Univerzita Karlova
1. lékařská fakulta

Studijní program: Molekulární a buněčná biologie, genetika a virologie

Mgr. Petra Zemánková

Analýza kvantitativních a kvalitativních genetických znaků v patogenezi hereditárních forem solidních nádorů

Analysis of quantitative and qualitative genetic features in the pathogenesis of hereditary solid tumors

Disertační práce

Školitel: prof. MUDr. Zdeněk Kleibl, Ph.D.
Školitel konzultant: Mgr. Viktor Stránecký, Ph.D.

Praha, 2019

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, 14. června 2019

Petra Zemánková

Podpis

Identifikační záznam:

ZEMÁNKOVÁ, Petra. *Analýza kvantitativních a kvalitativních genetických znaků v patogenezi hereditárních forem solidních nádorů. [Analysis of quantitative and qualitative genetic features in the pathogenesis of hereditary solid tumors]*. Praha, 2019. Počet stran 58, počet příloh 8. Disertační práce. Univerzita Karlova, 1. lékařská fakulta, Ústav biochemie a experimentální onkologie. Vedoucí závěrečné práce Kleibl, Zdeněk.

Abstrakt

Nádorová onemocnění patří mezi druhou nejčastější příčinu úmrtí v ČR. Nosiči mutací v genech predisponujících k dědičné formě onemocnění tvoří malou, ale klinicky velmi významnou skupinu vysoce rizikových osob. V současné době jsou známy desítky predispozičních genů pro vznik hereditárních nádorových syndromů, pro jejichž analýzu se cílené sekvenování nové generace (NGS) stalo metodou první volby. NGS umožňuje rapidní zrychlení určení příčinné mutace v oblasti diagnostiky hereditárních nádorových syndromů. K identifikaci mutací v genech predisponujících ke vzniku dědičných nádorových onemocnění jsme navrhli analýzu pomocí panelového NGS včetně bioinformatického zpracování, které umožňuje spolehlivou identifikaci jednonukleotidových záměn, krátkých inzercí/delecí i rozsáhlých intragenových přestaveb. Bioinformatické postupy, popsané v této dizertační práci, jsme následně využili k validaci panelového NGS, ale i pro identifikaci alterací v konkrétních genech, která umožnila nalézt jejich doposud nepopsané asociace s dědičnými nádorovými onemocněními. Bioinformatické analýzy se staly základem pro jednotné zpracování rozsáhlých souborů dat z CZECANCA konsorcia a umožňují konstrukci frekvenční databáze variant, která slouží pro zlepšení klinické diagnostiky nádorové predispozice u pacientů v ČR. Všestrannost NGS umožňuje jeho využití i na analýzu sestřihových variant nádorových predispozičních genů, která je nezbytným předpokladem pro identifikaci patogenních variant způsobujících aberantní sestřih.

Klíčová slova:

Hereditární nádorové syndromy, sekvenování nové generace, bioinformatická analýza

Abstract

Cancer the second most common causes of death in the Czech Republic. Carriers of mutations in genes predisposing to hereditary cancers represent a small but clinically significant group of high risk individuals. Today, dozens of predisposing genes for hereditary tumor syndromes are known and targeted next generation sequencing (NGS) has become a standard approach for their analysis. NGS allows rapid acceleration diagnostics of causal mutation in high-risk individuals. To identify mutations in genes predisposing to hereditary cancers, we designed a panel NGS analysis including subsequent bioinformatics analysis allowing a reliable identification of single nucleotide variants, insertions/deletions, and large intragenic rearrangements. The bioinformatics procedures described in this thesis were used for panel NGS validation, but also for identification of alterations associating with so far undescribed hereditary tumor types. Bioinformatics analyzes have become the basis for the unified processing of large datasets from the CZECANCA consortium and enable the construction of a population-specific database of genotypes that serve to improve clinical diagnostics of cancer predisposition in Czech patients. The versatility of NGS also allows its use for RNA (cDNA-based) analyzes of splicing variants in the genes of interest, which prerequisite for aberrant splicing identification.

Key words:

Hereditary cancer syndromes, next generation sequencing, bioinformatics analysis

Poděkování

Na tomto místě bych ráda poděkovala především Zdeňkovi Kleiblovi za velkou příležitost, pracovní výzvu, cenné rady a pomoc. V neposlední řadě Viktorovi Stráneckému za trpělivost, cenné rady a podporu. Dále bych chtěla poděkovat kamarádům za podporu a kolegům za skvělou spolupráci.

Významné poděkování patří také mé rodině a Honzovi za podporu při studiu.

Práce byla podpořena granty Agentury pro zdravotnický výzkum MZČR NR NV15-28830A, NV16-29959A, 16-30954A, 762216, NV18-03-00024, NV19-03-00279, projekty Univerzity PROGRES Q28/LF1, GAUK 762216, SVV2019/260367 a UNCE/MED/016. Analýza souboru neselektovaných kontrol byla umožněna díky existenci a podpoře vědecké infrastuktury Národního centra lékařské genomiky (LM2015091) a jeho projektu zaměřeného na vytvoření referenční databáze genetických variant České republiky (CZ.02.1.01/0.0/0.0/16_013/0001634).

Obsah

1 Úvod 9

1.1 Hereditární nádorové syndromy 10

1.2 Přístupy ke studiu genetické podstaty hereditárních nádorových syndromů 14

1.2.1 Vazebné a GWAS analýzy 15

1.2.2 Instrumentální přístupy pro analýzu nádorové predispozice 15

2 Bioinformatická analýza 21

2.1 Mapování 22

2.2 Genotypování 25

2.3 Analýza počtu kopií a velkých přestaveb 26

2.4 Anotace 27

3 Východiska a cíle práce 29

4 Seznam prací, sloužících jako podklad dizertační práce 31

5 Komentář k vybraným publikovaným pracím 34

5.1 Článek 1: Hereditary truncating mutations of DNA repair and other genes in *BRCA1*/*BRCA2*/*PALB2*-negatively tested breast cancer patients. 34

5.2 Článek 2: RE: Frameshift variant FANCL*c.1095_1099dupATTA is not associated with high breast cancer risk 37

5.3 Článek 3: Identification and Functional Testing of ERCC2 Mutations in a Multi-national Cohort of Patients with Familial Breast- and Ovarian Cancer 38

5.4 Článek 4: The c.657del5 variant in NBN gene predisposes to pancreatic cancer 39

5.5 Článek 5: CZECANCA: CZEch CAncer paNel for Clinical. Application – návrh a příprava cíleného sekvenačního panelu pro identifikaci nádorové predispozice u rizikových osob v České Republice 41

5.6 Článek 6: Validation of CZECANCA (Czech CAncer paNel for Clinical Application) for targeted NGS – base analysis of hereditary cancer syndromes 44

5.7 Článek 7: Identification of deleterious germline CHEK2 mutations and their association with breast and ovarian cancer. 47

5.8 Článek 8: Multiplex PCR and NGS-based identification of mRNA splicing variants: Analysis of BRCA1 splicing pattern as a model 49

6 Shrnutí a závěr 51

7 Literatura 54

8 Přílohy: vybrané publikované práce *in extenso*

Seznam zkratek

1000g:	1000 Genomes
Align-GVGD:	Grantham Variation, Grantham Deviation
BAM:	Binary Alignment Map
BRCA1:	Breast Cancer Gene 1
BRCA2:	Breast Cancer Gene 2
CADD:	Combined Annotation Dependent Depletion
CHEK2:	Checkpoint Kinase 2
ClinVar:	Clinical Variant
CNV:	Copy Number Variation
CYP2C19:	Cytochrome P450 Family 2 Subfamily C Member 19
CZECANCA:	CZech CAncer paNel for Clinical Application
dbNSFP:	Database for Nonsynonymous SNPs' Functional Predictions
ddNTP:	dideoxyribonucleotide triphosphat
DGGE:	Denaturing Gradient Gel Electrophoresis
dHLPC:	Denaturing High Performance Liquid Chromatography
EM-G3:	buněčná linie EM-G3
ERCC2:	ERCC excision repair 2
ESP6500:	Exome Sequencing Project v.6500
ExAC:	Exome Aggregation Consortium
EYS:	Eyes Shut Homolog
FA:	Fanconiho anemie
FANCA:	FA complementation group A
FANCC:	FA complementation group C
FANCD2:	FA complementation group D2
FANCI:	FA complementation group I
FANCL:	FA complementation group L
FANCW:	FA complementation group W
FM index:	Ferragina Manzini index
GATK:	Genome Analysis Toolkit
GERP:	Genomic Evolutionary Rate Profiling
gNOMAD:	The Genome Aggregation Database
GWAS:	Genome-Wide Association Study
HeLa:	buněčná linie HeLa
HRMA:	High Resolution Melting Analysis
LARGE:	LARGE xylosyl- and glucuronyltransferase 1
LOD:	Logarithm of Odds
LRT:	Log Ratio Test
MAF:	Minor Allele Frequency
MCF7:	buněčná linie MCF7
MDA-MB-231:	buněčná linie MDA-MB-231
MLH1:	mutL Homolog 1
MLPA:	Multiplex ligation-dependent probe amplification
MSH2:	mutS Homolog 2
MSH6:	mutS Homolog 6
NBN:	Nibrin
NER:	Nucleotide Excision Repair

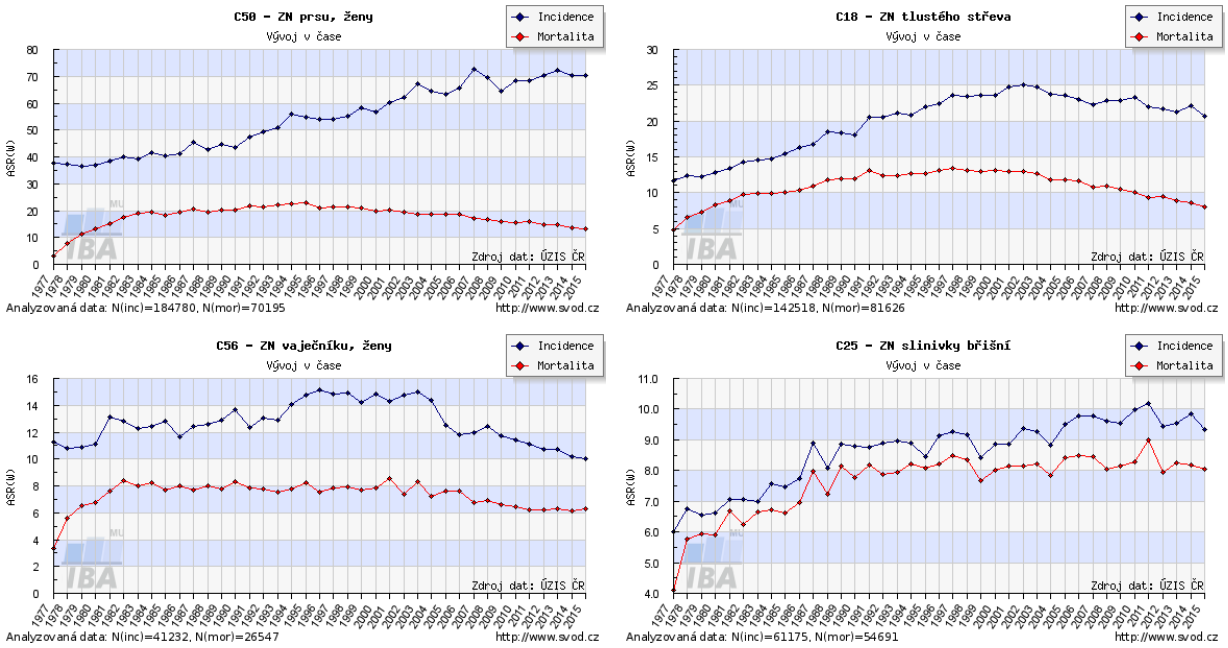
NGS:	Next Generation Sequencing
PALB2:	Partner and Localizer of BRCA2
PMS2:	PMS1 Homolog 2, mismatch repair system component
POLD1:	DNA Polymerase Delta 1, Catalytic Subunit
POLE:	DNA Polymerase Epsilon, Catalytic Subunit
PolyPhen-2:	Polymorphism Phenotyping v2
PPV:	Positive Predictive Value
RAD51:	RAD51 Recombinase
RAD51C:	RAD51 Paralog C
RAD51D:	RAD51 Paralog D
RefSeq:	The Reference Sequence
SAM:	Sequence Alignment Map
SIFT:	Sorting Intolerant From Tolerant
SLC01B1:	Solute Carrier Organic Anion Transporter Family Member 1B1
TP53:	Tumor Protein p53
UCSC:	University of California, Santa Cruz
VCF:	Variant Call Format

1 ÚVOD

Nádorová onemocnění představují po onemocněních kardiovaskulárního systému druhou nejčastější příčinu úmrtí v naší populaci (28% všech úmrtí v roce 2016; data: www.uzis.cz). Česká republika zaujímá přední místo v celoevropských statistikách incidence nádorů (Arnold M. et al., 2015).

V naší laboratoři se zabýváme především karcinomy prsu, ovaria, pankreatu a tlustého střeva, které zahrnují časté nebo prognosticky nepříznivé nádorové diagnózy (Obr. 1).

***Obr. 1:** **Trendy incidence a mortality vybraných nádorových onemocnění.** Převzato z www.svod.cz*



Celoživotní riziko **karcinomu prsu** dosahuje 10-12% v naší populaci (Kleibl Z., Kristensen V., 2016). Zatímco incidence onemocnění dlouhodobě roste (přes 7869 nových případů v roce 2016), mortalita postupně mírně klesá. Přes tento pozitivní vývoj zemřelo v roce 2016 na karcinom prsu 1921 žen, což z něj činí druhou nejčastější příčinu úmrtí na onkologická onemocnění v naší ženské populaci.

Oproti karcinomu prsu, se **karcinom ovaria** vyznačuje významně nižší incidencí (998 případů v roce 2016), avšak mortalita je značně vysoká (628 případů v roce 2016). Kromě biologických charakteristik onemocnění je jedním z důvodů nepříznivé prognózy špatná diagnostikovatelnost onemocnění, které na sebe upozorní obvykle až v pokročilých stadiích (Jessmon P. et al., 2017).

Velmi častým (a druhým nejčastějším nádorovým onemocněním u obou pohlaví) je **karcinom tlustého střeva**, v jehož incidenci patří ČR celosvětově šestá příčka. V posledních letech můžeme zaznamenat mírný pokles incidence i mortality, na kterém se pravděpodobně podílejí časnější záchyt onemocnění, zlepšení prevence a také úspěšnosti léčby (Zavoral M. et al., 2014).

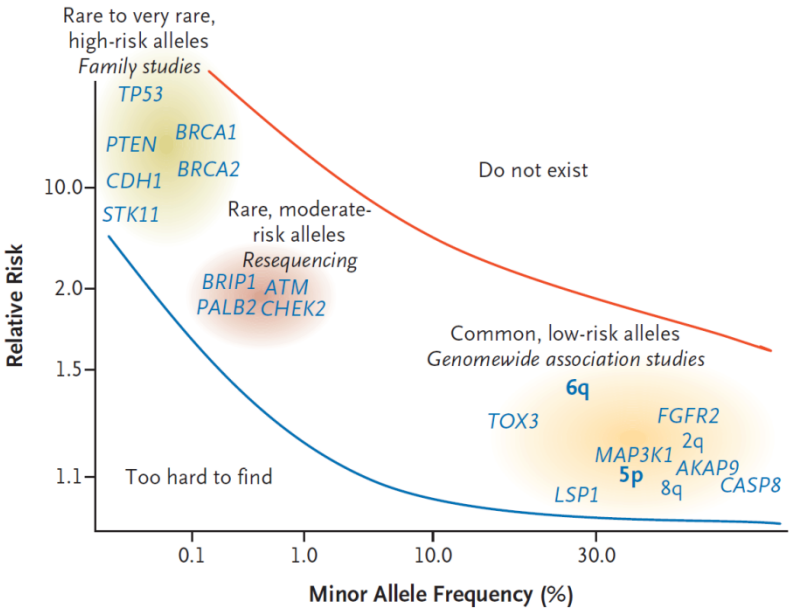
Karcinom pankreatu představuje jedno z nejzávažnějších onkologických onemocnění. Jeho incidence je ve vyspělých zemích trvale rostoucí, a protože onemocnění provází vysoká fatalita, dostává se karcinom pankreatu postupně na přední místa v příčinách úmrtí na onkologická onemocnění. Léčba karcinomu pankreatu, obzvláště v jeho pokročilých stádiích, ve kterých je diagnostikována většina případů, je po desetiletí zcela neuspokojivá a značně rezistentní i na nové trendy včetně cílené terapie (Choi M. et al., 2017).

1.1 Hereditární nádorové syndromy

U všech onkologických diagnóz převládají **sporadická onemocnění** vznikající na základě akumulace somatických mutací genomové DNA. Tyto mutace nejčastěji postihují protoonkogeny, kódující pozitivní regulátory růstu buněk a tkání, nebo tumor supresorové geny, které regulují buněčný růst negativně, indukují apoptózu nebo se podílejí na opravách poruch genomové DNA (Hanahan D., Weinberg R., 2000).

Přibližně 3% všech nádorových onemocnění však vznikají jako **hereditární nádory** (Rahman N., 2014A). Hereditární nádorové syndromy mají původ v germinálních mutacích zděděných od biologických rodičů nebo vzácněji vznikajících v germinálních buňkách *de novo*. Dědičnost hereditárních mutací nádorových predispozičních genů je převážně autosomálně dominantní. V některých případech je zastoupení dědičných nádorů podstatně vyšší než zmíněná 3%; hereditární původ můžeme vysledovat u více než 15% ovariálních karcinomů, 20% případů medulárního karcinomu štítné žlázy a více než 30% případů feochromocytomu (Rahman N., 2014A).

Dědičné mutace způsobující hereditární nádorová onemocnění postihují v naprosté většině tumor supresorové geny. Vrozená inaktivace jedné z alel zvyšuje pravděpodobnost inaktivace druhé alely, která následně podmiňuje vznik samotného onemocnění (Knudson A., 2001). Původ onemocnění v dědičnosti hraje klíčovou roli jak pro samotného probanda (ovlivňuje prognózu onemocnění a jeho léčbu), tak pro jeho příbuzné v rodině. Identifikace příčinné genetické změny je nezbytným předpokladem prediktivního testování, které v případě přítomnosti patogenní mutace u doposud asymptomatického nosiče, umožňuje jeho zařazení do preventivních programů, jejichž cílem je snížení rizika a v ideálním případě eliminace možnosti vzniku onkologického onemocnění. Klinický význam diagnostiky nádorových predispozičních genů má smysl v případě mutací se střední až vysokou penetrancí, která odpovídá alespoň dvojnásobně zvýšenému relativnímu riziku pro nosiče mutace ve srovnání s normální populací (Obr. 2, Foulkes W. D., 2008).



Obr. 2. ***Graf uvádějící vztah relativní rizika k frekvenci mutací v daných genech.***

Převzato z Foulkes W. D., 2008.

Přestože podíl hereditárních forem na celkovém počtu nádorových onemocnění je malý, vysoké riziko vzniku onemocnění činí diagnostiku dědičných forem klinicky závažným úkolem, jehož řešení významně přispívá ke zlepšení kvality života nosičů mutací (Tab. 1)

Tab. 1. ***Odhad počtu hereditárních forem nádorových onemocnění ve Velké Británii a ČR dle svod.cz s výčtem hlavních predispozičních genů.*** *Upraveno z Rahman N., 2014A (*odhad).*

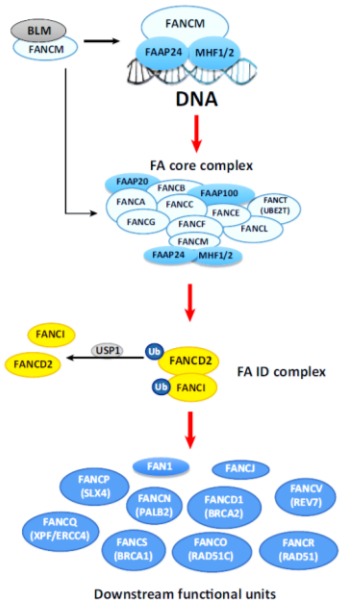
Nádorové onemocnění	Gen	Přibližný podíl, způsobený danými geny	Přibližný odhad výskytů nádorového onemocnění způsobený známými geny	
			Velká Británie	Ceská republika*
Prs	*BRCA1*	3-5%	~2000	~320
	BRCA2			
	ATM			
	BRIP1			
	CHEK2			
	CDH1			
	PALB2			
	PTEN			
	STK11			
	TP53			
Kolorektum	*APC*	3-5%	~2000	~320
	BMPR1A			
	MLH1			
	MSH2			
	MSH6			
	MUTYH			
	PMS2			
	POLD1			
	POLE			
	PTEN			
	SMAD4			
	STK11			
Ovarium	*BRCA1*	~15%	~1000	~150
	BRCA2			
	BRIP1			
	MLH1			
	MSH2			
	MSH6			
	PMS2			
	RAD51C			
	RAD51D			

Nádorových predispozičních genů bylo doposud popsáno několik set. Největší rozmach zaznamenala charakterizace „hlavních“ predispozičních genů v posledním desetiletí 20. století, kdy byly popsány i mutace v hlavních predispozičních genech pro vznik **hereditární formy karcinomu prsu a ovaria** *BRCA1* (Miky Y. et al., 1994) a *BRCA2* (Wooster R. et al., 1995). Dědičné patogenní mutace zvyšují riziko vzniku karcinomu prsu u nosičů téměř desetinásobně a zvyšují také riziko vzniku karcinomu ovarií a nádorů dalších tkání. Mutace v genu *BRCA1* zvyšují riziko vzniku karcinomu kolorekta, v genu *BRCA2* pak přispívá také ke vzniku karcinomu prsu u mužů, nádorů pankreatu a prostaty (Foulkes W. D., 2008). Oba poměrně rozsáhlé predispoziční geny kódují strukturně nepříbuzné, avšak funkčně podobné genové produkty. Proteiny BRCA1 i BRCA2 jsou velké jaderné fosfoproteiny, jejichž dominantní funkcí je podíl na opravách dvouřetězcových zlomů DNA. Jejich hlavní úlohou je formace rozsáhlých multiproteinových komplexů, vznikajících v procesu oprav cestou homologní rekombinace (Nielsen F. C. et al., 2016).

Po identifikaci genů *BRCA1* a *BRCA2* a jejich úloze v reparaci genomové DNA se karcinom prsu stal frekventovaným modelovým onemocněním, u kterého byl studován význam dědičných mutací ve stovkách různých genů, a především v genech, kódujících DNA reparační proteiny nebo regulátory buněčné odpovědi na přítomnost poškození genomové DNA v buňce. Tyto analýzy umožnily charakterizovat i další predispoziční geny s různorodou funkcí jejich proteinových produktů. Podobně jako *BRCA1* a *BRCA2*, je důležitým predispozičním genem pro vznik karcinomu prsu i *PALB2*, identifikovaný v roce 2007 (Rahman N., 2007), jehož genový produkt funkčně kooperuje s proteiny BRCA1 a BRCA2 při opravě dvouřetězcových zlomů. Další charakterizované predispoziční geny jako např. *CHEK2* a *TP53* kódují proteiny regulující odpověď na poškození DNA inhibicí buněčného cyklu v kontrolních bodech či aktivací apoptózy a senescence při selhání DNA reparačních dějů (Kleibl Z., Kristensen V., 2016). Akumulace případů karcinomu prsu a ovaria v nádorových rodinách a důležitost DNA reparačních genů v patogenezi hereditárních nádorů prsu umožnila charakterizovat i další predispoziční geny. Predispozice ke vzniku karcinomu ovaria byla před několika lety prokázána u nosičů mutací v genech *RAD51C* (Pelttari L. M. et al., 2011) a *RAD51D* (Loveday C. et al., 2011), které kódují paralogní proteiny podílející na opravě dvouřetězcových zlomů DNA cestou homologní rekombinace.

Asociace s dědičnými poruchami genů kódujících DNA reparační proteiny, které se podílejí na opravách dvouřetězcových zlomů cestou homologní rekombinace, je typickou, i když ne výlučnou, charakteristikou dědičných forem karcinomu prsu a ovaria. Velmi vzácně se vyskytující homozygotní mutace (či přítomnost *trans* patogenní mutací u složených heterozygotů) způsobují vrozená onemocnění provázená závažnými symptomy a časnou letalitou spojenou obvykle se vznikem maligních onemocnění. Typickým příkladem těchto syndromů je Fanconiho anémie, heterogenní onemocnění, která se projevuje mnohočetnými vrozenými vadami (atypické barvení kůže nebo kožní skvrny *café-au-lait*, deformované palce, růstová či mentální retardace), selháním kostní dřeně vedoucí

k pancytopenii a vysokou predispozicí k hematologickým i solidním malignitám (Che R. et al., 2017; Nalepa G. a Clapp D. W., 2018). Fanconiho anemii způsobují mutace v některém z 21 FA genů (*FANCA* až *FANCW*), které kódují proteiny podílející se na opravě meziřetězcových kovalentních spojení DNA (Obr. 3) a reparaci dvouřetězcových zlomů. Mezi pacienty s Fanconiho anemií dominují mutace ve třech genech: *FANCA* (64%), *FANCC* (12%) a *FANCG* (8%) (Che R. et al., 2017)



Obr. 3. Geny Fanconiho anemie kódují proteiny, které se účastní proximální části reparace dvouřetězcových zlomů DNA.
FA proteiny (FANCA, B, C, E, F, G, L, M, T, I) spolu s FA-asociovanými proteiny (FAAP 20/24/100 a proteiny MHF1/2) vytvářejí tzv. hlavní (core) komplex, katalyzující aktivitu ubikvitin E3 ligázy (proteinu FANCL) pro monoubikvitinizaci ID komplexu (složeného z proteinů FANCD2 a FANCI).
Ubikvitinylací aktivovaný FANCD2/FANCI komplex interaguje s řadou dalších proteinů angažovaných v procesu homologní rekombinace.

Převzato a upraveno z Che R. et al., 2017.

Na rozdíl od hereditárních forem karcinomu prsu a ovaria s dědičnými mutacemi v DNA reparačních genech pro proteiny homologní rekombinace, mutace v rodinách s **dědičným nepolypózním karcinomem kolorekta** – Lynchovým syndromem – jsou způsobeny dědičnými mutacemi mismatch-repair (MMR) genů – *MLH1*, *MSH2*, *MSH6*, *PMS2* (Lynch H. T. et al., 2015). Predilekci poruch MMR genů ke karcinomu kolorekta dokumentují i v nedávné době popsané hereditární mutace v genech pro DNA polymerázy (*POLE*, *POLD1*) účastnící se této reparační cesty, které zvyšují riziko oligopolypózních forem kolorektálního karcinomu (Palles C. et al., 2013).

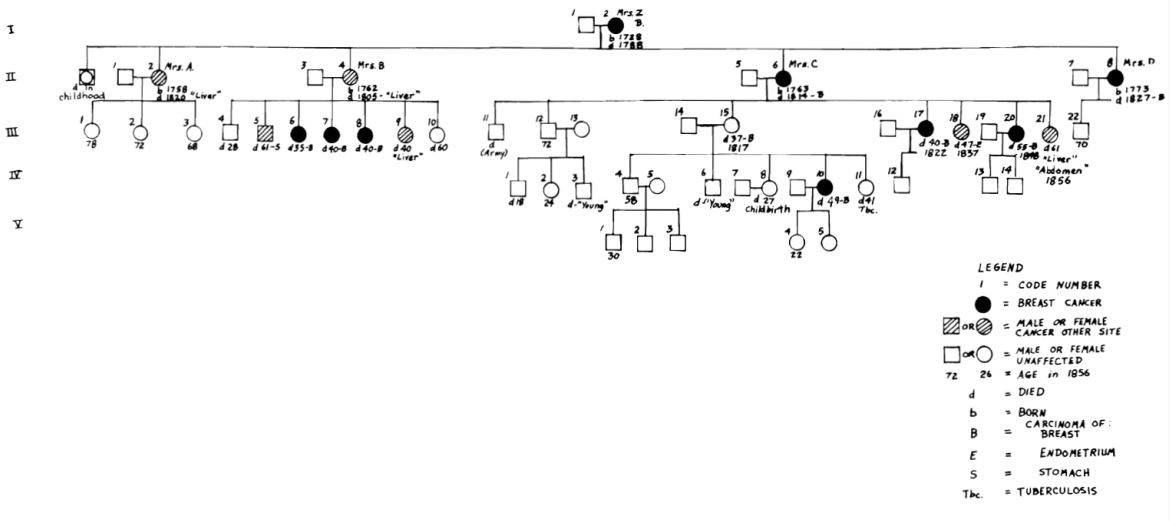
Biologický důvod, proč mutace postihující jednotlivé DNA reparační dráhy asociují především s karcinomem prsu a ovaria a jiné s kolorektálním karcinomem není doposud spolehlivě objasněn. Z klinického pohledu je však kauzalita postižení jednotlivých DNA reparačních pochodů (a v širší míře i obecně poruch tumor supresorových genů) důležitou okolností, která umožňuje u postižených osob intenzivní dohled nad vznikem místně-specifických nádorových onemocnění v jejich časném stádiu či případně chirurgickou prevenci s odstraněním rizikové tkáně u nosičů vysoce penetrantních mutací podmiňujících vznik onemocnění s vysokou mortalitou.

1.2 Přístupy ke studiu genetické podstaty hereditárních nádorových syndromů

Identifikace osob s dědičně podmíněnými nádorovými syndromy vyžaduje průkazné určení patogenní mutace. S ohledem na skutečnost, že nádorových predispozičních a kandidátních genů bylo identifikováno několik set, jejich mutace se liší **penetrancí** (mírou asociovaného rizika vzniku onemocnění), jedná se o poměrně náročný problém. Podstatnou komplikací je rovněž nedostatek informací, které máme o vztahu genotypu s fenotypem u postižených osob a rodin, protože, s výjimkou několika málo genů (jako je *BRCA1*, *BRCA2*, či mutátorové geny), jsou dědičné mutace v nádorových predispozičních genech vzácné a často populačně či geograficky specifické.

První vědecký popis dědičné predispozice ke vzniku nádorových onemocnění pochází od Paula Brocy (Krush A. J., 1979). V práci *Traite des Tumeurs* z roku 1866 Broca zdokumentoval rozsáhlý rodokmen rodiny postižené mnohočetným výskytem nádorů, především karcinomu prsu (Obr. 4).

Obr. 4. ***Rodokmen pacientky Paula Brocy****.Rodokmen popisuje 26 členů rodiny, kteří přesáhli 30 let, kdy 15 z nich vyvinulo tumor – z toho devět karcinom prsu (ženy) a šest jedinců s jiným karcinomem. Převzato z Krush A.J., 1979.*



O 50 let později Theodor Boveri navrhl, že ztráta buněčných klíčových znaků, v současné době známých jako tumor supresorové geny, je hlavní spouštěč při vývoji tumoru. Dle jeho názoru také hraje roli dědičnost (Hansford S., Huntsman D. G., 2014).

Důležitým mezníkem poznání dědičných nádorů je Knudsonova hypotéza z roku 1971, podle které hereditární nádory vznikají v důsledku dvou následných mutací - „dvou zásahů“. Prvním „zásahem“ je dědičná alterace tumor supresorového genu, která fenotypově nepůsobí problémy do fáze, než dojde ke druhému zásahu. Vznik mutace v druhé doposud zdravé alele v somatických buňkách vede k funkčnímu poškození klíčového proteinu a rozvoji nádorové transformace. Tato hypotéza byla vystavena na souboru 48 pacientů s retinoblastomem (Knudson A. G. 1971). Analýzou souboru dalších 13 pacientů s retinoblastomem, nesoucí deleci na chromosomu 13, Knudson předpověděl, že

musí existovat další (somatický) zásah ve stejném genu (Knudson A. G., 1971). Na základě jeho studií byl roku 1987 objeven predispoziční gen *RB1* (Rahman N., 2014B).

1.2.1 Vazebné a GWAS analýzy

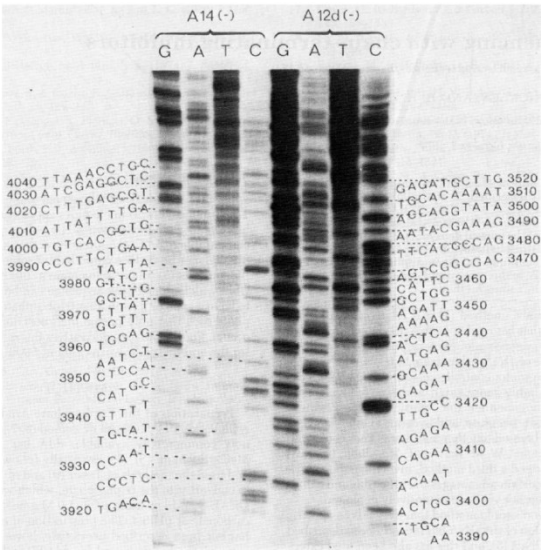
Charakterizace nádorových predispozičních genů však vyžadovala významný technologický pokrok v metodách molekulární biologie a DNA diagnostiky, ke kterému přispěl objev PCR a sekvenování. Zásadní posun v identifikaci nádorových predispozičních genů měly na počátku **vazebné analýzy**, kterými bylo v letech 1980-1990 identifikováno několik predispozičních genů. Vazebná analýza zahrnuje vyšetření stovek vysoce polymorfních markerů napříč genomem u členů rodiny (skupiny) s výskytem onemocnění rozdělených podle fenotypových charakteristik. Skupinám jedinců je pak přidělena pravděpodobnost, že daný fenotyp je vázán s jednou konkrétní alelou polymorfního markeru. Na základě těchto informací je pak vypočítán logaritmus pravděpodobnosti vazby mezi genetickým markerem a sledovaným fenotypem (logarithm of odds - LOD), hodnota ≥ 3 znamená, že je vysoce pravděpodobné, že hledaný gen se nachází blízko daného markeru, zatímco hodnota < -2 tuto možnost vazby vylučuje (Foulkes W. D., 2008).

Modifikací vazebných analýz umožněné dalším technologickým pokrokem a zavedením vysoko-kapacitních (a vysoce denzních) metod jsou **celogenomové asociační studie** (Genome-Wide Associations studies; GWAS), analyzující frekvenci jednonukleotidových polymorfismů napříč celým genomem v populaci u tisíců zdravých a nemocných jedinců. První GWAS studie proběhla v roce 2005, kdy byla odhalena příčina věkem podmíněné makulární degenerace. V současné době dokázaly GWAS studie identifikovat 450 variant, které jsou asociovány se zvýšeným rizikem vývoje nádorů (Sud A. et al., 2017), U karcinomu prsu se v poslední době podařilo identifikovat 65 nových nízkopenetrantních lokusů, které agregovaně zvyšují riziko vývoje onemocnění (Michailidou K. et al., 2017).

1.2.2 Instrumentální přístupy pro analýzu nádorové predispozice

Rozvoj GWAS analýz je spojen s rozvojem vysoko kapacitních analýz nukleových kyselin, kterou odstartovalo masivní používání čipových technologií. **Čipy (DNA arrays)** obsahují oligonukleotidové sondy, které jsou komplementární ke studovaným úsekům analyzovaných nukleových kyselin. Po hybridizaci analyzovaných vzorků genetického materiálu na sondy kovalentně ukotvené k povrchu čipu je měřena míra specifické interakce díky různým druhů značení. Kvantifikace signálu je vyhodnocena z digitálního snímku čipu. Metody založené na principu DNA/RNA čipů mají velké využití v měření genové exprese, analýze navázání transkripčních faktorů a genotypování. V současné době však nahrazuje čipové technologie sekvenování nové generace (Bumgarner R., 2013).

V roce 1977 byl popsán princip tzv. **Sangerova sekvenování**, za použití radioaktivně značných 2',3'-dideoxyribonukleotidů (ddNTPs) terminujících polymeraci templátu DNA vytvářeného DNA polymerázou *in vitro* (Sanger F. et al., 1977). Sangerovo sekvenování bylo prezentováno sekvenací DNA bakteriofága ØX174 s elektroforetickým rozdělením fragmentů na akrylamidovém gelu. Sekvenace probíhala ve čtyřech sekvenačních reakcích dle terminačních ddNTPs odděleně analyzovaných na sekvenační elektroforéze (Obr. 5).



Obr. 5. Ukázka sekvenování na akrylamidovém gelu z původního článku Sanger F. et al., 1977. *Sekvence determinovaného úseku při použití restrikčních fragmentů A12d a A14 jako primerů komplementárního vlákna ØX174. Použité inhibitory byly ddGTP, ddATP, ddTTP a araCTP.*

Sangerova metoda výrazně zrychlila sekvenci DNA a v roce 1980 byla oceněna Nobelovou cenou za chemii. V současné době se používají fluorescenčně značené ddNTP, sekvenace probíhá najednou pro všechny typy nukleotidů a detekce je v automatizované formě spolu s kapilární elektroforézou, přičemž sekvence jsou v podobě chromatogramu. Třebaže Sangerovo sekvenování je postupně nahrazováno sekvenováním nové generace, zůstává v oblasti diagnostiky dědičných nádorových onemocnění standardem pro konfirmace variant.

Sekvenování nové generace (Next Generation Sequencing - NGS; masivní paralelní sekvenování) je dalším generačním stupněm sekvenování genetické informace, umožňujícím paralelní analýzu milionů DNA templátů v jediné analýze, často z více vzorků pacientů současně. NGS umožňuje rychlou identifikaci variant v mnoha genech až celých genomech zároveň u mnoha pacientů. Limitace tohoto přístupu je vysoká (ale neustále se snižující) ekonomická náročnost analyzátorů a analýz, vysoké nároky na specializované bioinformatické zpracování rozsáhlých souborů dat a některá technická omezení (množství sekvenačních chyb), které se liší i v rámci různých sekvenačních platforem. Významným faktorem omezujícím klinickou využitelnost NGS je především interpretace nálezů.

Třebaže výkonnost NGS umožňuje získat sekvenci celého lidského genomu nebo exomu v reálném čase, při diagnostickém sekvenování, zaměřeném na určení prediktivních či prognostických parametrů konkrétního onemocnění, jsou některé informace nadbytečné, zbytečně zvyšující náklady na vyšetření

a ztěžující bioinformatické zpracování a interpretaci výsledků. Příkladem může být zjištění přítomnosti variant v genech s doposud neznámou funkcí při exomovém sekvenování nebo identifikace variant v genech podmiňujících fenotyp nesouvisející s vyšetřovanou diagnózou (tzv. diskordantní nálezy). Proto je většina současných postupů rutinní genetické diagnostiky založena na použití různě rozsáhlých skupin vyšetřovaných genů (genových panelů). V případě identifikace nádorové predispozice umožňuje **panelové NGS** vyšetření konkrétních nádorových predispozičních genů (Soukupová J., 2016). Navržené cílové oblasti můžou pokrývat desítky i stovky genů a jsou tak zaměřeny na identifikaci predispozice ke vzniku konkrétního nádorového onemocnění nebo širšího spektra dědičných nádorových syndromů. K selektivnímu **obohacení cílových templátů** (sequence capture) lze využít mnoha technologických a koncepčních přístupů (Ballester L. Y. et al., 2016; Kozarewa I. et al., 2015). Omezení velikosti cílové oblasti sekvenované DNA umožňuje paralelní zpracování vzorků od více vyšetřovaných osob (multiplexing), což zrychluje analýzy v reálném provozu a snižuje jejich finanční náročnost (Shearer A. E. et al., 2012).

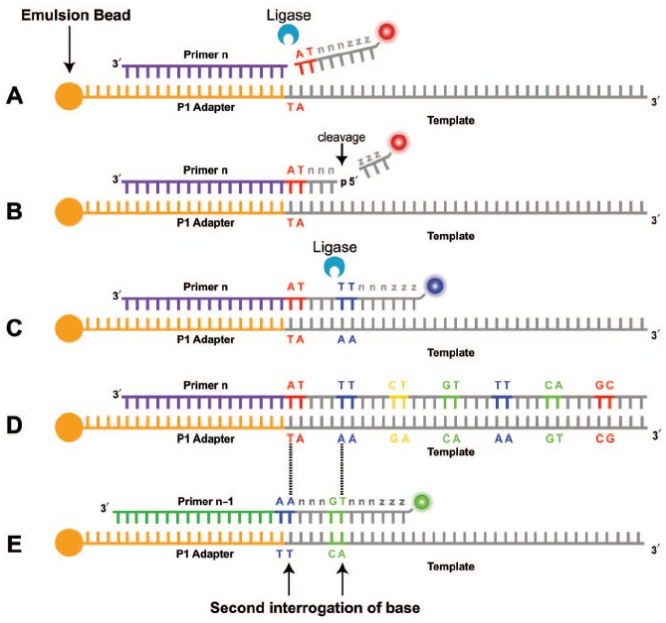
V této práci se budu zabývat dvěma panely, které jsme v naší laboratoři postupně vyvinuli pro analýzu nádorové predispozice. Oba panely zahrnovaly kódující oblasti, intron-exonové přechody a vybrané promotorové oblasti nádorových predispozičních genů nebo genů, asociovaných s prognózou nádorových onemocnění a sloužily k sekvenování na technologických platformách SOLiD a Illumina.

Základem každého NGS je **příprava sekvenační knihovny**. V případě analýz predispozice ke vzniku geneticky podmíněných onemocnění slouží jako obvyklý templát genomová DNA, která je nejčastěji izolována z leukocytů periferní krve. Zásadní limitací současných rutinně používaných sekvenačních technologií je délka čtení, která je omezena na úseky DNA nepřesahující 1 kb. Proto na počátku přípravy knihovny musí být vysokomolekulární genomová DNA štěpena pomocí ultrazvuku nebo enzymaticky směsí restrikčních enzymů (fragmentáz) na úseky požadované délky. Po štěpení jsou DNA fragmenty enzymaticky ošetřeny otupěním přesahujících konců, aby byly připraveny na navázání adaptorů, sloužících pro namnožení cílových fragmentů po obohacení a k rozlišení DNA fragmentů patřících ke konkrétnímu pacientovi (Obr. 6).

Obr. 6. Schéma DNA fragmentu připraveného k přípravě sekvenační knihovny.
Na upravené tupé konce fragmentu DNA jsou postupně ligovány oligonukleotidové adaptory. Tyto sekvence zajišťují identifikaci fragmentu DNA od konkrétní osoby (barcode) v multiplexní přípravě sekvenační knihovny, při které jsou smíchány a společně připravovány vzorky od řady (desítek) vyšetřovaných společně v jedné reakční směsi. Další oligonukleotidové úseky (P1/P2-adaptory) slouží jako univerzální sekvence pro nasedání primerů při amplifikaci cílové sekvence po jejím obohacení.
P1_ADAPTOR-XXXXXX_FRAGMENT_DNA_XXXXXX-INTERNÍ_ADAPTOR-BARCODE-P2_ADAPTOR

SOLiD sekvenování (Sequencing by Oligonucleotide Ligation and Detection) využívá k sekvenačnímu procesu hybridizaci fluorescenčně značených krátkých oligonukleotidů (prob) hybridizujících ke čtenému úseku DNA, které jsou spojovány pomocí DNA ligázy v několika cyklech (Shendure J. et al., 2005). Sekvenovaná DNA je po nabohacení cílových oblastí amplifikována pomocí

emulzní PCR na magnetických kuličkách, které jsou následně kovalentně navázány na sklíčko sekvenační komůrky. Na začátku sekvenování nasedá na sekvenovaný fragment univerzální primer, který je při sekvenačních cyklech následován sadou oktamerových sond rozpoznávajících dvoubázový motiv templátu, bezprostředně následující po posledním nukleotidu úvodního primeru (v prvním cyklu) nebo předchozí sondy (v cyklech následujících; Obr. 7).



Obr. 7. Schéma SOLiD sekvenování.

A) Ligace mezi primerem a sondou – dochází k určení sekvence AT.

B) Po zaznamenání signálu se odstraní část s fluoroforem.

C) Ligace mezi novou a předchozí sondou – dochází k určení sekvence TT.

D) Nově syntetizované vlákno po 7 cyklech.

E) Nasednutí nového primeru posunutého o jednu bázi. Zobrazení 2 cyklů, při kterých dochází k určení sekvence AA a ve druhém cyklu k určení sekvence GT.

Převzato a upraveno z Voelkerding K.V. et al., 2009.

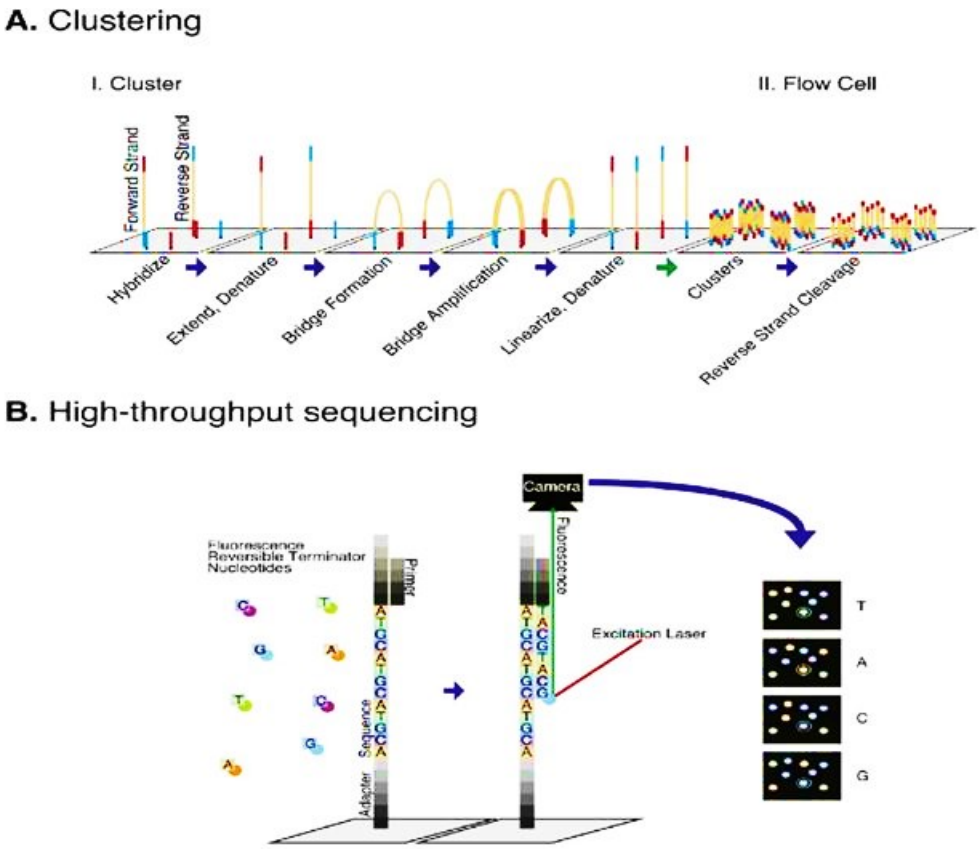
Sondy nesou na 5'-konci fluorescenční značení kodifikující rozpoznávaný dvoubázový motiv. Po hybridizaci sondy na komplementární vlákno, probíhá ligace mezi primerem a sondou (při první reakci) nebo mezi sondami (při následných cyklech). Po zaznamenání fluorescenčního signálu je fluorescenční značka odštěpena a sekvenování postupuje do následujícího ligačního cyklu. Na konci všech (obvykle 10 až 15) cyklů je syntetizované vlákno odstraněno a s posunem o jednu bázi nasedá nový primer, čímž je zajištěno dvojí čtení každé báze. V posledních verzích bylo možné pomocí SOliD sekvenování dosáhnout čtení o délce až 75bp.

Dvoubázové čtení a dvojitá identifikace každého sekvenovaného nukleotidu cílové sekvence při SOLiD sekvenování přinášelo značnou výhodu v podobě nízkého počtu sekvenačních chyb. Zásadní omezení této technologie však spočívaly v nemožnosti čtení delších fragmentů DNA, značné časové náročnosti ligačních cyklů a dosažení kapacitní meze platformy na hranici 100 Gb sekvenačního výstupu. Z důvodů těchto omezení nemohlo sekvenování technologie SOLiD čelit nástupu konkurenčních platforem a z dnešní perspektivy se jedná o ukončenou vývojovou větev NGS technologií (Shendure J. et al., 2017).

Dominantní platformou NGS v současnosti je **Illumina**, využívající k amplifikaci tzv. můstkovou PCR, kdy jsou jednořetězcové fragmenty templátů opatřené adaptory hybridizovány na primery

imobilizované na dně sekvenační komory (flowcell), (Obr. 8A). Každý fragment během amplifikace kolem sebe vytvoří tzv. cluster (Fedurco M. et al., 2006) - shluk identických kopií daného fragmentu DNA kovalentně vázaný k povrchu flowcelly. Před následujícím sekvenováním je odstraněno jedno z komplementárních vláken DNA. Vlastní Illumina sekvenování je založeno na principu sekvenování syntézou. Sekvenování probíhá za účasti sekvenačního primeru, hybridizujícího k adaptorové sekvenci v templátu a deoxynukleotidů značených fluorescenční značkou s navázaným terminátorem. V každém cyklu sekvenování se začlení právě jedna báze. Osvícením flowcelly laserem se získá signál z každého zařazeného nukleotidu, charakterizovaného fluoroforem emitujícím po ozáření jinou vlnovou délku (Obr. 8B). Po zaznamenání signálu je odstraněn terminátor s fluorescenční značkou, což uvolňuje volný 3' hydroxyl a začíná nový cyklus. Sekvenování může probíhat dvojím způsobem. Tzv. párové sekvenování (**pair end**) spočívá v postupné sekvenaci téhož fragmentu z obou konců. Při druhém způsobu (**single end**) se sekvenují fragmenty pouze z jednoho konce. V současnosti je u obou sekvenační modů maximální délka čtení až 300bp.

Obr. 8. Schéma sekvenačního procesu, probíhajícího na flowcelle. *A) Tvorba clusterů pomocí můstkové PCR. Modře a červeně jsou označeny komplementární adaptory k primerům imobilizovaným na flowcelle. Každý cluster představuje identické kopie fragmentu. B) Sekvenování syntézou – v každém cyklu se inkorporuje právě jeden nukleotid díky terminátoru s fluorescenční značkou, která je na konci cyklu po ozáření laserem odstraněna, uvolní volný 3' hydroxyl a začíná nový cyklus. Převzato z https://www.1010genome.com/illumina-sequencing/.*

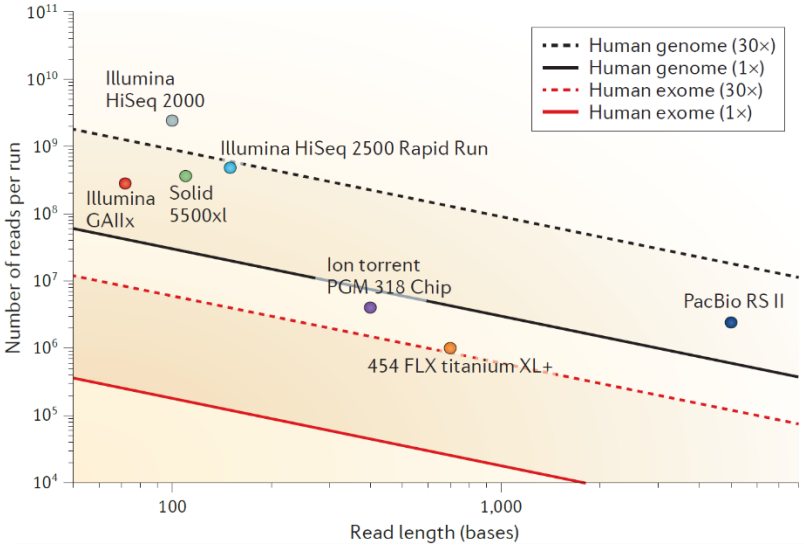


Illumina sekvenování představuje v současnosti nejrozšířenější sekvenační platformu. K dominantnímu postavení této technologie přispívá i množství sekvenačních analyzátorů pokrývající

širokou oblast požadované sekvenační kapacity v rozsahu mezi MiniSeq (7,5Gb) až NovaSeq (3Tb). Nejvíce zastoupeným analyzátorem je MiSeq, jehož sekvenační výstup (v závislosti na zvolené délce čtení) dosahuje až 15Gb.

Výhodou sekvenačních platforem Illumina je především rychlost a efektivnost. Jejich přesnost se udává kolem 99,9%. Další nespornou výhodou jsou lepší sekvenační výsledky v homopolymerních oblastech a to díky systému, který dovoluje začlenění právě jednoho nukleotidu v rámci cyklu a blokuje inkorporaci dalšího. Oproti tomu nevýhodou této metody jsou četné chyby v substitucích, kvůli šumu pozadí, který v každém dalším cyklu roste. (Ari S. a Arikan M., 2016).

Dalším stupněm technologického rozvoje je tzv. **sekvenování třetí generace**. Výhodou sekvenování třetí generace je výrazně větší délka čtení (Obr. 9) a vynechání amplifikačního kroku, což urychlí čas přípravy templátu a snižuje množství amplifikačních artefaktů.



Obr. 9. Délky a počet čtení na jednotlivých sekvenačních platformách. *Na ose X je délka čtení, na ose Y počet čtení za jeden run/analýzu a v grafu jsou zobrazeny jednotlivé sekvenační platformy.*

Převzato z Sims D. et al., 2014.

Tzv. real-time sekvenování je sekvenování pomocí nanostruktury **Zero Mode Waveguide**, které probíhá v jamkách o průměru 70 nm, umístěných na čipu s detekčním objemem 20 zeptolitrů (Korlach J., 2008). V každé jamce je navázán jeden polymerázový komplex inkorporující přesně jednu molekulu DNA (Levene M. J., 2003). Pro každý nukleotid je odpovídající fluorescenční barvivo, které umožňuje detekci začleněné báze. Tuto metodu uvedla na trh firma Pacific Biosciences.

Další představitelem technologie třetí generace je **nanoporové sekvenování**, jehož principem je translokace jednovláknové DNA přes pór o průměru jeden nm aplikací elektrického pole a měření následných fyzikálních změn, které jsou typické pro jednotlivé nukleotidy (Rusk N., 2014). Vývojem nanoporového sekvenování se v současné době zabývá společnost Oxford Nanopore Technologies.

2 BIOINFORMATICKÁ ANALÝZA

Sekvenování nové generace produkuje velké množství dat a vyžaduje rozsáhlou bioinformatickou analýzu. Zpracování sekvenačních dat probíhá dominantně na výkonných výpočetních serverech v operačním systému Linux, který je vhodný pro práci s velkými objemy dat (Gb – Tb informací). Přesto jsou bioinformatické analýzy sekvenačních dat časově náročné a trvají řádově desítky hodin strojového času. Analýza sekvenačních dat se skládá z řady na sebe navazujících kroků, které obvykle zahrnují kontrolu kvality, mapování sekvenačních čtení na referenční sekvenci, nalezení odchylek (variant) oproti referenční genomové sekvenci a jejich anotaci.

Primárním výstupem sekvenačních dat jsou soubory v podobě **FASTQ** (FASTA formát spolu s kvalitou každé báze) formátu, což jsou všechny sekvence jednotlivých analyzovaných fragmentů (reads - čtení), ohodnocené kvalitou, vyjadřující pravděpodobnost přítomnosti dané báze. FASTQ soubory představují vstupní formát pro úvodní kontrolu kvality (např. v softwaru FastQC, Andrews S. 2010), která zahrnuje kontrolu poměrného zastoupení bází (GC:AT), délku jednotlivých čtení, kvalitu každé jednotlivé báze a přítomnost sekvence adaptorů, které se mohou ve čtení objevit při sekvenování příliš krátkých fragmentů DNA. V případě přítomnosti nekvalitních bází nebo tzv. pročtení se do adaptorů lze sekvenci upravit tzv. trimmingem – odstranění nekvalitních bází na konci jednotlivých čtení nebo sekvencí adaptorů.

Upravené FASTQ soubory jsou následně mapovány na referenční genom (viz dále kapitola 2.1). Výsledkem mapování jsou data v **SAM** (sorted alignment map) formátu obsahující sekvenci čtení , její koordinátu (lokalizaci) v referenčním genomu a odchylky oproti referenční sekvenci. Pro další zpracování jsou SAM soubory nadměrně velké, proto jsou převedeny do binární podoby – **BAM** formátu, seřazeny a jsou z nich odstraněny PCR duplikáty (označující čtení totožné sekvence a délky), které by zkreslovaly kvalitu sekvenované oblasti. Volitelným krokem je rekalibrace kvality bází, která umožňuje detekci a odstranění systémových chyb sekvenačního procesu pomocí.

Takto upravené BAM soubory jsou vstupem pro vlastní hodnocení sekvenačních dat, genotypování, CNV analýzu (viz kapitola 2.3), analýza středně velkých inzercí, delecí a pro vizualizaci hodnocené oblasti (Obr. 10).

Finálním krokem bioinformatického zpracování sekvenačních dat je genotypování – nalezení odchylek od referenčního genomu s konkrétní koordinátou, kde se tato varianta nachází, poměr zastoupení referenční báze k alternativní a kvalitu, hodnotící skutečnost, zda se jedná o sekvenační chybu nebo reálnou variantu. Výstupem tohoto kroku je soubor ve formátu **VCF** (Variant Call Format), obsahující všechny nalezené varianty, které jsou následně funkčně anotovány (viz kapitola 2.4).

***Obr. 10.** **Ukázka vizualizace BAM souboru v softwaru IGV (Integrative Genome Viewer) (Robinson J.T. et al., 2011).** Oblast dlouhá 3099 párů bází zobrazuje exon genu BRCA1, který je na konkrétní koordinátě dle obrázku pokryt 117x (pokrytí zobrazuje šedá oblast). Červené a modré šipky představují jednotlivá čtení a barvy dva směry sekvenování v sekvenátoru. Svislá modročervená čára uprostřed šedé oblasti znamená přítomnost varianty. Modré čárky v jednotlivých čteních pak představují pozici alternativní báze v konkrétním čtení.*

2.1 Mapování

Základním krokem vyhodnocení sekvenačního výstupu z NGS je mapování získaných sekvencí na referenční genom. S ohledem na současný stav NGS technologií je typickou situací mapování velkého množství krátkých (stovky bp dlouhých) DNA fragmentů. Omezená délka čtení může komplikovat mapování v oblastech repetitivních sekvencí či pseudogenů.

Existuje velké množství mapovacích programů, které využívají rozdílné algoritmy. Některé z nich jsou přednostně určeny pro kratší čtení, jiné zase pro čtení o velikosti nad 100 bp.

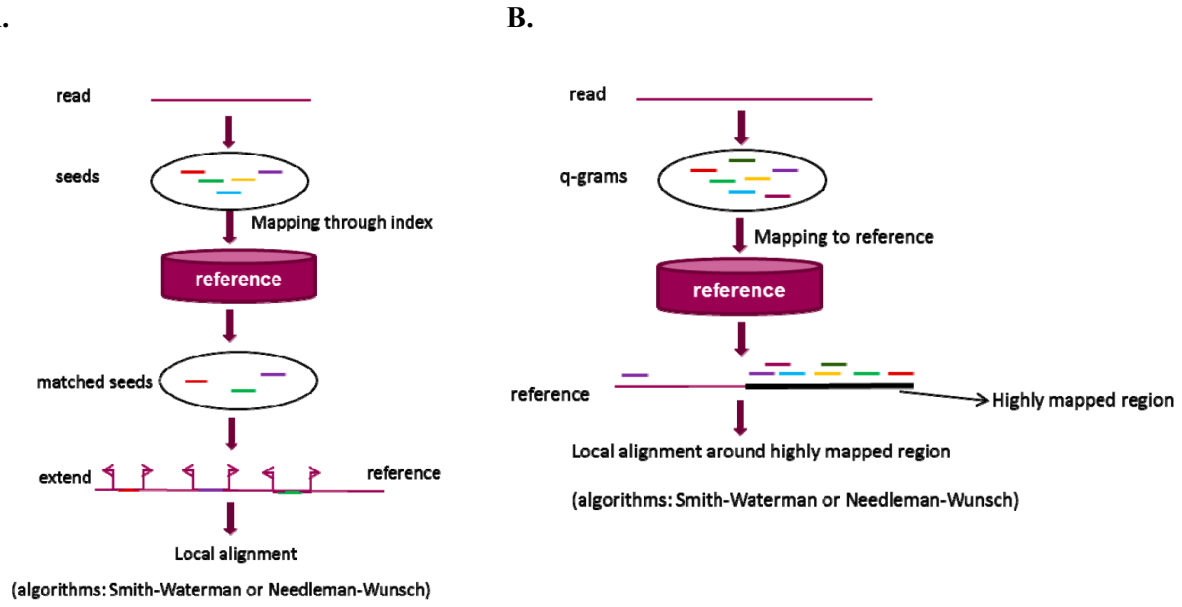
Úkolem mapovacích algoritmů je lokalizovat pozici sekvenačního čtení v referenčním genomu. Vyhledávání této podobnosti není založeno na přesné shodě sekvence čtení a referenčního genomu, protože v přítomnosti odchylky (varianty) od referenčního genomu přesná shoda nikdy nenastane. Každý algoritmus má proto nastavenou jistou toleranci vůči neshodě se vzorem v genomu.

Dva hlavní mapovací přístupy jsou tzv. „seed and extend“ algoritmus a „q-gram filter“ (Obr. 11).

Algoritmus „**seed and extend**“ zahrnuje několik kroků: vygenerování základních fragmentů – „seedů“ (což jsou části jednotlivých čtení), identifikace shodných pozic fragmentů s referenčním genomem, rozšíření fragmentů o celou sekvenci čtení a následným lokálním mapováním, které v případě přítomnosti inzerce nebo delece upřesňuje pozice začátku a konce dané změny.

Algoritmus „**q - gram filter**“ vygeneruje podobně jako u prvního algoritmu malé fragmenty – „q – gramy“, identifikuje shodné pozice s referenčním genomem díky více fragmentům a následuje opět lokální mapování. (Ye H. et al., 2015)

Obr. 11. ***Přístupy k mapování čtení na referenční genom.*** *A) Systém mapování za použití algoritmu „seed and extend“ mapuje jednotlivá čtení na referenční sekvenci. B) Systém mapování za použití algoritmu q-gram filter mapuje více q-gram fragmentů čtení najednou. Převzato z Ye H. et al., 2015.*



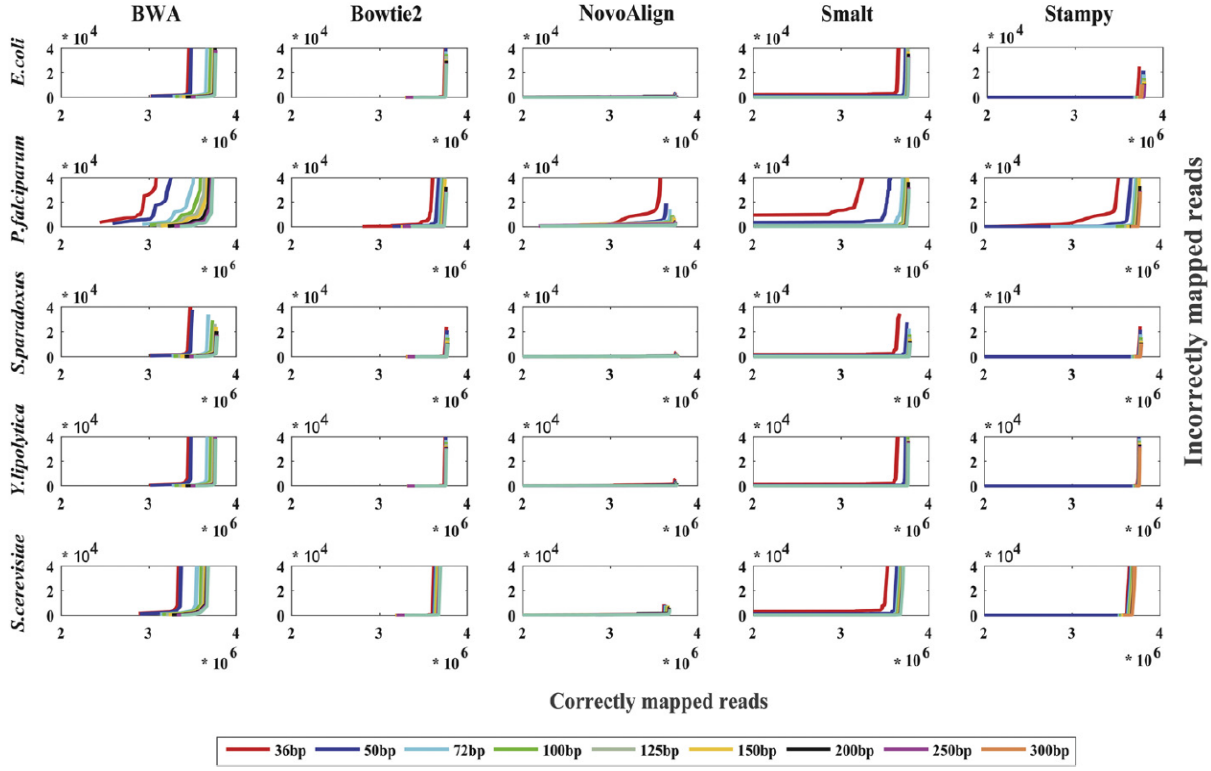
První krok mapování, tedy generování základních fragmentů (seedů nebo q-gramů), lze provést dvěma klíčovými algoritmy - pomocí tzv. **hashovací tabulky** nebo s využitím **FM indexu** (Ferragina P. a Manzini G., 2000). Oba procesy jsou založeny na vyhledávání podobnosti v textu. Hashovací tabulka zachovává v rámci čtení několik pevných pozic a následně hledá přesnou shodu. FM index využívá Burrows – Wheelerovy transformace, což je reverzibilní permutace písmen v textu. Originál je pak komprimován a v textu hledán jako shluk (Yu X. et al., 2012), proto je tento algoritmus obecně rychlejší (Ye H. et al., 2015).

Další kroky, jako rozšiřování pozic a následná shoda s genomem, se dějí pomocí různé kombinace zmíněných algoritmů s dalšími, které umožňují toleranci určitých neshod – záměn, delecí a inzercí. Každá neshoda je penalizována a stanovuje tak celkové skóre konkrétního čtení; čili tolerance mapovacího softwaru vůči odchylkám v referenčním genomu je vyjádřením kvality jednotlivých čtení. Na základě tohoto skóre kvality je sekvenační čtení buď namapováno, nebo nezařazeno. Příklady algoritmů jsou **Needleman – Wunschův** nebo **Smith – Watermanův** algoritmus.

Mapovací softwary fungují jako kombinace výše zmíněných algoritmů v rámci obecných tří kroků.

Při výběru nejvhodnějšího mapovacího programu, záleží na řadě parametrů. Z výsledků srovnávacích studií provedených na často využívaných programech jako **SOAP2** (Ruiqiang L. et al., 2009), **Bowtie** (Langmead B. et al., 2009), **BWA** (Li H. and Durbin R., 2009) a **Novoalign** (www.novocraft.com) vyplývá, že jejich shoda je v průměru větší než 90 % čtení namapovaných na stejné místo. Každý z použitých programů se vyznačuje určitými výhodami a nevýhodami, např. Novoalign podává lepší výsledky ze vstupních dat odstraněny kontaminace pomocí trimmování sekvenačních čtení (Yu X. et al., 2012). Z publikace Thankaswamy-Kosalai S. et al., vyplývá, že každý software je jinak citlivý na délku čtení při mapování na různé referenční genomy bakterií (Obr. 12).

Obr. 12. Porovnání mapovacích softwarů při různých délkách čtení. *Na ose X je počet správně namapovaných čtení, na ose Y počet nesprávně zařazených čtení. Každý řádek představuje konkrétní organismus bakterie a každý sloupec příslušný mapovací software. Devět barev v grafu znázorňuje devět kategorií velikostí sekvenačních čtení. Převzato z Thankaswamy-Kosalai S. et al., 2017.*



Výstupem mapovacích programů jsou obvykle soubory ve formátu SAM, které jsou poté převedeny na BAM soubory. Následuje odstranění PCR duplikátů. Takto upravené soubory jsou pak podrobeny dvěma volitelným analýzám – „**realignmentu**“ - znovuzhodnocení inzercí nebo delecí v oblasti, kde se varianta vyskytuje a **rekalibraci** kvality bazí – detekování sekvenačních chyb.

2.2 Genotypování

Stanovení genotypu jedinců je obecně sumarizace kvality bází a hloubky čtení z BAM souborů. Dva hlavní způsoby genotypování jsou pomocí bayesiánské pravděpodobnosti, která produkuje robustní odhady všech možných genotypů na příslušné koordinátě, nebo pomocí heuristických faktorů (minimální počet alel s alternativní bází, kvalita bází nebo hranice pokrytí v daném místě), díky nimž se vytvoří sada pravidel, které dávají vznik genotypům. Nejrozšířenějšími nástroji jsou **SAMTOOLS** (Li H., 2011) a **GATK** (Genome Analysis Toolkit, McKenna A. et al., 2010), pracující na základě bayesiánské pravděpodobnosti. Oba dokáží zajistit i více kroků dříve popsaných úprav (odstranění PCR duplikátů, realignment, rekalibrace kvality bází)

Studie porovnávající softwary pro genotypizaci přinesly rozporuplné výsledky. Hwang S. et al., 2015 udává doporučení pro kombinace genotypizačních programů s mapovacími softwary, odděleně pro identifikaci bodových záměn a inzercí/delecí. Sandmann S. et al., 2017 porovnává programy na základě PPV (positive predictive value) a F1 score (počítáno kombinací sensitivity a PPV). Vzhledem k rozdílným metodikám hodnocení a různým analyzovaným datasetům vykazovala hodnocení programů diskordantní závěry. Nejlépe s ohledem na všechny druhy analýz byl hodnocen program VarDict (Lai Z. et al., 2016), který ovšem nevykazoval dobré výsledky při zpracování vzorků z amplikonového sekvenování (Brouwer R.W.W. et al., 2018). Každý software pracuje na jiném výchozím nastavení. Z hlediska základních nastavení, je uživateli k dispozici manuál, jak dosáhnout nejlepších výsledků – odfiltrovat nekvalitní varianty a sekvenační chyby, který Sandmann S. et al ve studii nebrali v potaz, tudíž softwary, které mají doporučený postup, mohly mít lepší výsledky. Genotypizace variant je ovlivnitelná i sekvenační platformou. Každá generuje specifické druhy chyb ovlivňujících genotypizační proces (Shendure J. a Ji H., 2008).

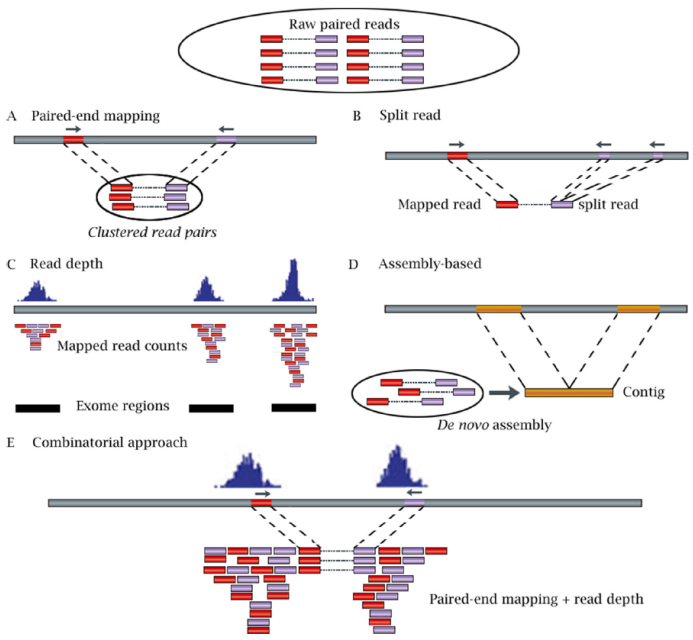
Analýza variant se liší také podle druhů vzorků, které jsou sekvenovány. Genotypizace somatických variant vyžaduje jiný přístup a nastavení v rámci samotného softwaru oproti hereditárním variantám. Totéž platí pro varianty ze sekvenování RNA.

Kritickým bodem genotypizačního kroku, jsou multialelické varianty s výskytem více alternativních bazí na jedné koordinátě (např. *CHEK2* IVS2+1G>A a IVS2+1G>T). Doporučuje se před samotným genotypováním provést normalizační krok, kdy se tato místa rozdělí na samostatná (http://annovar.openbioinformatics.org), jinak dochází ke ztrátě informací, jelikož anotační algoritmy mají s multialelickou anotací problém a neuvádějí dané místo vůbec nebo uvádějí jen polovinu informací. Ve většině případů se jedná o sekvenační chyby, které vznikají v repetitivních oblastech, nicméně je nutné odlišit sekvenační chybu od skutečných záměn na stejné koordinátě, ale dvou alelách (na jedné alele např. synonymní varianta, ale na druhé na stejné koordinátě nonsense mutace).

2.3 Analýza počtu kopií a velkých přestaveb

Důležitým krokem NGS analýzy dat je analýza velkých přestaveb tzv. „**CNV calling**“. Velké přestavby - CNV (copy number variation) varianty jsou příčinou vzniku až 2% hereditárních nádorových syndromů. Jde o velké přestavby (inzerce, delece, inverze, duplikace) v genomu, zasahující i kódující oblasti, tudíž mají následně negativní vliv na tvorbu proteinového produktu. V předchozích pracích jsme identifikovali rekurentní přestavby v genech *BRCA1*, *PALB2*, *CHEK2* (Pohlreich P. et al., 2005; Ticha I. et al., 2010; Janatova M. et al., 2013; Kleiblova P. et al., 2019).

K identifikaci CNV ze sekvenačních dat je možné využít několik možných přístupů. **Pair-endové mapování** je založeno na sekvenování pomocí pair-endového čtení. Identifikace variant pak spočívá v detekci rozdílné vzdálenosti mezi páry čtení, než je rozptyl knihovny (Obr. 13A). Vzdálenost mezi páry čtení určuje velikost samotného sekvenovaného fragmentu. Nevýhodou přístupu je neschopnost zachytit malé delece, které nepřekračují velikosti sekvenovaných fragmentů. Další přístup je založen na vyhledání oblastí, ve kterých se vyskytují tzv. rozdělená čtení (**split read-based**). Rozpoznání CNV závisí na mapování částí jednoho čtení na nesousední oblasti genomu lokalizující hranice přestavby (Obr. 13B). Nejčastěji používaný přístup je **analýza založená na hloubce pokrytí** daného segmentu, která je schopna detekovat oblasti s nižším či vyšším pokrytím než je průměr pokrytí v celé sekvenované oblasti u analyzovaného vzorku oproti skupině kontrolních vzorků, a tak určit místo delece nebo duplikace (Obr. 13C). Záleží tedy na hloubce pokrytí a uniformitě dat. Další možností je způsob identifikace CNV variant založený na novém sestavení genomové sekvence (de-novo assembly) a následném srovnáním s referenční genomovou sekvencí (Obr. 13D). Pro detekci CNV variant lze využít i kombinaci předchozích způsobů (Obr. 13E).



Obr. 13. Přehled metod umožňujících detekci CNV variant. *A) Analýza pomocí pair-endového mapování. B) split readová technika. C) Detekce pomocí algoritmu, založeného na rozdílu v pokrytí daných oblastí. D) Analýza pomocí vlastní reference. E) Kombinace předchozích algoritmů.*

Převzato z Zhao M. et al., 2013.

Každá z uvedených metod má své výhody i omezení a proto žádná z nich není schopná detekovat všechny strukturní varianty, které se v genomu mohou vyskytovat (Zhao M. et al., 2013). Zásadní zlepšení v rámci všech metod detekce CNV přinese zavedení NGS třetí generace, která používá delší čtení a umožní přesnější mapování a následně přesnější výsledky.

2.4 Anotace

Posledním krokem bioinformatického zpracování NGS dat je anotace variant – pojmenování biologické funkce, v jakém genu se varianta nachází, zda se jedná o kódující oblast a jaký má na ni dopad. Zmíněný proces se nazývá funkční anotace. V současnosti existuje několik databází poskytující referenční genomové sekvence a definice genů včetně jejich transkripčních variant. Jednotlivé transkripty stejných genů se mohou v databázích od sebe poměrně výrazně lišit. Nejpoužívanější databáze jsou **UCSC** (http://genome.ucsc.edu/; Fujita P. et al., 2011), **RefSeq** (https://www.ncbi.nlm.nih.gov/refseq/; Pruitt K. et al., 2012) a **ENSEMBL** (https://www.ensembl.org/; Flicek P. et al., 2012). Pro klinickou diagnostiku je většinou využívána databáze RefSeq nebo ENSEMBL, která obsahuje velké množství predikovaných transkripčních variant. Z tohoto důvodu při hodnocení dat dáváme přednost RefSeq databázi.

Dalším problémem při anotaci je samostatné „rozhodnutí“ anotačního softwaru, kdy si vybírá z různých transkriptů, na kterých by měla varianta ten nejhorší dopad. Při porovnání dvou anotačních softwarů – ANNOVAR (Wang K. et al., 2010) a VEP (McLaren W. et al., 2010) vyšla najevo shoda pouze v 65% variantách se ztrátou funkce (McCarthy D. J. et al., 2014).

Nalezené varianty jsou dále anotovány frekvencemi z populačních databází např. ESP6500 (https://esp.gs.washington.edu/), 1000g (Genomes Project C. et al., 2012), ExAC a gNOMAD (Das R. a Ghosh S.K., 2017). Tyto databáze obsahují frekvence jednotlivých variant v konkrétních populacích, což je praktické při hledání vzácných patogenních variant (MAF<0.05). K zachyceným variantám je s výhodou přiřadit i informace o klinické závažnosti, které jsou uvedeny např. v databázi ClinVar (Landrum M. J. et al., 2014).

Pro analýzu funkčního dopadu doposud nepopsaných variant nebo variant nejasného významu (VUS) byla vytvořena rozsáhlá kolekce predikčních programů, založených na různých algoritmech jako je **SIFT** (Kumar P. et al., 2009), **PolyPhen-2** (Adzhubei I. A. et al., 2010), **LRT** (Chun S. a Fay J. C., 2009), **MutationTaster** (Schwarz J. M. et al., 2010), **PhyloP**, (Pollard K. S. et al., 2010), **CADD** (Kircher M. et al., 2014) nebo **GERP** (Cooper G. M. et al., 2005). Tyto přístupy využívají fylogenetickou konzervovanost dané oblasti, chemické odlišnosti ve složení polypeptidového řetězce proteinové izoformy, či strukturní charakteristiky polypeptidového řetězce a umožňují odhadnout dopad varianty na daný proteinový produkt. Třebaže nejasná spolehlivost predikčních programů

neumožňuje hodnocení významu jednotlivých variant pro použití v jejich klinické interpretaci, jsou tyto nástroje vhodným doplňkem pro prioritizaci nalezených variant pro případné *in vitro* analýzy.

Pro filtraci variant je klíčová jejich klasifikace s ohledem na dopad na vznikající polypeptidový produkt. Jako patogenní lze automaticky hodnotit varianty významně posunující čtecí rámec, zkracující proteinový produkt ve vysoce konzervovaných či funkčně podstatných proteinových doménách nebo ovlivňující kanonický sestřih pre-mRNA. Význam nesynonymních variant je nezbytné hodnotit s opatrností na základě jejich popisu v dostupných databázích, případně literárních zdrojích. Nesynonymní VUS je možno hodnotit i výše uvedenými i predikčními programy. Synonymní varianty jsou zpravidla nepatogenní, některé ovšem mohou ovlivňovat sestřih pre-mRNA; pro *in silico* predikci poruch sestřihových signálů lze využít specializované predikční algoritmy (např. **spidex** https://www.deepgenomics.com/spidex; **dbNSFP** (Liu X. et al., 2011)).

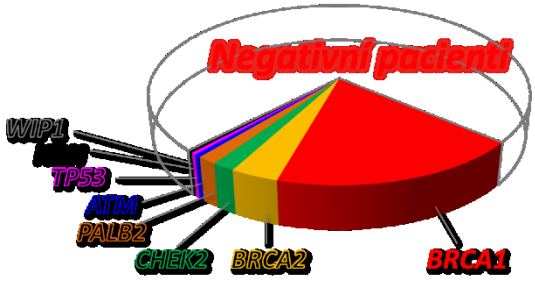
3 VÝCHODISKA A CÍLE PRÁCE

Naše pracoviště se tématem analýzy nádorové predispozice zabývá od roku 1997. Nejvíce studovanými nádory jsou karcinomy prsu a ovaria, ale s ohledem na překryv nádorových syndromů byly v průběhu let v naší laboratoři analyzovány i vzorky DNA od pacientů s dalšími malignitami. Spektrum vyšetřovaných nádorů umožnilo získat rozsáhlou kolekci vzorků a ve spolupráci s klinickými spolupracovníky i příslušných klinicko-patologických dat.

Prvotní genetické analýzy byly na našem pracovišti prováděny pomocí molekulárně-biologických metod, zahrnujících amplifikace kódujících úseků DNA, pre-screening amplifikovaných úseků na přítomnost trunkačních mutací pomocí *in vitro* transkripce/translace (protein truncation test) a ověření případné mutace pomocí Sangerova sekvenování na polyakrylamidovém gelu s radioaktivně značenými fragmenty. Postupem doby byly zavedeny další techniky pro pre-screening vrozených alterací typu malých delecí a inzercí (DGGE, dHPLC, HRMA). Radioaktivní sekvenování bylo v roce 2002 nahrazeno kapilárním automatickým sekvenováním s fluorescenčně značenými terminátory (ABI 310 a následně ABI3130). V roce 2005 byla zavedena MLPA pro stanovení rozsáhlých inzercí/delecí v analyzovaných genech.

Přes nesporný technologický posun byla však do nástupu NGS prováděna vyšetření na našem pracovišti postupně po jednotlivých genech na základě pravděpodobnosti výskytu patogenních variant. V případě mladých pacientek s karcinomem prsu či ovaria s pozitivní rodinou byly nejprve vyšetřeny nejčastější alterace v genu *BRCA1*, pokud nebyly nalezeny, byl vyšetřen celý *BRCA1* gen včetně MLPA. V případě negativity následovalo vyšetření *BRCA2* a následně vyšetření dalších genů (*PALB2*, *CHEK2*, *ATM*, *RAD51C/D, TP53*). Tyto analýzy však v průběhu 15 let ukázaly, že postupné vyšetřování po jednotlivých genech je neúnosně náročné časově i ekonomicky a počet vyšetřovaných genů negativně koreloval s výskytem jejich alterací. Zatímco nosičky mutací *BRCA1* tvořily kolem 18% indikovaných nemocných, nosičky mutací v *TP53* méně než 1% z vysoce rizikových osob ve vyšetřovaném souboru (Obr. 14).

.



Obr. 14. Analýza genů predisponujících ke vzniku karcinomu prsu u 1680 rodin s karcinomem prsu v naší laboratoři v letech 1997-2015.

BRCA1/BRCA2: Pohlreich P. et al., 2005; Ticha I. et al., 2010; Mateju M. et al., 2010
CHEK2: Kleibl Z. et al., 2005 & 2008
PALB2: Janatova M. et al., 2013
ATM: Soukupova J. et al., 2008
NBN: Mateju M. et al., 2012
WIP1: Kleiblova P. et al., 2013
Ostatní geny: nepublikovaná data

V roce 2010 proto byly v naší laboratoři zahájeny první pokusy se sekvenováním pomocí NGS. Výsledky komerčního sekvenování na panelu 80 genů u 60 nosičů známých mutací však především ukázaly, že pro další rozvoj NGS v laboratoři bude nezbytné zajistit laboratorní i bioinformatické analýzy vlastními silami. Zatímco změna laboratorních postupů, jakkoliv zásadní, umožňovala využít stávajících zkušeností molekulárně biologických technik, které byly v laboratoři dostupné, množství dat generovaných NGS vyžadovalo koncepční změnu interpretace vyšetření a především hodnocení získaných nálezů.

Cílem dizertační práce bylo:

- **vytvoření robustního bioinformatického postupu pro hodnocení NGS dat,**
- **vytvoření postupů pro kontrolu spolehlivosti analýz,**
- **zavedení postupů pro prioritizaci variant, které umožní charakterizovat kandidátní genetické prognostické a prediktivní faktory ovlivňující vznik a vývoj dědičných nádorových onemocnění analyzovaných v naší laboratoři,**
- **vytvoření databáze genotypů s fenotypovými charakteristikami pacientů (histopatologická a klinická data o onemocnění) umožňující jejich statistické zpracování.**

4 SEZNAM PRACÍ, SLOUŽÍCÍCH JAKO PODKLAD DIZERTAČNÍ PRÁCE

V časopisech s IF (bez IF – označeny)

1. Lhota F, **Zemankova P,** Kleiblova P, Soukupova J, Vocka M, Stranecky V, Janatova M, Hartmannova H, Hodanova K, Kmoch S, Kleibl Z. Hereditary truncating mutations of DNA repair and other genes in *BRCA1*/*BRCA2*/*PALB2*-negatively tested breast cancer patients. ***Clin Genet.*** 2016;90(4):324-33, (IF_{2016}= 3.326).
2. **Zemankova P,** Lhota F, Kleiblova P, Soukupova J, Vocka M, Janatova M, Kleibl Z. RE: Frameshift variant FANCL*c.1095_1099dupATTA is not associated with high breast cancer risk. ***Clin Genet.*** 2016;90(4):387-9, (IF_{2016}=3.326).
3. Rump A, Benet-Pages A, Schubert S, Kuhlmann JD, Janavičius R, Macháčková E, Foretová L, Kleibl Z, Lhota F, **Zemankova P**, Betcheva-Krajcir E, Mackenroth L, Hackmann K, Lehmann J, Nissen A, DiDonato N, Opitz R, Thiele H, Kast K, Wimberger P, Holinski-Feder E, Emmert S, Schröck E, Klink B. Identification and Functional Testing of ERCC2 Mutations in a Multi-national Cohort of Patients with Familial Breast- and Ovarian Cancer. ***PLoS Genet***. 2016;12(8):e1006248, (IF_{2016}=6.1).
4. Borecka M, **Zemankova P**, Lhota F, Soukupova J, Kleiblova P, Vocka M, Soucek P, Ticha I, Kleibl Z, Janatova M. The c.657del5 variant in NBN gene predisposes to pancreatic cancer. ***Gene***. 2016;587(2):169-72, (IF_{2016}=2.415).
5. Soukupová J, **Zemánková P**, Kleiblová P, Janatová M, Kleibl Z. CZECANCA: CZEch CAncer paNel for Clinical. Application – návrh a příprava cíleného sekvenačního panelu pro identifikaci nádorové predispozice u rizikových osob v České Republice. Klin Onkol. 2016;29 Suppl 1:S46-54.
6. Soukupova J, **Zemankova P**, Lhotova K, Janatova M, Borecka M, Stolarova L, Lhota F, Foretova L, Machackova E, Stranecky V, Tavandzis S, Kleiblova P, Vocka M, Hartmannova H, Hodanova K, Kmoch S, Kleibl Z. Validation of CZECANCA (Czech CAncer paNel for Clinical Application) for targeted NGS – based analysis of hereditary cancer syndromes. ***PLoS One***.;13(4):e0195761, (IF_{2018}=2.766).
7. Kleiblova P, Stolarova L, Krizova K, Lhota F, Hojny J, **Zemankova P**, Havranek O, Vocka M, Cerna M, Lhotova K, Borecka M, Janatova M, Soukupova J, Sevcik J, Zimovjanova M, Kotlas J, Panczak A, Vesela K, Cervenkova J, Schneiderova M, Burocziova M, Burdova K, Stranecky V, Foretova L, Machackova E, Tavandzis S, Kmoch S, Macurek L, Kleibl Z.

Identification of deleterious germline CHEK2 mutations and their association with breast and ovarian cancer. **Int J Cancer**. 2019 May 3.

8. Hojny J, **Zemankova P**, Lhota F, Sevcik J, Stranecky V, Hartmannova H, Hodanova K, Mestak O, Pavlista D, Janatova M, Soukupova J, Vocka M, Kleibl Z, Kleiblova P. Multiplex PCR and NGS-based identification of mRNA splicing variants: Analysis of BRCA1 splicing pattern as a model. ***Gene***. 2017; 30(637):41-49, (IF_{2017}=2.498).

Seznam dalších spoluautorských prací v časopisech s IF (bez IF – označeny)

9. Vocka M, Zimovjanova M, Bielcikova Z, Tesarova P, Petruzelka L, Mateju M, Krizova L, Kotlas J, Soukupova J, Janatova M, **Zemankova P**, Kleiblova P, Novotny J, Konopasek B, Chodacka M, Brychta M, Sochor M, Smejkalova-Musilova D, Cmejlova V, Kozevnikovova R, Miskarova L, Argalacsova S, Stolarova L, Lhotova K, Borecka M, Kleibl Z. Estrogen Receptor Status Oppositely Modifies Breast Cancer Prognosis in BRCA1/BRCA2 Mutation Carriers Versus Non-Carriers. ***Cancers*** (Basel). 2019 May 28;11(6).
10. Pejsova H, Hubacek JA, **Zemankova P**, Zlatohlavek L. Baseline Leptin/Adiponectin Ratio is a Significant Predictor of BMI Changes in Children/Adolescents after Intensive Lifestyle Intervention. ***Exp Clin Endocrinol Diabetes***. 2019 Mar 6. doi: 10.1055/a-0859-7041.
11. Lovecek M, Janatova M, Skalicky P, Zemanek T, Havlik R, Ehrmann J, Strouhal O, **Zemankova P**, Lhotova K, Borecka M, Soukupova J, Svebisova H, Soucek P, Hlavac V, Kleibl Z, Neoral C, Melichar B, Mohelnikova-Duchonova B. Genetic analysis of subsequent second primary malignant neoplasms in long-term pancreatic cancer survivors suggests new potential hereditary genetic alterations. ***Cancer Manag Res***. 2019;11:599-609.
12. Burke LJ, Sevcik J, Gambino G, Tudini E, Mucaki EJ, Shirley BC, Whiley P, Parsons MT, De Leeneer K, Gutiérrez-Enríquez S, Santamariña M, Caputo SM, Santana Dos Santos E, Soukupova J, Janatova M, **Zemankova P**, Lhotova K, Stolarova L, Borecka M, Moles-Fernández A, Manoukian S, Bonanni B; ENIGMA Consortium, Edwards SL, Blok MJ, van Overeem Hansen T, Rossing M, Diez O, Vega A, Claes KBM, Goldgar DE, Rouleau E, Radice P, Peterlongo P, Rogan PK, Caligo M, Spurdle AB, Brown MA. BRCA1 and BRCA2 5' noncoding region variants identified in breast cancer patients alter promoter activity and protein binding. ***Hum Mutat***. 2018;39(12):2025-2039.
13. 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.

14. Borecka M, **Zemankova P**, Vocka M, Soucek P, Soukupova J, Kleiblova P, Sevcik J, Kleibl Z, Janatova M. Mutation analysis of the PALB2 gene in unselected pancreatic cancer patients in the Czech Republic. ***Cancer Genet***. 2016;209(5):199-204.
15. Janatová M, Borecká M, Soukupová J, Kleiblová P, Stříbrná J, Vočka M, **Zemánková P**, Panczak A, Veselá K, Souček P, Foretová L, Kleibl Z. PALB2 jako další kandidátní gen pro genetické testování u pacientů s hereditárním karcinomem prsu v České republice. ***Klin Onkol***. 2016;29 Suppl 1:S31-4.
16. Janatova M, Soukupova J, Stribrna J, Kleiblova P, Vocka M, **Boudova P**, Kleibl Z, Pohlreich P. Mutation Analysis of the RAD51C and RAD51D Genes in High-Risk Ovarian Cancer Patients and Families from the Czech Republic. PLoS One. 2015 Jun 9;10(6):e0127711.

5 KOMENTÁŘ K VYBRANÝM PUBLIKOVANÝM PRACÍM

5.1 Článek 1: Hereditary truncating mutations of DNA repair and other genes in *BRCA1*/*BRCA2*/*PALB2*-negatively tested breast cancer patients.

Lhota F, **Zemankova P,** Kleiblova P, Soukupova J, Vocka M, Stranecky V, Janatova M, Hartmannova H, Hodanova K, Kmoch S, Kleibl Z. ***Clin Genet.*** 2016;90(4):324-33, (IF_{2016}= 3.326).

Uvedená práce shrnuje naše počáteční úsilí o analýzu nádorové predispozice pomocí NGS. V práci jsme navrhli sekvenační panel pro cílené obohacení vzorků zaměřené na analýzu 581 genů, pomocí kterého jsme analyzovali soubor 325 pacientů s karcinomem prsu negativně testovaných na hlavní predispoziční geny *BRCA1/BRCA2/PALB2* a 105 nenádorových kontrol na platformě SOLiD.

Panel 581 genů jsme navrhli tak, že 141 genů se účastní přímo DNA reparačních procesů a 449 genů bylo vybráno z databáze Phenopedia (https://phgkb.cdc.gov/PHGKB/startPagePhenoPedia.action; Yu W. et al., 2010) na základě slovního spojení „breast neoplasms“ s nejméně dvěma záznamy.

V rámci analýz jsme kodifikovali postup pro bioinformatické analýzy sekvenačních dat. Tento postup zahrnoval mapování pomocí softwaru Novoalign (CS1.01.08), odstranění duplikátů pomocí softwaru Picard tools a SAMtools (0.1.8) pro callování variant. Varianty byly anotovány pomocí ANNOVARu. Filtrace variant pak probíhala na základě kvality >150 a pokrytí >10x. Mezní hodnotu kvality jsme určili na základě validace různých SNP. Zatímco při kvalitě <100 se chybovost pohybovala nad hranicí 50% falešně pozitivních SNV, při kvalitě >150 klesla míra falešné pozitivity pod 5%. Díky frekvenčním databázím ESP6500 a 1000g jsme byli schopni vyloučit varianty s frekvencí >0.01. Dále byly odstraněny populačně časté varianty, u kterých není pravděpodobné, že by se podílely na podstatném zvýšení rizika vzniku nádorů (viz Obr. 2). Z tohoto důvodu jsme odstranili varianty, které se nevyskytovaly u pacientů nebo se vyskytovaly u >2 osob kontrolní populace. V posledním kroku byla provedena klasifikace variant a analyzovány varianty nejasného významu (VUS). Za patogenní jsme považovali varianty vedoucí ke zkrácení proteinového produktu z důvodu porušení čtecího rámce (pokud nebyly klasifikovány jinak v databázi ClinVar – např. rekurentní benigní varianta p.K3326* v genu *BRCA2*). Dále byly jako patogenní varianty označeny změny kanonických sestřihových míst (c.+/- 1 a 2) a intronové varianty způsobující aberantní sestřih při analýze z RNA. Jako patogenní byly rovněž hodnoceny varianty klasifikované jako patogenní a potenciálně patogenní v databázi ClinVar, které neměly konfliktní hodnocení. Nesynonymní nové varianty, které byly identifikovány predikčními programy jako škodlivé (SIFT, PolyPhen-2, LRT, MutationTaster a PhyloP), byly

označeny jako potenciálně patogenní. Za nevýznamné jsme považovali varianty, které byly jako benigní či pravděpodobně benigní uvedeny v databázích ClinVar a HGMD.

Za použití výše popsané filtrace jsme z počátečního počtu 491,385 variant získali 4,540 vzácných variant (2,647 unikátních variant). V souboru pacientů jsme identifikovali 127 trunkačních variant, 34 variant in-frame delecí/inzercí, z 1,599 nesynonymních unikátních variant jich 356 bylo identifikováno potenciálně patogenních (Obr. 15, označeny oranžově). Varianty zkracující proteinový produkt jsme našli celkem u 32% pacientů (Obr. 15), u 9% pacientů byly zaznamenány germinální varianty v klinicky významných predispozičních genech.

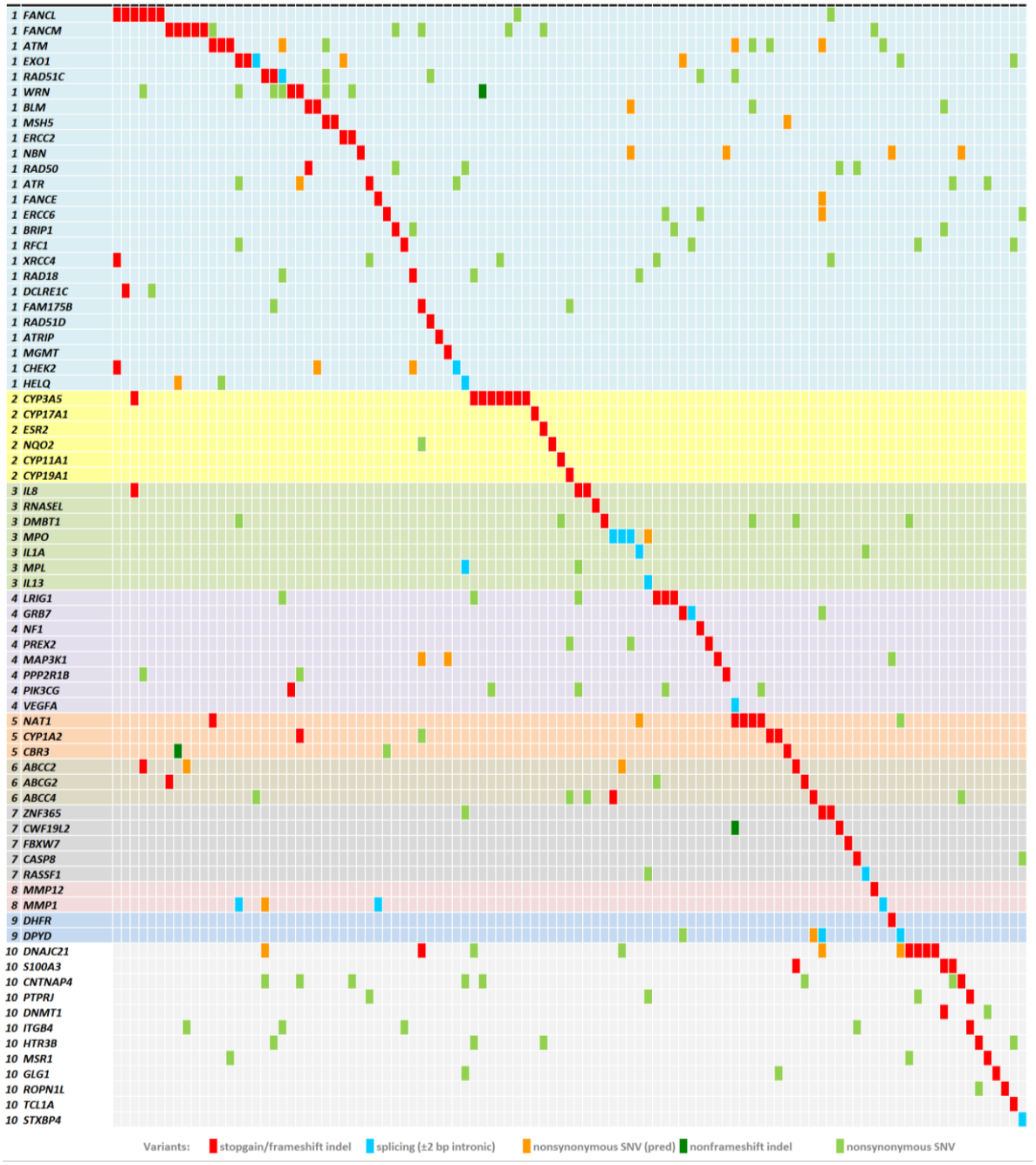
Největší pozornost jsme věnovali variantám vedoucím ke zkrácení proteinového produktu s posunem čtecího rámce. Celkem 36 těchto variant jsme nalezli u 25 genů kódujících proteiny zúčastněné v DNA reparačních pochodech. Mezi nejčastěji postižené geny patřily geny kódující proteiny komplexu Fanconiho anemie. V analyzovaném souboru jsme u šesti pacientek s karcinomem prsu a (žádné kontroly) zachytili variantu c.1096_1099dupATTA genu *FANCL*. Proteinový produkt genu *FANCL* je klíčová ubikvitin ligáza FA core complexu (viz Obr. 3). Varianta c.1096_1099dupATTA přiléhá k PHD/RING finger doméně katalyzující ubikvitinylaci a již dříve byla identifikována Alim A. M. et al. 2009 u pacienta s mírnými projevy Fanconiho anemie komplementační skupiny L. Proto jsme provedli dodatečnou genotypizaci této varianty na souboru 337 vysoce rizikových pacientů s karcinomem prsu, 673 pacientů se sporadickým karcinomem prsu a u 686 nenádorových kontrol pomocí High-Resolution Melting Analysis (HRMA). Varianta c.1096_1099dupATTA po dodatečné analýze identifikována u 9/662 (1,3%) vysoce rizikových pacientů, 3/673 (0,4%) pacientek se sporadickým karcinomem prsu a u 3/791 (0,4%) kontrol. Statisticky významný rozdíl ve frekvenci výskytu byl zachycen pouze mezi vysoce rizikovými pacienty a kontrolami (p_{Fisher}= 0.04). V rámci analýz jsme si povšimli, že všichni nosiči této mutace v populaci pacientů byly ženy, zatímco všichni nosiči v souboru kontrol byli muži.

Funkční vztahy mezi ostatními 48 geny postiženými 53 germinálními trunkačními mutacemi v ostatních („nereparačních“) genech u 74/325 (22,8%) pacientů a 10/105 (9,5%) kontrol jsme analyzovali pomocí funkčních anotací v nástrojích STRING (www.string-db.org) a KEGG (www.genome.jp/kegg). Tyto geny jsme zařadili do devíti funkčních skupin (gene group 2-10 v Obr. 15) a naznačili některé další možné funkční skupiny genů ovlivňující nádorovou predispozici ke vzniku karcinomu prsu.

Výsledky této studie nám umožnili vyhodnotit proveditelnost panelového NGS pro analýzu nádorové predispozice od návrhu panelu, přes optimalizace fragmentace DNA, obohacení cílových oblastí, konstrukci sekvenačních knihoven, bioinformatickou analýzu a interpretaci nálezů. Studie zároveň umožnila identifikovat některá slabá místa analýz, jako značnou velikost cílového panelu s obtížnou

interpretací variant v genech s málo známým funkčním vztahem ke karcinomu prsu a časově i ekonomicky náročné sekvenování na platformě SOLiD.

Obr. 15. ***Zobrazení variant v genech, u kterých byly identifikovány trunkační mutace.*** *Pacienty představují sloupce, v řádcích je uveden název genu, ve kterém se vyskytuje varianta u příslušného pacienta.*



5.2 Článek 2: RE: Frameshift variant FANCL*c.1095_1099dupATTA is not associated with high breast cancer risk.

Zemankova P, Lhota F, Kleiblova P, Soukupova J, Vocka M, Janatova M, Kleibl Z. ***Clin Genet.*** 2016;90(4):387-9, (IF_{2016}=3.326).

V reakci na předchozí článek Lhota et al., 2016 (Kapitola 5.1), ve kterém jsme identifikovali zvýšený výskyt varianty c.1096_1099dupATTA v genu *FANCL* u pacientek s dědičnou formou karcinomu prsu, provedli Pfeifer et al., 2016 genotypizaci této varianty u pacientů a kontrol z Německa a Makedonie. V analyzovaných souborech identifikovali variantu c.1096_1099ATTA u 3/887 pacientů a 5/976 kontrol z Německa a u 1/278 pacientů a 1/229 kontrol z Makedonie. Autoři tak zpochybňují klinický význam našich zjištění.

V naší odpovědi jsme k dříve publikovaným výsledkům rozšířené analýzy Lhota et al., 2016 (Kapitola 5.1), která identifikovala 15 nosičů c.1096_1099dupATTA varianty doplnili klinické a histopatologické charakteristiky nosičů mutací a identifikovali jsme dalších osm nosičů této varianty identifikovaných v rámci sekvenování panelem CZECANCA (viz Kapitola 5.5). Analýza klinicko-patologických dat ukázala, že na rozdíl od Pfeifer et al., kteří nalezli pozitivní rodinnou anamnézu s přítomností karcinomu prsu pouze u jednoho z 10 (10%) identifikovaných nosičů c.1096_1099dupATTA, v našem souboru byla tato varianta zachycena u 9/23 (39%) nosičů a přítomnost jakéhokoli nádoru jsme zaznamenali u 15/23 (65%) c.1096_1099dupATTA. Nápadná asociace s familiárním výskytem nádorových onemocnění u nosičů varianty tak naznačuje, že varianta může modifikovat riziko vzniku nádorového onemocnění u nosičů mutací, avšak ke zhodnocení tohoto účinku bude nezbytné provedení analýzy velmi rozsáhlých souborů pacientů a kontrol.

5.3 Článek 3: Identification and Functional Testing of ERCC2 Mutations in a Multi-national Cohort of Patients with Familial Breast- and Ovarian Cancer

Rump A, Benet-Pages A, Schubert S, Kuhlmann JD, Janavičius R, Macháčková E, Foretová L, Kleibl Z, Lhota F, **Zemankova P**, Betcheva-Krajcir E, Mackenroth L, Hackmann K, Lehmann J, Nissen A, DiDonato N, Opitz R, Thiele H, Kast K, Wimberger P, Holinski-Feder E, Emmert S, Schröck E, Klink B., ***PLoS Genet***. 2016;12(8):e1006248, (IF_{2016}=6.1).

Po publikování našeho prvního článku z panelového sekvenování (Článek 1), ve kterém jsme identifikovali dvě posunové mutace postihující gen *ERCC2*, jsme byli přizváni ke spolupráci na mezinárodní studii zaměřené na význam *ERCC2* mutací u pacientek s dědičným karcinomem prsu a ovarií. Gen *ERCC2* kóduje DNA helikázovou podjednotku (ERCC2/XPD) transkripčního faktoru IIH, která se podílí na nukleotidové excizní reparaci (NER) DNA. Dědičné bi-alelické mutace *ERCC2* genu se manifestují jako tři zcela rozdílná onemocnění: cerebro-okulo-facio-skeletální syndrom 2, fotosenzitivní trichothiodystrofie 1, nebo xeroderma pigmentózum D (XPD). Vznik XPD je spojen se zvýšeným výskytem kožních tumorů, avšak přítomnost monoalelických mutací *ERCC2* genu nebyla u pacientek s familiárním výskytem karcinomu prsu a ovaria studována.

Ve studii se podařilo analyzovat 1,345 pacientů (včetně 325 z naší laboratoře) a 2,400 kontrol (včetně 345 z naší laboratoře) z Německa, ČR a Litvy analyzovaných pomocí panelového NGS. U pacientů bylo identifikováno celkem pět trunkačních a raritních 20 missense variant, které kolegové z Německa podrobili funkčním in vitro analýzám.

Přesto, že trunkační a missense mutace klasifikované jako funkčně-defektní varianty byly uvedeny v databázi ExAC jako velmi raritní (s výjimkou jediné všechny se vyskytovaly s frekvencí <0,05%), analýzy populačně-porovnatelných kontrolních vzorků odhalily přítomnost i některých patogenních variant se značně prevalencí (např. frekvence posunové mutace p.F568fs* v naší populaci dosahovala 0,43% u pacientů, ale také 0,44% v kontrolách). Segregační analýza v dostupných rodinách nosičů mutací neprokázala jasnou asociaci fenotypu s přítomností patogenních variant *ERCC2* genu.

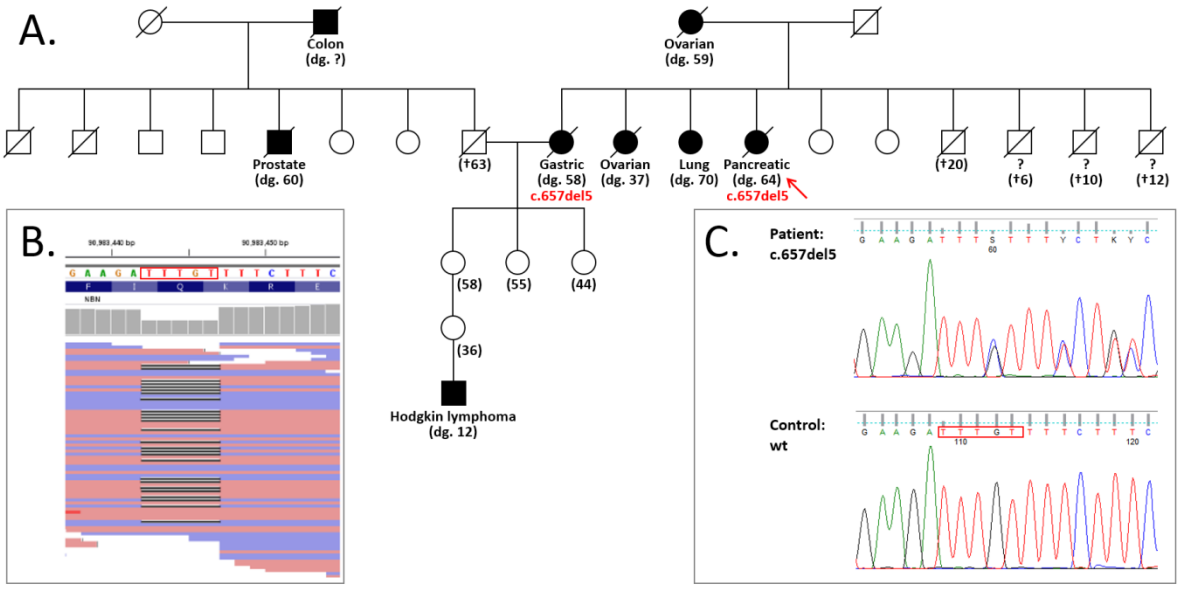
Výsledky studie tak nepotvrdily předpoklad, že nosičství patogenních mutací genu *ERCC2* je spojeno se zvýšeným rizikem vzniku karcinomu prsu a ovaria. Přinesly však cenné pozorování, které ukazuje, že analýza nádorové predispozice u genů s neúplnou penetrancí a s raritními dědičnými variantami bude vyžadovat pro porovnání analýzu dostatečně velkého souboru etnicky a geograficky srovnatelné kontrolní populace. Z důvodu možného „founder“ efektu se pro hodnocení frekvencí výskytu variant v těchto genech nelze spoléhat pouze na veřejné multietnické databáze.

5.4 Článek 4: The c.657del5 variant in NBN gene predisposes to pancreatic cancer

Borecka M, **Zemankova P**, Lhota F, Soukupova J, Kleiblova P, Vocka M, Soucek P, Ticha I, Kleibl Z, Janatova M., ***Gene***. 2016;587(2):169-72, (IF_{2016}=2.415).

Při přípravě dat k publikaci článku Lhota F. et al., 2016 (Článek 1) jsme provedli re-evaluaci klinických a histopatologických údajů analyzovaných pacientů, která ukázala, že 13 sekvenovaných vzorků pochází od pacientů s jinou diagnózou, než je karcinom prsu. Tyto vzorky byly z práce vyřazeny. Jedním z vyřazených byl vzorek pacientky s karcinomem pankreatu s mnohočetným nádorovým výskytem v rodině. Při analýze vzorku probandky byla zachycena mutace v genu *NBN* c.657del5, kterou jsme následně potvrdili i v histologickém bločku z nádoru u její sestry s karcinomem žaludku (Obr. 16).

Obr. 16. Rodokmen pacientky s karcinomem pankreatu (A) s mutací c.657del5 v genu NBN, identifikované pomocí NGS (B) a konfirmované Sangerovým sekvenování (C).



Gen *NBN* kóduje nibrin, součást MRN komplexu, katalyzujícího zahájení procesu homologní rekombinace při reparaci dvouřetězcových zlomů DNA v S a G2 fázi buněčného cyklu (Carney J. P. et al., 1998). Bialelické patogenní mutace v tomto genu způsobují vznik Nijmegen-breakage syndromu a heterozygotní mutace zvyšují riziko lymfoidních malignit a jiných nádorů (Varon R. et al., 1998), jako např. karcinomu prsu (Gorski B. et al., 2003), non-Hodgkinova lymfomu (Steffen J. et al., 2006) a karcinomu prostaty (Cybulski C. et al., 2013). Mutace c.657del5 v genu *NBN* byla popsána jako častá varianta v naší populaci (Varon R. et al., 1998).

Protože asociace přítomnosti této varianty s karcinomem pankreatu nebyla známa, provedli jsme genotypizaci c.657del5 varianty pomocí HRMA na souboru 241 neselektovaných pacientů s karcinomem pankreatu. Při této analýze jsme zachytili dalších pět nosičů mutace c.657del5 v souboru pacientů (2,07%). Výskyt u pacientů s karcinomem pankreatu byl signifikantně vyšší, než výskyt této varianty v kontrolním souboru (2/915; 0,23%), což naznačuje, že mutace v genu *NBN* mohou zvyšovat riziko vzniku karcinomu pankreatu (OR 9.7; 95% CI: 1,9 – 50,2; p=0,006) a reprezentují tak další predispoziční gen pro hereditární formu tohoto závažného onemocnění, které je šestým nejčastějším nádorem v České republice a tumorem s velmi nepříznivou prognózou (pětileté přežití je pouze 7%, s mediánem přežití 6 měsíců; Siegel R. L., et al., 2015).

Souběžně s publikací našich výsledků byla publikována studie Polských autorů (Lener M. R. et al 2016), která zaznamenala srovnatelnou frekvenci c.657del5 varianty u pacientů s karcinomem pankreatu v Polsku (8/383; 2.09%), která se rovněž lišila od frekvence v kontrolách (22/4,000; 0,55%; OR 3.80; p=0,002).

5.5 Článek 5: CZECANCA: CZEch CAncer paNel for Clinical. Application – návrh a příprava cíleného sekvenačního panelu pro identifikaci nádorové predispozice u rizikových osob v České Republice

Soukupová J, **Zemánková P**, Kleiblová P, Janatová M, Kleibl Z. Klin Onkol. 2016;29 Suppl 1:S46-54., (časopis bez IF).

Zkušenosti z přípravy sekvenačního panelu, sekvenování a bioinformatického zpracování popisovaném v Článku 1 (Lhota F. et al., 2016) jsme využili při návrhu panelu a testování, určenému ke klinickému využití v diagnostice nosičů patogenních mutací pro vznik dědičných nádorových onemocnění v ČR.

Cílové geny jsme identifikovali na základě důkladné literární rešerše, ze které vyplynul seznam cílových genů. Kromě genů s již stanovenou klinickou relevancí, jsme se rozhodli zahrnout i DNA reparační geny, u nichž je klinický význam zatím nejasný. Lze předpokládat, že identifikace genotypů s fenotypovými charakteristikami v průběhu užívání panelu, posune znalost genů s dosud neznámou funkcí, v klinickém významu. Panel CZECANCA umožňuje rychlejší a efektivnější detekci patogenních mutací v genech s klinickou utilitou. Panel byl navržen v online NimbleDesign softwaru (NimbleGen, Roche) za přísných podmínek, znemožňujících navrženým próbám nasednout více jak na tři cílové sekvence v genomu. Cílová oblast o velikosti přibližně 600 kb zahrnovala všechny popsané exony, intron-exonové přechody a pro vybrané geny i promotorové oblasti; příprava panelu ve verzi 1.0 s přísným požadavkem na unikátnost hybridizace cílové sekvence však neumožnila úplné pokrytí všech navržených genů (Obr. 17).

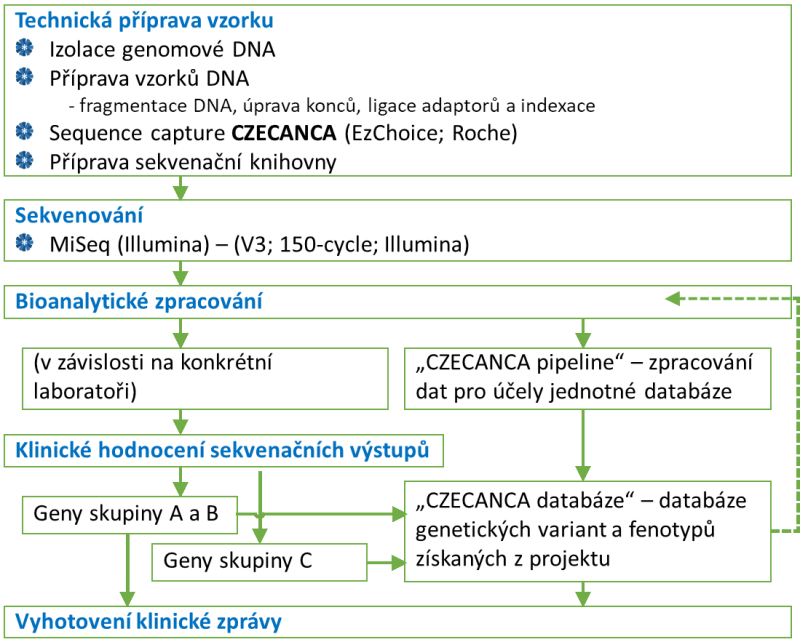
Obr. 17. Seznam genů zahrnutých do panelu CZECANCA.

AIP; ALK; APC; APEX1; **ATM**; ATMIN; ATR; ATRIP; AURKA; AXIN1; BABAM1; BAP1; **BARD1**; BLM; BMPR1A; BRAP; **BRCA1**; **BRCA2**; BRCC3; BRE; **BRIP1**; BUB1B; C11orf30; C19orf40; casp8; CCND1; CDC73; **CDH1**; CDK4; CDKN1B; CDKN1C; CDKN2A; CEBPA; CEP57; CLSPN; CSNK1D; CSNK1E; CWF19L2; CYLD; DCLRE1C; DDB2; DHFR; DICER1; ~~DIS3L2;~~ ~~DMBT1;~~ DMC1; DNAJC21; DPYD; EGFR; EPCAM; EPHX1; ERCC1; ERCC2; ERCC3; ERCC4; ERCC5; ERCC6; ESR1; ESR2; EXO1; EXT1; EXT2; EYA2; EZH2; FAM175A; FAM175B; FAN1; FANCA; FANCB; FANCC; FANCD2; FANCE; FANCF; FANCG; FANCI; FANCL; **FANCM**; FBXW7; FH; FLCN; GADD45A; GATA2; GPC3; GRB7; HELQ; HNF1A; HOXB13; HRAS; HUS1; CHEK1; **CHEK2**; KAT5; KCNJ5; KIT; LIG1; LIG3; LIG4; LMO1; LRIG1; MAX; MCPH1; MDC1; MDM2; MDM4; MEN1; MET; MGMT; MLH1; MLH3; MMP8; MPL; **MRE11A**; MSH2; MSH3; MSH5; MSH6; MSR1; MUS81; MUTYH; NAT1; **NBN**; NCAM1; NELFB; NF1; NF2; NFKBIZ; NHEJ1; NSD1; OGG1; **PALB2**; PARP1; PCNA; PHB; PHOX2B; PIK3CG; PLA2G2A; PMS1; ~~PMS2;~~ POLB; POLD1; POLE; PPM1D; PREX2; PRF1; PRKAR1A; PRKDC; **PTEN**; PTCH1; PTTG2; RAD1; RAD17; RAD18; RAD23B; **RAD50**; RAD51; RAD51AP1; RAD51B; **RAD51C**; **RAD51D**; RAD52; RAD54B; RAD54L; RAD9A; RB1; RBBP8; RECQL; RECQL4; RECQL5; RET; RFC1; RFC2; RFC4; RHBDF2; RNF146; RNF168; RNF8; RPA1; RUNX1; ~~SBDS;~~ ~~SDHA;~~ SDHAF2; SDHB; ~~SDHC;~~ ~~SDHD;~~ SETBP1; SETX; SHPRH; SLX4; SMAD4; SMARCA4; SMARCB1; SMARCE1; **STK11**; SUFU; TCL1A; TELO2; TERF2; TERT; TLR2; TLR4; TMEM127; TOPBP1; **TP53**; TP53BP1; TSC1; TSC2; TSHR; UBE2A; UBE2B; UBE2I; UBE2V2; UBE4B; UIMC1; VHL; WRN; WT1; XPA; XPC; XRCC1; **XRCC2**; XRCC3; XRCC4; XRCC5; XRCC6; ZNF350; ZNF365.

XXXX – kompletní gen (všechny kódující exony +10 bp do intronu)
XXXX – kompletní gen (všechny kódující exony +10 bp do intronu) + 5'UTR
XXXX – nekompletní gen (chybí některé exony z důvodu výskytu pseudogenů)
~~XXXX~~ – gen chybí (z důvodů pseudogenů, vysoce homologních míst, repetic apod.)

Od počátku jsme plánovali využití panelu i v dalších laboratořích v ČR, což by umožnilo analyzovat velké množství indikovaných osob a získat tak údaje o výskytu i raritních variant ve studovaných genech v naší populaci. Proto byl součástí projektu CZECANCA nejen sekvenační panel, ale zároveň optimalizovaný postup pro přípravu sekvenačních knihoven a jednotný postup pro bioinformatickou analýzu (Obr. 18) umožňující snazší interpretaci nalezených variant. Příprava knihovny CZECANCA panelu byla optimalizována na platformě Illumina MiSeq, jako nejrozšířenější NGS platformy současnosti.

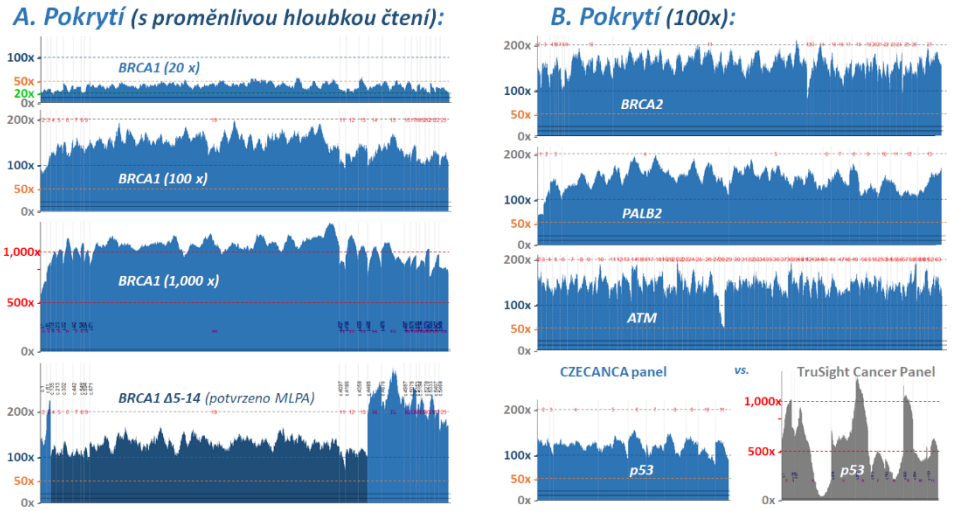
Bioinformatická analýza probíhá následujícími postupnými kroky. Mapování sekvencí zpracovává software Novoalign. Vygenerované SAM soubory, jsou převedeny do binární formy pomocí Picard tools, díky němuž se také odstraňují v těchto souborech duplikáty. Genotypizace variant se uskutečňuje pomocí nástroje GATK, následná anotace v ANNOVARu a ve frekvenčních databázích (ExAC, 1000g a ClinVar) a predikčních softwarech (např. SIFT, CADD, GERP, spidex, atd.). Pro detekci středně velkých inzercí a duplikací přesahujících polovinu délky sekvenačního čtení, jsme použili software Pindel. CNV analýza probíhá z BAM souborů v softwaru CNVkit.



Obr. 18. Projekt CZECANCA zahrnuje ucelený postup pro analýzu nádorové predispozice.

Analyzované geny byly roztříděny do skupiny A-C podle klesající klinické významnosti. Z analytického hlediska jsme velkou pozornost při přípravě panelu věnovali rovnoměrnosti pokrytí cílových oblastí. Tento požadavek umožnil racionální využití sekvenační kapacity a byl zásadním předpokladem pro analýzu CNV. Pro rutinní kontrolu pokrytí genů reportovaných v klinických zprávách jsme vyvinuli skript Boudalyzer (v prostředí R; http://www.R-project.org.) umožňující rychlou vizuální kontrolu celkového pokrytí hodnocených genů (Obr. 19).

Obr. 19. ***Zobrazení kódujících oblastí libovolných genů umožňuje vlastní skript Boudalyzer.*** *V panelu A ukazuje, že rovnoměrnost pokrytí při analýze BRCA1 není závislá na hloubce čtení a umožňuje dobré rozpoznání CNV (poslední panel vlevo). V panelu B ukazuje, že rovnoměrné pokrytí pří analýze cílené na pokrytí 100x je dosahováno s panelem CZECANCA i u jiných genů. V posledním obrázku je porovnáno pokrytí genu TP53 v porovnání s panelem TruSight.*



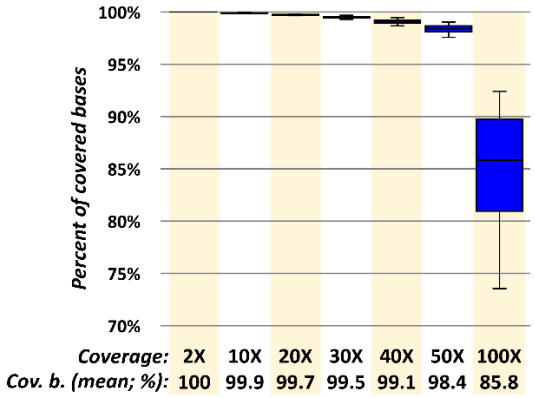
Koncept CZECANCA projektu umožnil vznik národního konsorcia, ke kterému se do současné doby připojilo devět laboratoří analyzujících vzorky pacientů s podezřením na nádorovou predispozici. Účastníci projektu společně sdílejí hrubá sekvenační data (prostřednictvím portálu BaseSpace), která zpracováváme v naší laboratoři jednotným postupem pro konstrukci frekvenční databáze. Genotypové údaje jsou účastníky doplněny o klinicko-patologické charakteristiky sekvenovaných pacientů s onkologickými onemocněními.

5.6 Článek 6: Validation of CZECANCA (Czech CAncer paNel for Clinical Application) for targeted NGS – base analysis of hereditary cancer syndromes

Soukupova J, **Zemankova P**, Lhotova K, Janatova M, Borecka M, Stolarova L, Lhota F, Foretova L, Machackova E, Stranecky V, Tavandzis S, Kleiblova P, Vocka M, Hartmannova H, Hodanova K, Kmoch S, Kleibl Z. ***PLoS One***;13(4):e0195761, (IF_{2018}=2.766).

Před zavedením sekvenačního panelu CZECANCA do rutinní klinické diagnostiky hereditárních nádorových syndromů jsme provedli důkladné testování parametrů panelu.

Testování ukázalo, že v rutinních podmínkách se sekvenováním cíleným na průměrné pokrytí 100x dosahuje sekvenování panelem CZECANCA 100x pokrytí v 85,8% a pokrytí 50x v 98,4% cílové oblasti (Obr. 20). Pouze 0,3% cílové oblasti je pokryto <20x, což je hodnota spolehlivosti při analýze heterozygotních germinálních variant.



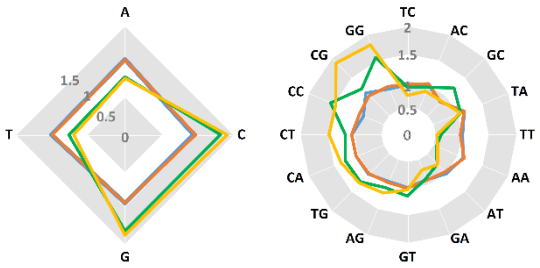
Obr. 20. Průměrné pokrytí při rutinním NGS cílícím na pokrytí 100x.

Uniformita pokrytí v datech je velmi vyhovující a umožňuje snadnější a spolehlivější detekci CNV variant. Uniformita panelu vzhledem k reprodukovatelnosti výsledků, byla ověřena několika způsoby. Do jednoho runu jsme zahrnuli DNA z jednoho vzorku, která byla zpracována v odlišných koncentracích (33 ng – 100%, 24.75 ng – 75% a 16.5 ng – 50%). Výsledky ukazují, že 289/293 (98,6%) variant bylo nalezeno ve všech třech replikátech.

Další analýza funkčnosti panelu proběhla testováním stejné DNA v rozdílných sekvenačních analýzách. Z 356 unikátních variant se jich v obou nezávislých sekvenačních analýzách nacházelo 354 (99,4%). Dodatečná analýza ukazuje na uniformitu zpracování ve čtyřech laboratořích, kde je CZECANCA rutinně používána. Bylo detekováno 332 unikátních variant. Jeden vzorek v rámci

čtyřech pracovišť tedy vykazuje shodu 331/332 (99,7%), 327/332 (98.5%), 329/332 (99.1%) a 329/332 (99.1%).

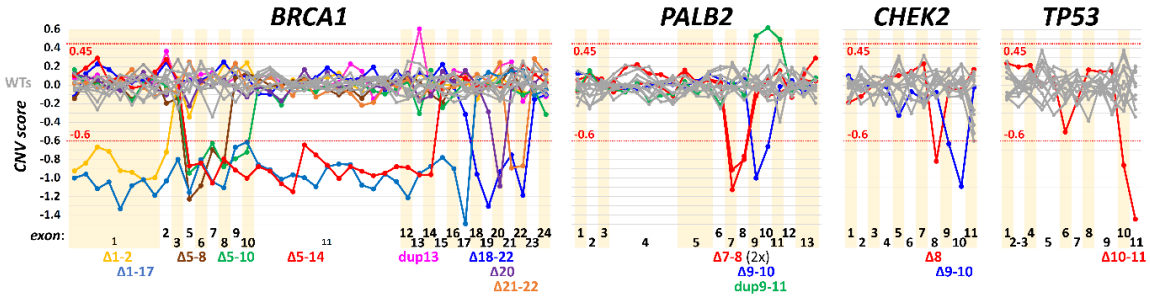
Mírné rozdíly mezi jednotlivými laboratořemi lze částečně přisoudit odlišnému způsobu fragmentace genomové DNA. Analýzou terminálních nukleotidů a dinukleotidů sekvenovaných fragmentů DNA po štěpení pomocí fragmentázy a ultrazvukové fragmentaci jsme prokázali, že ultrazvukovým štěpení lze dosílit rovnoměrnější fragmentace DNA (Obr. 21).



Obr. 21. ***Normalizovaný poměr výskytu terminálních nukleotidů a dinukleotidů sekvenovaných fragmentů štěpených ultrazvukem (červená a modrá) a enzymaticky (žlutá a zelená)*** *provedená ve čtyřech nezávislých laboratořích. Při zcela náhodném štěpení by poměr terminálních nukleotidů a dinukleotidů měl být identický (blízký 1). Tomu se blíží fragmentace pomocí ultrazvuku, zatímco enzymatické štěpení preferenčně lépe fragmentuje CG-bohaté oblasti.*

CNV analýzu jsme optimalizovali na panelu CZECANCA analýzou více než 300 vzorků souběžně analyzovaných pomocí MLPA. Po úpravě protokolu k dosažení uniformního a dostatečného pokrytí, což byl základ pro tuto analýzu, jsme dospěli k softwaru CNVkit (Talevich E., 2016), fungujícího na bázi Linuxu. Díky pozitivním kontrolám, jsme provedli dodatečnou analýzu pomocí vlastního skriptu v prostředí R. Teprve po sloučení obou postupů jsme dosáhli 100% záchytu všech sekvenovaných pozitivních kontrol s velkými přestavbami souběžně analyzovanými pomocí MLPA (Obr. 22).

Obr. 22. ***Výsledek kalibrace CNV analýzy na našem pracovišti.*** *Hodnota <-0.6 znamená přítomnost delece, hodnota > 0.5 znamená přítomnost duplikace. Panel BRCA1 obsahuje devět pozitivních kontrol – osm delecí a jednu duplikaci, panel PALB2 obsahuje čtyři pozitivní kontroly – tři delece a jednu duplikaci, panel CHEK2 obsahuje dvě delece a panel TP53 jednu deleci.*



Kvalitativní parametry a výsledky interních kontrol kvality detekce SNV a CNV jsme doplnili o nezávislé hodnocení European Molecular Genetics Quality Network (www.emqn.org), kdy jsme dosáhli dle externího hodnocení 100% senzitivity v detekci variant. Další kontrola kvality zahrnovala sekvenování pěti vzorků od Coriell Institute for Medical Research, pro které jsou dostupné vysoce kvalitní a verifikované BAM a VCF soubory umožňující vyčíslení falešné pozitivity/negativity.

Porovnání sekvenování v oblasti sekvenačního cíle CZECANCA panelu u vzorků Coriell Institute uvedeny v tabulce 2.

***Tab. 2.** Výsledky analýzy kontroly kvality na pěti vzorcích referenčních DNA (NA#) ze společnosti Coriell Institute for Medical Research.*

Reference standard no.	NA12878	NA24143	NA24149	NA24385	NA24631
True positive (TP)	355	341	332	351	348
True negative (TN)	627,672	627,674	627,678	627,658	627,671
False positive (FP)	42	49	56	57	48
False negative (FN)	0	5	3	3	2
Total	628,069	628,069	628,069	628,069	628,069
Sensitivity [TP/(TP+FN)]	100.000%	98.555%	99.104%	99.153%	99.429%
Specificity [TN/(TN+FP)]	99.993%	99.992%	99.991%	99.991%	99.992%
Accuracy [(TP+TN)/Total]	99.993%	99.991%	99.991%	99.990%	99.992%

Výsledky validace sekvenačního panelu CZECANCA prokázaly, že analýza pomocí tohoto systému je funkční a vyhovující pro současnou oblast diagnostiky v detekci variant s vysokou specificitou, senzitivitou a robustností z hlediska klinické praxe. Umožňuje analyzovat a spolehlivě detekovat nejen SNV a malé inzerce/delece, ale i CNV u většiny z vyšetřovaných genů. Analýzy rovněž prokázaly, že bioinformatické zpracování je plně srovnatelné se zavedenými standardy a bioinformatický postup analýz všech vzorků členů CZECANCA konsorcia povede ke konstrukci hodnotné frekvenční databáze germinálních variant u vysoce rizikových osob.

5.7 Článek 7: Identification of deleterious germline CHEK2 mutations and their association with breast and ovarian cancer.

Kleiblova P, Stolarova L, Krizova K, Lhota F, Hojny J, **Zemankova P**, Havranek O, Vocka M, Cerna M, Lhotova K, Borecka M, Janatova M, Soukupova J, Sevcik J, Zimovjanova M, Kotlas J, Panczak A, Vesela K, Cervenkova J, Schneiderova M, Burocziova M, Burdova K, Stranecky V, Foretova L, Machackova E, Tavandzis S, Kmoch S, Macurek L, Kleibl Z. **Int J Cancer**. 2019 May 3, doi: 10.1002/ijc.32385. (IF_{2017}=7.360).

Frekvenční databáze CZECANCA obsahuje v současnosti genotypová data od více než 6000 pacientů. Mezi nimi výrazně převažují pacientky s diagnózou karcinomu prsu a ovarií. Mutace u těchto pacientek dominantně postihují v geny *BRCA1* a *BRCA2*, třetím nejčastěji mutovaným genem je *CHEK2*.

Gen *CHEK2* kóduje kinázu CHK2 aktivovanou proteinem ATM v odpovědi na přítomnost dvouřetězcových zlomů v DNA, která následně fosforyluje řadu proteinů ovlivňujících DNA reparaci (např. BRCA1, KAP1) a zástavu buněčného cyklu (např. p53). Vrozené mutace v genu *CHEK2* byly asociovány se vznikem různých solidních i hematoonkologických nádorových onemocnění, včetně karcinomu prsu, kolorekta, štítné žlázy, prostaty, ledviny, nebo non-Hodgkinova lymfomu (Cybulski C. et al. 2004; Kleibl Z. et al. 2008; Havranek O. et al. 2015). Významné rozdíly v prevalenci alterací *CHEK2* genu v jednotlivých populacích vedou k diskrepantním odhadům rizika asociovaného s nosičstvím mutací *CHEK2* genu. Zatímco práce vyčíslující riziko u nosičů mutací *CHEK2* genu na základě porovnání s populačně-specifickými kontrolami uvádějí riziko na horní hranici genů se střední penetrancí (RR ~ 5), práce porovnávající výskyt u nosičů s výskytem alterací ve veřejných databázích docházejí k podstatně nižším rizikům na dolní hranici pásma pro geny střední penetrance (RR ~ 2). K nejasnému vztahu nosičství mutací v *CHEK2* genu ke skutečnému zvýšení rizika přispívá i nejasný význam naprosté většiny vzácných missense variant.

Cílem práce byla identifikace dědičných variant *CHEK2* genu v naší populaci u 1,928 vysoce rizikových pacientů analyzovaných na nádorovou predispozici ke karcinomu prsu a ovaria a u 3,360 nenádorových kontrol, funkční charakterizace významu nalezených missense variant, vyčíslení rizika vzniku karcinomu prsu a ovaria u nosičů trunkačních a funkčně-defektních missense variant a analýza klinických a histopatologických charakteristik nádorů u nosičů těchto alterací.

Analýza zahrnovala výsledky testování mutací *CHEK2* genu, které byly prováděny v období více než 10 let. Polovina vzorků však pocházela z dat zpracovaných panelovým NGS, které umožnilo i spolehlivou identifikaci dvou rozsáhlých delecí *CHEK2* genu, které tvořily třetinu všech nalezených trunkačních mutací. Protože v původním návrhu panelu CZECANCA nebyly z důvodů velkého

výskytu pseudogenů zahrnuty exony 12-15, byly tyto oblasti analyzovány dodatečně u všech vzorků analyzovaných pomocí NGS. (Poznámka: Zmiňované exony a nepokryté geny a oblasti dalších genů, viz Obr. 17, jsou v panelu CZECANCA zahrnuty od jeho verze 1.1, pokrývající všech 226 navržených genů).

V analyzovaném souboru jsme identifikovali 10 různých trunkačních mutací a 26 missense variant. Zatímco frekvence trunkačních mutací se významně lišila mezi souborem pacientů (2,39% nosičů) a kontrol (0.33% nosičů; p=1,1×10^{-14}), celková frekvence missense alterací byla srovnatelná (4,56% vs. 3,90%; p=0,42). Pod vedením Dr. Macůrka z ÚMG AVČR, bylo vyvinuto funkční vyšetření variant *CHEK2* genu, které umožnilo kvantifikovat enzymovou aktivitu CHK2 kinázy v lidských netransformovaných buňkách. Tato analýza umožnila funkčně klasifikovat 11 VUS jako varianty s podstatnou poruchou kinázové aktivity, pět variant s částečnou poruchou kinázové aktivity, a 10 variant se zachovanou kinázovou aktivitou.

Na základě funkčních analýz jsme byli schopni vyčíslit rizika spojená s nosičstvím germinálních mutací v *CHEK2* genu u pacientek s karcinomem prsu a ovaria, které ukazují, že riziko je klinicky významné pro trunkační mutace a vznik karcinomu prsu u žen, ale i u mužů, asociace s karcinomem ovaria je méně významná. Funkčně defektní missense varianty jsou spojeny pravděpodobně s nižším rizikem než trunkační mutace.

Práce dokumentuje zásadní význam v detekci CNV pomocí panelového NGS. Analýza funkčního významu umožní klasifikovat všechny nalezené VUS popsané CZECANCA konsorciem, což zvýší klinickou výpovědní hodnotu analýz u vysoce rizikových pacientů.

5.8 Článek 8: Multiplex PCR and NGS-based identification of mRNA splicing variants: Analysis of BRCA1 splicing pattern as a model

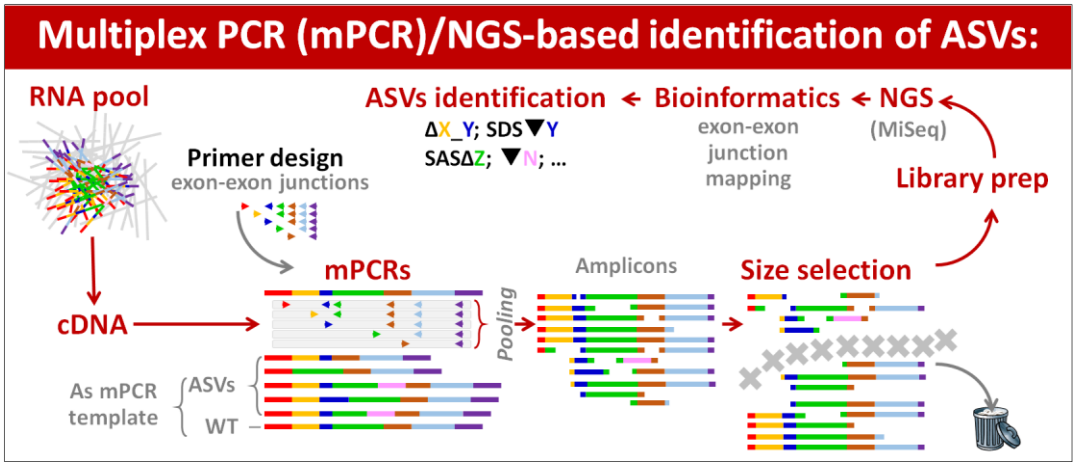
Hojny J, **Zemankova P**, Lhota F, Sevcik J, Stranecky V, Hartmannova H, Hodanova K, Mestak O, Pavlista D, Janatova M, Soukupova J, Vocka M, Kleibl Z, Kleiblova P. ***Gene***. 2017; 30(637):41-49, (IF_{2017}=2.498).

V naší laboratoři se dlouhodobě zabýváme analýzou genů predisponujících ke vzniku dědičných nádorových syndromů. Dlouhodobým zájmem jsou i analýzy alternativních a aberantních sestřihových variant. Fyziologicky se v buňkách vyskytují alternativně sestřižené formy proteinových produktů. V odlišných tkáních je zastoupeno různorodé procento konkrétní transkripční varianty. Varianty v DNA mohou ovlivňovat alternativní sestřih a vytvářet sestřih aberantní, který může mít značný vliv na funkci proteinu. Detekovat vliv variant na sestřih primárního transkriptu je tudíž zásadní a vyžaduje znalost fyziologicky se vyskytujících variant alternativního sestřihu.

Současné metody zahrnují různé strategie cílené amplifikace alternativních produktů, které však neumožňují kvantitativní hodnocení alternativních sestřihových forem. Výsledkem pokroku a rozvoje NGS je sekvenování RNA (RNAseq), avšak oblastí našeho zájmu jsou geny s velmi nízkou genovou expresí na úrovni ~1 TPM, a proto analýza jejich, často minoritních, alternativních sestřihových variant pomocí RNAseq je ekonomicky nevýhodná.

V naší práci jsme navrhli a na analýze sestřihových variant BRCA1 mRNA validovali nový systém pro identifikaci alternativních sestřihových variant (Obr. 23).

Obr. 23. Schéma analýzy alternativních sestřihových variant pomocí multiplexního PCR amplifikujících všechny exon-exonová spojení z cDNA a následného NGS velikostně nabohacených alternativně sestřižených (krátkých) izoforem. *Nabohacení probíhá pomocí velikostní selekce, po které krátké fragmenty slouží pro přípravu sekvenační knihovny.*



Navržením primerů na všechna exon-exonová spojení jsme docílili amplifikace pokrývající všechny možné varianty alternativního sestřihu BRCA1, který jsme použili jako modelový příklad. Vzorky RNA (resp. cDNA) jsme analyzovali od 32 jednotlivců – z toho 16 nenádorových kontrol, osmi pacientů s karcinomem prsu bez přítomnosti mutace v genu *BRCA1* a osmi pacientů s karcinomem prsu s *BRCA1* mutací. RNA byla izolována z leukocytů perifení krve, perimamární tukové tkáně a mamární tkáně. Paralelně jsme analyzovali RNA ze stabilních buněčných linií (MCF7, EM-G3, HeLa a MDA-MB-231).

Pro analýzu dat bylo nezbytné vyvinout vlastní postup. Data ze sekvenátoru jsme mapovali pomocí softwaru Novoalign dvojím procesem. V prvním kroku sloužil jako referenční genom fasta soubor, který jsme zkonstruovali ze sekvencí všech kombinatorických možností exon-exon spojení. V druhém kroku byly data namapovány na DNA sekvenci daného genu, abychom zachytili exonizované introny. Před mapováním vzhledem k nejednotné délce PCR produktů, bylo nutné provést trimming adaptorů a nekvalitních bází na konci všech čtení pomocí softwaru Trimmomatic. Konverze souborů ze SAM formátu do BAM byla provedena softwarem Picard tools. Statistiky o pokrytí jednotlivých oblastí byly spočítány díky softwaru SAMtools a soft-clipové báze ručně zhodnoceny ve vizualizačním softwaru IGV.

Takto navrženým systémem se nám podařilo detekovat celkem 94 variant, tedy více alternativních sestřihových variant než v dříve popsané studii (Colombo M. et al., 2014). Největší množství (72 variant) bylo detekováno v mamární tkáni, dále v leukocytech (67) a perimamární tukové tkáni (57). Ze 76 variant detekovaných v buněčných liniích jich 11 nebylo zachyceno v žádném jiném vzorku.

Postup analýzy je možné aplikovat na vyšetření libovolných transkriptů, včetně velmi málo exprimovaných genů, jako je *BRCA1*. Na rozdíl od RNAseq umožňuje naše analýza přibližně 1000x vyšší detekci kanonických exon-exonových spojení a tím i značně spolehlivější kvantifikaci i minoritních sestřihových variant (tvořících <1% z celkového sestřihu) za nesrovnatelně nižších nákladů, protože analýza BRCA1 transkriptů všech analyzovaných vzorků spotřebovávala v našem případě méně než 10% sekvenační kapacity při sekvenování na MiSeq.

6 SHRNUTÍ A ZÁVĚR

Nástup sekvenování nové generace v poslední dekádě znamenal revoluční posun v diagnostice genetických onemocnění. NGS se v současnosti stalo rutinní metodou používanou v diagnostických laboratořích. Největší výzva v této oblasti je sestavení bioinformatického pracovního postupu pro analýzu, ale zejména interpretaci nalezených variant.

Základní součástí bioinformatické analýzy je mapování hrubých dat na referenční genom nebo vlastní konsensus. Zde dochází již k prvnímu zkreslení a nedokonalosti dat, které pak může ovlivnit výsledky. Druhým krokem je z namapovaných souborů získat jednotlivé varianty, velké přestavby a střední inzerce a duplikace. Následuje filtrace variant v rámci dostupných databází a predikčních softwarů. Výsledné varianty s pravděpodobně patogenním dopadem na protein jsou reportovány diagnostikem. Všechny výše jmenované kroky nesou riziko zkreslení, které může negativně ovlivnit interpretaci výsledků. Proto je nutné bioinformatické zpracování důkladně validovat a dle nálezů referenčních vzorků pak upravit příslušné parametry nebo označit kritické oblasti.

Postup NGS analýz, který vyústil do projektu CZECANCA, byl na počátku testován na více než 500 vzorcích se známými alteracemi se 100% záchytem daných mutací. Nejproblémovější bylo nastavení CNV analýz a středně velkých inzercí a duplikací. Panel CZECANCA je vzhledem k uniformitě pokrytí spolehlivě zpracovatelný a oproti jiným komerčním panelům umožňuje analýzu SNV, indelů i CNV.

Panelové sekvenování je v současnosti nejefektivnější formou diagnostiky dle potřeby konkrétního pracoviště. S jeho nástupem se rapidně mění oblast diagnostiky příčiny hereditárních karcinomů. Jedná se o efektivní analýzu, která umožňuje diagnostiku pacientů s různými diagnózami, pro které je dostupná řada cílených panelů (Tab. 3).

Panely umožňují výběr genů analyzujících predispozici ke konkrétním nádorovým diagnózám, avšak u pacientů s nejasným fenotypem (např. mnohočetným výskytem rozdílných nádorových diagnóz v rodině probanda) je výběr konkrétního panelu je obtížný. Studie u pacientů s kolorektálním karcinomem ukázala, že třetina pozitivních pacientů fenotypem neodpovídala nálezům (Pearlman R. et al., 2017). Analýza 528 pozitivních pacientů s dědičnými mutacemi v MMR genech identifikovala karcinom prsu u 11,9% nosičů, avšak současná klinická doporučení pro nosiče mutací MMR genů nezohledňují preventivní opatření snižující riziko vzniku karcinomu prsu (Espenschied C. R. et al., 2017). Z těchto důvodů jsme se při návrhu panelu CZECANCA zaměřili na širokou oblast predispozičních genů všech častých nádorových onemocnění.

Tab. 3: Příklady panelů genů používaných pro analýzu nádorové predispozice (upraveno podle Soukupová J. 2016).

Název panelu	Cílení genů	Poskytovatel
Breast/Ovarian Cancer Panel	20 genů asociovaných s predispozicí ke karcinomu prsu	GeneDx
BROCA – Cancer Risk Panel	23 genů asociovaných s predispozicí ke karcinomu prsu	University of Washington
CancerNext	32 genů asociovaných s nádorovou predispozicí	Ambry Genetics
CancerNext-expanded	49 genů asociovaných s nádorovou predispozicí	Ambry Genetics
ClearSeq Comprehensive Cancer	151 genů asociovaných s nádorovou predispozicí	Agilent
ColoNext	17 genů asociovaných s predispozicí ke karcinomu střeva a konečníku	Ambry Genetics
Comprehensive Cancer Panel	32 genů asociovaných s nádorovou predispozicí	GeneDx
CZECANCA	226 genů asociovaných s nádorovou predispozicí/DNA reparačních genů	1. LF UK
Invitae Common Hereditary Cancers Panel (Breast, Gyn, GI)	42 genů asociovaných s predispozicí ke karcinomu prsu, vaječníků, dělohy, tlustého střeva a konečníku, žaludku, slinivky	Invitae
Invitae Multi-Cancer Panel	80 genů asociovaných s nádorovou predispozicí	Invitae
Invitae Pediatric Solid Tumors Panel	48 genů asociovaných s predispozicí k solidním nádorům dětského věku	Invitae
Myriad myRisk Hereditary Cancer test	28 genů asociovaných s nádorovou predispozicí	Myriad Genetics
OvaNext	24 genů asociovaných s predispozicí ke karcinomu prsu, vaječníků a/nebo dělohy	Ambry Genetics
Pediatric Tumor Panel	27 genů asociovaných s nádory dětského věku	GeneDx
TruSight Cancer	94 genů a 284 SNP asociovaných s nádorovou predispozicí	Illumina

Sekvenace exomů nebo genomů by v nejasných případech byla ideálním postupem, avšak není, s ohledem na počet vyšetření nádorové predispozice, ekonomicky reálná v současném rutinním provozu. Analýzy nadto komplikuje neúplná znalost funkčního dopadu nalezených variant.

Nejasné nálezy jsou také součástí panelového sekvenování. Díky spolupráci jednotlivých center, analýze nenádorové populace, používání jednotného panelu a způsobu zpracování, je možné vytvořit databázi, která usnadňuje interpretaci problematických variant a vytvořit následná kritéria péče pro nosiče těchto mutací. Spolupráce s laboratořemi preklinického významu je důležitým předpokladem pro časově náročné funkční testování prioritizovaných variant.

Obtížná je také intepretace tzv. diskordantních nálezů tj. přítomnost patogenní mutací v genech, které se zatížením rodiny hereditárními syndromy by teoreticky vůbec neměly souviset. Takové nálezy pak vrhají jiné světlo na problematiku spojitosti konkrétní oblasti DNA reparačních genů s konkrétním nádorovým syndromem.

V naší laboratoři jsme prováděli vyšetření klasickým způsobem do zavedení metody NGS. S mým nástupem a optimalizací zpracování dat jsme byli schopni zpracovat panel 581 genů a vytvořit nový panel CZECANCA, který je součástí diagnostiky již v několika centrech v České republice. Díky panelovému sekvenování jsme byli schopni zpětně detekovat mutace u pacientek, které byli vyšetřeny klasickými postupy.

7 LITERATURA

Adzhubei I.A., Schmidt S., Peshkin L. et al., 2010, A method and server for predicting damaging missense mutations. Nat Methods.; 7(4):248-9.

Andrews S., 2010, FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ali A.M., Kirby M., Jansen M. et al., 2009, Identification and characterization of mutations in FANCL gene: a second case of Fanconi anemia belonging to FA-L complementation group, Hum Mutat. Jul;30(7):E761-70.

Arı S. a Arıkan M., 2016, Next Generation Sequencing: Advantages, Disadvantages and Future, In "Plant Omics-Trends and Applications" Springer International Publishing, pp: 109-136.

Arnold M., Karim-Kos H.E., Coebergh J.W. et al., 2015, Recent trends in incidence of five common cancers in 26 European countries since 1988: Analysis of the European Cancer Observatory, Eur J Cancer.;51(9):1164-87.

Ballester L.Y., Luthra R., Kanagal-Shamanna R. et al., 2016, Advances in clinical next-generation sequencing: target enrichment and sequencing technologies, Expert Rev Mol Diagn.;16(3):357-72.

Brouwer R.W.W., van den Hout M.C.G.N., Kockx C.E.M. et al., 2018, Nimbus: a design-driven analyses suite for amplicon-based NGS data, Bioinformatics.;34(16):2732-2739.

Bumgarner R., 2013, DNA microarrays: Types, Applications and their future, Curr Protoc Mol Biol.;Chapter 22:Unit 22.1.

Carney J.P., Maser R.S., Olivares H. et al., 1998, The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response, Cell. May 1;93(3):477-86.

Che R., Zhang J., Nepal M. et al., 2017, Multifaceted Fanconi Anemia Signaling, Trends Genet;34(3): p.171-183.

Choi M., Bien H., Mofunanya A. et al., 2017, Challenges in Ras therapeutics in pancreatic cancer, Semin Cancer Biol, pii: S1044-579X(17)30235-3.

Chun S. a Fay J.C., 2009, Identification of deleterious mutations within three human genomes. Genome Res.; 19(9):1553-61.

Colombo M., Blok M.J., Whiley P. et al., 2014, Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium, Hum Mol Genet. Jul 15;23(14):3666-80.

Cooper G.M., Stone E.A., Asimenos G. et al., 2005, Distribution and intensity of constraint in mammalian genomic sequence. Genome Res.; 15(7):901-13.

Cybulski C., Gorski B., Huzarski T. et al., 2004, CHEK2 is a multiorgan cancer susceptibility gene. Am J Hum Genet;75:1131–5.

Cybulski C., Wokołorczyk D., Kluźniak W. et al., 2013, An inherited NBN mutation is associated with poor prognosis prostate cancer, Br J Cancer. Feb 5;108(2):461-8.

Das R. a Ghosh S.K.,2017, Genetic variants of the DNA repair genes from Exome Aggregation Consortium (EXAC) database: significance in cancer. DNA Repair (Amst).; 52:92-102.

Espenschied C.R., LaDuca H., Li S. et al., 2017, Multigene Panel Testing Provides a New Perspective on Lynch Syndrome, J Clin Oncol. Aug 1;35(22):2568-2575.

Fedurco M., Romieu A., Williams S., et al., 2006, BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34(3): p.e22.

Ferragina P. a Manzini G, 2000, Opportunistic data structures with applications, In proceedings of the 41st Annual Symposium on foundations of computer Science (FOCS 2000). IEEE computer society.

Flicek .P, Amode M., Barrell D. et al., 2012, Ensembl 2012. Nucleic Acids Res.;40:D84–D90.

Foulkes W.D., 2008, Inherited Susceptibility to Common Cancers, N Engl J Med 359(20): p.2143-53.

Fujita P., Rhead B., Zweig A. et al., 2011, The UCSC genome browser database: update 2011. Nucleic Acids Res.;39:D876–D882.

Genomes Project C., Abecasis G.R., Auton A. et al., 2012, An integrated map of genetic variation from 1,092 human genomes. Nature.; 491(7422):56-65.

Górski B., Debniak T., Masojć B. et al., 2003, Germline 657del5 mutation in the NBS1 gene in breast cancer patients, Int J Cancer. Sep 1;106(3):379-81.

Hanahan D. a Weinberg R.A, 2000, The Hallmarks of Cancer Review, Cell, Vol. 100: p.57-70.

Hansford S. a Huntsman D.G., 2014, Boveri at 100: Theodor Boveri and genetic predisposition to cancer, J Pathol.;234(2):142-5.

Havranek O., Kleiblova P., Hojny J. et al., 2015, Association of Germline CHEK2 gene variants with risk and prognosis of non-Hodgkin lymphoma. PLoS One;10:e0140819.

Hwang S., Kim E., Lee I. et al., 2015, Systematic comparison of variant calling pipelines using gold standard personal exome variants, Sci Rep.;5:17875.

Janatova M., Soukupova J., Stribrna J. et al., 2015, Mutation Analysis of the RAD51C and RAD51D Genes in High-Risk Ovarian Cancer Patients and Families from the Czech Republic. PLoS One. 10(6):e0127711.

Janatova M., Kleibl Z., Stribrna J. et al., 2013, The PALB2 gene is a strong candidate for clinical testing in BRCA1- and BRCA2-negative hereditary breast cancer, Cancer Epidemiol Biomarkers Prev. Dec;22(12):2323-32.

Jessmon P., Boulanger T., Zhou W. et al., 2017, Epidemiology and treatment patterns of epithelial ovarian cancer, Expert Rev Anticancer Ther.; 17(5):427-437.

Kircher M., Witten D.M., Jain P. et al., 2014, A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet.; 46(3):310-5.

Kleibl Z. a Kristensen V.N., 2016, Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management, Breast Aug;28:136-44.

Kleibl Z., Havranek O., Novotny J. et al., 2008, Analysis of CHEK2 FHA domain in Czech patients with sporadic breast cancer revealed distinct rare genetic alterations. Breast Cancer Res Treat; 112:159–64.

Kleibl Z., Novotny J., Bezdickova D. et al., 2005, The CHEK2 c.1100delC germline mutation rarely contributes to breast cancer development in the Czech Republic, Breast Cancer Res Treat. Mar;90(2):165-7.

Kleiblova P., Shaltiel I.A., Benada J. et al., 2013, Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint, J Cell Biol. May 13;201(4):511-21.

Knudson A.G., 1971, Mutation and Cancer: Statistical Study of Retinoblastoma, Proc. Nat. Acad. Sci. USA, Vol. 68, No. 4, pp. 820-823.

Knudson A.G.,2001, Two genetic hits (more or less) to cancer, Nat Rev Cancer. ;1(2):157-62.

Korlach J., MArks P.J., Cicero R.L. et al., 2008, Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci* U S A;105(4):1176-81.

Kozarewa I., Armisen J., Gardner A.F. et al., 2015, Overview of Target Enrichment Strategies, Curr Protoc Mol Biol. ;112:7.21.1-23.

Krush A.J., 1979, Contributions of Pierre Paul Broca to Cancer Genetics, Transactions of the Nebraska Academy of Sciences- Volume VII, 1979.

Kumar P., Henikoff S., Ng P.C., 2009, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc.; 4(7):1073-81.

Lai Z., Markovets A., Ahdesmaki M. et al., 2016, VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research, Nucleic Acids Res.;44(11):e108.

Landrum M.J., Lee J.M., Riley G.R. et al., 2014, ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Res. ;42(Database issue):D980-5.

Langmead B., Trapnell C., Pop M. et al., 2009, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol10:R25.

Lener M.R., Scott R.J., Kluźniak W. et al., 2016, Do founder mutations characteristic of some cancer sites also predispose to pancreatic cancer?, Int J Cancer. Aug 1;139(3):601-6.

Levene M.J., Korlach J., Turner S.W., et al., 2003, Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations, Science;299(5607): p.682-6.

Li H. a Durbin R., 2009, Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Li H., 2011, A statistical Framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics.;27(21):2987-93.

Li R., Yu C., Li Y. et al., 2009, SOAP2: an improved ultrafast tool for short read alignment, Bioinformatics. 2009 Aug 1;25(15):1966-7.

Liu X., Jian X., Boerwinkle E., 2011. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human Mutation. 32:894-899.

Loveday C., Turnbull C., Ramsay E. et al., 2011, Germline mutations in RAD51D confer susceptibility to ovarian cancer, Nat Genet. ;43(9):879-882.

Lynch H.T., Snyder C.L., Shaw T.G. et al., 2015, Milestones of Lynch syndrome: 1895-2015.

Mateju M., Kleiblova P., Kleibl Z. et al., 2012, Germline mutations 657del5 and 643C>T (R215W) in NBN are not likely to be associated with increased risk of breast cancer in Czech women, Breast Cancer Res Treat. Jun;133(2):809-11.

Mateju M., Stribrna J., Zikan M. et al., 2010, Population-based study of BRCA1/2 mutations: family history based criteria identify minority of mutation carriers, Neoplasma;57(3):280-5.

McCarthy D.J., Humburg P., Kanapin A. et al., 2014, Choice of transcripts and software has a large effect on variant annotation, Genome Med. 2014 Mar 31;6(3):26.

McKenna A., Hanna M., Banks E. et al., 2010The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data , GENOME RESEARCH 20:1297-303.

McLaren W., Pritchard B., Rios D. et al., 2010, Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 26:2069–2070.

Michailidou K., Lindström S., Dennis J. et al., 2017, Association analysis identifes 65 new breast cancer risk loci, Nature Nov 2; 551(7678):92-94.

Miki Y., Swensen J., Shattuck-Eidens D. et al., 1994, A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1, Science;266(5182):66-71.

Nalepa G. a Clapp D.W., 2018, Fanconi anaemia and cancer: an intricate relationship, Nat Rev Cancer. 2018 Mar;18(3):168-185.

Nielsen F.C., van Overeem Hansen T., Sørensen C.S., 2016, Hereditary breast and ovarian cancer: new genes in confined pathways, Nat Rev Cancer.;16(9):599-612.

Palles C., Cazier J.B., Howarth K.M .et al., 2013, Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas, Nat Genet.;45(2):136-44.

Pearlman R., Frankel W.L., Swanson B. et al., 2017, Prevalence and Spectrum of Germline Cancer Susceptibility Gene Mutations Among Patients With Early-Onset Colorectal Cancer, JAMA Oncol. Apr 1;3(4):464-471.

Pelttari L.M., Heikkinen T., Thompson D. et al., 2011, RAD51C is a susceptibility gene for ovarian cancer, Hum Mol Genet. ;20(16):3278-88.

Pfeifer K., Schürmann P., Bogdanova N. et al., 2016, Frameshift variant FANCL*c.1096_1099dupATTA is not associated with high breast cancer risk, Clin Genet. Oct;90(4):385-6.

Pohlreich P, Zikan M, Stribrna J, et al., 2005, High proportion of recurrent germline mutations in the BRCA1 gene in breast and ovarian cancer patients from the Prague area. *Breast cancer research : BCR.*;7(5):R728-736.

Pollard K.S., Hubisz M.J., Rosenbloom K.R. et al., 2010, Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res.; 20(1):110-21.

Pruitt K., Tatusova T., Brown G. et al., 2012,. NCBI Reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res.;40:D130–D135.

Rahman N., 2014A, Mainstreaming genetic testing of cancer predisposition genes, Clin Med (Lond). 14(4): p.436-9.

Rahman N., 2014B, Realizing the promise of cancer predisposition genes. Nature;505(7483):302-8.

Rahman N., Seal S., Thompson D. et al., 2007, PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene, Nat Genet.;39(2):165-7.

Robinson J.T., Thorvaldsdóttir H., Winckler W.et al., 2011, Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 .

Rusk N., 2014, Nanopores read long genomic DNA, Nat Methods;11(9):887.

Sandmann S., de Graaf A.O., Karimi M. et al., 2017, Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data, Sci Rep.;7:43169.

Sanger F., Nicklen S. a Coulson A.R., 1977, DNA sequencing with chain-terminating inhibitors, Proc. Nati. Acad. Sci. USA, Vol. 74, No. 12, pp. 5463-5467.

Schwarz J.M., Rodelsperger C., Schuelke M. et al.,2010, MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods.; 7(8):575-6.

Shearer A.E., Hildebrand M.S., Ravi H. et al., 2012, Pre-capture multiplexing improves efficiency and cost-effectiveness of targeted genomicenrichment, BMC Genomics. ;13:618.

Shendure J, Balasubramanian S, Church GM et al., 2017, DNA sequencing at 40: past, present and future, Nature.;550(7676): p.345-353.

Shendure J. a Ji H., 2008, Next-generation DNA sequencing, Nat Biotechnol;26(10): p.1135-45.

Shendure J., Porreca G. J., Reppas N. B., et al., 2005, Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 309(5741): p. 1728-1732.

Siegel R.L., Miller K.D., Jernal A., 2015 Multiplex single-tube screening for mutations in the Nijmegen breakage syndrome (NBS1) gene in Hodgkin's and non- Hodgkin's lymphoma patients of Slavic origin, Eur. J. Hum. Genet. 11, 416-419.

Sims D., Sudbery I., Ilott N.E., et al., 2014, Sequencing depth and coverage: key considerations in genomic analyses., Nat Rev Genet. Feb;15(2):121-32.

Soukupova J., 2016, Úskalí interpretace sekvenačních dat v diagnostice dědičných nádorových syndromů, Labor Aktuell 04/16, 23-26.

Soukupova J., Dundr P., Kleibl Z. et al., 2008, Contribution of mutations in ATM to breast cancer development in the Czech population, Oncol Rep. Jun;19(6):1505-10.

Steffen J., Maneva G., Popławska L. et al., 2006, Increased risk of gastrointestinal lymphoma in carriers of the 657del5 NBS1 gene mutation, Int J Cancer. Dec 15;119(12):2970-3.

Sud A., Kinnersley B. a Houlston R.S., 2017, Genome-wide association studies of cancer: current insights and future perspectives, Nat Rev Cancer. ;17(11): p.692-704.

Talevich E., Shain A.H., Botton T. et al., 2016, CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing, PLoS Comput Biol. 2016; 12(4):e1004873.

Thankaswamy-Kosalai S., Sen P., Nookaew I., 2017, Evaluation and assessment of read-mapping bymultiple next-generation sequencing aligners based on genome-wide characteristics, Genomics;109(3-4):186-191.

Ticha I., Kleibl Z., Stribrna J. et al., 2010, Screening for genomic rearrangements in BRCA1 and BRCA2 genes in Czech high-risk breast/ovarian cancer patients: high proportion of population specific alterations in BRCA1 gene, Breast Cancer Res Treat. Nov;124(2):337-47.

Varon R., Vissinga C., Platzer M. et al., 1998, Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome, Cell. May 1;93(3):467-76.

Voelkerding K.V., Dames S.A., Durtschi J.D., 2009, Next-Generation Sequencing: From Basic Research to Diagnostics, Clin Chem. ;55(4):641-58.

Wang K., Li M., Hakonarson H., 2010, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38:e164.

Wooster R1, Bignell G, Lancaster J et al., 1995, Identification of the breast cancer susceptibility gene BRCA2, Nature;378(6559):789-92.

Xu L., Hou Y., Bickhart D.M. et al., 2013, Comparative Analysis of CNV Calling Algorithms: Literature Survey and a Case Study Using Bovine High-Density SNP Data, Microarrays (Basel).;2(3):171-85.

Ye H., Meehan J., Tong W. et al., 2015, Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine, Pharmaceutics, 7, p. 523-541.

Yu X., Guda K., Willis J., et al., 2012, How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?, BioData Min.;5(1):6.

Yu W., Clyne M., Khoury M.J. et al., 2010 Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics. Jan 1;26(1):145-6.

Zavoral M., Suchanek S., Majek O. et al., 2014, Colorectal cancer screening: 20 years of development and recent progress, World J Gastroenterol.;20(14):3825-34.

Zhao M., Wang Q., Wang Q. et al., 2013, Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives, BMC Bioinformatics.;14 Suppl 11:S1.