emulzním PCR na kuličkách (454, SOLiD, Ion Torrent), nebo metodou amplifikace rotujícího kruhu (Rolling-Circle Amplification) v roztoku (BGI).

Metody obohacení DNA

Ač je možné sekvenovat celé genomy, je mnohdy výhodné zaměřit se pouze na určité konkrétní oblasti. Pro selektivní obohacení DNA knihovny o vybrané úseky genomu (cílené sekvenování, sequence capture, target capture) existují různé metody a každá je vhodná k jiným účelům v závislosti na velikosti cílené sekvence, počtu vzorků a finanční dostupnosti (Ballester et al., 2016; Kozarewa et al., 2015). Standardní **PCR** je vhodnou metodou v případě, kdy je počet oblastí zájmu malý. **Long-range PCR** (LR-PCR) lze použít v případě, kdy potřebujeme získat relativně dlouhou, nepřerušenou molekulu DNA, což je vhodné pro TGS, které umožňují číst molekuly o délce až 2 Mb (Payne et al., 2019). Pokud je oblastí zájmu více, je možné využít metod založených na multiplexním PCR. Nevýhodou metod založených na PCR je to, že v průběhu amplifikace dochází k zanesení chyb v důsledku chybovosti DNA polymerázy. **Hybridizační metody** využívají oligonukleotidové próby, které jsou komplementární k sekvencím v oblastech zájmu. K obohacení může docházet jak na čipu, tak v roztoku. U čipových technologií jsou próby imobilizovány na čipu a vychytávají požadované fragmenty DNA. Nezachycené fragmenty jsou z čipu odmyty a poté jsou uvolněny zachycené cílené sekvence. Principy obohacení v roztoku využívají biotinem značené próby (DNA či RNA), které jsou poté z roztoku společně se zachycenými fragmenty vychytány pomocí streptavidinem obalených magnetických kuliček. Metody obohacení založené na hybridizaci jsou nejvhodnější metodou pro cílené NGS sekvenování, protože mohou cílit až na tisíce různých oblastí, nedochází k zanášení chyb tak, jako u PCR a poskytují dobrou reprodukovatelnost (Mamanova et al., 2010).

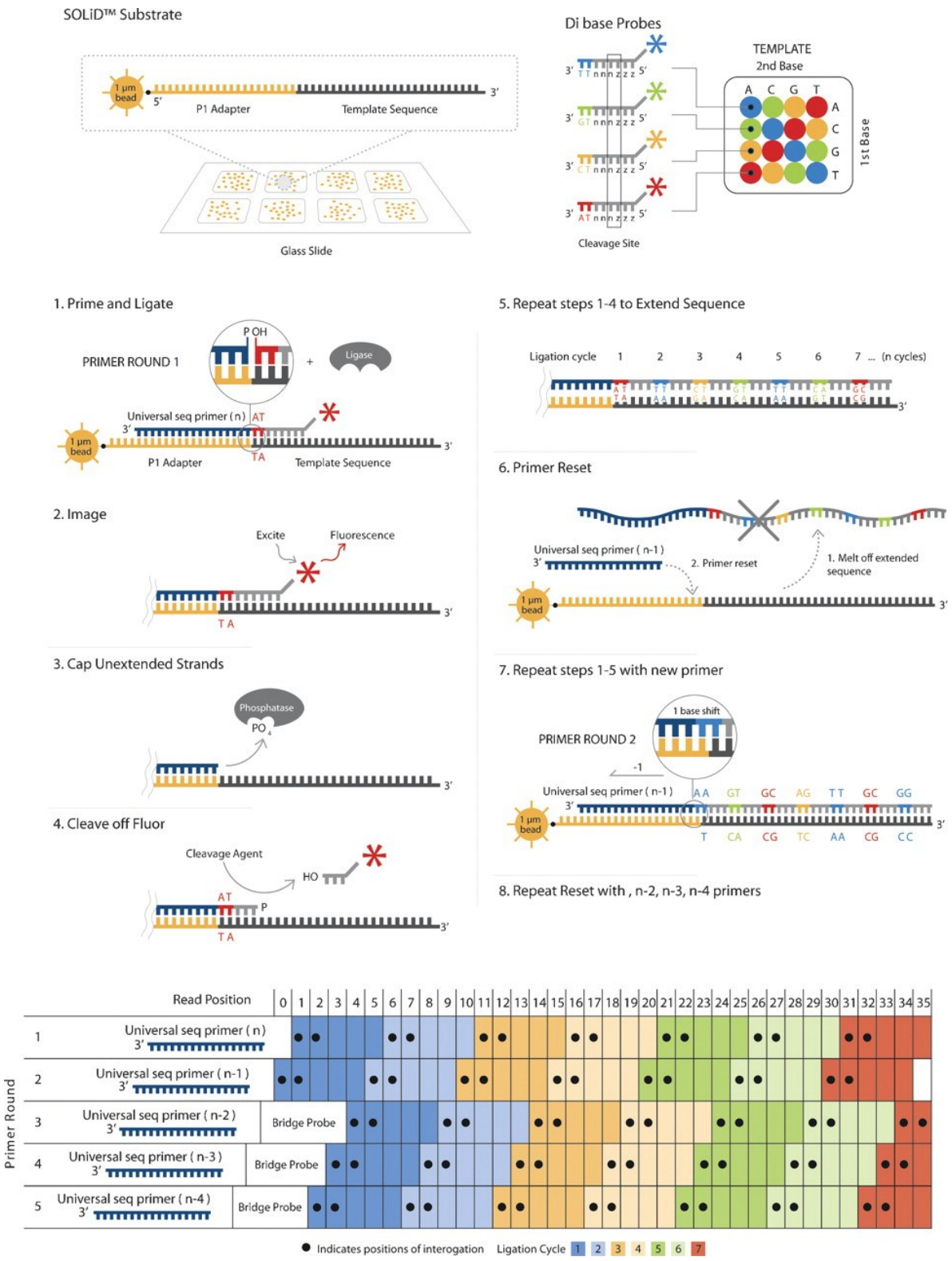
Cílené sekvenování je velmi efektivním způsobem, jak snížit sekvenační náklady a zpracovat vzorky od více pacientů najednou (multiplexing) v kratším čase. Díky tomu se SGS sekvenování postupně rozšířilo i do menších laboratoří a stalo se nedocenitelným nástrojem při odhalování kauzálních variant (Mamanova et al., 2010). Mezi nejčastější aplikace patří **exomové sekvenování** (Ng et al., 2009), kdy se sekvenují všechny kódující oblasti genomu (exony) a **panelové sekvenování**, při kterém se sekvenuje vybraný set několika desítek až tisícovek genů, či jejich exonů.

2.1.2.1.2 Sekvenování

Sekvenační principy SGS můžeme rozdělit do dvou hlavních kategorií: sekvenování ligací a sekvenování syntézou (Obrázek 1). **Sekvenování ligací (SBL)** je založené na cyklické hybridizaci fluorescenčně značených krátkých oligonukleotidových prób, které jsou DNA ligázou připojovány vždy k předchozí próbě. Následně dochází k uvolnění fluoroforu, který odpovídá specifické bázi či bázím na určitém místě próby. Tento princip využívají sekvenátory SOLiD (Valouev et al., 2008) a sekvenátory firmy BGI (dříve Complete Genomics; Drmanac et al., 2010). **Sekvenování syntézou (SBS)** je založené na postupném začleňování nukleotidů pomocí DNA polymerázy a detekci specifického signálu (fluorescence, změna pH, světlo), k jehož uvolnění dojde po začlenění nukleotidu do rostoucího řetězce. Sekvenování syntézou můžeme dále dělit na metody pracující na principu **přidávávání jednoho nukleotidu (SNA)**, kterou využívají sekvenátory firmy 454 (Margulies et al., 2005) a IonTorrent (Rothberg et al., 2011) a metody založené na **cyklické reverzibilní terminaci (CRT)**, kterou využívají sekvenátory firmy Illumina (Bentley et al., 2008).

2.1.2.1.3 SOLiD

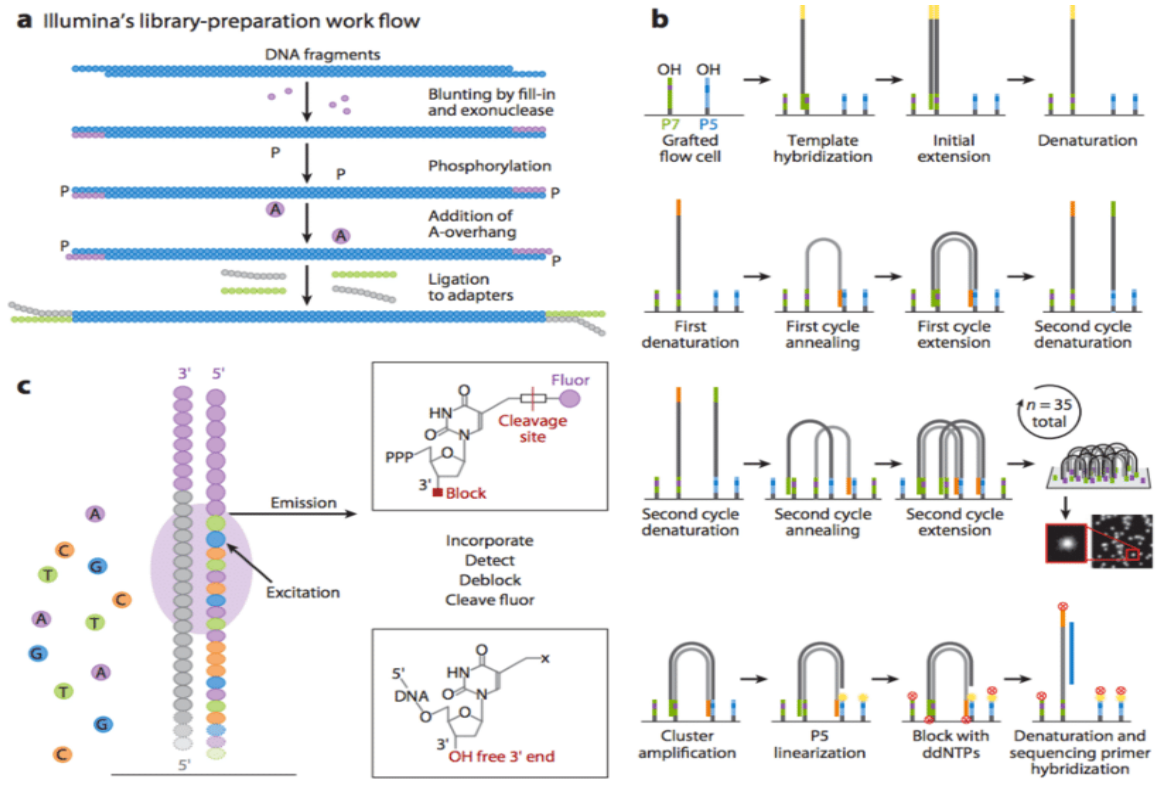
Sekvenátory firmy Life Technologies pracují na principu sekvenování ligací – odtud pochází i jejich název - SOLiD (Sequencing by Oligonucleotide Ligation and Detection; Obrázek 4). Vstupní DNA je rozštěpena na fragmenty dlouhé 150-200 bp. Na každý z konců je ligován odlišných adaptor – P1 a P2. Amplifikace probíhá pomocí emulzního PCR. Do zkumavky obsahující roztok oleje ve vodě je vložena DNA knihovna, DNA polymeráza, primery P1 a P2 a kuličky, které nesou oligonukleotidy komplementární k primeru P1. V ideálním případě by v každé vodní kapičce měla být právě jedna kulička s jednou molekulou DNA, která je pomocí PCR namnožena. Kuličky jsou následně kovalentně navázány na sekvenační destičku a na každé z nich dochází k oddělené sekvenační reakci. V prvním kroku sekvenační reakce dojde k navázání univerzálního primeru na adaptor P1. Za univerzální primer nasedá jedna ze sekvenačních prób, rozpoznávajících vždy dvoubázový motiv a nesoucí jemu odpovídající fluorescenční značku. Pomocí ligace dojde ke spojení próby a předchozího oligonukleotidu. Dochází k detekci fluorescence a odštěpení posledních třech bází spolu s fluorescenční značkou. Podle délky čtené sekvence následuje dalších pět až patnáct ligačních cyklů. Poté dojde k resetování vlákna, navázání univerzálního primeru vždy o jeden nukleotid kratší a následuje další kolo ligací. Následně dochází k přečtení sekvence s využitím kódování pomocí dvou bází. Jednou flurescenční značkou jsou označeny vždy čtyři různé dvojice rozpoznávaných bází. Abychom mohli identifikovat konkrétní bázi, potřebujeme znát vždy bázi předcházející. Výhodou tohoto systému je to, že je možné odlišit chybu systému (záměna jedné barvy oproti referenční sekvenci) od varianty reálné (záměna dvou barev oproti referenční sekvenci). Naopak nevýhodou SOLiD sekvenování je krátká délka čtení (maximálně 75 bází) a časová náročnost ligačních cyklů. Z těchto důvodů byl vývoj sekvenátorů SOLiD ukončen.



Obrázek 4: Princip sekvenátorů SOLiD. Vlevo) Zobrazení sekvenačního sklíčka, na kterém jsou kovalentně navázány kuličky s DNA fragmenty, amplifikovanými pomocí emulzního PCR. Vpravo) sekvenační próby rozpoznávající dvoubázový motiv a jim odpovídající fluorescenční značky. 1) univerzální primer nasedá na adaptor knihovny, první próba rozpoznávající motiv TA hybridizuje za univerzální primer a je k němu připojena ligázou. 2) detekce fluorescence 3) ošetření neprodloužených fragmentů fosfatázou 4) odstranění posledních třech bází s fluoroforem 5) opakování kroků 1-4 6) odstranění všech prób a navázání nového univerzálního primeru, který je o 1 nukleotid kratší 7) opakování kroku 1-5 8) odstranění všech prób a navázání nového univerzálního primeru postupně o 2, 3 a 4 nukleotidy kratším. Dole) Schéma ligačních cyklů a rozpoznávaných pozic. Převzato z Valouev et al., 2008

2.1.2.1.4 Illumina

Sekvenátory firmy Illumina pracují na principu sekvenování syntézou pomocí cyklické reverzibilní terminace (Obrázek 5). Vstupní DNA je rozštěpena na fragmenty dlouhé ~300 bp. Na každý z konců je ligován odlišný adaptor a DNA je následně amplifikována můstkovou amplifikací. Amplifikace probíhá přímo na sklíčku, jehož povrch je pokryt oligonukleotidy komplementárními k oběma typům adaptorů navázaným na DNA fragmenty v průběhu přípravy knihovny. Denaturované fragmenty jsou hybridizovány každým koncem k jednomu z adaptorů na sklíčku tak, že vytvoří „můstek". K fragmentu je dosyntetizováno druhé vlákno a poté jsou obě vzniklá vlákna oddělena denaturací. Výsledkem jsou dva identické fragmenty navázané na povrchu sklíčka. Tento proces je opakován 35x, dokud nevzniknou tisíce kopií původního templátu umístěných ve shlucích. Při následném sekvenování jsou v každém cyklu přidány čtyři typy odlišně fluorescenčně označených nukleotidů. Jejich 3´- OH konce jsou chemicky inaktivované, aby během cyklu nedošlo k inkorporaci více než jedné báze. Fluorescenční signál inkorporovaných nukleotidů je v jednotlivých shlucích detekován pomocí laserového paprsku a zaznamenán citlivou kamerou. Následuje odstranění fluorescenční značky a odblokování 3′ konce. Tento cyklus se podle délky sekvenovaných čtení opakuje 100 až 300x.



Obrázek 5: Princip sekvenátorů firmy Illumina a) příprava knihovny: enzymatické ošetření konců, ligace adaptorů b) můstková amplifikace: hybridizace templátů na sekvenační sklíčko, úvodní extenze, první denaturace, první cyklus annealingu, první cyklus extenze, druhý cyklus denaturace, druhý cyklus annealingu, druhý cyklus extenze, amplifikace shluků je opakovaná 35x, linearizace P5 adaptoru, zablokování oligonukleotidů pomocí ddNTPs, denaturace a hybridizace sekvenačního primeru c) sekvenování syntézou pomocí cyklické reverzibilní terminace: začlenění jednoho ddNTP, detekce fluorescence, odblokování 3'-OH a fluoroforu. Převzato z Mardis, 2013

2.1.2.1.5 Přínos a limity

S rozvojem a rozšířením SGS, klesající cenou sekvenování a rostoucími znalostmi o lidském genomu, byl nastolen **nový koncept identifikace onemocnění** podmiňujících genů, založený buď na porovnávání genomových sekvencí jedinců (postižení vs. zdraví) nebo na porovnávání genomových sekvencí mezi tkáněmi (nádorová vs. normální). Již se nesoustředíme na konkrétní, malé části genomu tak jako dříve (metody pozičního a funkčního klonování), ale čteme genom celý (či jeho velkou část) a ten porovnáváme s genomy zdravých lidí/tkání a hledáme vzácné varianty ve funkčně důležitých oblastech genomu, které by mohly vysvětlit studovaný fenotyp. Tento přístup vedl k určení nových kauzálních mutací u řady onemocnění. Nicméně u mnoha onemocnění stále zůstává jejich molekulární podstata neznámá. Jedním z důvodů je, že mnoho oblastí lidského genomu nelze sekvenačními technologiemi první a druhé generace správně analyzovat. Jedná se například o vysoce homologní oblasti, oblasti s vysokým obsahem GC či AT bází, repetitivní oblasti, expanze tandemových repetic, segmentální duplikace, transpozony či geny mající pseudogeny a homologní geny, je obtížné správně určit fázi haplotypu a identifikovat nové mRNA izoformy.

V některých případech je možné přijít s alternativním bioinformatickým řešením, díky kterému je možné ze SGS dat získat požadované informace, ale v jiných případech je potřeba využít alternativních technologií – například NGS technologií třetí generace.

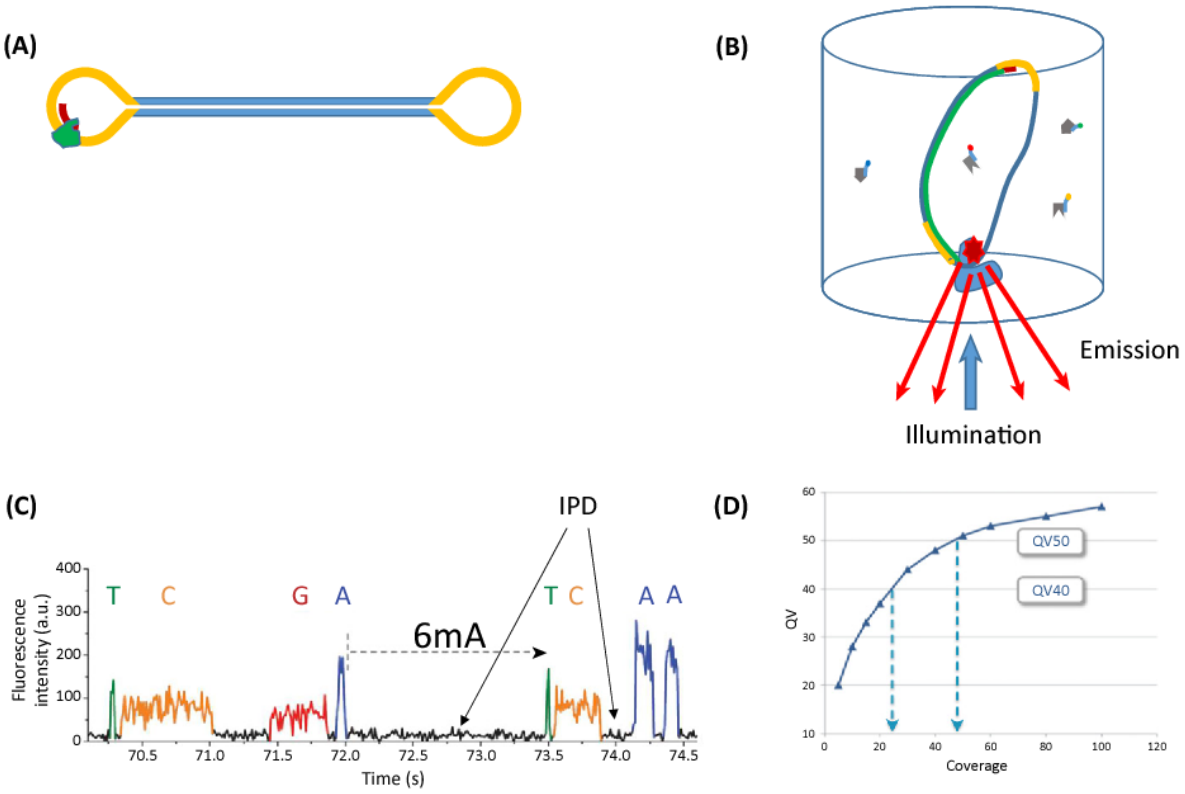
2.1.2.2 Třetí generace sekvenování – TGS

Metody sekvenování třetí generace jsou specifické tím, že umožňují čtení jednotlivých molekul v reálném čase. Oproti SGS tedy odpadá PCR amplifikační krok, ve kterém dochází k zanesení chyb, znevýhodňuje GC bohaté oblasti a odstraňuje epigenetické modifikace.

Prvním komerčně dostupným (2009) sekvenátorem čtoucím jednotlivé molekuly bez amplifikace byl přístroj Helicos od firmy Helicos Biosciences (Pushkarev et al., 2009) založený na stejném principu jako sekvenátory od firmy Illumina, pouze vynechávající můstkovou amplifikaci. Vzhledem k tomu, že se jednalo o metodu relativně pomalou, drahou a umožňující čtení pouze o délce 32 bp, neměla dlouhého trvání. Následovalo komerční představení dalších dvou metod, které již umožňovaly čtení dlouhých molekul.

2.1.2.2.1 Pacific Biosciences

První z nich byla technologie **SMRT** (Single-Molecule Real-Time, Eid et al., 2009), komerčně představená firmou Pacific Biosciences (PacBio) v roce 2011 (Obrázek 6). Sekvenační knihovna je připravena ligací vlásenkových adaptorů na dvouvláknovou molekulu DNA (dsDNA), čímž dojde k vytvoření cirkulární molekuly, která se nazývá SMRT bell. Následuje přidání polymerázy navázané na sekvenační primer, který hybridizuje na adaptor. Směs těchto komplexů je nanesena na sekvenační čip a pomocí difuze dojde k navázání vždy jednoho komplexu do jedné sekvenační jamky pomocí komplexu streptavidin-biotin. Tato metoda je založena na sekvenování syntézou – pomocí DNA polymerázy jsou začleňovány fluorescenčně značené nukleotidy uvolňující fluorescenční signál po excitaci laserem. Emitovaná fluorescence je snímána a nahrávána citlivou kamerou v reálném čase. Je zaznamenána nejen barva fluorescence, ale i doba potřebná k inkorporaci nukleotidu – doba mezi pulzy (IPD, InterPulse Duration). Z prodloužené IPD je možné určit epigenetické modifikace. Chybovost čtení jednotlivých molekul je vysoká, přibližně 13 %. Pro kratší fragmenty je možné využít přesnějšího módu sekvenování. Vzhledem k tomu, že knihovna je cirkulární, může polymeráza číst templát několikrát dokola. Tato několikanásobná čtení jsou poté bioinformaticky zkombinována do takzvané cirkulární konsensuální sekvence (CCS, Circular Consensus Sequence) a platí, že čím vícekrát je jednotlivá molekula přečtena dokola, tím větší je přesnost získaných čtení. Při ~25 cyklech je přesnost čtení srovnatelná s přesnosti čtení na sekvenátorech firmy Illumina. Délkový limit této technologie je dán procesivitou polymerázy. Průměrná délka získaných čtení je 10–16 kb (Ardui et al., 2018), což je až o dva řády více než u SGS.

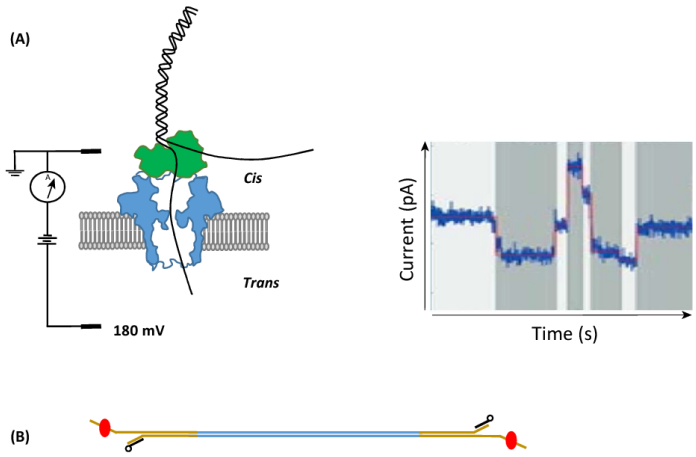


Obrázek 6: Princip SMRT sekvenování firmy PacBio. A) Sekvenační knihovna sestává z dsDNA (modrá), na kterou jsou naligovány vlásenkové adaptory (žlutá). Tím vzniká cirkulární molekula. Následně je přidána polymeráza (zelená) a sekenační adaptor (červená) B) Schéma sekvenační jamky, na jejíž dno je přichycen komplex sekvenační knihovny. Polymeráza inkorporuje fluorescenčně značené nukleotidy. Fluoresecence je zaznamenávána v reálném čase. C) Ze záznamu fluoresecence je možné vidět nejen jaká báze byla inkorporována, ale i jak dlouho inkorporace trvala (IPD, černá). Jednotlivé epigenetické modifikace vedou k charakteristickému prodloužení IPD. D) Díky tomu, že je knihovna cirkulární je možné fragment číst vícekrát dokola. Se zvyšujícím se počtem čtení se zvyšuje přesnost čtení. Již při 25 cyklech dosahuje přesnost čtení QV40 (Q skóre, viz. kapitola 2.2.1.1), což odpovídá kvalitě sekvenování na sekvenátorech Illumina. Převzato z van Dijk et al., 2018.

2.1.2.2.2 Oxford Nanopore

Druhou metodou sekvenování třetí generace je technologie **nanopórového sekvenování** (Clarke et al., 2009), komerčně představená v roce 2015 firmou Oxford Nanopore Technologies (ONT). Sekvenační cela sestává z dvou komor naplněných iontovým roztokem a oddělených membránou, ve které jsou zasazeny proteinové póry. Aplikace napětí způsobuje iontový tok, díky kterému jsou jednotlivé molekuly translokovány skrz póry. Při translokaci jednotlivých bází pórem dochází ke specifickým změnám elektrického proudu, které jsou pro každý jednotlivý nanopór snímány, zaznamenávány a následně je odečítána sekvence procházející molekuly (Obrázek 7A). Sekvenační knihovna je připravena z dsDNA, kdy jsou na oba konce naligovány adaptory, které mají na 5′ konci připevněný motor protein. Ten slouží ke zpomalení průchodu molekuly skrz nanopór. Na 3′ konci je navázán oligonukleotid s cholesterolem, který slouží k navázání molekuly na membránu, čímž zvyšuje efektivitu sekvenování (Obrázek 7B).

Výhodou této technologie je, že délkový limit je dán délkou molekul v sekvenační knihovně. Není limitován technologií samotnou, jako je tomu u technologie SMRT. Při izolaci DNA a přípravě knihovny je potřeba minimalizovat její fragmentaci. Z toho důvodu dochází například k návratu k izolaci DNA fenol-chloroformovou nebo vysolovací metodou, při kterých na rozdíl od izolace DNA na kolonkách nedochází k tak výrazné fragmentaci. Nejdelší reportované čtení mělo délku > 2 Mb (Payne et al., 2019). Naopak nevýhodou této technologie je vysoká chybovost čtení jednotlivých molekul (~15 %) a absence cirkulární molekuly, která může být čtená vícekrát za sebou pro zvýšení přesnosti čtení. V poslední době ONT představil tzv. $1D^2$ systém umožňující čtení obou vláken molekuly, díky kterému dochází ke snížení chybovosti na 3 %.



Obrázek 7: Princip nanopórového sekvenování. A) Dvě komory (Cis a Trans) naplněné iontovým roztokem jsou rozdělené lipidovou membránou, ve které jsou zasazeny proteinové póry (světle modrá). Nukleová kyselina (černá) je v průběhu translokace skrz pór zpomalována a odvíjena motor proteinem (zelená) tak, že prochází pouze jedno vlákno. Změny proudu v důsledku bází procházejících skrz pór jsou zaznamenávány v reálném čase. B) Schéma sekvenační knihovny. Fragment dsDNA (světle modrá) s navázanými adaptory (hnědá). 5' konec má na sobě navázaný motor protein (červená) a 3' konec oligonukleotid s cholesterolem (černá). Převzato z van Dijk et al., 2018.

2.1.3 Výběr vhodné sekvenační technologie

Každá z výše zmíněných technologií má své výhody a limitace. Výběr konkrétní z nich závisí na našich specifických potřebách a možnostech. Při zvažování může pomoci porovnání základních charakteristik jako je délka čtení, kapacita přístroje (počet čtení v jednom běhu), GC bias (nižší či vyšší pokrytí v GC bohatých oblastech), chybovost, primární typ chyb a možnost detekce modifikovaných bází (Tabulka 1).

Technologie	Délka čtení N50	Počet čtení/běh	GC bias	Chybovost (%)	Primární typ chyb	Detekce modifikace bází
FGS	< 1 kb	96	Ano	0,1-1	substituce	ne
SGS	50-500 bp	10^6-10^9	Ano	~0,1	substituce	ne
TGS	1-100 kb	10^5-10^6	žádný/nízký	3-15	inserce, delece	ano

Tabulka 1: Přehled charakteristik sekvenátorů první, druhé a třetí generace. Převzato a přeloženo z Ameur et al., 2019

Dále je vhodné zvážit pořizovací náklady, náklady na provoz, náročnost přípravy vzorku, velikost sekvenátoru, možnost jeho přenášení a další faktory. Kombinace uvedených charakteristik rozhoduje o vhodnosti jednotlivých technologií pro konkrétní aplikace (Tabulka 2).

Aplikace	Optimální technologie	Důvod
strukturní varianty	TGS	dlouhá čtení překlenující zlomy
mozaiky a minoritní varianty	SGS/PacBio	nízká chybovost (u PacBio pouze CCS čtení)
tandemové repetice	TGS	dlouhá čtení umožňují pročíst repetitivní sekvence po celé délce
odlišení pseudogenů	TGS	dlouhá čtení umožňují pročíst celou oblast
sekvenování lidského exomu a genomu	SGS	vysoká kapacita sekvenátorů umožňuje rutinní sekvenování mnoha lidských genomů v jednom běhu
určení haplotypu	TGS	dlouhá čtení umožňují přímé fázování
de novo sestavování genomů	TGS	dlouhá čtení umožňují dosáhnout vyšších kvality *de novo* sestavení
genomové asociační studie	SGS	vysoká kapacita za nízkou cenu
sekvenování v terénu	ONT	přenosnost (sekvenátor MinION), nízké pořizovací náklady
epigenetika	TGS	přímá detekce DNA modifikací
detekce RNA modifikací	ONT	přímé sekvenování RNA
sekvenování mikrobiálních genomů	TGS	dlouhá čtení umožňují sestavení do jednoho kontigu
metagenomika	SGS	lepší rozlišení a vysoká kapacita
exprese genů	SGS	lepší rozlišení a vysoká kapacita
RNA izoformy	TGS	dlouhá čtení umožňují pročtení RNA (cDNA) od začátku do konce a tím i identifikaci nových izoforem
ověřování variant u malého počtu vzorků	FGS	jednoduché navržení primerů pro nové cíle, dostupnost, jednodušší analýza

Tabulka 2: Přehled aplikací a vhodných sekvenačních technologií. Převzato, přeloženo a doplněno z Ameur et al., 2019

Hlavními výhodami FGS je nízká chybovost, přesnost čtení, nižší náklady na pořízení přístoje a jednoduché navržení primerů pro nové cíle. Je vhodné například pro ověřování kandidátních variant nalezených pomocí SGS/TGS u dalších členů rodiny, nebo při ověřování úspěšnosti využití metod genového inženýrství, jako je například klonování do bakteriálních vektorů či vytváření mutací pomocí CRISPR-Cas systému. FGS je vhodné v případech, kdy se zaměřujeme na konkrétní malou oblast DNA u malého množství vzorků (≤ 20). Sangerovo sekvenování tedy stále má a nejspíš i bude mít ve výzkumu i diagnostice své místo.

Hlavními výhodami SGS je nízká chybovost spolu s možností sekvenovat až miliardy čtení v jednom běhu přístroje. Mezi nevýhody pak patří GC bias a krátká délka čtení. SGS je vhodné v případě, kdy potřebujeme osekvenovat velké množství DNA. Hodí se tedy pro genomové a exomové sekvenování, metagenomiku a studium genové exprese. SGS jsou v současné době již velmi rozšířené a standardně se využívají nejen ve výzkumu, ale i v diagnostice.

Hlavními výhodami TGS je délka čtení v řádu kilobází, nízký GC bias a možnost detekovat modifikované báze. Mezi nevýhody patří vysoká chybovost a nižší kapacita než u SGS. TGS se tedy hodí pro sekvenování *de novo*, hledání nových RNA izoforem, identifikaci strukturních variant a tandemových repetic, odlišování pseudogenů a detekci modifikovaných bází. TGS jsou stále relativně nové a na rozdíl od SGS ještě neexistují standardizované postupy pro přípravu knihovny, sekvenaci a hlavně analýzu dat.

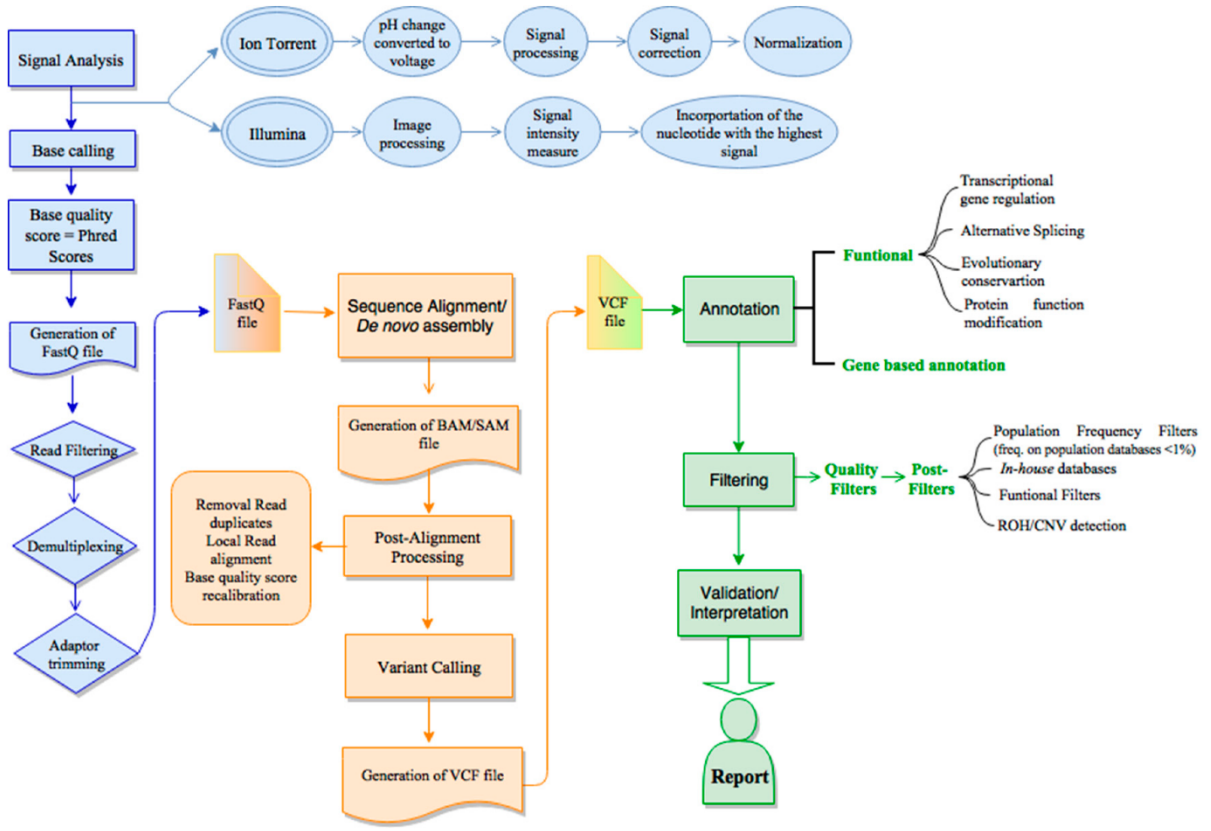
Kromě ONT a PacBio existují i další firmy, které pracují na nových technologiích. Jen budoucnost ukáže, které z nich budou komercializovány a rozšířeny. Uvedu jen některé příklady: GenapSys (www.genapsys.com) vyvíjí malý přenosný sekvenátor o velikosty iPadu, Roche (Genia, www.roche.com) se věnuje nanopórovému sekvenování a Stratos Genomics (www.stratosgenomics.com) vyvíjí sekvenátor založený na sekvenování expanzí.

2.2 Bioinformatická analýza sekvenačních dat

Nedílnou součástí sekvenačních technologií je následné zpracování získaných dat. Vzhledem k tomu, že NGS sekvenátory produkují velké množství dat (GB až TB), není možné je zpracovávat stejným postupem, jako data ze Sangerova sekvenování. Je nutné přistoupit ke specializovaným bioinformatickým postupům, které mnohdy vyžadují znalost práce v příkazové řádce unixového prostředí, znalost programovacích jazyků a schopnost hledání specifických řešení pro jednotlivé projekty. Vzhledem k vysokým nárokům na výpočetní a úložnou kapacitu probíhá většina výpočtů na tzv. clusterech, umožňujících výpočty paralelizovat a tím pádem zrychlit. Analýza genomu může v závislosti na typu úlohy a způsobu zpracování dat trvat tisíce hodin strojového času, což v závislosti na stupni paralelizace představuje několik dní až týdnů (Puckelwartz et al., 2014). V následujících kapitolách představím základní principy bioinformatické analýzy NGS dat, primárně zaměřené na data získaná ze sekvenátorů SGS. Budu se věnovat jednak základním krokům bioinformatické analýzy dat, tak následně analýzám speciálním.

2.2.1 Základní kroky bioinformatické analýzy

Základní bioinformatickou analýzu můžeme rozdělit na primární, sekundární a terciální (Obrázek 8).



Obrázek 8: Bioinformatická analýza dat se dělí na prmární (modrá), sekundární (červená) a terciální (zelená). Podrobnosti k jednotlivým krokům popisují následující kapitoly. Převzato z Pereira et al., 2020

2.2.1.1 Primární analýza

Primární analýza zahrnuje rekonstrukci sekvencí z detekovaných signálů (fluorescence, světlo, změna pH či elektrického proudu) zaznamenaných v podobě obrázků či videí a jejich uložení do textové podoby (tzv. hrubá sekvenační data). Data jsou ve většině případů uložena ve formátu FASTQ (Cock et al., 2010). FASTQ soubor obsahuje sekvence čtených úseků (A, C, T, G), spolu s číselným vyjádřením jejich kvality pomocí Q skóre, které je také někdy označováno jako Phred skóre či QV (Ewing and Green, 1998). Q skóre je záporný dekadický logaritmus pravděpodobnosti nesprávného určení konkrétní báze (P). Tedy $Q = -10 * \log_{10}(P)$. Pokud je například pravděpodobnost nesprávného určení dané báze 1/1000, je jí přirazena hodnota kvality 30. Vyšší hodnoty Q skóre vyjadřují nižší pravděpodobnost špatného určení báze a naopak. Dalším krokem je **kontrola kvality dat** ve FASTQ souborech, například

v programu FastQC (Andrews, 2010), zahrnující statistiky jako je počet neurčených bází (označených písmenem N), poměr zastoupení jednotlivých bází, distribuce délek čtení, průměrná kvalita bází dle sekvenačního cyklu, počet duplikovaných sekvencí či sekvencí, které jsou nadměrně zastoupeny. Následuje přiřazení čtení jednotlivým vzorkům (**demultiplexing**) v případě, že jich bylo sekvenováno více v jednom běhu a volitelný krok ořezání bází (**trimming**) s nízkou kvalitou a sekvencí odpovídajících adaptorům použitých v přípravě knihovny.

2.2.1.2 Sekundární analýza

Upravené FASTQ soubory jsou nadále zpracovávány v **sekundární analýze**, jejímž cílem je sestavení sekvencí do původní podoby. To může probíhat buď mapováním (mapping) získaných čtení na referenční sekvenci, nebo sestavováním sekvence *de novo* (*de novo* assembly) v případě, že neznáme referenční sekvenci studovaného organizmu (není tématem této práce). Jedná se o výpočetně nejnáročnější a zároveň nejdůležitější krok bioinformatické analýzy (Flicek and Birney, 2009).

Cílem **mapování** je přiřadit jednotlivé sekvence z FASTQ souborů ke správným místům referenční sekvence, a to s co největší přesností a rychlostí. Vzhledem k tomu, že každý genom má specifické odchylky od referenční sekvence, mapování je založené na hledání maximální, nikoliv přesné shody. Existuje mnoho algoritmických přístupů, které se mezi sebou dají různými způsoby kombinovat. Základní princip většiny programů však zůstává obdobný. Mapování probíhá ve dvou krocích (Flicek and Birney, 2009). V první fázi je pomocí rychlých, ale méně přesných heuristických algoritmů (Burrows-Wheelerova transformace, hašovací tabulky) pro každé čtení identifikováno několik kandidátních oblastí. Ve druhé fázi je pro identifikaci nejlepšího možného zarovnání využito přesnějších, ale výpočetně náročných algoritmů (optimální lokální zarovnávání - Smith-Waterman, optimální globální zarovnávání - Needelman-Wunsch). Příklady mapovacích programů jsou BWA (Li and Durbin, 2009), GEM (Marco-Sola et al., 2012) a Novoalign (Novocraft, Malaysia).

Výsledky jsou uloženy ve formátu SAM (Sequence Alignment/Map; Li et al., 2009), který obsahuje sekvenci čtení, chromosomální pozici, kvalitu mapování a informace o nalezených rozdílech oproti referenční sekvenci. Pro zrychlení dalších kroků analýzy jsou SAM soubory převedeny do binární podoby, tzv. BAM formátu (Binary Alignment/Map), **setříděny** podle chromosomální souřadnice a **indexovány.** Následuje **odstranění PCR duplikátů**, tedy čtení,

která mají stejnou sekvenci a délku. Volitelnými kroky jsou **rekalibrace kvality bází**, díky které je možné pomocí metod strojového učení a listu známých variant odstranit systematické chyby sekvenování a **lokální zarovnání čtení** kolem insercí a delecí. Klíčovými nástroji jsou SAMtools (Li et al., 2009), picard (http://broadinstitute.github.io/picard) a GATK (McKenna et al., 2010).

Posledním krokem je **detekce variant** (variant calling, genotypování). V tomto kroku identifikujeme všechny odchylky, kterými se analyzovaný vzorek liší oproti referenční sekvenci. Standardně detekujeme jednonukleotidové polymorfismy (SNPs, Single Nucleotide Polymorphisms) a krátké inserce a delece (indely). Je ale možné detekovat i strukturní varianty (kapitola 2.2.2.3) či expanze tandemových repetic (kapitola 2.2.2.4). Algoritmy pro detekci SNPs a indelů jsou založeny na různých principech, mezi něž patří detekce podle počtu čtení obsahujících báze odlišné od reference, metody založené na Bayesovské či pravděpodobnostní statistice a nově i metody strojového učení. Mezi nejznámější nástroje v daných kategoriích patří SAMtools/BCFtools (Li et al., 2009), Freebayes (Garrison and Marth, 2012), GATK (Poplin et al., 2017) a DeepVariant (Poplin et al., 2018). Jednotlivé varianty jsou definovány chromosomální pozicí, kvalitou, zastoupením čtení se sekvenační změnou a bez ní. Výsledky jsou ukládány v souboru VCF (Variant Calling Format; Danecek et al., 2011).

2.2.1.3 Terciální analýza

Cílem **terciální analýzy** je dodat získaným variantám biologický smysl. Použité metody a nástroje se liší v závislosti na studovaném projektu. Ve studiu molekulární podstaty vzácných onemocnění je terciální analýza založena na anotaci variant ve VCF souboru pomocí informací z rozličných databází, které jsou následně využity pro filtrování, prioritizaci a vizualizaci kandidátních variant.

Anotace je prvním a klíčovým krokem terciální analýzy a můžeme ji rozdělit na anotaci variant (tzv. funkční anotace) a anotaci genů, ve kterých se varianty nacházejí. Pomocí anotačních programů automaticky přiřazujeme k jednotlivým variantám a genům ve VCF souboru informace o jejich biologickém kontextu. Nejprve potřebujeme určit, v jakém genu a transkriptech se varianta nachází, v jaké oblasti (kódující, nekódující, regulační atd.) a jaký je její funkční dopad na vznikající protein (změna smyslu, ztráta smyslu, posun čtecího rámce, ztráta/nový stop kodon atd.). Mezi nejznámější nástroje pro funkční anotaci patří ANNOVAR (Wang et al., 2010), SnpEff (Cingolani et al., 2012b), SnpSift (Cingolani et al., 2012a) a Variant

Effect Predictor (McLaren et al., 2016). Tyto nástroje využívají informací o genech a transkripčních variantách z databází jako je Ensembl (Cunningham et al., 2019), UCSC (Haeussler et al., 2019) a RefSeq (O'Leary et al., 2016). Získané výsledky se mohou zásadně lišit v závislosti na použité databázi, anotačním programu i nastavení konkrétních parametrů programu (McCarthy et al., 2014).

U variant s neznámým funkčním dopadem na protein využíváme predikčních algoritmů jako je SIFT (Ng, 2003), GERP (Cooper, 2005), GERP++ (Davydov et al., 2010), PolyPhen-2 (Adzhubei et al., 2010), MutationTaster (Schwarz et al., 2014), LRT (Chun and Fay, 2009), CADD (Kircher et al., 2014) a PhyloP (Pollard et al., 2010). Každý z těchto algoritmů pracuje na jiném principu. Využívají například informací o mezidruhové evoluční konzervovanosti, regulačních elementech, epigenetických markerech, sekvenční homologii, struktuře proteinu, počtu známých mutací v daném genu a jiných charakteristikách. Výsledky predikčních algoritmů je třeba interpretovat obezřetně, protože mají velkou chybovost (Li et al., 2018).

Populační frekvence variant je určena z veřejně dostupných databází velkých sekvenačních projektů, jako například databáze projektu 1000 genomů (1000 Genomes Project Consortium et al., 2015), Exome Variant Server (http://evs.gs.washington.edu/EVS), ExAC (Lek et al., 2016) a gnomAD (Karczewski et al., 2019). Existují i národní a regionální databáze, které jsou důležité při diagnostice onemocnění s častějším výskytem v konkrétní populaci. Pro Českou republiku jsme pod záštitou Národního centra lékařské genomiky vytvořili veřejně dostupnou databází, která obsahuje varianty běžné pro českou populaci (https://ncmg.cz).

Následně při terciální analýze anotujeme geny, ve kterých se nalezené varianty nacházejí. Využíváme informací z rozličných biologických databází. Uvádím příklady pouze některých z nich, protože tato fáze anotace je značně variabilní a je možné přidávat i anotace vlastní. Zajímá nás, jaká je funkce genu a jeho proteinu a v jakém buněčném kompartmentu se nachází (Gene Ontology; Ashburner et al., 2000), jakých metabolických drah se účastní (Reactome; Wu and Haw, 2017 a KEGG; Kanehisa, 2000), v jakých tkáních je exprimován (GTEx; Lonsdale et al., 2013), jestli již byl spojen s nějakým onemocněním (ClinVar; Landrum et al., 2018 a HGMD; Stenson et al., 2003 a OMIM; Amberger et al., 2019), s jakými dalšími proteiny interaguje (STRING; Szklarczyk et al., 2019), jaká je intolerance genu k mutacím (RVIS; Petrovski et al., 2013) a další.

Výsledkem exomového sekvenování jsou desítky tisíc variant a výsledkem genomového sekvenování dokonce až stovky tisíc variant. Většinou ale pouze jedna či několik z nich jsou kauzální. Abychom nalezli kauzální variantu, **filtrujeme** varianty na základě informací

přiřazených v předchozích krocích. Pro snížení falešně pozitivních variant filtrujeme varianty podle kvality genotypu, počtu čtení pokrývající danou variantu (pokrytí, coverage) a procenta čtení obsahující variantu. Při filtrování se také zohledňují předpokládané modely dědičnosti, které je u každého studovaného případu nutné dobře stanovit na základě rodinné anamnézy, pokud je dostupná. Dále nás zajímají varianty, které se nacházejí ve funkčně významných oblastech genomu, jsou predikovány jako měnící funkci proteinu a jejich frekvence v populačních databázích je nízká (filtr MAF, Minor Allele Frequency). Vhodným nástrojem je například GEMINI (Paila et al., 2013)

Následně **prioritizujeme** ty varianty, které svou biologickou funkcí odpovídají klinickým projevům pacienta, laboratorním vyšetřením, údajům z literatury a databází a mohly by být tedy kauzální. Pro usnadnění této práce existují nástroje, které často využívají pro popis symptomů onemocnění ontologii lidských fenotypů (HPO, Human Phenotype Ontology). HPO umožňuje propojení a klasifikaci symptomů onemocnění a onemocnění způsobujících genů například pomocí porovnávání s výsledky experimentů na modelových organismech (PHIVE; Robinson et al., 2014), nebo s využitím dalších ontologií popisujících funkce genů (Phevor; Singleton et al., 2014). Vzhledem ke složitosti interpretace výsledků NGS je vhodné postupovat podle doporučení odborných organizací, jako jsou například doporučení Americké společnosti lékařské genetiky a genomiky (ACMG; Richards et al., 2015).

U každé kandidátní varianty je vhodné ověřit její kvalitu a sekvenční kontext **vizualizací** namapovaných čtení (BAM soubor) v prohlížeči genomických variant, jako je například IGV (Robinson et al., 2017).

Vybrané kandidátní varianty jsou následně **validovány** pomocí funkčních studií (kapitola 2.3).

2.2.2 Speciální analýzy z NGS dat druhé generace

Předchozí kapitola popisovala základní kroky bioinformatické analýzy SGS dat. Nicméně získaná sekvenační data se dají využít i k analýzám, které jsou/byly primárně prováděny pomocí jiných metod. Jedná se například o vazebnou analýzu a homozygotní mapování, které byly dříve prováděny analýzou polymorfních genetických markerů získaných například pomocí štěpení restrikčními endonukleázami, PCR či technologií DNA čipů. Dále se jedná o detekci změn genové dávky, která se provádí pomocí technologií DNA čipů a o detekci expanzí tandemových repetic, která se provádí metodou short-repeat PCR (Morling, 1998), nebo

Southern blot. Všechny zmíněné analýzy je možné provádět jak z genomových dat, tak s omezenou úspěšností i z dat exomových.

2.2.2.1 Vazebná analýza

Vazebná analýza je metoda používaná pro mapování dědičných znaků, jako jsou právě například onemocnění, na chromozomální oblasti. U jedinců sdílejících zkoumaný znak je potřeba stanovit genotypy stovek či tisíců genetických markerů po celém genomu. Využívají se vysoce polymorfní genetické markery jako například: krátké tandemové repetice (STRs, Short Tandem Repeats), variabilní počet tandemových repetic (VNTR, Variable Number of Tandem Repeats), polymorfismy v délce restrikčních fragmentů (RFLP, Restriction Fragment Length Polymorphism) nebo SNPs (právě ty jsou využívány v případě analýzy z SGS dat).

V případě, že známe model dědičnosti zkoumaného onemocnění, využíváme parametrickou vazebnou analýzu. Pokud model dědičnosti neznáme, využíváme analýzou neparametrickou. Model dědičnosti je možné zjistit analýzou rodokmenu, nebo pomocí segregační analýzy, kdy sledujeme segregaci určitého genetického markeru s ohledem na přítomnost či nepřítomnost zkoumaného onemocnění. Další faktory, které ovlivňují analýzu, jsou frekvence alely v populaci, předpokládaný vliv vnějšího prostředí a penetrance (podíl jedinců s alelou, které vykazují klinické příznaky zkoumané patologie).

Pro každý genetický marker u každého jedince vypočítáme pravděpodobnost vazby. Tato pravděpodobnost se vypočítá jako podíl pravděpodobnosti, že dva lokusy jsou ve vazbě (rekombinační frakce = θ) a pravděpodobnosti, že ve vazbě nejsou (rekombinační frakce = 0,5). Logaritmus pravděpodobnosti vazby se označuje jako LOD skóre (Log Of the Odds; Morton, 1955). Čím vyšší je hodnota LOD skóre, tím vyšší je pravděpodobnost vazby znaku a genetického markeru. Podle konvencí je za důkaz vazby považováno LOD skóre větší než 3 a za vyloučení vazby LOD skóre menší než -2.

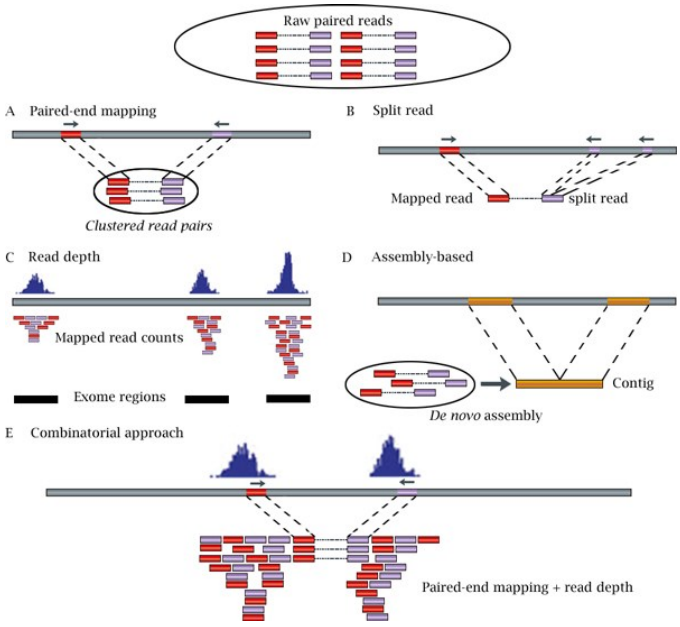
Příkladem nástrojů pro vazebnou analýzu je Merlin (Abecasis et al., 2002) a ALOHOMORA (Ruschendorf and Nurnberg, 2005). V případě exomového sekvenování, kdy využíváme jako genetické markery SNPs, nám umožní vazebná analýza snížit počet kandidátích variant až na polovinu. (Gazal et al., 2016).

2.2.2.2 Homozygotní mapování

Homozygotní mapování, neboli také autozygotní mapování, je metodou vhodnou při hledání kauzálních genů u vzácných recesivních onemocnění, především u dětí z příbuzenských svazků (Lander and Botstein, 1987). Tato metoda předpokládá, že kauzální varianta bude preferenčně ležet v některé z homozygotních oblastí, které pocházejí od společného předka (IBD, identity by desecent). Pro identifikaci těchto oblastí využíváme stejně jako u vazebné analýzy polymorfních genetických markerů získaných například s využitím genotypovacích čipů, nebo i SNPs získaných z genomového či exomového sekvenování (Pippucci et al., 2014). Příklady nástrojů pro homozygotní mapování jsou Homozygosity mapper (Seelow et al., 2009) a ROH nástroj obsažený v GEMINI (Paila et al., 2013).

2.2.2.3 Strukturní varianty

Strukturní varianty jsou obecně definovány jako oblasti DNA o velikosti větší než 1 kb (Freeman, 2006). Mohou zahrnovat inserce, duplikace a delece - nazývané souhrnně jako změny genové dávky (CNVs, Copy Number Variations), a dále translokace a inverze. I tyto varianty mohou být příčinou rozvoje onemocnění. Pro detekci CNVs jsou vhodnější TGS, které sekvenují dlouhá čtení. Nicméně je možné (byť s omezenou úspěšností) detekovat CNVs i s využitím SGS, která sekvenují pouze krátká čtení. Existuje pět různých přístupů, jak se dají detekovat CNVs z SGS dat: mapováním pair-end čtení fragmentové knihovny (pair-end mapping), detekcí rozdělených čtení (split-read), metodami založenými na hloubce čtení, metodami založenými na sestavování *de novo* a metodami kombinujícími předchozí přístupy (Obrázek 9). Příklady nástrojů pro detekci CNVs ze SGS dat jsou CONTRA (Li et al., 2012) a CNVkit (Talevich et al., 2016).



Obrázek 9: Algoritmické přístupy pro detekci CNVs z SGS dat. A) mapování pair-end čtení B) detekce rozdělených čtení (split-read) C) detekce rozdílů v hloubce čtení D)detekce pomocí de novo *zarovnávání E) Kombinace předchozích algoritmů. Převzato ze Zhao et al., 2013*

2.2.2.4 Expanze repetic

Krátké tandemové repetice (mikrosatelity), jsou definovány jako repetetivní DNA složené z opakujících se jednotek o délce 2-6 bází. Počet opakování se mezi jednotlivci liší a právě proto jsou oblasti STRs s výhodou využívány jako genetické markery. Pokud ale dojde k expanzi repetic, může docházet k rozvoji onemocnění, jako je například Huntingtonova choroba, spinocerebelární ataxie a syndrom fragilního X (Mirkin, 2007). U onemocnění způsobených expanzí tandemových repetic můžeme často pozorovat, že nástup onemocnění je v dalších generacích časnější, v důsledku narůstání počtu tandemových repetic (anticipace). STRs nebylo možné donedávna ze SGS dat detekovat. V posledních letech bylo vyvinuto několik algoritmů, které detekci STRs umožňují: ExpansionHunter (Dolzhenko et al., 2017), exSTRa (Tankard et al., 2018), STRetch (Dashnow et al., 2018) a TREDPARSE (Tang et al., 2017). Všechny metody vycházejí z analýzy souboru BAM, který musí obsahovat pair-end čtení. Jsou vyhledávána čtení, která leží částečně či plně v STR oblasti. Vyhledávání probíhá buď heuristicky (ExpansionHunter, exSTRa, STRetch), nebo pomocí pravděpodobnostních modelů (TREDPARSE). Alely s expanzí repetic přispívají větší měrou a mají vyšší pokrytí než alely normální. Většina těchto nástrojů je primárně určena pro analýzu z genomových dat – s výjimkou nástroje exSTRa, který je vhodný i pro analýzu z exomových dat.

2.2.3 Specifika bioinformatické analýzy NGS dat třetí generace

Mnoho programů, algoritmů a datových formátů vyvinutých pro SGS není adekvátních pro data TGS. Hlavní problém představují dlouhá čtení, vysoká chybovost a chyby specifické pro jednotlivé sekvenační platformy. Dochází k úpravám stávajících programů, algoritmů, datových formátů a k vývoji nových bioinformatických přístupů, které umožňují z dlouhých čtení získat i informace, které bylo velmi obtížné či nemožné získat z krátkých čtení SGS platforem. Dlouhá čtení umožňují detekci nových strukturních variant, izoforem, genových fúzí, segementálních duplikací, repetitivních oblastí, GC bohatých oblastí, epigenetických změn a fázování variant po celé délce chromosomů. Výpočetní nároky pro analýzu TGS dat jsou velmi vysoké. Analýza lidského genomu trvá desítky tisíc hodin centrální procesorové jednotky (Koren et al., 2017). Dobrý přehled problematiky bioinformatické analýzy TGS dat podává Sedlazeck et al. (2018).

2.3 Validace kandidátních variant

Nedílnou součástí studia onemocnění podmiňujících genů je i validace nalezených variant. Přítomnost dané varianty může a nemusí být ověřována v závislosti na technologii použité pro její určení a kvalitě čtení. Pokud například kandidátní varianta nemá ideální pokrytí, či se jedná o *de novo* variantu, je vhodné ji ověřit Sangerovým sekvenováním či specifickým štěpením restrikčními endonukleázami. Klíčové pro průkaz kauzality je ověření segregace varianty v rodině se studovaným fenotypem, průkaz rekurence u další rodiny a prokázání funkčního efektu varianty na rozvoj fenotypu. Výběr metod pro ověření funkčního efektu varianty závisí na mnoha faktorech, jako je dostupnost biologických materiálů od pacienta a jeho rodinných příslušníků, biologická funkce daného genu a jeho produktu a jeho lokalizace v buněčných kompartmentech. Metody můžeme rozdělit na *in silico* (predikce, simulace a analýza biologických dějů na počítačích), *in vitro* (studium buněčných modelů jako jsou tkáňové kultury zahrnující embryonální kmenové buňky, imortalizované buněčné linie a indukované pluripotentní kmenové buňky) a *in vivo* (zvířecí modely).

3 Cíle disertační práce

Cílem disertační práce bylo využít nových metod analýzy genomu k určení a charakterizaci kauzálních genů, genových mutací a genomových změn v případech vzácných geneticky podmíněných onemocnění neznámé etiologie a definovat základní biologické a patologické procesy podmiňující studované fenotypy.

Dílčí cíle disertační práce:

1. Vývoj a validace metody pro cílené sekvenování genů podmiňujících dědičné metabolické poruchy (METABO panel).

2. Identifikace kauzálních genů a mutací u vybraných vzácných geneticky podmíněných onemocnění pomocí NGS metod druhé generace (SOLiD, Illumina).

3. Identifikace variant v obtížně analyzovatelných oblastech genomu pomocí NGS metod druhé a třetí generace (SOLiD, Illumina, Oxford Nanopore, PacBio).

4 Seznam publikací, které jsou podkladem disertace

Prvoautorské (sdílené prvoautorství)

1. Živná M, Kidd K, **Přistoupilová A**, Barešová V, DeFelice M, Blumenstiel B, Harden M, Conlon P, Lavin P, Connaughton DM, Hartmannová H, Hodaňová K, Stránecký V, Vrbacká A, Vyleťal P, Živný J, Votruba M, Sovová J, Hůlková H, Robins V, Perry R, Wenzel A, Beck BB, Seeman T, Viklický O, Rajnochová-Bloudíčková S, Papagregoriou G, Deltas CC, Alper SL, Greka A, Bleyer AJ, Kmoch S. **Noninvasive Immunohistochemical Diagnosis and Novel *MUC1* Mutations Causing Autosomal Dominant Tubulointerstitial Kidney Disease**. *J Am Soc Nephrol*. 2018;29(9):2418-2431. doi:10.1681/ASN.2018020180. **IF= 8,547**

2. Jedlickova I, **Přistoupilová A**, Nosková L, Majer F, Stránecký V, Hartmannová H, Hodaňová K, Trešlová H, Hýblová M, Solár P, Minárik G, Giertlová M, Kmoch S. **Spinal muscular atrophy caused by a novel Alu-mediated deletion of exons 2a-5 in *SMN1* undetectable with routine genetic testing**. Journal: *Mol Genet Genomic Med 1*. 2020;na(na):na. (Accepted, in production). **IF= 2,448**

3. Jedlickova I, **Pristoupilova A**, Hulkova H, Vrbacka A, Stranecky V, Hruba E, Jesina P, Honzík T, Hrdlicka I, Fremuth J, Pivovarcikova K, Bitar I, Matej R, Kmoch S, Sikora J. ***NOTCH2NLC* CGG repeats are not expanded and skin biopsy was negative in an infantile patient with neuronal intranuclear inclusion disease.** (In review)

Spoluautorské

4. Bolar NA, Golzio C, Živná M, Hayot G, Van Hemelrijk C, Schepers D, Vandeweyer G, Hoischen A, Huyghe JR, Raes A, Matthys E, Sys E, Azou M, Gubler M-C, Praet M, Van Camp G, McFadden K, Pediaditakis I, **Přistoupilová A**, Hodaňová K, Vyleťal P, Hartmannová H, Stránecký V, Hůlková H, Barešová V, Jedličková I, Sovová J, Hnízda A, Kidd K, Bleyer AJ, Spong RS, Vande Walle J, Mortier G, Brunner H, Van Laer L, Kmoch S, Katsanis N, Loeys BL. **Heterozygous Loss-of-Function *SEC61A1* Mutations Cause Autosomal-Dominant Tubulo-Interstitial and Glomerulocystic Kidney Disease with Anemia.** *Am J Hum Genet*. 2016;99(1):174-187. doi:10.1016/j.ajhg.2016.05.028. **IF= 9,025**

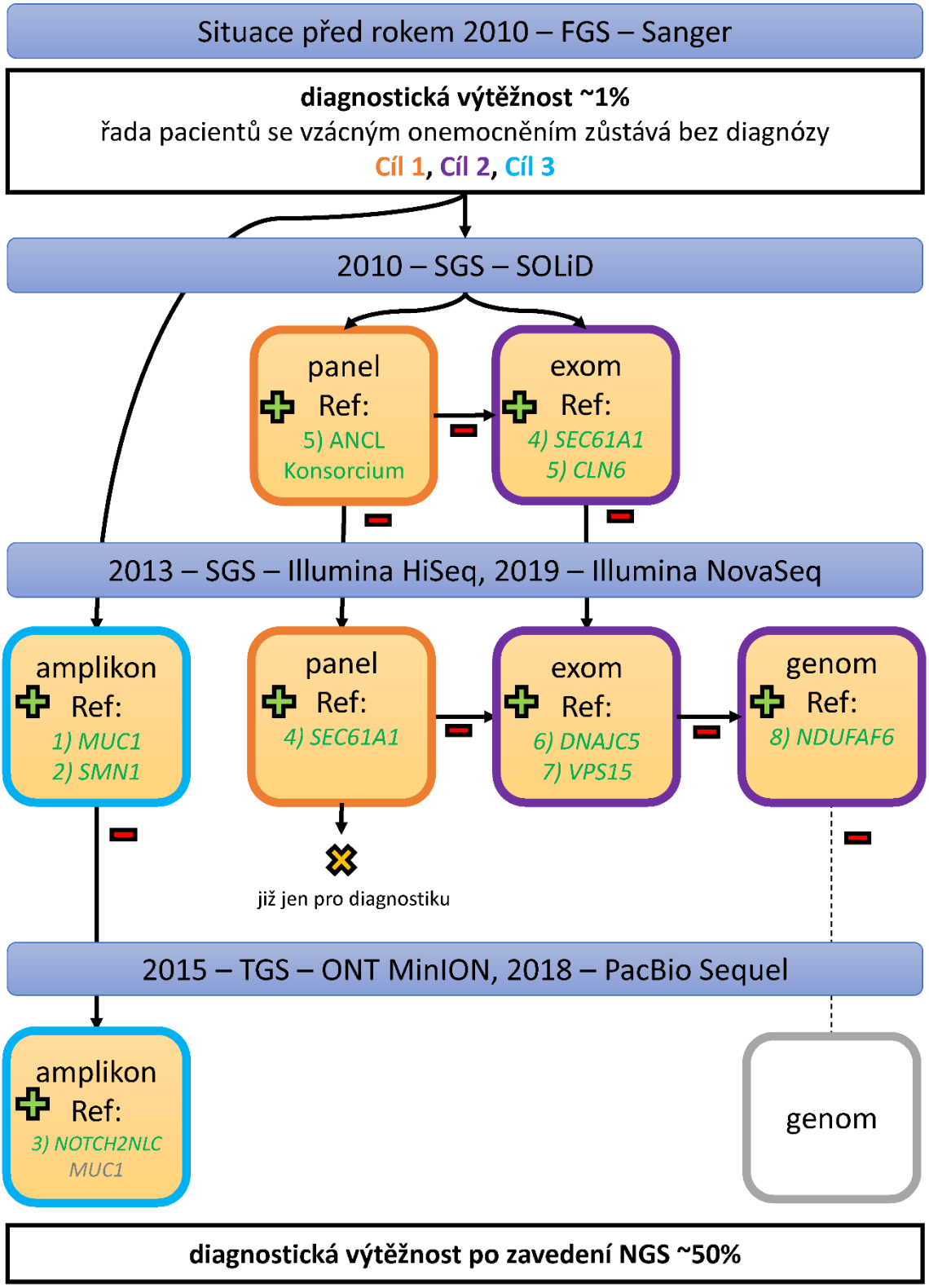
5. Berkovic SF, Staropoli JF, Carpenter S, Oliver KL, Kmoch S, Anderson GW, Damiano JA, Hildebrand MS, Sims KB, Cotman SL, Bahlo M, Smith KR, Cadieux-

Dion M, Cossette P, Jedličková I, **Přistoupilová A**, Mole SE, ANCL **Gene Discovery Consortium. Diagnosis and misdiagnosis of adult neuronal ceroid lipofuscinosis (Kufs disease)**. *Neurology*. 2016;87(6):579-584. doi:10.1212/WNL.0000000000002943. **IF= 7,592**

6. Jedličková I, Cadieux-Dion M, **Přistoupilová A**, Stránecký V, Hartmannová H, Hodaňová K, Barešová V, Hůlková H, Sikora J, Nosková L, Mušálková D, Vyleťal P, Sovová J, Cossette P, Andermann E, Andermann F, Kmoch S. **Autosomal-dominant adult neuronal ceroid lipofuscinosis caused by duplication in *DNAJC5* initially missed by Sanger and whole-exome sequencing**. *Eur J Hum Genet*. 2020;87(6):579-584. doi:10.1038/s41431-019-0567-2. **IF= 3,650**

7. Gstrein T, Edwards A, **Přistoupilová A**, Leca I, Breuss M, Pilat-Carotta S, Hansen AH, Tripathy R, Traunbauer AK, Hochstoeger T, Rosoklija G, Repic M, Landler L, Stránecký V, Dürnberger G, Keane TM, Zuber J, Adams DJ, Flint J, Honzik T, Gut M, Beltran S, Mechtler K, Sherr E, Kmoch S, Gut I, Keays DA. **Mutations in *Vps15* perturb neuronal migration in mice and are associated with neurodevelopmental disease in humans**. *Nat Neurosci*. 2018;21(2):207-217. doi:10.1038/s41593-017-0053-5. **IF= 21,126**

8. Hartmannová H, Piherová L, Tauchmannová K, Kidd K, Acott PD, Crocker JFS, Oussedik Y, Mallet M, Hodaňová K, Stránecký V, **Přistoupilová A**, Barešová V, Jedličková I, Živná M, Sovová J, Hůlková H, Robins V, Vrbacký M, Pecina P, Kaplanová V, Houštěk J, Mráček T, Thibeault Y, Bleyer AJ, Kmoch S. **Acadian variant of Fanconi syndrome is caused by mitochondrial respiratory chain complex I deficiency due to a non-coding mutation in complex I assembly factor *NDUFAF6***. *Hum Mol Genet*. 2016;25(18):4062-4079. doi:10.1093/hmg/ddw245. **IF= 5,340**



Obrázek 10: Přehled sekvenačních technologií využívaných v laboratoři pro studium vzácných nemocí KDDL, 1.LF UK a VFN. Rok označuje, kdy jsme danou technologii začali využívat. FGS – první generace sekvenování, SGS – druhá generace sekvenování, TGS – třetí generace sekvenování. Čtverce označují typ sekvenování - sekvenování PCR amplikonů, panelové sekvenování, exomové sekvenování a genomové sekvenování. „+" značí úspěšně vyřešené případy, „-" značí případy nevyřešené, „x" značí, že panelové sekvenování již ve výzkumu nevyužíváme – hodí se pro diagnostiku. Ref: uvádí odkaz na publikace v kapitole 4 - Seznam publikací, které jsou podkladem disertace (zelené písmo), které jsou v přípravě (šedé písmo). Barva ohraničení čtverce odpovídá jednotlivým cílům práce – cíl1 (oranžová), cíl2 (fialová), cíl 3 (modrá). Šedé ohraničení značí výhled do budoucnosti.

5 Výsledky a komentář k vybraným publikovaným pracím

5.1 Cíl 1) Vývoj a validace metody pro cílené sekvenování genů podmiňujících dědičné metabolické poruchy (METABO panel)

Dědičné metabolické poruchy (DMP) jsou dle současného stavu poznání specifickou heterogenní skupinou přibližně 850 geneticky podmíněných onemocnění (Kožich and Zeman, 2010), kde je z důvodu genetického defektu (mutace) postižena určitá klíčová součást metabolické dráhy. Její narušení vyvolává různé patologické projevy daného onemocnění, které jsou způsobeny enzymovým deficitem, dysfunkcí transportního proteinu či poruchou jiného proteinu souvisejícího s touto metabolickou dráhou (Fernandes, 2008). Diagnostika DMP je založena na využití tandemové hmotnostní spektrometrie, biochemických, enzymatických, histologických a genetických vyšetření jednotlivých genů.

V rámci grantového projektu Grantové agentury Univerzity Karlovy (1402213), jehož jsem byla řešitelem, jsme měli za cíl vyvinout a validovat metody cíleného sekvenování všech doposud známých genů podmiňujících DMP a dále všech jaderně kódovaných mitochondriálních genů. Tyto metody měly umožnit efektivnější a levnější diagnostiku DMP a vést k odhalení nových kauzálních genů u případů pacientů s postižením mitochondriálního energetického metabolismu.

Geny byly vybrány z databáze lidských genů a genetických poruch OMIM a databáze genetických asociací HuGE Navigator (Yu et al., 2008) pomocí thesauru Medical Subject Headings (MeSH, https://www.nlm.nih.gov/mesh) a klíčového slova „Metabolic Diseases“ [C18.452]) a dále pomocí klasifikace DMP podle Society for the Study of Inborn Errors of Metabolism (SSIEM, http://www.ssiem.org). Seznam známých jaderně kódovaných mitochondriálních genů byl získán ze specializovaných mitochondriálních databází MitoCarta (Pagliarini et al., 2008), Mitomap (Lott et al., 2013) a Mitoproteom (Cotter et al., 2004). Celkový počet genů byl 3616.

Sekvence regulačních a kódujících oblastí jednotlivých genů byly získány z databáze UCSC Genome Browser (Kent et al., 2002). Koordináty cílených sekvencí byly zaslány k pilotnímu návrhu oligonukleotidové knihovny pro cílené obohacení DNA pomocí obohacovacího kitu Roche Nimblegen SeqCap EZ Choice Library. Námi navržená knihovna cílí oblast lidského genomu o rozsahu 6,88 Mb.

Úspěšnost pilotního návrhu knihovny byla hodnocena na základě míry pokrytí cílených sekvencí a úspěšnosti detekce známých variant v testovacích vzorcích se známou genovou sekvencí. Použili jsme vzorky sekvenované v rámci projektu HapMap (Belmont et al., 2003) a 1000 Genomes (1000 Genomes Project Consortium et al., 2015). Průměrné pokrytí cílené oblasti pro jednotlivé testovací vzorky bylo více než 77x. Cílená oblast u jednotlivých testovacích vzorků byla pokryta více než 10x z 96 % a více než 2x z 98 %. Shoda genotypů v testovacích vzorcích s již známými variantami byla 98,4 %. Vyhodnocením testovacích vzorků jsme zjistili, že knihovna byla navržena optimálně, a proto již nebylo třeba navrhovat nové oligonukleotidy ani zvyšovat počet sekvenčně identických oligonukleotidů.

Pro dosažení co nejvyšší míry úspěšnosti detekce známých variant a redukci falešně pozitivních variant v testovaných vzorcích byly optimalizovány postupy analýzy sekvenačních dat. Porovnány byly aktuálně dostupné mapovací algoritmy pro rekonstrukci původní sekvence DNA a algoritmy sloužící k identifikaci sekvenčních variant. Pro analýzu sekvenačních dat byly zvoleny následující algoritmy: Novoalign v. 1.99 pro mapování čtení na referenční sekvenci, GATK v. 3.3-0-g37228af a Samtools v. 0.1.18 pro identifikaci sekvenčních variant, SnpSift v. 3.6 pro anotaci sekvenčních variant a SnpEff v. 3.6 pro predikci vlivu varianty na funkci proteinu.

V souladu s cíli projektu byla vyvinuta a validována metoda pro cílené sekvenování všech doposud známých genů podmiňujících DMP a jaderně kódovaných mitochondriálních genů. Námi vyvinutá metoda umožňuje současně vyšetřit vybraných 3616 genů u mnoha pacientů najednou, za cenu ~5000 Kč na jeden vzorek. Jen pro porovnání, za stejnou cenu bychom Sangerovou metodou osekvenovali přibližně pět genů.

Pomocí této metody jsme provedli sekvenaci u 185 vzorků od pacientů s podezřením na některou z DMP či na poruchu mitochondriálního energetického metabolismu. U 15 % vzorků byly nalezeny kauzální mutace a byla určena přesná diagnóza, což je nezbytným krokem ke kvalifikovanému poradenství, prevenci a případné léčbě. U 40 % vzorků jsme nalezli nové kandidátní mutace, jejichž kauzalitu bylo nutné funkčně ověřit. U 45 % vzorků nebyla nalezená žádná kandidátní mutace.

Tento projekt přispěl k rozšíření aplikací metod SGS a cíleného sekvenování jak ve výzkumu, tak diagnostice a vedl k vytvoření bioinformatických postupů nezbytných pro klinickou diagnostiku.

V následujících kapitolách uvádím dva konkrétní příklady, ve kterých jsme využili kombinace cíleného sekvenování genů METABO panelem a exomového sekvenování při identifikaci kauzálních variant u autosomálně dominantního tubulointersticiálního onemocnění ledvin a adultní neuronální ceroidní lipofuscinózy.

5.1.1 Autosomálně dominantní tubulointersticiální onemocnění ledvin - *SEC61A1*

Bolar NA, Golzio C, Živná M, et al (2016) **Heterozygous Loss-of-Function SEC61A1 Mutations Cause Autosomal-Dominant Tubulo-Interstitial and Glomerulocystic Kidney Disease with Anemia**. Am J Hum Genet 99:174–187. doi: 10.1016/j.ajhg.2016.05.028

***

Autosomálně dominantní tubulointersticiální onemocnění ledvin (ADTKD) sestává ze skupiny genetických onemocnění charakterizovaných renálně tubulárními a intersticiálními abnormalitami, vedoucími k progresivnímu renálnímu selhání vyžadujícími dialýzu a transplantaci ledvin (Bleyer et al., 2017; Devuyst et al., 2019; Eckardt et al., 2015). ADTKD jsou způsobeny mutacemi v různých jaderných genech zahrnujících *MUC1* (Kirby et al., 2013), *UMOD* (Hart et al., 2002), *HNF1B* (Lindner et al., 1999), *REN* (Živná et al., 2009), *DNAJB1* (Cornec-Le Gall et al., 2018), *NPHP1* (Snoek et al., 2018) a mitochondriální DNA (Connor et al., 2017). Nově zavedená nomenklatura onemocnění sestává z názvu ADTKD a přívlastku příslušné genetické příčiny, například ADTKD-*UMOD* atd. (Eckardt et al., 2015). U mnoha případů ADTKD zůstává genetická příčina zatím neznámá.

V tomto článku prezentujeme případ dvou rodin s ADTKD a vrozenou anémií doprovázenou buď intrauterinním zpomalením růstu, nebo neutropenií. Ultrazvukové vyšetření a ledvinná biopsie odhalili malé dysplastické ledviny s cystami, tubulární atrofii a sekundární glomerulární sklerózou.

Symptomy onemocnění byly u první rodiny podobné jako u jedinců s mutacemi v reninu (anémie s časným nástupem, hyperurikémie, progresivní renální selhání). Z toho důvodu jsme nejprve pomocí Sangerova sekvenování u probanda vyloučili mutace v reninu. Získali jsme DNA od 12 členů rodiny (7 postižených) a provedli jsme parametrickou vazebnou analýzu s využitím čipu Illumina Human CytoSNP12. Pod dominantním modelem dědičnosti jsme identifikovali kandidátní oblast s maximálním LOD skóre 2,7 na chromosomu 3q14-25.1. Kandidátní oblast o velikosti 32 Mb, chr3:120000–152000 (hg19 – referenční sekvence lidského genomu verze 19) zahrnovala 357 genů.

Provedli jsme exomové sekvenování probanda pomocí obohacovacího kitu Agilent Technologies 50 Mb SureSelect Enrichment Kit na sekvenátoru SOLiD 4 a zaměřili se na vzácné (MAF < 1 %) kódující a nesynonymní varianty. V oblasti vymezené vazebnou analýzou jsme nalezli dvě kandidátní varianty. Variantu c.1189C>T (p.Arg397Cys) v exonu 7 genu *NPHP3* a variantu c.553A>G (p.Thr185Ala) v exon 7 genu *SEC61A1*. Varianta v genu *NPHP3* v rodině s fenotypem nesegregovala, kdežto varianta v genu *SEC61A1* ano. Záměna konzervované aminokyseliny Thr185 byla predikovaná predikčními nástroji SIFT a MutationTaster jako patogenní a nástrojem PolyPhen jako benigní. Varianta nebyla přítomna v žádné z kontrolních databází – 1000 genomů, Exome Variant Server, ExAC, naší interní databázi (~600 exomů a ~140 panelů), ani interní databázi Nijmegen (~500 exomů, Radboud University Medical Center).

Abychom prokázali kauzalitu mutace, provedli jsme screening dalších jedinců s obdobným fenotypem, a to pomocí METABO panelu vytvořeném na našem ústavu (viz. úvod kapitoly 5.1). Celkem jsme vyšetřili 46 nepříbuzných jedinců s fenotypem odpovídajícím ADTKD. U jednoho z nich jsme nalezli heterozygotní variantu měnící smysl kodónu c.200T>G (p.Val67Gly) v *SEC61A1*. Varianta byla konzervovaná jak na úrovni nukleotidové, tak aminokyselinové. Varianta byla predikovaná jako škodlivá pomocí predikčních program SIFT a PolyPhen a jako onemocnění způsobující pomocí nástroje Mutation Taster. Varianta byla reportovaná v databázi ExAC (1/112410 alel). V ostatních databázích reportovaná nebyla. Segregace varianty s onemocněním v rodině probanda byla prokázána pomocí Sangerova sekvenování.

SEC61A1 kóduje alfa podjednotku heterotrimetrického kanálu SEC61. Výsledný kanál, složený z alfa, beta a gama podjednotek je součástí savčího translokonu. Ten je lokalizován v membráně drsného endoplasmatického retikula (ER) a umožňuje transport translatovaných proteinů mezi cytoplazmou a lumen ER. *In silico* predikce strukturálního a funkčního dopadu variant v SEC61 prokázala, že obě dvě varianty se nacházejí v konzervovaných oblastech translokonu. Varianta Thr185 se nachází v transmebránovém helixu 5, který je lokalizován v blízkosti translokovaných peptidů (Cannon et al., 2005) a pravděpodobně narušuje strukturní integritu kanálu (Gray and Matthews, 1984). Varianta Val67 je součástí domény, která stabilizuje translokon v uzavřeném stavu pomocí tzv. zátky (plug).

Vzhledem k tomu, že žádné mutace v *SEC61A1* nebyly u ADTKD dosud popsány, provedli jsme transientní expresi obou mutovaných proteinů v buněčných liniích. Zjistili jsme, že wild-type protein byl lokalizován výlučně v ER, kdežto oba dva mutované proteiny byly

lokalizovány nejen v ER, ale i v Golgi. Tento nález byl pomocí imunohistochemické analýzy potvrzen v ledvinných biopsiích pacientů.

Abychom zjistili, jaká je role *SEC61A1* ve vývoji ledviny, připravili jsme *in vivo* model embryí dánia pruhovaného (*Danio rerio*, zebrafish). Knockdown *sec61al2*, ortologu lidského *SEC61A1,* způsobil defekty při vývoji nefronů, kopírující tubulární fenotyp pozorovaný v ledvinných biopsiích pacientů s ADTKD. *In vivo* komplementace mutovaných buněk pomocí wild-type mRNA zvrátila nefrotický fenotyp u embryí, zatímco komplemetace jednou nebo druhou mutovanou variantou mRNA k reverzi fenotypu nevedla.

Tyto výsledky naznačují, že SEC61A1 je nezbytný pro správný vývoj nefronů, a že u dvou studovaných rodin je ADTKD s kongenitální anémií způsobená mutacemi v *SEC61A1*. Obě dvě mutace funkčně ovlivňují důležitou konzervovanou část SEC61A1 a vedou k defektu translokace proteinů přes membránu endoplazmatického retikula, což má za následek rozvoj fenotypu ADTKD.

5.1.2 Adultní neuronální ceroidní lipofuscinózy – ANCL Konsorcium

Berkovic SF, Staropoli JF, Carpenter S, et al (2016) **Diagnosis and misdiagnosis of adult neuronal ceroid lipofuscinosis (Kufs disease).** Neurology 87:579–84. doi: 10.1212/WNL.0000000000002943

***

Neuronální ceroidní lipofuscinózy (NCL) jsou heterogenní skupinou vzácných, geneticky podmíněných neurodegenerativních onemocnění. U NCL dochází k intralysosomálnímu střádání materiálu (ceroid, lipofuscin, lipopigment) jak v buňkách nervového systému, tak i v periferních tkáních. Střádaný materiál obsahuje převážně podjednotku c mitochondriální ATP syntázy, saposiny A a D. Je vysoce hydrofobní a vyžaduje speciální mechanismy pro jeho rozklad a eliminaci (Palmer et al., 2013). V důsledku střádání materiálu v buňkách dochází k neurodegeneraci vedoucí k rozvoji neurologických obtíží jako je postupné zhoršování kognitivních i pohybových schopností, poruchy vidění a epileptické záchvaty.

Podle věku nástupu onemocnění dělíme NCL na infantilní, pozdně infantilní, časně juvenilní, juvenilní a adultní formy. Adultní formy NCL (ANCL, Kufsova choroba) onemocnění jsou na rozdíl od dětských forem NCL velmi obtížně diagnostikovatelné, diagnóza

je povětšinou stanovena až *post mortem*. Většinou totiž nedochází ke střádání materiálu v periferních tkáních, ale pouze v některých neuronech. Navíc je obtížné odlišit patologické střádání od fyziologického střádání lipofuscinu, ke kterému dochází v pozdějším věku.

ANCL můžeme dále rozdělit na typ A projevující se progresivní myoklonickou epilepsií a typ B projevující se demencí a zhoršením motorických funkcí. Dědičnost ANCL může být jak autosomálně dominantní, tak recesivní. Bylo identifikováno několik kauzálních genů *CLN6* (Arsov et al., 2011), *CTSF* (Smith et al., 2013), *PPT1* (Van Diggelen et al., 2001), *CLN5* (Xin et al., 2010), *GRN* (Smith et al., 2012), *ATP13A2* (Bras et al., 2012) a *DNAJC5* (Nosková et al., 2011). Mutace v genu *DNAJC5* byly popsány na našem pracovišti pomocí kombinace vazebné analýzy, exomového sekvenování a analýzy genové exprese. Výsledky byly součástí mé diplomové práce (Přistoupilová, 2011).

I přesto, že byla identifikována řada kauzálních genů, zůstává mnoho pacientů s ANCL bez správné diagnózy. Z tohoto důvodu bylo založeno ANCL Gene Discovery Consortium (Konsorcium), které shromažďuje vzorky pacientů s podezřením na ANCL z celého světa s cílem identifikovat genetickou příčinu onemocnění. Naše laboratoř je součástí tohoto Konsorcia. Konsorcium vypracovalo diagnostická kritéria, podle kterých dělí pacienty na základě klinických a patologických nálezů do několika skupin: jednoznačná, pravděpodobná, možná a vyloučená diagnóza ANCL. Ze 47 shromážděných vzorků od pacientů bylo určeno 5 jako jednoznačných, 2 pravděpodobné a 9 možných. U 31 pacientů byla diagnóza ANCL vyloučena a následně byla u 10 z nich diagnostikováno onemocnění jiné. Byly nalezeny mutace v genech *HTT, NPC1, PLA2G2, c19orf12, PRNP, SACS* (Muona et al., 2015), *PSEN1* (Ehling et al., 2013), *SERPINI1* (Muona et al., 2015) a *MTND3*.

U 14 pacientů s jednoznačnou, pravděpodobnou a možnou ANCL jsem provedli vyšetření známých ANCL genů (*CLN6, CTSF, DNAJC5, GRN* a *PPT1*) pomocí Sangerova sekvenování a/nebo cíleného sekvenování za použití METABO panelu. U žádného z pacientů jsme v těchto genech nenašli patogenní ani pravděpodobně patogenní varianty. Následně jsme u 12 pacientů provedli exomové sekvenování a u jednoho z nich jsme nalezli dvě varianty v genu *CLN6*, jež byly predikovány jako pravděpodobně patogenní. Pacient byl pro tyto dvě varianty složeným heterozygotem. U ostatních pacientů jsme se zaměřili na vzácné patogenní varianty, ale nenašli jsme žádné kandidátní varianty, jejichž patogenitu bychom mohli dále ověřovat pomocí funkčních studií.

Obtížnost diagnostiky ANCL je dána několika faktory: vzácností onemocnění, variabilitou klinických příznaků a obtížností odlišení abnormálního lipopigmentu charakteristického pro

ANCL od normálního lipofuscinu, k jehož akumulaci dochází se stoupajícím věkem. V této studii jsme zjistili, že >1/3 případů byla špatně diagnostikována jako ANCL. U případů určených jako jednoznačná, pravděpodobná a možná ANCL jsme genetickou příčinu odhalili pouze u jednoho pacienta.

Souhrnné výsledky jsou součástí publikace uvedené v záhlaví této kapitoly. Po vydání této publikace se nám podařilo identifikovat genetickou příčinu u dalších 3 pacientů. U prvního pacienta byla nalezena nová mutace, 30 bp inzerce v genu *DNAJC5* (viz. kapitola 5.2.1). U druhého pacienta byla nalezena expanze tandemových repetic v genu *ATN1* způsobující Dentato-rubro-palido-luysiánskou atrofii. Expanze byla nalezena pomocí nástrojů pro identifikaci tandemových repetic z SGS dat (viz. kapitola 2.2.2.4) a byla ověřena cíleným genotypováním (publikace v přípravě). U třetí pacientky s progresivní myoklonickou epilepsií byla jako kauzální mutace určena expanze tandemových repetic v genu *C9orf72* (van den Ameele et al., 2018) pomocí metody short-repeat PCR (Gijselinck et al., 2018).

5.2 Cíl 2) Identifikace kauzálních genů a mutací u vybraných vzácných geneticky podmíněných onemocnění pomocí NGS metod druhé generace (SOLiD, Illumina)

5.2.1 Adultní neuronální ceroidní lipofuscinóza – *DNAJC5*

Jedličková I, Cadieux-Dion M, Přistoupilová A, et al (2020) **Autosomal-dominant adult neuronal ceroid lipofuscinosis caused by duplication in *DNAJC5* initially missed by Sanger and whole-exome sequencing.** Eur J Hum Genet. doi: 10.1038/s41431-019-0567-2

***

V této publikaci prezentujeme případ jedné z rodin zmiňovaných v souhrnném ANCL článku (kapitola 5.1.2), která byla Konsorciem klasifikovaná jako pravděpodobná ANCL s dominantní dědičností. V rámci Konsorcia byly Sangerovým sekvenováním a METABO panelem vyloučeny mutace ve všech známých genech, včetně genu *DNAJC5*. U dvou postižených bratrů bylo dále provedeno exomové sekvenování obohacovacím kitem Agilent Technologies Sure Select Human All Exon v4 na platformě Illumina HiSeq 2000. Vzhledem k autosomálně dominantnímu modelu dědičnosti jsme hledali heterozygotní varianty u obou bratrů, které měly pokrytí ≥ 10 čtení a MAF < 0,5 % v databázi ExAC. Nenalezli jsme žádnou funkčně relevantní kandidátní variantu. Při opětovné analýze výsledků jsme snížili filtrovací kritéria a hledali jsme varianty s pokrytím ≥ 5. Odhalili jsme 30 bp duplikaci ve čtvrtém exonu genu *DNAJC5*, chr20:g.62562252_62562281dup (hg19); NM_025219.2:c.370_399dup (p.(Cys124_Cys133dup)). Tato varianta při inspekci BAM souboru v prohlížeči IGV nebyla zobrazena. Pro zobrazení varianty bylo klíčové mít v IGV prohlížeči zapnuté zobrazení soft-clipped bází (*View>Preferences>Filter and Shading Options> Show soft-clipped bases*), které vede k vizualizaci konců čtení, které nemapují na referenční sekvenci v plné délce a jsou tzv. zastřihnuty (clipping). Vzhledem k nízkému pokrytí bylo potřeba variantu ověřit jinou metodu. Tato duplikace nebyla primárně detekována pomocí standardního PCR protokolu a Sangerova sekvenování kvůli preferenční amplifikaci krátké referenční alely. Modifikovali jsme původní PCR protokol a variantu se nám podařilo potvrdit pomocí Sangerova sekvenování.

DNAJC5 kóduje cystein-string protein alpha (CSPα), který v komplexu s dalšími proteiny působí jako molekulární chaperon při formaci presynaptických SNARE komplexů. SNARE komplexy jsou zásadní pro kotvení, fúzi a recyklaci synaptických váčků. Existuje stále více důkazů, že poruchy SNARE aparátu vedou k neurodegeneraci (Gorenberg and Chandra, 2017).

Stejně tak jako předchozí nalezené mutace (Benitez et al., 2011; Nosková et al., 2011; Velinov et al., 2012), i tato duplikace je lokalizovaná v centrální konzervované doméně proteinu CSPα, která je na cystein bohatá (CSD, cystein-string domain). Palmitoylace cysteinů umožňuje zakotvení CSPα do synaptické membrány (Chamberlain and Burgoyne, 1998; Greaves et al., 2008; Greaves and Chamberlain, 2006). Pomocí *in silico* analýzy jsme zjistili, že nalezená duplikace zvyšuje hydrofobicitu CSD a že přítomnost nadbytečných 7 cysteinů mění její palmitoylaci. Tyto změny mohou způsobit vyšší agregaci mutovaného proteinu (Ramadan et al., 2007). Změna buněčné lokalizace byla potvrzena v *in vitro* studii pomocí transientní exprese wild-type a mutovaných forem konstruktů v neuronálních buňkách. Mutované konstrukty zahrnovaly i námi dříve nalezené mutace Leu115Arg, Leu116del (Nosková et al., 2011). Metodou Western blot byla prokázána nepřítomnost palmitoylované formy Cys124_Cys133dup, narozdíl od wild-type CSPα, Leu115Arg a Leu116del, které byly přítomny jak v palmitoylované, tak nepalmitoylované formě. U všech tří mutovaných proteinů byla pozorována tvorba vysokomolekulárních nerozpustných agregátů.

Potvrdili jsme, že nalezená varianta Cys124_Cys133dup v genu *DNAJC5* vede k duplikaci centrálního motivu CSD domény, ovlivňuje na palmitoylaci závislé třídění CSPα v kulturách nervových buněk a je příčinou ANCL u této rodiny.

Nezávisle se objevující varianty v CSD doméně naznačují, že tato oblast může být více náchylná k chybám vznikajících při DNA replikaci a že by inzerce a duplikace v této oblasti měly být zvažovány u nevyřešených případů s ANCL. Zároveň může při diagnostice docházet k falešně negativním nálezům.

5.2.2 Neurodegenerativní onemocnění neznámé etiologie - *VPS15*

Gstrein T, Edwards A, Přistoupilová A, et al (2018) **Mutations in *Vps15* perturb neuronal migration in mice and are associated with neurodevelopmental disease in humans**. Nat Neurosci 21:207–217. doi: 10.1038/s41593-017-0053-5

***

V tomto článku prezentujeme případ chlapce s neurodegenerativním onemocněním neznámé etiologie. Onemocnění se manifestovalo již po narození hypotonií a opožděným psychomotorickým vývojem. Postupně docházelo k progresi stavu a projevil se těžký mozečkový syndrom, spastická kvadruparéza, pseudobulární syndrom, hypotrofie končetin, atrofie mozku a zrakových nervů. Byla provedena vyšetření na metabolická a neuromuskulární

onemocnění, nicméně všechny nálezy byly nespecifické. V 17 letech chlapce se objevil první epileptický záchvat a začaly se objevovat menší, ale časté generalizované myoklonie. Chlapec zemřel v 19 letech v důsledku dýchacího selhání.

V roce 2012 (v osmnácti letech chlapce) bylo během mé stáže na pracovišti Centro Nacional de Análisis Genómico (CNAG) provedeno exomové sekvenování všech členů rodiny (proband, matka, otec a tři zdraví sourozenci) obohacovacím kitem Roche Nimblegen SeqCap EZ v3 na sekvenátoru Illumina HiSeq 2000. Při filtrování jsme se zaměřili na všechny vzácné, protein kódující varianty odpovídající možným modelům dědičnosti (autosomálně recesivní MAF < 1 %, *de novo* MAF < 0,1 %, složený heterozygot MAF < 2 %, X-vázaný MAF < 1 %). Celkem jsme nalezli mutace ve 13 kandidátních genech. Bohužel ani jednu z nich jsme nedokázali při současném stavu poznání určit jako kauzální. V roce 2013 jsme provedli reanalýzu a reinterpretaci dat a zjistili jsme, že byl publikován článek, který spojuje mutace v jednom z kandidátních genů, *PIK3R4 (VPS15)*, s autofagickou vakuolární myopatií a s lysosomálním onemocnění v myším modelu (Nemazanyy et al., 2013).

Gen *VPS15* kóduje regulační podjednotku fosfatidylinositol-3-kinázového komplexu třetí třídy a je nezbytný pro stabilitu a aktivaci Vps34, který slouží jako katalytická podjednotka. Tento komplex přeměňuje fosfatidylinositol na fosfatidylinositol-3-fosfát a je nezbytný pro třídění proteinů u kvasinek (Schu et al., 1993).

Abychom zjistili, jestli neexistuje další pacient s obdobným fenotypem a mutací ve stejném genu, vložili jsme gen *VPS15* do databáze GeneMatcher (Sobreira et al., 2015), bohužel v té době bez výsledku. V roce 2015 se nám přes tuto databázi podařilo propojit s dr. Keaysem, který vytvořil myší model s mutací ve *Vps15*, která vedla v důsledku narušené neuronální migrace k poruchám mozkových struktur. Kompletní ablace *Vps15* vedla k akumulaci autofagických substrátů, indukci apoptózy a závažné kortikální atrofii – tedy ke stejnému fenotypu, jaký jsme pozorovali u našeho pacienta.

U našeho pacienta jsme nalezli v genu *VPS15* homozygotní mutaci (hg38; chr3:130681528A>C), která vede ke změně vysoce konzervovaného leucinu na arginin p.Leu1224Arg. Mapováním této mutace na strukturu kvasinkového komplexu Vps15–Vps34–Beclin1 jsme zjistili, že mutace leží uvnitř WD40 domény, části důležité pro folding a stabilitu (Rostislavleva et al., 2015).

Pro prokázání funkčního efektu varianty na rozvoj fenotypu byla provedena studie na fibroblastech pacienta a jeho rodičů. Ve fibroblastech pacienta bylo prokázáno snížené

množství proteinů VPS15, VPS34 a BECLIN1. Množství mRNA nebylo sníženo, což naznačuje, že snížené množství proteinů je dáno posttranskripční dysregulací. Dále bylo pozorováno zvýšené množství receptoru p62, což ukazuje na poruchu autofagie (Liu et al., 2016). Vliv mutace p.Leu1224Arg byl potvrzen komplementací pomocí wild-type VPS15, po které došlo ke stabilizaci množství proteinů BECLIN1 a VPS34.

Prokázali jsme, že mutace p.Leu1224Arg vede k poruše komplexu VPS15-BECLIN1-VPS34, akumulaci substrátů autofagie a je příčinou neurodegenerativního onemocnění u tohoto pacienta.

5.2.3 Akadská varianta Fanconiho syndromu – *NDUFAF6*

Hartmannová H, Piherová L, Tauchmannová K, et al (2016) **Acadian variant of Fanconi syndrome is caused by mitochondrial respiratory chain complex I deficiency due to a non-coding mutation in complex I assembly factor *NDUFAF6***. Hum Mol Genet 25:4062–4079. doi: 10.1093/hmg/ddw245

***

Renální Fanconiho syndrom je vzácné onemocnění projevující se sníženým vstřebáváním elektrolytů a organických sloučenin v proximálních tubulech, což vede k metabolické acidóze, glykosurii, fosfatourii, aminoacidurii a urikosurii (Bökenkamp and Ludwig, 2010; Klootwijk et al., 2015). Může být způsoben léky (Kitterer et al., 2015), toxiny (Hall et al., 2014) či genetickými mutacemi ovlivňujícími funkci proximálních tubulů (Klootwijk et al., 2015; Solano et al., 2014).

Akadská varianta Fanconiho syndromu (AVFS) se projevuje pouze u Akaďanů, populace s efektem zakladatele v kanadském Novém Skotsku. AVFS je charakterizován generalizovanou dysfunkcí proximálních tubulů vedoucí k pomalu progredujícímu chronickému selhání ledvin, spojenému s intersticiální fibrózou plic (Crocker et al., 1997; Wornell et al., 2007). Genetická a molekulární podstata tohoto onemocnění byla neznámá.

Provedli jsme retrospektivní studii na 12 pacientech z 8 rodin. Struktura rodokmenů a předpokládaný efekt zakladatele naznačovaly, že se jedná o autosomálně recesivní model dědičnosti. Nejprve jsme provedli exomové sekvenování pomocí obohacovacího kitu Roche NimbleGen SeqCap EZ Exome v2 na sekvenátoru SOLiD 4 u čtyř jedinců z první rodiny (tři postižení a jeden zdravý). Pomocí homozygotního mapování jsme určili homozygotní oblasti

genomu > 2 Mb. Nalezli jsme pouze jednu takovou oblast, chr8:90958422-96960058 (hg19). Vzhledem k předpokládanému autosomálně recesivnímu modelu dědičnosti jsme se v této oblasti zaměřili na homozygotní varianty (CNVs, SNPs, inserce a delece) sdílené všemi čtyřmi postiženými jedinci, ale nepřítomné u zdravého jedince. Nalezli jsme dvě varianty s nízkou frekvencí výskytu a to v genech *OTUD6B* (chr8:92097062G>A; rs3210518, hg19) a *RBM12B* (chr8: 94746049T>G; rs16916188, hg19). Pomocí Sangerova sekvenování jsme provedli genotypování varianty rs3210518 u všech členů této rodiny. Všichni postižení nesli tuto variantu v homozygotním stavu, všichni zdraví nesli tuto variantu heterozygotně nebo u nich nebyla přítomna vůbec. Segregační analýzou jsme tak potvrdili relevanci homozygotní oblasti v této rodině. Pro identifikaci kauzálních mutaci jsme dále provedli exomové sekvenování rozšířené o 5´ a 3´ nepřekládané oblasti obohacovacím kitem Roche NimbleGen SeqCap EZ Human Exome + UTR na sekvenátoru Illumina HiSeq 1500. Díky tomu jsme odhalili rekombinace u postiženého jedince z další rodiny a tím jsme upřesnili homozygotní oblast na chr8:94242350-97172487 (hg19). Pro určení nekódujících variant v kandidátní oblasti jsme provedli genomové sekvenování u dvou postižených na sekvenátoru Illumina HiSeq X Ten s pokrytím 30x. Homozygotní mapování dále zúžilo kandidátní oblast na chr8:94423201-96206283 (hg19), která obsahovala 322 variant. Při filtrování jsme upřednostnili varianty vzácné (MAF < 0,5 %), konzervované (GERP skóre) a predikované jako škodlivé (CADD skóre). Zbyla nám pouze jedna kandidátní varianta chr8:96046914T>C, rs575462405, v genu *NDUFAF6*.

Varianta se nachází v druhém intronu genu *NDUFAF6* (NM_152416.3; c.298-768T>C), 37 bází upstream od varianty chr8:96046951A>G (c.298-731A>G), rs74395342), která vytváří nové donorové místo sestřihu. *In silico* analýza provedená několika nástroji pro predikci efektu varianty na sestřih (Alamut® Visual, Interactive Biosoftware) předpověděla, že varianta c.298-768T>C vytváří nové akceptorové místo sestřihu. Pomocí Sangerova sekvenování jsme prokázali homozygotní stav varianty chr8:96046914T>C u postižených jedinců ze všech 8 rodin a absenci této varianty v homozygotním stavu u všech zdravých.

NDUFAF6 kóduje asemblační faktor 6 komplexu I respiračního řetězce NADH dehydrogenázy. *NDUFAF6* existuje v několika různých izoformách, včetně jedné, která je predikovaná jako mitochondriální (McKenzie et al., 2011). Izolovali jsme celkovou RNA z fibroblastů a z plicní biopsie pacientů. Produkty reverzní transkripce jsme osekvenovali paralelně na sekvenátoru Illumina HiSeq 1500 a pomocí Sangerova sekvenování. Nalezli jsme celkem 10 různých izoforem. Prokázali jsme, že rs575462405 - buď samostatně nebo

v kombinaci s rs74395342 ovlivňuje sestřih a syntézu *NDUFAF6* izoforem, vede ke ztrátě mitochondriální izoformy a následně k poruše tvorby a funkce respiračního komplexu I. Význam nalezené mutace pro rozvoj onemocnění byl potvrzen komplementací – po transfekci wild-type *NDUFAF6* do fibroblastů pacienta došlo k odstranění předchozího enzymového defektu.

Naše výsledky prokazují, že AVFS je způsobená deficitem mitochondriálního komplexu I respiračního řetězce v důsledku nekódující mutace v asemblačním faktoru *NDUFAF6*. Tato informace může být využita pro diagnostiku a prevenci onemocnění u jedinců a rodin akadského původu a významně rozšiřuje spektrum klinické prezentace mitochondriálních onemocnění, defektů respiračního řetězce a poruch respiračního komplexu I.

5.3 Cíl 3) Identifikace variant v obtížně analyzovatelných oblastech genomu pomocí NGS metod druhé a třetí generace (SOLiD, Illumina, Oxford Nanopore, PacBio)

5.3.1 Autosomálně dominantní tubulointersticiální onemocnění ledvin – *MUC1*

ADTKD-*MUC1* je dědičné onemocnění způsobené posunovými mutacemi v repetitivním genu mucin 1 (*MUC1)*, které v důsledku změny čtecího rámce vedou k syntéze abnormálního, cystein-bohatého a vysoce bazického proteinu MUC1-fs. Hromadění MUC1-fs v tubulárních buňkách ledvin vede k postupné ztrátě tubulárních funkcí a renálnímu selhání, vyžadujícím dialýzu a transplantaci ledvin. Věk, ve kterém dochází k renálnímu selhání, je u pacientů různý (17–75 let). Předpokládáme, že doba nástupu může souviset s pozicí mutace v rámci genu – tedy čím dříve dojde ke změně čtecího rámce, tím více patogenního MUC1-fs vzniká a tím dříve dochází k renálnímu selhání.

Identifikace posunových mutací a určení jejich přesné pozice v rámci genu *MUC1* je komplikováno několika faktory. Kódující sekvence genu *MUC1* obsahuje repetitivní oblast sestávající z 25-122 degenerovaných tandemových repetic (VNTR), které jsou 60 bází dlouhé, GC bohaté (82 %) a některé z nich obsahují homopolymerní oblast 7 cytosinů. Každá alela je navíc jinak dlouhá, protože obsahuje jiný počet tandemových repetic. Kombinace těchto faktorů znesnadňuje využití standardních molekulárně-diagnostických a bioinformatických postupů.

5.3.1.1 Identifikace mutací

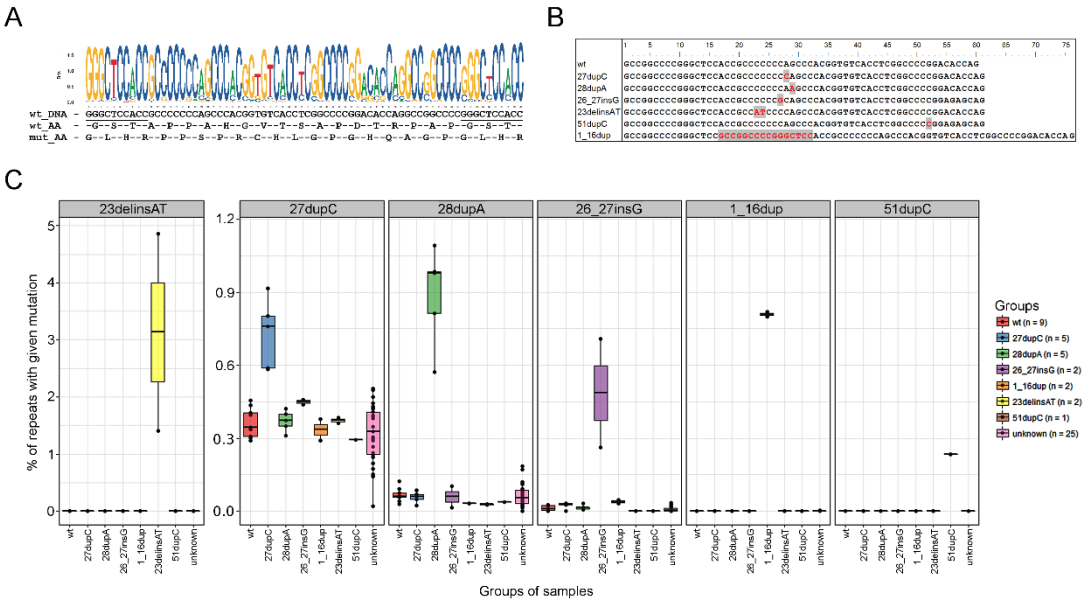
Živná M, Kidd K, Přistoupilová A, et al (2018) **Noninvasive Immunohistochemical Diagnosis and Novel MUC1 Mutations Causing Autosomal Dominant Tubulointerstitial Kidney Disease**. J Am Soc Nephrol 29:2418–2431. doi: 10.1681/ASN.2018020180

***

Náš postup pro identifikaci mutací v *MUC1* byl založen na obohacení VNTR oblasti *MUC1* z genomové DNA pomocí LR-PCR. Pro získání dostatečného množství materiálu a minimalizaci chyb vzniklých během PCR byl vzorek připraven kombinací několika individuálních reakcí. Délku jednotlivých alel jsme určili pomocí přístroje Fragment Analyzer a podle segregace v rámci rodokmenu jsme zjistili, která alela je nositelkou mutace.

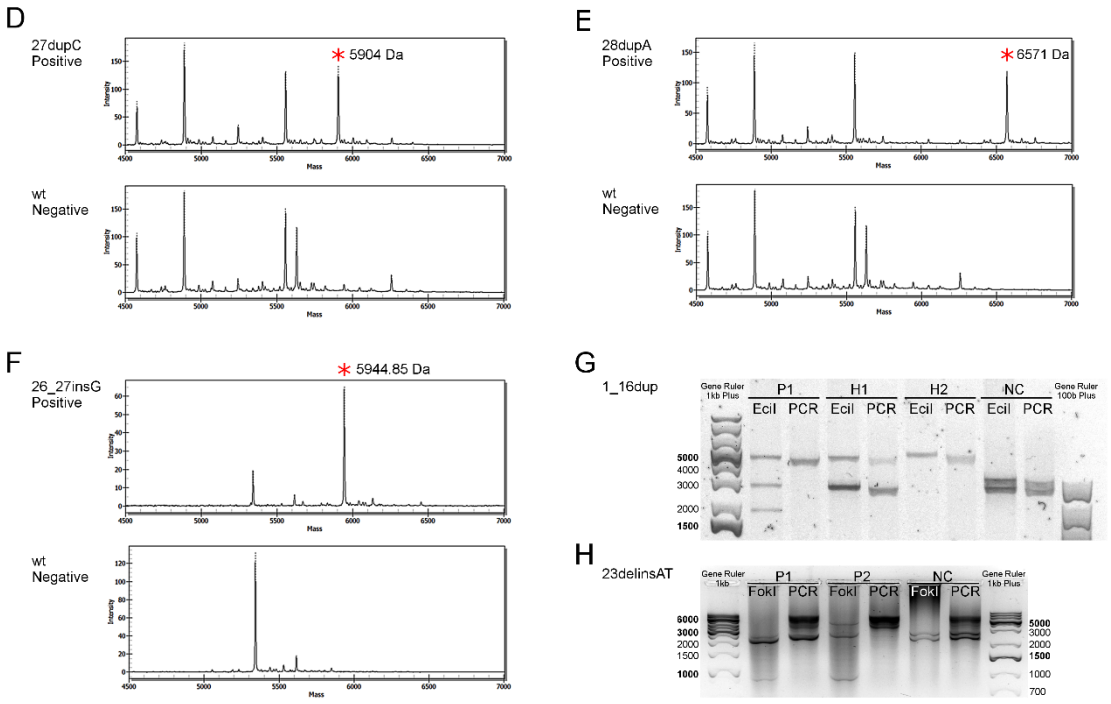
Pro identifikaci nových mutací jsme získanou VNTR oblast sekvenovali na přístroji Illumina HiSeq/NovaSeq. Vzhledem k repetitivní povaze VNTR oblasti a krátké délce získaných čtení (101 bp) nebylo možné použít standardní postup pro analýzu SGS dat založený na mapování čtení na referenční sekvenci. Vyvinuli jsme proto vlastní bioinformatický postup založený na analýze hrubých sekvenačních dat, ve kterých cíleně vyhledáváme posunové mutace.

Abychom našli nejvíce konzervovanou oblast degenerovaných repetic, provedli jsme mnohonásobné zarovnávání sekvencí (MSA, Multiple Sequence Alignment) získaných z FASTQ souborů. Sekvenci nejvíce konzervovaných oblastí jsme použili jako takzvané kotvy (Obrázek 11a). Pokud je vzdálenost mezi dvěma kotvami 3n nukleotidů (tři nukleotidy pro každou jednu aminokyselinu), nedochází k posunu čtecího rámce. Pro nalezení kandidátních repetic nesoucích posunové mutace vyhledáváme čtení, ve kterých je vzdálenost mezi dvěma kotvami 3n+1 nebo 3n-2 bází. Pro všechny vzorky poté spočítáme procentuální zastoupení těchto kandidátních repetic. Vzhledem k tomu, že mutace mohou náhodně vznikat během PCR, považujeme za reálné mutace pouze ty, které mají procentuální zastoupení větší, než je průměr + 2 SD u devíti zdravých kontrol. Například duplikace 27dupC je přítomna u zdravých kontrol v 0,36 % ± 0,06 % čtení. Proto pro určení mutace 27dupC jako reálné tedy požadujeme, aby byla zastoupena alespoň v 0,48 % čtení. (Obrázek 11c). Pomocí tohoto přístupu se nám podařilo identifikovat pět nových posunových mutací u celkem 6 rodin (Obrázek 11b).



Obrázek 11: A) analýza konzervovaných oblastí degenerovaných VNTR repetic, wt_DNA odpovídá nejčastěji zastoupené repetici, wt_AA uvádí aminokyselinovou sekvenci wild-type proteinu, mut_AA uvádí aminokyselinovou sekvenci mutovaného proteinu vzniklého v důsledku posunu čtecího rámce o 1 bázi, podtržením jsou vyznačené kotvy použité pro identifikaci nových posunových mutací B) Sekvence VNTR repetic, wild-type (wt) a indentifikovaných posunových mutací C) osa x uvádí skupiny vzorků, ve kterých bylo určováno procentuální zastoupení čtení (osa y) nesoucích jednotlivé mutace (uvedené v hlavičce každého z boxů).

Všechny tyto mutace byly ověřeny nezávislou metodou. Mutace 28dupA a 26_27insG byly ověřeny na Broad Institutu (MIT and Harvard) pomocí eseje, která se využívá pro identifikaci mutací 27dupC (Blumenstiel et al., 2016). Tato esej je založena na tom, že v důsledku mutace 27dupC dochází ke ztrátě rozpoznávacího místa pro restrikční endonukleázu *MwoI*. Mutovaná repetice tedy není rozštěpena a je namnožena pomocí PCR a metody prodlužování primeru (probe extension). Získané fragmenty jsou detekovány pomocí hmotnostní spektrometrie. Zjistili jsme, že tuto esej je možné s modifikovanými próbami použít i pro nové mutace 28dupA a 26_27insG (Obrázek 12E a Obrázek 12F), protože i tyto mutace narušují *MwoI* rozpoznávací místo. Mutace 1_16dup a 23delinsAT byly ověřeny štěpením PCR amplikonu pomocí restrikčních enzymů. *EciI* štěpí alelu specificky v místě mutace 1_16 dup a *FokI* v místě mutace 23delinsAT. Mutovaná alela je tedy rozštěpena na dvě části, a vizualizována pomocí agarózové elektroforézy (Obrázek 12G a Obrázek 12H).



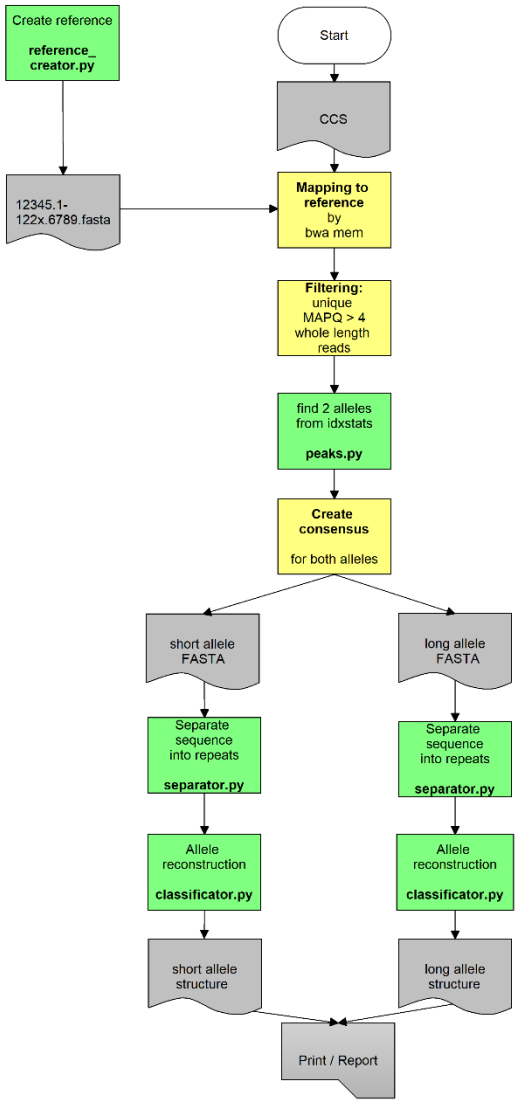
Obrázek 12: Potvrzení mutací nezávislou metodou (D, E, F) potvrzení mutací 27dupC, 28dupA a 26_27insG esejí založenou na hmotnostní spektrometrii, hvězdička znázorňuje mutovaný produkt (G, H) potvrzení mutací 1_16dup a 23delinsAT pomocí štěpění restrikčními endonukleázami EciI a FokI u pacientů P1 a P2, příbuzných kontrol H1 a H2 a nepříbuzné kontroly NC. U každého jedince je neprve zobrazen štěpený amplifikovaný produkt - označený názvem štěpícího enzymu a poté produkt neštěpený – označený jako PCR. U pacientů dochází k naštěpení mutované alely na dvě části.

5.3.1.2 Určení pozice mutací v rámci VNTR

Characterization of *MUC1* structure and its effect on progression of ADTKD (in preparation)

***

Pro určení přesné struktury obou alel a určení pozice mutace v rámci genu využíváme technologii SMRT. Tato unikátní technologie nám dovoluje pročtení celých molekul *MUC1* VNTR v jednom čtení, včetně homopolymerní oblasti 7 cytosinů a následné určení pozice mutace v rámci genu pomocí našich vlastních bioinformatických postupů, které znázorňuje Obrázek 13.



Obrázek 13: Bioinformatický postup pro analýzu VNTR oblasti MUC1, založený na kolekci skriptů naprogramovaných v Bashi (žlutá) a Pythonu (zelená).

Pro potřeby analýzy jsme nejprve vytvořili referenční sekvenci (**reference_creator.py**). Referenční sekvence byla vytvořena tak, aby odrážela strukturu VNTR oblasti, která sestává z 25-122 degenerovaných repetic, které jsou 60 bází dlouhé. Nejvíce konzervovanou repetici značíme písmenem X. Před VNTR oblastí se vždy nachází celkem pět konzervovaných repetic (označených čísly 1, 2, 3, 4, 5) a za VNTR oblastí se nacházejí čtyři konzervované repetice (6, 7, 8, 9). Referenční sekvence tedy obsahuje 122 různě dlouhých kontigů, lišících se počtem degenerovaných repetic X uprostřed. Sekvence zmíněných repetic ukazuje Obrázek 14.

Stejně jako v případě identifikace mutací na přístroji Illumina, i zde vycházíme z LR-PCR amplikonů. Ty sekvenujeme na přístroji Sequel (PacBio) v režimu CCS. Získané FASTQ soubory nejprve **mapujeme** na vlastní referenční sekvenci.

Namapovaná čtení **filtrujeme**, abychom se zbavili nekvalitních, chimerických a nekompletních čtení. Podle počtu čtení namapovaných na jednotlivé kontigy určíme délku obou dvou alel (**peaks.py**). Poté určíme konkrétní sekvence obou dvou alel (**Create consensus**). Následně rozdělíme sekvenci každé alely na jednotlivé repetice (**separator.py**), a ty klasifikujeme podle jejich sekvence (**classificator.py**). Každé repetici je přiděleno písmeno či číslo odpovídající konkrétní sekvenci repetic, které byly popsány dříve (Kirby et al., 2013; Wenzel et al., 2018) a nebo písmeno nové (Obrázek 14). Tímto způsobem je možné určit i nové mutace a určit jejich pozici v rámci VNTR.

```
AAGGAGACTTCCGCTACCCAGAGAAGTTCAGTGCCCAGCTCTACTGAGAAGAATGCTGTG – 1
AGTATGACCAGCAGCGTACTCTCCAGCCACAGCCCCGGTTCAGGCTCCTCCACCACTCAG – 2
GGACAGGATGTCACTCTGGCCCCGGCCACGGAACCAGCTTCAGGTTCAGCTGCCACCTGG – 3
GGACAGGATGTCACCTCGGTCCCAGTCACCAGGCCAGCCCTGGGCTCCACCACCCCGCCA – 4
GCCCACGATGTCACCTCAGCCCCGGACAACAAGCCAGCCCCGGGCTCCACCGCCCCCCCA – 5
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCGGGCCCCGGGCTCCACCCCGGCCCCG – 6
GGCTCCACCGCCCCCCCAGCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCG – 7
GGCTCCACCGCCCCCCCAGCCCATGGTGTCACCTCGGCCCCGGACAACAGGCCCGCCTTG – 8
GGCTCCACCGCCCCTCCAGTCCACAATGTCACCTCGGCCTCAGGCTCTGCATCAGGCTCA – 9
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCCCCCCA – X
GCCCACGGTGTCACCTCGGCCCCGGAGAGCAGGCCGGCCCCGGGCTCCACCGCGCCCGCA – A
GCCCACGGTGTCACCTCGGCCCCGGAGAGCAGGCCGGCCCCGGGCTCCACCGCCCCCCCA – B
GCCCACGGTGTCACCTCGGCCCCGGACACCAGGCCGGCCCCGGGCTCCACCGCCCCCCAA – C
```

Obrázek 14: Sekvence repetic před VNTR oblastí (1, 2, 3, 4, 5), za VNTR oblastí (6, 7, 8, 9), nejvíce konzervovaná repetice uvnitř VNTR oblasti (X) a příklady dalších repetic (A, B, C)

Následuje příklad zápisu sekvence obou alel u jednoho pacienta, nesoucích 35 a 42 degenerovaných repetic. Pozice posunové mutace je vyznačená červeně.

P1-35x : 1 2 3 4 5 C X A X C B X X A B A X C X X X A A B X X X X X X X X A B X X X X X X **6 7 8 9**

P1-42x : 1 2 3 4 5 C X A X C B X X A B A X C X X X A A B X X X X X X X A B **X** X X X X X A B B X X X X X **6 7 8 9**

Na závěr je pro každý analyzovaný vzorek vygenerován **report** s přesnou strukturou obou alel, nalezenými mutacemi včetně jejich pozice a statistikami umožňujícími zkontrolovat kvalitu výsledků.

Pomocí tohoto postupu jsme osekvenovali celkem 45 vzorků (29 pacientů a 16 kontrol). Kompletní sekvenci obou alel *MUC1* se nám podařilo určit u 41 vzorků (25 pacientů, 16 kontrol). U čtyř zbývajících vzorků nebylo kvůli nízkému počtu čtení reprezentujících dlouhou alelu možné určit její sekvenci. Ve skupině vzorků od pacientů se známou posunovou mutací (mutace byly dříve identifikovány metodami zmíněnými v kapitole 5.3.1.1), jsme osekvenovali 16 vzorků a mutace byla identifikována u všech. Ve skupině vzorků od pacientů bez známé posunové mutace, ale s prokázanou akumulací MUC1-fs, jsme osekvenovali 9 vzorků.

Posunová mutace byla identifikována pouze ve 2 z nich, a to i přesto, že pokrytí obou alel bylo vysoké. Ve skupině 16 kontrolních vzorků jsme nenašli žádné posunové mutace.

SMRT sekvenování následované speciální bioinformatickou analýzou umožňuje plně zrekonstruovat sekvenci VNTR oblasti *MUC1*, nalézt známé i nové mutace a spolehlivě určit jejich pozici. Zavedené metody umožňují genetickou diagnostiku ADTKD-*MUC1* a mohou přispět k odhalení genetických faktorů podmiňujících průběh renálního selhání u jednotlivých pacientů.

V současné době sekvenujeme vzorky od dalších kontrol a pacientů s podezřením na ADTKD-*MUC1* a připravujeme výsledky pro publikaci.

5.3.2 Spinální svalová atrofie – *SMN1*

Ivana Jedličková, Anna Přistoupilová, Lenka Nosková, (XXXX) **"Spinal muscular atrophy caused by a novel *Alu*-mediated deletion of exons 2a-5 in *SMN1* undetectable with routine genetic testing**" Mol Genet Genomic Med n/a:n/a-n/a. (Accepted)

* * *

Spinální svalové/muskulární atrofie (SMA) jsou skupinou autosomálně recesivních dědičných onemocnění postihujících 1 z 8000 novorozenců. Jsou charakterizované progresivní degenerací alfa motorických neuronů předních rohů míšních, často v kombinaci s degenerací motorických jader hlavových nervů vedoucí k ochablosti svalstva a paralýze. Podle doby nástupu a tíže onemocnění se SMA dělí na čtyři skupiny (typ I, MIM #253300; typ II, MIM #253550, MIM #253400, typ IV, MIM #271150; https://omim.org/).

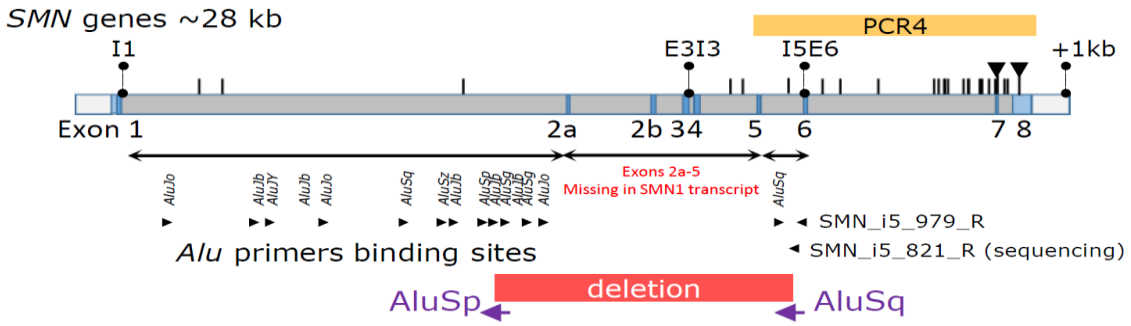
Příčinou SMA jsou bialelické mutace v genu *SMN1* (survival of motor neuron 1). V 95 % diagnostikovaných případů se jedná o delece exonů 7 a/nebo 8 (*SMN1Δ7, SMN1Δ(7-8)*). *SMN1* se nachází v komplexní oblasti 5q13, ve které také leží homologní pseudogen *SMN2* (survival of motor neuron 2), jež se liší od *SMN1* pouze v 5 nukleotidech a vyskytuje se v nula až šesti kopiích (Crawford et al., 2012). *SMN1* a *SMN2* (*SMN* geny) kódují stejný protein, nicméně *SMN2* tvoří jen 10-20 % účinného proteinu SMN ve srovnání s genem *SMN1*. *SMN2* nemůže tedy plně nahradit deficit *SMN1*, ale většinou platí, že čím více má pacient kopií genu *SMN2*, tím mírnější je projev jeho onemocnění (Butchbach, 2016).

V tomto článku prezentujeme pacientku s podezřením na SMA s nástupem prvních příznaků v prvním měsíci života, která byla zachycena ve Fakultní nemocnici v Košicích. Byla

provedena standardní diagnostika pomocí MLPA analýzy (Multiplex Ligation-dependent Probe Amplification; Mercuri et al., 2018), která ukázala, že pacientka je nositelkou maternálně zděděné heterozygotní delece *SMN1Δ(7-8)* a nese dvě kopie exonů 7 a 8 genu *SMN2*. Sangerovo sekvenování kódujících oblastí *SMN*, vyšetření karyotypu, komparativní genomová hybridizace, ani panelové SGS sekvenování genů asociovaných s neuromuskulárními onemocněními neodhalilo žádné patologické varianty ani chromosomální přestavby, které by vysvětlovaly SMA fenotyp. V této fázi byla pacientka zařazena do projektu zaměřeného na nevyřešené případy vzácných onemocnění (NV19-08-00137) řešeného v naší laboratoři.

Sangerovo sekvenování plnodélkových amplikonů SMN transkriptů odhalilo u matky a otce transkripty odpovídající *SMN1* a *SMN2*, kdežto u pacientky byl detekován pouze transkript odpovídající *SMN2*. U pacientky a jejího otce byl detekován abnormální zkrácený 600 bp produkt, odpovídající *SMN1* s deletovanými exony 2 až 5 (*SMN1Δ(2a-5)*). Pomocí imunodetekce metodou Western blot bylo oproti kontrole prokázáno snížené množství SMN proteinu u pacientky (na 10 %), matky (na 27 %) i otce (na 41 %).

Analýza genové dávky pomocí kvantitativní PCR potvrdila přítomnost intragenové delece genu *SMN1* u pacientky a jejího otce a naznačila celogenovou deleci *SMN1* u pacientky a její matky. Abychom identifikovali varianty způsobující deleci *SMN1Δ(2a-5)* u pacientky a jejího otce, provedli jsme SGS sekvenování 4 LR-PCR amplikonů pokrývajících celou délku genů *SMN*. Získaná čtení jsme namapovali na referenční sekvenci hg19 pomocí nástroje Novoalign. Abychom neztratili čtení v důsledku dvojího mapování na homologní pozice genů *SMN1* a *SMN2*, změnili jsme parametry mapovacího algoritmu tak, aby všechna čtení která mapují na více rozdílných pozic v genomu, byla namapována na všechny z nich. Pomocí SNPs odlišujících *SMN1* a *SMN2* jsme zjistili, že produkt odpovídající intronům 6 a 7 a exonům 7 a 8

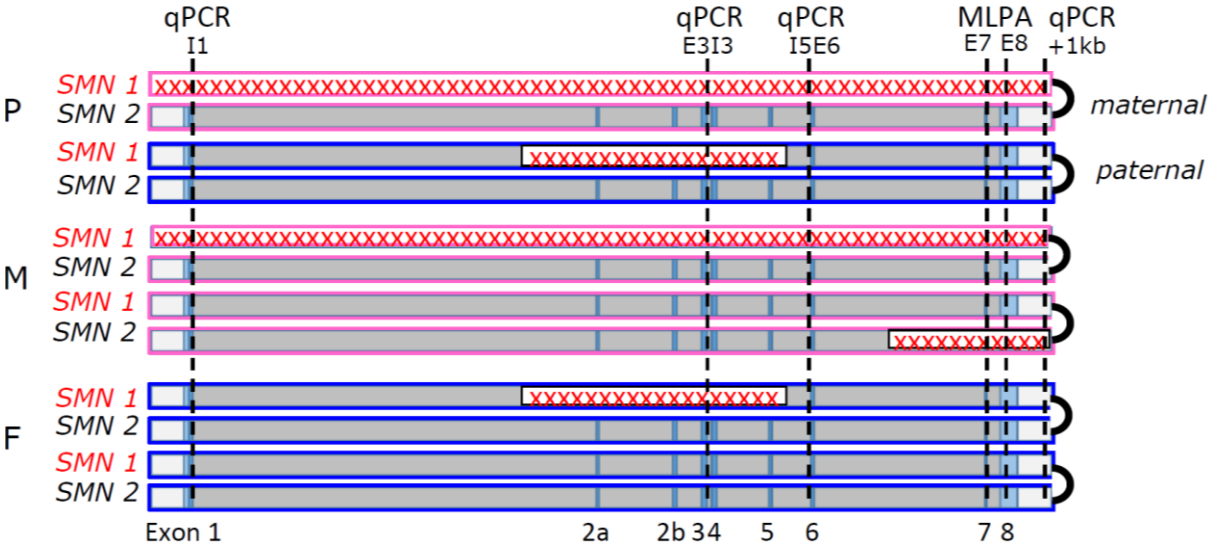


Obrázek 15: Schématická reprezentace genů SMN1 a SMN2 znázorňující exony (E) a introny (I). Pozice míst odlišujících SMN1 a SMN2 jsou označeny černými svislými čarami. Černé trojúhelníky označují specifické sekvenční varianty v exonech 7 a 8, které jsou cíleny MLPA próbami při rutinní diagnostice. Oranžově je označena pozice PCR4 produktu. Pozice odpovídající kandidátním Alu elementům jsou označeny vertikálním písmem a kauzální elementy jsou označeny fialovými šipkami a písmem (AluSp a AluSq). Mezi nimi je červeně vyznačená kauzální delece exonů 2a-5 chybějících v transkriptu 1.

(PCR4, Obrázek 15) byl amplifikován výhradně ze *SMN2*. Tento nález ukázal, že *SMN1* je narušen na obou alelách pacientky.

Zbývalo charakterizovat mechanismus vzniku a rozsah delece na paternální alele. Vzhledem k tomu, že *SMN1* se nachází v oblasti, která je bohatá na *Alu* elementy, považovali jsme za nejpravděpodobnější mechanismus *Alu* elementy zprostředkovanou přestavbu (Ottesen et al., 2017). Provedli jsme *in silico* analýzu *Alu* elementů a v kandidátní oblasti intronu 1 jsme identifikovali 21 *Alu* elementů. V kandidátní oblasti intronu 5 jsme identifikovali pouze jeden element *AluSq*. Provedli jsme 4 PCR reakce, vždy s primerem specifickým pro *AluSq* v intronu 5 s jedním z *Alu* univerzálních primerů (Obrázek 15). Sangerovo sekvenování odhalilo dva produkty obsahující jak sekvenci původního *AluSq*, tak sekvenci nového chimérického *Alu* vzniklého rekombinací s jedním z *AluSq* elementů v intronu 1. Pomocí této sekvence jsme navrhli PCR primery pro specifické překlenutí této delece a získali 8978 bp produkt odpovídající deleci *SMN1Δ(2a-5)* u pacientky a jejího otce. Sekvenací získaného produktu jsme určili rozsah této nové, *Alu* elementy zprostředkované delece jako NC_000005.9:g.70232118-70241095del; NM_000344.3:c.82-2548_723+515del.

Příčinou SMA u této pacientky je kompletní delece *SMN1* na maternální alele a *Alu* elementy zprostředkovaná delece exonů 2a-5 *SMN1* na paternální alele (Obrázek 16).



Obrázek 16: Schematická reprezentace genů SMN1 a SMN2 u pacientky (P), matky (M) a otce (F). Růžové boxy reprezentují mateřské alely, modré boxy reprezentují paternální alely. Červené křížky označují identifkované delece. Černé vertikální přerušované čáry reprezentují místa, do kterých nasedají próby pro qPCR a MLPA analýzu. U pacientky je označena fáze zděděných alel. Model ukazuje u pacientky i) kompletní deleci jedné alely SMN1, zděděné od matky a detekované pomocí qPCR a MLPA ii) deleci druhé alely SMN1, zděděné od otce a detekované pomocí E3I3 qPCR a analýzy transkriptů ii) deleci jedné SMN2 alely u matky, detekované pomocí MLPA a +1kb qPCR

Rutinní diagnostika pomocí MLPA analýzy a následná sekvenace kódujících oblastí *SMN1* u této pacientky odhalila pouze maternálně zděděnou heterozygotní deleci *SMN1Δ(7-8)*. MLPA

analýza testuje deleci exonů 7 a 8 a je schopná odlišit *SMN1* od *SMN2* na základě jednonukleotidového polymorfismu. MLPA neumožňuje určit rozsah delecí ani identifikovat delece v jiných oblastech *SMN* genů. Vzhledem k vysoké genomické komplexitě této oblasti a genetické heterogenitě SMA je pravděpodobně mnoho výsledků rutinní diagnostiky falešně negativních. V současné době zůstává ~50 % pacientů s podezřením na SMA bez genetické diagnózy (Karakaya et al., 2018). Identifikace genetické příčiny SMA je pro pacienty klíčová, protože pouze pacienti s bialelickými mutacemi v *SMN1* mohou podstoupit genovou terapii (Michelson et al., 2018). V současné době existují dva typy genetické léčby. Léčba pomocí opakovaného podávání antisense oligonukleotidů Spinraza® (nusinersen) a jednorázová genová terapie Zolgensma® (onasemnogene abeparvovec-xioi), kterou je v současné době možné podat pouze do 2 let věku dítěte a je schválena pouze v USA. Na schválení v EU se zatím čeká. Vzhledem k tomu, že naší pacientce bylo v době stanovení diagnózy 2 roky a osm měsíců, již léčbu přípravkem Zolgensma® podstoupit nemohla, rodičům však stanovení přesné genetické příčiny dává možnost využít prenatální testování v případě dalšího těhotenství.

U pacientů se silným podezřením na SMA a s negativním rutinním diagnostickým testem doporučujeme pro identifikaci/vyloučení genomických přestaveb a stanovení genetické diagnózy provést paralelní sekvenaci *SMN* transkriptů, analýzu genové dávky *SMN* a určení množství SMN proteinu v mononukleárních buňkách periferní krve.

5.3.3 Onemocnění s neuronálními intranukleárními inkluzemi – *NOTCH2NLC*

Ivana Jedlickova, Anna Pristoupilova, Helena Hulkova, (XXXX) CGG repeats in *NOTCH2NLC* are not expanded in a patient with infantile neuronal intranuclear inclusion disease. (In review)

***

Onemocnění s neuronálními intranukleárními inkluzemi (NIID, MIM #603472; https://omim.org/) je autosomálně dominantní, pomalu progredující neurodegenerativní onemocnění. Je charakterizované řadou klinických příznaků, jako je mozečková ataxie, pyramidální a extrapyramidální symptomy, periferní neuropatie a zhoršení kognitivních funkcí postupně vedoucích k demenci. Patologické nálezy ukazují eosinofilní intranukleární inkluze nejen v neuronech, ale i dalších tkáních. NIID se podle doby nástupu onemocnění, neurologických symptomů a výsledků magnetické reznonance mozku dělí na infantilní (iNIID), juvenilní (jNIID) a adultní (aNIID) (Sone et al., 2016; Takahashi-Fujigasaki, 2003). V poslední

době bylo diagnostikováno mnoho pacientů s aNIID a to díky zavedení efektivní *ante-mortem* diagnostiky založené na vyšetření kožní biopsie (Sone et al., 2016, 2014).

Infantilní forma NIID je extrémně vzácné neuropediatrické onemocnění, které se v několika klinických a (neuro)patologických aspektech liší od aNIID. Dodnes bylo popsáno pouze sedm iNIID pacientů (Mano et al., 2007; McFadden et al., 2005; Pilson et al., 2018). Šest z těchto pacientů mělo velmi podobnou klinickou prezentaci charakterizovanou náhlým zhoršením mentálního a motorického vývoje před čtvrtým rokem života vedoucí k úmrtí před rokem devátým. Převažujícími symptomy byla hypotonie a poruchy funkce mozečku. Všech sedm pacientů bylo diagnostikováno *post-mortem*, na základě nálezu neuronálních intranukleárních inkluzí v nervovém systému. *Ante-mortem* kožní biopsie byla provedena pouze u jednoho pacienta a byla negativní (Pilson et al., 2018).

Čtyři nezávislé studie nedávno identifikovaly expanzi repetic v genu *NOTCH2NLC* jako kauzální u familiálních a sporadických případů (>130 pacientů) s aNIID a jNIID (Deng et al., 2019; Ishiura et al., 2019; Sone et al., 2019; Tian et al., 2019). Žádná z těchto studií nezahrnovala pacienty infantilní.

V naší laboratoři jsme stanovili diagnózu iNIID u 7 letého pacienta pomocí *post-mortem* neuropatologické analýzy. *Ante-mortem* kožní biopsie byla negativní. Abychom zjistili, jestli u pacienta došlo k expanzi repetic v genu *NOTCH2NLC*, navrhli jsme PCR primery specifické pro repetitivní oblast genu *NOTCH2NLC (*primery byly navrhnuty tak, abychom se vyhnuli nespecifické amplifikaci homologních genů *NOTCH2NLA*, *NOTCH2NLB*, *NOTCHNLR* a *NOTCH2*). Provedli jsme LR-PCR u pacienta a zdravé kontroly a získaný ~1200 bp dlouhý produkt jsme osekvenovali pomocí technologie SMRT na sekvenátoru Sequel od firmy PacBio. Byly provedeny dva typy analýzy – CCS a analýza dlouhých amplikonů (LAA, Long amplicon analysis; Bowman and Ranade, 2014). Výsledky z LAA ukázaly rozdílné počty repetic v duplikátech. Francis et al. (2018) popisují obdobný fenomén a vyslovují hypotézu, že oblasti repetetivní a překrývající se mohou způsobit chyby v LAA algoritmu.

Podívali jsme se tedy zpět do hrubých sekvenačních dat obsahujících CCS čtení a provedli vlastní bioinformatickou analýzu. V první fázi jsme vybrali čtení odpovídající genu *NOTCH2NLC* a sekvence odpovídající reverznímu komplementu převedli na sekvenci dopřednou. Poté jsme extrahovali repetitivní oblast pomocí specifických sekvencí, které se nacházejí před a za touto oblastí a kvantifikovali počet CGG repetic. Zjistili jsme, že pacient nese na každé alele rozdílný počet repetic: $AGG(CGG)_9(AGG)_2CGG$/ $AGG(CGG)_{15}(AGG)_2CGG$. Počet repetic odpovídá hodnotám zjištěným u naší zdravé kontroly

i u dříve reportovaných zdravých jedinců (Deng et al., 2019; Ishiura et al., 2019; Sone et al., 2019; Tian et al., 2019). Onemocnění u našeho pacienta tedy není způsobené expanzí repetic v genu *NOTCH2NLC*.

Abychom nalezli případnou kauzální mutaci, provedli jseme celogenomové sekvenování na přístroji Illumina NovaSeq. Nepodařilo se nám určit žádnou konkrétní kandidátní variantu, u které bychom mohli provést funkční studie. Interpretace genomových dat je obtížná, a to obzvlášť v případě, kdy není dostupný genetický materiál od žádného z příbuzných ani od dalších pacientů se stejným onemocněním, a není tak možné snížit počet počet kandidátních variant.

Naše výsledky naznačují, že NIID může být geneticky heterogenním onemocněním. Při diagnostice pacientů s iNIID může docházet k falešně negativním výsledkům v důsledku nepřítomnosti neuronálních intracelulárních inkluzí v kožní biopsii a vyloučením expanze repetic v genu *NOTCH2NLC*. Negativní výsledky těchto testů nevylučují diagnózu NIID i přesto, že genetická příčina zůstává u těchto případů zatím neznámá. Určení genetické příčiny je komplikováno nedostatkem pacientů s iNIID, které souvisí s nemožností toto onemocnění diagnostikovat *ante-mortem*.

6 Souhrn výsledků

Tato disertační práce představuje využití nových metod analýzy genomu, konkrétně nových metod sekvenace genomu druhé a třetí generace, k určení a charakterizaci kauzálních genů, genových mutací a genomových změn v případech vzácných geneticky podmíněných onemocnění neznámé etiologie a definuje základní biologické a patologické procesy podmiňující studované fenotypy.

V rámci řešení projektu bylo dále využito již zavedených technik genotypování, technologie DNA čipů, vazebné analýzy, homozygotního mapování, analýzy změn počtu kopií a analýzy změn genové exprese.

Souhrnnými výsledky s ohledem na dílčí cíle disertační práce jsou:

Ad Cíl 1) Vývoj a validace metody pro cílené sekvenování genů podmiňujících dědičné metabolické poruchy (METABO panel) a jeho využití při studiu souboru pacientů s podezřením na některou z dědičných metabolických poruch či poruchu mitochondriálního energetického metabolismu.

a) Identifikace mutací v genu *SEC61A1* jako kauzální příčiny autosomálně dominantního tubulointersticiálního onemocnění ledvin (nově pojmenováno jako ADTKD-*SEC61A1*) u dvou rodin pomocí kombinace vazebné analýzy, cíleného sekvenování METABO panelem a exomového sekvenování. **Příloha 1a**
b) Vyšetření pacientů s podezřením na adultní neuronální ceroidní lipofuscinózu (ANCL) pomocí kombinace cíleného sekvenování METABO panelem a exomového sekvenování. Identifikace kauzálních mutací v genu *CLN6* u jednoho pacienta a vyloučení mutací ve známých ANCL genech u 13 pacientů. **Příloha 1b**

Ad Cíl 2) Identifikace kauzálních genů a mutací u vybraných vzácných geneticky podmíněných onemocnění pomocí NGS metod druhé generace (SOLiD, Illumina).

a) Identifikace kauzální mutace v genu *DNAJC5* u rodiny s ANCL pomocí exomového sekvenování. U této rodiny byly mutace v *DNAJC5* prvotně vyloučeny Sangerovým sekvenováním (v důsledku preferenční amplifikace krátké alely) a neodhaleny při první analýze exomových dat (nízké pokrytí a chybné nastavení prohlížeče genomických variant). **Příloha 2a**
b) Identifikace kauzálního genu *VPS15* u neurodegenerativního onemocnění neznámé etiologie pomocí exomového sekvenování. **Příloha 2b**

c) Identifikace kauzální nekódující mutace v genu *NDUFAF6*, a objasnění mechanismu vzniku Akadské varianty Fanconiho syndromu pomocí kombinace metod exomového a genomového sekvenování, homozygotního mapování a sekvenování transkriptů. **Příloha 2c**

Ad Cíl 3) Identifikace variant v obtížně analyzovatelných oblastech genomu pomocí NGS metod druhé a třetí generace (SOLiD, Illumina, Oxford Nanopore, PacBio).

a) Vývoj a aplikace bioinformatických postupů pro identifikaci a lokalizaci mutací v repetitivní oblasti genu *MUC1*, způsobujících ADTKD-*MUC1*. Identifikace pěti nových posunových mutací pomocí amplikonového sekvenování na sekvenátoru Illumina a speciální bioinformatické analýzy. **Příloha 3a**
b) Identifikace delecí na obou alelách genu *SMN1* u pacientky se spinální svalovou atrofií, u které byla pomocí standardní MLPA diagnostiky identifikována pouze mutace na maternální alele. Bylo využito amplikonového sekvenování gDNA a transkriptů, kvantitativní PCR, a *Alu*-specifického PCR. **Příloha 3b**
c) Vyloučení expanze tandemových repetic v genu *NOTCH2NLC* u onemocnění s neuronálními intranukleárními inkluzemi pomocí amplikonového sekvenování technologií SMRT. **Příloha 3c**

7 Praktický význam dosažených výsledků

Moje práce významně přispěla k zavedení nových metod analýzy genomu v laboratoři pro studium vzácných nemocí Kliniky dětského a dorostového lékařství na 1. lékařské fakultě UK a VFN. Konkrétně jsem se podílela na zavedení nových metod sekvenace genomu druhé a třetí generace (platformy SOLiD, Illumina, PacBio a Oxford Nanopore), bioinformatické analýze, interpretaci získaných dat a ověřování kandidátních variant.

Tyto postupy byly využity ve více než 30 projektech (nejen u vzácných onemocnění) a vedly k určení a charakterizaci kauzálních genů, genových mutací a genomových změn u autosomálně dominantní tubulointersticiální onemocnění ledvin ADTKD-*SEC61A1* (příloha 1a) a ADTKD-*MUC1* (příloha 3a), autosomálně dominantní neuronální ceoridní lipofuscinózy (*CLN6*, příloha 1b ; *DNAJC5,* příloha 2a), neurodegenerativního onemocnění neznámé etiologie (*VPS15,* příloha 2b), Akadské varianty Fanconiho syndromu (*NDUFAF6,* příloha 2c), spinální svalové atrofie (*SMN1,* příloha 3b), GAPO syndromu (*ANTXR1;* Stránecký et al., 2013), X-vázané formy hypertrofické kardiomyopatie (*FHL1;* Hartmannova et al., 2013), Oliverova-McFarlaneova syndromu (*PNPLA6;* Kmoch et al., 2015), myoklonické epilepsie (*C9orf72;* van den Ameele et al., 2018), deficitu lipoproteinové lipázy (*LPL1;* Kolářová et al., 2014), mitochondriálních onemocnění (*TK2* a *AARS2;* Mazurova et al., 2017), statinových myopatií (*SLCO1B;* Neřoldová et al., 2016; Stránecký et al., 2016) a impulzivního násilí (Vevera et al., 2019).

Identifikace kauzálních genů a mutací umožnila diagnostiku a prevenci onemocnění s využitím metod prenatální a preimplantační diagnostiky. Určení molekulární podstaty onemocnění dále přispělo k detailnějšímu porozumění základních biologických a patologických procesů podmiňujících studované fenotypy, jejichž pochopení je klíčové pro prevenci, léčbu onemocnění a vývoj nových terapeutických postupů. Důsledkem některých studí, bylo vylepšení stávajících metod DNA diagnostiky či zavedení metod nových. Před rokem 2011, kdy jsem začala pracovat v laboratoři pro studium vzácných nemocí KDDL 1. LF UK a VFN, byla diagnostická výtěžnost u vzácných onemocnění pouhé 1 %. Zavedením nových metod analýzy genomu se zvýšila na 50 %.

Zavedené technologie a postupy dnes tvoří základní instrumentální a metodickou platformu pro studium nejen vzácných, ale i komplexních a onkologických onemocnění. Tato platforma je v rámci výzkumné infrastruktury Národního centra lékařské genomiky (NCLG, www.ncmg.cz) využívaná i dalšími klinickými pracovišti v ČR a ve světě. V rámci NCLG jsme

vytvořili veřejně přístupnou referenční databáze genetických variant České republiky, která v současné době obsahuje 1055 vzorků. Tento soubor je dále rozšiřován s přibývajícím počtem sekvenovaných vzorků a představuje tak významný zdroj informací a údajů o populačně specifické variabilitě české populace.

8 Seznam publikací, které nejsou součástí disertace

Tort F, Ugarteburu O, Texidó L, Gea-Sorlí S, García-Villoria J, Ferrer-Cortès X, Arias Á, Matalonga L, Gort L, Ferrer I, Guitart-Mampel M, Garrabou G, Vaz FM, **Pristoupilova A**, Rodríguez MIE, Beltran S, Cardellach F, Wanders RJ, Fillat C, García-Silva MT, Ribes A. **Mutations in TIMM50 cause severe mitochondrial dysfunction by targeting key aspects of mitochondrial physiology**. *Hum Mutat*. 2019;40(10):1700-1712. doi:10.1002/humu.23779. **IF= 4,453**

Vevera J, Zarrei M, Hartmannová H, Jedličková I, Mušálková D, **Přistoupilová A**, Oliveriusová P, Trešlová H, Nosková L, Hodaňová K, Stránecký V, Jiřička V, Preiss M, Příhodová K, Šaligová J, Wei J, Woodbury-Smith M, Bleyer AJ, Scherer SW, Kmoch S. **Rare copy number variation in extremely impulsively violent males**. *Genes Brain Behav*. 2019;18(6):e12536. doi:10.1111/gbb.12536. **IF= 3,157**

van den Ameele J, Jedlickova I, **Pristoupilova A**, Sieben A, Van Mossevelde S, Ceuterick-de Groote C, Hůlková H, Matej R, Meurs A, Van Broeckhoven C, Berkovic SF, Santens P, Kmoch S, Dermaut B. **Teenage-onset progressive myoclonic epilepsy due to a familial C9orf72 repeat expansion.** *Neurology*. 2018;90(8):e658-e663. doi:10.1212/WNL.0000000000004999. **IF= 8,689**

Mazurova S, Magner M, Kucerova-Vidrova V, Vondrackova A, Stranecky V, **Pristoupilova A**, Zamecnik J, Hansikova H, Zeman J, Tesarova M, Honzik T. **Thymidine kinase 2 and alanyl-tRNA synthetase 2 deficiencies cause lethal mitochondrial cardiomyopathy: case reports and review of the literature**. *Cardiol Young*. 2017;27(05):936-944. doi:10.1017/S1047951116001876. **IF= 0,978**

Neřoldová M, Stránecký V, Hodaňová K, Hartmannová H, Piherová L, **Přistoupilová A**, Mrázová L, Vrablík M, Adámková V, Hubáček JA, Jirsa M, Kmoch S**. Rare variants in known and novel candidate genes predisposing to statin-associated myopathy.** *Pharmacogenomics*. 2016;17(13):1405-1414. doi:10.2217/pgs-2016-0071. **IF= 2,35**

Stránecký V, Neřoldová M, Hodaňová K, Hartmannová H, Piherová L, Zemánková P, **Přistoupilová A**, Vrablík M, Adámková V, Kmoch S, Jirsa M. **Large copy-number variations in patients with statin-associated myopathy affecting statin myopathy-related loci.** *Physiol Res*. 2016;65(6):1005-1011. doi:10.33549/physiolres.933284. **IF= 1,461**

Esteban-Jurado C, Vila-Casadesús M, Garre P, Lozano JJ, **Pristoupilova A**, Beltran S, Muñoz J, Ocaña T, Balaguer F, López-Cerón M, Cuatrecasas M, Franch-Expósito S, Piqué JM, Castells A, Carracedo A, Ruiz-Ponte C, Abulí A, Bessa X, Andreu M, Bujanda L, Caldés T, Castellví-Bel S. **Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer**. *Genet Med*. 2015;17(2):131-142. doi:10.1038/gim.2014.89. **IF= 7,71**

Kmoch S, Majewski J, Ramamurthy V, Cao S, Fahiminiya S, Ren H, MacDonald IM, Lopez I, Sun V, Keser V, Khan A, Stránecký V, Hartmannová H, **Přistoupilová A**, Hodaňová K, Piherová L, Kuchař L, Baxová A, Chen R, Barsottini OGP, Pyle A, Griffin H, Splitt M, Sallum J, Tolmie JL, Sampson JR, Chinnery P, Boycott K, MacKenzie A, Brudno M, Bulman D, Dyment D, Banin E, Sharon D, Dutta S, Grebler R, Helfrich-Foerster C, Pedroso JL, Kretzschmar D, Cayouette M, Koenekoop RK. **Mutations in PNPLA6 are linked to photoreceptor degeneration and various forms of childhood blindness**. *Nat Commun*. 2015;6:5614. doi:10.1038/ncomms6614. **IF= 11,329**

Esteban-Jurado C, Garre P, Vila M, Lozano JJ, **Pristoupilova A**, Beltrán S, Abulí A, Muñoz J, Balaguer F, Ocaña T, Castells A, Piqué JM, Carracedo A, Ruiz-Ponte C, Bessa X, Andreu

M, Bujanda L, Caldés T, Castellví-Bel S. **New genes emerging for colorectal cancer predisposition**. *World J Gastroenterol*. 2014;20(8):1961-1971. doi:10.3748/wjg.v20.i8.1961. **IF= 2,369**

García-Cazorla A, Oyarzabal A, Fort J, Robles C, Castejón E, Ruiz-Sala P, Bodoy S, Merinero B, Lopez-Sala A, Dopazo J, Nunes V, Ugarte M, Artuch R, Palacín M, Rodríguez-Pombo P, Alcaide P, Navarrete R, Sanz P, Font-Llitjós M, Vilaseca MA, Ormaizabal A, **Pristoupilova A**, Agulló SB. **Two Novel Mutations in the BCKDK (Branched-Chain Keto-Acid Dehydrogenase Kinase) Gene Are Responsible for a Neurobehavioral Deficit in Two Pediatric Unrelated Patients**. *Hum Mutat*. 2014;35(4):470-477. doi:10.1002/humu.22513. **IF= 5,144**

Kolářová H, Tesařová M, Švecová Š, Stránecký V, **Přistoupilová A**, Zima T, Uhrová J, Volgina SY, Zeman J, Honzík T**. Lipoprotein lipase deficiency: clinical, biochemical and molecular characteristics in three patients with novel mutations in the LPL gene.** *Folia Biol (Praha)*. 2014;60(5):235-243. **IF= 1**

Ehling R, Nosková L, Stránecký V, Hartmannová H, **Přistoupilová A**, Hodaňová K, Benke T, Kovacs GG, Ströbel T, Niedermüller U, Wagner M, Nachbauer W, Janecke A, Budka H, Boesch S, Kmoch S. **Cerebellar dysfunction in a family harboring the PSEN1 mutation co-segregating with a Cathepsin D variant p.A58V**. *J Neurol Sci*. 2013;326(1-2):75-82. doi:10.1016/j.jns.2013.01.017. **IF= 2,262**

Hartmannova H, Kubanek M, Sramko M, Piherova L, Noskova L, Hodanova K, Stranecky V, **Pristoupilova A**, Sovova J, Marek T, Maluskova J, Ridzon P, Kautzner J, Hulkova H, Kmoch S. **Isolated X-Linked Hypertrophic Cardiomyopathy Caused by a Novel Mutation of the Four-and-a-Half LIM Domain 1 Gene**. *Circ Cardiovasc Genet*. 2013;6(6):543-551. doi:10.1161/CIRCGENETICS.113.000245. **IF= 5,337**

Melia MJ, Kubota A, Ortolano S, Vilchez JJ, Gamez J, Tanji K, Bonilla E, Palenzuela L, Fernandez-Cadenas I, **Pristoupilova A**, Garcia-Arumi E, Andreu AL, Navarro C, Hirano M, Marti R**. Limb-girdle muscular dystrophy 1F is caused by a microdeletion in the transportin 3 gene.** *Brain*. 2013;136(5):1508-1517. doi:10.1093/brain/awt074. **IF= 10,226**

Stránecký V, Hoischen A, Hartmannová H, Zaki MS, Chaudhary A, Zudaire E, Nosková L, Barešová V, **Přistoupilová A**, Hodaňová K, Sovová J, Hůlková H, Piherová L, Hehir-Kwa JY, de Silva D, Senanayake MP, Farrag S, Zeman J, Martásek P, Baxová A, Afifi HH, St. Croix B, Brunner HG, Temtamy S, Kmoch S. **Mutations in ANTXR1 Cause GAPO Syndrome.** *Am J Hum Genet*. 2013;92(5):792-799. doi:10.1016/j.ajhg.2013.03.023. **IF= 10,987**

Nosková L, Stránecký V, Hartmannová H, **Přistoupilová A**, Barešová V, Ivánek R, Hůlková H, Jahnová H, van der Zee J, Staropoli JF, Sims KB, Tyynelä J, Van Broeckhoven C, Nijssen PCG, Mole SE, Elleder M, Kmoch S. **Mutations in DNAJC5, Encoding Cysteine-String Protein Alpha, Cause Autosomal-Dominant Adult-Onset Neuronal Ceroid Lipofuscinosis.** Am J Hum Genet. 2011;89(2):241-252. doi:10.1016/j.ajhg.2011.07.003. **IF= 10,603**

9 Použitá literatura

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015;526:68–74. https://doi.org/10.1038/nature15393.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101. https://doi.org/10.1038/ng786.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9. https://doi.org/10.1038/nmeth0410-248.

Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 2019;47:D1038–43. https://doi.org/10.1093/nar/gky1151.

van den Ameele J, Jedlickova I, Pristoupilova A, Sieben A, Van Mossevelde S, Ceuterick-de Groote C, et al. Teenage-onset progressive myoclonic epilepsy due to a familial C9orf72 repeat expansion. Neurology 2018;90:e658–63. https://doi.org/10.1212/WNL.0000000000004999.

Ameur A, Kloosterman WP, Hestand MS. Single-Molecule Sequencing: Towards Clinical Applications. Trends Biotechnol 2019;37:72–85. https://doi.org/10.1016/j.tibtech.2018.07.013.

Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 2018;46:2159–68. https://doi.org/10.1093/nar/gky066.

Arsov T, Smith KR, Damiano J, Franceschetti S, Canafoglia L, Bromhead CJ, et al. Kufs disease, the major adult form of neuronal ceroid lipofuscinosis, caused by mutations in cln6. Am J Hum Genet 2011. https://doi.org/10.1016/j.ajhg.2011.04.004.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.

Ballester LY, Luthra R, Kanagal-Shamanna R, Singh RR. Advances in clinical next-generation sequencing: target enrichment and sequencing technologies. Expert Rev Mol Diagn 2016;16:357–72. https://doi.org/10.1586/14737159.2016.1133298.

Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'Ang LY, et al. The international HapMap project. Nature 2003;426:789–96. https://doi.org/10.1038/nature02168.

Benitez BA, Alvarado D, Cai Y, Mayo K, Chakraverty S, Norton J, et al. Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. PLoS One 2011;6. https://doi.org/10.1371/journal.pone.0026741.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008. https://doi.org/10.1038/nature07517.

Bleyer AJ, Kidd K, Živná M, Kmoch S. Autosomal Dominant Tubulointerstitial Kidney Disease. Adv Chronic Kidney Dis 2017;24:86–93. https://doi.org/10.1053/j.ackd.2016.11.012.

Blumenstiel B, DeFelice M, Birsoy O, Bleyer AJ, Kmoch S, Carter TA, et al. Development

and Validation of a Mass Spectrometry–Based Assay for the Molecular Diagnosis of Mucin-1 Kidney Disease. J Mol Diagnostics 2016:1–6. https://doi.org/10.1016/j.jmoldx.2016.03.003.

Bowman B, Ranade S. A Novel Analytical Pipeline for de novo Haplotype Phasing and Amplicon Analysis using SMRT® Sequencing Technology. J Biomol Tech 2014.

Bras J, Verloes A, Schneider SA, Mole SE, Guerreiro RJ. Mutation of the parkinsonism gene ATP13A2 causes neuronal ceroid-lipofuscinosis. Hum Mol Genet 2012. https://doi.org/10.1093/hmg/dds089.

Butchbach MER. Copy Number Variations in the Survival Motor Neuron Genes: Implications for Spinal Muscular Atrophy and Other Neurodegenerative Diseases. Front Mol Biosci 2016;3. https://doi.org/10.3389/fmolb.2016.00007.

Cannon KS, Or E, Clemons WM, Shibata Y, Rapoport TA. Disulfide bridge formation between SecY and a translocating polypeptide localizes the translocation pore to the center of SecY. J Cell Biol 2005;169:219–25. https://doi.org/10.1083/jcb.200412019.

Chamberlain LH, Burgoyne RD. The cysteine-string domain of the secretory vesicle cysteine-string protein is required for membrane targeting. Biochem J 1998;335:205–9. https://doi.org/10.1042/bj3350205.

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009;19:1553–61. https://doi.org/10.1101/gr.092619.109.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 2012a;3. https://doi.org/10.3389/fgene.2012.00035.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 2012b;6:80–92. https://doi.org/10.4161/fly.19695.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 2009. https://doi.org/10.1038/nnano.2009.12.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010;38:1767–71. https://doi.org/10.1093/nar/gkp1137.

Connor TM, Hoer S, Mallett A, Gale DP, Gomez-Duran A, Posse V, et al. Mutations in mitochondrial DNA causing tubulointerstitial kidney disease. PLoS Genet 2017;13:e1006620. https://doi.org/10.1371/journal.pgen.1006620.

Cooper GM. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15:901–13. https://doi.org/10.1101/gr.3577405.

Cornec-Le Gall E, Olson RJ, Besse W, Heyer CM, Gainullin VG, Smith JM, et al. Monoallelic Mutations to DNAJB11 Cause Atypical Autosomal-Dominant Polycystic Kidney Disease. Am J Hum Genet 2018;102:832–44. https://doi.org/10.1016/j.ajhg.2018.03.013.

Cotter D, Guda P, Fahy E, Subramaniam S. MitoProteome: mitochondrial protein sequence database and annotation system. Nucleic Acids Res 2004;32:D463-7. https://doi.org/10.1093/nar/gkh048.

Crawford TO, Paushkin S V., Kobayashi DT, Forrest SJ, Joyce CL, Finkel RS, et al. Evaluation of SMN Protein, Transcript, and Copy Number in the Biomarkers for Spinal Muscular Atrophy (BforSMA) Clinical Study. PLoS One 2012;7:e33572.

https://doi.org/10.1371/journal.pone.0033572.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res 2019;47:D745–51. https://doi.org/10.1093/nar/gky1113.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8. https://doi.org/10.1093/bioinformatics/btr330.

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol 2018;19:121. https://doi.org/10.1186/s13059-018-1505-2.

Davydov E V, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol 2010;6:e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

Deng J, Gu M, Miao Y, Yao S, Zhu M, Fang P, et al. Long-read sequencing identified repeat expansions in the 5′UTR of the NOTCH2NLC gene from Chinese patients with neuronal intranuclear inclusion disease. J Med Genet 2019;56:758–64. https://doi.org/10.1136/jmedgenet-2019-106268.

Devuyst O, Olinger E, Weber S, Eckardt KU, Kmoch S, Rampoldi L, et al. Autosomal dominant tubulointerstitial kidney disease. Nat Rev Dis Prim 2019;5. https://doi.org/10.1038/s41572-019-0109-9.

Van Diggelen OP, Thobois S, Tilikete C, Zabot MT, Keulemans JLM, Van Bunderen PA, et al. Adult neuronal ceroid lipofuscinosis with palmitoyl-protein thioesterase deficiency: First adult-onset patients of a childhood disease. Ann Neurol 2001. https://doi.org/10.1002/ana.1103.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet 2018;34:666–81. https://doi.org/10.1016/j.tig.2018.05.008.

Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res 2017;27:1895–903. https://doi.org/10.1101/gr.225672.117.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 2010;327:78–81. https://doi.org/10.1126/science.1181498.

Eckardt KU, Alper SL, Antignac C, Bleyer AJ, Chauveau D, Dahan K, et al. Autosomal dominant tubulointerstitial kidney disease: Diagnosis, classification, and management - A KDIGO consensus report. Kidney Int 2015;88:676–83. https://doi.org/10.1038/ki.2015.28.

Ehling R, Nosková L, Stránecký V, Hartmannová H, Přistoupilová A, Hodaňová K, et al. Cerebellar dysfunction in a family harboring the PSEN1 mutation co-segregating with a Cathepsin D variant p.A58V. J Neurol Sci 2013;326:75–82. https://doi.org/10.1016/j.jns.2013.01.017.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science 2009;323:133–8. https://doi.org/10.1126/science.1162986.

European Union. Regulation (EC) N°141/2000 of the EuropeanParliament and of the Council of 16 December 1999 on orphanmedicinal products 2000. http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:018:0001:0005:EN:PDF.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error

probabilities. Genome Res 1998;8:186–94. https://doi.org/10.1101/gr.8.3.186.

Federal Food, Drug and CA. Orphan Drug Act. Wkly Compil Pres Doc 1983;19:2049–66.

Fernandes J. Diagnostika a léčba dědičných metabolických poruch. Prague: Triton; 2008.

Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods 2009;6:S6–12. https://doi.org/10.1038/nmeth.1376.

Francis F, Dumas MD, Davis SB, Wisser RJ. Clustering of circular consensus sequences: Accurate error correction and assembly of single molecule real-time reads from multiplexed amplicon libraries. BMC Bioinformatics 2018. https://doi.org/10.1186/s12859-018-2293-0.

Freeman JL. Copy number variation: New insights in genome diversity. Genome Res 2006;16:949–61. https://doi.org/10.1101/gr.3677206.

Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv Prepr ArXiv 2012;1207.3907.

Gazal S, Gosset S, Verdura E, Bergametti F, Guey S, Babron M-C, et al. Can whole-exome sequencing data be used for linkage analysis? Eur J Hum Genet 2016;24:581–6. https://doi.org/10.1038/ejhg.2015.143.

Gijselinck I, Cruts M, Van Broeckhoven C. The Genetics of C9orf72 Expansions. Cold Spring Harb Perspect Med 2018;8:a026757. https://doi.org/10.1101/cshperspect.a026757.

Gorenberg EL, Chandra SS. The Role of Co-chaperones in Synaptic Proteostasis and Neurodegenerative Disease. Front Neurosci 2017;11:248. https://doi.org/10.3389/fnins.2017.00248.

Gray TM, Matthews BW. Intrahelical hydrogen bonding of serine, threonine and cysteine residues within α-helices and its relevance to membrane-bound proteins. J Mol Biol 1984;175:75–81. https://doi.org/10.1016/0022-2836(84)90446-7.

Greaves J, Chamberlain LH. Dual Role of the Cysteine-String Domain in Membrane Binding and Palmitoylation-dependent Sorting of the Molecular Chaperone Cysteine-String Protein. Mol Biol Cell 2006;17:4748–59. https://doi.org/10.1091/mbc.e06-03-0183.

Greaves J, Salaun C, Fukata Y, Fukata M, Chamberlain LH. Palmitoylation and Membrane Interactions of the Neuroprotective Chaperone Cysteine-string Protein. J Biol Chem 2008;283:25014–26. https://doi.org/10.1074/jbc.M802140200.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 2019;47:D853–8. https://doi.org/10.1093/nar/gky1095.

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15:57–61. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G.

Hart TC, Gorry MC, Hart PS, Woodard AS, Shihabi Z, Sandhu J, et al. Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy. J Med Genet 2002;39:882–92. https://doi.org/10.1136/jmg.39.12.882.

Hartmannova H, Kubanek M, Sramko M, Piherova L, Noskova L, Hodanova K, et al. Isolated X-Linked Hypertrophic Cardiomyopathy Caused by a Novel Mutation of the Four-and-a-Half LIM Domain 1 Gene. Circ Cardiovasc Genet 2013;6:543–51. https://doi.org/10.1161/CIRCGENETICS.113.000245.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: Overviews and challenges. Biotechniques 2014;56. https://doi.org/10.2144/000114133.

Ignatov KB, Blagodatskikh KA, Shcherbo DS, Kramarova T V., Monakhova YA, Kramarov VM. Fragmentation Through Polymerization (FTP): A new method to fragment DNA for next-generation sequencing. PLoS One 2019;14:e0210374. https://doi.org/10.1371/journal.pone.0210374.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45. https://doi.org/10.1038/nature03001.

Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, et al. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. Nat Genet 2019;51:1222–32. https://doi.org/10.1038/s41588-019-0458-z.

Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.

Karakaya M, Storbeck M, Strathmann EA, Delle Vedove A, Hölker I, Altmueller J, et al. Targeted sequencing with expanded gene profile enables high diagnostic yield in non-5q-spinal muscular atrophies. Hum Mutat 2018;39:1284–98. https://doi.org/10.1002/humu.23560.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 2019. https://doi.org/10.1101/531210.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res 2002;12:996–1006. https://doi.org/10.1101/gr.229102.

Kirby A, Gnirke A, Jaffe DB, Barešová V, Pochet N, Blumenstiel B, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat Genet 2013;45:299–303. https://doi.org/10.1038/ng.2543.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5. https://doi.org/10.1038/ng.2892.

Kmoch S, Majewski J, Ramamurthy V, Cao S, Fahiminiya S, Ren H, et al. Mutations in PNPLA6 are linked to photoreceptor degeneration and various forms of childhood blindness. Nat Commun 2015;6:5614. https://doi.org/10.1038/ncomms6614.

Knief C. Analysis of plant microbe interactions in the era of next generation sequencing technologies. Front Plant Sci 2014;5. https://doi.org/10.3389/fpls.2014.00216.

Kolářová H, Tesařová M, Švecová Š, Stránecký V, Přistoupilová A, Zima T, et al. Lipoprotein lipase deficiency: clinical, biochemical and molecular characteristics in three patients with novel mutations in the LPL gene. Folia Biol (Praha) 2014;60:235–43.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Genome Res 2017;27:722–36. https://doi.org/10.1101/gr.215087.116.

Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL. Overview of target enrichment strategies. Curr Protoc Mol Biol 2015;2015:1–23. https://doi.org/10.1002/0471142727.mb0721s112.

Kožich V, Zeman J. Dědičné metabolické poruchy v pediatrii. Postgraduální Medicína

2010;12:793–9.

Lander ES, Botstein D. Homozygosity Mapping : A Way to Map Human Recessive Traits with the DNA of Inbred Children. Science 1987;236:1567–70.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921. https://doi.org/10.1038/35057062.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–7. https://doi.org/10.1093/nar/gkx1153.

Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91. https://doi.org/10.1038/nature19057.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9. https://doi.org/10.1093/bioinformatics/btp352.

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 2012;28:1307–13. https://doi.org/10.1093/bioinformatics/bts146.

Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res 2018;46:7793–804. https://doi.org/10.1093/nar/gky678.

Lindner TH, Njølstad PR, Horikawa Y, Bostad L, Bell GI, Søvik O. A novel syndrome of diabetes mellitus, renal dysfunction and genital malformation associated with a partial deletion of the pseudo-POU domain of hepatocyte nuclear factor-1β. Hum Mol Genet 1999;8:2001–8. https://doi.org/10.1093/hmg/8.11.2001.

Liu WJ, Ye L, Huang WF, Guo LJ, Xu ZG, Wu HL, et al. p62 links the autophagy pathway and the ubiqutin–proteasome system upon ubiquitinated protein degradation. Cell Mol Biol Lett 2016;21:29. https://doi.org/10.1186/s11658-016-0031-z.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5. https://doi.org/10.1038/ng.2653.

Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr Protoc Bioinforma 2013;44:1.23.1-26. https://doi.org/10.1002/0471250953.bi0123s44.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010;7:111–8. https://doi.org/10.1038/nmeth.1419.

Mano T, Takizawa S, Mohri I, Okinaga T, Shimono K, Imai K, et al. Neuronal Intranuclear Hyaline Inclusion Disease With Rapidly Progressive Neurological Symptoms. J Child Neurol 2007;22:60–6. https://doi.org/10.1177/0883073807299952.

Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods 2012;9:1185–8. https://doi.org/10.1038/nmeth.2221.

Mardis ER. Next-Generation Sequencing Platforms. Annu Rev Anal Chem 2013;6:287–303. https://doi.org/10.1146/annurev-anchem-062012-092628.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005. https://doi.org/10.1038/nature03959.

Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A 1977;74:560–4. https://doi.org/10.1073/pnas.74.2.560.

Mazurova S, Magner M, Kucerova-Vidrova V, Vondrackova A, Stranecky V, Pristoupilova A, et al. Thymidine kinase 2 and alanyl-tRNA synthetase 2 deficiencies cause lethal mitochondrial cardiomyopathy: case reports and review of the literature. Cardiol Young 2017;27:936–44. https://doi.org/10.1017/S1047951116001876.

McCarthy DJ, Humburg P, Kanapin A, Rivas M a, Gaulton K, Cazier J-B, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med 2014;6:26. https://doi.org/10.1186/gm543.

McFadden K, Hamilton RL, Insalaco SJ, Lavine L, Al-Mateen M, Wang G, et al. Neuronal Intranuclear Inclusion Disease Without Polyglutamine Inclusions in a Child. J Neuropathol Exp Neurol 2005;64:545–52. https://doi.org/10.1093/jnen/64.6.545.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. https://doi.org/10.1101/gr.107524.110.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4.

Mercuri E, Finkel RS, Muntoni F, Wirth B, Montes J, Main M, et al. Diagnosis and management of spinal muscular atrophy: Part 1: Recommendations for diagnosis, rehabilitation, orthopedic and nutritional care. Neuromuscul Disord 2018;28:103–15. https://doi.org/10.1016/j.nmd.2017.11.005.

Michelson D, Ciafaloni E, Ashwal S, Lewis E, Narayanaswami P, Oskoui M, et al. Evidence in focus: Nusinersen use in spinal muscular atrophy. Neurology 2018;91:923–33. https://doi.org/10.1212/WNL.0000000000006502.

Mirkin SM. Expandable DNA repeats and human disease. Nature 2007;447:932–40. https://doi.org/10.1038/nature05977.

Morling N. Amplification of Short Tandem Repeat Loci Using PCR. Forensic DNA Profiling Protoc., New Jersey: Humana Press; 1998, p. 173–80. https://doi.org/10.1385/0-89603-443-7:173.

Morton NE. Sequential tests for the detection of linkage. Am J Hum Genet 1955;7:277–318.

Muona M, Berkovic SF, Dibbens LM, Oliver KL, Maljevic S, Bayly MA, et al. A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. Nat Genet 2015;47:39–46. https://doi.org/10.1038/ng.3144.

Neřoldová M, Stránecký V, Hodaňová K, Hartmannová H, Piherová L, Přistoupilová A, et al. Rare variants in known and novel candidate genes predisposing to statin-associated myopathy. Pharmacogenomics 2016;17:1405–14. https://doi.org/10.2217/pgs-2016-0071.

Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4. https://doi.org/10.1093/nar/gkg509.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture

and massively parallel sequencing of 12 human exomes. Nature 2009;461:272–6. https://doi.org/10.1038/nature08250.

Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2019:165–73. https://doi.org/10.1038/s41431-019-0508-0.

Nosková L, Stránecký V, Hartmannová H, Přistoupilová A, Barešová V, Ivánek R, et al. Mutations in DNAJC5, Encoding Cysteine-String Protein Alpha, Cause Autosomal-Dominant Adult-Onset Neuronal Ceroid Lipofuscinosis. Am J Hum Genet 2011;89:241–52. https://doi.org/10.1016/j.ajhg.2011.07.003.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189.

Ottesen EW, Seo J, Singh NN, Singh RN. A Multilayered Control of the Human Survival Motor Neuron Gene Expression by Alu Elements. Front Microbiol 2017;8. https://doi.org/10.3389/fmicb.2017.02252.

Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, et al. A mitochondrial protein compendium elucidates complex I disease biology. Cell 2008;134:112–23. https://doi.org/10.1016/j.cell.2008.06.016.

Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Comput Biol 2013;9:e1003153. https://doi.org/10.1371/journal.pcbi.1003153.

Palmer DN, Barry LA, Tyynelä J, Cooper JD. NCL disease mechanisms. Biochim Biophys Acta - Mol Basis Dis 2013;1832:1882–93. https://doi.org/10.1016/j.bbadis.2013.05.014.

Payne A, Holmes N, Rakyan V, Loose M. Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics 2019;35:2193–8. https://doi.org/10.1093/bioinformatics/bty841.

Pereira R, Oliveira J, Sousa M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. J Clin Med 2020;9. https://doi.org/10.3390/jcm9010132.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet 2013;9:e1003709. https://doi.org/10.1371/journal.pgen.1003709.

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. Hum Mutat 2015;36:915–21. https://doi.org/10.1002/humu.22858.

Pilson K, Farrell M, Lynch B, Devaney D. A Case of Juvenile Onset Neuronal Intranuclear Inclusion Disease With a Negative Antemortem Skin Biopsy. Pediatr Dev Pathol 2018;21:494–6. https://doi.org/10.1177/1093526617724293.

Pippucci T, Magi A, Gialluisi A, Romeo G. Detection of Runs of Homozygosity from Whole Exome Sequencing Data: State of the Art and Perspectives for Clinical, Population and Epidemiological Studies. Hum Hered 2014;77:63–72. https://doi.org/10.1159/000362412.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110–21. https://doi.org/10.1101/gr.097857.109.

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 2018;36:983–7. https://doi.org/10.1038/nbt.4235.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA Van der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv 2017. https://doi.org/10.1101/201178.

Přistoupilová A. Využití nových sekvenačních technik v biomedicínském výzkumu. Karlova Univerzita, 2011.

Puckelwartz MJ, Pesce LL, Nelakuditi V, Dellefave-Castillo L, Golbus JR, Day SM, et al. Supercomputing for the parallelization of whole genome analysis. Bioinformatics 2014;30:1508–13. https://doi.org/10.1093/bioinformatics/btu071.

Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. Nat Biotechnol 2009. https://doi.org/10.1038/nbt.1561.

Ramadan H, Al-Din AS, Ismail A, Balen F, Varma A, Twomey A, et al. Adult neuronal ceroid lipofuscinosis caused by deficiency in palmitoyl protein thioesterase 1. Neurology 2007;68:387–8. https://doi.org/10.1212/01.wnl.0000252825.85947.2f.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17:405–23. https://doi.org/10.1038/gim.2015.30.

Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. Cancer Res 2017;77:e31–4. https://doi.org/10.1158/0008-5472.CAN-17-0337.

Robinson PN, Kohler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 2014;24:340–8. https://doi.org/10.1101/gr.160325.113.

Rostislavleva K, Soler N, Ohashi Y, Zhang L, Pardon E, Burke JE, et al. Structure and flexibility of the endosomal Vps34 complex reveals the basis of its function on membranes. Science 2015;350:aac7365. https://doi.org/10.1126/science.aac7365.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011. https://doi.org/10.1038/nature10242.

Ruschendorf F, Nurnberg P. ALOHOMORA: a tool for linkage analysis using 10K SNP array data. Bioinformatics 2005;21:2123–5. https://doi.org/10.1093/bioinformatics/bti264.

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 1985;230:1350–4. https://doi.org/10.1126/science.2999980.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 1977;74:5463–7. https://doi.org/10.1073/pnas.74.12.5463.

Schu P, Takegawa K, Fry M, Stack J, Waterfield M, Emr S. Phosphatidylinositol 3-kinase encoded by yeast VPS34 gene essential for protein sorting. Science 1993;260:88–91. https://doi.org/10.1126/science.8385367.

Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods 2014;11:361–2. https://doi.org/10.1038/nmeth.2890.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. Nat Rev Genet 2018. https://doi.org/10.1038/s41576-018-0003-4.

Seelow D, Schuelke M, Hildebrandt F, Nurnberg P. HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Res 2009;37:W593–9. https://doi.org/10.1093/nar/gkp369.

Singleton M V., Guthery SL, Voelkerding K V., Chen K, Kennedy B, Margraf RL, et al. Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families. Am J Hum Genet 2014;94:599–610. https://doi.org/10.1016/j.ajhg.2014.03.010.

Smith KR, Dahl HHM, Canafoglia L, Andermann E, Damiano J, Morbin M, et al. Cathepsin F mutations cause Type B Kufs disease, an adult-onset neuronal ceroid lipofuscinosis. Hum Mol Genet 2013. https://doi.org/10.1093/hmg/dds558.

Smith KR, Damiano J, Franceschetti S, Carpenter S, Canafoglia L, Morbin M, et al. Strikingly different clinicopathological phenotypes determined by progranulin-mutation dosage. Am J Hum Genet 2012. https://doi.org/10.1016/j.ajhg.2012.04.021.

Snoek R, Van Setten J, Keating BJ, Israni AK, Jacobson PA, Oetting WS, et al. NPHP1 (Nephrocystin-1) gene deletions cause adult-onset ESRD. J Am Soc Nephrol 2018;29:1772–9. https://doi.org/10.1681/ASN.2017111200.

Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: A Matching Tool for Connecting Investigators with an Interest in the Same Gene. Hum Mutat 2015;36:928–30. https://doi.org/10.1002/humu.22844.

Sone J, Kitagawa N, Sugawara E, Iguchi M, Nakamura R, Koike H, et al. Neuronal intranuclear inclusion disease cases with leukoencephalopathy diagnosed via skin biopsy. J Neurol Neurosurg Psychiatry 2014;85:354–6. https://doi.org/10.1136/jnnp-2013-306084.

Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat Genet 2019;51:1215–21. https://doi.org/10.1038/s41588-019-0459-y.

Sone J, Mori K, Inagaki T, Katsumata R, Takagi S, Yokoi S, et al. Clinicopathological features of adult-onset neuronal intranuclear inclusion disease. Brain 2016;139:3170–86. https://doi.org/10.1093/brain/aww249.

Stenson PD, Ball E V, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003;21:577–81. https://doi.org/10.1002/humu.10212.

Stránecký V, Hoischen A, Hartmannová H, Zaki MS, Chaudhary A, Zudaire E, et al. Mutations in ANTXR1 Cause GAPO Syndrome. Am J Hum Genet 2013;92:792–9. https://doi.org/10.1016/j.ajhg.2013.03.023.

Stránecký V, Neřoldová M, Hodaňová K, Hartmannová H, Piherová L, Zemánková P, et al. Large copy-number variations in patients with statin-associated myopathy affecting statin myopathy-related loci. Physiol Res 2016;65:1005–11.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13. https://doi.org/10.1093/nar/gky1131.

Takahashi-Fujigasaki J. Neuronal intranuclear hyaline inclusion disease. Neuropathology 2003;23:351–9. https://doi.org/10.1046/j.1440-1789.2003.00524.x.

Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Comput Biol 2016;12:e1004873. https://doi.org/10.1371/journal.pcbi.1004873.

Tambuyzer E, Vandendriessche B, Austin CP, Brooks PJ, Larsson K, Miller Needleman KI, et al. Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. Nat Rev Drug Discov 2019. https://doi.org/10.1038/s41573-019-0049-9.

Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, et al. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet 2017;101:700–15. https://doi.org/10.1016/j.ajhg.2017.09.013.

Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am J Hum Genet 2018;103:858–73. https://doi.org/10.1016/j.ajhg.2018.10.015.

The Lancet Diabetes & Endocrinology. Spotlight on rare diseases. Lancet Diabetes Endocrinol 2019;7:75. https://doi.org/10.1016/S2213-8587(19)30006-3.

Tian Y, Wang JPL, Huang W, Zeng S, Jiao B, Liu Z, et al. Expansion of Human-Specific GGC Repeat in Neuronal Intranuclear Inclusion Disease-Related Disorders. Am J Hum Genet 2019;105:166–76. https://doi.org/10.1016/j.ajhg.2019.05.013.

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 2008. https://doi.org/10.1101/gr.076463.108.

Velinov M, Dolzhanskaya N, Gonzalez M, Powell E, Konidari I, Hulme W, et al. Mutations in the gene DNAJC5 cause autosomal dominant kufs disease in a proportion of cases: Study of the parry family and 8 other families. PLoS One 2012;7:e29729. https://doi.org/10.1371/journal.pone.0029729.

Vevera J, Zarrei M, Hartmannová H, Jedličková I, Mušálková D, Přistoupilová A, et al. Rare copy number variation in extremely impulsively violent males. Genes Brain Behav 2019;18:e12536. https://doi.org/10.1111/gbb.12536.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164. https://doi.org/10.1093/nar/gkq603.

Wenzel A, Altmueller J, Ekici AB, Popp B, Stueber K, Thiele H, et al. Single molecule real time sequencing in ADTKD-MUC1 allows complete assembly of the VNTR and exact positioning of causative mutations. Sci Rep 2018;8:4170. https://doi.org/10.1038/s41598-018-22428-0.

Wu G, Haw R. Functional Interaction Network Construction and Analysis for Disease Discovery. Methods Mol Biol 2017;1558:235–53. https://doi.org/10.1007/978-1-4939-6783-4_11.

Xin W, Mullen TE, Kiely R, Min J, Feng X, Cao Y, et al. CLN5 mutations are frequent in juvenile and late-onset non-finnish patients with NCL. Neurology 2010. https://doi.org/10.1212/WNL.0b013e3181cff70d.

Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. Nat Genet 2008;40:124–5. https://doi.org/10.1038/ng0208-124.

Zhao M, Wang Qingguo, Wang Quan, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 2013;14:S1. https://doi.org/10.1186/1471-2105-14-S11-S1.

Živná M, Hůlková H, Matignon M, Hodaňová K, Vylet'al P, Kalbáčová M, et al. Dominant Renin Gene Mutations Associated with Early-Onset Hyperuricemia, Anemia, and Chronic Kidney Failure. Am J Hum Genet 2009;85:204–13. https://doi.org/10.1016/j.ajhg.2009.07.010.

10 Publikace *in extenso*, které jsou podkladem disertace