ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

FAKULTA BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

Katedra biomedicínské informatiky

Bioinformatická analýza RNA-seq dat

Bioinformatics analysis of RNA-seq data

Bakalářská práce

Studijní program:	Biomedicínská a klinická technika
Studijní obor:	Biomedicínská informatika

Autor bakalářské práce:	Irma Snášelová, DiS.
Vedoucí bakalářské práce:	Ing. Bohuslav Dvorský

Kladno 2022

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení:	**Snášelová**	Jméno:	**Irma**	Osobní číslo:	**492260**
Fakulta:	**Fakulta biomedicínského inženýrství**				
Garantující katedra:	**Katedra biomedicínské informatiky**				
Studijní program:	**Biomedicínská a klinická technika**				
Studijní obor:	**Biomedicínská informatika**				

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Bioinformatická analýza RNA-seq dat

Název bakalářské práce anglicky:

Bioinformatics analysis of RNA-seq data

Pokyny pro vypracování:

V posledním desetiletí došlo k významnému pokroku v oblasti analýzy sekvenačních dat, způsobeným rozvojem sekvenačních technologií, zejména tzv. sekvenováním nové generace (NGS). Spektrum analýzy NGS může sahat od malého počtu genů až po celý genom, v závislosti na cíli. Sekvenování celého genomu (WGS) a sekvenování celého exomu (WES) poskytuje sekvenci bází DNA napříč genomem a exomem. Sekvenování celého transkriptomu (RNA-seq) poskytuje sekvenční informace o kódujících a nekódujích formách RNA pro posouzení variací a úrovní genové exprese v celém transkriptomu. Cílem této práce je zmapovat aktuální sekvenační technochnologie s důrazem na transkriptomové sekvenování, dále pak popsat úkony bioinformatické analýzy sekvenačních dat a identifikovat aktuálně používané nástroje pro bioinformatickou analýzu, popřípadě procesy, do nichž jsou tyto nástroje zapojené. Dílčími podcíly této práce je: - Vysvětlení rozdílu mezi WES, WGS, RNA-seq. - Popsání úkonů bioinformatické analýzy sekvenačních dat. - Nalezení aktuálních best practises a technologií úkony RNA-seq analýzy (variant calling, fusion calling, analýza genové exprese) - Provedení vlastní bioinformatické analýzy RNA-seq a její interpretace.

Seznam doporučené literatury:

[1] PŘISTOUPILOVÁ Anna, Využití nových metod analýzy genomu ve studiu molekulární podstaty vzácných geneticky podmíněných onemocnění., 2020
[2] ANTAO Tiago, Bioinformatics with Python Cookbook, ed. Second, 2018, Packt Publishing, 9781789344691
[3] Friederike D¨undar, Luce Skrabanek, Paul Zumbo, Introduction to differential gene expression analysis using RNA-seq, November 14, 2019, https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf
[4] Luis A. Corchete, Elizabeta A. Rojas, Diego Alonso-López, Javier De Las Rivas, Norma C. Gutiérrez, Francisco J. Burguillo , Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis, Scientific Reports, ročník 10, číslo 19737, 2020

Jméno a příjmení vedoucí(ho) bakalářské práce:

Ing. Bohuslav Dvorský

Jméno a příjmení konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **14.02.2022**
Platnost zadání bakalářské práce: **18.09.2023**

doc. Ing. Zoltán Szabó Ph.D.
vedoucí katedry

prof. MUDr. Jozef Rosina, Ph.D., MBA
děkan

PROHLÁŠENÍ

Prohlašuji, že jsem bakalářskou práci s názvem „Bioinformatická analýza RNA-seq dat“ vypracovala samostatně a použila k tomu úplný výčet citací použitých pramenů, které uvádím v seznamu přiloženém k diplomové práci.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu § 60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů.

V Kladně dne ……. ……..……...

Irma Snášelová, DiS.

PODĚKOVÁNÍ

Chtěla bych poděkovat vedoucímu Ing. Bohuslavu Dvorskému za jeho odborné vedení, připomínky a starostlivost v průběhu psaní této práce.

ABSTRAKT

Bioinformatická analýza RNA-seq dat

Bakalářská práce se věnuje analýze sekvenačních dat se zvláštním zaměřením na sekvenování nové generace (next-generation sequencing), kam se rovněž řadí RNA-seq, během kterého dochází k sekvenování celého transkriptomu. Cílem práce bylo zmapovat vývoj sekvenačních metod s popisem jednotlivých kroků bioinformatické analýzy sekvenačních dat a identifikovat softwarové nástroje, na kterých je analýza prováděna. Dílčími podcíli práce bylo rozvedení pojmů WES, WGS a RNA-seq, dále nalezení současných best practices s RNA-seq daty a provedení vlastní bioinformatické RNA-seq analýzy s následnou interpretací výsledků.

Klíčová slova

DNA, RNA, genová exprese, sekvenování nové generace, RNA-seq

ABSTRACT

Bioinformatics analysis of RNA-seq data

The bachelor's thesis is devoted to the analysis of sequencing data with a special focus on next-generation sequencing, which also includes RNA-seq, during which the entire transcriptome is sequenced. The aim of the work was to map the development of sequencing methods with a description of the individual steps of bioinformatics analysis of sequencing data and to identify software tools on which the analysis is performed. Part of the work was to develop the concepts of WES, WGS and RNA-seq, to find current best practices with RNA-seq data and to perform its own bioinformatics RNA-seq analysis with subsequent interpretation of results.

Keywords

DNA, RNA, gene expression, next generation sequencing, RNA-seq

Obsah

Seznam zkratek

Zkratka	Význam
bp	Komplementární pár bází (*Base pair*)
DNA	Deoxyribonukleová kyselina
RNA	Ribonukleová kyselina
FGS	První generace sekvenování *(First-generation sequencing*)
NGS	Nová generace sekvenování (*Next-generation sequencing)*
SGS	Druhá generace sekvenování (S*econd-generation sequecing*)
TGS	Třetí generace sekvenování (*Third-generation sequencing*)
FASTA	Formát souboru výstupních dat ze sekvenátoru (*Fast alignment*)
FASTQ	Formát souboru výstupních dat ze sekvenátoru s odpovídajícím PHRED skóre kvality
SAM	Formát souboru po mapování (*Sequence alignment map*)
BAM	Binární formát souboru SAM (*Binary alignment map*)
VCF	Formát souboru po detekci variant (*Varint calling format*)
WGS	Sekvenování celého genomu (*Whole genome sequencing*)
WES	Sekvenování celého exomu (*Whole exome sequencing*)
RNA-seq	Sekvenování celého transkriptomu (*RNA-sequencing*)
Best practices	Osvědčené postupy

1 Úvod

Sekvenace je souhrnný termín pro biochemické metody, kterými je možné určit pořadí nukleotidů (A, C, G, U/T) v krátkých úsecích DNA nebo RNA. Během posledních dvaceti let došlo k vývoji sekvenačních metod, které na rozdíl od klasických metod umožňovaly zpracování tisíců až milionů sekvencí současně. Metody sekvenování nové generace (NGS) pomohly ke snížení ceny, a rovněž k urychlení procesu sekvenování. Obrovská produkce výstupních dat si proto žádá data následně utřídit a analyzovat.

Výstupní data zahrnují informace o úrovni genové exprese, jejíž výkyvy mohou mít závažné klinické důsledky. RNA-seq je aktuálně považována za nevhodnější metodu NGS pro měření genové exprese, nalezení fúzních genů či data dále zpracovávat pro modelace transkripčních regulačních sítí. O finálním léčebném postupu pro daného pacienta rozhodne ošetřující lékař na základě informací zpracovaných bioinformatikem. Složitost analýzy RNA-seq dat podnítila rozsáhlý výzkum v této oblasti a také vznik mnoha nástrojů, metod nebo algoritmů pro různé fáze analýzy.

Cílem bakalářská práce byla podrobná rešerše vývoje metod sekvenování DNA a RNA se zaměřením na metodu RNA-seq, která se řadí mezi metody sekvenování nové generace. Následně pak popis úkonů bioinformatické analýzy sekvenačních dat a identifikace aktuálně používaných softwarových nástrojů a procesů, kterou jsou součástí bioinformatické analýzy.

Dílčími podcíli práce bylo vysvětlení pojmů sekvenování celého genomu (WGS), exomu (WES) a transkriptomu (RNA-seq). Ve spolupráci s Ústavem hematologie a krevní transfuze v Praze (ÚHKT) došlo k identifikaci aktuálních best practices při dalších zpracování RNA-seq dat, kam se řadí analýza diferenciální genová exprese, genové fúze nebo metoda variant calling. V závěru práce je možné se seznámit s výsledky vlastní bioinformatické analýzy, kterou jsem zpracovala již zmíněnou metodou RNA-seq a poté výsledky interpretovala.

Úvodní první kapitola představuje teoretický základ o vybraných základních pojmech v rámci biologie a sekvenování. Druhá kapitola detailněji pojednává o vývoji sekvenačních metod, mapuje jednotlivé kroky bioinformatické analýzy a upřesňuje pojem RNA-seq. Třetí kapitola představuje praktickou část a popisuje vlastní provedenou bioinformatickou analýzu RNA-seq. Čtvrtá kapitola interpretuje výsledky bioinformatické analýzy. Pátá kapitola je věnována diskusi a možnosti návaznosti na tuto práci. V šesté kapitole je vyneseno rozhodnutí, zda projekt splnil požadavky zadání.

1.1 DNA

Všechny živé organismy disponují biologickou pamětí, která uchovává informace o své struktuře a funkci. Mateřské buňky tuto vlastnost předávají dál buňkám dceřiným. Tomuto jevu se říká dědičnost, tedy schopnost předávat informace obsažené v buňce do dalších generací. *„Ve většině organismů je dědičnost fyzicky zajišťována přenosem genetické informace ve formě molekuly deoxyribonukleové kyseliny (DNA). Informace je zakódována v pořadí nukleotidů. Informace obsažené v sekvencích nukleotidů jsou strukturované do jednotek informace, které nazýváme gen.“* [7]

U eukaryot nalezneme DNA v buněčném jádře a semiautonomních organelách (mitochondrie a chloroplasty), zatímco u prokaryot se nachází volně v cytoplazmě.

Jedná se o polymer složený z řetězce nukleotidů, který obsahuje dusíkaté báze (purinové (adenin, guanin) a pyrimidinové (cytosin, thymin)), monosacharid (deoxyribóza) a zbytky kyseliny fosforečné. Řetězce jsou k sobě navzájem antiparalelní, tzn. že jeden je orientován ve směru od 3′ do 5′ konce a druhý ve směru opačném.

Dvě vlákna DNA po spojení vytvářejí ikonickou dvoušroubovici, kde jsou jednotlivé řetězce vázány vodíkovými můstky (tzv. komplementárními bázemi, base pair, zkratka bp). Díky zákonu komplementarity jsou k sobě vždy vázány pouze specifické báze a to:

a) A–T (spojené dvěma vodíkovými můstky)
b) C–G (spojené třemi vodíkovými můstky)

Komplementární párování a zápis sledu nukleových bází tvoří kód zápisu genetické informace, která je klíčová pro sekvenační analýzu. [5]

1.2 RNA

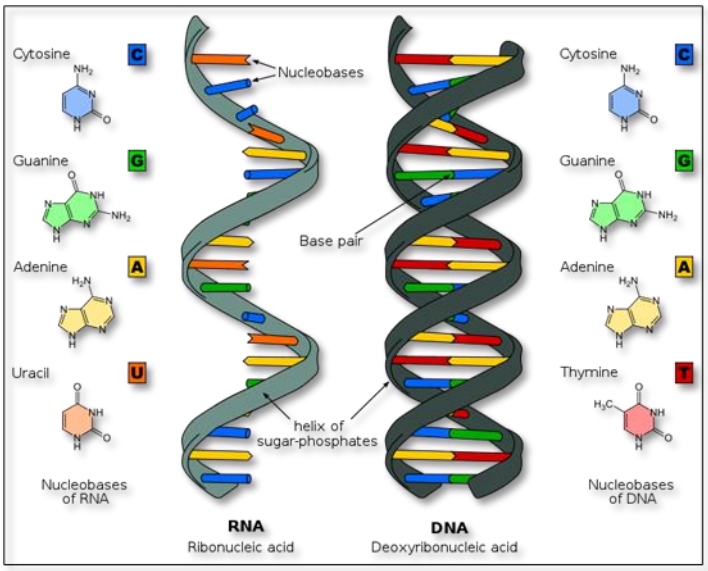
Ribonukleová kyselina (RNA) přenáší informace z nukleových kyselin do proteinů. Tento přenos probíhá v procesu translace a bude popsán v další kapitole. U některých virů je namísto DNA hlavním nositelem genetické informace.

Na Obrázku 1.1 je znázorněna odlišnost od DNA, a to v několika ohledech:

a) přítomností dusíkaté báze uracilu, která zde nahrazuje thymin. V řetězci je tedy tvořena vazba A–U, která je rovněž spojena dvěma vodíkovými můstky,

b) monosacharid deoxyribóza je nahrazen ribózou,

c) molekula RNA bývá zpravidla jedno vláknová. [5]



Obrázek 1.1: Rozdíl mezi strukturou DNA a RNA [4]

RNA vzniká během procesu transkripce a dále se dělí na mnoho podtypů, které plní rozdílné funkce. Pro potřeby sekvenační analýzy jsou klíčové tyto tři typy:

a) mRNA (Mediátorová, informační či messenger RNA)

Tento typ hraje důležitou roli během přenosu genetické informace, protože v průběhu procesu translace kóduje přesné pořadí aminokyselin v bílkovině.

b) tRNA (Transferová RNA)

Funkcí tohoto typu RNA je během procesu translace přinášet na správné místo komplementární antikodon odpovídající kodonu mRNA pro danou aminokyselinu.

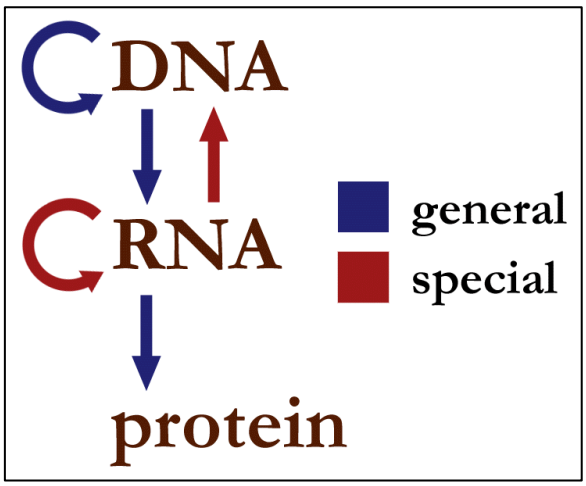
c) rRNA (Ribozomální RNA)

Jak už název napovídá, tento typ tvoří stavební složku ribozomu, což je nukleoprotein, který se u prokaryot nachází v cytoplazmě a u eukaryot na povrchu drsného endoplazmatického retikula, a na jehož povrchu dochází k translaci bílkovin. [6]

1.3 Genová exprese

„*Vnitřní biologická paměť buňky obsahuje všechny nezbytné informace pro její život a pro její reprodukci. Umožňuje buňce samu sebe regulovat, informace z paměti vyzvedávat, tuto paměť doplňovat a dědičně předávat. Většina těchto vnitřních informací je zapsaná ve struktuře nukleových kyselin jako genetická informace…Cestu, kterou se genetická informace vyjadřuje, realizuje z molekul DNA až po vznik různých znaků a vlastností organismů, nazýváme genová exprese.*“ [6]

Jednosměrný tok genetické informace během genové exprese byl popsán už v roce 1958 britským biologem Francisem Crickem jako tzv. ústřední dogma molekulární biologie.



Obrázek 1.2: Ústřední dogma molekulární biologie [8]

Na Obrázku 1.2 je znázorněno, že tok genetické informace především z DNA do RNA, ale některé RNA viry (zejména retroviry) jsou schopny pomocí reverzní transkripce přepsat svou genetickou informaci do DNA.

Vyjádření genetické informace z DNA do bílkoviny probíhá v několika krocích, a to replikací, transkripcí a následně translací. [6]

1.3.1 Replikace

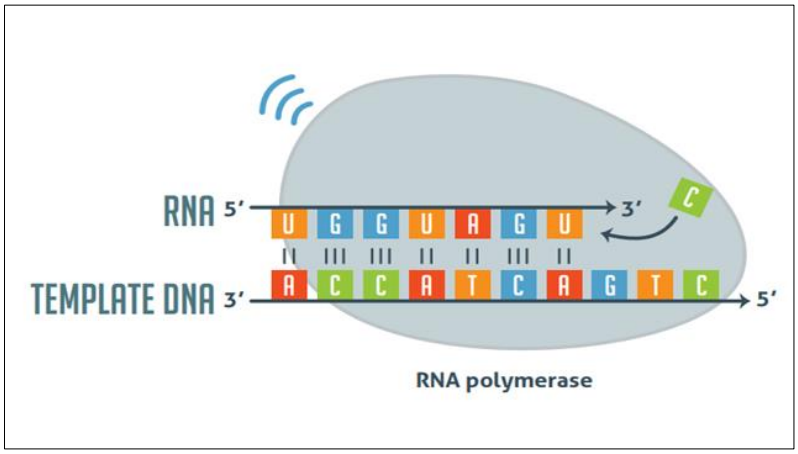
U replikace probíhá přenos genetické informace vždy buď z DNA do DNA nebo z RNA do RNA. Původní molekula se označuje jako templát a vzniklá kopie se označuje jako replika z důvodu vzniku dvou totožných molekul. Proces probíhá semikonzervativním způsobem, který je charakteristický tím, že vzniklá molekula DNA nebo RNA vždy obsahuje jeden řetězec z původní molekuly a jeden nový (syntetizovaný). [6]

Replikace začíná na místě zvaném ORI neboli replikační počátek, kde jej iniciuje enzym primáza. Molekula DNA je tzv. antiparalelní, protože každé vlákno původní molekuly je replikováno jiným způsobem z důvodu opačné orientace. DNA polymeráza je vedena pouze v jednom směru a to od 5' konce k 3' konci. V každém případě je však dle templátu původní DNA vytvářena nová DNA, která je k původnímu řetězci komplementární. Enzym DNA polymeráza syntetizuje nové vlákno jen ve směru 5′ – 3′, a na druhém řetězci 3′ – 5′, vzniknou úseky označované jako tzv. Okazakiho fragmenty, které jsou po odstranění primeru spojeny pomocí DNA ligázy v kontinuální vlákno. Oba řetězce jsou poté spojeny. [33]

1.3.2 Transkripce

Během transkripce dochází k přepisu genetické informace z DNA do RNA. Tento proces je závislý na působení enzymu RNA polymerázy, která urychlí syntézu RNA dle matricového (vzorového) vlákna DNA. Pořadí nukleotidů v nově syntetizované molekule RNA je tedy určeno sekvencí bází matričního vlákna DNA, jak je znázorněno na Obrázku 1.3.

„Na rozdíl od replikace DNA jsou při transkripci DNA přiřazovány k deoxyribonukleotidům s adeninem ribonukleotidy s uracilem, protože v molekulách RNA je thymin nahrazen uracilem. Vzájemným spojením ribonukleotidů pak vzniká polyribonukleotidový řetězec, který se postupně od molekuly DNA odpojuje.“ [6]



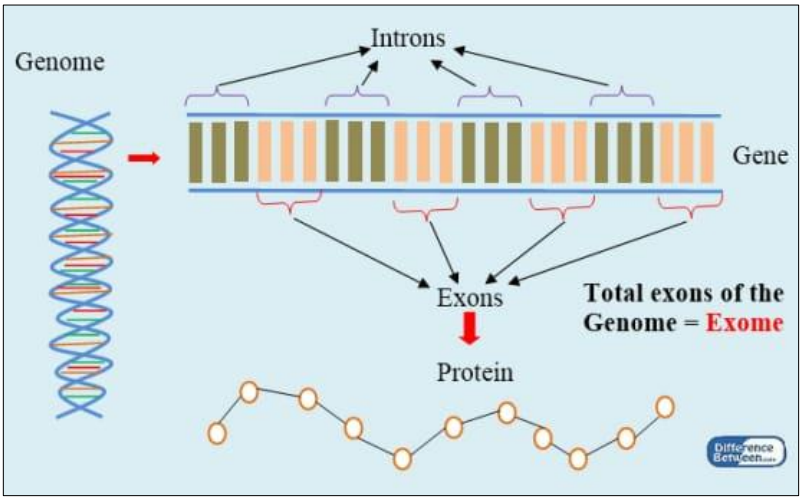
Obrázek 1.3: Průběh transkripce [9]

Transkripcí vznikají všechny druhy RNA, s tím rozdílem, že mRNA je syntetizována transkripcí strukturních (kódujících) genů a tRNa společně s rRNA transkripcí RNA (nekódujících) genů.

Princip transkripce je obdobný ve všech typech buněk, ale u prokaryot a eukaryot existují v určitých detailech rozdíly. [6]

V eukaryotické buňce se vyskytují tzv. složené strukturní geny, které se skládají z exonů a intronů. Část obsahující exony je sice minoritní (pouhých 5 %), ale za to velmi důležitá, obsahuje kódující sekvence s genetickou informací a pouze tyto části se následně na ribozomech zúčastní procesu translace čili můžeme tyto části označit jako proteinotvorné. Naopak introny tvoří zbývajících 95 % genu, za to nejsou překládány a označujeme je za neproteinotvorné, tedy nekódující bílkoviny. *„K čemu jsou tedy*

zdánlivě zbytečné části DNA potřebné? Svoji roli hrály zejména v počátku evoluce genů, často urychlovaly vznik nových bílkovin pomocí rekombinace exonů.“ [12]



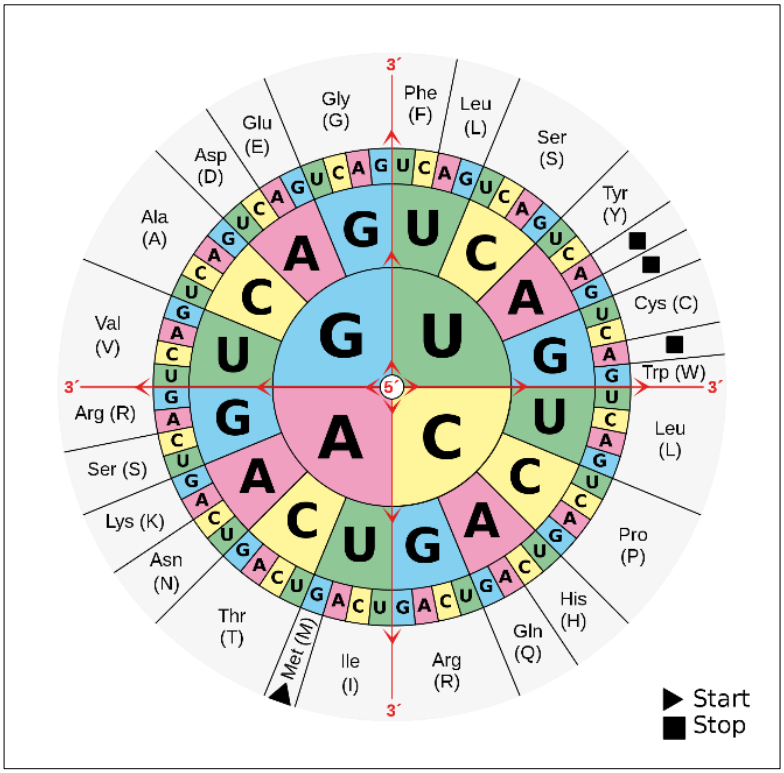
Obrázek 1.4: Znázornění exonů a intronů [28]

Primární transkript eukaryotických genů je velmi dlouhý a obsahuje kopie exonů i intronů, jak ilustruje Obrázek 1.4, avšak ještě v jádře probíhá jejich další úprava, tzv. sestřih, kdy jsou enzymaticky odstraněny (vystříhány) kopie intronů. V takto upravené molekule mRNA tedy zůstává pouze část s exony, které pak slouží jako předloha pro následnou translaci bílkovin. [6]

1.3.3 Translace

Neboli syntéza bílkovin je dalším krokem genové exprese, během něhož je překládáno pořadí nukleových bází v mRNA do pořadí aminokyselin, které tvoří stavební jednotky bílkovin. Překlad probíhá buď na membránách drsného endoplazmatického retikula nebo v cytoplazmě buněk na ribozomech dle pravidel genetického kódu.

Na Obrázku 1.5 je znázorněn genetický kód, kterým se překládá čtyřpísmenná abeceda nukleových bází v mRNA do abecedy aminokyselin, která obsahuje 20 písmen základních aminokyselin.



Obrázek 1.5: Překlad genetického kódu do aminokyselin [10]

Genetický kód má dále tyto vlastnosti:

a) je univerzální, tedy platí pro všechny skupiny organismů,

b) je tripletový, tzn. že teprve kombinace tří nukleových bází poskytuje dostatečný počet variant pro překlad aminokyselin. Trojici těchto bází nazýváme kodon,

c) je nepřekryvný, tj. informaci vyjádřenou pořadím nukleových bází čteme postupně vždy ve směru 5´ -> 3´konce vlákna mRNA,

d) je degenerovaný, to znamená, že pouze 61 tripletů kóduje aminokyseliny. Tři zbývající triplety nekódují žádnou aminokyselinu, ale značí ukončení procesu translace. Dle jejich funkce je nazýváme terminačními kodony (stop kodony), a jsou to triplety UAA, UAG, UGA. Zatímco triplet AUG jako jediný značí začátek průběhu translace (iniciační kodon) a zároveň kóduje aminokyselinu methionin (MET). [6]

Syntéza bílkovin probíhá na ribozomech na těchto třech místech:

a) aminoacylovém místě (A) pro navázání tRNA s komplementárním antikodonem,

b) peptidylovém místě (P) pro navázání polypeptidového řetězce,

c) místě pro připojení mRNA. [6]

Před zahájením samotné translace se ribozom naváže na 5´ konec mRNA a během procesu translace se posunuje směrem k jejímu 3´ konci. Na volný 5´ konec mRNA se ale opět naváže další ribozomy a tím vzniká přechodný útvar polyribozom (polyzom), který se po konci translace rozpadá.

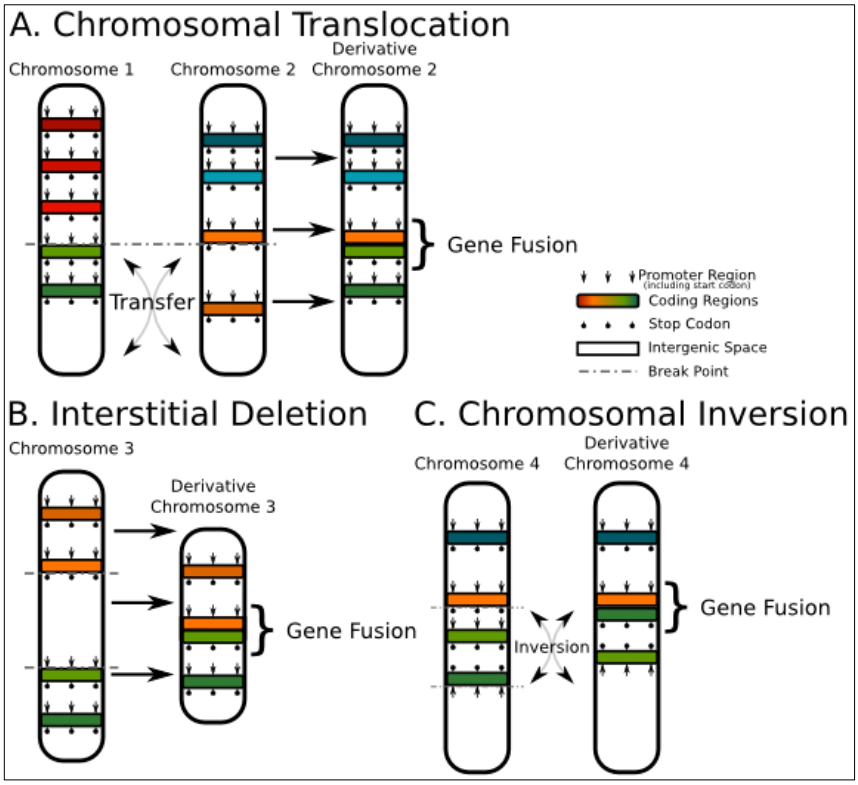
Translace je zahájena v momentě, kdy se do místa P dostane iniciační kodon AUG, to je signálem pro připojení tRNA s iniciační aminokyselinou methioninem (MET), znázorněno na Obrázku 1.5. Do vazebného místa A se dostává kodon následující po iniciačním kodonu a na něj se opět připojí tRNA molekula s odpovídajícím antikodonem. V tuto chvíli může mezi těmito dvěma aminokyselina vzniknout peptidová vazba a následně vybraná bílkovina Pořadí takto nově vzniklých aminokyselin závisí přesně na pořadí kodonů v mRNA. Proces probíhá stejným principem až do doby navázání terminačního kodonu, kdy je translace ukončena.

V prokaryotě probíhají procesy transkripce a translace v cytoplazmě vedle sebe podél chromozomu, ale u eukaryot dochází k transkripci v buněčném jádře a vytvořená mRNA se dostává jadernými póry do cytoplazmy, kde se na drsném endoplazmatické retikulu váže na ribozomy. [6]

1.4 Genová fúze

Termínem fúzní gen je označen hybridní gen, který vzniká spojením dvou dříve nezávislých genů. K této abnormalitě může dojít v důsledku translokace (výměna dvou odlomených segmentů ze dvou chromozomů), intersticiální delece (delece střední části některého z ramének chromozomu) nebo chromozomální inverze (přetočení segmentu mezi zlomy a následné spojení s distálními segmenty druhého chromozomu) genu, vše znázorněno na Obrázku 1.6. Bylo zjištěno, že fúzní geny převládají ve všech hlavních typech nádorových tkání. Identifikace těchto fúzních genů tedy hraje významnou roli pro diagnostiku a predikci nádorových onemocnění. [17]

Fúzní geny jsou často označované jako onkogeny, tedy druhy genu, které způsobují rakovinové bujení. Onkogen vzniká aktivací určitého genu, tzv.protoonkogenu, v důsledku mutací nebo zvýšené hladiny genové exprese. Tyto geny kódují proteiny, které pomáhají regulovat buněčný růst a diferenciaci. Většina normálních buněk podstoupí programovanou formu rychlé buněčné smrti (apoptózu), když jsou kritické funkce změněny a nesprávně fungují. Aktivované onkogeny mohou způsobit, že buňky určené pro apoptózu místo toho přežijí a proliferují (nekontrolovaně se množí). [18]



Obrázek 1.6: Způsoby vzniku fúzních genů [17]

První fúzní gen byl popsán v rakovinných buňkách na počátku 80. let 20. století. Toto zjištění bylo založeno na objevu malého abnormálního markerového chromozomu v roce 1960 Peterem Nowellem a Davidem Hungerfordem ve Filadelfii u pacientů s chronickou myeloidní leukémií – první konzistentní chromozomovou abnormalitou zjištěnou u lidského maligního onemocnění, později označeného jako Philadelphia chromozom. [17]

Přestože genové fúze byly po desetiletí uznávány jako důležité hnací síly rakoviny, naše chápání prevalence a funkce genových fúzí způsobilo revoluci v důsledku vzestupu sekvenování nové generace, pokroku v teorii bioinformatiky a rostoucí kapacity pro rozsáhlou počítačovou biologii. Výpočetní práce na genových fúzích byla značně různorodá a současný stav literatury je roztříštěný. Rostoucí synergie bude znamenat pokroky v identifikaci, charakterizaci a hodnocení významu genové fúze.

Včasná detekce onkogenů je velice důležitá neboť rakovina je celosvětově odpovědná za jedno z osmi úmrtí. [11] Detekce fúzních genů se tradičně spoléhala na techniky FISH (fluorescenční in situ hybridizace) nebo RT-PCR (real time polymer chain reaction). [18]

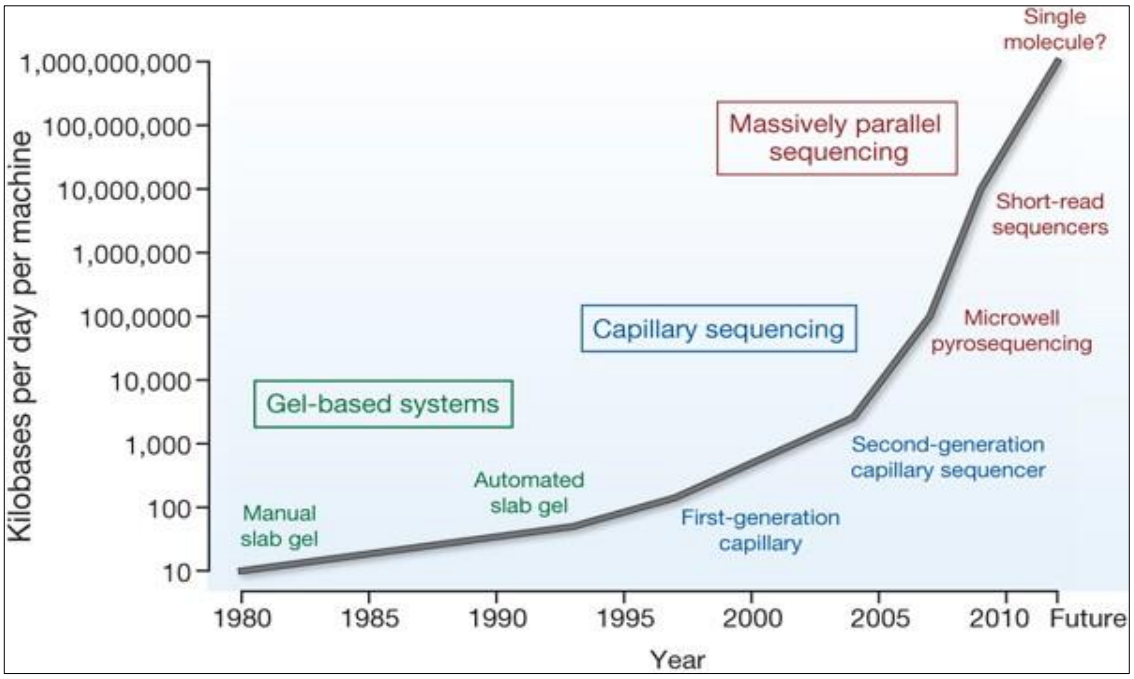
S rozvojem sekvenování nové generace se nově zpracovávají buď data vzniklá z celogenomového sekvenování (WGS) nebo pouze ze sekvenace transkriptomu, tedy RNA-seq, kterému se tato práce bude dále věnovat.

2 Přehled současného stavu

Tato kapitola přiblíží vývoj sekvenačních metod od 70. let 20. století až po současný stav.

2.1 Sekvenační metody

Na Obrázku 2.1 je zobrazen vývoj sekvenačních metod. Tato práce se věnuje metodám sekvenování nové generace, které reagovaly na poměrně finančně nákladnou a zdlouhavou analýzu první generace sekvenování.



Obrázek 2.1: Vývoj sekvenačních metod [11]

2.1.1 První generace sekvenování (FGS)

Do 70.let minulého století se většina onemocnění způsobená variacemi genů identifikovala pozičním a funkčním klonováním. Přelomovým se stal rok 1977, kdy byly shodou okolností představeny hned dvě nové metody sekvenace DNA. Jednalo se o Maxam-Gilbertovu a Sangerovu metodu sekvenování. Obě metody byly založeny na roztřídění jednotlivých sekvencí pomocí gelové elektroforézy, získání sekvencí se u těchto metod ale liší.

Poprvé v historii lidstva bylo možné číst a zkoumat až celé genomy organismů. Jak již bylo řečeno v úvodu kapitoly, jejich hlavní nevýhodou byla a je vysoká cena

za sekvenování a limitace v počtu párů bází (cca 1000 bp), které je možné přečíst najednou. [1]

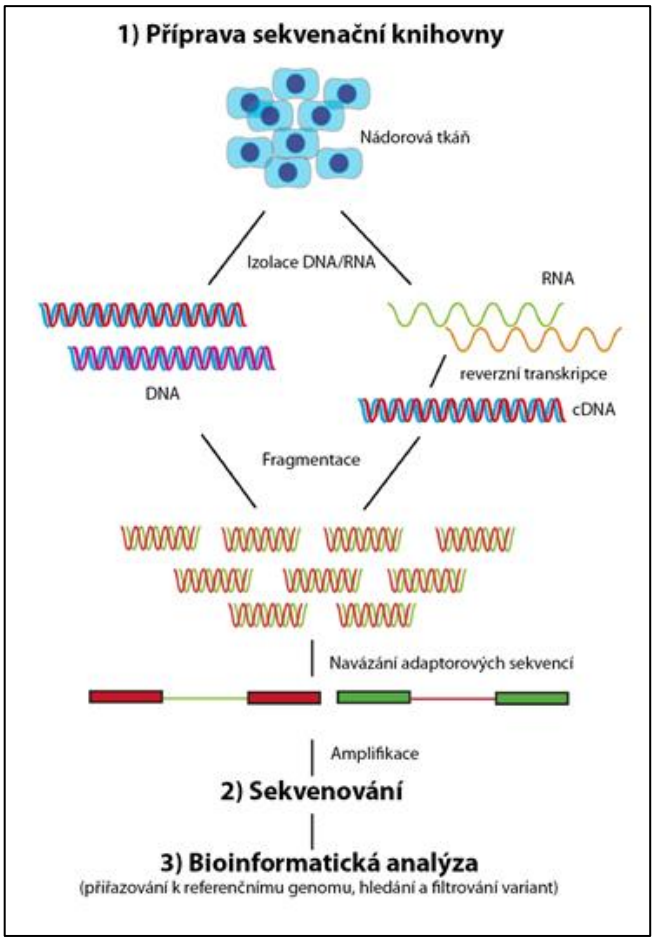
„*Hlavními výhodami FGS je nízká chybovost, přesnost čtení, nižší náklady na pořízení přístroje a jednoduché navržení primerů pro nové cíle. Je vhodné například pro ověřování kandidátních variant nalezených pomocí SGS/TGS u dalších členů rodiny, nebo při ověřování úspěšnosti využití metod genového inženýrství, jako je například klonování…Sangerovo sekvenování tedy stále má a nejspíš i bude mít ve výzkumu i diagnostice své místo.“* [1]

2.1.2 Sekvenování nové generace (NGS)

Po roce 2004 došlo k vyhlášení grantového projektu prostřednictvím National Human Genome Research Institute, který si kladl za cíl snížit cenu sekvenování do 10 let pod 1 000 USD. I díky tomuto tlaku se zrychlil vývoj nových sekvenačních metod, které měly být nejen výkonnější, ale také levnější a hlavně rychlejší.

Nově vznikající metody NGS jsou založeny na principu masivně paralelního sekvenování mnoha nukleových bází (během jednoho běhu jsou sekvenovány až miliardy bází). Následkem toho je velká produkce výstupních dat, které je potřeba utřídit a analyzovat. [1]

„Prvním krokem NGS je příprava vzorku DNA pro účely sekvenace neboli příprava tzv. sekvenační knihovny. V tomto kroku je vždy zahrnuta izolace DNA ze vzorku nádoru (případně přepis nádorové RNA do cDNA), fragmentace DNA, navázání tzv. adaptorových sekvencí na konce fragmentů DNA, což umožní jejich globální namnožení (amplifikaci), a dále jejich fixaci na solidní povrch. Následně, ve druhém kroku, dochází k vlastní sekvenaci syntézou. K imobilizovaným fragmentům DNA jsou komplementárně připojovány jednotlivé nukleotidy (ACTG), což je spojeno s emisí vždy jiného fyzikálního signálu zachycovaného detektorem. Takto získané sekvence jednotlivých fragmentů jsou ve třetím kroku bioinformaticky analyzovány.“ [27]



Obrázek 2.2: Proces sekvenování nové generace [27]

Pracovní postup metody NGS je znázorněn na Obrázku 2.2 a bližší popis zpracování výstupních dat ze sekvenování je popsán v následující kapitole, které se věnuje bioinformatické analýze.

Metody nové generace sekvenování se dále dělí na SGS a TGS.

2.1.3 Druhá generace sekvenování (SGS)

Tato metoda vyžaduje předchozí amplifikaci (namnožení) čtených bází, jejichž délka se pohybuje v rozmezí 35–500 bp. Vzhledem k tomu, že se jedná o báze kratší než u FGS, jsou často označovány jako metody sekvenování krátkých čtení.

Před samotných začátkem sekvenování je potřeba připravit knihovnu se vstupním materiálem (ten je potřeba „naštěpit“ na menší fragmenty), tato data amplifikovat (můstková amplifikace, emulzní PCR amplifikace na kuličkách oleje, případně amplifikace rotujícího kruhu) a následně sekvenovat.

Sekvenování je buď založeno na principu SBL (sekvenování ligací) nebo SBS (sekvenování syntézou). [1]

2.1.4 Třetí generace sekvenování (TGS)

Od roku 2009 vstupuje na komerční trh metoda TGS, která na rozdíl od SGS nevyžaduje předchozí amplifikaci vstupních dat, při které může docházet k zanesení chybných údajů a rovněž znevýhodňuje oblasti s vysokým počtem guanin-cytosin bází. Zároveň je možné číst báze v délce až milion bp v reálném čase, z tohoto důvodu se často uvádí jako sekvenování dlouhých čtení.

Jako první se na trhu objevila technologie SMRT (Single Molecule Real Time), která byla založena na sekvenování syntézou a kamera v reálném čase snímala fluorescenčně zabarvené nukleové báze.

O několik let později byla na trh uvedena další nanopórová technologie sekvenování, která snímá nanopóry zkoumaných molekul a jejich změny při průchodu elektrickým proudem.

Bohužel nevýhodou TGS je vysoká chybovost čtení (mezi 13 až 15 %), kterou je ale možné snížit opakováním několika cyklů za sebou. [1]

2.1.5 Základní aplikace sekvenování nové generace

„Za dvě základní aplikace NGS lze považovat sekvenaci genomové DNA nebo jejích částí a dále sekvenaci transkriptomu, tzv. RNA-seq; technicky se jedná o sekvenaci DNA získané přepisem nádorové RNA s cílem identifikovat přítomnost fúzních genů. Z hlediska genomové DNA lze potom sekvenovat zárodečnou DNA, získanou obvykle z periferních leukocytů či bukálního stěru, obyčejně pro účely klinické genetiky, případně DNA nádorovou, získanou izolací ze vzorku nádorové tkáně pro účely terapeutického plánování.“ [27]

Sekvenace genomové DNA se dále dělí na celogenomové a celoexomové sekvenování. Při použití celogenomového sekvenování (WGS, whole genome sequencing) je možné stanovit kompletní sekvenci genomové DNA. Tím pádem je možné detekovat mutace napříč celým genomem, ať už v kódujících oblastech (exomech) nebo nekódujících oblastech (intronech). Tento způsob umožňuje identifikaci také nových, dříve nepopsaných nádorových variant. Díky velkému množství získaných dat je pohled na specifické změny v nádorových oblastech komplexní, nicméně na druhé straně spektra se nachází velká časová a také finanční náročnost. Velký objem dat rovněž klade vysoké nároky na interpretaci nálezů, a to především v případě nalezení potenciálně rizikových, ale těžce interpretovatelných nálezů.

Oproti tomu celoexomové sekvenování (WES, whole exome sequencing) pokrývá sice pouze kódující oblasti genomu, tedy přibližně 1 % z jeho celkové sekvence, ale většina (přibližně 85 %) patogenních variant se nachází právě v exomech. Výstupy WES sekvenování jsou tedy v porovnání s WGS z hlediska interpretace méně náročnější, za to umožňují větší hloubku prováděných čtení a tím i vyšší senzitivitu zachycení nádorových

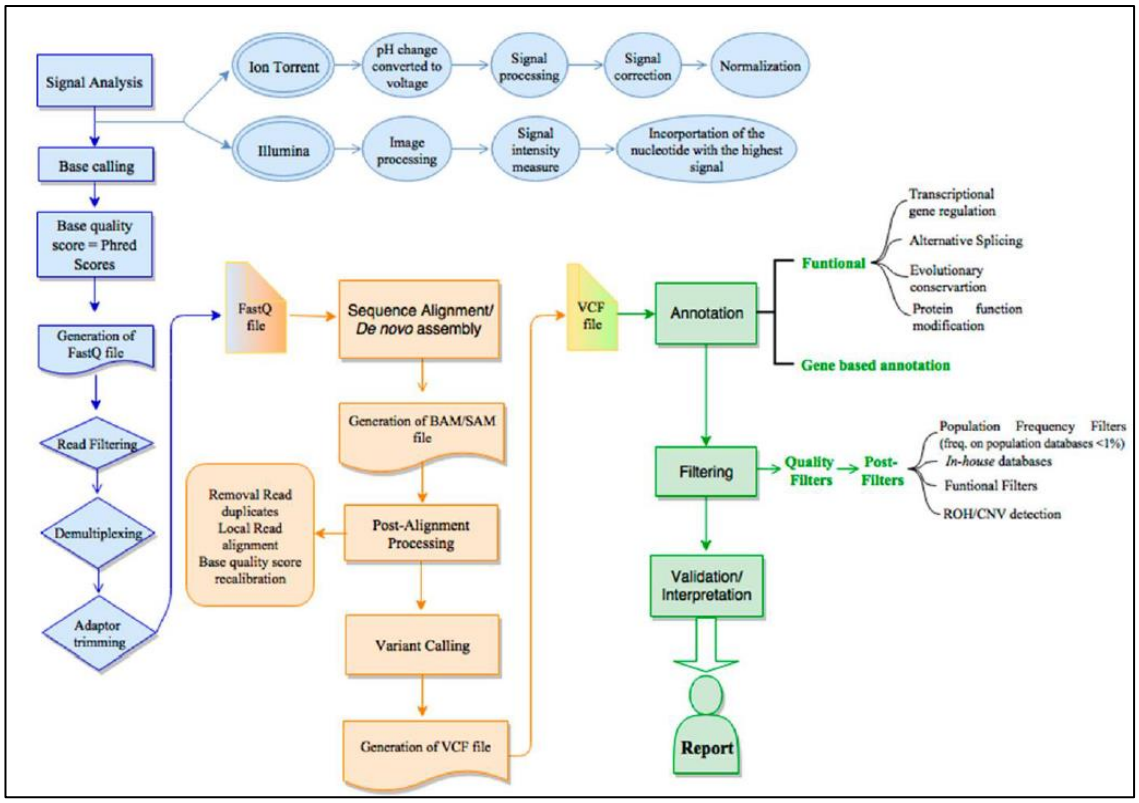
buněk. Vzhledem k tomu, že WES neumožňuje identifikaci variant v intronech, může nastat případný problém s interpretací potencionálně rizikových nálezů. Z hlediska klinicky významného informačního obsahu představuje WES velice dobrou alternativu k WGS. [27]

Pro sekvenaci celého transkriptomu se používá technika zvaná RNA-seq, kterou je možné stanovit hladinu genové exprese určitých biologických objektů za určitých podmínek. Jedná se o finančně efektivní způsob kvantifikace, který poskytuje jak vysokou reprodukovatelnost, vysokou přesnost tak i široký dynamický rozsah. RNA-seq může být aplikováno při studiu reakcí na léky, detekci biomarkerů, v základním lékařském výzkumu a vývoji léků. [27] Blíže se této metodě věnuje kapitola 2.3.

2.2 Bioinformatická analýza sekvenačních dat

Po získání dat prostřednictvím sekvenačních technologií přichází na řadu jejich zpracování. Vzhledem k obrovskému množství vyprodukovaných dat technologiemi NGS (až TB dat) je nutná znalost programovacích jazyků (např. Python), které následnou analýzu urychlí. [2]

Obrázek 2.3 znázorňuje všechny kroky potřebné ke správnému vyhodnocení dat. Samotná bioinformatická analýza začíná krokem base calling.



Obrázek 2.3: Proces bioinformatické analýzy [2]

2.2.1 Primární analýza

Base calling zpracovává výstupní data ze sekvenátorů, která obsahují markerem označené nukleové báze (dle různých sekvenačních metod můžou být báze označené fluorescenční barvou, světlem, změnou pH, případně elektrickým nábojem). Tato data se z formy videí a obrázků zpracují do tzv. hrubých sekvenačních dat a jsou uložena ve formátu FASTA.

FASTA soubor je v podstatě textový popis DNA nebo RNA, který obsahuje hlavičku (začíná znakem „>“) s informacemi o daném vzorku, jako např. ID, název, původ, popis organismu apod. Data jsou nejčastěji zapsána abecedou nukleových bází (A, C, G, T/U) nebo abecedou aminokyselin (A–X). Pro charakterizaci kvality a porovnání chybovosti různých metod sekvenování se přidává číselné hodnocení, tzv. Q skóre (Phred score). Jedná se o označení pravděpodobnosti chybného určení báze. Čím se hodnota Q skóre vyšší, tím nižší nesprávnost základního volání bází. Do doplnění Q skóre se formát FASTA přepíše na FASTQ formát, v hlavičce je úvodní symbolem nahrazen znakem „@“ a samotné hodnocení kvality je umístěno zpravidla na čtvrtém řádku.

V případě, že bylo v jednom cyklu sekvenování zahrnuto více vzorků, výsledná čtení (ready) jsou k nim zpětně přiřazena (demultiplexing) a posledním krokem primární analýzy je tzv. trimming, tedy doslova „ořezání“ bází s nízkou kvalitou. [2]

2.2.2 Sekundární analýzy

Finálně upravené FASTQ soubory se v rámci sekundární analýzy podrobují dalším úpravám, které slouží ke zpětnému sestavení všech sekvencí do původní podoby daného vzorku. K tomuto účelu slouží dvě metody: alignment (mapování) přiřadí data z FASTQ souboru k referenční sekvenci a různými algoritmy nehledá přesnou shodu ale naopak maximální možnou shodu vzhledem k tomu, že každý genom může mít specifické odchylky od dané referenční sekvence. Hledání shody probíhá ve dvou fázích, kdy se nejprve vytipují různé kandidátní oblasti, kde se může hledaná báze nacházet a ty se posléze podrobují užšímu zkoumání pomocí náročnějších algoritmů. Může ovšem nastat i situace, kdy není možné získaná data porovnat s referenční sekvencí. V tomto případě bude potřeba zkoumanou sekvenci sestavit tzv. de novo. [1]

Výstupem mapování jsou data uložena ve formátu SAM, která obsahují velké množství dat (až 70 GB), zejména díky 11 povinným údajům jako je přesná poloha báze vůči referenčním datům a další specifické informace sloužící pro porovnání. Pro efektivnější práci se tento soubor převede do binární podoby a výsledkem je formát BAM, který je výrazně menší a obsahuje cca 7 GB dat. [2]

Dalším krokem sekundární analýzy je tzv. Post Alignment Processing, kdy jsou data setříděna na základě chromozomálních souřadnic a následně označeny indexem.

Následuje krok, ve kterém jsou ze seřazených dat odstraněny PCR duplikáty (vzniklé amplifikací při přípravě vstupních dat do sekvenátoru během primární analýzy), které mají stejnou délku a sekvenci. Případně je možné provést opětovnou rekalibraci kvality bází a tím odstranit systematické chyby sekvenování nebo zarovnat ready kolem insercí a delecí (indely) zkoumaného vzorku.

Po těchto úpravách přichází na řadu tzv. variant calling, během kterého jsou identifikovány veškeré odchylky zkoumaného vzorku od referenční sekvence. Obvykle se pozornost soustředí na možné polymorfismy a indely. Veškeré změny či nesrovnalosti jsou označeny u chromozomálních pozic, kvalitou či informací, že k sekvenační změně nedošlo. Finálním výstupem sekundární analýzy jsou data ve VCF formátu. [1]

2.2.3 Terciální analýza

Pro další práci a porozumění získaným datům ve VCF souborech je jim potřeba dodat biologický kontext. Nejprve se provede tzv. anotace, kdy se pomocí anotačních programů (např. online nástroj Wannovar) určí v jaké genu, v jaké oblasti (kódující/nekódující) a jaký má vznikající bílkovina dopad na funkci (iniciační/terminační kodon, změna/ztráta smyslu apod.).

V dalším kroku je potřeba nalezené anotace filtrovat, protože získaná data jsou stále velmi objemná (exomové sekvenování může mít až desítky tisíc variant) a kauzálních variant bývá zpravidla mnohem méně (v některých případech i jedna jediná). Správně nastaveným filtrem se eliminují varianty s nízkou kvalitou genotypu, ready pokrývající oblast dané varianty, případně i předpokládané modely dědičnosti na základě rodinné anamnézy. Zajímavé jsou také varianty nacházející se ve funkčně významných částech genomu, protože jejich frekvence výskytu je v databázích populací velmi nízká.

Varianty se následně prioritizují a jsou vybírány ty, které mohou být kauzální (údaje z databází, odborné literatury, projevy pacienta). K tomuto účelu je možné použít online nástroj HPO, který obsahuje popis fenotypů. Výsledné varianty jsou poté validovány funkčními studiemi (průkazy rekurence, prokázání rozvoje fenotypu atd.). [1]

2.3 RNA-seq

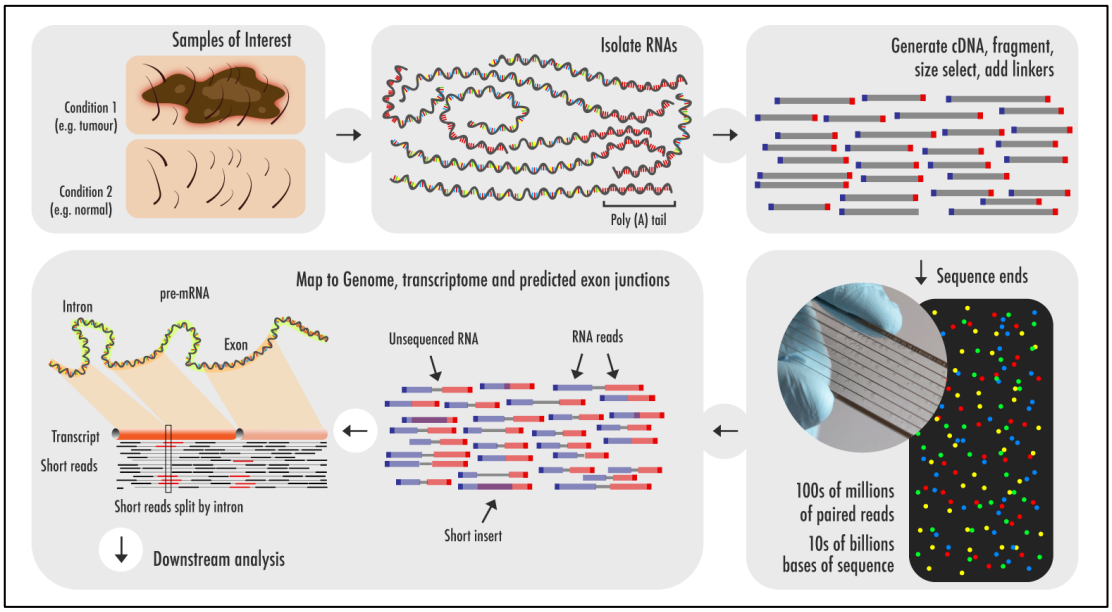
Tato práce se blíže zabývá bioinformatickou analýzou RNA-seq, kdy probíhá sekvenování celého transkriptomu (blíže popsáno v kapitole 2.1.5 výše). RNA-seq se řadí mezi metody NGS, ale jednotlivé kroky samotné bioinformatické analýzy se nepatrně liší.

Tato metoda byla představena v roce 2005 a RNA-seq zahrnuje informace o celém transkriptomu, který obsahuje kompletní informace o všech transkriptech v celé buňce. Díky tomu poskytuje unikátní informace nejen o zkoumaném genu, ale také o aktuálním stavu celé buňky za daných podmínek a informace o jejím daném vývojovém stádiu.

„Poznání transkriptomu přináší celou řadu informací a může tedy přispět k porozumění rozvoje chorob.“ [13]

Pro získání těch nejrelevantnějších výsledků, je potřeba ještě před samotným začátkem sekvenování promyslet jeho účel, tedy to, co přesně hledáme. Dle účelu se plánuje hloubka jednotlivých čtení (opakované čtení sekvence za účelem získání spolehlivých dat), případně množství fragmentů (to je potřeba navýšit u genů, které jsou méně přepisované).

Na Obrázku 2.4. je znázorněn pracovní postup u RNA-seq. Před zahájením samotného sekvenování dochází k přípravě knihovny, kdy se nejprve ze vzorku tkáně se izoluje RNA, které se rozštěpí na fragmenty a přepíše na cDNA (chromozomální DNA). Tento krok se provádí pomocí reverzní transkripce z důvodu následného použití pracovního postupu NGS. cDNA se poté fragmentuje a na každý konec fragmentů se přidají adaptéry. Tyto adaptéry obsahují funkční prvky, které umožňují sekvenování, například amplifikační prvek (který usnadňuje klonální amplifikaci fragmentů) a primární sekvenační primární místo. Po procesech amplifikace, výběru velikosti, čištění a kontroly kvality je pak knihovna cDNA analyzována pomocí NGS za vzniku krátkých sekvencí, které odpovídají celému nebo části fragmentu, ze kterého byla odvozena. Před zahájením bioinformatické analýzy je ještě zkontrolována kvalita upraveného vzorku. [29]



Obrázek 2.4: Proces RNA-seq [29]

Získaná RNA-seq data je možné dále zpracovávat při analýze diferenciální genové exprese, při analýze genové fúze či pro zjištění transkripčních regulačních sítí nebo

u metody variant calling, což je proces identifikace pozic nukleotidů vůči referenčnímu genomu.

2.3.1 Současné metody analýzy RNA-seq

Aktuální best practices jsem konzultovala na Ústavu hematologie a krevní transfúze v Praze na oddělení molekulární genetiky (ÚHKT), kde se mimo jiné provádí sekvenování pacientů s akutní lymfoblastickou leukemií (ALL). Nejčastější malignita se objevuje u dětí ve věku od 1-5 let, což znamená, že dětští pacienta jsou v raném věku diagnostikováni v Motolské nemocnici v laboratoři Childhood Leukaemia Investigation Prague (CLIP), která je určena pro pacienty do 18 let. Po dovršení této věkové hranice pacienti přecházejí do péče ÚHKT.

Jak již bylo zmíněno v kapitolách výše, RNA-seq se používá nejčastěji pro detekci fúzních genů nebo pro stanovení diferenciální genové exprese. Oba tyto jevy se zkoumají u onkologických pacientů, kdy je nejprve provedeno vstupní vyšetření, během kterého je za pomoci sekvenátorů určen podtyp onkologického onemocnění, na jehož základě může být lékařem zvolen vhodný léčebný postup. U onemocnění ALL se konkrétně jedná o potvrzení přítomnosti fúzního genu BCR-ABL1, který je přítomen cca u 20 % případů ALL ve starším věku.

Vzhledem k vysoké ceně sekvenování a bohužel také problémům s proplácením zdravotními pojišťovnami (složitá příprava podkladů a dokázání prospěchu pro daného pacienta), není možné provádět pravidelná kontrolní sekvenování pacienta, to je možné pouze v případě, že stanovená léčba selže.

Metoda RNA-seq je na ÚHKT prováděna buď interně pomocí HemaVision, což je řada in vitro diagnostických testů pro rychlou a citlivou detekci chromozomálních translokací spojených s leukémií. Tímto testem je možné analyzovat až 28 translokací a více než 145 klinicky relevantních translokačních zlomů v jediném testu. Výchozím materiálem pro testy je celková RNA extrahovaná z krve nebo kostní dřeně. Výsledky testu jsou známy cca do 4 hodin, ale jedná se pouze o rychlou analýzu a je zapotřebí další ověření a provedení potvrzujících testů. Další testy je potřeba provést externě, v případě ÚHKT se sekvenování provádí na přístroji NovaSeq, kdy se vzorky RNA posílají buď do laboratoře do Olomouce nebo do Plzně. Výstupy jsou z laboratoří dostupné většinou do 24 hodin a následně se výsledky zpracovávají v Metacentru (virtuální organizace, která sdružuje výpočetní a úložné kapacity hostované v několika institucích jako např. Fyzikální ústav AV ČR, CESNET, Západočeská univerzita v Plzni a jiné). Výstupní data se dále zpracují jedním z dostupných softwarových nástrojů na ÚHKT, jedná se o nástroje Arriba, StarFusion, Cicero nebo komerční SoftGenetics (placená licence). V dalších kapitolách této práce je zmíněn nástroj FusionCatcher, jehož nasazení UHKT také zvažovalo, ale po provedení rešerše jeho použití zamítli kvůli vysokému počtu falešně pozitivních nálezů. Finální výsledky jsou dostupné cca za týden a poté následuje

interpretacevýsledků, na které se podílí jak bioinformatik, tak lékaři. Při anotaci je potřeba zjistit i pozadí případně nalezeného fúzního genu, kdy se může objevit v databází fúzí, ale již chybí bližší popis toho, zda se jedná o klinicky relevantní mutaci.

Na ÚHKT by ocenili pořízení vlastního výpočetního zdroje, který by zkrátil dobu čekání na finální výsledky a také usnadnil administrativní náročnost sdílení citlivých údajů třetí straně. Rovněž by usnadnil získání grantu, prostřednictvím kterého by bylo možné snadnější proplácení sekvenování ze strany zdravotních pojišťoven.

Co je pro bioinformatiky naopak matoucí, je množství nově generovaných sekvenačních nástrojů na akademických půdách, které nemají komerční využití ani příslušnou zákaznickou podporu v případě poruchy nebo komplikací.

2.3.2 Analýza diferenciální genové exprese

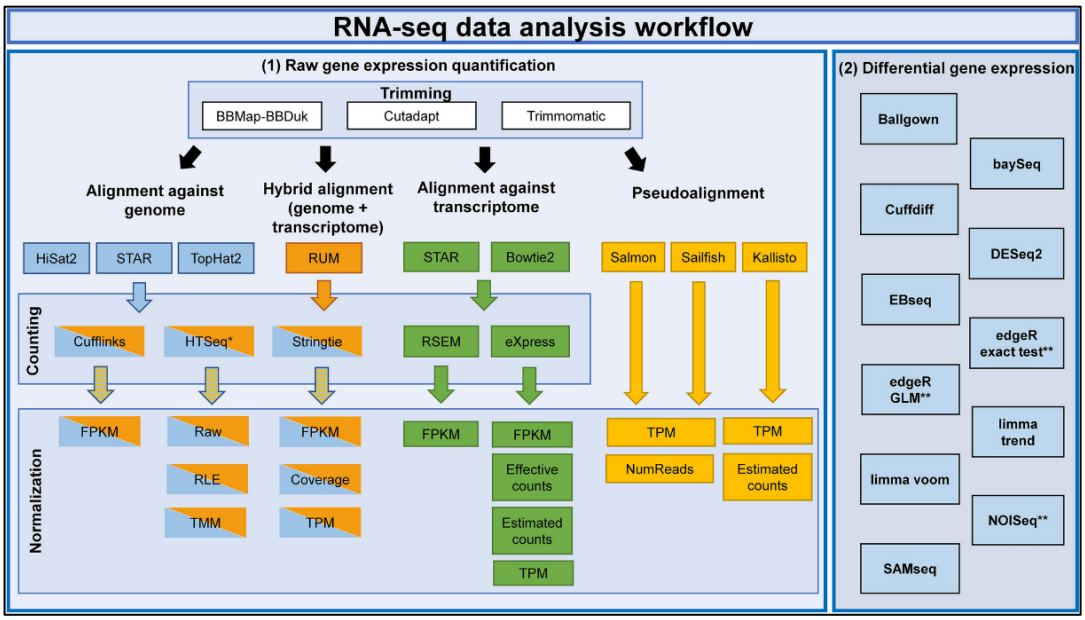
V kapitole 1.3 jsou popsány principy genové exprese. Tato kapitola se bude věnovat detailnějšímu popisu metod, zejména kvantitativní genové expresi, a následnému vyhodnocení výsledků pomocí tzv. diferenciální exprese jednotlivých transkriptů za pomocí vybraných programů.

Genová exprese je velmi komplikovaný děj, který probíhá ve všech buňkách. Krok po kroku jsou informace obsažené v DNA nejprve transkripcí přepsány do RNA a následně translací přeloženy do aminokyselin z jejichž peptidového řetězce vzniká výsledný produkt – bílkovina neboli protein. *„Poznání, jaké geny jsou přepisované v buňce v konkrétním okamžiku, nám pomůže odhalit, co se v buňce děje v různých vývojových stádiích, při vzniku nemocí, při působení cizorodých látek a podobně.“* [13]

Na Obrázku 2.5 je znázorněn proces analýzy RNA-seq dat, kdy se nejprve zpracují tzv. raw data (hrubá data). V rámci primární analýzy dochází standardně k tzv. trimmingu, který zvýší rychlost mapování čtení odstraněním nekvalitních nukleotidů. Musí být používán neagresivně, spolu s dobře zvolenou délkou čtení, aby se zabránilo nepředvídatelným změnám v genové expresi a transkriptomové sestavě. Trimming je nápomocný jak při porovnávání referenčních sekvencí, tak při metodě „de novo“. *„U nemodelových organismů, pro nedostupnost sekvence genomu, a tedy i čipů (které jsou založeny na imobilizovaných sondách o známé sekvenci), představuje RNA-seq jedinou možnost, jak získat nějaká data.“* [13]

Jakmile jsou ready zmapovány, musí být přiřazeny ke genu nebo transkriptomu v procesu známém jako counting neboli kvantifikace. Nejjednodušší způsob kvantifikace genové exprese pomocí RNA-seq je sečtení readů, které se mapují (tj. zarovnávají se) ke každému genu (počtu readů) pomocí programů, jako je např. HTSeq-count.

Poté následuje krok tzv. normalizace během něj se identifikují a následně odstraní báze, které by mohly finální výsledky zkreslovat (sequencing bias). [14]



Obrázek 2.5: Proces RNA-seq [14]

Následuje vyhodnocení diferenciální genové exprese, kde, jak už samotný název napovídá, zkoumáme určité rozdíly. Tyto rozdíly se týkají četnosti výskytu genových transkriptů v celém transkriptomu a jsou porovnávány s výskytem daného fenotypu i podmínek, za jakých sekvenování probíhalo. Cílem toho porovnání je přesné rozlišení genů, u kterých exprese probíhá za stejných podmínek, ale v odlišné míře. *„Je důležité si uvědomit, že gen je považován za diferenciálně exprimovaný, pokud je pozorovaný rozdíl hladiny jeho exprese mezi dvěma experimentálními podmínkami statisticky významný, tzn. pokud je rozdíl větší než to, co by se dalo očekávat jen kvůli náhodným změnám.“* [15]

Diferenciální genová exprese se svým charakterem řadí mezi statistické metody a z toho plyne, že k zajištění validních výsledků je nezbytné dodržovat několik pravidel:

a) Pro správné porovnání dat musí být zajištěny minimálně dvě skupiny vzorků s alespoň třemi replikáty na skupinu.
b) Jasné stanovení referenční a výsledné skupiny (neplatí v případě „assembly de novo“
c) V případě následného porovnání s referenční skupinou je vhodné tato data čerpat z veřejných databází (např. National Center for Biotechnology Information) [15]

K účelu zjištění hladiny rozdílů genové exprese bylo vyvinuto mnoho nástrojů. Tyto nástroje slouží k tomu, aby porovnaly výsledky většího množství vzorků napříč různými cDNA knihovnami. Zjednodušeně řečeno, porovnávají průměrné úrovně exprese různých genů a hledají místa, kde dochází k rozdílům, tedy k mutacím genů. [14]

Metody pro diferenciální analýzu genové exprese lze rozdělit do dvou hlavních podskupin: parametrické a neparametrické.

Parametrické metody zachycují všechny informace o datech v rámci parametrů. V těchto případech je možné předpovědět hodnotu neznámých dat z pozorování adoptovaného modelu a jeho parametrů. Pokud jsou parametrické metody aplikovány na diferenciální genovou expresi, předpokládá se, že obvykle po normalizaci je každá hodnota exprese pro daný gen mapována do určité distribuce nebo je negativně binomická. Nástroje založené na parametrických metodách s negativním binomických modelem jsou např. DESeq2 nebo baySeq.

Na druhou stranu neparametrické metody mohou zachytit více podrobností o distribuci dat, protože nevycházejí z pevně definovaných (tedy omezených) parametrů, takže množství informací o datech se může zvyšovat s jejich objemem. Tuto metodu využívá např. nástroj SAMseq.

V současné době neexistuje mezi vědci shoda o tom, která metodika je nejvhodnější nebo jaký přístup zajišťuje nejrelevantnější výsledky, pokud jde o robustnost, přesnost nebo reprodukovatelnost. [16]

Nástroje pro analýzu diferenciální genové exprese

V této práci jsem se zaměřila na tři nástroje: baySeq, SAMseq a DESeq2, a to z důvodu porovnání výsledků mezi parametrickými (baySeq a DESeq2) a neparametrickými (SAMseq) metodami. Ve většině případů všechny tři nástroje dosahují velmi dobrých výsledků. Přístroj DESeq2 se řadí mezi nejcitovanější nástroje pro vyhodnocení diferenciální genové exprese.

Prostřednictvím vybraných studií byly srovnávány výhody a nevýhody většiny dostupných nástrojů používaných k vyhodnocení diferenciální genové exprese.

První vybraná studie byla publikována v listopadu 2020 na odborných stránkách www.nature.com pod názvem Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Na celkem 17 různých nástrojích testovala vzorky dvou lidských buněčných linií. Jednalo se o vzorky mnohočetného myelomu, známého pod označením KMS12-BM, ošetřovaných dvěma různými léky. Na tomto vzorku bylo detekováno celkem 107 genů udržujících referenční expresi.

Z Obrázku 2.6 níže je zřejmé, že ze dvou buněčných linií byly namnoženy 3 replikáty, celkem bylo tedy použito 6 vzorků s označením CLA-T0 a CLB-T0 a ty byly následně porovnávány mezi sebou. Další kritériem zkoumání byla tzv. false discovery

rate (FDR) neboli míra falešných zjištění, která byla udávána v těchto hodnotách: FDR <0,05, FDR <0,01 a FDR <0,001. Z vybraných 3 nástrojů dopadl nejlépe baySeq, který společně ještě s jedním nástrojem, dosáhl nejlepšího možného skóre napříč všemi pozorováními.

Naopak další 2 vybrané nástroje, DESeq2 a SAMseq, skončily až na samém chvostu hodnotící škály. SAMseq dosáhl nejhoršího možného skóre (17 bodů) hned v 4 z osmi možných hodnocení. [14]

Method	Overall performance	Performance by number of DEG scenario					Performance by statistical significance cut-off		
		CLA-T0 *vs.* CLB-T0	CLA-T1 *vs.* CLA-T0	CLA-T2 *vs.* LCA-T0	CLB-T1 *vs.* CLB-T0	CLB-T2 *vs.* CLB-T0	FDR < 0.05	FDR < 0.01	FDR < 0.001
Ballgown	13	11	15	13	5	16	4	17	16
baySeq	1	9	15	4	1	3	4	5	1
Cuffdiff	10	9	3	7	15	15	15	9	11
DESeq2	13	1	11	15	12	8	17	7	12
EBSeq	3	13	8	16	3	1	13	10	8
edgeR exact RLE	13	8	11	13	15	13	16	6	14
edgeR exact TMM	16	6	13	10	12	13	11	10	13
edgeR exact UQ	12	11	6	11	15	11	1	13	15
edgeR GLM RLE	3	2	5	11	7	8	11	7	2
edgeR GLM TMM	3	2	3	8	9	12	8	3	5
edgeR GLM UQ	8	2	13	17	5	8	3	15	10
limma trend	3	7	8	5	2	4	6	2	2
limma voom	3	2	10	5	4	4	6	4	8
NOISeq FPKM	8	14	6	1	8	4	1	1	7
NOISeq TMM	10	15	1	8	11	2	10	13	6
NOISeq UQ	1	15	2	2	9	4	13	10	4
SAMseq	17	17	17	3	12	17	9	16	17

1 Good performance — 17 Bad performance

Obrázek 2.6: Výsledky první studie [14]

Jako druhá srovnávací studie byla zvolena RNA-seq differential expression analysis: An extended review and a software tool, která byla publikována v prosinci 2017 na stránkách www.journal.com. Testování probíhalo na 11 různých nástrojích, kde byly analyzovány dva biologické vzorky: Ambion (RNA nacházející se v lidském mozku) a Stratagene (univerzální lidská RNA). Na tomto vzorku bylo zjištěno celkem 997 unikátních genů.

K hodnocení byly použity celkem 4 kategorie a to: TPR (True positive rate neboli pravdivě pozitivní míra), SPC (Specifity v překladu Specifita), PPV (Positive Predict Value neboli Pozitivní predikční hodnota) a ACC (Accurancy v překladu přesnost).

Tool	TPR	SPC	PPV	ACC	F_1 measure
edgeR	0.71	0.94	0.90	0.85	0.79
baySeq	0.92	0.40	0.52	0.61	0.66
DESeq	0.44	0.59	0.43	0.53	0.44
NOIseq	**0.80**	**0.95**	**0.92**	**0.89**	**0.86**
SAMseq	0.44	0.52	0.39	0.49	0.42
limma+voom	**0.81**	**0.93**	**0.89**	**0.88**	**0.85**
EBSeq	0.68	0.55	0.52	0.60	0.59
DESeq2	**0.84**	**0.95**	**0.92**	**0.90**	**0.88**
sleuth	0.77	0.54	0.54	0.63	0.64

Obrázek 2.7: Výsledky druhé studie [16]

Na Obrázku 2.7 je možné si všimnout, že z vybraných nástrojů naopak na DESeq2 dosáhl konzistentních a kladných výsledků ve všech čtyřech zkoumaných kategoriích. Oproti tomu nástroje baySeq a SAMseq, které dosáhly skvělého hodnocení v předešlé studii, tak v této jejich hodnocení nebylo moc kladné. baySeq dosáhl nevyššího hodnocení v kategorii TPR (92 %), ale v ostatních třech kategoriích výsledky klesly mezi 40-60 %. SAMseq ani v jedné kategorii nepřesáhl 53 %. [16]

2.3.3 Analýza genových fúzí

V kapitole 1.4 jsou popsány principy genové fúze. Tato kapitola se bude věnovat detailnějšímu popisu analýzy výsledných hodnot a následnému filtrování relevantních nálezů za pomocí široké škály softwarových nástrojů, kdy každý z nich pracuje na základě rozlišných algoritmů.

Za posledních 10 letech byly vyvinuty desítky softwarových nástrojů pro detekci fúzních genů právě na základě výstupních dat RNA-seq. Tyto nástroje jsou zpravidla založeny na jedné z těchto metod: mapping-first approach nebo assembly-first approach. [19]

Mezi přední strategie pro detekci genové fúze se řadí metoda mapping-first, která byla vyvinuta jako jedna z prvních a stále nebyla překonána z hlediska spotřeby výpočetních zdrojů, rychlosti a přesnosti. Tato metoda je založena na zarovnání (alignment) readů k referenční sekvenci RNA a na vzorku identifikuje rozdílné alignmenty.

Můžou nastat dvě situace, první možnost se označuje pojmem spanning read a nastává v případě, kdy jsou spárované ready synchronizovány inverzními stranami a zároveň nedochází k přímému překrývání oblastí dvou rozdílných genů. Druhá možnost se označuje jako split (rozdělené) ready a označuje místa fúzního spojení (fusion junction), tedy místa, kde dochází k návaznosti readů z rozdílných vzorků. Rozdíly mezi spanning a split ready jsou znázorněny na Obrázku 2.8. [21]

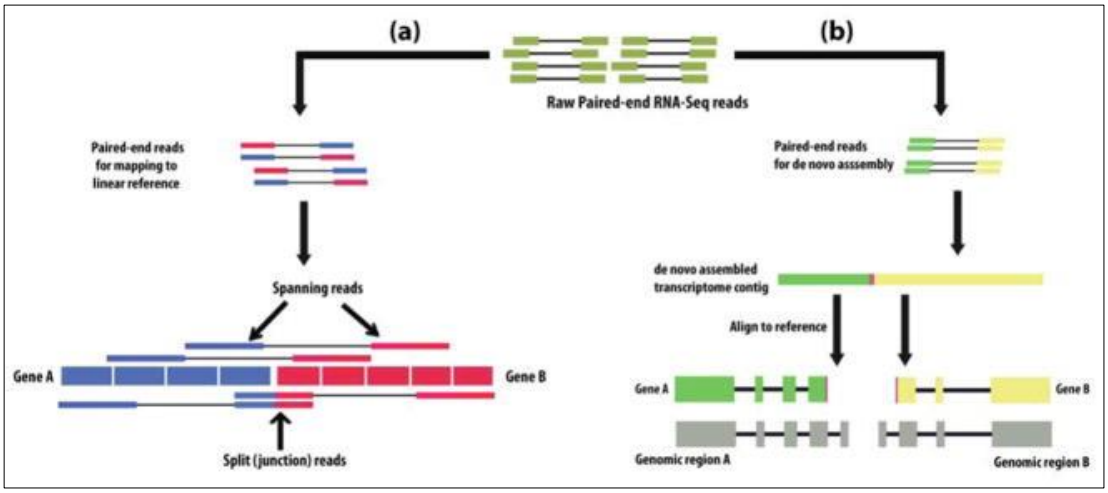
K identifikace fúzního spojení za pomoci split readů je potřeba nejprve ready rozdělit na menší fragmenty, nalézt vhodný alignment a ten opakovat do nalezení přesného

umístění fúzního spojení. Spanning ready lze rovněž použít k nalezení fúzních spojení, a to pomocí zpětného ověření pomocí split readů.

Pro eliminaci velkého množství falešně pozitivních readů obsahují nástroje pro detekci genové fúze velké množství různých filtrovacích funkcí.

Zjednodušeně řečeno, metoda mapping-first nejprve zarovná ready k referenčního genomu a mapované ready jsou pak sestavovány do exonů a případně transkriptů. Naproti tomu metoda assembly-first nejprve sestaví ready na základě jejich překrývání a takto sestavené sekvence (odpovídající sadám exonů) jsou pak zarovnány s referenčním genomem.

Dlouho se předpokládalo, že metoda mapping-first je omezena pouze na nemodelové druhy, kde není k dispozici žádný (nebo aspoň odpovídající) referenční genom, a navíc se zdála být nedostatečná i v případech, kdy byl k dispozici anotovaný referenční genom. Nedávný pokrok v sestavování transkriptomů formou „de novo“ tento pohled na věc jasně mění. Aplikace metody assembly-first s datasetem lidské RNA však stále zůstává vzácná, ačkoli některé studie již ukázaly její potenciál pro detekci nových biologicky relevantních variant splicingu, což je proces, kterým jsou introny, nekódující oblasti genů, vystřiženy z primárního transkriptu mRNA a exony (tj. kódující oblasti) jsou spojeny dohromady za vzniku zralé mRNA, který následně slouží jako předloha pro syntézu specifického proteinu. [21]



Obrázek 2.8: Srovnání Mapping-first approach (a) vs. Assembly-first approach (b)[20]

Nástroje pro analýzu genové fúze

Jak již bylo řečeno v předchozích kapitolách, přítomnost fúzních genů je spojována s iniciací a progresí nádorového bujení. Ovšem správné určení a hledání fúzních genů v datasetech RNA-seq komplikuje velká míra falešně pozitivních nálezů. Fúzní geny se ale většinou nacházejí v nemocných buňkách, a proto se očekává, že počet fúzních genů nalezených ve zdravých vzorcích bude nulový nebo velmi blízký nule. [22]

Pro identifikaci fúzních genů s využitím výstupních dat RNA-seq bylo vyvinuto velké množství výpočetních nástrojů, jako je např. FusionSeq, FusionMap, Tophat-fusion, PRADA, SOAPfuse, JAFFA, ChimPipe, FusionCatcher nebo Arriba. [18]

Všechny nástroje vyjmenované v odstavci výše, pracují na základě tzv. zarovnávačů neboli aligners, které jsou nejdůležitějším softwarem používaným v oblasti transkriptomických studií a příbuzných oborech. Všechny alignery mohou být nakonfigurovány tak, aby poskytovaly dobré výsledky, ale přesto výzkumníci a vědci čelí výzvám při výběru přesného, citlivého, vyžadujícího méně hardwarových zařízení, a nakonec vhodného pro jejich výzkumné cíle. [25]

V rámci mé bakalářské práce provádím analýzu genových fúzí z dat RNA-seq. Pro provedení této analýzy jsem si zvolila 2 algoritmy, provádějící hledání fúzí – Arriba a FusionCatcher. V následujících podkapitolách je blíže představím.

FusionCatcher

V této části představím softwarový nástroj FusionCatcher, který hledá nové ale také již známé fúzní geny, translokace a chiméry (jedinec se nevyvinul z jedné buňky (zygoty) ale ze dvou buněk) v datech RNA-seq (read na párovém konci (paired-end)) nebo na jednom konci(single-end)) a pracuje s anotačními daty dostupnými v databázi Ensembl.

Hlavními účely tohoto nástroje je rychlost validace PCR v reálném čase (tj. přesnost), která umožňuje praktické ověření kandidátních fúzních genů, ale také rychlost detekce (tj. senzitivita) fúzních genů.

Pro detekci genové fúze nástroj FusionCatcher nejprve na datech RNA-seq provede předběžné zpracování a filtrování. Kvalitní filtrování readů se provádí pomocí:

a) Odstranění readů, které se zarovnávají na ribozomální/transferovou RNA, mitochondriální DNA, HLA geny nebo známé genomy virů, fágů nebo bakterií,
b) oříznutí readů obsahujících adaptéry a poly-A/C/G/T koncovky,
c) ořezávání readů na základě skóre kvality,
d) odstranění readů, která jsou sekvenátorem označena jako nekvalitní. [22]

FusionCatcher automaticky vybere nejlepší parametry pro nalezení kandidátních fúzních genů, např. automatické vyhledání adaptérů, kvalitní oříznutí readů, automatické vytvoření spojení exon-exon na základě délky vstupních readů atd.

Nemapované ready, což jsou ready, které prošly filtrací kvality a nemapují se na transkriptom nebo na genom, se uchovávají pro další analýzy.

Ready, které prošli filtrací se dále používají k sestavení předběžného seznamu (tzv. preliminary list) fúzních genů, a to hledáním genových párů z genů A, B, kdy jeden read mapuje oblast transkriptomu genu A, která odpovídá readu na párovém konci (paired-end) genu B.

Z předběžného seznamu jsou následně odstraněny páry genů pomocí známých a nových kritérií, která dávají biologický smysl, jako např:

a) Oba geny jsou paralogem toho druhého v databázi Ensembl,
b) nalezený gen je pseudogen toho druhého v databázi Ensembl,
c) fúze je a priori známá jako falešně pozitivní událost,
d) již dříve nalezena ve vzorcích od zdravých osob, jako například z dat sekvenování RNA Illumina Body Map 2.0 nebo vlastní databáze RNA-seq zdravých vzorků,
e) oba geny se navzájem překrývají na stejném řetězci podle jedné z veřejně známých databází, jako jsou databáze Ensembl, UCSC nebo RefSeq,
f) pár genů s velmi vysokým množstvím simultánně mapovaných readů. [22]

K identifikaci fúzních spojení se dále používá jedna ze čtyř různých metod a čtyř různých alignerů. Každá metoda odpovídá jednomu aligneru (Bowtie, BLAT, STAR a Bowtie2).

Arriba

Nástroj Arriba byl vyvinut pro detekci genových fúzí z dat RNA-seq pro použití v prostředí klinického výzkumu. Proto byly krátké doby běhu a vysoká citlivost důležitými kritérii návrhu. Je založen na ultrarychlém zarovnávači STAR a doba alignmentu je obvykle několik desítek minut až hodinu. Kromě genových fúzí může Arriba detekovat další strukturální přestavby s potenciální klinickou relevancí, jako jsou virová integrační místa, interní tandemové duplikace, duplikace celých exonů, zkrácení genů (tj. body zlomu v intronech a mezi genových oblastech). [23]

Aligner STAR (Spliced Transcripts Alignment to a Reference) je založen na dříve nepopsaném algoritmu zarovnání RNA-seq, který používá sekvenční vyhledávání v nekomprimovaných oblastech. STAR překonává ostatní alignery faktorem >50 v rychlosti mapování, přirovnává k lidskému genomu 550 milionů 2 × 76 bp párovaných readů za hodinu a zároveň zlepšuje citlivost a přesnost alignmentu. Kromě de novo detekce je také schopen mapovat sekvence RNA v plné délce. [24]

Arriba je vítězem DREAM SMC-RNA Challenge, mezinárodní soutěže organizované ICGC, TCGA, IBM a Sage Bionetworks s cílem určit současný zlatý standard pro detekci genových fúzí z dat RNA-seq. [23]

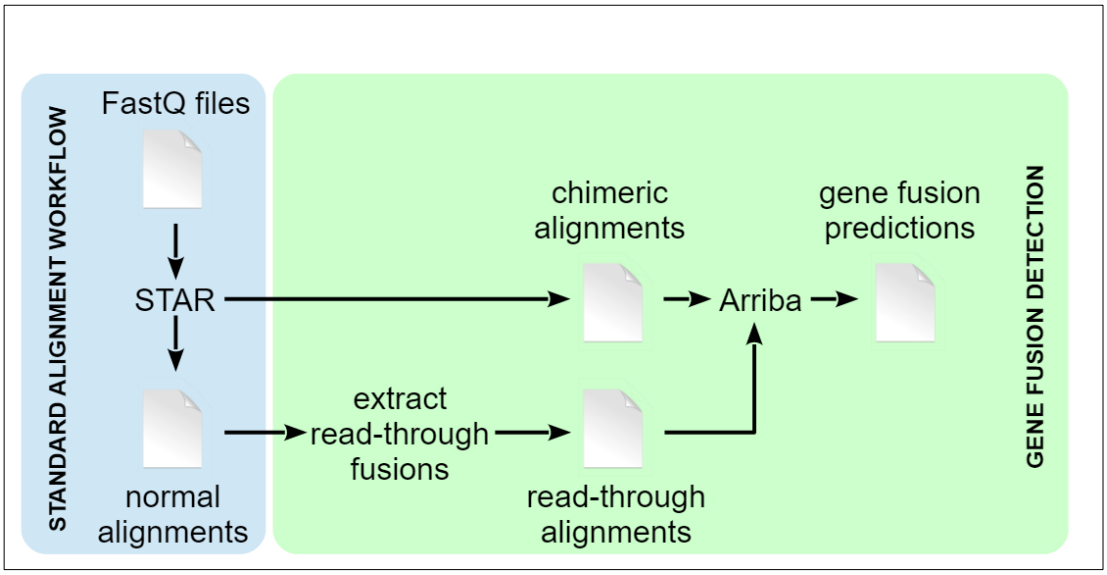
Pracovní postup nástroje Arriba, který je znázorněn na Obrázku 2.9 je o poznání jednodušší ve srovnání s nástrojem FusionCatcher, který kombinuje 4 různé metody a 4 alignery. Arriba provádí detekci fúzních genů ve třech krocích s použitím jediného aligneru STAR:

a) Nejprve dochází k detekci chimérických readů za použití aligneru STAR, tím se vygeneruje další výstupní soubor obsahující všechny chimérické alignmenty (rozdělené ready a nesouhlasné vazby). Tento soubor obsahuje důkazy

o translokacích, inverzích, duplikacích a delecích větších, než je obvyklá velikost intronu.

b) Soubor extract-read-through-fusions je výstupem skenu normální zarovnání/alignmentu pro ready, které potenciálně indikují přečtené transkripty nebo delece menší než obvyklá velikost intronu. Výstupy jsou do samostatného BAM souboru read-through alignments.

c) Nástroj Arriba zpracovává chimérický a read-throuh alignmenty. Aplikuje sadu filtrů k odstranění artefaktů (uměle vzniklá nebo pozměněná struktura buněk) a přepisů pozorovaných ve zdravé tkáni. Konečným výstupem je seznam vysoce spolehlivých předpovědí fúze. [23]



Obrázek 2.9: Workflow softwaru Arriba [23]

Tento výsledný seznam je řazen dle tzv. confidence scoring neboli tříd spolehlivosti, kdy je každý read označen buď nízkou, střední nebo vysokou hodnotou možnosti vzniky genové fúze. Je možné upravovat dle preferované rovnováhy mezi senzitivitou a specifičností výběrem událostí nad určitou třídou spolehlivosti. Toto hodnocení reflektuje tři aspekty, konkrétně pravděpodobnost, že:

a) Transkript je aberantní (nebyl zaznamenán ve zdravé tkáni),
b) jedná se o změnu v nekódující oblasti tzv. tichou mutaci, která se většinou neprojevuje,
c) nejedná se o artefakt. [26]

Při hledání recidivujících fúzovaných genů ve výstupních datech je vhodné brát v úvahu pouze středně a vysoce spolehlivé předpovědi, jinak budou výsledky obohaceny o falešně pozitivní výsledky. Ale v situacích, kdy je citlivost rozhodující, mohou mít předpovědi s nízkou spolehlivostí vysokou hodnotu. Například v onkologii založené

na HTS je přijatelný zvýšený počet falešně pozitivních předpovědí jako kompromis za vyšší citlivost, pokud jsou potenciálně relevantní předpovědi následně validovány.

Výsledný seznam obsahuje predikce fúze řazené od nejvyšší po nejnižší pravděpodobnost výskytu. Pořadí je nejvíce ovlivněno počtem podpůrných readů, ale bere v úvahu také mnoho dalších funkcí, jako je blízkost bodů zlomu (intragenní vs. read-through vs. distální), více variant transkriptů mezi stejným párem genů nebo hladina šumu pozadí v daném genu (e-hodnota). [23]

Přestože byly v průběhu let vyvinuty různé výpočetní nástroje pro detekci genové fúze, stále neexistuje žádný zlatý standard. Spolehlivá predikce genových fúzí z RNA-seq se ukázala být obtížná kvůli nesčetnému množství artefaktů, které byly zavedeny během přípravy knihovny a zarovnání sekvencí, jako je popsáno v předchozí kapitole o nástroji FusionCatcher. Aby byl počet falešně pozitivních předpovědí nízký, implementují algoritmy přísné filtry s nežádoucím vedlejším účinkem, kdy jsou bohužel občas vyřazeny drivery fúze a důkazy jsou následně v datech RNA-seq zcela ztraceny. Současná praxe aplikuje vždy alespoň dva nástroje a využívá spojení nebo průniku jejich předpovědí. Tento přístup je výpočetně nákladný, protože příprava každého nástroje obvykle trvá mnoho hodin, než se spustí. S technologií vysokokapacitního sekvenování (high-throughput sequencing, HTS), která je v klinické praxi stále běžnější pro identifikaci cílených změn, roste poptávka po algoritmech, které jsou přesné a účinné. Nepřesné předpovědi komplikují interpretaci výsledků založených na HTS a v případě onkologické studie se jedná o kritické zpomalení bioinformatických procesů zpracování.

2.3.4 Další využití RNA-seq dat

Získaná data z RNA-seq lze použít i pro množství dalších aplikací. Příkladem těchto aplikací je modelace transkripčních regulačních sítí (transcriptional regulatory networks) nebo sledování variací počtu kopií na genomu (copy number variation).

Změny v průběhu genové exprese mohou vyvolat řadu onemocnění, přesto víme stále relativně málo o tom, jak specifické transkripční faktory (bílkoviny, které iniciují proces transkripce genu) tyto změny nepřímo ovlivňují. Nedávné pokroky v genomických technologiích a výpočetním modelování způsobily revoluci ve schopnosti konstruovat tyto modely. Zevrubný popis transkripčních regulačních sítí souvisejících s daným onemocněním může pomoci objasnit potenciální mechanismy tohoto onemocnění a stanovit priority cílů pro vývoj nových léčiv či léčebných metod. [30]

Pokud se na genomu opakují určité úseky a počet opakování se v genomu liší, jedná se o strukturální variaci počtu kopií, a to buď o duplikace (i mnohonásobné) případně o delece. Varianty počtu kopií lze obecně rozdělit do dvou hlavních skupin, a to na krátká nebo dlouhá opakování. Krátká opakování zahrnují především dva opakující se nukleotidy (např. ACACAC...). Dlouhá opakování zahrnují repetice celých genů. Lidský genom se skládá přibližně až ze dvou třetin z repetic a 4,8–9,5 % lidského genomu

lze klasifikovat jako variace počtu kopií. U všech savců hrají tyto variace počtu kopií důležitou roli při vytváření nezbytných různorodých variací v populaci a také mohou ovlivnit fenotyp ve smyslu projevu určitého onemocnění. [31]

Rovněž využití RNA-seq dat k filtrování variant, tzv. variant calling má mnohostranné využití. Umožňuje používat snadno dostupná data RNA-seq k profilování vzorků pro známé varianty nebo umožňuje potvrzení variant, které byly detekovány sekvenováním genomu (např. validace somatických mutací souvisejících s rakovinou). Kromě toho umožňuje detekci dříve neznámých variant, které by mohly mít důležité funkční důsledky. [32]

3 Provedení RNA-seq analýzy dat

V této části bakalářské práce se zaměřím na provedení samotné bioinformatické analýzy RNA-seq, konkrétně na provedení hledání a interpretace fúzních genů, pro kterou jsem použila dva softwarové nástroje. Prvním vybraným byl nástroj Arriba, který se stal vítězem mezinárodní soutěže DREAM SMC-RNA Challenge a také první volbou i pro ÚHKT. Druhým nástrojem byl FusionCatcher, který ÚHKT zvažovalo, ale po pečlivé rešerši od jeho použití nakonec upustilo, a to z důvodu velkého množství falešně pozitivních nálezů. Bude tedy zajímavé porovnat výsledky z obou nástrojů a potvrdit či vyvrátit provedené rešerše.

K provedení analýzy byl vybrán dataset z volně dostupné online databáze obsahující čtyři vzorky, kdy každý z nich obsahoval několik již dříve klinicky ověřených fúzních genů, které byly přítomny u pacientek s karcinomem prsu. Jedná se o:

- Vzorek BT474 s ověřenými fúzními geny
  a) ACACA--STAC2, DIDO1--TTI1, CPNE1--PI3. GLB1--CMTM7, LAMP1--MCF2L, RAB22A--MYO9B, RPS6KB1--SNF8, SKA2--MYO19, STARD3--DOK5, VAPB--IKZF3, ZMYND8--CEP250 [34]
  b) AHCTF1—NAAA, MED1--ACSF2, MED1--STXBP4, MED13--BCAS3, PIP4K2B--RAD51C, STX16--RAE1, THRA--AC090627.1, TOB1—SYNRG, TRPC4AP--MRPL45, USP32--MED1 [35]
  c) LIMA1--USP22, ACACA--STAC2, FAM102A--CIZ1, GLB1--CMTM7, MED1--STXBP4, PIP4K2B--RAD51C, RAB22A--MYO9B, RPS6KB1--SNF8, STARD3--DOK5, TRPC4AP--MRPL45, ZMYND8--CEP250 [36]

- Vzorek KPL4 s ověřenými fúzními geny
  a) BSG—NFIX, PPP1R12A--SEPT10, NOTCH1--NUP214 [34]

- Vzorek MCF7 s ověřenými fúzními geny
  a) ARA46:A522--SULF2, BCAS4--BCAS3, RPS6KB1--VMP1 [34]
  b) AC099850.1--VMP1, GCN1L1--MSI1, SMARCA4--CARM1 [35]
  c) ADAMTS19--SLC27A6, ARFGEF2--SULF2, ATXN7L3--FAM171A2, BCAS4--BCAS3, GCN1L1--MSI1, RPS6KB1--DIAPH3, SULF2--PRICKLE2, MYH9--EIF3D [36]
  d) AHCYL1--RAD51C, ARFGEF2--SULF2, ARHGAP19--DRG1, BCAS4--BCAS3, PAPOLA--AK7, MYO9B--FCHO1 [37]

- Vzorek MCF7 s ověřenými fúzními geny
  a) ANKHD1--PCDH1, CCDC85C--SETD3, CSE1L--AL035685.1, CYTH1--EIF3H, DHX35—ITCH, NFS1--PREX1, PREX1--CPNE1,

RARA—PKIA, SUMF1--LRRFIP2, TATDN1—GSDMB, WDR67--ZNF704, KLHDC2--SNTB1 [34]

Seznam všech prokázaných fúzních genů v těchto vzorcích a odkazy na vědecké práce, kde byla potvrzena přítomnost u pacientek s rakovinou prsu jsou uloženy k nahlédnutí na přiloženém CD.

Analýza genových fúzí má v zásadě dvě fáze, samotné nalezení fúzních kandidátů (potencionální fúze se danou konfidencí) a jejich intepretace. V první fázi je tedy nutné pro daný vzorek spustit zvolené nástroje.

3.1 Spuštění nástrojů pro hledání fúzí

K oběma vybraným softwarovým nástrojům je možný přístup prostřednictvím technologie pro kontejnerizaci aplikací Docker. Ten umožňuje zabalení softwaru do izolovaného prostředí nezávislého na platformě a odstínění uživatelé od problémů spojených rozdílným prostředím OS, konfigurací a podobně.

V následujících kapitolách jsou prezentovány ukázky vzorového kódu pro spouštění jednotlivých nástrojů. Kompletní zdrojové kódy jsou k nahlédnutí na přiloženém CD.

3.1.1 Arriba

Na Obrázku 3.1 je uveden ukázkový příkaz pro spuštění nástroje Arriba za pomocí Dockeru. Uvedený příkaz je pro dataset vzorku s označením BT474 a mimo specifických konstruktů Dockeru obsahuje mapování souborů a složek pro běh nástroje Arriba: adresář, kam má být uložen výstup, adresář obsahující referenční genom a jeho anotace a cesty ke vstupním FASTQ souborům. Příkazy pro zbývající vybrané fúzní geny jsou vždy stejné, mění se jen název vstupního FASTQ souboru.

```
docker run --rm \
       -v /home/kbi/data2/irma/data/results/BT474:/output \
       -v/home/kbi/data2/irma/data/references/arriba:/references:ro \
       -v /home/kbi/data2/irma/data/BT474.Left.fq.gz:/read1.fastq.gz:ro \
       -v /home/kbi/data2/irma/data/BT474.Right.fq.gz:/read2.fastq.gz:ro \
       uhrigs/arriba:2.2.1 \
       arriba.sh
```

Obrázek 3.1: Příkazy pro spuštění nástroje Arriba

3.1.2 FusionCatcher

Na Obrázku 3.2 je uveden ukázkový příklad kódu pro spuštění nástroje FusionCatcher opět za pomocí Dockeru. Struktura příkazu pro spuštění FusionCatcheru je analogická k příkazu pro spuštění Arriby.

```
docker run -d –rm
 -v
/home/kbi/data2/references/fusioncatcher_ref/data/human_v102:/reference \
-v /home/kbi/data2/irma/data/BT474.Left.fq.gz:/input/r1.fastq.gz \
-v /home/kbi/data2/irma/data/BT474.Right.fq.gz:/input/r2.fastq.gz \
-v /home/kbi/data2/irma/data/results/BT474/fusioncatcher:/output
irma/fusioncatcher:1.33
fusioncatcher_run.sh
```

Obrázek 3.2: Příkazy pro spuštění nástroje FusionCatcher

4 Výsledky

Pro lepší interpretaci výsledků z obou nástrojů byl použit porovnávací software Fusion Combiner, který vytvořil vedoucí práce Ing. Bohuslav Dvorský. Tento software je volně dostupný ke stažení na stránkách GitLab (https://gitlab.com/bdvorsky/fusion-combiner). Kompletní vstupní, výstupová i srovnávací data všech vzorků jsou k nahlédnutí na přiloženém CD.

4.1 Vzorek BT474

Obrázek 4.1 znázorňuje ukázku výsledných data ze vzorku BT474. V prvním sloupci označeném jako „Detector“ je vypsán vždy ten nástroj, který fúzní gen zachytil, např. na prvním řádku je zřejmé, že stejný nález zachytily oba nástroje. Sloupec „breakpoint1“ značí čtení zleva a „breakpoint2“ čtení zprava. Ve sloupci „fusion“ jsou uvedeny jednotlivé fúzní geny, např. na prvním řádku se jedná o spojení genů THRA a AC090627.1, jejich spojením vznikl fúzní gen s označením THRA--AC090627.1. Sloupec „breakpoints by tool“ označuje přesné pozice bodu zlomu daného nálezu, což je přesnější než jen pouhé označení fúzního genu. Zbývající sloupce „spanning reads“, „split reads1“ a „split reads2“ uvádí zachycení počet těchto readů jednotlivými nástroji.

Detector	breakpoint1	breakpoint2	fusion	breakpoints by tool	spanning reads	split_reads1	split_reads	split_read
Arriba + FusionCa	17:40086853	17:48294347	Arriba:THRA--AC090627.1 ; FusionCatcher:THRA--THRA1/BTR	Arriba: 17:40086853 -- 17:48294347 ; FusionCatcher: 17:40086853 -- 17:48294347	Arriba: 106 ; FusionCatcher: 101	Arriba: 44	Arriba: 29	
Arriba + FusionCa	17:40086853	17:48307331	Arriba:THRA--AC090627.1 ; FusionCatcher:THRA--THRA1/BTR	Arriba: 17:40086853 -- 17:48307331 ; FusionCatcher: 17:40086853 -- 17:48307331	Arriba: 19 ; FusionCatcher: 101	Arriba: 16	Arriba: 6	
FusionCatcher	6:73492688	17:39715286	FusionCatcher:EEF1A1--ERBB2	FusionCatcher: 6:73492688 -- 17:39715286	FusionCatcher: 8			
FusionCatcher	20:34417239	17:39715286	FusionCatcher:ITCH--ERBB2	FusionCatcher: 20:34417239 -- 17:39715286	FusionCatcher: 2			
FusionCatcher	17:47592545	17:38191030	FusionCatcher:NPEPPS--TBC1D	FusionCatcher: 17:47592545 -- 17:38191030	FusionCatcher: 63			
Arriba	17:59893325	17:48943975	Arriba:RPS6KB1--SNF8	Arriba: 17:59893325 -- 17:48943975	Arriba: 48	Arriba: 8	Arriba: 12	
Arriba	17:37122531	17:39218173	Arriba:ACACA--STAC2	Arriba: 17:37122531 -- 17:39218173	Arriba: 47	Arriba: 33	Arriba: 6	
FusionCatcher	20:58389517	17:39788374	FusionCatcher:VAPB--IKZF3	FusionCatcher: 20:58389517 -- 17:39788374	FusionCatcher: 47			
Arriba	17:50866058	17:37520648	Arriba:TOB1--SYNRG	Arriba: 17:50866058 -- 17:37520648	Arriba: 45	Arriba: 7	Arriba: 8	
Arriba	20:58389517	17:39777767	Arriba:VAPB--IKZF3	Arriba: 20:58389517 -- 17:39777767	Arriba: 40	Arriba: 8	Arriba: 7	
Arriba	1:246931578	4:75925811	Arriba:AHCTF1--NoneAA	Arriba: 1:246931578 -- 4:75925811	Arriba: 31	Arriba: 1	Arriba: 3	

Obrázek 4.1: Ukázka porovnání výsledků vzorku BT474

Na Obrázku 4.1 jsou žlutě vyznačené ověřené fúzní geny, kterých se ve vzorku BT474 nacházelo celkem 23 a z nichž bylo nalezeno celkem 18 fúzních genů (z tohoto počtu jich 16 nalezl nástroj Arriba a pouhých 5 nástroj FusionCatcher, některé fúzní geny měly rozdílné body zlomu, není tedy možné výsledky obou nástrojů sčítat). Z čehož vyplývá, že 5 ověřených fúzních genů nebylo nalezeno ani jedním nástrojem.

Ve výsledcích je uvedeno dalších 28 nenalezených fúzí, 22 z nich identifikoval nástroj Arriba příznakem „high“ v hodnocení confidence score. 6 jich nalezl nástroj FusionCatcher a polovina z nich byla označena příznakem „oncogene“, druhá polovina jako „probably false positive“.

4.2 Vzorek KPL4

Obrázek 4.2 znázorňuje ukázku výsledných data ze vzorku KPL4. Žlutě jsou vyznačené jsou 3 ověřené fúzní geny ve vzorku KPL4 a jejich přítomnost byla analýzou potvrzena. Nástroj Arriba nalezl všechny 3, FusionCatcher pouze 2 ověřené fúzní geny.

Detector	breakpoint1	breakpoint2	fusion	breakpoints by tool	spanning reads	split_reads1	split_reads2
Arriba + FusionCatcher	19:580782	19:13025021	Arriba:BSG--NFIX ; FusionCatcher:BSG--NFIX	Arriba: 19:580782 -- 19:13025021 ; FusionCatcher: 19:580782 -- 19:13025021	Arriba: 23 ; FusionCatcher: 19	Arriba: 1	Arriba: 5
Arriba + FusionCatcher	9:136544024	9:131187289	Arriba:NOTCH1--NUP214 ; FusionCatcher:NOTCH1--NUP214	Arriba: 9:136544024 -- 9:131187289 ; FusionCatcher: 9:136544024 -- 9:131187289	Arriba: 7 ; FusionCatcher: 6	Arriba: 2	Arriba: 1
FusionCatcher	17:47592545	17:38191030	FusionCatcher:NPEPPS--TBC1D3	FusionCatcher: 17:47592545 -- 17:38191030	FusionCatcher: 6		
Arriba	18:2707920	11:92263300	Arriba:SMCHD1--RPL7AP57(101411),NDUFB11P1(72732)	Arriba: 18:2707920 -- 11:92263300	Arriba: 5	Arriba: 1	Arriba: 1
Arriba	9:134165399	9:137956771	Arriba:RNU6ATAC(835),BX649601.1(3370)--CACNone1B	Arriba: 9:134165399 -- 9:137956771	Arriba: 4	Arriba: 0	Arriba: 1
Arriba	12:79817394	2:109585838	Arriba:PPP1R12A--SEPT10	Arriba: 12:79817394 -- 2:109585838	Arriba: 2	Arriba: 3	Arriba: 1
Arriba	9:120957285	9:87156573	Arriba:C5--C9orf170	Arriba: 9:120957285 -- 9:87156573	Arriba: 1	Arriba: 1	Arriba: 1
Arriba	13:26613058	9:134822998	Arriba:WASF3--COL5A1	Arriba: 13:26613058 -- 9:134822998	Arriba: 1	Arriba: 0	Arriba: 1
Arriba	4:1904378	4:4183499	Arriba:NSD2--OR7E43P(8261),OTOP1(5304)	Arriba: 4:1904378 -- 4:4183499	Arriba: 1	Arriba: 0	Arriba: 1
Arriba	9:112716461	9:120957368	Arriba:INIP--C5	Arriba: 9:112716461 -- 9:120957368	Arriba: 1	Arriba: 1	Arriba: 0
Arriba	9:112716461	9:113256548	Arriba:INIP--CDC26	Arriba: 9:112716461 -- 9:113256548	Arriba: 1	Arriba: 1	Arriba: 0
Arriba	21:39293811	21:9077714	Arriba:BRWD1--TEKT4P2	Arriba: 21:39293811 -- 21:9077714	Arriba: 0	Arriba: 4	Arriba: 1

Obrázek 4.2: Ukázka porovnání výsledků vzorku KPL4

Ve výsledcích je uvedeno dalších 9 nenalezených fúzí, 8 z nich identifikoval nástroj Arriba opět příznakem „high“ v hodnocení confidence score. Pouze 1 nástroj FusionCatcher a tento nález byl označen jako „probably false positive“.

4.3 Vzorek MCF7

Na Obrázku 4.3 jsou žlutě vyznačené ověřené fúzní geny, kterých se ve vzorku MCF7 nacházelo celkem 16, polovina z nich byla nalezena a polovina nikoliv. Nástroj Arriba nalezl všech 8 fúzí, nástroj FusionCatcher pouhé 2.

Detector	breakpoint1	breakpoint2	fusion	breakpoints by tool	spanning reads	split_reads1	split_reads2
Arriba + FusionCatcher	20:50795173	17:61368327	Arriba:BCAS4--BCAS3 ; FusionCatcher:BCAS4--BCAS3	Arriba: 20:50795173 -- 17:61368327 ; FusionCatcher: 20:50795173 -- 17:61368327	Arriba: 136 ; FusionCatcher: 102	Arriba: 11	Arriba: 6
FusionCatcher	20:50795173	17:61353588	FusionCatcher:BCAS4--BCAS3	FusionCatcher: 20:50795173 -- 17:61353588	FusionCatcher: 102		
Arriba	3:63913225	1:106216305	Arriba:ATXN7--AL355306.2(91741),AL499605.1(328037)	Arriba: 3:63913225 -- 1:106216305	Arriba: 13	Arriba: 2	Arriba: 1
Arriba	3:63913225	1:106216305	Arriba:ATXN7--AL355306.2(91741),AL499605.1(328037)	Arriba: 3:63913225 -- 1:106216305	Arriba: 13	Arriba: 2	Arriba: 1
Arriba	20:48922010	20:47736942	Arriba:ARFGEF2--SULF2	Arriba: 20:48922010 -- 20:47736942	Arriba: 12	Arriba: 7	Arriba: 7
Arriba	20:47502019	1:106581397	Arriba:NCOA3--AL499605.1(36653),AL596327.1(198826)	Arriba: 20:47502019 -- 1:106581397	Arriba: 10	Arriba: 3	Arriba: 4
Arriba	17:59107591	17:59838295	Arriba:AC099850.1--VMP1	Arriba: 17:59107591 -- 17:59838295	Arriba: 6	Arriba: 2	Arriba: 2
Arriba	17:59914703	17:59839768	Arriba:RPS6KB1--VMP1	Arriba: 17:59914703 -- 17:59839768	Arriba: 4	Arriba: 2	Arriba: 3
Arriba	17:60747279	3:63718546	Arriba:BCAS3--SNTN(39526),AC104162.1(23281)	Arriba: 17:60747279 -- 3:63718546	Arriba: 3	Arriba: 1	Arriba: 4
Arriba	17:60600886	17:60265583	Arriba:PPM1D--USP32	Arriba: 17:60600886 -- 17:60265583	Arriba: 3	Arriba: 1	Arriba: 1
Arriba	19:10986593	19:10904951	Arriba:SMARCA4--CARM1	Arriba: 19:10986593 -- 19:10904951	Arriba: 3	Arriba: 1	Arriba: 1
Arriba	19:17102557	19:17770425	Arriba:MYO9B--FCHO1	Arriba: 19:17102557 -- 19:17770425	Arriba: 3	Arriba: 0	Arriba: 1

Obrázek 4.3: Ukázka porovnání výsledků vzorku MCF7

Ve výsledcích je uvedeno dalších 19 nenalezených fúzí, všech 15 z nich identifikoval nástroj Arriba s příznakem „high“ v hodnocení confidence score, a pouze 4 nástroj FusionCatcher, kdy jeden byl označen příznakem „probably false positive“, jeden příznakem „already known fusion“ a dva se již dříve nacházely v odlišných databázích, kde byly označeny jako „cancer“.

4.4 Vzorek SKBR3

Na Obrázku 4.4 jsou žlutě vyznačené ověřené fúzní geny, kterých se ve vzorku SKBR3 nacházelo celkem 12 a z nichž bylo nalezeno celkem 10 fúzních genů (z tohoto počtu jich 8 nalezl nástroj Arriba a pouhé 3 nástroj FusionCatcher, některé fúzní geny měly rozdílné body zlomu, není tedy možné výsledky obou nástrojů sčítat). Z čehož vyplývá, že 2 ověřené fúzní geny nebyly nalezeny ani jedním nástrojem.

Detector	breakpoint1	breakpoint2	fusion	breakpoints by tool	spanning reads	split_reads1	split_reads2
Arriba + FusionCatcher	8:124539025	17:39909924	Arriba:TATDN1--GSDMB ; FusionCatcher:TATDN1--GSDMB	Arriba: 8:124539025 -- 17:39909924 ; FusionCatcher: 8:124539025 -- 17:39909924	Arriba: 149 ; FusionCatcher: 144	Arriba: 2	Arriba: 170
Arriba + FusionCatcher	8:124539025	17:39906271	Arriba:TATDN1--GSDMB ; FusionCatcher:TATDN1--GSDMB	Arriba: 8:124539025 -- 17:39906271 ; FusionCatcher: 8:124539025 -- 17:39906271	Arriba: 13 ; FusionCatcher: 144	Arriba: 1	Arriba: 30
Arriba + FusionCatcher	8:124539025	17:39905985	Arriba:TATDN1--GSDMB ; FusionCatcher:TATDN1--GSDMB	Arriba: 8:124539025 -- 17:39905985 ; FusionCatcher: 8:124539025 -- 17:39905985	Arriba: 11 ; FusionCatcher: 144	Arriba: 5	Arriba: 53
FusionCatcher	6:73492688	17:39715286	FusionCatcher:EEF1A1--ERBB2	FusionCatcher: 6:73492688 -- 17:39715286	FusionCatcher: 4		
FusionCatcher	10:73377445	12:52898899	FusionCatcher:ANXA7--KRT8	FusionCatcher: 10:73377445 -- 12:52898899	FusionCatcher: 2		
FusionCatcher	8:124538928	17:39909924	FusionCatcher:TATDN1--GSDMB	FusionCatcher: 8:124538928 -- 17:39909924	FusionCatcher: 144		
Arriba	17:78782202	8:116756019	Arriba:CYTH1--EIF3H	Arriba: 17:78782202 -- 8:116756019	Arriba: 46	Arriba: 8	Arriba: 5
FusionCatcher	17:39730263	17:39742252	FusionCatcher:ERBB2--GRB7	FusionCatcher: 17:39730263 -- 17:39742252	FusionCatcher: 33		
Arriba	8:124538928	17:39905985	Arriba:TATDN1--GSDMB	Arriba: 8:124538928 -- 17:39905985	Arriba: 32	Arriba: 3	Arriba: 3
Arriba	20:49072453	20:49340320	Arriba:CSE1L--AL035685.1	Arriba: 20:49072453 -- 20:49340320	Arriba: 24	Arriba: 7	Arriba: 7
Arriba	8:124539025	17:39909042	Arriba:TATDN1--GSDMB	Arriba: 8:124539025 -- 17:39909042	Arriba: 20	Arriba: 0	Arriba: 1
Arriba	17:78782202	8:116726172	Arriba:CYTH1--EIF3H	Arriba: 17:78782202 -- 8:116726172	Arriba: 20	Arriba: 0	Arriba: 1

Obrázek 4.4: Ukázka porovnání výsledků vzorku SKBR3

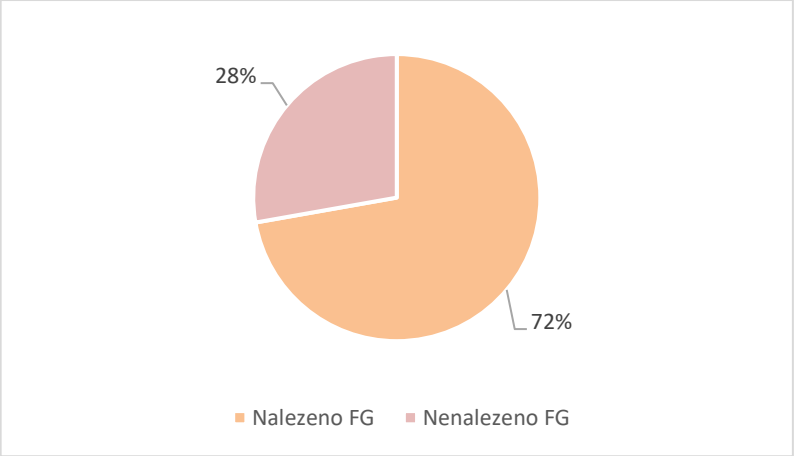
Ve výsledcích je uvedeno dalších 19 nenalezených fúzí, 12 z nich identifikoval nástroj Arriba opět s příznakem „high“ v hodnocení confidence score, a pouze 7 nástroj FusionCatcher, z nichž byly dva označeny příznakem „probably false positive“, dva příznakem „already known fusion“, jeden příznakem „probably novel fusion“ a zbývající dva se již dříve nacházely v odlišných databázích, kde byly označeny jako „oncogene“.

5 Diskuse

Provedená bioinformatická RNA-seq analýza měla za úkol nejen v praxi ověřit informace získané předchozí rešerší, ale také potvrdit best practices zjištěné z konzultace na ÚHKT, tedy že nástroj Arriba produkuje spolehlivější výstupní data před nástrojem FusionCatcher, který udává velké množství falešně pozitivních nálezů.

Veřejně dostupné vzorky sekvenačních dat, které obsahovaly i klinicky potvrzené fúzní geny byly použity jako vstup pro nástroje Arriba a FusionCatcher, které automaticky provedly zarovnání k referenci, detekci a selekci možných fúzních kandidátů.

Výstupní data z obou nástrojů byla následně ještě softwarově porovnána. Zjišťovala jsem, zda byly vybrané fúzní geny nalezeny oběma nástroji. Z celkového množství 54 ověřených fúzních genů jich bylo nalezeno celkem 39, celkem 15 jich zůstalo nenalezeno. V grafu na Obrázku 5.1 je vyjádřena procentuální úspěšnost nálezů.



Obrázek 5.1: Úspěšnost nálezu ověřených fúzních genů [autor práce]

Z celkového počtu 39 objevených fúzních genů ze seznamu jich celkem 35 objevil nástroj Arriba. FusionCatcher má v tomto ohledu o mnoho horší úspešnost nálezů a to pouhých 12. Některé fúzní geny byly identifikovány oběma nástroji, proto jejich součet neodpovídá celkovému počtu nalezených fúzních genů ze seznamu.

Výstupní data ovšem neobsahovala pouze vybrané fúzní geny, ale také další varianty, kdy se může jednat o zatím nenalezené fúzní geny. U těchto variant není možné jednoznačné určit, zda se jedná o relevantní nález, protože každý gen je jinak exprimován (má jiný počet čtení), proto relevantní fúze u málo exprimovaných genů může mít jenom jeden spanning read a budeme o relevanci uvažovat a u jiného genu nám jich nestačí ani 200. V případě, že je hodnota poměru součtu spanning reads a split reads vůči aplikované

coverage vyšší než 0,1, pak se dá o relevanci uvažovat, ale není to striktní pravidlo. Je důležité mít s výstupními daty předchozí zkušenost, protože spousta fúzí se odfiltruje na základě zkušeností.

V každém případě je nutné další zkoumání. To může být provedeno buď na dalším nástroji (např. CICERO) nebo případně laboratorní metodou pro potvrzení přítomnosti daného readu. Nástroj Arriba tento počet spanning a split readů započítává do tzv. confidence scoring. Nástroj FusionCatcher obsahuje knihovnu různých databází, dle kterých je možné kandidáty na fúzní gen před vybrat. V případě, že nález obsahuje malý počet důkazů, je možné se na něj zaměřit a přítomnost potvrdit jiným způsobem, např. laboratorně. Dále je potřeba ve spolupráci s ošetřujícím lékařem provést anotaci, tzn. zjistit z vědeckých publikací, zda byla prokázána přítomnost dané varianty genu u onkologických pacientů.

Z celkového počtu 75 nalezených fúzních genů mimo seznam, jich většinu, celkem 58 objevila Arriba. Všechny tyto nálezy byly hodnoceny příznakem high confidence score, což značí potencionální relevantní nález fúze. Z celkového počtu jich 18 nalezl přístroj FusionCatcher, což značí, že stejný gen byl oběma přístroji identifikován pouze v jednom případě a jednalo se o fúzní gen TBC1D31--ZNF704 ze vzorku SKBR3.

Nástroj FusionCatcher nalezl ve vzorcích BT474 a SKBR3 pokaždé fúzní gen EEF1A1--ERBB2 s označením „oncogene“, zatímco Arriba tento nález neidentifikovala.

Dále pak FusionCatcher ve vzorcích BT474 a KPL4 pokaždé označil fúzní gen NPEPPS--TBC1D3 jako falešně pozitivní nález.

Po provedení celé bioinformatické analýzy se mi potvrdila best practices používaná na ÚHKT, tedy že nástroj Arriba je v tomto ohledu spolehlivější. FusionCatcher v mnoha případech nenalezl ani přítomné ověřené fúzní geny ze seznamu, ani nálezy falešně pozitivních genů nebyly tak početné. Nástroj Arriba má tedy v případě provedení této bioinformatické analýzy vyšší senzitivitu než nástroj FusionCatcher, ať už při zjištění ověřených fúzních genů nebo při nálezech fúzních genů mimo seznam.

V práci se skrývá potenciál pro další směřování ať už bakalářských či diplomových prací. Poskytuje solidní základ pro výzkum v oblasti sekvenačních technologií či zdokonalování nástrojů a metod analýzy zejména v terciální analýze sekvenčních dat. Navazující práce by také mohly být zaměřeny na rozvoj metod pro zpřesňování terciální analýzy RNA-seq dat nebo implementaci a rozvoj technologií pro automatizaci bioinformatických procesů.

6 Závěr

Během druhé poloviny 20. století byly vyvinuty první metody sekvenování, které pomohly rozvoji molekulární biologie. Sekvenování DNA a RNA hraje důležitou roli nejen v jejich studiu, ale především v klinickém prostředí, kde může lékařům pomoci s určením vhodného léčebného postupu. Tato práce měla za cíl přiblížit vývoj sekvenačních metod a postup bioinformatické analýzy, při které je možné použít několik softwarových nástrojů. Rovněž měla za úkol vysvětlit rozdíly mezi WGS, WES a RNA-seq aplikacemi a objasnit best practices pro práci s RNA-seq daty. Všechna získaná fakta a postupy jsem následně aplikovala provedením vlastní bioinformatické analýzy a výstupní data interpretovala z pohledu bioinformatika a tím byly všechny cíle naplněny.

V současné době se metody sekvenování neustále vyvíjejí, a i díky rozvoji výpočetní techniky, která je schopna zpracovat nejen stále větší objem dat, ale pracovat i s větší přesností, bude v budoucnu role bioinformatika při stanovení diagnózy nebo návrhu léčebného postupu stále důležitější a více žádána. Je ovšem potřeba věnovat velký důraz výběru nástroje k provedení bioinformatické analýzy, aby práci bioinformatika a lékaře nezatěžoval, ale naopak ji ulehčil.

Seznam použité literatury

[1] PŘISTOUPILOVÁ, Anna. Využití nových metod analýzy genomu ve studiu molekulární podstaty vzácných geneticky podmíněných onemocnění. Praha, 2020. Disertační práce. Univerzita Karlova, 1. lékařská fakulta.

[2] BŮŽKOVÁ, Veronika. Přehled formátů uložení NGS dat a softwarových nástrojů pro jejich zpracování v jazyce Python. Kladno, 2021. Projekt I. České vysoké učení technické v Praze, Fakulta Biomedicínského inženýrství.

[3] RACLAVSKÝ, Vladislav. Metody molekulární genetiky [online]. Ústav biologie Lékařské fakulty Univerzity Palackého. Olomouc, 2003. [cit. 2022-03-27]. Dostupné z: https://web.archive.org/web/20100228153117/http://biologie.upol.cz/metody/Sekvenovani%20DNA.htm

[4] ŠTEFÁNEK, Jiří. Medicína, nemoci, studium na 1. LF UK [online], 2010. [cit. 2022-03-27]. Dostupné z: https://www.stefajir.cz

[5] ROSYPAL, Stanislav. Nový přehled biologie. Praha: Scientia, 2003. ISBN 978-80-86960-23-4.

[6] CHALUPOVÁ-KARLOVSKÁ, Vlastimila. Obecná biologie: evoluce, biologie buňky, genetika s 558 řešenými testovými otázkami: středoškolská učebnice. 3. opravené vydání. Olomouc: Nakladatelství Olomouc, 2018. ISBN 978-80-7182-305-6.

[7] PEARSON, Helen. What is a gene?. Nature [online]. 2006, 441(7092), 398-401 [cit. 2022-03-27]. ISSN 0028-0836. Dostupné z: doi:10.1038/441398a

[8] CRICK, FRANCIS. Central Dogma of Molecular Biology. Nature [online]. 1970, 227(5258), 561-563 [cit. 2022-03-27]. ISSN 0028-0836. Dostupné z: doi:10.1038/227561a0

[9] CAMPBELL, Molly. Transcription vs Translation Worksheet. Technology Networks [online]. 21. 08. 2019 [cit. 2022-03-27]. Dostupné z: http://www.news-courier.com/genomics/articles/transcription-vs-translation-worksheet-323080

[10] BAILEY, John. Nucleosides, Nucleotides, Polynucleotides (RNA and DNA) and the Genetic Code. Inventive Geniuses Who Changed the World [online]. Cham: Springer International Publishing, 2022, 2022-11-25, 313-340 [cit. 2022-05-11]. ISBN 978-3-030-81380-2. Dostupné z: doi:10.1007/978-3-030-81381-9_13

[11] STRATTON, Michael R., Peter J. CAMPBELL a P. Andrew FUTREAL. The cancer genome. Nature [online]. 2009, 458(7239), 719-724 [cit. 2022-03-27]. ISSN 0028-0836. Dostupné z: doi:10.1038/nature07943

[12] OTOVÁ, Berta, Milada KOHOUTOVÁ a Aleš PANCZAK. Lékařská biologie a genetika. Praha: Karolinum, 2013. ISBN ISBN978-80-246-1594-3.

[13] MATOUŠKOVÁ, Ph.D., Ing. Petra. Stanovení genové exprese. Hradec Králové, 2018. Habilitační práce. Univerzita Karlova. Farmaceutická fakulta v Hradci Králové. Dostupné z: https://dspace.cuni.cz/bitstream/handle/20.500.11956/105838/Habilita%C4%8Dn%C3%AD%20pr%C3%A1ce%20Matou%C5%A1kov%C3%A1.pdf?sequence=1&isAllowed=y

[14] CORCHETE, Luis A., Elizabeta A. ROJAS, Diego ALONSO-LÓPEZ, Javier DE LAS RIVAS, Norma C. GUTIÉRREZ a Francisco J. BURGUILLO. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Scientific Reports [online]. 2020, 10(1) [cit. 2022-03-27]. ISSN 2045-2322. Dostupné z: doi:10.1038/s41598-020-76881-x

[15] RNASeq a analýza diferenciální exprese. SEQme [online]. 2012–2022 [cit. 2022-03-28]. Dostupné z:

https://www.seqme.eu/cs/next-gen-sekvenovani/clanek/rnaseq-a-analyza-diferencialni-exprese

[16] COSTA-SILVA, Juliana, Douglas DOMINGUES, Fabricio Martins LOPES a Zhi WEI. RNA-seq differential expression analysis: An extended review and a software tool. PLOS ONE [online]. 2017, 12(12) [cit. 2022-03-28]. ISSN 1932-6203.

Dostupné z: doi: 10.1371/journal.pone.0190152

[17] LU, H a N VILLAFANE. Engineering and Functional Characterization of Fusion Genes Identifies Novel Oncogenic Drivers of Cancer. In: National Cancer Institute [online]. 2017 [cit. 2022-05-11]. Dostupné z:

https://ocg.cancer.gov/news-publications/publications/engineering-and-functional-characterization

[18] LATYSHEVA, Natasha S. a M. Madan BABU. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Research [online]. 2016, 44(10), 4487-4503 [cit. 2022-03-28]. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkw282

[19] HAAS, Brian J a Michael C ZODY. Advancing RNA-seq analysis. Nature Biotechnology [online]. 2010, 28(5), 421-423 [cit. 2022-03-28]. ISSN 1087-0156. Dostupné z: doi:10.1038/nbt0510-421

[20] KUMAR, Shailesh, Sundus Khalid RAZZAQ, Angie Duy VO, Mamta GAUTAM a Hui LI. Identifying fusion transcripts using next generation sequencing. WIREs RNA [online]. 2016, 7(6), 811-823 [cit. 2022-03-28]. ISSN 1757-7004.

Dostupné z: doi:10.1002/wrna.1382

[21] BENOIT-PILVEN, Clara, Camille MARCHET, Emilie CHAUTARD, et al. Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. Scientific Reports [online]. 2018, 8(1) [cit. 2022-03-28]. ISSN 2045-2322.

Dostupné z: doi:10.1038/s41598-018-21770-7

[22] NICORICI, Daniel a Mihaela ŞATALAN. FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data [online]. 2014 [cit. 2022-03-28]. Dostupné z: https://doi.org/10.1101/011650

[23] UHRIG, Sebastian, Julia ELLERMANN, Tatjana WALTHER, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. In: Genome Research [online]. 2021, s. 448-460 [cit. 2022-03-28]. ISSN 1088-9051.

Dostupné z: doi:10.1101/gr.257246.119

[24] DOBIN, Alexander, Carrie A. DAVIS, Felix SCHLESINGER, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [online]. 2013, 29(1), 15-21 [cit. 2022-03-28]. ISSN 1460-2059. Dostupné z: doi:10.1093/bioinformatics/bts635

[25] ABOLFAZL, Bahrami. Which Aligner Software is the Best for Our Study?. Journal of Genetics and Genome Research [online]. 2020, 7(1) [cit. 2022-03-28]. ISSN 23783648. Dostupné z: doi:10.23937/2378-3648/1410048

[26] Confidence scoring. Arriba [online]. [cit. 2022-03-28].

Dostupné z: https://arriba.readthedocs.io/en/latest/interpretation-of-results/#confidence-scoring

[27] SLABÝ, Ondřej. Technologie sekvenování nové generace: celogenomové, celoexomové a cílené – hotspot – sekvenování. ProLekare.cz [online]. 2018 [cit. 2022-03-28]. ISSN 1803-6597. Dostupné z: https://www.prolekare.cz/tema/precizni-medicina/detail/technologie-sekvenovani-nove-generace-celogenomove-celoexomove-a-cilene-hotspot-sekvenovani-105628

[28] UDAYANGANI, Samanthi. Rozdíl mezi genomem a exome [online]. 2017 [cit. 2022-03-28]. Dostupné z:

https://lafayettefirefighters.com/cs/difference-between-genome-and-vs-exome#Exome

[29] MACKENZIE, Ruairi J. RNA-Seq: Basics, Applications and Protocol. In: Technology Networks [online]. [cit. 2022-04-23]. Dostupné z: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461

[30] BABU, M Madan, Nicholas M LUSCOMBE, L ARAVIND, Mark GERSTEIN a Sarah A TEICHMANN. Structure and evolution of transcriptional regulatory networks. Current Opinion in Structural Biology [online]. 2004, 14(3), 283-291 [cit. 2022-04-23]. ISSN 0959440X. Dostupné z: doi:10.1016/j.sbi.2004.05.004

[31] DE KONING, A. P. Jason, Wanjun GU, Todd A. CASTOE, Mark A. BATZER, David D. POLLOCK a Gregory P. COPENHAVER. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. PLoS Genetics [online]. 2011, 7(12) [cit. 2022-04-24]. ISSN 1553-7404. Dostupné z: doi:10.1371/journal.pgen.1002384

[32] PISKOL, Robert, Gokul RAMASWAMI a Jin Billy LI. Reliable Identification of Genomic Variants from RNA-Seq Data. The American Journal of Human Genetics [online]. 2013, 93(4), 641-651 [cit. 2022-04-24]. ISSN 00029297. Dostupné z: doi:10.1016/j.ajhg.2013.08.008

[33] MURRAY, Robert K. Harperova Biochemie. Jinočany, 2002. Lange medical book. ISBN 80-731-9013-3.

[34] EDGREN, Henrik, Astrid MURUMAGI, Sara KANGASPESKA, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biology [online]. 2011, 12(1) [cit. 2022-05-04]. ISSN 1465-6906. Dostupné z: doi:10.1186/gb-2011-12-1-r6

[35] KANGASPESKA, Sara, Susanne HULTSCH, Henrik EDGREN, Daniel NICORICI, Astrid MURUMÄGI, Olli KALLIONIEMI a Janet SHIPLEY. Reanalysis of RNA-Sequencing Data Reveals Several Additional Fusion Genes with Multiple Isoforms. PLoS ONE [online]. 2012, 7(10) [cit. 2022-05-04]. ISSN 1932-6203. Dostupné z: doi:10.1371/journal.pone.0048745

[36] ASMANN, Yan W., Asif HOSSAIN, Brian M. NECELA, et al. Nucleic Acids Research [online]. 2011, 39(15) [cit. 2022-05-04]. ISSN 1362-4962. Dostupné z: doi:10.1093/nar/gkr362

[37] MAHER, Christopher A., Nallasivam PALANISAMY, John C. BRENNER, et al. Proceedings of the National Academy of Sciences [online]. 2009, 106(30) [cit. 2022-05-04]. ISSN 0027-8424. Dostupné z: doi:10.1073/pnas.0904720106

Obsah přiloženého CD

17KBIBP_492260_Zadání.pdf…….Zadání bakalářské práce

17KBIBP_492260_Irma_Snaselova.pdf……..Text bakalářské práce

Abstrakt_CZ.pdf……Abstrakt a klíčová slova v českém jazyce

Abstract_ENG.pdf…… Abstrakt a klíčová slova v anglickém jazyce

17KBIBP_492260RNA-seq data.zip……Vstupní a výstupní data provedené analýzy