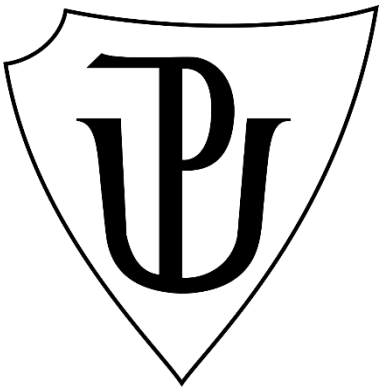
UNIVERZITA PALACKÉHO V OLOMOUCI

Přírodovědecká fakulta

Katedra biochemie



Bioinformatická analýza dat z formátu FASTQ

BAKALÁŘSKÁ PRÁCE

Autor:	**Natálie Haitlová**
Studijní program:	B1406 Biochemie
Studijní obor:	Bioinformatika
Forma studia:	Prezenční
Vedoucí práce:	**Ing. Květoslava Mahútová**
Rok:	2024

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

V Olomouci dne ....

Podpis studenta

Bibliografická identifikace

Jméno a příjmení autora	Natálie Haitlová
Název práce	Bioinformatická analýza dat z formátu FASTQ
Typ práce	Bakalářská
Pracoviště	Katedra biochemie
Vedoucí práce	Ing. Květoslava Mahútová
Rok obhajoby práce	2024

Abstrakt

V této bakalářské práci se zabývám bioinformatickou analýzou sekvenačních dat ze dvou platforem, konkrétně MGI a Illumina. Analyzovala jsem 7 pacientů s různými diagnózami a následně jsem výsledky analýzy porovnávala mezi platformami. Srovnávala jsem několik parametrů, mezi nimi počet kvalitních variant a sekvenačních chyb v reportech, dále třeba kvalitu genotypu a mapování. Následně jsem provedla výpočet parametrů potřebných k validaci NGS platformy. Cílem bylo zjistit, která platforma je lepší, k jasnému výsledku jsem však nedošla.

Klíčová slova	Bioinformatická analýza, validace, FASTQ, NEXTSEQ, MGI
Počet stran	58
Počet příloh	2
Jazyk	Český

Bibliographical identification

Autor’s first name and surname	Natálie Haitlová
Title	Bioinformatic analysis of data from the FASTQ format
Type of thesis	Bachelor
Department	Department of biochemistry
Supervisor	Ing. Květoslava Mahútová
The year of presentation	2024

Abstract

In this bachelor's thesis, I focused on the bioinformatic analysis of sequencing data from two platforms, specifically MGI and Illumina. I analyzed data from 7 patients with various diagnoses and subsequently compared the analysis results between the platforms. I compared several parameters, including the number of high-quality variants and sequencing errors in the reports, as well as genotype and mapping quality. Subsequently, I calculated parameters necessary for the validation of the NGS platform. The goal was to determine which platform is better, but I did not reach a clear conclusion.

Keywords	Bioinformatic analysis, validation, FASTQ, NEXTSEQ, MGI
Number of pages	58
Number of appendices	2
Language	Czech

OBSAH

CÍLE

Cílem mé bakalářské práce je porovnat výsledky NGS analýz reálného pacienta, které proběhly nezávisle na dvou odlišných platformách, konkrétně MGI a Illumina. Analyzovat budu parametry kvality sekvenačních dat včetně pokrytí, dále srovnám rozdíly v záchytech sekvenačních variant, pokud se vyskytnou. K analýze budu využívat data ve formátu FASTQ z rutinních běhů provedených dříve na platformách MGI a Illumina.

V teoretické části zpracuji přehled platforem pro NGS analýzy, popíšu jednotlivé kroky bioinformatické analýzy a objasním jednotlivé parametry pro srovnání daných platforem.

Praktickou částí pak bude samotná analýza a srovnání výsledků.

Tato bakalářská práce je realizována v Laboratořích AGEL v Novém Jičíně a bude využita k validaci nového NextSeq sekvenátoru.

1. Úvod

Bioinformatická analýza se stala nedílnou součástí biologického výzkumu v 21. století. Ve světě, kde jsou technologie pro získávání biologických dat na vzestupu, bioinformatici hrají klíčovou roli při zkoumání a porozumění obrovského množství informací obsažených v genomických, transkriptomických, proteomických a metabolomických datech. Bioinformatická analýza spojuje biologii s informačními technologiemi a statistickými metodami s cílem extrahovat, interpretovat a porovnávat biologické informace z rozsáhlých datových souborů. Tato disciplína umožňuje vědcům zkoumat genetické mechanismy onemocnění, porozumět biologickým procesům na molekulární úrovni a objevovat nové léky a terapeutické cíle.

V této práci se zaměříme na jednu z klíčových částí bioinformatické analýzy, což je zpracování a analýza dat z formátu FASTQ. Formát FASTQ je běžně používaným formátem pro uchování sekvenačních dat z moderních technik NGS. Porozumění tomuto formátu a schopnost provádět bioinformatickou analýzu dat v něm obsažených je klíčové pro většinu genetických a genomických studií prováděných v dnešní době.

NGS neboli sekvenování nové generace, představuje revoluční posun v oblasti biologického výzkumu, který umožnil rychlé a cenově dostupné sekvenování genetických informací v rozsahu, který byl dříve nepředstavitelný. Tato inovativní technologie se stala pilířem moderní genomiky a transkriptomiky a otevřela nové možnosti pro studium genetické diverzity, struktury genomů, exprese genů a jejich vztahů k fenotypickým charakteristikám. V průběhu posledních dvou desetiletí NGS změnilo paradigma v oblasti biologického výzkumu. Kde dříve bylo sekvenování genomu nákladné a časově náročné, dnes je možné získat miliony sekvencí DNA nebo RNA za několik dní nebo dokonce hodin, a to za zlomek původních nákladů. Tato technologie umožnila rychlý pokrok v oblastech jako je genetická diagnostika, lékařská genomika, evoluční biologie, zemědělský výzkum a mnoho dalších. Principy NGS spočívají v masivní paralelní sekvenaci velkého množství krátkých fragmentů DNA nebo RNA, které jsou následně mapovány na referenční sekvence nebo sestavovány do kompletních genomů nebo transkriptomů. Tato technologie umožňuje vědcům zkoumat genomické sekvence různých organismů, identifikovat genetické varianty, studovat genovou expresi a zjišťovat epigenetické modifikace s přesností a rozlišením, které bylo dříve nedosažitelné.

Tato práce bude poskytovat přehled o základních principech bioinformatické analýzy dat ve formátu FASTQ, včetně procesů jako je kontrola kvality, trimming a filtrace dat, mapování na referenční sekvence a analýza exprese genů. Představíme si také některé běžně používané nástroje a postupy v bioinformatické analýze a objasníme si význam a aplikace těchto technik v biologickém výzkumu. Cílem této práce je poskytnout čtenářům ucelený přehled o bioinformatické analýze dat ve formátu FASTQ a podnítit zájem o tuto stále se rozvíjející disciplínu, která hraje klíčovou roli v moderním biologickém výzkumu.

2. Současný stav řešené problematiky

2.1. Bioinformatika jako vědní disciplína

Na začátku 50. let 20. století nebylo mnoho informací o deoxyribonukleové kyselině a nebyla příliš uznávána jako molekula nesoucí genetickou informaci. Trojice vědců Avery, MacLeod a McCarty v roce 1944 ukázali, že je možné přenést virulenci z kmenové bakterie na nevirulentní kmen, ale jejich výsledek nebyl okamžitě přijat **(Avery et al., 1944)**. Tehdy totiž byly proteiny považovány za nositele genetické informace **(Griffiths et al., 2000)**. Role DNA byla potvrzena až v roce 1952.

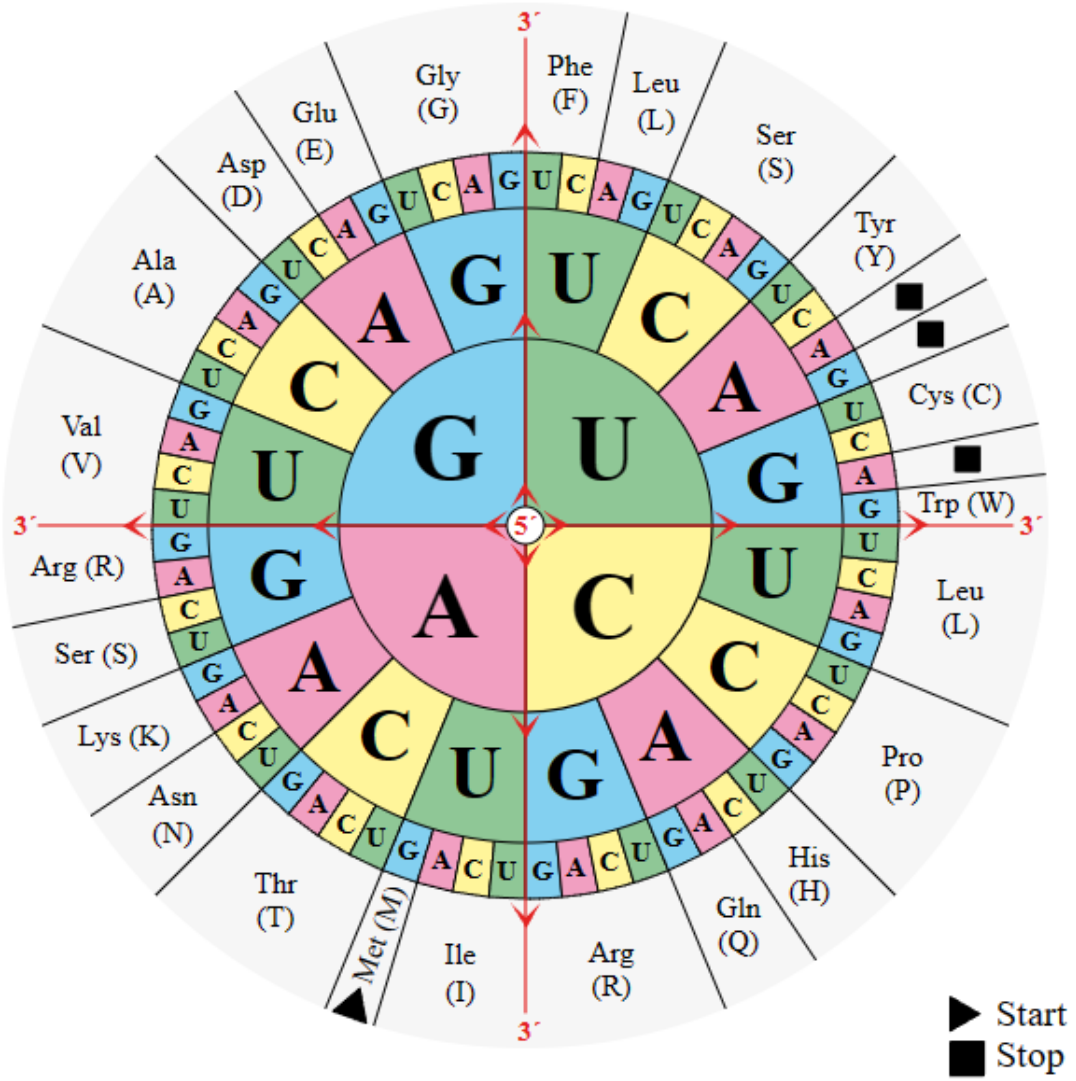
Uspořádání molekuly DNA však tehdy ještě nebylo známo. Vědělo se pouze, že nukleotidy v DNA byly v ekvimolárních poměrech, tedy že DNA obsahuje stejné množství thyminu jako adeninu a cytosinu jako guaninu **(Tamm et al., *n.d.*)**. Až v roce 1953 objevili James Watson a Francis Crick dvouvláknovou strukturu DNA **(Watson & Crick, 1953)**. Poté dalších 13 let trvalo rozluštit genetický kód a dalších 25 let, než se objevily první metody sekvenování DNA **(Nirenberg & Leder, 1964) (Sanger et al., 1977) (Maxam & Gilbert, 1977)**. V důsledku toho se analýza DNA zpozdila téměř o 20 let za analýzou proteinů, které byly tehdy lépe pochopeny, včetně jejich chemických vlastností.

Margaret Oakley Dayhoff je považována za skutečnou zakladatelku oboru bioinformatika **(Chen, 2004).** Prosazovala aplikaci výpočetních metod v oblasti biochemie a sestavila v roce 1965 první proteinovou databázi, tzv. Atlas proteinových sekvencí, který byl v knižní podobě. V tomto Atlase se už používaly jednopísmenné značky aminokyselin, jaké se používají dodnes **(IUPAC-IUB Comm. on Biochem. Nomenclature, 1968)**. První vydání Atlasu obsahovalo 65 proteinových sekvencí **(Dayhoff et al., 1965)**.

V roce 1970 byl vyvinut první algoritmus pro porovnávání dvojic proteinových sekvencí Needlemanem a Wunschem **(Needleman & Wunsch, 1970)**. Algoritmy pro porovnání více sekvencí, tzv. MSA z anglického Multiple Sequence Alignment, se však objevily až na počátku 80. let.

Tzv. centrální dogma molekulární biologie říká, že sekvence RNA transkribovaná z DNA určuje aminokyselinovou sekvenci proteinů. S tímto tvrzením přišel poprvé Francis Crick. Naopak sekvence aminokyselin určuje trojrozměrnou strukturu proteinu. Aby ale bylo možné předpovědět jakoukoli strukturu proteinu, muselo by být možno přečíst a přeložit DNA organismu, který obsahuje daný protein. V roce 1968 tak bylo rozluštěno

všech 64 kodonů genetického kódu **(Crick, 1968)**. DNA tak byla od teď čitelná a bylo potřeba najít způsob získávání DNA sekvencí.



Obr. 1: Genetický kód. Převzato z: (Genetický kód, n.d.*)*

2.2. Historie sekvenování

První metodou DNA sekvenování byla Maxam-Gilbertova metoda sekvenování z roku 1976, která byla založena na chemické modifikaci DNA a následném rozštěpení řetězce v místech modifikovaných nukleotidů **(Maxam & Gilbert, 1977)**. Její složitost a využívání radioaktivity a chemikálií však z velké části omezovala její praktické využití.

V roce 1977 byla vyvinuta tzv. metoda plus minus týmem Fredericka Sangera. Tato metoda spoléhala na syntézu s primerem pomocí DNA polymerázy. První získaný DNA

genom bakteriofága byl sekvenován pomocí této metody. Drobnými úpravami této metody jsme později získali Sangerovu metodu sekvenování, která je známá jako enzymatická metoda a využívá specifických vlastností DNA polymerázy při syntéze nového řetězce **(Sanger et al., 1977)**.

V roce 1987 byl na trh uveden první automatický sekvenátor AB370, který využíval kapilární elektroforézu k separaci jednotlivých úseků DNA. Umožňoval osekvenovat 500 kilobazí za den s délkou čteného fragmentu, tzv. readem, okolo 600 bazí. Pro srovnání současný model AB3730xl dokáže za den vyhodnotit až 2,88 megabazí a jeden read obsahuje 900 bazí **(Liu et al., 2012)**.

V roce 1995 byla provedena první kompletní sekvenace genomu žijícího organismu institutem TIGR **(Fleischmann et al., 1995)**. Skutečným zlomem však bylo zveřejnění lidského genomu na začátku 21. století. Projekt lidského genomu byl zahájen v roce 1991 Národním ústavem pro zdraví v USA a sekvenování probíhalo dlouhých 13 let **(Venter et al., 2001)**. Zároveň tento projekt poukázal na nedostatky Sangerova sekvenování a potřebu nových technologií. Pro srovnání, dnes by osekvenování celého lidského genomu trvalo zhruba týden a náklady by byly minimální.

První světová databáze nukleotidových sekvencí EMBL Nucleotide Sequence Data Library byla na webu zpřístupněna v roce 1993. Tato databáze zahrnovala několik dalších databází, jako je například databáze SWISS-PROT **(Rice et al., 1993)**. O rok později byla zpřístupněna dnes asi nejznámější databáze NCBI, včetně nástroje BLAST a databáze GenBank **(Benson et al., 1993)**. Poté následovalo zřízení několika dalších databází, které jsou používány dodnes. Jedná se o databáze Genomes, PubMed a Human Genome.

2.3. Sekvenování nové generace

Sekvenování nové generace, zkráceně NGS, označuje moderní metody sekvenování. Využívá se bioinformatických metod ke zpracování velkého množství sekvenačních dat, až miliony sekvencí současně, a jejich porovnání s referenčním genomem **(Behjati & Tarpey, 2013)**.

Základní myšlenkou sekvenování nové generace je tzv. masivně paralelní sekvenování, které funguje tak, že během jednoho okamžiku experimentu je osekvenováno velké množství molekul. Vstupní DNA se fragmentuje na úseky o délce typicky několika desítek až stovek párů bazí a v případě potřeby se amplifikuje pomocí PCR. Tyto fragmenty jsou poté čteny a zpracovávány. Na rozdíl od Sangerova sekvenování, při

kterém je čteno maximálně 96 molekul DNA, se u NGS čte velké množství fragmentů, a to v řádech milionů **(Koubková et al., 2014)**.

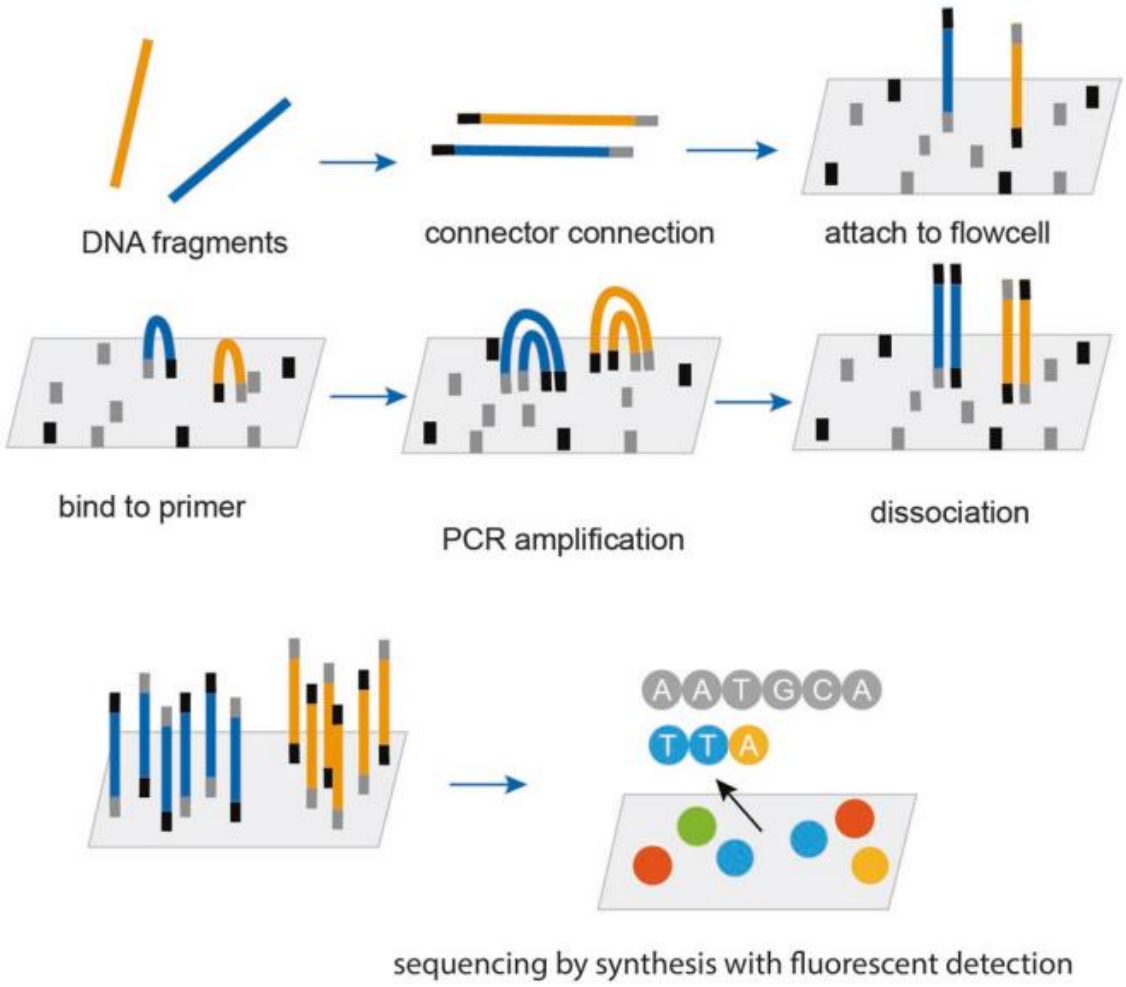
Všechny dostupné technologie pracují na podobném principu a mají společné kroky, konkrétně přípravu templátu neboli vytvoření knihovny, dále samotné sekvenování a analýzu dat. Rozdíly mezi technologiemi určuje unikátní kombinace jednotlivých kroků. Důležité jsou rozdíly ve výstupních datech. Všichni výrobci odhadují přesnost a kvalitu čtení, ale neexistuje tvrzení, které by dokazovalo, že kvalita čtení dvou platforem je ekvivalentní **(Metzker, 2010)**.

2.3.1. Roche 454

První technologií byla technologie pyrosekvenování '454', která umožnila sekvenování tisíců až milionů molekul DNA v jednom běhu **(Margulies et al., 2005)**. DNA se štěpí na fragmenty o délce 300 až 800 bází, na jejich koncích jsou odstraněny nespárované báze. DNA se následně rozdělí do jednořetězové formy a na každý konec jsou přidány adaptérové sekvence. Pomocí těchto adaptérů jsou fragmenty imobilizovány na kuličkách tak, aby každá kulička nesla jeden fragment. Kuličky jsou následně zapouzdřeny do emulze, což umožňuje následnou emulzní PCR **(Zhou et al., 2010)**. Tato technologie byla udržována společností Roche 454 až do jejího vyřazení v roce 2013.

2.3.2. Illumina

Další a dnes velmi rozšířenou společností na trhu se sekvenátory je Illumina, dříve Solexa. Ta používá metodu sekvenace syntézou společně s tzv. 'bridge' amplifikací. Prvním krokem této metody je fragmentování DNA na úseky o velikosti 200 až 800 bází. Na konce vzniklých fragmentů se následně navážou adaptéry tak, aby byl na každém konci fragmentu jiný adaptér. Fragmenty jsou následně denaturovány a připojeny k jednořetězovým nukleotidům, které jsou navázány na povrchu reakční komůrky, tzv. 'flow cell'. Každý fragment je tak jedním koncem imobilizován k povrchu reakční komůrky. Nukleotidy, na které jsou fragmenty vázány, následně slouží jako primery pro PCR amplifikaci. Po této amplifikaci jsou vzniklé dvouřetězcové molekuly DNA denaturovány na jednořetězcové, původní templát se odmyje a zůstává pouze nově nasyntetizované vlákno DNA, které je kovalentně navázáno k povrchu reakční komůrky. Volným koncem hybridizuje k primerům na povrchu, ohne se a dojde k přemostění, proto 'bridge' amplifikace. V dalším PCR cyklu je vytvořen dvouvláknový most a proces se opakuje.



Obr. 2: Princip Illumina sekvenování. Převzato z: (Pan & Tang, 2021)

2.3.2.1. Sekvenátory

Illumina Genome Analyzer, tedy první sekvenátor společnosti Illumina, generoval 1 gigabázi dat za běh, který trval 2 až 3 dny a byl schopen přečíst sekvence dlouhé 35 bází. V současnosti nabízí společnost Illumina několik různých přístrojů, mezi nimi MiSeq, MiniSeq a NextSeq. MiSeq je sekvenátor určen pro sekvenování menších genomů. Pracuje průměrně 4 až 55 hodin a generuje až 15 gigabazí za běh. Jeho v podstatě menší dvojče MiniSeq generuje přesně polovinu, tedy 7,5 gigabáze za běh. NextSeq je momentálně nejvýznamnějším sekvenátorem menších laboratoří. Doba jednoho sekvenačního běhu je 12 až 30 hodin a maximální výstup takového běhu je až 120 gigabází dat **(Illumina Sequencing Platforms, 2023)**.

Illumina sekvenátory jsou poněkud limitovány krátkou délkou sekvenačních čtení. Je to dáno zvýšenou nebo sníženou účinností inkorporace nukleotidu a selháním při odstranění

nebo přidání terminační skupiny. To může způsobit nekompletní prodloužení vlákna, a proto nejčastější chybou je substituce nukleotidu. Tato chyba vzrůstá s délkou čtení. Obecně jsou sekvenátory Illumina schopné produkovat více dat za méně peněz i času oproti Sangerovu sekvenování, ale za cenu vyšší chybovosti, která může vést k falešné pozitivitě při identifikaci sekvenačních variací **(Metzker, 2010)**. I přes chybovost se tyto sekvenátory řadí momentálně ke špičce díky vysoké výkonnosti a jsou využívány k většině celogenomových aplikací **(Shokralla et al., 2012)**.

Key specifications	 iSeq 100 System	 MiniSeq System	 MiSeq System	 NextSeq 550 System
Max output per flow cell	1.2 Gb[a]	7.5 Gb[b]	15 Gb[c]	120 Gb[b]
Run time (range)[e]	~9.5–19 hr	~5–24 hr	~5–56 hr	~11–29 hr
Max reads per run (single reads)	4M[a]	25M[b]	25M[c]	400M[b]
Max read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp

Obr. 3: Sekvenátory Illumina. Převzato z: (Illumina Sequencing Platforms, 2023)

2.3.2.2. MGI sekvenování

V této práci se zabývám platformou MGI a tak, i přesto že není příliš známá, by bylo dobré si ji trochu představit. Společnost MGI je jedním z poskytovatelů sekvenování s vysokou propustností. Jejich sekvenační technologie, známá jako MGISeq, patří mezi platformy NGS a je známá právě svou vysokou propustností, rychlostí a efektivitou. Je vhodná především pro sekvenace většího měřítka.

Platforma MGISeq využívá technologie DNA nanokuliček (DNB). Do těchto kuliček je amplifikována DNA a následuje sekvenace syntézou. Nukleotidy se připojují jeden po druhém a inkorporované nukleotidy jsou detekovány pomocí fluorescenčního značení kamerou. Společnost MGI zároveň vyvinula inovativní variantu sekvenace syntézou, při které se nepoužívá fluorescenčního značení nukleotidů barvami, ale využívá se fluorescenční značení protilátek, které jsou levnější na výrobu a dávají nám přesnější

údaje o sekvenování. Tato metoda se nazývá CoolMPS **(MGI Sequencers, *n.d.*) (How does MGI sequencing technology work?, *n.d.*).**

2.3.3. SOLiD

V roce 2007 byla představena platforma SOLiD, která je založena na sekvenování ligací. Příprava DNA knihovny probíhá pomocí emPCR a k fragmentům DNA jsou připojeny adaptéry komplementární k adaptérům imobilizovaným na povrchu magnetických kuliček. Tyto kuličky jsou po amplifikaci kovalentně navázány na sklíčko se speciálním povrchem. Tato metoda využívá osm nukleotidů dlouhé sondy, z nichž každá má známou sekvenci prvních dvou bází a je značena jednou z fluorescenčních barev, z nichž každá představuje jinou dinukleotidovou sekvenci. Tato technologie zaručuje, že každý nukleotid bude přečten dvakrát, což zvyšuje přesnost určení pořadí nukleotidů v sekvenci. Mezi všemi NGS technologiemi má tedy SOLiD nejmenší chybovost a zároveň platí, že jeho nejčastějším typem chyby je substituce nukleotidu **(Mardis, 2008)**.

2.3.4. Ion Torrent

Ion Torrent neboli iontové polovodičové sekvenování je založené na detekci vodíkových protonů uvolněných v průběhu syntézy nového řetězce. Jedná se o proces probíhající na polovodičovém čipu, který je hustě pokryt mikrojamkami, pod nimiž je umístěna citlivá vrstva, která detekuje změnu pH při uvolnění vodíkového iontu. Příprava knihovny opět zahrnuje emPCR. Kuličky se umisťují na čip, který je postupně zaplavován nukleotidy a probíhá syntéza DNA. Detektor zaznamená signál pouze v případě, že se jedná o komplementární nukleotid. Tato technologie je velice rychlá a levná **(Liu et al., 2012)**.

2.3.5. SMRT

Metoda SMRT neboli Single Molecule Real-Time Sequencing přinesla výraznou změnu ve vývoji nových sekvenačních metod. Bývá označována jako třetí generace sekvenování a liší se tím, že nevyužívá žádného amplifikačního kroku před vlastní sekvenací. To krátí dobu přípravy DNA, redukuje cenu a chybovost, umožňuje větší flexibilitu a přesnou kvantifikaci molekul, protože signál je zaznamenáván v reálném čase **(Xuan et al., 2013)**.

SMRT metoda používá nanostrukturu zvanou Zero Mode Waveguide, což je destička s 10 tisíci jamkami o průměru 10 nm. Na dně každé jamky je ukotvená DNA polymeráza, která syntetizuje komplementární vlákno. Fluorescenční značka vydá záblesk právě

tehdy, když dojde k inkorporaci nukleotidu. V současnosti tato metoda poskytuje největší délku čtení a nejvyšší přesnost.

2.3.6. Nanopore Sequencing

Tato metoda využívá biologických vlastností nanopórů, které jsou součástí proteinových kanálků, protéká jimi konstantní proud a dovolují výměnu iontů. Jednořetězcová molekula DNA prochází nanopórem, kde dojde k detekci nukleotidů. Pro každý nukleotid je předem určen proud. Tato technologie je velmi jednoduchá ve srovnání s ostatními metodami **(Oxford Nanopore Technologies, *n.d.*)**.

2.3.7. Aplikace NGS technologií

Díky produkci obrovského množství sekvenačních dat v krátké době jsou NGS technologie užitečným nástrojem pro řadu aplikací, například de novo sekvenování, objevování nových mutací nebo analýzu transkriptomu a DNA methylovaných oblastí. NGS technologie se uplatňují v molekulární diagnostice dědičných a infekčních onemocnění, nádorů a prenatální diagnostice **(Guan et al., 2012)**.

Nádorové onemocnění způsobují nahromaděné mutace v genetickém materiálu. Tyto mutace mohou být germinální, tedy předané od rodičů, nebo somatické, tedy získané v průběhu života. Studium profilů mutací jednotlivých nádorů významně pomáhá porozumění mechanismu kancerogeneze. V posledních letech bylo provedeno velké množství studií, díky kterým byly objeveny nové geny asociované s mnoha druhy nádorových onemocnění **(Shyr & Liu, 2013)**. Navíc jsou NGS technologie velmi slibným nástrojem pro diagnózu germinálních mutací.

2.3.7.1. Celogenomové sekvenování

Pojem celogenomové sekvenování se skládá z resekvenování a de novo sekvenování. Resekvenování je proces, při kterém se sekvenují ty genomy, pro které je k dispozici referenční sekvence, na kterou se nová sekvence zpětně mapuje. De novo sekvenování je kompletní sekvenování neznámých genomů, stavíme tedy sekvenci genomu úplně od začátku. Výhoda celogenomového sekvenování je možnost osekvenování celé chromozomální DNA a poskytnutí úplného mutačního profilu daného genomu. Tento typ sekvenování se využívá nejčastěji pro identifikaci nových a vzácně se vyskytujících mutací **(Guan et al., 2012)**.

2.3.7.2. Cílené sekvenování

Tzv. 'targeted sequencing' je technika umožňující sekvenaci pouze vybraných genů nebo částí genomu. To může výrazně ušetřit čas i potřebné finanční prostředky. Tato technika je typicky využívána při sekvenování velkého počtu vzorků při screeningu. Zároveň je zde možnost záchytu variant, které nezachytí klasické Sangerovo sekvenování, protože jejich identifikace je příliš drahá **(Chang & Li, 2013)**.

S cíleným sekvenováním souvisí také exomové sekvenování a sekvenování transkriptomu. U exomového sekvenování jsou sekvenovány pouze kódující oblasti neboli exomy. Velikost lidského exomu je rovna velikosti asi 1 % lidského genomu, jeho sekvenování je tedy snazší a levnější. Zároveň umožňuje vyšší pokrytí neboli coverage. Transkriptomem rozumíme všechny molekuly RNA, tedy mRNA, rRNA, tRNA a další nekódující RNA molekuly. Jeho analýza je velmi důležitá pro charakterizaci nádorů **(Jones et al., 2009)**.

2.4. Oblasti bioinformatického výzkumu

Bioinformatický výzkum zahrnuje širokou škálu disciplín a aplikací, které využívají principy bioinformatiky k analýze biologických dat a řešení biologických otázek. Jednotlivé oblasti bioinformatického výzkumu se neustále vyvíjejí a prolínají, což vytváří nové příležitosti pro objevování biologických znalostí a inovace například v oblasti lékařství.

2.4.1. Anotace genomu

Anotace genomu je proces, který identifikuje funkční elementy v rámci sekvence genomu a tím jim dává význam. Je to důležité, protože sekvenování DNA nám poskytuje sekvence, o jejichž funkci nemáme zprvu žádné informace. V posledních třech desetiletích se anotace genomu vyvíjela od výpočetní anotace dlouhých genů kódujících proteiny na jednotlivých genomech, přes experimentální označování krátkých regulačních elementů na malém vzorku z nich, až po populační označování jednotlivých nukleotidů na tisících individuálních genomech. Dalo by se říct, že se vyvinula od jednoho genu na druh, po mnoho genomů na druh. Tento pokrok v anotaci genomů, který zahrnuje větší rozlišení a rozšíření v populačních datových sadách, jako jsou genotypy a fenotypy, nám poskytuje detailnější pohled na biologii druhů, populací i jednotlivců **(Abril & Castellano, 2019)**.

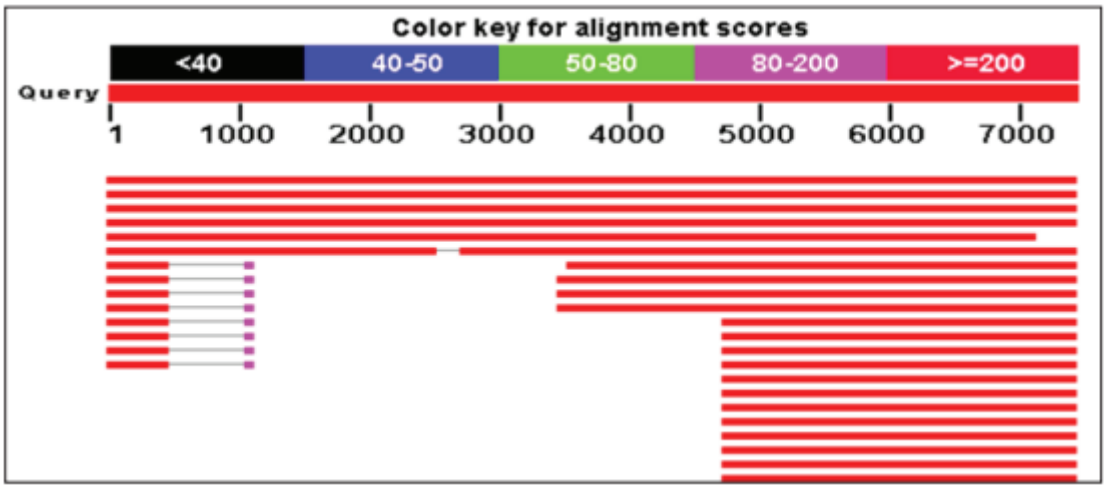
2.4.1.1. Identifikace genů

Identifikace genů se rychle vyvíjí díky pokroku týkajícího se molekulárně biologických metod a díky zvýšené dostupnosti genetických a funkčních genomických dat. Bioinformatika pomáhá identifikovat geny v dlouhých DNA sekvencích. Tato metoda umožňuje lokalizovat geny prostřednictvím analýzy sekvencí s využitím počítačových prostředků.

Důležitým aspektem bioinformatiky je predikce genů. Ta spočívá v identifikaci protein kódujících genů. Identifikace genu je zásadní, protože pomáhá vědcům rozlišit kódující a nekódující oblasti genomu a porozumět funkci genů. Takové poznání nám umožňuje vést výzkum zaměřený na detekci, léčbu a prevenci genetických chorob. Geny jsou obvykle identifikovány dvěma způsoby, a to hledáním podobných sekvencí a ab-initio predikcí.

Hledání na základě podobnosti je způsob identifikace genu, který využívá hledání podobných sekvencí v databázích. Tato metoda předpokládá, že exony jsou evolučně konzervovanější než introny. Nástrojem pro hledání na základě podobnosti je program BLAST.

BLAST je program pro prohledávání sekvencí, který hledá pouze vysoce signifikantní oblasti shody. Je k dispozici ve formě webové aplikace na stránkách https://blast.ncbi.nlm.nih.gov. Je možné jej použít jak pro prohledávání nukleotidových sekvencí, tak i proteinových sekvencí.



Obr. 4: Grafický výstup programu BLAST. Převzato z: (Syngai et al., 2013)

Ab-initio predikce využívá pouze nukleotidovou sekvenci k predikci struktury proteinu. Tento postup je nezbytný pro kompletní řešení problému predikce proteinové struktury. V současné době je ale přesnost modelování ab-initio relativně nízká a úspěšnost je omezena na malé proteiny **(Lee et al., 2017)**.

2.4.1.2. Funkční anotace

Identifikovaným genům je třeba přiřadit jejich biologickou funkci. Tato anotace spočívá v použití nástroje BLAST k nalezení podobných sekvencí a následné anotaci genů nebo proteinů. Analýza funkční anotace zahrnuje anotaci genů pomocí GO termínů a informací o drahách.

GO je zkratka pro genovou ontologii. Ontologie reprezentuje soubor znalostí v rámci daného oboru. Genová ontologie má za cíl sjednotit popis genů a genových produktů napříč organismy, protože podobné geny mají často v různých organismech stejnou funkci. Domény, které genová ontologie pokrývá, jsou molekulární funkce, buněčná složka a biologický proces. Termíny molekulární funkce popisují aktivity probíhající na molekulární úrovni, jako například katalýza. Buněčná složka popisuje součást buňky nebo jejího okolí. Biologický proces popisuje sled událostí na molekulární úrovni, který se vztahuje k živým jednotkám **(Gene Ontology Overview, *n.d.*)**.

2.4.1.3. Anotace nekódujících sekvencí

Kromě genů obsahuje genom také mnoho nekódujících sekvencí, které mohou mít regulační funkce, vliv na strukturu chromozomů nebo další biologické účinky. Tyto sekvence mohou být anotovány pro identifikaci promotorů, enhancerů, siRNA, mikroRNA a dalších.

Opakující se elementy, jako jsou transpozony, jsou identifikovány a anotovány, protože mohou hrát roli v genomové stabilitě, evoluci a regulaci genové exprese. Transpozony vznikají procesem transpozice, což je přesun sekvence na jiné místo genomu. Mají schopnost pohybu. Retrotranspozony se mohou kopírovat, DNA transpozony se nejprve vystřihnou z původního místa a pak se přesouvají. Transpozony nemají žádnou důležitou funkci v buňce, ale jejich mobilita může být dobrá pro plasticitu genomu **(Mills et al., 2007)**.

Strukturální anotace se zaměřuje na identifikaci a popis struktur genomu, jako jsou introny, exony, promotorové oblasti, úseky intergenového prostoru a další.

Exony jsou úseky DNA nebo RNA, které obsahují kódující informaci pro vytvoření bílkoviny. Tvoří části genů, které jsou transkribovány do pre-mRNA během procesu transkripce. Tyto části pre-mRNA jsou poté spojeny do konečné mRNA.

Introny jsou nekódující oblasti genu, které se nacházejí mezi exony. Během procesu transkripce jsou přepisovány do pre-mRNA, ale poté jsou vystřiženy v procesu zvaném 'splicing', který vede společně s dalšími k vytvoření mRNA. Introny jsou důležité pro regulaci genové exprese **(Kidd et al., 2008)**.

2.4.2. Sekvenční analýza

S vývojem programu BLAST se sekvenční analýza stala velmi populární a má mnoho oblastí využití. Sekvenční analýza se dá využít například pro anotaci nových oblastí, k nalezení konzervovaných a regulačních oblastí a k predikci fyzikálně chemických vlastností sekvencí. Dále analyzujeme sekvence, abychom zjistili podobnosti mezi nimi pomocí nástrojů pro zarovnání sekvencí. Známými nástroji pro zarovnání sekvencí jsou programy BLAST a FASTA Clustal.

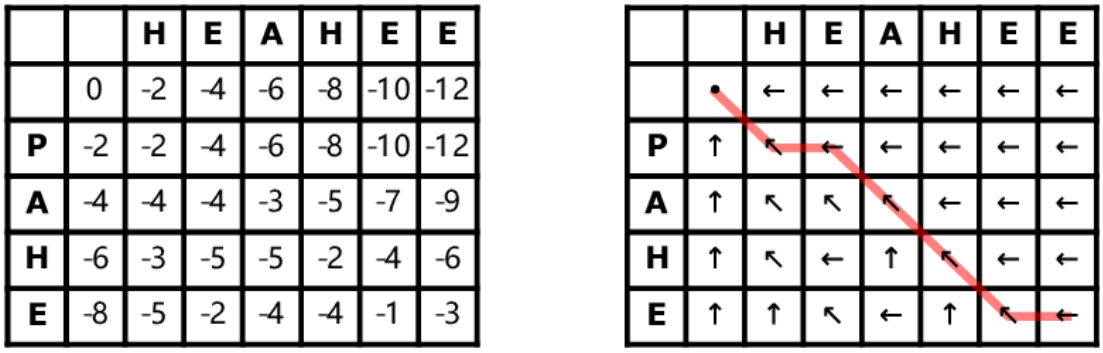
Program BLAST využívá algoritmus pro lokální zarovnání sekvencí. Porovnává nukleotidovou sekvenci se známými sekvencemi v databázích vytvořením tzv. 'seedů', což jsou krátké úseky vložené sekvence. Tyto úseky se následně porovnávají se známými sekvencemi a BLAST vrací procento podobnosti s jednotlivými sekvencemi **(BLAST: Basic Local Alignment Search Tool, *n.d.*) (Altschul et al., 1990)**.

Program FASTA také přijímá nukleotidovou nebo proteinovou sekvenci a prohledává databáze sekvencí pomocí lokálního přiřazení. Nejprve porovnává krátké úseky sekvence, označuje potenciální shody a následně využívá Smith-Waterman algoritmu **(FASTA, *n.d.*)**.

Smith-Waterman algoritmus je algoritmus pro lokální zarovnání a vychází z Needleman-Wunsch algoritmu pro globální zarovnání sekvencí. Principiálně jsou si velmi podobné, vstupem jsou vždy dvě zarovnávané sekvence, skórovací matice velikosti $((m+1) \times (n+1))$, kde m a n jsou délky sekvencí, a hodnota sankce za vložení mezery. Zásadním rozdílem mezi nimi však je fakt, že Smith-Waterman algoritmus nahrazuje všechny negativní skóre hodnotou 0.

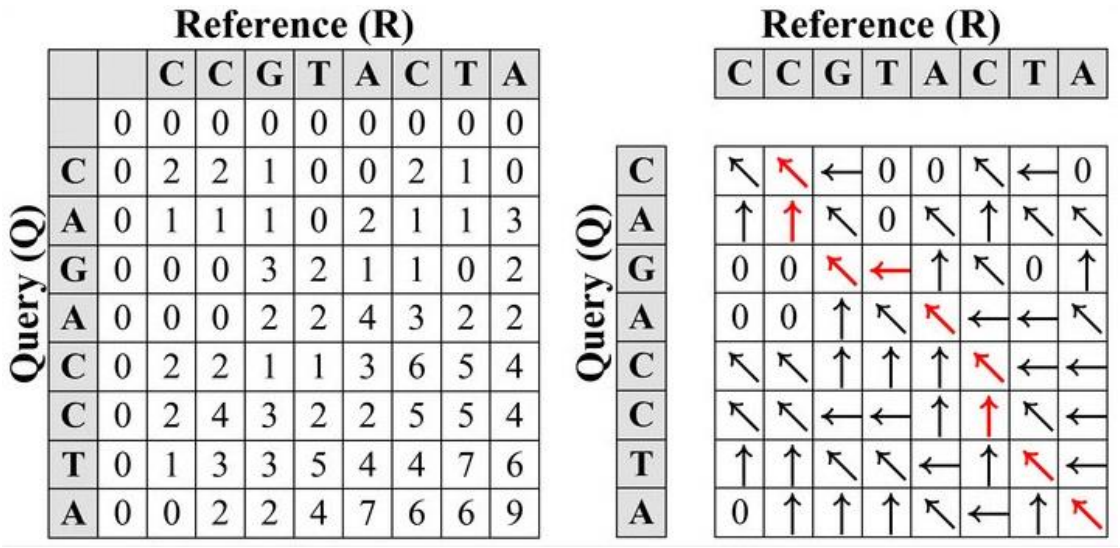
Postup algoritmů je tedy stejný, v prvním kroku zapíšeme do levého horního rohu hodnotu 0 a následně do každého dalšího pole zapíšeme maximum z hodnot skóre pole vlevo + sankce za vložení mezery, skóre pole nahoře + sankce za vložení mezery, a skóre

pole vlevo nahoře + hodnota za shodu či neshodu znaků podle skórovací matice. Následně jdeme zpětným chodem od nejvyššího skóre zpět k hodnotě 0. Tato zpětná cesta pak reprezentuje nejlepší zarovnání **(Smith-Waterman Algorithm, *n.d.*)**.



		H	E	A	H	E	E
	0	-2	-4	-6	-8	-10	-12
P	-2	-2	-4	-6	-8	-10	-12
A	-4	-4	-4	-3	-5	-7	-9
H	-6	-3	-5	-5	-2	-4	-6
E	-8	-5	-2	-4	-4	-1	-3

Obr. 5: Needleman-Wunsch algoritmus. Převzato z: (Nguyen et al., n.d.*)*



Reference (R)

Query (Q)		C	C	G	T	A	C	T	A
	0	0	0	0	0	0	0	0	0
C	0	2	2	1	0	0	2	1	0
A	0	1	1	1	0	2	1	1	3
G	0	0	0	3	2	1	1	0	2
A	0	0	0	2	2	4	3	2	2
C	0	2	2	1	1	3	6	5	4
C	0	2	4	3	2	2	5	5	4
T	0	1	3	3	5	4	4	7	6
A	0	0	2	2	4	7	6	6	9

Reference (R)

Query (Q)	C	C	G	T	A	C	T	A
C	↖	↖	←	0	0	↖	←	0
A	↑	↑	↖	0	↖	↑	↖	↖
G	0	0	↖	←	↑	↖	0	↑
A	0	0	↑	↖	↖	←	←	↖
C	↖	↖	↑	↑	↑	↖	←	←
C	↖	↖	←	←	↑	↑	↖	←
T	↑	↑	↖	↖	←	↑	↖	←
A	0	↑	↑	↑	↖	←	↑	↖

Obr. 6: Smith-Waterman algoritmus. Převzato z: (Liao et al., 2018)

2.4.3. Analýza úrovně genové exprese

Proces exprese genu lze rozdělit do dvou částí, a to transkripce a translace. Transkripce je přepis genetické informace z DNA do RNA. Translace je přepis sekvence RNA do sekvence aminokyselin proteinu. Tyto procesy se musejí kontrolovat a regulovat, aby bylo jasné, jaké proteiny a v jakém množství jsou přítomny v buňce.

Analýza genové exprese má za úkol srovnání úrovně exprese RNA u více genů, případně více vzorků zároveň a může být přínosná pro identifikaci fenotypových rozdílů a pro

cílenou expresi genu. Úkolem molekulárních biologů je měření úrovně genové exprese jistých genů, tedy měření množství transkribované mRNA. Metody využívané k tomuto měření jsou například reverzní transkripce do cDNA či SAGE. Úkolem bioinformatiků je následná úprava rozsáhlých souborů dat z tohoto měření **(Gene Expression, *n.d.*)**.

cDNA je DNA syntetizovaná z mRNA za působení enzymu reverzní transkriptázy. Na rozdíl od klasické DNA, cDNA obsahuje pouze kódující oblasti genu, tedy exony, protože je tvořena z již upravené mRNA po sestřihu **(CDNA (copy DNA), *n.d.*)**.

SAGE neboli sériová analýza exprese genů zahrnuje proces vytvoření fragmentů z cDNA, které jsou následně amplifikovány a sekvenovány technologií 'high-throughput' sekvenování. Výsledná analýza poskytuje jakýsi obraz transkriptomu včetně informace o původní sekvenci mRNA **(Reina, 2016)**.

2.4.4. Srovnávací genomika

Srovnávací neboli komparativní genomika má za cíl porovnávání kompletních genomových sekvencí a struktur napříč druhy. Využívá sekvencí genomů ke zjištění nových informací o biologickém druhu a zkoumá evoluční příbuznost jednotlivých druhů.

Základním porovnáním dvou genomů může být porovnání velikosti genomu, počtu genů a chromozomů. Zároveň platí, že velikost genomu neodpovídá celkové komplexitě organismu, a také že počet genů nesouvisí s velikostí genomu **(Comparative Genomics, *n.d.*)**.

2.4.5. Predikce struktury proteinů

Porozumění struktuře proteinu je důležité pro znalost jeho funkce, experimentální zjišťování struktury například pomocí rentgenové krystalografie je však poměrně drahé a nepraktické, proto se bioinformatici zabývají také předpovědí struktury proteinu.

Primární struktura proteinu je dána pořadím aminokyselin a lze ji jednoduše zjistit sekvenací mRNA. Sekundární a další struktury pak lze získat využitím predikčních programů. Tyto programy využívají k predikci například fyzikálně chemické vlastnosti jednotlivých aminokyselin nebo znalosti struktury homologního proteinu s podobnou funkcí. Používanými predikčními programy jsou například JPred či AlphaFold.

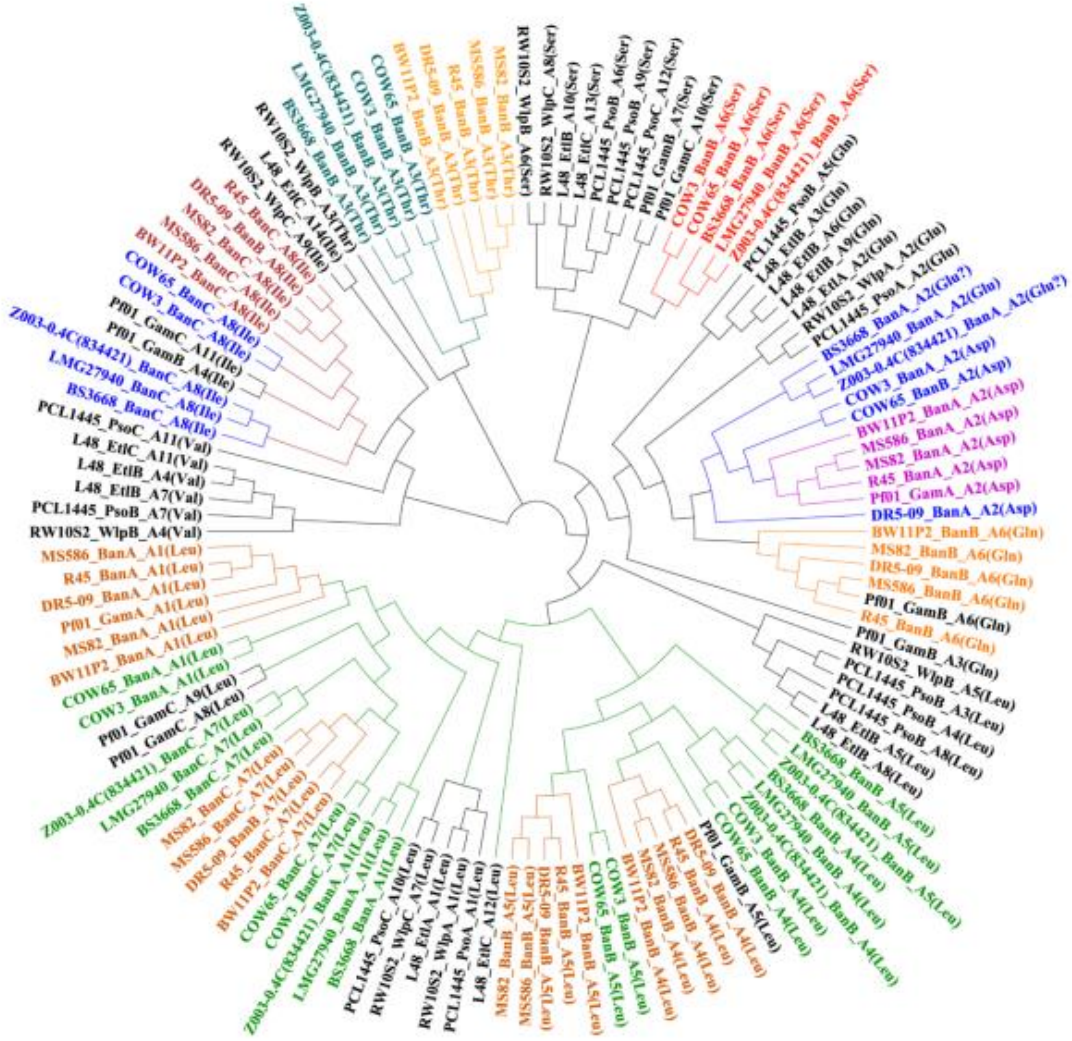
Obr. 7: Příklad výstupu z programu AlphaFold. Převzato z: (Probable disease resistance protein At1g58602, n.d.*)*

2.4.5.1. Fylogenetický strom

Fylogenetický strom se používá k vizuální reprezentaci vztahů mezi organismy a reprezentaci evolučního vývoje druhů. Typický fylogenetický strom je tzv. bifurkální, což znamená, že z jednoho uzlu stromu vycházejí dvě větve. Tyto větve reprezentují potomky a jejich společný uzel společného předka. Pokud je strom tzv. multifurkální, pak z jednoho uzlu vychází víc než dvě větve. To značí, že není možné určit posloupnost větvení. Kořenový uzel reprezentuje nejvzdálenějšího společného předka.

Existují dvě metody konstrukce fylogenetických stromů, a to metody vzdálenostní a znakové. Mezi vzdálenostní metody patří metody UPGMA, metoda minimální evoluce a

Neighbor Joining, mezi znakové metody pak patří metody 'Maximum Parsimony' a 'Maximum Likelihood' **(Phylogenetic Trees, *n.d.*)**.



Obr. 8: Příklad hvězdicovitého fylogenetického stromu vytvořeného metodou Neighbor Joining. Převzato z: (Omoboye et al., 2019)

2.5. Zpracování sekvenačních dat

Práce bioinformatika zahrnuje hlavně počítačové zpracování sekvenačních dat. Tyto data je třeba převést do čitelné podoby, jelikož ze sekvenátorů vycházejí například v podobě FASTQ souborů, které moc čitelné nejsou. Proto je třeba provést tzv. bioinformatickou analýzu, která zahrnuje zarovnání čtení k referenčnímu genomu i s kontrolou případných změn, kontrolu správnosti získaných dat a anotaci získaných variant.

První řádek FASTQ souboru začíná znakem zavináče a obsahuje záhlaví, které poskytuje dodatečné informace o záznamu, jako je například identifikátor sekvence. Po zalomení řádku následuje sekvence DNA nebo RNA. Po dalším zalomení následuje oddělovač v podobě znaku '+' a na posledním řádku se nachází tzv. Phred quality scores přiřazené každé bázi sekvence. To vyjadřuje, s jakou pravděpodobností je daná báze čtena špatně. Skóre je počítáno jako $Q = -10\ log_{10}\ P$, kde Q je Phred skóre a P je pravděpodobnost chybného čtení. Tyto hodnoty se kódují ASCII znaky.

```
@M04042:277:000000000-DD8L7:1:1101:13956:1769
1:N:0:TCCGGAGA+CCTATCCT
CCTGATATATGTTCTCTAGGCCTTTTAGAAAACTTGGTGT
+
1>A11D3DFFFFGGGDG3111AFGHGB0133001D1110BE0
```

Obr. 9: Příklad FASTQ souboru. Převzato z Laboratoří AGEL Nový Jičín.

2.5.1. Zarovnání čtení

Základní postup je vždy stejný. Nejprve je třeba najít tu část referenční sekvence, která odpovídá sekvenci čteného fragmentu, jinak řečeno mapujeme všechny ready na referenční genom. Dále je potřeba nalézt všechny odlišnosti nasekvenované DNA od referenčního genomu. S pomocí statistických výpočtů se určuje pravděpodobnost, s jakou se jedná o sekvenační chybu či s jakou pravděpodobností se jedná o skutečný polymorfismus.

Při mapování se používá technika indexování. Vytvoří se index nad sekvencí genomu, který je následně využit k prohledání genomu. Pro každý genom stačí vytvořit jen jeden index, který bude využit i pro budoucí mapování. Tvorba indexu se provádí nejčastěji Burrows-Wheelerovou transformací **(Li & Durbin, 2009)**.

2.5.1.1. Referenční genom

Na lidském referenčním genomu se pracuje již od roku 1990, ale i přesto jeho kompletní sekvence stále není známa. Kvůli nedostatečné délce sekvenačních readů je největší problém v určení správného počtu kopií ve vysoce repetitivních oblastech, například v okolí centromer. Další problémy jsou u větších skupin sekvenčně podobných genů.

Díky pokroku v technologiích se však délka readů pořád zvětšuje, a tedy i referenční genom se postupem času vyvíjí. První verze referenčního genomu obsahovala 150 000

neznámých regionů. Devatenáctá verze už jich měla v sobě pouze 357 a verze hg38 jako první obsahuje odhady počtu centromerických repetic **(Schneider et al., 2017)**.

2.5.2. Kontrola indexování

Ke kontrole indexování souboru se používá Picard tools. To je sada nástrojů vyvinutých společností Broad Institute pro manipulaci se sekvenačními daty. Tyto nástroje poskytují širokou škálu funkcí, které umožňují bioinformatickým analytikům spravovat, zpracovávat a analyzovat sekvenční data. Mezi běžné úlohy, které Picard tools umožňují, patří manipulace s formáty souborů jako je BAM, kontrola a oprava poruch v zarovnáních, získávání statistik z dat sekvenování a mnoho dalších.

Jedním z důležitých aspektů Picard tools je jeho schopnost zachovat integritu dat a metadata během zpracování, což je klíčové pro zachovámí spolehlivosti v bioinformatických analýzách. Tyto nástroje jsou běžně používány v bioinformatických laboratořích a výzkumných institucích po celém světě jako součást sekvenačních pracovních postupů.

2.5.3. Odstranění duplicitních čtení

Dále je potřeba odstranit PCR duplikáty a duplikáty způsobené optickým senzorem stroje. K odstranění těchto duplikátů je používán nástroj MarkDuplicates z balíčku nástrojů Picard tools. Ten slouží k identifikaci a označení duplikátů v souborech formátu BAM. MarkDuplicates identifikuje duplikáty na základě jejich umístění v genomu a poté je označí v metadatech souboru BAM. Tato označení mohou být důležitá pro další analýzy, kde se duplikáty mohou chovat nežádoucím způsobem. Výstupem je opět soubor ve formátu BAM. Tento soubor se musí indexovat, aby se následně dal použít ve vizualizačním prostředí IGV. K indexaci se používá knihovna Samtools, což je soubor nástrojů pro práci s BAM a SAM formáty. Tato knihovna nástrojů se používá z příkazového řádku.

IGV je vizualizační nástroj, který umožňuje prozkoumávat genetické varianty a další biologické informace v kontextu genomických dat. IGV poskytuje interaktivní uživatelské rozhraní, které umožňuje uživatelům pohodlně přibližovat a oddalovat, navigovat po genomu, vyhledávat konkrétní geny nebo oblasti a zobrazovat podrobné informace o genetických variantách a dalších biologických rysů. Díky své schopnosti zobrazovat data v kontextu genomických vlastností se IGV často používá v bioinformatickém výzkumu, genetice a klinické genetice k prohlížení, analýze a

interpretaci genomických dat. Je to mocný nástroj pro identifikaci potenciálně významných genetických variant spojených s různými fenotypy a nemocemi.

Obr. 10: Příklad IGV vizualizace. Převzato z DOI:10.1007/978-3-319-17157-9

2.5.4. Změny proti referenčnímu genomu

Pomocí knihoven Samtools a VarScan se provádí tzv. variant calling. Ten zjišťuje informaci o tom, zda se v analyzovaném vzorku nachází změny jako jsou SNP, inzerce nebo delece. Výstupem je soubor ve formátu VCF neboli Variant Call Format.

SNP neboli jednonukleotidový polymorfismus je sekvenční variace DNA běžně se vyskytující v populaci, ve které se jeden nukleotid v genomu liší mezi zástupci jednoho biologického druhu. Tyto polymorfismy se mohou nacházet v kódujících i nekódujících oblastech genů, i v oblastech mimo geny. SNP mohou být zděděné, de novo nebo získané v průběhu života.

VCF formát obvykle obsahuje hlavičku, jejíž řádky začínají znaky '#'. V hlavičce jsou obvykle informace o souboru včetně definice jednotlivých zkratek, které jsou následně zapsány jako názvy 'sloupců' v pomyslné 'tabulce'. Tyto sloupce obsahují informace o chromozomu, pozici koordináty, identifikační číslo varianty, poté alelu nalezenou v referenčním genomu a alelu nalezenou v zarovnávané sekvenci. Dále kvalitativní skóre a doplňující informace. Sloupce FORMAT obsahuje informace o hodnotách, které očekáváme v následujících sloupcích. Může tam být například informace o genotypu nebo kvalitě čtení. Toto jsou ty hodnoty, které jsme popisovali v hlavičce souboru.

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Obr. 11: Příklad VCF souboru. Převzato z https://samtools.github.io/hts-specs/VCFv4.2.pdf

2.5.5. Anotace variant

Pro anotaci se používá VEP neboli Ensembl Variant Effect Predictor, který slouží k popsání nalezených variant pomocí dostupných databází. V současnosti pilířem bioinformatiky je trojice vzájemně provázaných databází GenBank, EMBL a DDBJ. Ty jsou zároveň součástí konsorcia INSDC. Zdrojem veřejně dostupných databází je National Center of Biotechnology Information neboli NCBI.

2.5.6. Doplňující analýza

CNV neboli Copy Number Variation slouží k detekci duplikací či delecí celých exonů. Pro CNV analýzu se používá knihovna GATK neboli Genome Analysis Toolkit.

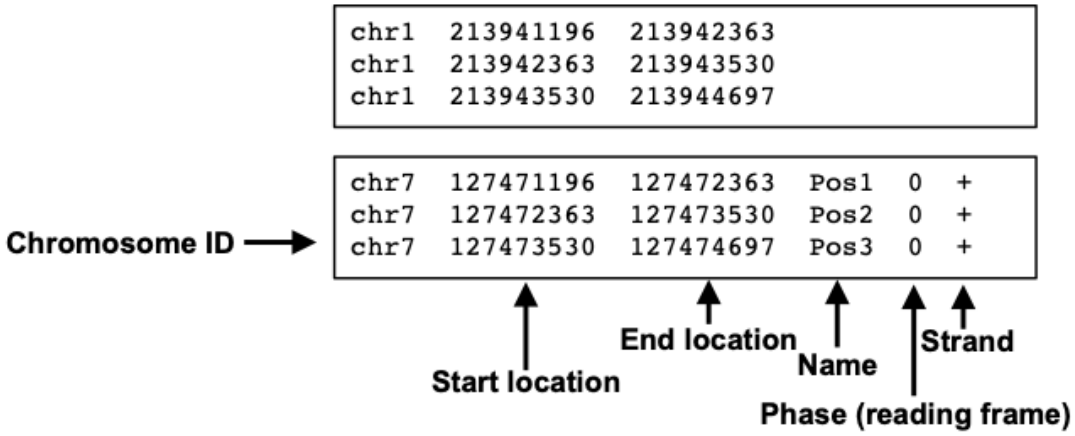
Pindel analýza zahrnuje detekci velkých inzercí, delecí, tandemových duplikací a jiných strukturálních variant ze sekvenačních dat. Pro tuto analýzu existuje přímo nástroj příkazového řádku *pindel*.

3. Experimentální část

3.1. Bioinformatické zpracování dat z Illumina sekvenátoru

Stažení dat z NGS stroje je možné z internetové stránky Illumina Basespace. Tyto data jsou ve formátu FASTQ. Pro správné spuštění analýzy je třeba mít k dispozici několik dalších souborů, které do analýzy vstupují. Jedná se o soubory BED, NM, ENST, soubor s exony a konfigurační soubor.

Soubor BED je formát souboru používaný v bioinformatických aplikacích pro reprezentaci genetických dat, jako jsou například genomické varianty. Soubory BED jsou běžně používány pro vizualizaci dat v genomických prohlížečích a jejich manipulaci v bioinformatických nástrojích. Typický soubor BED obsahuje informace o chromozomu, počáteční a koncové pozice variant společně s názvem a skóre pro danou variantu. V našem případě obsahuje soubor BED genomické pozice oblastí našeho zájmu.



Obr. 12: Příklad BED souboru. Převzato z: (Handling peak files with bedtools, n.d.*)*

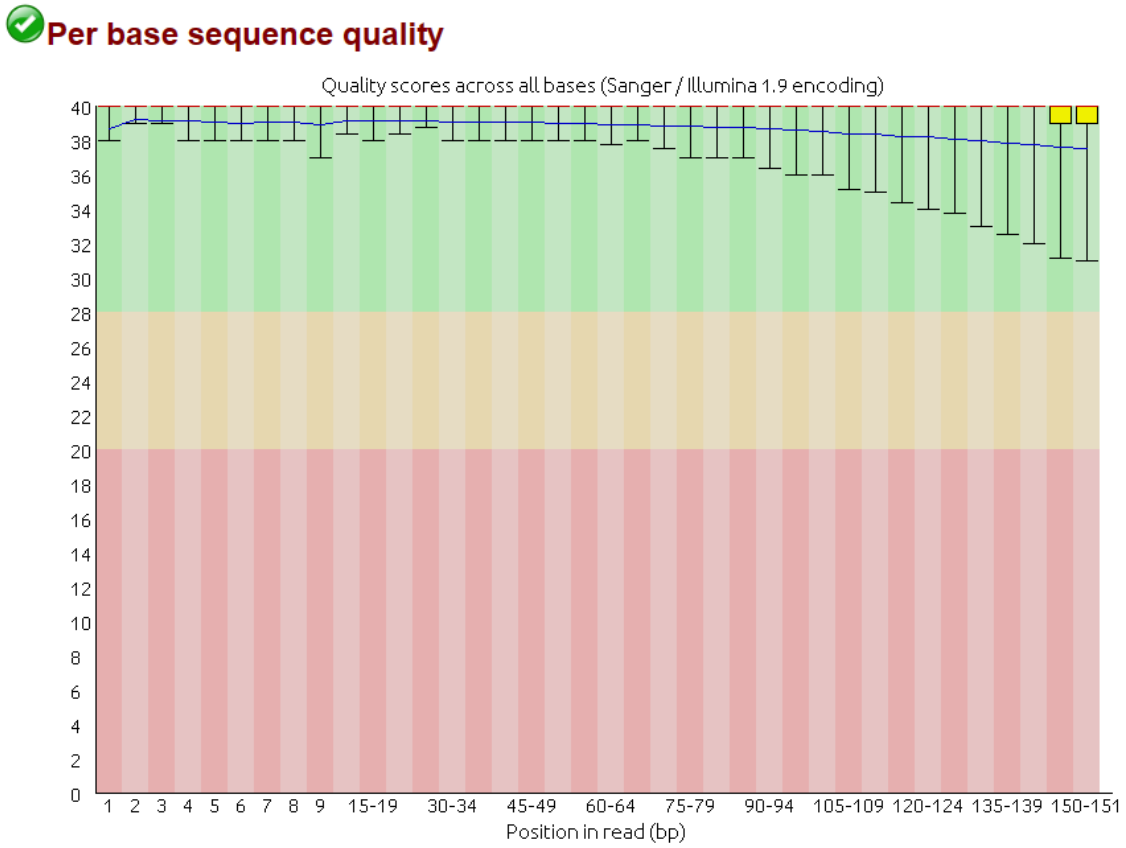
NM a ENST jsou soubory obsahující vybrané referenční sekvence. NM soubor obsahuje informace o referenčních mRNA transkriptech, které jsou kódovány geny. Každý transkript má přidělený unikátní identifikátor v podobě NM čísla. ENST obsahuje ENST čísla, což jsou identifikátory pro transkripty, které jsou kódovány geny v databázi Ensembl.

Soubor s exony obsahuje informace o pozicích jednotlivých exonů genů, které analyzujeme. V konfiguračním souboru jsou pak informace o názvu běhu, panelu genů, filtračních parametrech, použitém genomu a dalších.

3.1.1. Kontrola kvality dat

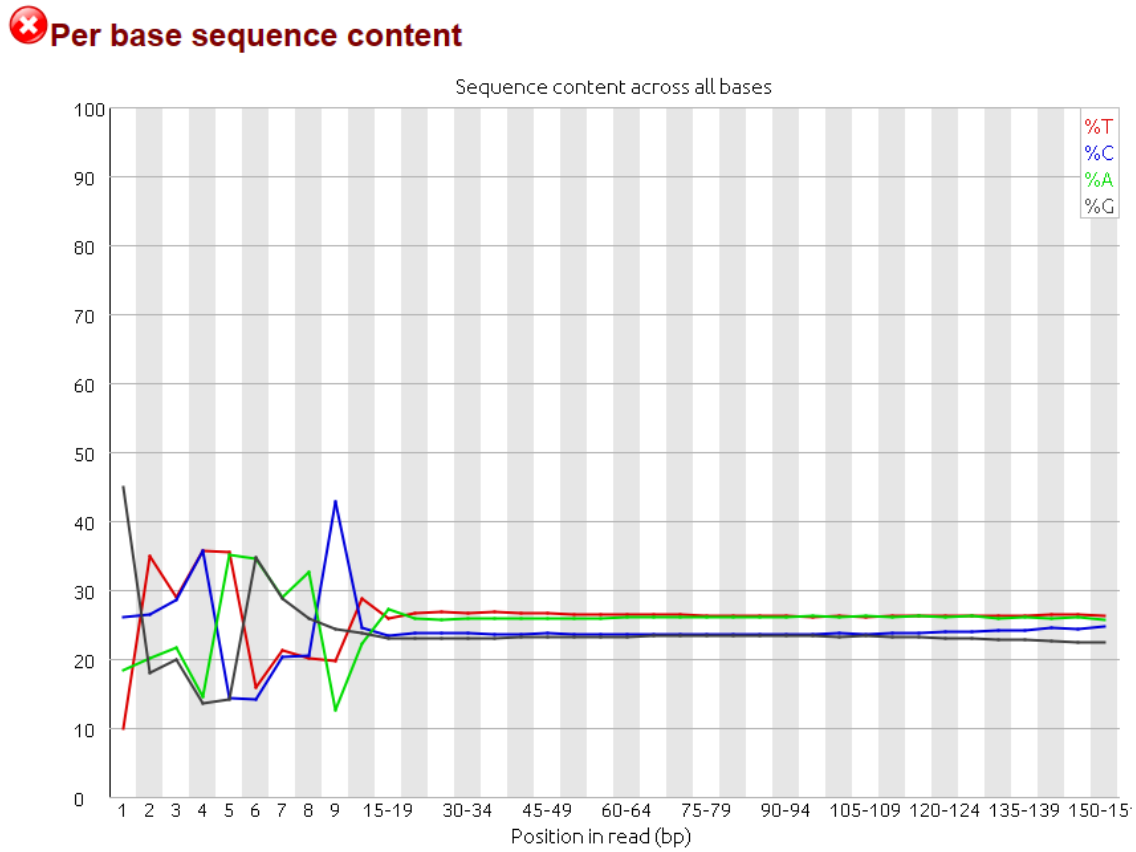
Kvalita dat se kontroluje v programu FastQC, což je nástroj používaný v bioinformatice k analýze kvality dat ze sekvenátorů. Tento nástroj poskytuje různé metriky a grafy, které umožňují uživatelům rychle zhodnotit kvalitu sekvencí získaných z experimentálního sekvenování.

Na obrázku můžete vidět jeden z výstupů programu FastQC. Jedná se o graf, na kterém jde krásně vidět, že všechny pozice čtení jsou dobré kvality, protože hodnoty kvality se vyskytují v zelené části grafu, konkrétně mezi hodnotami 30 a 40. Zároveň i sám program nám zelenou fajfkou naznačuje, že je podle něj v této části vše v pořádku.



Obr. 13: Příklad grafu z programu FastQC. Vygenerováno programem FastQC.

Na druhém grafu je naopak vidět, že ani sám program není s výstupem spokojený. Levá část grafu je to, co zde dělá problémy. Nejspíše se jedná o sekvenci adaptéru a měla by se správně vystřihnout.

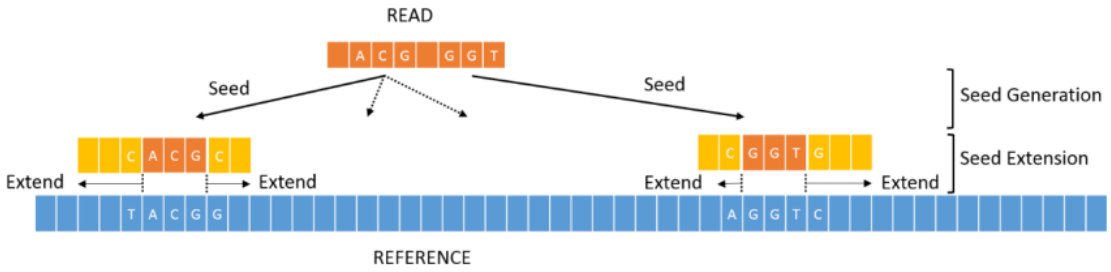


Obr. 14: Příklad grafu z programu FastQC se špatným vstupem. Vygenerováno programem FastQC.

3.1.2. Zarovnání čtení

Prvním krokem po kontrole dat je mapování na referenční genom. Používaným genomem je hg38. V našem případě se jedná o cílené sekvenování, stačí nám tedy zjistit přítomnost hledaných variant. Pro Burrows-Wheelerovu transformaci používáme softwarový balíček BWA, konkrétně BWA MEM. Výstupem je soubor ve formátu SAM.

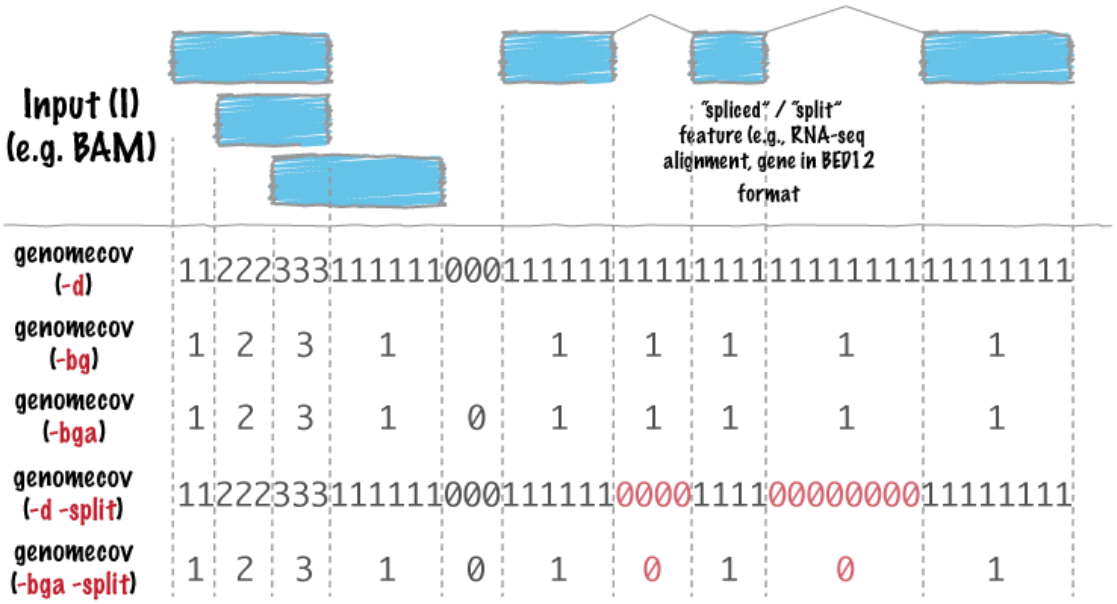
BWA MEM zpracovává čtení pomocí paradigmatu Seed and Extend. Pro každé čtení jsou pravděpodobné mapovací lokality nalezeny vyhledáním přesně odpovídajících podsekvencí, tzv. 'seedů'. Tyto 'seedy' jsou pak rozšířeny do obou směrů pomocí podobného Smith-Waterman algoritmu, který umožňuje neúplné shody. Z tohoto rozšíření je pak vybráno zarovnání s nejlepším skóre **(Houtgast et al., 2016)**.



Obr. 15: BWA MEM. Převzato z: (Houtgast et al., 2016)

3.1.3. Kontrola pokrytí

Pokrytí neboli coverage se stanovuje a kontroluje za účelem odstranění variant, které mají nízký počet čtení a mohou být tedy chybně označeny za pozitivní nebo naopak negativní nález. Jedná se o počet přečtení daného fragmentu. Pokrytí určuje nástroj bedtools. Oblasti s nízkým pokrytím jsou ukládány do samostatného souboru.



Obr. 16: Výpočet coverage pomocí nástroje bedtools. Převzato z: (Genomecov, n.d.*)*

3.1.4. Výstupní report

Výstupní report obsahuje informace, které molekulární biologové vyžadují a které lze získat pomocí popsaných metod.

NEXTSEQ_POOL	GENE	RefSeq	CHR	POS	EX/INT	COV	F	AF[%]	R	AR[%]	FREQ[%]	GT	HGVSc	HGVSp	AMINO	dbSNP	gnomAD_AF[%]	CLIN	SIFT	POLYPHEN	SO_term	IMPACT
744/1948	ADAMTS2	NM_014244.5	chr5	179140053	-/10	79	16	62,5	63	47,62	50,63	HET	c.1630-1BT>C	-	-	rs2303638	26,97	benign	-	-	intron_variant	MODIFIER
647/1948	CCER2	NM_001243212.2	chr19	38911815	2/-	72	41	56,1	31	45,16	51,39	HET	c.99G>C	p.Glu33Asp	E/D(33)	rs76973734	19,86	benign	deleterious	possibly_dama	missense_variant	MODERATE
12/1948	COL11A1	NM_001854.4	chr1	102913704	-/52	103	43	37,2	60	53,33	46,6	HET	c.3979-14A>T	-	-	rs186245518	0,5214	benign,li	-	-	splice_polypyrimidine_	LOW
863/1948	COL11A1	NM_001854.4	chr1	102979123	-/32	113	58	46,6	55	41,82	44,25	HET	c.2611-19A>C	-	-	rs11164649	69,65	benign	-	-	intron_variant	MODIFIER
533/1948	COL1A2	NM_000089.4	chr7	94427037	47/-	149	94	47,9	55	61,82	53,02	HET	c.3135C>T	p.Gly1045=	G(1045)	rs1800248	12,17	benign	-	-	synonymous_variant	LOW
588/1948	COL4A2	NM_001846.4	chr13	110445879	17/-	88	34	100	54	100	100	HOM	c.1008C>T	p.Pro336=	P(336)	rs4103	51,12	benign	-	-	synonymous_variant	LOW
489/1948	COL4A4	NM_000092.5	chr2	227057388	-/29	67	26	100	41	100	100	HOM	c.2545+51A>G	-	-	rs7567789	63,32	benign	-	-	intron_variant	MODIFIER
118/1948	COL9A2	NM_001852.4	chr1	40304111	-/24	14	7	57,1	7	42,86	50	HET	c.1288-12C>T	-	-	rs77695700	5,152	benign	-	-	splice_polypyrimidine_	LOW
962/1948	CRTAP	NM_006371.5	chr3	33132664	5/-	87	42	45,2	45	35,56	40,23	HET	c.1032T>G	p.Thr344=	T(344)	rs1135127	36,94	benign	-	-	synonymous_variant	LOW
816/1798	ERBB4	NM_005235.3	chr2	211387139	27/-	82	49	49	33	54,55	51,22	HET	c.3195A>G	p.Val1065=	V(1065)	rs3748962	30,03	benign	-	-	synonymous_variant	LOW
925/1948	EVC	NM_153717.3	chr4	5783715	12/-	56	18	55,6	38	39,47	44,64	HET	c.1727G>A	p.Arg576Gln	R/Q(576)	rs1383180	34,92	benign	deleterious	probably_dama	missense_variant	MODERATE
139/1948	FLNB	NM_001164317.2	chr3	58106801	12/-	85	43	32,6	42	57,14	44,71	HET	c.1869C>T	p.Asp623=	D(623)	rs2140104	3,586	benign	-	-	synonymous_variant	LOW
572/1422	LTBP1	NM_206943.4	chr2	33263352	15/-	140	70	51,4	70	40	45,71	HET	c.2577G>A	p.Thr859=	T(859)	rs12468099	24,81	-	-	-	synonymous_variant	LOW
232/1948	NBAS	NM_015909.4	chr2	15402228	26/-	126	84	38,1	42	42,86	39,68	HET	c.3011G>A	p.Arg1004Gln	R/Q(1004)	rs16862653	6,268	benign	deleterious_l	probably_dama	missense_variant	MODERATE
877/1948	NF1	NM_000267.3	chr17	31181757	7/-	184	117	44,4	67	38,81	42,39	HET	c.702G>A	p.Leu234=	L(234)	rs1801052	62,72	benign	-	-	synonymous_variant	LOW
825/1948	NOTCH1	NM_017617.5	chr9	136513480	14/-	130	78	51,3	52	57,69	53,85	HET	c.2265T>C	p.Asn755=	N(755)	rs2229971	40,07	benign	-	-	synonymous_variant	LOW
512/1948	NOTCH1	NM_017617.5	chr9	136523808	3/-	75	55	100	20	100	100	HOM	c.312T>C	p.Asn104=	N(104)	rs4489420	58,74	benign	-	-	synonymous_variant	LOW
893/1948	NPHS2	NM_014625.4	chr1	179551371	8/-	80	41	61	39	51,28	56,25	HET	c.954C>T	p.Ala318=	A(318)	rs1410592	61,9	benign	-	-	synonymous_variant	LOW
536/1948	NSD1	NM_022455.5	chr5	177210575	5/-	93	74	51,4	19	42,11	49,46	HET	c.2176T>C	p.Ser726Pro	S/P(726)	rs28932178	20,24	benign	tolerated	possibly_dama	missense_variant	MODERATE
411/1948	P3H1	NM_001243246.2	chr1	42747198	14/-	79	49	53,1	30	56,67	54,43	HET	c.2129T>A	p.Phe710Tyr	F/Y(710)	rs3738496	13,09	benign	tolerated_lov	benign	missense_variant	MODERATE
1001/1948	SCN9A	NM_002977.3	chr2	166288464	10/-	157	118	50	39	35,9	46,5	HET	c.1287T>A	p.Arg429=	R(429)	rs6747673	49,33	benign	-	-	synonymous_variant	LOW
893/1948	SCN9A	NM_002977.3	chr2	166311583	2/-	96	77	54,6	19	47,37	53,12	HET	c.174G>A	p.Gln58=	Q(58)	rs6432901	57,57	benign	-	-	synonymous_variant	LOW
369/1948	TGFBR3	NM_003243.5	chr1	91712381	13/-	133	86	100	47	100	100	HOM	c.2028T>C	p.Phe676=	F(676)	rs1805113	35,92	-	-	-	synonymous_variant	LOW
279/1948	TGFBR3	NM_003243.5	chr1	91861488	2/-	180	106	39,6	74	59,46	47,78	HET	c.44C>T	p.Ser15Phe	S/F(15)	rs1805110	13,13	-	tolerated	benign	missense_variant	MODERATE

Obr. 17: Příklad výstupního reportu. Převzato z Laboratoří AGEL Nový Jičín.

Sloupec POOL obsahuje informaci o počtu pacientů analyzovaných konkrétně v Laboratořích AGEL v Novém Jičíně, u kterých se objevila stejná varianta jako v analyzovaném vzorku. Dále vidíme název genu a jeho NM číslo. Jedná se o identifikační číslo z databáze RefSeq. V dalších sloupcích je informace o chromozomu, na které se daná varianta vyskytuje, společně s konkrétními pozicemi a informací, zda se jedná o exonovou či intronovou variantu.

Sloupec COV nám dává informaci o pokrytí, tedy o celkovém počtu čtení dané varianty. Ve sloupcích F, AF[%], R a AR[%] jsou pak hodnoty pokrytí na konkrétním vlákně.

Sloupec GT obsahuje informaci o tom, zda se jedná o homozygotní či heterozygotní variantu genu. V dalších sloupcích je psána změna báze a aminokyseliny. Číslo 'rs' je specifický identifikátor, který je používán pro danou variantu v databázi jednonukleotidových polymorfismů dbSNP. GnomAD je veřejně dostupná databáze genetických variant, která shromažďuje informace o genetické variabilitě v lidském genomu. Tento sloupec obsahuje informaci o frekvenci mutace v celkové populaci.

Sloupce CLIN, SIFT a POLYPHEN obsahují výsledky analýzy dané varianty z těchto predikčních programů. SO_term je soubor výrazů a vztahů, které popisují rysy biologické mutace a dopad na fenotyp. Fenotyp je soubor všech znaků znaků jedince. Je možné jej popsat jako projev genotypu. Mezi znaky morfologického typu patří například výška, hmotnost a IQ. Znaky fyziologického typu jsou třeba krevní tlak a hladina cukru v krvi.

3.2. Srovnání výstupů

Jedním z hlavních úkolů mé bakalářské práce bylo srovnat výsledky analýz ze dvou různých platforem. Pacienti tedy byli vybráni z běhů analyzovaných na obou platformách, tedy MGI i NEXTSEQ. Při srovnávání jsem pracovala jak s výstupními reporty, tak i s VCF a BAM soubory jednotlivých pacientů.

3.2.1. Pacienti

Analyzovala jsem celkem 7 pacientů, z nichž každý má jinou potvrzenou diagnózu. Ke každé diagnóze se vztahuje jiná patogenní varianta a každou diagnózu určuje jiný soubor genů. Značení pacientů je irelevantní pro tuto práci a je v souladu s GDPR.

L	ADPKD
S	HCM
O1	Retinitis pigmentosa
O2	Stargardtova makulární dystrofie
H	Nesyndromová hluchota
M	LDS, EDS, LFS
DD	Hereditární sférocytóza

Tab. 1: Pacienti a jejich diagnózy.

Pacient L má potvrzenou diagnózu ADPKD, což je autosomálně dominantní polycystická ledvinová nemoc. Jedná se o genetickou poruchu, která postihuje ledviny a může vést k postupnému zhoršování funkce ledvin až k selhání ledvin. Mezi hlavní rysy ADPKD patří tvorba cyst, což jsou tekutinou naplněné dutiny, v ledvinách. Tyto cysty mohou způsobit zvětšení ledvin a poškození zdravé tkáně. Tato choroba je způsobena mutacemi v genech PKD1 a PKD2 **(Audrézet et al., 2012)**.

Pacientovi S byla potvrzena diagnóza HCM, což je hypertrofická kardiomyopatie. Je to geneticky podmíněné onemocnění srdce charakterizované zesílením srdeční svaloviny, a to zejména v levé komoře **(Bonaventura & Veselka, 2019)**. Gen MYBPC3 kóduje protein zvaný kardiální myosinový vazebný protein C, který je důležitý pro regulaci kontrakce svalových vláken srdce. Mutace v tomto genu mohou mít různé dopady na funkci tohoto proteinu a mohou vést k různým klinickým projevům HCM, například k srdeční arytmii, angině pectoris nebo dokonce k náhlému srdečnímu úmrtí **(MYBPC3 gene, *n.d.*)**.

Onemocnění pacienta O1, Retinitis pigmentosa, je geneticky podmíněné onemocnění sítnice, které postupně vede k degeneraci fotoreceptorů v sítnici, což způsobuje postupnou ztrátu zraku. Jedná se o poměrně vzácné onemocnění, které ale může být způsobeno mnoha různými mutacemi v různých genech, mezi které patří například

RPGR, RHO a RP1. Klíčové rysy zahrnují ztrátu nočního, periferního až centrálního vidění. Na tuto nemoc neexistuje zatím lék na úplné vyléčení, avšak některé léky mohou pomoci zpomalit progresi onemocnění **(Genová terapie dědičných onemocnění SÍTNICE A ZRAKOVÉHO NERVU: současný stav poznání, *n.d.*)**.

Pacient O2 trpí Stargardtovou makulární dystrofií, což je geneticky podmíněné onemocnění očí, které postihuje makulu, což je část sítnice zodpovědná za centrální vidění a detailní vnímání. Tato nemoc je způsobena mutacemi v genech spojených s transportem a metabolismem vitaminu A v retinálních buňkách. Projevuje se ztrátou centrálního vidění a barev a citlivostí na světlo. Zatím pro ni neexistuje žádná léčba **(Stargardtova choroba, *n.d.*)**.

Pacient H má diagnostikovánu nesyndromovou hluchotu. Jedná se o formu hluchoty, která není spojena s žádnými dalšími významnými fyziologickými či anatomickými abnormalitami. Může být dědičná nebo získaná. Například infekční meningitida či zranění hlavy mohou vést k poškození sluchového ústrojí a mohou způsobit nesyndromovou hluchotu. U tohoto pacienta se však jedná o genetickou variantu, která může být způsobena mutacemi v různých genech, například OTOA **(Genetické příčiny percepčních sluchových vad, 2015)**.

Diagnóza pacienta M obsahuje Loeys-Dietz syndrom, Ehlers-Danlos syndrom a Li-Fraumeniho syndrom. LDS a EDS jsou geneticky podmíněné poruchy pojivové tkáně a cév. LDS se projevuje aneurysmaty či deformací kostí **(Pacienti s dědičnými aortopatiemi vyžadují multioborovou péči, *n.d.*)**. EDS se projevuje například náchylností k modřinám či hypermobilitou kloubů **(Diagnostika, *n.d.*)**. Li-Fraumeniho syndrom je vzácný genetický syndrom spojený s vysokým rizikem vzniku různých typů rakoviny v mladém věku. Je způsoben mutacemi v genech spojených s kontrolou buněčného růstu **(Foretová et al., *n.d.*)**.

Pacient DD má potvrzenou hereditární sférocytózu, což je geneticky podmíněné onemocnění červených krvinek, které způsobuje abnormálně tvarované červené krvinky, též sférocyty. Tyto sférocyty jsou mírně sférické a mají sníženou schopnost pružně se deformovat, což je důležité pro jejich průchod cévami. Příčinou hereditární sférocytózy jsou mutace v genech, které regulují strukturu a funkci membrány červených krvinek. Příznaky tohoto onemocnění mohou zahrnovat anémii, žloutenku, zvětšenou slezinu a močové kameny **(Hereditární sférocytóza, *n.d.*)**.

3.2.2. Vyřazení sekvenčních chyb

Jako první jsem hledala sekvenční chyby. Tyto chyby je možné najít ve výstupním reportu ve sloupcích F, AF[%], R a AR[%], kde se vyznačují číslem 0. Pokud například je u nějaké varianty ve sloupci R číslo 0, znamená to, že daná varianta není ani jednou čtena na dané pozici ve směru 3‘-5‘, tedy na reverse vlákně. Podobně je tomu u sloupce F, kde se zaměřujeme na pozice ve směru 5‘-3‘. Toto hledání se dá jednoduše provést přímo v excelové tabulce, do které je report formátován, a to pomocí seřazení řádků vzestupně podle hodnoty v těchto sloupcích.

Počet sekvenčních chyb jsem následně u každého pacienta porovnala mezi platformami. Čím méně sekvenčních chyb, tím lepší výsledek pro danou platformu. Počty jsou zaznamenány v tabulce. Můžeme si všimnout, že u většiny pacientů je výsledek celkem jednotný, a to lepší, tedy nižší počet sekvenčních chyb, u MGI. Výjimka je pouze u pacienta M.

pacient	MGI	NEXTSEQ
L	45	69
S	3	11
O1	10	25
O2	8	21
H	5	23
M	69	34
DD	8	25

Tab. 2: Počet sekvenčních chyb v reportech.

3.2.3. Porovnání variant v reportech

Dalším krokem byla filtrace variant, které jsou na dané platformě lépe čteny neboli které se v reportu z druhé platformy nevyskytují. Srovnání jsem prováděla v programu Meld, který je volně ke stažení na stránkách www.meldmerge.org. Tento program zvýrazňuje rozdíly v souborech, případně umožňuje soubory rovnou upravovat.

V tomto kroku jsem porovnávala tři sloupce v reportu, konkrétně název genu, jeho variantu, tedy NM číslo, a pozici na které se daná varianta vyskytuje. To mi umožnilo účinně selektovat pouze ty varianty, které jsou v obou reportech stejné. Zároveň tím umožňuji změny v jiných informacích, které nejsou jednoznačné pro danou variantu, jako je pokrytí nebo frekvence.

Zajímal mě opět především počet těchto lépe čtených variant. Výsledky jsou zaznamenány v tabulce. Tentokrát šlo o to, aby byl počet co nejvyšší, a opět si lépe vedla platforma MGI. Opět můžeme vidět jediný rozdíl u pacienta M, kde NEXTSEQ dopadl lépe.

pacient	MGI	NEXTSEQ
L	108	60
S	20	9
O1	55	24
O2	30	24
H	38	11
M	32	82
DD	62	34

Tab. 3: Počet lépe čtených variant v reportech.

pacient	MGI původní	NEXTSEQ původní	jednotné varianty
L	1930	1906	1777
S	576	573	553
O1	1214	1198	1149
O2	1248	1255	1210
H	688	679	645
M	651	666	550
DD	1100	1089	1030

Tab. 4: Rozdíl v počtu variant v původních a upravených reportech.

Po nalezení všech sekvenčních chyb i všech lépe čtených variant jsem dostala pro každého pacienta oba reporty se stejným počtem variant. Jednalo se o stejné varianty, co se týče verze genu a pozice. Mezi pacienty byl ale tento počet jiný, a to především z důvodu panelu genů, které jsou pro dané pacienty zkoumány. Geny jsou totiž před analýzou rozděleny podle toho, k jakému typu onemocnění se vztahují, například geny pro onemocnění očí, ledvin, srdce a další. Pacient je pak analyzován na tu sadu genů, která odpovídá jeho pravděpodobné diagnóze. Rozdíl v počtu variant před a po úpravě reportů je vidět v tabulce 4 výše.

3.2.4. Hledání kvalitních sekvenčních variant

Z laboratoří jsem dále dostala seznam kvalitních sekvenčních variant, což jsou varianty, které jsou správně čtené z obou stran. Tyto varianty byly vybrány z analýzy z NEXTSEQ běhu. Hledala jsem tedy tyto varianty nejprve v reportech z NEXTSEQ platformy a ty co jsem našla jsem následně hledala v reportech z MGI platformy. Účelem bylo, aby se obsažené varianty shodovaly u každého pacienta.

I v tomto bodě určitým způsobem vynikal pacient M, protože jedna ze sekvenčních variant nalezená v jeho NEXTSEQ reportu nebyla obsažena i v jeho MGI reportu. Jedná se o variantu zobrazenou na obrázku.

NEXTSEQ_POOL	GENE	RefSeq	CHR	POS	EX/INT	COV	FREQ[%]	GT	HGVSc	HGVSp	dbSNP	gnomAD_AF[%]	CLIN	SO_term	IMPACT
57/162	FLNB	NM_001164317.2	chr3	58123435	21/-	72	62,5	HET	c.3469G>A	p.Asp1157Asn	rs1131356	37,67	benign	missense_variant	MODERATE

Obr. 18: Varianta, která nebyla nalezena v MGI reportu pacienta M.

pacient	Počet kvalitních sekvenčních variant
L	5
S	3
O1	7
O2	8
H	4
M	17
DD	13

Tab. 5: Počet nalezených kvalitních sekvenčních variant u jednotlivých pacientů.

U ostatních pacientů všechny varianty nalezené v NEXTSEQ reportech byly i v MGI reportech. Počet těchto kvalitních sekvenčních variant nalezených u jednotlivých pacientů je zapsán v tabulce 5.

3.2.5. Srovnání pokrytí a frekvence

Pro validaci přístroje je důležitá mimo jiné kvalita čtení, proto v dalším kroku bylo potřeba srovnat hodnoty s ní úzce související. K tomuto srovnání jsem si napsala skript v jazyce Python, který na vstupu přijímá dva excelové soubory a čísla sloupců určených ke srovnání. Na výstupu pak vydá všechny srovnávané hodnoty a dále počet hodnot, které jsou vyšší v prvním souboru než hodnoty ve druhém souboru, počet hodnot nižších a počet stejných hodnot. Skript naleznete v příloze této práce. Výsledky jeho výpočtu jsou uvedeny v tabulkách níže.

Pacient	Vyšší u MGI	Vyšší u NEXTSEQ	stejné
L	608	1122	47
S	542	11	0
O1	911	221	17
O2	1131	74	5
H	635	9	1
M	95	444	11
DD	956	69	5

Tab. 6: Srovnání pokrytí mezi reporty pacientů.

V tomto bodě už se nám výsledky docela rozcházejí. Nelze s jistotou říct, zda je lepší MGI nebo NEXTSEQ. Především u pokrytí si můžeme všimnout, že počty vyšších hodnot mezi platformami si nejsou vůbec blízko.

Pacient	Vyšší u MGI	Vyšší u NEXTSEQ	stejné
L	651	576	550
S	189	192	172
O1	420	405	324

O2	433	419	358
H	240	225	180
M	192	186	172
DD	350	358	322

Tab. 7: Srovnání frekvence mezi reporty pacientů.

3.2.6. Srovnání kvality genotypu

Význam kvality genotypu při sekvenční analýze je klíčový, protože přesnost a spolehlivost získaných genotypů ovlivňuje správnost interpretace výsledků. Nízká kvalita genotypu může vést k chybám ve vyhodnocování genetických dat. To může způsobit falešně pozitivní nebo falešně negativní výsledky v analýzách genetických variant, což může mít důsledky na diagnózu chorob.

Při sekvenční analýze je tedy nutné zajistit, aby genotypy byly získány s co nejvyšší přesností a spolehlivostí. Toho lze dosáhnout pomocí kvalitních vzorků DNA či RNA, technických postupů, jako je příprava knihoven a sekvencování, a sofistikovaných bioinformatických metod pro analýzu a zpracování genetických dat. Důkladné kontroly kvality jsou klíčové pro minimalizaci chyb a zajištění důvěryhodnosti výsledků analýzy.

Kvalita genotypu je zjistitelná z VCF souboru každého pacienta. Zjišťovala jsem pro každého pacienta pro obě platformy maximální hodnotu kvality genotypu a nejčastěji opakovanou hodnotu kvality genotypu. Tyto hodnoty jsou zapsány v tabulce níže. Můžeme si všimnout, že maximální hodnota je vždy stejná a zároveň nejčastěji se opakující, což je dobře.

Pro zjištění těchto hodnot jsem použila příkaz v Bashi, který na vstupu přijímá VCF soubor a na výstupu vydá nejčastěji opakovanou hodnotu s počtem opakování. Pro maximální hodnotu kvality genotypu je příkaz upraven pouze chybějícím parametrem '-c' u příkazu 'uniq', které v původním příkazu udává počet výskytů každé jedinečné hodnoty, což v tomto případě nepotřebujeme.

```
cut -f 10 soubor.vcf | awk -F ':' '{print $2}' | sort $1 | uniq -c | sort -nr | head -n 1
cut -f 10 soubor.vcf | awk -F ':' '{print $2}' | sort $1 | uniq | sort -nr | head -n 1
```

Obr. 19: Příkazy v Bashi pro zjištění hodnot kvality genotypu.

pacient	MGI maximální	MGI nejčastější (počet opakování)	NEXTSEQ maximální	NEXTSEQ nejčastější (počet opakování)
L	255	255 (1553)	255	255 (1638)
S	255	255 (3072)	255	255 (1954)
O1	255	255 (1672)	255	255 (1339)
O2	255	255 (3055)	255	255 (2081)
H	255	255 (2776)	255	255 (1660)
M	255	255 (849)	255	255 (1144)
DD	255	255 (2260)	255	255 (1581)

Tab. 8: Maximální hodnoty kvality genotypu, nejčastěji opakované hodnoty kvality genotypu.

3.2.7. Srovnání kvality mapování

Kvalita mapování je klíčovým prvkem při sekvenční analýze, zejména při analýze genomických dat. Mapování se týká procesu přiřazování krátkých čtecích fragmentů získaných ze sekvenátoru na referenční genomickou sekvenci. Tento proces umožňuje identifikaci, kde se nacházejí čtecí fragmenty v genomu nebo v jiné referenční sekvenci.

Kvalita mapování ovlivňuje spolehlivost a přesnost výsledků sekvenční analýzy. Správné přiřazení čtecích fragmentů na referenční genom umožňuje správně identifikovat genetické varianty, jako jsou SNP, indely nebo strukturální varianty. Nesprávně zarovnané čtecí fragmenty mohou vést k chybám v identifikaci variant.

Kvalita mapování dále ovlivňuje pokrytí genomu, což je podíl genomu, který je pokrytý čtecími fragmenty. Vyšší kvalita mapování obvykle vede k vyššímu pokrytí genomu, což umožňuje robustnější analýzu. Celkově lze říci, že kvalita mapování přímo ovlivňuje spolehlivost a přesnost výsledků sekvenční analýzy.

Hodnota kvality mapování je zjistitelná z BAM souboru každého pacienta. Zjišťovala jsem průměrnou kvalitu mapování, počet výskytů maximální hodnoty kvality mapování a počet výskytů nižších hodnot. Toto všechno jsem opět zjišťovala pomocí příkazů v Bashi, tentokrát s využitím nástroje samtools. Aby byly nalezené hodnoty přehlednější,

vypočítala jsem jejich procentuální zastoupení v celkovém počtu čtení zjištěném ze statistiky níže. Výsledek je zapsán v tabulce.

```
samtools view soubor.bam | awk '{sum+=$5} END {print "průměr:", sum/NR}'
samtools view -q 60 soubor.bam | wc -l
samtools view soubor.bam | awk '{if ($5<60) count ++} END {print "menší než 60:", count}'
```

Obr. 20: Příkazy v Bashi pro zjištění hodnot kvality mapování.

	MGI			NEXTSEQ		
pacient	MQ Ø	MQ = max (60)	MQ <60	MQ Ø	MQ = max (60)	MQ <60
L	59,6115	98,04 %	1,96 %	59,4819	97,5 %	2,5 %
S	59,5776	97,89 %	2,11 %	59,4221	97,24 %	2,76 %
O1	59,5710	97,85 %	2,15 %	59,4206	97,18 %	2,82 %
O2	59,5817	97,90 %	2,10 %	59,4291	97,26 %	2,74 %
H	59,6119	98,05 %	1,95 %	59,4779	97,47 %	2,53 %
M	59,6495	98,27 %	1,73 %	59,5644	97,87 %	2,13 %
DD	59,5904	97,97 %	2,03 %	59,4607	97,41 %	2,59 %

Tab. 9: Průměr hodnot kvality mapování, procentuální výskyt vysokých a nízkých hodnot.

Dále jsem zjišťovala počet výskytů maximální hodnoty na každém vlákně zvlášť, tedy na forward vlákně čteném ve směru 5'-3' a na reverse vlákně čteném ve směru 3'-5'. Tato informace nám ukázala, jaké jsou rozdíly v mapování mezi vlákny. Počet čtení na každém vlákně zvlášť je opět zjišťován ve statistice níže. V BAM souboru se tato čtení od sebe odlišují znaménky '+' a '-'. Zjištěné hodnoty jsem opět vyjádřila jako procentuální podíl v tabulce.

```
samtools view -q 60 soubor.bam | awk '$9>0' | wc -l
samtools view -q 60 soubor.bam | awk '$9<0' | wc -l
```

Obr. 21: Příkazy v Bashi pro zjištění hodnot kvality mapování pro každé vlákno zvlášť.

	MGI		NEXTSEQ	
pacient	Forward MQ = max (60)	Reverse MQ = max (60)	Forward MQ = max (60)	Reverse MQ = max (60)
L	98,01 %	98,00 %	97,20 %	97,50 %
S	97,87 %	97,86 %	96,91 %	97,24 %
O1	97,83 %	97,82 %	96,85 %	97,19 %
O2	97,87 %	97,86 %	96,95 %	97,26 %
H	98,03 %	98,03 %	97,14 %	97,48 %
M	98,50 %	98,52 %	97,66 %	98,07 %
DD	97,95 %	97,95 %	97,10 %	97,43 %

Tab. 10: Procentuální výskyt vysokých hodnot kvality mapování.

3.2.8. Statistika

Pro celkovou statistiku analýzy pacienta jsem použila příkaz *samtools flagstat soubor.bam*. Tento příkaz slouží k analýze formátu BAM a poskytuje statistické informace o mapování sekvencí na referenční genom. Vypisuje informace o celkovém počtu porovnání sekvence s referenčním genomem a o počtu porovnání, které byly úspěšně zarovnány k referenčnímu genomu.

Dále o počtu přečtených párových porovnání a o počtu jednotlivých přečtení v párově porovnaných sekvencích. Pokud jsou totiž data vygenerována z párového sekvenování, každá DNA molekula je fragmentována a každý fragment je sekvenován z obou stran. Výsledkem jsou dvě sekvence, které jsou vzájemně opačné.

Další je informace o počtu zdvojených porovnání, které jsou mapované na více místech v referenčním genomu a jejich procentuální podíl. A nakonec počet porovnání, které nebyly vůbec zarovnány k referenčnímu genomu. Všechny tyto informace jsou zapsány pro každého pacienta v tabulkách v sekci přílohy.

3.3. Validační protokol

Účelem validace platformy je stanovení, že masivní paralelní sekvenační systém je schopen správně přečíst sekvenci DNA. Zároveň umožní vyhodnotit, jak přesně může být detekována každá varianta. Platforma nezahrnuje jen NGS sekvenátor, ale také izolaci DNA, obohacování metody, přípravu knihoven a analýzu dat. Součástí validace není interpretace nalezených variant.

3.3.1. Parametry validace

Specifičnost je pravděpodobnost negativního výsledku testu v případě nepřítomnosti hledané varianty testovaného znaku. Je vyjadřována jako poměr mezi správnou negativitou (TN) a součtem správné negativity a falešné pozitivity (FP) podle vzorce *TN/(TN+FP)*. Výsledná hodnota je vyjádřena jako relativní číslo, jehož hodnota se pohybuje v rozmezí od 0 do 1 a v ideálním případě se blíží 1.

Citlivost je pravděpodobnost pozitivního výsledku testu v případě přítomnosti hledané varianty testovaného znaku. Je vyjadřována jako poměr mezi správnou pozitivitou (TP) a součtem správné pozitivity a falešné negativity (FN) podle vzorce *TP/(TP+FN)*. Výsledná hodnota je vyjádřena jako relativní číslo, jehož hodnota se pohybuje v rozmezí od 0 do 1 a v ideálním případě se blíží 1.

Robustnost je schopnost metody poskytovat přijatelné výsledky měření i v případě, že dojde k malým odchylkám od nastavených podmínek jejího provádění. Udává spolehlivost metody při běžném používání.

4. Výsledky, diskuze

4.1. Výpočet

K výpočtu validačních parametrů jsem potřebovala nevytříděné reporty pacientů, tzv. snpindel reporty. Tyto reporty obsahují všechny varianty všech genů nalezených v genomu daného pacienta. Reporty, které používám výše, už jsou po úpravě a obsahují pouze ty varianty, které se týkají panelu genů, na kterém je pacient analyzován, tedy například panel genů pro onemocnění očí.

4.1.1. Specifičnost

Výpočet specifičnosti zahrnuje kontrolu skutečně negativních a falešně pozitivních pacientů. Podle diagnózy víme, kterou nemoc pacienti mají. Pro výpočet specifičnosti budeme předpokládat, že žádný z pacientů netrpí onemocněním jiného pacienta. U každého pacienta provedeme kontrolu přítomnosti variant genů všech onemocnění ostatních pacientů. Pro negativní výsledek píšeme 'wt', pro pozitivní výsledek píšeme 'patho'.

K hledání variant jsem použila příkaz v Bashi, který na vstupu přijímá snpindel report a textový soubor s geny, jejichž mutace kódují onemocnění některého z pacientů. Prohledává snpindel report a vybírá z něj varianty těchto genů. Následně vybrané varianty prohledá znovu a vypíše pouze ty, které obsahují slovo 'pathogenic'. Toto slovo v reportu značí, že daná varianta je patogenní podle některého z predikčních programů. V těchto variantách poté vyhledáváme jednotlivé varianty, které jsou patogenní u nemocného pacienta.

```
cat snpindel.txt | grep -Fwf gene.txt | grep -w "pathogenic" | grep -w ".*c.varianta.*" | less -S
```

Obr. 22: Příkaz v Bashi pro zjištění variant potřebných k výpočtu specifičnosti.

Výsledek zde byl vcelku jednoznačný. U všech pacientů byla potvrzena jejich negativní diagnóza na nemoci ostatních pacientů. U některých byly nalezeny patogenní varianty v genech některého z onemocnění, nejednalo se však o varianty, které jsem hledala. Výsledky hledání i výpočty jsou zapsány v tabulce 18.

pacient	diagnóza	MGI	NEXTSEQ	Výpočet MGI	Výpočet NEXTSEQ
L	HCM	wt	wt	TN/(TN+FP) = 42/(42+0) = 1	TN/(TN+FP) = 42/(42+0) = 1
	Retinis pigmentosa	wt	wt		
	Stargardtova makulární dystrofie	wt	wt		
	Nesyndromová hluchota	wt	wt		
	LDS, EDS, LFS	wt	wt		
	Hereditární sférocytóza	wt	wt		
S	ADPKD	wt	wt		
	Retinis pigmentosa	wt	wt		
	Stargardtova makulární dystrofie	wt	wt		
	Nesyndromová hluchota	wt	wt		
	LDS, EDS, LFS	wt	wt		
	Hereditární sférocytóza	wt	wt		
O1	ADPKD	wt	wt		
	HCM	wt	wt		
	Stargardtova makulární dystrofie	wt	wt		
	Nesyndromová hluchota	wt	wt		
	LDS, EDS, LFS	wt	wt		
	Hereditární sférocytóza	wt	wt		
O2	ADPKD	wt	wt		
	HCM	wt	wt		
	Retinis pigmentosa	wt	wt		
	Nesyndromová hluchota	wt	wt		
	LDS, EDS, LFS	wt	wt		
	Hereditární sférocytóza	wt	wt		
H	ADPKD	wt	wt		
	HCM	wt	wt		
	Retinis pigmentosa	wt	wt		

	Stargardtova makulární dystrofie	wt	wt		
	LDS, EDS, LFS	wt	wt		
	Hereditární sférocytóza	wt	wt		
M	ADPKD	wt	wt		
	HCM	wt	wt		
	Retinis pigmentosa	wt	wt		
	Stargardtova makulární dystrofie	wt	wt		
	Nesyndromová hluchota	wt	wt		
	Hereditární sférocytóza	wt	wt		
DD	ADPKD	wt	wt		
	HCM	wt	wt		
	Retinis pigmentosa	wt	wt		
	Stargardtova makulární dystrofie	wt	wt		
	Nesyndromová hluchota	wt	wt		
	LDS, EDS, LFS	wt	wt		

Tab. 11: Výpočet specifičnosti.

4.1.2. Citlivost

K výpočtu citlivosti jsem potřebovala zjistit kterým pacientům na které platformě vyšla jejich skutečná diagnóza. Pro každého pacienta jsem tedy hledala v jeho nevytříděném reportu varianty obsahující ty geny, jejichž mutace ovlivňují potvrzené onemocnění pacienta. Pro negativní výsledek opět píšeme 'wt' a pro pozitivní 'patho'. Následně počítáme citlivost podle vzorce, do kterého dosadíme počet skutečně pozitivních nálezů a falešně negativních nálezů.

Použila jsem podobný příkaz jako u specifičnosti, jen s drobnými úpravami. Příkaz přijímá opět snpindel report pacienta, ale hledá v něm pouze geny, které souvisí s jeho onemocněním. U těchto genů vyhledává jak slovo "pathogenic", tak i "HIGH", což

znamená, že má daná mutace velký dopad na fenotyp a zdraví jedince, a tedy může být patogenní. Ve výstupu následně hledáme konkrétní variantu.

```
cat snpindel.txt | grep -Fwf gene.txt | grep -w "pathogenic\|HIGH" | less -S
```

Obr. 23: Příkaz v Bashi pro zjištění variant potřebných k výpočtu citlivosti.

Výsledek zde nebyl tak uspokojivý jako u specifičnosti. U dvou pacientů nebyla nalezena žádná z variant, které se vztahují k jejich onemocnění, výsledek byl však na obou platformách vždy stejný. Je ale potřeba říct, že ty varianty, které jsem nenašla, v reportu daného pacienta jsou. Jen nesplňují podmínku, že by jejich řádek obsahoval slovo 'pathogenic' nebo 'HIGH', nemohla jsem je tedy do výpočtu zařadit a byla jsem nucena je považovat za falešně negativní nález. Nicméně to nemusí být, a pravděpodobně ani není, chyba sekvenátorů, jelikož právě tyto varianty nemusí být za každou cenu patogenní. Může u nich záviset na tom, zda se jedná o dominantní či recesivní alelu daného genu.

pacient	diagnóza	MGI	NEXTSEQ	Výpočet MGI	Výpočet NEXTSEQ
L	ADPKD	patho	patho	TP/(TP+FN) = 5/(5+2) = 0,714	TP/(TP+FN) = 5/(5+2) = 0,714
S	HCM	patho	patho		
O1	Retinis pigmentosa	patho	patho		
O2	Stargardtova makulární dystrofie	patho	patho		
H	Nesyndromová hluchota	wt	wt		
M	LDS, EDS, LFS	patho	patho		
DD	Hereditární sférocytóza	wt	wt		

Tab. 12: Výpočet citlivosti.

4.1.3. Robustnost

U výpočtu robustnosti chceme zjistit, zda je daný postup spolehlivý na obou platformách stejně. Chceme tedy nalézt na obou platformách, pokud možno všechny patogenní varianty pro onemocnění jednotlivých pacientů.

Příkaz pro nalezení patogenních variant nemocných pacientů je dost podobný oběma předchozím. Na vstupu přijímá snpindel report pacienta a soubor genů pro jeho onemocnění. V těchto genech pak vyhledává slovo 'pathogenic' a konkrétní varianty, které se vztahují k danému onemocnění. Pro negativní nález opět píšeme 'wt' a pro pozitivní 'patho'.

```
cat snpindel.txt | grep -Fwf gene.txt | grep -w "pathogenic" | grep -w ".*c.varianta.*" | less -S
```

Obr. 24: Příkaz v Bashi pro hledání variant potřebných k zjištění robustnosti.

Bohužel, většina patogenních variant nebyla v snpindel reportech nalezena. Výsledek mezi platformami je však stejný, takže se dá říct, že jsou obě platformy stejně spolehlivé. Opět ale platí to stejné, co u předchozího bodu, a tedy že ty varianty v reportech pacientů jsou, pouze je predikční programy nepovažují za jednoznačně patogenní.

pacient	geny	varianty	MGI	NEXTSEQ
L	PKD2	c.2356_2357 del	wt	wt
	SLC34A1	c.1223 T>A	wt	wt
	BBS10	c.273 C>G	wt	wt
S	MYBPC3	c.3697 C>T	patho	patho
O1	ABCA4	c.5882 G>A	patho	patho
	ABCA4	c.5603 A>T	wt	wt
	RHO	c.263 T>C	wt	wt
	TYR	c.164 G>A	wt	wt
	TYR	c.575 C>A	wt	wt
O2	ABCA4	c.5318 C>T	patho	patho
	HMCN1	c.5603 A>T	wt	wt
	CNGA3	c.3278 C>A	wt	wt
H	OTOA	c.3074 A>C	wt	wt
	ESPN	c.2405 G>A	wt	wt
M	ARHGEF17	c.3421 del	patho	patho
DD	ANK1	c.5097-33 G>A	wt	wt

Tab. 13: Zjišťování robustnosti.

5. Závěr

V první části práce jsem se zabývala bioinformatickou analýzou sekvenačních dat z platforem MGI a Illumina. Bylo mi vybráno 7 pacientů s různými diagnózami, jejichž data jsem srovnávala mezi platformami s cílem dozvědět se, která z platforem je lepší.

Nyní můžu říct, že se určité rozdíly mezi platformami vyskytly. Například v záchytech sekvenčních variant dominovala platforma MGI. Zjistila jsem, že výstup platformy NEXTSEQ obsahuje více sekvenčních chyb a zároveň méně kvalitních sekvenčních variant než výstup platformy MGI. Navíc výstupní report platformy MGI obsahoval celkově více záchytů, které nezachytil sekvenátor NEXTSEQ, než tomu bylo naopak.

Jediný pacient, u kterého byly výsledky opačné, byl pacient M. Vzhledem k ojedinělosti tohoto výsledku se ale můžeme nejspíš přiklonit k variantě, že to mohlo být způsobeno něčím jiným než chybou sekvenátoru. Každý pacient byl totiž analyzován na úplně jinou sadu genů.

U srovnání frekvence a pokrytí se nám hodnoty začaly docela rozcházet. Tento trend pak pokračoval i u srovnání kvalit genotypu a mapování, kde sice nebyly žádné závratné rozdíly mezi hodnotami, ale byly docela kolísavé. Ze zjištěných hodnot se ale nedá s jistotou o některé z platforem prohlásit, že by dopadla lépe než ta druhá.

Následně jsem se přesunula k výpočtům validačních parametrů. Hodnotila jsem specifičnost, citlivost a robustnost platforem. U specifičnosti byly výsledky podle předpokladů 100 % na obou platformách. U citlivosti jsem u dvou pacientů neobjevila patogenní varianty jejich onemocnění a musela jsem je tedy považovat za falešně negativní. Dané varianty však v reportech byly, jen se podle predikčních programů nejednalo o jednoznačně patogenní varianty. U robustnosti se pak vyskytlo více takových variant, které sice pacienti v reportu mají, ale nejsou specifikovány jako jednoznačně patogenní. Bohužel, toto kritérium jsem měla nastavené pro hledání daných variant, a nemohla jsem ho tedy přehlížet.

6. Literatura

Abril, J. F., & Castellano, S. (2019). Genome Annotation. *Encyclopedia of Bioinformatics and Computational Biology*, 195-209. https://doi.org/10.1016/B978-0-12-809633-8.20226-4

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-410. https://doi.org/10.1016/S0022-2836(05)80360-2

Audrézet, M. -P., Cornec-Le Gall, E., Chen, J. -M., Redon, S., Quéré, I., Creff, J., Bénech, C., Maestri, S., Le Meur, Y., & Férec, C. (2012). Autosomal dominant polycystic kidney disease: Comprehensive mutation analysis of PKD1 and PKD2 in 700 unrelated patients. *Human Mutation*, *33*(8), 1239-1250. https://doi.org/10.1002/humu.22103

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *Journal of Experimental Medicine*, *79*(2), 137-158. https://doi.org/10.1084/jem.79.2.137

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing?, *98*(6), 236-238. https://doi.org/10.1136/archdischild-2013-304340

Benson, D., Lipman, D. J., & Ostell, J. (1993). GenBank. *Nucleic Acids Research*, *21*(13), 2963-2965. https://doi.org/10.1093/nar/21.13.2963

BLAST: Basic Local Alignment Search Tool. National Center for Biotechnology Information. Retrieved May 13, 2024, from https://blast.ncbi.nlm.nih.gov/Blast.cgi

Bonaventura, J., & Veselka, J. (2019). Genetic testing in patients with hypertrophic cardiomyopathy. *Vnitřní lékařství*, *65*(10), 652-658. https://doi.org/10.36290/vnl.2019.113

CDNA (copy DNA). National Human Genome Research Institute. Retrieved May 13, 2024, from https://www.genome.gov/genetics-glossary/Copy-DNA

Comparative Genomics. Nature. Retrieved May 13, 2024, from https://www.nature.com/scitable/knowledge/library/comparative-genomics-13239404

Crick, F. H. C. (1968). The origin of the genetic code. *Journal of Molecular Biology*, *38*(3), 367-379. https://doi.org/10.1016/0022-2836(68)90392-6

Dayhoff, M. O., Chang, M. A., Eck, R. V., & Sochard, M. R. (1965). *Atlas of protein sequence and structure*.

Diagnostika. Ehlers-Danlosův syndrom a syndrom hypermobility. Retrieved May 13, 2024, from https://www.ehlers-danlosuv-syndrom.org/diagnostika?fbclid=IwZXh0bgNhZW0CMTAAAR3gzwPS8OsYXx582k7NHQX0JKLUXn0X-7oL6U90pnF0gBlLHtGpRiApanw_aem_AYJQL8KPm1RorL6ofksinicjI8K6oz2lU0XD6nZezQsuPExlTTNpPltyb6Qk8ca89ghfy4PcR8KUqhVlRElCI9cI

FASTA. Wikipedia. Retrieved May 13, 2024, from https://en.wikipedia.org/wiki/FASTA

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. -F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh,, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. -lng, Glodek, A., et al. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496-512. https://doi.org/10.1126/science.7542800

Foretová, L., Sedláček, Z., & Křepelová, A. Syndrom Li-Fraumeni - diagnostické a preventivní možnosti. Kazuistika pacientky s delecí celého genu TP53. https://www.linkos.cz/lekar-a-multidisciplinarni-tym/kongresy/po-kongresu/databaze-tuzemskych-onkologickych-konferencnich-abstrakt/syndrom-li-fraumeni-diagnosticke-a-preventivni-moznosti-kazuistika-pacientky-s-d/?fbclid=IwZXh0bgNhZW0CMTAAAR1EnWtb7tClGr_YDTdTY-uuvy6VxDKIrgu98GIQmo0u5LAVM4mkiTJCsr0_aem_AYJJBJUA8ttZN3hN4HN9eDFakDUA4ggPgeHAYHSfIc80wr723Ms7Eye6XdWSZ9IyqyvBb7KrgnBJiSQtF25i7YPr

Gene Expression. Nature. Retrieved May 13, 2024, from https://www.nature.com/scitable/topicpage/gene-expression-14121669/

Gene Ontology Overview. Gene Ontology. Retrieved May 13, 2024, from https://geneontology.org/docs/ontology-documentation

Genetické příčiny percepčních sluchových vad. (2015). IDětskýSluch. Retrieved May 13, 2024, from https://www.idetskysluch.cz/sluchove-vady/priciny/geneticke-priciny-percepcnich-sluchovych-vad-8/?fbclid=IwZXh0bgNhZW0CMTAAAR22HSt-oMhqb-M_H9YuQVKd9yB2QTbmNH_SftzE5u-whaU3rUuNZOsqZ28_aem_AYL4F6W7cZwg7Pj8GFohLbqpIoRmXuFveDDQ2bOhxVi6gicnw4Kxm8cQPMXqTt1nacaQ61e3HBk21K7pOdpoUYUJ

Genetický kód. Wikipedie. Retrieved May 13, 2024, from https://cs.wikipedia.org/wiki/Genetick%C3%BD_k%C3%B3d

Genomecov. Bedtools: a powerful toolset for genome arithmetic. Retrieved May 13, 2024, from https://bedtools.readthedocs.io/en/latest/content/tools/genomecov.html

Genová terapie dědičných onemocnění SÍTNICE A ZRAKOVÉHO NERVU: současný stav poznání. Pro Lékaře. Retrieved May 13, 2024, from https://www.prolekare.cz/casopisy/ceska-slovenska-oftalmologie/2016-4/genova-terapie-dedicnych-onemocneni-sitnice-a-zrakoveho-nervu-soucasny-stav-poznani-59509?fbclid=IwZXh0bgNhZW0CMTAAAR1EnWtb7tClGr_YDTdTY-uuvy6VxDKIrgu98GIQmo0u5LAVM4mkiTJCsr0_aem_AYJJBJUA8ttZN3hN4HN9eDFakDUA4ggPgeHAYHSfIc80wr723Ms7Eye6XdWSZ9IyqyvBb7KrgnBJiSQtF25i7YPr

Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R., & Gelbart, W. M. (2000). *An Introduction to Genetic Analysis*. https://doi.org/10.1023/A:1015187026471

Guan, Y. -F., Li, G. -R., Wang, R. -J., Yi, Y. -T., Yang, L., Jiang, D., Zhang, X. -P., & Peng, Y. (2012). Application of next-generation sequencing in clinical oncology to

advance personalized treatment of cancer. *Chinese Journal of Cancer*, *31*(10), 463-470. https://doi.org/10.5732/cjc.012.10216

Handling peak files with bedtools. GitHub. Retrieved May 13, 2024, from https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/07_handling_peaks_bedtools.html

Hereditární sférocytóza. WikiSkripta. Retrieved May 13, 2024, from https://www.wikiskripta.eu/w/Heredit%C3%A1rn%C3%AD_sf%C3%A9rocyt%C3%B3za?fbclid=IwZXh0bgNhZW0CMTAAAR3gzwPS8OsYXx582k7NHQX0JKLUXn0X-7oL6U90pnF0gBlLHtGpRiApanw_aem_AYJQL8KPm1RorL6ofksinicjI8K6oz2lU0XD6nZezQsuPExlTTNpPltyb6Qk8ca89ghfy4PcR8KUqhVlRElCI9cI#cite_ref-1

Houtgast, E. J., Sima, V. -M., Bertels, K., & Al-Ars, Z. (2016). GPU-Accelerated BWA-MEM Genomic Mapping Algorithm Using Adaptive Load Balancing. *Architecture of Computing Systems – ARCS 2016*, 130-142. https://doi.org/10.1007/978-3-319-30695-7_10

How does MGI sequencing technology work? Alithea Genomics. Retrieved May 13, 2024, from https://alitheagenomics.com/blog/how-does-mgi-sequencing-technology-work

Chang, F., & Li, M. M. (2013). Clinical application of amplicon-based next-generation sequencing in cancer. *Cancer Genetics*, *206*(12), 413-419. https://doi.org/10.1016/j.cancergen.2013.10.003

Chen, J. (2004). Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business. *Briefings in Bioinformatics*, *5*(3), 305-307. https://doi.org/10.1093/bib/5.3.305

Illumina Sequencing Platforms. (2023). Illumina. Retrieved February 9, 2023, from https://emea.illumina.com/systems/sequencing-platforms.html

IUPAC-IUB Comm. on Biochem. Nomenclature, . (1968). A one-letter notation for amino acid sequences. Tentative rules. *Biochemistry*, *7*(8), 2703-2705. https://doi.org/10.1021/bi00848a001

Jones, S., Hruban, R. H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D. W., Lin, J. C. -H., Palmisano, E., Brune, K., Jaffee, E. M., Iacobuzio-Donahue, C. A., Maitra, A., Parmigiani, G., Kern, S. E., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Eshleman, J. R., Goggins, M., et al. (2009). Exomic Sequencing Identifies PALB2 as a Pancreatic Cancer Susceptibility Gene. *Science*, *324*(5924), 217-217. https://doi.org/10.1126/science.1171202

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, *453*(7191), 56-64. https://doi.org/10.1038/nature06862

Koubková, L., Vyzula, R., & Vojtěšek, B. (2014). Sekvenování nové generace a možnosti jeho využití v onkologické praxi. *Klinicka onkologie*, *27*(Suppl 1). https://www.linkos.cz/files/klinicka-onkologie/395/4484.pdf

Lee, J., Freddolino, P. L., & Zhang, Y. (2017). Ab Initio Protein Structure Prediction. *From Protein Structure to Function with Bioinformatics*, 3-35. https://doi.org/10.1007/978-94-024-1069-3_1

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Liao, Y. -L., Li, Y. -C., Chen, N. -C., & Lu, Y. -C. (2018). Adaptively Banded Smith-Waterman Algorithm for Long Reads and Its Hardware Accelerator. *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, 1-9. https://doi.org/10.1109/ASAP.2018.8445105

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., & Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, *2012*, 1-11. https://doi.org/10.1155/2012/251364

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, *9*(1), 387-402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. -J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380. https://doi.org/10.1038/nature03959

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560-564. https://doi.org/10.1073/pnas.74.2.560

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31-46. https://doi.org/10.1038/nrg2626

MGI Sequencers. Retrieved May 13, 2024, from https://en.mgi-tech.com/products/

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, *23*(4), 183-191. https://doi.org/10.1016/j.tig.2007.02.006

MYBPC3 gene. MedlinePlus. Retrieved May 13, 2024, from https://medlineplus.gov/genetics/gene/mybpc3/?fbclid=IwZXh0bgNhZW0CMTAAAR1wnprhnwkg2Bhvd6Fc4hKdwlB_R7XJFpohiNdVQJYjDZzSnNAlw0Oz1mw_aem_AYIEoHRx-Db-te0lemOqvqjjidbo0ZnQqChGS-8rf476fuaElJ587D7Eu4Y0DX3Bg2AAejE3QJmZYiYIu2fJPDCJ#references

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443-453. https://doi.org/10.1016/0022-2836(70)90057-4

Nguyen, A., Schwinn, L. M., Eskofier, B. M., & Zhang, W. *Conformance Checking for a Medical Training Process Using Petri net Simulation and Sequence Alignment*.

Nirenberg, M., & Leder, P. (1964). RNA Codewords and Protein Synthesis. *Science*, *145*(3639), 1399-1407. https://doi.org/10.1126/science.145.3639.1399

Omoboye, O. O., Geudens, N., Duban, M., Chevalier, M., Flahaut, C., Martins, J. C., Leclère, V., Oni, F. E., & Höfte, M. (2019). Pseudomonas sp. COW3 Produces New Bananamide-Type Cyclic Lipopeptides with Antimicrobial Activity against Pythium myriotylum and Pyricularia oryzae. *Molecules*, *24*(22). https://doi.org/10.3390/molecules24224170

Oxford Nanopore Technologies. Retrieved May 13, 2024, from https://nanoporetech.com/

Pacienti s dědičnými aortopatiemi vyžadují multioborovou péči. Medical Tribune. Retrieved May 13, 2024, from https://www.tribune.cz/medicina/pacienti-s-dedicnymi-aortopatiemi-vyzaduji-multioborovou-peci/?fbclid=IwZXh0bgNhZW0CMTAAAR18F_3RBLVv2NRv2q9UiXkUpZM_BVWPey_3kuDmJasQUoZfKOq20iSpAEQ_aem_AYK-zrCW8FUYblkWm4Ll6abC1gO9Mj-8Lj_F-LOnJv5Gm2hn2LlroLQ2ze-sBym0OI1HZ5opMB3CEi8NqmZprzJD

Pan, S., & Tang, J. (Eds.). (2021). *Clinical Molecular Diagnostics*. Springer Singapore. https://doi.org/10.1007/978-981-16-1037-0

Phylogenetic Trees. Biological Principles. Retrieved May 13, 2024, from https://bioprinciples.biosci.gatech.edu/module-1-evolution/phylogenetic-trees

Probable disease resistance protein At1g58602. AlphaFold Protein Structure Database. Retrieved May 13, 2024, from https://alphafold.ebi.ac.uk/entry/Q8W3K0

Reina, O. (2016). *Vital for Soup, Vital for Labs: Serial Analysis of Gene Expression (SAGE), part 1*. BiteSize Bio. Retrieved May 13, 2024, from https://bitesizebio.com/30076/serial-analysis-of-gene-expression-sage-part-1

Rice, C. M., Fuchs, R., Higgins, D. G., Stoehr, P. J., & Cameron, G. N. (1993). The EMBL data library. *Nucleic Acids Research*, *21*(13), 2967-2971. https://doi.org/10.1093/nar/21.13.2967

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463-5467. https://doi.org/10.1073/pnas.74.12.5463

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*(8), 1794-1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Shyr, D., & Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biological Procedures Online*, *15*(1). https://doi.org/10.1186/1480-9222-15-4

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. -C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849-864. https://doi.org/10.1101/gr.213611.116

Smith-Waterman Algorithm. Retrieved May 13, 2024, from https://cs.stanford.edu/people/eroberts/courses/soco/projects/computers-and-the-hgp/smith_waterman.html

Stargardtova choroba. WikiSkripta. Retrieved May 13, 2024, from https://www.wikiskripta.eu/w/Stargardtova_choroba?fbclid=IwZXh0bgNhZW0CMTAAAR2S1oTiupE3gl2zEHmT6RXWe2naYd-BJHCHZtC0UgnO0n8l7cEEzStgjLo_aem_AYKa2iACdb-DKIH4FP7gcrMcdBcnFhvQxfpqWiZQD76iqXLYL9nT8dJ9_mKUGgOxcJppK0vFGDjtW6ZWACZCdmi6

Syngai, G., Barman, P., Bharali, R., & Dey, S. (2013). BLAST: An introductory tool for students to Bioinformatics Applications. *Keanean Journal of Science*, (2).

Tamm, C., Shapiro, H. S., & Lipshitz, R. Distribution density of nucleotides within a deoxyribonucleic acid chain.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., et al. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304-1351. https://doi.org/10.1126/science.1058040

Watson, J. D., & Crick, F. H. C. (1953). THE STRUCTURE OF DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*, 123-131. https://doi.org/10.1101/SQB.1953.018.01.020

Xuan, J., Yu, Y., Qing, T., Guo, L., & Shi, L. (2013). Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters*, *340*(2), 284-295. https://doi.org/10.1016/j.canlet.2012.11.025

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology and application, *1*(6), 520-536. https://doi.org/10.1007/s13238-010-0065-3

7. Seznam použitých symbolů a zkratek

BAM – Binary Alignment Map

BED – Browser Extensible Data

BLAST – Basic Local Alignment Search Tool

BWA – Burrows-Wheeler Alignment

cDNA – komplementární DNA

DDBJ – DNA DataBank of Japan

DNA – deoxyribonukleová kyselina

EMBL – European Molecular Biology Laboratory

GATK – Genome Analysis Toolkit

gnomAD – The Genome Aggregation Database

GO – Gene Ontology

ChIP-Seq - Chromatin ImmunoPrecipitation followed by Sequencing

IGV – Integrative Genomics Viewer

INSDC – International Nucleotide Sequence Database Collaboration

mRNA – messenger RNA

MSA - Multiple Sequence Alignment

NCBI – National Center for Biotechnology Information

NGS – Next Generation Sequencing (sekvenování nové generace)

PCR – Polymerase Chain Reaction (polymerázová řetězová reakce)

RMSD – Root-Mean-Square Deviation

RNA – ribonukleová kyselina

rRNA – ribosomal RNA

SAGE – Serial Analysis of Gene Expression

SAM – Sequence Alignment Map

SMRT – Single Molecule Real-Time sequencing

SNP – Single Nucleotide Polymorfism

SOLiD – Sequencing by Oligonucleotide Ligation and Detection

TIGR – The Institute for Genomic Research

tRNA – transfer RNA

UPGMA – Unweighted Pair Group Method with Arithmetic mean

VCF – Variant Call Format

VEP – Variant Effect Predictor

8. Seznam obrázků

9. Seznam tabulek

10.Přílohy

10.1. Tabulky se statistikou každého pacienta

	MGI	NEXTSEQ
Celkový počet porovnání	7 658 179	10 419 787
Úspěšně zarovnaná porovnání	100 %	100 %
Podíl párových porovnání	99,98 %	99,99 %
Úspěšná párová porovnání	99,67 %	99,61 %
Správně mapovaná párová porovnání	99,97 %	99,92 %
Jednočetná porovnání	0,0065 % (496)	0,0620 % (6 464)
Špatně mapovaná párová porovnání	0,0533 % (4 080)	0,1171 % (12 199)

Tab. 14: Statistika analýzy pacienta L.

	MGI	NEXTSEQ
Celkový počet porovnání	15 674 445	12 681 687
Úspěšně zarovnaná porovnání	100 %	100 %
Počet párových porovnání	99,98 %	99,99 %
Úspěšná párová porovnání	99,35 %	99,39 %
Správně mapovaná párová porovnání	99,97 %	99,92 %
Jednočetná porovnání	0,0048 % (751)	0,0619 % (7 848)
Špatně mapovaná párová porovnání	0,0521 % (8 167)	0,1384 % (17 556)

Tab. 15: Statistika analýzy pacienta S.

	MGI	NEXTSEQ
Celkový počet porovnání	9 115 043	9 437 745
Úspěšně zarovnaná porovnání	100 %	100 %
Počet párových porovnání	99,98 %	99,98 %
Úspěšná párová porovnání	99,83 %	99,73 %
Správně mapovaná párová porovnání	99,97 %	99,91 %
Jednočetná porovnání	0,0068 % (624)	0,0771 % (7 278)
Špatně mapovaná párová porovnání	0,0529 % (4 824)	0,1252 % (11 816)

Tab. 16: Statistika analýzy pacienta O1.

	MGI	NEXTSEQ
Celkový počet porovnání	16 091 305	14 407 970
Úspěšně zarovnaná porovnání	100 %	100 %
Počet párových porovnání	99,98 %	99,99 %
Úspěšná párová porovnání	99,18 %	99,31 %
Správně mapovaná párová porovnání	99,97 %	99,93 %
Jednočetná porovnání	0,0049 % (791)	0,0584 % (8 409)
Špatně mapovaná párová porovnání	0,0564 % (9 074)	0,1330 % (19 166)

Tab. 17: Statistika analýzy pacienta O2.

	MGI	NEXTSEQ
Celkový počet porovnání	13 952 937	10 935 021
Úspěšně zarovnaná porovnání	100 %	100 %
Počet párových porovnání	99,97 %	99,98 %
Úspěšná párová porovnání	99,64 %	99,58 %
Správně mapovaná párová porovnání	99,97 %	99,92 %
Jednočetná porovnání	0,0038 % (527)	0,0598 % (6 542)
Špatně mapovaná párová porovnání	0,0488 % (6 809)	0,1344 % (14 697)

Tab. 18: Statistika analýzy pacienta H.

	MGI	NEXTSEQ
Celkový počet porovnání	7 132 618	12 582 080
Úspěšně zarovnaná porovnání	100 %	100 %
Počet párových porovnání	98,99 %	98,85 %
Úspěšná párová porovnání	99,26 %	98,72 %
Správně mapovaná párová porovnání	98,98 %	98,81 %
Jednočetná porovnání	0,0122 % (869)	0,0433 % (5 449)
Špatně mapovaná párová porovnání	0,5189 % (37 014)	0,8801 % (110 732)

Tab. 19: Statistika analýzy pacienta M.

	MGI	NEXTSEQ
Celkový počet porovnání	11 598 356	10 716 516
Úspěšně zarovnaná porovnání	100 %	100 %
Počet párových porovnání	99,97 %	99,98 %
Úspěšná párová porovnání	99,67 %	99,62 %
Správně mapovaná párová porovnání	99,96 %	99,93 %
Jednočetná porovnání	0,0049 % (570)	0,0537 % (5 759)
Špatně mapovaná párová porovnání	0,0482 % (5 589)	0,1282 % (13 741)

Tab. 20: Statistika analýzy pacienta DD.