Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Bioinformatika

Studijní obor: B-BINF

Arnošt Polák

Identifikace sekvenačních chyb v NGS vznikajících na podkladě sekvenčního kontextu, analytických postupů a mapovacích nástrojů

Identification of NGS sequencing errors caused by sequencing context, analytical procedures, and mapping tools

Bakalářská práce

Vedoucí práce:
Mgr. Kateřina Matějková

Konzultant:
prof. MUDr. Zdeněk Kleibl, Ph.D.

Praha, 2025

Prohlášení

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně, a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce, ani její podstatná část, nebyla předložena k získání jiného nebo stejného akademického titulu. Pro práci byla použita umělá inteligence a technologie podporované umělou inteligencí.

V Praze, 28.4. 2025

Arnošt Polák

Poděkování

Chtěl bych zde poděkovat výbornému a trpělivému duu vedoucí a konzultanta, kteří vždy byli ochotni pomoci a práce s nimi byla na té nejvyšší úrovni. Také bych chtěl poděkovat mé přítelkyni za ochotu a nekonečnou podporu. Nakonec bych chtěl poděkovat i rodině a mým kamarádům za prostředí, ve kterém se velmi dobře tvořilo.

Abstrakt

Sekvenování DNA je základním pilířem moderní diagnostiky pro současné potřeby molekulární medicíny. Ačkoliv soudobé technologie sekvenování druhé generace (NGS) umožňují efektivní přečtení celého lidského genomu, jsou – od přípravy vzorku po bioinformatickou analýzu – stále zatíženy různými chybami. Ty mohou mít závažné důsledky pro výstupy výsledného genetického vyšetření.

Práce se zaměřuje na systematické chyby specifické pro nejrozšířenější diagnostickou platformu NGS, technologii Illumina, při analýze germinální DNA. Práce se věnuje chybám vyplývajícím ze sekvenčního kontextu, vznikajícím v důsledku sekvenační technologie, chybám souvisejícím s bioinformatickým zpracování, včetně mapování krátkých čtení na referenční genom a chybám vznikajícím během identifikace variantních alel (*variant callingu*).

Cílem práce je přispět k lepšímu porozumění zdrojům těchto chyb a podpořit přesnější a bezpečnější využití NGS v klinickém prostředí. Výsledky mohou sloužit jako referenční rámec pro výběr vhodných nástrojů, metod a parametrů v bioinformatické analýze genomových dat.

Klíčová slova: Illumina, DNA sekvenování, mapování, *variant calling*, chyby

Abstract

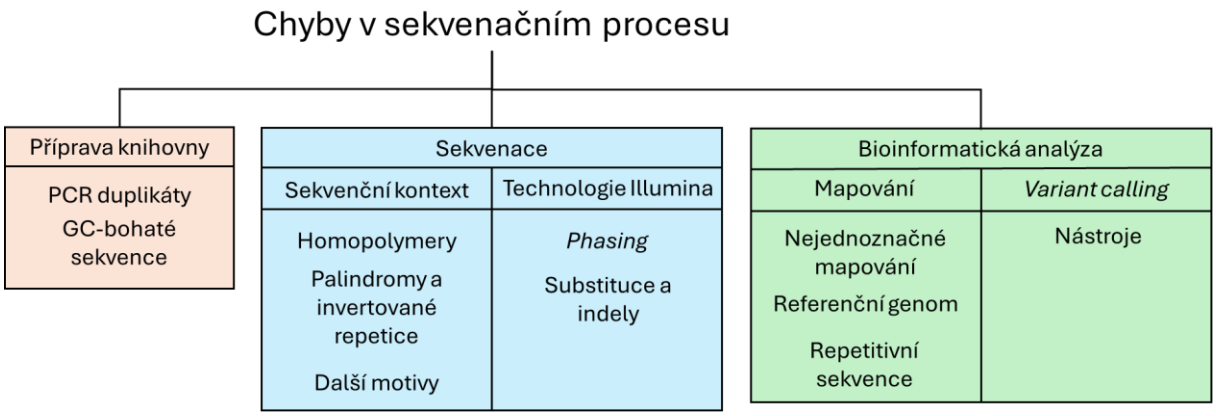
DNA sequencing represents a fundamental pillar of modern diagnostics for current needs of molecular medicine. Although today's next-generation sequencing (NGS) technologies enable efficient reading of the entire human genome, they are — from sample preparation to bioinformatic analysis — still affected by various types of errors. These errors can have serious consequences for the results of genetic testing.

This thesis focuses on systematic errors specific to the most widely used diagnostic NGS platform, the Illumina technology, in the analysis of germline DNA. It addresses errors arising from the sequence context, errors caused by the sequencing technology itself, errors associated with bioinformatic processing — including the mapping of short reads to the reference genome — and errors emerging during variant allele identification (variant calling).

The aim of this thesis is to contribute to a better understanding of the sources of these errors and to support the more accurate and safer application of NGS in the clinical setting. The results can serve as a reference framework for the selection of appropriate tools, methods, and parameters in the bioinformatic analysis of genomic data.

Keywords: Illumina, DNA sequencing, mapping, variant calling, errors

Grafický abstrakt



Seznam zkratek

bp – *base pair* – pár bází

BWT – *Burrows-Wheeler Transform* – Burrows-Wheelerova transformace

CNV – *Copy Number Variant* – variace počtu kopií

ddNTP(s) – dideoxynukleosid trifosfát(y)

DNA – deoxyribonukleová kyselina

dNTP(s) – deoxynukleosid trifosfát(y)

dsDNA – *double stranded DNA* – dvouvláknová DNA

HLA – Human leukocyte antigen

IGV – Integrative Genomics Viewer

indel – inzerce nebo delece

MAPQ – *mapping quality* – mapovací kvalita

mRNA – *messenger RNA* – mediátorová RNA

NGS – *next-generation sequencing* – sekvenování druhé generace

ONT – Oxford Nanopore Technologies

PacBio – Pacific Biosciences

PCR – *Polymerase Chain Reaction* – polymerázová řetězová reakce

RNA – ribonukleová kyselina

ssDNA – *single strand DNA* – jednovláknová DNA

TCEP – tris(2-karboxyetyl)fosfin

tRNA – transferová RNA

UMI – *unique molecular identifier* – jednoznačný molekulární identifikátor

VC – *Variant Calling*

Obsah

Úvod

Sekvenování nukleových kyselin je základním nástrojem k našemu pochopení smyslu genetického kódu. V posledních dekádách se stalo nepostradatelným nástrojem nejen ve výzkumu, ale i v klinické diagnostice. Díky prudkému rozvoji sekvenačních technologií lze dnes během několika dní přečíst celý lidský genom s vysokou přesností a za zlomkovou cenu oproti prvním sekvenačním projektům vrcholícím na přelomu milénia. Přes tento technologický pokrok však proces sekvenování není bezchybný.

Cílem této práce je systematicky identifikovat a popsat chyby, které vznikají v průběhu sekvenačního procesu DNA – od přípravy vzorku po bioinformatickou analýzu. Zvláštní důraz je kladen na technologii Illumina, která je v současnosti nejrozšířenější metodou sekvenování v klinické praxi. Analyzovány jsou chyby plynoucí ze sekvenčního kontextu (např. homopolymery nebo oblasti s extrémním zastoupením GC párů), dále chyby vznikající při mapování krátkých čtení na referenční genom (např. nejednoznačné mapování a nedostatky referenčního genomu) a při následné identifikaci genetických variant. Součástí práce je také srovnání s technologiemi Oxford Nanopore Technologies a Pacific Biosciences, které nabízejí alternativní přístup k sekvenaci.

V klinické praxi sekvenování DNA patří mezi klíčové nástroje genetické analýzy jedince pro identifikaci geneticky podmíněných onemocnění. Na světě jsou každý rok osekvenovány milióny lidí (Stephens et al., 2015) a pravděpodobnost chybné diagnózy není nulová – může být chybně diagnostikováno neexistující onemocnění, nebo naopak skutečné onemocnění může být zcela opomenuto. Obě situace mohou vést k vážným až fatálním důsledkům. Proto je zcela nutné porozumět chybám, které v sekvenačním procesu vznikají. Dobrý klinický genetik by měl vědět, jak některým chybám předejít, jak jiné identifikovat nebo jak je systematicky odstranit.

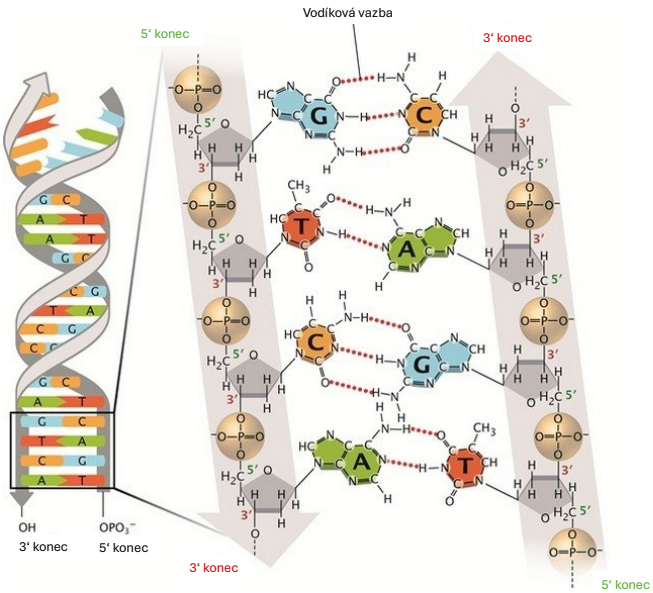
1 Teoretický úvod

1.1 Nukleové kyseliny

Nukleové kyseliny jsou lineární polynukleotidy složené z ribonukleotidů (RNA), respektive deoxyribonukleotidů (DNA), které jsou kovalentně spojeny fosfodiesterovou vazbou. Nukleotidy (a od nich odvozené přirozeně se vyskytující 2ˈ-deoxynukleotidy) jsou základními stavebními kameny nukleových kyselin, složené z nukleosidů s navázanými fosfátovými zbytky. Podle obsažené dusíkaté báze se dělí na purinové a pyrimidinové (Watson & Crick, 1953).

Purin je dusíkatá heterocyklická sloučenina, která sestává ze dvou aromatických dusíkatých heterocyklů – pyrimidinu a imidazolu. Mezi základní purinové báze se řadí adenin (A) a guanin (G). Mezi další typy purinových bází, které se mohou vyskytovat v genetickém materiálu, patří jejich prekurzory, modifikované báze či degradační produkty (např. hypoxanthin v transferové RNA – tRNA) (Christman, 1952; Schmalle et al., 1988).

Pyrimidin je aromatická sloučenina s jedním dusíkatým heterocyklem. Mezi základní deriváty pyrimidinů, které najdeme v nukleových kyselinách, patří thymin (T), cytosin (C) a uracil (U) (Christman, 1952). Dále existují další deriváty, které se přirozeně vyskytují



Obrázek 1 - Struktura DNA. Vlevo je znázorněna dvoušroubovice DNA. Vpravo je znázorněn detail chemické struktury nukleotidů, jejich vzájemných vazeb a standardního Watson-Crickova párování. Upraveno podle Pray (2008).

v buňce (např. pseudouracil nebo 5‘-metylcytosin v RNA) nebo mají další využití ve farmacii (X. Li et al., 2016; Rani et al., 2016).

Bází obecně nazýváme derivát purinu nebo pyrimidinu, který najdeme v nukleosidech. Nukleosid je sloučenina dusíkaté báze, která je N-glykosidovou vazbou kovalentně navázaná na C1’ ribózy (v RNA) nebo deoxyribózy (postrádá hydroxylovou skupinu na C2’ v DNA). U purinových bází vytváří tuto vazbu dusík N9, u pyrimidinových N1. Nukleotidy (obr. 1) jsou fosforylované formy nukleosidů, které mají na C5‘ ribózy/deoxyribózy navázány jeden až tři fosfátové zbytky. Na 5’ konci řetězce polynukleotidů je volný fosfát a na 3’ konci je volná OH skupina (Saenger, 1973).

Strukturu nukleových kyselin má hierarchické uspořádání. Primární struktura je sekvence (pořadí) nukleotidů v polynukleotidovém řetězci. Sekundární struktura zahrnuje v DNA dvoušroubovicovou strukturu antiparalelních řetězců polynukleotidů. Vyšší organizační struktury DNA zahrnují nejen samotný polynukleotid, ale také asociovaný nukleoproteinový komplex, složený z proteinových komplexů (včetně histonů a nehistonových proteinů). Vyšší organizační struktury v DNA vznikají především v důsledku kondenzace genetického materiálu při buněčném dělení.

Molekuly RNA jsou v principu jednořetězcové. Při výskytu komplementárních bází jsou schopny vytvářet antiparalelní šroubovicové sekundární struktury v rámci jednoho řetězce, nebo i dvou nezávislých řetězců RNA/RNA nebo i RNA/DNA. Terciární struktura RNA je 3D konformace, která vychází z interakcí mezi jednotlivými sekundárními strukturami (Batey et al., 1999). Kvarterní struktura RNA je soubor interagujících terciárních struktur a proteinů (Jones & Ferré-D’Amaré, 2015).

1.1.1 Struktura DNA

DNA je sestavena ze čtyř základních deoxynukleotidů – dAMP, dGMP, dCMP a dTMP. Pro genomovou DNA platí, že v důsledku Chargaffova pravidla je poměr purinových a pyrimidinových bází 1:1 (Chargaff et al., 1952; Zamenhof et al., 1950).

Za fyziologických podmínek v živých organismech je sekundární struktura DNA obvykle organizována do pravotočivé dvoušroubovice, což určuje většinu jejích vlastností (jak biologických, tak chemických a fyzikálních). DNA je však velmi flexibilní a dynamická

struktura, která má mnoho forem, mezi kterými může přecházet (Travers & Muskhelishvili, 2015). Dvoušroubovice DNA se skládá ze dvou antiparalelních řetězců, orientovaných ve směru 3 5'–3' (obr. 1). Řetězce DNA ve dvoušroubovici interagují pomocí vodíkových můstků. V nejčastějším – Watson-Crickově – párování jsou vodíkové můstky formovány mezi A a T (dva vodíkové můstky) a G a C (tři vodíkové můstky) (Watson & Crick, 1953).

1.1.2 Nukleové kyseliny v buňce

DNA je nositelkou genetické informace v buňce. V buněčném jádře je v rámci buněčného dělení replikovaná pomocí DNA-dependentních DNA polymeráz. Replikace začíná v replikačních počátcích, kam nasednou enzymy helikázy, které rozplétají dvoušroubovici DNA (dsDNA – *double stranded DNA*) na dvě vlákna jednovláknové DNA (ssDNA – *single-stranded DNA*) v dynamické struktuře replikační vidličky. Aby mohla DNA polymeráza zahájit replikaci, musí být syntetizovány *primery*, což jsou krátké RNA oligonukleotidy, syntetizované pomocí enzymu primázy (DNA-dependentní RNA polymerázy). Na hybridní dsDNA/RNA molekulu nasedá DNA polymeráza a podle templátu ssDNA (DNA-dependentní) polymeráza syntetizuje nové komplementární vlákno ssDNA ve směru 5‘-3‘. Replikace DNA je semikonzervativní – každá nově vzniklá dvoušroubovice obsahuje jedno původní a jedno nově syntetizované vlákno (O’Donnell et al., 2013).

RNA je transkribována dle DNA templátu pomocí komplexu DNA-dependentní RNA polymerázy. V eukaryotických buňkách geny obsahují exony (kódující sekvence) a introny (nekódující). Geny jsou transkribovány do pre-mRNA (mediátorové RNA), ale během *splicingu* jsou introny vystřiženy a v mRNA zůstanou jenom sekvence exony. U prokaryot se často transkribují operony, skupiny genů sdílející jeden promotor, jejichž exprese je regulována společně (Lee & Young, 2000).

Po transkripci je mRNA rovněž opatřena čepičkou a poly-A úsekem na 5’ konci finálního transkriptu a následně transportována do cytoplazmy, kde slouží v ribozomech jako templát pro translaci do proteinů. Sekvence mRNA je při translaci čtena po trojicích nukleotidů, tzv. kodónech. Každý kodón specifikuje jednu aminokyselinu nebo signalizuje začátek či konec translace. Ribozom při translaci využívá molekuly tRNA, které přinášejí

specifické aktivované aminokyseliny odpovídající jednotlivým kodónům, a tím dochází k syntéze proteinu s přesně definovanou sekvencí aminokyselin (Zhang et al., 2023).

DNA je neustále vystavována vnějším a vnitřním vlivům, které mohou indukovat poruchy její struktury způsobující mutace či varianty. Mezi vnější mutageny patří fyzikální (např. ionizující a neionizující záření), chemické (např. látky modifikující báze, interkalační činidla) a biologické (např. viry). Vnitřní mutagenní pochody zahrnují působení endogenních procesů j (např. chyby při replikaci, oprava DNA, oxidativní poškození). Z hlediska ontogeneze můžeme poruchy DNA rozdělit na germinální (pocházející ze zárodečných buněk) a somatické (vznikající v průběhu života v buňkách tkání). Germinální mutace jsou nejčastěji děděny od rodičů (kombinace variant obou rodičů) a všechny buňky jedince jsou přenašeči. (Beichman et al., 2024). Dědičné varianty DNA z pohledu klasické genetiky dělíme na mutace (varianty s frekvencí <1 % v populaci) a polymorfismy (varianty s vyšší frekvencí). Zatímco významná pravděpodobnost ovlivnění fenotypu přítomností polymorfismu je zanedbatelná, mutace mohou mít na výsledný fenotyp vliv značný. Efekty mutací lze rozdělit do tří základních kategorií: 1) benigní – mutace, které nezpůsobují onemocnění, ale mohou ovlivnit fenotyp (vnější pozorovatelné znaky); 2) patogenní – mutace, které způsobují onemocnění; 3) letální – mutace, které nejsou slučitelné se životem (smrt nastává před porodem nebo v raném věku). Bodové mutace (mutace, které ovlivní jednu bázi v sekvenci) rozdělujeme na synonymní (mutace nezmění translatovanou aminokyselinu), nesynonymní (mutace změní translatovanou aminokyselinu), *nonsense* (mutace změní původní kodón na STOP kodón), *nonstop* (mutace změní STOP kodón na jiný než STOP). Inzerce nebo delece nukleotidu, jejichž délka není násobkem tří, způsobují posun čtecího rámce (angl. *frameshift*), jenž může ovlivnit translaci podobně jako bodové mutace (Gorlov et al., 2018; Mignogna et al., 2022).

1.2 Sekvenace nukleových kyselin

Sekvenování nukleových kyselin je metoda sloužící k určení primární struktury DNA nebo RNA. V klinické diagnostice se tyto analýzy většinou zabývají identifikací variant a určením jejich dopadu na fenotyp (např. onemocnění) vyšetřované osoby. Proces

sekvenování je vždy rozdělen nejdříve na přípravu sekvenační knihovny (příprava vzorku) a až pak následuje sekvenační reakce.

1.2.1 Sangerovo sekvenování

První všeobecně využívanou metodu sekvenování DNA vyvinul Sanger a jeho spolupracovníci (Sanger & Coulson, 1975). Sangerovo sekvenování patří mezi technologie první generace sekvenování. V přibližně stejnou dobu Maxam a Gilbert vynalezli alternativní sekvenační technologii založenou na specifické chemické modifikaci bází a jejich následném štěpení (Maxam & Gilbert, 1977). Tato metoda nebyla tak dobře škálovatelná jako Sangerova, tudíž se od jejího využití upustilo.

Sangerovo sekvenování lze rozdělit na čtyři kroky: příprava, syntéza, terminace a čtení DNA. Pro sekvenační reakci je potřeba amplifikovaná DNA, *primer*, termostabilní DNA-dependentní DNA polymeráza, pufr, dNTPs (2‘-deoxynukleotid trifosfáty) a ddNTPs (2‘, 3‘-dideoxynukleotid trifosfáty) s fluoroforem. V ddNTP tedy chybí OH skupina na 3‘ uhlíku, tudíž se na tento nukleotid již nemůže navázat další nukleotid.

Amplifikace DNA je prováděna metodou PCR (polymerázová řetězová reakce, angl. *Polymerase Chain Raction*) (Mullis et al., 1986). PCR probíhá v cyklech, přičemž každý cyklus obsahuje tři kroky: 1) denaturace (při 94-98 °C) – DNA je denaturována na dvě ssDNA vlákana; 2) hybridizace (při 50–65°C) – na každé vlákno ssDNA se naváží krátké *primery* (*forward* a *reverse primery*) ohraničující z obou stran amplifikovaný úsek DNA, které jsou přítomné v extrémním nadbytku; 3) polymerace (při 72°C) – termostabilní Taq polymeráza nasedne na úsek s *primerem* a syntetizuje nové vlákno v 5‘-3‘ směru (Kuno, 1998). Tento cyklus je opakován do dosažení vhodného množství nukleových kyselin (typicky 30–35krát dle množství vstupní DNA).

Při Sangerově sekvenování se využívá jen jediný *primer* pro postupnou inkorporaci nukleotidů jenom v jednom směru. V momentě, kdy se místo dNTP inkorporuje do prodlužujícího se řetězce ddNTP, syntéza skončí. dNTP je v roztoku 10krát až 300krát více než ddNTP. Délky syntetizovaných fragmentů jsou tak náhodné a v ideálním případě rovnoměrně rozdělené. Čtení DNA v moderních sekvenátorech probíhá pomocí kapilární elektroforézy (původně pomocí gelové elektroforézy), kde se terminované fragmenty

s ddNTP rozdělují podle velikosti a následně analyzují pomocí laserové fluorescence (Huang et al., 1992).

Sangerovo sekvenování bylo použito k získání první kompletní sekvence lidského genomu, která byla publikována v roce 2001 (International Human Genome Sequencing Consortium, 2001). S nástupem druhé generace sekvenování poklesl význam Sangerova sekvenování, nicméně metoda se hojně používá do současnosti pro účely přesné identifikace malých úseků DNA (s délkou kolem 1000 bp) (Kopernik et al., 2025).

1.2.2 Sekvenování druhé generace

Technologie sekvenování druhé generace (NGS) přináší řadu výhod oproti Sangerově metodě. Hlavní výhoda je možnost masivně paralelního sekvenování, tedy paralelní analýzy tisíců až milionů různých fragmentů DNA v rámci jednoho sekvenačního běhu. Sangerovou metodou může být v rámci jedné reakce analyzován pouze jeden fragment DNA s délkou maximálně kolem 1 kbp. První lidský genom dokončený v roce 2001 trvalo osekvenovat přes deset let (International Human Genome Sequencing Consortium, 2001). Pomocí druhé generace sekvenování lze analyzovat celý lidský genom v řádech dnů (McCombie et al., 2019).

Mezi technologie druhé generace sekvenování se řadí Illumina (Bentley et al., 2008), Roche 454 (Margulies et al., 2005), SOLiD (McKernan et al., 2009) a Ion Torrent (Rothberg et al., 2011). Technologie Illumina se stala dominantní díky kombinaci vysoké přesnosti, kapacity a nízké ceny (viz tab. 1). Navzdory pokroku v sekvenování zůstává Illumina *de facto* standardem v klinické diagnostice, kde je sekvenování využíváno pro analýzu nejen lidského genomu, a proto se tato práce věnuje zejména této technologii.

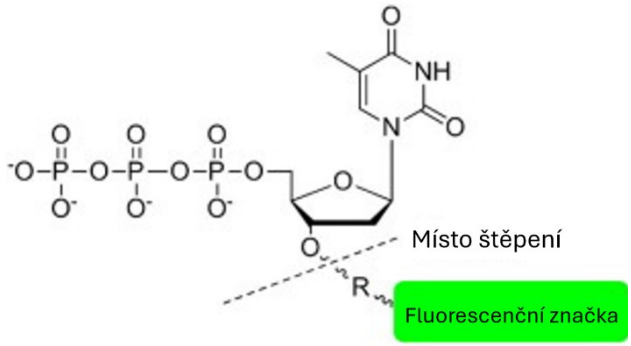
1.2.3 Illumina

Sekvenační technologie Illumina je založena na principu sekvenování syntézou (angl. *sequencing by synthesis*) (Bentley et al., 2008). Při přípravě sekvenační knihovny je DNA nejdříve fragmentována – sonikací nebo restrikčními enzymy (Soukupova et al., 2018). Na oba konce fragmentů jsou následně navázány ligázou dvojice oligonukleotidů s *primery* (na každém konci jsou oba typy oligonukleotidů, které nejsou vzájemně

komplementární). Ty jsou následně linearizovány pomocí PCR amplifikace. Nakonec vznikne fragment, jehož schematické znázornění je na obrázku 3b.

Na začátku sekvenačního běhu jsou tyto fragmenty denaturovány a vloženy na *flow cell*, která má na svém povrchu komplementární sekvence buď k P5 nebo P7 (také označovány jako adaptéry, viz obr. 3), na které jednotlivá vlákna nasednou. Na povrchu *flow cell* probíhá můstková amplifikace, při níž se jednotlivá vlákna DNA navazují na komplementární sekvence adaptérů P5 a P7. Tento proces vytváří můstkové struktury DNA, které jsou následně amplifikovány. Pokud se sekvenuje sekvence od P5 k P7 (použije se *primer* pro čtení 1 – angl. *read*), tak se můstek přeruší štěpením adaptéru P5; pokud se sekvenuje opačným směrem (čtení 2), tak se můstek přeruší štěpením adaptéru P7. Nakonec jsou vlákna denaturována na jednořetězcovou DNA a fragmenty, které nejsou navázány na *flow cell*, jsou odmyty.

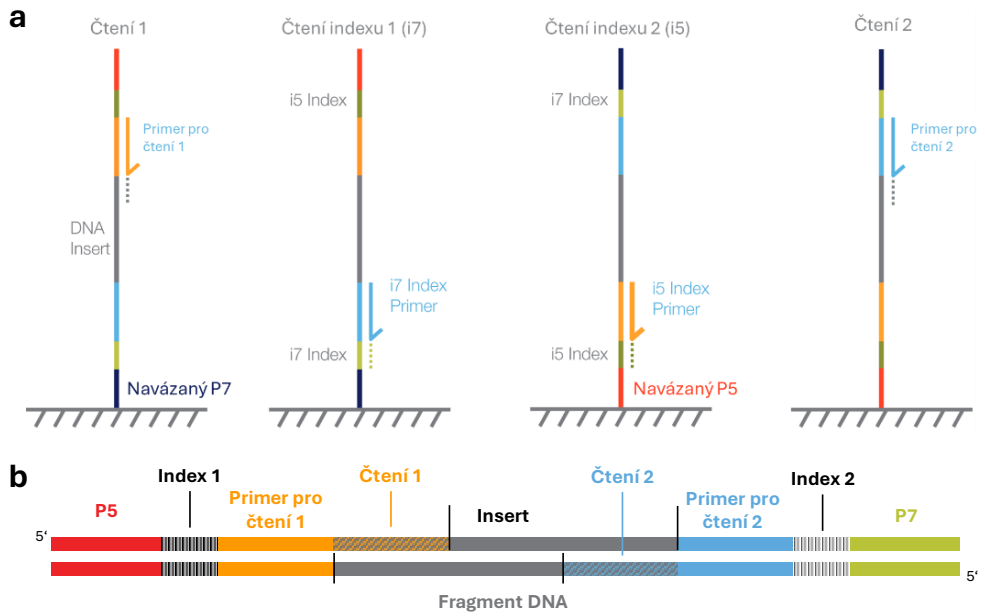
Pro sekvenování DNA musí na hybridizovaný *primer* nasednout DNA-dependentní DNA polymeráza, která v cyklech inkorporuje upravenou bázi (3‘-O-azidomethyl-2‘-deoxynukleotidtrifosfát) se specifickým fluoroforem pro každou bázi (viz obr. 2). Modifikace na 3‘nukleotidu zabraňuje náhodné inkorporaci dalšího nukleotidu . Na *flow cell* je namířen laser, který excituje fluorofor, čímž se identifikuje právě inkorporovaná báze. TCEP (tris(2-karboxyetyl)fosfin) odstraní 3‘ fluorofor a azidomethyl obnoví 3‘ hydroxylovou skupinu (místo štěpení na obr. 2), což umožní navázání dalšího nukleotidu v následném cyklu. Cyklus se pak opakuje 75krát nebo 150krát (pro čtení délky 75bp-150bp) (Illumina, b.r.-b). Pro párové sekvenování (angl. *paired-end*), kdy se sekvenuje fragment DNA z obou konců, je třeba znovu nasyntetizovat jedno vlákno pomocí *primeru* pro čtení 1, toto vlákno se komplementárně naváže pomocí P7 adaptéru



Obrázek 2 – Deoxythymidintrifosfát s upraveným 3' koncem a fluoroforem. Po inkorporaci a excitaci fluoroforu (fluorescenční značky) je pomocí TCEP fluorofor odštěpen a 3‘ OH skupina je obnovena, aby se na ni mohl navázat další nukleotid. Upraveno podle Kim et al. (2014).

a původní vlákno se po denaturaci odmyje. Nesekvenovaná část mezi čtením a adaptérem se nazývá *insert* (viz obr. 3b). Na *primer* pro čtení 2 nasedne DNA polymeráza a proces se opakuje.

Sekvenci bází, které identifikuje sekvenační analyzátor na základě emisí světla z fluoroforů, se nazývá čtení. Každé bázi je přiřazeno skóre, které vyjadřuje její kvalitu. Počet čtení, které bylo přiřazeno jedné bázi v referenčním genomu se popisuje jako sekvenační hloubka (angl. *depth*) a průměrný počet čtení, které byly přiřazeny všem bázím referenčního genomu je pokrytí (angl. *coverage*) (International Human Genome Sequencing Consortium, 2001). Díky indexům v navázaných oligonukleotidech (obr. 3) je možné najednou (v různých *flow cell*, ale ve stejném přístroji) sekvenovat fragmenty DNA z více genomů (jedinců), které jsou při následné analýze ke konkrétnímu jedinci přirazeny na základě indexů (Shokralla et al., 2015). *Primery* pro čtení 1 a 2 také fungují jako *primery* v opačném směru pro samotné indexy (viz obr. 3a). Sekvenování indexů je prováděno až po sekvenování fragmentů DNA.



Obrázek 3 – **a)** Schéma sekvenování DNA fragmentu z obou stran s vyznačenými *primery*. Během čtení 1 je fragment na *flow cell* navázán pomocí P7. Po dočtení sekvenovaného vzorku je pomocí i7 index *primeru* (stejného jako pro čtení 2) sekvenován i7 index. Při čtení z opačné strany je postup stejný, jenom je fragment navázán pomocí adaptéru P5, používá primer pro čtení 2 a i5 index použije vazebné místo pro primer pro čtení 1 z druhé strany. Upraveno podle Illumina (2020). **b)** Schéma fragmentu DNA pro sekvenování pomocí technologie Illumina. Na obou koncích fragmentu jsou P5 a P7 adaptéry pro vazbu na flow cell. Za nimi jsou indexy, které identifikují vzorek při následné analýze, což umožňuje v jednom přístroji naráz sekvenovat vzorky z více jedinců. Nakonec je v rámci navázaného oligonukleotidu vazebné místo pro primer jak pro syntézu fragmentu, tak pro syntézu indexu v opačném směru. Mezi oligonukleotidy je samotný sekvenovaný vzorek (fragment DNA), který je většinou 150 bp až 500 bp dlouhý. Čtení jsou většinou kratší než celý fragment, a tudíž vznikne *insert* – nesekvenovaná část fragementu. Upraveno podle Illumina (2024).

1.2.4 Technologie třetí generace sekvenování

Mezi technologie třetí generace sekvenování se řadí technologie Oxford Nanopore Technologies (ONT) (Clarke et al., 2009; H. Lu et al., 2016) a Pacific Biosciences (PacBio) (Rhoads & Au, 2015). Tyto technologie se vyznačují zejména delšími čteními, které mohou dosahovat až milionu bází. Dále se vyznačují řádově vyšší chybovostí oproti technologii Illumina. Jak ONT, tak PacBio mají chybovost přibližně 0,5-2 %. Sekvenační technologie PacBio HiFi umožňuje čtení velmi dlouhých úseků (v průměru 13,5 tisíc bází) a zároveň vykazuje vysokou míru přesnosti (99,95 %) (PacBio, b.r.; Wenger et al., 2019). Rozšířené porovnání všech generací sekvenování je v tabulce 1.

ONT a PacBio mají výrazně jiný způsob detekování jednotlivých bází než Illumina. ONT sekvenování využívá proteinové nanopóry, skrze které prochází jednovláknová DNA. Elektrický signál generovaný změnami iontového toku přes nanopór je dekódován a přiřazen odpovídajícím nukleotidovým sekvencím (Clarke et al., 2009). PacBio SMRT sekvenování využívá cirkularizované DNA templáty a dlouhodobé sledování polymerázové aktivity v mikroskopických jamkách. Tato metoda umožňuje opakované čtení stejného fragmentu DNA, čímž se zvyšuje přesnost sekvenování (Rhoads & Au, 2015).

Tabulka 1 – Porovnání všech generací sekvenačních technologií podle výhod (+) a nevýhod (-) (Complete Genomics, 2024; Illumina, b.r.-a; NC State University, 2024; PacBio, b.r.; Whiteford, 2022).

Generace sekvenační technologie	**Přesnost**	**Délka čtení**	**Kapacita**	**Finanční náročnost**	**Další**
1. generace (Sanger)	99,95 % (+)	Až 1kbp (+)	Nízká (-)	$2500 – $8000 / Mbp (-)	Časově náročné (-)
2. generace (Illumina)	99,9 % (+)	75bp – 150bp (-)	Až 16Tb / běh (+)	$5 – $30 / Gbp (+)	Většinou vyžaduje PCR (-)
3. generace (ONT, PacBio)	98-99,5 % (-)	Až Mbp (+)	50 – 500Gb / běh (-)	$9 – $500 / Gbp (-)	Není potřeba PCR (+)

2 Sekvenační chyby na základě sekvenčního kontextu

Ke vzniku chyb v určení primární sekvence analyzované DNA může dojít na všech úrovních analýzy. Sekvenování je přirozeně závislé již jen na kvalitě analyzované DNA, která je ovlivněna celou řadou parametrů a pre-analytických kroků při získávání sekvenačního templátu. Kvalita vstupního genetického materiálu je ovlivněna stářím analyzované DNA. Dalším faktorem je heterogenita vzorku nukleových kyselin, která v případě analýzy germinální DNA může být komplikována přítomností DNA z klonální hematopoézy či nerozpoznané (pre)maligní cirkulující populace (Nix et al., 2020). Tyto "pre-analytické" chyby však nejsou předmětem této práce.

2.1 Substituce vs. inzerce/delece

Jedním z nejznámějších typů chyb, které se vyskytují v technologii Illumina, jsou tzv. substituční chyby, tedy situace, kdy je daná báze přečtena nesprávně (Stoler & Nekrutenko, 2021). Druhým typem chyb jsou inzerce/delece (indel), kdy se buď chybně vloží báze, která do sekvence nepatří, anebo báze není přečtena vůbec.

Nejčetnější substitucí na platformě Illumina je substituce A>C (Dohm et al., 2008; Ross et al., 2013; Schirmer et al., 2015; Stoler & Nekrutenko, 2021). Další časté substituce jsou znázorněny v tabulce 2. Kromě těchto substitucí bylo také ukázáno, že substituce závisí i na bázi, která předchází substituci (nejčastěji G nebo A) (Stoler & Nekrutenko, 2021), a že nejvíce náchylné na substituce je G a T (Schirmer et al., 2016). Literatura (Tab. 2) se shoduje jen v jedné substituci (A>C) a ostatní četné substituce mají velmi malý nebo žádný překryv mezi různými pracemi.

Dle Jeon a jeho spolupracovníků jsou tranziční substituce (purin>purin nebo pyrimidin>pyrimidin) častější než transverzní (purin>pyrimidin nebo pyrimidin>purin) (Jeon et al., 2021). Kdežto ostatní autoři (Dohm et al., 2008; Ross et al., 2013; Schirmer et al., 2015; Stoler & Nekrutenko, 2021) k tomuto závěru nedošli, většina substitucí, které identifikovali jsou transverzní substituce (tab. 2). Vznik substitučních chyb tudíž pravděpodobně nezávisí pouze na úzkém sekvenčním kontextu, ale také na dalších vlivech, jako jsou rozdílné způsoby přípravy knihovny, rozdílné sekvenační metody různých produktů Illumina a další.

Indel chyby jsou méně časté na platformě Illumina[1] (Schirmer et al., 2016). Na platformách ONT a PacBio převládají nad ostatními typy sekvenačních chyb, což patrně odráží sekvenační technologii náchylnou na nuance v kontinuálně detekovaném signálu (Goodwin et al., 2016; Karst et al., 2021).

Tabulka 2 - Přehled nejčastějších substitucí. V horním indexu všech substitucí je vyznačen typ – transverzní substituce (TV) nebo tranziční substituce (TI).

Substituce	Četnost	Literatura
$A>C^{TV}$	*****	Dohm et al. (2008), Ross et al. (2013), Stoler a Nekrutenko (2021), Schirmer et al. (2015)
$A>G^{TI}$	****	Stoler a Nekrutenko (2021), Jeon et al. (2021)
$A>T^{TV}$	***	Stoler a Nekrutenko (2021)
$T>G^{TV}$	****	Ross et al. (2013)
$T>C^{TI}$	***	Schirmer et al. (2015), Jeon et al. (2021)
$G>T^{TV}$	****	Dohm et al. (2008), Ross et al. (2013), Schirmer et al. (2015)
$G>A^{TI}$	**	Jeon et al. (2021)
$C>A^{TV}$	**	Schirmer et al. (2015)
$C>T^{TI}$	**	Jeon et al. (2021)

2.2 Homopolymery

Homopolymery (polynukleotidové sekvence složené z identických bází) jsou dalším typem sekvence, která je pro sekvenování obtížná. Homopolymery v genomové DNA mají většinou strukturní funkci při organizaci chromatinu (Dechering et al., 1998). Dále mohou mít regulační funkci v rámci transkripčních faktorů (de Oliveira Martins et al., 2022). Homopolymery mají většinou nižší pokrytí než jiné sekvence, ale G/C-homopolymery mývají ještě nižší pokrytí, než A/T-homopolymery (Modlin et al., 2021).

[1] Substituční chyby se objevovaly s přibližnou frekvencí 0,005-0,01 na bázi, indel chyby se objevovaly s přibližnou frekvencí $2{,}8\times10^{-6}$-$5{,}1\times10^{-6}$. Indel chyby jsou tedy přibližně tisíckrát až deset tisíckrát méně časté.

Se zvyšující se délkou homopolymeru (2 až 8 bází dlouhé) vzniká vyšší pravděpodobnost substitučních a indel chyb způsobujících chybné určení délky homopolymeru. Substituční chyby v homopolymerech jsou častěji generovány na platformách Illumina, indel chyby častěji na platformě ONT (Chen et al., 2024; Karst et al., 2021). G/C-homopolymery jsou k těmto chybám více náchylné než A/T-homopolymery (Shin & Park, 2016). Ross et al. (2013) dále ukázali, že na platformě Illumina mají delší homopolymery (až 20bp) vyšší chybovost na rozdíl od kratších homopolymerů (kratší než 10bp). Na platformě PacBio byla chybovost velmi podobná napříč různými délkami čtení. Relativní chybovost substitučních chyb na obou platformách byla přibližně stejná (10^{-2} na bázi).

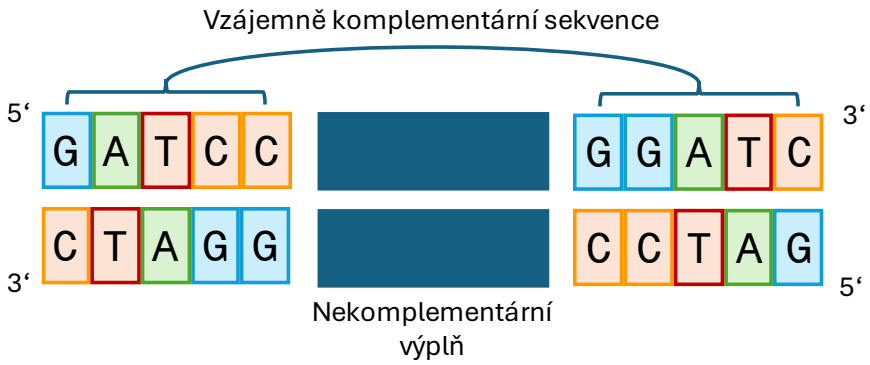
Často se také objevují chyby, které vznikají těsně za homopolymerem. Zde vznikne substituční chyba, kdy je „přečtena“ báze jako báze homopolymeru (skutečná sekvence: GGGGAC, přečtená sekvence: GGGGGC). Tyto chyby se vyskytují také zejména u G/C-homopolymeru a méně častěji i u A/T-homopolymerů (Stoler & Nekrutenko, 2021).

V repetitivních regionech genomu, jako jsou homopolymery nebo také tandemové repetice může během jejich replikace sklouznout DNA polymeráza. Po sklouznutí dochází ke vzájemnému “posunutí” nascentního a templátového vlákna v repetitivní oblasti jehož výsledkem je buď inkorporace jedné nebo několika repetic navíc (inzerce) nebo naopak jejich chybění (delece).Sklouznutí DNA polymerázy je pravděpodobně hlavní důvod pro vznik chyby v homopolymerních oblastech (Levinson & Gutman, 1987; Murat et al., 2020; Shinde et al., 2003).

2.3 Palindromy a invertované repetice

Sekvence, jejichž reverzní komplement je identický se samotnou sekvencí, jsou nazývány palindromy. Palindromy, které mají přesně v půlce své sekvence vloženou libovolně dlouhou mezerníkovou sekvenci, se označují jako invertované repetice (viz obr. 4) (Warburton et al., 2004). Tyto sekvenční motivy často zapříčiňují vznik různých sekundárních struktur jako jsou vlásenky (L. Lu et al., 2007).

Kvůli možnosti indukovat sekundární struktury mohou invertované repetice inhibovat elongaci nukleotidů nejen během sekvenování, ale také během PCR (Nakamura et al.,



Obrázek 4 - Schéma invertované repetice. Barevně vyznačené báze na koncích tvoří palindrom – na 5‘ a na 3‘ koncích obou sekvencí jsou identické sekvence. V rámci jednoho vlákna DNA je posledních 5 bází na obou koncích vzájemně komplementární. Mezi vzájemně komplementárními sekvencemi je nekomplementární výplň, která může být libovolně dlouhá. Pokud je dlouhá alespoň jednu bázi tak může na denaturovaném vlákně vzniknout vlásenka, která využije vzájemně komplementárními sekvence pro navázání. Výplň na sebe sama navázána nebude, jelikož se jedná o nekomplementární úsek. Autorovo vlastní dílo.

2011). Tudíž mohou tyto sekvence být jak nedostatečně zastoupené v sekvenovaném vzorku, tak i ve výsledku sekvenování.

2.4 Vliv obsahu GC na pokrytí

Sekvence, které jsou bohaté na GC páry (60 % a více) mají nerovnoměrné pokrytí – oproti průměrnému pokrytí celé sekvence. V eukaryotním genomu se obecně GC bohatší úseky vyskytují v exonech, zatímco oblasti s vyšším zastoupení AT párů se nacházejí v intronech a intergenových sekvencích. Sekvence bohaté, na GC páry mají relativně nižší pokrytí než jiné části sekvence (Barbitoff et al., 2022; Dohm et al., 2008; Hillier et al., 2008; Modlin et al., 2021; Ross et al., 2013). Nižší pokrytí CG-bohatých úseků DNA je nejpravděpodobněji způsobeno sníženou efektivitou amplifikace během PCR, která je zapříčiněna vyšší teplotou denaturace dsDNA v GC-bohatých oblastech (Benjamini & Speed, 2012; Dohm et al., 2008; Hillier et al., 2008; Modlin et al., 2021; Panjkovich & Melo, 2005; Ross et al., 2013). Platformy, které nepotřebují PCR amplifikaci (např. PacBio či ONT), nemají pokrytí ovlivněné relativním množstvím GC párů (Browne et al., 2020). Pokud se při přípravě knihovny pro platformu Illumina vynechá krok amplifikace pomocí PCR, výrazně se tím sníží riziko nerovnoměrného pokrytí (Kozarewa et al., 2009). Ross et al. (2013) ukázali, že technologie PacBio je výrazně lepší při čtení GC-bohatých i GC-chudých oblastí, protože technolgie Illumina má přibližně padesátkrát horší pokrytí GC-bohatých sekvencí než PacBio a pokrytí GC-chudých bylo jen desetkrát horší než na PacBio. Výkyvy v pokrytí

těchto sekvencí má velký dopad na mapování a *de novo assembly*[2], kde se kvůli špatnému pokrytí těžce identifikují opakující se sekvence (Kozarewa et al., 2009).

2.5 Další zdroje sekvenačních chyb vyplývající ze sekvenčního kontextu

Kromě dobře definovatelných motivů jako jsou homopolymery, mohou sekvenování na platformě Illumina komplikovat i jiné druhy motivů (tabulka 3). Z výsledků řady pozorování (Tab. 3) plyne, že mnoho takových motivů začíná guanosinem nebo obsahují „GG“, „GGT“ potažmo „GGC“, a že přibližně 30 % chyb vzniká, když před chybovou pozicí je guanosin (Dohm et al., 2008). Výsledky výzkumu se shodují jen v některých kratších a jednodušších motivech. Identifikace dalších motivů (mimo GG atp.) mohou být zapříčiněny artefakty způsobenými malým množstvím zkoumaných vzorků i přes jednotnou statistickou analýzu napříč studiemi. Motivy „AAA“ a „TTT“ jsou pravděpodobně zmíněny jen Schirmerem et al. (2016) protože indel chyby jsou řádově méně časté než substituční chyby (viz kapitola 2.2). Většina zmíněných motivů je GC-bohatá, což může být další z důvodů, proč se u těchto motivů vyskytují chyby.

Na platformě PacBio byl ukázán motiv generující chyby „CCDG“ a „GRTRA“ a na ONT motivy „AAAAADD“ nebo „RGTGVTA“ (Nasrin & Rahman, 2019). Tyto motivy jsou podobném těm, které přímo indukují indel chyby na platformě Illumina (Tab. 3).

Chyby se také častěji vyskytují na koncích krátkých čtení. Zejména tomu tak je pro 3‘ konce kvůli tzv. *phasing* efektu (jenom u technologie Illumina). Čtení fragmentu je mimo fázi se správným pořadím inkorporace nukleotidů. Pokud se v rámci cyklu neodštěpí blokující molekula, nemůže dojít k inkorporaci dalšího nukleotidu. V rámci jednoho cyklu se také mohou chybně inkorporovat dvě a více bází najednou, čímž se čtení některých nukleotidů přeskočí nebo bude zároveň excitováno více fluoroforů. V malém množství případů může docházet i ke kombinacím zmíněných defektů. Tento jev se sice děje s malou pravděpodobností, ale s prodlužujícím se čtením se pravděpodobnost podobných chyb zvyšuje, což je jeden z faktorů limitující délku čtení platformou Illumina

[2] *Assembly* je proces sestavení delší sekvence nebo celého genomu na základě čtení bez využití předešlých znalostí jiných referenčních genomů.

(Dohm et al., 2008; Schirmer et al., 2015; Star et al., 2014). Bylo pozorováno, že se chyby vyskytují blízko sebe (vedle sebe nebo mají mezi sebou jednu další bázi) nebo na stejných čteních (Dohm et al., 2008; Stoler & Nekrutenko, 2021).

Tabulka 3 - Motivy, které indukují chyby při sekvenování pomocí technologie Illumina.

Motiv	**Efekt motivu**	**Literatura**
GGT, CGT, AGT, NGGT, CTGRH, GGYRR	Asociace s chybou	Stoler a Nekrutenko (2021), Allhoff et al. (2013), Nasrin a Rahman (2019)
GGC, GGT	Chyba *downstream* od motivu	Nakamura et al. (2011), Shin a Park (2016)
GG, GGT, ACGGCGGT, GTGGCGGT	Chyba nastane bezprostředně po motivu	Meacham et al. (2011), Allhoff et al. (2013)
AAA, TTT	Způsobuje indel chybu	Schirmer et al. (2016)
GGG, CGG, AGG	Způsobuje substituční chybu	Schirmer et al. (2016)

3 Chyby vznikající při bioinformatické analýze

3.1 Mapování na referenční genom

3.1.1 Alignment

Nejstarší používaný algoritmus pro *pairwise alignment* (zarovnání dvou sekvencí, obr. 5b) je Needleman-Wunsch (Needleman & Wunsch, 1970). Tento algoritmus hledá optimální globální *alignment*. Obě sekvence zarovná tak, aby všechny báze z jedné sekvence byly přiřazeny k bázi z druhé sekvence nebo k mezeře. Na podobném principu je založen algoritmus Smith-Waterman(-Gotoh) (Gotoh, 1982; Smith & Waterman, 1981), který hledá optimální lokální *alignment*. Algoritmus najde jenom krátké podsekvence, které mají vysokou podobnost a nesnaží se zarovnat celé sekvence. Optimálnost *alignmentu* hodnotí skórovací funkce, která vychází ze skórovací matice. Algoritmus je založen na principu dynamického programování. Algoritmus postupně od začátku do konce prochází obě sekvence a u toho si zapisuje do tabulky skóre, které přiřadil všem dvojicím bází (Smith & Waterman, 1981).

Skórovací funkce rozhoduje o skóre, které přiřadí každé dvojici bází z obou sekvencí. Vstupem funkce jsou dvě báze, skóre předešlých dvojic bází z tabulky, skórovací matice (obr. 5a) a srážka za mezeru (*gap penalty*). Výstupem je optimální (maximální) skóre pro danou dvojici. Skórovací matice určuje, jak má skórovací funkce ohodnotit specifickou dvojici bází. Nejjednodušší matice obsahuje pouze hodnoty, které popisují shodu (*match*) a neshodu (*mismatch*) dvou bází. Složitější skórovací matice mají hodnoty ke všem dvojicím bází, které jsou specifické pro porovnávané organismy (Chiaromonte et al.,

a

	A	C	G	T
A	1	-2	-2	-2
C	-2	1	-2	-2
G	-2	-2	1	-2
T	-2	-2	-2	1

b

											Celkem
Sekvence 1	A	A	G	T	T	-	-	T	T	G	-
Sekvence 2	A	A	C	T	T	T	C	T	T	G	-
Skóre dle matice	1	1	-2	1	1			1	1	1	5
Srážka za mezeru						-3	-1				-4
											1

Obrázek 5 - Příklad skórovací matice a alignmentu. **a)** Skórovací matice pro porovnání dvou bází; shoda: 1, neshoda: -2. **b)** Alignment dvou sekvencí s jednou substitucí a mezerou o délce dva. Substituce je penalizována dle matice. Srážka za mezeru je počítána následovně; začátek mezery: -3 a prodloužení mezery: -1. Výsledné skóre je součet skóre dle matice a srážek za mezery, zde 5-4=1. Autorovo vlastní dílo.

2001). Výsledkem *alignmentu* je optimální skóre, které vyjadřuje počet přesných shod bází, počet neshod bází a počet mezer.

Skórovací funkce mohou používat tři různé typy srážek za mezeru. Nejjednodušší typ je konstantní srážka za každou mezeru bez rozlišení její délky. Další typ je lineární srážka – za každou zařazenou mezeru místo báze (Needleman & Wunsch, 1970). Afinní srážka (obr. 5b) je momentálně nejvyužívanější a biologicky nejvěrnější (Marco-Sola et al., 2023). Tento typ různě penalizuje začátek mezery (*gap-open*) a prodloužení mezery (*gap-extend)* (Altschul, 1998; Altschul & Erickson, 1986; Gotoh, 1982).

3.1.2 Mapování

Cílem mapování je přiřadit co nejvíce čtení na konkrétní místo v referenčním genomu. K tomu je zaprvé potřeba indexovat referenční genom anebo všechna čtení, kdy algoritmus vloží do paměti polohové informace sekvencí referenčního genomu anebo čtení, aby v nich mohl následně rychle vyhledávat. Následně algoritmus určí, kam všude se mohou jednotlivá čtení přiřadit na referenční genom. V dalším kroku algoritmus použije *pairwise alignment*, aby rozhodl, která místa jsou nejvhodnější a jestli vzniknou indely nebo substituce. Většina algoritmů používá algoritmy Needleman-Wunsch nebo Smith-Watermann pro konečný *alignment* (Alser et al., 2021).

Referenční genom je genom, který slouží jako standardní sekvence pro jeden živočišný druh (Kaye & Wasserman, 2021). První rozšířený lidský referenční genom GRCh37 (Church et al., 2011) je lineární. Populace (jedinci, jejichž genom byl použit pro sestavení reference) je znázorněna jako jedna sekvence (linearita), která bere v potaz jedinou nejčastější bázi na každé své pozici. Novější verze – GRCh38 – přinesla kromě nových oblastí, které nebyly v GRCh37 zahrnuty, také výrazně více alternativních *scaffoldů* (sestavení), které obsahují sekvence, které se vyskytly v populaci (Schneider et al., 2016). Nejúplnější lidský lineární referenční genom je T2T-CHM13 (Nurk et al., 2022), kde T2T znamená *telomere-to-telomere* (od telomery k telomeře) (Nurk et al., 2022). Nejmodernější referenční genomy jsou grafové pangenomy, které udržují informace o různých alelách genů a variacích v populaci pomocí struktury grafu (W.-W. Liao et al., 2023).

Mapovací algoritmy lze rozdělit do dvou kategorií podle datové struktury, kterou používají pro vyhledávání v referenčním genomu a pro hledání místa, kam se čtení nejlépe přiřadí: hashovací tabulka a sufixové pole.

Mezi algoritmy používající hashovací tabulky se řadí MAQ (H. Li et al., 2008), NovoAlign (Novocraft, 2023) a Minimap2 (H. Li, 2018). Hashovací tabulka je datová struktura, která kterékoli jedinečné hodnotě přiřadí jedinečnou hodnotu pomocí hashovacího algoritmu, čímž lze následně v tabulce velmi rychle vyhledávat (Maurer & Lewis, 1975). Pomocí hashovací tabulky algoritmus indexuje referenční genom a následně hledá zkrácené čtení v referenčním genomu (těmto krátkým sekvencím se říká *seed*). Aby bylo možné vyhledávat čtení se substitucemi na začátku, *spaced seedy* mají definovaná místa, kde se se nemusí perfektně shodovat s referenčním genomem. Tyto *seedy* jsou následně prodlužovány pomocí *alignment* algoritmů, aby pokryly co největší část čtení (H. Li & Homer, 2010). *Seedy* i *spaced seedy* nepodporují indely na začátku čtení. Toto je řešeno q-gram filtrem, který dovoluje k sobě přiřazovat sekvence, které mají mezi sebou mezery (Singh et al., 2018). Q-gram filtr je implementován například v nástrojích SHRiMP (Rumble et al., 2009) a RazerS 3 (Weese et al., 2012).

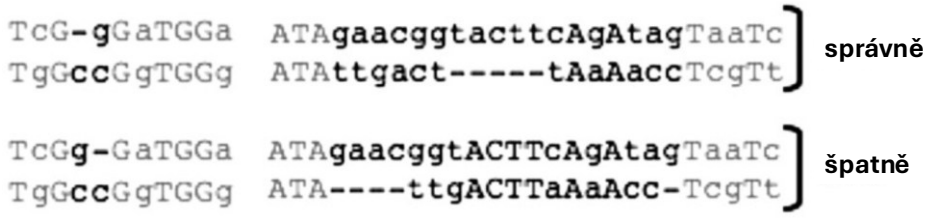
Další metodou, která umožňuje rychlé vyhledávání v sekvenci znaků, je vyhledávání pomocí sufixových polí. Sufixové pole je alternativní, paměťově úspornější struktura k sufixovému stromu, která ukládá pouze setříděné počátky všech sufixů. Sufixový strom je stromová datová struktura, ve které jsou uloženy všechny možné sufixy celé sekvence, aby šlo v sekvenci rychle vyhledávat (Alser et al., 2021). Na podobném principu, jako je sufixové pole, je také založen FM-index, který je odvozen od Burrows-Wheelerovy transformace (BWT). FM-index neukládá přímo počátky sufixů, ale umožňuje efektivní vyhledávání podřetězců pomocí operací *rank/select*; pozice nalezených sufixů lze zpětně rekonstruovat pomocí vzorkování sufixového pole (Ferragina & Manzini, 2005; H. Li & Durbin, 2009). Tyto datové struktury využívají například algoritmy BWA (H. Li & Durbin, 2009), BWA-MEM2 (Vasimuddin et al., 2019), Bowtie2 (Langmead & Salzberg, 2012) a HISAT2 (Kim et al., 2019). Podobně jako u algoritmů používající hashovací tabulky, i tyto algoritmy hledají nejdříve krátké sekvence (*seedy*), které jsou následně rozšiřovány algoritmy pro *pairwise alignment* (Vasimuddin et al., 2019).

3.1.3 Nedostatky skórovacích funkcí

Princip skórovacích funkcí umožní najít více než jednu část na referenčním genomu, kde její zarovnání vede ke stejně vysokému skóre. Těmto zarovnáním se říká „vzájemně optimální zarovnání“ nebo „nejednoznačné mapování“. Mapovaná čtení jsou pak vyloučena (Marioni et al., 2008), náhodně přiřazena na jedno z míst (MAQ, BWA), nebo jedno z míst, kde se mohou zarovnat, je vybráno jako primární a další jako sekundární (BWA-MEM, Bowtie2) (Aldawiri et al., 2022; Landan & Graur, 2009; Wilton & Szalay, 2023). Dále je možné, že skórovací funkce vytvoří chybné zarovnání – nukleotidy jsou zarovnány tak, že mají mezi sebou špatně určené mezery, i když je skóre optimální (viz obr. 6). Správnost alignmentu je hodnocena z evolučního pohledu (Landan & Graur, 2009).

Pro kvantifikaci nejednoznačného mapování byla zavedena mapovací kvalita (MAPQ), která přiřazuje nízkou kvalitu čtením, které se mohou stejně dobře mapovat na více míst, a vysokou kvalitu čtením, které mají až jen jedno místo, kam se mohou namapovat (H. Li et al., 2008). Výpočet MAPQ je odvozen od pravděpodobnosti špatného zarovnání sekvence (angl. *misalignment*), a její výpočet se liší mezi mapovacími algoritmy. Pravděpodobnost lze upravit nebo lépe odhadnout na základě sekvenovaného vzorku například pomocí strojového učení, které bere v potaz sekvenční kontext, kvalitu *alignmentu* a další proměnné. Tyto dodatečné odhady mají často efekt na zvýšený počet správně detekovaných bodových mutací, ale ne na větší strukturní varianty (Cline et al., 2020).

Skórovací funkce jsou parametrizovány hodnotami, které definují její funkcionalitu – jak má vybudovat zarovnání a jaké skóre mu dát, velikosti *seedů* atp. Mapovací programy mají svoje předdefinované hodnoty těchto parametrů, které je možné manuálně změnit



Obrázek 6 - Příklady možných chyb alignmentu. Skórovací funkce může přiřadit dvěma zcela odlišným zarovnáním dvou sekvencí identické (zároveň optimální) skóre a je na náhodě, které z nich bude vybráno. Na těchto dvou příkladech je ilustrováno, jak může vypadat jednoduchý (vlevo) a složitý (vpravo) špatně a správně sestavený alignment. Nakonec pracovník vyhodnotí, který alignment je nejlepší nebo upraví parametry skórovací funkce, aby dosáhl nejlepšího výsledku. Upraveno dle Landan a Gaur (2009).

pro zlepšení výsledku pro daný organismus, což je výhodné zejména pro nemodelové nebo vysoce mutované genomy (Nielsen et al., 2011; Smolka et al., 2015). Existuje velký výběr mapovacích nástrojů a algoritmů a je běžné, že se pro specifické využití porovnávají mezi sebou pro dosažení co nejlepšího výsledku. Programy, jako je Teaser (Smolka et al., 2015), automaticky vyhodnotí sekvenovaný vzorek a zjistí nejvhodnější nastavení parametrů specifických mapovacích nástrojů (např. Bowtie2 má 34 různých parametrů). Tato optimalizace může zvýšit přesnost mapování a výrazně zrychlit jeho proces. Během optimalizace často záleží na sledované hodnotě (např. počet falešně pozitivních mapování) a podle toho je nutné parametry upravit (Hatem et al., 2013; Wilton & Szalay, 2023).

3.1.4 Nedostatky referenčního genomu

Referenční genomy, kvůli sekvenačním problémům nastíněným výše a nedostatkům při *de novo assembly,* nejsou kompletní a jsou vytvořeny z velmi malého vzorku analyzovaných individuálních genomů. GRCh38 je sestaven z genomů několika desítek jedinců a T2T-CHM13 je dokonce výstupem genomu jednoho člověka, což není reprezentativní pro celou lidskou populaci. Verze GRCh38 rozšířila GRCh37 o mnoho chybějících sekvencí v duplikovaných, telomerických a centromerických oblastech a opravuje mnoho dalších menších chyb (Schneider et al., 2016). Přesto v GRCh38 stále chybí přibližně 250 Mbp sekvence (W.-W. Liao et al., 2023). T2T-CHM13 je jediný kompletní lidský referenční genom, který rozšiřuje GRCh38 o 236 Mbp sekvence zejména v centromerických oblastech, mimosatelitních duplikacích a chromozomu Y (Nurk et al., 2022; Rhie et al., 2023).

Vysoce polymorfní oblasti představují další problém referenčních genomů. Mezi vysoce polymorfní patří například oblasti kódující geny hlavního histokompatibilního systému (HLA; *Human Leukocyte Antigen*). Mapování čtení na tyto regiony představuje výzvu, protože standardní lineární referenční genom nemůže pojmout všechny varianty, které mohou v populaci vzniknout (Dilthey, 2021). Mimo jiné HLA geny mají blízké paralogy na chromozomech 1, 9 a 19, kvůli kterým se mohou čtení mapovat nejednoznačně (Brandt et al., 2015; Kasahara et al., 2004). Pro mapování do HLA lokusů byla vytvořena jedna z prvních lidských grafových referenčních sekvencí (PRG), která na rozdíl od lineárních

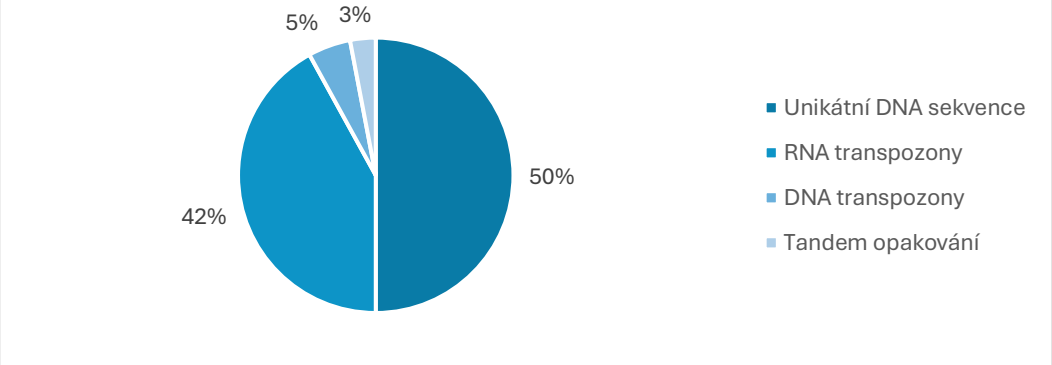
referenčních genomů umí zahrnout i strukturní a další populačně specifické varianty (Dilthey et al., 2015).

Mapovací algoritmy mají těžko odstranitelnou zaujatost vůči alele, která je v referenčním genomu. Algoritmy fungují na principu přiřazování čtení na část genomu, které se nejvíce podobá. Nejpoužívanější genomy (GRCh37 a GRCh38) prakticky fungují jako homozygotní. Pokud je analyzovaný jedinec v jednom místě heterozygotní, tak čtení budou obsahovat jak referenční alelu, tak alternativní (vůči referenčnímu genomu). Referenční alela se bude mapovat vždy (pokud opomeneme mutace), kdežto alternativní alela se bude mapovat méně pravděpodobně, protože je jiná, a algoritmy ji mohou mapovat s nízkou MAPQ nebo ji nenamapují vůbec (Brandt et al., 2015; Lunter & Goodson, 2011).

3.1.5 Repetitivní části genomu

Přibližně 50 % lidského genomu tvoří repetitivní sekvence (International Human Genome Sequencing Consortium, 2001; W.-W. Liao et al., 2023), jako jsou genové duplikace, transpozony nebo kratší homopolymery a repetice (obr. 7). Při mapování čtení, které pochází z repetitivních sekvencí, se často stává, že vznikne nejednoznačné mapování a výsledná MAPQ celé repetice je řádově nižší (Ross et al., 2013).

Sekvenování pomocí technologie Illumina má hlavní nevýhodu v krátkých čteních, jejichž délka je často kratší (i při použití párového sekvenování) než délka sekvenované repetitice. V GRCh37 existuje přibližně 23 000 000 úseků o délce 1 kbp, která nemají jednoznačné mapování a většina z nich (> 99 %) je součástí transpozonů a velkých segmentových duplikací (W. Li & Freudenberg, 2014). Pro efektivní sekvenaci repetitivní



Obrázek 7 - Rozdělení repetitivních sekvencí v lidském genomu. Přibližně 50 % lidského genomu je unikátní, 42 % je složeno z RNA transpozonů, 5 % z DNA transpozonů a 3 % z tandemových repetic. Upraveno dle Liao et al. (2023).

části genomu je vhodné používat technologie PacBio nebo ONT, která mají delší čtení (Gall-Duncan et al., 2022; Reinert et al., 2015; Tanudisastro et al., 2024). Je však nutno podotknout, že většina diagnostických sekvenací obvykle na repetitivní oblasti necílí.

Pro vyřešení nejednoznačných mapování nebo pro dodatečné mapování čtení, která nebyla namapována vůbec, zejména v oblastech s krátkými repeticemi, je možné využít nástrojů pro *remapping* (Tanudisastro et al., 2024). *Remapping* je metoda, která identifikuje repetitivní sekvence a čtení, které mají nízkou MAPQ, nejsou namapovány nebo mají výrazně nenormální délku *insertu* či orientaci. Následně jsou tato čtení znovu *alignována* pomocí specializovaných programů jako HipSTR (Willems et al., 2017) nebo GangSTR (Mousavi et al., 2019).

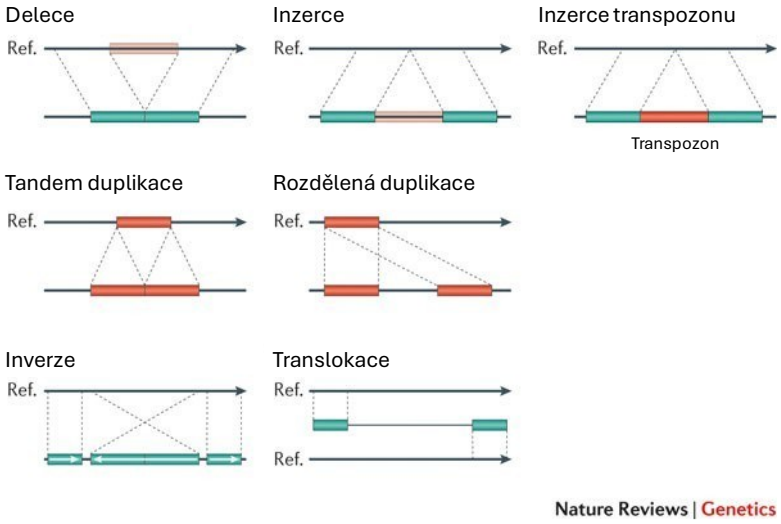
3.1.6 Strukturní varianty

Stejně jako bodové mutace i strukturní varianty (varianty o délce větší 50 bp) hrají důležitou roli ve variabilitě genomu. Mezi strukturní varianty se řadí inzerce (vložení nové sekvence nebo transpozonu), delece (ztráta úseku sekvence), duplikace (zmnožení úseku sekvence), inverze (otočení úseku sekvence) a translokace (přemístění kusu sekvence). Variace počtu kopií (angl. *Copy Number Variant, CNV*) je speciální kategorie strukturních variant, které mění počet jedné sekvence v genomu (duplikace a delece) (Alkan et al., 2011). Detekce strukturních variant pomocí krátkých čtení lze rozdělit na tři základní způsoby: 1) párová čtení mají mezi sebou *insert*, který je netypicky dlouhý nebo je jedno čtení ve špatném směru vůči druhému; 2) čtení jsou rozdělena na dvě části; 3) pokrytí sekvence je v daném místě abnormální (Alkan et al., 2011; Hanlon et al., 2022; Mahmoud et al., 2019).

Každý typ strukturních variant má svůj typický způsob mapování na referenční genom nebo mají typický profil pokrytí (obr. 8). Například duplikace výrazně zvýší pokrytí duplikované sekvence, inverze změní orientaci čtení a delece zvětší mezeru mezi párovými čteními (Mahmoud et al., 2019). I přes tyto identifikátory může být mapování strukturních variant náročné, obzvláště v případech, kdy dochází k balancovaným přestavbám s hranicemi v místech nepokrytých dostatečným počtem čtení nebo v místech repetitivních sekvencí.

Zejména při používání sekvenování technologií Illumina, která má čtení většinou kratší než většina strukturních variant, se stává, že varianta nebude detekována. Při komplexních strukturních variantách, které kombinují více než jeden typ variant, je mapování pomocí krátkých čtení skoro nemožné. Dále, krátké delece nebo inzerce, které vzniknou mezi páry čtení (v rámci *insertu*) se mohou velmi jednoduše ztratit, protože touto metodou lze nalézt jen delší varianty, které výrazněji vyčnívají z rozdělení[3] délky *insertu*. Detekce strukturních variant pomocí krátkých čtení má vysokou pravděpodobnost falešně pozitivních identifikací a algoritmy naráží na strop toho, co je technologicky možné (Mahmoud et al., 2019).

Sekvenování s delšími čteními (ONT a PacBio) tyto nedostatky do velké míry odstraňuje. Jedno čtení může pojmout i velmi složité varianty a přestavby a je lepší konkrétně při identifikaci inzercí (Mahmoud et al., 2019; Y. Pei et al., 2024).



Obrázek 8 – Znázornění efektu delece, inzerce, transpozonů, duplikace, inverze a translokace na sekvenci vůči referenčnímu genomu (Ref.). Při deleci sekvence zmizí, při inzerci se sekvence vloží mezi namapované části, při tandemové duplikaci se nová sekvence zařadí za starou (to je během mapování většinou znázorněno znásobením pokrytí duplikované části), při rozdělené duplikaci se nová sekvence zařadí jinam v genomu, při inverzi se sekvence otočí a při translokaci se sekvence posune jinam v genomu. Upraveno dle Alkan et al. (2011).

3.1.7 PCR duplikáty

Pro sekvenování a následné bioinformatické zpracování je vhodné, aby byl celý sekvenovaný vzorek rovnoměrně amplifikován. Během PCR amplifikace však dochází ke

[3] Fragmenty DNA nebývají všechny stejně dlouhé, zejména, když jsou fragmentovány sonikací. Tudíž při sekvenování jedné délky čtení vzniknou různě dlouhé *inserty*. Tyto délky mají nějaké matematické rozdělení, na jehož základě lze odhadnout, zda insert není moc dlouhý či krátký, což by indikovalo strukturní variantu.

vzniku tzv. PCR duplikátů – identických fragmentů DNA, které vznikly nadměrnou amplifikací jednoho templátu. Tyto duplikáty se vyskytují napříč celou knihovnou a přispívají k umělému navýšení sekvenační hloubky a velikosti výsledných dat, aniž by přinesly nové biologické informace. Jejich výskyt je ovlivněn očekávaným pokrytím a omplexitou vzorku – vzorky s nižší komplexitou nebo velmi hlubokým sekvenováním mívají vyšší podíl PCR duplikátů (Rochette et al., 2023).Mapování PCR duplikátů na referenční genom uměle zvýší pokrytí postižených částí, což může negativně ovlivnit zejména identifikaci duplikací, které jsou detekovány zvýšeným pokrytím. PCR duplikáty uměle navyšují pokrytí analyzované sekvence oproti očekávanému pokrytí, čímž se nadhodnotí kvalita sekvenování. Proto by bylo vhodnější pokrytí počítat nikoliv jako průměr hloubky sekvenování přes každou bázi, ale jako maximální hloubku sekvenování, které dosahují všechny báze v.

PCR duplikáty jsou jednoduše počítačově detekovatelné (jsou to identická čtení o stejné délce, která se v datech objevují nadměrně častokrát), a proto existuje mnoho deduplikačních nástrojů, které duplikáty identifikují a ze vzorku odstraní. Nejpoužívanější nástroje jsou Picard MarkDuplicates (Broad Institute, 2019) a SAMTools (H. Li et al., 2009). Efektivita obou nástrojů je podobná a jejich využití se snižuje kvůli vyšší efektivitě neustále se vyvíjejících sekvenančních technologií a nástrojů pro identifikaci variant (Ebbert et al., 2016). Deduplikační nástroje mohou však někdy chybně odstranit i skutečné duplikáty (duplikovaný segment DNA) nebo některé duplikáty minou (repetitivní segment DNA) (Zvěřinová & Guryev, 2022). Výskyt a identifikaci PCR duplikátů je možné ovlivnit i pomocí úpravy v přípravě sekvenovaných vzorků. Během přípravy sekvenační knihovny lze na fragmenty navázat UMI (*unique molecular identifier*), které jsou sekvenovány společně s fragmentem DNA a jednoznačně identifikují každý analyzovaný fragement DNA. Na jejich základě lze odstranit PCR duplikáty, protože kvantifikace UMI určí fragmenty, které byly nadmíru amplifikovány (Rochette et al., 2023).

3.2 *Variant calling*

Posledním krokem po mapování čtení na referenční genom je identifikace variant (*variant calling* – VC) a určení efektu těchto variant na fenotyp analyzované osoby. Kvalita VC je

nejvíce opřena o kvalitu nástroje pro VC, ale i přes výrazné zvýšení kvality mapovacích nástrojů, tak hrají roli v kvalitním mapování (Barbitoff et al., 2022; Olson et al., 2023).

3.2.1 Nástroje

Nástroje, které varianty identifikují, se nazývají *callery*. Podobně jako u nástrojů pro mapování existuje i řada různých *callerů* pro VC, často specializovaných na konkrétní typ variant – bodové varianty, strukturní varianty, CNV a další varianty. Rozsáhlý přehled nástrojů publikovali Zvěřinová a Guryev (2022), Koboldt (2020) a Nielsen et al. (2011). Níže je uveden přehled nejpoužívanějších.

Nejčastěji používaným *callery* jsou GATK HaplotypeCaller (McKenna et al., 2010), FreeBayes (Garrison & Marth, 2012) a Bcftools (Daněček et al., 2021). Tyto tři *callery* mají rozdílné přístupy k VC a všechny jsou primárně určeny pro VC krátkých variant – bodové mutace a krátké indely. GATK HaplotypeCaller nejdříve identifikuje aktivní oblasti, na nich vytvoří grafovou strukturu a identifikuje možné varianty pomocí Smith-Waterman algoritmu. Následně využije skryté Markovovy modely pro vypočítání pravděpodobnosti jednotlivých alel a z těch je vyvozen nejpravděpodobnější genotyp (GATK, 2023). FreeBayes identifikuje varianty na základě modelu Bayesovy podmíněné pravděpodobnosti, kde ke každému genotypu daného lokusu na základě čtení přiřadí pravděpodobnost výskytu. Přestože postup je většinou používán pro kratší varianty, lze tento algoritmus aplikovat i pro delší strukturní varianty jako CNV (Garrison & Marth, 2012). V balíčku Bcftools je nejdříve aplikován příkaz `mpileup`, který vygeneruje pravděpodobnosti pro genotypy přes všechny přečtené báze. Následně se provede příkaz `call`, který varianty na základě pravděpodobností identifikuje. V rámci Bcftools jsou i nástroje pro VC CNV (Daněček et al., 2021).

V poslední době vznikají *callery*, které využívají strojové učení a je zřejmé, že tyto nástroje budou postupně nahrazovat nástroje, které na něm založené nejsou (Olson et al., 2023). Mezi nejvyužívanější nástroje tohoto typu se řadí DeepVariant, který je založen na principu hlubokého strojového učení. DeepVariant prvně identifikuje možná místa variant, ze kterých vytvoří speciální datovou strukturu. Ta je následně vložena do předtrénované konvoluční neuronové sítě, která vrátí pravděpodobnosti genotypů pro každou variantu (Poplin et al., 2018).

Výběr vhodného *calleru* je velmi důležitý – každý *caller* je založen na jiném principu, proto je překryv mezi identifikovanými variantami mezi různými *callery* jen částečný. Překryvy mezi výše zmíněnými *callery* se pohybují od 57 % až po 92 % (Hwang et al., 2015; Lin et al., 2022), i když vždy záleží na vzorku, který je pro tato porovnávání používán. V komplexních částech genomu (množství repetic, homopolymerů, GC-bohatých sekvencí) může být překryv mezi *callery* nízký (75 % a méně), ale v méně komplexních částech genomu tento překryv může být až 99,7 % (Krusche et al., 2019). Jak bylo uvedeno, jsou *callery* určené k identifikaci různých typů variant a je tudíž časté, že je jeden vzorek analyzován několika různými *callery* pro identifikaci více typů variant. Některé nové dokonce kombinují několik různých *callerů* pro přesnější VC (Alkan et al., 2011).

Hodnocení kvality *callerů* je v různých studiích velmi rozdílné , což je pravděpodobně dáno studovanými vzorky a množinou porovnávaných *callerů* (Hwang et al., 2015; Koboldt, 2020; Krusche et al., 2019; Lin et al., 2022; S. Pei et al., 2021). Nelze tedy říct, že by jeden *caller* byl lepší než ostatní. Pro určení nejvhodnějšího *calleru* existují doporučení (Krusche et al., 2019) a optimalizační nástroje jako RecallME, které určí nejlepší *caller* včetně relevantních parametrů (Vozza et al., 2023).

Některé varianty mohou být z výsledného VC odstraněny (filtrovány) na základě pokrytí, MAPQ nebo genetických ukazatelů (Hardy-Weinbergova rovnováha, vazebná nerovnováha a další) (Koboldt, 2020; Nielsen et al., 2011), což může zvýšit přesnost VC (S. Pei et al., 2021). Uvedený postup ovšem není univerzální, protože při jeho nesprávném použití může dojít k chybnému odstranění skutečných složitých varianty s nízkým pokrytí, což vede k falešně negativním výsledkům (Olson et al., 2023). Kvalitní *callery* umí pracovat s nízkým pokrytím nebo nízkým MAPQ, a tudíž jimi nejsou do takové míry ovlivňovány (Barbitoff et al., 2022).

Callery dosahují velmi vysoké přesnosti – až 99,9 %. Přesto i při této úrovni přesnosti vznikají falešně pozitivní nálezy. V typickém lidském genomu je identifikováno přibližně 4–5 milionů zděděných variant a zhruba 70 *de novo* variant. Pokud bychom identifikovali 5 milionů variant s 0,1 % chybovostí, objevilo by se přibližně 5000 falešně pozitivních variant – tedy asi 70krát více, než je očekávaný počet skutečných *de novo* variant

(Koboldt, 2020; S. Pei et al., 2021). Proto je vždy doporučováno provést i vizuální kontrolu každé varianty s potenciálním dopadem na výsledný fenotyp pomocí vizualizačního nástroje jako je Integrative Genomics Viewer (IGV) (Koboldt, 2020; Lin et al., 2022; Olson et al., 2023; S. Pei et al., 2021). Přesnost detekce varianty lze také ovlivnit již při přípravě knihovny nebo při výběru analytické metody. Při pokrytí menší než 10 je přesnost VC výrazně menší než při vyšších pokrytí, proto je obvyklé používat pro identifikaci bodových mutací alespoň pokrytí 10×, ale pokrytí alespoň 30× je doporučováno pro celogenomové sekvenování. Od pokrytí 40× se již přesnost VC heterozygotních variant v germinálním genomu nezvyšuje (Hwang et al., 2015; Koboldt, 2020). Sekvenování pomocí dlouhých čtení také výrazně zvyšuje přesnost VC nejen u strukturních variant, ale také menších variant ve vysoce polymorfních regionech (Olson et al., 2023; S. Pei et al., 2021; Zvěřinová & Guryev, 2022).

Kromě VC pomocí sekvenování lze využít řadu jiných metod. Tyto metody se spíše zabývají rozsáhlými strukturními variantami – CNV, chromozomové přestavby a další. Mezi tyto metody patří *Array CGH*, *SNP microarray*, FISH a optické mapování (Alkan et al., 2011; Balachandran & Beck, 2020).

3.2.2 Anotace variant

Pro přiřazení efektu varianty se varianty anotují. Efekty dělíme na několik kategorií (viz kapitola 1.1.2). Důsledky bodových mutací, potažmo strukturních variant, na genovou expresi lze odhadnout pomocí počítačových programů, ale predikovat jejich skutečný důsledek vyžaduje složitější nástroje.

Nejpoužívanější přístup k anotaci variant je skrze rozsáhlé kontrolované databáze variant jako je např. ClinVar (Landrum et al., 2018). Tyto databáze obsahují tisíce germinálních variant s popisem jejich klinického významu. Přesto je nezbytné uvedené informace pečlivě ověřovat, protože mohou obsahovat chyby nebo nepřesnosti (Koboldt, 2020; Najafi et al., 2020). Specializované databáze jsou zaměřeny na hodnocení somatických variant.

Populační databáze (např. gnomAD), které zaznamenávají informace o frekvenci variant ve světových populacích, jsou také používány pro filtrování variant (většinou jenom krátké indely nebo bodové varianty). Nejčastěji jsou varianty filtrovány na populační frekvenci

nižší než 0,1 %, což nemusí být vždy dostatečné, zejména v klinické diagnostice , kdy jsou hledány velmi vzácné varianty (Najafi et al., 2020).

Závěr

Bakalářská práce se zaměřuje na identifikaci chyb, které vznikají během sekvenování germinální DNA pomocí NGS a v následných bioinformatických analýzách, a které mohou poměrně zásadním způsobem ovlivnit klinickou interpretaci výsledků. Práce se zejména věnuje technologii Illumina, která je v současnosti nejrozšířenější platformou pro klinickou diagnostiku dědičných poruch zárodečné DNA v humánní medicínské praxi. Práce ukazuje, že přesnost výsledků NGS je ovlivněna řadou faktorů napříč celým sekvenačním řetězcem – od biochemických vlastností DNA, přes limity sekvenační technologie až po zpracování dat pomocí bioinformatických algoritmů. Proto je při návrhu experimentu nutno uvažovat o tom, jak se volby v každém kroku procesu kumulativně promítají do výsledku.

V práci byly vynechány chyby vznikající na pre-analytické úrovni samotné DNA. Chyby vznikající při přípravě knihovny, zejména při PCR amplifikaci, mohou být částečně eliminovány vhodným experimentálním návrhem. Naopak chyby spojené se samotnou analýzou pomocí NGS jsou často technologicky podmíněné a nelze jim zcela předejít. O to důležitější je s nimi počítat při interpretaci dat. V rámci bioinformatického zpracování, tedy při mapování a variant callingu, hraje klíčovou roli výběr vhodných softwarových nástrojů a parametrů, které by měly být optimalizovány podle typu vzorku a zamýšleného klinického využití.

V klinické praxi je však nutné, aby celý sekvenační a analytický proces byl co nejvíce bezchybný. Dlouhodobou slabinou je první krok analýzy – mapování na referenční genom – od něhož se další chyby odvíjí. Referenční genom není kompletní a není populačně reprezentativní. Ideální referenční genom by měl být sestaven z dostatečně velkého vzorku genomů z definované populace (země, oblast, kontinent nebo celý svět), aby v sobě uchovával i populačně specifickou variabilitu. Takový referenční genom by výrazně zvýšil přesnost jakékoli genetické analýzy, protože filtrování variant a určení jejich efektu by bylo opřené o silnější datový model. Kvůli mapování krátkých čtení dochází také k obtížím při mapování v repetitivních nebo silně polymorfních oblastech genomu.

V návaznosti na chyby v mapování vznikají chyby i ve *variant callingu*, kde mohou být skutečné varianty přehlédnuty nebo naopak dochází k detekci umělých, neexistujících variant. Kvalita výsledné variantní analýzy tak závisí nejen na samotné kvalitě vstupních dat, ale i na parametrech požitých nástrojů, které musí být pečlivě optimalizovány dle konkrétního typu analýzy.

V současnosti technologie Illumina zůstává zlatým standardem pro klinickou diagnostiku germinální DNA. Její největší slabinou zůstává mapování krátkých čtení na referenční genom, které je zdrojem mnoha dalších chyb v dalších analýzách. Avšak vysoká přesnost v oblasti detekce malých variant, robustnost technologie a široká dostupnost bioinformatických nástrojů ji stále činí preferovanou metodou v běžné praxi.

Sekvenační technologie třetí generace představují v mnoha ohledech vhodnější technologii pro sekvenování v klinickém prostředí. Pro přípravu knihovny není potřeba PCR, čímž nevznikají mnohé artefakty, které negativně ovlivňují sekvenování a analýzu. Díky jejich dlouhým čtením lze přesněji analyzovat i velmi komplexní a polymorfní části genomu. Největší výhoda dlouhých čtení je v mapování a VC rozsáhlých strukturních variant, které jsou více méně nedetekovatelné pomocí krátkých čtení. Ač v prvních letech byly tyto technologie značně méně přesné, tak s jejich soustavným rozvojem již dosahují přesnosti první i druhé generace sekvenování. V momentě, kdy se i jejich finanční náročnost sníží, tak mohou velmi efektivně nahradit předešlé generace sekvenování. Jejich finanční náročnost je mimo jiné velkou překážkou v rozsáhlejší tvorbě referenčních genomů, které potřebují přesné a dlouhé čtení.

Budoucí práce se mohou věnovat vytvoření počítačových programů soustředících se na identifikaci chyb popsaných v této práci. V současnosti jsou vyvíjeny nové a modernější technologie sekvenování jako je technologie Roche *Sequencing-by-Expansion*, které slibují vylepšení oproti technologii Illumina (delší čtení, nízká finanční náročnost a nízká časová náročnost). V budoucnu bude potřeba je systematiky zanalyzovat a prozkoumat jejich chybovost.

Seznam literatury

Aldawiri, T., Nanduri, B., Ramkumar, M., & Perkins, A. D. (2022). A Novel Approach for Mapping Ambiguous Sequences of Transcriptomes. *Proceedings of 14th International Conference on Bioinformatics and Computational Biology*, *83*, 76–85.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping *. *Nature Reviews Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958

Allhoff, M., Schönhuth, A., Martin, M., Costa, I. G., Rahmann, S., & Marschall, T. (2013). Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, *14*(S5), S1. https://doi.org/10.1186/1471-2105-14-S5-S1

Alser, M., Rotman, J., Deshpande, D., Taraszka, K., Shi, H., Baykal, P. I., Yang, H. T., Xue, V., Knyazev, S., Singer, B. D., Balliu, B., Koslicki, D., Skums, P., Zelikovsky, A., Alkan, C., Mutlu, O., & Mangul, S. (2021). Technology dictates algorithms: Recent developments in read alignment *. *Genome Biology*, *22*(1), 249. https://doi.org/10.1186/s13059-021-02443-7

Altschul, S. F. (1998). Generalized affine gap costs for protein sequence alignment. *Proteins*, *32*(1), 88–96.

Altschul, S. F., & Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bulletin of Mathematical Biology*, *48*(5–6), 603–616. https://doi.org/10.1007/BF02462326

Balachandran, P., & Beck, C. R. (2020). Structural variant identification and characterization. *Chromosome Research*, *28*(1), 31–47. https://doi.org/10.1007/s10577-019-09623-z

Barbitoff, Y. A., Abasov, R., Tvorogova, V. E., Glotov, A. S., & Predeus, A. V. (2022). Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. *BMC Genomics*, *23*(1), 155. https://doi.org/10.1186/s12864-022-08365-3

Batey, R. T., Rambo, R. P., & Doudna, J. A. (1999). Tertiary Motifs in RNA Structure and Folding *. *Angewandte Chemie (International Ed. in English)*, *38*(16), 2326–2343. https://doi.org/10.1002/(sici)1521-3773(19990816)38:16<2326::aid-anie2326>3.0.co;2-3

Beichman, A. C., Zhu, L., & Harris, K. (2024). The Evolutionary Interplay of Somatic and Germline Mutation Rates *. *Annual Review of Biomedical Data Science*, *7*(1), 83–105. https://doi.org/10.1146/annurev-biodatasci-102523-104225

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, *40*(10), e72–e72. https://doi.org/10.1093/nar/gks001

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in

the 1000 Genomes Project Phase I Data. *G3 Genes|Genomes|Genetics*, *5*(5), 931–941. https://doi.org/10.1534/g3.114.015784

Broad Institute. (2019). *Picard Toolkit*. Broad Institute, GitHub repository. https://broadinstitute.github.io/picard/

Browne, P. D., Nielsen, T. K., Kot, W., Aggerholm, A., Gilbert, M. T. P., Puetz, L., Rasmussen, M., Zervas, A., & Hansen, L. H. (2020). GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience*, *9*(2), giaa008. https://doi.org/10.1093/gigascience/giaa008

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, *4*(4), 265–270. https://doi.org/10.1038/nnano.2009.12

Cline, E., Wisittipanit, N., Boongoen, T., Chukeatirote, E., Struss, D., & Eungwanichayapant, A. (2020). Recalibration of mapping quality scores in Illumina short-read alignments improves SNP detection results in low-coverage sequencing data. *PeerJ*, *8*, e10501. https://doi.org/10.7717/peerj.10501

Complete Genomics. (2024, září 16). *Next-Generation Sequencing Costs: The Sub $100 Genome*. https://www.completegenomics.com/next-generation-sequencing-costs/

Daněček, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Dechering, K. J., Konings, R. N. H., Cuelenaere, K., & Leunissen, J. A. M. (1998). Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Research*, *26*(17), 4056–4062. https://doi.org/10.1093/nar/26.17.4056

de Oliveira Martins, L., Bloomfield, S., Stoakes, E., Grant, A. J., Page, A. J., & Mather, A. E. (2022). Tatajuba: Exploring the distribution of homopolymer tracts. *NAR Genomics and Bioinformatics*, *4*(1), lqac003. https://doi.org/10.1093/nargab/lqac003

Dilthey, A. (2021). State-of-the-art genome inference in the human MHC *. *The International Journal of Biochemistry & Cell Biology*, *131*, 105882. https://doi.org/10.1016/j.biocel.2020.105882

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, *47*(6), 682–688. https://doi.org/10.1038/ng.3257

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105. https://doi.org/10.1093/nar/gkn425

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Kauwe, J. S. K., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17*(S7), 239. https://doi.org/10.1186/s12859-016-1097-3

Ferragina, P., & Manzini, G. (2005). Indexing compressed text. *Journal of the ACM*, *52*(4), 552–581. https://doi.org/10.1145/1082036.1082039

Gall-Duncan, T., Sato, N., Yuen, R. K. C., & Pearson, C. E. (2022). Advancing genomic technologies and clinical awareness accelerates discovery of disease-

associated tandem repeat sequences *. *Genome Research*, *32*(1), 1–27. https://doi.org/10.1101/gr.269530.120

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing* (Verze 2). arXiv. https://doi.org/10.48550/ARXIV.1207.3907

GATK. (2023, leden 25). *HaplotypeCaller*. GATK. https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies *. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gorlov, I. P., Pikielny, C. W., Frost, H. R., Her, S. C., Cole, M. D., Strohbehn, S. D., Wallace-Bradley, D., Kimmel, M., Gorlova, O. Y., & Amos, C. I. (2018). Gene characteristics predicting missense, nonsense and frameshift mutations in tumor samples. *BMC Bioinformatics*, *19*(1), 430. https://doi.org/10.1186/s12859-018-2455-0

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, *162*(3), 705–708. https://doi.org/10.1016/0022-2836(82)90398-9

Hanlon, V. C. T., Lansdorp, P. M., & Guryev, V. (2022). A survey of current methods to detect and genotype inversions *. *Human Mutation*, *43*(11), 1576–1589. https://doi.org/10.1002/humu.24458

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(1), 184. https://doi.org/10.1186/1471-2105-14-184

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., Magrini, V. J., Richt, R. J., Sander, S. N., Stewart, D. A., Stromberg, M., Tsung, E. F., Wylie, T., Schedl, T., Wilson, R. K., & Mardis, E. R. (2008). Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods*, *5*(2), 183–188. https://doi.org/10.1038/nmeth.1179

Huang, X. C., Quesada, M. A., & Mathies, R. A. (1992). DNA sequencing using capillary array electrophoresis. *Analytical Chemistry*, *64*(18), 2149–2154. https://doi.org/10.1021/ac00042a021

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, *5*(1), 17875. https://doi.org/10.1038/srep17875

Chargaff, E., Lipshitz, R., & Green, C. (1952). COMPOSITION OF THE DESOXYPENTOSE NUCLEIC ACIDS OF FOUR GENERA OF SEA-URCHIN. *Journal of Biological Chemistry*, *195*(1), 155–160. https://doi.org/10.1016/S0021-9258(19)50884-5

Chen, H., Wang, B., Cai, L., Zhang, Y., Shu, Y., Liu, W., Leng, X., Zhai, J., Niu, B., Zhou, Q., & Cao, S. (2024). The performance of homopolymer detection using dichromatic and tetrachromatic fluorogenic next-generation sequencing platforms. *BMC Genomics*, *25*(1), 542. https://doi.org/10.1186/s12864-024-10474-0

Chiaromonte, F., Yap, V. B., & Miller, W. (2001). SCORING PAIRWISE GENOMIC SEQUENCE ALIGNMENTS. *Biocomputing 2002*, 115–126. https://doi.org/10.1142/9789812799623_0012

Christman, A. A. (1952). Purine and Pyrimidine Metabolism. *Physiological Reviews*, *32*(3), 303–348. https://doi.org/10.1152/physrev.1952.32.3.303

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M.,

Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing Reference Genome Assemblies. *PLoS Biology*, *9*(7), e1001091. https://doi.org/10.1371/journal.pbio.1001091

Illumina. (b.r.-a). *NovaSeq X Series | Production scale, ultra-high-throughput sequencers*. Získáno 14. březen 2025, z https://emea.illumina.com/systems/sequencing-platforms/novaseq-x-plus.html

Illumina. (b.r.-b). *Sequencing Read Length | How to calculate NGS read length*. Získáno 14. březen 2025, z https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html

Illumina. (2020). *Indexed Sequencing Overview Guide*. https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/indexed-sequencing-overview-guide-15057455-08.pdf

Illumina. (2024, prosinec 12). *Illumina adapter portfolio | Illumina Knowledge*. https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000003275

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jeon, S. A., Park, J. L., Park, S.-J., Kim, J. H., Goh, S.-H., Han, J.-Y., & Kim, S.-Y. (2021). Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. *Genes & Genomics*, *43*(7), 713–724. https://doi.org/10.1007/s13258-021-01096-x

Jones, C. P., & Ferré-D'Amaré, A. R. (2015). RNA quaternary structure and global symmetry *. *Trends in Biochemical Sciences*, *40*(4), 211–220. https://doi.org/10.1016/j.tibs.2015.02.004

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., Knight, R., & Albertsen, M. (2021). High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nature Methods*, *18*(2), 165–169. https://doi.org/10.1038/s41592-020-01041-y

Kasahara, M., Suzuki, T., & Pasquier, L. D. (2004). On the origins of the adaptive immune system: Novel insights from invertebrates and cold-blooded vertebrates *. *Trends in Immunology*, *25*(2), 105–111. https://doi.org/10.1016/j.it.2003.11.005

Kaye, A. M., & Wasserman, W. W. (2021). The genome atlas: Navigating a new era of reference genomes *. *Trends in Genetics*, *37*(9), 807–818. https://doi.org/10.1016/j.tig.2020.12.002

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing *. *Genome Medicine*, *12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Kopernik, A., Sayganova, M., Zobkova, G., Doroschuk, N., Smirnova, A., Molodtsova-Zolotukhina, D., Sagaydak, O., Ryzhkova, O., Kutsev, S., Groznova, O., Melikyan, L., Bondarchuk, E., Woroncow, M., Albert, E., Bogdanov, V., & Volchkov, P. (2025). Sanger validation of WGS variants. *Scientific Reports*, *15*(1), 3621. https://doi.org/10.1038/s41598-025-87814-x

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., & Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, *6*(4), 291–295. https://doi.org/10.1038/nmeth.1311

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., & Zook, J. M. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560. https://doi.org/10.1038/s41587-019-0054-x

Kuno, G. (1998). Universal diagnostic RT-PCR protocol for arboviruses. *Journal of Virological Methods*, *72*(1), 27–41. https://doi.org/10.1016/S0166-0934(98)00003-2

Landan, G., & Graur, D. (2009). Characterization of pairwise and multiple sequence alignment errors. *Gene*, *441*(1), 141–147. https://doi.org/10.1016/j.gene.2008.05.016

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., … Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, *46*(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lee, T. I., & Young, R. A. (2000). TRANSCRIPTION OF EUKARYOTIC PROTEIN-CODING GENES. *Annual Review of Genetics*, *34*(1), 77–137. https://doi.org/10.1146/annurev.genet.34.1.77

Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: A major mechanism for DNA sequence evolution *. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a040442

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing *. *Briefings in Bioinformatics*, *11*(5), 473–483. https://doi.org/10.1093/bib/bbq015

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/gr.078212.108

Li, W., & Freudenberg, J. (2014). Characterizing regions in the human genome unmappable by next-generation-sequencing at the read length of 1000 bases.

Computational Biology and Chemistry, *53*, 108–117. https://doi.org/10.1016/j.compbiolchem.2014.08.015

Li, X., Ma, S., & Yi, C. (2016). Pseudouridine: The fifth RNA nucleotide with renewed interests *. *Current Opinion in Chemical Biology*, *33*, 108–116. https://doi.org/10.1016/j.cbpa.2016.06.014

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., & Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome *. *Communications Biology*, *6*(1), 954. https://doi.org/10.1038/s42003-023-05322-y

Lin, Y.-L., Chang, P.-C., Hsu, C., Hung, M.-Z., Chien, Y.-H., Hwu, W.-L., Lai, F., & Lee, N.-C. (2022). Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports*, *12*(1), 1809. https://doi.org/10.1038/s41598-022-05833-4

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279. https://doi.org/10.1016/j.gpb.2016.05.004

Lu, L., Jia, H., Dröge, P., & Li, J. (2007). The human genome-wide distribution of DNA palindromes. *Functional & Integrative Genomics*, *7*(3), 221–227. https://doi.org/10.1007/s10142-007-0047-6

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it *. *Genome Biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Marco-Sola, S., Eizenga, J. M., Guarracino, A., Paten, B., Garrison, E., & Moreto, M. (2023). Optimal gap-affine alignment in *O*(*s*) space. *Bioinformatics*, *39*(2), btad074. https://doi.org/10.1093/bioinformatics/btad074

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., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517. https://doi.org/10.1101/gr.079558.108

Maurer, W. D., & Lewis, T. G. (1975). Hash Table Methods. *ACM Computing Surveys*, *7*(1), 5–19. https://doi.org/10.1145/356643.356645

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

McCombie, W. R., McPherson, J. D., & Mardis, E. R. (2019). Next-Generation Sequencing Technologies *. *Cold Spring Harbor Perspectives in Medicine*, *9*(11), a036798. https://doi.org/10.1101/cshperspect.a036798

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., … Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, *19*(9), 1527–1541. https://doi.org/10.1101/gr.091868.109

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I., Singer, M., & Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, *12*(1), 451. https://doi.org/10.1186/1471-2105-12-451

Mignogna, M. L., Ficarella, R., Gelmini, S., Marzulli, L., Ponzi, E., Gabellone, A., Peschechera, A., Alessio, M., Margari, L., Gentile, M., & D'Adamo, P. (2022). Clinical characterization of a novel *RAB39B* nonstop mutation in a family with ASD and severe ID causing RAB39B downregulation and study of a *Rab39b* knock down mouse model. *Human Molecular Genetics*, *31*(9), 1389–1406. https://doi.org/10.1093/hmg/ddab320

Modlin, S. J., Robinhold, C., Morrissey, C., Mitchell, S. N., Ramirez-Busby, S. M., Shmaya, T., & Valafar, F. (2021). Exact mapping of Illumina blind spots in the Mycobacterium tuberculosis genome reveals platform-wide and workflow-specific biases. *Microbial Genomics*, *7*(3). https://doi.org/10.1099/mgen.0.000465

Mousavi, N., Shleizer-Burko, S., Yanicky, R., & Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Research*, *47*(15), e90–e90. https://doi.org/10.1093/nar/gkz501

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*(0), 263–273. https://doi.org/10.1101/SQB.1986.051.01.032

Murat, P., Guilbaud, G., & Sale, J. E. (2020). DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biology*, *21*(1), 209. https://doi.org/10.1186/s13059-020-02124-x

Najafi, A., Caspar, S. M., Meienberg, J., Rohrbach, M., Steinmann, B., & Matyas, G. (2020). Variant filtering, digenic variants, and other challenges in clinical sequencing: A lesson from fibrillinopathies. *Clinical Genetics*, *97*(2), 235–245. https://doi.org/10.1111/cge.13640

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, Md., Ogasawara, N., & Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, *39*(13), e90–e90. https://doi.org/10.1093/nar/gkr344

Nasrin, S., & Rahman, A. (2019). Exploring Systematic Errors in Sequencing Technologies. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 132–137. https://doi.org/10.1109/BIBE.2019.00032

NC State University. (2024, květen 30). Pricing | Genomic Sciences Laboratory. *Genomic Sciences Laboratory*. https://research.ncsu.edu/gsl/pricing/

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

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data *. *Nature Reviews Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Nix, D. A., Hellwig, S., Conley, C., Thomas, A., Fuertes, C. L., Hamil, C. L., Bhetariya, P. J., Garrido-Laguna, I., Marth, G. T., Bronner, M. P., & Underhill, H. R. (2020). The stochastic nature of errors in next-generation sequencing of circulating cell-free DNA. *PLOS ONE*, *15*(2), e0229063. https://doi.org/10.1371/journal.pone.0229063

Novocraft. (2023). *novoAlign | Novocraft*. https://www.novocraft.com/products/novoalign/

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

O'Donnell, M., Langston, L., & Stillman, B. (2013). Principles and Concepts of DNA Replication in Bacteria, Archaea, and Eukarya *. *Cold Spring Harbor Perspectives in Biology*, *5*(7), a010108–a010108. https://doi.org/10.1101/cshperspect.a010108

Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., & Zook, J. M. (2023). Variant calling and benchmarking in an era of complete human genome sequences *. *Nature Reviews Genetics*, *24*(7), 464–483. https://doi.org/10.1038/s41576-023-00590-0

PacBio. (b.r.). *HiFi Reads—Highly accurate long-read sequencing—PacBio*. Získáno 14. březen 2025, z https://www.pacb.com/technology/hifi-sequencing/

Panjkovich, A., & Melo, F. (2005). Comparison of different melting temperature calculation methods for short DNA sequences. *Bioinformatics*, *21*(6), 711–722. https://doi.org/10.1093/bioinformatics/bti066

Pei, S., Liu, T., Ren, X., Li, W., Chen, C., & Xie, Z. (2021). Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Briefings in Bioinformatics*, *22*(3), bbaa148. https://doi.org/10.1093/bib/bbaa148

Pei, Y., Tanguy, M., Giess, A., Dixit, A., Wilson, L. C., Gibbons, R. J., Twigg, S. R. F., Elgar, G., & Wilkie, A. O. M. (2024). A Comparison of Structural Variant Calling from Short-Read and Nanopore-Based Whole-Genome Sequencing Using Optical Genome Mapping as a Benchmark. *Genes*, *15*(7), 925. https://doi.org/10.3390/genes15070925

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep

neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Pray, L. A. (2008). *Discovery of DNA structure and function: Watson and Crick*. Nature Education. http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397

Rani, J., Kumar, S., Saini, M., Mundlia, J., & Verma, P. K. (2016). Biological potential of pyrimidine derivatives in a new era *. *Research on Chemical Intermediates*, *42*(9), 6777–6804. https://doi.org/10.1007/s11164-016-2525-8

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of Next-Generation Sequencing Reads *. *Annual Review of Genomics and Human Genetics*, *16*(1), 133–151. https://doi.org/10.1146/annurev-genom-090413-025358

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Fungtammasan, A., … Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978), 344–354. https://doi.org/10.1038/s41586-023-06457-y

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and its Applications *. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rochette, N. C., Rivera-Colón, A. G., Walsh, J., Sanger, T. J., Campbell-Staton, S. C., & Catchen, J. M. (2023). On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity. *Molecular Ecology Resources*, *23*(6), 1299–1318. https://doi.org/10.1111/1755-0998.13800

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5), R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. https://doi.org/10.1038/nature10242

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., & Brudno, M. (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Computational Biology*, *5*(5), e1000386. https://doi.org/10.1371/journal.pcbi.1000386

Saenger, W. (1973). Structure and Function of Nucleosides and Nucleotides *. *Angewandte Chemie International Edition in English*, *12*(8), 591–601. https://doi.org/10.1002/anie.197305911

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Shin, S., & Park, J. (2016). Characterization of sequence-specific errors in various next-generation sequencing systems. *Molecular BioSystems*, *12*(3), 914–922. https://doi.org/10.1039/C5MB00750J

Shinde, D., Yinglei, L., Fengzhu, S., & Norman, A. (2003). Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and

(A/T)n microsatellites. *Nucleic Acids Research*, *31*(3), 974–980. https://doi.org/10.1093/nar/gkg178

Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., Golding, G. B., & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, *5*(1), 9687. https://doi.org/10.1038/srep09687

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., & Quince, C. (2016). Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, *17*(1), 125. https://doi.org/10.1186/s12859-016-0976-y

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, *43*(6), e37–e37. https://doi.org/10.1093/nar/gku1341

Schmalle, H. W., Hänggi, G., & Dubler, E. (1988). Structure of hypoxanthine. *Acta Crystallographica Section C Crystal Structure Communications*, *44*(4), 732–736. https://doi.org/10.1107/S0108270188000198

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., … Church, D. M. (2016). *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly*. https://doi.org/10.1101/072116

Singh, R., Rai, D., & Prasad, R. (2018). A review on parameterized string matching algorithms *. *Journal of Information and Optimization Sciences*, *39*(1), 275–283. https://doi.org/10.1080/02522667.2017.1374730

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Smolka, M., Rescheneder, P., Schatz, M. C., von Haeseler, A., & Sedlazeck, F. J. (2015). Teaser: Individualized benchmarking and optimization of read mapping results for NGS data. *Genome Biology*, *16*, 235. https://doi.org/10.1186/s13059-015-0803-1

Soukupova, J., Zemankova, P., Lhotova, K., Janatova, M., Borecka, M., Stolarova, L., Lhota, F., Foretova, L., Machackova, E., Stranecky, V., Tavandzis, S., Kleiblova, P., Vocka, M., Hartmannova, H., Hodanova, K., Kmoch, S., & Kleibl, Z. (2018). Validation of CZECANCA (CZEch CAncer paNel for Clinical Application) for targeted NGS-based analysis of hereditary cancer syndromes. *PLOS ONE*, *13*(4), e0195761. https://doi.org/10.1371/journal.pone.0195761

Star, B., Nederbragt, A. J., Hansen, M. H. S., Skage, M., Gilfillan, G. D., Bradbury, I. R., Pampoulie, C., Stenseth, N. C., Jakobsen, K. S., & Jentoft, S. (2014). Palindromic Sequence Artifacts Generated during Next Generation Sequencing Library Preparation from Historic and Ancient DNA. *PLoS ONE*, *9*(3), e89676. https://doi.org/10.1371/journal.pone.0089676

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *. *PLOS Biology*, *13*(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*, *3*(1), lqab019. https://doi.org/10.1093/nargab/lqab019

Tanudisastro, H. A., Deveson, I. W., Dashnow, H., & MacArthur, D. G. (2024). Sequencing and characterizing short tandem repeats in the human genome. *Nature Reviews Genetics*, *25*(7), 460–475. https://doi.org/10.1038/s41576-024-00692-3

Travers, A., & Muskhelishvili, G. (2015). DNA structure and function *. *The FEBS Journal*, *282*(12), 2279–2295. https://doi.org/10.1111/febs.13307

Vasimuddin, Md., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Vozza, G., Bonetti, E., Tini, G., Favalli, V., Frigè, G., Bucci, G., De Summa, S., Zanfardino, M., Zapelloni, F., & Mazzarella, L. (2023). Benchmarking and improving the performance of variant-calling pipelines with RecallME. *Bioinformatics*, *39*(12), btad722. https://doi.org/10.1093/bioinformatics/btad722

Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y., & Benson, G. (2004). Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeats That Contain Testes Genes. *Genome Research*, *14*(10a), 1861–1869. https://doi.org/10.1101/gr.2542904

Watson, J. D., & Crick, F. H. C. (1953). THE STRUCTURE OF DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, *18*(0), 123–131. https://doi.org/10.1101/SQB.1953.018.01.020

Weese, D., Holtgrewe, M., & Reinert, K. (2012). RazerS 3: Faster, fully sensitive read mapping. *Bioinformatics*, *28*(20), 2592–2599. https://doi.org/10.1093/bioinformatics/bts505

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Whiteford, N. (2022, září 27). Cost per Gigabase. *41J Blog*. https://41j.com/blog/2022/09/cost-per-gigabase/

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., & Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, *14*(6), 590–592. https://doi.org/10.1038/nmeth.4267

Wilton, R., & Szalay, A. S. (2023). Short-read aligner performance in germline variant identification *. *Bioinformatics*, *39*(8), btad480. https://doi.org/10.1093/bioinformatics/btad480

Zamenhof, S., Chargaff, E., & Brawerman, G. (1950). DISSYMMETRY IN NUCLEOTIDE SEQUENCE OF DESOXYPENTOSE NUCLEIC ACIDS. *Journal of Biological Chemistry*, *187*(1), 1–14.

Zhang, D., Zhu, L., Wang, F., Li, P., Wang, Y., & Gao, Y. (2023). Molecular mechanisms of eukaryotic translation fidelity and their associations with diseases. *International*

Journal of Biological Macromolecules, *242*, 124680. https://doi.org/10.1016/j.ijbiomac.2023.124680

Zvěřinová, Š., & Guryev, V. (2022). Variant calling: Considerations, practices, and developments *. *Human Mutation*, *43*(8), 976–985. https://doi.org/10.1002/humu.24311