

Univerzita Karlova

2. lékařská fakulta

Studijní program: Molekulární a buněčná biologie, genetika a virologie



Mgr. Petr Vojta

*Zpracování dat z vysokokapacitního DNA sekvenování pro studium
variability genomu a transkriptomu*

*Study of genome and transcriptome variability employing data
processing from massive parallel DNA sequencing*

Dizertační práce

Vedoucí závěrečné práce/Školitel: **Doc. MUDr. Marián Hajdúch, Ph.D.**

Konzultant: **Prof. MUDr. Milan Macek, DrSc.**

Praha, 2018

Prohlášení:

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

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

V Olomouci, 2018

Petr Vojta

Poděkování:

Na tomto místě bych chtěl poděkovat především svému školiteli a konzultantovi za cenné rady a podporu během studia. Dále bych chtěl poděkovat celému týmu Ústavu molekulární a translační medicíny LF UP v Olomouci za spolupráci na řešených projektech v rámci svého studia.

Vojta, Petr. *Zpracování dat z vysokokapacitního DNA sekvenování pro studium variability genomu a transkriptomu. [Study of genome and transcriptome variability employing data processing from massive parallel DNA sequencing]*. Praha, 2018. 110 s., 25 příl. Dizertační práce (Ph.D.). Univerzita Karlova, 2. lékařská fakulta, Ústav molekulární a translační medicíny LF UP v Olomouci.

Vedoucí práce: Doc. MUDr. Marián Hajdúch, Ph.D.

Klíčová slova: MPS, DBA, *RPS7*, anotace DNA variant, RNASeq, stres suchem u rostlin

Obsah

1 Obecný úvod do problematiky.....	8
1.1 Sekvenační technologie.....	8
1.1.1 Sekvenační technologie firmy Illumina.....	8
1.1.2 Sekvenační platforma Roche - 454.....	9
1.1.3 Technologie SOLID.....	10
1.1.4 Platforma Ion Torrent.....	10
1.1.5 Technologie PacBio.....	11
1.1.6 Sekvenace technologií Oxford Nanopore.....	11
1.2 Formáty.....	12
1.2.1 FASTQ.....	12
1.2.2 SAM/BAM a MPILEUP.....	13
1.2.3 GTF/GFF a BED.....	15
1.2.4 VCF.....	16
1.3 Bioinformatika: Analýza a ukládání dat.....	17
1.4 Klinické využití MPS.....	19
1.4.1 Laboratorní zpracování materiálu.....	20
1.4.2 Bioinformatické zpracování.....	21
1.5 RNASeq.....	23
1.5.1 Návrh RNASeq experimentu.....	23
1.5.2 Zpracování RNASeq dat.....	24
2 Hypotézy a cíle práce.....	27
3 MOLDIMED - integrovaná platforma pro zpracování MPS dat.....	28
3.1 Úvod.....	28
3.2 Metody.....	30
3.2.1 Formáty vstupních souborů pro MOLDIMED.....	30
3.2.2 MOLDIMED v3.....	31
3.2.3 MOLDIMED v4.....	34

3.2.4 MOLDIMED v5.....	34
3.2.5 Anotační program.....	35
3.2.5.1 Základní anotace.....	36
3.2.5.2 Poziční a intervalová anotace.....	38
3.2.5.3 Anotace pomocí genových ontologií.....	40
3.2.6 Evaluace.....	41
3.2.6.1 Srovnání detekcí variant ve verzích MOLDIMED.....	41
3.2.6.2 Evaluace anotačního programu.....	42
3.3 Výsledky.....	43
3.3.1 Srovnání detekce variant ve verzích MOLDIMED.....	43
3.3.2 Evaluace anotačního programu (základní, poziční a intervalové anotace).....	44
3.3.3 Rekonstrukce ontologií a testování WES pokrytí ontologických kategorií.....	47
3.3.4 Fenotypová prioritizace a záchyt variant.....	48
3.4 Diskuze.....	49
3.5 Závěr.....	51
4 Charakteristika nové mutace RPS7 u Diamond-Blackfanovy anémie.....	53
4.1 Úvod.....	53
4.2 Materiál a metody.....	54
4.2.1 Klinický popis.....	54
4.2.2 RP panel.....	54
4.2.3 RNASeq.....	56
4.3 Výsledky.....	56
4.3.1 Charakteristika nové mutace v RPS7.....	57
4.3.2 Analýza diferenciální genové exprese.....	58
4.3.2.1 RPS7 transkriptom.....	60
4.3.2.2 RPS19 transkriptom.....	61
4.3.2.3 DBA transkriptom.....	62
4.4 Diskuze.....	67

4.5 Závěr.....	69
5 Transkriptomová analýza transgenního ječmene se zvýšenou odolností vůči stresu suchem	71
5.1 Úvod.....	71
5.2 Materiál a metody.....	73
5.2.1 Rostlinný materiál, podmínky kultivace a aplikace stresu.....	73
5.2.3 RNASeq.....	73
5.3 Výsledky.....	74
5.3.1 Diferenciální exprese genů za optimálních podmínek v nadzemní části rostlin.....	75
5.3.2 Diferenciální exprese genů za stresových podmínek a během revitalizace v nadzemní části transgenních rostlin.....	77
5.3.3 Diferenciální exprese genů za stresových podmínek a během revitalizace v kořenech transgenních rostlin.....	78
5.4 Diskuze.....	80
5.4.1 Vliv deficiencie cytokininů na nadzemní část <i>vAtCKX1</i> pěstovaných za optimálních podmínek.....	80
5.4.2 Odpověď v nadzemních částech <i>vAtCKX1</i> během revitalizace.....	82
5.4.3 Odpověď v kořenech <i>vAtCKX1</i> a <i>cAtCKX1</i> během stresu a revitalizace.....	83
5.4.4 Fenylpropanoidová biosyntetická dráha.....	84
5.5 Závěr.....	85
6 Souhrn.....	87
7 Summary.....	88
8 Seznam zkratk.....	89
9 Seznam citované literatury.....	97
10 Přílohy.....	111
10.1 Životopis.....	121
10.2 Příložené publikace.....	125

1 Obecný úvod do problematiky

1.1 Sekvenační technologie

Metody masivního paralelního DNA sekvenování (MPS) označované též jako sekvenování nové generace (NGS) během poslední dekády velmi ovlivnily přístupy v molekulární biologii, jak v diagnostice onemocnění, tak v oborech zabývajících se analýzami rozsáhlých souborů molekul („Omics“), jejichž rozvoj je právě podmíněn existencí detekčních metod produkujících velké množství dat. MPS se primárně uplatňuje v analýze transkriptomu (transkriptomika), genomu (genomika) a epigenetické regulace (epigenomika), případně ve studiu biologických komunit v konkrétním prostředí (metagenomika). Za dobu používání MPS vznikly rozsáhlé studie s přínosem pro porozumění molekulární podstaty mnoha onemocnění, specifických buněčných procesů, populačních a evolučních fundamentů. Na studium analýzy dat z MPS se zaměřuje bioinformatika. Produkce velkého množství sekvenačních dat z MPS sekvenačních platforem je umožněná paralelizací jednotlivých sekvenačních reakcí.

První generace MPS DNA sekvenátorů vyžaduje specifické klonální obohacení sekvenačních DNA templátů, které je umožněno několika přístupy. Nejznámější jsou přístupy PCR emulzní reakce (Dressman *et al.*, 2003), která je typická pro 454 sekvenační technologii a amplifikace templátu na pevném povrchu (Adessi, 2000). Pro MPS sekvenátory druhé generace (3. generace DNA sekvenování) je charakteristická absence klonální amplifikace DNA templátu, přičemž inkorporace nukleotidů dle komplementárního DNA templátu je detekována v reálném čase. Nové technologie jsou také označovány jako „single-molecule“ sekvenování vyznačující se delším čtením a kratší dobou sekvenačního běhu. Celkový výtěžek sekvenovaných nukleotidů je však nižší a s mnohem vyšší chybivostí, než v případě první generace MPS sekvenátorů. Mezi zástupce DNA „single-molecule“ sekvenačních technologií patří PacBio a Oxford Nanopore.

1.1.1 Sekvenační technologie firmy Illumina

Sekvenační technologie byla vyvinuta původně společností Solexa, která byla odkoupena firmou Illumina (Illumina, Kalifornie, USA). V prvních fázích vývoje byla krátká čtení nepřesahující délku 100 nt značný limitující faktor rozsáhlejšího etablování technologie. Během několika let se podařilo systematicky čtení prodlužovat a v současnosti je maximální délka udáváná společností Illumina 300 nt pro MiSeq a 150 nt pro platformy s větším výstupem řady HiSeq a NovaSeq (www.illumina.com). Celkový výstup nejvýkonnější sekvenační platformy Illumina NovaSeq 6000 je 3 Tb, přičemž sekvenační běhy trvají několik

dnů. Knihovny (specificky upravené DNA fragmenty templátu) jsou klonálně amplifikovány do podoby klastrů na čipu s navázanými adaptorovými oligonukleotidy. Po tvorbě klastrů probíhá sekvenační reakce cyklickou reverzibilní terminací, pomocí fluorescenčně značených nukleotidů (Bentley *et al.*, 2008). Nevýhodou technologie je tzv. „dephasing effect“ způsobený variabilní účinností inkorporace nukleotidů do čtecího vlákna, který způsobuje umělé prodloužení nebo naopak opoždění nově syntetizovaného DNA vlákna od templátu. Efekt kromě toho může být způsobený inkorporací nukleotidu s nenavázanou fluorescenční značkou nebo opožděným fluorescenčním signálem, případně abnormalitami během odstraňování fluorescenčního značení nukleotidu (Metzker, 2010).

Technologie Illumina má v současnosti jednoznačnou dominanci na trhu s nejnižší cenou na sekvenovaný nukleotid a s přijatelnou chybovostí pro resekvenanční projekty, diagnostiku, populační studie GWAS, metagonomiku, transkriptomiku i sekvenování genomů *de novo*, které je prováděno často v kombinaci s dlouhým „single-molecule“ čtením platformami druhé generace MPS sekvenátorů. Navíc bylo firmou Illumina vytvořeno několik sekvenátorů, které jsou vhodné na různé sekvenační projekty, od sekvenování genových panelů nebo hlubokého „ultra-deep“ sekvenování onkologických markerů v diagnostice chorob s nižším celkovým sekvenačním výstupem (MiSeq) až po sekvenátory s dostatečným výstupem pro resekvenování celého lidského genomu několika desítek jedinců v jednom běhu (HiSeq, NovaSeq).

1.1.2 Sekvenační platforma Roche - 454

Technologie 454 (Roche, Švýcarsko) je založena na pyrosekvenování v kombinaci s emulzní PCR přípravou knihovny. Emulzní PCR probíhá na specificky obalených nosičích krátkými oligonukleotidovými sekvencemi po navázání jedné molekuly DNA templátu. PCR reakce se uskutečňuje v lipidických mikroreaktorech, které zabraňují kontaminaci jinou templátovou molekulou DNA. Templátová molekula je obohacena do několika milionů kopií a kovalentně vázána na nosiči. Vlastní sekvenační reakce probíhá v pikotitrační destičce - skleněném čipu s reakčními komorami, ve kterých se nachází klonálně amplifikovaný templát na nosiči. Inkorporace nukleotidů do nově syntetizovaného vlákna dle templátu je detekována pomocí luminiscence. Enzymatická spřažená reakce luminiscence probíhá pomocí ATP, sulfurylázy, luciferázy, apyrázy a DNA polymerázy (Margulies *et al.*, 2005).

454 sekvenační technologie byla jedna z prvních široce používaných technik MPS a to především z důvodu dlouhého čtení s průměrem 700 nt. Souhrnný výstup z GS FLX

sekvenátoru je 700 Mb v jednom dni. V současnosti je metoda 454 na ústupu zejména kvůli vysoké sekvenační ceně a chybovosti čtení v homopolymerních oblastech způsobující zvýšený výskyt chybných inserčních a delečních (INDEL) variant (Liu *et al.*, 2012). Zmíněné nedostatky jsou způsobeny limitací pyrosekvenační reakce a nezdá se, že by v nejbližší době byly překonatelné. I přes uvedené nedostatky se metoda nadále uplatňuje v sekvenování genomů *de novo* a metagenomického profilování.

1.1.3 Technologie SOLID

SOLID sekvenační systém (Life Technologies, Kalifornie, USA) je založený na přístupu sekvenace ligací krátkých oligonukleotidových řetězců s navázanou fluorescenční značkou, které dle komplementarity nasedají na DNA templátovou předlohu. Templátové vlákno je sekvenováno dvakrát, což zvyšuje přesnost čtení, která je u platformy SOLID nejvyšší ve srovnání všech MPS sekvenačních technologií (>99.9 %) (Shendure *et al.*, 2005). Knihovny jsou připravovány obdobným způsobem pomocí emulzní PCR, jak již bylo zmíněno. Nevýhodou SOLID systému je velmi krátké čtení 50-75 nt s časově dlouhým sekvenačním během.

Life Technologies nabízí pro SOLID technologii sekvenátory s vyšším celkovým výstupem (240 Gb). Použití SOLID sekvenační platformy je vhodné zejména tam, kde je vyžadována vysoká přesnost např. v detekci SNP polymorfizmů a diagnostice jednonukleotidových variant (SNV).

1.1.4 Platforma Ion Torrent

Ion Torrent sekvenační technologie (Life Technologies, Kalifornie, USA) je založená na konstrukci knihovny s ligací specifických adaptorů na fragmenty DNA templátu stejně jako v předchozích technologiích. Namísto optické detekce fluorescenčního signálu jsou sekvenované nukleotidy odečítány pomocí změny pH okolního roztoku (Rothberg *et al.*, 2011). Sekvenační reakce probíhá ve specifickém mikročipu, kde jsou umístěny i elektrické sensory. Délka čtení je v závislosti na použitém čipu buď 200 bp nebo 400 bp s celkovou maximální výtěžností 10 Gb (pouze pro 200 bp dlouhé čtení). Výhodou sekvenátorů je minimalizovaná délka běhu, která se i v případě maximální výtěžnosti 10 Gb pohybuje kolem 4 hodin. Stejně jako v případě 454 technologie je zde vyšší chybovost v homopolymerních regionech. Vzhledem k velmi rychlému běhu může technologie najít uplatnění např. v detekci rozsáhlých duplikací a delecí genomu (CNV polymorfismus) a v preimplantační genetické

diagnostice.

1.1.5 Technologie PacBio

Pacific Biosciences (Kalifornie, USA) poskytuje sekvenační technologii „single-molecule real time“ (SMRT) s průměrnou délkou čtení >15,000 nukleotidů a s maximem vyšším jak 100,000 nukleotidů (Eid *et al.*, 2009). DNA polymeráza je pomocí biotin-streptavidinové vazby ukotvena ve speciálních reaktorech tzv. „zero-mode waveguide“ (ZMW), ve kterých probíhá DNA syntéza pomocí templátu a fluorescenčně značených nukleotidů. Po inkorporaci nukleotidu do nově syntetizovaného vlákna dojde k enzymatickému štěpení fluorescenční značky, jejíž signál je detekován pomocí CCD („charge-coupled device“) kamery. Odečítání signálu probíhá paralelně s DNA replikací, což také vede k vysoké chybovosti technologie, která s udávanou hodnotou kolem 15 % vysoce přesahuje konvenční MPS sekvenační technologie. Cirkularizace templátové DNA pomocí specifických hairpin adaptorů (SMRTbell) umožňuje několikanásobné čtení DNA templátu, což vede ke zvýšení přesnosti čtení (Travers *et al.*, 2010). Celkový výtěžek sekvenačního běhu se pohybuje kolem 50-100 Mb s dobou trvání několika hodin. V kombinaci s konvenčními technikami MPS je PacBio účinné v hybridním skládání genomů *de novo*, kdy při dosahované délce čtení umožňuje překlenout repetitivní úseky genomu.

1.1.6 Sekvenace technologií Oxford Nanopore

Oxford Nanopore (Oxford Nanopore Technologies, Spojené království) je druhým zástupcem „single-molecule“ čtení. Oproti ostatním představovaným technologiím je Oxford Nanopore založen na unikátní technologii DNA sekvenování průchodem vlákna proteinovým nanopórem (Branton *et al.*, 2008). DNA templát se uchytil na strukturu nanopóru, kde dochází k rozvinutí vlákna. Jedno vlákno dvoušroubovice je procesováno nanopórem uchyceném v membráně. Specifická konformace jednotlivých nukleotidů vede ke změně průtoku iontů skrz nanopór, což vede ke změně elektrického signálu, který je odečítán. Ve zmíněném případě tedy nedochází k syntéze nového DNA vlákna dle templátu a technologie tudíž není závislá na DNA polymerázové aktivitě. Nanopórová technologie je zatím ve vývoji a v současné době nemá velký praktický přínos. Zajímavostí je také to, že sekvenátor je modifikovaný USB flash disk, který může být připojen k notebooku případně mobilnímu telefonu. K dispozici je několik řešení založených vždy na základní jednotce tzv. MinION, která má sekvenační výstup 10-20 Gb během 48 hodin při maximální délce čtení několik

stovek kb. Ačkoliv je technologie stále ve vývoji, miniaturizace sekvenátoru a bezproblémová příprava vzorku může najít uplatnění například v terénní detekci biologických agens.

1.2 Formáty

Datové bioinformatické výstupy jsou zaznamenány ve specifických formátech s určenými standardy zápisu. Většina formátů má základní podobu v textových formátech v kódování „unicode“, které je multiplatformní, jednoznačné a univerzální, s maximálním využitím snadné zpracovatelnosti textu strojovými metodami. V této kapitole popíšeme několik základních formátů v datové analýze MPS, na které bude v dalším textu často odkazováno. Rozhodně se nejedná o kompletní výčet většiny používaných formátů v bioinformatice.

1.2.1 FASTQ

FASTQ formát (Obrázek 1.1) vychází ze základního nukleotidového zápisu FASTA jeho rozšířením o další dodatečné informace. Stejně jako v případě FASTA je FASTQ textový soubor zahrnující nukleotidovou sekvenci jednotlivých čtení a název sekvence, který je automaticky udáván sekvenátorem. Oproti FASTA je navíc poskytována informace o kvalitě jednotlivých nukleotidů. Identifikátor sekvence je udáván speciálním znakem (@), kde na následujícím řádku je nukleotidový zápis sekvence čtení oddělený na dalším řádku speciálním znakem (+), pod kterým se nachází kvalitativní zápis ve formátu ASCII v analogickém řazení nukleotidového zápisu sekvence. Kvalita je udávána PHRED skórovací maticí. PHRED kvalita (Q) je počítána jako $Q = -10\log_{10}P$, kde P je pravděpodobnost chyby. Q je tedy logaritmické vyjádření pravděpodobnosti chyby, kdy Q(10) odpovídá přesnosti 90 %, Q(20) 99 %, Q(30) 99.9 %, atd. Kvalitativní zápis umožňuje provést souhrnnou analýzu kvality sekvenačního běhu včetně úprav jednotlivých čtení odstraněním regionů s nízkou kvalitou. Nejčastější ASCII zápis je v současnosti ASCII +33, některé starší sekvenační platformy Illumina používají starší ASCII +64. Hodnoty ASCII znaků jsou konvertovány do číselných hodnot, které odpovídají kvalitativní hodnotě Q.

[illegible]

Obrázek 1.1: FASTQ

Ukázka zápisu sekvence ve formátu FASTQ.

1.2.2 SAM/BAM a MPILEUP

SAM formát („Sequence Alignment Map“) je textový soubor se zápisem alignmentu FASTQ čtení na referenční sekvenci (Li *et al.*, 2009) (Obrázek 1.2). Předepsaný formát tvoří 11 následně vysvětlených informativních sloupců oddělených tabulátorem:

1. „QNAME“ obsahuje identifikátor sekvence převzatý z FASTQ.
2. „FLAG“ udává číselnou hodnotu, která v binárním rozkladu poskytuje dodatečné informace, např. zda se jedná o PCR duplikát, jestli bylo správně mapováno příslušné druhé čtení u „paired-end“ konfigurace, orientace alignmentu, zda-li je sekvence alignována ještě na jiném místě v referenci tzv. sekundární alignment, atd. Pro rozklad „FLAG“ čísla do binární podoby je možné použít některé softwarové nástroje např. (<https://broadinstitute.github.io/picard/explain-flags.html>). Na základě informací ve „FLAG“ sloupci lze provádět i selektivní výběr sekvencí podle zvolených atributů programem SAMtools, který byl vytvořen pro manipulaci se SAM/BAM formátem.
3. „RNAME“ je sloupec vyhrazený pro název referenční sekvence (obvykle číslo chromozomu).
4. „POS“ levá pozice (-1) alignmentu oproti referenční sekvenci.
5. „MAPQ“ číselná hodnota kvality mapování alignmentu; výpočet probíhá podle stejné rovnice jako v případě PHRED Q hodnoty. V případě, že MAPQ hodnota není k dispozici, je uváděna hodnota 255 (neplatí u některých RNASeq alignerů).
6. „CIGAR“ textové pole doplňuje informaci struktury alignmentu o ne/shodách s referenční sekvencí číselnou hodnotou odpovídající počtu nukleotidů daného atributu např. zápis 10S62M29S indikuje, že 10 prvních bází bylo odstraněno ([S] „soft-clip“), 62 následujících nukleotidů odpovídá referenci ([M] „matched“) a posledních 29 nukleotidů také není zahrnuto v alignmentu [S]. Technika odstraňování resp. maskování nukleotidových sekvencí na koncích čtení v případě nekorelující sekvence s referencí často zabraňuje reportování variant v případě, že z hrubých sekvenačních dat nebyly kompletně odstraněny adaptorové sekvence, které se nachází na konci čtení. Obvykle jsou varianty reportovány tehdy, kdy minimální okrajová část alignmentu odpovídá referenční sekvenci (např. 81M1119M

- naznačuje inserci [I] jednoho nukleotidu v 82. pozici čtení oproti referenci).
7. „RNEXT“ název referenční sekvence (obvykle číslo chromozomu) pro druhé čtení v případě konfigurace „paired-end“.
 8. „PNEXT“ udává levou (start) pozici alignmentu druhého čtení na referenční sekvenci.
 9. „TLEN“ délka mezi prvním a druhým čtením v případě, že jsou obě čtení alignovány na stejné referenční sekvenci.
 10. „SEQ“ sekvence čtení převzatá z FASTQ.
 11. „QUAL“ ASCII kvalita převzatá z FASTQ.

```
@HD      VN:1.0  SO:coordinate
@SQ      SN:chr20      LN:64444167
@PG      ID:TopHat      VN:2.0.14      CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714      16      chr20      190930      3      100M      *      0      0
      CCGTGTTTAAAGGTGGATGCGGTACCTTCCAGCTAGGCTTAGGGATTCTTAGTGGCTTAGGAAATCCAGCTAGTCTGTCTCTCAGTCCCCCTCT
C      BBDCDDCCDDDDDDDDDDDDCCDDCCDBC?DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDHFFFDCC@@
      AS:i:-15      XM:i:3      XO:i:0      XG:i:0      MD:Z:55C20C13A9      NM:i:3      NH:i:2      CC:Z:=      CP:i:55352714      HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961      16      chr20      193953      50      100M      *      0      0
      TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G      DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDHFFFDCC@@
      AS:i:-16      XM:i:3      XO:i:0      XG:i:0      MD:Z:60G16T18T3      NM:i:3      NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030      16      chr20      270877      50      100M      *      0      0
      GGCTTTATTGGTAAAAAAGGAATAGCAGATTAAATCAGAAATTCACCTGGCCAGCAGACCAACAGAAAGAAGGGAAGAAGACAGGAAAAACCA
C      DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDHFFFDCC@@
      AS:i:-11      XM:i:2      XO:i:0      XG:i:0      MD:Z:0A85G13      NM:i:2      NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699      0      chr20      271218      50      50M4700N50M      *      0
      0      GTGGCTCTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAGCAGAACATCTCAATATGACCTCTCG
```

Obrázek 1.2: SAM formát

Ukázka zápisu alignmentu ve formátu SAM; speciální znak (@) definuje pole hlavičky souboru.

Kromě uvedených sloupců je na začátku souboru uvedena hlavička, která obsahuje další doplňující informace např. verze SAM formátu, informace o použité referenci, název vzorku, knihovny, použité programy a jejich nastavení pro tvorbu a optimalizaci alignmentu, atd. Pro práci se soubory alignmentu je upřednostňována binární podoba SAM tzv. BAM formát s mnohem menší velikostí souboru. V případě archivace dat z MPS je doporučována archivace právě BAM souborů, které umožňují případnou extrakci kompletních originálních FASTQ dat. V případě archivace BAM souborů je ovšem nutné neodstraňovat žádné sekvence např. PCR duplikáty a pouze je značit v „CIGAR“ poli. V případě odstranění některých sekvencí může být narušena integrita zpětně extrahovaných FASTQ dat zejména v „paired-end“ konfiguraci, kdy během alignmentu je vyžadován kompletní pár čtení. Integrita

BAM souboru je kontrolována tzv. EOF („end-of-file“) znakem, což je několik desítek posledních bitů souboru, které musí být ve správné hodnotě. Absence EOF je typická pro nekompletní BAM soubor vzniklý například během přerušeného kopírování dat. BAM soubory mohou být tříděné buď podle názvu identifikátoru sekvence, což znamená, že párové čtení se nachází pod sebou, anebo dle pozic alignmentu odpovídající referenci. Pro archivaci dat se následně zavádí alternativní formát (CRAM) s lepšími parametry komprimace pro datové ukládání. Specifikace formátu CRAM je uvedena na (<https://samtools.github.io/hts-specs/CRAMv3.pdf>).

Další možností zápisu alignmentu je mpileup, textový formát tříděných pozic se specificky zapsaným alignmentem a kvalitou bází v ASCII s hodnotami oddělenými tabulátorem. Mpileup patří mezi formáty s vysokými nároky na velikost souboru a není vhodný k datové archivaci. Standard formátu ve sloupcích je následující:

1. Název referenční sekvence (obvykle název chromozomu).
2. Referenční pozice.
3. Referenční nukleotid.
4. Pokrytí pro reportovanou pozici (DP).
5. Specifický zápis alignmentu pro reportovanou pozici.
6. Kvalitativní zápis alignovaných nukleotidů pro reportovanou pozici v ASCII.

1.2.3 GTF/GFF a BED

GTF/GFF („General Feature Format“) (Obrázek 1.3) a BED („Browser Extensible Data“) jsou textové intervalové soubory, které se používají k popisu regionů, funkčních genomových celků nebo definují oblast genomu zájmu uživatele. Intervalové soubory jsou navázány vždy na určitý referenční genom případně DNA kontig. Specifikace formátu GTF/GFF definuje textový formát s poli oddělenými tabulátorem. Standardně je používáno 9 sloupců o následujících parametrech:

1. Název referenční sekvence (obvykle název chromozomu dle reference).
2. Zdroj, kterým byly souřadnice a anotace získány.
3. Název/definice regionu.
4. Start referenční pozice ohraničující region.
5. Stop referenční pozice ohraničující region.
6. Číselná hodnota skóre.
7. Orientace regionu na DNA vlákne (+/-).
8. Čtecí rámec (0, 1 nebo 2).

9. Atributy regionu, kde doplňující informace jsou odděleny středníkem (;).

```
X      Ensembl Repeat  2419108 2419128 42      .      .      hid=trf; hstart=1; hend=21
X      Ensembl Repeat  2419108 2419410 2502    -      .      hid=AluSx; hstart=1; hend=303
X      Ensembl Repeat  2419108 2419128 0       .      .      hid=dust; hstart=2419108; hend=2419128
X      Ensembl Pred.trans. 2416676 2418760 450.19 -      2      genscan=GENSCAN00000019335
```

Obrázek 1.3: GFF formát

Část intervalového zápisu ve formátu GTF/GFF .

BED formát je oproti GTF/GFF formátu více variabilní a jeho hlavní úlohou je definování regionů zájmů. Stejně jako v případě GTF/GFF jsou sloupce odděleny tabulátorem. Intervalový BED soubor slouží k filtraci oblastí alignmentu případně jiných pozičních formátů, na které se uživatel chce primárně zaměřit, anebo je naopak odstranit. Předepsané jsou pouze tři sloupce, přičemž rozšířené možnosti BED formátu jsou volitelné a jejich definice je uvedena např. (<http://bedtools.readthedocs.io>):

1. Název referenční sekvence.
2. Start referenční pozice ohraničující region.
3. Stop referenční pozice ohraničující region.

1.2.4 VCF

VCF („Variant calling format“) je textový formát zápisu variant rozdělených dle pozic s hlavičkou souboru, která obsahuje metadatové informace včetně definice formátu (<https://github.com/samtools/hts-specs>) (Obrázek 1.4). Současné používané standardy VCF jsou verze 4 a vyšší. Základní pole oddělené tabulátorem definující VCF jsou:

1. „CHROM“ pro název referenční sekvence (obvykle číslo chromozomu s „chr“ notací nebo pouze s číselným značením chromozomu včetně X, Y a MT).
2. „POS“ pozice reportované varianty v referenční sekvenci.
3. „ID“ identifikátor varianty; uvádí se v případě referenčních VCF souborů.
4. „REF“ pro referenční nukleotidový zápis.
5. „ALT“ pro alterovaný nukleotidový zápis specifický pro analyzovaný vzorek.
6. „QUAL“ kvalita genotypu; uvádí pouze některé programy pro volání variant.
7. „FILTER“ informativní pole, které je definováno při úspěšné kvalitativní filtraci variant obvykle hodnotou „PASS“.

8. „INFO“ Pole vyhrazené pro doplňující informace, které jsou obvykle vysvětleny v hlavičce souboru s hodnotami oddělenými středníkem (;).
9. „FORMAT“ definující formát a řazení příslušných informací, které jsou poté uvedeny ke vzorku nebo vzorkům v případě sloučených VCF souborů ve sloupcích řazených od pozice 10.

V posledních letech se začíná preferovat odvozená forma tzv. gVCF, kde kromě variantních pozic jsou zaneseny i intervalové pozice, které vykazují shodu s referenční sekvencí.

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NAO0001 NAO0002 NAO0003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0/0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Obrázek 1.4: VCF formát

Ukázka zápisu variant složeného VCF souboru (v příkladu 3 vzorky NAX); speciální znak (#) definuje pole hlavičky souboru s informacemi pro příslušný VCF soubor.

1.3 Bioinformatika: Analýza a ukládání dat

Bioinformatika je hlavním limitujícím krokem v technologiích MPS především kvůli náročným úkonům jako je ukládání, analýza a interpretace získaných dat. Většinou existují až čtyři úrovně analýzy MPS dat. Prvním krokem je produkce sekvencí čtení softwarem zabudovaným v sekvenační platformě, který konvertuje získaný sekvenační signál do krátkých čtení nukleotidů ve formátu FASTQ. Druhým krokem je sestavování alignmentů s následnou detekcí variant. Třetím krokem je anotace, integrace a vizualizace dat. Posledním krokem je pak sloučení všech dat získaných z různých MPS platforem nebo kombinací bioinformatických analýz do jednoho, koherentního výstupu umožňující biologický nebo forenzní náhled na zpracovaná data. Archivace sekvenačních dat a metadat je problém týkající se výpočetní kapacity v mnoha výzkumných laboratořích. Skladování dat je důležitou

akvizicí pro bioinformatickou analýzu, umožňující případnou opravu a reprocessování dat. Kontrola kvality je základním pilířem pro správné zpracování a následné hodnocení dat ve všech krocích metadatové analýzy. MPS data ze sekvenačních studií bývají většinou poskytována s veřejným přístupem do databází. V případě DNA sekvenačních studií do NCBI SRA (Sequence read archive) databáze a v případě RNASeq nebo ChIP-Seq studií do GEO (Gene Expression Omnibus). Požadavky pro získání a optimalizaci alignmentu a detekci variant jsou závislé na specifitě zadaného projektu. Pro analýzu a převod datových formátů existuje mnoho programů, jak na komerční bázi, tak bezplatných s různými druhy licencí pro svobodné použití programu. Obecně není jediný univerzální software pro všechny typy analýz, a je nutné mezi jednotlivými kroky zpracování dat volit různé nástroje. Bioinformatickou analýzu znesnadňuje fakt, že drtivá většina programů je vázaná na UNIX/LINUX platformy vyžadující od uživatele znalost těchto pracovních prostředí, včetně schopnosti ovládání výpočetních úkonů přes příkazovou řádku tzv. terminál pomocí textových příkazů linuxového shellu.

Procesování sekvenačních dat může být v zásadě dvojího typu. V prvním případě se provádí resekvenace již popsaných genomů za účelem detekce genetických variant (mutace, populační polymorfizmy, asociované studie polygenních onemocnění, detekce aneuploidií, atd.) nebo popisu biologické regulace genové exprese ve specifických podmínkách (analýzy genové exprese a epigenetického stavu genomu). Problematice hodnocení genové exprese a detekce genetických variant budou detailněji věnovány další kapitoly.

Druhým případem je procesování sekvenačních dat za účelem popisu genomu nebo jeho částí *de novo*. V tomto případě nejsou FASTQ alignována, ale překrývající čtení je skládáno do kontigů či scaffoldů pomocí assemblerů. Sekvenační assembly mohou využívat různé algoritmy pro sestavování grafů překryvných sekvencí. Pro krátká čtení se nejčastěji používají assembly založené na konstrukci de Bruijního grafů (Idury & Waterman, 1995; Zerbino & Birney, 2008; Pevzner *et al.*, 2001), kdy k sestavování kontigu dochází na základě hledání shod sekvenčních délkových jednotek, jejichž volená délka je označována jako tzv. „k-mer“ (k), přičemž přesahy jsou velikostně definovány ($k - 1$). S nástupem MPS sekvenačních platform s dlouhým čtením se v současnosti uplatňují hybridní přístupy skládání krátkých čtení de Bruijního metodou společně s identifikací překryvných částí sekvence metodou OLC („overlap-layout consensus“) (Ye *et al.*, 2016). Anotace scaffoldů či kontigů je poté většinou prováděna kombinací přístupů heuristických analýz genomu (programy RAST (Aziz *et al.*, 2008), AUGUSTUS (Stanke *et al.*, 2006), MAKER (Cantarel *et al.*, 2008) a další), včetně identifikace homologických částí proteinového překladu oproti databázovým zdrojům jako jsou prohledávací databáze BLASTu nebo Pfam

(Altschul *et al.*, 1990; Finn *et al.*, 2016). Při anotaci eukaryotních genomů se také uplatňují programy rozeznávající repetitivní oblasti genomu (retrotranspozony, transpozony, satelitní sekvence), např. RepeatMasker (<http://www.repeatmasker.org>), CENSOR (Kohany *et al.*, 2006) a další. Doprovodné anotace se provádí na základě znalostí ontologií a signálních drah prohledáváním databází.

Obecně lze tvrdit, že většinu molekulárně biologické problematiky lze pokrýt bioinformatickými přístupy v rámci „omik“ se vzrůstajícím trendem nahrazování konvenčních přístupů vysokokapacitními technologiemi. Limitujícími faktory stále zůstává cena velkých sekvenačních platforem, včetně jejich údržby, výpočetní infrastruktura a personální zajištění.

1.4 Klinické využití MPS

Metody MPS se uplatňují v klinické praxi v diagnostice germinálních i somatických mutací a stále více pronikají do oblasti cytogenetického hodnocení aneuploidních změn genomu a CNV polymorfizmů. V diagnostice dědičných onemocnění analýzou germinálních mutací jsou standardně prováděna vyšetření genových panelů, exomu (WES), celého genomu (WGS), mitochondriálně vázaných onemocnění a prenatální diagnostiky (Brown & Meloche, 2016; Plutino *et al.*, 2018; Dunn *et al.*, 2018; Jelin & Vora, 2018). Z důvodu dostupnosti malých sekvenačních platforem a snažší interpretace je dnes nejčastěji uplatňovaná strategie sekvenování genových panelů i s ohledem na zvyšující se nabídku komerčních kitů, případně plně automatizovaného servisu zpracování kitu pro panel dle požadavků klienta. Široká nabídka zahrnuje například genové panely pro dědičné vázané nádorové syndromy, kardiomyopatie, vrozenou amyloidózu, slepotu a další (Audo *et al.*, 2012; Voelkerding *et al.*, 2010; DiNardo *et al.*, 2018; Chyra Kuřova *et al.*, 2018). Existují i cílené možnosti sekvenace pouze jednoho chromozomu, pokud jsou známy asociované fenotypové abnormality s vazbou na daný chromozom, jako v případě X vázaných poruch intelektu (Moortgat *et al.*, 2018). MPS metody se mohou uplatnit i v detekci HLA haplotypů nebo preimplantační genetické diagnostice monogenních onemocnění (Bergmann, 2018; Carapito *et al.*, 2016). Ačkoliv je v současnosti doporučováno u hereditárních onemocnění provádět nejdříve specifické vyšetření genového panelu a až v případě negativního nálezu provést vyšetření WES, domníváme se, že vzhledem ke klesající ceně vyšetření exomu, vzrůstající technologické dostupnosti a rapidnímu vývoji bioinformatických nástrojů, bude v budoucnosti větší příklon k WES a WGS metodám genetického vyšetření i s ohledem na možný retrospektivní náhled na získaná komplexní data. U detekce somatických variant získaných onemocnění je kladen důraz na co nejvyšší hloubku pokrytí hotspotových míst. Oproti panelovým a WES technikám se v případě detekce somatických variant uplatňuje

amplikonové sekvenování, kdy jsou knihovny připravovány pomocí PCR. Nejrozšířenější je použití v detekci mutací RAS signální dráhy, *TP53*, *PIK3CA* a *JAK2* (Jennings *et al.*, 2017).

1.4.1 Laboratorní zpracování materiálu

DNA jako zdrojový materiál může být izolována z periferní krve, tkání, buňkálního stěru, ale také z biologického materiálu fixovaného formalínem a zalitého do parafinu (FFPE). Pro všechny kroky kvantifikace DNA během přípravy materiálu se preferují citlivé metody založené buď na interkalaci detekčního barviva, nebo na přístupu „real-time” PCR hodnocení. Knihovny pro MPS jsou připravovány vždy v režimu dle instrukcí výrobce, avšak některá specifika přípravy knihovny jsou obecného charakteru. U sekvenačních platform MPS první generace jsou knihovny soubory fragmentované templátové DNA, jejíž délka je závislá na délce čtení sekvenační platformy. Fragmenty DNA mohou být získány několika přístupy. V případě amplikonového sekvenování jsou délky DNA fragmentů závislé na designu PCR primerů.

Pro techniky cíleného vylučování DNA (genové panely, WES) se nejdříve DNA fragmentuje pomocí nespecifického enzymatického štěpení nebo mechanicky sonikací. K DNA fragmentům jsou následně ligovány adaptorové sekvence s indexy. Do doby indexování DNA knihoven je manipulace se vzorky citlivá na křížovou kontaminaci mezi vzorky a manipulace by měla probíhat v pre-PCR režimu. Po fragmentaci a adaptorové ligaci následuje hybridizace templátu pro obohacení DNA oblastí zájmu. Po hybridizaci následuje PCR reakce za účelem zvýšení vstupního materiálu. PCR reakce v každé fázi přípravy knihovny bývají největším zdrojem nepřesností (bias) generovaných chybami DNA polymeráz (Van Dijk *et al.*, 2014). U hybridizačních technik je PCR bias odstraněn bioinformatickým zpracováním duplikovaných čtení. V případě „ultra-deep” amplikonového sekvenování odstranění duplikátů není možné a PCR bias je hlavní limitující faktor senzitivity v detekci somatických mutací. V případě amplikonových sekvenování je vhodné zavedení specifických adaptorových sekvencí s vlastními tagy, které uvedený problém překonávají a silně zvyšují senzitivitu „ultra-deep” amplikonových sekvenování v detekci somatických variant (Kinde *et al.*, 2011). Pro redukci PCR biasu je také vhodné během PCR reakcí používat DNA polymerázy s nízkou chybivostí.

Většina MPS platform v klinických laboratořích tvoří sekvenátory Illumina (MiSeq, HiSeq a NextSeq). Sekvenační běhy se provádí v libovolné konfiguraci čtení „single-end” nebo „paired-end”. V případě „single-end” čtení je sekvenován DNA templát pouze z jedné strany v případě „paired-end” konfigurace dochází ke čtení templátu z obou stran. Délka

čtení se navrhuje vždy specificky pro daný typ experimentu, přičemž obvyklé délky čtení se pohybují od 50 do 300 nukleotidů. Hrubá data odečtu signálu během sekvenování jsou uložena ve formátu BCL a je nutné je konvertovat do FASTQ.

1.4.2 Bioinformatické zpracování

Základní schéma zpracování bioinformatických dat u resekvenačních studií lze rozvrhnout do několika na sebe navazujících analytických fází:

Předzpracování/Pre-procesování dat je první fází, která zahrnuje demultiplexování FASTQ podle indexových sekvencí v adaptorech, kontrolu kvality FASTQ, odstranění adaptorových sekvencí a části čtení s nízkou PHRED kvalitou.

Složení sekvenačního čtení (alignment) na referenční genomovou sekvenci. Nejaktuálnější verze lidského genomu je verze hg38/GRCH38. Velmi často se v laboratořích provádí alignment oproti starší verzi lidského genomu hg19. Jednotlivé verze genomu nemají shodné pozice a nejsou mezi sebou kompatibilní, což se týká i dalších datových souborů generovaných po alignmentu. Při mezilaboratorním sdílení dat by měla být používána jednotná verze genomu. Jeden z nejčastěji používaných alignerů je BWA, který alignment provádí mechanismem Burrowsovy-Wheelerovy transformace (BWT). Principem metody je, že alignment není oproti referenčnímu genomu prováděn ve FASTA formátu, ale používá se indexovaná BWT verze genomu, která umožňuje rychlejší prohledávání a kompresi genomové sekvence, ve které jsou pak shody jednotlivých čtení specifickým algoritmem prohledávána. Hlavními kroky optimalizace alignmentu je rekalibrace PHRED kvality bází a odstranění PCR duplikátů. Rekalibrace kvality bází předchází systematickým chybám v přiřazení kvality bází sekvenátorem pomocí srovnání známých variant v genomu s variantami v alignmentu a určení nových empirických hodnot kvality bází. Odstranění resp. značení PCR duplikátů probíhá stanovením všech čtení pravděpodobně klonálně odvozených z jedné molekuly templátu během PCR. Chyby vzniklé v průběhu PCR v počátečních cyklech mohou dosahovat vysokých frekvencí v celkovém pokrytí a vést k falešně pozitivním výsledkům. Značení probíhá na základě shodných pozic v alignmentu a orientace čtení, v takových případech pak dochází k redundanci těchto typů čtení. Kroky optimalizace alignmentu je ovšem nutné vždy vážit podle typu experimentu. Některé studie naznačují, že například rekalibrace kvality bází snižuje senzitivitu reportovaných variant ve variabilních místech genomu a místech s nízkým pokrytím (Tian *et al.*, 2016). Stejně tak i odstranění PCR duplikátů má minimální vliv na přesnost získaných variant (Ebbert *et al.*, 2016). Optimalizace alignmentu může být časově značně náročná a v analýze vzorku může

pokryt většinu času celého procesu hodnocení MPS dat, proto je potřeba vždy vážit efektivitu a nutnost provádění všech metadatových analýz.

Volání variant zahrnuje identifikaci genetických variant vzorku oproti referenci. Jako vstupní soubory slouží formáty alignmentu nebo jeho modifikace. K detekci malých variant (SNV a INDEL) se používá zápis VCF formátu, který je generován programem. Mezi nejpopulárnější programy pro volání variant patří VarScan2 (Koboldt *et al.*, 2012), GATK-HaplotypeCaller (McKenna *et al.*, 2010), FreeBayes (Garrison & Marth, 2012), VarDict (Lai *et al.*, 2016), SAMtools (Li *et al.*, 2009), Platypus (Rimmer *et al.*, 2014) a další. Některé programy provádí vlastní re-alignment například GATK, Platypus nebo VarDict, který je nutné zohlednit v případě provádění analýzy korelace variant s vizualizací vstupního alignmentu. Při volbě atributů programu je třeba brát v potaz, že některé programy generují při standardním nastavení velké množství variant, ovšem za cenu snížené specifiity. Z velké části v tomto ohledu záleží i na osobní preferenci uživatele. Část uživatelů preferuje přístup komplexního zachytu variant i za cenu vyššího množství falešně pozitivních nálezů. Domnívám se, že vhodnější strategií je nastavení přísnějších parametrů a zaměřit se na kvalitu sekvenčního výstupu (vyšší pokrytí, minimalizace PCR cyklů, důraz na PHRED kvalitu čtení), anebo genetický nález vždy confirmovat alternativní metodou. K detekci CNV polymorfismu je možné použít alternativní programy jako např. ExomeDepth nebo CoNIFER (Krumm *et al.*, 2012; Plagnol *et al.*, 2012).

Anotace variant společně s interpretací a prioritizací tvoří finální fázi klinického zpracování dat. Anotování variant slouží k interpretaci biologické funkce zachycených variant na transkripční, genové i proteinové úrovni. Kromě základní anotace se dále uplatňuje prohledávání zachycených variant s popsáním biologickým vlivem podle databázových zdrojů, které mohou být i součástí anotačního programu. Populárními anotačními programy jsou ANNOVAR (Wang *et al.*, 2010), VEP (McLaren *et al.*, 2016), snpEff (Cingolani *et al.*, 2012) a mnoho komerčních platforem. Klíčové veřejné databáze známých genetických variant jsou dbSNP, což je katalog genomových variant souhrnně obsahující varianty z NCBI (Sherry, 2001). Další databází variant je ExAC, která je primárně zaměřena na exomová data laboratoří zahrnutých v konsorciu. ExAC zahrnuje exomové profily 60,706 jedinců (Lek *et al.*, 2016). Další databáze jsou více zaměřeny na klinické informace daných variant ve spojení s lidskými onemocněními jako například LOVD (Fokkema *et al.*, 2011), SNPedia (Cariaso & Lennon, 2012), ClinVar (Landrum *et al.*, 2014) a další. Některé vybrané databázové anotační zdroje a jejich funkce budou předmětem následujících kapitol.

1.5 RNASeq

V posledních letech nízká cena MPS sekvenačních technik spolu s vysokým sekvenačním výkonem některých platform umožnila rozvoj použití MPS pro analýzu diferenciální genové exprese (DGE). Oproti čipovým technologiím má sekvenační přístup hodnocení DGE několik výhod. Kromě analýzy exprese jednotlivých genů je navíc umožněna detekce fúzovaných genů (Teixeira, 2006), rekonstrukce sestřihových izoforem a analýza exprese na úrovni jednotlivých transkriptů (Pertea *et al.*, 2016) nebo analýza sekvenčních variant (Barbazuk *et al.*, 2007). Navíc analýza transkriptomu může být provedena bez znalosti referenčního genomu i u druhů s komplexními genomy (Diao *et al.*, 2014; Oono *et al.*, 2013). Anotace transkriptomu složeného *de novo* umožňuje charakterizovat také nové geny a jejich funkce na základě analýzy homologie a predikce.

1.5.1 Návrh RNASeq experimentu

Klíčovým krokem pro konstrukci molekulární knihovny cDNA je selekce mRNA přes poly-dT oligonukleotidy kovalentně vázané na nosičích, které se v roztoku separují na magnetech. Důležitou volbou je rozhodnutí, zda knihovna bude připravena v orientovaném režimu „strand-specific“. Tento režim umožňuje specifičtější analýzu transkriptomu díky znalosti orientace cDNA fragmentu a správné přiřazení čtení i v případech překryvných genů s opačnou orientací v genomu (Zhang *et al.*, 2012). Pro orientované RNASeq knihovny se používá dUTP metoda. V posledních letech se začíná pozornost upínat směrem k cílené analýze transkriptomu na buněčných úrovních nebo u malého počtu buněk, které umožňují studování specifických signálních drah a buněčné nebo tkáňové heterogenity (Griffiths *et al.*, 2018). Pomocí RNASeq metod je také možné studovat expresi malých RNA a jejich biologický vliv na funkční fenotyp (Bortolomeazzi *et al.*, 2017). Dalšími fundamenty pro správné provedení RNASeq experimentu je volba materiálu, počet replikátů nebo například zvolená délka čtení a konfigurace čtení. Volba materiálu se řídí obecnými doporučeními, jako jsou jednotné laboratorní podmínky během přípravy materiálu, volba stejného pohlaví, konstantní doba sběru materiálu, použití stejné délky čtení, ale například i randomizace skupin vzorků do jednotlivých sekvenačních běhů, aby došlo k minimalizování dávkového efektu („batch effect“). Pro minimalizaci dávkového efektu bylo představeno i několik statistických metod (Risso *et al.*, 2014; Leek *et al.*, 2012). Vzhledem k tomu, že transkripční aktivita nekoreluje se složitostí genomu a u vyšších organismů je velikost transkriptomu do určité míry podobná, existují obecná doporučení pro minimální počet čtení pro analýzu transkriptomu. Tento by neměl být menší, než 10 milionů čtení na vzorek. Pro optimální návrh RNASeq experimentu v závislosti na předpokládaném získání počtu

čtení a počtu biologických replikátů je možné použít veřejně dostupné nástroje (Busby *et al.*, 2013). Minimální doporučená konfigurace čtení pro RNASeq je 50 bp v „single-end“ konfiguraci (Chhangawala *et al.*, 2015).

1.5.2 Zpracování RNASeq dat

Prvním krokem je zhodnocení kvality hrubých sekvenačních dat, které je do značné míry analogické jako při zpracování MPS dat pro gDNA sekvenační experimenty. U hrubých sekvenačních dat se kontroluje průměrná distribuce PHRED kvality bází v celé délce čtení, přítomnost kontaminace adaptorových sekvencí, GC obsah a samozřejmě celkový počet získaných čtení. Vizualizaci a úpravu hrubých FASTQ je možné provést v programech FASTQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) a Trimmomatic (Bolger *et al.*, 2014). Alignment se provádí buď na transkriptomovou referenci při použití reference *de novo* zpracované specifickými programy pro skládání transkriptomu např. Trinity (Haas *et al.*, 2013), anebo na referenční genom programy určenými pro RNASeq alignment (Tophat2 [Kim *et al.*, 2013], HISAT [Kim *et al.*, 2015], STAR [Dobin *et al.*, 2013] a další). Obecně by mělo být dosaženo úrovně >90 % mapovaných čtení na genomovou referenci; nižší hodnoty mapovaných čtení většinou nasvědčují kontaminaci materiálu. Po alignmentu následuje kvantifikace čtení pro jednotlivé geny/transkripty oproti referenčnímu GTF/GFF souboru programem HTSEQ (Anders *et al.*, 2015) nebo SubReads/FeatureCounts (Liao *et al.*, 2014), případně je alternativně provedena analýza transkripčních isoform *de novo*, kterou umožňují některé programy jako StringTie (Pertea *et al.*, 2015) nebo Cufflinks (Trapnell *et al.*, 2012).

Kvantifikaci čtení je nutné pro výpočet genové exprese normalizovat. První zavedenou normalizací pro RNASeq data je tzv. RPKM („reads per kilobase per million“), která zohledňuje počet získaných čtení pro transkript k délce transkriptu a celkovému počtu čtení na vzorek vyjádřené v upravené rovnici:

$$RPKM_i = 10^9 \times \frac{n_i}{l_i \times \sum_j n_j}$$

kde $RPKM_i$ je hodnota genové exprese pro transkript (i), n_i počet čtení přiřazených pro transkript (i) s l_i délkou transkriptu (i) a n_j souhrnným počtem čtení zahrnutým v analýze (Mortazavi *et al.*, 2008). Někdy se namísto délky transkriptu (l_i) používá tzv. efektivní délka,

kteřá je získána rozdílem délky transkriptu (l) průměrnou délkou získaných čtení s navýšenou hodnotou +1. Efektivní délka zohledňuje možnost mapovatelných míst v závislosti na zvolené délce čtení a délky transkriptu (Rehrauer *et al.*, 2013). Od RPKM normalizace se odvozuje tzv. FPKM („fragments per kilobase per million“) normalizace, která je počítána analogicky pro čtení v konfiguraci „paired-end“ jako RPKM používanou pro „single-end“ konfiguraci čtení. V FPKM jsou namísto čtení počítány správně namapované páry. Korekce na délku transkriptu není nutná v případě porovnávání genové exprese konkrétního genu napříč vzorky, ale neumožňuje porovnání exprese genů v rámci vzorku. Zmíněný fakt, je nutné brát v potaz u programů, které normalizují pouze na velikost knihovny a neprovádí normalizaci na délku nebo ještě lépe efektivní délku transkriptu jako je například DESeq2 nebo edgeR (Robinson *et al.*, 2010; Love *et al.*, 2014), které se používají na hodnocení změny genové exprese mezi skupinami vzorků statistickým hodnocením pomocí zobecněných lineárních modelů (GLM) vykazujících velmi dobrou sensitivitu pro hodnocení DGE (Lin *et al.*, 2016).

Ačkoliv RPKM/FPKM normalizace je stále používána, od tohoto typu normalizace se začíná ustupovat. Hodnoty RPKM/FPKM jsou závislé na relativní molární koncentraci vstupní RNA a preferenčně je doporučována TPM („transcripts per kilobase million“) normalizace (Wagner *et al.*, 2012), jejíž hodnotu je snadné získat konverzí RPKM/FPKM hodnot podle následujícího vzorce:

$$TPM_i = 10^6 \times \frac{FPKM_i}{\sum FPKM_j}$$

kde TPM_i normalizovaná hodnota exprese odpovídá poměru $FPKM_i$ normalizované hodnoty transkriptu (i) oproti sumě všech FPKM normalizovaných hodnot transkriptu ($\sum FPKM_j$) násobenou o zvolený koeficient.

Funkční anotace diferencielně exprimovaných genů oproti kontrolnímu souboru se provádí prohledáváním databázových zdrojů, jakými jsou například aktuální verze Gene ontology (GO) (Ashburner *et al.*, 2000; Carbon *et al.*, 2017), nebo databáze signálních drah KEGG (Ogata *et al.*, 1999). Několik nástrojů pro funkční analýzu v prostředí programovacího jazyka R je lokalizováno v repozitáři Bioconductor (Huber *et al.*, 2015). K dispozici jsou i nástroje dostupné přes webové rozhraní jako DAVID (Huang *et al.*, 2009), GOrilla (Eden *et al.*, 2009), nebo Reactome (Fabregat *et al.*, 2018). Pro funkční anotaci lze také využít komerční nástroje s uživatelsky přívětivým ovládáním jako je Babelomics (Medina *et al.*, 2010) nebo Blast2GO (Conesa *et al.*, 2005). Pomocí databázových zdrojů je

obvykle možné anotovat většinu diferenciatně exprimovaných genů. V případě, že anotace není k dispozici, je nutné u genů provést funkční a predikční analýzu pomocí homologie nebo homologie funkčních domén proteinu již popsáných genů, například pomocí nástrojů InterPro (Hunter *et al.*, 2009) nebo HMMER (Finn *et al.*, 2011).

2 Hypotézy a cíle práce

Primárním cílem doktorské práce bylo zlepšit stávající schopnosti analýzy a interpretace biologických dat z vysokokapacitního sekvenování v oblasti variability na genomové a transkriptomové úrovni.

Specifické cíle v představované práci byly následující:

1. Vývoj integrované platformy pro zpracování MPS dat pro resekvenační studie se zaměřením na lidský genom.
2. Vývoj anotačního software s účinnou fenotypovou prioritizací variant.
3. Použití dostupných i vlastně vyvíjených nástrojů pro detekci genetických variant u lidských onemocnění.
4. Analýza a zhodnocení transkriptomových studií v modelových studiích u Diamond-Blackfanovy anémie a transgenních rostlin ječmene se zvýšenou rezistencí na sucho.

3 MOLDIMED - integrovaná platforma pro zpracování MPS dat

3.1 Úvod

Metody masivního paralelního sekvenování se v současnosti začínají široce uplatňovat v klinicky orientovaném výzkumu a diagnostice DNA molekulárních patologií. Sekvenační běhy na DNA sekvenátorech nové generace produkují velké množství dat v řádech 10^3 až 10^{12} sekvenovaných bází na jeden běh. MPS umožňuje naplňovat koncept personalizovaného přístupu genetického testování, kdy v rámci jednoho běhu je možné provést sekvenční analýzu celého genomu u několika desítek vzorků. Především u polygenních dědičných onemocnění nebo u genů, kde se mutace mohou vyskytovat náhodně v rámci délky celého genu, je potřeba kompletní analýzy dlouhých DNA regionů, u kterých by konvenční metody kapilárního DNA sekvenování byly časově i finančně náročné. V uvedených případech je výhodnější použít vysokokapacitní metody MPS. Exomové sekvenování (WES), ampikonové sekvenování nebo sekvenování genových panelů jsou nejčastěji používané přístupy pro identifikaci genetických rizikových faktorů (Xuan *et al.*, 2013). WES charakterizuje získání velkého množství dat a metadat v rámci jednoho vzorku, které je nutné cíleně dále zpracovat. Oproti tomu panelová nebo ampikonová sekvenování jsou většinou prováděna na větším souboru vzorků, které je potřeba paralelně zpracovat uniformním způsobem. Současné nástroje pro zpracování dat jsou většinou primárně určeny pro Unix/Linux operační systémy a jejich řízení je ovládáno pomocí terminálu bez grafického rozhraní (GUI). Výpočetní požadavky a nároky na datová úložiště pro zpracování dat z MPS splňují výhradně serverové architektury, na které je nutné se vzdáleně přihlašovat, což opět komplikuje dostupnost pro většinového uživatele. Nástroje pro zpracování dat, dále obsahují řadu uživatelsky definovaných parametrů, které často komplikují mezilaboratorní reprodukovatelnost výsledků. Navzdory možnosti uživatelsky volitelných parametrů a různých způsobů metadatového zpracování, je k dispozici několik všeobecně uznávaných doporučení pro zpracování MPS dat, která umožňují do určité míry automatizovaný přístup zpracování dat (DePristo *et al.*, 2011; Weiss *et al.*, 2013; Richards *et al.*, 2015; Matthijs *et al.*, 2016; Tian *et al.*, 2016).

V současnosti existuje několik uživatelsky přívětivých platforem pro integrovanou datovou analýzu MPS, které jsou většinou na komerční bázi a zajišťovány soukromými společnostmi například CIC bio (Qiagen, Německo) nebo BaseSpace (Illumina, USA). V rámci nabízených služeb uvedené společnosti poskytují i přístup k výpočetním zdrojům pomocí cloudu, takže uživatel nemusí řešit vlastní výpočetní infrastrukturu. Samotné výpočty jsou ve většině případů realizovány softwarem třetích stran, které jsou vyvíjeny v akademickém prostředí se svobodnou licencí, ovšem častokrát právě s uživatelsky

nepríznivým pracovním rozhraním. Oproti komerčnímu přístupu, existují i nekomerční, svobodné platformy; jedním z nejznámějších a etablovaných projektů je projekt Galaxy (Giardine *et al.*, 2005; Afgan *et al.*, 2016). Navzdory všem výhodám cloudového řešení jako je kromě již uvedených i vyšší míra reprodukovatelnosti dat nebo zajištění aktualizovaných databázových zdrojů, použití jak komerčních, tak svobodných platform v analýze patientských vzorků je spojeno s diskutabilními právními a etickými aspekty, kdy charakterem vlastní služby se analyzovaná data dostávají mimo instituci zajišťujících lékařskou péči.

Platformy integrované datové bioinformatické analýzy pro DNA resekvenační studie většinou zahrnují nástroje dělící se do několika základních kategorií: 1) Alignery skládající FASTQ na referenční genom (BWA [Li & Durbin, 2009]), Bowtie [Langmead *et al.*, 2009], Novoalign [<http://www.novocraft.com>] a další). 2) Nástroje pro úpravu a optimalizaci alignmentu, které pracují s datovými formáty alignmentu BAM/SAM (SAMtools [Li *et al.*, 2009], GATK-Picard Tools [McKenna *et al.*, 2010, <http://broadinstitute.github.io/picard/>] a další). 3) Programy generující zápis genetických variant ve standardizovaných formátech VCF/gVCF (VarScan2 [Koboldt *et al.*, 2012], Platypus [Rimmer *et al.*, 2014], GATK-HaplotypeCaller [McKenna *et al.*, 2010], FreeBayes [Garrison & Marth, 2012], VarDict [Lai *et al.*, 2016] a další). 4) Anotační nástroje interpretující význam zachycených genetických variant (ANNOVAR [Wang *et al.*, 2010], snpEff [Cingolani *et al.*, 2012], VEP [McLaren *et al.*, 2016] a další).

V této kapitole představujeme novou platformu MOLDIMED pro analýzu MPS dat z resekvenačních studií lidského genomu, exomu nebo genového panelu. Platforma MOLDIMED zahrnuje i webovou aplikaci s grafickým rozhraním, která umožňuje uživateli kompletní zpracování hrubých FASTQ dat na supervýpočetním klastru (HPC). Platforma může být realizována v cloudovém řešení, ale v případě potřeby instalována i na lokální serverové architektuře. Představovaná platforma sestává ze dvou nezávislých aplikačních částí: V prvním případě nabízí možnost kompletního zpracování FASTQ hrubých dat po vygenerování zápisu variant, kde integrujeme několik běžně užívaných nástrojů třetích stran. Druhou představovanou aplikací je pak samostatný nový anotační software běžící i v aplikaci MOLDIMED, který kromě anotace umožňuje i prioritní a selektivní výběr variant na základě uživatelsky definovaných fenotypových parametrů oproti ontologickým vztahům. Anotační program provádí spárování známých ontologických kategorií ke genům a navíc provádí rekonstrukci všech ancestrálních vztahů, čímž výrazně rozšiřuje a zpřesňuje ontologické genové vazby. Obě aplikační části ve webové aplikaci v současné době umožňují zpracování dat z lidských resekvenačních studií oproti referenčnímu genomu hg38.

3.2 Metody

Vnitřní část MOLDIMED je napsaná v jazyce Python 3 a zahrnuje řízení a paralelizaci integrovaných nástrojů s otevřenou licenci nebo licenci svobodného použití pro akademické účely pro zpracování MPS dat, tak i přímé napojení na anotační software, který je jeho součástí. V současnosti jsou dostupné tři varianty MOLDIMED (varianta v3, v4 a v5), které se liší v použitých programech pro volání variant a procesování alignmentu (Tabulka 3.1). MOLDIMED v3 je dostupný přes webovou aplikaci s ovládáním přes GUI (vyvíjeno ve spolupráci s IT4i, VŠB v Ostravě). Zbylé varianty (v4, v5) ještě nejsou v současnosti implementovány na webové aplikaci. V MOLDIMED není zahrnutá kontrola kvality FASTQ. Uživatel je o kvalitě použitých sekvenčních datech informován v reportu, až na konci výpočtu, což v některých případech může vést k nežádoucímu využití výpočetních zdrojů a proto doporučujeme provést kontrolu kvality FASTQ ještě před použitím aplikace.

Verze programu	GUI	Software třetích stran	Verze	Zajišťované procesy
MOLDIMED v3	+	BWA MEM	0.7.17-r1188	alignment
		SAMtools/BCFtools/HTSlib	1.6	konverze formátu SAM / BAM, sortování alignmentu, odstranění PCR duplikátů, konverze do mpileup, odstranění čtení s nízkou mapovací kvalitou, statistické hodnocení alignmentu
		VarScan2	2.3	volání variant
MOLDIMED v4	-	BWA MEM	0.7.17-r1188	alignment
		SAMtools/HTSlib	1.6	statistické hodnocení alignmentu
		GATK	4.0.2.1	konverze formátu SAM / BAM, sortování alignmentu
		Platypus	0.8.1	volání variant, odstranění PCR duplikátů
MOLDIMED v5	-	BWA MEM	0.7.17-r1188	alignment
		SAMtools/HTSlib	1.6	statistické hodnocení alignmentu
		GATK	4.0.2.1	konverze formátu SAM / BAM, sortování alignmentu, odstranění PCR duplikátů, rekalibrace kvality bazí
		GATK-HaplotypeCaller	4.0.2.1	volání variant

Tabulka 3.1: Seznam integrovaného software třetích stran ve variantách MOLDIMED

V tabulce jsou pro jednotlivé varianty MOLDIMED uvedeny použité programy, jejich verze a zajišťované procesy zpracování dat. Dostupné webové rozhraní s GUI je prozatím dostupné (+) pouze v MOLDIMED v3.

3.2.1 Formáty vstupních souborů pro MOLDIMED

Základními vstupními daty pro všechny varianty MOLDIMED jsou FASTQ soubory strukturovaně uspořádané ve složkách dle názvu vzorků. Při použití webové varianty doporučujeme používat FASTQ v komprimované podobě s příponou .gz programem gzip. FASTQ musí být organizovány ve složkách podle počtu vzorků. Počet FASTQ souborů je závislý na typu čtení, v případě „single-end” čtení je přítomný jeden FASTQ soubor ve složce a pro „paired-end” čtení jsou přítomny dva soubory s příponou R1.fastq (R1.fastq.gz) pro první čtení a R2.fastq (R2.fastq.gz) pro druhé čtení. Pro transformaci BCL strojových sekvenančních dat do FASTQ na přístrojích Illumina doporučujeme do programu bcl2fastq zahrnout atribut „--no-lane-splitting” z důvodu zachování kompatibility s MOLDIMED nebo provést sloučení FASTQ programem cat na UNIX/LINUX systémech při dodržení pořadí jednotlivých čtení separátně pro první a druhé čtení v případě „paired-end” dat. Jednotlivé

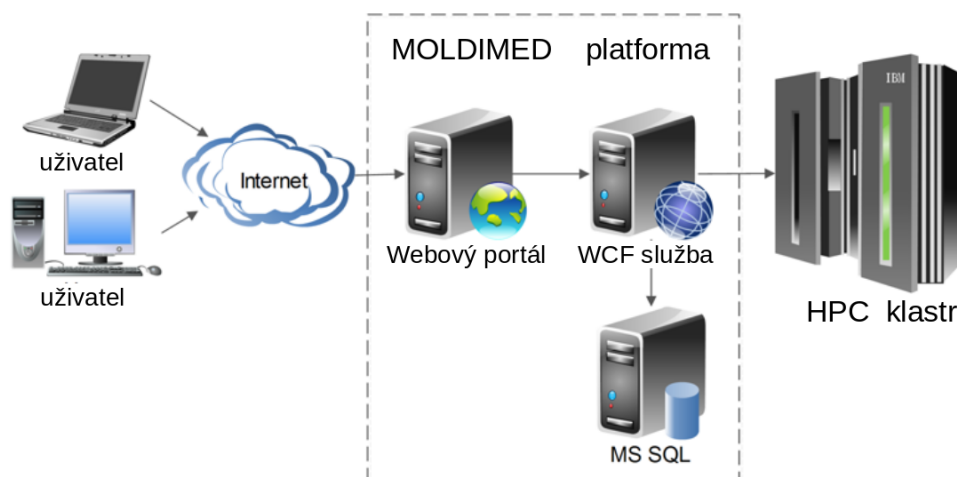
složky dle vzorků musí být organizovány v kořenovém adresáři, jehož cesta je zadávána atributem „--fastq”. Program automaticky rozpozná na základě počtu FASTQ souborů o jaký typ čtení se jedná a podle toho zvolí příslušné atributy programu pro alignment. Různé typy čtení je možné kombinovat i v rámci jedné analýzy.

Dále je nutné poskytnout vyplněný průvodní textový soubor CSV. Formát tabulkového souboru byl navrhnut tak, aby bylo možné použít rozšířenou formu tzv. „sample sheet”, který je zadáván pro sekvenační běhy na přístrojích Illumina. Dodatečné informace se vyplňují v poli definovaným textem [Data]. Názvy sloupců musí být ve správném formátu (Příloha 3.1), ale mohou být zadány v libovolném pořadí (načítání dat probíhá po indexování hlavičky). Vyplněná pole pro (**Sample_ID**) musí obsahovat shodné názvy jako složky s FASTQ. Během načtení a procesování tabulky dochází i ke kontrole stromového adresáře pro podsložky s FASTQ. Pro doplnění hlavičky SAM/BAM souborů slouží sloupce (**Platform**), (**Platform_unit**), (**Library**) a (**Barcode**). Pro prioritizaci variant na základě ontologické anotace je určen sloupec (**Phenotype**), kde se vyplňují identifikátory jednotlivých ontologií, které jsou použity v anotačním programu. V případě zadání více fenotypových kategorií dle identifikátoru, se oddělují jednotlivé kategorie středníkem (;). Pokud uživatel nemá zájem o fenotypovou prioritizaci variant, vyplní do sloupce (g). Sloupec (**Material**) definuje zdroj DNA a je reportován v html výsledkovém formuláři.

Při lokálním běhu programu musí být zadán atribut „--config” konfiguračního souboru používaného všemi variantami MOLDIMED, ve kterém jsou definovány specifické atributy pro filtrování variant, zvolené sloupce anotačních tabulek do výstupního html formuláře, cesty pro ukládání dat a cesty k proprietárním souborům a anotačním databázím (Příloha 3.2). Proměnné a hodnoty mohou být odděleny mezerou nebo tabulátorem. V konfiguračním souboru doporučujeme měnit pouze parametry (Form), (Genome) a (ncpu - udávající počet jader pro paralelizaci výpočtů) dle konfigurace programu MOLDIMED na lokální výpočetní stanici.

3.2.2 MOLDIMED v3

MOLDIMED v3 je prozatím jediná dostupná varianta ve webové aplikaci MOLDIMED s grafickým rozhraním pro uživatele - GUI. S nasazením ostatních variant je počítáno s obdobným grafickým prostředím s pouze pozměněnými atributy na filtraci reportovaných variant ve VCF při použití jiných programů pro jejich volání. MOLDIMED v3 běží na webovém portálu komunikujícím s supervýpočetním klastrem, který poskytuje uživatelům výpočetní zdroje. Schéma infratraktury je uvedeno na Obrázku 3.1.



Obrázek 3.1: Schéma infrastruktury MOLDIMED v3

Služba MOLDIMED v3 je dostupná přes web pomocí šifrovaného spojení HTTPS. Přihlašování do aplikace je pomocí přihlašovacího jména a hesla, takže každý uživatel vlastní svůj účet. Serverová propojení přes „Windows communication Foundation“ (WCF) pro webovou aplikaci zajišťuje komunikaci mezi aplikacemi GUI, MOLDIMED v3, výpočetním klastrem (HPC) a správou databází (SQL).

Po přihlášení do aplikace MOLDIMED zvolí uživatel nahrání souborů pomocí [Resumable uploader], kliknutím na odkaz [Select from your computer]. [Resumable uploader] umožňuje paralelní nahrávání více souborů. Do aplikace doporučujeme nahrávat komprimovaná FASTQ především z rozsáhlejších sekvenančních projektů (WES/WGS) z důvodu kapacitního omezení internetového připojení. Po nahrání veškerých FASTQ a průvodní tabulky v poli [SampleSheet] rozklikneme nabídku, kde se nám automaticky zobrazí námi nahraný průvodní CSV soubor po automatickém rozpoznání .csv přípony a nabídku potvrdíme. Soubor je poté převeden do tabulky s možností editace (Příloha 3.3). Tabulkový editor v GUI je možné samostatně použít pro průvodní CSV dokument bez potřeby jej předem nahrávat. Potvrzením odkazu [Create folders] dojde k automatickému vygenerování stromu se složkami s názvy vzorků zadaných v (**Sample_ID**) sloupci. Uživatel následně spáruje FASTQ do příslušných složek; opět jako v případě použití MOLDIMED přes příkazovou řádku je nutné dodržet správný formát názvu pro „paired-end“ čtení s příponou R1 resp. R2. Standardní nastavení pro procesování dat je určeno pro záchyt germinálních variant se standardním nastavením aligneru BWA MEM. V alignmentu jsou pomocí balíku SAMtools odstraněny PCR duplikáty a je provedeno řazení čtení alignmentu dle pozic

a konverze do formátu MPILEUP, který slouží jako vstupní formát pro program VarScan2. Standardní nastavení volání variant bylo adjustováno dle (Warden *et al.*, 2014) (minimální PHRED mapovací kvalita 10, podpora minimálně čtyř čtení k reportování varianty, filtrace na řetězcové zkrácení ignorující varianty se zachytem čtení pouze v jedné orientaci s vyšší frekvencí než 90 %, minimální pokrytí 10, minimální PHRED kvalita bází 20, minimální frekvence pro reportování varianty 30 %, minimální frekvence pro reportování homozygotní varianty 75 %, p-hodnota pro VarScan2 0.99) Reportovány jsou pouze variantní pozice (v nastavení použita hodnota 1; pro případ reportování všech sekvenovaných pozic splňujících kvalitativní kritéria volíme hodnotu 0). V případě analýzy jiného typu dat je uživateli umožněno nastavení měnit například vypnutí odstranění PCR duplikovaných čtení pro ampikonové sekvenování nebo snížení minimální frekvence pro reportování variant u analýz somatických mutací. Přehled nastavitelných atributů je znázorněn v Příloze 3.4. Analýza je zahájena potvrzením tlačítka [Run FastqPipeline]. Zprávy z průběhu analýzy jsou v reálném čase zaznamenávány v poli [Service Response] včetně výpisů informací jednotlivých programů v pipeline. Po ukončení analýzy je možné celý výpis stáhnout v textovém formátu k případné kontrole analýzy. Ke konci analýzy proběhne strukturované uložení dat a metadatových souborů (alignmenty a jejich optimalizované podoby včetně kompletních anotačních tabulek v CSV formátu). Veškeré soubory lze stáhnout pod příslušným odkazem v komprimované podobě. Schéma zpracování grafického rozhraní pro řízení programu s odkazy na uložení metadatového výstupu jsou zobrazeny v Příloze 3.4.

Anotační tabulky jsou řazeny v MOLDIMED (ve všech variantách) do tří kategorií. Kompletní anotace sdružují všechny anotované varianty. Filtrované anotace zahrnují varianty s patogenní nebo pravděpodobně patogenní klasifikací dle ClinVar (Landrum *et al.*, 2014) bez ohledu na další filtrační kritéria; dále varianty které mají funkční efekt na proteinové úrovni (nesynonymní varianty, mutace v sestřihových místech, delece v CDS, inserce v CDS s posunem i bez posunu čtecího rámce) a nejsou vedeny jako běžné varianty v populaci a také v rámci proteinové funkční predikce obsahují pouze tolerované nebo benigní predikce. Do podsložky klinických variant jsou podstoupeny pouze filtrované varianty, které jsou dále rozděleny do dvou kategorií: patogenní nález (dle Clinvar (Landrum *et al.*, 2014) s patogenní klasifikací a zároveň bez benigní nebo konfliktní klasifikace varianty bez ohledu na fenotypovou prioritizaci pomocí zadaných ontologických identifikátorů a fenotypově prioritizovaný nález filtrovaný podle ontologických identifikátorů v případě, že frekvence varianty <5 % v celosvětové populaci dle dbSNP (Sherry, 2001) a ExAC (Lek *et al.*, 2016). Ve všech anotačních skupinách jsou vždy dva soubory pro každý vzorek. V základním souboru jsou anotovány varianty vůči kompletnímu souboru izoforem transkriptů, které často duplikují nález. Ve druhém souboru s označením „removed_duplicates” jsou duplikáty

na základě shody proteinové pozice, aminokyseliny a sledované záměny v CDS oblasti odstraněny. Sloupec pro název transkriptu obsahuje v případě souboru s odstraněnými duplikáty hodnotu „no_duplicates”.

Po ukončení anotační fáze se pro každý vzorek generuje výsledková zpráva v html formátu obsahující kvalitativní metriku pro optimalizovaný (finální) alignment, který byl použit pro volání variant, použitý referenční genom a doplňující informace převzaté z průvodního CSV souboru. Výsledky anotace do html reportu jsou převzaty z klinických anotací s odstraněnými duplikáty. Duplikované nálezy zůstávají pouze v případě variant s patogenní klasifikací dle ClinVar. Reportovaný nález je rozdělen tedy do dvou tabulek. V první tabulce jsou všechny nalezené varianty s patogenním nálezem dle předchozí definice s omezeným počtem sloupců oproti kompletní anotační tabulce. Do nálezu jsou přeneseny pouze hodnoty ze sloupců: název genu, zápis varianty na genomové úrovni, funkční oblast genu s názvem (exon, intron, CDS, sestřihové místo, atd.) a jejich pořadím v genu, záměna na proteinové úrovni, genotyp, frekvence v populaci dle ExAC, ClinVar rs číslo a fenotyp udávaný ClinVar. Ve druhé tabulce jsou varianty, které obsahují ontologické kategorie definované uživatelem v průvodním CSV dokumentu resp. fenotypově prioritizovaný nález. Do tabulky fenotypově prioritizovaného nálezu html reportu jsou také přenesena jen vybraná pole kompletních anotačních tabulek obdobně jako v případě patogenního nálezu.

3.2.3 MOLDIMED v4

V MOLDIMED v4 je zahrnut minimální počet výpočetních úloh. Alignment je prováděn pomocí BWA MEM se standardním nastavením a jeho řazení dle pozic pomocí GATK. V této variantě není prováděna žádná dodatečná úprava alignmentu. Platypus na volání variant, který je v MOLDIMED v4 použit, je navrhnut s aspektem tzv. „on-the-fly” použitím, kdy značení PCR duplikátů, lokální re-alignment, identifikace variant a její filtrace probíhá integrovaně v programu pro volání variant. Pro volání variant jsou v programu MOLDIMED některé parametry volitelné s adjustovanými hodnotami ve standardním nastavení MOLDIMED (odstranění PCR duplikátů, minimální počet čtení pro podporu varianty 4, odstranění adaptorů, minimální frekvence pro reportování varianty 0.3 a odstranění čtení v případě nenamapování druhého čtení pro konfiguraci „paired-end” čtení).

3.2.4 MOLDIMED v5

Analýza je založena na doporučeních pro použití GATK softwaru pro volání germinálních variant. Alignment je proveden pomocí BWA MEM ve standardním nastavení. Dále je alignment upraven značením PCR duplikátů a rekalicací kvality bází programem GATK (dodatečné doporučené soubory pro podporu úpravy alignmentu byly získány

z (<https://software.broadinstitute.org/gatk/download/bundle>). Pro volání variant je použit GATK-HaplotypeCaller (součást GATK) standardně oproti BED souboru definující celý rozsah genomu. Volba BED souboru je uživatelsky volitelná pomocí atributu „--bed”. Základní VCF soubor je dále adjustován pomocí uživatelsky volitelných parametrů se standardními hodnotami (minimální frekvence variantní alely pro heterozygota 0.3, minimální pokrytí reportované varianty 10, minimální genotypová kvalita (GQ) 0).

3.2.5 Anotační program

Anotační program je dostupný jako součást aplikace MOLDIMED nebo jej lze použít samostatně s ovládáním přes příkazovou řádku na pracovních stanicích nebo serverech, kde je dostupná základní verze Python 3, případně i ve webové aplikaci. Vstupní soubory pro anotaci je VCF nebo gVCF soubor zpracovaný pro jeden vzorek (hromadné VCF soubory sloučených vzorků nejsou podporovány). Základními vstupními soubory pro tvorbu anotační databáze je FASTA soubor genomu, který má být anotován s příslušnou verzí GFF souboru (<http://gmod.org/wiki/GFF3>). V současnosti je na Ensembl (Zerbino *et al.*, 2018) (<https://www.ensembl.org>) dostupných více jak 120 genomů živočichů včetně aktuální verze lidského genomu, přes 800 kmenů hub (<https://fungi.ensembl.org>) a 50 druhů rostlin (<https://plants.ensembl.org>). V případě projektů sekvenování genomu *de novo*, je možné použít programy na generování GFF, např. AUGUSTUS (Stanke *et al.*, 2006). V případě použití referenčních souborů z Ensembl databáze je nejprve nutné GFF soubor upravit tak, aby oblasti stop kodonů byly součástí CDS regionů, což je umožněno skriptem, který je součástí programu. Anotace je rozdělena do třech základních kategorií: 1. základní, 2. poziční a intervalová, 3. anotace na úrovni ontologických vztahů. Při použití anotačního programu přes příkazovou řádku je povinná vždy základní úroveň anotace, volitelné jsou potom kombinace poziční, intervalové a ontologické anotace. Anotace v anotačním programu integrovaném v MOLDIMED určené pro lidský genom ve verzi hg38 anotace probíhají na všech třech úrovních bez možnosti volby. Anotační program umožňuje anotaci polyploidních genomů, MNP variant a rozeznává fázový i nefázový zápis genotypu. Základní zdrojové soubory pro anotační program v MOLDIMED a jejich verze jsou shrnuty v Tabulce 3.2.

Soubor / Databáze	Verze / Soubor	Zdroj
Genom (fasta)	Homo_sapiens.GRCh38.dna.primary_assembly_titled.fa	https://www.ensembl.org/
GFF	Homo_sapiens.GRCh38.88.gtf (tag=basic)	https://www.ensembl.org/
Clinvar	(2018-02-25)	ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/
dbSNP	dbSNP_BUILD_ID=150 (2017-07-10)	ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/
ExAC	ExAC.r0.3.1.sites.vep.vcf	ftp://ftp.broadinstitute.org/pub/ExAC_release
GDI	NA	http://lab.rockefeller.edu/casanova/assets/file/GDI_full_10282015.txt
RVIS	NA	http://genic-intolerance.org/data/GenicIntolerance_v3_12Mar16.txt
TTD	NA	https://db.idrblab.org/ttd/sites/default/files/ttd_database
PolyPhen2 (HDIV, HVAR), SIFT	dbNSFP3.3a	https://sites.google.com/site/jpopgen/dbNSFP
GO (OBO)	go-basic.obo (2018-03-31)	http://geneontology.org/page/download-ontology
GO asociace	goa_human.gaf (2018-02-26)	http://www.geneontology.org/page/download-go-annotations
HPO (OBO)	(2018-03-08)	http://purl.obolibrary.org/obo/hp.obo
HPO asociace	ALL_SOURCES_ALL_FREQUENCIES_genes_to_phenotype.txt (2017-07-10)	http://compbio.charite.de/jenkins/job/hpo.annotations.monthly

Tabulka 3.2: Přehled současných verzí vstupních souborů

Přehled základních databázových zdrojových souborů v současných variantách MOLDIMED; NA - verze není udávána.

3.2.5.1 Základní anotace

Prvním krokem je vytvoření databáze funkčních celků všech transkriptů a od nich odvozených proteinů s příslušnými genomovými souřadnicemi. Jako vstupní soubory slouží referenční genom ve FASTA formátu a příslušný GFF soubor s řádky zvláště pro exonové a CDS oblasti s vyplněným informačním sloupcem, kde jsou definovány názvy transkriptů a genů (obvykle „transcript_id” a „gene_id”). V případě alternativních názvů je schopen uživatel definovat klíčová slova pomocí atributů. Genom ve FASTA formátu musí mít v názvu sekvencí uvedený pouze název chromozomu nebo kontigu (v případě nekompletních genomů) ve stejném formátu jako v prvním sloupci GFF souboru. Během modelování *in silico* veškerých transkriptů probíhá i jejich kontrola, kdy je testována přítomnost nepřerušovaného otevřeného čtecího rámce (ORF) od první pozice kódující oblasti, přítomnost methioninu v 1. pozici proteinu, nepřekrývající se funkční celky transkriptu (CDS, exony, introny, 5'UTR, 3'UTR). Pokud není možné transkript správně zrekonstruovat, do databáze je přenesen s atributem NA. Nalezené varianty budou spárovány s takto zrekonstruovaným transkriptem/genem s NA atributem pro funkční oblast transkriptu/genu bez provedené anotace na proteinové úrovni. V aktuální verzi anotačního programu je prozatím použita pouze univerzální verze genetického kódu, proto některé geny s odchylkami v genetickém kódu (např. některé mitochondriální geny) nebudou anotovány na proteinové úrovni. Použití univerzálního genetického kódu může být limitující faktor i při volbě anotace některých organismů se specifickými odchylkami v genetickém kódu. Během rekonstrukce transkriptů je vytvořen slovníkový soubor pro propojení názvů transkriptů a genů, ve volitelné kombinaci dle zadaných atributů, jak již bylo zmíněno. Lze využít databázových identifikátorů názvu genů dle databází nebo obecně známých názvů genů. Na platformě MOLDIMED jsou použity obecné názvy genů („gene_name” atribut v GFF) namísto Ensembl „ENSG” nomenklatury („gene_id” atribut v GFF). Na MOLDIMED v3 platformě jsou prováděny anotace oproti referenčnímu genomu

hg38/GRCh38 a oproti upravenému GFF ve verzi GRCh38.p10. K anotaci byly použity pouze protein kódující transkripty s tagem „basic“, který označuje referenční soubor transkriptů dle GENCODE doporučený pro anotační účely. Pro základní anotaci je vypracována databáze (**FeaturesList.annotdb_db**) funkčních regionů transkriptů rozdělených do souborů s názvy dle chromozomů/kontigů a slovníkový textový soubor (**GeneList.annotdb**) pro propojení názvů transkriptů s názvy genů. Položky rozdělené v databázi podle chromozomů a řazených dle pozic umožňují paralelizaci výpočetních zdrojů během anotace. Stejný princip paralelizace výpočtů díky distribuci položek do souborů podle chromozomů/kontigů je využit i v případě prohledávání VCF a GFF souborů během poziční a intervalové anotace.

Základní anotace zahrnuje 22 informativních sloupců. V prvním sloupci anotační tabulky (**An**) je uvedeno, zda-li se varianta nachází v genové oblasti („gene“) či nikoliv („non-genic“); ohraničení genových oblastí je určeno počáteční pozicí prvního exonu a konečnou pozicí posledního exonu transkriptu. Určení „up-stream“ a „down-stream“ regionů genů je potenciálně možné na základě intervalové anotace. Rozpoznávání uvedených regionů není ovšem ve standardních GFF anotačních databázích zahrnuto a muselo by být zpracováno uživatelem. Informace ve čtyřech následujících sloupcích jsou převzaty z VCF, jedná se o lokalizaci varianty na chromozomu (**VCFchro**), konkrétní pozici (**VCFpos**), referenční nukleotid případně řetězec nukleotidů u MNP nebo delecí (**VCFref**) a reportovaná nukleotidová záměna (**VCFalt**). Složení variant v genotypu (**GenotypeConform**) je v kompletní ploidii udávané VCF, přičemž dělicí znak zápisu alel je dvojího typu; (/) označuje nefázový zápis, (I) označuje fázový zápis v přesném pořadí jednotlivých alel. Typy reportovaných polymorfizmů (**Type**) zahrnují kategorie INDEL, MNP a SNV. Genotyp (**Genotype**) je klasifikován třemi kategoriemi. V případě heterozygotních variant zapisujeme do dvou kategorií 1) Referenční heterozygot s alespoň jednou reportovanou genotypovou referenční variantou v pozici. 2) Nereferenční heterozygot, kde ani jedna alelická varianta není referenční; homozygot pro variantu je reportován jako nereferenční homozygot. Ve sloupcích (**Start**) a (**Stop**) je interval variantní oblasti minimálního zápisu varianty nebo biologicky relevantní souřadnice INDEL variant, které se liší oproti zápisu ve VCF (obvykle počáteční pozice + 1). Od detekce minimální pozice se odvíjí další dva zápisy ve sloupcích (**RefMin**) a (**AltMin**), u kterých v případě MNP variant odpovídá zápis referenční (**RefMin**) a alterované alely (**AltMin**) v minimálním zápisu, ale v případě INDEL variant je ve sloupci pro referenční zápis uvedeno „ins“ pro inzerci nebo „del“ pro delecí, kde delece nukleotidového řetězce je ve sloupci (**AltMin**). V případě, že ve VCF formátu je uvedeno pokrytí a kvalita genotypu, jsou příslušné hodnoty přeneseny i do anotačních tabulek do sloupců (**GQ**) a (**DP**).

3.2.5.2 Poziční a intervalová anotace

Poziční anotace variant je prováděna na základě srovnání pozice varianty, chromozomu a nukleotidové záměny oproti VCF anotační databázi, které jsou distribuovány do souborů značených dle chromozomů/kontigů s řazenými pozicemi. Informace do výsledných anotačních tabulek je předávána ze sloupce pro identifikátor (ID), jak je určený pro formáty VCF. Cesta k adresářovým strukturám pozičních databázových zdrojů je zadávána pomocí atributu „--vcf_db” do adresáře se složkami se souborovými systémy jednotlivých anotačních zdrojů. Pro reportování nálezu oproti VCF databázím musí dojít ke shodě pozice a DNA alterace ve VCF analyzovaném souboru a VCF anotační databázi. Informačně nejrozsáhlejší položkou anotačního programu pro poziční anotaci je transformovaná databáze ClinVar, kdy na základě shody anotované varianty a pozice oproti ClinVar je reportováno rs číslo varianty. ClinVar je rozdělen do jednotlivých kategorií, pro které jsou v anotační tabulce generovány sloupce pro případné nálezy variantních shod. Kategorie reflektují oficiální dělení ClinVar dle NCBI a zahrnují:

1. Varianty způsobující fenotypy, které nejsou vedeny jako onemocnění např. laktózová intolerance (**clinvar_affects**).
2. Varianty s potenciální klinickou signifikancí na základě rozsáhlých populačních studií (**clinvar_association**).
3. Varianty s benigní klasifikací (**clinvar_benign**).
4. Varianty s konfliktní interpretací (**clinvar_conflicting_interpretations_of_pathogenicity**).
5. Varianty ovlivňující odpověď na léčiva dle doporučení konsorcia klinických farmakogenetických studií - CPIC (**clinvar_drug_response**).
6. Varianty pravděpodobně benigní (**clinvar_likely_benign**).
7. Varianty pravděpodobně patogenní (**clinvar_likely_pathogenic**).
8. Varianty bez poskytnuté klinické signifikance (**clinvar_not_provided**).
9. Varianty klinicky neodpovídající oficiálním kategoriím (**clinvar_other**).
10. Varianty s patogenní klasifikací (**clinvar_pathogenic**).
11. Varianty s protektivním účinkem k onemocněním (**clinvar_protective**).
12. Varianty zvyšující riziko onemocnění (**clinvar_risk_factor**).
13. Varianty s neurčitou klinickou signifikancí (**clinvar_uncertain_significance**).
14. Souhrnné všechny rs varianty v databázi ClinVar (**clinvar_all**).
15. Specificky transformovaná databáze, ze které jsou do anotační tabulky čerpány informace o fenotypu, dědičnosti a názvu onemocnění uvedených v ClinVar (**clinvar_disease**).

Další proprietární položkou v anotačním programu běžícího na MOLDIMED platformě jsou zdroje z populačních sekvenačních studií dbSNP a ExAC. dbSNP je rozdělena do tří databází poskytujících různé druhy informací. Prvním je souhrnná databáze všech rs čísel vedených variant v rámci dbSNP (**dbSNP_all**). Druhá dbSNP databáze (**dbSNP_CAF**) poskytuje frekvenci penetrace varianty převzatých z projektu 1000 genome (Gibbs *et al.*, 2015). Databáze (**dbSNP_common**) spravuje pouze varianty vedené jako běžné v celosvětové populaci, která je založena na variantách se sledovaným minimálním výskytem 1 % ve většinových lidských populacích. Databáze z projektu ExAC je transformována do dvou zdrojů poskytujících informace o frekvenci varianty v populaci (**ExAC_adj_AF**) a počtu, kolikrát byla daná varianta v projektu ExAC reportována (**ExAC_adj_AC**). V obou případech byla použita adjustovaná hodnota, kde jsou započítány pouze varianty splňující minimální kvalitativní kritéria genotypové kvality (GQ) ≥ 20 a pokrytím (DP) ≥ 10 . Pro vyhodnocení predikce funkčního efektu nesynonymních variant na stabilitu jsou převzaté z databáze dbNSFP (Liu *et al.*, 2016) a pro anotační účely byly vybrány pouze tři prediktory Polyphen 2 HDIV (**predict_Polyphen2_HDIV**), Polyphen 2 HVAR (**predict_Polyphen2_HVAR**) a SIFT (**predict_SIFT**). Vyhodnocení predikce je klasifikováno písmeny (B a T) pro benigní a tolerovanou aminokyselinovou záměnu v proteinu, (P) pro potenciálně škodlivou záměnu a (D) pro záměnu s poškozující predikcí na stabilitu proteinu. Doplnujícím databázovým zdrojem pro poziční anotaci je databáze vypracovaná z vlastních WES profilů české a slovenské populace (**frequency_48.frq.num_db**) pro frekvence zachycených variant v rámci diagnostiky vzácných dědičných onemocnění, která nyní zahrnuje 48 patientských vzorků nebo jejich rodinných příslušníků, u kterých byl udělen informovaný souhlas s anonymním zveřejněním exomového DNA profilu (v současnosti je v přípravě rozšíření o dalších více než 50 vzorků).

Pro intervalovou anotaci je nabídnuta možnost anotace vůči GFF intervalovým souborům. Cesta k adresářovým strukturám je zadávána pomocí atributu „--gff_db“. Shoda je nalezena v případě polohy pozice anotované varianty v intervalu nebo rozkladu pozic komplexních variant mezi intervalové pozice položek GFF databáze (povoleny jsou i jednopoziciční intervaly v GFF databázi se shodnou start a stop pozicí). Kontrola shody probíhá i na úrovni orientace v případě, že v základní anotaci proběhlo přiřazení varianty k funkční genomové oblasti s udávanou orientací. V situaci bez přiřazené orientace je reportována shoda s GFF anotační intervalovou databází ve všech případech s orientací (+ nebo -) anebo bez orientace (značeny 0 nebo tečkou [.]) ve sloupci pro řetězcovou orientaci v GFF souboru. Naopak pokud není orientace v GFF souboru uvedena (0 nebo .) je shoda vždy bez ohledu na řetězcovou orientaci funkční genomové oblasti. Stejně jako v případě VCF poziční anotace je do anotační tabulky předávána informace z pole identifikátoru (ID) pro GFF formát. Některé databáze pro intervalové anotace na platformě

MOLDIMED jsou převzaty z databáze BioMart (Kinsella *et al.*, 2011), jež zahrnují intervaly cytogenetických chromozomových regionů (**Cytoband**), informace o typu genů (protein kódující gen, procesovaný pseudogen, miRNA, a další) (**GeneType**), informace o fenotypu (**Phenotype**), popisy genů převzaté z WikiGene (**WikiGene**). Součástí jsou dále hodnotící databáze konzervovanosti jednotlivých genů „Gene Damage Index” (GDI) (Itan *et al.*, 2015) a „Residual Variation Intolerance Score” (RVIS) (Petrovski *et al.*, 2013) (**GDI_Phred**, **GDI_Score**, **RVIS_Percentile**, **RVIS_Score**). Pro identifikaci terapeuticky senzitivních genů byla zvolena databáze Therapeutic Target Database (TTD) (Li *et al.*, 2018) rozdělená do cílů, u kterých je popsána úspěšná léčba (**TTD_success_db**), terapeutik ve fázi klinického testování (**TTD_clinical_db**) a experimentálních terapeutik (**TTD_research_db**). Doplnujícím zdrojem pro námi zpracovaný patientský soubor z WES je frekvence nalezených variant a databáze číselných hodnot u kolika námi zpracovaných vzorků byla daná pozice v intervalu dostatečně pokryta při DP ≥ 10 (**frequency_48.AC.num_db**).

K dispozici je uživateli skript Annot_db_build.py, který umožňuje konstrukci databáze do anotačního programu z VCF i GFF souborů i bez řazených pozic. V současnosti existuje v anotačním programu 24 VCF databází pro poziční anotaci a 12 GFF databází pro intervalovou anotaci. Složení databázových anotačních zdrojů tedy není fixní a uživateli je umožněno nahrání vlastního souborového systému do příslušné adresářové struktury definovaným atributem „--vcf_db” nebo „--gff_db” v dodržném formátu oproti kterému je následně možné anotaci provádět. Předepsaný formát je složka (název databáze) se soubory s názvy chromozomů dle použité reference genomu s řazenými pozicemi v předepsaném formátu (VCF nebo GFF) pro zvolený typ anotace. Podle názvů složky/složek jsou poté automaticky pojmenovány sloupce v anotační tabulce.

3.2.5.3 Anotace pomocí genových ontologií

Mimo základní, poziční a intervalovou anotaci nabízí anotační software rozšíření o genové ontologie vůči Gene Ontology databázi (GO) (Carbon *et al.*, 2017) a Human Phenotype Ontology databázi (HPO) (Köhler *et al.*, 2017). Obě použité ontologie pracují se základními soubory rekonstrukce ontologických grafů s formátem OBO a asociačními soubory názvů genů a přiřazených ontologií. Mimo známé asociace je anotační software schopen rekonstrukce všech chybějících ancestrálních procesů, které nejsou uvedeny v asociačním souboru. Nejprve program na pozadí provede kompletní rekonstrukci všech ontologických kategorií a jejich vazeb, poté jsou ke genům přiděleny kategorie dle asociačního souboru a doplněny veškeré kategorie, až ke kořenovému procesu. Asociační soubory jsou pro anotační program zpracovány pro GO (**GO_assoc.annotdb**)

a HPO (**HPO_assoc.annotdb**). Podmínkou výběru OBO souboru pro rekonstrukci ontologického grafu je absence cyklických vztahů, které se mohou vyskytovat v pracovních verzích spravovaných příslušnými konsorciemi. Spojnice mezi jednotlivými kategoriemi pro GO jsou definovány příbuzností „relationship” a přímým napojením „is_a” ve formátu OBO. V případě GO se vztahy mohou volně křížit mezi základními kategoriemi biologické procesy, molekulární funkce a buněčné komponenty a rekonstruovat tak ke všem kořenům v rámci základních kategorií. HPO využívá hierarchické uspořádání pouze s jedním kořenem a spojnici grafu definovanými „is_a”.

V anotační tabulce jsou GO rozděleny do čtyř sloupců: pro přiřazené ontologické kategorie na podkladě asociačního souboru genů a kategorií všech základních GO skupin (**GO_all_unreconstructed**), rekonstruované kategorie pro biologické procesy (**GO_BP_reconstructed**), rekonstruované kategorie pro molekulární funkce (**GO_MF_reconstructed**) a rekonstruované kategorie pro buněčné komponenty (**GO_CC_reconstructed**). HPO jsou v anotacích zpracovány do dvou sloupců pro nerekonstruované (**HP_unreconstructed**) a rekonstruované (**HP_reconstructed**) kategorie.

3.2.6 Evaluaace

3.2.6.1 Srovnání detekcí variant ve variantách MOLDIMED

Referenční data pro srovnání jednotlivých variant a nastavení platform MOLDIMED byly získány z konsorcia genome-in-a-bottle (<https://github.com/genome-in-a-bottle>) (Zook *et al.*, 2014), ze kterého byl použit VCF soubor (pro hg38) HG005_NA24631 referenčních genomových variant muže čínského etnika Han získaných sekvenační rodinnou studií. Pro komparaci s MOLDIMED byly vybrány pouze varianty v CDS regionech získaných pomocí programu BEDTools v.2.27.1 (Quinlan & Hall, 2010) oproti GFF pro hg38. FASTQ byla získána z BAM souboru (HG005-EEogPU) exomového sekvenování HG005_NA24631 na platformě Illumina pomocí SAMtools v.1.6. FASTQ byly zpracovány všemi dostupnými variantami MOLDIMED. Pro hodnocení byly vybrány VCF soubory generované s adjustovanými i neadjustovanými parametry - defaultní hodnoty všech použitých programů pro volání variant, z nichž byly vybrány opět jen varianty CDS oblastí stejným přístupem jako v případě referenčního VCF. Analýza shod variant byla získána programem BCFtools v.1.6 (Narasimhan *et al.*, 2016). Hodnoty pozitivní predikce (PPV) byly získány dělením počtu shodných variant s referencí celkovým počtem reportovaných variant a vyjádřeny v procentech.

3.2.6.2 Evaluace anotačního programu

K demonstraci základní, poziční a intervalové anotace jsme použili dvě hypotetické varianty, přičemž první tri-alelická varianta odpovídá triploidní konfiguraci ve fázovém zápisu a bez tzv. minimálního zápisu varianty, se kterým se můžeme setkat např. v haplotypizačním programu FreeBayes u některých MPS variant. Druhá varianta je jednoduchý polymorfismus SNV, na kterém demonstrujeme vyhledávání pro poziční a intervalovou anotaci. Pro poziční a intervalovou anotaci jsme vytvořili jednoduché databáze. Zápis všech vstupních souborů je uveden v Příloze 3.5. Komplexní trialelickou variantu jsme vyhodnotili i pomocí VEP anotačního programu a vizualizovali v IGV (Thorvaldsdóttir *et al.*, 2013).

Ontologické anotace byly testovány na dvou úrovních. 1) V rámci testování úspěšnosti a efektivity rekonstrukce ontologických vztahů *de novo* na základě známých asociací genů s jednotlivými ontologickými kategoriemi a jejich rozšíření o veškeré ancestrální nadřazené kategorie bylo provedeno porovnání obou přístupů rozšířeného i nerozšířeného ontologického hodnocení. K rekonstrukci byl využit algoritmus anotačního programu, asociační soubory konsorcií HPO i GO s příslušnými OBO formáty struktury ontologií. V rámci obou skupin byl hodnocen průměrný počet genů v kategoriích, které i v základní, nerozšířené ontologii, asociují s geny a průměrný počet těchto kategorií na jeden gen. Navíc bylo vyhodnoceno množství nově získaných kategorií s genetickou asociací. 2) Vzhledem k faktu, že prioritizace variant je anotačním programem prováděna ve velké míře právě na základě řazení genů do ontologických kategorií, bylo analyzováno pokrytí genů v rámci těchto kategorií na patientském WES vzorku. Knihovna byla připravena pomocí kitu Nextera Rapid Capture Exome (Illumina, USA) dle instrukcí výrobce a sekvenována na přístroji HiSeq2500 (Illumina, USA) v konfiguraci čtení 2x 100. FASTQ byla analyzována MOLDIMED v5. Na základě souřadnic CDS v referenčním GFF byla pro každý gen spočítána velikost kódující oblasti bez redundance transkripčních izoform. Souhrnná velikost regionu pro každou ontologickou kategorii byla spočítána součtem velikostí CDS všech genů v ontologické kategorii. Obdobně byl zpracován i finální optimalizovaný testovací alignment, kde byly filtrovány oblasti s minimálním pokrytím 10x. Senzitivita pro každou ontologickou kategorii byla spočítána poměrem pokryté CDS oblasti alignmentu k referenční hodnotě CDS pro hodnocenou ontologickou kategorii.

Dále byl testován anotační program pomocí vybraných patientských vzorků, úspěšnost detekce fenotypově kauzálních variant pomocí prioritizace a hodnocení variant mitochondriálního genomu z VCF transkriptomových dat. Veškeré testovací výpočty byly prováděny na serveru (AMD Opteron 6212 - 32 CPU, 120 GB RAM).

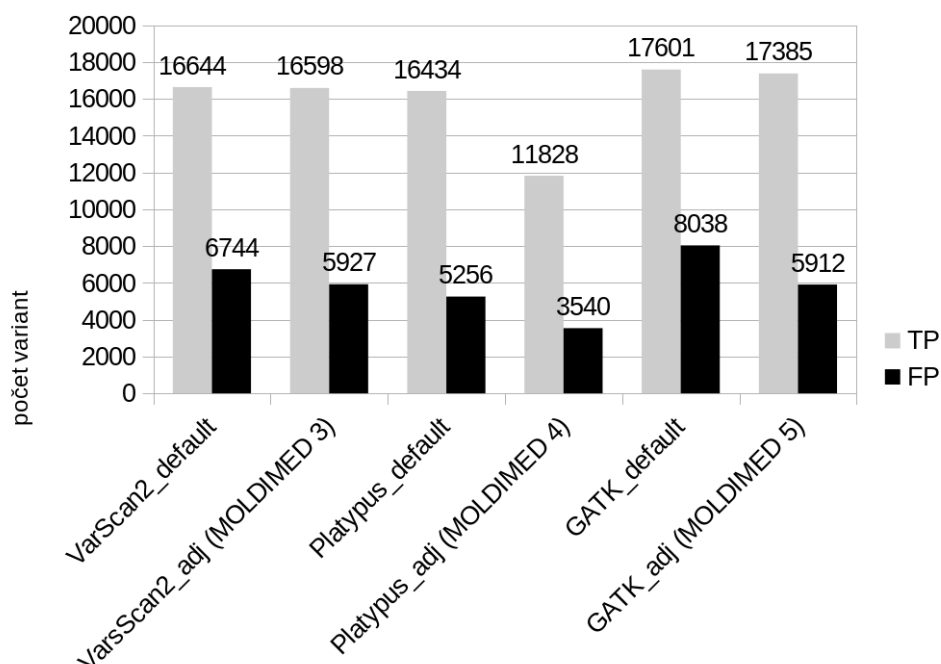
3.3 Výsledky

Anotační program i všechny varianty MOLDIMED jsou napsány v programovacím jazyce Python 3. Platforma obsahuje 3 knihovny (GenomicsFunctions.py pro obecné funkce, Pipelines.py pro funkce MOLDIMED a VO.py pro funkce anotačního programu), 4 řídicí skripty (3 pro spuštění různých variant MOLDIMED a jeden pro nezávislé anotace VCF souborů) a 7 skriptů pro generování databází anotačního programu. Celkově vnitřní část platformy (bez webové aplikace a GUI) obsahuje více jak 4 tisíce řádků zdrojového kódu.

3.3.1 Srovnání detekce variant ve variantách MOLDIMED

Rozkladem BAM souboru HG005-EEogPU z WES studie bylo získáno přes 122 milionů FASTQ čtení v konfiguraci „paired-end“ 2x 126. 120 milionů čtení (99.08 % bylo zahrnuto v primárním alignmentu) ve všech MOLDIMED variantách s minimálním pokrytím 100x odpovídajících 45 MB targetovaného regionu exomu. Referenční VCF HG005_NA24631 obsahuje 3,378,413 genomových variant z toho 17,768 variant bylo lokalizováno v CDS oblastech. Ve srovnání všech dostupných verzí MOLDIMED ve standardním (adjustovaném) nastavení byla nejvyšší konkordance reportovaných variant s referenčním souborem dosažena v MOLDIMED v5 (GATK), kdy bylo nalezeno 17,385 společných variant (TP) a 5,921 falešně pozitivních variant (FP) s celkovou hodnotou PPV 75 %. Nejnižší konkordanci s referencí vykazoval MOLDIMED v4 (Platypus) s 11,828 TP variantami, ale také s nejnižším počtem FP variant 3,540 a s nejvyšší dosaženou PPV hodnotou 77 % v rámci všech testovaných nastavení MOLDIMED verzích. Ve webovém rozhraní aplikace MOLDIMED v3, kde volání variant je zajišťováno programem VarScan2 bylo nalezeno 16,598 TP variant, což je o 787 variant méně, než bylo rozpoznáno adjustovanou verzí s GATK (MOLDIMED v5) a horší výsledek vykazoval i v počtu falešně pozitivních variant (16 variant více) s celkovou PPV hodnotou 74 %. Pokud byly při generování VCF souboru použity standardní nastavení bez adjustace, bylo ve všech případech dosaženo zvýšení konkordance s referenčním souborem, ale také s nepoměrně zvýšeným výskytem falešně pozitivních nálezů, např. GATK bez adjustovaných parametrů reportoval o 217 TP variant více oproti adjustovanému nastavení, zatímco falešně pozitivních variant bylo o 2,127 více (PPV = 69 %), než ve standardním nastavení v MOLDIMED. Lepších proporčních hodnot TP a FP oproti adjustovaným atributům byl zaznamenán u programu Platypus ve standardním nastavení s hodnotami TP o 4,607 a FP o 1,716 vyššími, než v případě adjustace. Obecně úprava parametrů u programů pro volání variant vedla ke slabší konkordanci s referenčním VCF souborem, ale k nižšímu počtu reportování falešně pozitivních nálezů a tím pádem k vyšším PPV hodnotám s adjustací atributů pro volání variant. Srovnání dosažených TP a FP hodnot ve všech variantách MOLDIMED včetně neadjustovaných parametrů u programů pro volání variant je uvedeno

na Obrázku 3.3. Časová náročnost výpočtů pro zpracování FASTQ testovacího vzorku pro jednotlivé varianty MOLDIMED byly pro MOLDIMED v3 (350 min; 5,657 min/CPU), MOLDIMED v4 (236 min; 2,854 min/CPU), MOLDIMED v5 (1,117 min; 3,991 min/CPU) včetně anotací, přičemž anotačním programem byl VCF zpracován za 87 min; 807 min/CPU.

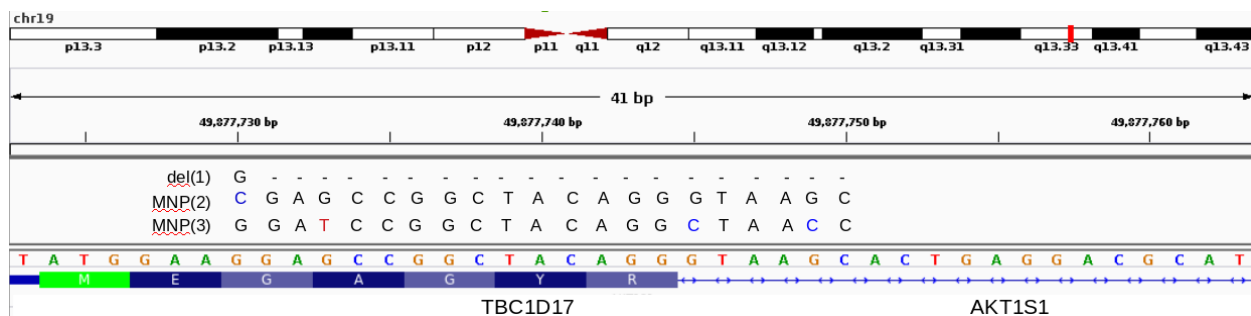


Obrázek 3.3: Úspěšnost detekce variant v jednotlivých variantách MOLDIMED

Konkordance zachycených variant oproti referenčnímu VCF souboru HG005_NA24631 při použití všech dostupných variantách MOLDIMED s rozdílnými programy pro volání variant s adjustovanými parametry (adj) a defaultními parametry (default); shodné varianty s referencí (TP) a falešně pozitivní nálezy (FP).

3.3.2 Evaluace anotačního programu (základní, poziční a intervalové anotace)

Anotační program anotuje běžné sekvenční varianty (SNV, MNP a INDEL) zapsané ve VCF souboru a není určen k anotaci strukturálních změn typu CNV. Pro demonstraci základní, poziční a intervalové anotace byly zvoleny dvě pozice na chromozomu 19 v oblasti 49,877,730 (varianta 1) a 49,878,356 (varianta 2). Varianta 1 je složená triploidní varianta s delecí del(1) a dvěma komplexními variantami MNP (MNP(2) a MNP(3)) ve fázovém zápisu. Vizualizace variant je na Obrázku 3.4. Varianta 2 je jednoduchá homozygotní transverze SNV hg38[chr19:g.49,878,356 G>C] bez fázového zápisu ve VCF.



Obrázek 3.4: Vizualizace složené komplexní varianty (Varianta 1) v IGV pro genom hg38

Triploidní varianta v oblasti s překryvem dvou genů v opačné orientaci *AKT1S1* (-) a *TBC1D17* (+). Varianta 1 zahrnuje delecii del(1) oblasti 49,877,731 až 49,877,750, variantu MNP(2) bez minimálního zápisu s jednoduchou záměnou 49,877,730 G>C v 1. pozici kodonu pro glycin a komplexní variantu MNP(3) také bez minimálního zápisu se třemi alteracemi: 49,877,733 G>T v 1. pozici kodonu pro alanin, 49,877,745 G>C v sestřihovém místě a 49,877,749 G>C v oblasti intronu.

Upravená verze vyhodnocení variant pomocí VEP je uvedena v Příloze 3.6. Varianty hodnocené ve VEP byly rozloženy do 76 nálezů. Delece del(1) byla vyhodnocena jako varianta sestřihového místa, kódující oblasti, 5'UTR, intronu 1 a exonu 1 pro gen *TBC1D17*; intronu, 5'UTR, exonu 1 a intronu 1 pro *AKT1S1*; intronu 1 nekódujícího transkriptu pro gen *PNKP*. Varianta MNP(2) byla správně vyhodnocena jako varianta kódující oblasti a 5'UTR pro exon 1 genu *TBC1D17* bez reportování varianty v intronu. Pro gen *AKT1S1* a *PNKP* byla varianta vyhodnocena se stejným nálezem jako pro předchozí delecii. MNP(3) byla správně vyhodnocena jako varianta sestřihu, intronové a kódující oblasti genu *TBC1D17* a se stejným nálezem v ostatních genech jako u předchozích variant. Varianta 2 byla vyhodnocena v asociaci s geny *TBC1D17* (varianta intronu 1 a 2), *AKT1S1* (5'UTR varianta exonu 1), *PNKP* („up-stream“ genetická varianta). U žádné MNP varianty ani delecce nebyly vyhodnoceny aminokyselinové záměny v případě nesynonymních variant. VEP anotace nabízí možnost prioritizace variant ve sloupci s označením „IMPACT“, kde kategorie „HIGH“ byla přidělena del(1) a MNP(3) pro *TBC1D17*. Varianty ve VEP nejsou anotovány s ohledem na genotyp ani na fázi zápisu. Kromě základních funkčních celků byly varianty anotovány i v kontextu regulačních oblastí resp. promotorů.

Základní anotace testovacích variant námi představovaným anotačním programem je uvedena v Příloze 3.7. Varianty byly rozloženy do 27 nálezů pro anotaci s odstraněnými duplikáty a anotace jsou v korelaci s nálezem ve VEP mimo asociaci variant s *PNKP* pro nekódující transkript. Nekódující transkripty jsou ve standardní verzi anotačního programu vyřazeny. Genotyp triploidní fázové varianty byl ve správném řazení variant, jak je reportováno ve VCF souboru. Varianty bez minimálního zápisu byly korigovány včetně

posunu souřadnic. Pro CDS oblasti byly vyhodnoceny aminokyselinové záměny uvedené ve sloupci (**ACh**) ve formátu: pozice v proteinu : pozice kodonu_reference>alterace. V rámci evaluace intervalové anotace oproti testovacímu GFF souboru byly správně přiřazeny intervalové regiony pro pozice v minimálním zápisu a se správnou orientací, případně bez uvedené orientace ([0] nebo [.]). Poziční anotací, která je vyhodnocena na základě shody variantní pozice a alterace v analyzovaném VCF oproti anotačnímu VCF byly přiřazeny pouze absolutní shody analyzovaných variant oproti referenčním variantám bez ohledu na minimální zápis varianty, nebo rozkladu zápisu varianty, takže po variantu MNP(3) nebyla reportována shodná alterace v minimálním zápisu ani SNV varianta hg38[chr19:g.49,877,745_G>C], která je v MNP(3) zahrnuta. Výsledky pro poziční a intervalovou anotaci oproti testovacím VCF a GFF souborům jsou uvedeny v Příloze 3.8.

Použití univerzálního genetického kódu v interpretaci čtecích rámců pro anotační program vede k limitacím ve správném hodnocení genů s odchylkami v genetickém kódu. V průběhu anotace nedochází ovšem k žádnému vyloučení variant z VCF souboru. V případě, že není možné interpretovat gen na úrovni proteinu, je anotován pouze název genu s „NA” značením ve sloupcích pro anotaci funkčních celků genu. V intervalové, poziční i ontologické anotaci je varianta dále vyhodnocena. Jako příklad anotace variant v mitochondriálním genomu jsme použili VCF soubor pro mitochondriální genom získaný analýzou variant z transkriptomové studie, kde bylo reportováno 9 variant v mitochondriálním genomu. Zjednodušený anotační výstup je uveden v Tabulce 3.2.

Genotype	Start	Stop	HGV	Strand	Feature	ExFeature	Sense	ACh	GeneName	dbSNP_all
non-reference homozygous	3992	3992	g[MT:3992_C>T]	+	NA	NA	.	.	MT-ND1	rs879051705
non-reference homozygous	4024	4024	g[MT:4024_A>G]	+	NA	NA	.	.	MT-ND1	rs41504646
non-reference homozygous	5004	5004	g[MT:5004_T>C]	+	NA	NA	.	.	MT-ND2	.
non-reference homozygous	8269	8269	g[MT:8269_G>A]	+	exon	exon:1	.	.	MT-CO2	.
non-reference homozygous	8269	8269	g[MT:8269_G>A]	+	CDS	CDS:1	synonymous	228:3_*>*	MT-CO2	.
non-reference homozygous	9123	9123	g[MT:9123_G>A]	+	exon	exon:1	.	.	MT-ATP6	rs28358270
non-reference homozygous	9123	9123	g[MT:9123_G>A]	+	CDS	CDS:1	synonymous	199:3_L>L	MT-ATP6	rs28358270
non-reference homozygous	14365	14365	g[MT:14365_C>T]	-	exon	exon:1	.	.	MT-ND6	rs2853815
non-reference homozygous	14365	14365	g[MT:14365_C>T]	-	CDS	CDS:1	synonymous	103:3_V>V	MT-ND6	rs2853815
non-reference homozygous	14582	14582	g[MT:14582_A>G]	-	exon	exon:1	.	.	MT-ND6	rs41354845
non-reference homozygous	14582	14582	g[MT:14582_A>G]	-	CDS	CDS:1	nonsynonymous	31:2_V>A	MT-ND6	rs41354845
non-reference homozygous	15172	15172	g[MT:15172_G>A]	+	NA	NA	.	.	MT-CYB	rs367572771
non-reference homozygous	15326	15326	g[MT:15326_A>G]	+	NA	NA	.	.	MT-CYB	rs2853508

Tabulka 3.2: Anotace variant mitochondriálního genomu

V případě, že není možné správně zrekonstruovat proteinový produkt genu na základě referenčního GFF, genomu ve FASTA a univerzálního genetického kódu, jsou sloupce vyhrazené pro funkční genové oblasti (**Feature**) a (**ExFeature**) doplněny „NA”. Následující anotace (poziční, intervalové a ontologické) nejsou ovlivněny, čímž je umožněno reportování známých variant (rs) včetně patogenních klasifikací; v tabulce jsou uvedeny rs čísla variant ve srovnání s dbSNP databází.

3.3.3 Rekonstrukce ontologií a testování WES pokrytí ontologických kategorií

Z asociačních souborů a OBO souborů bylo po rekonstrukci ontologických vztahů získáno 19,672 genů řazených aspoň do jedné z kategorií pro GO a 3,717 genů s výskytem aspoň v jedné z kategorií pro HPO. Na základě asociačních souborů genů s přiřazenými identifikátory pro jednotlivé ontologické kategorie byly následně do všech ontologických kategorií přiřazeny geny z asociačních souborů k vypracování tzv. nerekonstruovaných ontologií pro GO a HPO. Souhrnně bylo pro GO získáno 22,470 kategorií s minimálně jedním přiřazeným genem a 24,721 prázdných kategorií bez žádného přiřazeného genu. HPO generuje 8,000 kategorií s minimálně jedním přiřazeným genem a 8,907 prázdných kategorií s žádnou genetickou asociací. V žádné ze sledovaných ontologických kategorií GO i HPO nedošlo k disproporci vyššího výskytu genu nerekonstruovaným přístupem oproti zrekonstruovanému.

V rámci GO rekonstrukce bylo získáno 4,744 kategorií s přiřazenými geny, které v nerekonstruované verzi neobsahovaly žádný gen. Mezi 5 nejvíce obohacených kategorií po rekonstrukci patří obecné ontologie (GO:0044464 „cell part” - 17,175 genů, GO:0005488 „binding” - 15,005 genů, GO:0043226 „organelle” - 13,749 genů, GO:0065007 „biological regulation” 12,286 genů, GO:0071704 „organic substance metabolic process” - 11,254 genů). Průměrně na jednu kategorii bylo v GO kategoriích, které obsahovaly geny v obou demonstrováných přístupech (17,727 kategorií) přiřazeno 16 genů bez rekonstrukce a 87 genů s rekonstrukcí. Průměrný počet asociovaných kategorií na gen pro skupinu kategorií, ve kterých byly zjištěny genetické asociace i bez rekonstrukce je 16 kategorií bez rekonstrukce a 87 kategorií s rekonstrukcí. Pro ilustraci rozšíření počtu kategorií GO po rekonstrukci uvádíme příklad pro *BYSL* gen, u kterého bylo nalezeno 18 kategorií bez rekonstrukce a 91 kategorií po rekonstrukci (Příloha 3.9).

V rekonstruované HPO bylo sledováno 1,211 kategorií, které po rekonstrukci získaly genetickou asociaci a bez rekonstrukce neobsahovaly žádný gen. Obdobně jako v případě GO mezi nejvíce 5 obohacených kategorií patří kategorie obecnějšího charakteru včetně kořenového uzlu (HP:0000001 „All” - 3,717 genů, HP:0000118 „Phenotypic abnormality” - 3,699 genů, HP:0000005 „Mode of inheritance” - 3,379 genů, HP:0000152 „Abnormality of head or neck” - 2,215 genů, HP:0011842 „Abnormality of skeletal morphology” - 2,077 genů). Průměrně na jednu kategorii bylo v HPO kategoriích s minimálním počtem jednoho genu v obou analyzovaných přístupech (6,789 kategorií) přiřazeno 18 genů bez rekonstrukce a 47 genů s rekonstrukcí, což odpovídá průměrnému počtu 32 asociovaných kategorií na gen pro skupinu s nerekonstruovanými vztahy a 85 kategorií na gen pro vztahy rekonstruované pro skupiny ontologických kategorií, ve kterých byly nalezeny genové asociace v obou analytických přístupech. Například pro *EGFR* bylo

po rekonstrukci přiřazeno 65 ontologických kategorií, zatímco v asociačním souboru je pro *EGFR* vedeno pouze 11 kategorií (Příloha 3.10). Rekonstruované a nerekonstruované ontologie pro GO jsou uvedeny v Příloze 3.11 a pro HPO v Příloze 3.12. v CSV formátu se sloupci oddělenými tabulátorem.

Vzhledem k tomu, že ontologické anotace tvoří jeden z hlavních pilířů prioritizace variant v anotačním procesu u detekce kauzálních mutací ve spojení s fenotypovým projevem z WES nebo WGS studií v anotačním programu, byla provedena analýza pokrytí kódujících oblastí protein kódujících genů řazených v jednotlivých ontologických kategoriích u modelového vzorku z WES studie. Celkově byla získána *in silico* souhrnná referenční CDS oblast pro 19,608 genů pokrývající 35 MB z referenčního genomu hg38 oproti GFF pro protein kódující geny referenčního souboru dle GENCODE. Pro testovaný vzorek zpracovaný WES bylo získáno 64.5 milionů čtení (2x 100), z toho 99.29 % bylo mapováno s referenčním genomem při minimálním pokrytí (10x ~ 77MB, 30x ~ 52.6 MB).

V rámci HP bylo rekonstruováno celkově 16,907 ontologických kategorií, z nichž 7,979 tvoří kategorie, které obsahují alespoň jeden gen. 15,988 genů nespádá do žádné ontologické kategorie v HP (3,620 genů spadá alespoň do jedné kategorie). HP je zaměřena pouze na anotaci genů, které již byly popsány v souvislosti s lidskými onemocněními nebo abnormalitami a proto většina lidských genů není v této ontologii zahrnuta. Testovaný vzorek vykazoval pro geny zahrnuté v HP 98% pokrytí. Nejnižší pokrytí bylo 36 % pro kategorie HP:0004732 „Impaired renal uric acid clearance“, HP:0002048 „Renal cortical atrophy“ a HP:0012585 „Renal atrophy“, do těchto kategorií spadá pouze jeden gen *MUC1*. Pokrytí pod 80 % bylo naměřeno ještě u dalších 28 kategorií souhrnně obsahujících 17 genů (*BHLHA9*, *B3GALT6*, *SOX18*, *CLIC5*, *ESPN*, *RUNX2*, *GPX4*, *KCNJ10*, *HCRT*, *HOXA13*, *SMOC2*, *PRKACA*, *MXN1*, *MUC1*, *NPHP1*, *TMEM67* a *XPNPEP3*).

Pro GO bylo určeno 47,191 kategorií z toho 22,438 je v asociaci alespoň s jedním genem a 1,311 genů není určeno ani v jedné z GO kategorií. GO je největší současná genová ontologie a zahrnuje většinu známých genů. Testovaný vzorek vykazoval pro kompletní soubor genů zahrnutých v GO kategoriích pokrytí 97 %, při 98% průměrném pokrytí pro jednotlivé GO kategorie. Mírně zvýšená průměrné pokrytí GO kategorií oproti celkovému pokrytí je způsobeno vyšším množstvím kategorií s absolutním pokrytím obsahujících nízké počty genů. Pod hranicí 80% pokrytí bylo v testovaném vzorku stanoveno 157 kategorií.

3.3.4 Fenotypová prioritizace a záchyt variant

Funkčnost fenotypové prioritizace a filtračních kritérií anotačního programu představujeme na dvou reálných patientských vzorcích zpracovaných WES přístupem.

Vzorky byly zpracovány MOLDIMED v4 s integrovaným programem pro volání variant Platypus.

Prvním analyzovaným pacientem je dívka ve věku 10 let s udávanou diagnózou nespecifické stigmatizace a opožděným psychomotorickým vývojem indikovaná k vyšetření exomu. U pacientky bylo reportováno 42,074 variant ve VCF souboru. Po kompletní anotaci bylo získáno 104,190 anotačních nálezů na transkriptomové úrovni s odstraněnými duplicitami. Do filtrované anotace bylo zaneseno 596 anotačních nálezů a v klinických anotacích bylo reportováno ve finálním html formátu 17 patogenních variant a 28 variant ve spojení s fenotypem HP:0001263 „Global developmental delay”. Použitá ontologická kategorie „Global developmental delay” zahrnuje 1,089 genů v nerekonstruované podobě a 1,136 genů v zrekonstruované formě, což je zhruba třetina všech genů v HPO ontologii. Po vyloučení relevance všech patogenních variant s onemocněním byla v rámci fenotypově filtrovaných anotací stanovena molekulární diagnóza heterozygotní mutace hg38[chr22:g.24,322,728_T>C] *SPECC1L* p.L583P v asociaci s autozomálně dominantním Opitz G/BBB syndromem dle ClinVar. Jedná se o dosud nepopsaný genetický nález.

Ve druhém případě byl indikován na sekvenační vyšetření exomu muž ve věku 19 let s agranulocytózou. S ohledem na uváděnou diagnózu byl u pacienta zvolen fenotypový filtr HP:0012234 „Agranulocytosis”. Ve VCF bylo stanoveno 40,389 variant, které po základní anotaci s vyloučenými duplicitami byly zpracovány do 94,882 nálezů z toho ve filtrovaných anotacích bylo zahrnuto 844 nálezů. V klinických anotacích s patogenní klasifikací bylo reportováno 15 nálezů a 1 nález dle fenotypové filtrace. Po vyloučení fenotypové korelace s nálezem s patogenní klasifikací byla u pacienta stanovena molekulární diagnóza heterozygotní mutace hg38[chr19:g.853,279_G>C] *ELANE* p.R81P. Diagnostikovaná mutace nebyla vedena v databázích ClinVar, dbSNP a ExAC. Gen *ELANE* je v asociaci s autozomálně dominantní neutropenií dle ClinVar. Na uvedeném příkladu by fenotypová filtrace bez rekonstrukce neproběhla úspěšně, protože gen není v asociační souboru pro HPO veden v kategorii HP:0012234 „Agranulocytosis”, ale pouze v odvozených kategoriích (HP:0001875 „Neutropenia” a HP:0040289 „Cyclic neutropenia”). U pacienta s agranulocytózou je soubor všech filtrovaných anotovaných variant bez odstraněných duplikátů ve formátu XLSX přiložen v Příloze 3.13 a výstupní html report převedený do pdf v Příloze 3.14.

3.4 Diskuze

Navzdory existenci několika integrovaných platforem pro zpracování MPS dat a anotačních nástrojů, představujeme vlastní řešení pro zpracování sekvenačních dat s uživatelsky přívětivým grafickým rozhraním ve webové aplikaci a komplexním anotačním nástrojem, který umožňuje kromě základní, poziční a intervalové anotace i rozšířenou

ontologickou anotaci s účinnou prioritizací variant jako klíčovou vlastností v diagnostickém přístupu použití MPS metod. Kromě webové aplikace byl kladen důraz na jednoduché ovládání platformy přes příkazovou řádku ve všech variantách MOLDIMED s potenciálem použití na uzavřených výpočetních infrastrukturách. Vzhledem k existenci několika populárních programů pro volání variant lišících se v účinnosti detekce variant nebo rychlosti generování VCF souboru (Hasan *et al.*, 2015; Sandmann *et al.*, 2017), je platforma dostupná v několika variantách umožňující volbu nástroje pro volání variant. Adjustování standardních parametrů programů pro volání variant vykazovaly lepší poměr senzitivity a specifity reportovaných variant oproti referenční kontrole. Naše testování citlivosti zachytu variant ukázalo, že vysoká senzitivita detekce variant je na úkor snížené specifity. Nejlepších výsledků bylo dosaženo pomocí programu GATK-HaplotypeCaller s adjustovanými parametry (MOLDIMED v5). Na druhou stranu GATK patří mezi nejpomalejší programy pro detekci variant, i když v současné době je k dispozici beta verze pro paralelizaci výpočtů.

Anotační program nebo jeho část byly použity v detekci lidských somatických i germinálních genetických variant, analýze transkriptomu ječmene a genetických variant hub rodu *Claviceps* (Majeska-Cudejkova *et al.*, 2016; Pospíšilová *et al.*, 2016; Vojta *et al.*, 2016; Kokas *et al.*, 2016; Zatopkova *et al.*, 2017; Volejnikova *et al.*, 2018; Chyra Kufova *et al.*, 2018). Anotační program je schopný zpracovat genotypy variant fázového typu ve správném řazení alel včetně vyhodnocení komplexních variant a jejich úpravu pro minimální zápis s omezením pro poziční anotaci. Vzhledem k tomu, že vždy dochází ke korekci pozic v případě INDEL variant je algoritmus vyhledávání pozic oproti intervalovým regionům upraven tak, aby vyhledávání probíhalo na základě minimálního zápisu varianty v případě intervalové anotace. Úprava vyhledávacího algoritmu pro minimální zápis i v případě poziční anotace oproti VCF není obdobným způsobem řešena. Domníváme se, že je úlohou programů pro volání variant reportovat variantu vždy ve správném minimálním zápisu varianty, stejně jako v případě referenčních databázových VCF souborů. Referenční databáze by měly obsahovat veškeré modifikace variantní pozice, tak jako je uvedeno v Tabulce 3.4 pro pozici hg38 chr13:g32,338,916 převzaté z databáze ClinVar. Vzhledem k násobně vyššímu počtu položek v databázových VCF souborech, které je nutné testovat, by úprava vyhledávacího algoritmu pro poziční anotaci vedla ke zvýšené časové náročnosti výpočtů. Volba standardních vstupních formátů pro databázové anotace (VCF, GFF, OBO) umožňuje snadnou implemetaci dalších databázových zdrojů v anotování z veřejně dostupných nebo interních zdrojů. Ontologické anotace, její rozšíření o rekonstruované vztahy a specifická fenotypová filtrace dovoluje uživateli jednoduše zpracovávat MPS data a interpretovat klinická data variabilních diagnóz nebo fenotypových projevů i s ohledem na vysokou úroveň pokrytí kódujících oblastí genomu knihovnami pro cílené sekvenování. Nižší počet nalezených ontologických kategorií s genovou asociací v analýze pokrytí genů

řazených v ontologiích z WES bylo způsobeno volbou analýzy pouze pro referenční protein kódující transkripty dle GENCODE.

Chromosom	Pozice	Číslo rs	Reference	Alterace
13	32338916	rs370723514	C	T
13	32338916	rs1064795258	CTA	TTG
13	32338916	rs1064795258	CTA	ATG
13	32338916	rs483353115	CTA	C

Tabulka 3.4: Zápis překryvných variant dle ClinVar

Ačkoliv varianty rs370723514 a rs1064795258 sdílí stejnou alteraci v první pozici, jsou obě varianty vedeny v referenčním VCF nezávisle z důvodu rozdílné biologické interpretace.

Strukturované řazení anotačních výsledků do kategorií („complete“, „filtred“, „clinical“) umožňuje v případě potřeby detailnější analýzu variant. Zjednodušená forma zápisu anotace včetně hodnocení kvality alignmentu a pokrytí je pak automaticky generována v html formátu. Jednou z hlavních předností představovaného anotačního programu je řešení fenotypové prioritizace na základě rozšířených ontologických vztahů, které výrazně navyšují průměrný počet ontologických kategorií na gen oproti počtu kategorií vydávanými příslušnými konsorciemi. Použití MOLDIMED platformy bylo akreditováno ISO 15189:2013 Českým institutem pro akreditaci ke zpracování WES dat.

3.5 Závěr

V první kapitole je představována integrovaná platforma MOLDIMED pro zpracování MPS dat včetně nového anotačního programu. MOLDIMED tvoří pipeline pro zpracování FASTQ dat do finálního produktu zápisu variant ve formátu VCF. V současnosti jsou k dispozici tři varianty platformy, z nichž prozatím k modelové variantě (v3) bylo zpracováno grafické rozhraní pro snadnou obsluhu a zapojení veškerých výpočtů na serverové nebo klastrové výpočetní infrastruktuře. Výpočetní operace probíhají paralelizovaně a automaticky po nadefinování základních atributů uživatelem. Účinnost detekce variant je závislá na jednotlivých programech volání variant, z nichž každý má své přednosti i nedostatky, proto ponecháváme ve vývoji všechny tři oddělené větve platformy a volbu varianty platformy tak necháváme na rozhodnutí uživatele. Jádrem všech platformy MOLDIMED je nový anotační program pro hodnocení variant v zápisu VCF/gVCF. Anotování variant probíhá na třech úrovních: základní, poziční/intervalové a na základě ontologických vztahů k hodnoceným genům. Všechny zdrojové formáty pro anotační program využívají

standardizovaných bioinformatických formátů zápisu a umožňují individuální rozšíření uživatelem. Anotační program byl primárně testován v hodnocení lidských genetických variant, ale funkčnost byla úspěšně ověřována i na jiných typech genomů s různou ploidií. Jednou z výhod oproti stávajícím anotačním programům je účinná fenotypová prioritizace variant na základě ontologických vztahů umožňujících efektivní anotaci genetických variant v plné šíři lidských onemocnění a genetického testování na základě aktuálně dostupných informací o molekulárních patologiích, známých genetických variantách a biologických procesech, ve kterých jsou hodnocené geny zahrnuty.

4 Charakteristika nové mutace RPS7 u Diamond-Blackfanovy anémie

4.1 Úvod

Diamond-Blackfanova anémie (DBA) je vzácná autozomálně dominantní dědičná aplázie červené krevní řady způsobená mutacemi v genech ribozomálních proteinů (RP) nebo jejich haploinsuficiencí. Prozatím byly mutantní formy v asociaci s DBA popsány u 7 genů v rodině genů pro malou ribozomální podjednotku (*RPS19*, *RPS26*, *RPS7*, *RPS24*, *RPS17*, *RPS10*, *RPS15A*) a 3 genů z rodiny pro velkou ribozomální podjednotku (*RPL5*, *RPL11* a *RPL35A*) (Ball, 2011; Ikeda *et al.*, 2016; Errichiello *et al.*, 2017). Mutace v RP genech ovlivňují buněčnou translační aktivitu a správné skládání proteinů (Cmejlova *et al.*, 2006). Některé práce uvádějí nalezené mutace u DBA fenotypu i v genech mimo RP jako je *GATA1* (Sankaran *et al.*, 2012), *TSR2* (Gripp *et al.*, 2014) nebo *MCM2* (O'Brien *et al.*, 2013). Řazení neribozomálních genů do DBA zůstává ovšem nadále diskutabilní. Molekulární patologie DBA způsobuje ribozomální stres, který vede k aktivaci TP53 signální dráhy, zástavě buněčného cyklu a indukci apoptózy (Badhai *et al.*, 2009). Některé RP geny (*RPS3*, *RPS7*, *RPS14*, *RPS27*, *RPL5*, *RPL11*, *RPL23*, *RPL26*) vyvolávají ribosomální stres vazebnou afinitou k MDM2 (Danilova *et al.*, 2008; Deisenroth & Zhang, 2010; Riley & Lozano, 2012). Po navázání volných RP proteinů na MDM2 dochází k zablokování ubikvitinylační aktivity proteinu, která přirozeně vede k degradaci. Deficience *TP53* na modelovém organismu „zebrafish“ pro ribosomální stres vede k částečné normalizaci fenotypu (Azuma *et al.*, 2006). Na „zebrafish“ modelu se sníženou expresí v RP genech (*RPS19*, *RPS24*, *RPL5* a *RPL11*) technikou „morpholino“ byly provedeny transkriptomové studie popisující další molekulární mechanismy a signální dráhy ribozomálního stresu (Jia *et al.*, 2013; Zhang *et al.*, 2013; Song *et al.*, 2014; Wan *et al.*, 2016). Ribozomálního stres také vede ke zvýšenému oxidačnímu stresu v erytrocytech u DBA pacientů (Utsugisawa *et al.*, 2016).

DBA se ve většině případech manifestuje do prvního roku po narození. Poruchy hemopoézy jsou u DBA pacientů silně variabilní od bezpříznakových průběhů přes středně těžké, typicky makrocytární anémie, až po transfúzně dependentní s často doprovodnými fenotypovými abnormalitami jako je malý vzrůst, mentální retardace a zvýšené riziko rozvoje neoplázií (Gustavsson *et al.*, 1997; Cario *et al.*, 1999; Willig *et al.*, 1999; Goudarzi & Lindström, 2016). Navzdory intenzivnímu hledání molekulárních příčin u DBA pacientů, zůstává kolem 20 % případů s neobjasněnou molekulární příčinou onemocnění.

V předkládané studii jsme u pacientů z Českého Národního registru DBA (Pospisilova

et al., 2012) s nediagnostikovanou molekulární příčinou DBA provedli DNA sekvenční analýzu RP genového panelu metodou MPS. Podařilo se nám diagnostikovat novou nesynonymní mutaci v *RPS7* u tří členů rodiny s variabilním fenotypem. Impakt nově nalezené *RPS7* mutace jsme analyzovali v *in vivo* modelu transfekovaných buněk (není součástí dizertační práce) a dále provedli genovou expresní profilaci u všech členů rodiny sekvenováním transkriptomu oproti zdravým kontrolám a DBA pacientům se známou mutací v *RPS19*.

4.2 Materiál a metody

4.2.1 Klinický popis

U pacientky byla mutace v *RPS7* objevena v 18 letech. Porodní váha dítěte byla 2,600 gramů a narodila se po nekomplikovaném těhotenství. U pacientky byla po porodu diagnostikována silná anémie během prvního měsíce. Analýza aspirátu kostní dřeně potvrdila závažnou erytrocytární aplázii diagnostikovanou jako DBA. Během prvního roku života byla pacientka transfuzně dependentní. Od 1. roku do 3. let pacientka podstupovala transfuze již nepravidelně, po tuto dobu úspěšně odpovídala na léčbu nízkými dávkami kortikosteroidů (prednisolon). Pacientka je v současnosti v remisi s hladinou hemoglobinu v krvi (Hb) 101 g/l s objemem erytrocytů (MCV) 112 fL. Žena má normální výšku a nejsou u ní přítomny žádné vývojové abnormality. Matka probanda a starší sestra jsou asymptomatické bez vývojových malformací s normální hodnotou hemoglobinu v krvi (120 g/l a 127 g/l). U obou vyšetřovaných rodinných příslušníků byla laboratorním vyšetřením stanovena mírná makrocytóza (hodnoty MCV: 97 fl a 96 fl). Ostatní členové rodiny jsou zdraví a v rodinné historii neuvádí poruchy krvev tvorby ani vývojové abnormality.

4.2.2 RP panel

DNA byla izolována z periferní krve sadou Genomic DNA Whole Blood Kit (Qiagen, Německo) pomocí izolátoru MagCore HF116 (RBC Bioscience Corp., Tchaj-wan) dle instrukcí výrobce. Periferní krev pro RNA izolaci byla odebrána do zkumavek Tempus Blood RNA (Thermo Fisher Scientific, USA) a izolována z leukocytů sadou Tempus Spin RNA Isolation kit (Thermo Fisher Scientific, USA) dle instrukcí výrobce. Koncentrace nukleových kyselin byla měřena na Nanodrop ND 1000 (Thermo Scientific, USA). Integrita RNA (RIN) byla analyzována na přístroji Agilent Bioanalyzer (Agilent Technologies, USA).

Pro DNA sekvenční analýzu byl navrhnuty cílené sondy pro všechny exonové regiony RP genů pro přípravu molekulární knihovny na DNA sekvenování metodou „sequence capture“ systémem NimbleGen SeqCap EZ custom targeting probe (Roche, Švýcarsko).

Knihovna byla připravena z 500 ng DNA dle instrukcí výrobce a sekvenována na přístroji 454 GS Junior (Roche, Švýcarsko).

Hrubá sekvenační data FASTQ byla kvalitativně kontrolována a upravena programem Trimmomatic (Bolger *et al.*, 2014). Pro alignment byl použit BWA MEM v.0.7.5a (Li & Durbin, 2009) oproti referenčnímu genomu hg38. Z alignmentu byly odstraněny PCR duplikáty a čtení s nízkou mapovací kvalitou (MQ <10) a optimalizovaný alignment byl převeden do mpileup formátu pomocí SAMtools v.1.5 (Li *et al.*, 2009). Varianty byly volány v programu VarScan2 (Koboldt *et al.*, 2012) s upravenými parametry (minimální pokrytí 10x, minimální frekvence varianty 0.3, varianta reportována minimálně ve 4 čteních, minimální Phred kvalita bází 20) a anotovány platformou MOLDIMED. S ohledem na známou chybovost metody 454 DNA sekvenování byly z analýzy odstraněny všechny délkové varianty (INDEL) a evaluovány byly pouze jednonukleotidové záměny (SNV). Získané varianty byly dále srovnány s databází dbSNP (Sherry, 2001), kde byly odfiltrovány všechny známe (rs) varianty s nepatogenní klasifikací. BedTools v.2.27.1 (Quinlan & Hall, 2010) byl použit na analýzu pokrytí cílového DNA regionu. Vybrané varianty byly dále kontrolovány a vizualizovány v IGV prohlížeči (Thorvaldsdóttir *et al.*, 2013). Pro proteinovou funkční predikci nesynonymních variant byly použity Polyphen2 (Adzhubei *et al.*, 2010), I-Mutant2.0-Seq (Capriotti *et al.*, 2005) a MUp (Cheng *et al.*, 2005).

Sekvenování RP genového panelu odhalilo mutaci hg38[chr2:g.3,580,153G>T], která byla následně potvrzena kapilárním DNA sekvenováním pomocí primerů navrhnutých do místa mutace exonu 6 *RPS7* (cattttgacttaaagaggtgc, cactaaaatccactctcactg). Elektroforetogramy byly následně editovány a vizualizovány v Unipro UGENE (Okonechnikov *et al.*, 2012).

Dále byla provedena sekvenační analýza cDNA *RPS7* pacientky, matky a sestry z části materiálu periferní krve. 3 µg celkové RNA bylo inkubováno s 0.3 µg náhodných hexamerních primerů (Promega, USA) při 70°C po dobu 5 minut. Reverzní transkripce byla připravena pomocí kitu RevertAid reverzní transkriptázy (Fermentas, Lithuania) v reakci (6 µl RevertAID 5x RT pufr, 150 U RevertAid „Moloney Murine Leukemia Virus” reverzní transkriptázy, 3 µl 10mM dNTP, 0.75 µl 40 U/µl RNAsin ribonukleasového inhibitoru (Promega, USA) doplněné do finálního objemu 30 µl DEPC vodou (Ambion, USA)), inkubace probíhala při 42°C po dobu 60 minut a následně 10 minut při 72°C. Kompletní cDNA *RPS7* byla získána PCR s primery navrhnutými do 5'UTR a 3'UTR oblastí genů (tctcgcgagatttgggtctctt, ttacaattgaaactctgggaattcaaaat). Z cDNA byla připravena knihovna pomocí Nextera DNA Library Preparation Kitu (Illumina, USA) dle instrukcí výrobce a sekvenována na sekvenátoru MiSeq (Illumina, USA). Alignmenty byly získány analogickým přístupem jako v kapitole 4.2.3 RNASeq.

4.2.3 RNASeq

Pro hodnocení genové exprese byly použita RNA z krevních odběrů rozdělených do tří skupin: kontrolní skupina (4 zdraví jedinci: 2 muži, 2 ženy), kontrolní skupina RPS19 (4 DBA pacienti: 2 muži a 2 ženy s popsanou mutací v *RPS19*) a RPS7 skupina (tři členové rodiny: 3 ženy s nově nalezenou mutací *RPS7*).

Periferní krev pro RNA izolaci byla odebrána do zkumavek Tempus Blood RNA (Thermo Fisher Scientific, USA) a izolována z leukocytů sadou Tempus Spin RNA Isolation kit (Thermo Fisher Scientific, USA) dle instrukcí výrobce. Koncentrace nukleových kyselin byla měřena na přístroji Nanodrop ND 1000 (Thermo Scientific, USA). Integrita RNA (RIN) byla analyzována na přístroji Agilent Bioanalyzer (Agilent Technologies, USA).

2.5 µg celkové RNA bylo použito pro přípravu mRNA knihovny sadou TruSeq Stranded mRNA Sample Preparation Kit (Illumina, USA) a sekvenováno na přístroji HiSeq 2500 (Illumina, USA) v konfiguraci „single-end” 1x 100.

FASTQ čtení byly alignovány programem STAR v.2.4.2a (Dobin *et al.*, 2013) oproti referenčnímu genomu hg38. Z následně generovaných alignmentů byly dále filtrovány jen unikátně mapovaná čtení (MQ=255). Pokrytí genů bylo počítáno balíkem Subread v.1.6.1 (Liao *et al.*, 2014). oproti referenčnímu GFF souboru Homo_sapiens.GRCh38.88.gtf (Ensembl). Analýza genové exprese byla paralelně zpracována programy DESeq2 (Love *et al.*, 2014) a Cufflinks v.2.2.1 (Trapnell *et al.*, 2012). V případě použití Cufflinks nebyla prováděna rekonstrukce transkriptů *de novo*, ale byl použit uvedený GFF referenční soubor. Pro další analýzu byly vybrány pouze geny se změněnou expresí, které byly reportovány jako signifikantně změněné (p-hodnoty <0.05) shodně v programech DESeq2 a Cufflinks a které vykazovaly stejný trend změny exprese v obou vyhodnoceních. Ovlivněné genové interakce, biologické procesy a signální dráhy byly zjišťovány pomocí webové aplikace Reactome (Fabregat *et al.*, 2018) a vlastními skripty psanými v jazyce Python 3 ve srovnání se zdrojovými soubory Gene ontology (GO) (Ashburner *et al.*, 2000), kde byly filtrovány nejvíce ovlivněné GO kategorie s minimální asociací 30 genů v rámci kompletního souboru genů lidského genomu.

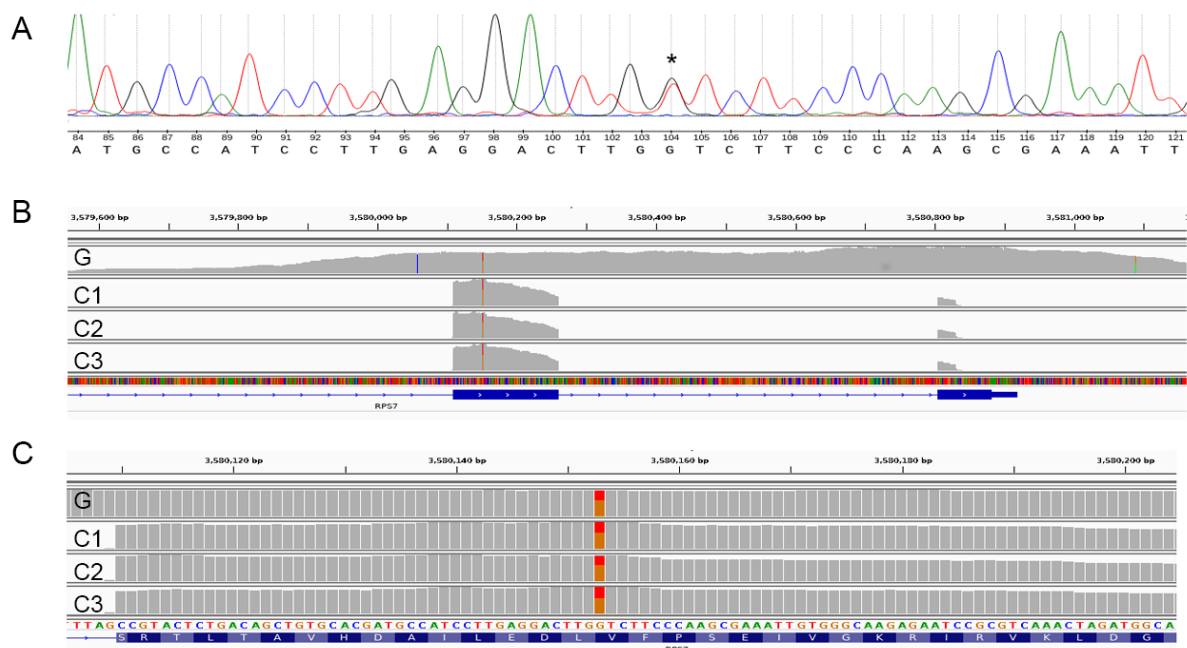
4.3 Výsledky

Český Národní DBA registr v současnosti zahrnuje 52 pacientů. Molekulární příčina onemocnění byla diagnostikována u 33 jedinců a to v genech *RPS19*, *RPS26*, *RPL5*, *RPL11* a *RPS17*.

4.3.1 Charakteristika nové mutace v RPS7

V této studii jsme provedli sekvenování genového panelu pro RP u vybraných pacientů bez nalezené mutace. U pacientky s nově nalezenou heterozygotní mutací hg38[chr2:g.3,580,153G>T], RPS7 p.V134F jsme získali celkově 108,774 FASTQ čtení pokrývající 200,696 nukleotidů zahrnující RP exonové regiony s minimálním pokrytím 10x a s celkovým průměrným pokrytím 52x. 74,924 RP exonových pozic nesplňovalo minimální pokrytí; z toho pouze 3,379 pozic bylo určeno pro kódující DNA oblast (CDS). Nalezená germinální heterozygotní transverze hg38[chr2:g.3,580,153G>T]; RPS7 p.V134F byla lokalizována v *RPS7* exonu 6 byla dále potvrzena kapilárním sekvenováním a ampikonovým sekvenováním cDNA (při průměrném pokrytí 78,285), která potvrdila stanovením penetrace mutované alely heterozygotní konformaci genotypu u všech testovaných (pacientka, matka, sestra) s následnými hodnotami zachytu: proband: 55 % *RPS7^{wt}* / 45 % *RPS7^{mut}*, matka: 58 % *RPS7^{wt}* / 42% *RPS7^{mut}*, sestra: 63% *RPS7^{wt}* / 37% *RPS7^{mut}* (Obrázek 4.1).

Nově identifikovaná SNV nesynonymní mutace v *RPS7* není popsána v databázích lidských DNA variant. Záměna konzervovaného valinu v pozici 134 byla hodnocena v Polyphen2 (skóre 0.965) jako potencionálně poškozuující. Aminokyselinová záměna p.V134F byla hodnocena *in silico* jako substituce snižující stabilitu proteinu (I-Mutant index 9, MUpro score - 0.97 metodou neurálních sítí a - 1 metodou SVM (support vector machines)).



Obrázek 4.1: RPS7 mutace u DBA pacientky a dvou asymptomatických členů rodiny

Potvrzení heterozygotní mutace hg38[chr2:g.3,580,153G>T] kapilárním sekvenováním panel (A). Mutace byla u DBA pacientky objevena MPS RP panelu genů (G). Alelická rovnováha byla prověřena amplikonovým „ultra-deep” sekvenováním cDNA *RPS7* u pacientky (C1), matky pacientky (C2) a její sestry (C3). Vizualizace alignmentu exonu 6 a 7 genu *RPS7* v IGV panelu (B). Panel (C) zobrazuje detailní oblast exonu 6 *RPS7* s detekovanou mutací u všech vyšetřovaných členů rodiny.

4.3.2 Analýza diferenciální genové exprese

Transkriptomovou analýzu jsme provedli za účelem pochopení molekulárních důsledků mutace *RPS7* genu ve srovnání s nejčastěji mutovaným genem u DBA (*RPS19*) a kontrolními profily zdravých jedinců. Ze sekvenování transkriptomu bylo získáno od 27 do 122 milionů unikátně mapovaných čtení na vzorek. Z unikátně mapovaných čtení 79 až 92 % bylo s jednoznačným významem zahrnuto ke kvantifikaci genové exprese. Výsledky alignmentu jsou shrnuty v Tabulce 4.1.

Celkově bylo programy rozeznáno více jak 55 tisíc genů (Subread 56,505; Cufflinks 58,219). V případě srovnání skupiny *RPS7* se zdravými kontrolami bylo rozpoznáno 1,555 genů se signifikantně změněnou expresí adj. p-hodnota <0.05 programem DESeq2 a 3,632 genů bylo reportováno v Cufflinks se signifikantně změněnou expresí (p-hodnota <0.05) oproti zdravým kontrolám. 627 genů bylo shodně reportováno v obou programech a z toho 600 genů mělo i stejný trend změny exprese vůči kontrolám (Příloha 4.1).

V případě srovnání skupiny *RPS19* se zdravými kontrolami bylo identifikováno 5,419 genů se signifikantní změnou exprese (DESeq2), 1,285 bylo reportováno programem

Cufflinks. 1,065 genů bylo nalezeno analýzou průniku a 1,049 bylo se shodným trendem změny exprese (Příloha 4.2). 294 genů vykazovalo signifikantně změněnou expresi se stejným trendem shodně ve skupině RPS7 a RPS19 oproti kontrolnímu souboru (Příloha 4.3).

Ve srovnání skupiny RPS7 oproti RPS19 bylo hodnoceno 197 genů se signifikantně změněnou expresí, zatímco DESeq2 reportoval 29 genů; 7 genů bylo hodnoceno shodně a ve všech případech byl zachován trend změny exprese (Příloha 4.4). Do hodnocení ovlivnění signálních drah a procesů, ve kterých jsou zahrnuty dysregulované geny, byly použity pouze soubory, které byly shodně reportovány oběma programy Cufflinks i DESeq2.

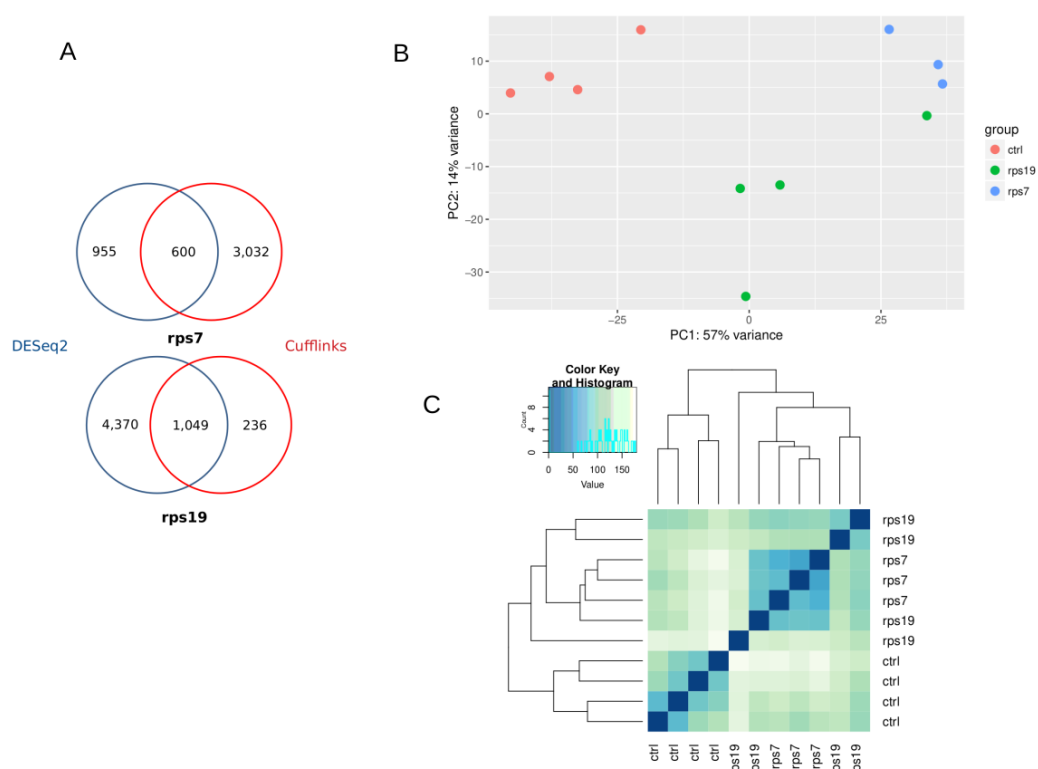
Číslo vzorku	zdravé kontroly			
	ctrl_1	ctrl_2	ctrl_3	ctrl_4
Přiřazená čtení	70,652,954	122,042,088	109,302,390	76,722,987
Nejednoznačně přiřazená čtení	1,864,958	3,132,073	2,737,289	1,738,775
Nepřiřazená čtení	13,938,392	26,715,833	25,871,835	11,416,639
Přiřazená čtení (%)	82	80	79	85
Nepřiřazená čtení (%)	16	18	19	13

Číslo vzorku	RPS19			
	rps19_1	rps19_2	rps19_3	rps19_4
Přiřazená čtení	45,078,325	57,240,169	27,205,710	76,855,421
Nejednoznačně přiřazená čtení	1,350,785	1,493,583	826,802	1,574,002
Nepřiřazená čtení	3,970,800	6,648,025	3,230,725	8,513,705
Přiřazená čtení (%)	89	88	87	88
Nepřiřazená čtení (%)	8	10	10	10

Číslo vzorku	RPS7		
	rps7_1	rps7_2	rps7_3
Přiřazená čtení	91,051,448	112,748,395	114,516,238
Nejednoznačně přiřazená čtení	2,143,768	2,680,731	2,587,373
Nepřiřazená čtení	9,225,062	8,755,412	7,975,500
Přiřazená čtení (%)	89	91	92
Nepřiřazená čtení (%)	9	7	6

Tabulka 4.1: Počet získaných čtení pro jednotlivé knihovny a úspěšnost alignmentu

Souhrn získaných čtení pro testované skupiny a úspěšnost alignmentu.



Obrázek 4.2: Souhrnné statistické hodnocení analyzovaných skupin v transkriptomové studii

Panel (A) zobrazuje průniky expresně dysregulovaných genů na hladině významnosti p -hodnota ≤ 0.05 hodnocených programy DESeq2 a Cufflinks. Panel (B) PCA analýza testovaných skupin hodnocena programem DESeq2. Panel (C) zobrazuje heatmapu analýzy „sample-to-sample” vzdáleností „variance-stabilizing” transformovaných hodnot hodnocené programem DESeq2.

4.3.2.1 RPS7 transkriptom

Analýzou v Reactome portálu bylo identifikováno 52 signálních drah jako signifikantně změněných (p -hodnota < 0.05) (Příloha 4.5). Rozpoznané dráhy lze shrnout do několika základních procesů, které jsou v asociaci s WNT signální dráhou (*TLE2*, *MYC*, *TCF7*, *LEF1*, *AXIN2*, *AES*), biogenezí ribosomů a translační aktivitou (*RPL3*, *EIF1AY*, *RPL36A*, *RPS4Y1*, *RPS16*, *RPS19*, *EIF3K*, *EIF3I*, *RPL13*, *RPLP2*, *RPL38*, *RPL28*, *RPL19*), rRNA procesováním (*NOP14*, *RPL3*, *RPL36A*, *SNU13*, *RPS4Y1*, *BYSL*, *FBL*, *RPS16*, *RPS19*, *PES1*, *RPL13*, *RPLP2*, *RPL38*, *RPL28*, *RPL19*), NOTCH signální dráhou (*PTCRA*, *FCER2*, *ADAM17*, *TLE2*, *ELF3*, *AGO3*, *EGF*, *FBXW7*, *MYC*, *E2F1*, *MIB1*), regulací buněčného cyklu (*BTG2*, *PLAGL1*, *E2F1*, *MYC*, *MYBL2*, *CASP8*, *CFLAR*, *PMAIP1*) a imunitním systémem (*FCER2*, *CXCL8*, *IL1B*, *CCL3L3*, *CCL3*, *CCL2*, *CXCL1*, *PTGS2*, *TLR1*, *MAL*, *TLR6*, *TLR4*, *CASP8*, *IL1B*).

Obdobné výsledky byly získány také GO analýzou (Tabulka 4.2), kdy pro skupinu RPS7 byly identifikovány procesy ovlivňující imunitní odpověď; v rámci imunoglobulinových komplexů je sledována zvýšená exprese oproti kontrolám (*CD79A*, *IGHD*, *IGHM*, *IGHV1-2*, *IGHV1-46*, *IGHV1-58*, *IGHV1-69D*, *IGHV4-39*, *TRBC1*, *TRBC2*). V rámci procesu regulace chemotaxe neutrofilních granulocytů je sledována snížená exprese u chemokinů (*CXCL1*, *CXCL8*) a zvýšená exprese (*CCR7*, *IL23A*, *PF4V1*). Další procesy zahrnují vazebné mechanismy na pre-mRNA, kde byly identifikovány 4 geny se sníženou expresí (*CELF2*, *DDX5*, *RNPC3*, *U2AF1L5*) a *SNRPC* se zvýšenou expresí. Zvýšené hladiny exprese byly zaznamenány u genů podílejících se na rRNA procesování (SSU-rRNA, 5.8S rRNA, LSU-rRNA) a to konkrétně u 4 genů *BYSL*, *NOP14*, *RPS16*, *RPS19* a *TSR2*. U drah asociovaných s RP geny sledujeme zvýšenou expresi *RPS19*, *RPL13*, *RPL19*, *RPL28*, *RPL3*, *RPL38*, *RPLP0P6* a *RPLP2*. Naopak snížená exprese v reportovaných drahách je pouze u jednoho ribozomálního genu *RPL36A*. 7 genů bylo asociováno s vazebnou specifitou k ribosomům, z toho 3 se sníženou expresí (*CPEB4*, *FMR1*, *RICTOR*) a 4 se zvýšenou expresí (*EIF3K*, *EIF5AL1*, *RACK1*, *RPN2*).

4.3.2.2 RPS19 transkriptom

Programem Reactome bylo reportováno 72 signálních drah, ovlivněných expresí genovou dysregulací v RPS19 skupině (Příloha 4.6). Mezi nejvíce ovlivněné procesy patří skupiny signálních drah podílejících se na imunitní odpovědi: interleukin-10 signalizace (*CXCL8*, *IL1R1*, *IL1R2*, *PTAFR*, *CXCL1*, *PTGS2*, *TNF*, *FCER2*, *CCL5*, *IL1B*, *CCL4*, *CCL3*, *CCL2*), kaskáda komplementu (imunoglobulinové geny - *IGHV*, *C5AR1*, *CD19*, *CD81*, *PROS1* a další), signalizace B-lymfocytárních receptorů, FCER1 NF-κB aktivace (*FCER1A*, *LYN*, *TAB2*, *MS4A2*, *UBA52*, *CARD11* a imunoglobulinové geny), degranulace neutrofilních granulocytů (*DDX3X*, *TNFAIP6*, *RAB3D*, *CXCL1*, *C4BPA*, *LRMP*, *SIRPB1*, *FCAR*, *MMP25*, *PNP*, *LAMP2*, *PGM2*, *SPTAN1*, *DSP*, *HLA-H*, *CR1*, *CD93*, *MME*, *APAF1*, *TUBB*, *KRT1*, *APRT*, a.j.). Mezi další ovlivněné dráhy patří regulace buněčného cyklu a apoptózy přes TP53 signalizační kaskádu (*BTG2*, *CDKN1A*, *PRMT1*, *PLAGL1*, *CARM1*, *E2F1*, *SFN*), aktivaci NOXA (*E2F1*, *PMAIP1*) a WNT signální dráhy (*TLE2*, *MYC*, *TCF3*, *AXIN2*).

Ve srovnání s GO v souboru 30 procesů obsahující minimálně 30 genů v rámci kompletního souboru lidského genomu bylo identifikováno celkem 18 GO kategorií spojených s imunitním systémem (Tabulka 4.3). Jedná se například o stimulaci T-lymfocytů, kde byly nalezeny 3 geny se sníženou expresí (*LYN*, *MAP3K8*, *TNFSF13B*) a 19 genů se zvýšenou expresí (*CARD11*, *CCR7*, *CD28*, *CD3D*, *CD3E*, *CD40LG*, *CD5*, *GRAP2*, *HLA-DRB5*, *ICOSLG*, *LCK*, *TMIGD2*, *TNFRSF13C*, *TRAC*, *TRAV29DV5*, *TRBC1*). V genech

pro imunoglobulinové komplexy nebyl nalezen žádný gen se sníženou expresí, ale 20 genů vykazovalo zvýšenou expresi (*CD79A*, *CD79B*, *IGHD*, *IGHG1*, *IGHG3*, *IGHM*, *IGHV1-18*, *IGHV1-2*, *IGHV1-24*, *IGHV1-69D*, *IGHV2-26*, *IGHV2-5*, *IGHV4-39*, *IGHV4-59*, *IGHV5-51*, *IGKC*, *IGKV3-20*, *IGLC3*, *IGLC6*, *TRBC1*, *TRBC2*). Zvýšená exprese genů imunoglobulinové rodiny variabilní oblasti těžkého řetězce IGHV byla pozorována i v případě skupiny RPS7. Podobně jako ve skupině RPS7 byla snížená exprese u některých chemokinů regulujících chemotaxi neutrofilních granulocytů (*C5AR1*, *CXCL1*, *CXCL8*, *CXCR2*, *DAPK2*, *IL1R1*, *MOSPD2*). V rámci tohoto procesu byla naopak zjištěna zvýšená exprese u *CCR7*, *CD74*, *IL23A*, *PPBP*. Nekróptóza, programovaná buněčná smrt iniciovaná vazbou ligandu na receptor např. signální dráhou Toll, byla také identifikována jako jeden z obohacených procesů v rámci GO analýzy. U nekroptózy bylo zahrnuto 6 genů (*CFLAR*, *FAS*, *LY96*, *PPIF*, *TLR4*, *TNF*) se sníženou expresí a 4 geny se zvýšenou expresí (*TRAF2*, *UBA52*, *UBB*, *YBX3*). Ve spojitosti s nekroptózou v rámci dalších obohacených procesů byla zaznamenána MyD88 toll-like signální dráha, kde je řazeno 8 genů se zvýšenou expresí (*IRAK3*, *LY96*, *TAB2*, *TLR1*, *TLR2*, *TLR4*, *TLR6*, *TLR8*) a 2 geny se sníženou expresí (*UBA52*, *UBB*). V rámci hodnoceného souboru byly přítomny dva procesy týkající se hematopoézy, a to proces regulace aktivace trombocytů s 5 geny se sníženou expresí (*CD9*, *LYN*, *PLEK*, *THBD*, *TLR4*) a dvěma geny se zvýšenou expresí (*DMTN*, *NOS3*) a dále proces vývoje erytrocytů s jedním genem se sníženou expresí (*FLVCR1*) a 5 geny se zvýšenou expresí (*DMTN*, *EPB42*, *FAM210B*, *TAL1*, *TRIM58*). Dalších 6 genů bylo zařazeno do kategorie odpovědi na mineralokortikoidy, 3 geny se zvýšenou expresí (*FOS*, *FOSB*, *FOXO3*) a 3 geny se sníženou expresí (*CDKN1A*, *CTGF*, *PARP1*).

4.3.2.3 DBA transkriptom

U společně dysregulovaných genů, jak pro skupinu RPS7, tak pro RPS19, bylo programem Reactome rozpoznáno 59 signálních drah (Příloha 4.7). Stejně jako v předchozích případech největší část tvoří signální dráhy imunitního systému (*FCER2*, *CXCL8*, *IL23A*, *MYC*, *IL1B*, *S1PR1*, *CCL2*, *FOS*, *PTGS2*, *EGR1*, *HLA-H*, *CXCL8*, *CXCL1*, *FOS*, *IL1RAP*, *PTGS2*, *IL27RA*, *FCER2*, *MT2A*, *CA1*, *IFI27*, *IL23A*, *LCK*, *MYC*, *IL1B*, *CCL3*, *PELI1*, *IL21R*, *S1PR1*, *CCL2*, *MAP3K8*, *LTB*, *SOS2*), WNT signální dráhy (*TLE2*, *MYC*, *AXIN2*), NOTCH signalizace (*PTCRA*, *FCER2*, *TLE2*, *EGF*, *MYC*, *E2F1*), TP53 signální dráhy (*BTG2*, *PLAGL1*, *E2F1*, *TP53INP1*, *PMAIP1*), regulace buněčného cyklu a apoptózy (*MYC*, *E2F1*, *MYBL2*, *TK1*, *CCNL1*, *CFLAR*, *TLR4*) a faktory ovlivňující vývoj megakaryocytů a trombocytů (*DOCK5*, *TUBB2A*, *HBG2*, *HBG1*, *TUBB*, *HBD*, *KIF27*).

V hodnocení společných dysregulovaných genů skupin RPS7 a RPS19 pomocí GO analýzy (Tabulka 4.4) se obdobně vyskytují procesy spojené s imunitním systémem, jako je

regulace chemotaxe neutrofilních granulocytů se zvýšenou expresí chemokinových genů (*CXCL1*, *CXCL8*) a se sníženou expresí chemokinového receptoru *CCR7* a interleukinu *IL23A*. V procesu regulace biosyntézy cytokinů je zařazeno 5 genů se sníženou expresí (*EGR1*, *IL1B*, *TLR1*, *TLR4*, *TLR6*) a dva geny s expresí zvýšenou (*CD3E*, *LTB*). V obou skupinách sledujeme zvýšenou expresi genu *MYC*, který je ve filtrované GO analýze zařazen do procesu negativní regulace myeloidní leukocytární diferenciace společně s dalšími třemi geny se sníženou expresí (*CCL3*, *TLR4*, *TRIB1*). Reportovány byly také některé molekulární přenašeče zejména skupiny hemoglobinových genů (*HBD*, *HBG1*, *HBG2*) se zvýšeným expresním trendem oproti kontrolní skupině, naopak snížená exprese byla zaznamenána u přenašeče ATPázového tranportéru mědi *ATP7A*. Klastrovací analýza DBA společných genů s dysregulovanou expresí je zobrazena na Obrázku 4.5.

GO ID	Kategorie	Název	Celkový #	% ovlivněných genů	Název Genu
Snížená exprese					
GO:0042788	cellular_component	polysomal ribosome	5	17%	FMRI1,NUPF2,RPL36A
GO:0090023	biological_process	positive regulation of neutrophil chemotaxis	5	17%	CXCL1,CXCL8
GO:0045737	biological_process	positive regulation of cyclin-dependent protein serine/threonine kinase activity	6	16%	ADAM17,CNN1L,CCNT2,PKD2
GO:0071624	biological_process	positive regulation of granulocyte chemotaxis	5	16%	CXCL1,CXCL8
GO:0045058	biological_process	T cell selection	7	16%	.
GO:0032735	biological_process	positive regulation of interleukin-12 production	5	16%	TLR4
GO:1904637	biological_process	beta-catenin-TCF complex assembly	5	16%	.
GO:1904031	biological_process	positive regulation of cyclin-dependent protein kinase activity	6	15%	ADAM17,CNN1L,CCNT2,PKD2
GO:1902624	biological_process	positive regulation of neutrophil migration	5	15%	CXCL1,CXCL8
GO:0090184	biological_process	positive regulation of kidney development	6	15%	CFLAR,EGR1
GO:0031295	biological_process	T cell costimulation	11	14%	MAP3K8
GO:0090022	biological_process	regulation of neutrophil chemotaxis	5	14%	CXCL1,CXCL8
GO:0031294	biological_process	lymphocyte costimulation	11	14%	MAP3K8
GO:0036002	molecular_function	pre-mRNA binding	5	13%	CELF2,DDX5,RNPC3,U2AF1L5
GO:0030515	molecular_function	snoRNA binding	4	13%	.
GO:0016417	molecular_function	S-acyltransferase activity	4	13%	ZDHHC17,ZDHHC21
GO:0000462	biological_process	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	5	13%	.
GO:0045429	biological_process	positive regulation of nitric oxide biosynthetic process	5	13%	IL1B,PKD2,PTGS2,TLR4,TLR6
GO:0043368	biological_process	positive T cell selection	4	13%	.
GO:0006882	biological_process	cellular zinc ion homeostasis	5	13%	.
GO:0002762	biological_process	negative regulation of myeloid leukocyte differentiation	6	13%	CCL3,FBXW7,TLR4,TRIB1
GO:1904407	biological_process	positive regulation of nitric oxide metabolic process	5	13%	IL1B,PKD2,PTGS2,TLR4,TLR6
GO:1902622	biological_process	regulation of neutrophil migration	5	13%	CXCL1,CXCL8
GO:0051385	biological_process	response to mineralocorticoid	4	13%	FOS,FOSB
GO:0071353	biological_process	cellular response to interleukin-4	4	13%	.
GO:0071276	biological_process	cellular response to cadmium ion	5	13%	ATP7A,FOS
GO:0042088	biological_process	T-helper 1 type immune response	5	13%	TLR4,TLR6
GO:0022625	cellular_component	cytosolic large ribosomal subunit	9	12%	RPL36A
GO:0019814	cellular_component	immunoglobulin complex	10	12%	.
GO:0043022	molecular_function	ribosome binding	7	12%	CPEB4,FMRI1,RICTOR
Zvýšená exprese					
GO:0042788	cellular_component	polysomal ribosome	5	17%	RPL19,RPL38
GO:0090023	biological_process	positive regulation of neutrophil chemotaxis	5	17%	CCR7,IL23A,PF4V1
GO:0045737	biological_process	positive regulation of cyclin-dependent protein serine/threonine kinase activity	6	16%	COR2D,FAM58A
GO:0071624	biological_process	positive regulation of granulocyte chemotaxis	5	16%	CCR7,IL23A,PF4V1
GO:0045058	biological_process	T cell selection	7	16%	BCL11B,CCR7,CD3D,CD3E,GATA3,IL23A,LY9
GO:0032735	biological_process	positive regulation of interleukin-12 production	5	16%	CCR7,CD40,IL23A,LTB
GO:1904637	biological_process	beta-catenin-TCF complex assembly	5	16%	AXIN2,LEF1,MYC,TCF7,TLE2
GO:1904031	biological_process	positive regulation of cyclin-dependent protein kinase activity	6	15%	COR2D,FAM58A
GO:1902624	biological_process	positive regulation of neutrophil migration	5	15%	CCR7,IL23A,PF4V1
GO:0090184	biological_process	positive regulation of kidney development	6	15%	GATA3,MYC,NOG,PDGFA
GO:0031295	biological_process	T cell costimulation	11	14%	CCR7,CD3D,CD3E,DPP4,GRAP2,HLA-DPB1,ICK,TMIGD2,TRAC,TRBC1
GO:0090022	biological_process	regulation of neutrophil chemotaxis	5	14%	CCR7,IL23A,PF4V1
GO:0031294	biological_process	lymphocyte costimulation	11	14%	CCR7,CD3D,CD3E,DPP4,GRAP2,HLA-DPB1,ICK,TMIGD2,TRAC,TRBC1
GO:0036002	molecular_function	pre-mRNA binding	5	13%	SNRPC
GO:0030515	molecular_function	snoRNA binding	4	13%	BYSL,NOP14,NUDT1,SNU13
GO:0016417	molecular_function	S-acyltransferase activity	4	13%	MCAT,ZDHHC4
GO:0000462	biological_process	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	5	13%	BYSL,NOP14,RPS16,RPS19,TSR2
GO:0045429	biological_process	positive regulation of nitric oxide biosynthetic process	5	13%	.
GO:0043368	biological_process	positive T cell selection	4	13%	BCL11B,CD3D,IL23A,LY9
GO:0006882	biological_process	cellular zinc ion homeostasis	5	13%	LCK,MT1F,MT1X,MT2A,SLC39A8
GO:0002762	biological_process	negative regulation of myeloid leukocyte differentiation	6	13%	C1QC,MYC
GO:1904407	biological_process	positive regulation of nitric oxide metabolic process	5	13%	.
GO:1902622	biological_process	regulation of neutrophil migration	5	13%	CCR7,IL23A,PF4V1
GO:0051385	biological_process	response to mineralocorticoid	4	13%	MAOB,PARP1
GO:0071353	biological_process	cellular response to interleukin-4	4	13%	GATA3,LEF1,RPL3,TCF7
GO:0071276	biological_process	cellular response to cadmium ion	5	13%	MT1F,MT1X,MT2A
GO:0042088	biological_process	T-helper 1 type immune response	5	13%	IL23A,IL27RA,LEF1
GO:0022625	cellular_component	cytosolic large ribosomal subunit	9	12%	RPL13,RPL19,RPL28,RPL3,RPL38,RPLP0P6,RPLP2,SNU13
GO:0019814	cellular_component	immunoglobulin complex	10	12%	CD79A,IGHD,IGHM,IGHV1-2,IGHV1-46,IGHV1-58,IGHV1-69D,IGHV4-39,TRBC1,TRBC2
GO:0043022	molecular_function	ribosome binding	7	12%	EIF3K,EIFSAL1,RACK1,RPN2

Tabulka 4.2: Třicet nejvíce ovlivněných GO kategorií pro skupinu RPS7

Nejvíce ovlivněné termíny genové ontologie (GO) ve skupině RPS7. Procento přiřazených genů v ontologické kategorii se změnou exprese na hladině významnosti p-hodnoty ≤ 0.05 z celkového počtu genů pro ontologickou kategorii. Celková (#) hodnota udává počet přiřazených genů se změnou exprese.

GO ID	Kategorie	Název	Celkový #	% ovlivněných genů	Název Genu
Snižená exprese					
GO:1902622	biological_process	regulation of neutrophil migration	11	28%	CSAR1,CXCL1,CXCL8,CXCR2,DAPK2,IL1R1,MOSPD2
GO:0001772	cellular_component	immunological synapse	9	27%	AL,CAM,PRKARIA,SNX27
GO:0002755	biological_process	MyD88-dependent toll-like receptor signaling pathway	10	27%	IRAK3,LY96,TAB2,TLR1,TLR2,TLR4,TLR6,TLR8
GO:0070266	biological_process	necroptotic process	10	27%	CFIAR,FAS,LY96,PIIF,TLR4,TNF
GO:0031663	biological_process	lipopolysaccharide-mediated signaling pathway	14	26%	CARD8,CCL2,CCL3,IL1B,LY96,LYN,PTAFR,TLR2,TLR4,TNF,TRIB1
GO:0019814	cellular_component	immunoglobulin complex	21	25%	FAM212B,PRKARIA,SFN
GO:0030291	molecular_function	protein serine/threonine kinase inhibitor activity	8	24%	LYN,MAP3K8,TNFSF13B
GO:0031295	biological_process	T cell costimulation	19	24%	
GO:0045058	biological_process	T cell selection	11	24%	
GO:0097300	biological_process	programmed necrotic cell death	10	24%	CFIAR,FAS,LY96,PIIF,TLR4,TNF
GO:0042571	cellular_component	immunoglobulin complex, circulating	19	23%	
GO:0030449	biological_process	regulation of complement activation	25	23%	C4BPA,CSAR1,CR1
GO:0071622	biological_process	regulation of granulocyte chemotaxis	11	23%	CSAR1,CXCL1,CXCL8,CXCR2,DAPK2,MOSPD2
GO:0005547	molecular_function	phosphatidylinositol-3,4,5-trisphosphate binding	8	22%	ARAP3,DAPP1,GAB2,OGT,ZFYVE16
GO:0002756	biological_process	MyD88-independent toll-like receptor signaling pathway	7	22%	LY96,TLR4,TLR6,USB2D1
GO:0010543	biological_process	regulation of platelet activation	7	22%	CD9,LYN,PLEK,THBD,TLR4
GO:2000257	biological_process	regulation of protein activation cascade	25	22%	C4BPA,CSAR1,CR1
GO:0002920	biological_process	regulation of humoral immune response	28	22%	C4BPA,CSAR1,CR1,TNF
GO:0071470	biological_process	cellular response to osmotic stress	7	22%	DDX3X,PTGS2,SLC12A6,ZFP36L1
GO:0034987	molecular_function	immunoglobulin receptor binding	18	21%	
GO:0090218	biological_process	positive regulation of lipid kinase activity	7	21%	FLT3,KIT,LYN
GO:0072604	biological_process	interleukin-6 secretion	7	21%	DDX58,F2RL1,IL1B,IL1RAP,LEP,TLR8
GO:0030890	biological_process	positive regulation of B cell proliferation	9	21%	IRS2,PIEL1,TLR4,TNFSF13B
GO:0006910	biological_process	phagocytosis, recognition	20	21%	CLCE7A
GO:0004896	molecular_function	cytokine receptor activity	19	20%	CSF2RB,CXCR1,CXCR2,CXCR4,FLT3,IFNGR1,IFNGR2,IL13RA1,IL13RA2,IL13RA3,IL13RA4,IL13RA5,IL13RA6,IL13RA7,IL13RA8,IL13RA9,IL13RA10,IL13RA11,IL13RA12,IL13RA13,IL13RA14,IL13RA15,IL13RA16,IL13RA17,IL13RA18,IL13RA19,IL13RA20,IL13RA21,IL13RA22,IL13RA23,IL13RA24,IL13RA25,IL13RA26,IL13RA27,IL13RA28,IL13RA29,IL13RA30,IL13RA31,IL13RA32,IL13RA33,IL13RA34,IL13RA35,IL13RA36,IL13RA37,IL13RA38,IL13RA39,IL13RA40,IL13RA41,IL13RA42,IL13RA43,IL13RA44,IL13RA45,IL13RA46,IL13RA47,IL13RA48,IL13RA49,IL13RA50,IL13RA51,IL13RA52,IL13RA53,IL13RA54,IL13RA55,IL13RA56,IL13RA57,IL13RA58,IL13RA59,IL13RA60,IL13RA61,IL13RA62,IL13RA63,IL13RA64,IL13RA65,IL13RA66,IL13RA67,IL13RA68,IL13RA69,IL13RA70,IL13RA71,IL13RA72,IL13RA73,IL13RA74,IL13RA75,IL13RA76,IL13RA77,IL13RA78,IL13RA79,IL13RA80,IL13RA81,IL13RA82,IL13RA83,IL13RA84,IL13RA85,IL13RA86,IL13RA87,IL13RA88,IL13RA89,IL13RA90,IL13RA91,IL13RA92,IL13RA93,IL13RA94,IL13RA95,IL13RA96,IL13RA97,IL13RA98,IL13RA99,IL13RA100
GO:0043552	biological_process	positive regulation of phosphatidylinositol 3-kinase activity	6	20%	FLT3,KIT,LYN
GO:0050853	biological_process	B cell receptor signaling pathway	26	20%	ELF2,LYN
GO:0050701	biological_process	interleukin-1 secretion	10	20%	ABCA1,CARD8,CCL3,F2RL1,IL1IR2,TLR4,TLR6,TLR8
GO:0051385	biological_process	response to mineralocorticoid	6	20%	FOS,FOSB,FOXO3
GO:0048821	biological_process	erythrocyte development	6	20%	FLVCR1
Zvýšená exprese					
GO:1902622	biological_process	regulation of neutrophil migration	11	28%	CCR7,CD74,IL23A,PPBP
GO:0001772	cellular_component	immunological synapse	9	27%	CARD11,CD28,CD3E,CD81,EZR,LCK
GO:0002755	biological_process	MyD88-dependent toll-like receptor signaling pathway	10	27%	UBA52,UBB
GO:0070266	biological_process	necroptotic process	10	27%	TRAF2,UBA52,UBB,YBX3
GO:0031663	biological_process	lipopolysaccharide-mediated signaling pathway	14	26%	CCL5,CD180,NOS3
GO:0019814	cellular_component	immunoglobulin complex	21	25%	CD79A,CD79B,IGHD,IGHG1,IGHG3,IGHM,IGHV1-18,IGHV1-2,IGHV1-24,IGHV1-69D,IGHV2-26,IGHV2-5,IGHV4-39,IGHV4-59,IGHV5-51,IGKC,IGKV3-20,IGLC3,IGLC5,TRBC1,TRBC2
GO:0030291	molecular_function	protein serine/threonine kinase inhibitor activity	8	24%	CDKN1A,CIB1,HSPB1,PKIA,PRKAR2B
GO:0031295	biological_process	T cell costimulation	19	24%	CARD11,CCR7,CD28,CD3D,CD3E,CD40LG,CD5,GRAP2,HLA-DRB5,ICOSLG,LCK,TMIGD2,TNFRSF13C,TRAC,TRAV29DV5,TRBC1
GO:0045058	biological_process	T cell selection	11	24%	BATF,BCL11B,CARD11,CCR7,CD28,CD3D,CD3E,CD74,IL23A,LY9,THEMIS
GO:0097300	biological_process	programmed necrotic cell death	10	24%	TRAF2,UBA52,UBB,YBX3
GO:0042571	cellular_component	immunoglobulin complex, circulating	19	23%	IGHD,IGHG1,IGHG3,IGHM,IGHV1-18,IGHV1-2,IGHV1-24,IGHV1-69D,IGHV2-26,IGHV2-5,IGHV4-39,IGHV4-59,IGHV5-51,IGKC,IGKV3-20,IGLC3,IGLC5,TRBC1,TRBC2
GO:0030449	biological_process	regulation of complement activation	25	23%	CD19,CD81,CLU,IGHG1,IGHG3,IGHV1-2,IGHV2-5,IGHV4-39,IGHV4-59,IGKC,IGKV3-15,IGKV3-20,IGKV4-1,IGLC6,IGLV1-40,IGLV1-51,IGLV2-23,IGLV2-8,IGLV3-1,IGLV3-19,PROS1
GO:0071622	biological_process	regulation of granulocyte chemotaxis	11	23%	CCL5,CCR7,CD74,IL23A,PPBP
GO:0005547	molecular_function	phosphatidylinositol-3,4,5-trisphosphate binding	8	22%	HIP1R,MYO1B,NRGN
GO:0002756	biological_process	MyD88-independent toll-like receptor signaling pathway	7	22%	IKBKE,UBA52,UBB
GO:0010543	biological_process	regulation of platelet activation	7	22%	DMTN,NOS3
GO:2000257	biological_process	regulation of protein activation cascade	25	22%	CD19,CD81,CLU,IGHG1,IGHG3,IGHV1-2,IGHV2-5,IGHV4-39,IGHV4-59,IGKC,IGKV3-15,IGKV3-20,IGKV4-1,IGLC6,IGLV1-40,IGLV1-51,IGLV2-23,IGLV2-8,IGLV3-1,IGLV3-19,PROS1
GO:0002920	biological_process	regulation of humoral immune response	28	22%	MYLK,SLC2A1,YBX3
GO:0071470	biological_process	cellular response to osmotic stress	7	22%	IGHD,IGHG1,IGHG3,IGHM,IGHV1-18,IGHV1-2,IGHV1-24,IGHV1-69D,IGHV2-26,IGHV2-5,IGHV4-39,IGHV4-59,IGHV5-51,IGKC,IGLC3,IGLC5,TRBC1,TRBC2
GO:0034987	molecular_function	immunoglobulin receptor binding	18	21%	CCR7,CD81,PTK2,VAV2
GO:0090218	biological_process	positive regulation of lipid kinase activity	7	21%	BANK1
GO:0072604	biological_process	interleukin-6 secretion	9	21%	CARD11,CD74,CD81,CDKN1A,TNFRSF13C
GO:0030890	biological_process	positive regulation of B cell proliferation	20	21%	IGHD,IGHG1,IGHG3,IGHM,IGHV1-18,IGHV1-2,IGHV1-24,IGHV1-69D,IGHV2-26,IGHV2-5,IGHV4-39,IGHV4-59,IGHV5-51,IGKC,IGLC3,IGLC5,PEAR1,TRBC1,TRBC2
GO:0006910	biological_process	phagocytosis, recognition	20	21%	CCR7,CD74,CXCR3,IFNLR1,IL21R,IL27RA
GO:0004896	molecular_function	cytokine receptor activity	19	20%	CCR7,PTK2,VAV2
GO:0043552	biological_process	positive regulation of phosphatidylinositol 3-kinase activity	6	20%	BLK,CD19,CD79A,CD79B,IGHD,IGHG1,IGHG3,IGHM,IGHV1-18,IGHV1-2,IGHV1-24,IGHV1-69D,IGHV2-26,IGHV2-5,IGHV4-39,IGHV4-59,IGHV5-51,IGKC,IGLC3,IGLC5,LCK,PAX5,TRBC1,TRBC2
GO:0050853	biological_process	B cell receptor signaling pathway	26	20%	CCR7,IGHD
GO:0050701	biological_process	interleukin-1 secretion	10	20%	CDKN1A,CTGF,PARP1
GO:0051385	biological_process	response to mineralocorticoid	6	20%	DMTN,EPB42,FAM210B,TAL1,TRIM58
GO:0048821	biological_process	erythrocyte development	6	20%	

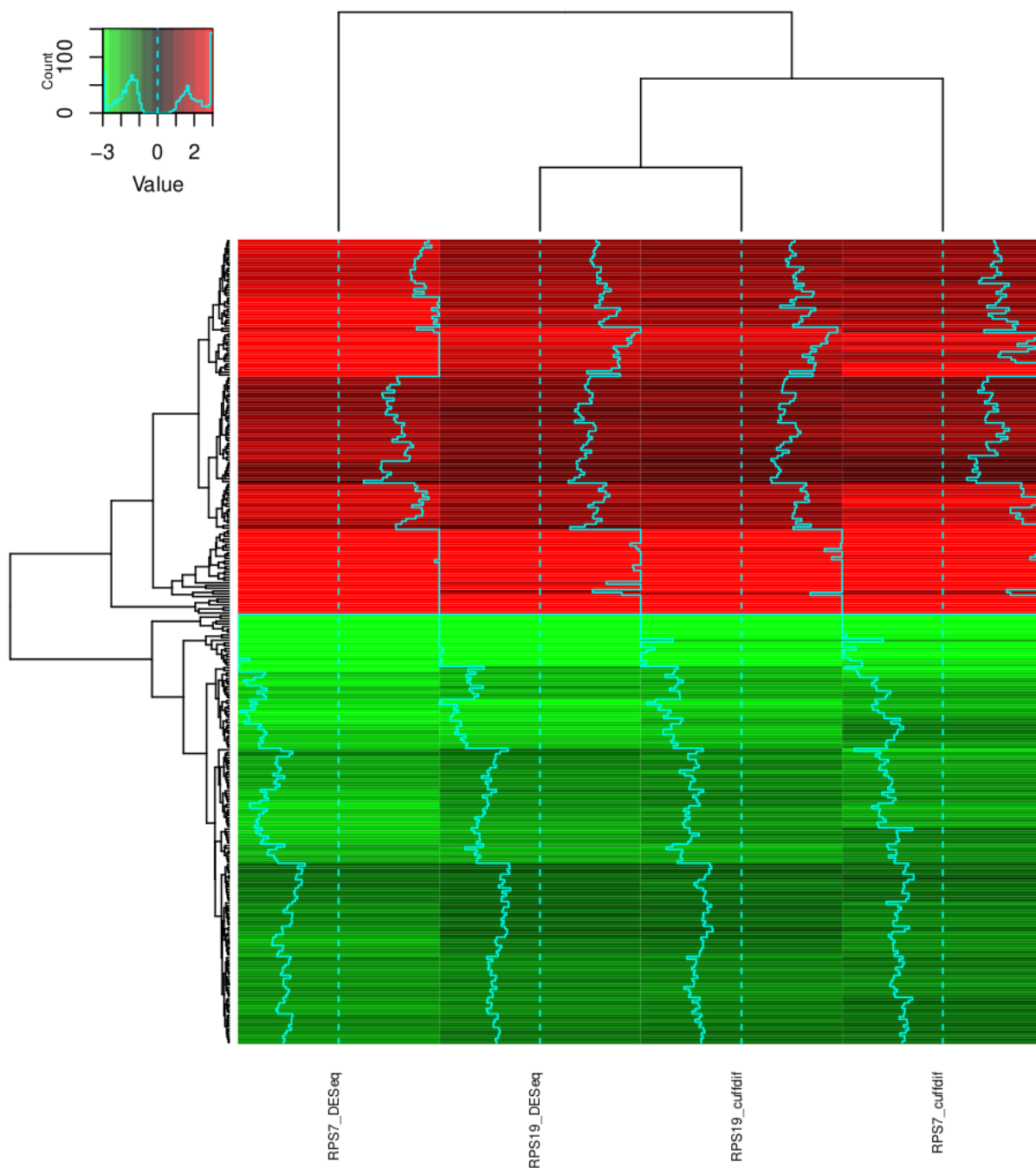
Tabulka 4.3: Třicet nejvíce ovlivněných GO kategorií pro skupinu RPS19

Nejvíce ovlivněné termíny genové ontologie (GO) ve skupině *RPS19*. Procento přiřazených genů v ontologické kategorii se změnou exprese na hladině významnosti p-hodnoty ≤ 0.05 z celkového počtu genů pro ontologickou kategorii. Celková (#) hodnota udává počet přiřazených genů se změnou exprese.

GO ID	Kategorie	Název	Celkový #	% ovlivněných genů	Název Genu
Snížená exprese					
GO:0090023	biological process	positive regulation of neutrophil chemotaxis	4	13%	CXCL1,CXCL8
GO:0045058	biological process	T cell selection	6	13%	.
GO:0140104	molecular function	molecular carrier activity	5	12%	ATP7A
GO:0031295	biological process	T cell costimulation	9	12%	MAP3K8
GO:1902624	biological process	positive regulation of neutrophil migration	4	12%	CXCL1,CXCL8
GO:0071624	biological process	positive regulation of granulocyte chemotaxis	4	12%	CXCL1,CXCL8
GO:0042108	biological process	positive regulation of cytokine biosynthetic process	7	12%	EGR1,IL1B,TLR1,TLR4,TLR6
GO:0032735	biological process	positive regulation of interleukin-12 production	4	12%	TLR4
GO:0000289	biological process	nuclear-transcribed mRNA poly(A) tail shortening	4	11%	BT G2,PAN3,ZCCHC6,ZFP36
GO:0090022	biological process	regulation of neutrophil chemotaxis	4	11%	CXCL1,CXCL8
GO:0045429	biological process	positive regulation of nitric oxide biosynthetic process	4	11%	IL1B,PTGS2,TLR4,TLR6
GO:1904407	biological process	positive regulation of nitric oxide metabolic process	4	11%	IL1B,PTGS2,TLR4,TLR6
GO:0031294	biological process	lymphocyte costimulation	9	11%	MAP3K8
GO:0071276	biological process	cellular response to cadmium ion	4	11%	ATP7A,FOS
GO:0004712	molecular function	protein serine/threonine/tyrosine kinase activity	4	10%	CLK1,CLK4,LRRK2,SGK1
GO:0050716	biological process	positive regulation of interleukin-1 secretion	3	10%	CCL3,TLR4
GO:0050701	biological process	interleukin-1 secretion	5	10%	CCL3,TLR4,TLR6
GO:1902622	biological process	regulation of neutrophil migration	4	10%	CXCL1,CXCL8
GO:0051491	biological process	positive regulation of filopodium assembly	3	10%	FMR1,RHOQ
GO:0051385	biological process	response to mineralocorticoid	3	10%	FOS,FOSB
GO:1904837	biological process	beta-catenin-TCF complex assembly	3	10%	.
GO:1904645	biological process	response to amyloid-beta	3	10%	TLR4,TLR6
GO:0042088	biological process	T-helper 1 type immune response	4	10%	TLR4,TLR6
GO:0042571	cellular component	immunoglobulin complex, circulating	7	9%	.
GO:0019814	cellular component	immunoglobulin complex	8	9%	.
GO:0000186	biological process	activation of MAPKK activity	5	9%	LRRK2,MAP3K2,MAP3K8,RAF1
GO:0050704	biological process	regulation of interleukin-1 secretion	4	9%	CCL3,TLR4
GO:1902373	biological process	negative regulation of mRNA catabolic process	4	9%	FMR1,ZFP36
GO:0050702	biological process	interleukin-1 beta secretion	4	9%	CCL3,TLR4,TLR6
GO:0002762	biological process	negative regulation of myeloid leukocyte differentiation	4	9%	CCL3,TLR4,TRIB1
Zvýšená exprese					
GO:0090023	biological process	positive regulation of neutrophil chemotaxis	4	13%	CCR7,IL23A
GO:0045058	biological process	T cell selection	6	13%	BCL11B,CCR7,CD3D,CD3E,IL23A,LY9
GO:0140104	molecular function	molecular carrier activity	5	12%	HBD,HBG1,HBG2,ZNF593
GO:0031295	biological process	T cell costimulation	9	12%	CCR7,CD3D,CD3E,GRAP2,LCK,TMIGD2,TRAC,TRBC1
GO:1902624	biological process	positive regulation of neutrophil migration	4	12%	CCR7,IL23A
GO:0071624	biological process	positive regulation of granulocyte chemotaxis	4	12%	CCR7,IL23A
GO:0042108	biological process	positive regulation of cytokine biosynthetic process	7	12%	CD3E,LTB
GO:0032735	biological process	positive regulation of interleukin-12 production	4	12%	CCR7,IL23A,LTB
GO:0000289	biological process	nuclear-transcribed mRNA poly(A) tail shortening	4	11%	.
GO:0090022	biological process	regulation of neutrophil chemotaxis	4	11%	CCR7,IL23A
GO:0045429	biological process	positive regulation of nitric oxide biosynthetic process	4	11%	.
GO:1904407	biological process	positive regulation of nitric oxide metabolic process	4	11%	.
GO:0031294	biological process	lymphocyte costimulation	9	11%	CCR7,CD3D,CD3E,GRAP2,LCK,TMIGD2,TRAC,TRBC1
GO:0071276	biological process	cellular response to cadmium ion	4	11%	MTIX,MT2A
GO:0004712	molecular function	protein serine/threonine/tyrosine kinase activity	4	10%	.
GO:0050716	biological process	positive regulation of interleukin-1 secretion	3	10%	IGHD
GO:0050701	biological process	interleukin-1 secretion	5	10%	CCR7,IGHD
GO:1902622	biological process	regulation of neutrophil migration	4	10%	CCR7,IL23A
GO:0051491	biological process	positive regulation of filopodium assembly	3	10%	CCR7
GO:0051385	biological process	response to mineralocorticoid	3	10%	PARP1
GO:1904837	biological process	beta-catenin-TCF complex assembly	3	10%	AXIN2,MYC,TLE2
GO:1904645	biological process	response to amyloid-beta	3	10%	PARP1
GO:0042088	biological process	T-helper 1 type immune response	4	10%	IL23A,IL27RA
GO:0042571	cellular component	immunoglobulin complex, circulating	7	9%	IGHD,IGHM,IGHV1-2,IGHV1-69D,IGHV4-39,TRBC1,TRBC2
GO:0019814	cellular component	immunoglobulin complex	8	9%	CD79A,IGHD,IGHM,IGHV1-2,IGHV1-69D,IGHV4-39,TRBC1,TRBC2
GO:0000186	biological process	activation of MAPKK activity	5	9%	EGF
GO:0050704	biological process	regulation of interleukin-1 secretion	4	9%	CCR7,IGHD
GO:1902373	biological process	negative regulation of mRNA catabolic process	4	9%	AXIN2,E2F1
GO:0050702	biological process	interleukin-1 beta secretion	4	9%	CCR7
GO:0002762	biological process	negative regulation of myeloid leukocyte differentiation	4	9%	MYC

Tabulka 4.4: Třicet nejvíce ovlivněných GO kategorií pro dysregulované geny v obou skupinách (RPS7 a RPS19)

Nejvíce ovlivněné termíny genové ontologie (GO) společných pro *RPS7* a *RPS19*. Procento přiřazených genů v ontologické kategorii se změnou exprese na hladině významnosti p-hodnoty ≤ 0.05 z celkového počtu genů pro ontologickou kategorii. Celková (#) hodnota udává počet přiřazených genů se změnou exprese.



Obrázek 4.5: Klastrovací heatmapa genů v obou porovnávaných souborech (RPS19 a RPS7)

Klastrovací analýza exprese všech společných genů pro skupinu *RPS7* a *RPS19* se změněnou expresí ($\log_2$) oproti zdravým kontrolám na hladině významnosti $p\text{-hodnota} \leq 0.05$. Maximum reportovaného rozsahu změny exprese bylo upraveno do intervalu $(-3, 3)$.

4.4 Diskuze

V této kapitole prezentujeme první případ nesynonymní varianty v genu *RPS7* jako molekulární příčinu DBA. Do současnosti bylo publikováno 5 případů mutací v *RPS7* způsobujících lidského onemocnění DBA (Gazda *et al.*, 2008; Gerrard *et al.*, 2013; Smetanina *et al.*, 2015; Ichimura *et al.*, 2017). Ve všech publikovaných případech byly mutace v *RPS7* lokalizovány v sestřihových místech a klinické projevy DBA byly mírnější než v případě našeho probanda. U myšího modelu heterozygotní mutantní formy *RPS7* p.V156G se vyskytovaly analogické fenotypové projevy jako u člověka, jako např. menší vzrůst a abnormality skeletu, ale i další odlišnosti - malformace očí a bílá pigmentace srsti na ventrální straně těla (Watkins-Chow *et al.*, 2013). Identifikovali jsme nesynonymní mutaci způsobující aminokyselinovou záměnu *RPS7* p.V134F u pacientky s DBA, která byla při následném skríninku detekována i u dvou asymptomatických členů rodiny (matka, sestra). Mutace byla použitými *in silico* metodami predikována jako poškozující stabilitu proteinu. Oba asymptomatictí členové rodiny vykazovali pouze mírně zvýšené hodnoty eADA (erytrocytární adenosindeamináza) jako jednoho z diagnostických markerů DBA (Blume *et al.*, 1977). Stanovení hodnot eADA probíhalo na externím pracovišti a není součástí této práce.

Kromě identifikace mutace jsme provedli analýzu transkriptomu z periferní krve ve srovnání se zdravým kontrolním souborem a DBA pacienty z Českého Národního DBA registru s potvrzenou mutací v *RPS19* (2 pacienti *RPS19* p.R56Q a 2 pacienti s mutací v sestřihovém místě *RPS19*). Po zhodnocení transkriptomu bylo pozorováno výrazné klastrování všech členů rodiny spolu s *RPS19* skupinou odlišné od zdravých kontrol. V minulosti již byly popsány značně heterogenní fenotypové projevy DBA od asymptomatických přenašečů až k závažným průběhům (Farrar & Dahl, 2011).

Mezi geny s nejvíce sníženou expresí v obou skupinách *RPS7* a *RPS19* se vyskytovaly mitogenní a diferenciacní faktory *EGR1* a *EGR2*. *EGR1* může být transaktivován mnoha cytokiny a hormony včetně inzulínu přes *ERK1/2* dráhu (Keeton *et al.*, 2003). Nedávné studie naznačují, že deficece *RP* genů vedou k dysregulaci inzulínové dráhy připomínající inzulínovou rezistenci (Heijnen *et al.*, 2014). V souladu se sníženou expresí *EGR1* jsme v obou skupinách DBA zaznamenali sníženou expresi genu *TNF*, který je primárně produkován monocyty a podílí se na diferenciacních buněčných procesech a imunitní odpovědi, jako je například obrana proti infekci a tvorbě nádorů, a aktivátorem jeho exprese je právě *EGR1* (Collart *et al.*, 1990). Dalším z významných tumor supresorových genů, který byl ovšem identifikován se sníženou expresí pouze ve skupině *RPS19*, je *PTEN*, jenž je také expresně regulován *EGR1* (Virolle *et al.*, 2001). Navíc haploinsuficience *EGR1* vede k rozvoji myeloidních onemocnění u myší, jako je akutní myeloidní leukémie (AML) i myelodysplastický syndrom (MDS), a společně s lokalizací na 5q31 je zvažován jako jeden

z kandidátních tumor supresorových genů pro syndromy del(5q) (Joslin *et al.*, 2007). Bylo popsáno, že DBA pacienti mají větší predispozice právě pro rozvoj malignit typu AML/MDS (Vlachos *et al.*, 2008).

Ačkoliv nejčastěji zmiňovanou změnou signálních drah vedoucích k zástavě buněčného cyklu a apoptóze v případě ribozomopatií je TP53 hyperaktivace způsobená blokací aktivity jeho antagonisty MDM2 volnými ribozomálními proteiny, byly popsány i mechanismy způsobující ribozomální stres nezávisle na aktivaci TP53 (Donati *et al.*, 2012). Ve všech testovaných DBA skupinách pozorujeme zvýšenou expresi *MYC* genu, která je u DBA zaznamenána v souvislosti s *RPL11* deficiencí na „zebrafish“ modelu (Chakraborty *et al.*, 2017). Autoři studie prokázali, že zvýšená exprese *MYC* nesouvisí s aktivací TP53, ale byla vyvolána samotnou *RPL11* deficiencí. Bylo zjištěno, že za specifických podmínek nepřítomnosti cyklin-dependentních kináz může zvýšená exprese *MYC* vést k apoptóze (Van Riggelen & Felsher, 2010). Dalším regulátorem buněčného cyklu s výrazně zvýšenou expresí v obou DBA skupinách byl *E2F1*. Ektopická exprese *E2F1* indukuje stabilizaci a zvýšenou aktivitu TP53 inhibicí E3 ubiquitin ligázové aktivity MDM2 (Ginsberg, 2002). Navíc bylo ukázáno, že *E2F1* indukce apoptózy probíhá nezávisle na ARF (Russell *et al.*, 2002). V hodnocení transkriptomových dat analýzou v programu Reactome byly identifikovány další geny, které se podílejí na regulaci buněčného cyklu. V případě RPS7 skupiny byly dysregulovány geny *BTG2*, *PLAGL1*, *MYBL2*, *CASP8*, *CFLAR*, *PMAIP1* a v případě RPS19 se jednalo o geny *BTG2*, *CDKN1A*, *PRMT1*, *PLAGL1*, *CARM1*, *SFN*. V obou skupinách byla nalezena signifikantní dysregulace WNT signální dráhy, která se účastní základních vývojových procesů během embryogeneze. Dysregulace WNT dráhy byla sledována i v případě transkriptomové studie na *RPS24* „zebrafish“ modelu (Song *et al.*, 2014).

Vzhledem k volbě materiálu, ze kterého byly vytvořeny RNASeq knihovny pro analýzu transkriptomu (mononukleární buňky z periferní krve) mezi nejvíce ovlivněné procesy patří dráhy spjaté s imunitním systémem. Obecně vyšší hladiny exprese v DBA skupinách vykazovaly geny rodiny imunoglobulinů a sníženou hladinu exprese genové rodiny chemokinů, chemokininových ligandů a TLR receptorů. Opačná zjištění byla publikována v analýze imunitního systému DBA modelu „zebrafish“, kde byla popsána zvýšená exprese TLR receptorů (TLR3, TLR9) s indukcí interferonové signalizace a systému komplementu (Danilova *et al.*, 2018). Velká pozornost byla v problematice DBA a ribozomálního stresu věnována především mechanismům poruch hematopoézy v souvislosti s TP53 aktivací, zatímco rozsáhlejší soubor prací věnujících se imunitnímu systému u ribozomálního stresu chybí.

Navzdory použití mononukleárních buněk pro transkriptomovou studii byla

pozorována výrazná změna exprese genů specifických pro erytroidní progenitory oproti kontrolnímu souboru. Výrazně zvýšená byla u DBA exprese fetálních hemoglobinových genů *HBG1*, *HBG2* a delta podjednotky *HBD*. Kromě obecné funkce hemoglobinu jako přenašeče O_2 a CO_2 byla u hemoglobinu sledována ochrana před oxidativním stresem. Antioxidační funkce hemoglobinu se projevuje snížením produkce intracelulárních peroxidových radikálů a zvýšená exprese hemoglobinu zlepšuje životaschopnost mezangiálních buněk potkana vystavených oxidativnímu stresu (Nishi *et al.*, 2008). Kromě toho může lidský hemoglobin detoxikovat vysoce oxidující radikály železa (Goldstein & Samuni, 2005). Přetížení železem je jednou ze závažných komplikací DBA u polytransfundovaných pacientů (Roggero *et al.*, 2009). Je tedy otázkou, zda je zvýšená produkce hemoglobinu v leukocytech je pouze vyvolána hypoxií v důsledku anémie, nebo může sloužit jako částečný protektivní mechanismus proti oxidativnímu stresu. Nicméně zvýšená exprese hypoxií inducibilního genu *HIF1A* nebyla sledována v žádné z analyzovaných skupin DBA a tuto hypotézu považujeme za nepravděpodobnou. Z pohledu problematiky přetížení železem jsme dále sledovali u všech DBA skupin sníženou expresi transportéru hemu FLVCR1, který se kromě ochrany před toxicitou hemu podílí na správném průběhu erythropoézy (Chiabrando *et al.*, 2012).

Mezi další ovlivněné procesy se signifikantním hodnocením byly zaznamenány geny rRNA procesování a translace (*EIF1AY*, *EIF3K*, *EIF3I*, *NOP14*, *RPL3*, *RPL36A*, *SNU13*, *RPS4Y1*, *BYSL*, *FBL*, *RPS16*, *RPS19*, *PES1*, *RPL13*, *RPLP2*, *RPL38*, *RPL28*, *RPL19*). Uvedená zjištění jsou v korelaci se získanými výsledky na buněčném RPS7 p.V134F modelu, kde byla zjištěna snížená úroveň translace, abnormální morfologie nukleolů, abnormální procesování rRNA a hyperaktivace TP53 (výsledky z biologického modelu nejsou součástí předkládané práce a jsou detailně rozvedeny v příloženém rukopisu publikace).

4.5 Závěr

Ve druhé kapitole popisujeme první nesynonymní variantu genu malé ribozomální podjednotky RPS7 p.V134F jako molekulární příčinu s heterogenní fenotypovou expresí v rámci jedné rodiny. Kromě funkčního popisu a analýzy funkční predikce mutace jsme provedli transkriptomovou studii mononukleárních buněk periferní krve u všech členů rodiny ve srovnání se zdravou kontrolní skupinou a se skupinou DBA s mutací formou *RPS19*. Transkriptomový profil všech členů rodiny (včetně asymptomatických přenašečů) v RPS7 souboru překvapivě silně koreloval s RPS19 souborem a vykazoval společné dysregulované skupiny genů. Byla sledována snížená exprese *EGR1* a *EGR2*, které jsou mimo jiné součástí inzulinové dráhy, která je v současnosti studována ve spojitosti s ribozomopatiemi.

Naopak zvýšená exprese byla sledována u genů účastnících se procesů apoptózy

a aktivace TP53 (*MYC*, *E2F1*), což je v souladu s teoriemi příčin ribozomálního stresu a jím způsobené patologie. Mezi dalšími ovlivněnými buněčnými procesy byly identifikovány dysregulace signálních drah rRNA procesování a proteinové translace, což je opět v souladu se současným stavem poznání molekulární patologie u DBA.

Kromě již popsaných molekulárních souvislostí s DBA jsme detekovali výrazně zvýšenou expresi hemoglobinových genů. Důvodem takto zvýšené exprese může být jednak hypoxie způsobená anemickým stavem DBA pacientů, nebo tato může hrát protektivní roli při zvýšeném oxidativním stresu, což již bylo popsáno u neerytroidních buněk. Pravděpodobně i z důvodu volby materiálu pro transkriptomovou komparativní analýzu tvořily geny imunitního systému největší skupinu genů s dysregulovanou expresí ve srovnání se zdravými kontrolami v obou případech DBA (*RPS7* i *RPS19*). Ačkoliv byla u DBA modelu sledována dysregulace genů imunitního systému, v našem případě byly u popsaných genů sledovány opačné trendy exprese než v publikované studii. Pro potvrzení hypotéz biologických funkcí genové exprese u námi zjištěných markerů by byla potřeba rozsáhlejší analýza těchto markerů na biologických modelech a také jejich confirmaci na proteinové úrovni.

5 Transkriptomová analýza transgenního ječmene se zvýšenou odolností vůči stresu suchem

5.1 Úvod

Cytokinininy jsou rostlinné hormony, které společně s auxiny ovlivňují morfologii rostlin a fyziologické procesy jako je senescence a mobilizace živin (Zalabák *et al.*, 2013). Pozdější studie na modelu *Arabidopsis thaliana* naznačují také význam cytokininů v regulacích odpovědí na environmentální stres (Ha *et al.*, 2012). Cytokinin deficientní rostliny *Arabidopsis* vykazují stresu odolný fenotyp asociovaný se zvýšenou integritou buněčných membrán a zvýšenou citlivostí vůči stresovému hormonu, kyselině abscisové (Nishiyama *et al.*, 2011). Role kyseliny abscisové spočívá především v indukci zavírání průduchů, která předchází ztrátě vody v době jejího nedostatku. Rostliny deficientní pro cytokininové receptory vykazují zvýšenou odolnost vůči stresu suchem a salinitnímu stresu (Tran *et al.*, 2007; Nishiyama *et al.*, 2013). Kromě *Arabidopsis* byl stejný fenomén sledován i u rýže (*Oryza sativa*) (Sun *et al.*, 2014). Aktivované cytokininové receptory naopak negativně odpovídají na osmotický stres, a tím potvrzují roli sníženého vnímání cytokininového signálu na toleranci vůči suchu.

Na druhou stranu zvýšená tolerance vůči suchu vyvolaná akumulací cytokininů byla pozorována u několika rostlinných druhů (Rivero *et al.*, 2007; Zhang *et al.*, 2009; Ghanem *et al.*, 2011; Merewitz *et al.*, 2011; Peleg *et al.*, 2011; Kuppu *et al.*, 2013). Byly připraveny transgenní rostliny biosyntetizující cytokinininy pod kontrolou indukovatelného promotoru stresem. Po vystavení suchu transgenní rostliny udržovaly oproti kontrolním vysokou fotosyntetickou aktivitu díky prolongaci senescence a příjmu dusičnanů z půdy (Reguera *et al.*, 2013). Souhrnně, stresem indukovaná syntéza cytokininů má přímý pozitivní vliv na asimilaci uhlíku a příjem dusíku během období sucha.

Během vystavení abiotickému stresu mohou cytokinininy působit v součinnosti s dalšími rostlinnými hormony. Role auxinů v toleranci sucha byla prokázána zvýšenou aktivitou auxin-deaktivačních enzymů, která snižovala auxinové maxima v listech, což mělo za následek zvýšenou expresi LEA (late-embryogenesis abundant) proteinů zodpovědných za přechod mezi elongačním růstem rostliny a adaptací na stres (Zhang *et al.*, 2009). Auxin je schopen vyvolávat expresi genů participujících v biosyntéze dalšího stresového hormonu etylénu a opačně etylén indukuje místně specifickou biosyntézu auxinů vedoucí k redukci elongace kořenových buněk (Tsuchisaka & Theologis, 2004; Ruzicka *et al.*, 2007). Jelikož je známo, že cytokinininy mohou ovlivňovat produkci jak auxinů, tak etylénu, koordinovaná regulace hormonálních enzymových drah může hrát klíčovou roli v adaptacích rostlin na abiotický stres (Peleg & Blumwald, 2011). Rostliny s indukovanou produkcí cytokininů

vykazují pozitivní regulaci genů zapojených do biosyntézy a signalizace dalších hormonů ze skupiny brassinosteroidů (Rivero *et al.*, 2010; Peleg & Blumwald, 2011). Brassinosteroidy působí synergicky s dalšími fytohormony gibberelliny díky sdílení některých komponent signální kaskády (Wang *et al.*, 2009). Transgenní rostliny *Arabidopsis* konstitutivně nadprodukující geny gibberellinové odpovědi vykazují zvýšenou toleranci vůči různým stresovým faktorům, která je doprovázená vyšší produkcí kyseliny salicylové, dalšího stresového hormonu zapojeného do mechanismu tolerance vůči stresu (Alonso-Ramirez *et al.*, 2009).

Rostliny vystavené stresu suchem vykazují změny v obsahu cytokininů oproti rostlinám v kontrolních podmínkách. Například, rostliny kukuřice (*Zea mays*) vystavené suchu měly ve svých listech méně cytokininů a gibberellinů a jejich obsah kontinuálně klesal po dobu trvání stresu (Wang *et al.*, 2008). Podobná studie ukázala, že pokles cytokininů v listech kukuřice je dán hlavně zvýšenou aktivitou deaktivačního enzymu. Nicméně během aklimatizace pozměněný metabolismus vedl k mírné akumulaci aktivních cytokininových forem (Vyroubalova *et al.*, 2009). Zvýšená akumulace všech cytokininových forem byla pozorována také v rostlinách tabáku (*Nicotiana tabacum*) vystavených silnému stresu suchem (Havlova *et al.*, 2008). Akumulace aktivních cytokininů tak může mezi jinými procesy znamenat mechanismus, kterým se rostlina adaptuje na stresové podmínky inhibicí růstu.

Vysoká fotosyntetická aktivita je důležitým předpokladem pro udržení výnosů zemědělských plodin za nepříznivých environmentálních podmínek. Ačkoliv zvýšený obsah cytokininů způsobený senescenčně regulovanou expresí isopentenyl transferázy je účinným způsobem jak prodloužit fotosyntetickou aktivitu listu (Gan & Amasino, 1995), tento přístup u pšenice (*Triticum aestivum*) nevedl ke zvýšení výnosů za optimálních ani stresových podmínek (Sýkorová *et al.*, 2008). Inhibice růstu nadzemní části a intenzivní růst kořenového systému je považován za výhodný pro udržení stability plodiny za nepříznivých podmínek a zvyšuje stresovou toleranci rostliny na suchu (Sharp & Lenoble, 2002; Albacete *et al.*, 2014). V souladu s předchozím zjištěním, rostliny s redukováným poměrem nadzemní versus kořenové části jako následek deficiencie cytokininů, vykazovaly vyšší toleranci vůči stresu (Werner *et al.*, 2010; Macková *et al.*, 2013). Obecně, snížení či zvýšení hladin endogenních cytokininů může mít synergetický pozitivní efekt na zvýšenou toleranci k deficitu vody přes změnu morfologie kořenového systému, jehož architektura je lépe přizpůsobena deprivaci vody a aktivací fotosyntetických procesů a remobilizace živin.

V této práci byly studovány transgenní rostliny ječmene (*Hordeum vulgare*) exprimující gen pro cytokinin dehydrogenasu, což je cytokinin deaktivační enzym, pod kontrolou slabého kořenově specifického promotoru kukuřičné β -glukosidázy. Transgenní rostliny mají vyšší toleranci vůči suchu především díky změněné anatomii kořenového

systému a silnější lignifikaci vodivých pletiv kořene. Byla provedena detailní srovnávací transkriptomová analýza kořenů a nadzemní části během optimálního pěstování rostlin a během revitalizace transgenních rostlin po vystavení stresu suchem mezi transgenními a kontrolními rostlinami.

5.2 Materiál a metody

5.2.1 Rostlinný materiál, podmínky kultivace a aplikace stresu

Transgenní rostliny byly připraveny v rámci předchozí práce (Pospíšilová *et al.*, 2016). Kontrolní soubor netransgenních rostlin jarního kultivaru ječmene Golden Promise byl pěstován ve fotokomoře s fotoperiodou 15°C/16 hodin světla a 12°C/8 hodin tmy. Zdrojem světla byla kombinace wolframo-rtuťnaté výbojky a sodíkové lampy poskytující intenzitu ozáření 300 $\mu\text{mol m}^{-2} \text{s}^{-1}$. Rostliny byly pěstovány ve směsi rašelinového substrátu (Rašelina Soběslav, Česká republika) smíchaného v poměru 1:1 s hnědozemí původem z Olomouckého regionu a provzdušněné jednou objemovou třetinou perlitem (Perlit s.r.o., Česká republika).

Pro transkriptomovou analýzu kořenového systému byly rostliny pěstovány v hydroponických vanách s modifikovaným Hoaglandovým roztokem (Vlamis & Williams, 1962) ve stejném prostředí a podmínkách jak je uvedeno výše. Experiment byl prováděn v následujícím uspořádání: 2/3 vany bylo zaplněno dvěma transgenními genotypy a 1/3 rostlinami kontrolního souboru divokého typu. Celkově bylo pro každý genotyp pěstováno 27 rostlin ve třech vanách. Tři směsné vzorky pletiv rostlin byly součástí vždy jednoho biologického opakování. Stres suchem byl aplikován na 4 týdny staré rostliny odebráním nutričního roztoku a po 24 hodinách byl odebrán vzorek kořenového systému. Rostliny byly vráceny do nutričního roztoku a vzorek kořenového systému byl odebrán pro následnou analýzu po dalších 24 hodinách a dále po 14 dnech revitalizace.

Pro účely transkriptomové studie nadzemní části byly rostliny pěstovány v nádobách (30×20×5cm) se substrátem. Rostliny byly zalévány denně až do stáří 4 týdnů, kdy byl aplikován stres suchem omezením záливky na čtyři následující dny. Poté byla záливka opět obnovena. Vzorky celé nadzemní části byly vždy sbírány před aplikací stresu, 12 hodin a 14 dnů po obnovení záливky, kdy se rostliny dostaly do fáze revitalizace.

5.2.3 RNASeq

Izolace celkové RNA byla provedena pomocí RNAqueous kitu (Life Technologies, USA). Eliminace gDNA byla provedena TURBO DNA-free kitem (Life Technologies, USA)

s následnou RNA purifikací kitem Agencourt RNA-CLEAN XP (Beckman Coulter, USA). Pro přípravu knihoven bylo použito 2.5 µg celkové RNA. cDNA knihovny se specifickou orientací byly připraveny pomocí kitu TruSeq Stranded mRNA (Illumina, USA). Koncentrace knihovny byla stanovena kitem Kapa Library Quantification (Kapa Biosystems, USA) a knihovny byly sekvenovány na HiSeq 2500 (Illumina, USA). Vždy 2 knihovny směsných vzorků (3 rostliny) byly připraveny jako biologický replikát pro hodnocenou skupinu.

FASTQ byly alignovány na referenční genom *Hordeum vulgare* v.25 (Ensembl) programem TopHat2 v.2.0.12 (Kim *et al.*, 2013) ve standardním nastavení bez úpravy parametrů. Základní kvantifikace čtení byla zpracována v HTSeq v.0.6.0 (Anders *et al.*, 2015). Analýza diferenciální genové exprese a statistické vyhodnocení bylo provedeno programem DESeq2 (Love *et al.*, 2014).

Anotace genové ontologie (GO) referenčního genomu byla vylepšena programem Blast2GO v.3.0 (Conesa *et al.*, 2005), nt databáze (b2g_Jan15) a ncbi-blast+ v.2.2.28 programem (Camacho *et al.*, 2009) a dále databázemi UniProtKB (<http://www.uniprot.org/>, 2015_02) a PGSB (<http://pgsb.helmholtz-muenchen.de/plant/>; 2014_07_31) (Kokas *et al.*, 2016). Tato doplňující anotace GO termínů pomohla zvýšit počet anotovaných genů na 17,885 z celkového počtu 26,074 predikovaných genů v ječmeni. Analýza ovlivněných biologických procesů byla provedena pomocí vlastních skriptů s rekonstrukcí ancestrálních procesů pro GO. Analýza GO obohacených procesů byla prováděna vždy v kategoriích 6. úrovně.

5.3 Výsledky

Pro detailní transkriptomovou analýzu byly vybrány dva genotypy transgenních rostlin ječmene, které nejsou ovlivněny ve výnosových parametrech. První s rekombinantním proteinem s vakuolární signální sekvencí (*vAtCKX1*). U druhého genotypu byla odstraněna vakuolární signální sekvence s výslednou lokalizací v cytosolu (*cAtCKX1*). Celkově bylo za účelem srovnávací transkriptomové studie analyzováno 28 „single-end“ sekvenačních knihoven, které byly sekvenovány na sekvenátoru HiSeq 2500 (Illumina, USA). Základní charakteristika alignmentu je uvedena v Tabulce 5.1. Analýza integrity směsných knihoven jako nezávislých měření neprokázala odchylky ve směsných knihovnách (Kokas *et al.*, 2016).

Knihovna	Počet čtení	Počet alignovaných čtení	V (%)
Nadzemní část			
WT před aplikací stresu 1	84,517,106	78,995,347	93.50%
WT před aplikací stresu 2	63,270,045	59,274,446	93.70%
vAtCKX1 před aplikací stresu 1	75,279,326	70,471,879	93.60%
vAtCKX1 před aplikací stresu 2	71,354,832	66,706,602	93.50%
WT stresovány 1	77,801,469	70,486,367	90.60%
WT stresovány 2	46,104,470	43,271,563	93.90%
vAtCKX1 stresovány 1	71,985,223	68,031,123	94.50%
vAtCKX1 stresovány 2	73,507,253	69,533,881	94.60%
WT 12 h revitalizace 1	51,142,352	48,101,143	94.10%
WT 12 h revitalizace 2	65,959,107	62,277,842	94.40%
vAtCKX1 12 h revitalizace 1	69,436,337	64,832,308	93.40%
vAtCKX1 12 h revitalizace 2	71,082,657	66,986,453	94.20%
WT 14 dnů revitalizace 1	66,446,892	62,470,573	94.00%
WT 14 dnů revitalizace 2	68,521,784	64,326,131	93.90%
vAtCKX1 14 dnů revitalizace 1	59,187,683	54,355,201	91.80%
vAtCKX1 14 dnů revitalizace 2	71,287,568	66,620,894	93.50%
Kořenová část			
WT stresovány 1	45,007,428	37,581,752	83.50%
WT stresovány 2	200,564,800	168,454,522	84.00%
vAtCKX1 stresovány 1	152,220,463	121,895,816	80.10%
vAtCKX1 stresovány 2	78,643,428	65,854,656	83.70%
cAtCKX1 stresovány 1	75,960,243	64,038,286	84.30%
cAtCKX1 stresovány 2	63,499,681	52,087,213	82.00%
WT 14 dnů revitalizace 1	138,937,849	104,864,042	75.50%
WT 14 dnů revitalizace 2	90,088,238	65,540,634	72.80%
vAtCKX1 14 dnů revitalizace 1	64,437,018	48,016,522	74.50%
vAtCKX1 14 dnů revitalizace 2	104,760,752	77,622,260	74.10%
cAtCKX1 14 dnů revitalizace 1	84,791,919	66,888,893	78.90%
cAtCKX1 14 dnů revitalizace 2	124,010,156	99,545,337	80.30%

Tabulka 5.1: Počet získaných čtení pro jednotlivé knihovny a úspěšnost alignmentu

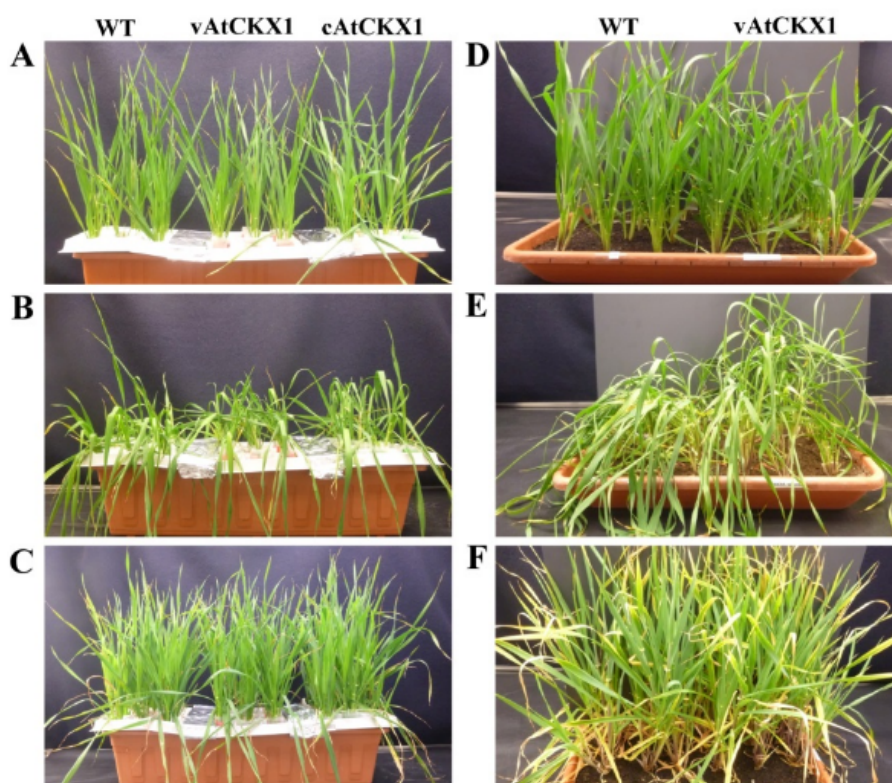
Souhrn získaných čtení pro testované skupiny a úspěšnost alignmentu.

5.3.1 Diferenciální exprese genů za optimálních podmínek v nadzemní části rostlin

Slabá exprese genu *AtCKX1* pod kontrolou promotoru pro β -glukosidázu má pozitivní vliv na vývoj kořenového systému zatímco nadzemní část transgenních rostlin není zásadně ovlivněna (Pospíšilová *et al.*, 2016). Výška je mírně redukována u *vAtCKX1* rostlin (Obrázek 5.1), zatímco rostliny *cAtCKX1* se viditelně neodlišují do 4 týdne vývoje od rostlin divokého typu. Transkriptomová srovnávací studie nadzemní části *vAtCKX1* rostlin v šestém týdnu vývoje ukázala statisticky významné vychýlení přibližně u 400 genů. Daleko více vychýlených genů (cca 2,400) bylo pozorováno v kořenech těchto rostlin pěstovaných v hydroponickém systému. Výsledky jsou detailně shrnuty v předchozí studii (Pospíšilová *et al.*, 2016).

Transkriptomová studie provedená v rámci této práce na *vAtCKX1* rostlinách ve stejném stádiu vývoje pěstovaných v substrátu byla provedena za účelem potvrzení genů,

které jsou vychýleny změnou statusu endogenních cytokininů a nejsou ovlivněny podmínkami pěstování (hydroponie/substrát) před aplikací stresu. Na rozdíl od rostlin pěstovaných v hydroponii bylo u rostlin pěstovaných v substrátu mezi kontrolou a transgenní rostlinou vychýleno zhruba čtyřikrát více genů v nadzemní části. Z celkového počtu 26,067 anotovaných genů bylo 988 identifikováno se statisticky významně sníženou mírou exprese a 609 se zvýšenou expresí (adjustovaná p-hodnota ≤ 0.05) (Vojta *et al.*, 2016). GO termíny na 6. úrovni ontologických kategorií statisticky dysregulovaných genů ve srovnání podmínek (hydroponie/substrát) byly porovnány a patnáct nejvíce ovlivněných procesů je shrnuto v Tabulce 5.2.



Obrázek 5.1: Morfologie nadzemních částí rostlin ječmene v substrátu/hydroponii v testovaných podmínkách

Transgenní rostliny a rostliny divokého typu (WT) ječmene pěstované v hydroponii (vlevo) nebo substrátu (vpravo). (A, D) optimálně zalévané rostliny 4 týdny po vyklíčení, (B, E) rostliny vystavené silnému stresu suchem, (C, F) revitalizované rostliny 2 týdny po období silného sucha.

GO číslo	Kategorie	GO termín	Celkový #	% ovlivněných genů		Identifikátory genů MLOC
				Hydrop.	Substrát	
SE ZVÝŠENOU EXPRESÍ						
GO:0004664	MF	prephenate dehydratase activity	4	50.0%	25.0%	23316, 65725, 56414
GO:0008131	MF	primary amine oxidase activity	6	16.7%	50.0%	70980, 4986, 17390
GO:0016165	MF	linoleate 13S-lipoxygenase activity	16	31.3%	18.8%	64972, 54031, 55029, 71275, 51884, 69572
GO:0005544	MF	calcium-dependent phospholipid binding	10	20.0%	20.0%	54932, 40592, 55134, 15770
GO:0004834	MF	tryptophan synthase activity	5	20.0%	20.0%	59863, 61188
GO:0009963	BP	positive regulation of flavonoid biosynthetic process	9	11.1%	22.2%	81070, 54366, 19988
GO:0004034	MF	aldose 1-epimerase activity	7	14.3%	14.3%	5638
GO:0005337	MF	nucleoside transmembrane transporter activity	7	14.3%	14.3%	55464
GO:0009718	BP	anthocyanin-containing compound biosynthetic process	11	18.2%	9.1%	61512, 65788, 64248
GO:0009407	BP	toxin catabolic process	27	7.4%	18.5%	17760, 73593, 68101, 57709, 72489
GO:0047262	MF	polygalacturonate 4-α-galacturonosyltransferase activity	8	12.5%	12.5%	11661, 57229
GO:0031418	MF	L-ascorbic acid binding	13	15.4%	7.7%	77814, 64248
GO:0004806	MF	triglyceride lipase activity	41	9.8%	12.2%	17298, 18031, 80878, 80586, 58940
GO:0006569	BP	tryptophan catabolic process	14	14.3%	7.1%	12847, 57323, 69262
GO:0015996	BP	chlorophyll catabolic process	33	3.3%	15.2%	80455, 34851, 55009, 21175, 64277
SE SNÍŽENOU EXPRESÍ						
GO:0008937	MF	ferredoxin-NAD(P) reductase activity	4	25.0%	50.0%	7761, 53537, 40355
GO:0010027	BP	thylakoid membrane organization	106	2.8%	61.3%	58382, Nezobrazeno
GO:0051667	BP	establishment of plastid localization	63	4.8%	55.6%	Nezobrazeno
GO:0045548	MF	phenylalanine ammonia lyase activity	7	28.6%	28.6%	79728, 62322, 477, 4684
GO:0009658	BP	chloroplast organization	137	2.2%	54.0%	Nezobrazeno
GO:0019682	BP	glyceraldehyde-3-phosphate metabolic process	202	1.5%	51.5%	Nezobrazeno
GO:0006720	BP	isoprenoid metabolic process	237	1.3%	46.0%	Nezobrazeno
GO:0010310	BP	regulation of hydrogen peroxide metabolic process	15	6.7%	26.7%	33774, 1518, 15501, 65632, 1340
GO:0030855	BP	epithelial cell differentiation	7	14.3%	14.3%	38181, 54366
GO:0016054	BP	organic acid catabolic process	153	3.9%	14.4%	Nezobrazeno
GO:0008544	BP	epidermis development	9	11.1%	11.1%	38181, 54366
GO:0009699	BP	phenylpropanoid biosynthetic process	22	9.1%	9.1%	4684, 477, 57736, 79728
GO:0070726	BP	cell wall assembly	11	9.1%	9.1%	52864, 67760
GO:0032870	BP	cellular response to hormone stimulus	205	8.8%	5.7%	Nezobrazeno
GO:0007166	BP	cell surface receptor signaling pathway	30	3.3%	10.0%	63541, 17680, 44275, 72162

Tabulka 5.1: Nejvíce ovlivněné kategorie GO v různých podmínkách kultivace substrát/hydroponie

Nejvíce ovlivněné termíny genové ontologie (GO) v nadzemní části *vAtCKX1* rostlin pěstovaných hydroponicky nebo v substrátu ve srovnání s rostlinami divokého typu. Procento z celkového počtu genů v genomu těch genů, které mají rozdílnou expresi na hladině významnosti p-hodnoty ≤ 0.05 pro GO termíny na 6. úrovni. MF, molekulární funkce; BP, biologický proces. Celková (#) hodnota udává počet všech přiřazených genů v genomu.

5.3.2 Diferenciální exprese genů za stresových podmínek a během revitalizace v nadzemní části transgenních rostlin

Kromě vzorků nadzemní části rostlin pěstovaných za optimálních podmínek byly další sekvenační knihovny připraveny z materiálu *vAtCKX1* a rostlin divokého typu vystaveného stresu s následnou revitalizací. První série knihoven byla připravená ze vzorků nadzemní části vystavené čtyřdennímu období sucha (Obrázek 5.1(E)), druhá pak z rostlinného materiálu 12 hodin po obnovení závlivky a třetí po období dvou týdenní revitalizace (Obrázek 5.1(F)). Mezi těmito skupinami bylo pouze pět genů statisticky významně ovlivněno (adjustovaná p-hodnota ≤ 0.05) mezi transgenním a kontrolním materiálem za podmínek sucha, což ukazuje na velmi silné ovlivnění transkriptomu obou genotypů suchem, které je daleko výraznější než rozdíl mezi oběma genotypy (Kokas *et al.*, 2016).

Deset genů mělo zvýšenou a devět sníženou expresi v transgenních rostlinách oproti divokému typu dvanáct hodin po obnovení závlivky (Vojta *et al.*, 2016). Navíc pět z těchto devatenácti genů nebylo ovlivněno mezi knihovnami, které byly připraveny z materiálu pěstovaného za optimálních podmínek. Mezi nimi byly nalezeny tři zajímavé geny, dva kódující F-box podobné proteiny (*MLOC_75620*, *MLOC_43997*), které hrají důležitou roli

ve vzájemné komunikaci mezi signálními dráhami různých fytohormonů (Moon *et al.*, 2004) a jeden kódující receptor podobný protein kinase (*MLOC_17138*). Protein *MLOC_17138* má kromě domény pro kinázovou aktivitu doménu pfam01657 s anotací role v odpovědi na salinitní stres a antifungální aktivitu.

Transgenní rostliny ektopicky exprimující *AtCKX1* gen vykazují lepší růstové parametry, když projdou určitým obdobím sucha (Pospíšilová *et al.*, 2016). Pro lepší pochopení vlastností těchto rostlin a molekulárních principů podílejících se na větší toleranci vůči suchu byla provedená komparativní transkriptomika nadzemní části dva týdny během revitalizace po periodě silného sucha. Z celkového počtu 26,067 genů 301 genů vykazovalo statisticky významně zvýšenou expresí a 31 sníženou expresí v revitalizované nadzemní části *vAtCKX1* rostlin oproti divokému typu (adjustovaná p-hodnota ≤ 0.05) (Vojta *et al.*, 2016). Obohacené GO kategorie genů se zvýšenou expresí jsou shrnuty v Tabulce 5.2.

GO číslo	Kategorie	GO termín	Celkový #	% ovlivněných genů	Identifikátor genu
GO.0016984	MF	ribose-bisphosphate carboxylase activity	5	40.0%	EPIHVUG0000010074, MLOC_21811
GO.0009765	BP	photosynthesis, light harvesting	32	34.4%	Nezobrazeno
GO.0030076	CC	light-harvesting complex	8	25.0%	EPIHVUG0000010021, MLOC_57061
GO.0009718	BP	anthocyanin-containing compound biosynthetic process	10	20.0%	MLOC_5324, 19814
GO.0016165	MF	linoleate 13S-lipoxygenase activity	16	18.8%	MLOC_37378, 51884, 71948
GO.0045259	CC	proton-transporting ATP synthase complex	27	14.8%	EPIHVUG0000010007, 10016, 10047, MLOC_26730
GO.0009767	BP	photosynthetic electron transport chain	56	14.3%	EPIHVUG0000010010, 10072, 10065, 10026, 10021, MLOC_52515, 22512, 39436
GO.0046271	BP	phenylpropanoid catabolic process	20	10.0%	MLOC_15203, 61189
GO.0052716	MF	hydroquinone:oxygen oxidoreductase activity	20	10.0%	MLOC_15203, 61189
GO.0009579	CC	Thylakoid	331	9.9%	Nezobrazeno
GO.0016597	MF	amino acid binding	31	9.7%	MLOC_62844, 19879, 80634
GO.0004499	MF	N,N-dimethylaniline monooxygenase activity	22	9.1%	MLOC_11897, 11896
GO.0009637	BP	response to blue light	48	8.3%	MLOC_43394, 22512, 11312, 52515
GO.0055082	BP	cellular chemical homeostasis	47	6.4%	MLOC_22808, 65878, 69460
GO.0034754	BP	cellular hormone metabolic process	38	5.3%	MLOC_6666, 73942

Tabulka 5.2: Nejvíce ovlivněné kategorie GO genů během revitalizace 2. týdne *vAtCKX1*

Nejvíce ovlivněné kategorie GO u genů se zvýšenou expresí v nadzemní části *vAtCKX1* rostlin ve 2. týdně revitalizace po obnovení závlivky. Procento udává poměr ovlivněných genů ku všem genům pro danou GO kategorii s rozdílnou expresí na hladině významnosti p-hodnoty ≤ 0.05 pro GO kategorie/termíny na šesté úrovni. MF, molekulární funkce; BP, biologický proces; CC, buněčná komponenta. Celková (#) hodnota udává počet všech přiřazených genů v genomu.

5.3.3 Diferenciální exprese genů za stresových podmínek a během revitalizace v kořenech transgenních rostlin

Kořeny rostlin pěstovaných v substrátu je problematické v použití pro transkriptomovou studii, protože během sběru dochází k mechanickému stresu, který by mohl ovlivnit výsledky transkriptomové studie. Rostliny pro analýzu transkriptomu kořene byly vždy pěstovány v hydroponickém systému. Z kořenů dvou transgenních genotypů *vAtCKX1* a *cAtCKX1* a rostlin divokého typu bylo připraveno celkem 12 sekvenačních knihoven a to vždy pro stresované vzorky (Obrázek 5.1(B)) a 14 dní po revitalizaci na tento stres (Obrázek 5.1(C)).

Stejně jako v nadzemní části stres indukoval silnou odpověď na transkriptomové

úrovni napříč všemi třemi genotypy. Mezi transgenními rostlinami a rostlinami divokého typu byla během stresu změněná exprese jen několika genů, zahrnující například domnělou *nikotinamin syntázu* (MLOC_71596) a *4-kumarát CoA ligázu* (MLOC_18901) hrající roli v lignifikaci. Celkem 57 genů bylo detekováno se signifikantně sníženou expresí v obou transgenních genotypech oproti divokému typu během stresu (Kokas *et al.*, 2016).

Během revitalizace došlo k výrazně vyšší dysregulaci genů mezi transgenními genotypy a divokým typem. Z celkového počtu 26,067 genů bylo 2,526 statisticky významně dysregulováno v genotypu *vAtCKX1* (1,973 genů s vyšší hladinou exprese a 553 genů s nižší hladinou exprese) a u genotypu *cAtCKX1* to bylo 4,153 genů (1,451 více a 2,702 méně) (Vojta *et al.*, 2016).

GO ID	Kategorie	Název	Celkový #	% ovlivněných genů		Identifikátory genů (MLOC)
				vAtCKX1	cAtCKX1	
SE ZVÝŠENOU EXPRESÍ						
GO:0008792	MF	arginine decarboxylase activity	2	100.0%	100.0%	58866, 39205
GO:0047987	MF	hydroperoxide dehydratase activity	2	100.0%	100.0%	8106, 21933
GO:0004664	MF	prephenate dehydratase activity	4	75.0%	100.0%	23316, 65725, 56414, 60716
GO:0009916	MF	alternative oxidase activity	4	75.0%	50.0%	34173, 82029, 53632
GO:0045548	MF	phenylalanine ammonia lyase activity	7	71.4%	71.4%	79728, 62322, 4684, 19798, 67067
GO:0003979	MF	UDP-glucose 6-dehydrogenase activity	3	66.7%	100.0%	5287, 70967, 63077
GO:0005345	MF	purine nucleobase transmembrane transporter activity	3	66.7%	66.7%	66246, 32910
GO:0004470	MF	malic enzyme activity	6	50.0%	83.3%	11548, 75667, 35785, 64502, 51144
GO:0009805	BP	coumarin biosynthetic process	9	55.6%	55.6%	66898, 20110, 17364, 19988, 52497
GO:0009699	BP	phenylpropanoid biosynthetic process	26	53.9%	46.2%	Nezobrazeno
GO:0016165	MF	linoleate 13S-lipoxygenase activity	16	50.0%	56.3%	Nezobrazeno
GO:0052542	BP	defense response by callose deposition	10	50.0%	50.0%	57937, 51297, 42826, 17648, 59580
GO:0006986	BP	response to unfolded protein	25	40.0%	40.0%	Nezobrazeno
GO:0035967	BP	cellular response to topologically incorrect protein	25	40.0%	40.0%	Nezobrazeno
GO:0042538	BP	hyperosmotic salinity response	10	40.0%	40.0%	69262, 52084, 42826, 2170
GO:0004348	MF	glucosylceramidase activity	5	40.0%	40.0%	398, 53162, 13522
GO:0010942	BP	positive regulation of cell death	5	40.0%	40.0%	12681, 75133
GO:0003885	MF	D-arabinono-1,4-lactone oxidase activity	9	33.3%	44.4%	31769, 34835, 68610, 59091
GO:0003978	MF	UDP-glucose 4-epimerase activity	6	33.3%	33.3%	70713, 10406
GO:0015020	MF	glucuronosyltransferase activity	12	25.0%	41.7%	4722, 54026, 65730, 8254, 56928
SE SNÍŽENOU EXPRESÍ						
GO:0004650	MF	polygalacturonase activity	32	21.9%	12.5%	6444, 4738, 68357, 67885, 51158, 53562, 75889, 13213, 72199
GO:0001666	BP	response to hypoxia	13	7.7%	23.1%	36714, 1340, 65221
GO:0042886	BP	amide transport	28	3.6%	25.0%	22335, 59508, 20029, 71333, 10510, 56891, 58935
GO:0009735	BP	response to cytokinin	13	7.7%	15.4%	23250, 58762
GO:0008375	MF	acetylglucosaminyltransferase activity	36	2.8%	19.4%	38958, 37085, 65593, 74430, 5087, 63430, 60533
GO:0016307	MF	phosphatidylinositol phosphate kinase activity	18	5.6%	16.7%	81640, 62872, 5875
GO:0008509	MF	anion transmembrane transporter activity	104	4.8%	16.4%	Nezobrazeno
GO:0042594	BP	response to starvation	44	6.8%	13.6%	56127, 7416, 44452, 16652, 4685, 57969
GO:0005667	CC	transcription factor complex	30	3.3%	16.7%	76757, 67781, 7755, 36554, 62730
GO:0045786	BP	negative regulation of cell cycle	25	8.0%	12.0%	57670, 65158, 62665
GO:0031669	BP	cellular response to nutrient levels	45	6.7%	13.3%	56127, 7416, 44452, 16652, 4685, 57969
GO:0042559	BP	pteridine-containing compound biosynthetic process	15	6.7%	13.3%	74131, 10075
GO:0008643	BP	carbohydrate transport	41	4.9%	14.6%	65088, 13612, 17903, 280, 63767, 67524, 10342, 59161
GO:0019901	MF	protein kinase binding	36	5.6%	13.9%	46471, 57670, 66940, 51179
GO:0010565	BP	regulation of cellular ketone metabolic process	16	6.3%	12.5%	36714, 14398
GO:0004672	MF	protein kinase activity	1099	4.3%	14.5%	Nezobrazeno
GO:0000272	BP	polysaccharide catabolic process	43	7.0%	11.6%	40915, 54306, 59047, 49756, 4022
GO:0009751	BP	response to salicylic acid	54	5.6%	13.0%	21464, 1340, 36714, 14398, 43518, 71936, 10787
GO:0003950	MF	NAD+ ADP-ribosyltransferase activity	11	9.1%	9.1%	66554, 72444
GO:0009723	BP	response to ethylene	35	8.6%	8.6%	44452, 53881, 64636

Tabulka 5.3: Nejvíce ovlivněné kategorie GO během revitalizace v kořenech transgenních rostlin.

Nejvíce ovlivněné kategorie/termíny genové ontologie (GO) v kořenech *vAtCKX1* a *cAtCKX1* rostlin pěstovaných hydroponicky 2 týdny revitalizaci po obnovení závlivky ve srovnání s rostlinami divokého typu. Procento udává poměr ovlivněných genů ku všem genům pro danou GO kategorii s rozdílnou expresí na hladině významnosti p-hodnoty ≤ 0.05 pro GO kategorie/termíny na šesté úrovni. MF, molekulární funkce; BP, biologický proces; CC, buněčná komponenta. Celková (#) hodnota udává počet všech přiřazených genů v genomu.

5.4 Diskuze

Transkriptomová studie na cytokinin deficientních rostlinách ječmene odhalila další mechanismy získané tolerance na stres suchem. Byla zjištěna umlčená odpověď přes cytokininový receptor HvHK3 a zvýšená exprese dvou transkripčních faktorů podílejících se na stresové odpovědi a senzitivitě ke stresovému hormonu, kyselině abscisové, v kořenech. Byla pozorována zvýšená exprese několika genů zapojených do biosyntézy fenypropanoidových prekurzorů a arogenát dehydratázy/lyázy, enzymu zodpovědného za syntézu fenyalaninu, v kořenech transgenních rostlin během revitalizace. Přestože bylo ve stresu vystavených kořenech detekováno více prekurzorů ligninu, celkové množství ligninu mezi rostlinami vystavenými stresu a těmi pěstovanými v kontrolních podmínkách změněno nebylo. Zvýšené hodnoty ligninu vykazovaly všechny testované skupiny transgenních rostlin oproti divokému typu nezávisle na vnějších podmínkách. Změny v transkriptomu kořene dále indikují na zvýšený obrát v metabolismu dalšího morfogenního hormonu – auxinu. Komparativní studie transkriptomu poukázala na zvýšenou expresi genů zapojených do syntézy dalšího stresového hormonu jasmonátu a dalších těkavých látek v transgenních rostlinách. Oproti podmínkám, kdy byly rostliny pěstovány za optimálního přísunu vody, byla v transgenních listech během revitalizace na suchu detekována zvýšená exprese genů kódující proteiny, které jsou zapojené do fotosyntézy, především ty obsažené v genomu chloroplastů. Toto zjištění koresponduje s faktem, že fotosyntetické parametry transgenních rostlin se vrátily k normálu rychleji během revitalizace než ty měřené u rostlin kontrolních (divokého typu). Částečná necitlivost na morfogenní hormon cytokinin u rostlin ektopicky exprimujících jeho deaktivční hormon vede ke zvýšené toleranci na stres suchem.

5.4.1 Vliv deficiencie cytokininů na nadzemní část *vAtCKX1* pěstovaných za optimálních podmínek

Čtyři nejvíce negativně ovlivněné procesy v listech *vAtCKX1* rostlin jsou spojeny s fotosyntézou. Měření fluorescenčních fotosyntetických parametrů ukázalo na pokles v přenosu energie mezi komplexy fotosystému II, měření obsahu chlorofylu pak nižší hladiny v transgenních listech (Vojta *et al.*, 2016). Tato zjištění naznačují, že fotosyntetický aparát a fotosyntéza celkově bude mírně negativně ovlivněna v transgenních rostlinách oproti kontrolám pěstovaných za optimálních podmínek.

V listech *vAtCKX1* rostlin byla zvýšená exprese tří z celkového množství všech čtyř genů kódovaných v ječmenném genomu kódující prefenát/arogenát dehydratázu, enzymu podílejícím se na posledním kroku syntézy aromatických aminokyselin tyrosinu a fenyalaninu (Rippert & Matringe, 2002). Fenyalanin je primárním substrátem

pro fenylpropanoidovou syntetickou dráhu vedoucí k produkci například ligninu, flavonoidů a antokyanů. Souhlasně s tím jsou mezi nejvíce ovlivněnými procesy ty odpovídající GO termínům GO:0009963 „Positive regulation of flavonoid biosynthetic process” a GO:0009718 „Anthocyanin-containing compound biosynthetic process”. Zvýšená produkce fenylalaninu může sloužit jako zdroj pro vyšší množství flavonoidů a antokyanů v listech *vAtCKX1* transgenních rostlin, kde se tyto látky mohou účastnit ochranných mechanismů proti rozličným stresovým faktorům. Expresi čtyř z osmi genů kódujících fenylalanin amoniak lyázu, jejíž aktivita je považována za přechod mezi primárním metabolismem aminokyselin a fenylpropanoidovou biosyntetickou dráhou (Cass *et al.*, 2015), byla výrazně snížena. Nicméně všechny čtyři geny (*MLOC_4684*, *MLOC_62322*, *MLOC_79728* a *MLOC_477*) mají zvýšenou expresi v kořenech *vAtCKX1* rostlin. Vzhledem k tomu, že metabolismus aromatických aminokyselin není změněn v kořenech, nadbytek fenylalaninu v nadzemní části transgenních rostlin může být translokován do kořenů, kde může sloužit jako zdroj pro zvýšenou produkci ligninu.

Třetím nejvíce ovlivněným procesem v nadzemní části transgenních rostlin je aktivita lipoxygenáz, které z intracelulárních lipidů produkují těkavé látky včetně rostlinného hormonu jasmonátu (Feussner & Wasternack, 2002). Tyto látky jsou většinou uvolňovány během odpovědi rostliny na napadení nějakým patogenem. V případě, že transkriptomová data jsou výsledkem čtyř biologických opakování, z čehož vždy dvě a dvě jsou z různého prostředí, je téměř vyloučeno, že zvýšená produkce lipogenáz a těkavých látek bude odezvou na nějakého přehlédnutého patogenního činitele. Kromě odpovědi na abiotický stres se jasmonáty podílejí na několika vývojových procesech jako je tvorba trichomů doprovázená zvýšenou produkcí sekundárních metabolitů jako jsou flavonoidy, antokyany a terpenoidy (Tian *et al.*, 2012; Wasternack, 2014). Ačkoliv neexistují důkazy o tom, že cytokininy se vzájemně ovlivňují s jasmonáty, předpokládá se, že jejich efekt bude antagonistický vzhledem k tomu, že jasmonát silně inhibuje cytokininem indukovanou tvorbu kalusu (Ueda & Kato, 1982). Nicméně, vzájemné ovlivňování cytokininů a jasmonátu pravděpodobně ovlivňuje koncentrace dalších hormonů (O'Brien & Benková, 2013). Celkem 12 z 55 a 8 z 39 genů kategorizovaných jako GO:0009753 „Response to jasmonic acid” a GO:0010026 „Trichome differentiation”, bylo reportováno se statisticky významnou zvýšenou expresí v kořenech dvou nezávislých liniích rostlin *vAtCKX1*.

Zvýšená produkce jasmonátu v nadzemní části transgenních rostlin by mohla ovlivňovat především detailní architekturu kořene jako je tvorba kořenového vlášení. Těkavý metyl-jasmonát je translokován jako rychlý chemický signál z nadzemní části do kořenů, kde funguje jako aktivátor genové exprese (Baldwin *et al.*, 1994). GO termín GO:0016165 „Linoleate 13S-lipoxygenase activity” patří mezi ty se zvýšeným zastoupením v transgenních

kořenech během revitalizace, proto je možné, že dochází i ke zvýšené produkci jasmonátu přímo v kořeni.

5.4.2 Odpověď v nadzemních částech *vAtCKX1* během revitalizace

Produkty mnoha genů, které jsou statisticky významně více exprimovány v *vAtCKX1* rostlinách oproti kontrolám, patří mezi strukturní komponenty a enzymy fotosyntetického aparátu ve shodě s měřenými fotosyntetickými parametry, které vykazovaly lepší hodnoty u transgenních rostlin v časných stádiích revitalizace po období silného sucha, podobně dosáhly hodnoty obsahu chlorofylu v transgenních listech hladin rostlin divokého typu během dvou týdnů revitalizace po obnovení pravidelné závlivky (Vojta *et al.*, 2016).

Mezi nejvíce nadexprimovanými geny byly ty kódované na ječmenném chloroplastovém genomu (anotované s prefixem EPIHVUG). Z celkového počtu 112 všech přepisovaných chloroplastových genů byla u čtrnácti exprese statisticky významně zvednutá 2-3x (na hladině významnosti adjustované p-hodnoty ≤ 0.05). Byla provedena „real-time“ PCR analýza šesti vybraných genů pro validaci spolehlivosti transkriptomové analýzy a získaných výsledků (Vojta *et al.*, 2016). Chloroplasty jsou již dávno popsáným cílem aktivity cytokininů. Exogenně aplikovaný cytokinin je schopen cíleně aktivovat expresi několika chloroplastových genů v listech ječmene, ve kterých se taktéž akumuluje stresový hormon kyselina abscisová (Zubo *et al.*, 2008). Zatím není jisté zda jsou cytokininy schopny přímo aktivovat transkripci RNA v chloroplastech. Je tedy diskutabilní, že zvýšená transkripce některých genů v *vAtCKX1* listech je zapříčiněna zvýšenou akumulací cytokininů po stresu suchem. Zmíněná hypotéza je ovšem podpořena faktem, že v listech rostlin se zvýšenou aktivitou cytokinin dehydrogenázy dochází k silné aktivaci exprese některých isopentenyltransferáz (cytokinin biosyntetický enzym) v několika vývojových stádiích. Ektopické zvýšení hladin cytokininů, proto může vést k posílení exprese chloroplastových genů (Pospíšilová *et al.*, 2016). Analýza endogenních cytokininů v chloroplastech by mohla potvrdit/vyvrátit navrhovanou hypotézu.

Mezi další zajímavé geny významně více exprimované v revitalizované nadzemní části *vAtCKX1* rostlin patří čtyři, které kódují domnělé akvaporiny (*MLOC_56278*, *MLOC_71237*, *MLOC_552*, *MLOC_22808*). Akvaporiny jsou kanálové proteiny zajišťující transport vody přes plazmatickou a další vnitrobuněčné membrány. Zvýšená exprese několika genů kódující ječmenné akvaporiny byla již pozorována v rostlinách ječmene vystaveným salinitnímu stresu (Wei *et al.*, 2007).

Shrňme-li dohromady výše uvedené poznatky, srovnávací expresní studie mezi transgenními *vAtCKX1* rostlinami a rostlinami divokého typu před a po periodě silného sucha odhalila několik genů a procesů, které jsou vyvolány nerovnováhou v hladinách cytokininů

a mohly by vést k lepší toleranci vůči suchu či rychlejšímu růstu ve fázi revitalizace.

5.4.3 Odpověď v kořenech *vAtCKX1* a *cAtCKX1* během stresu a revitalizace

Gen s nejvíce sníženou expresí v obou genotypch oproti kontrole během působení stresu je domnělý protein s F-box doménou (*MLOC_75620*, 12.6-krát ve *vAtCKX1* a 5.3x ve *cAtCKX1*). Tento gen byl také detekován mezi geny s nejvíce sníženou expresí v nadzemní části během 12-hodinové a 14-denní revitalizace po období sucha. Výrazně snížená exprese tohoto genu naznačuje, že by tento transkripční faktor mohl hrát zásadní roli v komunikaci s dalšími hormony v rostlinách ektopicky nadprodukujících cytokinin dehydrogenázu. Proteiny s F-box doménou reprezentují jednu z největších genových rodin v rostlinách, která je zapojena do procesu ubikvitinylace a degradace proteinů. Do současnosti bylo funkčně charakterizováno jen minimální množství proteinů s F-box doménou a to ty, které jsou zapojeny do regulace hormonálních signálních drah, kde mohou degradovat represory nebo aktivátory signalizace auxinů, gibberellinů, etylénu nebo jasmonátu (Moon *et al.*, 2004).

Analýza odlišně exprimovaných genů v revitalizovaných kořenech ukázala, že jeden z genů kódující cytokininový receptor (*HvHK3*; *MLOC_44452*) měl významně sníženou expresi (Vojta *et al.*, 2016). Gen měl také sníženou expresi v šest týdnů starých kořenech pěstovaných celou dobu za optimálních podmínek (Pospíšilová *et al.*, 2016). Pravděpodobně snížené vnímání přes *HvHK3* receptor je spíše vývojovým důsledkem v transgenních rostlinách se sníženou hladinou cytokininů v kořeni než odpovědí na prodělanou periodu sucha. Přidáním jednoho genu pro cytokinin dehydrogenázu do genomu ječmene vede k hormonální dysbalanci, kterou se snaží rostlina kompenzovat regulací ostatních endogenních cytokinin dehydrogenáz a isopentenyl transferáz velmi jemným mechanismem (Pospíšilová *et al.*, 2016). Rostlinné isopentenyl transferázy jsou exprimovány velmi slabě a jejich enzymová aktivita může být regulována pomocí farnesylace (Galichet *et al.*, 2008). Významné zvýšení exprese dvou abundantních endogenních cytokinin dehydrogenáz *HvCKX4* a *HvCKX5* bylo pozorováno v obou nezávislých experimentech rostlinách transgenního ječmene, zatímco další isoformy byly sníženy nebo neregulovány ve srovnání s rostlinami divokého typu (Pospíšilová *et al.*, 2016; Vojta *et al.*, 2016). Je možné, že lokální minima v koncentracích cytokininů mohou vést ke snížení transkripce *HvHK3* a oslabení vnímání přes tento receptor.

5.4.4 Fenylpropanoidová biosyntetická dráha

Prvotní transkriptomová charakterizace *vAtCKX1* transgenních kořenů odhalila silně zvýšenou expresi mnoha genů kódujících enzymy zahrnuté ve fenylpropanoidové biosyntetické dráze (Pospíšilová *et al.*, 2016). Obsah ligninu byl kvantifikován acetylbromidovou metodou v mase buněčných stěn zbavených proteinů ze vzorků celých kořenů a nadzemních částí hydroponicky pěstovaných rostlin (Vojta *et al.*, 2016). Oba transgenní genotypy ukázaly systematické zvýšení obsahu celkového ligninu v kořenech rostlin dva týdny po vyklíčení. Přibližně o 20 % více ligninu bylo detekováno v transgenních kořenech obou genotypů oproti kontrolám kultivovaným za optimálních podmínek šest týdnů po vyklíčení. Obsah ligninu v listech nebyl mezi vzorky významně ovlivněn.

Analýza prekurzorů ligninu ve 4-týdny starých rostlinách *cAtCKX1* odhalila, že tento genotyp akumuluje významně větší množství monolignolů a cinamátů než pletivo kontrolních rostlin. Čtyři týdny staré rostliny genotypu *vAtCKX1* pěstované v optimálních podmínkách obsahovaly větší množství sinapyl alkoholu, stejně jako o dva vývojové týdny později. Vystavení rostlin stresu suchem a následná revitalizace znamenala výraznější akumulaci všech kvantifikovaných intermediátů fenylpropanoidové metabolické dráhy, stejně tak vyšší obsah ligninu ve frakci buněčných stěn z kořene v obou transgenních genotypech oproti netransgenní kontrole. Avšak lignifikace v transgenních kořenech, které prošly periodou sucha a následnou revitalizací nebyla výraznější než u kořenů pěstovaných v optimálních podmínkách.

Měření potvrdila, že snížená citlivost na cytokinin v transgenních rostlinách měla pozitivní efekt na lignifikaci kořene. Účinek ale není posílen vystavením rostlin stresu i přesto, že je produkováno více prekurzorů ligninu. Neschopnost produkovat více ligninu během revitalizace po stresu může spočívat v nedostatečnosti specifických peroxidáz podílejících se na polymeraci monolignolů. Nejvíce obohacené GO termíny v revitalizovaných kořenech obou transgenních genotypů korelují se zvýšenou hladinou intermediátů fenylpropanoidové biosyntetické dráhy a silnější lignifikací.

Zhruba polovina genů připisovaných fenylpropanoidové biosyntetické dráze byla významně více exprimovaná stejně jako pět ze sedmi genů anotovaných jako fenylalanin:amoniak lyáza a čtyři geny kódující preferát dehydratázu. Zvýšená exprese čtyř vybraných genů byla potvrzena pomocí „real-time“ PCR (Vojta *et al.*, 2016). Syntéza ligninu je závislá na stejných prekursorech, jako syntéza derivátů benzoové kyseliny. Hladiny tří nejvíce abundantních derivátů kyseliny benzoové byly významně redukovány v kořenech *vAtCKX1* rostlin kultivovaných za optimálních podmínek. Proto se zdá být metabolický přechod mezi syntézou ligninu a derivátů kyseliny benzoové jako důsledek změněné homeostázy cytokininů.

Mezi dalšími procesy, jejichž geny mají zvýšenou expresi byly identifikovány aktivity hydroperoxid dehydratáz a linoleát 13S-lipoxygenáz, které mohou stimulovat produkci jasmonátu či dalších těkavých látek, nebo geny zapojené do syntézy kumarinů a odpovědi na hyperosmotický stres.

5.5 Závěr

Ve třetí kapitole je prezentována jedna z prvních transkriptomových studií provedená na rostlinách ječmene. Ačkoli je hrubý draft genomu ječmene dostupný již několik let, jeho přesná anotace a klasifikace predikovaných genů do funkčních kategorií, jako jsou například GO termíny je stále neúplná a v některých ohledech neuspokojivá. Například ke GO termínu „response to cytokinin“ je přiřazeno pouze 13 domnělých genů, ale žádný ze skutečných regulátorů primární cytokininové odpovědi není zahrnut. Nicméně, metoda RNASeq umožňuje komplexně hodnotit všechny procesy, které probíhají v rostlinách s cytokininovou nerovnováhou a poskytuje nové úhly pohledu na dosud neznámé interakce například s dalšími hormony. Mimo fakt, že popisované transgenní rostliny vykazují vlastnost lepší rezistence ke stresu suchem díky modifikované morfologii kořene, která je pravděpodobně důsledkem v posunu rovnováhy mezi dvěma morfogeny cytokininem a auxinem, částečná necitlivost k cytokininům způsobená nižší expresí receptoru HvHK3 může bezpochyby ovlivňovat další procesy vedoucí k vyšší toleranci k suchu. Důkazem toho je třeba zvýšená exprese čtyř genů kódujících akvaporiny, jejímž důsledkem by mohla být schopnost transgenních rostlin zvyšovat vodní potenciál rychleji než rostliny divokého typu. Proces revitalizace nadzemní části po období silného sucha je doprovázen zvýšenou expresí genů zapojených do fotosyntézy, zvláště těch kódovaných na genomu chloroplastů. Tento aspekt vede k rychlejší regeneraci transgenních rostlin, která je pozorována jako vyšší kumulace biomasy po prodělaném období sucha. Změněná homeostáza cytokininů dále vede k akceleraci sekundárního metabolismu vycházejícího z fenylalaninu, který má za následek akumulaci některých meziproduktů fenylpropanoidové biosyntetické dráhy v kořenech. Z toho pohledu není překvapivé masivnější ukládání ligninu v pletivech kořene a tvorba dalších sloučenin jako jsou antokyany nebo flavonoidy. Komparativní transkriptomová analýza odhalila několik genů, které mohou hrát klíčovou roli ve stresu odolnému fenotypu rostlin s ektopicky nadprodukovanou *AtCKX1*. Kromě dvou transkripčních faktorů z rodiny MYB nebo NAC, které již dříve ukázaly zvýšenou citlivost vůči abscisové kyselině, byl nalezen jeden domnělý protein s F-box doménou (*MLOC_75620*), jehož exprese byla silně snížena ve všech transgenních pletivech. Tento regulační protein by mohl být zapojen do ubiquitinylace represorů nebo aktivátorů signálních drah dalších fytohormonů jako jsou auxiny, gibberelliny nebo jasmonáty a zprostředkovávat tak vzájemnou komunikaci cytokininů s nimi. Posílená produkce jasmonátu a dalších těkavých látek by mohla být

z dalších procesů, které jsou charakteristické pro *AtCKX1* nadprodukující rostliny, vzhledem k tomu, že aktivita linoleát 13S-lipoxygenáz patří mezi nejvíce obohacené GO termíny jak v transgenních listech, tak kořenech během revitalizace. Závěrem lze tvrdit, že vložením jednoho genu pro degradaci cytokininů pod ektopicky řízeným slabým promotorem má za následek cytokinin necitlivý fenotyp, který aktivuje procesy, jenž umožňují rostlině lépe regenerovat po období deficitu vody.

6 Souhrn

Procesy analýzy dat z masivního paralelního sekvenování (MPS) jsou často náročné na výkon výpočetní infrastruktury a dobu počítání. Existuje tedy potřeba automatizace a paralelizace jednotlivých úloh na výkonných počítačových klastrech (HPC). V první kapitole představujeme nově vytvořenou platformu určenou pro resekvenační analýzu MPS dat-MOLDIMED a nový anotační nástroj, který je určen k nasazení na HPC.

Druhá kapitola popisuje použití MPS v případě Diamond-Blackfanovy anémie (DBA), které je zapříčiněno mutacemi v genech ribozomálních proteinů (RP). Přes známou molekulární patologii, zhruba 25 % případů DBA zůstává neobjasněno na molekulární úrovni. U pacientky s neznámou příčinou DBA jsme provedli RP panelové sekvenování a popsali jsme novou nesynonymní mutacikódující *RPS7* p.V134F. Mutace byla následně potvrzena i u dvou asymptomatických rodinných příslušníků, u kterých následně byla potvrzena slabá anémie. Dále jsme provedli komparativní transkriptomovou analýzu všech nositelů *RPS7* mutace ve srovnání se zdravým kontrolním souborem a DBA pacientů se známou mutací v *RPS19*, kde byla sledována dysregulace signálních drah především v procesech translace, rRNA procesování, regulace buněčného cyklu a imunitního systému.

Třetí kapitola představuje transkriptomovou studii transgenních odrůd ječmene s ektopickou expresí cytokinin dehydrogenázy 1 z *Arabidopsis* modelu. Transgenní rostliny vykazují zvýšenou rezistenci na stres suchem pozměněnou architekturou kořene a silnější lignifikací. Za optimálních podmínek v revitalizaci byla u listů sledována vyšší hladina exprese chloroplastových genů účastnících se procesů fotosyntézy.

7 Summary

Massive parallel sequencing (MPS) data analysis tasks are often computationally demanding and their execution time would take too long using standard computing machines. Thus there is a need for parallelization of this tasks and ability to execute them on a sufficiently powerful computing machines. In the first chapter we describe a newly created platform for resequencing analysis of MPS data - MOLDIMED and novel annotation tool, which is ready to deploy on HPC infrastructure.

The second chapter describes MPS approaches in Diamond-Blackfan anaemia (DBA), which is predominantly underlined by mutations in genes encoding ribosomal proteins (RP); however, its etiology remains unexplained in approximately 25% of patients. We performed panel sequencing of all ribosomal genes in DBA patient without previously known molecular pathology. A novel heterozygous *RPS7* mutation coding *RPS7* p.V134F was found in one female patient and subsequently confirmed in two asymptomatic family members, in whom mild anemia were detected on further examination. Subsequently, we performed whole transcriptome analysis in all family members and patient with *RPS7* mutation in comparison with healthy control group and with DBA patients with known mutation in *RPS19*. We observed dysregulation mainly in signal pathways of translation, cell cycle, rRNA processing and in genes of immune system.

The third chapter demonstrate MPS deployment in whole transcriptome analysis of novel transgenic barley lines overexpressing cytokinin dehydrogenase 1 gene from *Arabidopsis*. These lines showed drought-tolerant phenotype mainly due to alteration of root architecture and stronger lignification of root tissue. In optimal conditions of revitalized leaves, comparative transcriptomic analysis displayed up-regulation of genes encoding enzymes and proteins involved in photosynthesis, and especially those encoded by the chloroplast genome.

8 Seznam zkratek

ADAM17	ADAM metaloproteázová doména 17
AES	Amino-terminální transkripční korepresor
AGO3	Argonaute 3, RISC katalytická komponenta
AKT1S1	AKT1 substrát 1
AML	Akutní myeloidní leukémie
APAF1	Apoptotický proteázový aktivační faktor 1
APRT	Adenin fosforibozyltransferáza
AtCKX1	Arabidopsis cytokinin dehydrogenáza 1
ATP	Adenosintrifosfát
ATP7A	ATPázový transportér mědi alfa
AXIN2	Axin 2
B3GALT6	Beta-1,3-Galaktozyltransferáza 6
BED	Formát zápisu intervalových pozic genomu
BHLHA9	Bazický Helix-Loop-Helix transkripční faktor A9
BLAST	Prohledávací program sekvenčních vzorů
BTG2	BTG anti-proliferační faktor 2
BWA	Aligner Burrows-Wheelerovy transformace
BWT	Burrows-Wheelerova transformace
BYSL	Bystin-like
C4BPA	Vazebný protein komponenty komplementu 4 alfa
C5AR1	C5a Receptor 1
CA1	Karbonická anhydráza 1
CARD11	Kaspáza vazebný protein rodiny 11
CARM1	Arginin methyltransferázový koaktivátor 1
CASP8	Kaspáza 8
cAtCKX1	Cytosolická forma Arabidopsis cytokinin dehydrogenázy 1
CCD	Zařízení pro snímání obrazu (charge-coupled device)
CCL	Chemokinový ligand s C-C motivem

CCNL1	Cyklin L1
CCR7	Chemokinový receptor s C-C motivem 7
CDKN1A	Inhibitor cyklin dependentní kinázy 1A
cDNA	Kódující DNA
CDS	kódující sekvence
CELF2	Protein rodiny CELF/BRUNOL
CFLAR	Regulátor apoptózy kaspázy 8 a FADD
ChIP-Seq	Chromatin imunoprecipitační sekvenování
CLIC5	Vnitrobuněčný chloridový kanál
CNV	Početní dysbalance DNA
CPEB4	Vazebný protein polyadenylačního elementu
CR1	Receptor 1 (Complement C3b/C4b, Knops Blood Group)
CRAM	komprimovaný formát alignmentu
CSV	Specifický textový formát zápisu s dělenými sloupci čárkou
CTGF	Růstový faktor pojivové tkáně
CXCL	Chemokinový ligand C-X-C motivu
CXCR2	Chemokinový receptor 2 C-X-C motivu
DAPK2	Kináza asociovaná s buněčnou smrtí
DBA	Diamond-Blackfanova anémie
DDX	DEAD-Box helikáza
DEPC	Diethylpyrokarbonát
DGE	Diferenciální genová exprese
DMTN	Aktin vazebný protein
DNA	Deoxyribonukleová kyselina
dNTP	Deoxynukleotid trifosfát
DOCK5	Dedikátor cytokineze 5
DP	Hloubka čtení, pokrytí
DSP	Dezmoplakin
dUTP	Deoxyuridintrifosfát
E2F1	Transkripční faktor E2F1

eADA	Erytrocytární adenosindeamináza
EGF	Epidermální růstový faktor
EGFR	Receptor epidermálního růstového faktoru
EGR1	Mitogenní a diferenciační faktor EGR1
EIF x	Eukaryotní iniciační faktory translace
ELANE	Neutrofilní elastáza
ELF3	Transkripční faktor (E74 Like ETS)
EOF	Značka konce souboru
EPB42	Erytrocytární membránový protein Band 42
ERK1/2	Kinázové receptory vnějších signálů
ESPN	Espin
FAM210B	Protein (Sequence Similarity 210 Member B)
FAS	Povrchový receptor buněčné smrti
FBL	Fibrillarin
FBXW7	Protein 7 rodiny F-Box s WD doménovým motivem
FCAR	Fc fragment IgA receptoru
FCER	Fc fragment IgE receptoru
FFPE	Vzorky fixované v parafínu
FLVCR1	Transportérový protein hemu FLVCR1
FMR1	Protein FRAXA-syndromu fragilního X chromozomu
FOS	Protoonkogen, podjednotka AP-1 transkripčního faktoru
FOSB	Protoonkogen, podjednotka AP-1 transkripčního faktoru
FOXO3	Protein (Forkhead Box O3)
FPKM	Fragmentů na kilobázi na milion
GATA1	GATA vazebný protein 1
gDNA	Genomová DNA
GEO	Databáze (Gene Expression Omnibus)
GLM	Zobecněné lineární modely
GO	Genová ontologie
GPX4	Glutathion peroxidáza 4

GRAP2	Adaptorový protein 2 (GRB2-Related)
GTF/GFF	Formát zápisu funkčních genomových regionů
GUI	Uživatelské grafické rozhraní
gVCF	Zápis variant v genomovém formátu
GWAS	Populační genomové studie
Hb	Hemoglobin
HBD	Hemoglobinová podjednotka Delta
HBG	Hemoglobinová podjednotka Gamma
HCRT	Hypocretin Neuropeptide Precursor
HIF1A	Alfa podjednotka hypoxií inducibilního faktoru 1
HLA	Hlavní histokompatibilní komplex
HOXA13	Homeodoménový protein A13
HPC	Supervýpočetní klastr
HPO	Databáze (Human Phenotype Ontology)
HvCKX	Cytokinin oxidáza/dehydrogenáza ječmene
HvHK3	Cytokininový receptor ječmene HvHK3
ICOSLG	Inducibilní T buněčný ligand
IFI27	Inducibilní protein 27 interferonem alfa
IGHD	Konstatní oblast těžkého řetězce imonoglobulinu D
IGHG	Konstatní oblast těžkého řetězce imonoglobulinu G
IGHM	Konstatní oblast těžkého řetězce imonoglobulinu M
IGHV	Těžký řetězec imunoglobulinu
IGKC	Imunoglobulin kappa konstantní region
IGKV3-20	Imunoglobulin kappa variabilní region 3-20
IGLC	Imunoglobulin lambda konstantní region
IL1	Interleukin 1
IL21R	Interleukinový receptor 21
IL23A	Podjednotka alfa interleukinu 23
IL27RA	Podjednotka alfa interleukinového receptoru 27
INDEL	Inzerce/delece

IRAK3	Interleukinový receptor 1 asociovaný s kinázou 3
JAK2	Janus kináza 2
KCNJ10	Draslíkový kanál rodiny J 10
KEGG	Databáze (Kyoto Encyclopedia of Genes and Genomes)
KIF27	Člen rodiny kinezinů 27
KRT1	Keratin 1
LAMP2	Protein asociovaný s membránou lyzozómu 2
LCK	Protoonkogen rodiny Src
LEA	Protein (Late-embryogenesis abundant)
LEF1	Lymfoidní vazebný faktor 1
LRMP	Lymfoidní membránový protein
LSU-rRNA	Velká podjednotka ribozomální rRNA
LTB	Lymfotoxin beta
LY96	Lymfocyt antigen 96
LYN	LYN protoonkogen rodiny Src
MAL	T buněčný diferenciační faktor
MAP3K8	Mitogen aktivovaná protein kináza kináza kináza 8
MCM2	MCM2 pre-replikační komplex
MCV	Střední objem erytrocytu
MDM2	Protoonkogen MDM2
MDS	Myelodysplastický syndrom
MIB1	Mindbomb E3 Ubiquitin Protein Ligase 1
miRNA	mikroRNA
MME	Membránová metaloproteáza
MMP25	Matrix metaloproteáza 25
MNP	Komplexní nukleotidová varianta
MNX1	Motor Neuron And Pancreas Homeobox 1
MOSPD2	Motile Sperm Domain Containing 2
MPS	Masivní paralelní sekvenování
MQ	Mapovací kvalita

mRNA	Mediátorová RNA
MS4A2	MS4A2 membránový protein
MT2A	Metallothionein 2A
MUC1	Mucin 1
MYB	MYB protoonkogen
MYBL2	MYB like 2 protoonkogen
MYC	MYC protoonkogen
NAC	NAC transkripční faktory
NF-kB	Nukleární faktor kappa B
NGS	Sekvenování nové generace
NOP14	Jadérový protein NOP14
NOS3	Syntáza oxidu dusnatého 3
NOTCH	Transkripční faktor NOTCH
NOXA	Protein NOXA
NPHP1	Nefrocystin 1
OLC	Metoda skládání sekvencí na základě shody překryvu přesahů
ORF	Otevřený čtecí rámec
PARP1	Poly(ADP-ribóza) polymeráza 1
PCA	Analýza hlavních komponent
PCR	Polymerázová řetězová reakce
PELI1	Pellino E3 ubikvitin protein ligáza 1
PES1	Pescadillo ribozomální faktor 1
PF4V1	Trombocytární faktor 4 varianta 1
PGM2	Fosfoglukomutáza 2
PIK3CA	Fosfatidylinositol-4,5-bisfosfát 3-kinázová katalytická podjednotka alfa
PLAGL1	PLAG1 Like Zinc Finger 1
PLEK	Pleckstrin
PMAIP1	Phorbol-12-myristoyl-13-acetyl indukovaný protein 1
PNKP	polynukleotid kináza 3'-fosfatáza

PNP	Fosforyláza purinových nukleosidů
PPBP	Pro-trombocytární základní protein
PPIF	Peptidylprolyl izomeráza F
PPV	Hodnota pozitivní predikce
pre-mRNA	Primární transkript
PRKACA	cAMP-aktivovaná kinázová katalytická podjednotka alfa
PRMT1	Protein Arginine Methyltransferase 1
PROS1	Protein S
PTAFR	Receptor trombocytárního aktivačního faktoru
PTCRA	T buněčný antigen receptor alfa
PTEN	Fosfatáza a homolog tensinu
PTGS2	Cyklooxygenáza
RAB3D	Protein RAS protoonkogenní rodiny
RACK1	Receptor pro aktivovanou C kinázu 1
RAS	RAS protoonkogen
RICTOR	RPTOR nezávislý člen MTOR komplexu
RIN	Hodnota integrity RNA
RNA	Ribonukleová kyselina
RNASeq	RNA sekvenování
RNPC3	U11/U12 malý jaderný ribonukleoprotein
RP	Ribozomální proteiny
RPKM	Čtení na sto bází na milion
RPL	Protein/y velké ribozomální podjednotky
RPLP0P6	Ribozomální protein P0 - pseudogen 6
RPLP2	Ribozomální protein LP2
RPN2	Riboforin II
RPS	Protein/y malé ribozomální podjednotky
RPS4Y1	Ribozomální protein S4 na Y chromozomu
rRNA	Ribosomální rRNA
RUNX2	Transkripční faktor (Runt Related) 2

RVIS	Skóre varianční intolerance
S1PR1	Sfingosin-1-fosfát receptor 1
SFN	Stratifin
SIRPB1	Signál regulující protein beta 1
SMOC2	SPARC vápník vázající protein 2
SMRT	Molekulové sekvenování v reálném čase
SNP	Jednonukleotidový polymorfismus
SNRPC	Malý jaderný ribonukleoprotein C
SNU13	Malý jaderný ribonukleoprotein 13
SNV	Jednonukleotidová varianta
SOS2	SOS Ras/Rho guanin výměnný faktor 2
SOX18	SRY-Box 18
SPECC1L	Antigen s kalponin homologií
SPTAN1	Spektrin alfa 1, neerytrocytární
SQL	Standardizovaný strukturovaný dotazovací jazyk
SSU-rRNA	Malá podjednotka ribozomální ribonukleové kyseliny
vAtCKX1	Vakuolární forma Arabidopsis cytokinin dehydrogenázy 1
VCF	Formát zápisu genetických variant
WCF	Sada knihoven (Windows Communication Foundation)
WES	Celoexomové sekvenování
WGS	Celogenomové sekvenování
WNT	Wingless/Integration
XPNPEP3	X-Prolyl aminopeptidáza 3
YBX3	Y-Box vazebný protein 3
ZMW	Optická nanostruktura v čipu

9 Seznam citované literatury

- [1] Adessi C. (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research* 28, 87e–87.
- [2] Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., et al. (2010) A method and server for predicting damaging missense mutations. *Nature Methods* 7, 248–249.
- [3] Afgan E., Baker D., van den Beek M., Blankenberg D., Bouvier D., Čech M., et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research* 44, W3–W10.
- [4] Alonso-Ramirez A., Rodriguez D., Reyes D., Jimenez J.A., Nicolas G., Lopez-Climent M., et al. (2009) Evidence for a Role of Gibberellins in Salicylic Acid-Modulated Early Plant Responses to Abiotic Stress in Arabidopsis Seeds. *PLANT PHYSIOLOGY* 150, 1335–1344.
- [5] Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215, 403–410.
- [6] Anders S., Pyl P.T. & Huber W. (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169.
- [7] Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., et al. (2000) Gene ontology: Tool for the unification of biology. *Nature Genetics* 25, 25–29.
- [8] Audo I., Bujakowska K.M., Léveillard T., Mohand-Sad S., Lancelot M.E., Germain A., et al. (2012) Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases. *Orphanet Journal of Rare Diseases* 7, 8.
- [9] Aziz R.K., Bartels D., Best A., DeJongh M., Disz T., Edwards R.A., et al. (2008) The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* 9, 75.
- [10] Azuma M., Toyama R., Laver E. & Dawid I.B. (2006) Perturbation of rRNA synthesis in the *bap28* mutation leads to apoptosis mediated by p53 in the zebrafish central nervous system. *Journal of Biological Chemistry* 281, 13309–13316.
- [11] Badhai J., Fröjmark A.S., J. Davey E., Schuster J. & Dahl N. (2009) Ribosomal protein S19 and S24 insufficiency cause distinct cell cycle defects in Diamond-Blackfan anemia. *Biochimica et Biophysica Acta - Molecular Basis of Disease* 1792, 1036–1042.
- [12] Baldwin I.T., Schmelz E.A. & Ohnmeiss T.E. (1994) Wound-induced changes in root and shoot jasmonic acid pools correlate with induced nicotine synthesis in *Nicotiana sylvestris* *spiegazzini* and *comes*. *Journal of Chemical Ecology* 20, 2139–2157.
- [13] Ball S. (2011) Diamond Blackfan anemia. *Hematology Am Soc Hematol Educ Program* 2011, 487–491.
- [14] Barbazuk W.B., Emrich S.J., Chen H.D., Li L. & Schnable P.S. (2007) SNP discovery via 454 transcriptome sequencing. *Plant Journal* 51, 910–918.
- [15] Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59.
- [16] Bergmann C. (2018) Genetics of Autosomal Recessive Polycystic Kidney Disease and Its Differential

Diagnoses. *Frontiers in Pediatrics* 5, 221.

- [17] Blume K.G., Kaplan J.C., Löhr G.W., Ramot B. & Valentine W.N. (1977) International Committee for Standardization in Haematology: Recommended Methods for Red-Cell Enzyme Analysis. *British Journal of Haematology* 35, 331–340.
- [18] Bolger A.M., Lohse M. & Usadel B. (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- [19] Bortolomeazzi M., Gaffo E. & Bortoluzzi S. (2017) A survey of software tools for microRNA discovery and characterization using RNA-seq. *Briefings in Bioinformatics*.
- [20] Branton D., Deamer D.W., Marziali A., Bayley H., Benner S.A., Butler T., et al. (2008) The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26, 1146–1153.
- [21] Brown T.L. & Meloche T.M. (2016) Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* 108, 109–114.
- [22] Busby M.A., Stewart C., Miller C.A., Grzeda K.R. & Marth G.T. (2013) Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* 29, 656–657.
- [23] Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10, 421.
- [24] Cantarel B.L., Korf I., Robb S.M.C., Parra G., Ross E., Moore B., et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18, 188–196.
- [25] Capriotti E., Fariselli P. & Casadio R. (2005) I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* 33, W306-10.
- [26] Carapito R., Radosavljevic M. & Bahram S. (2016) Next-Generation Sequencing of the HLA locus: Methods and impacts on HLA typing, population genetics and disease association studies. *Human Immunology* 77, 1016–1023.
- [27] Carbon S., Dietze H., Lewis S.E., Mungall C.J., Munoz-Torres M.C., Basu S., et al. (2017) Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research* 45, D331–D338.
- [28] Cariaso M. & Lennon G. (2012) SNPedia: A wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research* 40, D1308-12.
- [29] Cario H., Bode H., Gustavsson P., Dahl N. & Kohne E. (1999) A microdeletion syndrome due to a 3-Mb deletion on 19q13.2--Diamond-Blackfan anemia associated with macrocephaly, hypotonia, and psychomotor retardation. *Clinical genetics* 55, 487–92.
- [30] Cass C.L., Peraldi A., Dowd P.F., Mottiar Y., Santoro N., Karlen S.D., et al. (2015) Effects of PHENYLALANINE AMMONIA LYASE (PAL) knockdown on cell wall composition, biomass digestibility, and biotic and abiotic stress responses in *Brachypodium*. *Journal of Experimental Botany* 66, 4317–4335.
- [31] Chakraborty A., Uechi T., Nakajima Y., Gazda H.T., O'Donohue M.F., Gleizes P.E., et al. (2017) Cross talk between TP53 and c-Myc in the pathophysiology of Diamond-Blackfan anemia: Evidence from RPL11-deficient in vivo and in vitro models. *Biochemical and Biophysical Research Communications* 495, 1839–1845.

- [32] Cheng J., Randall A. & Baldi P. (2005) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics* 62, 1125–1132.
- [33] Chhangawala S., Rudy G., Mason C.E. & Rosenfeld J.A. (2015) The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology* 16, 131.
- [34] Chiabrando D., Marro S., Mercurio S., Giorgi C., Petrillo S., Vinchi F., et al. (2012) The mitochondrial heme exporter FLVCR1b mediates erythroid differentiation. *Journal of Clinical Investigation* 122, 4569–4579.
- [35] Chyra Kufova Z., Sevcikova T., Januska J., Vojta P., Boday A., Vanickova P., et al. (2018) Newly designed 11-gene panel reveals first case of hereditary amyloidosis captured by massive parallel sequencing. *Journal of Clinical Pathology*, jclinpath-2017-204978.
- [36] Cingolani P., Platts A., Wang L.L., Coon M., Nguyen T., Wang L., et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80–92.
- [37] Cmejlova J., Dolezalova L., Pospisilova D., Petrtlyova K., Petrak J. & Cmejla R. (2006) Translational efficiency in patients with Diamond-Blackfan anemia. *Haematologica* 91, 1456–1464.
- [38] Collart M.A., Baeuerle P. & Vassalli P. (1990) Regulation of tumor necrosis factor alpha transcription in macrophages: involvement of four kappa B-like motifs and of constitutive and inducible forms of NF-kappa B. *Molecular and cellular biology* 10, 1498–506.
- [39] Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M. & Robles M. (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676.
- [40] Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.
- [41] Danilova N., Sakamoto K.M. & Lin S. (2008) Ribosomal protein S19 deficiency in zebrafish leads to developmental abnormalities and defective erythropoiesis through activation of p53 protein family. *Blood* 112, 5228–5237.
- [42] Danilova N., Wilkes M., Bibikova E., Youn M.Y., Sakamoto K.M. & Lin S. (2018) Innate immune system activation in zebrafish and cellular models of Diamond Blackfan Anemia. *Scientific Reports* 8, 5165.
- [43] Deisenroth C. & Zhang Y. (2010) Ribosome biogenesis surveillance: Probing the ribosomal protein-Mdm2-p53 pathway. *Oncogene* 29, 4253–4260.
- [44] DePristo M.A., Banks E., Poplin R., Garimella K. V., Maguire J.R., Hartl C., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43, 491–8.
- [45] Diao Y., Yang C., Yan M., Zheng X., Jin S., Wang Y., et al. (2014) De novo transcriptome and small RNA analyses of two *Amorphophallus* species. *PLoS ONE* 9, e95428.
- [46] Van Dijk E.L., Jaszczyszyn Y. & Thermes C. (2014) Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* 322, 12–20.
- [47] DiNardo C.D., Routbort M.J., Bannon S.A., Benton C.B., Takahashi K., Kornblau S.M., et al. (2018) Improving the detection of patients with inherited predispositions to hematologic malignancies using next-generation sequencing-based leukemia prognostication panels. *Cancer*.
- [48] Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., et al. (2013) STAR: Ultrafast universal

RNA-seq aligner. *Bioinformatics* 29, 15–21.

- [49] Donati G., Montanaro L. & Derenzini M. (2012) Ribosome biogenesis and control of cell proliferation: p53 is not alone. *Cancer Research* 72, 1602–1607.
- [50] Dressman D., Yan H., Traverso G., Kinzler K.W. & Vogelstein B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences* 100, 8817–8822.
- [51] Dunn P., Albury C.L., Maksemous N., Benton M.C., Sutherland H.G., Smith R.A., et al. (2018) Next generation sequencing methods for diagnosis of epilepsy syndromes. *Frontiers in Genetics* 9, 20.
- [52] Ebbert M.T.W., Wadsworth M.E., Staley L.A., Hoyt K.L., Pickett B., Miller J., et al. (2016) Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* 17, 239.
- [53] Eden E., Navon R., Steinfeld I., Lipson D. & Yakhini Z. (2009) GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10, 48.
- [54] Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323, 133–138.
- [55] Errichiello E., Vetro A., Mina T., Wischmeijer A., Berrino E., Carella M., et al. (2017) Whole exome sequencing in the differential diagnosis of Diamond-Blackfan anemia: Clinical and molecular study of three patients with novel RPL5 and mosaic RPS19 mutations. *Blood Cells, Molecules, and Diseases* 64, 38–44.
- [56] Fabregat A., Jupe S., Matthews L., Sidiropoulos K., Gillespie M., Garapati P., et al. (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Research* 46, D649–D655.
- [57] Farrar J.E. & Dahl N. (2011) Untangling the phenotypic heterogeneity of diamond Blackfan anemia. *Seminars in Hematology* 48, 124–135.
- [58] Feussner I. & Wasternack C. (2002) THE LIPOXYGENASE PATHWAY. *Annual Review of Plant Biology* 53, 275–297.
- [59] Finn R.D., Clements J. & Eddy S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 39, W29–37.
- [60] Finn R.D., Coggill P., Eberhardt R.Y., Eddy S.R., Mistry J., Mitchell A.L., et al. (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research* 44, D279–D285.
- [61] Fokkema I.F.A.C., Taschner P.E.M., Schaafsma G.C.P., Celli J., Laros J.F.J. & den Dunnen J.T. (2011) LOVD v.2.0: The next generation in gene variant databases. *Human Mutation* 32, 557–563.
- [62] Galichet A., Hoyerova K., Kaminek M. & Gruissem W. (2008) Farnesylation Directs AtIPT3 Subcellular Localization and Modulates Cytokinin Biosynthesis in Arabidopsis. *PLANT PHYSIOLOGY* 146, 1155–1164.
- [63] Gan S. & Amasino R.M. (1995) Inhibition of Leaf Senescence by Autoregulated Production of Cytokinin. *Science* 270, 1986–1988.
- [64] Garrison E. & Marth G. (2012) Haplotype-based variant detection from short-read sequencing. *arxiv:1207.3907*.
- [65] Gazda H.T., Sheen M.R., Vlachos A., Choesmel V., O'Donohue M.F., Schneider H., et al. (2008)

Ribosomal Protein L5 and L11 Mutations Are Associated with Cleft Palate and Abnormal Thumbs in Diamond-Blackfan Anemia Patients. *American Journal of Human Genetics* 83, 769–780.

- [66] Gerrard G., Valgañón M., Foong H.E., Kasperaviciute D., Iskander D., Game L., et al. (2013) Target enrichment and high-throughput sequencing of 80 ribosomal protein genes to identify mutations associated with Diamond-Blackfan anaemia. *British Journal of Haematology* 162, 530–536.
- [67] Ghanem M.E., Albacete A., Smigocki A.C., Frébort I., Pospíšilová H., Martínez-Andújar C., et al. (2011) Root-synthesized cytokinins improve shoot growth and fruit yield in salinized tomato (*Solanum lycopersicum* L.) plants. *Journal of Experimental Botany* 62, 125–140.
- [68] Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., et al. (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* 15, 1451–1455.
- [69] Gibbs R.A., Boerwinkle E., Doddapaneni H., Han Y., Korchina V., Kovar C., et al. (2015) A global reference for human genetic variation. *Nature* 526, 68–74.
- [70] Ginsberg D. (2002) E2F1 pathways to apoptosis. *FEBS Letters* 529, 122–125.
- [71] Goldstein S. & Samuni A. (2005) Intra- and intermolecular oxidation of oxymyoglobin and oxyhemoglobin induced by hydroxyl and carbonate radicals. *Free Radical Biology and Medicine* 39, 511–519.
- [72] Goudarzi K.M. & Lindström M.S. (2016) Role of ribosomal protein mutations in tumor development (Review). *International Journal of Oncology* 48, 1313–1324.
- [73] Griffiths J.A., Scialdone A. & Marioni J.C. (2018) Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology* 14, e8046.
- [74] Gripp K.W., Curry C., Olney A.H., Sandoval C., Fisher J., Chong J.X.L., et al. (2014) Diamond-Blackfan anemia with mandibulofacial dystostosis is heterogeneous, including the novel DBA genes TSR2 and RPS28. *American Journal of Medical Genetics, Part A* 164, 2240–2249.
- [75] Gustavsson P., Willig T.N., Haeringen A. V., Tchernia G., Dianzani I., Donner M., et al. (1997) Diamond-blackfan anaemia: Genetic homogeneity for a gene on chromosome 19q13 restricted to 1.8 Mb. *Nature Genetics* 16, 368–371.
- [76] Ha S., Vankova R., Yamaguchi-Shinozaki K., Shinozaki K. & Tran L.S.P. (2012) Cytokinins: Metabolism and function in plant adaptation to environmental stresses. *Trends in Plant Science* 17, 172–179.
- [77] Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J., et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8, 1494–1512.
- [78] Hasan M.S. habbi., Wu X. & Zhang L. (2015) Performance evaluation of indel calling tools using real short-read data. *Human genomics* 9, 20.
- [79] Havlova M., Dobrev P.I., Motyka V., Storchova H., Libus J., Dobr?? J., et al. (2008) The role of cytokinins in responses to water deficit in tobacco plants over-expressing trans-zeatin O-glucosyltransferase gene under 35S or SAG12 promoters. *Plant, Cell and Environment* 31, 341–353.
- [80] Heijnen H.F., van Wijk R., Pereboom T.C., Goos Y.J., Seinen C.W., van Oirschot B.A., et al. (2014) Ribosomal Protein Mutations Induce Autophagy through S6 Kinase Inhibition of the Insulin Pathway. *PLoS Genetics* 10, e1004371.
- [81] Huang D.W., Sherman B.T. & Lempicki R.A. (2009) Systematic and integrative analysis of large gene lists

using DAVID bioinformatics resources. *Nature Protocols* 4, 44–57.

- [82] Huber W., Carey V.J., Gentleman R., Anders S., Carlson M., Carvalho B.S., et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* 12, 115–121.
- [83] Hunter S., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., et al. (2009) InterPro: The integrative protein signature database. *Nucleic Acids Research* 37, D211–5.
- [84] Ichimura T., Yoshida K., Okuno Y., Yujiri T., Nagai K., Nishi M., et al. (2017) Diagnostic challenge of Diamond-Blackfan anemia in mothers and children by whole-exome sequencing. *International journal of hematology* 105, 515–520.
- [85] Idury R.M. & Waterman M.S. (1995) A New Algorithm for DNA Sequence Assembly. *Journal of Computational Biology* 2, 291–306.
- [86] Ikeda F., Yoshida K., Toki T., Uechi T., Ishida S., Nakajima Y., et al. (2016) Exome sequencing identified RPS15A as a novel causative gene for Diamond-Blackfan anemia. *Haematologica* 102, e93–e96.
- [87] Itan Y., Shang L., Boisson B., Patin E., Bolze A., Moncada-Vélez M., et al. (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences* 112, 13615–13620.
- [88] Jelin A.C. & Vora N. (2018) Whole Exome Sequencing: Applications in Prenatal Genetics. *Obstetrics and Gynecology Clinics of North America* 45, 69–81.
- [89] Jennings L.J., Arcila M.E., Corless C., Kamel-Reid S., Lubin I.M., Pfeifer J., et al. (2017) Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *Journal of Molecular Diagnostics* 19, 341–365.
- [90] Jia Q., Zhang Q., Zhang Z., Wang Y., Zhang W., Zhou Y., et al. (2013) Transcriptome analysis of the zebrafish model of Diamond-Blackfan anemia from RPS19 deficiency via p53-dependent and -independent pathways. *PloS one* 8, e71782.
- [91] Joslin J.M., Fernald A.A., Tennant T.R., Davis E.M., Kogan S.C., Anastasi J., et al. (2007) Haploinsufficiency of EGR1, a candidate gene in the del(5q), leads to the development of myeloid disorders. *Blood* 110, 719–726.
- [92] Keeton A.B., Bortoff K.D., Bennett W.L., Franklin J.L., Venable D.Y. & Messina J.L. (2003) Insulin-Regulated Expression of Egr-1 and Krox20: Dependence on ERK1/2 and Interaction with p38 and PI3-Kinase Pathways. *Endocrinology* 144, 5402–5410.
- [93] Kim D., Langmead B. & Salzberg S.L. (2015) HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* 12, 357–360.
- [94] Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R. & Salzberg S.L. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14, R36.
- [95] Kinde I., Wu J., Papadopoulos N., Kinzler K.W. & Vogelstein B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences* 108, 9530–9535.
- [96] Kinsella R.J., Kähäri A., Haider S., Zamora J., Proctor G., Spudich G., et al. (2011) Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database* 2011, bar030–bar030.

- [97] Koboldt D.C., Zhang Q., Larson D.E., Shen D., McLellan M.D., Lin L., et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22, 568–576.
- [98] Kohany O., Gentles A.J., Hankus L. & Jurka J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7, 474.
- [99] Köhler S., Vasilevsky N.A., Engelstad M., Foster E., McMurry J., Aymé S., et al. (2017) The human phenotype ontology in 2017. *Nucleic Acids Research* 45, D865–D876.
- [100] Kokas F., Vojta P. & Galuszka P. (2016) Dataset for transcriptional response of barley (*Hordeum vulgare*) exposed to drought and subsequent re-watering. *Data in Brief* 8, 334–341.
- [101] Krumm N., Sudmant P.H., Ko A., O'Rourke B.J., Malig M., Coe B.P., et al. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome Research* 22, 1525–1532.
- [102] Kufova Z., Sevcikova T., Growkova K., Vojta P., Filipova J., Adam Z., et al. (2017) Biomarkers in immunoglobulin light chain amyloidosis. *Klinicka Onkologie* 30.
- [103] Kuppu S., Mishra N., Hu R., Sun L., Zhu X., Shen G., et al. (2013) Water-Deficit Inducible Expression of a Cytokinin Biosynthetic Gene IPT Improves Drought Tolerance in Cotton. *PLoS ONE* 8, e64190.
- [104] Lai Z., Markovets A., Ahdesmaki M., Chapman B., Hofmann O., Mcewen R., et al. (2016) VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research* 44.
- [105] Landrum M.J., Lee J.M., Riley G.R., Jang W., Rubinstein W.S., Church D.M., et al. (2014) ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* 42, D980–5.
- [106] Langmead B., Trapnell C., Pop M. & Salzberg S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10, R25.
- [107] Lapin V., Mighion L.C., da Silva C.P., Cuperus Y., Bean L.J.H. & Hegde M.R. (2016) Regulating whole exome sequencing as a diagnostic test. *Human Genetics* 135, 655–673.
- [108] Leek J.T., Johnson W.E., Parker H.S., Jaffe A.E. & Storey J.D. (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883.
- [109] Lek M., Karczewski K.J., Minikel E. V., Samocha K.E., Banks E., Fennell T., et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291.
- [110] Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.
- [111] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- [112] Li Y.H., Yu C.Y., Li X.X., Zhang P., Tang J., Yang Q., et al. (2018) Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research* 46, D1121–D1127.
- [113] Liao Y., Smyth G.K. & Shi W. (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930.
- [114] Lin Y., Golovnina K., Chen Z.-X., Lee H.N., Negron Y.L.S., Sultana H., et al. (2016) Comparison of

normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics* 17, 28.

- [115] Liu L., Li Y., Li S., Hu N., He Y., Pong R., et al. (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* 2012, 1–11.
- [116] Liu X., Wu C., Li C. & Boerwinkle E. (2016) dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human Mutation* 37, 235–241.
- [117] Love M.I., Huber W. & Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15, 550.
- [118] Majeska-Cudejkova M., Vojta P., Valarik J. & Galuszka P. (2016) Quantitative and qualitative transcriptome analysis of four industrial strains of *Claviceps purpurea* with respect to ergot alkaloid production. *New Biotechnology* 33, 743–754.
- [119] Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.
- [120] Matthijs G., Souche E., Alders M., Corveleyn A., Eck S., Feenstra I., et al. (2016) Guidelines for diagnostic next-generation sequencing. *European Journal of Human Genetics* 24, 2–5.
- [121] McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., et al. (2010) The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303.
- [122] McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R.S., Thormann A., et al. (2016) The Ensembl Variant Effect Predictor. *Genome Biology* 17.
- [123] Medina I., Carbonell J., Pulido L., Madeira S.C., Goetz S., Conesa A., et al. (2010) Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Research* 38, W210–W213.
- [124] Merewitz E.B., Gianfagna T. & Huang B. (2011) Photosynthesis, water use, and root viability under water stress as affected by expression of SAG12-ipt controlling cytokinin synthesis in *Agrostis stolonifera*. *Journal of Experimental Botany* 62, 383–395.
- [125] Metzker M.L. (2010) Sequencing technologies the next generation. *Nature Reviews Genetics* 11, 31–46.
- [126] Moon J., Parry G. & Estelle M. (2004) The ubiquitin-proteasome pathway and plant development. *The Plant cell* 16, 3181–3195.
- [127] Moortgat S., Lederer D., Deprez M., Buzatu M., Clapuyt P., Boulanger S., et al. (2018) Expanding the phenotypic spectrum associated with OPHN1 mutations: Report of 17 individuals with intellectual disability but no cerebellar hypoplasia. *European Journal of Medical Genetics*.
- [128] Mortazavi A., Williams B.A., McCue K., Schaeffer L. & Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5, 621–628.
- [129] Narasimhan V., Danecek P., Scally A., Xue Y., Tyler-Smith C. & Durbin R. (2016) BCFTools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32, 1749–1751.
- [130] Nishi H., Inagi R., Kato H., Tanemoto M., Kojima I., Son D., et al. (2008) Hemoglobin Is Expressed by Mesangial Cells and Reduces Oxidant Stress. *Journal of the American Society of Nephrology* 19, 1500–

- [131] Nishiyama R., Watanabe Y., Fujita Y., Le D.T., Kojima M., Werner T., et al. (2011) Analysis of Cytokinin Mutants and Regulation of Cytokinin Metabolic Genes Reveals Important Regulatory Roles of Cytokinins in Drought, Salt and Abscissic Acid Responses, and Abscissic Acid Biosynthesis. *The Plant Cell* 23, 2169–2183.
- [132] Nishiyama R., Watanabe Y., Leyva-Gonzalez M.A., Van Ha C., Fujita Y., Tanaka M., et al. (2013) Arabidopsis AHP2, AHP3, and AHP5 histidine phosphotransfer proteins function as redundant negative regulators of drought stress response. *Proceedings of the National Academy of Sciences* 110, 4840–4845.
- [133] O'Brien J.A. & Benková E. (2013) Cytokinin cross-talking during biotic and abiotic stress responses. *Frontiers in Plant Science* 4, 451.
- [134] O'Brien K.A., Seidel N.E., Farrar J., Vlachos A., Anderson S.M., An X., et al. (2013) Autosomal recessive diamond-blackfan anemia: Identification of mutations in MCM2 and FLNB. In: *Blood*. .
- [135] Ogata H., Goto S., Sato K., Fujibuchi W., Bono H. & Kanehisa M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 27, 29–34.
- [136] Okonechnikov K., Golosova O. & Fursov M. (2012) Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167.
- [137] Oono Y., Kobayashi F., Kawahara Y., Yazawa T., Handa H., Itoh T., et al. (2013) Characterisation of the wheat (*triticum aestivum* L.) transcriptome by de novo assembly for the discovery of phosphate starvation-responsive genes: Gene expression in Pi-stressed wheat. *BMC Genomics* 14, 77.
- [138] Peleg Z. & Blumwald E. (2011) Hormone balance and abiotic stress tolerance in crop plants. *Current Opinion in Plant Biology* 14, 290–295.
- [139] Peleg Z., Reguera M., Tumimbang E., Walia H. & Blumwald E. (2011) Cytokinin-mediated source/sink modifications improve drought tolerance and increase grain yield in rice under water-stress. *Plant Biotechnology Journal* 9, 747–758.
- [140] Pertea M., Kim D., Pertea G.M., Leek J.T. & Salzberg S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature protocols* 11, 1650–1667.
- [141] Pertea M., Pertea G.M., Antonescu C.M., Chang T.C., Mendell J.T. & Salzberg S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 33, 290–295.
- [142] Petrovski S., Wang Q., Heinzen E.L., Allen A.S. & Goldstein D.B. (2013) Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics* 9, e1003709.
- [143] Pevzner P.A., Tang H. & Waterman M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* 98, 9748–9753.
- [144] Plagnol V., Curtis J., Epstein M., Mok K.Y., Stebbings E., Grigoriadou S., et al. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 28, 2747–2754.
- [145] Plutino M., Chausseot A., Rouzier C., Ait-El-Mkadem S., Fragaki K., Paquis-Flucklinger V., et al. (2018) Targeted next generation sequencing with an extended gene panel does not impact variant detection in

mitochondrial diseases. *BMC Medical Genetics* 19, 57.

- [146] Pospisilova D., Cmejlova J., Ludikova B., Stary J., Cerna Z., Hak J., et al. (2012) The Czech National Diamond-Blackfan Anemia Registry: Clinical data and ribosomal protein mutations update. *Blood Cells, Molecules, and Diseases* 48, 209–218.
- [147] Pospíšilová H., Jiskrová E., Vojta P., Mrázová K., Kokáš F., Čudejková M.M., et al. (2016) Transgenic barley overexpressing a cytokinin dehydrogenase gene shows greater tolerance to drought stress. *New Biotechnology* 33, 692–705.
- [148] Quinlan A.R. & Hall I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.
- [149] Reguera M., Peleg Z., Abdel-Tawab Y.M., Tumimbang E.B., Delatorre C.A. & Blumwald E. (2013) Stress-Induced Cytokinin Synthesis Increases Drought Tolerance through the Coordinated Regulation of Carbon and Nitrogen Assimilation in Rice. *PLANT PHYSIOLOGY* 163, 1609–1622.
- [150] Rehrauer H., Opitz L., Tan G., Sieverling L. & Schlapbach R. (2013) Blind spots of quantitative RNA-seq: The limits for assessing abundance, differential expression, and isoform switching. *BMC Bioinformatics* 14, 370.
- [151] Richards S., Aziz N., Bale S., Bick D., Das S., Gastier-Foster J., et al. (2015) Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* 17, 405–424.
- [152] Van Riggelen J. & Felsher D.W. (2010) Myc and a Cdk2 senescence switch. *Nature Cell Biology* 12, 7–9.
- [153] Riley M.F. & Lozano G. (2012) The Many Faces of MDM2 Binding Partners. *Genes and Cancer* 3, 226–239.
- [154] Rimmer A., Phan H., Mathieson I., Iqbal Z., Twigg S.R.F., Wilkie A.O.M., et al. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* 46, 912–918.
- [155] Rippert P. & Matringe M. (2002) Molecular and biochemical characterization of an *Arabidopsis thaliana* arogenate dehydrogenase with two. *Plant molecular biology* 48, 361–368.
- [156] Risso D., Ngai J., Speed T.P. & Dudoit S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* 32, 896–902.
- [157] Rivero R.M., Gimeno J., Van Deynze A., Walia H. & Blumwald E. (2010) Enhanced cytokinin synthesis in tobacco plants expressing PSARK::IPT Prevents the degradation of photosynthetic protein complexes during drought. *Plant and Cell Physiology* 51, 1929–1941.
- [158] Rivero R.M., Kojima M., Gepstein A., Sakakibara H., Mittler R., Gepstein S., et al. (2007) Delayed leaf senescence induces extreme drought tolerance in a flowering plant. *Proceedings of the National Academy of Sciences* 104, 19631–19636.
- [159] Robinson M.D., McCarthy D.J. & Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26, 139–40.
- [160] Roggero S., Quarello P., Vinciguerra T., Longo F., Piga A. & Ramenghi U. (2009) Severe iron overload in Blackfan-Diamond anemia: A case-control study. *American Journal of Hematology* 84, 729–732.

- [161] Rothberg J.M., Hinz W., Rearick T.M., Schultz J., Mileski W., Davey M., et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475, 348–352.
- [162] Russell J.L., Powers J.T., Rounbehler R.J., Rogers P.M., Conti C.J. & Johnson D.G. (2002) ARF differentially modulates apoptosis induced by E2F1 and Myc. *Molecular and cellular biology* 22, 1360–1368.
- [163] Ruzicka K., Ljung K., Vanneste S., Podhorska R., Beeckman T., Friml J., et al. (2007) Ethylene Regulates Root Growth through Effects on Auxin Biosynthesis and Transport-Dependent Auxin Distribution. *THE PLANT CELL ONLINE* 19, 2197–2212.
- [164] Sandmann S., De Graaf A.O., Karimi M., Van Der Reijden B.A., Hellström-Lindberg E., Jansen J.H., et al. (2017) Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports* 7, 43169.
- [165] Sankaran V.G., Ghazvinian R., Do R., Thiru P., Vergilio J.A., Beggs A.H., et al. (2012) Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *Journal of Clinical Investigation* 122, 2439–2443.
- [166] Shendure J., Porreca G.J., Reppas N.B., Lin X., McCutcheon J.P., Rosenbaum A.M., et al. (2005) Molecular biology: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732.
- [167] Sherry S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29, 308–311.
- [168] Smetanina N.S., Mersiyanova I. V., Kurnikova M.A., Ovsyannikova G.S., Hachatrian L.A., Bobrynnina V.O., et al. (2015) Clinical and genomic heterogeneity of Diamond Blackfan anemia in the Russian Federation. *Pediatric Blood and Cancer* 62, 1597–1600.
- [169] Song B., Zhang Q., Zhang Z., Wan Y., Jia Q., Wang X., et al. (2014) Systematic transcriptome analysis of the zebrafish model of diamond-blackfan anemia induced by RPS24 deficiency. *BMC Genomics* 15, 759.
- [170] Stanke M., Keller O., Gunduz I., Hayes A., Waack S. & Morgenstern B. (2006) AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Research* 34, W435-9.
- [171] Sun L., Zhang Q., Wu J., Zhang L., Jiao X., Zhang S., et al. (2014) Two Rice Authentic Histidine Phosphotransfer Proteins, OsAHP1 and OsAHP2, Mediate Cytokinin Signaling and Stress Responses in Rice. *PLANT PHYSIOLOGY* 165, 335–345.
- [172] Sýkorová B., Kurešová G., Daskalova S., Trčková M., Hoyerová K., Raimanová I., et al. (2008) Senescence-induced ectopic expression of the *A. tumefaciens* *ipt* gene in wheat delays leaf senescence, increases cytokinin content, nitrate influx, and nitrate reductase activity, but does not affect grain yield. *Journal of Experimental Botany* 59, 377–387.
- [173] Teixeira M.R. (2006) Recurrent fusion oncogenes in carcinomas. *Critical reviews in oncogenesis* 12, 257–271.
- [174] Thorvaldsdóttir H., Robinson J.T. & Mesirov J.P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14, 178–192.
- [175] Tian D., Tooker J., Peiffer M., Chung S.H. & Felton G.W. (2012) Role of trichomes in defense against herbivores: Comparison of herbivore response to woolly and hairless trichome mutants in tomato (*Solanum lycopersicum*). *Planta* 236, 1053–1066.

- [176] Tian S., Yan H., Kalmbach M. & Slager S.L. (2016) Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics* 17, 403.
- [177] Tran L.-S.P., Urao T., Qin F., Maruyama K., Kakimoto T., Shinozaki K., et al. (2007) Functional analysis of AHK1/ATHK1 and cytokinin receptor histidine kinases in response to abscisic acid, drought, and salt stress in *Arabidopsis*. *Proceedings of the National Academy of Sciences* 104, 20623–20628.
- [178] Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D.R., et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7, 562–578.
- [179] Travers K.J., Chin C.S., Rank D.R., Eid J.S. & Turner S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* 38, e159–e159.
- [180] Tsuchisaka A. & Theologis A. (2004) Unique and Overlapping Expression Patterns among the *Arabidopsis* 1-Amino-Cyclopropane-1-Carboxylate Synthase Gene Family Members. *Plant Physiology* 136, 2982–3000.
- [181] Tulpová Z., Luo M.C., Toegelová H., Visendi P., Hayashi S., Vojta P., et al. (2018) Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. *New Biotechnology*.
- [182] Ueda J. & Kato J. (1982) Identification of jasmonic acid and abscisic acid as senescence-promoting substances from *Cleome ochracea* dc. *Agricultural and Biological Chemistry* 46, 1975–1976.
- [183] Utsugisawa T., Uchiyama T., Toki T., Ogura H., Aoki T., Hamaguchi I., et al. (2016) Erythrocyte glutathione is a novel biomarker of Diamond-Blackfan anemia. *Blood Cells, Molecules, and Diseases* 59, 31–36.
- [184] Virolle T., Adamson E.D., Baron V., Birlé D., Mercola D., Mustelin T., et al. (2001) The Egr-1 transcription factor directly activates PTEN during irradiation-induced signalling. *Nature Cell Biology* 3, 1124–1128.
- [185] Vlachos A., Ball S., Dahl N., Alter B.P., Sheth S., Ramenghi U., et al. (2008) Diagnosing and treating Diamond Blackfan anaemia: Results of an international clinical consensus conference. *British Journal of Haematology* 142, 859–876.
- [186] Vlamis J. & Williams D.E. (1962) Ion Competition in Manganese Uptake by Barley Plants. *PLANT PHYSIOLOGY* 37, 650–655.
- [187] Voelkerding K. V., Dames S. & Durtschi J.D. (2010) Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: A Paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. *Journal of Molecular Diagnostics* 12, 539–551.
- [188] Vojta P., Kokáš F., Husíčková A., Grúz J., Bergougnoux V., Marchetti C.F., et al. (2016) Whole transcriptome analysis of transgenic barley with altered cytokinin homeostasis and increased tolerance to drought stress. *New Biotechnology* 33, 676–691.
- [189] Volejníková J., Zapletalová J., Jarosová M., Urbanková H., Vojta P., Klasková E., et al. (2018) Acute lymphoblastic leukemia in a child with Ieri-Weill syndrome and complete SHOX gene deletion: A case report. *Biomedical Papers* 162, 65–70.
- [190] Vyroubalová S., Vaclavíková K., Turecková V., Novák O., Smehilová M., Hluska T., et al. (2009) Characterization of New Maize Genes Putatively Involved in Cytokinin Metabolism and Their Expression

during Osmotic Stress in Relation to Cytokinin Levels. *PLANT PHYSIOLOGY* 151, 433–447.

- [191] Wagner G.P., Kin K. & Lynch V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* 131, 281–285.
- [192] Wan Y., Zhang Q., Zhang Z., Song B., Wang X., Zhang Y., et al. (2016) Transcriptome analysis reveals a ribosome constituents disorder involved in the RPL5 downregulated zebrafish model of Diamond-Blackfan anemia. *BMC Medical Genomics* 9, 13.
- [193] Wang C., Yang A., Yin H. & Zhang J. (2008) Influence of water stress on endogenous hormone contents and cell damage of maize seedlings. *Journal of Integrative Plant Biology* 50, 427–434.
- [194] Wang K., Li M. & Hakonarson H. (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 38.
- [195] Wang L., Wang Z., Xu Y., Joo S.H., Kim S.K., Xue Z., et al. (2009) OsGSR1 is involved in crosstalk between gibberellins and brassinosteroids in rice. *Plant Journal* 57, 498–510.
- [196] Warden C.D., Adamson A.W., Neuhausen S.L. & Wu X. (2014) Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ* 2, e600.
- [197] Wasternack C. (2014) Action of jasmonates in plant stress responses and development - Applied aspects. *Biotechnology Advances* 32, 31–39.
- [198] Watkins-Chow D.E., Cooke J., Pidsley R., Edwards A., Slotkin R., Leeds K.E., et al. (2013) Mutation of the Diamond-Blackfan Anemia Gene *Rps7* in Mouse Results in Morphological and Neuroanatomical Phenotypes. *PLoS Genetics* 9, e1003094.
- [199] Wei W., Alexandersson E., Golldack D., Miller A.J., Kjellbom P.O. & Fricke W. (2007) HvPIP1;6, a barley (*Hordeum vulgare* L.) plasma membrane water channel particularly expressed in growing compared with non-growing leaf tissues. *Plant and Cell Physiology* 48, 1132–1147.
- [200] Weiss M.M., Van der Zwaag B., Jongbloed J.D.H., Vogel M.J., Brüggewirth H.T., Lekan Deprez R.H., et al. (2013) Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: A national collaborative study of dutch genome diagnostic laboratories. *Human Mutation* 34, 1313–1321.
- [201] Willig T.N., Draptchinskaia N., Dianzani I., Ball S., Niemeyer C., Ramenghi U., et al. (1999) Mutations in ribosomal protein S19 gene and diamond blackfan anemia: wide variations in phenotypic expression. *Blood* 94, 4294–4306.
- [202] Xuan J., Yu Y., Qing T., Guo L. & Shi L. (2013) Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters* 340, 284–295.
- [203] Ye C., Hill C.M., Wu S., Ruan J. & Ma Z. (2016) DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific Reports* 6, 31900.
- [204] Zalabák D., Pospíšilová H., Šmehilová M., Mrázová K., Frébort I. & Galuszka P. (2013) Genetic engineering of cytokinin metabolism: Prospective way to improve agricultural traits of crop plants. *Biotechnology Advances* 31, 97–117.
- [205] Zatopkova M., Filipova J., Jelinek T., Vojta P., Sevcikova T., Simicek M., et al. (2017) Whole exome sequencing of aberrant plasma cells in a patient with multiple myeloma minimal residual disease. *Klinicka Onkologie* 30.

- [206] Zerbino D.R., Achuthan P., Akanni W., Amode M.R., Barrell D., Bhai J., et al. (2018) Ensembl 2018. *Nucleic Acids Research* 46, D754–D761.
- [207] Zerbino D.R. & Birney E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18, 821–829.
- [208] Zhang S.-W., Li C.-H., Cao J., Zhang Y.-C., Zhang S.-Q., Xia Y.-F., et al. (2009) Altered Architecture and Enhanced Drought Tolerance in Rice via the Down-Regulation of Indole-3-Acetic Acid by TLD1/OsGH3.13 Activation. *PLANT PHYSIOLOGY* 151, 1889–1901.
- [209] Zhang Z., Jia H., Zhang Q., Wan Y., Zhou Y., Jia Q., et al. (2013) Assessment of hematopoietic failure due to Rpl11 deficiency in a zebrafish model of Diamond-Blackfan anemia by deep sequencing. *BMC Genomics* 14, 896.
- [210] Zhang Z., Theurkauf W.E., Weng Z. & Zamore P.D. (2012) Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. *Silence* 3, 9.
- [211] Zook J.M., Chapman B., Wang J., Mittelman D., Hofmann O., Hide W., et al. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology* 32, 246–251.
- [212] Zubo Y.O., Yamburenko M. V., Selivankina S.Y., Shakirova F.M., Avalbaev A.M., Kudryakova N. V., et al. (2008) Cytokinin Stimulates Chloroplast Transcription in Detached Barley Leaves. *PLANT PHYSIOLOGY* 148, 1082–1093.

10 Přílohy

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	[Header]		5															
2	Investigator Version	P2																
3	Investigator Name	S295-180104-113																
4	Experiment Name	15.2.2018																
5	Date	GenerateFASTQ																
6	Workflow	HiSeq FASTQ Only																
7	Application	HiSeq 1500/2500																
8	Instrument Type	TruSeq DNA Exome Enrichment																
9	Assay	TruSeq DNA CD Indexes (96 Indexes)																
10	Index Adapters	WES-TruSeq DNA Exome																
11	Description	Angiogen																
12	Chemistry																	
13																		
14	[Reads]																	
15		101																
16		101																
17																		
18	[Settings]																	
19	ReverseComplement	0																
20	Adapter	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA																
21	AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT																
22																		
23	[Data]																	
24	Lane	Sample_ID	Sample_Name	Index_Plate_Well	I7_Index_ID	Index	I5_Index_ID	Index2	Sample_Proj	Description	Material	Phenotype	Platform	Platform_unit	Library	Barcode		
25		1	test_sample_01	MG10045	A01	D701	ATTACTCG	D501	TATAGCCT	WES	WES	MG10045	GO-0042254-HP-0001903	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
26		2	test_sample_02	MG10046	B01	D701	ATTACTCG	D502	ATAGAGGC	WES	WES	MG10046	HP-0012234	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
27		3	test_sample_03	MG10047	C01	D701	ATTACTCG	D503	CCTATGCT	WES	WES	MG10047	HP-0002086	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
28		4	test_sample_04	MG10048	D01	D701	ATTACTCG	D504	GGCTCTGA	WES	WES	MG10048	HP-0004322-HP-0040064-HP-0001903	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
29		1	test_sample_01	MG10045	E01	D701	ATTACTCG	D505	AGGCGAAG	WES	WES	MG10045	GO-0042254-HP-0001903	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
30		2	test_sample_02	MG10046	F01	D701	ATTACTCG	D506	TATCTTGA	WES	WES	MG10046	HP-0012234	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
31		3	test_sample_03	MG10047	G01	D701	ATTACTCG	D507	CAGGACGT	WES	WES	MG10047	HP-0002086	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	
32		4	test_sample_04	MG10048	H01	D701	ATTACTCG	D508	GTACTGAC	WES	WES	MG10048	HP-0004322-HP-0040064-HP-0001903	ILLUMINA	HARMVADXX	WES	150330_D00183_0031_AHARMVADXX	

Příloha 3.1: Upravená verze průvodní tabulky („Sample sheet“) pro analýzu v anotačním programu a v platformě MOLDIMED

Modře a červeně jsou znázorněny proprietární sloupce vyžadovány anotačním programem, přičemž červeně zvýrazněný sloupec je definován pro průvodní dokument k sekvenačním běhům na sekvenačních platformách Illumina.

```

#Form
form_folder      MOLDIMED/templates/protokol_html/properties
form_template    MOLDIMED/templates/protokol_html/template.html

#Genome
genome_default   MOLDIMED/Annot_DB/GRCH38_v83/input_files/Homo_sapiens.GRCh38.dna.primary_assembly_titled.fa
output_default   ./metadata
output_folder_default complete_annotation
filtered_folder  filtered_annotation
clinical_folder  clinical_annotation
html_output      html

ncpu              32
FeatureList_db_default MOLDIMED/Annot_DB/GRCH38_v83/vo_db/FeaturesList.annotdb_db
GeneList_db_default   MOLDIMED/Annot_DB/GRCH38_v83/vo_db/GeneList.annotdb
vcf_db_default         MOLDIMED/Annot_DB/GRCH38_v83/vo_db/vcf
gff_db_default         MOLDIMED/Annot_DB/GRCH38_v83/vo_db/gff
GO_list_file_default   MOLDIMED/Annot_DB/GRCH38_v83/vo_db/GO_assoc.annotdb
HP0_list_file_default  MOLDIMED/Annot_DB/GRCH38_v83/vo_db/HP0_assoc.annotdb
GO_obo_file_default    MOLDIMED/Annot_DB/GRCH38_v83/vo_db/go-basic.obo
HP0_obo_file_default   MOLDIMED/Annot_DB/GRCH38_v83/vo_db/hp.obo
known_sites           MOLDIMED/Annot_DB/GRCH38_v83/input_files/dbsnp_146.hg38.vcf.gz

####filtering criteria
####filtered in fields -> list define by field=mark, dbsnp=.,exac=.,...
obligatory_fields    clinvar_pathogenic,clinvar_likely_pathogenic
filtered_in_fields    dbsnp_common=.
included_fields       Sense=deletion,Sense=nonsynonymous,Sense=spliceAlteration,Sense=nonFS_insertion,Sense=FS_insertion
predictors_pos        predict_Polyphen2_HDIV,predict_Polyphen2_HVAR,predict_SIFT
predictors_tolerated_marks B,T

####clinical filtration
#pathogenic_output
pathogenic_field      clinvar_pathogenic
pathogenic_excluded_fields clinvar_conflicting_interpretations_of_pathogenicity,clinvar_benign
#freq_filter
frequency_field       dbSNP_CAF,ExAC_adj_AF
freq                  0.05
freq_excluded         dbSNP_common

####Form_filtering
form_table_conf_pathogenic      GeneName,HGV,ExFeature,ACh,Genotype,ExAC_adj_AF,clinvar_all,clinvar_disease
form_table_conf_searched       GeneName,HGV,ExFeature,ACh,Genotype,ExAC_adj_AF,dbSNP_all,Phenotype

```

Příloha 3.2: Konfigurační textový soubor pro anotační programu a MOLDIMED

Formát konfiguračního souboru v textovém formátu s vyplněnými přístupovými cestami pro databázové anotační zdroje a referenční soubory.

SampleSheet

Available files:

SampleSheet.csv

Load file

Load saved file

Add column option

Data type

string

Column name

Enter name of column

Add column

Remove column

	Phenotype	Platform	Platform_unit	Library	Material	Barcode	
	GO:0042254;HP:0001903	ILLUMINA	HARMVADXX	WES	MG10045	150330_D00183_0031_AHARMVADXX	
	HP:0012234	ILLUMINA	HARMVADXX	WES	MG10046	150330_D00183_0031_AHARMVADXX	
	HP:0002086	ILLUMINA	HARMVADXX	WES	MG10047	150330_D00183_0031_AHARMVADXX	
	HP:0004322;HP:0040064;HP:0001903	ILLUMINA	HARMVADXX	WES	MG10048	150330_D00183_0031_AHARMVADXX	

Pages: 1 1 of 1

Save file

Create folders

Příloha 3.3: GUI pro práci s průvodním dokumentem v aplikaci MOLDIMED 3
Grafické zpracování rozhraní pro zadání vstupních informací pro datovou analýzu.

Options

Off Remove duplicates

PHRED Mapping Quality

10

STRAND_FILTER

1

MIN_COVERAGE

10

MIN_READS2

4

MIN_AVG_QUAL

20

MIN_VAR_FREQ

0.3

MIN_FREQ_FOR_HOM

0.75

P_VALUE

0.99

VARIANTS

1

Run pipelineFastq

Stop pipelineFastq

Output files

clinical_annotation

complete_annotation

filtered_annotation

Metadata

select all

Prepare Output files

Delete Output files

Import Data to Database

Reports .html files

2756-17-A.html

2775-17-A.html

2780-17-A.html

2781-17-A.html

2782-17-A.html

280-18-A-I.html

2895-17-A.html

Příloha 3.4: GUI pro nastavení parametrů zpracování MPS dat v aplikaci MOLDIMED 3

Ukázka grafického zpracování pro úpravu parametrů optimalizace alignmentu a volání variant (na levé straně); uspořádání výstupních souborů v grafickém rozhraní (na pravé straně).

(A)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SampleX
19	49877730	.	GGAGCCGGCTACAGGGTAAGC	GGATCCGGCTACAGGCTAACC,G,CGAGCCGGCTACAGGGTAAGC	.	.	.	GT:GL:GOF:GQ:NR:NV	2 13:-14.59:0.0,-15.39:0.99:10:5
19	49878356	.	G	C	.	.	.	GT:GL:GOF:GQ:NR:NV	1/1:-14.59:0.0,-15.39:0.99:10:5

(B)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
19	49877730	rs_del(19:49877730_GGAGCCGGCTACAGGGTAAGC)	GGAGCCGGCTACAGGGTAAGC	G	.	.	.	.
19	49877733	rs(19:49877733_GCCGGCTACAGGGTAAG>TCCGGCTACAGGCTAAC)	GCCGGCTACAGGGTAAG	TCCGGCTACAGGCTAAC	.	.	.	.
19	49877745	rs_SNV(19:49877745_G>C)	G	C	.	.	.	.
19	49878356	rs_SNV(19:49878356_G>A)	G	A	.	.	.	.
19	49878356	rs_SNV(19:49878356_G>T,C)	G	T,C	.	.	.	.
19	49878356	rs_SNV(19:49878356_G>C)	G	C	.	.	.	.

(C)

#SeqName	Source	Feature	Start	End	Score	Strand	Frame	Attribute
19	annot	AKT1S1_gene	49869033	49878356	0	.	0	0
19	annot	left_aligned_region	49877730	49877730	0	.	0	0
19	annot	TBC1D17_spliceSite	49877745	49877746	0	+	0	0
19	annot	AKT1S1+TBC1D17_region(0)	49877745	49877746	0	0	0	0
19	annot	AKT1S1+TBC1D17_region(.)	49877745	49877746	0	.	0	0
19	annot	right_aligned_region	49877750	49877751	0	.	0	0

Příloha 3.5: Testovací vstupní soubory pro anotaci komplexních variant

Vstupní soubory pro evaluaci komplexní varianty; (A) testované varianty ve VCF zápisu, (B) VCF zápis pro pozíční anotaci, (C) GFF zápis pro intervalovou anotaci.

Příloha 3.6: Základní anotace komplexních variant programem VEP
Výstup základní anotace komplexní varianty programem VEP.

An	VCfdrh	VCfpos	HGV	Strand	Feature	ExFeature	Sense	ACh	GeneName
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	-	UTRS	UTRS	-	-	AKT1S1
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	-	intron	intron:1	-	-	AKT1S1
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	-	exon	exon:1	-	-	AKT1S1
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	+	UTRS	UTRS	-	-	TBC1D17
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	+	exon	exon:1	-	-	TBC1D17
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	+	CDS	CDS:1	-	3,2	GAGYR>del
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	+	intron	intron:1	-	-	TBC1D17
gene	19:49877730	gdel(19:49877731..49877750)	GAGCCGGGTACAGGGTAAGC]	+	spliceSite	spliceSite:1	-	-	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877730..G-C]	-	UTRS	UTRS	-	-	AKT1S1
gene	19:49877730	g19:49877730..G-C]	49877730..G-C]	-	intron	intron:1	-	-	AKT1S1
gene	19:49877730	g19:49877730..G-C]	49877730..G-C]	-	exon	exon:1	-	-	AKT1S1
gene	19:49877730	g19:49877730..G-C]	49877730..G-C]	+	UTRS	UTRS	-	-	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877730..G-C]	+	exon	exon:1	-	-	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877730..G-C]	+	CDS	CDS:1	nonsynonymous	3,1, G>R	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	UTRS	UTRS	-	-	AKT1S1
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	intron	intron:1	-	-	AKT1S1
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	exon	exon:1	-	-	AKT1S1
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	UTRS	UTRS	-	-	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	exon	exon:1	-	-	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	CDS	CDS:1	nonsynonymous	4,1	LAGYR>SGYR
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	intron	intron:1	-	-	TBC1D17
gene	19:49877730	g19:49877730..G-C]	49877749..GCCGGGTACAGGGTAAG-TCCGGGTACAGGGTAAC]	+	spliceSite	spliceSite:1	-	-	TBC1D17
gene	19:49878356	g19:49878356..G-C]	49878356..G-C]	-	UTRS	UTRS	-	-	AKT1S1
gene	19:49878356	g19:49878356..G-C]	49878356..G-C]	-	UTRS	UTRS	-	-	TBC1D17
gene	19:49878356	g19:49878356..G-C]	49878356..G-C]	+	intron	intron:1	-	-	TBC1D17
gene	19:49878356	g19:49878356..G-C]	49878356..G-C]	+	exon	exon:1	-	-	AKT1S1
gene	19:49878356	g19:49878356..G-C]	49878356..G-C]	+	intron	intron:2	-	-	TBC1D17

Výstup základní anotace představovaným anotačním programem pro testovanou komplexní variantu (tabulka je rozdělena do dvou částí řazených pod sebou).

[illegible]

Poziční a intervalová anotace komplexní varianty představovaného anotačního programu; vícečetné nálezy jsou v řádku odděleny speciálním znakem (^).

Biologické procesy		Buněčné komponenty		Molekulární funkce	
ID	Název kategorie	ID	Název kategorie	ID	Název kategorie
GO:0000462	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	GO:0005634	nucleus	GO:0003723	RNA binding
GO:0001829	trophoblast cell differentiation	GO:0005654	nucleoplasm	GO:0005515	protein binding
GO:0006364	rRNA processing	GO:0005730	nucleolus	GO:0030515	snoRNA binding
GO:0007155	cell adhesion	GO:0005737	cytoplasm	GO:0003674	molecular function
GO:0007565	female pregnancy	GO:0005829	cytosol	GO:0003676	nucleic acid binding
GO:0008283	cell proliferation	GO:0016020	membrane	GO:0005488	binding
GO:0000003	reproduction	GO:0030688	pre-ribosome, small subunit precursor	GO:0097159	organic cyclic compound binding
GO:0001701	in utero embryonic development	GO:0043231	intracellular membrane-bounded organelle	GO:1901363	heterocyclic compound binding
GO:0001824	blastocyst development	GO:0045177	apical part of cell		
GO:0001825	blastocyst formation	GO:0005575	cellular component		
GO:0006139	nucleobase-containing compound metabolic process	GO:0005622	intracellular		
GO:0006396	RNA processing	GO:0005623	cell		
GO:0006725	cellular aromatic compound metabolic process	GO:0030684	pre-ribosome		
GO:0006807	nitrogen compound metabolic process	GO:0031974	membrane-enclosed lumen		
GO:0007275	multicellular organism development	GO:0031981	nuclear lumen		
GO:0008150	biological process	GO:0032991	protein-containing complex		
GO:0008152	metabolic process	GO:0043226	organelle		
GO:0009653	anatomical structure morphogenesis	GO:0043227	membrane-bounded organelle		
GO:0009790	embryo development	GO:0043228	non-membrane-bounded organelle		
GO:0009792	embryo development ending in birth or egg hatching	GO:0043229	intracellular organelle		
GO:0009987	cellular process	GO:0043232	intracellular non-membrane-bounded organelle		
GO:0010467	gene expression	GO:0043233	organelle lumen		
GO:0016070	RNA metabolic process	GO:0044422	organelle part		
GO:0016072	rRNA metabolic process	GO:0044424	intracellular part		
GO:0022414	reproductive process	GO:0044428	nuclear part		
GO:0022610	biological adhesion	GO:0044444	cytoplasmic part		
GO:0022613	ribonucleoprotein complex biogenesis	GO:0044446	intracellular organelle part		
GO:0030154	cell differentiation	GO:0044464	cell part		
GO:0030490	maturation of SSU-rRNA	GO:0070013	intracellular organelle lumen		
GO:0032501	multicellular organismal process	GO:1990904	ribonucleoprotein complex		
GO:0032502	developmental process				
GO:0034470	ncRNA processing				
GO:0034641	cellular nitrogen compound metabolic process				
GO:0034660	ncRNA metabolic process				
GO:0042254	ribosome biogenesis				
GO:0042274	ribosomal small subunit biogenesis				
GO:0043009	chordate embryonic development				
GO:0043170	macromolecule metabolic process				
GO:0044085	cellular component biogenesis				
GO:0044237	cellular metabolic process				
GO:0044238	primary metabolic process				
GO:0044703	multi-organism reproductive process				
GO:0044706	multi-multicellular organism process				
GO:0046483	heterocycle metabolic process				
GO:0048646	anatomical structure formation involved in morphogenesis				
GO:0048856	anatomical structure development				
GO:0048869	cellular developmental process				
GO:0051704	multi-organism process				
GO:0071704	organic substance metabolic process				
GO:0071840	cellular component organization or biogenesis				
GO:0090304	nucleic acid metabolic process				
GO:1901360	organic cyclic compound metabolic process				

Příloha 3.9: Rekonstruované GO kategorie pro BYSL

Všechny nalezené GO kategorie pro BYSL gen po rekonstrukci ancestrálních vztahů anotačním programem; šedě vyznačeny kategorie, které jsou vedeny v asociačním souboru pro BYSL GO konsorciem.

ID	Název kategorie
HP:0000007	Autosomal recessive inheritance
HP:0000527	Long eyelashes
HP:0000822	Hypertension
HP:0001508	Failure to thrive
HP:0002013	Vomiting
HP:0006519	Alveolar cell carcinoma
HP:0006532	Recurrent pneumonia
HP:0025092	Epidermal acanthosis
HP:0100501	Recurrent bronchiolitis
HP:0200034	Papule
HP:0200039	Pustule
HP:0000001	All
HP:0000005	Mode of inheritance
HP:0000118	Phenotypic abnormality
HP:0000152	Abnormality of head or neck
HP:0000234	Abnormality of the head
HP:0000271	Abnormality of the face
HP:0000315	Abnormality of the orbital region
HP:0000366	Abnormality of the nose
HP:0000492	Abnormal eyelid morphology
HP:0000499	Abnormality of the eyelashes
HP:0000600	Abnormality of the pharynx
HP:0000951	Abnormality of the skin
HP:0001507	Growth abnormality
HP:0001574	Abnormality of the integument
HP:0001595	Abnormality of the hair
HP:0001626	Abnormality of the cardiovascular system
HP:0001739	Abnormality of the nasopharynx
HP:0002017	Nausea and vomiting
HP:0002086	Abnormality of the respiratory system
HP:0002087	Abnormality of the upper respiratory tract
HP:0002088	Abnormal lung morphology
HP:0002090	Pneumonia

ID	Název kategorie
HP:0002205	Recurrent respiratory infections
HP:0002664	Neoplasm
HP:0002715	Abnormality of the immune system
HP:0002719	Recurrent infections
HP:0002783	Recurrent lower respiratory tract infections
HP:0002788	Recurrent upper respiratory tract infections
HP:0002837	Recurrent bronchitis
HP:0004323	Abnormality of body weight
HP:0004325	Decreased body weight
HP:0005607	Abnormal tracheobronchial morphology
HP:0010978	Abnormality of immune system physiology
HP:0011025	Abnormality of cardiovascular system physiology
HP:0011121	Abnormality of skin morphology
HP:0011122	Abnormality of skin physiology
HP:0011123	Inflammatory abnormality of the skin
HP:0011124	Abnormality of epidermal morphology
HP:0011138	Abnormality of skin adnexa morphology
HP:0011355	Localized skin lesion
HP:0011458	Abdominal symptom
HP:0011793	Neoplasm by anatomical site
HP:0011947	Respiratory tract infection
HP:0012252	Abnormal respiratory system morphology
HP:0012387	Bronchitis
HP:0012647	Abnormal inflammatory response
HP:0012649	Increased inflammatory response
HP:0025031	Abnormality of the digestive system
HP:0025032	Abnormality of digestive system physiology
HP:0025426	Abnormal bronchus morphology
HP:0030669	Abnormal morphology of the ocular adnexa
HP:0030972	Abnormal systemic blood pressure
HP:0100526	Neoplasm of the lung
HP:0100552	Neoplasm of the tracheobronchial system
HP:0100606	Neoplasm of the respiratory system

Příloha 3.10: Rekonstruované HPO kategorie pro *EGFR*

Všechny nalezené HPO kategorie pro *EGFR* gen po rekonstrukci ancestrálních vztahů anotačním programem; šedě vyznačeny kategorie, které jsou vedeny v asociačním souboru pro *EGFR* HPO konsorciem.

Příloha 3.11: Elektronická příloha ne/rekonstruovaných ontologických vztahů pro GO

Tabulka v CSV formátu se sloupci oddělenými tabulátorem.

Příloha 3.12: Elektronická příloha ne/rekonstruovaných ontologických vztahů pro HPO

Tabulka v CSV formátu se sloupci oddělenými tabulátorem.

Příloha 3.13: Elektronická příloha anotovaných variant u pacienta s agranulocytózou

Tabulka filtrovaných variant bez odstraněných duplikátů ve formátu XLSX.

Příloha 3.14: Elektronická příloha finálního reportu u pacienta s agranulocytózou

Finální html report v pdf formátu.

Příloha 4.1: Elektronická příloha signifikantně dysregulovaných genů skupiny RPS7

Tabulka signifikantně dysregulovaných genů skupiny RPS7 oproti zdravým kontrolám reportována shodně programy Cufflinks a DESeq2 v XLSX formátu.

Příloha 4.2: Elektronická příloha signifikantně dysregulovaných genů skupiny RPS19

Tabulka signifikantně dysregulovaných genů skupiny RPS19 oproti zdravým kontrolám reportována shodně programy Cufflinks a DESeq2 v XLSX formátu.

Příloha 4.3: Elektronická příloha signifikantně dysregulovaných genů skupiny DBA

Tabulka signifikantně dysregulovaných genů společných pro skupinu RPS7 a RPS19 reportována shodně programy Cufflinks a DESeq2 v XLSX formátu.

Příloha 4.4: Elektronická příloha signifikantně dysregulovaných genů skupiny RPS7 oproti RPS19

Tabulka signifikantně dysregulovaných genů skupiny RPS7 oproti RPS19 reportována shodně programy Cufflinks a DESeq2 v XLSX formátu.

Příloha 4.5: Analýza dysregulovaných signálních drah RPS7 skupiny v programu Reactome

Tabulka ve formátu XLSX signifikantně dysregulovaných signálních drah pro skupinu RPS7.

Příloha 4.6: Analýza dysregulovaných signálních drah RPS19 skupiny v programu Reactome

Tabulka ve formátu XLSX signifikantně dysregulovaných signálních drah pro skupinu RPS19.

Příloha 4.7: Analýza dysregulovaných signálních drah DBA skupiny v programu Reactome

Tabulka ve formátu XLSX signifikantně dysregulovaných signálních drah společných pro skupiny RPS7 a RPS19.

10.1 Životopis

CURRICULUM VITAE

Mgr. Petr Vojta

Email: petr.vojta@upol.cz

Mobilní telefon: +420 776 840 217

Datum narození: 16.2.1984

Národnost: česká

Vzdělání:

2009 - 2. LF UK v Praze, postgraduální studium
obor: Molekulární a buněčná biologie, genetika a virologie

2007 - 2009 PŘF UK v Praze, magisterský studijní program
obor: Molekulární a buněčná biologie, genetika a virologie

2004 - 2007 PŘF UP v Olomouci, bakalářský studijní program
obor: Molekulární a buněčná biologie

Zaměstnavatel:

2011 - Ústav molekulární a translační medicíny, UP

2010 - 2011 Ústav biologie a lékařské genetiky, 2. LF UK

2008 - 2010 Ústav molekulární genetiky, AV ČR

Seznam publikací:

- Chyra Kuřova Z., Sevcikova T., Januska J., Vojta P., Boday A., Vanickova P., *et al.* (2018) Newly designed 11-gene panel reveals first case of hereditary amyloidosis captured by massive parallel sequencing. *Journal of Clinical Pathology*, jclinpath-2017-204978.
- Hroudova M., Vojta P., Strnad H., Krejčík Z., Ridl J., Paces J., *et al.* (2012) Diversity, phylogeny and expression patterns of pou and six homeodomain transcription factors in hydrozoan jellyfish *Craspedacusta sowerbyi*. *PLoS ONE* 7.
- Kokas F., Vojta P. & Galuszka P. (2016) Dataset for transcriptional response of barley (*Hordeum vulgare*) exposed to drought and subsequent re-watering. *Data in Brief* 8, 334–341.
- Kuřova Z., Sevcikova T., Growkova K., Vojta P., Filipova J., Adam Z., *et al.* (2017) Biomarkers in immunoglobulin light chain amyloidosis. *Klinická Onkologie* 30.
- Majeska-Cudejkova M., Vojta P., Valarik J. & Galuszka P. (2016) Quantitative and qualitative transcriptome analysis of four industrial strains of *Claviceps purpurea* with respect to ergot alkaloid production. *New Biotechnology* 33, 743–754.
- Pospíšilová H., Jiskrová E., Vojta P., Mrázová K., Kokáš F., Čudejková M.M., *et al.* (2016) Transgenic barley overexpressing a cytokinin dehydrogenase gene shows greater tolerance to drought stress. *New Biotechnology* 33, 692–705.
- Stranska J., Jancik S., Slavkovsky R., Holinkova V., Rabcanova M., Vojta P., *et al.* (2015) Whole genome amplification induced bias in the detection of KRAS-mutated cell populations during colorectal carcinoma tissue testing. *Electrophoresis* 36, 937–940.
- Tulpová Z., Luo M.C., Toegelová H., Visendi P., Hayashi S., Vojta P., *et al.* (2018) Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. *New Biotechnology*.
- Vojta P., Kokáš F., Husičková A., Grúz J., Bergougnoux V., Marchetti C.F., *et al.* (2016) Whole transcriptome analysis of transgenic barley with altered cytokinin homeostasis and increased tolerance to drought stress. *New Biotechnology* 33, 676–691.
- Volejnikova J., Zapletalova J., Jarosova M., Urbankova H., Vojta P., Klaskova E., *et al.* (2018) Acute lymphoblastic leukemia in a child with Ieri-Weill syndrome and complete SHOX gene deletion: A case report. *Biomedical Papers* 162, 65–70.
- Zatopkova M., Filipova J., Jelinek T., Vojta P., Sevcikova T., Simicek M., *et al.* (2017) Whole exome sequencing of aberrant plasma cells in a patient with multiple myeloma minimal residual disease. *Klinická Onkologie* 30.

Abstrakta:

Large Deletions and Novel Mutations in Genes Coding for Ribosomal Proteins (RPL11, RPL35A, RPS19) in the Czech National Registry of Diamond-Blackfan Anemia (DBA)	Hajdúch M, Horváthová M, Kořalková P, Ludíková B, Macečková Z, Pospíšilová D, Suková M, Timr P, Urbánková H, Vojta P, Volejníková J	Blood (print)	2017
Bioinformatic analysis of massive parallel sequencing data for cancer research	Hajdúch M, Slavkovský R, Vojta P	Abstract Book: XIII. Diagnostic, Predictive and Experimental Oncology Days	2017
MOLDIMED-analytical pipeline solution for DNA massive parallel sequencing	Vojta P, Svaton V, Hajdúch M	BIOSPOT	2017
Vliv léčby prednisolonem na transkriptomové profily u dětí s akutní lymfoblastickou leukémií.	Džubák P, Hajdúch M, Mihál V, Slavík H, Srovnal J, Špenerová M, Vojta P, Volejníková J	Klinická onkologie	2016
Two Novel Mutations in Ribosomal Proteins in the Czech National Diamond-Blackfan Anemia Registry	Hadacova I, Hajdúch M, Horváthová M, Jarošová M, Jungova A, Kořalková P, Macečková Z, Mojžíková R, Pospíšilová D, Saxová Z, Vojta P, Volejníková J	Blood (print)	2016
A novel mutation of ribosomal protein S7 in three family members with markedly different severity of Diamond-Blackfan anemia in a single family	Hajdúch M, Horváthová M, Konečný P, Kořalková P, Macečková Z, Pospíšilová D, Vojta P	Transfúze a hematologie dnes	2016
Nové mutace u pacientů s Diamondovou-Blackfanovou anémií (DBA)	Hajdúch M, Horváthová M, Jarošová M, Kořalková P, Macečková Z, Pospíšilová D, Vojta P, Volejníková J	Sborník abstrakt	2016

Vliv léčby prednisonem na transkriptomovou mapu u dětí s akutní lymfoblastickou leukémií	Slavík H, Vojta P, Volejníková J	Prague onco journal	2016
Ultra deep amplicon sequencing of RAS genes for colorectal carcinoma diagnostics	Drábek J, Hajdúch M, Jančík S, Slavkovský R, Stránská J, Šimová J, Věnsková V, Vojta P	XI. Diagnostic, Predictive and Experimental Oncology Days, Abstract book	2015
5-Azacytidine Nucleosides and their Derivates: Molecular Hallmarks of Drug Resistance	Agrawal K, Džubák P, Frydrych I, Hajdúch M, Holub D, Macečková Z, Otmar M, Vojta P	Abstract Book, X. Diagnostic, Predictive and Experimental Oncology Days	2014
System for reliable identification of alterations using protein mass spectrometry	Agrawal K, Džubák P, Hajdúch M, Holub D, Hruška M, Varanasi L, Vojta P, Voller J	Abstract Book, X. Diagnostic, Predictive and Experimental Oncology Days	2014
5-azacytidine nucleosides and their derivatives: Molecular hallmarks of drug resistance	Agrawal K, Džubák P, Frydrych I, Hajdúch M, Holub D, Krečmerová M, Otmar M, Vojta P	Proceedings of the 105th Annual Meeting of the American Association for Cancer Research	2014
5-azacytidine nucleosides and their derivatives: Molecular hallmark of drug resistance	Agrawal K, Džubák P, Frydrych I, Hajdúch M, Holub D, Krečmerová M, Otmar M, Vojta P	Zborník prednášok: Súťaž mladých onkológov 2014	2014
A system for identification and quantification of mutations in proteomic data	Agrawal K, Džubák P, Hajdúch M, Holub D, Hruška M, Macečková Z, Vojta P, Voller J	IX. Diagnostic, Predictive and Experimental Oncology Days: Abstract Book	2013
Diagnostics of minor mutations in RAS oncogenes in Mrcr with next-gen sequencing	Drábek J, Hajdúch M, Jančík S, Slavkovský R, Stránská J, Vojta P	IX. Diagnostic, Predictive and Experimental Oncology Days: Abstract Book	2013
Epigenetic study of 5-azacytidine nucleosides and their derivatives	Agrawal K, Džubák P, Frydrych I, Hajdúch M, Holub D, Krečmerová M, Otmar M, Vojta P	IX. Diagnostic, Predictive and Experimental Oncology Days: Abstract Book	2013

Použití polymerázové řetězové reakce v reakčním objemu 1 µl pro detekci mutací genu EGFR	Drábek J, Hajdúch M, Holinková V, Stránská J, Vojta P	Edukační sborník	2012
Vysokokapacitní DNA sekvenování v molekulární diagnostice onkologických markerů EGFR signální dráhy	Drábek J, Hajdúch M, Jančík S, Macečková Z, Stránská J, Vojta P	Optics Communications	2012
Použití polymerázové řetězové reakce v reakčním objemu 1 µl pro detekci mutací genu EGFR	Drábek J, Hajdúch M, Holinková V, Rabčanová M, Stránská J, Vojta P	Optics Communications	2011

10.2 Příložené publikace