UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie

Software pro analýzu nukleotidových modifikací

BAKALÁŘSKÁ PRÁCE

Autor:	Karel Vrabka
Studijní program:	B1406 Biochemie
Studijní obor:	Biochemie
Forma studia:	Prezenční
Vedoucí práce:	**Mgr. Bc. Filip Zavadil Kokáš**
Rok:	2019

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně s vyznačením všech použitých pramenů a spoluautorství. Souhlasím se zveřejněním bakalářské práce podle zákona č. 111/1998 Sb., o vysokých školách, ve znění pozdějších předpisů. Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, ve znění pozdějších předpisů.

V Olomouci dne 12.5.2019 ……

podpis bakaláře

Tímto způsobem bych rád poděkoval vedoucímu bakalářské práce Mgr. Bc. Filipu Zavadilovi Kokášovi za odborné rady a pomoc při psaní této práce.

Bibliografická identifikace

Jméno a příjmení autora	Karel Vrabka
Název práce	Software pro analýzu nukleotidových modifikací
Typ práce	Bakalářská
Pracoviště	Katedra biochemie
Vedoucí práce	Mgr. Bc. Filip Zavadil Kokáš
Rok obhajoby práce	2019

Abstrakt

Tato práce je zaměřena na tvorbu softwaru pro funkční anotaci jednonukleotidových polymorfismů a exportu proteinových sekvencí ovlivněných těmito DNA modifikacemi v podobě FASTA souboru. V práci je rovněž popsáno testování souboru na reálném datovém setu, který představují vzorky sekvenování sklerocii u dvou kmenů *Claviceps purpurea* (20.1 a Gal 404).

Teoretická část práce je zaměřena na popis současných sekvenačních metod a rovněž popis datových formátů využívaných v průběhu bioinformatické analýzy. Praktická část práce je věnována vývoji softwaru prezentovaném v této práci a jeho následnému testování na reálném datovém setu společně se srovnáním s dalším dostupným softwarem ANNOVAR poskytujícím anotaci jednonukleotidových polymorfismů. Výsledky anotace jednonukleotidových polymorfismů získané pomocí tohoto testování ukázaly, že prezentovaný software dosahuje srovnatelných výsledků se softwarem ANNOVAR. Software navíc poskytuje možnost vytvoření peptidových knihoven dále využitelných v proteomické analýze.

Klíčová slova	software, bioinformatika, *Claviceps*, anotace, jednonukleotidové polymorfizmy
Počet stran	44
Počet příloh	5
Jazyk	Český

Bibliographical identification

Autor's first name and surname	Karel Vrabka
Title	Software for analysis of nucleotide modifications
Type of thesis	Bachelor
Department	Department of biochemistry
Supervisor	Mgr. Bc. Filip Zavadil Kokáš
The year of presentation	2019

Abstract

Main goal of this thesis is development of software for annotating of single nucleotide polymorphisms and exporting of protein sequences, affected by these DNA modifications, in FASTA format. Thesis also contains description of software testing with real dataset on input. The dataset includes samples produced by sequencing two species of *Claviceps purpurea* (20.1 and Gal 404) sclerotia.

Theoretical part focuses on the description of most used sequencing methods and specification of data formats, that are used in bioinformatics analysis. Practical part aims on description of development and testing of created software followed by comparation with different software ANNOVAR with similar functions. Results of single nucleotide polymorphisms annotation showed that created software output is similar to ANNOVARs output. Created software also provides the option of creating peptid libraries, that can be used in proteomics analysis.

Keywords	software, bioinformatics, *Claviceps*, annotation, single nucleotide polymorphisms
Number of pages	44
Number of appendices	5
Language	Czech

OBSAH

Cíle bakalářské práce

1. Vypracování literární rešerše na téma kvalitativní bioinformatická analýza jednonukleotidových polymorfismů a současné sekvenační metody.
2. Srovnání softwarů sloužících k detekci jednonukleotidových polymorfismů.
3. Vytvoření softwaru sloužícího k anotaci DNA modifikací a následném exportu výsledků v podobě modifikovaných proteinových sekvencí ve formě FASTA formátu.
4. Zhodnocení efektivity vytvořeného softwaru srovnáním se softwarem poskytujícím podobné funkce.

1 Úvod

DNA tvořící základní složku buněk všech organismů podléhá spontánním případně indukovaným mutacím. Mezi nejčastější změny malého rozsahu ve struktuře DNA náleží zejména jednonukleotidové polymorfismy (substituce), inzerce a delece. Jednonukleotidový polymorfismus (SNP) je označením variace jedné nukleotidové báze v sekvenci DNA, která může způsobovat jak různé onemocnění, tak genetickou výhodu. Z tohoto důvodu je SNP častým objektem zkoumání. k získání informací o SNP v určitém organismu je potřeba sekvenačních metod, jejichž cílem je určení struktury daného genomu. Data získaná sekvenací DNA/RNA jsou následně bioinformaticky analyzována a porovnána s referenční DNA uloženou v databázi. Výsledkem je obraz rozdílů mezi genomem zkoumaného jedince a typickým genomem daného organismu, který je převážně způsoben SNP. Pro reprezentaci variací v genomu je využíván textový formát VCF.

Hlavním cílem práce je vytvoření softwaru, který analýzou VCF, GFF3 a FASTA souborů umožní uživateli přehledně vytvořit seznam SNP vyskytujících se v organismu a podat základní informace o jejich pozici v genomu a významu na úrovni translace. Další funkcí softwaru je generování proteinových sekvencí vzniklých v důsledku SNP a jejich export ve FASTA formátu. Bakalářská práce rovněž zahrnuje srovnání dvou softwarů pro kvalitativní detekci SNP. Vytvořený software je srovnáván s volně dostupným softwarem ANNOVAR poskytujícím anotaci jednonukleotidových polymorfismů.

Bakalářská práce je rozčleněna do několika kapitol. V první kapitole je prezentována literární rešerše na téma popis SNP a dalších modifikací DNA a jejich vlivu na genom organismu. Literární rešerše rovněž obsahuje přehled sekvenačních metod s důrazem na nejvyužívanější metody, a především specifikaci textových formátů využívaných v bioinformatice spolu se softwary určenými k jejich zpracování. Druhá kapitola obsahuje popisy programů a bioinformatických dat využitých v této práci. Kapitola s výsledky zahrnuje část obsahující deskripci softwaru implementovaného v této práci, jeho testování na reálném datovém setu a výsledky testování.

2 Současný stav řešené problematiky

2.1 Jednonukleotidové polymorfismy

Jednonukleotidové polymorfismy, zkráceně SNP jsou mutace v DNA, vyskytující se alespoň u 1 % populace, vzniklé nahrazením (substitucí), přidáním (inzercí), nebo odstraněním (delecí) nukleotidové báze na specifické pozici v genomu. Těmito variacemi se DNA sekvence zkoumaného organismu odlišuje od obvyklé (referenční) sekvence.

Většina detekovaných SNP se zpravidla nachází v nekódujících oblastech DNA (u člověka až 90 %, Chen *et al.,* 2016), tedy v úsecích, které nejsou využity pro syntézu bílkovin. Z pohledu dopadu SNP na fenotyp zkoumaného organismu se za důležitější považují SNP vyskytující se v oblastech DNA, jež kóduje proteiny. Ty mohou být dále rozděleny dle vlivu na potenciálně syntetizovaný protein (Clancy, 2008). Vzhledem ke skutečnosti, že genetický kód je degenerovaný (více tripletů kóduje stejnou aminokyselinu), jsou změny v sekvenci DNA, nezpůsobující přetvoření aminokyseliny, označovány jako synonymní. Pokud ovšem SNP ovlivní vznikající protein, jsou tyto mutace označovány jako nesynonymní. Mutace tohoto typu mohou vést k závažnějším problémům, u člověka např. k srpkovité anémii (Lettre *et al.,* 2008). Nicméně i SNP v nekódujících oblastech DNA mohou mít velký vliv, např. SNP rs1625579 ovlivňuje sekvenci miRNA, související s lidskou schizofrenií (Hrdlickova *et al.,* 2014).

2.2 Metody sekvencování nukleových kyselin

Metody sloužící k zjištění sekvence DNA se nazývají sekvenační metody a jsou nedílnou součástí analýzy genomu a transkriptomu. Proces sekvencování nukleových kyselin probíhá v několika krocích, které v sobě nesou kritické faktory nezbytné pro získání kvalitních dat pro následnou bioinformatickou analýzu (Morey *et al*., 2013).

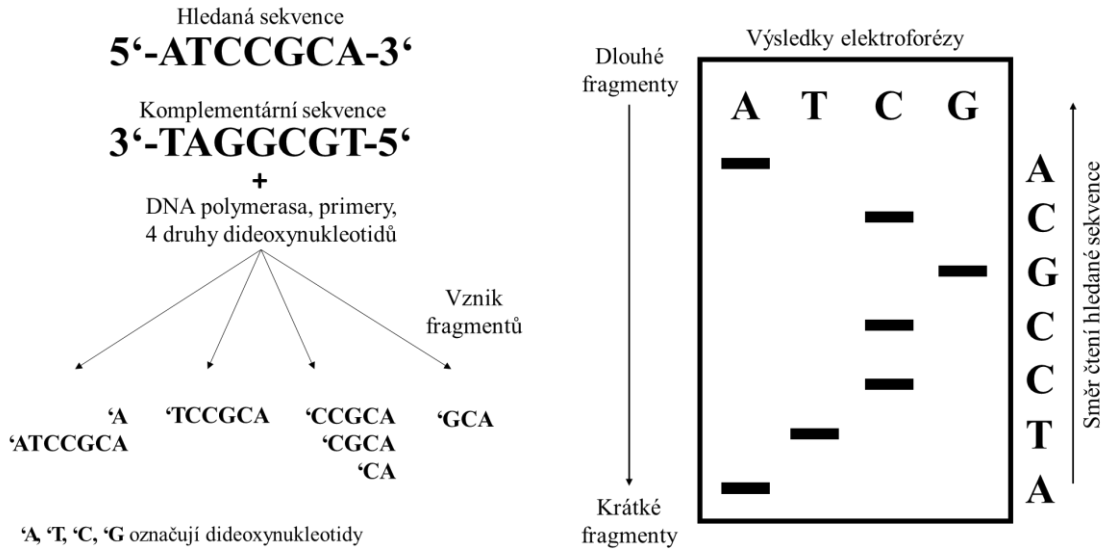
Prvním krokem sekvenace bývá zpravidla rozdělení sekvence DNA na kratší úseky, tato fragmentace není nutná v případě, kdy analyzovaná sekvence není příliš dlouhá. Rozštěpení může být prováděna různými způsoby, např. využitím restrikčních enzymů, nebo mechanickými metodami např. sonifikací (Knierim *et al*., 2011). Jednotlivé části DNA jsou následně ve většině případů namnoženy za účelem kvalitnější detekce jednotlivých bází. Současné sekvenační metody jsou často založeny na snímání chemických, nebo fyzikálních vlastností specifických pro každou bázi, a díky namnožení sekvencí probíhá současně stejná fyzikální/chemická operace u více totožných úseků

nukleové kyseliny, což způsobuje zesílení signálu a ulehčuje detekčním systémům zpracovat informaci o dané bázi (Ansorge, 2009). Každá sekvenační metoda má své specifické vlastnosti a provedení sekvenace. Dle doby vzniku lze sekvenační metody rozdělit na metody první, druhé a třetí generace (Morey *et al*., 2013). Některé z nich jsou stručně popsány v následujících kapitolách s důrazem na jejich současné využití.

2.2.1 Metody první generace

První významná sekvenační metoda byla vyvinuta v 70. letech 20. století týmem Fredericka Sangera (Sanger *et al.*, 1977), dnes známá jako Sangerova metoda, která se stala na další desítky let nejpoužívanější sekvenační metodou.

Při sekvenační reakci se využívá replikace DNA, při které je nové vlákno vytvářeno z deoxynukleotidů a dideoxynukleotidů. V reakční směsi jsou dideoxynukleotidy zastoupeny ve velmi malém množství a jejich začlenění do nově vytvářeného vlákna způsobuje zastavení replikace. Původní sekvenace probíhala ve čtyřech reakčních nádobách, v každé z nich bylo spolu s jednovláknovou DNA, vzniklou namnožením a denaturací původní DNA, velké množství deoxynukleotidů, malé množství vždy jednoho typu dideoxynukleotidu, primery a DNA polymerasa (Obr. 1). Náhodné začleňování dideoxynukleotidů způsobilo vznik různě dlouhých dvouvláknových segmentů DNA, u kterých byla známa poslední báze. Takto vzniklé úseky byly pomocí



Obrázek 1: Schéma sekvenace pomocí Sangerovy metody.

gelové elektroforézy seřazeny dle velikosti, a následně analýzou byla získána hledaná sekvence (Sanger *et al*., 1977; Obr. 1).

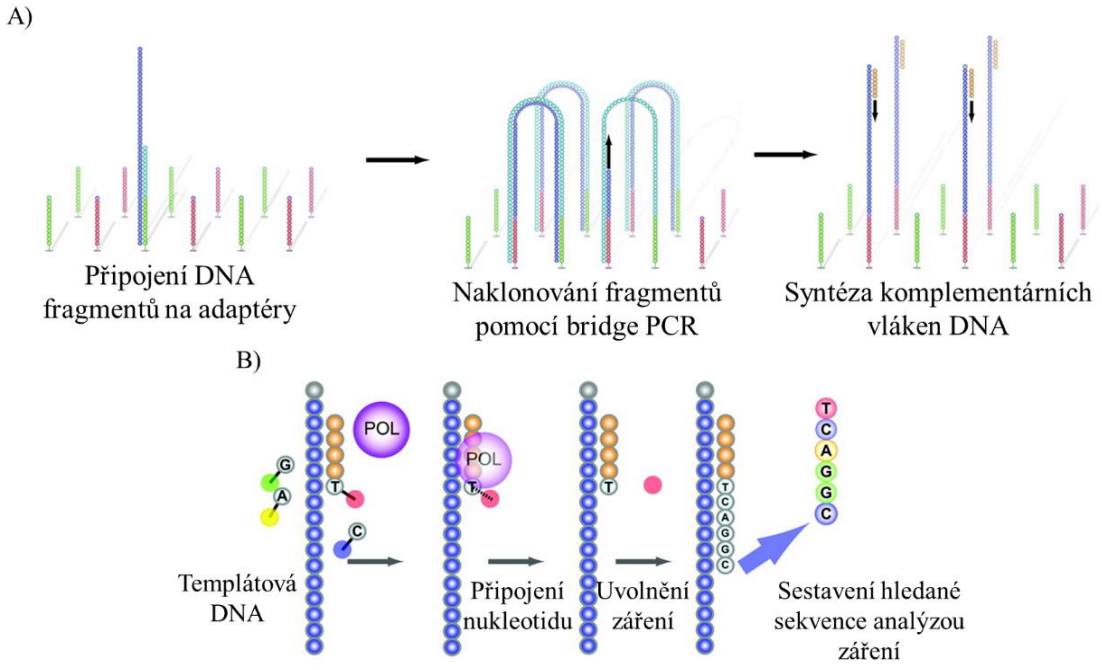
Ačkoli původní Sangerova metoda patří do metod první generace, během následujících let prošla několika vylepšeními, díky kterým je užívána i dnes (Totomoch-Serra *et al*., 2017). Mezi tyto vylepšení patří např. využití fluorescenčního značení specifického pro každý dideoxynukleotid, díky němuž může sekvenační reakce probíhat v jedné reakční nádobě (Smith *et al*., 1985). Další inovací je výměna gelové elektroforézy za kapilární (Swerdlow *et al*., 1990). Sekvencování dnes také není prováděno mechanicky, ale využívá se automatizace umožňující průběh sekvenační reakce v nanolitrových reakčních nádobách.

Ačkoli metoda prošla od počátku mnoha změnami, v dnešní době je ve srovnání s ostatními novějšími metodami výrazně dražší, nevyplatí se ji tedy využívat pro velké projekty. Využití nachází pro sekvencování delších úseků DNA, protože umožňuje generovat sekvence DNA („ready“) délky až 900 bp, čímž se liší oproti metodám druhé generace (Morozova a Marra, 2008).

2.2.2 Metody sekvencování druhé generace

Metody druhé generace byly vyvíjeny od 90. let 20. století a první komerční využití našly začátkem 21. století. V dnešní době se k sekvenaci DNA často používají sekvenační stroje společností – Illumina, 454 Life Sciences nebo SOLiD (Morey *et al*., 2013). Nejpoužívanější metodu představuje vzhledem k současné dominanci na trhu Illumina sekvencování.

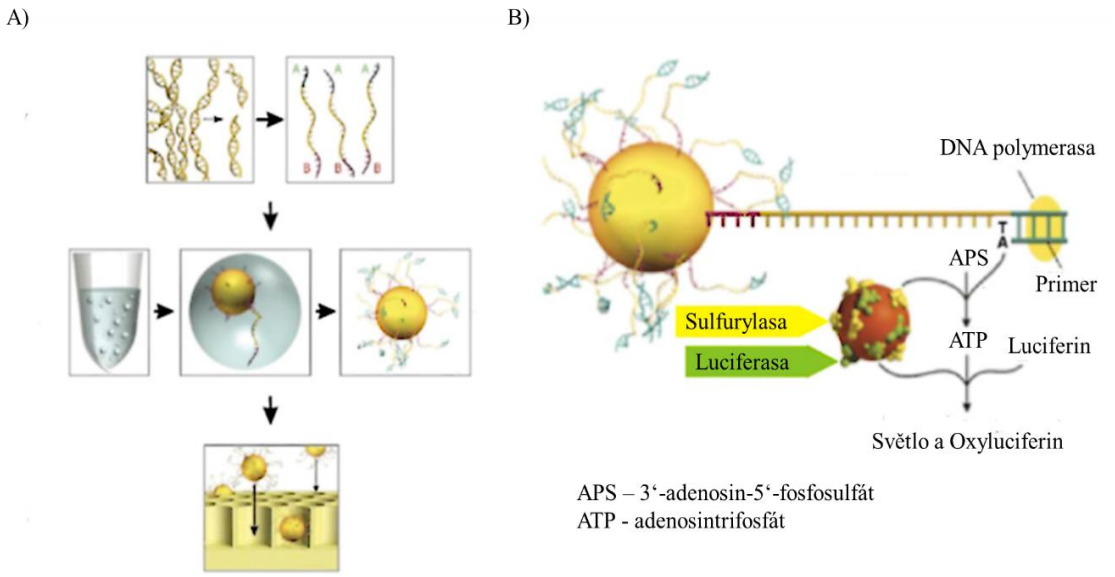
Sekvenace metodou Illumina je založena na detekci fluorescenčně značených deoxynukleotidů při syntéze komplementárního vlákna DNA za katalýzy DNA polymerasou. Před samotnou sekvenací musí proběhnout namnožení DNA (za účelem snadnější detekce signálu při syntéze komplementárního vlákna), což u této metody probíhá metodou „bridge-PCR“ (Obr. 2, Adessi *et al*., 2000). Specifické adaptéry jsou připojeny na oba konce DNA fragmentů, které po denaturaci slouží jednovláknovým fragmentům DNA pro přichycení k reakční destičce („flow cell“), na které sekvenace probíhá (Bentley *et al*., 2009). Povrch destičky je pokryt adaptéry, které jsou komplementární k adaptérům přichyceným k fragmentům (Morey *et al*., 2013). Po hybridizaci komplementárních fragmentů dochází k replikaci, zformovaná dvouvláknová DNA je zdenaturována a vzniklé jednovláknové fragmenty se znovu napojí na komplementární adaptéry a znovu proběhne replikace (Obr. 2).



Obrázek 2: Průběh sekvenace metodou Illumina (upraveno podle Voelkerding, 2009). A) Příprava DNA pro sekvenaci metodou „bridge-PCR". B) Schéma sekvenační reakce u metody Illumina.

Tímto způsobem je získán řádově větší počet fragmentů imobilizovaných na malém prostoru. Na konci procesu proběhne vymytí všech deoxynukleotidů a dalších reaktantů (Voelkerding *et al.*, 2009). K sekvenaci je následně užito upravených specificky fluorescenčně označených deoxynukleotidů mající schopnost reversibilní terminace syntézy díky O-azidomethylové skupině na 3' konci (Guo *et al.*, 2008). Přidání deoxynukleotidů na „flow cell" způsobí začlenění pouze jedné komplementární báze ke každému jednovláknovému fragmentu (Obr. 2). Pomocí CCD („couple-charge device") senzorů je každá přidaná báze detekována, následně proběhne vymytí všech deoxynukleotidů a odstranění O-azidomethylové skupiny. Proces se opakuje, dokud nedojde k syntéze všech komplementárních vláken na požadovanou délku. Poté je analýzou výsledků získána hledaná sekvence. Illumina patří mezi nejvíce využívané metody, je výhodná jak díky nízké ceně (0,07 $ za milion bází), tak díky své rychlosti (až 600 Gb za 3–10 dní). Nevýhodou této metody je generování pouze krátkých „readů" o délce 100 bp (Morey *et al.*, 2013).

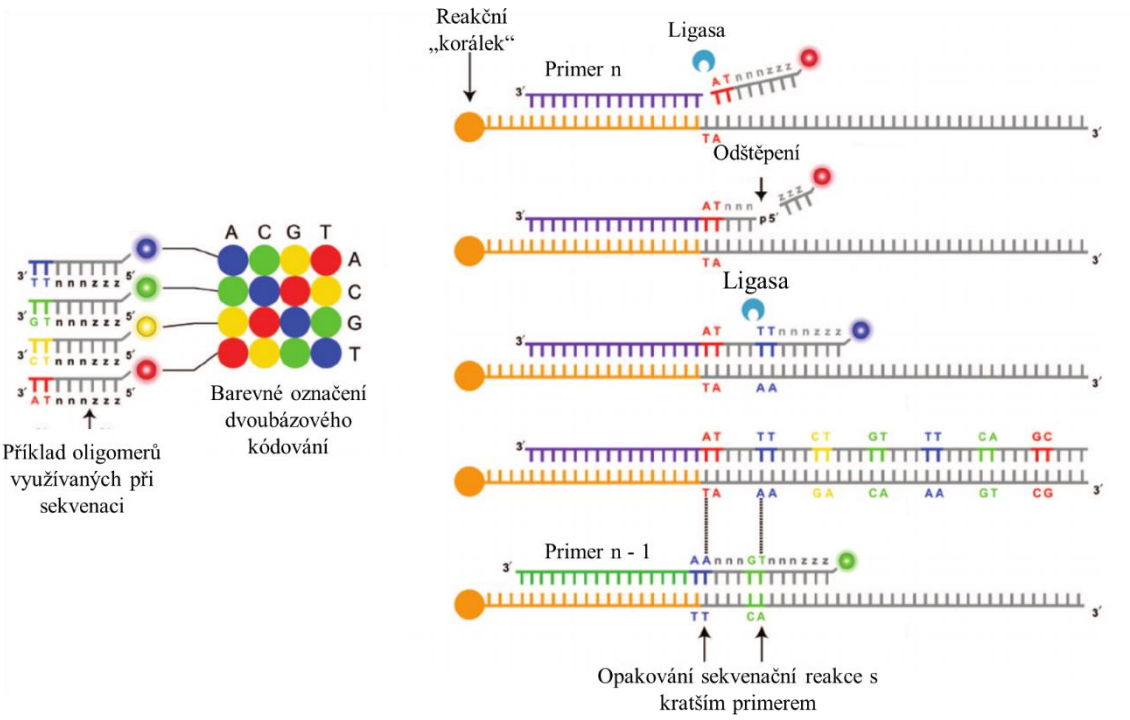
Pyrosekvencování užívané společností 454 Life Sciences je založeno na podobném principu jako Illumina sekvencování a představuje další metodu z druhé generace



Obrázek 3: Průběh sekvenace metodou pyrosekvencování (upraveno podle Ansorge, 2009). A) Znázornění fragmentace, připojení a naklonování DNA na „korálcích" a jejich následné umístění na reakční destičce. B) Schéma sekvenační reakce.

sekvenačních metod. Místo klonování metodou „bridge PCR" ovšem využívá emulzní PCR, při které dochází k namnožení každého fragmentu v odděleném prostředí na „korálcích". Samotná sekvenace probíhá v pikolitrových reakčních jamkách na reakční destičce, kdy v ideálním případě je v každé z nich umístěn jeden „korálek" pokrytý kopiemi jednoho fragmentu DNA (Obr. 3). Na reakční destičku je vždy přidán jeden typ deoxynukleotidů, a následně se měří záření v každé jamce. Po začlenění báze do syntetizovaného vlákna se uvolní pyrofosfát, který reaguje s ATP sulforylasou za vzniku ATP, něhož následně využívá luciferasa k emitaci světelného záření (Obr. 3). Enzym apyrasa je poté využit k degradaci nevyužitých deoxynukleotidů. Výhodou pyrosekvencování je generování poměrně dlouhých fragmentů (až 700 bp) a doba trvání běhu sekvenátoru je průměrně jeden den, nicméně cena sekvenace milionu bází je 10 $ (Ronaghi *et al*., 1996).

SOLiD metoda představuje třetí metodu sekvencování. Tato metoda není založena, na rozdíl od předchozích, na syntéze DNA pomocí DNA polymerasy, ale spočívá v ligování krátkých oligonukleotidů. Základem jsou oligomery, vyznačující se speciální strukturou. Oligomer obsahuje dva striktně definované nukleotidy a na dalších pozicích se nacházejí degenerované báze (Ambardar *et al*., 2016). Oligo mer je zároveň označen specifickou fluorescenční značkou korespondující s definovanými bázemi. Před začátkem sekvenace je potřeba k fragmentům DNA připojit adaptéry a amplifikovat vzorky pomocí „emulsion PCR". Pro každý takto připravený vzorek následně proběhne vlastní sekvenace

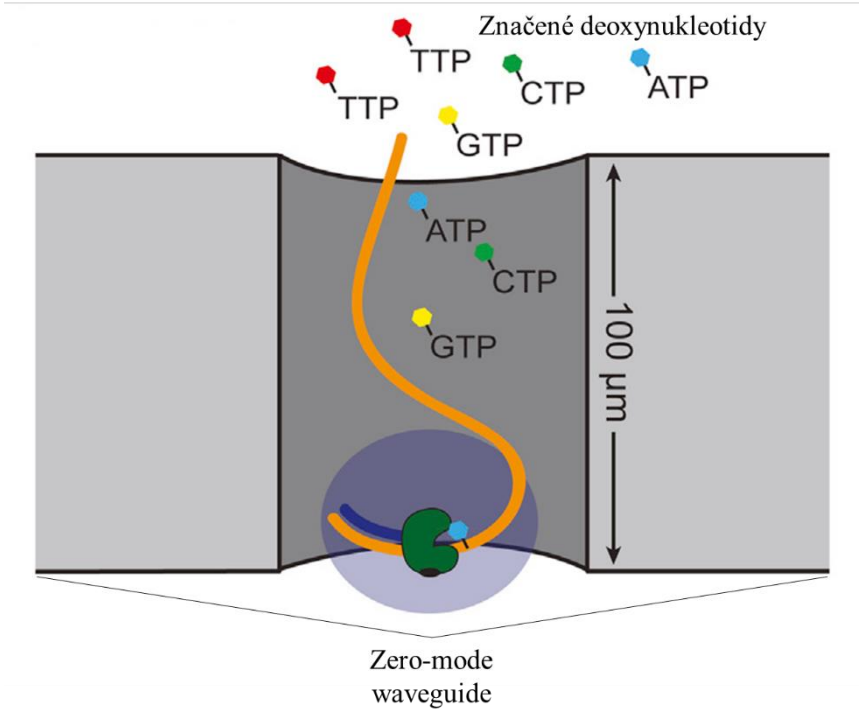


Obrázek 4: Schéma sekvenace metodou SOLiD (upraveno podle Voelkerding, 2009).

spočívající v připojení specificky dlouhého primeru. Po tomto kroku následuje navázání oligomerů a jejich připojení pomocí enzymu DNA ligasy. Vzhledem ke skutečnosti, že oligomer má definovány dvě báze, jedná se o tzv. dvoubázové dekódování. Značený oligomer je následně s pomocí fluorescence detekován a dochází k odštěpení značené části spolu se třemi koncovými bázemi (Mardis, 2008). Opakováním tohoto procesu dochází k sekvencování celého fragmentu až do produkované délky „readu“. Po dokončení procesu sekvenace je nasyntetizovaný řetězec odštěpen a celý proces je opakován s primerem o rozdílné délce (Obr. 4). Tímto způsobem je v rámci této metody každá báze sekvencována dvakrát, což signifikantně zvyšuje spolehlivost metody. SOLiD metoda ovšem může vytvářet chyby při sekvencování palindromických úseků DNA a běh sekvenátoru trvá podstatně delší dobu ve srovnání s ostatními technikami (až 14 dní), na druhou stranu je velice levná (0,17 $ za milion bází) a přesná (99,9 %, Buermans a Dunnen, 2014).

2.2.3 Metody sekvencování třetí generace

V posledních letech probíhá rozsáhlý vývoj metod tzv. třetí generace (Schadt *et al.*, 2010). Ve srovnání s předchozími metodami se vyznačují vysokou rychlostí a využitím nejmodernějších technologií. Pro správnou sekvenaci navíc není nutná amplifikace DNA.

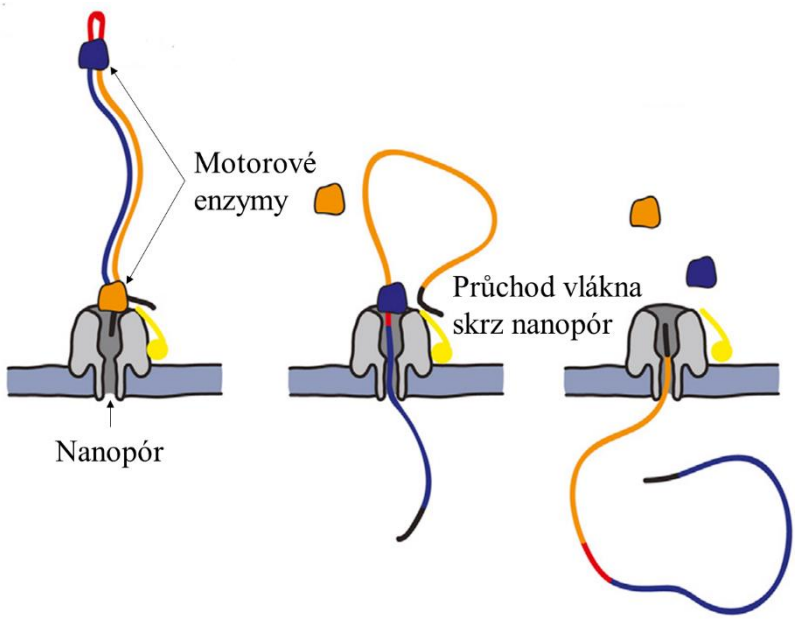


Obrázek 5: Schéma průběhu sekvenační metody SMRT (upraveno podle Reuter, 2015).

Mezi hlavní dva zástupce patří technika SMRT („Single-molecule-real-time“) sekvenace a „Nanopore“ sekvenace. Technika sekvencování pomocí nanopórů je v některých publikacích připisována již do metod sekvencování čtvrté generace (Feng *et al*., 2015).

SMRT sekvenace probíhá v tzv. „zero-mode waveguides“ (ZMV), speciálních dutinách o zeptolitrovém objemu, obsahujících senzor pro detekci záření a DNA polymerasu. SMRT čip obsahuje desítky tisíc ZMV a v každém probíhá katalyzovaná syntéza komplementárního vlákna DNA (Obr. 5) začleňováním fluorescenčně značených deoxynukleotidů (Reuter *et al*., 2015). Před sekvencováním je třeba segmenty upravit do cyklického tvaru připojením vlásenkovitých adaptérů, tato změna struktury umožní několikanásobnou sekvenaci každého fragmentu. Metodu není nutné v průběhu zastavovat, jak tomu bylo u předchozích metod, její rychlost je ekvivalentní s rychlostí DNA polymerasy, tedy až tisíc bází za sekundu. Vzhledem k tomu, že takto rychle v současnosti nelze detekovat záření, dochází k několikanásobnému sekvencování každého fragmentu a k umělému zpomalování reakce (Reuter *et al*., 2015). Průměrná délka „readu“ se může pohybovat okolo 14 kbp, avšak sekvenace může probíhat až s 11 % chyb. Přesnost metody se zvyšuje vícenásobnou sekvenací každého segmentu (Ardui *et al*., 2018).

Metoda „nanopore“ sekvencování využívá skutečnost, že při průchodu vlákna DNA skrze pór v membráně vzniká rozdíl mezi napětím na obou stranách membrány, který je



Obrázek 6: Schéma průběhu sekvenační metody „nanopore“ (upraveno podle Reuter, 2015).

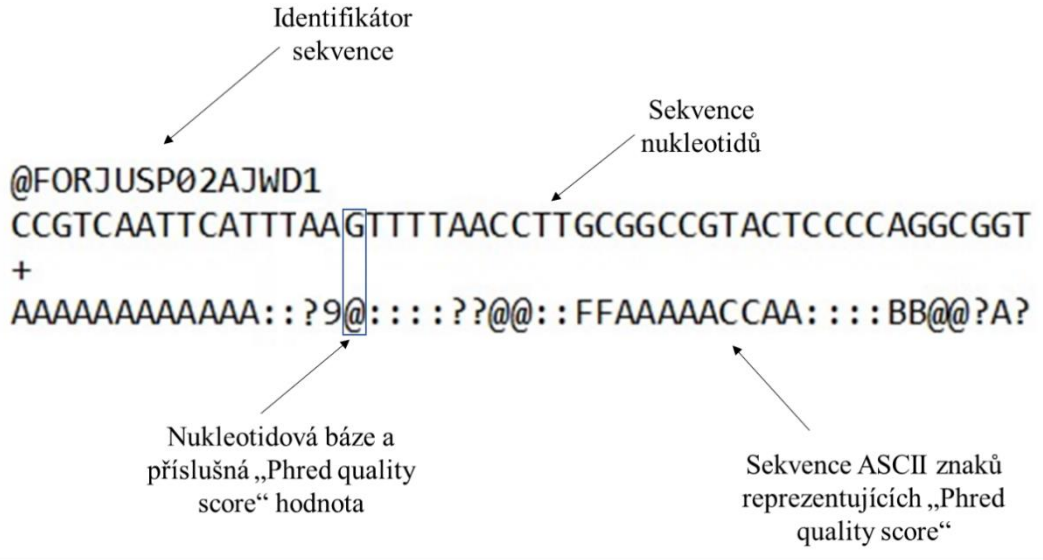
specifický pro každou bázi. Užívané póry mohou být biologického nebo syntetického původu a svým průměrem v řádu nanometrů jsou uzpůsobeny k tomu, aby umožnily průchod právě jednomu vláknu DNA (Wang *et al*., 2015). Sekvenaci předchází připojení adaptérů na fragmenty DNA umožňující následné ukotvení a průchod skrze nanopóry (Quick *et al*., 2014). Pro průchod je využíváno motorových enzymů a vzniklý rozdíl v membránovém napětí je detekován a zpracován (Obr. 6). Poté je s využitím počítačové analýzy zrekonstruována hledaná sekvence. Podobně jako metoda SMRT, rovněž „nanopore“ metoda umožňuje sekvenaci dlouhých fragmentů DNA (až 60 kb), nicméně vyznačuje se nižší přesností, která musí být kompenzována vícenásobným sekvencováním segmentů (Goodwin *et al*., 2015).

2.3 Kvalitativní bioinformatická analýza

2.3.1 Formáty v kvalitativní bioinformatické analýze

Bioinformatická analýza sekvenačních dat zahrnuje velké množství formátů velmi často specifických v závislosti na typu prováděné analýzy. Prvním široce používaným formátem je FASTQ, který je běžně užívaný k manipulaci s daty vzniklými sekvenací. Formát reprezentuje sekvence, které byly získány v průběhu sekvenačního procesu, a kromě nich také míru kvality s jakou byla každá báze sekvencována. Vzhledem k tomu, že formát nebyl dlouhou dobu formálně definován, vzniklo několik vzájemně

nekompatibilních variant, z nichž Sanger FASTQ formát patří k nejvíce používaným (Cock *et al*., 2010).



Obrázek 7: Struktura Sanger FASTQ formátu.

Převoditelnost mezi jednotlivými verzemi FASTQ formátu lze zajistit s pomocí různých softwarů, mezi které náleží např. Biopython, domovská stránka: https://biopython.org/ (20. 4. 2019).

Sanger FASTQ format byl poprvé vyvinut na přelomu tisíciletí (Cock *et al*., 2010) ve Wellcome Trust Sanger Institute Jimem Mullikinem. Data každé sekvence jsou reprezentována čtyřmi typy řádků (Obr. 7). První řádek začíná znakem ‚@‘, a obsahuje unikátní identifikátor zaznamenané sekvence a volitelný popis. Následují sekvenční řádky, obsahující samotnou sekvenci, skládající se z ASCII znaků, přičemž se řádky zpravidla pro přehlednost zalamují po určitém počtu znaků. Záznam o získané nukleotidové sekvenci je ukončen řádkem začínajícím znakem ‚+‘, který často obsahuje pouze tento znak, ale mohou zde být zapsány upřesňující informace. Poslední typ řádku koresponduje se sekvencí nukleotidů a označuje kvalitu (přesnost) sekvenace (Obr. 7).

K reprezentaci kvality získané sekvence je využíváno tzv. „Phred quality score“, založené na přesnosti, s jakou je Phred softwarem správně určena báze na základě hodnot získaných sekvenačním strojem (např. intenzity fluorescenčního záření). Phred hodnota je definována vztahem: $Q_{Phred} = -10 \times \log_{10}(P_e)$, kde P_e označuje pravděpodobnost chyby. K její reprezentaci je využíváno ASCII znaků, jejichž dekadický kód určuje přesnost (Cock *et al*., 2010). U Sanger FASTQ je využíváno znaků s kódem 33 až 126, tedy rozpětí 93 hodnot. To pokrývá pravděpodobnosti chyb od 1,0 (chybná báze) do $10^{-9,3}$ (přesně určená báze).

V roce 2004 přišla společnost Solexa se svojí verzí FASTQ formátu, která není kompatibilní se Sanger FASTQ formátem. Kvůli produkci souborů reprezentujícími kvalitu každé ze čtyř bází využívají jinou definici PHRED hodnoty, aby lépe pokryla velké pravděpodobnosti chyb. $Q_{Solexa} = -\log_{10}\left(\frac{P_e}{1-P_e}\right)$ je vzorec pro výpočet přesnosti pro každou bázi, kde P_e označuje pravděpodobnost chyby. Pro převod mezi jednotlivými PHRED hodnotami obou formátů platí vztahy: $Q_{Solexa} = 10 \times \log_{10}\left(10^{Q_{Phred}/10} - 1\right)$ $Q_{PHRED} = 10 \times \log_{10}\left(10^{Q_{Solexaa}/10} + 1\right)$. Solexa využívá ASCII znaků jejichž dekadický kód je 59 až 126 k reprezentaci hodnot -5 až 62 (Cock *et al.,* 2010).

V roce 2006 byla společnost Solexa převzata společností Illumina. Solexa FASTQ formát dále využívá, avšak přišla i s novým formátem, tzv. Illumina FASTQ formátem, který lze převést na Solexa FASTQ formát, avšak je znovu nekompatibilní se Sanger FASTQ formátem. Využívá ASCII znaků dekadického kódu od 64 do 126 ke kódování hodnot od 0 do 62 (Cock *et al.,* 2010).

Kromě FASTQ formátu existují další široce využívané formáty, mezi které náleží SAM a BAM formát. SAM je textový formát sloužící k uchování fragmentů biologických sekvencí a informací o jejich „alignmentu" (Handsaker *et al.,* 2009). BAM je ekvivalentní binární formát, který může uchovat stejnou informaci v komprimované podobě (SAM/BAM specifikace). Mapovací softwary často využívají SAM/BAM jako formát výstupních souborů, hlavně díky široké škále vlastností, které lze v souboru uchovat. Za účelem manipulace a tvorby SAM/BAM souborů byl vytvořen softwarový balíček Samtools, domovská stránka: https://samtools.github.io/ (20. 4. 2019). Každý SAM soubor lze rozdělit do dvou oddílů, záhlaví a sekci s „alignmenty" (SAM/BAM specifikace: https://samtools.github.io/hts-specs/SAMv1.pdf, navštíveno 20. 4. 2019).

Záhlaví není u těchto formátu povinné, ale pokud je v souboru obsaženo, tak musí být vždy na začátku. Řádky záhlaví slouží k popisu vlastností daného souboru. Základním nástrojem pro rozlišení informace v řádcích jsou dva typy dvoupísmenných „tagů" (Tab. 1). „Tag" označený znakem ‚@' slouží k určení kategorie informace – @HD obsahuje informace o základních vlastnostech souboru a „alignmentů", @SQ obsahuje informace o referenčních sekvencích, @RG obsahuje informace o „readech", @PG obsahuje informace o mapovacím softwaru a @CO slouží k volitelnému komentáři. Druhý typ „tagů" upřesňuje druh popisované vlastnosti dané kategorie (Tab. 1).

Tabulka 1: Příklady „tagů“ sloužících k roztřídění informací uchovaných v záhlaví souborů ve formátu SAM, spolu s jejich zařazení do kategorie a popisem informace, kterou označují.

Kategorie	Tag	Uchovaná informace
@HD	VN*	Verze formátu SAM
	SO	Styl seřazení „alignmentů“, platné hodnoty jsou – unknown pro neznámé, „unsorted“ pro neseřazené, „queryname“ pro lexikografické řazení podle QNAME a „coordinate“ pro řazení podle RNAME v pořadí @SQ řádků popisujících referenční sekvence
@SQ	SN*	Jméno referenční sekvence, hodnoty jsou využívány v RNAME a RNEXT polích v „alignment“ sekci
	LN*	Délka referenční sekvence
	AN	Alternativní jména pro referenční sekvence, které mohou být užity jinými nástroji. Tyto jména však nesmí být použita v „alignment“ sekci.
@RG	ID*	Identifikátor skupin readů. Každý řádek kategorie '@RG musí mít unikátní ID. ID je používáno v RG poli v „alignment“ sekci.
	DT	Datum spuštění programu.
@PG	ID*	Identifikátor záznamu programu. Každý řádek kategorie @PG musí mít unikátní ID. ID je používáno v PG poli v „alignment“ sekci
	PN	Jméno programu
@CO		Řádky sloužící k volitelnému komentáři.

*jsou označeny povinné „tagy“ v kategorii

Po sekci reprezentující záhlaví následuje „alignment“ sekce, kde obvykle jeden řádek reprezentuje lineární „alignment“ jednoho segmentu („readu“) a skládá se z 11 nebo více polí oddělených tabulátory, z nichž prvních 11 je povinných (Obr. 8). Pokud není nějaká hodnota pole známá, podle typu dané hodnoty je přiřazen znak ‚*‘ nebo ‚0‘. Všechny namapované segmenty se vyskytují na „forward“ vláknu, pokud je segment z „reverse“ vlákna, je uvedena jeho komplementární sekvence z „forward“ vlákna. Povinná pole jsou v tomto pořadí:

1. QNAME: Jméno DNA fragmentu, který byl sekvencován a z něhož pochází daná sekvence. Znak ‚*‘ označuje neznámé jméno.
2. FLAG: Celé číslo zapsané v dekadické soustavě, po převedení do binární soustavy každý bit označuje jednu vlastnost souboru.
3. RNAME: Jméno referenční sekvence daného „alignmentu“, pokud se v záhlaví nachází řádek kategorie @SQ, musí být RNAME napsáno pod „tagem“ SN alespoň v jednom řádku. Neznámá hodnota je označena znakem ‚*‘.
4. POS: Pozice na sekvenci první CIGAR hodnoty, která způsobí posunutí v referenční sekvenci, indexovaná od 1. Hodnotou je ‚0‘, pokud není známa.
5. MAPQ: Kvalita mapování. Definována jako $-\log_{10}(P)$, kde P označuje pravděpodobnost, že je segment špatně zmapovaný.
6. CIGAR: Sekvence znaků obsahující čísla a písmena, číslo označuje počet bází a následující znak vlastnost, kterou splňují. Hodnota ‚*‘ je použita, pokud nejsou známy (Tab. 2).
7. RNEXT: Jméno referenční sekvence následujícího „readu“. Pro poslední „read“ platí, že RNEXT je název referenční sekvence prvního „readu“. Pokud je RNEXT identické s RNAME, využívá se hodnota =, pokud je neznámé pak *.
8. PNEXT: Pozice v sekvenci dalšího „readu“, indexovaná od 1. Hodnota 0, pokud informace není známa. Informace je ekvivalentní hodnotě POS následujícího „readu“.
9. TLEN: Délka zmapované sekvence, namapovaný segment nejvíce nalevo má kladnou hodnotu, napravo má zápornou.
10. SEQ: Sekvence daného segmentu. Hodnota ‚*‘, pokud není sekvence známa. ‚=‘ značí báze identické k referenční sekvenci.
11. QUAL: PHRED „quality score“ pro každou bázi, ekvivalentní tomu ze Sangerova FASTQ formátu

```
@SQ     SN:Chromosome   LN:6413771
@SQ     SN:pCC7120alpha LN:408101
@SQ     SN:pCC7120beta  LN:186614
@SQ     SN:pCC7120delta LN:55414
@SQ     SN:pCC7120epsilon       LN:40340
@SQ     SN:pCC7120gamma LN:101965
@SQ     SN:pCC7120zeta  LN:5584
@PG     ID:TopHat       VN:2.1.1
M02925:111:000000000-AKGUY:1:1101:1724:15565    329     Chromosome
M02925:111:000000000-AKGUY:1:1101:1724:15565    89      Chromosome
M02925:111:000000000-AKGUY:1:1101:1724:15565    345     Chromosome
M02925:111:000000000-AKGUY:1:1101:1724:15565    329     Chromosome
M02925:111:000000000-AKGUY:1:1101:1757:16152    355     Chromosome
```

Obrázek 8: Ukázka záhlaví a středové části SAM souboru.

Kromě povinných polí, může být zahrnuta řada volitelných polí poskytující dodatečné informace o daném „alignmentu“. Všechna volitelná pole jsou ve formátu tag:typ:hodnota, kde „tag“ je dvoupísmenná zkratka, odlišná od „tagů“ užitých v záhlaví, každý se může vyskytovat v řádku pouze jednou. Typ je reprezentován jedním ze znaků – ‚A‘ pro vytisknutelný znak, ‚i‘ pro celé číslo, ‚f‘ pro desetinné číslo s jedním desetinným místem, ‚Z‘ pro vytisknutelný řetězec, ‚H‘ pro pole bytů v hexadecimálním formátu a ‚B‘ pro posloupnost čísel. Hodnota je poté ve formátu daného typu.

Mezi další formáty využívané v bioinformatické analýze náleží formáty výstupních souborů z bioinformatické analýzy. Jedním z nich je textový formát VCF sloužící k uchování informací o genových variacích daného organismu vůči referenční DNA. Ke každé pozici uvedené v souboru tohoto typu je možné přidat dodatečný popis. Ekvivalentním formátem v binární podobě je formát BCF, využívající binární kompresi dat. Tvorba souborů je nejsnadnější pomocí softwarového balíčku samtools, domovská stránka: https://samtools.github.io/ (20. 4. 2019), případně BCFTools, domovská stránka: https://samtools.github.io/bcftools/ (20. 4. 2019). Tyto balíčky vznikly za účelem tvorby

Tabulka 2: Přehled znaků v CIGAR sekvenci SAM formátu spolu s jejích popisem. Posunutí v sekvenci označuje skutečnost, kdy se při porovnání sekvenci posunuje pozice současně porovnávané báze.

Znak	Vlastnost	Posunutí v přiřazené sekvenci	Posunutí v referenční sekvenci
M	„alignment“ souhlasí	ano	ano
I	inserce vůči referenční sekvenci	ano	ne
D	delece vůči referenční sekvenci	ne	ano
N	přeskočený úsek z reference	ne	ano
S	„soft clipping“	ano	ne
H	„hard clipping“	ne	ne
P	„padding“	ne	ne
=	sekvence se shoduje	ano	ano
X	sekvence se neshoduje	ano	ano

```
##FORMAT=<ID=RDR,Number=1,Type=Integer,Description="Depth of
##FORMAT=<ID=ADF,Number=1,Type=Integer,Description="Depth of
##FORMAT=<ID=ADR,Number=1,Type=Integer,Description="Depth of
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
scaffold00001   1293    .       G       A       .       PASS
scaffold00001   1322    .       G       T       .       PASS
scaffold00001   1331    .       T       G       .       PASS
scaffold00001   1359    .       T       G       .       PASS
```

Obrázek 9: Ukázka záhlaví a středové části VCF souboru

a manipulace se SAM, BAM, VCF a BCF soubory. Data uložená v souboru lze rozdělit do dvou částí, část s meta-informacemi a část s daty o variacích.

Data uložená v části s genovými variacemi (meta-informacemi) jsou rozděleny do polí, z nichž každé má předem definované typy hodnot, které mohou nabývat. Nicméně formát dovoluje vytvářet i vlastní typy hodnot, jejichž popis spolu se základními informacemi o souboru se musí nacházet v části s meta-informacemi. Většina řádků má podobný formát, po znacích ‚##' se nachází název pole do kterého patří popisovaný typ hodnot. Popis je složen z párů (Obr. 9), které přesně definují typ hodnot.

Po části s meta-informacemi následuje část obsahující data o variacích. Každá variace je rozdělena do osmi povinných sloupců, a každý řádek popisuje jednu genovou variaci (Obr. 9).

Mezi povinné pole patří po řadě:

1. CHROM: Sekvence znaků sloužící jako identifikátor chromozomu z referenčního genomu. Pokud je ve formátu <ID>, odkazuje na id popsaný v části s meta-informacemi. Všechny položky se stejným CHROM musí být vzestupně seřazené podle POS. Povinný údaj, nesmí obsahovat dvojtečky ani bílé znaky.
2. POS: Nezáporné celé číslo označující pozici v referenčním genomu, kde dochází k variaci. V souboru se nesmí nacházet více položek se stejnou pozicí. Pro označení telomér slouží 0 a N+1, kde N je počet bází korespondujícího chromozómu. Povinný údaj.
3. ID: Sekvence znaků reprezentující identifikátor dané variace, středník může sloužit pouze k oddělení více identifikátorů. Jeden identifikátor nemůže být použit pro více položek. Pokud se jedná o SNP z dbSNP, může být použit identifikátor z databáze. Nepovinný údaj, nesmí obsahovat bílé znaky.
4. REF: Sekvence složena z písmen – ‚A' pro adenin, ‚G' pro guanin, ‚T' pro thymin, ‚C' pro cytosin, ‚N' pro neznámou bázi. Sekvence představuje báze z referenční sekvence, které podléhají variaci a první báze je na pozici POS. Pro

reprezentaci inzerce, nebo delece, při které by REF nebo ALT byly prázdné je nutné zapsat i bázi před variací. Povinný údaj.

5. ALT: Posloupnosti bází oddělených čárkami reprezentující alternativní báze k bázím referenčním. Může být ve stejném formátu jako REF, nebo <ID> kde ID je popsáno v části s meta-informacemi. Využívá se stejných znaků jako u REF a navíc znaku ‚*' reprezentujícího deleci. Nejsou povoleny bílé znaky a čárky.
6. QUAL: Číselná hodnota „Phred-scaled quality".
7. FILTER: Hodnota „PASS" pokud pozice prošla všemi filtry kvality a variace je platná. Pokud neprošla všemi filtry obsahuje kódy těchto filtrů oddělené středníky.
8. INFO: Slouží k dodatečném popisu každého záznamu. Středníky oddělené páry ve tvaru klíč:hodnota, jejich popis je v záhlaví.

Dále se zde může nacházet pole FORMAT, které slouží k dodatečnému popisu genomu. Hodnoty jsou ve tvaru ID=hodnota, kde ID i formát hodnoty je popsán v části s meta-informacemi.

2.3.2 Mapovací softwary

Sekvenací DNA, nebo RNA vznikají různě dlouhé „ready" složené z příslušných nukleotidů, které je nutné pro následnou analýzu sestavit do podoby, v jaké se nacházejí ve zkoumaném organismu. k tomuto účelu slouží mapovací softwary umožňující na základě porovnání fragmentů s referenčními sekvencemi vytvořit tzv. „alignment", tedy seřazení fragmentů tak, aby co nejvíce odpovídaly referenci (Conesa *et al*.,2016). Mezi hlavní dva přístupy mapování sekvenačních dat patří „unspliced align" a „spliced align" přístup (Schbath *et al*., 2012).

Tabulka 3: Přehled přístupů, metod a softwarů k mapování „readů", a jejich využití (upraveno dle Garber 2011).

Přístup	**Metoda**	**Software**	**Využití**
„Unspliced align"	„Seed"	SHRiMP	Mapování DNA „readů" a „readů" z exomového sekvenování.
	„Burrows-Wheeler transform"	Bowtie	
„Spliced align"	„Exon-first"	TopHat	Mapování RNA „readů".
	„Seed-extend"	GSNAP	

„Unspliced align“ nachází využití při mapování „readů“ vzniklých sekvencováním genomu (DNA „ready“), u kterých je známa referenční sekvence daného, nebo příbuzného organismu k přiřazení. Přístup totiž neumožňuje lokalizovat místa alternativního sestřihu a slouží hlavně ke kvantifikaci dat.

Dle typu zkoumaného organismu se „unspliced aligners“ rozdělují do dvou kategorií (Tab. 3). „Burrows-Wheeler transform“ metody tvoří první kategorii metod a jsou využívány, pro případy sekvencování, jejichž vyhodnocení nezahrnuje alternativní sestřih. Algoritmus těchto metod spočívá v převedení readů pomocí Burrows-Wheelerovy transformace na datovou strukturu. Takto transformované „ready“ lze poté velmi rychle porovnávat s referenční sekvencí (Garber *et al.,* 2011). Příkladem softwaru je Bowtie, domovská stránka: http://bowtie-bio.sourceforge.net/ (20. 4. 2019).

Druhou skupinu tvoří „Seed“ metody jejichž algoritmus je založen na rozdělení „readů“ na kratší úseky „semínka“ (z angl. „seeds“), přičemž se počítá s tím, že v každém „readu“ je alespoň jedno „semínko“, které lze přesně přiřadit k referenci. Tím se se určí alespoň přibližná pozice „readu“ na referenční sekvenci (Garber *et al.,* 2011). Na přiřazení zbytku „readu“ je poté využito jiných algoritmů, majících větší citlivost, ale jejichž využití na celý „read“ by bylo neefektivní, např. Smith-Watermanův algoritmus. Mezi software využívající tento typ metody patří SHRiMP, domovská stránka: http://compbio.cs.toronto.edu/shrimp/ (20. 4. 2019).

„Spliced align“ slouží k přiřazení „readů“ vzniklé sekvencováním celého transkriptomu. Jejich výhodou je identifikace nově vzniklých exonů, popřípadě „readů“ zasahujících do více exonů. Představené metody se rozdělují do dvou kategorií (Tab. 3).

První z nich je tvořena „Exon-first“ metodami. Jejich algoritmus je rozdělen do dvou částí. Nejprve jsou přiřazeny „ready“ zasahující pouze do jednoho exonu, pomocí „unspliced aligner“ metod. V druhém kroku jsou rozděleny zbylé „ready“ na krátké fragmenty a samostatně přiřazeny k referenčnímu genomu, následně jsou analyzovány části na koncích exonů, aby byly nalezeny spojení mezi těmito exony (Yang a Kim, 2015). Druhá část algoritmu je výrazně časově náročnější, proto jsou metody využívány hlavně v případech, kdy málo „readů“ zasahuje do více exonů (Garber *et al.,* 2011). Typickým příkladem „unspliced aligner“ softwaru je např. TopHat, domovská stránka: https://ccb.jhu.edu/software/tophat/ (20. 4. 2019).

Druhá skupina je tvořena metodami „Seed-extend“. Algoritmus těchto metod funguje podobně jako u „seed“ metod ve skupině „unspliced aligner“ přístupu. Všechny „ready“ jsou rozděleny na krátké fragmenty, které jsou následně přiřazeny k referenčními genomu

a tím jsou lokalizovány jejich pozice (Yang a Kim, 2015). Ke přiřazení zbylých úseků je využíván jiný algoritmus, který je více časově náročný. Následně jsou analyzovány „ready“, které byly rozděleny vlivem DNA sestřihu (Garber *et al.*, 2011). Příkladem softwaru je GSNAP, domovská stránka: https://www.gvst.co.uk/ (20. 4. 2019).

3 Experimentální část

3.1 Materiál a metody

3.1.1 Vstupní data pro analýzu

Pro bioinformatickou analýzu popisovanou v této práci byly použity data získaná v průběhu celotranskriptomového sekvencování sklerocia *Claviceps purpurea* industriálního kmene Gal404 (TEVA Czech Industries, Česká republika) a divokého kmene 20.1 (Centrum regionu Haná pro biotechnologický a zemědělský výzkum, Česká republika) rostoucích za standardních podmínek a posbíraných v roce 2014. Příprava knihovny byla provedena pomocí Illumina TruSeq Stranded mRNA Sample Preparation Kitu (Illumina, San Diego, CA, USA). Výsledná knihovna byla sekvencována s použitím MiSeq Reagent Kitu v3 (Illumina, San Diego, CA, USA) na přístroji MiSeq. Transkriptom byl sekvencován přístupem „paired-end“ za účelem zvýšení spolehlivosti při detekci jednonukleotidových polymorfismů a pro dosažení kvalitnější dat ze sekvencování.

3.1.2 Kontrola kvality a mapování “readů” na referenční genom

Kontrola kvality byla provedena s pomocí programu FastQC (verze 0.10.0), domovská stránka: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (20. 4. 2019). Proces mapování byl realizován pomocí programu TopHat2 (verze 2.0.12), domovská stránka: https://ccb.jhu.edu/software/tophat/index.shtml (20. 4. 2019). Sekvence byly mapovány na referenční genom *Claviceps purpurea* kmene 20.1 (Schardl *et al.*, 2013). Mapování sekvencí bylo provedeno příkazem s následující syntaxí:
tophat2 -o vystupní_adresář -G anotační_soubor.gff3 –p 10 -N 6 -read-edit-dist 6 indexovaný_genom R1.fastq R2.fastq

Parametry využité v tomto příkazu jsou shrnuty v následujícím přehledu:

-o	Výstupní adresář.
-G	Anotační soubor ve formátu GFF3.
-p	Počet jader CPU využitých v mapovacím procesu.
-N	Maximální počet neshodujících se nukleotidů v jediném “readu” pro jeho mapování na referenční genom.

-read-edit-dist	Maximální počet chybějících bazí v jediném „readu“.
indexovaný_genom	Index referenčního genomu.
R1.fastq, R2.fastq	FASTQ soubory obsahující „ready“.

3.1.3 Kvalitativní analýza transkriptomu

K vytvoření vstupních VCF souborů, obsahujících informace o genetických variacích ve zkoumaném organismu, bylo využito volně dostupných softwarů VarScan2, domovská stránka: http://dkoboldt.github.io/varscan/ (20. 4. 2019), a BCFTools, domovská stránka: https://samtools.github.io/bcftools/ (20. 4. 2019). Jelikož se algoritmy obou softwarů liší, výstupem jsou dva soubory obsahující data, která nejsou úplně totožná. Tab. 4 obsahuje popis jednotlivých parametrů využitých v příkazech.

Příkaz užitý pro vytvoření VCF souboru prostřednictvím softwaru BCFTools:

bcftools mpileup -f ref_soubor.fa alignment_soubor.bam -o vystupni_soubor.vcf

Software VarScan2 neumožňuje vytvoření VCF souboru, který by obsahoval data jak o

Tabulka 4: Popis parametrů, využívajících se v příkazech pro tvorbu VCF souboru.

Software	Parametr	Funkce parametru
BCFTools	*-o*	Výstupní soubor.
	-f	Indexovaný soubor s referenčními sekvencemi.
BCFTools a Samtools	*mpileup*	Funkce pro vytvoření souboru s daty o genomových variacích.
Samtools	*--reference*	Soubor s referenčními sekvencemi.
VarScan2	*mpileup2snp*	Funkce pro vytvoření souboru s genomovými substitucemi.
	mpileup2indel	Funkce pro vytvoření souboru s genomovými inzercemi a delecemi.
	--min-coverage	Minimální hloubka „readu“ na konkrétní pozici potřebná pro určení genomové variace.
	--min-reads2	Minimální množství podporujících „readů“ na konkrétní pozici potřebné pro určení genomové variace.
	--min-var-freq	Minimální frekvence variace.
	--p-value	Nejmenší hladina významnosti variace.
	--output-vcf	Výstup ve formě VCF souboru.

substitucích, tak o inzercích a delecích jedním příkazem. Bylo tedy nutné vytvořit dva soubory, využíván byl software samtools k vytvoření vstupního MPILEUP souboru.

Příkazy byly zadány v následujícím tvaru:

samtools mpileup --reference ref_soubor.fa alignment_soubor.bam | java -jar VarScan.jar mpileup2snp --min-coverage 2 --min-reads2 2 --min-var-freq 0.05 --p-value 0.05 --output-vcf 1 vystupni_soubor.vcf

samtools mpileup --reference ref_soubor.fa alignment_soubor.bam | java -jar VarScan.jar mpileup2indel --min-coverage 2 --min-reads2 2 --min-var-freq 0.05 --p-value 0.05 --output-vcf 1 vystupni_soubor.vcf

Data získaná programem BCFTools byla následně filtrována, aby bylo docíleno shodných parametrů jako v případě programu VarScan. Nukleotidové modifikace vykazující pokrytí menší než 4x, byly z výstupního souboru odstraněny. Po filtrování byla pro výsledné VCF z programu annovar provedena analýza jednonukleotidových polymorfismů na organismus pomocí implementovaného softwaru a komerčně dostupného programu ANNOVAR.

Vliv jednonukleotidových polymorfismů na organismus lze pomocí vytvořeného softwaru zjistit jediným příkazem, avšak je nutné mít k dispozici všechny potřebné vstupní soubory. Po spuštění hlavního souboru *snp_analyze.py* je uživatel dotázán na umístění GFF3, VCF a FASTA souboru. Následně proběhne zpracování těchto souborů a poté stačí v hlavní nabídce zadat možnost 7, *Zjištění vlivu nukleotidových variací na celý organismus*. Tímto příkazem je vytvořen v uživateli vybraném adresáři FASTA soubor s alternativními nukleotidovými sekvencemi vzniklými vlivem jednonukleotidových polymorfismů. Mimoto je v hlavním adresáři vytvořen soubor *log.txt*, který obsahuje popis průběhu programu. Toho lze využít při srovnávání s výsledky jiných softwarem.

Za účelem porovnání výsledků se softwarem ANNOVAR, domovská stránka: http://annovar.openbioinformatics.org/ (20. 4. 2019) bylo potřeba zpracovat log soubor do podoby textového souboru, jehož řádky obsahují počet variací v každém genu. Toho bylo docíleno možností 8, *Zpracování log souboru,* v hlavní nabídce vytvořeného softwaru a následně vybráním možnosti 1, *Soubor s počtem variací podle genomových oblastí*. Příkaz využitý pro generování anotace programem ANNOVAR je následující:

table_annovar.pl input.vcf databaze/ -buildver CP -out myanno -remove -protocol refGene -operation g,r -nastring . -vcfinput

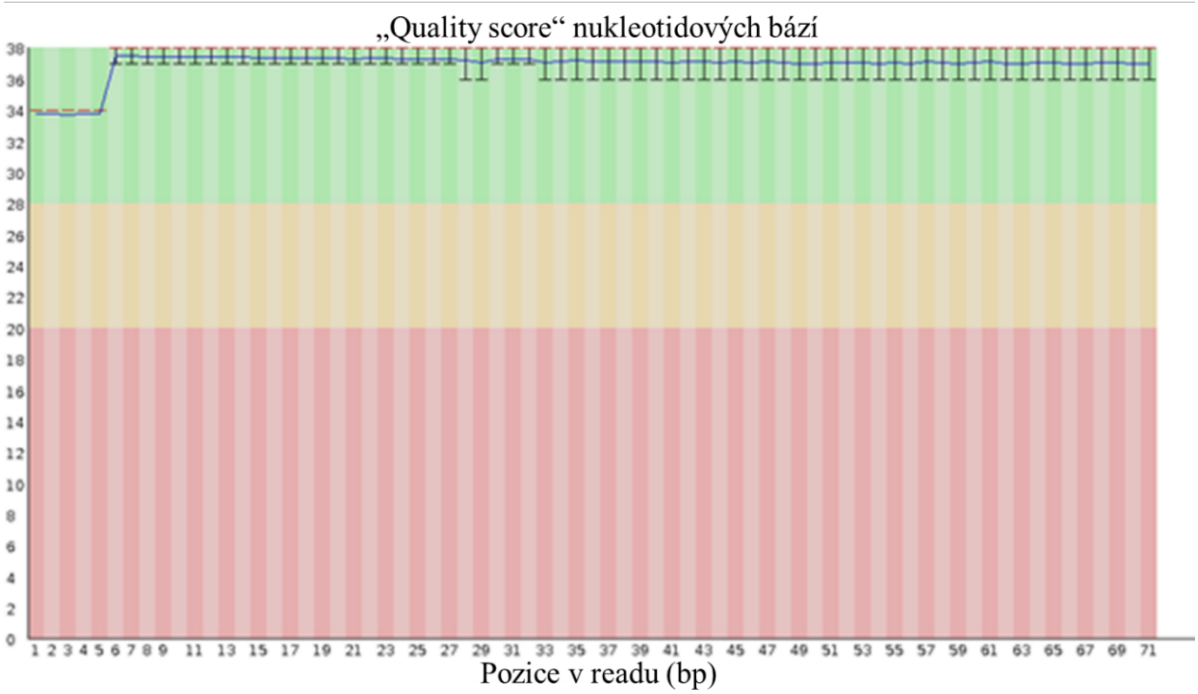
3.1.4 Zhodnocení zarovnání "readů" a kvality sekvenačních dat

Data získaná ze sekvenátoru Illumina MiSeq v rámci sekvencování transkriptomu *Claviceps purpurea* Gal 404 a divokého kmene 20.1 byla analyzováná s využitím programu FastQC (verze 0.10.0). Získané základní charakteristiky z datových souborů ukázaly, že u obou vzorku byla délka readů stanovena na 51 bp. Ve zkoumaných datových souborech nebyly nalezeny "ready" s odlišnou délkou "readů" než je tato hodnota. Celkový počet získaných readů z obou souborů byl stanoven na 9 414 479 pro kmen Gal 404 a 10 119 211 pro divoký kmen 20.1.

Celkový procentuální obsah bází GC byl stanoven na 52 % pro divoký kmen 20.1 a 51 % pro Gal 404. Uvedená hodnota je shodná s obsahem GC bází v genomu *Claviceps purpurea* uváděným v literatuře (Schardl *et al.*, 2013). Při porovnání těchto hodnot s obsahem uváděným pro kódující sekvence (GC 55 %; Schardl *et al.*, 2013), se hodnota liší o 3-4 %. Detekovaná odlišnost ovšem není takového rozsahu, aby data nemohla být v tomto ohledu interpretována jako kvalitní.Jednou ze základních charakteristik hodnotících kvalitu sekvenačních "readů" je „Phred quality score", které je přímo úměrné kvalitě báze na dané pozici v "readu" (Obr.10).

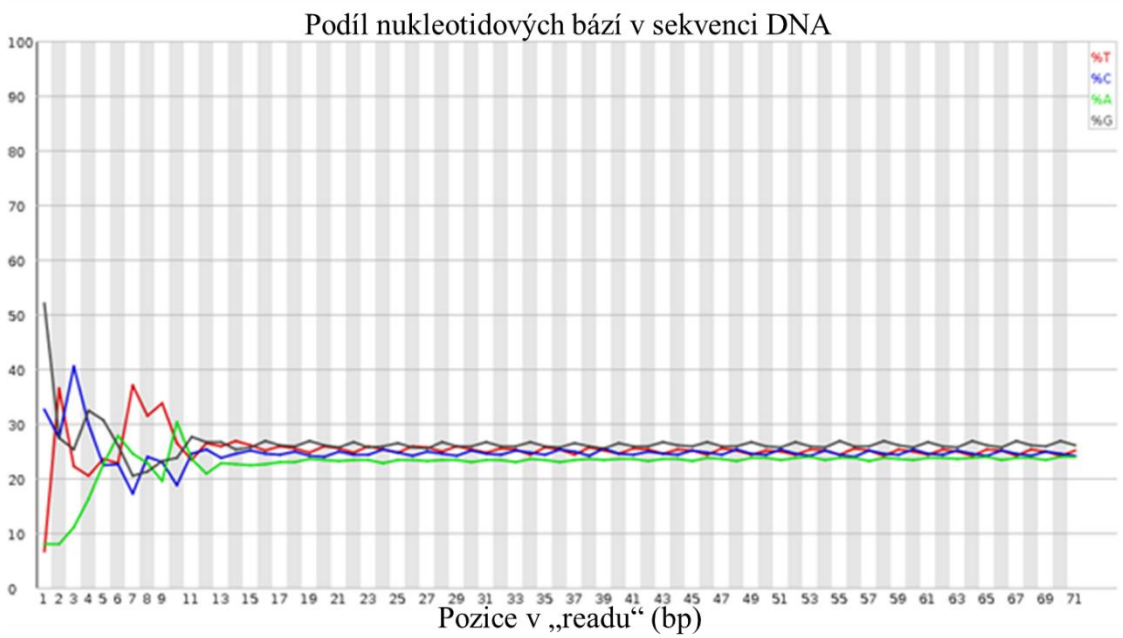
Skutečnost, že všechny ready se vyznačují vysokou kvalitou a analogický výsledek byl zjištěn, také pro divoký kmen 20.1. prezentovaném grafu se vyskytují tři barevně odlišné oblasti s vysoce kvalitními bázemi nacházejícími se v zelené oblasti. Data prezentována pro kmen Gal 404 (Obr. 10) zobrazují, že „ready" lze považovat za kvalitní v celé délce, vzhledem ke skutečnosti, že žádná báze nezasahuje do červené ani do žluté oblasti grafu což by znamenalo její sníženou kvalitu. Pro kmen 20.1 byly nalezeny podobné výsledky jako u kmene Gal 404.

Další možností, jak zhodnotit kvalitu sekvenačních "readů" je poměr mezi jednotlivými druhy nukleotidů ze kterých jsou "ready" tvořeny. Ukázka obsahu bází pro kmen Gal404 je zobrazen na Obr. 11.



Obrázek 10: Zobrazení „Phred skóre“ pro kmen *Claviceps purpurea* Gal 404 jako funkce pozice báze v „readu“. u každé báze je centrální červená čára zobrazující medián, žlutý sloupec zobrazující dolní a horní kvartil a úsečky zobrazující minimum a maximum. Modrá čára představuje průměrnou hodnotu „Phred skóre“. Žlutá a červená oblast grafu značí sníženou, respektive špatnou kvalitu.

Procentuální obsah jednotlivých nukleotidů ukazuje u bází 1-12 v “readu” anomálie v obsahu jednotlivých druhů nukleotidů. Ve zkoumané oblasti grafu se vyskytuje téměř



Obrázek 11: Procentuální obsah nukleotidů pro kmen Gal 404 *Claviceps purpurea* přítomných ve studovaných „readech“ v závislosti na poloze báze v „readu“.

30% odchylka ve srovnání s očekávanou hodnotou, která by se měla pohybovat kolem 25 %. Tato zhoršená kvalita bází vyskytující se na začátku „readů“ je častou komplikací pro sekvenátory firmy Illumina (Hansen *et al.*, 2010). U druhého studovaného kmene byly v tomto ohledu nalezeny obdobné výsledky.

Poslední charakteristikou zkoumanou v rámci hodnocení kvality bylo zkoumání obsahu sekvenačních adaptérů, které byly použity v průběhu přípravy sekvenační knihovny, a které by měly být odstraněny v průběhu zpracování surových dat přístrojem Illumina, což je zprostředkováno vestavěným softwarem. U žádného ze zkoumaných kmenů nebyla v tomto ohledu nalezena žádná kontaminace sekvenačními adaptéry. Na základě uvedených skutečností, lze usoudit, že získaná data vykazují dostatečnou kvalitu pro následnou bioinformatickou analýzu.

Bezprostředně po provedení kontroly kvality u studovaných “readů” bylo provedeno zarovnání „readů“ na referenční genom s pomocí programu TopHat2 (v. 2.0.12). u divokého kmene 20.1 bylo docíleno úspěšného zarovnání u 82,5 % (z celkového počtu 10 119 211) vstupních “readů”. Podobného výsledku bylo docíleno také v případě kmene Gal 404, kde bylo úspěšně namapováno 83,2 % (z celkového počtu 9 414 479). Pro ověření úspěšnosti mapování se používá podíl “readů”, které se vyznačují vícenásobným zarovnáním v referenčním genomu. Tento podíl by měl dosahovat minimálních hodnot, s přihlédnutím ke skutečnosti že některé “ready” mohou být zarovnány na více míst v referenčním genomu z důvodu výskytu repetitivních sekvencí. U vzorku divokého kmene 20.1 bylo nalezeno 2,0 % “readů”, pro kmen Gal 404 1,8 % “readů”, které vykazovaly vícenásobné zarovnání na referenční genom. Obě tyto hodnoty jsou vzhledem k celkovému počtu “readů” zanedbatelné a lze tedy v tomto ohledu usoudit, že se jedná o vysoce kvalitní mapování. Poněkud nižší počet zarovnaných „readů“ (80-85 %) je pravděpodobně způsoben neúplností referenčního genomu, a nikoliv nevyhovující kvalitou mapovaných readů.

3.1.5 Popis vývoje softwaru

Software byl vytvářen ve vývojovém prostředí Microsoft Visual Studio 2017 na počítačové platformě s operačním systémem Microsoft Windows 10. Využity byly dva programovací jazyky – Python 3.6, ve kterém je napsána převážná část zdrojového kódu a jazyk SQL, jenž je využíván pro uložení a práci s daty vzniklými zpracováním velkých souborů s biologickými informacemi.

Hlavní funkcí programu je zprostředkování přehledného popisu změn v genomu, zapříčiněnými jednonukleotidovými polymorfismy, případně inzercemi nebo delecemi spolu s jejich vlivem na translatované proteiny. Software lze rovněž využít i k zjištění základních informací o struktuře genomu např. pozice genů, exonů a intronů referenčního organismu. Dále také umožňuje uživateli zpracovávat samotné výstupní soubory (převedení nukleotidových sekvencí ve FASTA souboru do aminokyselinových sekvencí, popřípadě kvantifikace variací zpracováním log souboru). Výsledků je docíleno zpracováním dat ze tří vstupních souborů:

1. Soubor formátu GFF3 – obsahuje informace o struktuře genomu zkoumaného organismu, např. pozice genů a exonů v DNA.
2. Soubor formátu VCF – obsahuje pozici a podobu každé genomové variace ve zkoumaném organismu.
3. Soubor formátu FASTA – obsahuje DNA sekvence referenčního organismu.

Hlavním typem výstupního souboru je FASTA soubor, který obsahuje DNA nebo aminokyselinové sekvence vzniklé vlivem variací v genomu. Některé výsledky, především informativní, jsou zobrazeny v textové podobě v příkazovém řádku. Vedlejším výstupním souborem je tzv. log soubor, který obsahuje stručný popis a počet variací, nacházejících se v každém genu.

Programovací jazyk Python umožňuje rozčlenění zdrojového kódu do jednotlivých souborů tzv. modulů. Moduly slouží k lepší přehlednosti zdrojového kódu, každý modul obsahuje funkce vztahující se k určité problematice. Funkce definované v jednom modulu mohou být použity napříč ostatními. Samotný software se skládá ze sedmi modulů:

1. snp_analyze.py – hlavní modul, který slouží ke spuštění celého softwaru.
2. vcf_apps.py – modul obsahující funkce pro zpracování VCF souboru, hlavní funkcí je *process_vcf* umožňující uložení variací, nacházejících se v kódujících sekvencích organismů, do databáze. Pro práci s databází je využito zabudovaného modulu sqlite3, jenž umožňuje využití jazyka SQL.
3. gff_apps.py – modul obsahující funkce pro zpracování GFF3 souboru, hlavní funkcí je *process_gff* umožňující uložení informací z GFF3 souboru do datových typů Pythonu, kvůli rychlejšímu využívání dat v chodu programu.

4. snp_apps.py – modul s funkcemi umožňujícími splnění hlavních cílů programu, funkce *show_snp*, *show_sequence* a *show_protein* tvoří jeho hlavní kostru.

5. protein_apps.py – modul s funkcemi pro převod nukleotidových sekvencí do sekvencí aminokyselin.

6. fasta_apps.py – modul obsahující funkce pro práci s FASTA soubory, hlavní funkce *fasta_specific_read* slouží k přečtení určitých pozic z FASTA souboru s referenčními sekvencemi.

7. sequence_apps.py – modul s funkcemi upravujícími nukleotidové sekvence, např. *plus_strand* k převedení „reverse“ vlákna DNA na „forward“.

8. log_apps.py – modul s funkcemi sloužícími ke zpracování log souborů vytvořených pomocí tohoto softwaru.

Celý software je poskytnut jako elektronická příloha k bakalářské práci (Příloha 1). Manuál k implementovanému softwaru je popsán v následující kapitole.

4 Výsledky a diskuze

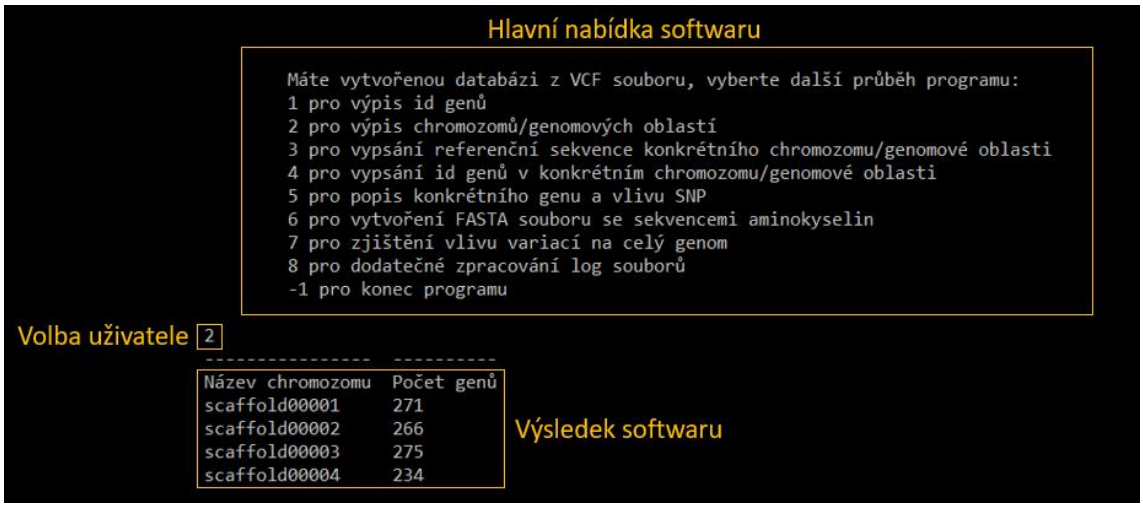
4.1 Manuál pro správné užití softwaru

Pro správnou funkčnost softwaru je třeba mít nainstalovaný Python kompilátor. Toho lze docílit instalací balíčku, volně stažitelného z oficiálních stránek jazyka Python: https://www.python.org/downloads/ (20. 4. 2019), pomocí instrukcí, které jsou na stránkách poskytnuty.

Pro správně zobrazení textu je vyžadován balíček Python modul Tabulate, domovská stránka: https://bitbucket.org/astanin/python-tabulate/src/master/ (20. 4. 2019). Instalace probíhá spuštěním příkazu *pip install tabulate* v příkazovém řádku. Dále stačí otevřít hlavní soubor programu *snp_analyze.py*. Při běhu softwaru několikrát dochází k vytvoření databází, souborů s koncovkou .db, tyto soubory lze otevírat např. pomocí open source programu SQLiteStudio, domovská stránka: https://sqlitestudio.pl/index.rvt (20. 4. 2019), v případě nesmyslných výsledků softwaru je doporučeno smazat databáze a nechat software vytvořit nové.

Uživateli je ovládání softwaru umožněno pomocí jednoduchého grafického rozhraní v prostředí příkazového řádku. Běh programu lze volit vždy zadáním daného znaku ze zobrazené nabídky (Obr. 12).

Při prvním spuštění softwaru je uživatel dotázán na umístění vstupních souborů. Pokud se jedná o dosud nevyužité soubory, dojde k vytvoření databáze obsahující zpracovaná data z VCF souboru. Postupný průběh je uživateli vypsán na obrazovce v podobě aktuálně zpracovávané hodnoty sloupce CHROM ve VCF souboru, vytvoření databáze je uživateli oznámeno v hlavní nabídce (Obr. 12).

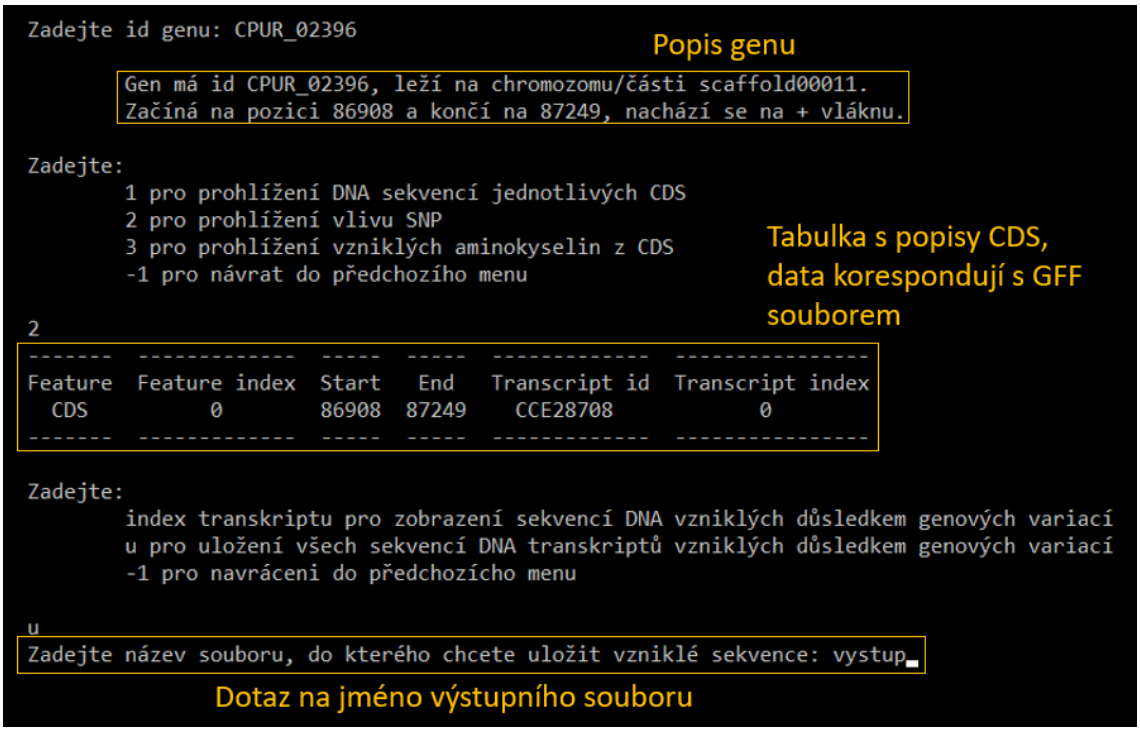


Obrázek 12: Hlavní nabídka implementovaného softwaru spolu s výpisem chromozomů.

Hlavní nabídka umožňuje výběr z osmi funkcí, dle jejich typu lze program rozdělit na tři části – popis genomu a vlivu variací na jednotlivé geny (možnosti 1, 2, 3, 4 a 5), vliv genových variací na celý organismus (možnost 7) a dodatečné zpracování výstupních souborů programu (možnosti 6 a 8) Znak ‚-1' slouží k ukončení programu. V následujícím textu jsou popsány všechny funkce programu (celý název funkce je vždy zapsán kurzívou), v závorce u jména funkce je znak, který slouží ke spuštění funkce v aktuální nabídce.

Funkce pro popis genomu a vlivu variací na jednotlivé geny jsou následující:

- *Výpis id genů* (1): Funkce slouží k vypsání všech identifikátorů genů v genomu na obrazovku. Uživatel si může vybrat kolik hodnot má být vypsáno v jednom řádku. Údaje odpovídají genovým identifikátorům, které jsou využity v GFF3 souboru.
- *Výpis chromozomů/genomových oblastí* (2): Funkce umožňuje výpis chromozomů, případně jiných částí, do kterých je rozdělen genom. Hodnoty korespondují s rozdělením genomu, dle vstupních souborů. Data se vypíší na obrazovku v podobě dvou sloupců, první označuje název chromozomu/části, druhý poté odpovídající počet genů.
- *Výpis referenční sekvence konkrétního chromozomu* (3): Funkce pro vypsání nukleotidové sekvence určitého chromozomu/části genomu na obrazovku. Jedná se o údaj ze vstupního FASTA souboru. Formát výpisu je ve standardní podobě FASTA souboru (standardně 70-80 znaků na řádek).
- *Výpis id genu v konkrétním chromozomu* (4): Funkce, jež vypíše identifikátory genů určitého chromozomu/části genomu. Uživatel musí zadat název



Obrázek 13: Ukázka softwarové funkce *Popis konkrétního genu a vlivu genových variací.*

chromozomu/části genomu ve tvaru v jakém se hodnota nachází ve výpisu pomocí funkce *Výpis chromozomů/genomových oblastí*, také si může zvolit počet hodnot na řádku. Znak ‚-1' slouží k navrácení do předchozí nabídky.

- *Popis konkrétního genu a vlivu genových variací* (5): Funkce zobrazující stručný popis struktury určitého genu a následně vypíše další funkce vztahující se k němu. Uživatel musí zadat gen ve tvaru v jakém se nachází ve výpisu pomocí funkce *Výpis id genů*, nebo *Výpis id genu v konkrétním chromozomu*. Popis struktury obsahuje název genu, jeho pozici v genomu a vlákno, na kterém se nachází kódující sekvence (Obr. 13). Znak ‚-1' slouží k navrácení do předchozí nabídky. Dále jsou uživateli nabídnuty tyto funkce:
  - *Prohlížení DNA sekvencí* (1): Funkce zobrazující DNA sekvenci kódující sekvence daného genu. Uživateli je zobrazena tabulka s informacemi o jednotlivých kódujících sekvencích (pozice, název transkriptu, indexy CDS a transkriptů), následně je možno vybrat mezi výpisem DNA sekvence jedné CDS zadáním znaku ‚1' a následně zadáním indexu transkriptu a CDS, nebo sekvence DNA konkrétního transkriptu zadáním znaku ‚2' a následně indexu transkriptu. Data v tabulce korespondují s informacemi z GFF3 souboru. Znak -1 slouží k navrácení do předchozí nabídky.

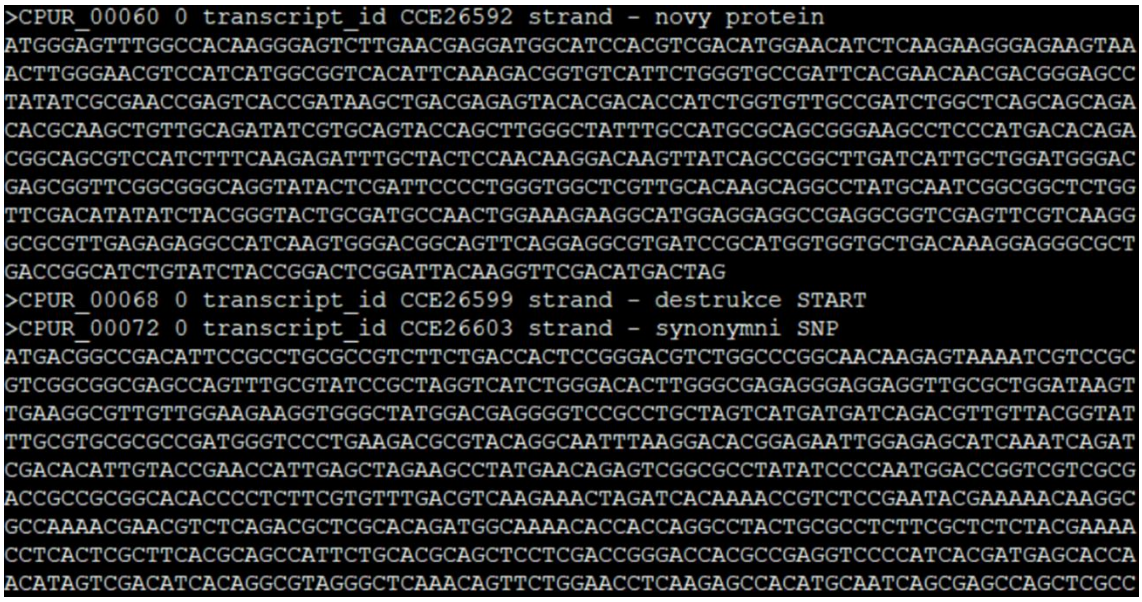
CHROM	GENE ID	TRANSCRIPT ID	SUBS	DELS	INS	INFO	
scaffold00001		CPUR_00005.0	CCE26537	0	2	0	destrukce STOP
scaffold00001		CPUR_00006.0	CCE26538	0	0	0	synonymni SNP
scaffold00001		CPUR_00007.0	CCE26539	0	1	0	novy protein
scaffold00001		CPUR_00014.0	CCE26546	0	1	0	novy protein
scaffold00001		CPUR_00015.0	CCE26547	0	0	0	synonymni SNP
scaffold00001		CPUR_00018.0	CCE26550	0	0	1	novy protein
scaffold00001		CPUR_00019.0	CCE26551	0	1	1	novy protein
scaffold00001		CPUR_00020.0	CCE26552	0	0	0	synonymni SNP
scaffold00001		CPUR_00022.0	CCE26554	0	3	0	novy protein
scaffold00001		CPUR_00023.0	CCE26555	0	1	0	novy protein
scaffold00001		CPUR_00024.0	CCE26556	0	1	0	novy protein

Obrázek 14: Ukázka dat v log souboru.

- *Prohlížení vlivu nukleotidových variací* (2): Funkce sloužící pro výpis nukleotidových variací a sekvencí DNA vzniklých důsledkem nukleotidových variací na gen. Uživateli je zobrazena tabulka s informacemi o jednotlivých kódujících sekvencích (pozice, název transkriptu, indexy CDS a transkriptů), následně je možno vybrat mezi výpisem variací a vzniklých sekvencí jednoho transkriptu DNA zadáním indexu transkriptu, nebo uložením všech alternativních sekvencí transkriptů, vzniklých důsledkem genových variací, zadáním znaku u. Sekvence jsou uloženy ve FASTA formátu do adresáře určeného uživatelem. Znak -1 slouží k navrácení do předchozí nabídky.
- *Prohlížení aminokyselin* (3): Funkce umožnuje výpis sekvencí aminokyselin, které jsou kódovány transkripty v genu. Aminokyseliny jsou reprezentovány jejich jednopísmennými zkratkami, stop kodon je reprezentován znakem *. Znak -1 slouží k navrácení do předchozí nabídky.

Funkce pro zjištění vlivu nukleotidových variací na celý organismus:

- *Zjištění vlivu nukleotidových variací na celý organismus* (7): Funkce vytvoří výstupní FASTA (obr. 15) soubor (pokud uživatel nezadá jinak) a log soubor (Obr. 14). FASTA soubor obsahuje všechny DNA sekvence vzniklé v důsledku jednonukleotidových polymorfismů, delecí a inzercí. Umístění výstupních souborů je voleno uživatelem. Dle formátu FASTA (Obr. 15) souborů patří ke každé sekvenci popisný řádek, který v tomto pořadí obsahuje – identifikátor genu, identifikátor transkriptu společně s číslem udávajícím



Obrázek 15: Ukázka dat ve výstupním FASTA souboru s DNA sekvencemi.

číslo sekvence v rámci transkriptu (v důsledku více možných variací na jedné pozici, dochází ke vzniku více sekvencí pro jeden transkript), druh vlákna DNA, na kterém se nachází kódující sekvence (+ nebo -) a nakonec stručný popis vlivu variací (např. nový protein, synonymní SNP, nebo destrukce start, popřípadě stop kodonu). Log soubor obsahuje data rozdělené tabulátory do sloupců. První řádek obsahuje definice hodnot v daných sloupcích – CHROM pro název chromozomu/části genomu, GENE_ID pro identifikátor genu, TRANSCRIPT_ID pro identifikátor trankriptu, SUB pro počet substitucí, DEL pro počet delecí, INS pro počet inzercí a INFO pro dodatečný popis. Každý následující řádek obsahuje kvantifikaci variací pro daný gen a transkript.

Dodatečné zpracování výstupních souborů umožňují následující funkce:

- *Vytvoření FASTA souboru aminokyselinovými sekvencemi* (6): Funkce, která zpracuje FASTA soubor se sekvencemi DNA a vytvoří korespondující soubor s aminokyselinovými sekvencemi, jež jsou kódovány sekvencemi z prvního souboru (Obr. 16). Uživatel musí zadat umístění vstupního souboru. Umístění výstupních souborů je voleno uživatelem. Vstupní soubor by měl být vytvořený pomocí tohoto softwaru. Sekvence aminokyselin jsou popsány stejně jako byly DNA sekvence vstupního souboru, pro reprezentaci aminokyselin je využíváno standardních jednopísmenných zkratek a pro stop kodon znak *.

```
>CPUR_00090 0 transcript_id CCE26621
MSIPGLGQITAQTVALPTRKISLKPFWEWRFQVPIGQTVVLKILSGSAEKDGTELAVQNAYSFTGIKSKILTWHGCELEI
EGQTEEESVATYASPQENPATSHINLHDWLGEMRAAAAREKCEGPRVLVAGPAATGKTTLVRTLTSYATRQGFQPVVVSA
DPKEGMLSLPGSLTASVFATIMDPESVDGWGSTPTSGPSVVPVKLPLVYYYGHSSPESDIDFYRELTCIIAGTVSGRFSE
DEEVRSSGVIIDSMGVGEDSSVGMDLLAHVVDEMSVNIIVVIGSTTMSNEFSKRFAGEQTSLGEPIHVIGLDRSTGAVER
SEAFLEHAREQAIKEYFFGDARTTLCPQIQQVEFESLVIYKASECPPGEQSSLTRHDPSSLMQHWTLAVMHASLEDAPDV
VRASTVMGFVYVSDVDEEKRRVKLLAPVGGRIDNRPLVFGRWPEPFMNLLG*
>CPUR_00092 0 transcript_id CCE26623
MGRSLHARFKRWDGMESDTSGEQQDAQGGWGSCSPRTSKEVVGRGLGWPSVLQFWERHKKYKKAFLARSLEATDSVSVSN
YPTQLLTPEYSTSNTFFRKTTNTLFTLSKCLPVAAHLAPAPAAAIVAPARAAATKLPSQCANPPPSARHWEATTGRTSTG
PSRSDEGPMAMQSHVQFIKIVRYLP*
>>CPUR_00101 0 transcript_id CCE26632
MAQGPPAGDGPPMNYERKQSQPENPFAELIPDQQIAIVPEFTLESGITLHNVPVAYTTRGKLNEEGNNVMVICHALTGSA
DLSDWWGPLLGGPGRVFDTSRFFIVCMNSLGSPYGTASPVTAKDNDASSGRYGPEFPLTTIRDDVNLHKLLLDDLGVKQI
```

Obrázek 16: Ukázka dat ve výstupním FASTA souboru s aminokyselinovými sekvencemi.

- *Zpracování log souboru* (8). Funkce pro kvantifikaci dat z log souborů. Znak -1 slouží k navrácení do předchozí nabídky. Uživatel má na výběr ze tří druhů výstupů:
    - *Soubor s počtem variací podle genomových oblastí* (1): Funkce sloužící k vytvoření nového souboru, jenž obsahuje data rozdělená tabulátory do sloupců. Na prvním řádku jsou definovány hodnoty jednotlivých sloupců – GEN pro název chromozomu/oblasti genomu, INS pro počet inzercí, DEL pro počet delecí, SUB pro počet substitucí. Každý řádek slouží ke kvantifikaci nukleotidových variací pro jeden chromozom/oblast genomu. Uživatel musí zadat umístění log souboru vytvořeného tímto softwarem. Umístění výstupního souboru je voleno uživatelem.
    - *Soubor s počtem všech variací v organismu* (2): Funkce sloužící k vytvoření nového souboru, jenž obsahuje data rozdělená tabulátory do sloupců. Uživatel musí zadat umístění VCF a GFF3 souboru, na základě jejichž zpracování jsou data získána. Umístění výstupního souboru je voleno uživatelem. V řádcích souboru jsou obsaženy data o počtech jednotlivých variací v odlišných oblastech genomu.
    - *Soubor s porovnáním dvou log souborů* (3): Funkce slouží k vytvoření nového souboru obsahujícího porovnání dat ze dvou log souborů vzniklých funkcí *Zjištění vlivu nukleotidových variací na celý organismus* nebo dvou souborů vzniklých funkcí *Soubor s počtem variací podle genomových oblastí*, uživatel si zvolí typ vstupního souboru. Uživatel musí zadat umístění dvou vstupních souborů a adresář, do kterého se má uložit výstupní soubor. Data jsou ve

výstupním souboru uložena do tabulátory oddělených sloupců. První řádek obsahuje názvy dvou vstupních log souborů. Druhý řádek slouží k definici hodnot ve sloupcích. – CHROM pro název chromozomu/oblasti genomu, SUB pro počet substitucí, INS pro počet inzercí, DEL pro počet delecí. Následující řádky obsahují pro každý chromozom/oblast genomu kvantifikaci variací pro oba log soubory.

4.2 Detekce nukleotidových modifikací

Po provedení kontroly kvality a následném mapovaní „readů“ na referenční genom byla provedena kvalitativní analýza transkriptomu dvou kmenů *Claviceps purpurea* (kmen 20.1 a Gal404). Analýza zahrnovala detekci jednonukleotidových polymorfismů, inzercí a delecí (souhrnně modifikací DNA) pomocí dvou softwaru (VarScan domovská stránka: http://dkoboldt.github.io/varscan/ navštíveno 20. 4. 2019 a BCFTools domovská stránka: https://samtools.github.io/bcftools/ navštíveno 20. 4. 2019), které jsou určeny pro tento účel. Jednotlivé události nalezené s pomocí těchto softwarů byly následně kvantifikovány a shrnuty do tabulky (Tab. 5; Přílohy 2-5).

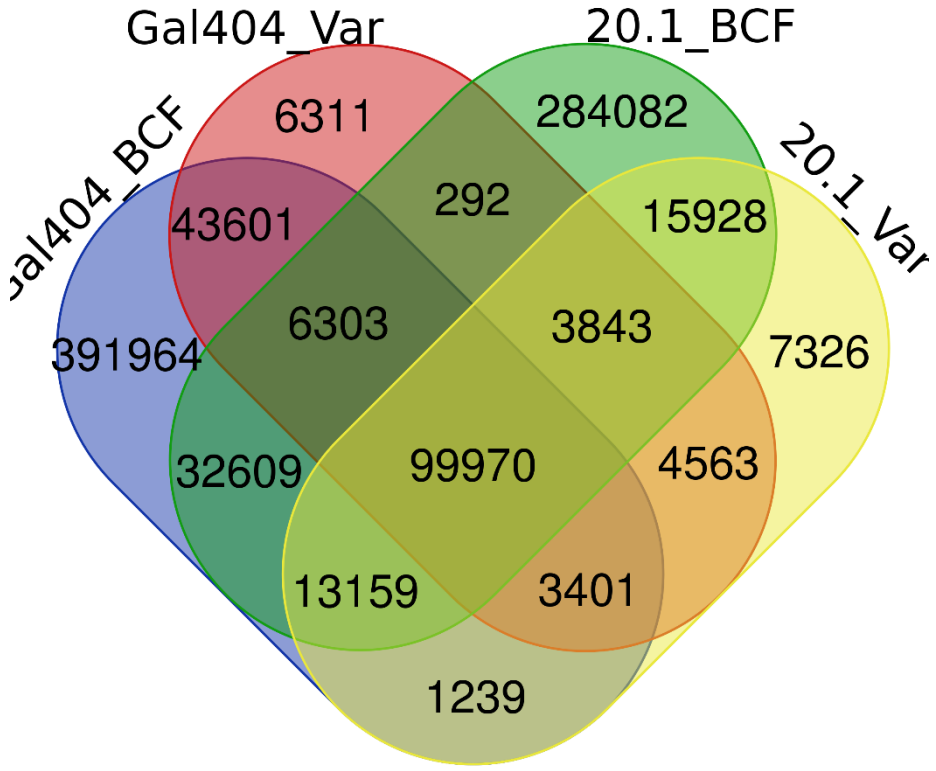
Nejvíce jednonukleotidových polymorfismů bylo nalezeno s pomocí programu BCFTools u kmenu *Claviceps purpurea* Gal404. U studovaných kmenů lze rovněž pozorovat trend, že program BCFTools poskytuje obecně více jednonukleotidových polymorfismů, stejně jako inzercí, nebo delecí ve srovnání s programem VarScan. Tento trend si lze vysvětlit využitím statistického vyhodnocení u programu VarScan ve srovnání s programem BCFTools, což ukazuje u VarScanu na konzervativní přístup k detekci zkoumaných nukleotidových modifikací, ovšem s větším důrazem na spolehlivost výsledků. Další skutečností, která byla sledována v případě srovnávaných dvou programu je zastoupení jednotlivých studovaných DNA modifikací vůči celkovému počtu. U obou programů byl v tomto ohledu nalezen majoritní podíl substitucí vůči ostatním

Tabulka 5: Detekce substitucí, inzercí a delecí programy VarScan a BCFTools u kmenů Claviceps purpurea 20.1 a Gal404.

Modifikace	Gal404		Kmen 20.1	
	BCFTools	**VarScan**	**BCFTools**	**VarScan**
Substituce	571 022	160 665	439 267	142 297
Inzerce	6 515	4 145	5 473	3 831
Delece	14 709	3 474	11 446	3 301

modifikacím DNA (inzerce a delece). u programu BCFTools podíl substitucí na detekovaných modifikacích DNA činil přibližně 97,5 % u obou studovaných kmenů, v případě programu VarScan podíl substitucí činil dokonce přes 99 %. Mezi programy tedy nebyly nalezeny zásadní rozdíly v rozdělení detekovaných modifikacích (substituce, inzerce, delece), ovšem byly nalezeny rozdíly mezi programy co do počtu detekovaných DNA modifikací. Program VarScan detekoval pouze cca 30 % mutací ve srovnání s programem BCFTools, což koresponduje s předpokladem, že u programu VarScan byl zahrnut statistický test dodávající detekovaným modifikacím větší spolehlivost ve srovnání s programem BCFTools.

Ve snaze docílit podrobnějšího srovnání studovaných kmenů *Claviceps purpurea* a použitých programů (BCFTools a VarScanu) byly detekované modifikace DNA zpracovány do podoby Vennova diagramu (Obr. 17). U studovaných kmenů bylo nalezeno celkem 99 970 DNA modifikací, které jsou společné pro oba kmeny a rovněž byly detekovány oběma programy. V kontrastu s tímto zjištěním bylo nalezeno 43 601 unikátních modifikací DNA, které byly detekovány oběma zkoumanými programy v kmenu Gal404 ve srovnání s kmenem 20.1. Podobná situace byla nalezena u kmene 20.1, kde bylo ovšem nalezeno pouze 15 928 DNA modifikací, které nebyly nalezeny v kmenu Gal404.



Obrázek 17: Vennuv diagram zobrazující distribuci detekovaných DNA modifikací ve kmenech *Claviceps purpurea* 20.1 a Gal404 s pomocí programů BCFTools a VarScan.

Detekované modifikace mohou mít zásadní vliv na fenotypové projevy u studovaných kmenů *Claviceps purpurea* zejména pokud se nacházejí v exonech, a ještě razantnější vliv, pokud tyto mutace zasahují do aktivních míst enzymů, které jsou kódovány těmito exony. Pro přesnější interpretaci těchto výsledků by ovšem byly vyžadovány rozsáhlejší studie s fenotypovým pozorováním a v případě kmenů *Claviceps purpurea* rovněž s měřením obsahu produkovaných sekundárních metabolitů.

Studie, která by našla, případně vyvrátila korelaci mezi modifikacemi DNA a obsahem sekundárních metabolitů by mohla být velmi přínosná pro objasnění případného rozdílného obsahu těchto metabolitů. Popisovaný experiment nicméně není součástí této práce, ani jejím hlavním cílem, nicméně data získaná v průběhu této práce mohou posloužit pro další experimenty zahrnující studované kmeny *Claviceps purpurea*.

V rámci srovnání programu implementovaného v rámci této práce a komerčně dostupného softwaru s podobnou funkcí ANNOVAR, domovská stránka: http://annovar.openbioinformatics.org/ (20. 4. 2019), byly pro srovnání vybrány mutace s větší spolehlivostí (detekované programem VarScan).

4.3 Anotace nukleotidových modifikací

DNA modifikace detekované programem VarScan, domovská stránka: http://dkoboldt.github.io/varscan/ (20. 4. 2019), byly anotovány s pomocí programu implementovaného v této práci a rovněž pomocí komerčně dostupného softwaru ANNOVAR, domovská stránka: http://annovar.openbioinformatics.org/ navštíveno 20. 4. 2019, (Tab. 6). Dle předpokladu, bylo docíleno stejných výsledků s ohledem na zdrojový anotační GFF soubor, ve kterém jsou striktně definovány rozmezí pozic pro introny, exony i genové oblasti. Vzhledem k celkovému počtu detekovaných mutací pro

Tabulka 6: Anotace modifikací DNA pomocí programu ANNOVAR a vyvíjeného softwaru u studovaných kmenů Claviceps purpurea 20.1 a Gal404 které se nacházejí v kódujících oblastech.

Pozice	Gal404			Kmen 20.1		
	Substituce	**Inzerce**	**Delece**	**Substituce**	**Inzerce**	**Delece**
Exon	90 249	451	160	75 710	397	184
Genová oblast bez exonu	4 325	224	189	5 234	278	232
Mezigenová oblast	66 091	3 470	3 125	61 353	3 156	2 885

kmen Gal404 (celkem 168 284 mutací) a pro kmen 20.1 (celkem 149 429 mutací) je patrné, že kmen Gal404 dosahuje většího počtu DNA modifikací ve srovnání s kmenem 20.1. S ohledem na zjištěné počty substitucí pro oba kmeny a jejich podíl vzhledem k celkovému počtu detekovaných modifikací DNA rovněž zde byly nalezeny majoritní podíly substitucí ve srovnání s inzercemi a delecemi. Získané mutace nacházející se v oblastech genů kódujících proteiny byly následně převedeny do proteinové sekvence s ohledem na detekované modifikace DNA. Tímto způsobem byly získány transkripty pro celkový počet 7 604 genů pro kmen 20.1 a 7 831 genů pro Gal404.

Výsledky obou softwarů pro anotaci jednonukleotidových polymorfismů (ANNOVAR a program implementovaný v této práci) byly srovnány a potvrdila se jejich shoda. Výhodou zde implementovaného softwaru je zejména skutečnost, že ve srovnání s programem ANNOVAR poskytuje údaje o počtech mutací pro každý jednotlivý zkoumaný transkript/gen. Program ANNOVAR na druhé straně poskytuje údaje pro každou jednotlivou mutaci, která se v sekvenci vyskytuje bez ohledu na to, jestli se sekvence nachází v jednom nebo ve více transkriptech. Další výhodou implementovaného softwaru je, že pro práci s ním není zapotřebí provádět tvorbu reference ve formátu specifickým pro daný software, jako tomu je u programu ANNOVAR. Pro práci se softwarem ANNOVAR je rovněž nutné provést registraci na internetových stránkách, které se tímto softwarem zabývají a teprve následně je uživateli zaslán zdrojový kód pro instalaci softwaru. Tato strategie může do budoucna vést k tomu, že se ANNOVAR stane placeným softwarem, což bude podstatným způsobem limitovat jeho dostupnost pro vědeckou komunitu.

Poslední výhodou implementovaného softwaru je výstup obsahující modifikované proteiny, který může sloužit jako důležitý nástroj pro následnou proteomickou analýzu, (Faktor *et al.,* 2018), která vyžaduje tvorbu peptidové knihovny pro správnou identifikaci a následnou kvantifikaci proteinů detekovaných v průběhu experimentu. Takový výstup program ANNOVAR nepodporuje.

Jak již bylo dříve diskutováno, hlavním cílem této práce je implementace softwaru poskytujícím nástroj pro anotaci DNA modifikací a následnou proteomickou analýzu, nikoliv biologická interpretace dat získaných v průběhu kvalitativní bioinformatické analýzy transkriptomu. S ohledem na získána data by ovšem pro budoucí experimenty mohlo být zajímavé lokalizovat mutace nacházející se kupříkladu v genovém klastru zodpovědném za syntézu alkaloidů (Haarmann *et al.,* 2005) v *Claviceps Purpurea* a pokusit se najít korelaci mezi detekovanými modifikacemi DNA a obsahem alkaloidů

v těchto kmenech, kdy některé proteiny se mohou stát vlivem mutací v DNA nefunkčními, což může značně ovlivnit produkci souvisejících metabolitů (Kück a Hoff, 2010).

5 Závěr

V průběhu této bakalářské práce byla vypracována literární rešerše, která je obsahem první kapitoly, a která je zaměřena zejména na současné metody sekvencování nukleových kyselin a následnou bioinformatickou analýzu. Literární rešerše rovněž přehledně popisuje základní datové formáty, které jsou využívány v bioinformatice a které jsou klíčové pro tuto práci.

Druhá část práce je věnována zejména implementaci softwaru vytvořeného za účelem zpracování dat získaných kvalitativní bioinformatickou analýzou sekvenačních dat s důrazem na detekování modifikací DNA malého rozsahu. Vytvořený program pro svou práci využívá soubory GFF3, FASTA a VCF, které jsou poskytnuty uživatelem. Soubory datového typu FASTA a GFF3 poskytují informace o referenční sekvenci, na kterou bylo provedeno mapování „readů" získaných ze sekvenátoru. Soubor VCF obsahuje informace o modifikacích DNA detekovaných prostřednictví bioinformatické analýzy. Prezentovaný software dovoluje uživateli dále zpracovávat výstupní soubory za účelem kvantifikace nukleotidových variací. Hlavním výstupním souborem je FASTA soubor obsahující sekvence DNA vzniklé důsledkem genomových variací, spolu s log souborem, který obsahuje informace o počtech variací v jednotlivých genech.

Poslední část bakalářské práce je věnována kvalitativní bioinformatické analýze na vybraném datovém setu zahrnujícím bioinformatická data získaná sekvencováním sklerocii dvou kmenů *Claviceps Purpurea* (kmen 20.1 a Gal404) a jejich následné komparativní analýze. Text se rovněž věnuje srovnání implementovaného programu s dalším dostupným softwarem ANNOVAR, který poskytuje má podobnou funcionalitu.

Na základě analýzy provedené v této práci bylo v kmenu Gal404 nalezeno podstatně více mutací ve srovnání s kmenem 20.1. Za účelem kvalitní detekce modifikací DNA bylo provedeno srovnání dvou volně dostupných softwarů (BCFTools a VarScan). Výsledky a myšlenková analýza algoritmu ukázala, že software VarScan poskytuje více kvalitní detekci jednonukleotidových polymorfismů ve srovnání se softwarem BCFTools navzdory skutečnosti, že poskytuje menší počet detekovaných nukleotidových modifikací. Pro následnou analýzu byly tedy vybrána data získaná programem VarScan, která byla následně zpracována programem ANNOVAR a softwarem publikovaným v této práci.

Výsledky analýzy ukázaly, že softwary poskytují shodné výsledky z hlediska anotace modifikací DNA. Prezentovaný software ve srovnání s programem ANNOVAR

poskytuje dostupnější nástroj pro anotaci modifikací DNA. Kromě toho publikovaný software poskytuje několik funkcí, které v ANNOVARu dostupné nejsou, zejména možnost generování proteinových sekvencí, které mohou být následně využity ke konstrukci peptidové knihovny a pro následnou proteomickou analýzu.

6 Literatura

Adessi C., Matton G., Ayala G., Turcatti G., Mermod J., Mayer P., Kawashima E. (2000): Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research*, **28** (20), https://doi.org/10.1093/nar/28.20.e87.

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., Vakhlu, J. (2016): High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian journal of microbiology* **56** (4), 394–404, https://doi.org/10.1007/s12088-016-0606-4.

Ansorge W. J. (2009): Next-generation DNA sequencing techniques. *New Biotechnology* **25** (4), 195-203, https://doi.org/10.1016/j.nbt.2008.12.009.

Ardui, S., Ameur, A., Vermeesch, J. R., Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research*, **46** (5), 2159-2168.

Bentley D. R., Balasubramanian S., Swerdlow, H. P., Smith G.P., Milton J., Brown C. G., Hall K. P., Evers D. J., Barnes C. L., Bignell H. R., *et al*. (2008): Accurate whole human genome sequencing using reversible terminatorchemistry. *Nature* **456**, 53–59, https://doi.org/10.1038/nature07517.

Buermans H. P. J., den Dunnen J. T. (2014): Next generation sequencing technology: Advances and applications. *Biochimica et biophysica acta – Molecular basis of disease* **1842**, 1932-1941.

Clancy, S. (2008): Genetic mutation. *Nature Education* **1** (1), 187.

Cock P. J. A, Fields Ch. J., Goto N., Heuer M. L, Rice P. M. (2010): The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, *Nucleic Acids Research*, **38** (6), 1767–1771, https://doi.org/10.1093/nar/gkp1137.

Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szczesniak M. W., Gaffney D. J., Elo L. L., Zhang X., Mortavazi A. (2016): A survey of best practices for RNA-seq data analysis. *Genome Biology* **17** (13), https://doi.org/10.1186/s13059-016-0881-8

Faktor J., Hernychová L., Vojtěšek B., Hupp T. (2018): Proteogenomic platform for identification of tumor specific antigens. *Clinical Oncology* **31**, 102-107, http://dx.doi.org/10.14735/amko20182S102.

Feng Y., Zhang Y., Ying C., Wang D., Du Ch. (2015): Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics* **13**:1, 4-16, https://doi.org/10.1016/j.gpb.2015.01.009.

Garber M., Grabherr M. G., Guttman M., Trapnell C. (2011): Computional methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* **8**:6, 469-477.

Goodwin S., Gurtowski J., Ethe-Sayers S., Deshpande P. Schatz C. M., Mccombie W. (2015): Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research* **25** (11), 1750-1756, https://doi.org/ 10.1101/gr.191395.115.

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., Marma M. S., Meng Q., Cao H., Li X., Shi S., Yu L., Kalachikov S., Russo J. J., Turro N. J., Ju, J. (2008): Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(27), 9145–9150. https://doi.org/10.1073/pnas.0804023105

Haarmann T., Machado C., Lübbe Y., Correia T., Schardl Ch., Panaccione D., Tudzynski P. (2005): The ergot alkaloid gene cluster in Claviceps purpurea. Extension of the cluster sequence and intra species evolution. *Phytochemistry* **66**, 1312-1320.

Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., Li H. (2009): The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25:12**, Issue 16, 15 August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

Hrdlickova B., de Almeida R. C., Borek Z., Withoff (2014): Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease* **1842:**10, 1910-1922.

Chen L., Jin P., Qin Z. (2016): DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biology* **201617:**252**,** https://doi.org/10.1186/s13059-016-1112-z.

Knierim E., Lucke B., Schwarz J. M, Schuelke M., Seelow D. (2011): Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *PLoS ONE* **6** (11), https://doi.org/10.1371/journal.pone.0028240.

Kück U., Hoff B. (2010): New tools for the genetic manipulation of filamentous fungi. *Applied Microbiology and Biotechnology* **86**, 51-62.

Lettre G., Vijay G. Sankaran V. G., Marcos André C. Bezerra M. A. C., Aderson S. Araújo A. S., Uda M., Sanna S., Cao A., Schlessinger D., Costa F. F., Hirschhorn J. N., Orkin S. H. (2008): DNA polymorphisms at the BCL11A, HBS1L-MYB, and β-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *PNAS***, 105:**33**,** 11869-11874, https://doi.org/10.1073/pnas.0804799105.

Mardis R. E. (2008): The impact of next-generation sequencing technology on genetics. *Trends in genetic* **24** (3), https://doi.org/10.1016/j.tig.2007.12.007.

Morey M., Fernández-Marmiesse A., Castineiras D., Fraga J. M., Couce M. L., Cocho J. A. (2013): a glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism* **107,** https://doi.org/10.1016/j.ymgme.2013.04.024.

Morozova O., Marra M. A. (2008): Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92** (5), 255-264.

Quick J., Quinlan A. R., Loman N. J. (2014): a reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience* **3** (1), https://doi.org/10.1186/2047-217X-3-22.

Reuter J. A., Spacek D. V., Snyder M. P. (2015): High-Throughput Sequencing Technologies. *Molecular Cell* **58,** http://dx.doi.org/10.1016/j.molcel.2015.05.004.

Ronaghi M., Karamohamed S., Pettersson B., Uhlén M., Nyrén P. (1996): Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242**, 84-89.

Sanger F., Nicklen S., Coulson A. R. (1977): DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74** (12), 5463-5467.

Schadt E. E., Turner S., Kasarskis A. (2010) window into third-generation sequencing. *Human Molecular Genetics, **19**, 227–240.*

Schbath S., Martin V., Zytnicki M., Fayolle J., Loux V., Gibrat J. F. (2012): Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *J Comput Biol*. 2012 ;**19** (6), :796-813.

Smith L. M., Fung S., Hunkapiller M. W., Hunkapiller T. J., Hood L. E. (1985): The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids* **13,** 2399–2412.

Swerdlow H., Gesteland R. (1990): Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids* **18**, 1415–1419.

Totomoch-Serra A., Marquez M. F., Cervantes-Barragán D. E. (2017): Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome. *F1000Research* **6:**1016, https://f1000research.com/articles/6-1016/v1.

Voelkerding K. V., Dames S. A., Durtschi J. D. (2009): Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry* **55** (4), 641-658, https://doi.org/10.1373/clinchem.2008.112789.

Voelkerding V., Dames K., S. D. Durtschi, Jacob. (2009): Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical chemistry* **55**, 641-58, http://clinchem.aaccjnls.org/content/55/4/641.long.

Wang Y., Yang Q., Wang Z. (2015): The evolution of nanopore sequencing. *Frontiers in genetics* **5**:449, https://doi.org/10.3389/fgene.2014.00449.

Yang I. S., Kim S. (2015): Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics & Informatics* **13** (4), 119-125, http://dx.doi.org/10.5808/GI.2015.13.4.119.

7 Seznam zkratek

ASCII	Americký standardní kód pro výměnu informací
bp	Pár bází
CDS	Kódující sekvence
dbSNP	Databáze jednonukleotidových polymorfismů
kbp	1000 párů bází
PCR	Polymerásová řetězová reakce

8 Seznam elektronických příloh

Příloha 1 – ZIP soubor obsahující všechny potřebné moduly vytvořeného softwaru.

Příloha 2 – Soubor s textovou tabulkou s počty nukleotidových variací rozdělených podle genů. Soubor vznikl zpracováním VCF souboru vytvořeného softwarem BcfTools obsahujícího data o nukleotidových variacích v *Claviceps purpurea* 20.1.

Příloha 3 – Soubor s textovou tabulkou s počty nukleotidových variací rozdělených podle genů. Soubor vznikl zpracováním VCF souboru vytvořeného softwarem BcfTools obsahujícího data o nukleotidových variacích v *Claviceps purpurea* Gal 404.

Příloha 4 – Soubor s textovou tabulkou s počty nukleotidových variací rozdělených podle genů. Soubor vznikl zpracováním VCF souboru vytvořeného softwarem Varscan obsahujícího data o nukleotidových variacích v *Claviceps purpurea* 20.1.

Příloha 5 – Soubor s textovou tabulkou s počty nukleotidových variací rozdělených podle genů. Soubor vznikl zpracováním VCF souboru vytvořeného softwarem Varscan obsahujícího data o nukleotidových variacích v *Claviceps purpurea* Gal 404.