Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Biologie

Studijní obor: Molekulární biologie, genetika a virologie

Bc. Barbora Peková

Vyšetření rekombinací mezi genem a pseudogenem pro β-glukocerebrosidasu vedoucích ke vzniku patogenních alel

Detection of β-glukocerebrosidase gene/pseudogene recombination events leading to pathogenic alleles

Diplomová práce

Vedoucí závěrečné práce: MUDr. Martin Hřebíček, Ph.D.

Praha, 2017

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 2.5.2017

Barbora Peková

Poděkování:

Chtěla bych poděkovat MUDr. Martinu Hřebíčkovi, Ph.D. za odborné vedení a poskytnutou pomoc při zpracovávání této diplomové práce. Ráda bych poděkovala Ing. Lence Mrázové, Ph.D. za svědomitý přístup, cenné rady a ochotu zabývat se vzniklými problémy. Mé poděkování patří i pracovníkům laboratoře ÚDMP VFN a 1. LF UK v Praze za rady a podněty.

Abstrakt

Cílem této diplomové práce bylo vypracovat přehled o genové konverzi, její roli v patogenezi lidských onemocnění a zhodnotit využití metod založených na sekvenování nové generace (NGS) pro detekci vzácných změn sekvence DNA. Jedná se o pilotní studii využití NGS pro detekci bodových mutací vznikajících *de novo* v genu pro β-glukocerebrosidasu genovou konverzí mezi ním a jeho pseudogenem v meiotických a mitotických buňkách kontrolních osob. Primery specifické pro aktivní gen byly použity pro selektivní amplifikaci úseku devátého a desátého exonu genu, kde se „rekombinantní" změny vyskytují nejčastěji. Byla využita metoda značení cílových molekul DNA pomocí náhodných sekvencí v primeru.

Bioinformatickým zpracováním byly detekovány bodové mutace v sekvencích získaných z 20 vzorků genomové DNA na platformě Illumina MiSeq. Sekvence byly filtrovány, tříděny podle unikátního značení jednotlivých molekul DNA a byly vytvořeny alignmenty těchto sekvencí. Softwarovou detekcí se záměrně nestriktními kritérii bylo v jednotlivých vzorcích nalezeno 12-48 potenciálních bodových mutací, které byly následně ověřovány přehodnocením alignmentů sekvencí nesoucích shodnou značící sekvenci. Počet alignmentů s unikátním značením byl v rozmezí 7-15 tisíc na vzorek. Ve třech vzorcích byly nalezeny v genu pro β-glukocerebrosidasu mutace potenciálně vzniklé genovou konverzí, což svědčí pro nižší frekvenci, než jaká byla detekována staršími metodami. Analýza značících sekvencí v primeru naznačuje možné způsoby zlepšení citlivosti metody.

Dalším cílem bylo testování metod přípravy templátu pro NGS pomocí obohacování konvertovaných sekvencí. Jednalo se o metodu založenou na principu zámkových sond, mutačně specifického PCR a izotermické amplifikace zprostředkované smyčkou. Při aplikaci na lidskou genomovou DNA tyto metody však nevedly k získání dostatečného množství kvalitního templátu pro NGS.

Klíčová slova: genová konverze, gen pro β-glukocerebrosidasu, pseudogen pro β-glukocerebrosidasu, sekvenování nové generace, bodová mutace

Abstract

This diploma thesis provides an overview of gene conversion, its role in the pathogenesis of human diseases and the use of methods based on next-generation sequencing (NGS) for detection rare variants of DNA sequence. Labeling of target DNA molecules by random nucleotides in primer and NGS were used for detection point mutations arising *de novo* in the β-glucocerebrosidase gene by gene conversion between it and its pseudogene in meiotic and mitotic cells of control subjects. Primers specific for the active gene were used to selectively amplify the ninth and tenth exon of the gene where "recombinant" variants occur most frequently.

Sequences generated from 20 genomic DNA samples on Illumina MiSeq platform were quality filtered, sorted by unique labels and consensus sequences were created from alignments of sequences carrying the same DNA tag. The number of potential point mutations in the samples ranged between 12 and 48. The mutations were manually re-evaluated from the alignments. The number of alignments with unique labeling was in the range of 7-15 thousand per sample. Only three samples carried possible recombinant mutations, suggesting a lower frequency of conversion in the region than reported by other techniques. Analysis of unique sequences in primer indicated possible ways to improve the sensitivity of the method.

Another goal was to test methods of preparing the template for NGS by enriching the converted sequences. Methods based on padlock probes, mutational-specific PCR and loop-mediated isothermal amplification yield sufficient amounts of good quality template for NGS.

Key words: gene conversion, β-glucocerebrosidase gene, β-glucocerebrosidase pseudogene, next-generation sequencing, point mutation

Obsah

Seznam zkratek 8

Úvod 10

Cíle práce 11

Přehled literatury 12

1 Genová konverze 12

1.1 Frekvence genové konverze 14

2 Oprava dvouřetězcového zlomu 16

3 Pseudogeny 19

4 Gen a pseudogen pro β-glukocerebrosidasu 21

4.1 Mutace 21

4.2 Gaucherova choroba 23

5 Detekce vzácných mutací 24

5.1 Původní studie genové konverze mezi aktivním genem a pseudogenem 24

5.2 Nové metody pro zachycení vzácných mutací 26

6 Zámkové sondy 30

6.1 Ligační metoda 32

7 Asymetrická extenze 34

8 Izotermální amplifikace zprostředkovaná smyčkou 34

Materiál a metody 36

9 Materiál 36

10 Metody 39

10.1 Elektroforéza 39

10.2 Purifikace PCR produktů 40

10.3 Příprava standardů - pozitivních a negativních kontrol použitých metod 43

10.4 Klonování PCR produktu 44

10.5 Asymetrická extenze....47
10.5.1 Příprava templátu pro NGS - bez obohacení konvertovanými sekvencemi....47
10.5.2 Příprava templátu pro NGS - s obohacením konvertovanými sekvencemi....49
10.6 Příprava templátu pro NGS....51
10.7 Ligační metoda....52
10.8 LAMP....57
11 Bioinformatické zpracování výsledků....58
11.1 Filtrování....59
Výsledky....61
12 Optimalizace a výsledky jednotlivých metod....61
13 Asymetrická extenze....61
13.1 Příprava templátu pro NGS - bez obohacení konvertovanými sekvencemi....61
13.2 Příprava templátu pro NGS - s obohacením konvertovanými sekvencemi....63
14 Ligační metoda....65
15 LAMP....68
16 Výsledky NGS....69
16.1 Testovací sekvenování na platformě Illumina MiSeq....69
16.2 Finální sekvenování na platformě Illumina MiSeq....70
Diskuse....76
Závěr....81
Seznam literatury....82
Příloha 1....88

Seznam zkratek

AA	akrylamide	akrylamid
APS	ammonium persulfate	persíran amonný
ARMS	amplification refractory mutation system	amplifikační refrakční mutační systém
BIR	break-induced replication	zlomem indukovaná replikace
cDNA	complementary deoxyribonucleic acid	komplementární deoxyribonukleová kyselina
dATP	deoxyadenosine triphosphate	deoxyadenosin trifosfát
DMSO	dimetyl sulfoxide	dimetylsulfoxid
dNTP	deoxynucleotide triphosphate	deoxynukleotid trifosfát
DSB	double-strand break	dvouřetězcový zlom
DSBR	double-strand break repair	oprava dvouřetězcového zlomu
dsDNA	double-stranded deoxyribonucleic acid	dvouřetězcová deoxyribonukleová kyselina
GBA	β-glukocerebrosidase	β-glukocerebrosidasa
GBA	gene for β-glukocerebrosidase	gen pro β-glukocerebrosidasu
GBAP	pseudogene for β-glukocerebrosidase	pseudogen pro β-glukocerebrosidasu
gDNA	genomic deoxyribonucleic acid	genomová deoxyribonukleová kyselina
HLA	human leukocyte antigen	hlavní histokompatibilní komplex
LAMP	loop mediated isothermal amplification	izotermální amplifikace zprostředkovaná smyčkou
MEPS	minimum efficient processing segment	minimální velikost sekvence pro účinnou konverzi
mRNA	messenger RNA	mediátorová RNA
NGS	next-generation sequencing	sekvenování nové generace

PAGE	polyacrylamide gel electroforesis	polyakrylamidová elektroforéza
PCR	polymerase chain reaction	polymerázová řetězová reakce
RCA	rolling circle amplification	amplifikace mechanismem valivé kružnice
Rec	recombinant	rekombinantní
RNAi	RNA interference	RNA interference
SDSA	synthesis-dependent strand annealing	hybridizace řetězce druhotnou syntézou
SMART	spacer multiplex amplification reaction	
SNP	single nucleotide polymorphism	jednonukleotidový polymorfismus
SSA	single-strand annealing	jednořetězcová hybridizace
ssDNA	single-stranded DNA	jednořetězcová DNA
TEMED	N,N,N′,N′-tetramethylethylendiamine	N,N,N′,N′-tetramethylethylendiamin

Úvod

Homologní rekombinace může probíhat odlišnými mechanismy, překřížením nebo genovou konverzí. Při genová konverzi dochází k jednostrannému přenosu fragmentu DNA mezi donorem a vysoce homologním akceptorem. Dárcovská molekula zůstává na rozdíl od mechanismu překřížení během tohoto procesu nezměněna. Konverze probíhá při opravě dvouřetězcového zlomu DNA molekul, ke kterému dochází s vysokou frekvencí jak v prokaryotických, tak v eukaryotických buňkách. Genová konverze hraje roli v evoluci genomu, protože vede ke změnám v genech a nekódujících sekvencích a má také vliv na rozvoj genetických chorob. Genová konverze by mohla mít v budoucnu využití v genové terapii. Pomocí tohoto typu rekombinace by bylo možné vnášet do DNA s defektním genem gen funkční.

Genová konverze se odehrává *de novo* v somatických a pohlavních buňkách zdravého jedince, což znamená, že k tomuto rekombinačnímu procesu dochází, jak během meiózy, tak i během mitózy. Význam konverze u zdravých jedinců je v somatických buňkách zanedbatelný, u pohlavních buněk je však riziko, že dají vznik novému jedinci nesoucímu patogenní alely. Genová konverze byla v buňkách zdravých jedinců v minulosti studována. V současné době však disponujeme novými přístroji a metodickými postupy, díky kterým jsme schopni přesněji odhadnout, s jakou frekvencí ke konverzi dochází.

V této práci se konkrétně zabývám využitím metod sekvenování nové generace pro studium genové konverze mezi genem a pseudogenem pro β-glukocerebrosidasu, díky které dochází ke vzniku patogenních alel, které mohou být příčinou vzniku Gaucherovy choroby.

Cíle práce

1) Zhodnotit využití metod založených na sekvenování nové generace pro detekci vzácných bodových mutací lidské genomové DNA.
2) Zavést metodu značení jednotlivých molekul DNA v templátu pro sekvenování nové generace a sekvenovat touto technikou úsek devátého a desátého exonu genu pro β-glukocerebrosidasu na sekvenátoru MiSeq.
3) Na základě výsledků zpřesnit odhad frekvence *de novo* genové konverze mezi genem a pseudogenem pro β-glukocerebrosidasu u meiotických a mitotických buněk kontrolních osob a porovnat jej s dřívějšími odhady.
4) Vyšetřit vhodnost použití metod na principu obohacování konvertovanými sekvencemi.

Přehled literatury

1 Genová konverze

Genová konverze je typem homologní rekombinace, ke které dochází během meiózy i mitózy. Genová konverze probíhá při opravě DNA poškozené vlivem vnějšího prostředí (UV záření, chemické látky, viry), vnitřního prostředí (poruchy replikace DNA, reaktivní kyslíkové formy) nebo v meióze během výměny genetické informace. Jedná se o jednostranný přenos genetické informace mezi vysoce sekvenčně homologními úseky DNA - donorem a akceptorem. Přenos se odehrává mezi homologními chromozomy, sesterskými chromatidami nebo homologními sekvencemi na stejném nebo jiném chromozomu. U lidí jsou přenášené úseky dlouhé obvykle několik stovek párů bází (Elliott et al. 1998).

Homologní rekombinace v meióze a mitóze

Meióza u eukaryot probíhá během spermatogeneze a oogeneze. K homologní rekombinaci dochází v profázi prvního meiotického dělení, konkrétně v pachytenu mezi dvěma duplexy DNA. Frekvence rekombinace se u gamet liší. Uvádí se, že frekvence rekombinace u oocytů je 1,5krát vyšší než u spermatocytů (Tease, Hartshorne and Hultén 2002).

U mitoticky dělících se buněk dochází k homologní rekombinaci především při opravě DNA v důsledku jejího poškození. Spontánně probíhá mitotická rekombinace v porovnání s meiotickou rekombinací vzácněji, konkrétně 25 tisíckrát méně často (Barbera and Petes 2006). Nejčastěji dochází k mitotické rekombinaci mezi sesterskými chromatidami, ale může probíhat i mezi homologními chromozomy. Z důvodu reciproké mitotické rekombinace mezi homologními chromozomy s heterozygotními alelami mohou z mateřské buňky vzniknout dvě homozygotní dceřiné buňky. Tento jev je nazýván ztráta heterozygozity. Pokud dochází k reciproké mitotické rekombinaci mezi nealelickými (ektopickými) homologními sekvencemi, mohou vznikat translokace, delece, duplikace, inverze a centrické nebo dicentrické chromozomy. Mitotická rekombinace mezi homologními chromozomy byla poprvé popsána roku 1936 na modelovém organismu *Drosofila melanogaster* (Stern 1936). K mitotické rekombinaci nedochází v samotné mitóze, ale v buněčné interfázi (Fabre 1978, Esposito 1978). Úseky přenášené mitotickou genovou konverzí jsou o mnoho delší než úseky přenášené v meióze. Jedná se přibližně o několik kilobází dlouhý úsek v mitóze a několik stovek bází dlouhý úsek během meiotické genové konverze (Yim et al. 2014). Genová konverze během meiózy má velký vliv na diversifikaci alel. Dochází ke sjednocování

příbuzných a homologních sekvencí, což má v určitých případech za následek vznik komplexních alel způsobujících dědičná onemocnění (Jeffreys and May 2004).

Interalelická a nealelická genová konverze

Ke genové konverzi dochází při opravě dvouřetězcového zlomu (Double-Strand Break, DSB), dělíme ji na interalelickou a nealelickou. K interalelické genové konverzi dochází při meióze mezi alelami na homologních chromozomech. Interalelická genová konverze má velký význam při tvorbě nových kombinací polymorfismů v duplikovaných genech a zvyšuje množství haplotypů, například u hlavního histokompatibilního komplexu (Ohta 1999). Ektopická genová konverze probíhá v meióze i mitóze, jedná se o přenos genetické informace mezi různými lokusy na různých chromozomech (Galtier et al. 2001).

Chybné párování bází a genová konverze

Ke genové konverzi dochází nejen při opravě DSB, ale také při opravě chybného párování bází, kdy dochází k vystřižení a nahrazení báze na základě komplementarity. Pokud je opravnými mechanismy buňky zjištěna záměna (mismatch) A/T, je vysoce pravděpodobné, že bude báze opravena na C/G a povede ke konverzi. Nálezem svědčícím pro význam tohoto mechanismu může být vysoký obsah GC páru v oblastech, kde dochází ke genové konverzi s vysokou frekvencí (Galtier et al. 2001).

Koheziny

Koheziny jsou proteinové komplexy, jejichž funkcí je regulace separace sesterských chromatid. V mitóze koheziny zajišťují, že k opravě DSB dochází přednostně mezi sesterskými chromatidami. Koheziny hrají také roli v meióze, kdy drží sesterské chromatidy pohromadě do pozdních fází profáze. Koheze sesterských chromatid je na určitých místech uvolněná během DSB-zprostředkované rekombinace v meióze, aby mohly spolu interagovat homologní chromozomy, a nikoliv sesterské chromatidy (Klein et al. 1999). V profázi meiózy dochází ke vzniku mnoha DSB, které nepokračují do překřížení. K opravě těchto DSB pomocí rekombinací nejsou preferované sesterské, ale homologní molekuly. Předpokládá se, že v genomu zůstávají stopy těchto zastavených rekombinací jako genové konverze (Cortés-Ledesma and Aguilera 2006).

1.1 Frekvence genové konverze

Frekvenci genové konverze ovlivňuje několik faktorů: specifické DNA motivy a struktury, stupeň homologie, vzdálenost, poziční efekt a délka homologních sekvencí. Dále frekvenci ovlivňuje, jestli se jedná o genovou konverzi v meióze nebo mitóze.

Specifické DNA motivy a struktury

Místa, ve kterých dochází k rekombinaci během meiózy nebo mitózy častěji, se nazývají hotspoty. Tyto hotspoty se v genomu nevyskytují náhodně, prakticky se nevyskytují v oblasti centromer a telomer (Su, Barton and Kaback 2000, Rockmill, Voelkel-Meiman and Roeder 2006). Naopak vysoké zastoupení hotspotů je v subtelometrické oblasti (Linardopoulou et al. 2005). Hotspoty se nacházejí v místech se specifickými sekvenčními motivy a DNA strukturami, které jsou náchylné k DSB. Jedná se o sekvence, které mají potenciál k tvorbě jiné než B-konformaci DNA, chi-like sekvence a purin-pyrimidinové úseky vytvářející levotočivou Z-konformaci DNA, dále sekvence tvořící trojvláknovou DNA (H-DNA), např. u genů kódujících transkripční faktory (c-myc) nebo sekvence bohaté na guanin, které tvoří G-kvadruplexy (Chuzhanova et al. 2009). Kromě toho je vysoká pravděpodobnost genové konverze u repetetivní sekvencí, např. u *Alu* sekvencí. Přímé repetice mohou tvořit tzv. klouzavou DNA (slipped DNA). Invertované repetice mohou tvořit stabilní vlásenkovou strukturu nebo strukturu křížovou, což je dvojitá vlásenka (Lilley 1980). Vlásenky se mohou vytvořit během replikace na opožďujícím se vlákně a způsobit tak zastavení replikační vidlice (Gordenin and Resnick 1998). Hotspoty se nacházejí také často v CpG bohatých oblastech, např. hotspot se sekvencí CCTCCCCT (Myers et al. 2005).

Nedávný výzkum dokazuje, že většina DSB v meióze vzniká u myši a člověka v místech, na která se váže histon-lysin N-metyltransferáza PRDM9 (Baudat et al. 2010). Vazba tohoto proteinu na DNA vede k vazbě dalších proteinů včetně Spo11, které vytvoří iniciační rekombinační komplex. Tandemová pole DNA-vazebných „zinkových prstů“ PRDM9 jsou velmi polymorfní a jejich varianty mají afinitu pro různé motivy DNA. Místa vzniku DSB u člověka jsou tímto definována preferenční afinitou variant PRDM9. Byly vytvořeny rekombinační mapy meiotických buněk u vzorků ze spermií pěti osob a ukázaly, že pozice a frekvence využití většiny hotspotů DSB se shodují u osob nesoucích stejné alely PRDM9 (Pratto et al. 2014).

Kromě genové konverze může u duplikovaných genů dojít na jednom vlákně ke vzniku tzv. fúzní alely (Obr. 7) jako je tomu u genu *CYP21* a jeho pseudogenu *CYP21P*. Ke spojení

genu a pseudogenu zde dochází ve dvou různých oblastech. První z oblastí byla lokalizována v chi-like sekvenci GCTGGGC ve třetím intronu a druhá oblast byla lokalizována v tandemové repetici minisatelitní sekvence TGGCAGGAGG v pátém exonu pseudogenu (L'allemand et al. 2000).

Stupeň homologie

Frekvence genové konverze také záleží na stupni homologie dotčených sekvencí. Genová konverze obecně probíhá, pokud je sekvenční identita vyšší než 80 %, nejčastěji se jedná o 95% homologii. Jsou známy i případy, kdy ke konverzi dochází, i když je sekvenční identita nižší než 80 %, ale s o mnoho nižší frekvencí. Jestliže u *Escherichie coli* klesne homologie sekvencí ze 100 % na 90 %, tak frekvence konverze klesne 40krát (Reiter et al. 1998).

Vzdálenost a poziční efekt homologních sekvencí

Frekvence genové konverze je vyšší intrachromozomální než interchromozomální a to jak v meióze, tak i v mitóze (Lichten, Borts and Haber 1987). U intrachromozomální konverze záleží na vzdálenosti homologních úseků. Čím je větší vzdálenost mezi homologními úseky, tím je nižší frekvence genové konverze. Z pokusů na myších vyplynulo, že u vzdálenosti menší než 55 kilobází mezi homologními sekvencemi ke genové konverzi došlo ve 30 % případů. Pokud byla vzdálenost větší než 370 kilobází, frekvence klesla na 10 % (Ezawa, OOta and Saitou 2006). Významný je také poziční efekt. Jestliže se nachází homologní sekvence v blízkosti centromery, dochází ke genové konverzi až s 40krát nižší frekvencí (Lichten et al. 1987).

Délka homologních sekvencí

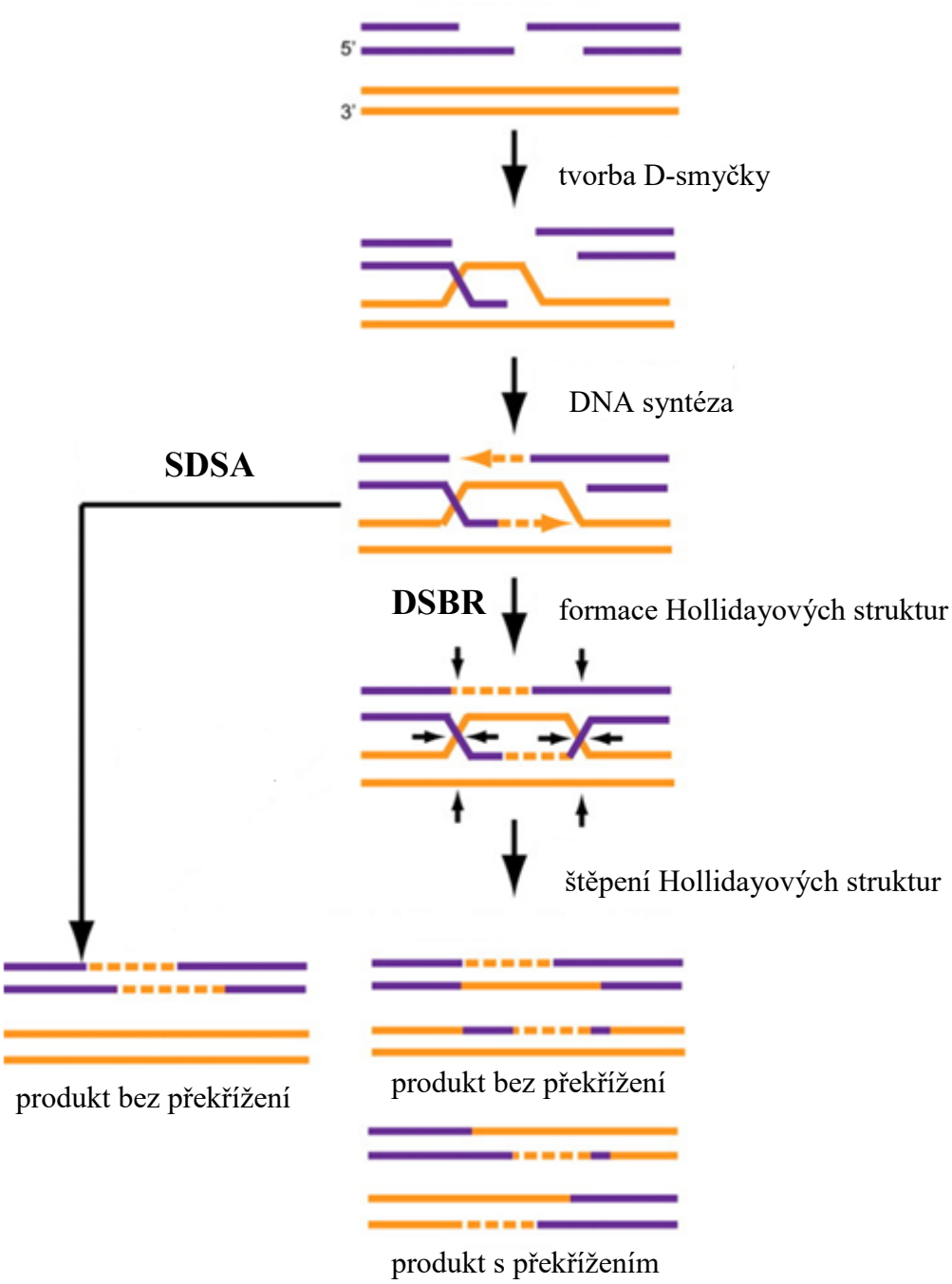
Délka homologních sekvencí je dalším faktorem ovlivňujícím frekvenci konverzí. Byla experimentálně stanovena minimální velikost sekvence pro účinnou konverzi (tzv. Minimum Efficient Processing Segment, MEPS) (Shen and Huang 1986). Délka MEPS záleží na organismu, u *Saccharomyces cerevisae* bylo MEPS stanoveno přibližně na 250 bází (Jinks-Robertson, Michelitch and Ramcharan 1993), u lidí na 350-450 bází (Reiter et al. 1998). V ojedinělých případech byla genová konverze pozorována i mezi menšími úseky o velikosti cca 10 bází (Mézard, Pompon and Nicolas 1992). Pravděpodobnost genové konverze je vyšší u delších homologních úseků. U 0,7 kilobází dlouhých homologních sekvencí savčího genomu dochází ke genové konverzi v 17-60 % případů homologní rekombinace vyvolané DSB. U dvojnásobně dlouhého úseku vzrůstá četnost genové konverze na 97 % (Taghian and Nickoloff 1997).

2 Oprava dvouřetězcového zlomu

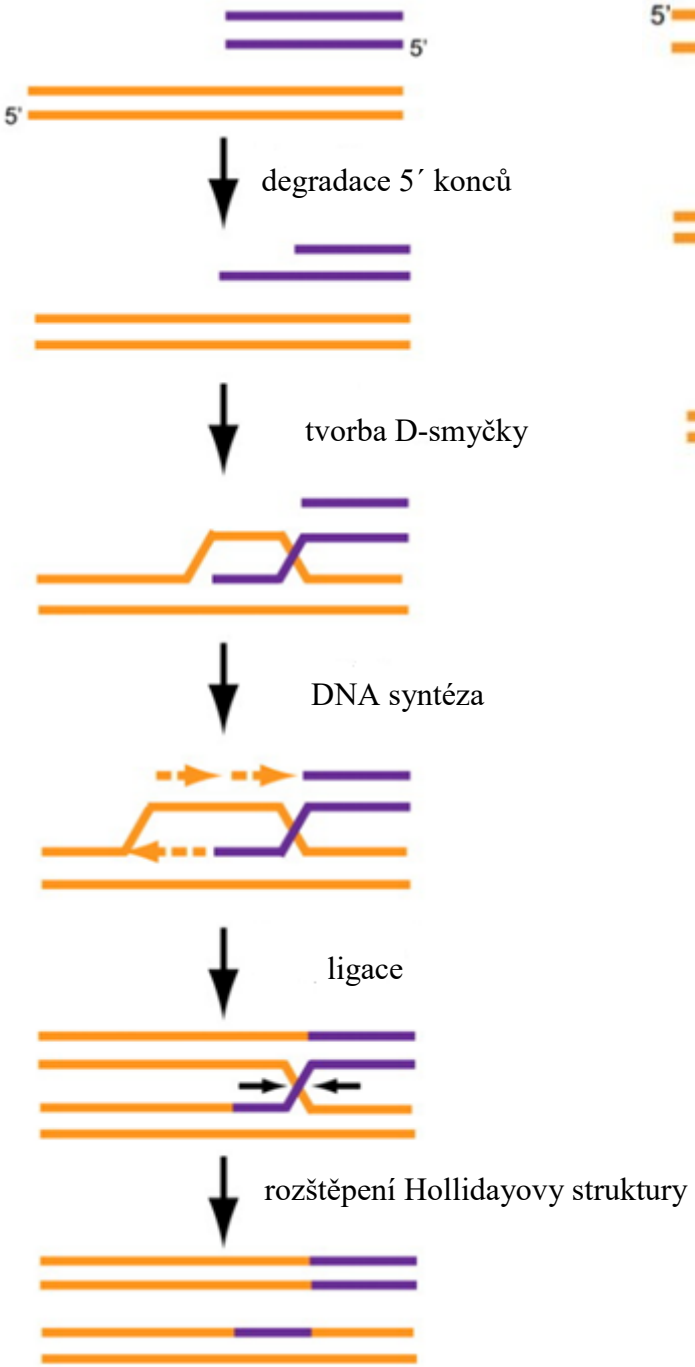
Ke genové konverzi dochází v místech DSB. V meióze je zlom iniciován v profázi, kdy se homologní chromozomy spárují za tvorby synaptonemálního komplexu a za účasti PRDM9 a Spo11 proteinů se vytvoří DSB (Baudat et al. 2010). V meióze je nejčastější mechanismus opravy dvouřetězcového zlomu DSBR (Double-Strand Break Repair) model, jehož výsledkem jsou jak produkty s překřížením, tak produkty bez překřížení (Obr. 1). Alternativní cestou genové konverze je oprava DSB hybridizací (Synthesis-Dependent Strand Annealing, SDSA), který je upřednostňován v mitóze, a výsledkem jsou jen produkty bez překřížení (Obr. 1) (Allers and Lichten 2001).

Podobným mechanismem může dojít ke genové konverzi při opravě zastavených nebo zkolabovaných replikačních vidlic, kdy je zlom v jedné části vidlice opraven nejčastěji podle neporušené sesterské chromatidy, kdy nedojde ke změně genetické informace. Mechanismus se nazývá zlomem indukovaná replikace (Break-Induced Replication, BIR) (Obr. 2) (Morrow, Connelly and Hieter 1997). BIR model se vyznačuje na rozdíl od ostatních modelů opravy DSB tím, že pouze jeden konec DSB je homologní se sekvencí donora. Pokud je jako templát pro opravu BIR použit stejný gen na homologním chromozomu nebo jiná homologní sekvence, může dojít ke konverzi. Tento model pomáhá také mimo jiné udržovat délku telomer (Roumelioti et al. 2016).

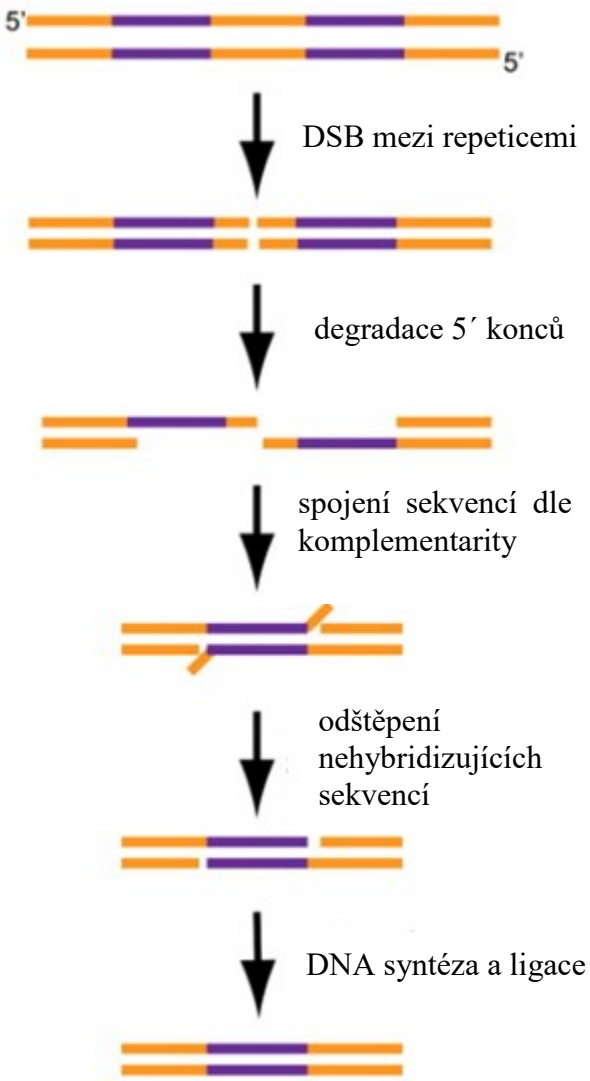
Dalším mechanismem opravy DSB rekombinací je jednořetězcová hybridizace (Single Strand Annealing, SSA) (Obr. 3). Dochází k ní v případě dvou přímých repetic, které mohou být vzdáleny až desítky kilobází. Při opravě je nenávratně vystřižena jedna z repetic a také oblast mezi repeticemi. Tento model je proto označován jako mutagenní (Schildkraut, Miller and Nickoloff 2005).



Obrázek 1: Mechanismus opravy DSB modelem DSBR a SDSA. Po vytvoření DSB dochází k degradaci 5′ konců. Jeden volný 3′ konec akceptorového řetězce putuje k duplexu donora, kde dojde odsunutím jednoho donorového řetězce k tvorbě D-smyčky. Volné 3′ konce akceptora slouží jako primery a prodlužují se DNA syntézou na základě komplementarity podle řetězců donora. SDSA model: invadující vlákno se na 3′ konci uvolní od D-smyčky a nasedne na komplementární 3′ sekvenci na druhé straně zlomu. Po doplnění chybějících bází a ligaci vznikne opravená molekula nesoucí donorovou sekvenci. Donorová DNA je nezměněna. DSBR model: D-smyčka zůstává zachována a dochází k ligaci na obou stranách zlomu. Tím vznikají dvě Holliadayovy struktury. Podle způsobu štěpení Holliadových struktur vzniká produkt s překřížením (reciproká rekombinace) nebo produkt bez překřížení (genová konverze). Převzato a upraveno z (Svendsen and Harper 2010).



Obrázek 2: Mechanismus opravy DSB modelem BIR. Po degradaci 5´ konce 3´ konec invaduje k sekvenci donora za tvorby D-smyčky. Po DNA syntéze a ligaci se vytvoří Hollidayova struktura, která je rozštěpena nukleázami. Převzato a upraveno z (Svendsen and Harper 2010).



Obrázek 3: Mechanismus opravy DSB modelem SSA. Po degradaci 5´ konců jsou přesahující 3´konce odstřiženy a následně DNA ligázou spojeny. Převzato a upraveno z (Svendsen and Harper 2010).

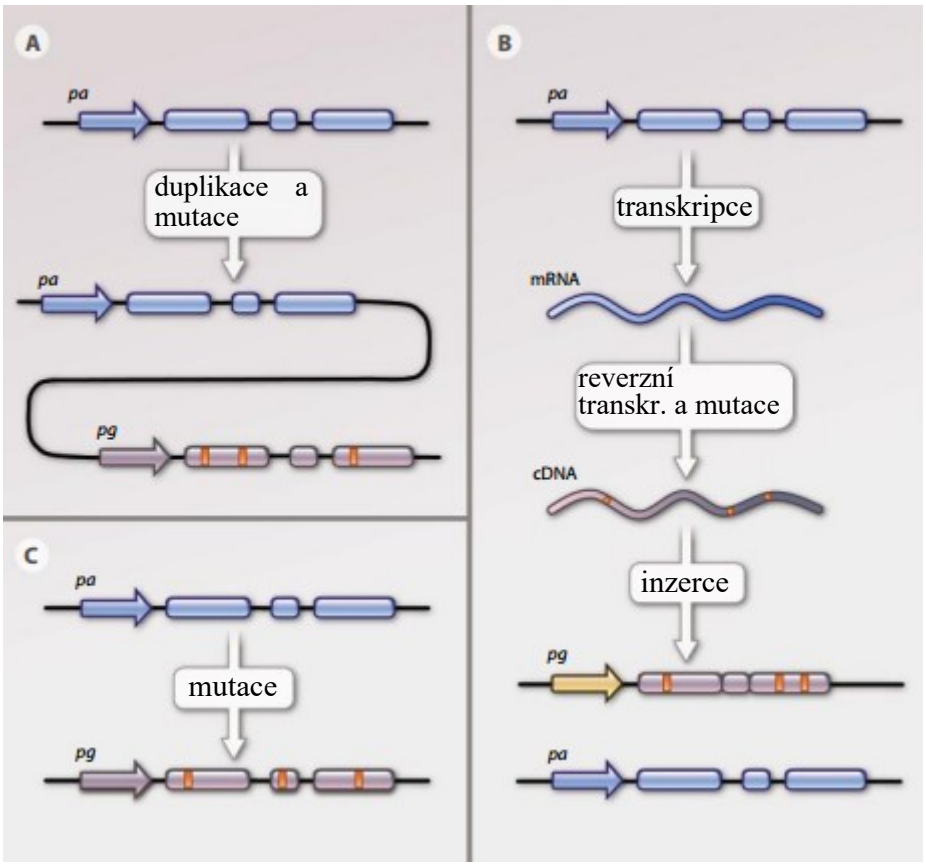
3 Pseudogeny

Roku 1977 C. Jacq objevil zkrácenou formu genu pro 5S ribozom u *Xenopus laevis*. Tato vysoce homologní sekvence k aktivnímu genu byla poprvé pojmenována pseudogen (Jacq, Miller and Brownlee 1977). Pseudogeny jsou vysoce homologní sekvence k aktivním genům, které ztratily schopnost tvořit funkční protein. Vyskytují se v genomu komplexních organismů, jako jsou obratlovci a vyšší rostliny, ale je možné je najít i u bakterií, hmyzu nebo například u organismů z říše Nematoda (Podlaha and Zhang 2010). Lidský a myší genom obsahuje téměř stejný počet pseudogenů jako aktivních genů (Zhang and Gerstein 2004).

Pseudogeny postrádají funkčnost a donedávna se myslelo, že i význam (Podlaha and Zhang 2010). Důvodem ztráty funkčnosti je předčasný stop-kodón nebo posunutí čtecího rámce. Nicméně nedávná studie pseudogenu u organismu *Drosophila sechellia* ukázala, že navzdory předčasnému stop-kodónu může vznikat funkční protein (Stensmyr 2016). Význam pseudogenů spočívá mimo jiné v tom, že RNA transkripty mohou ovlivňovat funkci nejenom mateřského genu, ale i úplně jiných genů. RNA transkripty pseudogenů mohou být procesovány na krátké interferující RNA, které RNAi drahou regulují aktivní geny. Pseudogenní protein může alostericky inhibovat protein kódovaný aktivním genem. Dokáží tedy zvyšovat nebo snižovat genovou expresi a mohou hrát i roli v patogenezi, např. u lidských nádorů (Poliseno 2012). Pseudogen *PTENP1* ovlivňuje funkci tumor supresorového genu *PTEN* tak, že transkripty genu a pseudogenu kompetují o navázání na stejné miRNA molekuly. Na buněčných liniích karcinomu prsu a tlustého střeva byl zjištěn deficit pseudogenu *PTENP1*. Transkripty tumor supresorového genu byly umlčeny miRNA v důsledku deficitu transkriptů pseudogenu a došlo tak k rozvoji rakovinného bujení (Poliseno et al. 2010).

Rozdělení

Pseudogeny se dělí na tři hlavní skupiny, na pseudogeny jednotné (unitary), procesivní a neprocesivní (Obr. 4). Jednotné pseudogeny vznikají spontánními mutacemi v aktivních genech. Procesivní pseudogeny vznikají retrotranspozicí mRNA transkriptů a jejich následným vmezeřením na nové místo v DNA. Mají polyA konec a nemají promotorové oblasti a introny (Maestre et al. 1995). Neprocesivní pseudogeny vznikají duplikací aktivních genů a následnou mutací. Obsahují exony, introny a promotorové oblasti. Příkladem neprocesivního pseudogenu je pseudogen pro β-glukocerebrosidasu.



Obrázek 4: Vznik pseudogenů. A. Neprocesivní pseudogeny vznikají duplikací a následnými mutacemi v duplikovaném genu. **B**. Procesivní pseudogeny vznikají transkripcí genu do mRNA, která je reverzní transkriptázou přepsána do kódující DNA (cDNA), ve které vznikají nové mutace a tato cDNA je zpět inzertována do genomu. **C.** U tzv. unitárních pseudogenů dochází v důsledku mutací k přetvoření genu na pseudogen, tudíž aktivní gen k takto vzniklému pseudogenu chybí, na rozdíl od předchozích dvou skupin. *pa* označení pro aktivní gen, *pg* pro pseudogen. Převzato a upraveno z (Poliseno 2012)

Onemocnění způsobená genovou konverzí mezi genem a pseudogenem

V důsledku genové konverze mezi aktivním genem a příslušným pseudogenem dochází k vnášení patogenních alel, které mohou způsobovat dědičná onemocnění. Příkladem může být genová konverze mezi kationickým trypsinogenem (*PRSS1*) a pseudogenem trypsinogenem 6 (*PRSS3P2*) způsobující chronickou pankreatitidu (Rygiel et al. 2015), nebo mezi beta-crystallinem B2 a jeho pseudogenem způsobující katarakty (Sarhadi et al. 2001). U Blackfan-Diamondovy anémie dochází ke genové konverzi mezi geny pro ribozomální proteiny a jejich pseudogeny (Boria et al. 2010). Mezi další onemocnění, která mohou být způsobená genovou konverzí mezi aktivním genem a pseudogenem patří například kongenitální andrenální dysplazie, von Willebrandova choroba, agamaglobulinémie, spinální muskulární atrofie, Gaucherova choroba a další (Bischof et al. 2006).

4 Gen a pseudogen pro β-glukocerebrosidasu

Gen a pseudogen pro β-glukocerebrosidasu jsou lokalizovány na chromozomu 1q21 (Ginns et al. 1985). Tento na geny bohatý region obsahuje sedm genů, dva pseudogeny a je dlouhý 85 kilobází. Nejblíže aktivnímu genu pro β-glukocerebrosidasu *(GBA)* a pseudogenu pro β-glukocerebrosidasu *(GBAP)* se nachází gen *MTX1* a pseudogen *MTXP* pro metaxin (Obr. 5). Sekvence *GBA* a *GBAP* byla popsána Horowitzovou roku 1989 (Horowitz et al. 1989). *GBA* je dlouhý 7,6 kilobází a je rozdělen na 11 exonů a 10 intronů, v intronových oblastech se nachází 4 *Alu* repetice. *GBAP* leží 16 kilobází od genu, jeho délka činí 5,7 kilobází a má stejný počet exonů a intronů jako *GBA*, avšak obsahuje pouze 1 *Alu* repetici v sedmém intronu (Zimran et al. 1990) a 55 bázovou deleci v devátém exonu. Sekvence *GBAP* je z 96 % homologní k sekvenci *GBA*. *GBAP* je transkribován díky zachovanému aktivnímu promotoru, který má stejně jako promotor *GBA* dva TATA boxy a dva CAT boxy. Jediný rozdíl v sekvenci promotoru *GBA* a *GBAP* je v substituci v druhém CAT boxu (Horowitz et al. 1989). Aktivita promotoru *GBAP* je slabší než u *GBA* (Reiner and Horowitz 1988).

Obrázek 5: Znázornění na geny bohaté oblasti na chromozomu 1q21. Pořadí sedmi genů (*clk2, propin1, cote1, GBA, MTX1, THBS3, MUC1*) a dvou pseudogenů (*MTXP*, *GBAP*) v 1q21 oblasti chromozomu. *MTX1* a *MTXP* jsou transkribovány v opačném směru než ostatní geny a pseudogen. Převzato a upraveno z (Winfield et al. 1997).

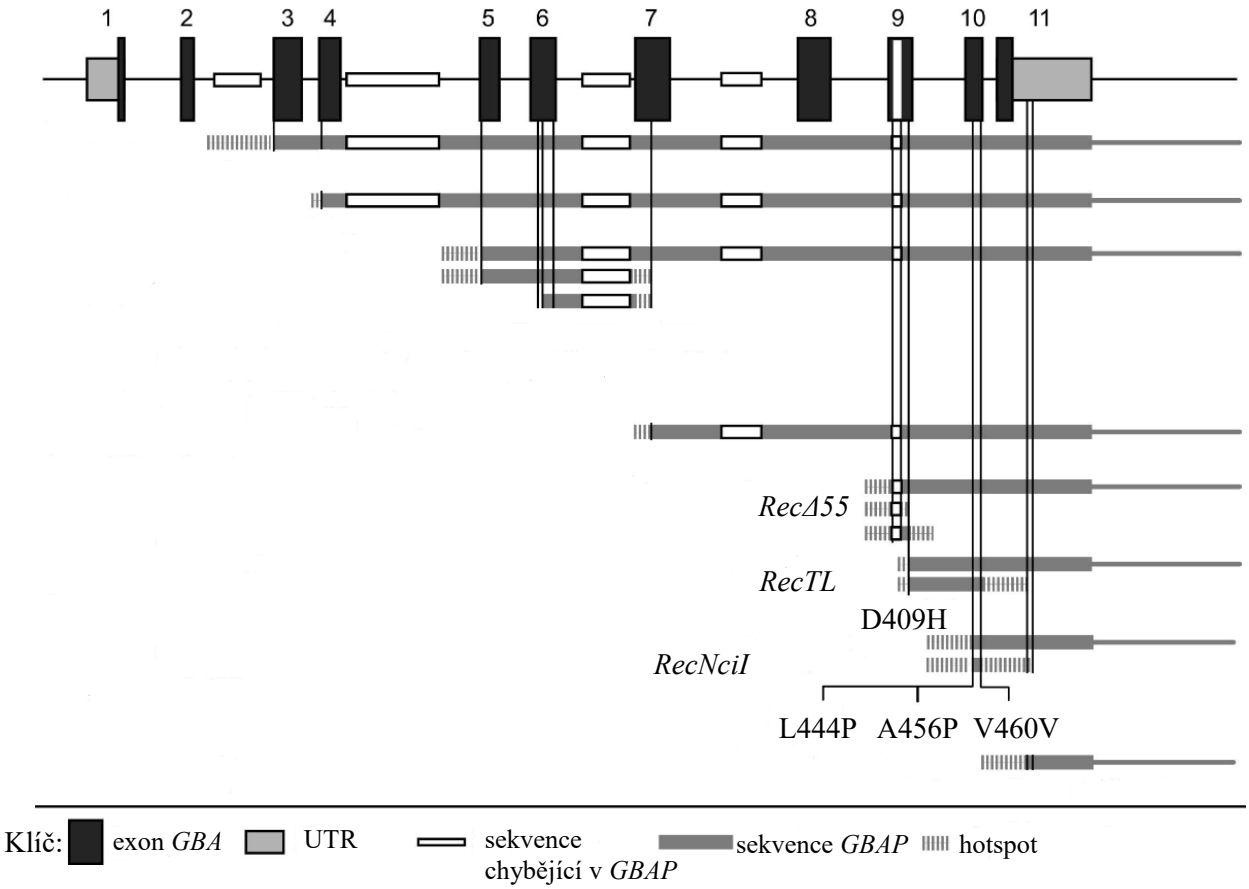
4.1 Mutace

U pacientů s Gaucherovou chorobou bylo zaznamenáno přes 200 mutací. Mezi nejčastější mutace se řadí 1226A>G (N370S), 1448T>C (L444P), 1342G>C (D409H), 1504C>T (R463C). Substituce L444P se nachází v desátém exonu a objevuje se u všech typů Gaucherovy choroby (Tsuji et al. 1987).

Jako rekombinantní (*Rec*) se označují komplexní alely, které obsahují dvě a více mutací, které se vyskytují v *GBAP* (Obr. 6). Mutace L444P spolu s mutacemi A456P a V460V tvoří rekombinantní alelu *RecNciI* začínající v rozmezí devátého intronu a desátého exonu, kdy je sekvence genu nahrazena sekvencí pseudogenu (Eyal, Wilder and Horowitz 1990). L444P a A456P jsou mutace měnící smysl kodónu (missense) a V460V je tichým polymorfismem.

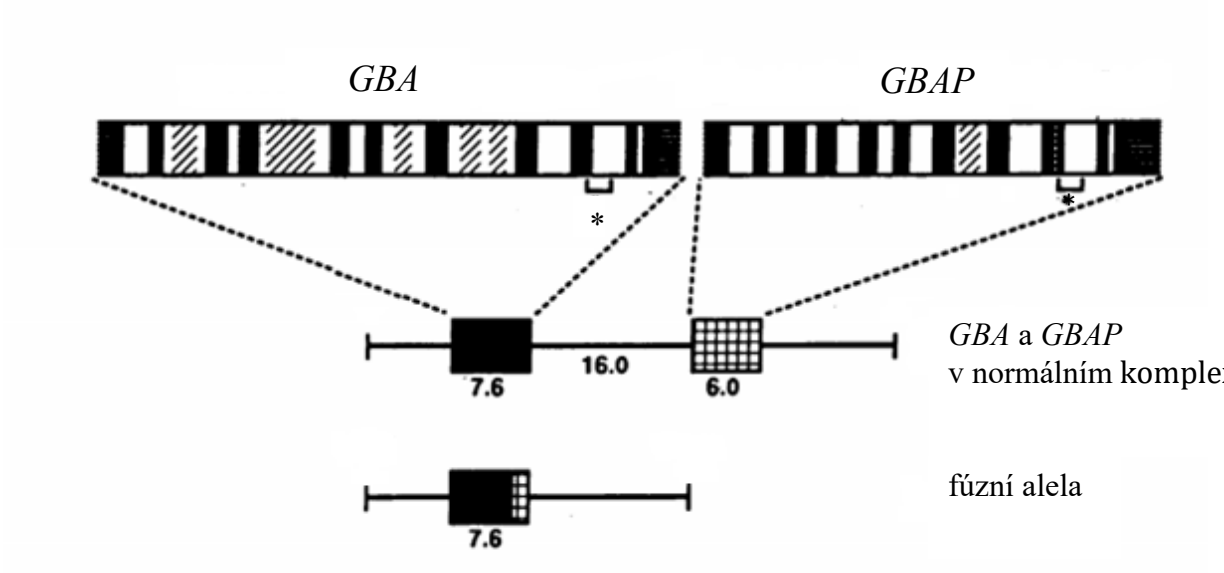
Vzdálenosti mezi těmito třemi mutacemi jsou malé, dělí je jen několik bází. Byly popsány i další rekombinantní alely *RecΔ55* a *RecTL* začínající v rozmezí osmého intronu a devátého exonu (Tayebi et al. 2003). Alela *RecTL* zahrnuje mimo L444P, A456P a V460V také mutaci D406H z devátého exonu. Alela *RecΔ55* obsahuje oproti *RecTL* navíc 55 bázovou deleci z devátého exonu (Beutler, Gelbart and West 1993a). V naší laboratoři byla objevena alela *RecFS*, která je zajímavá tím, že zahrnuje jen mutaci L444P a V460V, přitom neobsahuje mutaci A456P, která se nachází mezi nimi (Hodaňová et al. 1999).

Sekvence *GBA* je mezi intronem 8 a 3´ nepřekládanou oblastí z 98 % homologní k sekvenci *GBAP*, což je o 2 % více, než je tomu u zbylé sekvence (96 %). Dokonce tento region obsahuje 5 úseků menších než 200 bází, kde je homologie 100%. Proto v této oblasti dochází s nejvyšší frekvencí ke genové konverzi (Tayebi et al. 2003).



Obrázek 6: Znázornění oblastí v *GBA*, ve kterých dochází ke genové konverzi. *Rec* alely se nacházejí v devátém a desátém exonu, kde nejčastěji dochází k rekombinacím. K rekombinacím dochází také v menší míře v úseku třetího exonu a druhého intronu (Cormand et al. 2000), dále také ve výjimečných případech v oblasti mezi čtvrtým a sedmým exonem a ve 3´ nepřekládané oblasti. Převzato a upraveno z (Velayati et al. 2011).

Kromě rekombinantních alel byly popsány také fúzní alely, u kterých 3′ konec *GBA* byl nahrazen 5′ koncem *GBAP* (Obr. 7). U fúzní alely 5′ konec odpovídá sekvenci *GBA* a 3′ konec sekvenci *GBAP*. Oblast mezi *GBA* a *GBAP* spolu s *MTXP* je vystřižena mechanismem SSA (Obr. 3) (Zimran et al. 1990).



Obrázek 7: Fúzní alela. V horní části obrázku je znázorněn *GBA* a *GBAP* v normálním komplexu, kde tmavé pruhy znázorňují exony, světlé introny a šrafované *Alu* repetice. V dolní části obrázku je znázorněna fúzní alela, která má stejnou velikost 7,6 kilobází jako *GBA*. Úsek na 3′ konci dlouhý 1,6 kilobází je nahrazen sekvencí *GBAP* a sekvence mezi *GBA* a *GBAP* je nenávratně vystřižena. Hvězdičkou je na *GBA* i *GBAP* znázorněna oblast, ve které dochází k rekombinaci. Převzato a upraveno z (Zimran et al. 1990).

4.2 Gaucherova choroba

Gaucherova choroba patří mezi nejčastější lysosomální onemocnění. Je způsobená deficitem aktivity enzymu β-glukocerebrosidasy (GBA). Glukocerebrosidy nejsou enzymem štěpeny na ceramidy a glukózu a hromadí se zejména v buňkách makrofágového původu (Kattlove et al. 1969). Onemocnění se dědí autosomálně recesivně. U prvního typu choroby se incidence celosvětově odhaduje zhruba na 1 : 40 000 a u druhého a třetího typu na 1 : 100 000 (Charrow et al. 2000). V populaci Aškenázských Židů je incidence podstatně vyšší, uvádí se 1 : 855 a četnost přenašečů choroby se odhaduje na 1 : 14 u prvního typu choroby (Beutler et al. 1993b). Mezi klinické příznaky patří zvětšení jater a sleziny. Dále je postižena kosterní soustava z důvodu hromadění glukocerebrosidů v kostní dřeni. Většina pacientů trpí osteoporózou, což vede ke zvýšené lámavosti kostí. Jedinci s Gaucherovou chorobou mají nedostatek červených krvinek (anémie), krevních destiček (trombocytopenie) a mohou trpět zvýšenou krvácivostí. Dále mohou být postiženy plíce a centrální nervová soustava (Cox and

Schofield 1997). Gaucherova choroba se dělí na tři typy. Většina pacientů spadá do prvního typu bez neurologických komplikací (viscerální typ). Nejvyšší výskyt jedinců postižených Gaucherovou chorobou prvního typu je v populaci Aškenázských Židů. Pacienti s druhým typem nemoci (akutní neuronopatický) mají těžce postiženou nervovou soustavu a umírají v kojeneckém věku (Stone et al. 2000). Pro třetí typ (chronický neuronopatický) je charakteristické neurologické postižení s pomalým vývojem, jehož první příznaky se objevují v dětském školním věku nebo adolescenci. U tohoto typu nemoci se mohou vyskytovat dýchací obtíže v důsledku infiltrace plic střádajícími makrofágy. Typickým neurologickým příznakem je ztráta schopnosti pohybu očí ze strany na stranu (okulomotorická apraxie). Nejvyšší incidence třetího typu onemocnění je v oblasti Švédska zvané Norbotten, kde jsou pacienti obvykle homozygoti pro mutaci L444P v důsledku efektu zakladatele (Dreborg, Erikson and Hagberg 1980). Viscerální příznaky se v současné době mohou léčit enzymovou substituční terapií, kdy je pacientům intravenózně podávána modifikovaná rekombinantní β-glukocerebrosidasa (Barton et al. 1991). Další možností je substrát redukční terapie pomocí inhibice glukosyl transferázy, která katalyzuje tvorbu glukocerebrosidu z glukózy a ceramidu, a tím dochází ke snížení zátěže mutantního enzymu substrátem. Nejnovější studie prováděné na jedincích s Gaucherovou chorobou typu 1 ukázaly, že substrát redukční terapie je účinnější než enzymová substituční terapie (Smid et al. 2016). V současné době probíhá výzkum léčby pomocí chaperonů nebo genové terapie (Migdalska-Richards et al. 2017, Dahl et al. 2015).

5 Detekce vzácných mutací

5.1 Původní studie genové konverze mezi aktivním genem a pseudogenem

V minulých letech byly pokusy o zachycení vzácných mutací DNA vzniklých překřížením nebo genovou konverzí pomocí alelově specifické polymerázové řetězové reakce (alelově specifické PCR, ARMS) (Fan et al. 2001). To mělo své nevýhody. Nebylo možné zjistit rozsah konverzí z důvodu neschopnosti detekovat jednotlivé rekombinantní molekuly. Dalším úskalím je chybovost DNA polymerázy, která zvyšuje riziko falešné pozitivity. Jako materiál se většinou používaly meiotické buňky, a to spermie od anonymních dárců. V průběhu let došlo k vývoji nových metod umožňujících sekvenaci tisíců až milionů molekul najednou. Při využití polymeráz s nižší chybovostí a technik umožňujících rozpoznat chyby způsobené polymerázami, lze zvýšit přesnost těchto metod.

Detekce konverzí u kongenitální adrenální hyperplazie

Jedna ze starších studií se zabývá genovou konverzí mezi genem pro 21-hydroxylázu a jejími pseudogeny (Tusié-Luna and White 1995). Deficit enzymu 21-hydroxylázy vede k rozvoji onemocnění kongenitální adrenální hyperplazie. Pacienti trpící touto chorobou nejsou schopni syntetizovat kortizol.

Mutantní alely mohou obsahovat 30 kilobázovou deleci na rozhraní intronu 2 a exonu 3 vznikající nerovnoměrným překřížením nebo bodové mutace vzniklé genovou konverzí mezi genem a pseudogenem. Jako materiál byly použity spermie a leukocyty zdravých osob, pro porovnání frekvence genové konverze během meiózy a mitózy.

Frekvence genové konverze byla stanovena pomocí dvoufázového semi-nested PCR. V prvním kole PCR byly použity primery specifické pro gen na rozhraní intronu 2 a exonu 3. Molekuly nesoucí bodovou mutaci před koncem druhého intronu (i2g mutace) byly detekovány v druhém alelově-specifickém kole PCR. Sekvence horního primeru byla stejná jako v prvním kole, sekvence dolního primeru byla komplementární k i2g na 3´ konci. Produkty 2. kola PCR byly vizualizovány autoradiografií na polyakrylamidovém gelu díky radioaktivnímu značení jednoho z primerů. K ověření senzitivity a specifity metody, byla použita ředící řada od 500 ng do 1 pg DNA vzorku přenašeče i2g mutace. Pozitivní signál byl získán i při množství 1 pg DNA, z čehož lze usoudit, že je metoda senzitivní na úrovni téměř jedné molekuly. Frekvence genové konverze se lišila přibližně o 2 řády mezi jednotlivci. Nejnižší množství DNA, při kterém bylo možné detekovat mutaci, bylo od 1 do 100 ng. Takové množství DNA odpovídá frekvenci genové konverze 1 : 1 000 až 1 : 100 000, a to jak u spermatických buněk, tak u leukocytů. *De novo* delece byla zjištěna s frekvencí 1 : 100 000 až 1 : 1 000 000 a to pouze u spermatických buněk. Z toho vyplývá, že k nerovnoměrnému překřížení dochází pouze během meiózy (Tusié-Luna and White 1995).

U jednoho anonymního dárce bylo zjištěno, že u něho dochází s vysokou frekvencí ke genové konverzi a s nízkou frekvencí k deleci. Příčinou mohou být negenetické faktory nebo polymorfismus spjatý s *CYP21* a *CYP21P*. Jako negenetický faktor ovlivňující frekvenci konverze může vystupovat věk a pohlaví jednotlivce (Halldorsson et al. 2016).

Frekvence genové konverze mezi *GBA* a *GBAP*

V naší laboratoři byla roku 2000 vypracována diplomová práce, jejímž cílem bylo stanovení frekvence genové konverze mezi *GBA* a *GBAP*. Frekvence konverze byla měřena pomocí dvou technik založených na PCR: kvantitativní kompetitivní PCR s fluorescenční detekcí a

diluční metodou s radioaktivní detekcí. Byla stanovena frekvence jednotlivých mutací, které nese rekombinantní alela *RecΔ55* (mutace Δ55, D409H, L444P, A456P, V460V) v genu v souboru 10 vzorků mitotických a 10 vzorků meiotických buněk. Nejvyšší frekvence byla zaznamenána 1 : 200 kopií genomu u mutací L444P, A456P a V460V. S nižší frekvencí 1 : 2 000 až 1 : 40 000 byly detekovány mutace Δ55 a D409H (Mrázová 2000).

Frekvence meiotické genové konverze v HLA a PAR1 oblasti

Další studie se zabývala frekvencí genové konverze pouze u meiotických buněk, spermiích. Detekce probíhala ve 3 oblastech; dvě v HLA oblasti a třetí v pseudoautozomální oblasti pohlavního chromozomu PAR1. Nejprve se amplifikoval jeden haplotyp alelově specifickým PCR mimo hotspot. Poté proběhla hybridizace alelově specifickými oligonukleotidy, které hybridizují k sekvenci vnesené genovou konverzí nebo překřížením z jiného haplotypu. Frekvence genové konverze v úseku, kde docházelo k rekombinacím nejčastěji, byla udávána 1,3-3,4 × 10^{-3} na spermii. Frekvence překřížení byla nižší než frekvence genové konverze a byla stanovena na 0,9-1,2 × 10^{-3}. Frekvence konverze se snižovala se vzdáleností mezi homologními sekvencemi. Nejmenší vzdálenost byla 300 bází a největší 1 091 bází (Jeffreys and May 2004).

5.2 Nové metody pro zachycení vzácných mutací

Sekvenování nové generace

S rozmachem sekvenování nové generace (Next-Generation Sequencing, NGS) vzrůstají možnosti detekce vzácných mutací. Nyní je možné sekvenovat celý genom s relativně nízkými náklady a s malou časovou náročností. Díky RNA sekvenování je možné zkoumat nové sestřihové varianty nebo stanovit množství mRNA k určení genové exprese. Dále je možné určit interakce mezi DNA a proteiny nebo metylace DNA. NGS je dnes široce využíváno pro výzkum i DNA diagnostiku.

Chybovost NGS

Nevýhodou NGS je poměrně vysoká chybovost, která komplikuje zachycení vzácných mutací. Chyby nemusejí být způsobeny jen amplifikací, ale i při přípravě vzorku, fázemi před amplifikací, během cyklického sekvenování a analýzy obrazu. Polymeráza tvoří bodové mutace, které jsou způsobeny chybnou inkorporací komplementárního nukleotidu nebo „přeskokem“ polymerázy na jiný vysoce homologní templát. Konkrétně u platformy od firmy Illumina je chybně identifikováno přibližně 0,1 % bází (Manley, Ma and Levine 2016). Některé zdroje uvádějí rozmezí 0,05-1 % (Quail et al. 2008, Fuellgrabe et al. 2015). Záleží na

mnoha faktorech, např. na délce sekvence, použitých chemikáliích a typu detekovaných mutací. Dalším faktorem je typ použité platformy. Každá platforma má sklon k určitému typu mutací, platforma od firmy Illumina k substitucím, Oxford nanopórové sekvenování k delecím, 454 Roche, Ion Torrent a PacBio k indelovým mutacím (Fuellgrabe et al. 2015). Proto byla na úrovni biochemie nebo zpracovávání dat vyvinuta řada opatření zvyšujících přesnost sekvenování.

Unikátní značení molekul DNA - tagy

Ke zvýšení citlivosti sekvenování je možné použít unikátní značení molekul tzv. tagy a zjistit, jestli mutace vznikla chybou metody, nebo jestli jde o skutečnou mutaci.

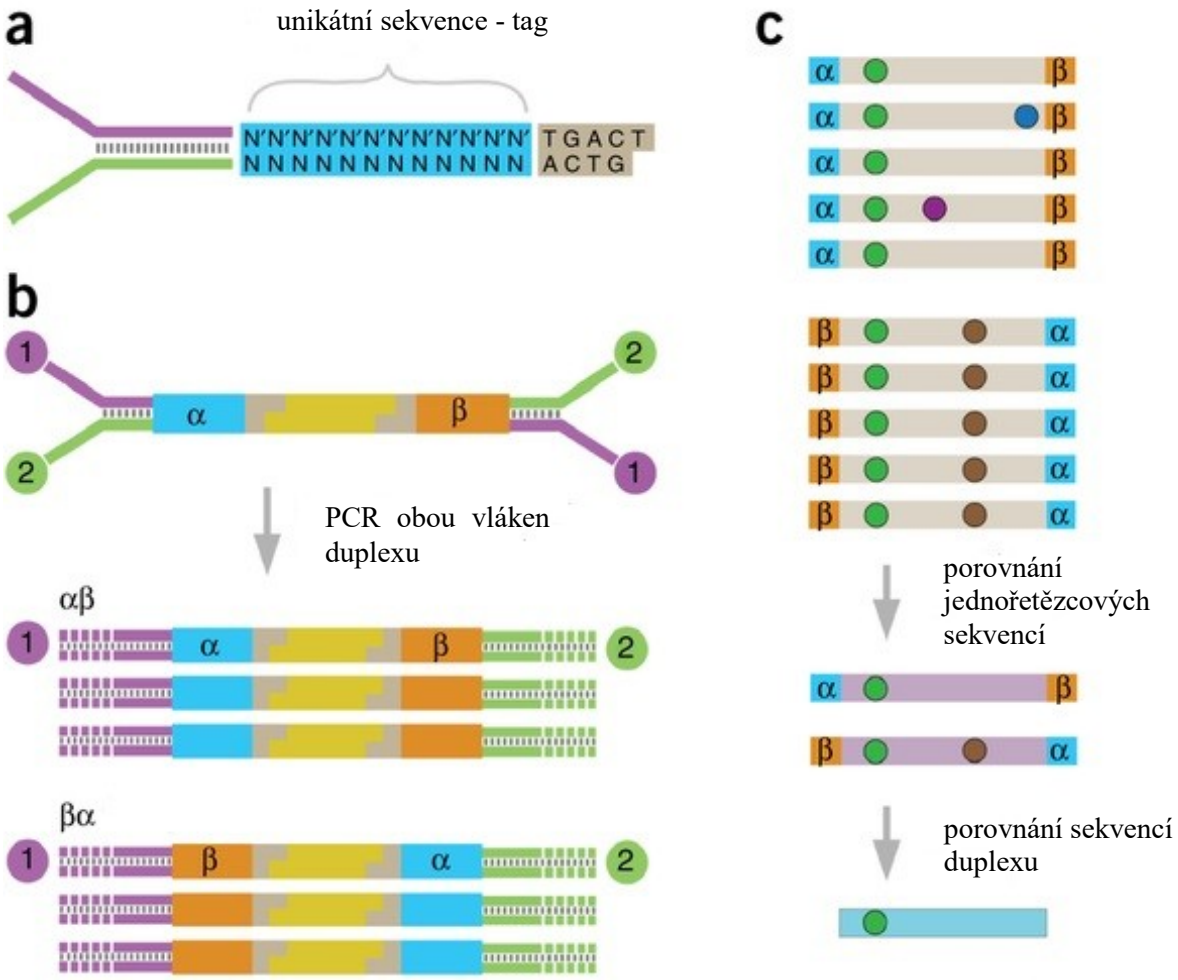
Unikátní sekvence – tag je přiřazen ke konkrétní molekule DNA: 1) jako součást primerů během lineární amplifikace nebo PCR, 2) jako součást primeru během reverzní transkripce nebo 3) ligací jako součást oligonukleotidů, které zpravidla nesou sekvence adaptérů pro sekvenování. Tagem se rozumí sekvence 10-14 náhodných nukleotidů (př. NNNNNNNNNN) nebo částečně degenerovaných nukleotidů (př. NNNNRNNYNN) (Kou et al. 2016). Počet náhodných nukleotidů záleží na počtu cílových molekul v templátu. V případě 10 náhodných nukleotidů je kapacita značení 4^{10}, což se rovná $1{,}05 \times 10^6$ unikátních sekvencí. Značené molekuly jsou obvykle následně amplifikovány v dalším kole pomocí PCR a produkty sekvenovány NGS. Výsledné sekvence jsou rozděleny do skupin podle shodných unikátních značení molekuly a mutace nalezené na jednotlivých molekulách jsou porovnány (Obr. 8). Pokud se určitá mutace nachází na všech nebo převažující většině porovnávaných molekul, jedná se s největší pravděpodobností o skutečnou mutaci, i když mutace v časných cyklech amplifikace může být přítomna na všech sekvencích nesoucích tag (tzv. jackpot mutace). Citlivost metody závisí na mnoha faktorech: na počtu cyklů v prvním kole amplifikace, na chybovosti užité polymerázy a na počtu cílů v templátu (Kinde et al. 2011).

U molekuly DNA značené tagy bylo zjištěno, že záměny vznikající během amplifikace jsou mnohem častější u substituce G>T než u reciproké záměny C>A. Důvodem je oxidativní poškození vzorku (Schmitt et al. 2012).

Duplexové sekvenování

Amplifikace značených molekul pomocí PCR může vést k falešně pozitivním výsledkům, jestliže vlivem polymerázy dojde k chybné inkorporaci nukleotidu již v prvním kole PCR. K rozpoznání takto vzniklých změn se užívá metoda duplexového sekvenování, která těží z komplementarity bází obou řetězců DNA (Obr. 8). NGS generuje data z jednovláknových DNA (ssDNA) fragmentů a sekvenace těchto fragmentů je limitující. Srovnáním sekvencí z individuálně označených amplikonů jednořetězcových fragmentů odvozených z jednoho řetězce dvouřetězcové DNA (dsDNA) je možné identifikovat většinu forem sekvenačních chyb. U této metody je udávána frekvence zachycení artefaktu $1 : 10^9$ osekvenovaných nukleotidů (Schmitt et al. 2012). Jiné zdroje uvádějí, že je možné detekovat jednu mutaci na $> 1 \times 10^7$ nukleotidů (Kennedy et al. 2014).

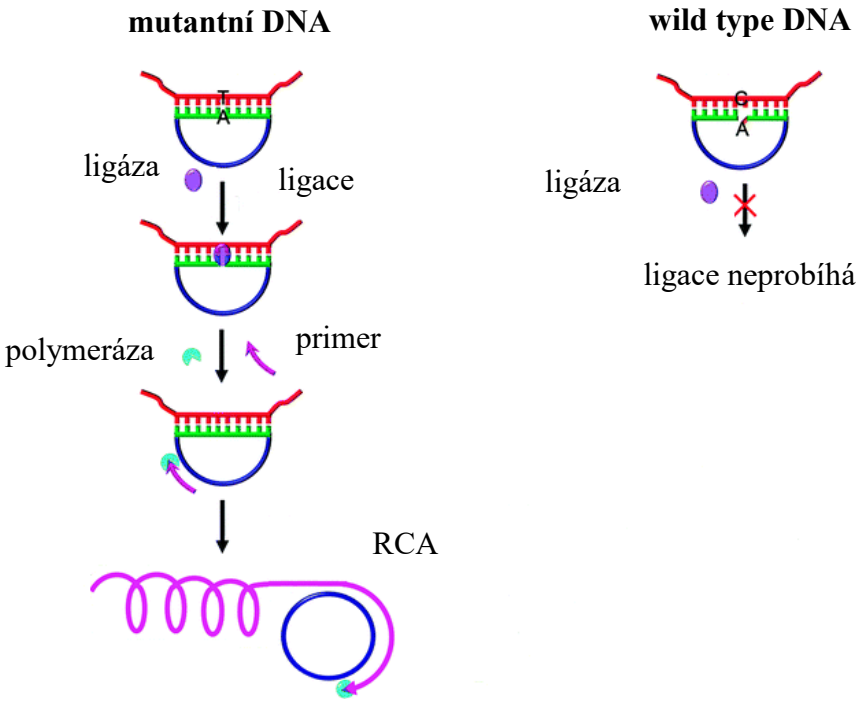
Jedinou možností u duplexové analýzy, kdy by nebyla odhalena mutace v důsledku chyby sekvenace, je v tom případě, že by se stejná mutace vyskytla u obou vláken na stejné pozici nukleotidu. Tato pravděpodobnost je ale velice nízká.



Obrázek 8: Duplexové sekvenování. A) Příprava adapterů s unikátními sekvencemi - tagy. Standardní sekvenační adapter pro Illuminu nese na 5´ konci jednoho vlákna 12 náhodných nukleotidů. Komplementární adapterová sekvence je následně prodloužena DNA polymerázou a vzniká dvouřetězcová komplementární sekvence, která je unikátní. **B)** Následuje ligace unikátně značených adapterů na naštěpenou dsDNA, která je takto individuálně značená na obou svých koncích. Vlákna jsou amplifikována z asymetrických primerových míst na adaptérovém konci a jsou podrobeny sekvenování párových konců. Každý PCR duplikát, který vznikne ze ssDNA nese unikátní sekvenci. **C)** Porovnáním sekvencí obou vláken z dsDNA umožňuje odhalení skutečné mutace od chyby metody. Pokud se jedná o chybu metody, tak se změna objeví jen na jednom vlákně DNA. Pokud se jedná o skutečnou mutaci, změnu uvidíme na obou vláknech. Výhodou duplexového sekvenování je možnost eliminace chyb i v prvním kole PCR. Převzato a upraveno (Kennedy et al. 2014).

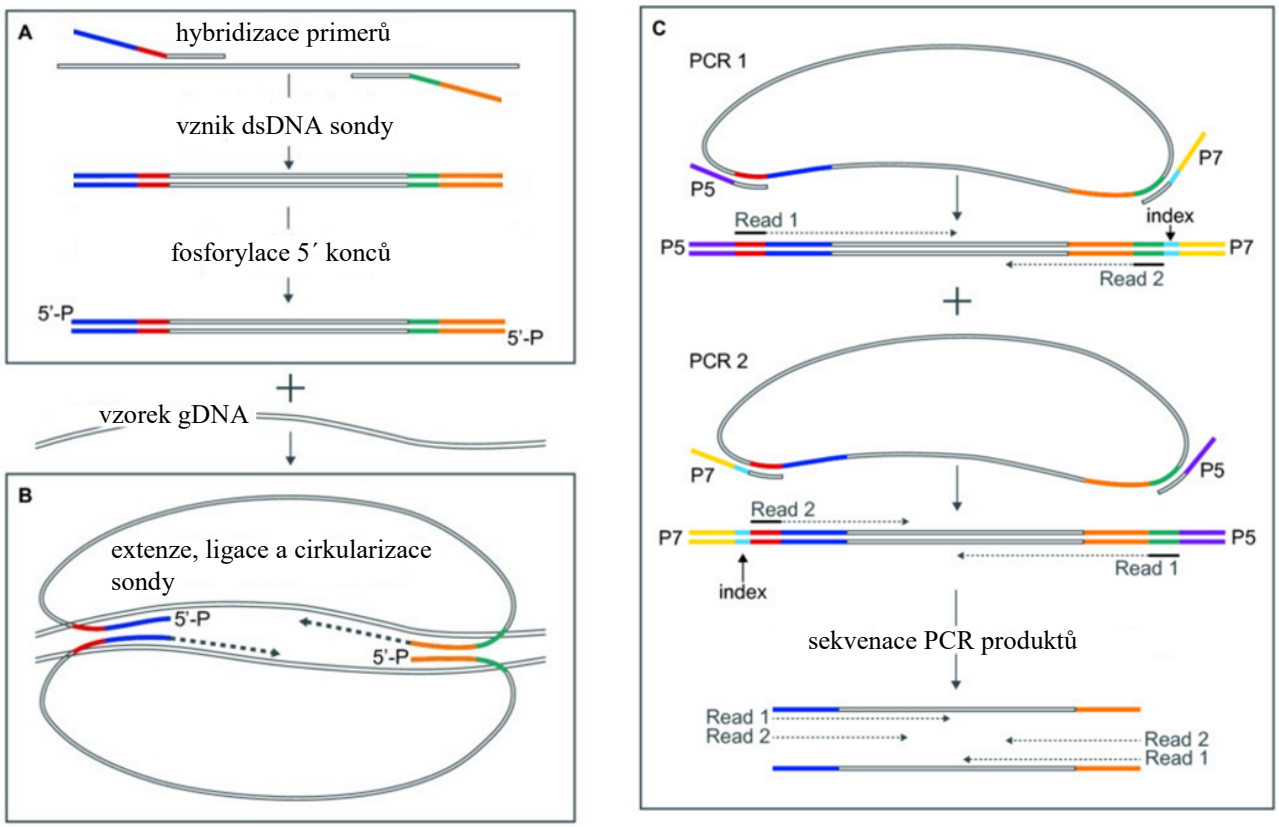
6 Zámkové sondy

Zámkové sondy jsou dlouhé oligonukleotidy, na jejichž koncích jsou specifické sekvence, které hybridizují k cílovým místům na stejném vlákně DNA. Mezi těmito sekvencemi se nacházejí univerzální amplifikační místa. 5´ konec sondy je fosforylovaný. V případě, že specifické sekvence hybridizují k jednomu vláknu DNA přímo vedle sebe, dojde mezi krajními nukleotidy na jejich 3´ a 5´ konci k ligaci a tím cirkularizaci sondy. Následně je možné tyto kruhové molekuly detekovat amplifikací pomocí mechanismu valivé kružnice (Rolling Circle Amplification, RCA), k čemuž slouží univerzální amplifikační místa. Této metody se mimo jiné využívá k zachycení vzácných mutací v populaci zdravých buněk. Metoda je schopná odhalit jednonukleotidovou záměnu, díky vysoké specifitě ligázové reakce. Jestliže je sekvence na 3´ konci sondy alelově specifická, k ligaci dojde pouze v případě hybridizace na tomto 3´ konci templátu. Metoda byla použita například k detekci jednobázových záměn u mitochondriální DNA (Obr. 9) (Larsson et al. 2004).



Obrázek 9: Použití zámkových sond při detekci jednobázových záměn. Zámková sonda s adeninem na 3´ konci je navržena pro detekci jednobázové záměny C>T. Zámková sonda je ligována a cirkularizována jen v případě mutantní DNA jako templátu. Následuje amplifikace RCA pomocí univerzálního primeru. RCA amplifikace je založená na tvorbě ssDNA pomocí polymerázy s tzv. dislokázovou aktivitou (odděluje řetězec DNA a nahrazuje jej nově syntetizovaným řetězcem). Wild type DNA není komplementární ke 3´ konci zámkové sondy a ligace neprobíhá. Převzato a upraveno z (Zhou et al. 2015).

V roce 2008 publikoval Krishnakumar práci, v níž využil vysoké specifity zámkových sond k obohacení vzorků DNA pro NGS o sekvence oblastí, které byly předmětem jejich zájmu (Krishnakumar et al. 2008). V jednom kroku bylo amplifikováno 5 471 exonů z 524 genů (Shen et al. 2011). Metodu nazval SMART (Spacer Multiplex Amplification Reaction). Specifické sekvence v tomto případě hybridizovaly ke genomové DNA (gDNA) v krajních oblastech intronů a byly od sebe vzdáleny 100-500 bází. Sonda byla proto prodloužena přidáním tzv. „linkeru“, což je asi 300 bází dlouhá sekvence, která není komplementární k lidské DNA a spojuje důležité koncové oblasti sondy. Obvykle se používá DNA sekvence z bakteriofága lambda. Po nasednutí zámkové sondy na jedno vlákno gDNA dochází nejprve k extenzi vlákna pomocí DNA polymerázy a poté v tom samém kroku k ligaci a k cirkularizaci sondy (Obr. 10). Necirkularizované sondy a gDNA jsou následně degradovány pomocí exonukleáz a kruhové molekuly jsou detekovány amplifikací pomocí PCR (Krishnakumar et al. 2008, Shen et al. 2013). V našem případě je vzdálenost specifických míst pro zachycení rekombinantních molekul dlouhá okolo 900 bází. Proto jsme navrhli novou modifikovanou metodu, kterou jsme nazvali ligační metoda.

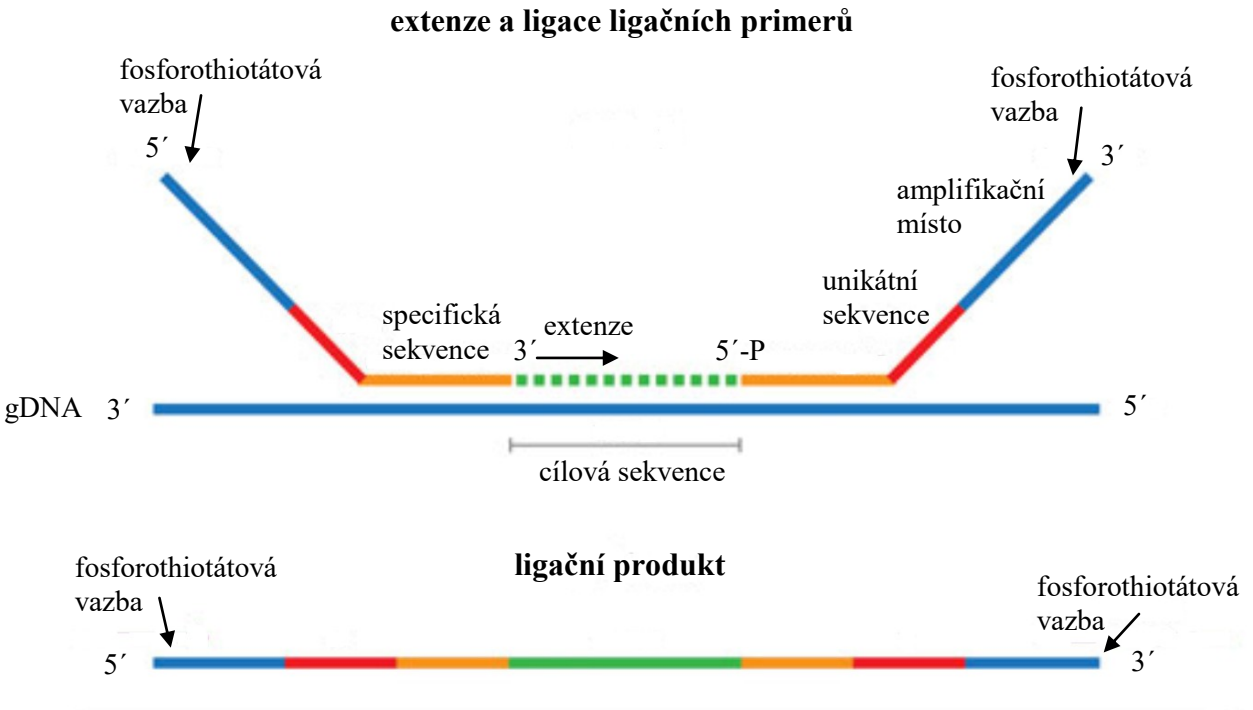


Obrázek 10: Použití zámkových sond k obohacení vzorků DNA, SMART. A) Příprava sondy. Pomocí dvou navržených primerů se amplifikuje dvouřetězcová sonda. Šedivá část primeru je komplementární k sekvenci např. bakteriofága lambda, modrá a oranžová část primeru je sekvenčně specifická ke genomové DNA a zelená a červená část jsou univerzální amplifikační místa. 5´ konce sondy jsou fosforylovány. **B) Vznik kruhové molekuly.** Ke vzorku gDNA se přidají vytvořené sondy, které nasednou na specifické sekvence DNA molekul. Dochází k extenzi vlákna ve směru šipek a následně k ligaci k 5´ konci sondy. **C) Amplifikace a příprava molekul na NGS.** Následuje PCR amplifikace. Primery pro PCR amplifikaci obsahují univerzální sekvenci komplementární k zámkovým sondám (šedě/červeně a šedě/zeleně) a sekvenci s indexy pro následné NGS. Převzato a upraveno z (Shen et al. 2013).

6.1 Ligační metoda

V naší laboratoři byla zavedena nová modifikovaná metoda, která spojuje výhody unikátních tagů, zámkových sond a metody SMART. Primery stejně jako u zámkových sond obsahují specifickou sekvenci komplementární k cílové sekvenci gDNA, univerzální sekvenci pro následnou amplifikaci a mezi tyto sekvence byl přidán 10 bází dlouhý tag (Obr. 11). Dále jsou primery upraveny tak, že pět nukleotidů na 5´ konci prvního primeru a na 3´ konci druhého primeru je spojeno fosforothiotátovou vazbou a 5´ konec druhého primeru je fosforylovaný. Fosforothiotátová vazba zabraňuje účinku exonukleáz a fosforylace 5´ konce umožňuje ligaci. V prvním kole po nasednutí primerů probíhá prodloužení prvního primeru DNA polymerázou

následované ligací s druhým primerem. Díky unikátnímu značení tagy je každé vlákno vzorku gDNA v tomto kroku jednoznačně identifikováno. Následuje odstranění nevyužitých primerů a gDNA směsí exonukleáz, stejně jako u zámkových sond. Degradaci ligačních produktů v našem případě nebrání cirkularizace, ale fosforothiotátová vazba na koncích modifikovaných primerů. Následuje PCR amplifikace pomocí univerzálních primerů a sekvenace NGS.



Obrázek 11: Vznik jednovláknového ligačního produktu. Ligační primery nasedající sekvenčně specifickou částí (oranžově) na cílovou oblast gDNA. Primery obsahují nehybridizující sekvenci s unikátní sekvencí – tagy (červeně) a univerzální amplifikační sekvencí (tmavě modře). Primery mají na vyznačených koncích 5 posledních nukleotidů spojeno fosforothiotátovou vazbou. Po nasednutí primerů na gDNA dochází k extenzi prvního primeru (zeleně) a následně k ligaci druhého primeru, který má 5´ konec fosforylovaný. Převzato a upraveno z (Hrdlickova, Lewis and Nehyba 2016).

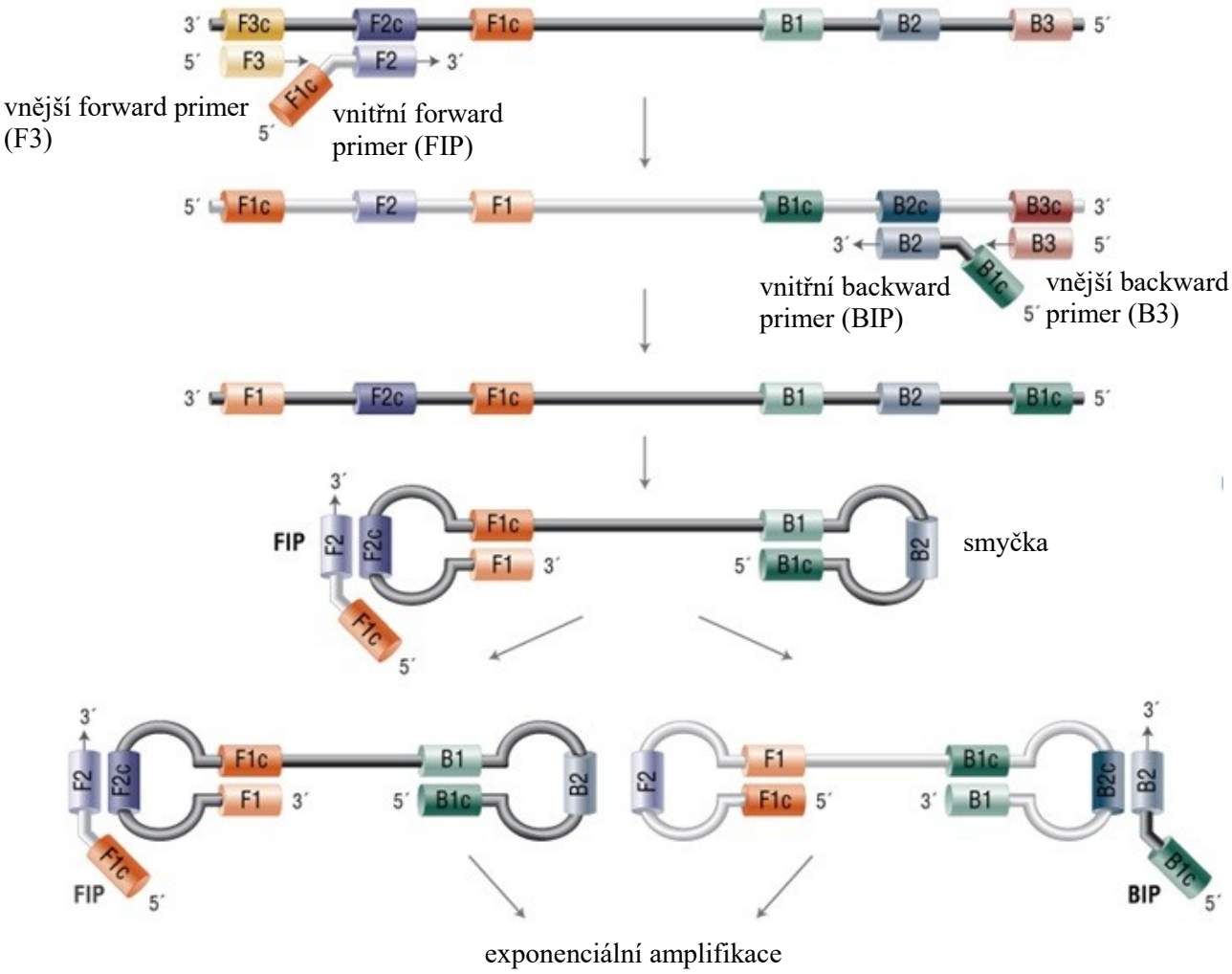
7 Asymetrická extenze

Metoda je založena na polymerázové reakci, při které je použit pouze jeden primer. Přítomnost jednoho primeru v reakci vede k tvorbě pouze jednoho vlákna. Délka vlákna závisí na době extenze a 3´ konec není jednoznačně ohraničen. Použitý primer je opět složený ze tří částí. Na svém 3´ konci nese specifickou sekvenci komplementární ke sledovanému úseku na gDNA, na 5´ konci nese univerzální amplifikační sekvenci nekomplementární ke gDNA a mezi těmito sekvencemi je unikátní 10 bázový tag. V případě pouze jednoho cyklu nasedání primeru a extenze vlákna je počet výsledných molekul menší nebo roven (v ideálním případě) počtu molekul gDNA. V případě následné amplifikace produktu asymetrické extenze, je každá molekula, která nese stejný tag, kopií stejného vlákna gDNA. V případě více cyklů je produkce vláken lineární nikoliv exponencionální, jako je tomu u obvykle používané PCR reakce se dvěma primery. Avšak informace o původu amplifikovaných molekul se ztrácí. Po následné PCR amplifikaci jsou sice všechny molekuly se stejným tagem kopií stejného vlákna gDNA, ovšem jedno vlákno gDNA mohlo být předlohou pro více molekul s rozdílným tagem v prvním kole.

8 Izotermální amplifikace zprostředkovaná smyčkou

Izotermální amplifikace zprostředkovaná smyčkou (Loop mediated isothermal amplification, LAMP) je vysoce specifická a efektivní amplifikační metoda, při které nedochází k teplotním krokům a cyklům. Amplifikace probíhá za konstantní teploty přibližně 65 °C. Standardně jsou používány 2 páry primerů, vnější forward a backward a vnitřní forward a backward (Obr. 12). Reakce probíhá pomocí DNA polymerázy s vysokou tzv. dislokázovou aktivitou, což znamená, že je schopna oddělit řetězec DNA a nahradit jej nově syntetizovaným řetězcem. Většinou se používá Bst DNA polymeráza. Množství kopií amplifikované DNA je u LAMP vyšší než u PCR. Produkt se detekuje na agarózovém gelu, kde je patrný neostře ohraničený pás DNA (smír) vznikající z důvodu velkého množství různě dlouhých fragmentů DNA. Dalším způsobem detekce produktu amplifikace může být přítomnost bílé sraženiny pyrofosforečnanu hořečnatého, která vzniká jako vedlejší produkt reakce (Mori et al. 2001). Výhodou této metody jsou nízké finanční náklady, rychlost, nižší citlivost k inhibitorům komplexních vzorků a možnost zpracování vzorků horší kvality. LAMP našlo široké využití v detekci cizorodé DNA i RNA, díky reverzní transkripci předcházející LAMP (Liu et al. 2017, Carter et al. 2017). Metoda je vysoce citlivá, je možné zachytit již 6 molekul cizorodé gDNA, respektive cDNA ve vzorku (Notomi et al. 2000). Metoda je také vysoce specifická

díky 6 úsekům DNA, které primery rozpoznávají. Nevýhodou je problematické navrhování primerů a nemožnost detekovat množství a velikost produktu amplifikace jako tomu je u PCR (Notomi et al. 2000).



Obrázek 12: Mechanismus izotermální amplifikace zprostředkované smyčkou. Prodlužováním vnitřního forward primeru (FIP) dochází k dislokaci řetězce dsDNA nově syntetizovaným řetězcem. Poté vnější forward primer (F3) dislokuje nově nasyntetizovaný řetězec DNA, který slouží jako templát pro druhý vnitřní backward (BIP) a vnější backward (B3) primer. Na obou koncích cílové sekvence se vytvoří smyčka z důvodu reverzně komplementárních sekvencí (F1 a B1). Vnitřní primery FIP a BIP nasedají na smyčku a iniciují další amplifikaci za vzniku složitých konkatemerů (Notomi et al. 2000). Převzato a upraveno z https://www.neb.com/applications/dna-amplification-and-pcr/isothermal-amplification.

Materiál a metody

9 Materiál

Vzorky biologického materiálu

Genomová DNA byla získána ze vzorků periferní krve 5 mužů a 5 žen (mitotické buňky) a ze vzorků spermatu 10 mužů (meiotické buňky). Jednalo se o nepříbuzné osoby, u kterých nebyla diagnostikována Gaucherova choroba. Dále byla získána gDNA z periferní krve od dvou nepříbuzných dárců s diagnózou Gaucherovy choroby, heterozygoti pro mutace *RecΔ55/*N370S a *RecNciI/*N370S. Vzorky byly získány z laboratoře Ústavu dědičných metabolických poruch a Sexuologického ústavu VFN a 1. LF.

Použité pufry a média

Fosfátový pufr (PBS) – 37 mM NaCl, 2,7 mM KCl, 10 mM Na_2HPO_4, 1,8 mM K_2HPO_4, pH 7,4

Borátový pufr (BB) – 100 mM $Na_2B_4O_7 \cdot 10H_2O$

TBE pufr - 89 mM Tris, 89 mM H_3BO_3, 2 mM EDTA

NEB2 pufr – 50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 100 μg/ml BSA

TE pufr – 10 mM Tris-Cl, 1 mM EDTA (pH 8)

Vázací pufr – 0,1 M TrisEDTA (pH 7,5), 4 M NaCl

Promývací pufr – 1 M TrisEDTA (pH 7,5), 4 M NaCl

LB médium – 1% pepton, 0,5% kvasničný autolyzát, 1% NaCl

SOC médium – 2% trypton, 0.5% kvasničný autolyzát, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM glukóza

Lyzační roztok – 50 mM Tris, 100 mM $MgSO_4 \cdot 7H2O$, 100 mM KCl, 10 mM DTT

Komerční soupravy

TOPO® XL PCR Cloning Kit (Invitrogen)

Kompetentní buňky *E.coli* TOP10 (Invitrogen)

FastPlasmid Mini-Prep Kit (5 Prime)

High Pure PCR Product Purification Kit (Roche)

Streptavidine Magnetic Particles (Roche)

QIAamp DNA Blood Mini Kit (QIAGEN)

SeqCap Pure Capture Bead Kit (Roche)

Nextera® XT Library Prep Kit (Illumina)

KAPA Library Quantification Kit (Kapa Biosystems)

Použitá laboratorní technika

Automatický sekvenátor ABIprism A3100 (Life Technologies)

Centrifuga (Sigma 4K15)

Cyklér (Bio-Rad ALS 1296)

Elektroforetický přístroj (Bioanalyzer Agilent)

Gelové dokumentační zařízení (Syngene)

Inkubátor (Biological termostat BT 120 M)

Laminární box (Holten Lamin Air)

Minicentrifuga (Eppendorf 5415D)

Předvážky (Scout SC 2020)

Sekvenátor (Illumina MiSeq)

Spektrofotometr (Nanodrop Technologies ND 1000)

Souprava pro agarózovou elektroforézu (Scie-Plas)

Souprava pro polyakryamidovou elektroforézu (Hoefer)

Termoblok (Techne Dri-Block DB-3)

Třepačka (Schoeller)

UV-transluminátor (Vilber Lourmat)

Vortex (Labnet)

Zdroj napětí (Consort EV 265)

Použitý software

Sequencing Analysis 5.1

Microsoft Excel

Primer explorer V5 (https://primerexplorer.jp/e/)

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Samtools v. 1.4 (https://github.com/samtools/samtools)

bwa v. 0.7.15 (Burrows-Wheeler aligner, http://bio-bwa.sourceforge.net/)

IGV 2.4 (http://software.broadinstitute.org/software/igv/)

Seznam primerů

Tab. I: Seznam použitých primerů a jejich sekvencí

Označení primeru	Sekvence
FS-S-3	5′-AACCATGATTCCCTATCTTC-3′
FS-S-5	5′-TTGCATTCTTCCCGTCACCCACCTC-3′
FS-A-3dl	5′-CCCAAGACTGGTTTTTCTACACACATGCA-3′
Biot_Lig_FS-S-3	5′- /biot/CTCGAGGTAAAACGACGGCCAGTGNNNNNNNNNNAACCATGATTCCCT-3′
L444PplusM13R	5′-CAGGAAACAGCTATGACGGATGCATCAGTGCCACTGCGTACG-3′
L444PminusM13R	5′-CAGGAAACAGCTATGACGGATGCATCAGTGCCACTGCGTACA-3′
PSG-S2	5′-GGGGGTGGTAGCTCATGGCTAT-3′
PSG-S3	5′-AACCATGATTCCCTGTCTTG-3′
PSG-A2	5′-CCCAAGACTGGTTTTTCTACTCTCATGACT-3′
Lig5S_FS-S-3	5′-C*A*T*G*T*AAAACGACGGCCAGTGNNNNNNNNNNAACCATGATTCCCTATCTTC-3′
Lig5S_del55	5′-C*A*T*G*T*AAAACGACGGCCAGTGNNNNNNNNNNCTGGACCGACTGGAACCTTGC-3′
Lig3S_L444P	5′-/5Phos/CGGACGCAGTGGCACTGATGCATCCNNNNNNNNNNGTCATAGCTGTTTCC*T*G*T*A*C-3′
Lig3A_L444P	5′-/5Phos/GGGTCGTTCTTCTGACTGGCAANNNNNNNNNNCACTGGCCGTCGTTTT*A*C*T*A*C-3′
Lig5A_FS-A-3	5′-C*A*T*C*A*GGAAACAGCTATGACNNNNNNNNNNGGTTTTTCTACTCTCATGCA-3′
Ampli1	5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTAAAACGACGGCCAGTG-3′
Ampli2	5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGGAAACAGCTATGAC-3′
Lig_L444P_sense	5′-CTGGTTGCCAGTCAGAAGAACGACC-3′
Del55S_M13S_biot	5′-/biot/ACCGGAATTCGTAAAACGACGGCCAGTGNNNNNNNNNNTTGGGTGCGTAACTTGTCG-3′
GBAe10_read2	5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCCCAGACCTCACCATTGC-3′
F3_LAMP_Uset	5′-TGAACCCCGAAGGAGGAC-3′
FIP_LAMP_Uset	5′-GGTGGTAGAACATGGGCTGTTTGTNNNNNNGTGCGTAACTTTGTCGACAGT-3′
B3_LAMP_Uset	5′-AGCAGAGCCATCGGGATG-3′
BIP_LAMP_Uset	5′-TCAGCAAGTTCATTCCTGAGGGCNNNNNNCATCAGTGCCACTGCGTACG-3′
F3_LAMP_Lset	5′-GAGGGCTCCCAGAGAGTG-3′
FIP_LAMP_Lset	5′-GACCTCACCATTGCCCTCACCNNNNNNCCAGTCAGAAGAACGAACC-3′
B3_LAMP_Lset	5′-TGAGTCACCCAAACCATTGC-3′
BIP_LAMP_Lset	5′-AGCCTGGGCATTAAAGGGACAGNNNNNNAAGCTCACACTGGCCCAGT-3′

Vysvětlivky k Tab. I: * fosforothiotátová vazba, /5Phos/ fosforylovaný 5′ konec, /biot/ biotinylovaný konec, N náhodný nukleotid

10 Metody

Izolace gDNA z periferní krve a ze spermatu

Genomová DNA z periferní krve a ze spermatu byla izolována pomocí kitu QIAamp DNA Blood Mini Kit (QIAGEN). Sperma 400 µl bylo centrifugováno 10 min při 300 × g za laboratorní teploty. Supernatant byl odstraněn a peleta byla resuspendována v 200 µl PBS. Dále byl postup stejný pro krev i resuspendované sperma. Ke 200 µl vzorku bylo přidáno 20 µl protézy a 200 µl lyzačního AL pufru. Poté byl vzorek inkubován 10 min při 56 °C v termobloku. Po inkubaci bylo ke vzorku přidáno 200 µl 96% etanolu. Vzorek byl přenesen na kolonku a centrifugován při 6 000 × g po dobu 1 min, eluát byl odstraněn. Poté byla kolonka 2krát promyta 500 µl pufry AW1 a AW2, pokaždé byla centrifugována 1 min při 6 000 × g a eluát byl odstraněn. Nakonec byla centrifugována 3 min na sucho a poté přemístěna do čisté zkumavky. Na kolonku bylo naneseno 50 µl elučního AE pufru a vzorek byl centrifugován 1 min při 6 000 × g. Eluát byl uschován a jeho koncentrace byla změřena na spectrofotometru (Nanodrop Technologies ND 1000).

Purifikace gDNA pomocí etanolového srážení

Genomová DNA byla v případě štěpení endonukleázami přečištěna etanolovým srážením. K vodnému roztoku gDNA byla přidána 1/10 objemu 3 M octanu sodného a poté 3krát více objemu 96% etanolu (tzn. k 100 µl gDNA bylo přidáno 10 µl 3 M octanu sodného a poté 300 µl 96% etanolu). Poté byla směs opatrně promísena a inkubována 30 min při -80 °C. Následovala centrifugace 20 min při 14 000 × g při +4°C. Supernatant byl odstraněn. Peleta gDNA byla promyta 500 µl ledového 70% etanolu, vysušena na vzduchu a rozpuštěna v 80 µl TE pufru (pH 8).

10.1 Elektroforéza

Horizontální agarózová elektroforéza

K separaci a následné vizualizaci fragmentů DNA byl použit 1% agarózový (SERVA) gel s fluorescenčním barvivem 1× GelGreen (Biotinum) v 1× borátovém pufru (BB). Gel byl umístěn do elektroforetické komory s 0,5× BB pufrem. Do jamek v gelu bylo naneseno 5 µl vzorku smíchaného s 2 µl 6× vzorkovacím pufrem. Do poslední jamky bylo naneseno 5 µl velikostního standardu GeneRuler 100 bp Plus (Fermentas). Komora byla připojena ke zdroji

elektrického napětí, elektroforéza probíhala při napětí 10 V/cm. Vizualizace probíhala pomocí UV-transluminátoru (Vilber Lourmat).

Vertikální polyakrylamidová elektroforéza

Pro detekci velmi krátkých úseků DNA (>100 bp) byla použita polyakrylamidová elektroforéza (PAGE). Složení 12% polyakrylamidového gelu je uvedeno v Tab. II. Persíran amonný (APS) a N,N,N′,N′- tetramethylethylendiamin (TEMED), které v součinnosti způsobují polymeraci akrylamidu (AA), byly přidány až jako poslední. Směs byla urychleně nalita do připravené aparatury, tloušťka gelu byla 1 mm a gel byl ponechán 2 hodiny při pokojové teplotě. Po polymeraci gelu byla aparatura umístěna do vertikální elektroforetické komory a zalita 1× TBE pufrem. Vzorky o 24 µl byly smíchány s 5 µl 6× vzorkového pufru, poté byly denaturovány 5 min při 95 °C a přemístěny na led. Do promytých jamek byly naneseny vzorky a do poslední jamky bylo naneseno 5 µl velikostního standardu GeneRuler 50 bp (Fermentas). Komora byla připojena ke zdroji elektrického napětí. Elektroforéza probíhala 1 hod při napětí 70 V, poté bylo napětí zvýšeno na 120 V po dobu 2 hod a 10 min. Aparatura byla rozebrána a gel po obarvení v roztoku 3× GelRed (Biotinum) po dobu 3 min byl vizualizován pomocí UV-transluminátoru.

Tab. II: Složení polyakrylamidového gelu

Složení polyakrylamidového gelu	Množství
urea	4,8 g
40% AA (29 : 1)	3,125 ml
10x TBE	1 ml
10% APS	42 µl
TEMED	6,5 µl
H_2O	1 ml

10.2 Purifikace PCR produktů

Purifikace pomocí kolonek se skelným papírem

PCR produkty byly přečištěny pomocí komerční soupravy High Pure PCR Product Purification Kit (Roche) dle protokolu. Směs 100 µl PCR produktu (v případě menšího množství PCR produktu, byl vzorek doplněn vodou do 100 µl) a 500 µl vázacího pufru byla nanesena na kolonku se skelným papírem a centrifugována 1 min při 14 000 × g, eluát byl odstraněn. Poté byla kolonka 2krát promyta 500 µl a 200 µl promývacího pufru a pokaždé centrifugována při 14 000 × g, eluát byl odstraněn. Nakonec byla kolonka centrifugována

naprázdno 3 min při 14 000 × g. Poté byla kolonka přenesena do čisté 1,5 ml zkumavky, na kolonku bylo naneseno 50 µl elučního pufru a kolonka byla inkubována 3 min při pokojové teplotě. Kolonka byla centrifugována při 14 000 × g a koncentrace vzorku změřena na spektrofotometru.

Purifikace pomocí magnetických streptavidinových částic

Příprava magnetických streptavidinových částic

K 40 µl magnetických streptavidinových částic v 1,5 ml zkumavce byl přidán 1 ml vázacího pufru a obsah zkumavky byl lehce promíchán. Poté byla zkumavka umístěna do magnetického stojánku, magnetické částice se shlukly na stěnu zkumavky a pufr byl odpipetován. Promývání proběhlo celkem 3×. Nakonec bylo k částicím přidáno 100 µl vázajícího pufru.

Purifikace PCR produktu

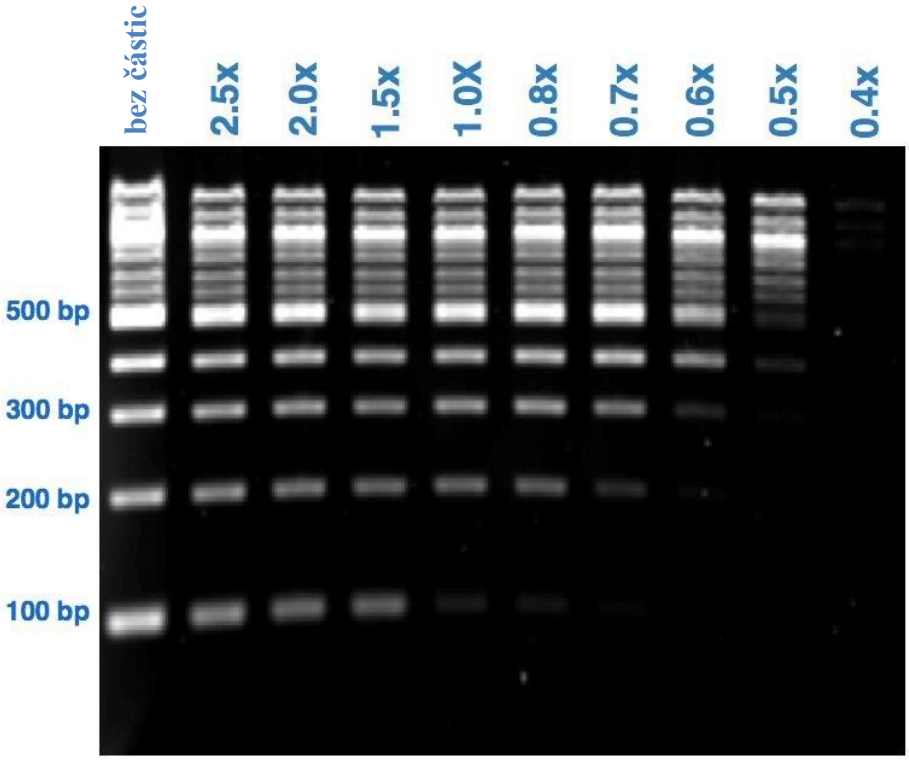
K 25 µl PCR produktu přečištěného přes kolonku bylo přidáno 25 µl magnetických streptavidinových částic ve vázacím pufru, směs byla ponechána 30 min za občasného lehkého promíchání při laboratorní teplotě. Poté byla zkumavka umístěna do magnetického stojánku a roztok byl odstraněn. Následně byly částice 2× promyty 1 ml promývacího pufru pomocí magnetického stojánku, poté 2× 1 ml TE pufru. Nakonec bylo k magnetickým částicím s navázaným PCR produktem přidáno 25 µl vody.

Purifikace pomocí magnetických částic

Výsledný PCR produkt byl purifikován pomocí komerční soupravy SeqCap Pure Capture Bead Kit (Roche), AMPure XP Beads. K 100 µl PCR produktu bylo přidáno 68 µl magnetických částic v polyetynelglykolu. Směs byla promíchána a ponechána 5 min při laboratorní teplotě. Poté byla zkumavka umístěna do magnetického stojánku, kde byla ponechána 10 min. Roztok byl odstraněn a magnetické částice shluklé na stěně zkumavky byly 2× promyty 200 µl 70% etanolu. Při promývání zůstávala zkumavka v magnetickém stojánku. Pro úplné odstranění etanolu byla zkumavka vložena na 7 min do inkubátoru při 37 °C. Nakonec bylo k magnetickým částicím přidáno 40 µl vody, do které se uvolnil purifikovaný PCR produkt a roztok byl pomocí magnetického stojánku přenesen do čisté zkumavky.

U tohoto typu purifikace je možná selekce na základě délky fragmentů DNA. Selekce je založena na poměru objemu PCR produktu ku objemu polyetylenglykolu, ve kterém jsou

magnetické částice. Z důvodu použití primerů delších než 50 bází vznikaly po PCR dimery dlouhé 100-200 bází, které již nebylo možné odstranit přes klasické purifikační metody jako je purifikace přes kolonku se skelným papírem. Na základě Obr. 13 jsme se rozhodli pro poměr objemu PCR produktu a objemu magnetických částic v polyetylenglykolu 1 : 0,68.



Obrázek 13: Selekce fragmentů DNA podle délky pomocí magnetických částic. V horní části obrázku - násobek objemu PCR produktu a polyetylenglykolu, v levé části obrázku - délka fragmentů DNA po purifikaci. Převzato a upraveno z https://www.broadinstitute.org/genome-sequencing/broadillumina-genome-analyzer-boot-camp.

10.3 Příprava standardů - pozitivních a negativních kontrol použitých metod

PCR amplifikace DNA úseku genu a pseudogenu pro GBA

Úseky genu a pseudogenu pro GBA byly amplifikovány ve dvou PCR produktech P5 (2232, gen) a PSG2 (2826 bází, pseudogen). Jako templát byla použita gDNA zdravého jedince a gDNA dvou složených heterozygotů pro mutace (*RecΔ55/*N370S*, RecNciI/*N370S). Primery byly převzaty z literatury a délka antisensového primeru FS-A-3dl byla upravena s ohledem na teplotu nasedání. U všech produktů byla použita High-Fidelity Q5 polymeráza s 3´-5´ exonukleázovou aktivitou. Výrobcem udávaná přesnost inkorporace komplementárních nukleotidů v řetězci je 280× vyšší než u Taq DNA polymerázy (Potapov and Ong 2017). Složení reakční směsi pro amplifikaci aktivního genu udává Tab. III a teplotní program Tab. IV. Složení reakční směsi pro amplifikaci pseudogenu udává Tab. V a teplotní program Tab. VI.

Tab. III: Složení reakční směsi při amplifikaci aktivního genu

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	4 ng/µl	1
Q5 pufr	1×	5
dNTPs	0,2 mM	2,5
FS-S-5	0,4 µM	1
FS-A-3dl	0,4 µM	1
Q5 polymeráza	0,4 U	0,2
H_2O		13,3
celkem		25

Tab. IV: Teplotní program

Teplota	Čas	Cykly
95 °C	3 min	
95 °C 66 °C 72 °C	10 s 20 s 2 min	35 cyklů
72 °C	5 min	

Tab. V: Složení reakční směsi při amplifikaci pseudogenu

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	4 ng/µl	1
Q5 pufr	1×	5
dNTPs	0,2 mM	2,5
PSG-S2	0,4 µM	1
PSG-A2	0,4 µM	1
Q5 polymeráza	0,4 U	0,2
H_2O		13,3
Celkem		25

Tab. VI: Teplotní program

Teplota	Čas	Cykly
95 °C	3 min	
95 °C 62 °C 72 °C	10 s 20 s 2 min	35 cyklů
72 °C	5 min	

10.4 Klonování PCR produktu

PCR produkty P5 zdravého jedince, P5 heterozygotních pacientů a PSG2 zdravého jedince sloužily k zaklonování těchto úseků do vektoru TOPO XL. Klonované úseky (P5, P5-*RecΔ55*, *RecNciI*, PSG2) byly následně použity jako pozitivní a negativní kontroly.

Prodloužení 3´ konců PCR produktu

Klonování PCR produktů do vektoru TOPO XL je založeno na TA klonování pomocí kovalentně vázané topoizomerázy vyžadující přesahující nukleotid adeninu na 3´ konci PCR produktu (Geng et al. 2006). Q5 polymeráza s 3´-5´ exonukleázovou aktivitou odstraňuje přesahující 3´ konce, proto je nutné prodloužit 3´ konce PCR produktu o několik deoxyadeninových zbytků pomocí terminálně transferázové aktivity Klentaq polymerázy. Složení reakční směsi udává Tab. VII.

Tab. VII: Složení reakční směsi při prodlužování 3´ konců PCR produktů

Reakční směs	Konečná koncentrace	Objem (µl)
PCR produkt		4
PC2 pufr	1×	1
dATPs	0,2 mM	2
Klentaq	2,5 U	0,5
H_2O		2,5
celkem		10

Reakce probíhala 30 min při 70 °C.

PCR produkt s prodlouženými 3´ konci byl následně přečištěn přes kolonku High Pure PCR Product Purification Kit (Roche).

Ligace

Směs 2 µl přečištěného PCR produktu s prodlouženými 3´ konci a 0,5 µl komerčního vektoru TOPO XL (Invitrogen) byla inkubována 20 min při pokojové teplotě. Reakce byla zastavena přidáním 0,5 µl Stop solution.

Transformace buněk

Transformace chemicky kompetentních buněk proběhla teplotním šokem. *E.coli* TOP 10 kompetentní buňky 4 × 50 µl byly rozmraženy na ledové lázni. Opatrně byla přidána ligační směs a buňky byly ponechány dalších 30 min na ledové lázni. Poté byly buňky rychle zahřáty ve vodní lázni na 42 °C po dobu 30 s a opět vráceny do ledové lázně minimálně po dobu 2 min. Následoval přídavek 130 µl SOC média předehřátého na 37 °C. Inkubace probíhala v 37 °C po dobu 1 hod při třepání 225 rpm.

Selekce pozitivních buněčných klonů

Po inkubaci byly transformované buňky naneseny na misky s LB agarem a kanamycinem (0,1%). Vektor TOPO XL nese selekční marker pro rezistenci k tomuto antibiotiku. Buňky byly inkubovány přes noc při 37 °C.

Selekce pozitivních buněčných kolonií byla provedena pomocí PCR amplifikujícího vkládaný úsek DNA. Složení reakční směsi jsou uvedeny v Tab. III a V, teplotní programy jsou uvedeny v Tab. IV a VI.

Kolonie buněk byly jemně nabrány párátkem a vloženy do připravené očíslované zkumavky se směsí pro PCR, poté byly buňky na párátku naneseny na novou očíslovanou misku

s LB agarem a kanamycinem (rozčárkování), kde byly inkubovány přes noc při 37 °C. PCR produkty z kolonií byly vizualizovány na 1% agarózovém gelu a u pozitivních kolonií byly z rozčárkováné misky založeny tekuté kultury, 4 ml tekutého LB média s kanamycinem inkubované přes noc při 37 °C a třepání při 225 rpm v 10 ml zkumavkách.

Izolace plazmidové DNA

Pro izolaci plazmidů byla použita komerční souprava FastPlasmid Mini Prep Kit (5 Prime).

Zkumavky s buněčnou kulturou byly centrifugovány 7 min při 3 000 rpm. Supernatant byl odstraněn a buněčná peleta byla resuspendována ve 400 µl ledového lyzačního roztoku, přenesena do 1,5 ml zkumavky, ve které byla promíchána po dobu 30 s a poté ponechána 3 min při laboratorní teplotě. Buněčný lyzát byl přenesen na kolonku a centrifugován 1 min při maximálních otáčkách, eluát byl odstraněn. Kolonka byla promyta 400 µl promývacího pufru, centrifugována 1 min při maximálních otáčkách, eluát byl odstraněn. K odstranění zbytků promývacího roztoku následovala centrifugace prázdné kolonky 1 min při maximálních otáčkách. Poté byla kolonka přenesena do nové 1,5 ml zkumavky a plazmidová DNA byla uvolněna 50 µl elučního pufru a centrifugací po dobu 1 min. Koncentrace izolované plazmidové DNA byla změřena na spektrofotometru.

Sekvenování a analýza plazmidů

K sekvenaci byl použit primer odpovídající vloženému úseku DNA; pro úsek P5 primery FS-S-S5 a FS-A-3dl, pro úsek PSG2 primery PSG-S3 a PSG-A2. Sekvenování proběhlo na obou vláknech vloženého úseku.

Vzorky byly sekvenovány na automatickém sekvenátoru ABIprism A3100 (Life Technologies). Sekvenaci prováděla diagnostická laboratoř Ústavu dědičných metabolických poruch. Analýza sekvencí probíhala v programu Sequencing Analysis 5.1. Sekvence vložených úseků P5 a PSG2 byly porovnávány se sekvencemi uvedenými v databázi NCBI. Plazmidy nesoucí bezchybné sekvence aktivního genu a pseudogenu pro GBA a rekombinantního genu *RecΔ55* a *RecNciI* byly použity jako pozitivní a negativní kontroly následných metod.

Uchování vybraných kultur

Současně byla vytvořena zásoba transformovaných buněk s genem a pseudogenem pro GBA a rekombinantní alelou *RecΔ55* a *RecNciI*. Čerstvá buněčná kultura 300 µl byla přenesena do 1,5 ml zkumavky s 300 µl směsi 65% glycerolu, 0,1 M $MgSO_4$, 0,025 M Tris/HCl pH 8 a byla

opatrně promísena. Poté byla celá směs prudce zchlazena v lázni suchého ledu a etanolu a uchována v -80 °C.

10.5 Asymetrická extenze

10.5.1 Příprava templátu pro NGS - bez obohacení konvertovanými sekvencemi

Extenze primeru – 1. kolo

Složení reakční směsi při extenzi primeru udává Tab. VIII a teplotní program Tab. IX.

Tab. VIII: Složení reakční směsi při extenzi primeru – 1. kolo

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	5,6 ng/µl	2
Q5 pufr	1×	5
dNTPs	2 mM	2,5
Del55S_M13S_biot	0,08 µM	0,2
Q5 polymeráza *	0,4 U	0,2
H_2O		15,1
celkem		25

* Q5 polymeráza byla přidána do jednotlivých vzorků až po nasednutí primerů

Tab. IX: Teplotní program

Teplota	Čas	Cykly
95 °C	5min	
72 °C – 1°C/cyklus	30 s	12 cyklů
60 °C * přídavek Q5 polymerázy 72 °C	45 min 50 s	

* Vzorky byly přeneseny do ledové lázně a ke každému bylo přidáno 0,4 U Q5 polymerázy, poté byly vráceny zpět do termocykléru na 50 s při 72 °C (extenze).

Primer Del55S_M13S_biot je biotinylovaný na 5´ konci a skládá se ze tří částí: univerzální sekvence M13S na 5´ konci, specifická sekvence komplementární k sekvenci aktivního genu v místě 55 bázové delece na 3´ konci a tagu 10 bází mezi těmito sekvencemi.

Produkt 1. kola, ssDNA, byl přečištěn přes kolonku se skelným papírem pomocí komerční soupravy High Pure PCR Product Purification Kit (Roche). Metoda je založena na vazbě nukleových kyselin na skelný papír v přítomnosti chaotropních solí (guanidinium HCl). Při purifikaci dojde k odstranění solí a enzymů přítomných v PCR reakci a k odstranění krátkých DNA fragmentů (primery, >100 bp). Genomová DNA není odstraněna.

PCR – 2. kolo

Složení reakční směsi při PCR – 2. kolo udává Tab. X a teplotní program Tab. XI.

Tab. X: Složení reakční směsi při PCR – 2. kolo

Reakční směs	Konečná koncentrace	Objem (µl)
Přečištěný produkt 1. kola		5
Q5 pufr	1×	8
dNTPs	0,2 mM	4
Ampli1	0,04 µM	0,1
GBAe10_read2	0,04 µM	0,1
Q5 polymeráza	0,4 U	0,2
H_2O		22,6
celkem		40

Tab. XI: Teplotní program

Teplota	Čas	Cykly
95 °C	3 min	
95 °C 60 °C 72 °C	10 s 20 s 1,5 min	35 cyklů
72 °C	5 min	

Primer Ampli1 je na 3´ konci komplementární k sekvenci M13S a na 5´ konci nese sekvenci komplementární k sekvenačním primerům NGS Illumina. Primer GBAe10_read2 je na 3´ konci reverzně komplementární k sekvenci aktivního genu v oblasti desátého exonu a na 5´ konci komplementární k sekvenačním primerům NGS Illumina. PCR produkt je 733 bází dlouhý a zahrnuje celý úsek potenciální *RecTL* alely.
PCR produkt byl vizualizován na 1% agarózovém gelu.

PCR produkt byl přečištěn pomocí komerční soupravy magnetických částic AMPure XP Beads v polyetylenglykolu (Roche). Poměr objemu PCR produktu a objemu magnetických částic v polyetylenglykolu byl 1 : 0,68.

Délka a kvalita fragmentů byla stanovena na fluorescenčním přístroji Bioanalyzer Agilent, který pracuje na principu kapilární elektroforézy (Obr. 15).

10.5.2 Příprava templátu pro NGS - s obohacením konvertovanými sekvencemi

Lineární PCR - 1. kolo

Složení reakční směsi při lineárním PCR udává Tab. XII a teplotní program Tab. XIII.

Tab. XII: Složení reakční směsi při lineárním PCR – 1. kolo

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	5,6 ng/µl	2
PC2 pufr	1×	2,5
dNTPs	0,2 mM	2,5
Biot_Lig_FS-S-3	0,6 µM	1,5
Klentaq polymeráza	0,75 U	0,15
H_2O		16,35
celkem		25

Tab. XIII: Teplotní program

Teplota	Čas	Cykly
95 °C	5 min	
95 °C 60 °C 72 °C	10 s 20 s 2 min	30 cyklů
72 °C	5 min	

Primer Biot_Lig_FS-S-3 je biotinylovaný na 5´ konci a skládá se ze tří částí: univerzální sekvence M13S na 5´ konci, specifická sekvence komplementární k sekvenci aktivního genu v oblasti intronu 8 na 3´ konci a tagu 10 bází mezi těmito sekvencemi.

Produkt 1. kola, ssDNA, byl přečištěn nejprve přes kolonku se skelným papírem pro odstranění nevyužitých primerů a poté přečištěn pomocí streptavidinových magnetických částic pro odstranění gDNA.

Přečištěný produkt 1. kola navázaný na streptavidinové částice sloužil jako templát v 2. kole při amplifikaci aktivního genu *GBA*.

PCR – 2. kolo

Složení reakční směsi při PCR – 2. kolo udává Tab. XIV a teplotní program Tab. XV.

Tab. XIV: Složení reakční směsi při PCR – 2. kolo

Reakční směs	Konečná koncentrace	Objem (µl)
Přečištěný produkt 1. kola		5
PC2 pufr	1×	4
dNTPs	0,2 mM	4
Ampli 1	0,4 µM	0,2
Lig5A_FS-A-3	0,4 µM	0,2
Klentaq polymeráza	0,75 U	0,15
H_2O		26,45
celkem		40

Tab. XV: Teplotní program

Teplota	Čas	Cykly
95 °C	3 min	
95 °C 60 °C 72 °C	10 s 20 s 1,5 min	35 cyklů
72 °C	5 min	

Primer Ampli1 je na 3´ konci komplementární k sekvenci M13S a na 5´ konci nese sekvenci komplementární k sekvenčním primerům NGS Illumina. Primer Lig5A_FS-A-3 je na 3´ konci reverzně komplementární k sekvenci aktivního genu v oblasti desátého exonu. PCR produkt je více než 1700 bází dlouhý a zahrnuje celý úsek potenciální *RecΔ55* alely.

PCR–ARMS – 3. kolo

Ve 3. kole PCR byla pro obohacení PCR produktu 2. kola o sekvence nesoucí rekombinantní alely použita metoda PCR-ARMS (Amplification Refractory Mutation System). Stručně, jeden z páru primerů pro PCR je navržen tak, že na svém 3´ konci nese bázi komplementární k sledované změně, např. bodové mutaci. Amplifikace za optimálních podmínek probíhá pouze v případě přítomnosti mutace. Pro zvýšení specifičnosti reakce se navíc na třetí nukleotid od 3´ konce primeru vkládá nekomplementární báze (Little 2001). Složení reakční směsi při PCR-ARMS udává Tab. XVI a teplotní program Tab. XVII.

Tab. XVI: Složení reakční směsi při PCR-ARMS – 3. kolo

Reakční směs	Konečná koncentrace	Objem (µl)
plazmid	$2{,}5 \times 10^{-3}$ ng/µl, $2{,}5 \times 10^{-4}$ ng/µl	1
PC2 pufr	1×	2,5
dNTPs	0,2 mM	2,5
Ampli 1	0,4 µM	1
L444PplusM13R	0,4 µM	1
DMSO	2%	1,25
Klentaq polymeráza	0,75 U	0,15
H_2O		15,6
celkem		25

Tab. XVII: Teplotní program

Teplota	Čas	Cykly
95 °C	3 min	
95 °C 69 °C 72 °C	10 s 20 s 2 min	35 cyklů
72 °C	5 min	

Pro optimalizaci nasedání ARMS primeru L444PplusM13 a L444PminusM13 specificky na vlákna nesoucí mutaci L444P byly použity zaklonované produkty aktivního genu a pseudogenu pro GBA a *Rec* alel *RecΔ55* a *RecNciI*.

Primer L444PplusM13 je komplementární na 3´ konci k alele nesoucí mutaci L444P a primer L444PminusM13 k alele aktivního genu bez L444P mutace. Primery jsou na 5´ konci komplementární k sekvenčním primerům NGS Illumina.

10.6 Příprava templátu pro NGS

Templát pro NGS byl dále zpracován podle standardních protokolů firmy Illumina. Indexy označující jednotlivé vzorky byly připojeny k templátům komerční soupravou Nextera® XT Library Prep Kit. Koncentrace takto upraveného templátu byla měřena kvantitativní PCR, reakce byla připravena komerční soupravou KAPA Library Quantification Kit. Templáty byly poté smíchány ve stejném poměru, denaturovány a sekvenovány na přístroji MiSeq.

10.7 Ligační metoda

Ligace

Ligace probíhala ve dvou krocích. V prvním kroku docházelo k nasedání primerů na stejné vlákno DNA (gDNA, respektive plazmidová DNA) během postupného snižování teploty z 95 °C na 56 °C. V druhém kroku probíhala extenze vlákna z prvního primeru a ligace prodlouženého vlákna k druhému primeru. Složení reakční směsi při 1. kroku ligace udává Tab. XVIII a teplotní program Tab. XX. Do reakční směsi 1. kroku ligace byla vybrána jedna z kombinací primerů dle Tab. XIX. Složení reakční směsi při 2. kroku ligace udává Tab. XXI a teplotní program Tab. XXII.

Tab. XVIII: Složení reakční směsi – 1. krok ligace

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	50 ng/µl	7,5
primer forward *	0,4 µM	0,5
primer reverse **	0,4 µM	0,5
PC2 pufr	1×	1
H_2O		0,5
celkem		10

Tab. XIX: Kombinace forward a reverse primerů

primer forward *	primer reverse **
Lig5S_FS-S-3	Lig3S_L444P
Lig5S_del55	Lig3S_L444P
Lig3A_L444P	Lig5A_FS-A-3

Tab. XX: Teplotní program

Teplota	Čas	Cykly
95 °C	5 min	
95 °C – 2 °C/cyklus	2 min	10 cyklů
75 °C – 1 °C/cyklus	1 min	23 cyklů
56 °C	1 hod	

Dvojice primerů (Obr. 11) je vždy komplementární k jednomu vláknu DNA (S – sense vlákno, A – antisensse vlákno). 5´ konce druhých primerů (Lig3S_L444P a Lig3A_L444P)

jsou fosforylovány a pět nukleotidů na 3´ konci těchto primerů je spojeno fosforothiotátovou vazbou. Pět nukleotidů na 5´ konci prvních primerů (Lig5S_FS-S-3, Lig5S_del55 a Lig5A_FS-A-3) je spojeno fosforothiotátovou vazbou. První primery nesou na 5´ konci univerzální amplifikační sekvenci a na 3´ konci specifickou sekvenci k DNA, mezi sekvencemi je 10 bází dlouhý tag. Druhé primery nesou na 3´ konci univerzální amplifikační sekvenci a na 5´ konci specifickou sekvenci k DNA, mezi sekvencemi je 10 bází dlouhý tag.

Tab. XXI: Složení reakční směsi – 2. krok ligace

Reakční směs	Konečná koncentrace	Objem (µl)
dNTPs	0,2 mM	1
NAD+	2 mM	2,8
PC2 pufr	1×	1
Hifi T4 DNA ligáza	4 U	1
Klentaq polymeráza	2 U	0,4
H_2O		3,8
Celkem		10

Tab. XXII: Teplotní program

Teplota	Čas
56 °C	10 min
68 °C	20 min
56 °C	20 min

Restrikce

Odstranění nevyužitých primerů a gDNA: k 20 µl ligačního produktu byly přidány exonukleázy: 3,5 U exonukleázy I, 18 U exonukleázy III, 4 U exonukleázy T7, 0,4 U exonukleázy T, 3 U RecJf, 0,2 U lambda exonukleázy.

V případě optimalizace reakce s templátem plazmidové DNA, byly do reakce zároveň přidány endonukleázy, které štěpí plazmid, ale neštěpí v oblasti zájmu: 10 U EcoRI, 10 U EcoRV, 5 U SpeI, 5 U MluI, 10 U XhoI, 10 U XbaI, 10 U HindIII, 5 U NotI.

Teplotní program restrikce je zaznamenán v Tab. XXIII.

Tab. XXIII: Teplotní program

Teplota	Čas
25 °C	30 min
37 °C	2 hod
80 °C	10 min
95 °C	5 min

PCR

Nepřečištěný produkt ligační reakce a restrikce byl použit jako templát pro PCR. Složení reakční směsi při PCR udává Tab. XXIV a teplotní program Tab. XXV.

Tab. XXIV: Složení reakční směsi při PCR

Reakční směs	Konečná koncentrace	Objem (µl)
Produkt ligační reakce		7
PC2 pufr	1×	5
dNTPs	0,2 mM	5
Ampli 1	0,4 µM	1
Ampli 2	0,4 µM	1
Klentaq polymeráza	2 U	0,4
H_2O		20,6
Celkem		40

Tab. XXV: Teplotní program

Teplota	Čas	Cykly
95 °C	2 min	
95 °C 60 °C 70 °C	10 s 20 s 1 min	35 cyklů
70 °C	5 min	

Po amplifikaci byl PCR produkt vizualizován na 1% agarózovém gelu.

Ligace s více cykly

Ke zvýšení efektivity ligační metody byl původní protokol upraven: ligace probíhala v jednom kroku, nasedání primerů, extenze i ligace prodlouženého vlákna probíhala při jedné teplotě ve 20 cyklech. Upraveno podle (Szemes et al. 2005). Složení reakční směsi při ligaci s více cykly udává Tab. XXVI a teplotní program Tab. XXVII.

Tab. XXVI: Složení reakční směsi při ligaci s více cykly

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	50 ng/µl	7,5
primer forward *	0,4 µM	0,5
primer reverse *	0,4 µM	0,5
PC2 pufr	1×	2
dNTPs	0,2 mM	1
NAD+	2 mM	2,8
Hifi T4 DNA ligáza	4 U	1
Klentaq polymeráza	2 U	0,4
H_2O		4,3
celkem		20

*Do reakční směsi byla vybrána jedna z kombinací primerů dle Tab XIX.

Tab. XXVII: Teplotní program

Teplota	Čas	Cykly
95°C	5 min	
95°C 65°C	30 s 5 min	 20 cyklů

Štěpení gDNA

Ke zvýšení efektivity nasedání primerů byla gDNA předem naštěpena enzymem HindIII, který neštěpí ve sledovaných oblastech *GBA* a *GBAP*. Složení reakční směsi při štěpení gDNA udává Tab. XXVIII.

Tab. XXVIII: Složení reakční směsi při štěpení gDNA

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	8,3 ng/µl	48
NEB2 pufr	1 ×	6
HindIII	120 U	6
celkem		60

Inkubace 16 hod při 37 °C. Následně byla naštěpená gDNA přečištěna etanolovou precipitací.

Ligace primerů

Pro ověření funkčnosti ligázy byla provedena ligační metoda se dvěma sousedními primery, které nasedají na stejné vlákno DNA. Složení reakční směsi při 1. kroku ligace primerů udává Tab. XXIX a teplotní program Tab. XXX. Složení reakční směsi při 2. kroku ligace primerů udává Tab. XXXI a teplotní program Tab. XXXII.

Tab. XXIX: Složení reakční směsi – 1. krok ligace primerů

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	50 ng/µl	7,5
Lig3S_L444P	0,4 µM	0,5
Lig_L444P_sense	0,4 µM	0,5
Ampligázový pufr	1×	1
H_2O		0,5
celkem		10

Tab. XXX: Teplotní program

Teplota	Čas	Cykly
95 °C	5 min	
95 °C – 2 °C/cyklus	2 min	10 cyklů
75 °C – 1 °C/cyklus	1 min	23 cyklů
56 °C	1 hod	

Primer Lig_L444P_sense nasedá na stejné vlákno gDNA těsně před primer Lig3S_L444P.

Tab. XXXI: Složení reakční směsi – 2. krok ligace primerů

Reakční směs	Konečná koncentrace	Objem (µl)
NAD+	2 mM	2,8
Ampligázový pufr	1×	1
Ampligáza	4 U	1
H_2O		5,2
celkem		10

Tab. XXXII: Teplotní program

Teplota	Čas
56 °C	10 min
68 °C	20 min
56 °C	20 min

Produkt ligace byl vizualizován na 12% polyakrylamidovém gelu.

10.8 LAMP

Byly navrženy dva sety primerů pro amplifikaci dvou oblastí genu *GBA* (U_set, L_set) v programu Primer explorer V5, (http://primerexplorer.jp/elamp4.0.0/index.html). Názvy a koncentrace LAMP primerů v mixu jsou zaznamenány v Tab. XXXIV. Složení reakční směsi při LAMP udává Tab. XXXIII a teplotní program Tab. XXXV.

Tab. XXXIII: Složení reakční směsi při LAMP

Reakční směs	Konečná koncentrace	Objem (µl)
gDNA	5,6 ng/µl	0,5
Izotermální amplifikační pufr	1×	2,5
dNTPs	10 mM	3,5
Mg^{2+}	6 mM	1,5
LAMP mix primerů *		1
Bst 2.0 polymeráza **	8U	1
H_2O		15
celkem		25

** Bst 2.0 polymeráza byla přidána do jednotlivých vzorků až po denaturaci po poklesu teploty na 65 °C

* **Tab. XXXIV:** Složení LAMP mix primerů

Název primeru	Koncentrace v LAMP mix primerů	Konečná koncentrace v reakční směsi
FIP	40 µM	1,6 µM
BIP	40 µM	1,6 µM
F3	5 µM	0,2 µM
B3	5 µM	0,2 µM

Tab. XXXV: Teplotní program

Teplota	Čas
95 °C	5 min
** přídavek Bst 2.0 polymerázy	
65 °C	1 hod
80 °C	10 min

Produkty amplifikace byly vizualizovány na 1,5% agarózovém gelu.

11 Bioinformatické zpracování výsledků

Výsledkem sekvenování každého ze vzorků byly dva soubory ve formátu fastq, první obsahující sekvence čtené z 5´ konce templátu (Read1), druhý z 3´ konce templátu (Read2). Sekvence Read1 obsahuje tag a část devátého exonu a intronu *GBA*. Sekvence Read2 obsahuje část desátého exonu a intronu. Díky tomuto uspořádání je možné detekovat mutaci D409H z devátého exonu a mutace L444P, A456P a V460V z desátého exonu.

Běžný postup pro vyhledávání mutací zahrnuje mapování na genom a párování Read1 a Read2 sekvencí. Sekvence jsou poté filtrovány podle kvality, kterou pro každou bázi stanoví sekvenční software, setříděny a obvykle indexovány v binárním formátu bam. Pro analýzu tagem označených sekvencí nejsou běžné bioinformatické postupy využitelné. Proto byl připraven v jazyce perl skript, který provádí zpracování sekvencí automaticky. Podílela jsem se na vytváření a optimalizaci bioinformatického postupu, vlastní skript byl napsán mým školitelem. Je součástí Přílohy 1.

Skript provádí následující kroky:

1) čtení hlaviček sekvencí a kódů kvality sekvence ve formátu fastq
2) filtrování sekvencí
3) setřídění sekvencí se shodnými tagy (ze sekvence Read1), uložení hlaviček sekvencí se shodnými tagy
4) vyhledání párových sekvencí z Read2 k sekvencím Read1 podle hlaviček, přiřazení tagu k sekvenci
5) setřídění sekvencí se shodnými tagy
6) nalezení mutací ve srovnávaných sekvencích se shodným tagem, vytvoření alignmentů sekvencí se shodným tagem a s mutacemi v sekvenci
7) určení počtu jednotlivých mutací nalezených po srovnání sekvencí se shodným tagem

Výstupem skriptu byly alignmenty sekvencí nesoucích shodný tag, údaje o počtu sekvencí, filtrovaných sekvencí, tagů, sekvencí připadajících na jeden tag a údaje o detekovaných záměnách.

11.1 Filtrování

Sekvence s vyšším počtem chyb a zjevné artefakty bylo třeba ještě před analýzou výsledků odfiltrovat, aby nedošlo k jejich zkreslení.

Filtrování pomocí kvality sekvence

Firma Illumina používá skóre kvality Q pro hodnocení kvality čtených sekvencí. Skóre Q pro každou bázi odpovídá pravděpodobnosti, že je báze čtena chybně. Pro zjištění kvality sekvence byl hodnocen očekávaný počet chyb, který odpovídal součtu pravděpodobností chyby u jednotlivých bází (Edgar and Flyvbjerg 2015). Dle zjištěné kvality sekvence byly filtrovány ty, které nesplňovaly kritérium. Kritériem bylo, že v námi sledovaném úseku, kde očekáváme bodové mutace, musí být očekávaný počet chyb menší než jedna.

Filtrování podle podobnosti s referenční sekvencí

Mapování nefiltrovaných sekvencí na genom ukázalo, že kromě malého množství fragmentů (pravděpodobně primery-dimery), se sekvence mapovaly do oblasti exonu 9 až intronu 10 *GBA*. Bylo tedy možné míru podobnosti testované sekvence s referenční sekvencí stanovit jako kritérium pro filtrování nežádoucích sekvencí. Tento způsob filtrování vedl i k tomu, že potenciální sekvence vzniklé fúzí *GBA* a *GBAP* byly odstraněny. Na tyto sekvence nebyla však analýza zaměřena a vzhledem k tomu, že jsou pravděpodobně velice vzácné, nebyly dále zvažovány. Jako kritérium byla konkrétně použita míra shody 100 bází (za 3´ koncem primeru u obou Readů) testované sekvence s referenční sekvencí. V analýze bylo empiricky stanoveno 7 povolených odchylek u Read1 sekvencí a 10 povolených odchylek u Read2 sekvencí. Vyšší počet povolených odchylek u Read2 sekvencí byl z důvodu očekávaných bodových mutací způsobených genovou konverzí.

Některé sekvenované artefakty neobsahovaly celou sekvenci primeru. Proto byla jako další kritérium pro filtrování stanovena možná odchylka sekvence od konstantních částí primerů. V analýze byly stanoveny až 4 povolené odchylky, což znamená, že až 4 báze konstantní části primeru testované sekvence se mohly odlišovat od sekvence referenční.

Filtrování podle délky sekvence

Dalším možným způsobem bylo filtrování podle délky sekvence. Očekávaná délka sekvence byla 250 bází. Pokud byly vyšetřované sekvence kratší, jednalo se ve většině případů o artefakty. Při finální analýze nebyl tento způsob filtrování použit z důvodu eliminace kratších sekvencí již pomocí předchozích dvou typů filtrování.

Kritéria srovnávání sekvencí a detekce bodových mutací

Kritériem pro detekci bodových mutací byl nadpoloviční výskyt mutace v daném místě alignmentu sekvencí, které nesou shodné tagy. Pro zohlednění detekovaných záměn musely alignmenty obsahovat minimálně tři sekvence.

Výsledky

12 Optimalizace a výsledky jednotlivých metod

Optimalizace PCR

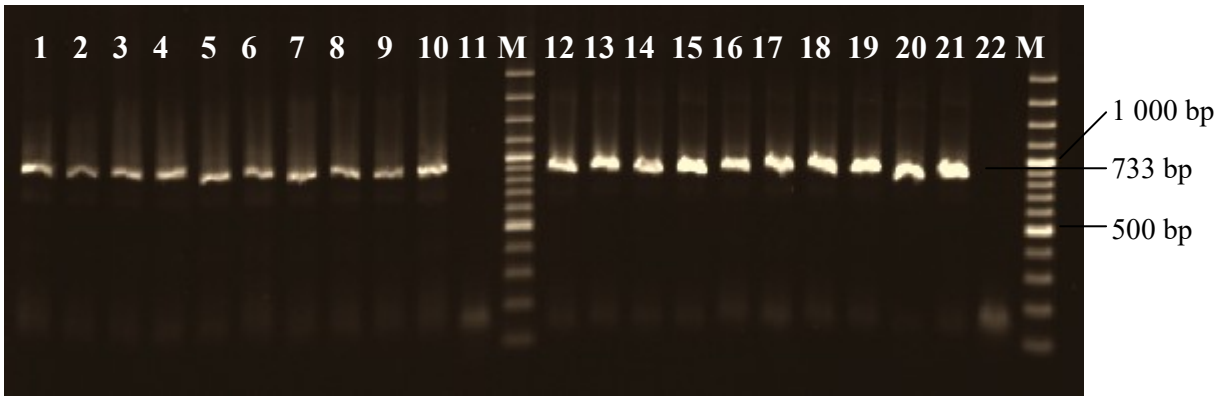
PCR u všech použitých metod bylo optimalizováno z několika hledisek. Použití standardní polymerázy (Klentaq, Red mix) nebo polymerázy (Q5) s 3´- 5´exonukleázovou aktivitou, teplota nasedání primerů, délka použitých primerů, koncentrace hořečnatých iontů, koncentrace betainu, koncentrace dimetylsulfoxidu (DMSO). Byla vybrána taková reakce, při které je nejvyšší výtěžnost, a nevznikají nespecifické produkty. Konečné podmínky jsou uvedeny v protokolech u jednotlivých metod v části Materiál a metody.

13 Asymetrická extenze

13.1 Příprava templátu pro NGS - bez obohacení konvertovanými sekvencemi

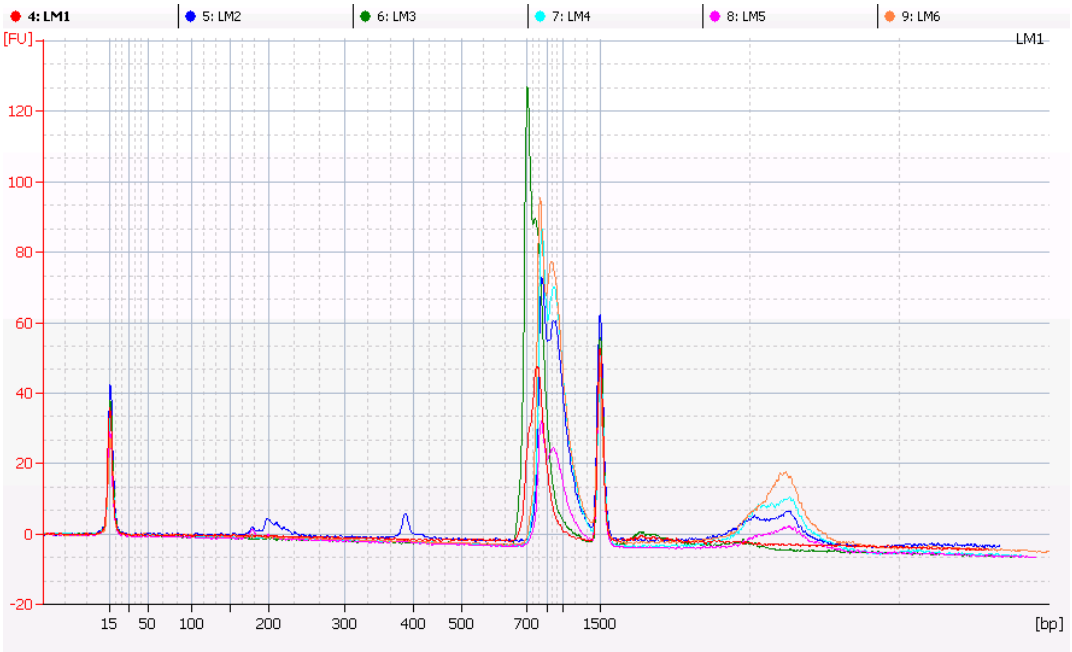
Deset vzorků gDNA z meiotických buněk od 10 mužů a 10 vzorků gDNA z mitotických buněk od 5 žen a 5 mužů bylo použito k přípravě templátu pro stanovení frekvence genové konverze u mezi *GBA* a *GBAP*. Templát pro NGS bez obohacení konvertovanými sekvencemi se připravuje ve dvou kolech. V prvním kole proběhne extenze za použití jednoho primeru, který obsahuje 10 bázový tag, tím je jednoznačně označena každá vytvořená molekula DNA. Použitý primer je specifický pro amplifikaci aktivního genu *GBA* v místě 55 bázové delece. Nevyužité primery byly odstraněny před 2. kolem.

V druhém kole probíhala amplifikace aktivního genu *GBA*, výhradně z produktů 1. kola, nesoucích unikátní tag. K amplifikaci byl použit primer Ampli1, který byl komplementární k univerzální nehybridizující sekvenci primeru z prvního kola a primer GBAe10_read2 komplementární k sekvenci desátého exonu *GBA*, přibližně 50 bází downstream od poslední mutace *RecTL* alely. K ověření specificity metody byly provedeny slepé kontroly bez přidání templátu a dále slepá kontrola bez Ampli1 primeru v druhém kole. PCR produkty všech vzorků a slepé kontroly byly vizualizovány na 1% agarózovém gelu (Obr. 14).



Obrázek 14: Výsledné PCR produkty. Vzorky 1-10: PCR produkty po druhém kole, gDNA z meiotických buněk S1-S10. Vzorek 11: slepá kontrola bez Ampli1 primeru. Vzorky 12-16: PCR produkty po druhém kole, gDNA z mitotických buněk žen Z1-Z5. Vzorky 17-21: PCR produkty po druhém kole, gDNA z mitotických buněk mužů M1-M5. Vzorek 22: slepá kontrola bez templátu. M: velikostní standard.

Pro purifikaci PCR produktu a odstranění nevyužitých primerů po druhém kole byly použity magnetické částice AMPure XP Beads v polyetylenglykolu. Po purifikaci proběhlo připojení indexů pro NGS na platformě Illumina podle standardního protokolu. Délka fragmentů a kvalita templátu byla změřena na fluorescenčním přístroji Bioanalyzer Agilent, který funguje na principu kapilární elektroforézy (Obr. 15).



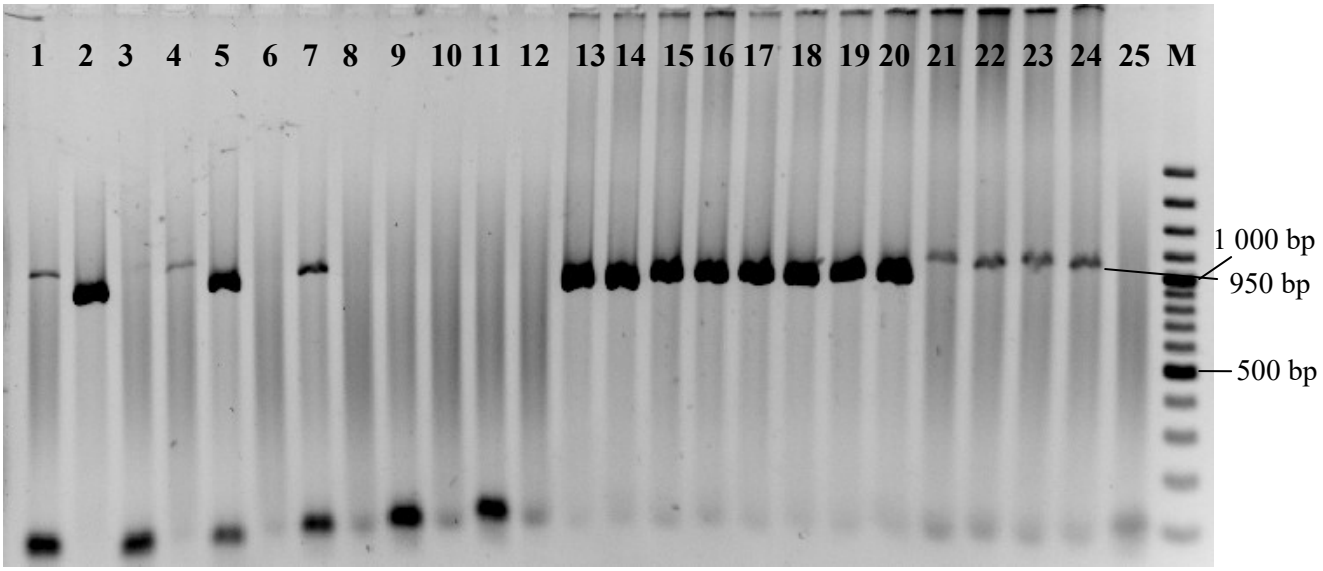
Obrázek 15: Měření délky a kvality templátu 6 vybraných vzorků na elektroforetickém fluorescenčním přístroji Bioanalyzer Agilent. Délka templátu po druhém kole odpovídá 733 bázím (vzorky S6 - červeně a M1 – zeleně), délka templátu po připojení indexů odpovídá 802 bázím (vzorky S6 – modře, M1 – tyrkysově, M5 – růžově, Z3 – oranžově) Osa X: délka DNA fragmentu, osa Y: jednotky fluorescence [FU]. Dimery byly úspěšně purifikační metodou odstraněny a templát prodloužen o indexy pro NGS.

13.2 Příprava templátu pro NGS - s obohacením konvertovanými sekvencemi

Vzhledem k předpokládané nízké frekvenci genové konverze v řádech jedna ku tisíců až statisíců kopií byl templát pro NGS obohacen o konvertované sekvence pomocí PCR-ARMS ve 3. kole.

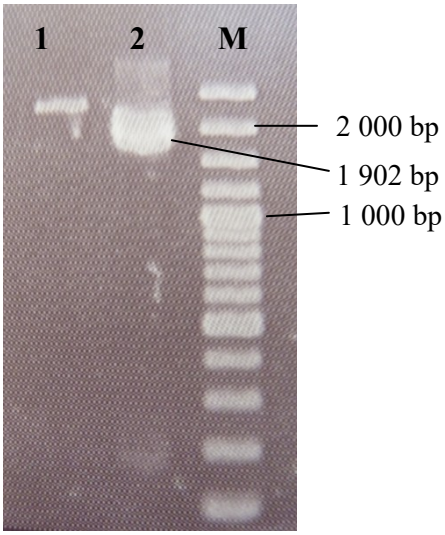
K optimalizaci PCR-ARMS byly jako templát použity zaklonované úseky *GBA*, *GBAP* a *RecNciI* alely. Plazmidová DNA byla ředěna tak, aby počet cílových molekul přibližně odpovídal 500 ng gDNA a více. V reakci byla použita Q5 polymeráza s 3´-5´exonukleázovou aktivitou a Klentaq polymeráza. Q5 polymeráza není vhodná k PCR-ARMS amplifikaci. Specifická báze pro mutovanou či wild type alelu na 3´ konci primeru je opravena vzhledem 3´-5´exonukleázové aktivitě Q5 a vzniká nespecifický PCR produkt (Obr. 16). Klentaq polymeráza se jevila jako dostatečně specifická pro PCR-ARMS reakci, při relativně vysoké teplotě nasedání (69 °C) a 2% DMSO. Na druhou stranu, Klentaq polymeráza má výrazně vyšší chybovost v inkorporaci nukleotidů při tvorbě řetězce.

Vizualizace PCR produktu probíhala na 1% agarózovém gelu.



Obrázek 16: Optimalizace PCR-ARMS u zaklonovaných kontrolních úseků *GBA*, *GBAP* a *RecNciI* alely. Vzorky 1-12: Klentaq polymeráza. Vzorky 13-25: Q5 polymeráza. Vzorek 1-4, 13-16: templátem je zaklonovaný úsek *GBA*. Vzorek 5-8, 17-20: templátem je zaklonovaná rekombinantní alela *RecNciI*. Vzorek 9-12, 21-24: templátem je zaklonovaný úsek *GBAP*. Vzorek s lichým číslem: primer L444PplusM13. Vzorek se sudým číslem: primer L444PminusM13. Vzorek 1,2,5,6,9,10,13,14,17,18,21,22: konečná koncentrace templátu v reakční směsi 2,5 × 10^{-3} ng/µl. Vzorek 3,4,7,8,11,12,15,16,19,20: konečná koncentrace templátu v reakční směsi 2,5 × 10^{-4} ng/µl. Vzorek 25: slepá kontrola bez templátu. M: velikostní standard.

Při přípravě obohaceného templátu o konvertované sekvence byla metoda postupně upravována oproti původně zamýšlenému postupu. Důvodem byla nízká koncentrace a špatná kvalita výsledného produktu u původního postupu. V 1. kole proběhla lineární PCR s 20 cykly, místo původně zamýšlené extenze primeru během 1. cyklu. Tím došlo k významnému navýšení počtu molekul templátu pro 2. kolo, ovšem na úkor unikátního značení pomocí 10 bázového tagu. Následovalo přečištění produktu 1. kola nejprve přes kolonky se skelným papírem pro odstranění nevyužitých primerů a poté pomocí streptavidinových magnetických částic pro odstranění gDNA (primer 1. kola byl na svém 5´ konci biotinylován). Ve 2. kole byl amplifikován celý úsek zahrnující rekombinantní alelu *RecΔ55*, PCR produkt byl dlouhý 1902 bází (Obr. 17). Ve 3. kole došlo ke specifické amplifikaci pouze molekul nesoucích mutantní alelu L444P, případně rekombinantní alely obsahující tuto mutaci (*RecΔ55*, *RecTL*, *RecNciI*). Přes veškerou snahu o optimalizaci metody, se nepodařilo připravit templát dostatečně koncentrovaný a kvalitní pro NGS.



Obrázek 17: Srovnání PCR produktů po 2. kole. Vzorek 1: PCR produkt purifikovaný pomocí streptavidinových magnetických částic. Vzorek 2: PCR produkt nepurifikovaný pomocí streptavidinových magnetických částic. M: velikostní standard.

14 Ligační metoda

Ligační metoda byla složena ze dvou kroků. V prvním kroku docházelo k nasedání primerů na gDNA (popř. na plazmidovou DNA). V druhém kroku bylo vlákno pomocí DNA polymerázy extendováno a současně navázáno k druhému primeru pomocí DNA ligázy. Pro ověření byly používány slepé zkoušky bez ligázy a bez DNA polymerázy.

K odstranění primerů a gDNA bylo do vzorku po ligaci přidáno 6 exonukleáz. Aby docházelo k účinnějšímu štěpení, zároveň bylo přidáváno dalších 8 endonukleáz, které v oblasti sledovaného úseku neštěpí. Jednovláknový ligační produkt již zbavený gDNA a primerů byl amplifikován pomocí univerzálních amplifikačních primerů. Vizualizace PCR produktu probíhala na 1% agarózovém gelu.

První pokusy ligační metody s gDNA dopadly neúspěšně, proto jsme přistoupili k optimalizaci metody na kontrolních plazmidech, kde jsme očekávali, že s templátem o vyšším počtu cílů bude výsledek detekován i při nižší účinnosti provedeného postupu.

Optimalizace 1. kola proběhla z hlediska výběru vhodných polymeráz, ligáz a pufrů. Z DNA polymeráz byly vyzkoušeny: Q5, Klentaq, Dreamtaq, Ampli Taq Gold, Phusion a HiQu. Z ligáz byly vyzkoušeny: termostabilní Ampligáza, T4 ligáza a HiFi Taq DNA ligáza. Byly testovány jak ligázové, tak polymerázové pufry, neboť oba enzymy museli být funkční v jednom prostředí. Dále byl vyzkoušen betain, který byl často zmiňován v původních pracích metod SMART (Krishnakumar et al. 2008), neboť může zlepšit extenzi v GC bohatých oblastech.

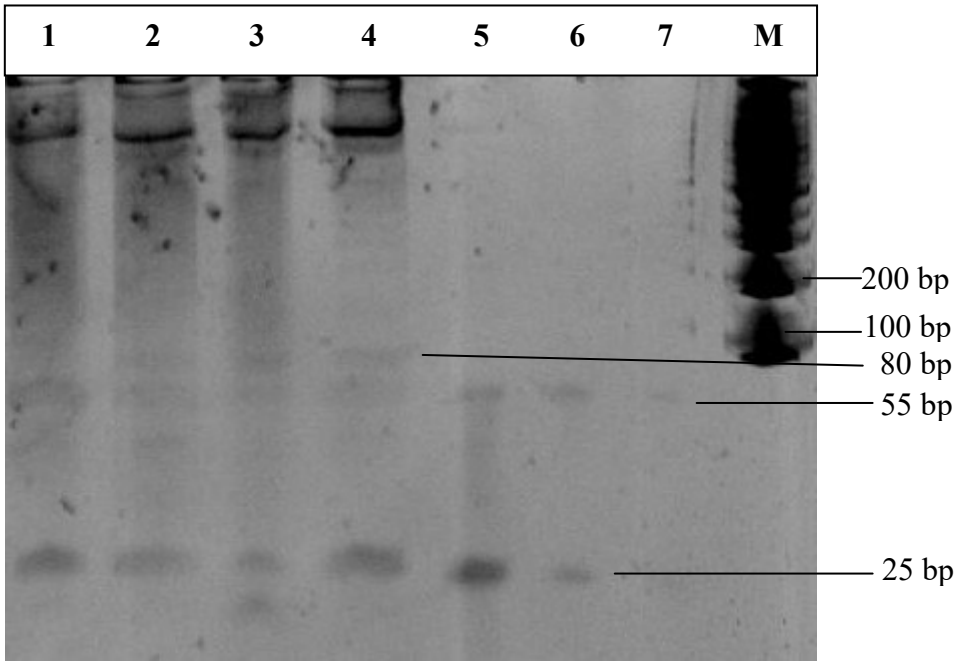
Z DNA polymeráz jsme získali PCR produkt u ligační metody jen za použití Klentaq polymerázy. Z ligáz se nám osvědčila Ampligáza a HiFi Taq DNA ligáza. Ligační metody jsme optimalizovali s PC2 pufrem, ve kterém byla funkční jak Klentaq polymeráza, tak také Ampligáza nebo HiFi Taq DNA ligáza. Betain přidaný do reakční směsi zhoršoval účinnost ligační metody.

Dále bylo optimalizováno množství exonukleáz a doba působení k odbourání lineární ssDNA a dsDNA po 1. kole. Také byly vyzkoušeny různé pufry, ve kterých by byla účinnost štěpení vyšší. Bylo zjištěno, že působení exonukleáz na ligační produkt přes noc není dostačující a stále byla na agarózovém gelu vizualizovaná gDNA a primery. Naopak množství exonukleáz příliš neovlivňovalo kvalitu štěpení. Pro vyšší účinnost štěpení gDNA a primerů byly do reakce přidávány také endonukleázy, které významnou měrou přispěly k vyšší efektivitě

štěpení exonukleázami. Přidání endonukleáz k ligačnímu produktu mělo také zásadní význam v případě, kdy jako templát byla použita plazmidová DNA, která je cirkulární a bylo nutné jí pro působení exonukleáz rozštěpit. Optimální prostředí pro směs štěpících enzymů byl 1× pufr lambda exonukleázy.

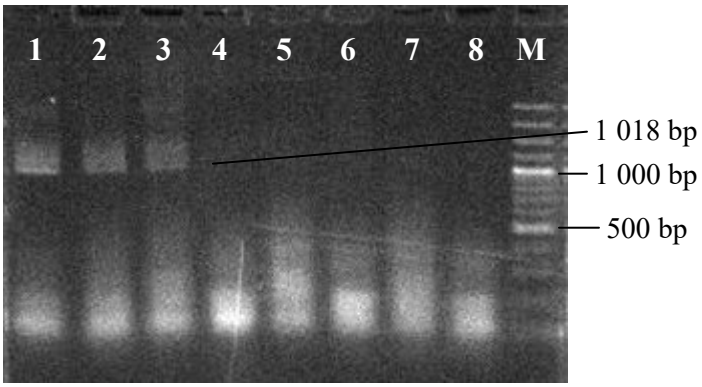
K ověření dostatečné ochrany ligačních produktů proti exonukleázovému štěpení, bylo provedeno PCR s primery, které mají na svém konci nukleotidy s fosforothiotátovými vazbami. Výsledným PCR produktem byla dsDNA, na jejíchž koncích byly nukleotidy s fosforothiotátovými vazbami. Byly vybrány exonukleázy ExoI, ExoT, RecJf a lambda exonukleáza, které štěpí pouze ssDNA a byly přidány k PCR produktu. PCR produkt nebyl degradován a zachoval si i stejnou koncentraci.

K ověření funkčnosti ligázy v prostředí PC2 pufru byl navržen primer Lig_L444P_sense (dlouhý 25 bází), který leží na stejném vlákně vedle (ve směru 5´-3´) primeru Lig3S_L444P (dlouhý 55 bází). Nejprve bylo nutné zjistit, kolik ng DNA je možné vizualizovat na polyakrylamidovém gelu. Byla připravena ředící řada primerů, kdy byly smíchány oba výše zmíněné primery a to tak, že první směs obsahovala dohromady 17 ng DNA, druhá 1,7 ng DNA a třetí 0,8 ng DNA. Ligační reakce byla provedena s 2,5 ng primerů a Ampligázou. Produkt ligace spolu s ředící řadou primerů byly naneseny na 12% polyakrylamidový gel (Obr. 18). Z Obrázku 18 je patrné, že došlo k ligaci primerů. Zdá se, že účinnost ligace není příliš vysoká z důvodu velkého množství neligovaných primerů.



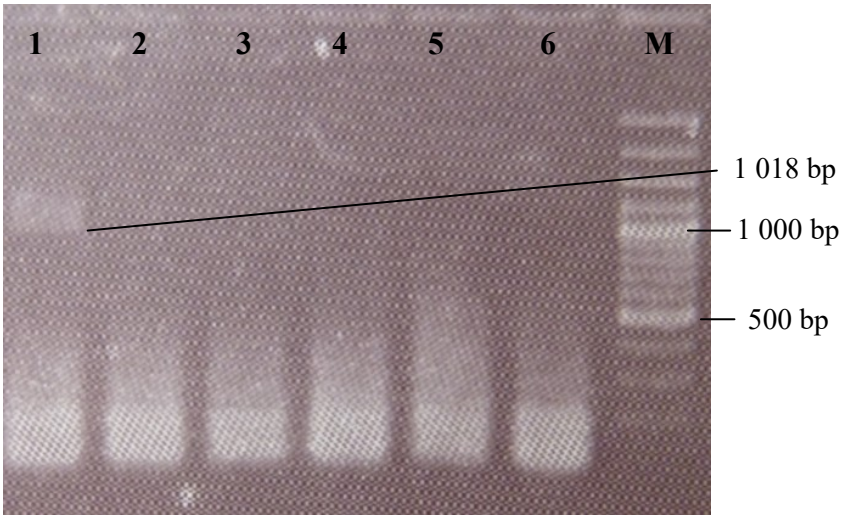
Obrázek 18: Optimalizace ligační reakce. Zobrazení produktů ligace primerů (velikost 80 bp) a ředící řada směsi primerů Lig_L444P_sense (25 bp) a Lig3S_L444P (55 bp). Vzorek 1: slepá kontrola bez přídavku ligázy. Vzorky 2-4: produkty ligace. Vzorky 5-7: ředící řada směsi primerů. Vzorek 5: směs primerů o 17 ng DNA, Vzorek 6: směs primerů o 1,7 ng DNA. Vzorek 7: směs primerů o 0,8 ng DNA. M: velikostní standard.

Optimalizace ligační metody na kontrolních plazmidech proběhla za výše uvedených podmínek. Jako slepé kontroly reakce sloužily vzorky bez přídavku polymerázy, ligázy a templátu, dále vzorky zaklonovaných *GBA-P5* a *GBAP-PSG2* (Obr. 19). PCR produkt byl purifikován a sekvenován na sekvenátoru ABIprism A3100. Sekvence odpovídala cílovému úseku.



Obrázek 19: Optimalizace ligační metody. Vzorek 1-3: pozitivní kontrola, templátem zaklonovaná *RecNciI* alela. Vzorky se liší koncentrací primerů, která se snižuje zleva doprava (0,4 μM, 0,2 μM, 0,1 μM). Vzorek 4: negativní kontrola, templátem zaklonovaný *GBA*-P5. Vzorek 5: negativní kontrola, templátem zaklonovaný *GBAP*-PSG2. Vzorek 6: slepá kontrola, bez přidání ligázy. Vzorek 7: slepá kontrola, bez přidání DNA polymerázy. Vzorek 8: slepá kontrola, bez přidání templátu. M: velikostní standard.

Dále byl vytvořen gradient koncentrace plazmidu se zaklonovanou *RecNciI* alelou, abychom zjistili, při jakém počtu cílů jsme schopni PCR produkt po amplifikaci vizualizovat. Zjistili jsme, že se jedná o 20 fmol (Obr. 20).



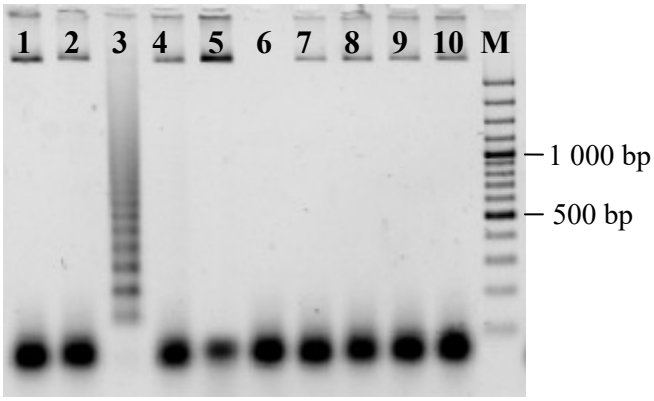
Obrázek 20: Gradient koncentrace plazmidu se zaklonovanou *RecNciI* alelou. Vzorky 1-4: koncentrace plazmidu se snižuje zleva doprava (20 fmol, 2 fmol, 200 attomol, 0,02 attomol). Vzorek 5: slepá kontrola bez přidané DNA ligázy. Vzorek 6: slepá kontrola bez přidaného templátu. M: velikostní standard.

Po optimalizaci reakce na kontrolních plazmidech jsme opět přešli ke gDNA. Z literatury o zámkových sondách bylo zjištěno, že vyšší efektivita nasedání primerů je u ligace s více cykly. PCR produkt nebyl vizualizován ani po ligaci s 20 cykly. Dospěli jsme k názoru, že příčinou nefunkčnosti ligační metody není nízká efektivita nasedání primerů na cílová místa v DNA.

Další příčinou nízké efektivity ligační metody mohla být špatná dostupnost cílů pro primery u gDNA. gDNA byla proto štěpena endonukleázou HindIII, která neštěpí ve sledované oblasti. Po naštěpení gDNA nebyla zaznamenána vyšší efektivita ligační metody.

15 LAMP

Byly navrženy 2 sety primerů U_set a L_set. Každý ze setů se skládá ze 2 párů primerů. Primery z U_setu nasedají na *GBA* v místě, kde je u *GBAP* 55 bázová delece a na mutantní alelu L444P. Primery L_setu nasedají na mutantní alelu L444P a úsek intronu 10, který je v tomto místě specifický pro aktivní gen *GBA*. Produkt amplifikace byl vizualizován na 1,5% agarózovém gelu (Obr. 21). Pozitivní výsledky byly i po optimalizaci výjimečné a nahodilé. Z tohoto důvodu jsme již dále tuto metodu nerozvíjeli.



Obrázek 21: Optimalizace LAMP. Vzorky 1-5: U_set primerů Vzorek 6: slepá kontrola bez přidaného templátu. Vzorek 7-10: L_set primerů. M: velikostní standard.

16 Výsledky NGS

16.1 Testovací sekvenování na platformě Illumina MiSeq

Byl připraven testovací templát pro NGS bez obohacení konvertovanými sekvencemi. Jednalo se o PCR produkt po 3 kolech – 1. kolo: extenze primeru, 2. kolo: PCR a 3. kolo rePCR. PCR produkt po třetím kole byl purifikován přes kolonku se skelným papírem. Primery posledního kola PCR obsahují na nehybridizujícím konci sekvenci potřebnou pro přidání indexů na NGS. Po přidání indexů byl PCR produkt sekvenován na platformě Illumina MiSeq. Bylo získáno více než 200 000 sekvencí z obou konců PCR produktu, tzn. Read1 a Read2. Po filtrování nevyhovujících sekvencí, zůstalo méně než 20 000 sekvencí s 4 043 hodnotitelnými tagy v Read1. Při hodnocení kvality sekvence pomocí FastQC bylo zřejmé, že jsou přítomny ve větší míře kratší sekvence než očekávaných 250 bází a s nízkou kvalitou. Analýza sekvencí potvrdila četnou přítomnost PCR artefaktů. Nebyly však detekovány žádné sekvence, které by plně odpovídaly sekvenci pseudogenu, což potvrdilo, že byla selektivně amplifikována sekvence aktivního genu.

Pro finální sekvenování byl upraven pracovní postup přípravy templátu pro NGS. Nebylo prováděno rePCR, koncentrace primerů byla snížena o více než polovinu a výsledný PCR produkt byl purifikován pomocí magnetických částic AMPure XP Beads. Konečný protokol je zaznamenán v části Materiál a metody. Při finálním sekvenování byl zaznamenán výrazně nižší počet krátkých sekvencí a PCR artefaktů.

16.2 Finální sekvenování na platformě Illumina MiSeq

Celkově bylo analyzováno více než 30 milionů čtených sekvencí z dohromady 20 vzorků, jejichž podíl odpovídal 1,2-7,8 % z celkového počtu (Tab. XXXVI). Kvalita jednotlivých sekvencí a délky čtení byla hodnocena pomocí programu FastQC. Celková kvalita byla nižší, pouze 70 % sekvencí mělo Q skóre větší než 30, z důvodu sekvenace templátu se stejnou sekvencí.

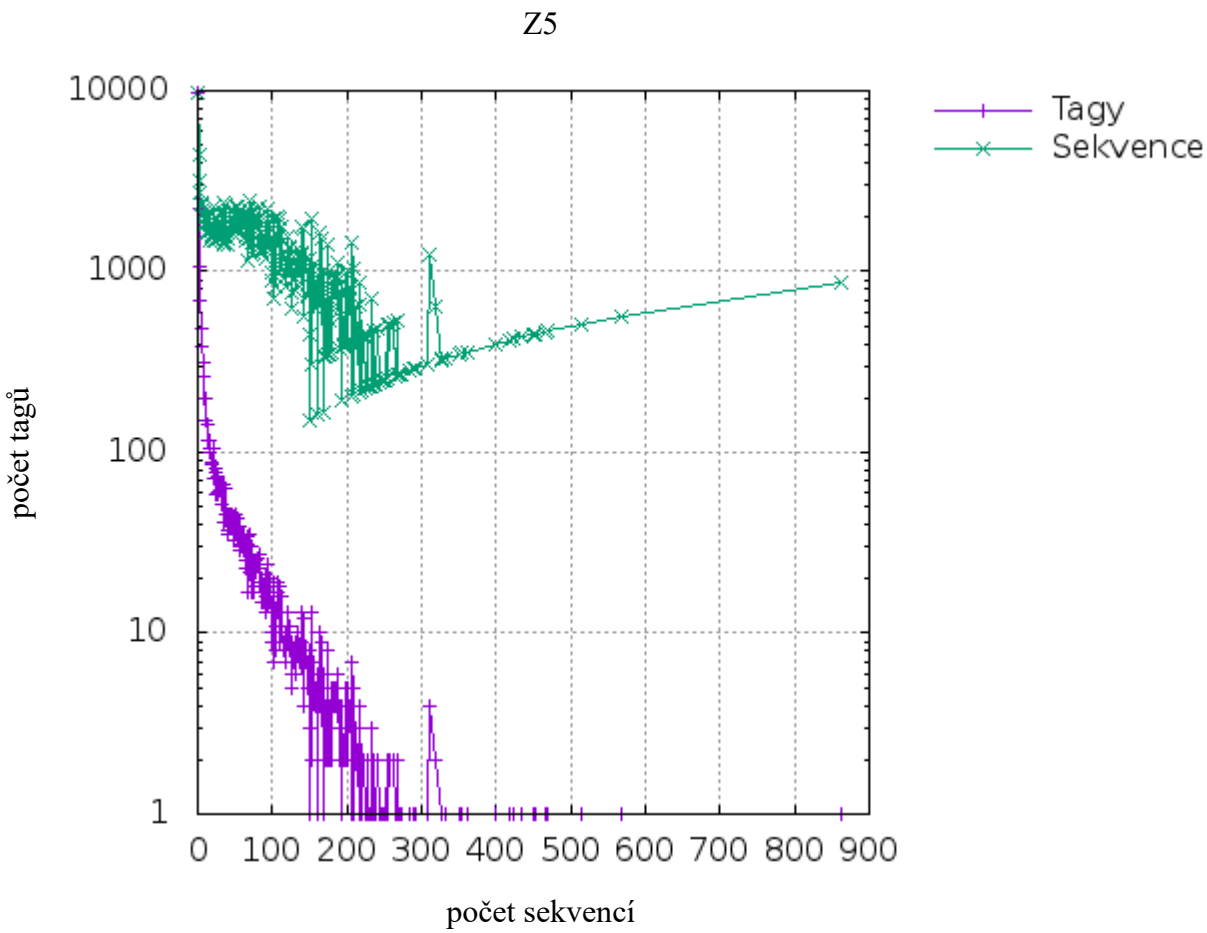
Tab. XXXVI: Srovnání jednotlivých vzorků z mitotických a meiotických buněk je možné díky indexům 1 a 2. Procentuální zastoupení jednotlivých vzorků bylo srovnatelné

Pořadí	Označení vzorku	Index 1 (I7)	Index 2 (I5)	Procentuální zastoupení (%)
1	S1	TAAGGCGA	CTCTCTAT	7.7886
2	S2	CGTACTAG	CTCTCTAT	5.6855
3	S3	TCCTGAGC	CTCTCTAT	5.7429
4	S4	GGACTCCT	CTCTCTAT	5.1357
5	S5	TAGGCATG	CTCTCTAT	2.7079
6	S6	TAAGGCGA	TATCCTCT	4.4961
7	S7	CGTACTAG	TATCCTCT	5.1008
8	S8	TCCTGAGC	TATCCTCT	4.2628
9	S9	GGACTCCT	TATCCTCT	3.9808
10	S10	TAGGCATG	TATCCTCT	3.5659
11	M1	TAAGGCGA	GTAAGGAG	2.8316
12	M2	CGTACTAG	GTAAGGAG	5.1594
13	M3	TCCTGAGC	GTAAGGAG	1.1956
14	M4	GGACTCCT	GTAAGGAG	2.8611
15	M5	TAGGCATG	GTAAGGAG	3.2279
16	Z1	TAAGGCGA	ACTGCATA	2.4760
17	Z2	CGTACTAG	ACTGCATA	4.2478
18	Z3	TCCTGAGC	ACTGCATA	2.3489
19	Z4	GGACTCCT	ACTGCATA	3.5573
20	Z1	TAGGCATG	ACTGCATA	3.2061

Pomocí skriptu byl analyzován počet sekvencí ve vztahu k počtu tagů ve všech vzorcích. Většina sekvencí měla tag, který se vyskytoval jen jednou. Byly také zaznamenány alignmenty s jedním shodným tagem, které obsahovaly mnoho set i tisíc sekvencí (Tab. XXXVII, Obr. 22).

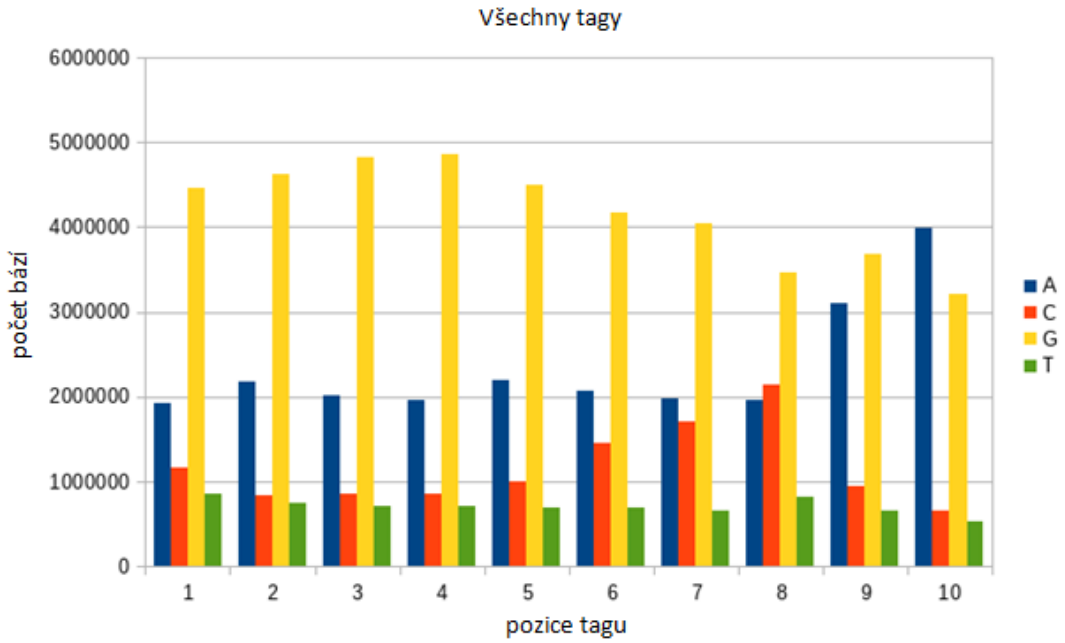
Tab. XXXVII: Vybrané počty sekvencí se shodným tagem, jejich četnost výskytu a procentuální zastoupení dle počtu sekvencí na shodný tag a procentuální zastoupení všech tagů

Počet sekvencí se shodným tagem	Četnost výskytu	Procentuální zastoupení počtu sekvencí (%)	Procenta všech tagů (%)
1	10 658	0,1272	0,13
299	1 701	0,0203	10,01
1 053	797	0,0095	20,08
2 473	456	0,0054	30,00
4 886	295	0,0035	40,31
8 155	211	0,0025	50,00
12 848	152	0,0018	60,00
19 443	107	0,0013	70,00
29 228	68	0,0008	80,00
46 431	33	0,0004	90,00
196 952	1	0,0000	0,00



Obrázek 22: Graf znázorňující počet sekvencí v alignmentu se shodným tagem (zeleně) a počet shodných tagů na počet sekvencí (fialově). Konkrétně se zde jedná o vzorek (Z5) z mitotických buněk ženy. Osa X: počet sekvencí, osa Y: počet tagů.

Unikátních tagů bylo v našem případě okolo 200 000. U všech tagů bylo zjištěno zastoupení nukleotidů na každé pozici (Obr. 23). Dále byly seřazeny nejčastěji se vyskytující tagy, byla vypsána jejich sekvence a procentuální zastoupení (Tab. XXXVIII). Sekvence odpovídá grafickému znázornění nejčastěji zastoupených nukleotidů v tagu na Obr. 21.



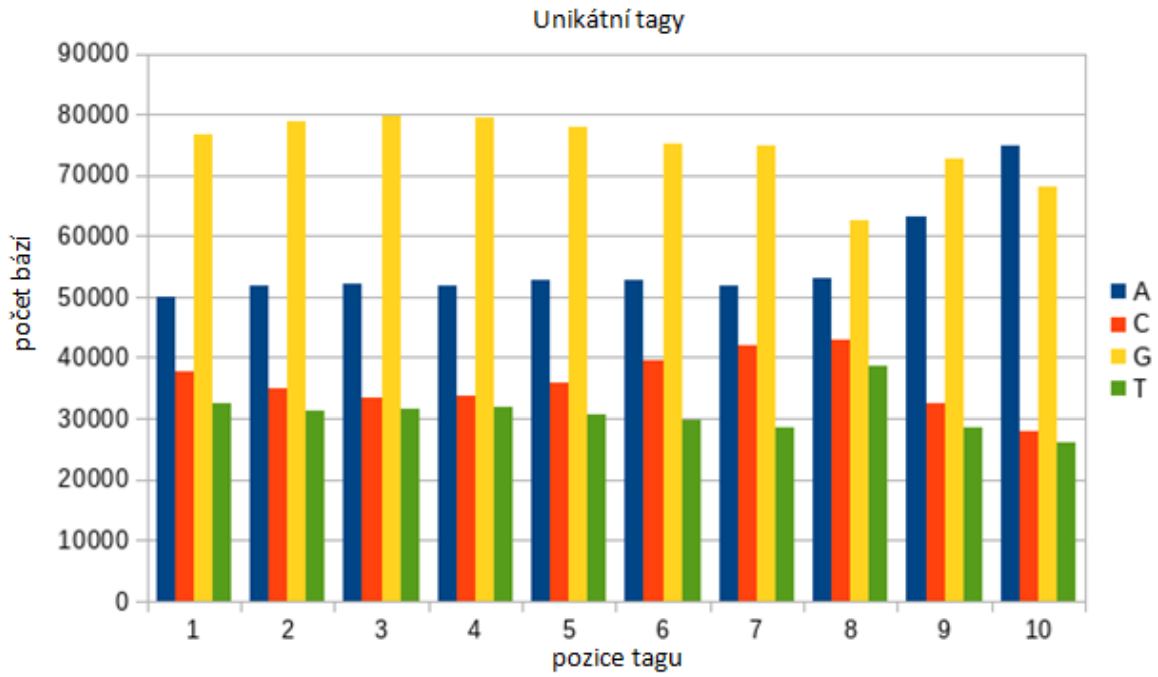
Obrázek 23: Graf znázorňující zastoupení bází na jednotlivých pozicích všech tagů. Na prvních osmi pozicích od 5´ konce tagu se s převahou vyskytuje guanin. Ostatní báze se průměrně vyskytují 2× až 10× méně často. Po guaninu je další často zastoupenou bází adenin, který se na desáté pozici od 5´ konce tagu vyskytuje nejčastěji. Osa X: pozice báze v tagu, osa Y: počet bází.

Tab. XXXVIII: Pořadí a sekvence tagů, které se vyskytují s největší četností. V sekvenci 10 nejčastějších tagů převládá guanin, v menší míře se vyskytuje adenin, cytosin se vyskytuje ojediněle a tymin nebyl v sekvenci obsažen

Pořadí	Sekvence tagu	Četnost výskytu	Procentuální zastoupení (%)
1	GGGGGGGCAA	10658	0,12725
2	GGAGGGGCAA	9195	0,10978
3	GGGGAGGCAA	8158	0,09740
4	GAGGGGGCAA	7763	0,09268
5	GGGGGGGGGA	7435	0,08877
6	GGGGGGGGAA	6200	0,07402
7	GGGGAGGGGA	6011	0,07177
8	GGGAGGGGGA	5939	0,07091
9	GGGGGCGCAA	5925	0,07074
10	GGGGGGACAA	5817	0,06945

Dále bylo srovnáno zastoupení bází na jednotlivých pozicích v sekvenci u unikátních tagů. Opakující se shodný tag byl započítán pouze jednou. Četnost výskytu bází byla opět nevyrovnaná (Obr. 24). Rozdíl byl však menší, než byl zaznamenán u všech tagů, kde se počítaly i opakující se shodné tagy (Obr. 23). Podíl na tom může mít i chybné čtení sekvence některého z častých tagů, které by vedlo k tomu, že tag by pak byl rozpoznán jako „unikátní“.

Tagy s určitými sekvencemi se tedy vyskytovaly výrazně častěji než tagy s jinými sekvencemi, což svědčí pro to, že značení sekvencí nemuselo být u nejčastějších tagů náhodné. Jen 299 sekvencí se shodným tagem zaujímalo 10 % všech sekvencí, 2 473 sekvencí se shodným tagem 30 % (Tab. XXXVII).



Obrázek 24: Graf znázorňující zastoupení bází na jednotlivých pozicích unikátních tagů. Na prvních devíti pozicích od 5´ konce tagu se s převahou vyskytuje guanin. Po guaninu je další často zastoupenou bází adenin, který na desáté pozici od 5´ konce tagu vyskytuje nejčastěji. Osa X: pozice báze v tagu, osa Y: počet bází.

Na ukázku byla náhodně vybrána část alignmentu Read1, která oproti referenční sekvenci obsahuje odchylku v nadpolovičním počtu srovnávaných sekvencí (Obr. 25).

```
R1 TATATCTCAA 1 110A

GTAAAACGACGGCCAGTGNNNNNNNNNNTTGGGTGCGTAACTTTGTCGACAGTCCCATCATTGTAGACATCACCAAGGACACGTTTTACAAACAGCCCATGTTCTACCACCTTGGC

..................TATATCTCAA.........................................T...................C...................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAAG..T.G...............................................................................A.....

..................TATATCTCAA.....................................................................................

..................TATATCTCAA.....................................................................................

..................TATATCTCAA.....................................................................................A.....

..................TATATCTCAA.....................................................................................A.....

-----------------TATATCTCAA-----------------------------------------------------------------------------A-----
```

Obrázek 25: Znázornění alignmentu, kde jsou sekvence se shodným tagem srovnávány s referenční sekvencí. Mezi bázemi 18 a 28 se nachází tag. Poslední řádek alignmentu ukazuje programem rozpoznanou odchylku oproti referenční sekvenci. Je patrná mutace C>A v nadpolovičním počtu srovnávaných sekvencí.

Detekce bodových mutací

Počet sekvencí Read1 byl na vzorek v rozmezí 119 586 až 672 305. Počet alignmentů s unikátním tagem na vzorek byl v rozmezí 7 804 až 15 290 a dohromady u všech vzorků čítal 217 191. V jednotlivých vzorcích bylo nalezeno 12-48 potenciálních bodových mutací, které byly následně ověřovány přehodnocením alignmentů sekvencí. Výsledně byly detekovány dvě bodové mutace L444P, jedna byla zachycena u vzorku z mitotických buněk ženy a druhá u vzorku z meiotických buněk muže. Dále byla detekována jedna bodová mutace D409H u vzorku z mitotických buněk muže. Frekvenci genové konverze mezi *GBA* a *GBAP* lze odhadnout menší než 1 : 10 000.

Diskuse

Detekce velmi vzácných změn v gDNA, jakými jsou například bodové mutace v *GBA* vzniklé genovou konverzí, je obtížná. I při využití současných NGS technik vznikají chyby, zejména kvůli nezbytnosti vícenásobných amplifikačních kroků. Ty nevyhnutelně vedou ke vzniku chyb z důvodu chybovosti použitých DNA polymeráz. Proto byly navrženy techniky, které se snaží výskyt chyb způsobený laboratorními postupy snížit nebo alespoň zjistit míru chybovosti, aby bylo možné výsledky smysluplně interpretovat.

Součástí studie bylo testování způsobu obohacení konvertovanými sekvencemi, což by umožnilo zjišťování rozsahu rekombinovaného úseku při sekvenování. Ačkoli NGS sekvenování PCR produktu bylo navrženo tak, aby umožnilo identifikaci mutací v kopiích jedné molekuly DNA, obohacení vzorku rekombinovanými molekulami by umožnilo získat jich mnohem větší počet.

U ligační metody byla snaha spojit výhody specifity ligační reakce, a tím rozeznání jenonukleotidového polymorfismu (SNP), dále zachycení delšího úseku (okolo 900 bází) na jedné molekule DNA a unikátní značení jednotlivých molekul. Z důvodu nefunkčnosti ligační metody byla snaha optimalizovat jednotlivé kroky postupu. Byly testovány různé DNA polymerázy, DNA ligázy a pufry. Metodu se podařilo zoptimalizovat jen za využití templátu o vysokém počtu cílových molekul (plazmidové DNA) s Klentaq polymerázou, která není pro detekci vzácných mutací vhodná z důvodu vysoké chybovosti. Použití Q5 polymerázy pravděpodobně nebylo možné z důvodu 3′-5′ exonukleázové aktivity. Dále nebyl optimalizován pufr pro DNA ligázu a zároveň pro DNA polymerázu, aby s vysokou účinností fungovaly oba enzymy. Byla testována ligace s více cykly, štěpení gDNA, restrikce a funkčnost ligační reakce. Přes veškeré optimalizace nebyl zaznamenán žádný výsledek na gDNA. Pozitivní výsledek s plazmidovou DNA byl zaznamenán teprve při 20 fmol vstupního templátu, tj. 12 miliard cílových molekul. Při 500 ng gDNA heterozygota byl počet cílů 75 000. Ligační metoda byla pro naše účely z tohoto pohledu málo citlivá.

Z důvodu velice nízké citlivosti ligační metody byla snaha zavést další metodu, která také funguje na principu obohacení konvertovanými molekulami. U této metody zahrnující ARMS PCR opět nebylo možné použití Q5 polymerázy z důvodu 3′-5′ exonukleázové aktivity, díky které dochází k odstranění nekomplementárního nukleotidu na 3′ konci ARMS primeru. Úskalím této metody byla také 3 kola PCR o 30 nebo více cyklech, kdy je vysoké riziko zanesení chyb z důvodu chybovosti Klentaq polymerázy. Rizikem jsou také nedokončené

produkty extenze z předchozích kol amplifikace, které mohou sloužit jako primery v následujících kolech a diferenciální amplifikace templátů může vést k přednostní amplifikaci jen některých templátů, což obojí vede ke změně poměrů mezi sledovanými sekvencemi. Dalším rizikovým krokem byla purifikace biotinylovaného PCR produktu přes streptavidinové částice, kdy docházelo po purifikaci k úbytku koncentrace PCR produktu.

Metoda LAMP je převážně používána pro detekci cizorodé DNA ve vzorku. Díky vysoké specifitě a senzitivitě je možné LAMP použít i k zachycení SNP. LAMP byla úspěšně aplikována na detekci SNP a následně tak rozpoznání alely cytochromu P450 (Iwasaki et al. 2003) nebo k detekci SNP, který je spojován s odolností vůči lékům proti malárii (Yongkiettrakul et al. 2017). Vhodná délka amplikonu je menší než 300 bází, pokud je delší než 500 bází, senzitivita metody je nižší (Li et al. 2016). V našem případě se jednalo o délku dvou amplikonů o 518 bázích a o 388 bázích. Pravděpodobně z důvodu delšího amplikonu než je doporučováno, a zároveň vysokého nároku na senzitivitu a specifitu z důvodu detekce SNP byly pozitivní výsledky nestabilní a ojedinělé i po optimalizaci metody.

Po neúspěšné optimalizaci ligační metody, mutačně specifického PCR a LAMP byla zavedena nová metoda, která již nefunguje na principu obohacení konvertovanými molekulami. Tato metoda využívá jen jeden cyklus nasedání primeru a extenze vlákna, tudíž byl počet výsledných molekul menší nebo roven (v ideálním případě) počtu molekul templátu. Další výhodou bylo, že extenze primeru a PCR probíhala za účasti Q5 polymerázy a PCR probíhalo jen v jednom kole.

Pro značení jednotlivých molekul DNA tagem jsme se rozhodli z důvodu poměrně vysoké chybovosti NGS (Manley et al. 2016). Již v dřívějších studiích bylo zaznamenáno úspěšné odfiltrování chyb vzniklých při amplifikaci a sekvenaci díky tagům. Příkladem může být studie, kdy v genomové DNA fága M13mp2 bylo pomocí senzitivních genetických stanovení zjištěna substituční frekvence 3×10^{-6} (McBride et al. 1992). M13mp2 DNA byla běžnou metodou bez použití tagů sekvenována na platformě Illumina HiSeq. Substituční frekvence byla udána jako $3{,}8 \times 10^{-3}$, což je tisíckrát víc, než je frekvence reálná, a to z důvodu vysoké chybovosti sekvenace. Poté byly na naštěpenou ssDNA M13mp2 ligací připojeny tagy. Zjištěná frekvence substitucí byla v tomto případě stanovena na $3{,}4 \times 10^{-5}$. Toto číslo je také falešně vyšší, ale méně než desetkrát. Tento rozdíl je dán mutacemi pravděpodobně zanesenými v prvním kole PCR. Dále byly vyzkoušeny ligací připojené tagy na duplexu DNA. Frekvence mutací byla určena na $2{,}5 \times 10^{-6}$, což už je srovnatelné číslo s reálnou

substituční mutací (Schmitt et al. 2012). V této práci byl použit tag jen pro jedno vlákno DNA a výsledné sekvence nemohly být srovnány s druhým tagem na druhém vlákně DNA, jako je tomu u duplexového sekvenování. Do budoucna by bylo zajímavé vyzkoušet aplikaci duplexového sekvenování na konvertovaných sekvencích *GBA*.

Předpoklad úspěšného označení jednotlivých molekul DNA byl tag skládající se z náhodných nukleotidů. Odhadované množství cílů v gDNA, která byla přidávána do reakce, bylo 75 000, teoretický počet tagů 4^{10} tedy výrazně převyšoval počet cílů v templátu. Při sekvenování bylo v jednotlivých vzorcích nalezeno 7-15 tisíc hodnotitelných alignmentů nesoucích unikátní tag. Toto číslo by mělo představovat počet kopií templátu, které byly amplifikovány. To však nezahrnuje tagy, které byly reprezentovány pouze jednou nebo dvěma sekvencemi, protože kriteriem pro vytvoření platného alignmentu byla přítomnost nejméně tří sekvencí. Podíl tagů s počtem sekvencí menším než tři byl velice vysoký, obvykle srovnatelný s počtem platných alignmentů. To bylo pozorováno i jinými autory (Brodin et al. 2015), kteří jako možnou příčinu zvažují chyby čtení tagů způsobené sekvenováním a chybami polymerázy nebo posun v poměrech počtu tagů způsobený amplifikací. U čtených sekvencí nebylo náhodné zastoupení bází tagů, ale byla zjištěna převaha bází, konkrétně guaninu, která byla patrná i u nejčastěji zastoupených tagů. Nepoměr v počtu shodných tagů na počet sekvencí by mohl být teoreticky způsoben přednostní amplifikací určitých tagů, což je méně pravděpodobné, protože sekvence s tagem nebyla součástí nasedací části primeru. Dalším možným vysvětlením by mohla být nenáhodná syntéza sekvence tagu, kdy by byly na některých pozicích přednostně zařazovány určité nukleotidy. Posledním možným vysvětlením je tzv. převzorkování, ke kterému dochází v případě nízkého počtu cílových molekul templátu následované mnohonásobnou amplifikací. Výsledky poté představují spíše poměry v PCR produktech po prvních kolech amplifikace než v původním vzorku DNA (Jabara et al. 2011).V každém případě se tím snižuje počet hodnotitelných cílů. Z důvodu nízkého počtu zohledňovaných alignmentů nebylo možné dosáhnout vysokého počtu pozitivních výsledků. Kritéria pro srovnávání sekvencí byla stanovena mírně, aby příliš přísné podmínky nevedly k tomu, že by některá z mutací nebyla detekována. Podobné podmínky využívali i jiní (Kennedy et al. 2014).

Dále je možná optimalizace metody z hlediska snížení koncentrace primerů, účinnější purifikace výsledného PCR produktu a snížení počtu cyklů PCR. Po testovacím sekvenování byla provedena tato opatření, což ve finálním sekvenování výrazně snížilo počet PCR

artefaktů. Aby se předešlo případnému převzorkování, bude vhodné pro další sekvenování ještě více snížit celkový počet cyklů PCR.

Chyby způsobené polymerázami a chyby v sekvenování tagů mohou vést k falešně vysokému počtu unikátních tagů. Brodinová a spolupracovníci seskupovali dohromady tagy, které se lišily jen o jednu bázi na základě předpokladu, že s největší pravděpodobností představují shodný tag se zanesenou chybou při sekvenaci (Brodin et al. 2015). V této práci nebyl zmíněný postup proveden. Tagy lišící se jednou bází mohou představovat skutečný rozdíl a seskupování sekvencí může tak vést k vnesení další chyby.

Ve třech vzorcích DNA z mitotických i meiotických buněk byly detekovány bodové mutace vyskytující se v pseudogenu. Frekvence *de novo* genové konverze mezi *GBA* a *GBAP* byla odhadnuta jako menší než 1 : 10 000. Frekvence konverze dřívější práce ze stejné laboratoře je u mutace L444P 1 : 200 a u mutace D409H 1 : 2 000 až 1 : 40 000 (Mrázová 2000). Naměřená frekvence u mutace L444P je ve srovnání s námi odhadnutou frekvencí poměrně vysoká vzhledem k počtu pacientů s mutacemi, což naznačuje možnou nepřesnost použitých metod. Frekvence *de novo* genové konverze mezi *CYP21* a *CYP21P* byla odhadnuta 1 : 1 000 až 1 : 100 000 (Tusié-Luna and White 1995). Počet námi nalezených potenciálně konvertovaných molekul odpovídá řádově odhadnuté frekvenci v předchozích studiích. Nízký počet nalezených mutací neumožňuje přesnější odhad frekvence.

Pro lepší odhad frekvence je třeba vyšší počet hodnotitelných sekvencí se shodným tagem. Pro zvýšení počtu cílů je možné použít většího množství výchozí gDNA. Ačkoli počet čtených sekvencí na vzorek byl přijatelný (stovky tisíc až více než milion sekvencí), byl přítomen vysoký počet alignmentů s unikátním tagem, kdy alignment tvořila jen jedna nebo dvě sekvence. Tento problém může částečně vyřešit vyšší počet degenerovaných bází v tagu, kterým budou také kompenzovány možné nepravidelnosti při syntéze primerů (Liang et al. 2014).

Nicméně tato studie využití NGS pro nalezení rekombinantních alel ověřila, že použití tagů umožňuje významně eliminovat chyby způsobené polymerázami a sekvenováním a je tak slibným přístupem. Použití náhodných sekvencí jako tagů může však vést jak k falešně negativním, tak i falešně pozitivním výsledkům (Liang et al. 2014). Tato studie není definitivní, ale jejím významným výsledkem je identifikace způsobu, jakým je třeba modifikovat laboratorní a bioinformatické postupy tak, aby bylo možné dosáhnout výrazně vyššího počtu hodnotitelných sekvencí. V tomto směru je jí nutné považovat za pilotní.

Studium *de novo* genové konverze má také určitý význam pro genetické poradenství. V rodinách, kde jeden rodič je pacientem s diagnostikovanou Gaucherovou chorobou nebo přenašečem pro patogenní mutaci, je možné upřesnit riziko vzniku onemocnění v důsledku vzniku *de novo* mutace konverzí u druhého z rodičů. V literatuře byla popsána pacientka s Gaucherovou chorobou a genotypem *RecNciI*/N370S. U rodičů dotyčné ženy nebyla alela *RecNciI* v buňkách periferní krve zaznamenána. Otec byl přenašečem mutace N370S. *RecNciI* alela tak vznikla *de novo* genovou konverzí v maternální zárodečné buňce (Alfonso, Pocovi and Giraldo 2008). Další studie pojednává o dvou pacientech s diagnostikovanou Gaucherovou chorobou z nepříbuzných rodin, u kterých byla zaznamenána mutace L444P, která vznikla *de novo* v maternální zárodečné buňce. V tomto případě šlo o bodovou mutaci, ne o komplexní rekombinantní alelu, jak tomu bylo v předchozím případě (Saranjam et al. 2013).

Závěr

- Technikou sekvenování nové generace využívající značení jednotlivých molekul DNA pomocí náhodných sekvencí v primeru (tagu) byly detekovány vzácné bodové mutace vznikající rekombinací mezi genem a pseudogenem pro β-glukocerebrosidasu. Výzkum byl prováděn na genomové DNA z 10 vzorků meiotických a 10 vzorků mitotických buněk zdravých osob.
- Využití tagů eliminovalo většinu chyb vzniklých při amplifikaci a sekvenování díky porovnání sekvencí z alignmentu nesoucích shodný tag. Poměr bází tagů nebyl zcela náhodný, možným vysvětlením jsou nepravidelnosti při syntéze primeru, preferenční amplifikace nebo převzorkování při malém počtu molekul templátu. Metoda neodliší mutace, které vznikly v časných fázích amplifikace.
- V jednotlivých vzorcích DNA bylo přítomno 7-15 tisíc alignmentů s unikátním tagem odpovídající předpokládaným původním molekulám templátu. Z malého počtu nalezených mutací lze odhadnout frekvenci genové konverze v devátém a desátém exonu genu pro β-glukocerebrosidasu menší než 1 : 10 000, což je nižší hodnota, než jaká byla detekována staršími metodami.
- Výsledky těchto pilotních experimentů vedou k návrhu metod, které umožní větší počet hodnotitelných sekvencí a tím i vyšší záchyt mutací.
- Na vzorcích genomové DNA se technikami založenými na principu zámkových sond, mutačně specifického PCR a LAMP nepodařilo obohacování konvertovaných molekul DNA pro přípravu templátu pro sekvenování nové generace

Seznam literatury

Alfonso, P., M. Pocovi & P. Giraldo (2008) Gaucher disease: report of de novo GBA mutation in a Spanish family. *Blood Cells, Molecules, and Diseases,* 40**,** 444-445.

Allers, T. & M. Lichten (2001) Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell,* 106**,** 47-57.

Barbera, M. A. & T. D. Petes (2006) Selection and analysis of spontaneous reciprocal mitotic cross-overs in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences,* 103**,** 12819-12824.

Barton, N. W., R. O. Brady, J. M. Dambrosia, A. M. Di Bisceglie, S. H. Doppelt, S. C. Hill, H. J. Mankin, G. J. Murray, R. I. Parker & C. E. Argoff (1991) Replacement therapy for inherited enzyme deficiency—macrophage-targeted glucocerebrosidase for Gaucher's disease. *New England Journal of Medicine,* 324**,** 1464-1470.

Baudat, F., J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop & B. De Massy (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science,* 327**,** 836-840.

Beutler, E., T. Gelbart & C. West (1993a) Identification of six new Gaucher disease mutations. *Genomics,* 15**,** 203-205.

Beutler, E., N. Nguyen, M. Henneberger, J. Smolec, R. McPherson, C. West & T. Gelbart (1993b) Gaucher disease: gene frequencies in the Ashkenazi Jewish population. *American journal of human genetics,* 52**,** 85.

Bischof, J. M., A. P. Chiang, T. E. Scheetz, E. M. Stone, T. L. Casavant, V. C. Sheffield & T. A. Braun (2006) Genome-wide identification of pseudogenes capable of disease-causing gene conversion. *Hum Mutat,* 27**,** 545-52.

Boria, I., E. Garelli, H. T. Gazda, A. Aspesi, P. Quarello, E. Pavesi, D. Ferrante, J. J. Meerpohl, M. Kartal & L. Da Costa (2010) The ribosomal basis of diamond-blackfan anemia: mutation and database update. *Human mutation,* 31**,** 1269-1279.

Brodin, J., C. Hedskog, A. Heddini, E. Benard, R. A. Neher, M. Mild & J. Albert (2015) Challenges with using primer IDs to improve accuracy of next generation sequencing. *PloS one,* 10**,** e0119123.

Carter, C., K. Akrami, D. Hall, D. Smith & E. Aronoff-Spencer (2017) Lyophilized visually readable loop-mediated isothermal reverse transcriptase nucleic acid amplification test for detection Ebola Zaire RNA. *Journal of Virological Methods,* 244**,** 32-38.

Charrow, J., H. C. Andersson, P. Kaplan, E. H. Kolodny, P. Mistry, G. Pastores, B. E. Rosenbloom, C. R. Scott, R. S. Wappner, N. J. Weinreb & A. Zimran (2000) The Gaucher registry: demographics and disease characteristics of 1698 patients with Gaucher disease. *Arch Intern Med,* 160**,** 2835-43.

Chuzhanova, N., J. M. Chen, A. Bacolla, G. P. Patrinos, C. Férec, R. D. Wells & D. N. Cooper (2009) Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. *Hum Mutat,* 30**,** 1189-98.

Cormand, B., A. Díaz, D. Grinberg, A. Chabás & L. Vilageliu (2000) A new gene-pseudogene fusion allele due to a recombination in intron 2 of the glucocerebrosidase gene causes Gaucher disease. *Blood Cells Mol Dis,* 26**,** 409-16.

Cortés-Ledesma, F. & A. Aguilera (2006) Double-strand breaks arising by replication through a nick are repaired by cohesin-dependent sister-chromatid exchange. *EMBO reports,* 7**,** 919-926.

Cox, T. M. & J. P. Schofield (1997) 3 Gaucher's disease: clinical features and natural history. *Bailliere's clinical haematology,* 10**,** 657-689.

Dahl, M., A. Doyle, K. Olsson, J. E. Månsson, A. R. Marques, M. Mirzaian, J. M. Aerts, M. Ehinger, M. Rothe, U. Modlich, A. Schambach & S. Karlsson (2015) Lentiviral Gene Therapy Using Cellular Promoters Cures Type 1 Gaucher Disease in Mice. *Mol Ther*.

Dreborg, S., A. Erikson & B. Hagberg (1980) Gaucher disease-norrbottnian type. *European journal of pediatrics,* 133**,** 107-118.

Edgar, R. C. & H. Flyvbjerg (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics***,** btv401.

Elliott, B., C. Richardson, J. Winderbaum, J. A. Nickoloff & M. Jasin (1998) Gene conversion tracts from double-strand break repair in mammalian cells. *Molecular and cellular biology,* 18**,** 93-101.

Esposito, M. S. (1978) Evidence that spontaneous mitotic recombination occurs at the two-strand stage. *Proceedings of the National Academy of Sciences,* 75**,** 4436-4440.

Eyal, N., S. Wilder & M. Horowitz (1990) Prevalent and rare mutations among Gaucher patients. *Gene,* 96**,** 277-83.

Ezawa, K., S. OOta & N. Saitou (2006) Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Molecular biology and evolution,* 23**,** 927-940.

Fabre, F. (1978) Induced intragenic recombination in yeast can occur during the G1 mitotic phase. *Nature,* 272**,** 795.

Fan, C. Y., T. K. Lian, W. Y. Ching, W. Y. Chui, Y. S. Fan & T. J. A. M. Anne (2001) The use of the amplification refractory mutation system (arms) in the detection of rare beta-thalassemia mutations in the Malays and Chinese in Malaysia.

Fuellgrabe, M. W., D. Herrmann, H. Knecht, S. Kuenzel, M. Kneba, C. Pott & M. Brüggemann (2015) High-throughput, amplicon-based sequencing of the CREBBP gene as a tool to develop a universal platform-independent assay. *PloS one,* 10**,** e0129195.

Galtier, N., G. Piganeau, D. Mouchiroud & L. Duret (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics,* 159**,** 907-11.

Geng, L., W. Xin, D.-W. Huang & G. Feng (2006) A universal cloning vector using vaccinia topoisomerase I. *Molecular biotechnology,* 33**,** 23-28.

Ginns, E. I., P. V. Choudary, S. Tsuji, B. Martin, B. Stubblefield, J. Sawyer, J. Hozier & J. A. Barranger (1985) Gene mapping and leader polypeptide sequence of human glucocerebrosidase: implications for Gaucher disease. *Proc Natl Acad Sci U S A,* 82**,** 7101-5.

Gordenin, D. & M. Resnick (1998) Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 400**,** 45-58.

Halldorsson, B. V., M. T. Hardarson, B. Kehr, U. Styrkarsdottir, A. Gylfason, G. Thorleifsson, F. Zink, A. Jonasdottir, A. Jonasdottir & P. Sulem (2016) The rate of meiotic gene conversion varies by sex and age. *Nature Genetics*.

Hodaňová, K., M. Hřebíček, M. Červenková, L. Mrázová, L. Vepřeková & J. Zeman (1999) Analysis of the β-glucocerebrosidase gene in Czech and Slovak Gaucher patients: mutation profile and description of six novel mutant alleles. *Blood Cells, Molecules, and Diseases,* 25**,** 287-298.

Horowitz, M., S. Wilder, Z. Horowitz, O. Reiner, T. Gelbart & E. Beutler (1989) The human glucocerebrosidase gene and pseudogene: structure and evolution. *Genomics,* 4**,** 87-96.

Hrdlickova, R., A. Lewis & J. Nehyba (2016) Designing Amplicon Panels. *Clinical OMICs,* 3**,** 25-27.

Iwasaki, M., T. Yonekawa, K. Otsuka, W. Suzuki, K. Nagamine, T. Hase, K.-I. Tatsumi, T. Horigome, T. Notomi & H. Kanda (2003) Validation of the loop-mediated isothermal

amplification method for single nucleotide polymorphism genotyping with whole blood. *Genome letters,* 2**,** 119-126.

Jabara, C. B., C. D. Jones, J. Roach, J. A. Anderson & R. Swanstrom (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences,* 108**,** 20166-20171.

Jacq, C., J. Miller & G. Brownlee (1977) A pseudogene structure in 5S DNA of Xenopus laevis. *Cell,* 12**,** 109-120.

Jeffreys, A. J. & C. A. May (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet,* 36**,** 151-6.

Jinks-Robertson, S., M. Michelitch & S. Ramcharan (1993) Substrate length requirements for efficient mitotic recombination in Saccharomyces cerevisiae. *Molecular and Cellular Biology,* 13**,** 3937-3950.

Kattlove, H. E., J. C. Williams, E. Gaynor, M. Spivack, R. M. Bradley & R. O. Brady (1969) Gaucher cells in chronic myelocytic leukemia: an acquired abnormality. *Blood,* 33**,** 379-390.

Kennedy, S. R., M. W. Schmitt, E. J. Fox, B. F. Kohrn, J. J. Salk, E. H. Ahn, M. J. Prindle, K. J. Kuong, J.-C. Shen & R.-A. Risques (2014) Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature protocols,* 9**,** 2586-2606.

Kinde, I., J. Wu, N. Papadopoulos, K. W. Kinzler & B. Vogelstein (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences,* 108**,** 9530-9535.

Klein, F., P. Mahr, M. Galova, S. B. Buonomo, C. Michaelis, K. Nairz & K. Nasmyth (1999) A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. *Cell,* 98**,** 91-103.

Kou, R., H. Lam, H. Duan, L. Ye, N. Jongkam, W. Chen, S. Zhang & S. Li (2016) Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. *PloS one,* 11**,** e0146638.

Krishnakumar, S., J. Zheng, J. Wilhelmy, M. Faham, M. Mindrinos & R. Davis (2008) A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proceedings of the National Academy of Sciences,* 105**,** 9296-9301.

Larsson, C., J. Koch, A. Nygren, G. Janssen, A. K. Raap, U. Landegren & M. Nilsson (2004) In situ genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. *Nature methods,* 1**,** 227-232.

Li, J.-j., C. Xiong, Y. Liu, J.-s. Liang & X.-w. Zhou (2016) Loop-Mediated Isothermal Amplification (LAMP): Emergence As an Alternative Technology for Herbal Medicine Identification. *Frontiers in Plant Science,* 7.

Liang, R. H., T. Mo, W. Dong, G. Q. Lee, L. C. Swenson, R. M. McCloskey, C. K. Woods, C. J. Brumme, C. K. Ho & J. Schinkel (2014) Theoretical and experimental assessment of degenerate primer tagging in ultra-deep applications of next-generation sequencing. *Nucleic acids research***,** gku355.

Lichten, M., R. H. Borts & J. E. Haber (1987) Meiotic gene conversion and crossing over between dispersed homologous sequences occurs frequently in Saccharomyces cerevisiae. *Genetics,* 115**,** 233-246.

Lilley, D. (1980) The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proceedings of the National Academy of Sciences,* 77**,** 6468-6472.

Linardopoulou, E. V., E. M. Williams, Y. Fan, C. Friedman, J. M. Young & B. J. Trask (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature,* 437**,** 94-100.

Little, S. (2001) Amplification-Refractory Mutation System (ARMS) Analysis of Point Mutations. *Current protocols in human genetics***,** 9.8. 1-9.8. 12.

Liu, M.-L., Y. Xia, X.-Z. Wu, J.-Q. Huang & X.-G. Guo (2017) Loop-mediated isothermal amplification of Neisseria gonorrhoeae porA pseudogene: a rapid and reliable method to detect gonorrhea. *AMB Express,* 7**,** 48.

L'allemand, D., V. Tardy, A. Gruters, D. Schnabel, H. Krude & Y. Morel (2000) How a patient homozygous for a 30-kb deletion of the C4-CYP 21 genomic region can have a nonclassic form of 21-hydroxylase deficiency. *The Journal of Clinical Endocrinology & Metabolism,* 85**,** 4562-4567.

Maestre, J., T. Tchenio, O. Dhellin & T. Heidmann (1995) mRNA retroposition in human cells: processed pseudogene formation. *The EMBO journal,* 14**,** 6333.

Manley, L. J., D. Ma & S. S. Levine (2016) Monitoring Error Rates In Illumina Sequencing. *Journal of Biomolecular Techniques: JBT,* 27**,** 125.

McBride, T., J. Schneider, R. Floyd & L. A. Loeb (1992) Mutations induced by methylene blue plus light in single-stranded M13mp2. *Proceedings of the National Academy of Sciences,* 89**,** 6866-6870.

Migdalska-Richards, A., W. K. D. Ko, Q. Li, E. Bezard & A. H. Schapira (2017) Oral ambroxol increases brain glucocerebrosidase activity in a nonhuman primate. *Synapse*.

Mori, Y., K. Nagamine, N. Tomita & T. Notomi (2001) Detection of loop-mediated isothermal amplification reaction by turbidity derived from magnesium pyrophosphate formation. *Biochemical and biophysical research communications,* 289**,** 150-154.

Morrow, D. M., C. Connelly & P. Hieter (1997) "Break copy" duplication: a model for chromosome fragment formation in Saccharomyces cerevisiae. *Genetics,* 147**,** 371-382.

Mrázová, L. (2000). Genová converse mezi genem a pseudogenem pro β-glukocerebrosidasu. Praha. Diplomová práce. Vysoká škola chemicko-technologická v Praze, Fakulta potravinářské a biochemické technologie, Ústav biochemie a mikrobiologie.

Myers, S., L. Bottolo, C. Freeman, G. McVean & P. Donnelly (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science,* 310**,** 321-4.

Mézard, C., D. Pompon & A. Nicolas (1992) Recombination between similar but not identical DNA sequences during yeast transformation occurs within short stretches of identity. *Cell,* 70**,** 659-670.

Notomi, T., H. Okayama, H. Masubuchi, T. Yonekawa, K. Watanabe, N. Amino & T. Hase (2000) Loop-mediated isothermal amplification of DNA. *Nucleic acids research,* 28**,** e63-e63.

Ohta, T. (1999) Effect of gene conversion on polymorphic patterns at major histocompatibility complex loci. *Immunological reviews,* 167**,** 319-325.

Podlaha, O. & J. Zhang (2010) Pseudogenes and their evolution. *eLS*.

Poliseno, L. (2012) Pseudogenes: newly discovered players in human cancer. *Sci Signal,* 5**,** 5.

Poliseno, L., L. Salmena, J. Zhang, B. Carver, W. J. Haveman & P. P. Pandolfi (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature,* 465**,** 1033-1038.

Potapov, V. & J. L. Ong (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing. *PloS one,* 12**,** e0169774.

Pratto, F., K. Brick, P. Khil, F. Smagulova, G. V. Petukhova & R. D. Camerini-Otero (2014) Recombination initiation maps of individual human genomes. *Science,* 346**,** 1256442.

Quail, M. A., I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow & D. J. Turner (2008) A large genome center's improvements to the Illumina sequencing system. *Nature methods,* 5**,** 1005-1010.

Reiner, O. & M. Horowitz (1988) Differential expression of the human glucocerebrosidase-coding gene. *Gene,* 73**,** 469-478.

Reiter, L. T., P. J. Hastings, E. Nelis, P. De Jonghe, C. Van Broeckhoven & J. R. Lupski (1998) Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *The American Journal of Human Genetics,* 62**,** 1023-1033.

Rockmill, B., K. Voelkel-Meiman & G. S. Roeder (2006) Centromere-proximal crossovers are associated with precocious separation of sister chromatids during meiosis in Saccharomyces cerevisiae. *Genetics,* 174**,** 1745-1754.

Roumelioti, F. M., S. K. Sotiriou, V. Katsini, M. Chiourea, T. D. Halazonetis & S. Gagos (2016) Alternative lengthening of human telomeres is a conservative DNA replication process with features of break-induced replication. *EMBO reports,* 17**,** 1731-1737.

Rygiel, A. M., S. Beer, P. Simon, K. Wertheim-Tysarowska, G. Oracz, T. Kucharzik, A. Tysarowski, K. Niepokój, J. Kierkus & M. Jurek (2015) Gene conversion between cationic trypsinogen (PRSS1) and the pseudogene trypsinogen 6 (PRSS3P2) in patients with chronic pancreatitis. *Human mutation,* 36**,** 350-356.

Saranjam, H., S. S. Chopra, H. Levy, B. K. Stubblefield, E. Maniwang, I. J. Cohen, H. Baris, E. Sidransky & N. Tayebi (2013) A germline or de novo mutation in two families with Gaucher disease: implications for recessive disorders. *European Journal of Human Genetics,* 21**,** 115-117.

Sarhadi, V., A. Reis, M. Jung, D. Singh, K. Sperling, J. R. Singh & J. Bürger (2001) A unique form of autosomal dominant cataract explained by gene conversion between β-crystallin B2 and its pseudogene. *Journal of medical genetics,* 38**,** 392-396.

Schildkraut, E., C. A. Miller & J. A. Nickoloff (2005) Gene conversion and deletion frequencies during double-strand break repair in human cells are controlled by the distance between direct repeats. *Nucleic acids research,* 33**,** 1574-1580.

Schmitt, M. W., S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt & L. A. Loeb (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences,* 109**,** 14508-14513.

Shen, P. & H. V. Huang (1986) Homologous recombination in Escherichia coli: dependence on substrate length and homology. *Genetics,* 112**,** 441-457.

Shen, P., W. Wang, A.-K. Chi, Y. Fan, R. W. Davis & C. Scharfe (2013) Multiplex target capture with double-stranded DNA probes. *Genome medicine,* 5**,** 50.

Shen, P., W. Wang, S. Krishnakumar, C. Palm, A.-K. Chi, G. M. Enns, R. W. Davis, T. P. Speed, M. N. Mindrinos & C. Scharfe (2011) High-quality DNA sequence capture of 524 disease candidate genes. *Proceedings of the National Academy of Sciences,* 108**,** 6549-6554.

Smid, B. E., M. J. Ferraz, M. Verhoek, M. Mirzaian, P. Wisse, H. S. Overkleeft, C. E. Hollak & J. M. Aerts (2016) Biochemical response to substrate reduction therapy versus enzyme replacement therapy in Gaucher disease type 1 patients. *Orphanet journal of rare diseases,* 11**,** 28.

Stensmyr, M. C. (2016) Evolutionary Genetics: Smells like a Pseudo-pseudogene. *Current Biology,* 26**,** R1294-R1296.

Stern, C. (1936) Somatic crossing over and segregation in Drosophila melanogaster. *Genetics,* 21**,** 625.

Stone, D. L., N. Tayebi, E. Orvisky, B. Stubblefield, V. Madike & E. Sidransky (2000) Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. *Hum Mutat,* 15**,** 181-8.

Su, Y., A. B. Barton & D. B. Kaback (2000) Decreased meiotic reciprocal recombination in subtelomeric regions in Saccharomyces cerevisiae. *Chromosoma,* 109**,** 467-475.

Svendsen, J. M. & J. W. Harper (2010) GEN1/Yen1 and the SLX4 complex: Solutions to the problem of Holliday junction resolution. *Genes & development,* 24**,** 521-536.

Szemes, M., P. Bonants, M. de Weerdt, J. Baner, U. Landegren & C. D. Schoen (2005) Diagnostic application of padlock probes—multiplex detection of plant pathogens using universal microarrays. *Nucleic acids research,* 33**,** e70-e70.

Taghian, D. G. & J. A. Nickoloff (1997) Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells. *Molecular and Cellular Biology,* 17**,** 6386-6393.

Tayebi, N., B. K. Stubblefield, J. K. Park, E. Orvisky, J. M. Walker, M. E. LaMarca & E. Sidransky (2003) Reciprocal and nonreciprocal recombination at the glucocerebrosidase gene region: implications for complexity in Gaucher disease. *Am J Hum Genet,* 72**,** 519-34.

Tease, C., G. M. Hartshorne & M. A. Hultén (2002) Patterns of meiotic recombination in human fetal oocytes. *The American Journal of Human Genetics,* 70**,** 1469-1479.

Tsuji, S., P. V. Choudary, B. M. Martin, B. K. Stubblefield, J. A. Mayor, J. A. Barranger & E. I. Ginns (1987) A mutation in the human glucocerebrosidase gene in neuronopathic Gaucher's disease. *New England Journal of Medicine,* 316**,** 570-575.

Tusié-Luna, M. T. & P. C. White (1995) Gene conversions and unequal crossovers between CYP21 (steroid 21-hydroxylase gene) and CYP21P involve different mechanisms. *Proc Natl Acad Sci U S A,* 92**,** 10796-800.

Velayati, A., M. A. Knight, B. K. Stubblefield, E. Sidransky & N. Tayebi (2011) Identification of Recombinant Alleles Using Quantitative Real-Time PCR: Implications for Gaucher Disease. *The Journal of Molecular Diagnostics,* 13**,** 401-405.

Winfield, S. L., N. Tayebi, B. M. Martin, E. I. Ginns & E. Sidransky (1997) Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease. *Genome Res,* 7**,** 1020-6.

Yim, E., K. E. O'Connell, J. St Charles & T. D. Petes (2014) High-resolution mapping of two types of spontaneous mitotic gene conversion events in Saccharomyces cerevisiae. *Genetics,* 198**,** 181-92.

Yongkiettrakul, S., J. Kampeera, W. Chareanchim, R. Rattanajak, W. Pornthanakasem, W. Kiatpathomchai & D. Kongkasuriyachai (2017) Simple detection of single nucleotide polymorphism in Plasmodium falciparum by SNP-LAMP assay combined with lateral flow dipstick. *Parasitology International,* 66**,** 964-971.

Zhang, Z. & M. Gerstein (2004) Large-scale analysis of pseudogenes in the human genome. *Current opinion in genetics & development,* 14**,** 328-335.

Zhou, H., H. Wang, C. Liu, H. Wang, X. Duan & Z. Li (2015) Ultrasensitive genotyping with target-specifically generated circular DNA templates and RNA FRET probes. *Chemical Communications,* 51**,** 11556-11559.

Zimran, A., J. Sorge, E. Gross, M. Kubitz, C. West & E. Beutler (1990) A glucocerebrosidase fusion gene in Gaucher disease. Implications for the molecular anatomy, pathogenesis, and diagnosis of this disorder. *J Clin Invest,* 85**,** 219-22.

Příloha 1

Skript pro filtrování a třídění sekvencí se shodným tagem

```
#! /usr/bin/perl -w

# v. 0.01 M. Hřebíček, B. Peková, L. Mrázová


$filename = $ARGV[0];
$filename =~ /([A-Za-z0-9_]+)\.\w/;

$basename = "$1";
$summaryheader = "basename";
$summary = "$basename";

$conminseq = $ARGV[1];                # minimal number of sequences needed for valid consensus
$mislimit = $ARGV[2];                 # max numberf of mismatches in R1 primer

# counters
$linecounter = 0;                    # counts lines in fastq reads
$readcounter1 = 0;                   # R1 counts reads in fastq file
$readcounter2 = 0;                   # R2 counts reads in fastq file
$taggedcounter1 = 0;                 # number of correctly tagged reads
$taggedcounter2 = 0;                 # number of correctly tagged reads
$untaggedcounter1 = 0;               # R1 number of garbage reads
$untaggedcounter2 = 0;               # R2 number of garbage reads
$shortreadcounter1 = 0;              # R1 reads shorter than $readlenlimit
$shortreadcounter2 = 0;              # R1 reads shorter than $readlenlimit
$headercounter = 0;
$r1emptygenocounter = 0;             # R1 reads, where there we no difference from the consensus
$r2emptygenocounter = 0;             # R2 reads, where there we no difference from the consensus
$r1validbincounter = 0;                # how many tag bins have higher number of reads than
minimal - conminseq
$r2validbincounter = 0;                # how many tag bins have higher number of reads than
minimal - conminseq

# read parameters
$read1ref = "GTAAAACGACGGCCAGTGNNNNNNNNNNTTGGGTGCGTAACTTTGTCG";
$read1from1 = 0;                    # first constant part of the primer limits; counts from 0,
not 1
$read1to1 = 17;
$read1from2 = 37;                   # second constant part of the primer limits; counts from 0,
not 1
$read1to2 = 47;
# $mislimit = 2;                      # tolerated number of mismatches in primer sequnces
$taglength = 10;
$readlenlimit = 100;                # minimal read length
$r1trim = 0;                        # trim primer sequences
$r1trimlimit = 47;                  # trim sequence from 0 to the limit
$r2trimlimit = 47;
$seqcycles = 260;                   # run length set at the sequencer
$r1gbaseqshift = 6814 - 38;         # add this to read position to get the position in GBA
sequence

# difference calling
$r1diffrom = 48;                    # call differences only from this position in the read
$r1diffto = 120;                    # call differences only to this position in the read
$r2diffrom = 20;                    # call differences only from this position in the read
$r2diffto = 120;                    # call differences only to this position in the read

$r1difflimit = 0;                   # report difference between the ref and consensus only if
there are more diffs than this
$r2difflimit = 0;



# control variables
#$conminseq = 1;                  # minimal number of sequences needed for valid consensus
$conresult = "";                 # string sums up properties of consensus sequence
```

```
#  quality filtering parameters
#  int codes for ascii error codes and corresponding probabilities for Illumina 1.18+ and
Phred+33
%asciip = qw(33 1 34 0.79433 35 0.63096 36 0.50119 37 0.39811 38 0.31623 39 0.25119 40 0.19953
41 0.15849 42 0.12589 43 0.1 44 0.07943 45 0.0631 46 0.05012 47 0.03981 48 0.03162 49 0.02512
50 0.01995 51 0.01585 52 0.01259 53 0.01 54 0.00794 55 0.00631 56 0.00501 57 0.00398 58
0.00316 59 0.00251 60 0.002 61 0.00158 62 0.00126 63 0.001 64 0.00079 65 0.00063 66 0.0005 67
0.0004 68 0.00032 69 0.00025 70 0.0002 71 0.00016 72 0.00013 73 0.0001 74 0.00008);

$r1qlimit = 1;                  # number of expected errors in the
$r1qfrom = 47;                  # calculate expected error in the region from
$r1qto = 147;                   # calculate expected error in the region ending at

$r2qlimit = 1;                  # the same for read2
$r2qfrom = 20;
$r2qto = 120;

# length filtering parameters
$r1minlength = 249;
$r2minlength = 249;

# filtering by similarity to reference sequence parameters
$r1maxmismatch = 7;
$r1from = 47;
$r1to = 147;

$r2maxmismatch = 10;
$r2from = 20;
$r2to = 120;


# files
$fastq1 = "$basename.fastq";
$read1inhand = "READ1IN";
open ($read1inhand, $fastq1) or die "Could not open $fastq1:$!\n";

$readtagdist1 = $basename.".readtagdist1.dat";                                  #
distribution of reads per tag
open (RDIST1, ">$readtagdist1") or die "Could not open $readtagdist1:$!\n";

$readtagdist2 = $basename.".readtagdist2.dat";                                  #
distribution of reads per tag
open (RDIST2, ">$readtagdist2") or die "Could not open $readtagdist2:$!\n";

$read1other = $basename.".read1other.fq";
open (READ1OTHER, ">$read1other") or die "Could not open $read1other:$!\n";

$tagfile = "all_tags.txt";
open (TAGS, ">>$tagfile") or die "Could not open $tagfile:$!\n";

$gnuplotfile = "gnuplotcommands.txt";                                           #
commands for printing of charts by gnuplot
open (GPLOT, ">>$gnuplotfile") or die "Could not open $gnuplotfile:$!\n";

$tablefile = "table.tsv";                                                       #
produces table with run parameters
open (TABLE, ">>$tablefile") or die "Could not open $tablefile:$!\n";

$read2in = "$basename.fastq";
$read2in =~ s/_R1_/_R2_/;
$read2inhand = "READ2IN";
open ($read2inhand, "$read2in") or die "Could not open $read2in:$!\n";

$read2out = $basename."read2out.fq";
open (READ2OUT, ">$read2out") or die "Could not open $read2out:$!\n";

$read2other = $basename.".read2other.fq";
open (READ2OTHER, ">$read2other") or die "Could not open $read2other:$!\n";

$analysis = $basename.".analysis.txt";
open (ANAL, ">$analysis") or die "Could not open $analysis:$!\n";

$diffdotalign1 = $basename.".diffdotalign1.txt";                                #
reads forming alignments with consensus bases differing from reference
open (DIFALIGN1, ">$diffdotalign1") or die "Could not open $diffdotalign:$!\n";
```

```
$diffdotalign2 = $basename.".diffdotalign2.txt";                                    #
reads forming alignments with consensus bases differing from reference
open (DIFALIGN2, ">$diffdotalign2") or die "Could not open $diffdotalign:$!\n";


# read reference sequence from plain text file
# NOTE tag bases have to to be written as Ns in the sequence for correct counting of
differences from the reference
$seqref = "delE10productB.txt";
open (SEQREF, "$seqref") or die "Could not open $seqref:$!\n";

# read reference sequence as plain text
$reference = "";
  while ($line = <SEQREF>) {
   chomp($line);
   $reference = $reference.$line;
  }
$reference = uc($reference);                              # convert to uppercase
$revreference = revcom($reference);                       # reverse and complement
$reference = substr($reference, 0, $seqcycles);           # cut reference to the length of the
run
$revreference = substr($revreference, 0, $seqcycles);   # cut reference to the length of the
run

# hashes
%r1tagged = ();                   # hash with sequences ; tag is the key, value is a reference
to 2d array with reads
%r1header = ();                   # hash: key is the shortened header, value is the tag
%r2tagged = ();                   # hash with sequences ; tag is the key, value is a reference
to 2d array with reads
%tag1genotype = ();               # hash holding tags as keys and string with R1 genotype
summary as values; used for matching genotypes with same tags
%r1variants = ();
%r2variants = ();
%tagnumreads = ();                # number of tags with the same number of reads; number of
reads is the key


# no changeable parameters below this line
*********************************************************************************

print ANAL "Run: $basename\n\nTolerated number of mismatches in primer sequences:
$mislimit\nTrim : $r1trim\nShort read length limit: $readlenlimit\nMinimal number of sequences
needed for valid consensus: $conminseq\n";
print ANAL "\nRead 1 file: $fastq1\n";

$summaryheader = $summaryheader."\tMax primer mismatch\tMin con seqs";
$summary = $summary."\t$mislimit\t$conminseq";


@read = ();

# read Read 1 fastq file into hashes
$endfile = 0;

do {{
   ($endfile, @read) = getfqread ($read1inhand);          # read one fastq read
    $readcounter1++;                                      # increase read counter

    if ( filter($read[1], $read[3], $reference, $r1minlength, $r1maxmismatch, $r1from, $r1to,
$r1qfrom, $r1qto, $r1qlimit) == 0) { next;}

    if (tagged($read[1]) == 1) {
    #process tagged reads
         $tag = substr($read[1],$read1to1 + 1, $taglength);
         print TAGS "$tag\n";
         $taggedcounter1++;
         if (length($read[1]) < $readlenlimit) {$shortreadcounter1++;}

         # add read to r1tagged hash; tag is the key, the value is array of arrays
         push @{$r1tagged{$tag}}, [ @read ];                  # pushing anonymous array
reference to the read into the array

         # add the part of the header (before the space) as a key and tag as value to
%r1header hash
```

```
         if ($read[0] =~ /^(\S+)\s/) {                                    # collect non-white
header text before the space
         if (exists($r1header{$1})) {print ANAL "\nRead 1 Duplicate header $1\n";}
           else {$r1header{$1} = $tag;}
          }
       }
     else {
      # process untagged reads - print them to a fastq file
            print READ1OTHER "$read[0]\n$read[1]\n$read[2]\n$read[3]\n";
            $untaggedcounter1++;
          }

}} until ($endfile == 1);

print ANAL "Read 1 reads: $readcounter1\nRead 1 filtered tagged reads: $taggedcounter1\nRead 1
non-tagged reads: $untaggedcounter1\nRead 1 reads shorter than $readlenlimit:
$shortreadcounter1\nTag number: ".scalar(keys(%r1tagged))."\n\n";

$summaryheader = $summaryheader."\tR1 reads\tR1 fitered tagged\tR1 nontagged reads\tTags";
$summary =
$summary."\t$readcounter1\t$taggedcounter1\t$untaggedcounter1\t".scalar(keys(%r1tagged));

# call difference between the reference and consensus sequences
$taggedcounter1 = 0;
$diffcallcounter =0;

while(($tag, $reads) = each(%r1tagged)) {         # $reads hold a reference to 2D array with
reads (sharing the same tag)
   $taggedcounter1++;

  # Count how many tags  have 1, 2, 3 ,  etc. reads; Add numbers to a hash
  if (exists ($tagnumreads{@$reads})) {$tagnumreads{@$reads} = $tagnumreads{@$reads} + 1;}
# add the number of reads in @reads to hash
   else {$tagnumreads{@$reads} =  1;}

   # Determine consensus
  ($cc,$conresult) = consensusmaker($reads, 250, 0, $conminseq);
  if ($conresult == 1) {
   $r1validbincounter++;
   ($diff, $diffc) = diffcaller($reference, $cc, $r1diffrom, $r1diffto);

   if ($diff =~ /\w+/) { $diff =~ s/\s+//;} else { $diff = " ";}

   $tag1genotype{$tag} = $diff.":";                    # add string with diffrences to the
hash, tag is the key.

   if ($diffc > $r1difflimit) {                    # if more than tolerated number of
differences
     $diffcallcounter++;
     $qq = dottedconsensus($reads,$reference,$cc);
     print DIFALIGN1 "R1 $tag $diffc $diff\n$qq\n\n";
  }
 }
}
print ANAL "Called consensus for Read 1 reads ...\nRead 1 tagged reads: $taggedcounter1\nRead
1 consensus reads  with variant sequence: $diffcallcounter\n";

$summaryheader = $summaryheader."\tR1 con reads\tR1 con reads with variants\tR1 valid bins";
$summary = $summary."\t$taggedcounter1\t$diffcallcounter\t$r1validbincounter";

# Now print numbers of tags with different numbers of reads
print RDIST1 "##Reads_per_tag Tags Reads\n";
foreach $n (sort { $a <=> $b } keys (%tagnumreads)) { print RDIST1 "$n ".$tagnumreads{$n}."
".$n*$tagnumreads{$n}."\n";}

# Erase the $tagnumreads hash
%tagnumreads = ();

# process Read 2 reads
# read the fastq file into a hash with tags read from r1header hash as keys; same as Read1
only tags come from the hash

$endfile = 0;
do {{
   ($endfile, @read) = getfqread ($read2inhand);
```

```
   # filter : seq .. sequence, qual .. quality line, ref .. reference seq, minlength ..
minimal length, maxmismatch .. max of tolerated mismatches,
   # from  .. mismatch check from base, to .. dtto, qfrom .. quality dtto, qto .. dtto, qlimit
.. limit on expected number of errors
   if ( filter( $read[1], $read[3], $revreference, $r2minlength, $r2maxmismatch, $r2from,
$r2to, $r2qfrom, $r2qto, $r2qlimit) == 0) {next;}
    $readcounter2++;                                          # increase read counter

   $read[0] =~ /^(\S+)\s/;                                    # collect part of header text
before the space
   if (exists($r1header{$1})) {                               # if the part of the header
before the space exists in header hash, get coresponding tag
    #process tagged reads
         $tag = $r1header{$1};
         #print "$tag\n";
         $taggedcounter2++;
         if (length($read[1]) < $readlenlimit) {$shortreadcounter2++;}
         # trim sequence and quality lines
         if ($r1trim == 1) {
           $read[1] = substr($read[1], $r2trimlimit + 1, length($read[1]) - $r2trimlimit-1);
# trim sequence
           $read[3] = substr($read[3], $r2trimlimit + 1, length($read[3]) - $r2trimlimit-1);
# trim quality
          }

         # add read to r2tagged hash; tag is the key, the value is array of arrays
         push @{$r2tagged{$tag}}, [ @read ];                  # pushing anonymous array
reference to the read into the array
      }
    else {
     # process untagged reads - print them to a fastq file
          print READ2OTHER "$read[0]\n$read[1]\n$read[2]\n$read[3]\n";
          $untaggedcounter2++;
         }

}} until ($endfile == 1);

print ANAL "Read 2 reads: $readcounter2\nRead 2 tagged reads: $taggedcounter2\nRead 2 non-
tagged reads: $untaggedcounter2\nRead 2 reads shorter than $readlenlimit:
$shortreadcounter2\nTag number: ".scalar(keys(%r2tagged))."\n\n";


$summaryheader = $summaryheader."\tR2 reads\tR2 fitered tagged\tR2 nontagged reads\tTags";
$summary =
$summary."\t$readcounter2\t$taggedcounter2\t$untaggedcounter2\t".scalar(keys(%r2tagged));


# call differences between the reference and consensus sequences

$diffcallcounter =0;
while(($tag, $reads) = each(%r2tagged)) {        # reads hold a reference to 2D array with
reads (sharing the same tag)
  $taggedcounter2++;

  # Count how many tags  have 1, 2, 3 ,  etc. reads; Add numbers to a hash
  if (exists ($tagnumreads{@$reads})) {$tagnumreads{@$reads} = $tagnumreads{@$reads} + 1;}
# add the number of reads in @reads to hash
   else {$tagnumreads{@$reads} =  1;}

  ($cc,$conresult) = consensusmaker($reads, 250, 0, $conminseq);
  if ($conresult == 1) {             # if the number of reads in consensus is lower than
conminseq
   $r2validbincounter++;
   ($diff, $diffc) = diffcaller($revreference, $cc, $r2diffrom, $r2diffto);

   if ($diff =~ /\w+/) { $diff =~ s/\s+//;} else { $diff = " ";}

   if (exists($tag1genotype{$tag})) {$tag1genotype{$tag} = $tag1genotype{$tag}.$diff;}  # R2
diffcodes are separated by ":"
   else {print ANAL "Error: Tag $tag from r2tagged hash not found in tag1genotype\n";}

   if ($diffc > r2difflimit) {                                    # if more than tolerated
number of differences
     $diffcallcounter++;
     $qq = dottedconsensus($reads,$revreference,$cc);
```

```
      print DIFALIGN2 "R2 $tag $diffc $diff\n$qq\n\n";
    }
   }
}
print ANAL "Called consensus for Read 2 reads ...\nRead 2 tagged reads: $taggedcounter2\nRead
2 consensus reads with variant sequence: $diffcallcounter\n";
$summaryheader = $summaryheader."\tR2 con reads\tR2 con reads with variants\tR2 valid bins";
$summary = $summary."\t$taggedcounter2\t$diffcallcounter\t$r2validbincounter";

# Now print numbers of tags with different numbers of reads
print RDIST2 "##Reads_per_tag Tags Reads\n";
foreach $n (sort { $a <=> $b } keys (%tagnumreads)) { print RDIST2 "$n ".$tagnumreads{$n}."
".$n*$tagnumreads{$n}."\n";}


# sort and count all variants

while(($tag, $genotype) = each(%tag1genotype)) {
   ($r1,$r2) =  split(/:/,$genotype);
   #print "$genotype  $tag \n";


   if (($r1 eq "") || ($r1 eq " ") || ($r1 eq "  ")) { $r1emptygenocounter++;}
   else {
         @r1var = split(/-/,$r1);
         foreach my $v (@r1var) {
         if (exists($r1variants{$v})) { $r1variants{$v} = $r1variants{$v} + 1;}
         else {$r1variants{$v} = 1;}
        }
       }

   if (($r2 eq "") || ($r2 eq " ") || ($r2 eq "  ")) { $r2emptygenocounter++;}
   else {
         @r2var = split(/-/,$r2);
         foreach $v (@r2var) {
         if (exists($r2variants{$v})) { $r2variants{$v} = $r2variants{$v} + 1;}
            else {$r2variants{$v} = 1;}
         }
     }
 }

print ANAL "Reads without variants in genotype hash : R1 $r1emptygenocounter, R2
$r2emptygenocounter\n";

print ANAL "\n$summaryheader\n$summary\n";
print TABLE "$summary\n";

print ANAL "\nRead 1 variants:\n";
 foreach $q (sort {$r1variants{$a} <=> $r1variants{$b}} keys %r1variants) {
    print ANAL "$q\t".$r1variants{$q}."\t".r1genomic($q)."\n";
 }

print ANAL "\nRead 2 variants:\n";
foreach $q (sort {$r2variants{$a} <=> $r2variants{$b}} keys %r2variants) {
    print ANAL "$q\t".$r2variants{$q}."\t".r2genomic($q)."\n";
 }

 # print graphs using gnuplot
 print "gnuplot -c test.pg \"R1 - $basename\"  \"$readtagdist1\"  > R1$basename.png\n";
 system "gnuplot -c test.pg \"R1 - $basename\"  \"$readtagdist1\"  \> R1$basename.png";
 print "gnuplot -c test.pg \"R2 - $basename\"  \"$readtagdist2\"  > R2$basename.png\n";
 system "gnuplot -c test.pg \"R2 - $basename\"  \"$readtagdist2\"  \> R2$basename.png";

close $read1inhand;
close RDIST1;
close RDIST2;
close READ1OTHER;
close $read2inhand;
#close READ2OUT;
close READ2OTHER;
close SEQREF;
close ANAL;
close DIFALIGN1;
close DIFALIGN2;
close TAGS;
close GPLOT;
```

```
CLOSE TABLE;

# END
*******************************************************************************************


sub  r1genomic {
  my ( $variant ) = @_ ;
  my $warning = "";
     $variant =~ /(\d+)([ACGTNacgtn])/;
  my $pos = $1;
  my $base = $2;
     if (($pos == 77) && ($base eq "C")) {$warning = "\tD409H";}
     $pos =  6767 + $pos;
     return "$pos$base$warning";
}

sub  r2genomic {
  my ( $variant ) = @_ ;
  my $warning = "";
     $variant =~ /(\d+)([ACGTNacgtn])/;
  my $pos = $1;
  my $base = $2;
     if (($pos == 34) && ($base eq "C")) {$warning = "\tV460V";}
     if (($pos == 48) && ($base eq "C")) {$warning = "\tA456P";}
     if (($pos == 83) && ($base eq "C")) {$warning = "\tL444P";}
     $pos =  6767 + $pos;
     $base =~  tr/ACGTNacgta/TGCANtgcan/;
     return "$pos$base$warning";
}

# filter
*******************************************************************************************
# filters sequence by various criteria, if sequencce passes the filter it returns 1, otherwise
0
# newly added filtering by quality
# seq .. sequence, qual .. quality, ref .. reference, minlength .. minimal length, maxmismatch
.. max of tolerated mismatches,
# from  .. mismatch check from base, to .. dtto, qfrom .. quality dtto, qto .. dtto
sub filter {
my ($seq, $qual, $ref, $minlength, $maxmismatch, $from, $to, $qfrom, $qto, $qlimit) = @_;

  $ result = 1;

  # filtering by length
  # if (length($seq) < $minlength) { $result = 0;}
  # filtering by similarity to reference sequence
  if (length($seq) < $to) { $to = length($seq);}                  # if the read sequence is
shorter than the limit, cut the limit
  if (seqfit($ref,$seq,$from,$to,$maxmismatch) == 0) {$result = 0;}

  # filtering by quality
  # Expected number of errors is the sum of probabilities in the region

  if (length($qual) < $qto) { $qto = length($qual);}                  # if the read sequence is
shorter than the limit, cut the limit
  my $psum = 0;
  # print "$qual\nqfrom $from qto $qto\n";
  for my $p (0 .. $qto - $qfrom) {
    my $ascode = ord(substr($qual, $p, 1));
    if (exists($asciip{$ascode})) { $q = $asciip{$ascode}; $psum = $psum + $q;}
# $z = substr($qual, $p, 1); print "$z \tcode $ascode \tint $q  pos $p \n"
    else {print ANAL "Error converting ascii quality code $ascode int $q  pos $p \n";}
   }
  if ($psum >= $qlimit) { $result = 0;}
  return $result;
}

# diffcaller
*******************************************************************************************
# compares reference and tested (i.e. consensus) sequences, returns a CIGAR-like summary (e.g.
145A170C200T).
# Calls only substitutions,not deletions and insertions.
# Added a range in which the differences are called
sub diffcaller {
```

```
   my ($ref, $tested, $from, $to) = @_;
   my $diffstring = "";
   my $diffcounter = 0;

   my $lastpos = 0;
   if (length($tested) <= length($ref)) {$lastpos = length($tested) - 1;} else { $lastpos =
length($ref) - 1;}  # strings start at 0
   if ($from > $lastpos) {$from = $lastpos;}
   if ($to > $lastpos) {$to = $lastpos;}

   for my $a ($from .. $to) {
      if ( substr($ref,$a,1) ne (substr($tested,$a,1))) {
       $diffstring = $diffstring.$a.substr($tested,$a,1)."-";
       $diffcounter++;
      }
    }
    $diffstring = substr($diffstring, 0, length($diffstring) - 1);     # remove the last extra
'_'
    return ($diffstring." ", $diffcounter);
}

# consensusmaker
*****************************************************************************
# iterates over a set of sequences with the same tag (received as a reference to the array of
arrays sharing the same tag)
# base that has most counts in the same position is added to the consensus sequence that is
returned
# if there are less that $conminseq (usually 3 or more ) sequence reads in a given position,
it breaks and returns consensus
#  consensusmaker will return result code
sub consensusmaker {
  my ($readsarrayref, $len, $posindex, $conminseq) = @_;
    my $consensus = "";                                  # consensus sequence
    my $con = "";                                        # local storage for consensus
    my $result = "";
    my $percent = 0;
    my $pos = 0;
    my $lastpos = 0;
    my $sum = 0;
    my $valid = 1;

    # if low number of reads in array
    if (@$readsarrayref <= $conminseq)    {$result = $result."L".@$readsarrayref;}
# NOTE quick reporter that the number of reads is sufficient
    if (@$readsarrayref < $conminseq)    {$valid = 0;}
    # check length of sequences
    my $c = 0;
    foreach my $a (0 .. @$readsarrayref - 1) {

      if (length ($readsarrayref->[$a][1]) <= $readlenlimit) { $c++;}
    }
    $result = $result."S$c";

    # iterate over all positions in sequences
   POS : foreach $pos (0 .. $len) {                 # for each position
      my $Acount = 0;
      my $Ccount = 0;
      my $Gcount = 0;
      my $Tcount = 0;
      my $othercount = 0;
      my $highest = 0;
      my $conreads = 0;
      $con = 'X';

      # iterate over sequences from reads and count bases
      RD : foreach my $a (0 .. @$readsarrayref - 1) {

        #print $readsarrayref->[$a][1]." ".length($readsarrayref->[$a][1])."\n";
        if ($pos > (length ($readsarrayref->[$a][1]) - 1)) { next RD;}
        if   (substr($readsarrayref->[$a][1],$pos,1) eq 'A') {$Acount++;}
        elsif(substr($readsarrayref->[$a][1],$pos,1) eq 'C') {$Ccount++;}
        elsif(substr($readsarrayref->[$a][1],$pos,1) eq 'G') {$Gcount++;}
        elsif(substr($readsarrayref->[$a][1],$pos,1) eq 'T') {$Tcount++;}
        elsif(substr($readsarrayref->[$a][1],$pos,1) =~ /\S/) {$othercount++;}  # if other
non-white character
```

```
        #print "$Acount + $Ccount + $Gcount + $Tcount + $othercount\n"
      }
      if    ($Acount > $highest) { $highest = $Acount; $con = 'A';}
      if($Ccount > $highest) { $highest = $Ccount; $con = 'C';}
      if($Gcount > $highest) { $highest = $Gcount; $con = 'G';}
      if($Tcount > $highest) { $highest = $Tcount; $con = 'T';}
      if($othercount > $highest) { $con = 'N';}

      #print "$con $highest ";print "$Acount + $Ccount + $Gcount + $Tcount + $othercount\n";
      $conreads = $Acount + $Ccount + $Gcount + $Tcount + $othercount;
      if ($conreads < $conminseq) { last POS;} #break if the bases are counted from less than
$conminseq reads
      #else {
         $consensus = $consensus.$con;                                    # append the base
to consensus string
         $sum = $sum + 100*$highest/$conreads;                           # sum percentage of
sequences that carry the consensus base
         $lastpos = $pos;
         $highest = 0;
      #}
    }
   #print "sum $sum pos $pos\n";
   my $perc = int($sum/($lastpos + 1));                                      # divide the
sum of percentages by the last position
   $result = $result."C$perc";                                      # append the
percentage to result
   #print "cons $consensus $result\n";
   return ($consensus, $valid);
}

# getfqread
*********************************************************************************************
# skips lines that do not start with @, then reads and chomps 4 lines into an array, returns 0
if not eof and the array
# otherwise 1 and an empty array
sub getfqread {
 my ($fhandle) = @_;
 my $endfile = 0;
 my $linecounter = 0;
 my @read = ();
 my $line = "";

 while ($linecounter ne 4) {
   if (defined ($line = <$fhandle>)) {
      chomp ($line);
      if(($line =~ /^\@/) && ($linecounter == 0))  { $read[0]= $line; $linecounter++; next;}  #
header line
      elsif($linecounter==0)  {next;}
      elsif($linecounter == 1)  { $read[1] = $line; $linecounter++; next;}
# sequence
      elsif($linecounter == 2)  { $read[2] = $line; $linecounter++; next;}
# +
      elsif($linecounter == 3)  { $read[3] = $line; $linecounter = 4;}
# quality, last line in fastq record
    }
    else {$endfile = 1; last; }
 }
 return ($endfile, @read);
}


# tagged
*********************************************************************************************
# test if there there are correct primer sequences flanking the tag in read1
sub tagged{
   my ($test) = @_;
   if ((seqfit($read1ref, $test, $read1from1, $read1to1, $mislimit) == 1) &&
       (seqfit($read1ref, $test, $read1from2, $read1to2, $mislimit) == 1))
   {return 1;}  else {return 0;}
}

# seqfit
*********************************************************************************************
# Checks if the start of the sequence is (almost) identical with the primer,
# assumes two reference and tested strings are equal or almost equal and have the same starts
# Returns 1 (passed) or 0 (failed).
```

```
sub seqfit {
  my($reference,$tested,$from,$to,$mislimit) = @_;

  # sanity checks
  my $lntest = length($tested)-1;
  my $lnref = length($reference)-1;

  if (($from > $lntest) || ($to > $lntest))  { return 0;}  # sequence too short

  my $mismatches = 0;
  for my $a ($from..$to) {
    if (substr($reference,$a,1) ne substr($tested,$a,1)) {$mismatches++;}
   }
  if ($mismatches <= $mislimit) {return 1;}  else {return 0;}
}

sub dottedconsensus{
  my ($reftoarray, $reference,$consensus) = @_;
  my $output = "$reference\n";
  foreach my $a (0 .. @$reftoarray - 1) {

     foreach my $b (0 .. length($reftoarray->[$a][1]) - 1) {
       if (substr($reftoarray->[$a][1], $b, 1) eq (substr($reference, $b, 1))) { $output =
$output.".";}
         else { $output = $output.substr($reftoarray->[$a][1], $b, 1);}
       }
      $output = $output."\n";
    }
  for $a (0 .. length($consensus) - 1) {
     if (substr($consensus, $a, 1) eq (substr($reference, $a, 1))) { $output = $output."-";}
         else { $output = $output.substr($consensus, $a, 1);}
   }
  return $output;
}

sub revcom {
  my ($sequ) = @_;
  $sequ = reverse $sequ;
  $sequ =~ tr/ABCDGHMNRSTUVWXYabcdghmnrstuvwxy/TVGHCDKNYSAABWXRtvghcdknysaabwxr/;
  return $sequ;
}
```