ZÁPADOČESKÁ UNIVERZITA V PLZNI
FAKULTA APLIKOVANÝCH VĚD
KATEDRA KYBERNETIKY

Bakalářská práce

Automatické měření metrik NGS zarovnávacích nástrojů

Plzeň, 2023 David Staníček

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2022/2023

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**David STANÍČEK**
Osobní číslo:	**A19B0384P**
Studijní program:	**B0714A150005 Kybernetika a řídicí technika**
Specializace:	**Umělá inteligence a automatizace**
Téma práce:	**Automatické měření metrik NGS zarovnávacích nástrojů**
Zadávající katedra:	**Katedra kybernetiky**

Zásady pro vypracování

1. Seznamte se s metodami získávání DNA dat (zaměřte se na Next-Generation Sequencing).
2. Prostudujte přístupy a nástroje pro zarovnání NGS dat (výhody, nevýhody, HW nároky, vhodnost použití s ohledem na lidská data).
3. Definujte metriky pro porovnání zarovnávacích nástrojů (přesnost zarovnání, časová/výpočetní náročnost, ...).
4. Vytvořte nástroj pro automatické hodnocení definovaných metrik.
5. Proveďte zhodnocení vybraných nástrojů.

Rozsah bakalářské práce: **30-40 stránek A4**
Rozsah grafických prací:
Forma zpracování bakalářské práce: **tištěná**

Seznam doporučené literatury:

Dle doporučení vedoucí práce

Vedoucí bakalářské práce: **Ing. Lucie Houdová, Ph.D.**
Katedra kybernetiky

Datum zadání bakalářské práce: **17. října 2022**
Termín odevzdání bakalářské práce: **22. května 2023**

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Prof. Ing. Josef Psutka, CSc.
vedoucí katedry

V Plzni dne 17. října 2022

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného akademického titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Západočeská univerzita v Plzni má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V textu jsou použity názvy produktů, technologií, služeb, aplikací, společností apod., které mohou být ochrannými známkami nebo registrovanými ochrannými známkami příslušných vlastníků.

V Plzni dne 29. května 2023 David Staníček

Poděkování

Rád bych tímto poděkoval Ing. Lucii Houdové, Ph.D. za odborné vedení, ochotu, trpělivost a věcné připomínky k této bakalářské práci. Dále bych rád poděkoval Ing. Kateřině Kratochvílové za připomínky k praktickým částem této práce.

Abstrakt

Tato práce se zabývá vytvořením nástroje pro automatické hodnocení zarovnávacích nástrojů pro data, která jsou výstupem sekvenačních systémů nové generace (NGS). V první části jsou shrnuty základní teoretické poznatky o získávání těchto dat. Dále jsou shrnuty některé nástroje a jejich vlastnosti a aktuální situace poznání v tomto směru. V další části jsou definovány hodnotící metriky, poté následuje sada experimentů spolu s výsledným zhodnocením konkrétních nástrojů. Cílem této práce je vytvořit nástroj, který automaticky podrobuje zarovnávací software zkouškám za účelem zhodnocení jejich výkonu, předností a nedostatků.

Klíčová slova

NGS, DNA data, zarovnání NGS dat, automatické hodnocení, editační vzdálenost, NGS zarovnávací nástroje

Abstract

This work deals with the creation of a tool for automatic evaluation of alignment tools for data generated by next-generation sequencing (NGS) systems. The first part provides a summary of the basic theoretical knowledge about obtaining such data. It further summarizes some tools and their characteristics, as well as the current state of knowledge in this field. The next part defines evaluation metrics, followed by a series of experiments and the resulting assessment of specific tools. The outcome of this work is a tool that automatically subjects alignment software to tests to evaluate their performance, strengths, and weaknesses.

Keywords

NGS, DNA data, NGS data alignment, automatic evaluation, editing distance, NGS alignment tools

Obsah

1 Úvod

DNA data v současné době hrají velkou roli v široké škále biomedicínských oborů. Díky výzkumu založenému na analýze DNA dat jsme schopni identifikovat důležité biomarkery hrající roli ve výběru vhodného dárce pro transplantaci, a zvýšit tak pravděpodobnost její úspěšnosti a také snížit riziko následných komplikací. Od odběru vzorku po výsledek analýzy vede dlouhá řada nepřímo navazujících procesů od přípravy vzorku přes proces sekvenace a přípravy dat až po identifikace genových variant a analýzy výsledků.

Tato práce se zabývá oblastí následující po sekvenaci, tedy přípravou dat k další analýze a procesům, tj. problémem sestavení výsledné sekvence z velkého množství dat, která jsou výstupem sekvenačního přístroje využívajícího často používanou technologii next generation sequencing. Pro tyto účely byla vyvinuta celá řada komerčních i volně dostupných softwarových nástrojů. Každý z těchto nástrojů má jinou charakteristiku, ať už se jedná o princip samotné funkce nebo o účel jeho použití. Tato práce byla napsána za účelem porovnání těchto softwarových nástrojů, popisu výhod, úskalí a vhodnosti jejich použití v konkrétních oblastech výzkumu především lidské DNA.

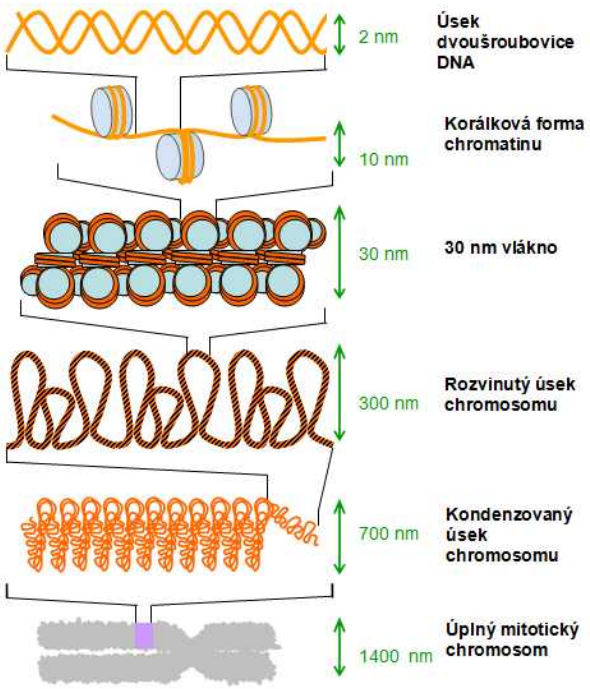
Za tímto účelem byly definovány metriky, vybrány specifické softwarové nástroje, definována vstupní data a byly provedeny experimenty, na jejichž základě byly nástroje ohodnoceny s důrazem na jejich přesnost, hardwarovou náročnost a časovou náročnost pomocí automatického hodnotícího programu.

Získané výsledky mohou být dále využity pro rozhodování ve věci výběru softwaru ke konkrétnímu výzkumnému účelu nebo jako základ pro větší a rozsáhlejší projekty týkající se této problematiky.

2 Metody získávání DNA dat

2.1 Co jsou DNA data

DNA daty se rozumí sekvence za sebou jdoucích písmen označující nukleotidové báze, ze kterých je složena molekula DNA. Molekula DNA se skládá ze 4 bází. Jsou to adenin, cytosin, thymin a guanin (A,C,T,G). Tyto molekuly DNA se nacházejí vždy v buněčném jádře, kde jsou součástí větších shluků - chromosomů. Každý chromosom je tvořen jednou molekulou DNA a komplexem bílkovin - histonů, které jsou navázané na DNA, aby umožnily svinutí DNA do kompaktnější struktury.[1]



Obrázek 1: Struktura chromosomu (převzato [1])

Vlastní molekula DNA je stočena do tzv. dvoušroubovice. Ve dvoušroubovici jsou tedy vždy obsaženy 2 navzájem antiparalelní řetězce bází. Báze jsou navzájem spojeny tak, že respektují tzv. komplementaritu bází, což znamená, že adenin se bude v DNA

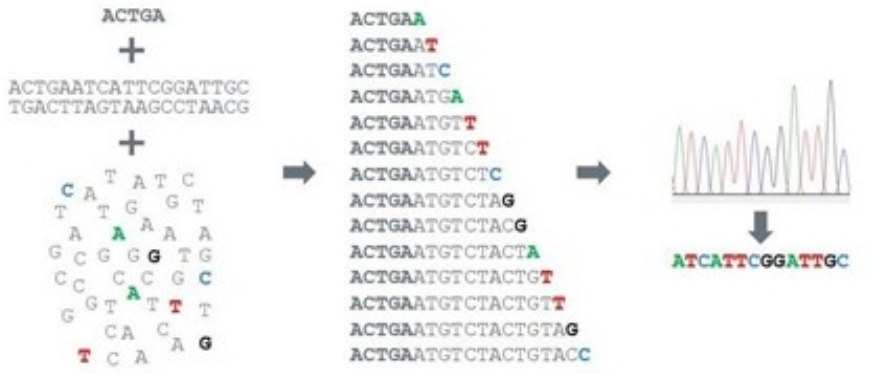
párovat pouze s thyminem, guanin pouze s cytosinem a naopak. Díky tomuto pravidlu nám tedy k reprodukci celé dvoušroubovice stačí znát jen její jednu polovinu.

Vlastní sekvence DNA tvoří tzv. genetický kód. Genetický kód je univerzální pro většinu živých organismů. Genetický kód je tripletový. Triplet neboli kodon je tvořen sekvencí 3 za sebou jdoucích bází. Speciální kodony jsou iniciační kodon (většinou AUG), podle kterého poznáme začátek genové sekvence, a stop kodon (terminační kodon, UAG, UGA, nebo UAA), který sekvenci ukončuje.

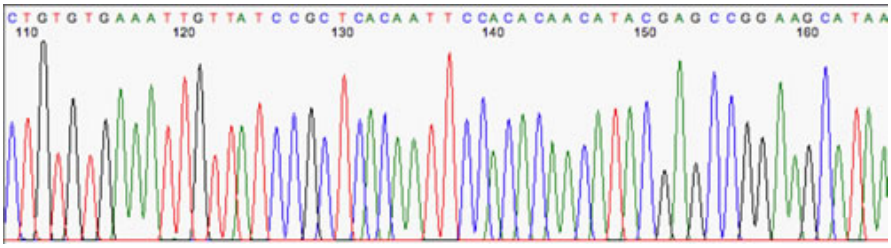
DNA data jsou tedy označení, které souhrnně definuje pořadí bází, jež tvoří jednotlivé kodony, a dohromady tak genetický kód. [2]

2.2 Přístupy k získávání DNA dat 1. generace

V roce 1977 objevil Frederick Sanger metodu sekvenování, která spočívala v opakované syntéze nových řetězců DNA podle stejného vzoru - templátu, přičemž při konci syntézy daného řetězce je do něj náhodně přiřazena modifikovaná nukleotidová báze, která zapříčiní pokračování další syntézy nově vznikajícího vlákna. Tyto modifikované báze lze navázat na zbytek řetězce, ale na ně samotné již další navázat nejdou, tudíž jejich navázáním se syntéza zastavuje. Při Sangerově sekvenování nese každá modifikovaná báze fluorescenční barvu, podle které poznáme, o kterou bázi se jedná. Tímto náhodným přerušením získáme různě dlouhé molekuly v řádech desítek, stovek až jednoho tisíce bází. Na základě velikosti se pak jednotlivé syntetizované řetězce řadí podle velikosti. Při dostatku různě dlouhých náhodně zakončených řetězců pak lze na základě obarvení přečíst sekvenci bází, a získat tak DNA data. Výhodami Sangerovy metody je poměrná přesnost. Nevýhodou této metody je její vysoká nákladnost a nízká rychlost. Navíc pomocí Sangerovy metody lze číst pouze jeden předem daný úsek DNA, a je tedy nevhodná pro čtení např. celého genomu najednou. [3]



Obrázek 2: Schéma Sangerovy metody (převzato, upraveno [4])



Obrázek 3: Výstup čtení - chromatogram (převzato [5])

2.3 Přístupy k získávání DNA dat nové generace

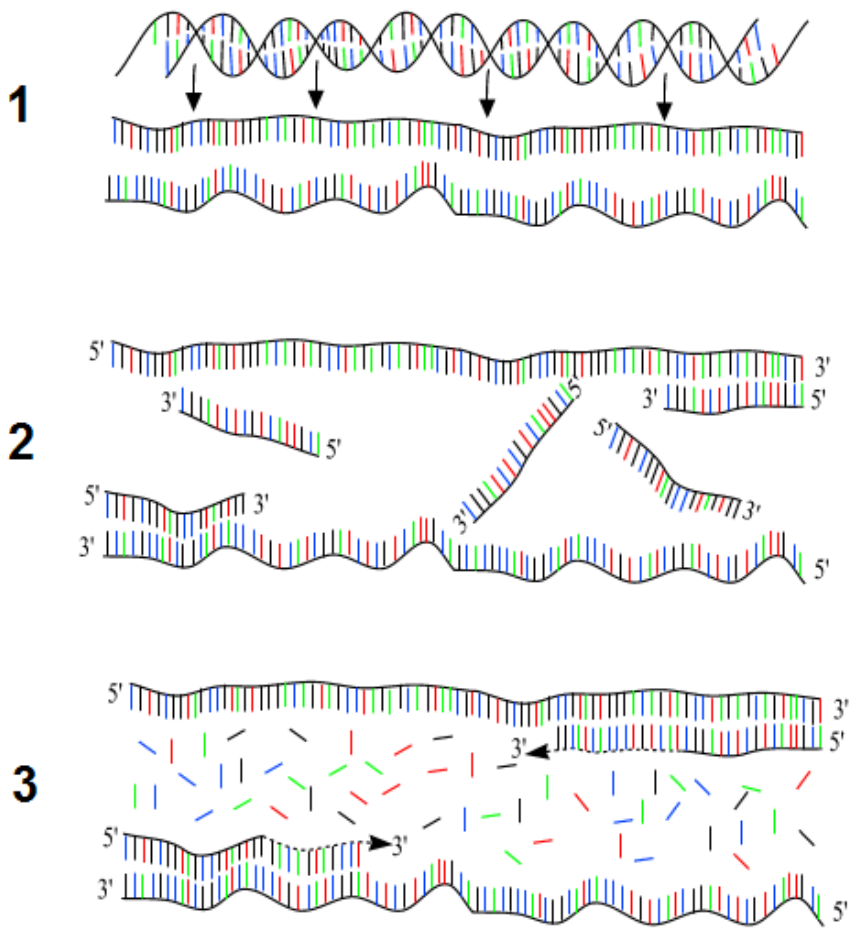
2.3.1 Základní principy

Zásadní rozdíl mezi první a druhou generací sekvenování genomických dat spočívá především v jejich množství. Zatímco Sangerovou metodou lze sekvenovat a získat data pouze z malého fragmentu DNA, pomocí metod nové generace lze paralelně sekvenovat velké množství částí genomu najednou. Pokud pro srovnání použijeme dobu nutnou pro sekvenování celého lidského genomu, pak Sangerovou metodou by toto zabralo desetiletí, zatímco s použitím metod nové generace je to možné za jediný den.[6] Základní princip spočívá ve fragmentaci vstupní DNA na krátké úseky (desítky až stovky bází) a jejich amplifikace PCR metodou. Tyto fragmenty jsou poté čteny paralelně. Princip

samotného určování bází je pak velmi podobný tomu, který je používán v Sangerově metodě. [7]

2.3.2 PCR amplifikace

Amplifikace se dá volně přeložit jako zesílení, nebo významově přesněji namnožení daného vzorku. PCR - polymerázová řetězová reakce - probíhá v přístroji jménem termocykler, což je zařízení schopné velmi rychle měnit vnitřní teplotu. Jedná se o levnou a účinnou metoda amplifikace DNA, nebo RNA. Základní princip PCR spočívá v řetězové reakci. Jedna molekula DNA je okopírována na dvě nové, ty jsou dále okopírovány, čímž vznikají 4 nové a takto proces pokračuje dále, dokud nevznikne požadované množství stejných molekul. PCR se dá rozdělit na 3 základní kroky, kterými jsou denaturace, nasednutí primerů a syntéza DNA. V prvním kroku je molekula dvoušroubovice podélně rozdělena na úrovni bázových párů na 2 samostatné molekuly za teploty 90-97 stupňů Celsia. Ve druhém kroku dojde k přilnutí tzv. primerů (počáteční prvky určující začátek replikace), a to za teploty snížené na 50-60 stupňů Celsia. Ve třetím kroku se na konec připojených primerů začnou za teploty 72 stupňů Celsia připojovat nové báze při respektování pravidla komplementarity. Tento proces trvá 2 až 5 minut. Proces pak může být libovolně opakován pro dosažení potřebného množství kopií.[8]



Obrázek 4: Ilustrace PCR metody (převzato [9])

2.3.3 Čtení bází

Samotné čtení bází velmi závisí na použité technologii. V současnosti je asi nejvíce používaná technologie vyvinutá společností Illumina, která dokáže sekvenovat až 900 miliard bází najednou, a stojí tedy nyní na první příčce ve smyslu maximálního možného výkonu. Při Illumina sekvenování jsou jednotlivé molekuly DNA nastříhány a přichyceny na destičku, kde jsou následně amplifikovány pro vznik homogenních shluků. Podobně jako u Sangerovy metody jsou do rostoucích řetězců průběžně přidávány báze s fluorescenční barvou, jejichž napojení zastavuje syntézu. Tato barva je pak díky amplifikaci do shluku čtena kamerou. Po přečtení barvy dojde k odstranění barvení i ter-

minační báze a reakce takto iterativně pokračuje. Toto čtení, které funguje na základě metod automatického rozpoznávání obrazu, probíhá najednou pro všechny molekuly, které jsou přichyceny na destičce. Každý krok, kdy je použito obarvení a odbarvení, je zaznamenáván a díky tomu je pak počítačově možné rekonstruovat celou sekvenci. Metoda společnosti Illumina se vyznačuje velmi vysokou přesností.[3]

Dalšími používanými metodami jsou metody 454 nebo Ion Torrent, které byly průkopnické ve smyslu sekvenování nové generace. Fungují na obdobném principu jako novější Illumina s tím rozdílem, že molekuly jsou amplifikovány na miniaturních kuličkách. Čtení probíhá u sekvenování 454 na principu sledování světelných změn, u metody Ion Torrent na principu sledování intenzity změn pH.[10][11]

V poslední době se principy čtení bází začínají ubírat směrem tzv. metod sekvenování třetí generace. Existují 3 aktuálně rozšířené dostupné komerční platformy, mezi které patří PacBio Single Molecule Real Time, Illumina Tru-seq Synthetic Long-Read technology a Oxford Nanopore Technologies sequencing platform. Velký rozdíl oproti metodám druhé generace spočívá v tom, že původní molekula není nijak amplifikována, a báze jsou tedy čteny jen z jedné molekuly. Metoda vyvinutá společností PacBio využívá principu fluorescenčního značení nukleotidů s velmi vysokou citlivostí. Metoda společnosti Oxford Nanopore využívá principu průchodu molekuly porézní strukturou, přičemž při průchodu každé báze se měří, jak moc se daná báze "vešla"do zmíněné struktury. Na základě toho se pak vyhodnotí, o jakou bázi se může jednat. Tyto metody se momentálně v praxi ještě příliš nevyužívají, a to zejména kvůli jejich chybovosti a prozatím nedostatečnému technologickému zázemí. [12]

2.3.4 Vstupní formáty

Přečtené báze počítač ukládá ve formátech, které jsou k tomu účelu vhodné vzhledem k další práci. Nejčastějšími formáty pro uchování dat jsou formáty FASTA a FASTQ. Oba formáty dokáží v jednom souboru uchovávat více než jednu sekvenci. Formát FASTQ uchovává sekvenci a informace o její kvalitě a obvykle potřebuje 4 řádky pro zápis jedné sekvence. Znakem @ začíná řetězec obsahující informace a identifikátory

sekvence. Na dalším řádku následuje samotná sekvence a za ní na dalším řádku znak +. Na poslední řádek pak formát FASTQ ukládá informace o kvalitě předešlé sekvence. [13] Data ve formátu FASTQ jsou typicky výstupem sekvenačních systémů.

```
@HLA:HLA01013-5100/1
ATGCAGACTGCCTGCAGGAACTACGGCGATATCTAAAATCCGGCGTAGTCCTGAGGA
GAACAGGTACCGACGCTGGCCAGGGGCTCTCCTCTCCCTCCAATTCTGCTAGAGTTG
CCTCACCTCCCAGATGTGTCCAGGGAAACCCTCCCTGTGCTATGGATGAAGGCATTT
CCTGTTGGCACATCGTGTCCTGATTTTCCTCTATTGTTAGAGCCACTGGATAAAGAC
AGTGGGTCAGGGACTGGACCAT
+
?A???BBBB<D+DDDDGFFFGGICIIIHFIFHIGIHHDIHHIHIIHHHCHHIFEHH*
IHIHIIIHEFHHHIIFGIIIIIIBIHHH7E?IIIHIIHHHCHGGHGFIHHFHHHFGH
-DFHEFHFFEDFGBDGG?FGGGFIEFGFGGGGGC8EG<GFE*ECGEEGGCGAEFEE4
)FEGGG;GEGGGGE:GE:GEEGFG;EGDGGGGCGG?EE?EAHCEH;6EEE:?GGGGE
GEG(GCCGEFE?GG?EGEGG?G
```

Listing 1: Vizualizace syntetických readů alely HLA01013 ve formátu FASTQ

Formát FASTA obsahuje jednořádkový popis sekvence, který je na dalším řádku následován samotnou sekvencí. Popisový řádek vždy začíná symbolem >.[14] Ve formátu FASTA jsou často uchovávány referenční sekvence.

```

>IPDMHCgen:DLA04814 DLA04814.1
GACCATGTTGCCTACTACGGCATAAATGTCTACCAGTCTTACGGTCCCTCTGGCCA
GTACACCCATGAATTTGATGGCGATGAGGAATTCTACGTGGACCTGGAGAAGAAGG
AAACTGTCTGGCGGCTGCCTGTGTTTAGCACATTTGCAAGTTTTGACCCACAGGGT
GCACTGAGAAACTTGGCTATAGCAAAACAAAACTTGAACATCATGACTAAAAGGTC
CAACCAAACTGCTGCTACCAATNN
```

Listing 2: Vizualizace alely DLA04814.1 ve formátu FASTA

2.3.5 Výstupní formáty

Formát SAM (Sequence Alignment Map) je textový formát, který uchovává zarovnané sekvence. Je používán zejména pro ukládání dat získaných metodami sekvenování

nové generace (NGS), kterými se tato práce zabývá. Většina NGS zarovnávacích nástrojů produkuje výstup ve formátu SAM. Formát BAM (Binary alignment map) je obdobou k formátu SAM s tím rozdílem, že data uchovává v binární komprimované podobě. [15]

```
1
2 HLA:HLA26896-8  167 * 1388  99  119=1X57=1X42=1X29= * 1414  276
     AGTGCTGGAGCATATGACAGTGCTGAGCATCTTTCCCAAGCCCCACCCTCCCCCAGA
3 GCACCCTCCCCTCCTGTCCTCACCCTACCCCAAGTTCTCCCACAGTCACTCCTGCCC
4 CATGCACATGCCGCCCTCCAGTTCTTGCTCTGCCCATCTCCCCTCCCCAACCCAGAC
5 CTAAAAAAGGCTGTTGGGCCAACTGTTCCTTGACCTTCCTTCTTTTCTTGTGGTTCC
6 TTGACCCCAGTGGGCTCTCACT  ???,?BB3DD-BDDDD>GCGGF;>FFHI/HCHE
7 FI>C.EGIHFI@GIFIFHAGFAHE8H=G+IDFBGF
8 FHHIDIHDFDEIACHFHHHAECFEFHI5FID8HDFGHHHEHHGEHDHGC:E,FGD=F
9 EBEEHDHG*DG:GEEBEEGFE(GFCEG=EF(CEGGE?AECE?GCGGGFGCEG+E>G;
10 FCGG?/CG:GG?GG/A?8EGGG=GH?2H*?(A*FE(,8*GG4/A:CE:EE:E;EA?E
11 (G:CA-?*/:6
```

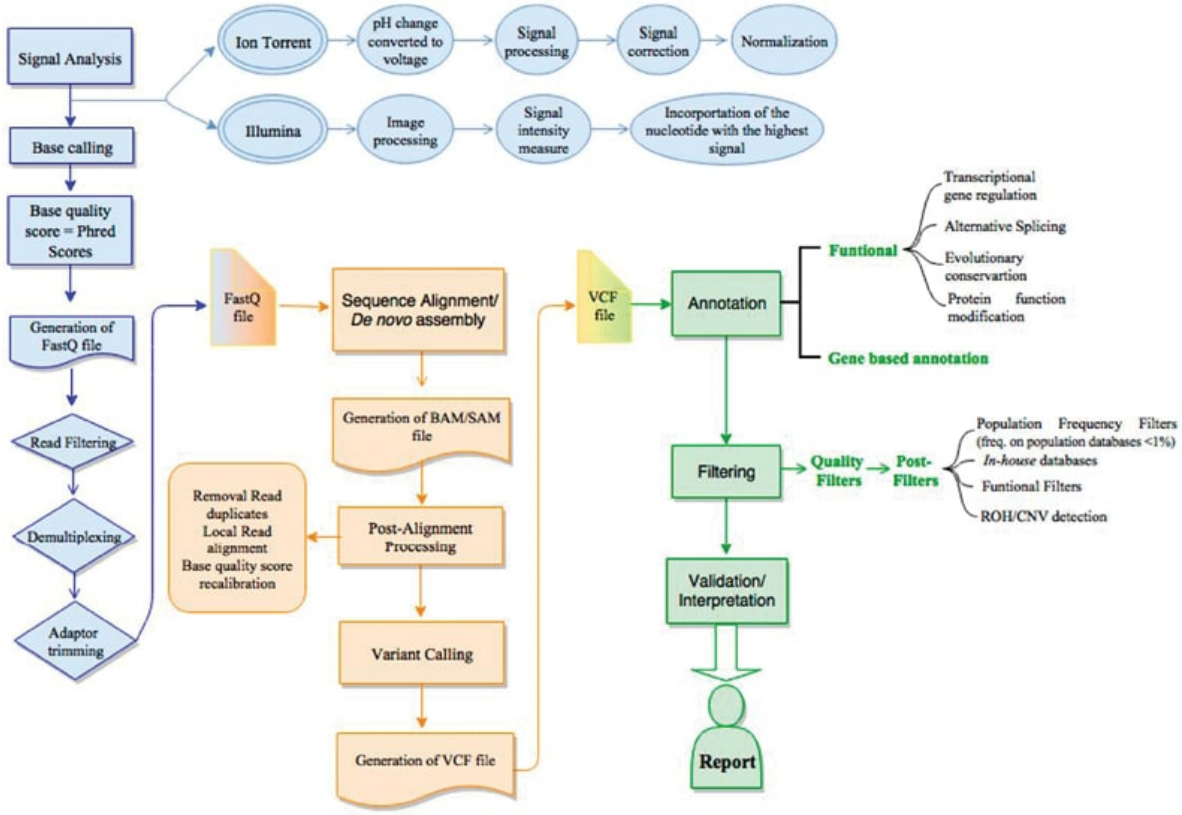
Listing 3: Příklad vizualizace výstupu z generátoru syntetických NGS dat zobrazených pomocí nástroje SAMTools

```
1
2 HLA:HLA26896-12 147 HLA:HLA26896  1179  1 250M  = 1137  -292
     CTCTGGACCCTGACTCATTGTGGCCCCTTCCCCACTCCTTCATCCTCAGCCTCACCCT
3 CTTGAGGACCCCACCCTCCAGCCCACAGGTGCTGGACCATCCCTCCCTGGTCCCTCCG
4 CCCCTCTCCACCTTGTGACCTTGTGCTGCTCCTATCTCTTGCCCAGCTGCCTTGGGCC
5 CTCAGCACGTTCTCATCTTTCAGTGGGAAAGTGGGAGTGCTGGAGCATATGACAGTGC
6 TGAGCATCTTTCCCAAGC;:CE*CG0ECC*(6E?:EGDC6::CE8.EFC:0EECGC
7 E6HG.C(0G1AG(4EEE-D(4GGFCEGG:EE(FE(G@D6EE(EC;<GEG8??EGCEEG
8 E;G;G0EG9@FGGBGDEGEC:GHGGDHEDEHHEGEHCHGHH#IGHE=3H,HC
9 EIEH3GHIHH?IHIHI8=HHHHGF-HIIICGECHIHIHEIIEHCCFH@H=EFHF,HIH
10 CEFHHD>FGGCHEAHFEIIHHIGGF
11 FF@D@DDD<+BAB-????5 AS:i:-2 XS:i:-2 XN:i:0  XM:i:1  XO:i:0
12 XG:i:0  NM:i:1  MD:Z:12C237 YS:i:0  YT:Z:CP
```

Listing 4: Vizualizace výstupu z nástroje Bowtie2 ve formátu SAM

2.3.6 Zpracování dat

Zpracování DNA dat obvykle začíná analýzou vzorku příslušnou technologií, které byly již popsány v samostatných kapitolách o způsobech získávání DNA dat. Takto získaná data jsou příslušným systémem navíc ohodnocena na základě jejich předpokládané kvality, je provedena filtrace a předzpracování dat a je z nich vygenerován soubor ve formátu FASTQ. Takovýto soubor obsahuje vstupní data, která jsou podstatná v rovině této práce. Na těchto datech je provedeno zarovnání vůči referenční sekvenci, pokud takovou máme k dispozici. Jestliže není referenční sekvence k dispozici, je možné použít metodu de novo assembly, což je sestavení a uspořádání sekvence bez apriorní znalosti referenční sekvence. Výstupem obou těchto procesů je soubor ve formátu BAM, nebo SAM. Následná práce s daty může zahrnovat další zpracování dat jako např. seřazení, filtrace duplicitních readů, identifikace konkrétních variant daných genů atd. Takto připravená data mohou být následně předána k použití ke konkrétním klinickým/výzkumným účelům.



Obrázek 5: Obecná formulace práce s DNA daty(převzato [16])

3 Přístupy a nástroje pro zarovnání NGS dat

Při počítačovém zpracování výstupů ze sekvenačních přístrojů je očekávaným výsledkem souvislá sekvence za sebou jdoucích bází. Vzhledem k tomu, že principem NGS je sekvenování velkého množství menších molekul, je nutné z nich souvislý řetězec sestavit. Tento proces v drtivé většině případů začíná zarovnáním získaných readů vzhledem k referenční sekvenci.

3.1 Obecná formulace problému zarovnání readů

Obecný problém zarovnání spočívá v tom, aby se read zarovnal na správné místo vzhledem k referenci na základě shody řetězců readu a reference. Je nicméně nutné, aby

algoritmus respektoval i možnost, že byl read přečten s chybou nebo nějakou chybu či genetickou nesrovnalost sám skutečně obsahoval, a přesto se zařadil na správné místo podle reference. Je také nutné počítat s tím, že eukaryotický genom obsahuje opakující se sekvence. Pokud tedy náhodou dostaneme 2 různé ready, které ale obsahují totožnou sekvenci, je nemožné určit, kam přesně je zařadit. [17]

3.2 Rozdělení algoritmů pro zarovnání dat

Nástroje pro zarovnání dat se dají rozdělit podle principu jejich funkce na nástroje využívající Burrowsovu-Wheelerovu transformaci (BWT) a na nástroje založené na principu hash tabulek. Nově se se stoupajícími nároky na výkon objevují i hybridní nástroje využívající současně obou těchto principů.

3.2.1 Burrowsova-Wheelerova transformace

BWT je metoda využívaná pro vyhledávání v textu. Metoda restrukturalizuje řetězec tak, aby byl snadněji komprimovatelný. V prvním kroku jsou vytvořeny všechny permutace vstupního řetězce. V dalším kroku jsou tyto rotace lexikograficky seřazeny s tím, že se předpokládá, že symbol $ je v abecedě jako první. Tímto vznikne tzv. Burrowsova-Wheelerova matice. Její poslední sloupec je potom výstup BWT. Následující mnou zpracované vizualizace ukazují základní kroky algoritmu BWT.

$	g	o	o	g	l	e
e	$	g	o	o	g	l
l	e	$	g	o	o	g
g	l	e	$	g	o	o
o	g	l	e	$	g	o
o	o	g	l	e	$	g
g	o	o	g	l	e	$

Obrázek 6: Vizualizace BWT, vytvoření všech permutací řetězce

$	g	o	o	g	l	**e**
e	$	g	o	o	g	**l**
g	l	e	$	g	o	**o**
g	o	o	g	l	e	**$**
l	e	$	g	o	o	**g**
o	g	l	e	$	g	**o**
o	o	g	l	e	$	**g**

Obrázek 7: Vizualizace BWT, abecední seřazení, zvýrazněný sloupec reprezentuje výstup: BWT($google)=elo$gog

Princip reverzibility BWT spočívá v její vlastnosti, která se nazývá First-Last mapping. Tato vlastnost přiřazuje každému znaku v řetězci číslo reprezentující počet předchozích výskytů daného znaku v řetězci. Posloupnost těchto čísel se zachovává v prvním i v posledním sloupci B-W matice.[18]

3.2.2 Hash tabulky

Hash tabulka (rozptylová tabulka) je datová struktura pro uchovávání datových dvojic ve schématu klíč-hodnota. Hash tabulka je postavena nad polem omezené velikosti, pole tedy nepopisuje celý stavový prostor klíče, díky čemuž se razantně snižují paměťové nároky při provádění úkonů vyhledávání v tabulce podle klíče. Pro adresaci využívá hash tabulka tzv. hash (rozptylovou) funkci. Požadavky na hash funkci jsou následující:

1. funkce vrací konzistentně pro stejný objekt vždy stejnou adresu
2. funkce nezaručuje, že pro dva různé objekty vrátí různou adresu
3. funkce využije celého prostoru adres, a to se stejnou pravděpodobností
4. výpočet adresy proběhne velmi rychle.

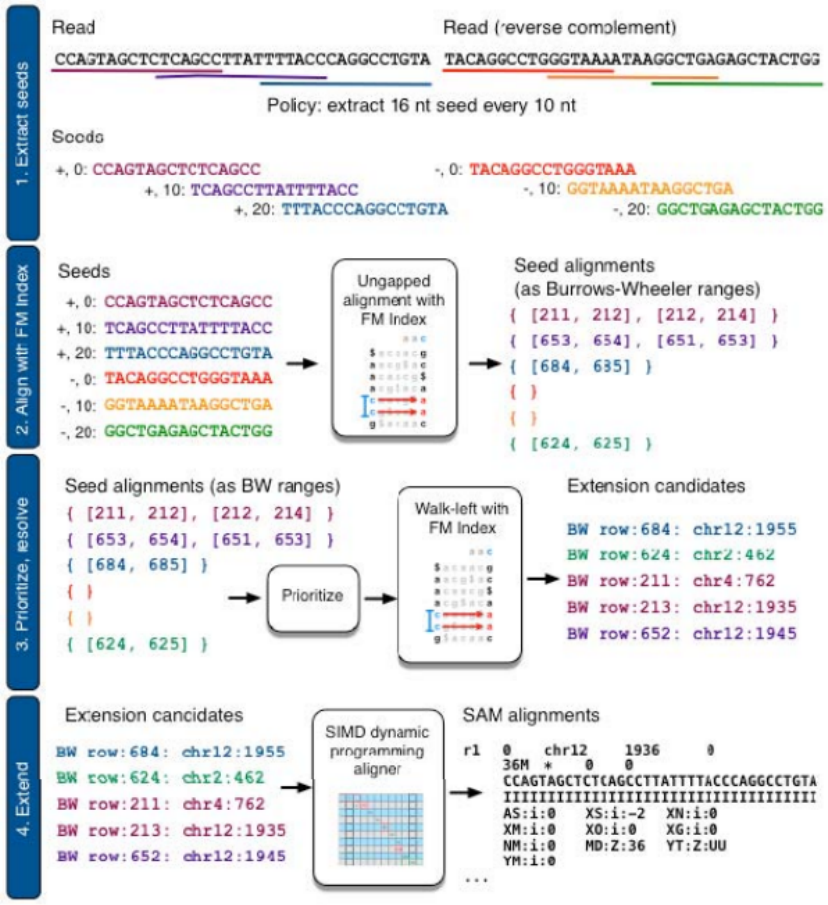
Z druhé vlastnosti vyplývá, že na jednu adresu lze uložit více objektů. Této vlastnosti se říká kolize, záleží však na definici účelu, pro kterou byla tabulka vytvořena. Např. pro náš účel zarovnání readů z NGS se využívají výhradně tzv. bezkolizní tabulky.[19]

3.3 Nástroje na zarovnávání sekvenovaných dat

3.3.1 Bowtie2

Bowtie2 je rychlý a paměťově efektivní nástroj pro zarovnání sekvenovaných dat. Je vhodný k zarovnávání readů o rozměrech od 50 do řádově tisíců párů bází. Pro zarovnávání využívá BWT, FM indexaci a backtracking. [20] Bowtie2 zpracovává každý read ve čtyřech krocích. Nejprve z readu extrahuje tzv. seedy. V dalším kroku extrahované seedy zarovná podle reference pomocí full-text minute index, a to bez mezer. V kroku 3 jsou seedy ohodnoceny na základě Burrows-Wheelerovy vzdálenosti, přičemž vyšší priorita je přiřazena sloupcům s nižším ohodnocením. Algoritmus poté sloupce náhodně prochází a na základě jejich offsetu je řadí na referenčním genomu. V kroku 4 dochází k

úplnému zarovnání provedením dynamického akcelerovaného programování SIMD. [21]



Obrázek 8: Schéma zarovnání pomocí Bowtie2 (převzato [21])

3.3.2 BWA

BWA (Burrows-Wheeler aligner) je zarovnávací nástroj založený na výše zmíněném principu BWT. Jedná se o poměrně jednoduchý nástroj pro zarovnávání různě dlouhých readů, který lze získat v různých verzích, a to např. BWA-MEM, BWA-backtrack a BWA-SW, přičemž každá verze má své jedinečné využití. BWA-MEM je vhodný pro použití na výstupech z technologií Illumina, IonTorrent, Roche 454 na readech, které jsou v průměru delší než 70 bp. Pro ready, které jsou kratší, je vhodnější verze BWA-

backtrack. BWA-SW je vhodnější použít pro takové výstupy, kde jsou časté mezery mezi ready. Díky využití principu Burrowsovy-Wheelerovy transformace a využití prohledávání řetězců pomocí stromových struktur je algoritmus paměťově efektivní, což obecně platí pro všechny nástroje založené na BWT včetně výše zmíněného Bowtie2. Ohledně časové efektivity je BWA méně efektivní, než Bowtie2, ale i tak jsou časové nároky únosné. Výstupem z BWA jsou zarovnané sekvence ve formátu SAM, který lze vizualizovat, případně upravovat např. pomocí nástroje SAMTools.[18][22]

3.3.3 mrFAST

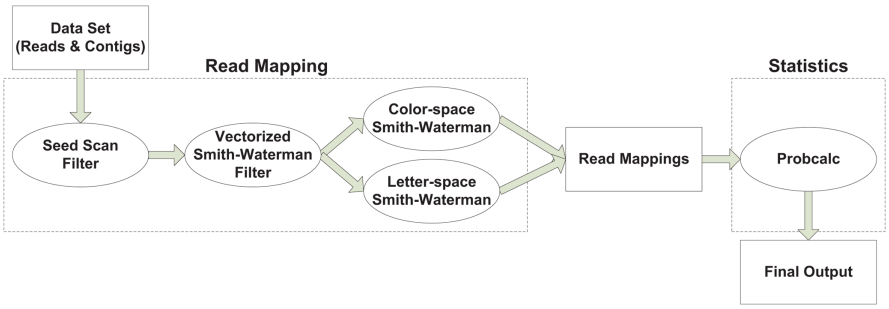
mrFAST (micro-read Fast Alignment Search Tool) je zarovnávací nástroj pro zarovnávání krátkých readů (kratších než 25 bp) se speciálními funkcemi pro vyhledávání strukturálních variací a duplikací segmentů. mrFAST je primárně vyvinut pro výstupy z technologie Illumina. Na rozdíl od již zmíněných algoritmů Bowtie2 a BWA není mrFAST založen na BWT, ale využívá k mapování na referenční sekvenci hash tabulku. Nástroj mrFAST převádí 4 různé možné páry bází do 2 bitů, kde A je kódováno jako 00, C jako 01, G jako 10 a T jako 11. Takto zakódovanou hodnotu pak použijeme jako hodnotu hash daného umístění v tabulce. Hashovací tabulka je bezkolizní, nabízí tedy velmi rychlé vyhledávání klíče. Jak bylo zmíněno, mrFAST je vytvořen speciálně pro výstupy z přístroje Illumina, jehož výstupní ready mají vždy jednotnou délku. Proto je nutné, aby i případné syntetické ready pro testování nástroje měly jednotnou délku. [23]

3.3.4 MAQ

MAQ (Mapping and Assembly with Qualities) je nástroj pro zarovnávání krátkých readů, který je navržen pro práci s výstupy z technologie Illumina a také z technologie SOLiD. Nástroj MAQ při zarovnání odhaduje chybu zarovnání každého readu. U každého readu, který je zarovnán podle referenční sekvence, je tedy vyhodnoceno, na kolik procent byl zařazen (ne)správně a podle tohoto skóre lze tedy v budoucnu špatně zařazené ready vyřadit. MAQ je schopen zarovnání celého lidského genomu.[24]

3.3.5 SHRiMP

SHRiMP (SHort Read Mapping Package) je nástroj, který byl původně vyvinut pro zarovnávání krátkých readů. Software byl navržen pro zpracování výstupů z technologie SOLiD společnosti Applied Biosystems. Nástroj SHRiMP je účelově navržen pro práci s daty, ve kterých se vyskytuje velké množství polymorfismů (polymorfismus = stav, kdy pro určitý genetický znak existuje 2 a více alel - genetických variant). [25]



Obrázek 9: Schéma algoritmu SHRiMP (převzato [25])

3.3.6 Novoalign

Novoalign je zarovnávací software společnosti NovoCraft, který pracuje na základě principu rozptylové tabulky. Software je dostupný pod komerční licencí, nebo pod volně dostupnou licencí pro akademické účely avšak jen do verze Novoalign v3 s omezením (aktuální je verze Novoalign v4). [26]

3.3.7 GEM3

GEM-Mapper v3 je výkonný mapovací nástroj navržený pro zarovnávání sekvenovaných readů na velké referenční genomy (např. celý lidský genom). Je navržen zejména pro velmi dlouhé ready (až 1000 bp). GEM3 funguje na principu Burrows-Wheelerovy transformace a FM indexace. GEM3 je distribuován pod licencí GPLv3 a je dostupný pro operační systémy Linux a MacOS.[27]

3.3.8 Kart

Kart je zarovnávací nástroj, který byl vyvinut především kvůli narůstajícím nárokům na softwarové nástroje s postupným příchodem sekvenačních metod třetí generace. Výše popsané nástroje pracovaly buď s principem BWT, nebo s principem využití hash tabulky, přičemž oba přístupy mají své klady a zápory. Nástroj Kart tyto dva přístupy kombinuje za účelem optimalizace jak paměťových, tak i časových nároků kladených na zarovnávací nástroje pro NGS data. Kart rozdělí ready do dvou skupin, a to na skupinu se snadným zarovnáním a na skupinu s nutností zarovnání s mezerami. Každou z těchto skupin zarovná nezávisle a až poté provede finální zarovnání. Oproti předchozím nástrojům je Kart vhodný spíše pro využití na delších readech. S ohledem na časovou a výpočetní náročnost je nástroj Kart oproti nástrojům založeným na BWT (BWA, Bowtie2) časově značně efektivnější, nicméně má větší paměťové nároky.[28]

3.3.9 Cloudové nástroje

V poslední době se začínají stále častěji využívat cloudové nástroje, jejichž potenciál tkví především ve zvýšení výpočetního výkonu a paměti hardwaru, na kterém je zarovnávání prováděno. Nejvíce cloudových nástrojů je momentálně hostováno na cloudových službách společnosti Amazon. [29]

4 Metriky pro hodnocení zarovnávacích nástrojů

Většina publikací zabývající se problematikou NGS zarovnávacích nástrojů, např. [30], [31], [32] se v různém pojetí zabývají třemi metrikami: čas, přesnost, paměťová náročnost. Následující podkapitoly popisují možné použití a pojetí těchto vybraných metrik.

4.1 Hammingova vzdálenost

Hammingova vzdálenost definuje počet pozic, na kterých jsou dva porovnávané řetězce rozdílné.[33] Například Hammingova vzdálenost řetězců RAIN a SHINE se rovná 3. Tato metrika není příliš vhodná pro účely porovnávání genomických sekvencí, což lze demonstrovat na následujícím příkladu porovnání dvou sekvencí:
v=”GCGCGCGC”, w=”CGCGCGCG”
$d_H = (v, w) = 8$
x=”GCGCGCGC-”, y=”-CGCGCGCG”
$d_H = (x, y) = 2$

4.2 Levenshteinova vzdálenost

Levenshteinova vzdálenost (jinak také editační vzdálenost) definuje míru odlišnosti řetězců pomocí počítání minimálního množství potřebných operací vložení, mazání a substituce tak, aby výsledné řetězce byly totožné, jako například při výpočtu vzdálenosti řetězců RAIN a SHINE :

$$RAIN \rightarrow SAIN \rightarrow SHIN \rightarrow SHINE$$

Při transformaci řetězce RAIN na řetězec SHINE byla dvakrát použita operace substituce a jednou operace vložení. Levenshteinova vzdálenost těchto řetězců je tedy 3.[34]

4.2.1 Definice

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases}$$

Převzato z [35]

4.3 Damerau-Levenshteinova vzdálenost

Damerau-Levenshteinova vzdálenost se oproti Levenshteinově vzdálenosti liší v tom, že navíc zavádí operaci transpozice, tedy záměnu dvou vedlejších znaků. Tato vlastnost sice nemá vliv na průměrnou relativní odchylku, která je zásadní pro tuto práci, ale je užitečná z hlediska dalšího možného využití či rozšíření nástroje pro práci s reálnými daty a získávání validních informací. Operace transpozice vedlejších znaků má význam z hlediska biologické podstaty vstupních dat. Vezmeme-li příklad zmiňovaný v předchozím textu, Damerau-Levenshteinova vzdálenost slov RAIN a SHINE je stále 3. Pokud ale uvažujeme jiný příklad, třeba řetězce CTCG a TCCG, což mohou být reálně hodnocené části genomické sekvence, dostaneme pro klasickou Levenshteinovu vzdálenost výsledek 2 (smazání, vložení), ale pro Damerau-Levenshteinovu vzdálenost výsledek 1, jelikož se použila pouze jedna operace transpozice.

4.3.1 Definice

$$d_{a,b}(i,j) = \begin{cases} max(i,j) \\ min \begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 ... i,j > 1, a_i = b_{j-1} \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \\ d_{a,b}(i-2,j-2)+1 \end{cases} \end{cases}$$

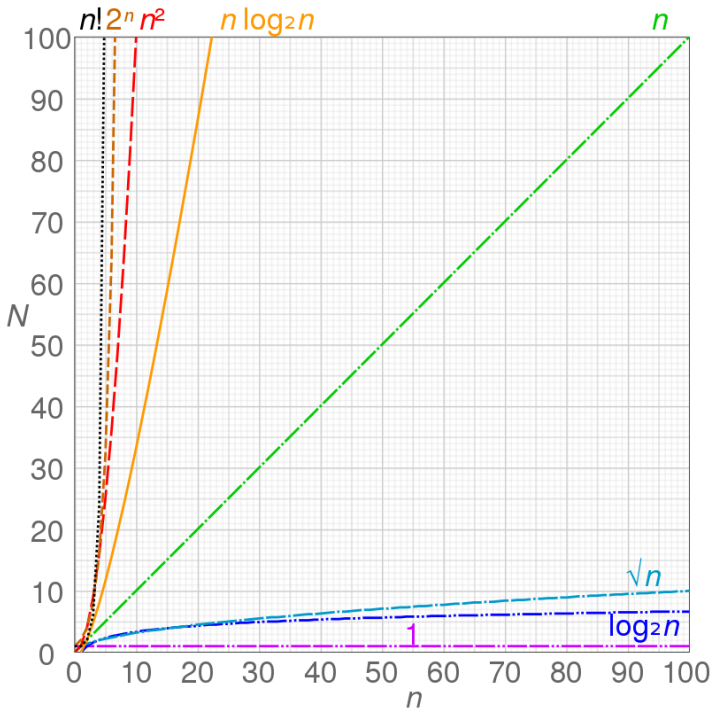
Převzato z [35]

4.4 Využití vzdálenosti řetězců v hodnocení zarovnávacích nástrojů

Spočítáním vzdálenosti řetězců dokážeme určit míru nepřesnosti zarovnání, které daný nástroj provedl. Ačkoliv ready mohou obsahovat chyby např. v podobě mutací, díky identitě vstupních dat na tyto chyby nemusíme brát ohled. Pokud by nástroje měly naprosto stejnou přesnost, vyšla by průměrná vzdálenost řetězců u všech nástrojů stejná, ale nejspíše nebude nulová právě kvůli možnosti výskytu těchto chyb. Hodnocení nepřesnosti nástroje tedy provádíme na základě odchylky od průměrné hodnoty vzdálenosti řetězců ve výstupech jednotlivých nástrojů. Pro vlastní implementaci do porovnávacího nástroje byla vybrána metrika Damerau-Levenshtein kvůli vlastnostem, které jsou popsány v kapitole 4.3, tedy že nejvíce vyhovuje obecnějšímu využití v dalších možných aplikacích a také z důvodu možného rozšíření do další práce.

4.5 Výpočetní složitost

Výpočetní složitost se obecně definuje jako algoritmická (asymptotická) složitost programu, která udává, jak se bude náročnost algoritmu měnit na základě zvyšování velikosti vstupních dat. V případě NGS zarovnávacích nástrojů se jedná o velmi složité algoritmy, u nichž je určení asymptotické složitosti z apriorní znalosti kódu netriviální, a autor nástroje je neuvádí v dokumentaci. Možností, jak zjistit přibližnou algoritmickou složitost, je její experimentální identifikace na základě postupného zvyšování množství vstupních dat a měření délky běhu daného nástroje.



Obrázek 10: Porovnání určitých tříd algoritmické složitosti (převzato[36])

Výpočetní složitost nástroje pro náš případ NGS zarovnávacích nástrojů budeme definovat jako nároky na množství paměti, které nástroj pro svůj běh potřebuje. Tato metrika je pro porovnání jedna z rozhodujících, protože přímo definuje, zda je daný nástroj vhodný k použití na dostupném hardware.

4.6 Časová náročnost

Časová náročnost je chápána jako potřebný čas, během kterého je nástroj schopný provést indexaci a zarovnání poskytnutých readů, a souvisí s algoritmickou složitostí popsanou v předchozí kapitole. Předpokládá se, že s rostoucím poškozením vstupních dat bude narůstat i čas nutný pro jejich zarovnání s odpovídající přesností. Na základě této metriky pak bude možné vybrat pro konkrétní účel nástroj, který bude kompromisem mezi časovou náročností a kvalitou zarovnání readů.

5 Návrh nástroje pro automatické hodnocení

Nástroj pro automatické hodnocení byl navržen v jazyce Python. Pro účely testování nástrojů byly generovány syntetické ready vzhledem k možné kontrole výstupu, škálování experimentů, zkoumání vlivu různých počátečních podmínek atd., což by na reálných datech nebylo možné vzhledem k jejich omezené dostupnosti.

5.1 Generování syntetických readů

Pro účely hodnocení zarovnávacích nástrojů byl použit generátor syntetických readů ART ve verzi s označením MountRainier. ART na základě vložené referenční sekvence vygeneruje ready tak, jako kdyby byly výstupem reálného sekvenačního systému, a data poškodí zanesením chyb. Charakteristiku vygenerovaných readů lze ovlivnit úpravou parametrů ARTu. ART dokáže simulovat výstupy ze sekvenátorů Illumina, Roche 454 nebo SOLiD. Pro účely této práce byl využit generátor simulující výstupy ze systému Illumina, a to vzhledem k jeho častému využití v praxi a dostupnosti reálných dat pro možné testování nástrojů na těchto datech. Parametry nástroje určují, zda se bude jednat o single-end, paired-end, nebo mate-pair ready. Dále můžeme ovlivnit profil sekvenátoru, tedy jaký sekvenační systém budeme simulovat (parametr -ss). V závislosti na profilu sekvenátoru lze nastavit požadovanou délku readů (parametr -l). Dále nastavujeme parametr -f, který vyjadřuje amplifikaci sekvencí (hloubku pokrytí daného úseku), a další parametry určující např. průměrnou délku fragmentu vzorku, který byl sekvenován. Pro účely této práce byl ART nastaven podle níže uvedených parametrů reálných readů, které má k dispozici vedoucí práce.

-p (paired-end)

-ss MSv1 (sekvenační systém MiSeq 1)

-l 250 (délka readů 250 bp)

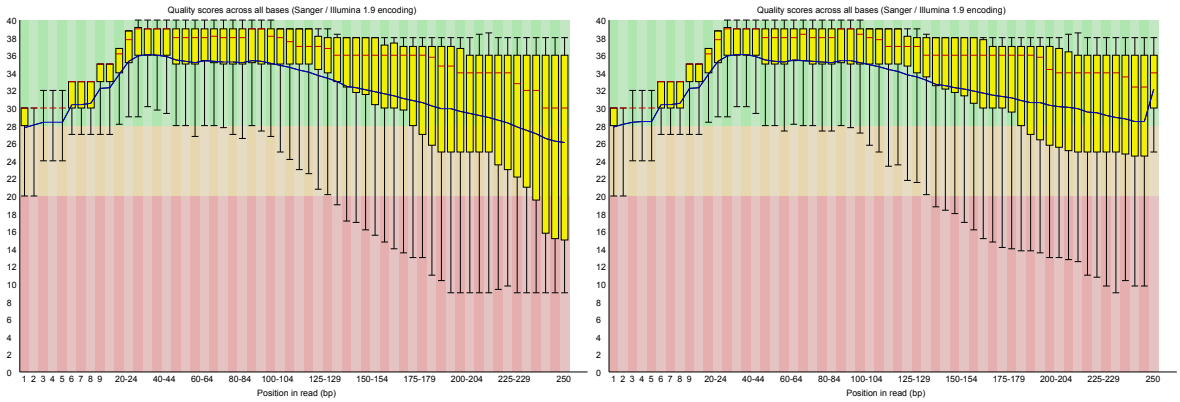
-f 100 (stonásobná amplifikace sekvence)

5.2 Knihovny pro práci se sekvencemi

V první části programu dochází k importu knihoven nezbytných pro funkci nástroje. Nástroj využívá knihovnu pysam, která usnadňuje práci se SAM formátem, jenž je běžným výstupem ze zarovnávacích nástrojů. Knihovna subprocess je zde využívána ke spouštění jednotlivých zarovnávacích nástrojů a generátoru syntetických readů ART tak, jako by se spouštěly pomocí terminálu. Knihovna time zajišťuje práci s aktuálním strojovým časem pro účely měření doby běhu jednotlivých zarovnávačů. Knihovna Bio zavádí do skriptu funkce Biopythonu pro jednoduchou práci se sekvencemi (objekt Seq...). Knihovna Levenshtein a fastDamerauLevenshtein zavádí funkce pro výpočet editační vzdálenosti mezi řetězci, dále knihovna numpy pro matematické funkce a knihovna matplotlib pro vizualizace. Knihovna msparse zavádí funkce pro snadnou práci s výstupy nástroje valgrind. Pokyny pro instalaci knihoven jsou uvedeny v uživatelské dokumentaci. Dalšími použitými knihovnami jsou configparser umnožňující čtení z konfiguračního souboru, knihovna csv pro práci s .csv soubory a knihovna sys pro přesměrování výstupů z konzole do souboru.

5.3 Předzpracování vstupních dat

Součástí většiny pipeline pro práci s NGS daty je určitá forma jejich předzpracování. Předzpracování je důležité při práci s reálnými daty vzhledem k jejich omezené kvalitě. Sekvenační systém určuje každé bázi v daném výstupním readu kvalitativní ohodnocení toho, do jaké míry je daná báze pravděpodobná. Tuto pravděpodobnost lze vizualizovat pomocí nástroje FastQC. Obrázek 11a dokládá, jak s délkou readů klesá jejich kvalita (pravděpodobnost správnosti umístění bází). Tato klesající kvalita může představovat problém pro možné budoucí využití nástroje. K vyřešení tohoto problému je na vygenerované ready použit nástroj Trimmomatic, který zajistí ořez nekvalitních bází, případně odstranění celého readu, pokud jeho kvalita nevyhovuje nastaveným kriteriím. Pro ořez bylo použito posuvné okno o velikosti 4 bp, které ořízne každou bázi, jejíž kvalita klesne pod hodnotu 20.



(a) Před ořezem

(b) Po ořezu

Obrázek 11: Efekt předzpracování dat

5.4 Měření metrik a vizualizace výsledků

Měření časové a výpočetní náročnosti probíhá přímo při postupném spouštění jednotlivých nástrojů pomocí knihovny subprocess. Metriky nejsou měřeny pro ART, protože pro výsledek nejsou relevantní. Stejně tak neprobíhá měření pro vytváření indexů jednotlivých nástrojů, protože indexace slouží pouze k urychlení běhu nástrojů a probíhá jako samostatný krok před zarovnáním. Měření času a spotřeby paměti probíhá za účelem zlepšení výpovědní hodnoty ve dvou oddělených krocích, jelikož nástroj Valgrind, jímž je zprostředkováno měření paměťové náročnosti, citelně zpomaluje měřený proces a to by mohlo vést ke zkreslení údajů o časové náročnosti nástrojů. Nástroj Valgrind měří nejvyšší momentální množství alokované paměti (peak), jelikož tento údaj přímo udává, zda bude hardware, na kterém bude zarovnání provedeno, dostačující. Časová náročnost se měří uchováním strojového času před spuštěním nástroje, od kterého se odečte aktuální strojový čas po ukončení běhu nástroje. Jednotlivé hodnoty jsou uchovány pro následnou vizualizaci a srovnání.

Přesnost zarovnání se vyhodnotí pomocí připraveného algoritmu, který načte zarovnanou sekvenci (read) a jí odpovídající úsek o stejné (automaticky určené) délce z původní referenční sekvence. Dále se určí Damerau-Levenshteinova vzdálenost těchto

dvou řetězců. Tato vzdálenost se uloží do seznamu vzdáleností a pokračuje se další iterací nad další sekvencí v pořadí podle seřazeného výstupního souboru. Nakonec se určí průměrná Damerau-Levenshteinova vzdálenost nad celým takto vytvořeným seznamem. Tento postup se opakuje tolikrát, kolik zarovnávacích nástrojů je předloženo k porovnání. Použitá knihovna pro výpočet Damerau-Levenshteinovy vzdálenosti umožňuje určit podobnost řetězců i v relativním měřítku, tedy to, na kolik procent jsou si řetězce podobné.

Pro názornost jsou výstupy nástroje zobrazovány ve formě grafů vytvořených pomocí knihovny matplotlib.

6 Zhodnocení vybraných nástrojů

6.1 Použitá HW konfigurace

Nástroj byl primárně testován na počítači HP Pavilion x360 14-dh1xxx s následující hardwarovou konfigurací:

Procesor	Intel Core i5, 1,6 GHz - 2 fyzická jádra
RAM	8 GB
GPU	Intel UHD Graphics 6000, 1536 MB
Operační systém	Ubuntu 20.04.1 LTS

Tabulka 1: HW konfigurace

6.2 Zdůvodnění výběru nástrojů ke srovnání

K porovnání metrik byly vybrány nástroje, jejichž hlavní vlastnosti jsou popsány v následující tabulce.

Název	**Verze**	**Kompatibilní operační systémy**	**Princip zarovnání**
Bowtie2	2.5.1	Linux, macOS, Windows	BWT
BWA-SW	0.7.17-r1188	Linux, macOS, Windows	BWT
BWA-MEM	0.7.17-r1188	Linux, macOS, Windows	BWT
GEM3	3.6.0	Linux, macOS	BWT
Kart	2.5.6	Linux, macOS	Hybridní (BWT/Hash)

Tabulka 2: Vybrané zarovnávací nástroje a jejich vlastnosti

Nástroj Bowtie2 byl vybrán z důvodu jeho velké rozšířenosti a vyhovující úrovně dokumentace. Nástroje BWA-SW a BWA-MEM byly vybrány jednak proto, že jsou

stejně jako Bowtie2 velmi rozšířené a dobře zdokumentované, ale také proto, aby byl demonstrován rozdíl výsledků vzhledem k vhodnosti použití konkrétní verze BWA na daná vstupní data. Nástroj GEM3 byl vybrán na základě výsledků článku [37], který uvádí, že GEM3 je z hlediska časové náročnosti mnohem rychlejší než např. Bowtie2 nebo BWA. Nástroj Kart byl vybrán z důvodu hybridního principu zarovnání a jako zástupce zarovnávacího software pro velké objemy dat získávané především novými sekvenačními postupy.

6.3 Volba testovacích dat

Za účelem zhodnocení nástrojů byla vybrána data dostupná z databáze imunopolymorfismů IPD[38] s ohledem na další možné rozšíření práce a na dostupnost reálných dat, která má k dispozici vedoucí práce. Vzhledem k tomu, že práce se zabývá pouze zhodnocením výkonu zarovnávacích nástrojů, nemusejí vstupní data odpovídat realitě a dávat smysl z biologického hlediska. Biologická podstata dat byla z výše uvedených důvodů zachována pouze z hlediska samotné struktury DNA (GC kontent atd.), aby báze nebyly řazeny naprosto náhodně. Pro navýšení množství testovacích dat bylo zařazeno více genů za sebe bez oddělujících struktur, čímž byla vytvořena delší jednotná sekvence. Za účelem testování nástroje na "objemných datech"byl využit soubor obsahující celý lidský genom s označením hg19, konkrétně úsek chromosomu 6 - dostupné z [39]. Výčet sekvencí a jejich charakteristiky jsou uvedeny u prvního experimentu. Pro zbytek experimentů jsou charakteristiky a původ vstupních dat k dispozici v příloze.

6.4 Provedené experimenty

Pro účely zhodnocení vybraných nástrojů byly nástroje podrobeny experimentům s celkem třemi různými sadami dat. Při testování a postupném připojování nástrojů bylo zjištěno, že některé z nich nedokáží zarovnat různě dlouhé ready, které vznikají při předzpracování dat. Vzhledem k tomu, že jejich zhodnocení je i tak přínosné, byl každý experiment proveden s předzpracovanými daty a dále s daty, kde k žádnému ořezu nedo-

šlo, a mohly tak být připojeny i tyto nástroje. Každý experiment byl proveden dvakrát, aby ve výsledcích byla zřejmá odchylka způsobená generátorem syntetických dat. Opakované experimenty pro malou a střední sadu vstupních dat (označené 1/2.2.N/T) jsou k dispozici v příloze. Další provedený experiment ukazuje, jaký vliv má počet vstupních vzorků na rychlost daného nástroje. Poslední provedený experiment a jeho vyhodnocení se zaměřuje na výsledné pokrytí reference po provedení zarovnání. Následující tabulka popisuje provedené experimenty a jejich označení.

Označení	**Popis**
1.1.T	Porovnání metrik, malá sada vstupních dat, předzpracovaná data, č. 1
1.1.N	Porovnání metrik, malá sada vstupních dat, nepředzpracovaná data č. 1
1.2.T	Porovnání metrik, malá sada vstupních dat, předzpracovaná data, č. 2
1.2.N	Porovnání metrik, malá sada vstupních dat, nepředzpracovaná data, č. 2
2.1.T	Porovnání metrik, střední sada vstupních dat, předzpracovaná data, č. 1
2.1.N	Porovnání metrik, střední sada vstupních dat, nepředzpracovaná data, č. 1
2.2.T	Porovnání metrik, střední sada vstupních dat, předzpracovaná data, č. 2
2.2.N	Porovnání metrik, střední sada vstupních dat, nepředzpracovaná data, č. 2
3.1.T	Porovnání metrik, velká sada vstupních dat, předzpracovaná data, č. 1
3.1.N	Porovnání metrik, velká sada vstupních dat, nepředzpracovaná data, č. 1
3.2.T	Porovnání metrik, velká sada vstupních dat, předzpracovaná data, č. 2
3.2.N	Porovnání metrik, velká sada vstupních dat, nepředzpracovaná data, č. 2
	Vliv počtu vzorků na výkon nástroje z hlediska času
	Vliv charakteru vstupních dat na rovnoměrnost pokrytí reference

Tabulka 3: Soupis provedených experimentů

6.4.1 Experiment 1.1.T

Původ dat	KIR:KIR2DL1*007
Délka ref. sekvence	14749
Počet vygenerovaných readů - 1. z páru	2900 (725000 bp)
Počet vygenerovaných readů - 2. z páru	2900 (725000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	702200 bp
Počet bp po ořezu nekvalitních bází - 2. z páru	507000 bp

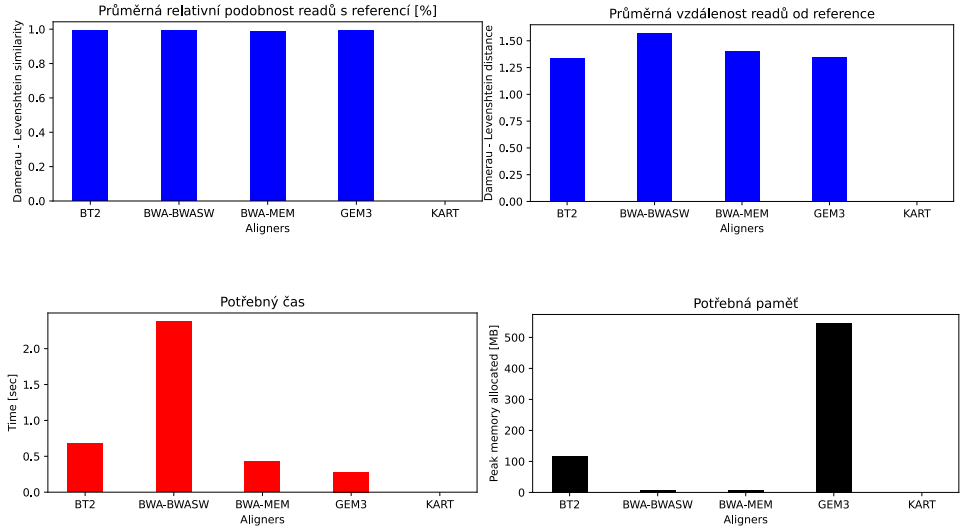
Tabulka 4: Charakteristika vstupních dat

Nástroj	**Čas běhu [s]**	**Potřebná paměť [MB]**	**Průměrná vzdálenost readů od reference**	**Průměrná vzdálenost readů od reference [%]**
Bowtie2	0.679	116.323	1.342	0.993
BWA-SW	7.703	8.134	1.571	0.992
BWA-MEM	2.378	7.703	1.400	0.987
Kart	-	-	-	-
GEM3	0.283	546.634	1.352	0.992

Tabulka 5: Porovnání získaných hodnot pro malá vstupní data

Malá vstupní data reprezentující jeden gen skupiny KIR, předzpracováno. Z hlediska času vykazuje nejlepší výsledky GEM3, nejpřesnější je nástroj Bowtie2. Z hlediska ná-

roků na paměť jsou nejúspornější oba nástroje ze skupiny BWA. Kart nebyl testován.



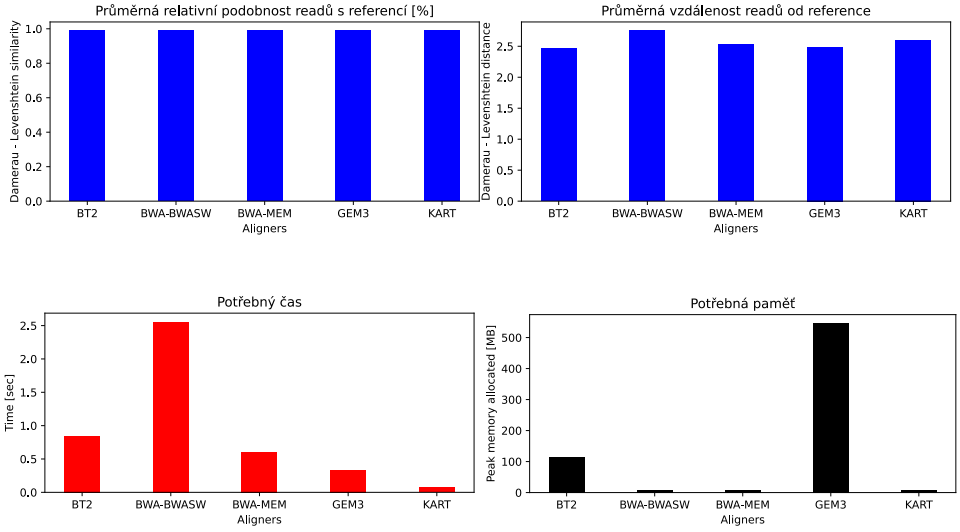
Obrázek 12: Vizualizace výsledků experimentu 1.1.T

6.4.2 Experiment 1.1.N

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	0.84	114.394	2.465	0.990
BWA-SW	2.559	9.334	2.755	0.989
BWA-MEM	0.608	8.411	2.525	0.990
Kart	0.072	8.276	2.603	0.990
GEM3	0.331	546.634	2.487	0.990

Tabulka 6: Porovnání získaných hodnot pro malá vstupní data

Malá vstupní data reprezentující jeden gen skupiny KIR, nepředzpracováno. Z hlediska času a nároků na paměť vykazuje nejlepší výsledky Kart, z hlediska přesnosti je nejlepší nástroj Bowtie2.



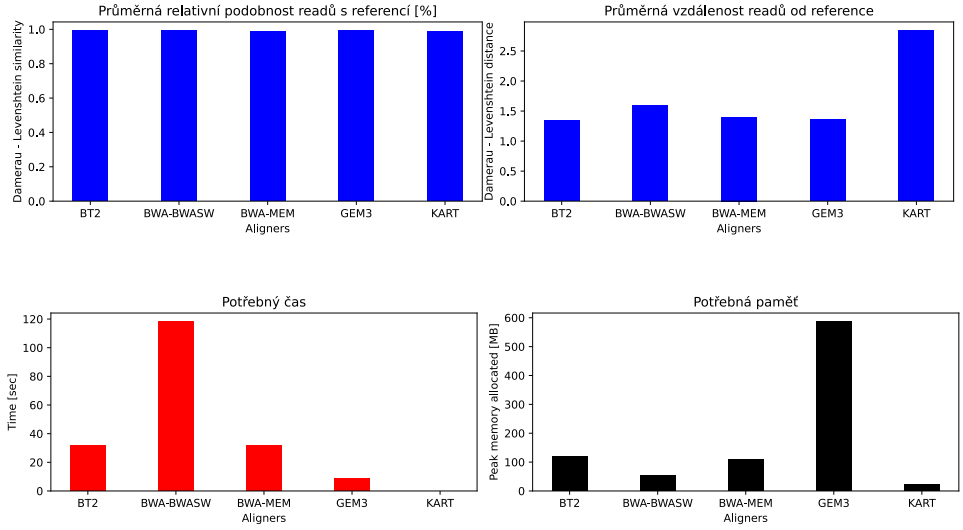
Obrázek 13: Vizualizace výsledků experimentu 1.1.N

6.4.3 Experiment 2.1.T

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	31.986	118.676	1.345	0.993
BWA-SW	118.294	52.149	1.595	0.992
BWA-MEM	31.471	107.761	1.403	0.988
Kart	-	-	-	-
GEM3	8.499	587.043	1.356	0.993

Tabulka 7: Porovnání získaných hodnot pro střední vstupní data

Středně velká vstupní data reprezentující 18 genů skupiny KIR, předzpracováno. Z hlediska času vykazuje nejlepší výsledky GEM3, nejpřesnější je nástroj Bowtie2. Z hlediska nároků na paměť jsou nejúspornější nástroj BWA-SW. Kart nebyl testován.



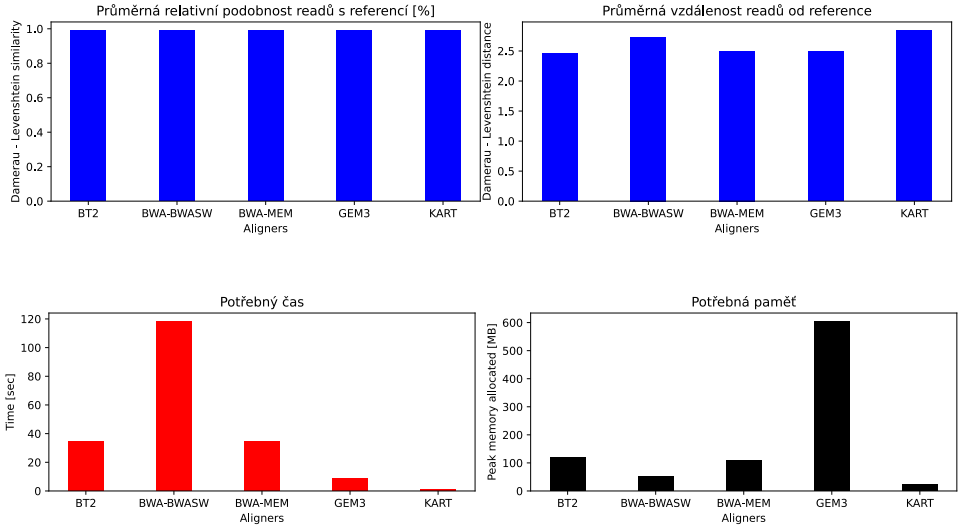
Obrázek 14: Vizualizace výsledků experimentu 2.1.T

6.4.4 Experiment 2.1.N

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	34.432	118.054	2.460	0.990
BWA-SW	118.212	53.082	2.732	0.989
BWA-MEM	34.299	108.289	2.498	0.990
Kart	0.793	21.249	2.841	0.990
GEM3	8.569	603.872	2.487	0.990

Tabulka 8: Porovnání získaných hodnot pro střední vstupní data

Střední vstupní data reprezentující 18 genů skupiny KIR, nepředzpracováno. Z hlediska času a nároků na paměť vykazuje nejlepší výsledky Kart, z hlediska přesnosti je nejlepší nástroj Bowtie2.



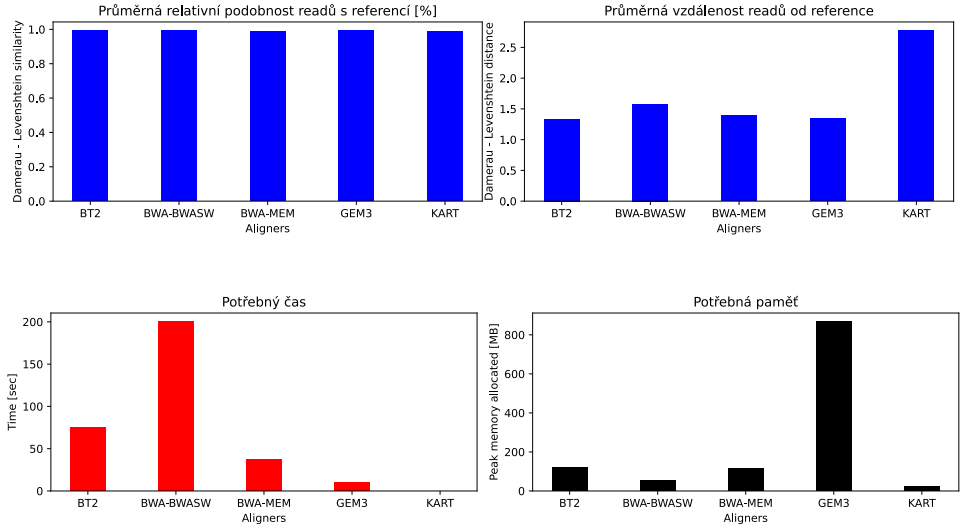
Obrázek 15: Vizualizace výsledků experimentu 2.1.N

6.4.5 Experiment 3.1.T

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	74.91	120.365	1.336	0.993
BWA-SW	200.417	56.023	1.579	0.992
BWA-MEM	37.583	115.498	1.389	0.988
Kart	-	-	-	-
GEM3	9.674	868.388	1.343	0.993

Tabulka 9: Porovnání získaných hodnot pro velká vstupní data

Velká vstupní data reprezentující úsek lidského chromosomu 6, předzpracováno. Z hlediska času vykazuje nejlepší výsledky GEM3, nejpřesnější je nástroj Bowtie2. Z hlediska nároků na paměť je nejúspornější nástroj BWA-SW. Kart nebyl testován.



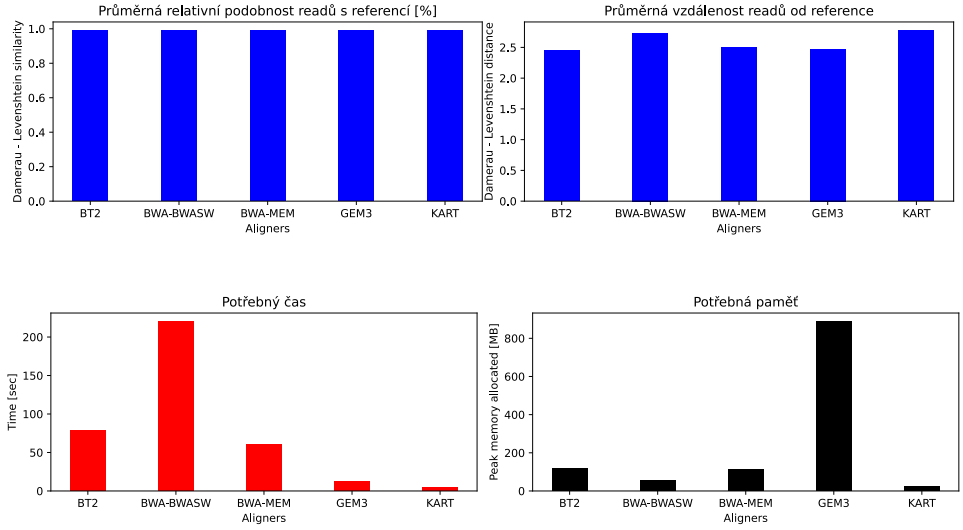
Obrázek 16: Vizualizace výsledků experimentu 3.1.T

6.4.6 Experiment 3.1.N

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	78.685	119.657	2.454	0.990
BWA-SW	220.183	56.093	2.732	0.989
BWA-MEM	60.373	114.092	2.493	0.990
Kart	5.002	25.185	2.775	0.989
GEM3	12.665	888.125	2.474	0.990

Tabulka 10: Porovnání získaných hodnot pro velká vstupní data

Velká vstupní data reprezentující úsek lidského chromosomu 6, nepředzpracováno. Z hlediska času a nároků na paměť vykazuje nejlepší výsledky Kart, z hlediska přesnosti je nejlepší nástroj Bowtie2.



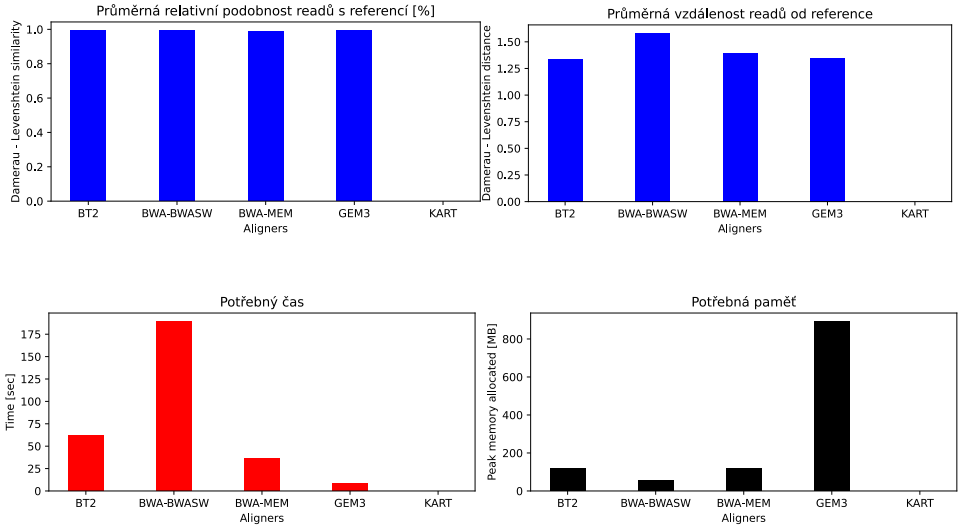
Obrázek 17: Vizualizace výsledků experimentu 3.1.N

6.4.7 Experiment 3.2.T

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	61.454	120.150	1.337	0.993
BWA-SW	189.199	55.795	1.579	0.992
BWA-MEM	36.747	117.403	1.391	0.988
Kart	-	-	-	-
GEM3	8.749	892.027	1.344	0.993

Tabulka 11: Porovnání získaných hodnot pro velká vstupní data

Velká vstupní data reprezentující úsek lidského chromosomu 6, předzpracováno. Z hlediska času vykazuje nejlepší výsledky GEM3, nejpřesnější je nástroj Bowtie2. Z hlediska nároků na paměť je nejúspornější nástroj BWA-SW. Kart nebyl testován.



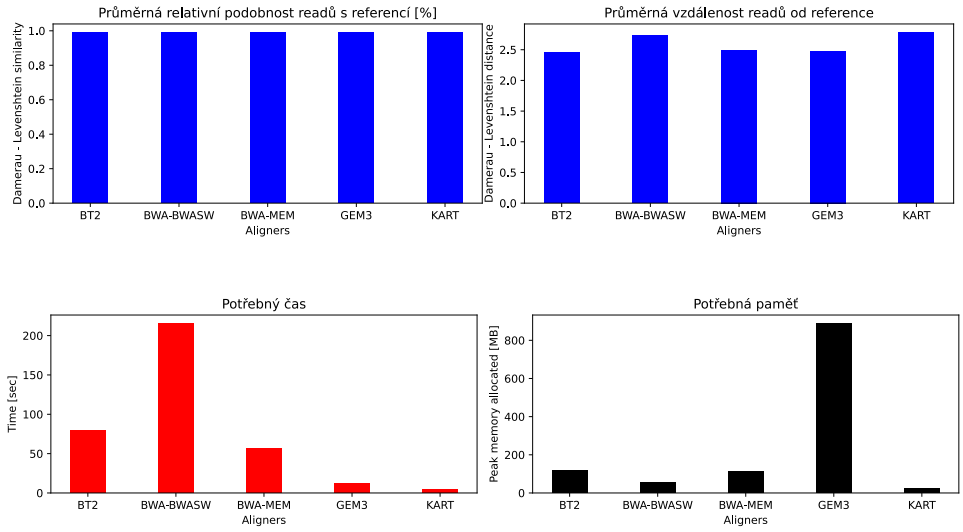
Obrázek 18: Vizualizace výsledků experimentu 3.2.T

6.4.8 Experiment 3.2.N

Nástroj	Čas běhu [s]	Potřebná paměť [MB]	Průměrná vzdálenost readů od reference	Průměrná vzdálenost readů od reference [%]
Bowtie2	79.255	118.989	2.458	0.990
BWA-SW	215.382	56.268	2.731	0.989
BWA-MEM	56.414	114.969	2.496	0.990
Kart	3.976	25.281	2.781	0.989
GEM3	12.047	888.129	2.478	0.990

Tabulka 12: Porovnání získaných hodnot pro velká vstupní data

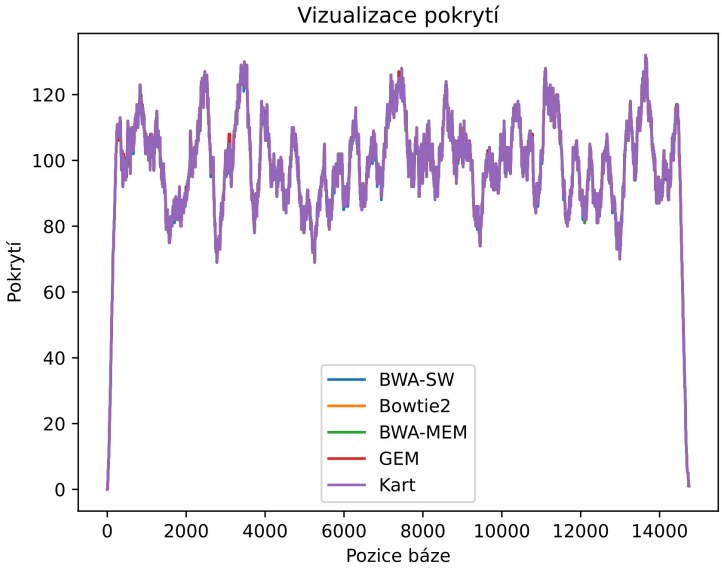
Velká vstupní data reprezentující úsek lidského chromosomu 6, nepředzpracováno. Z hlediska času a nároků na paměť vykazuje nejlepší výsledky Kart, z hlediska přesnosti je nejlepší nástroj Bowtie2.



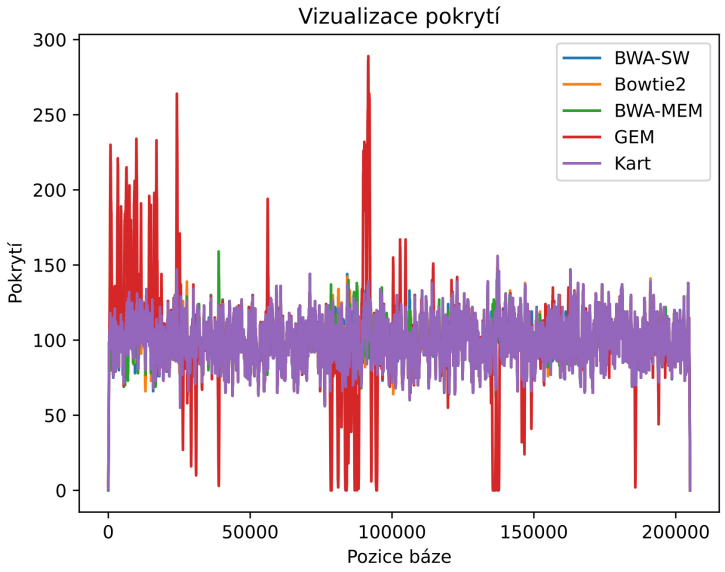
Obrázek 19: Vizualizace výsledků experimentu 3.2.N

6.5 Vizuální ověření správnosti zarovnání z hlediska rovnoměrného pokrytí

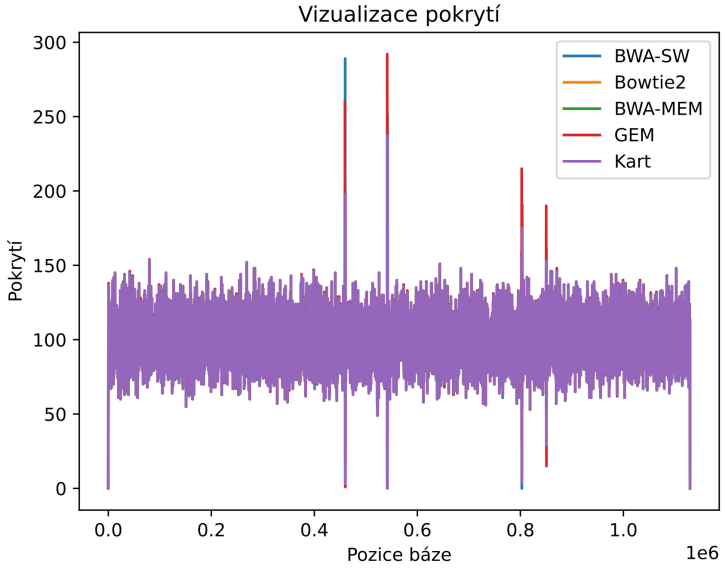
Při testování nástrojů na různých datech bylo zjištěno, že některé nástroje si nedokáží poradit s příliš podobnými daty (varianty téhož genu v jedné dlouhé sekvenci). Z toho důvodu bylo přidáno ověření výsledků z hlediska rovnoměrnosti pokrytí. Pro každý výstup ze zarovnávacího nástroje bylo provedeno seřazení výstupu pomocí nástroje samtools a pomocí téhož nástroje byla určena hloubka pokrytí každé jednotlivé báze v referenční sekvenci. Hloubka pokrytí každé jednotlivé báze je vizualizována graficky. Ve výstupním souboru lze najít procentuální hodnotu pokrytí pro každý nástroj. Procentuální hodnota se počítá s ohledem na směrodatnou odchylku 5 %, tedy báze je označena za dostatečně pokrytou tehdy, je-li pokryta alespoň pěti ready.



Obrázek 20: Pokrytí reference - malá sada vstupních dat

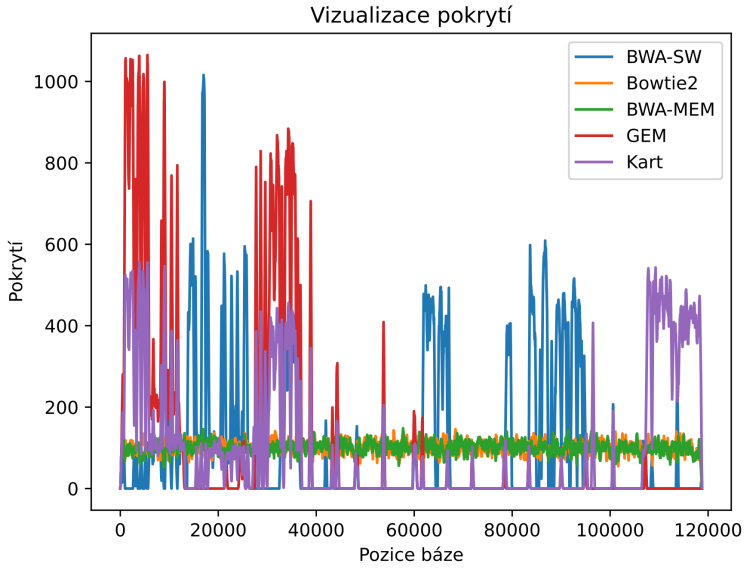


Obrázek 21: Pokrytí reference - střední sada vstupních dat



Obrázek 22: Pokrytí reference - velká sada vstupních dat

Následující obrázek ukazuje nežádoucí chování některých nástrojů při použití příliš podobných dat v referenční sekvenci. Je zřejmé, že pokud jsou v jedné referenci obsaženy pouze varianty téhož genu, dokáží je správně zarovnat pouze nástroje Bowtie2 a BWA-MEM.



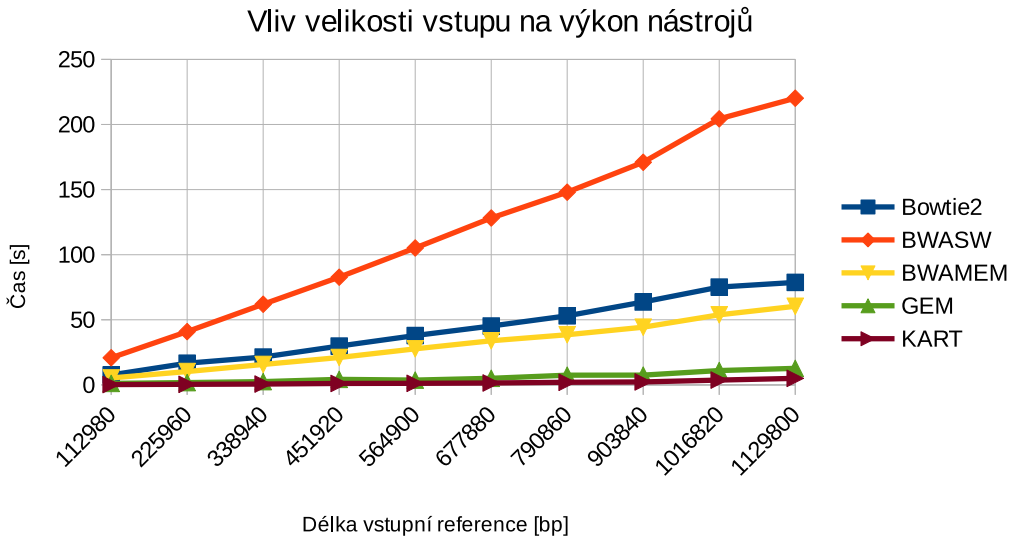
Obrázek 23: Pokrytí reference - nežádoucí výstup

6.6 Vliv počtu vzorků na rychlost nástrojů

Tento experiment byl navržen a proveden, aby ukázal, jak se mění výkon (z hlediska rychlosti) zarovnávacích nástrojů v závislosti na velikosti vstupní referenční sekvence. Jako reference byla použita stejná sekvence jako v experimentech 3.1 a 3.2, pouze byla postupně krácena vždy o přibližně stejný úsek, aby se z měření dala vyvodit představa o rychlosti daného nástroje na konkrétní velikosti vzorku. Nástroj byl za tímto účelem pomocí konfiguračního souboru nastaven tak, aby provedl generování readů, vytvoření indexů a zarovnání s měřením času. Zbytek funkcí byl vypnut.

Délka reference [bp]	**Čas běhu Bowtie2 [s]**	**Čas běhu BWA-SW [s]**	**Čas běhu BWA-MEM [s]**	**Čas běhu GEM3 [s]**	**Čas běhu Kart [s]**
112980	7.65	20.79	5.29	1.12	0.24
225960	16.67	40.91	10.25	1.74	0.45
338940	21.36	61.94	15.68	2.50	0.65
451920	29.82	82.73	21.00	4.22	1.07
564900	37.89	105.19	27.67	3.66	1.21
677880	45.21	128.18	33.88	5.07	1.44
790860	53.03	148.02	38.49	7.47	1.99
903840	63.68	170.98	44.28	7.48	2.27
1016820	75.08	204.35	53.81	11.02	3.70
1129800	78.69	220.18	60.37	12.67	5.00

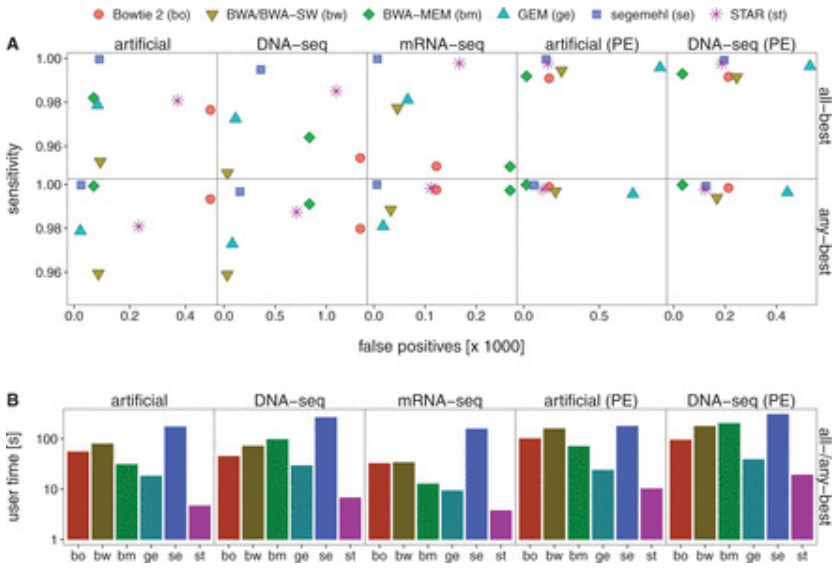
Tabulka 13: Vliv počtu vzorků na rychlost nástrojů



Obrázek 24: Graf vlivu velikosti vstupu na rychlost nástrojů

6.7 Validace výsledků

Za účelem validace dosažených výsledků byly získané výstupy nástroje porovnány s výsledky v článku [37]. Tyto výsledky je velmi obtížné přesně validovat, jelikož studie, které se tímto tématem zabývaly, jsou zastaralé nebo používají příliš odlišné metriky, odlišný hardware, jinak poškozená či dokonce naprosto jiná data (jiná než lidská data, RNA data...). Z údajů obsažených v obrázku 29, konkrétně z grafu "artificial(PE)", který odpovídá syntetickým paired-end readům použitým v této práci vyplývá, že zhodnocení přesnosti nástrojů i změřená časová náročnost zhruba odpovídají dosaženým výsledkům experimentů navržených v této práci (nutno přihlédnout k jiné použité metrice).



Obrázek 25: Výsledek hodnocení nástrojů z článku [37] [převzato]

6.8 Shrnutí

Na základě provedených experimentů lze konstatovat, že z hlediska časové náročnosti je jednoznačně nejrychlejším nástrojem testovaný Kart. Nejhorší výsledek vykázal nástroj BWA, konkrétně jeho varianta BWA-SW, což bylo ovšem způsobeno nevhodností použití tohoto nástroje pro data užitá v této práci. Naproti tomu varianta BWA-MEM

má velmi podobné časové nároky jako Bowtie2 a u těchto nástrojů se doba zpracování ukázala jako přijatelná. Nástroj GEM3 potřeboval oproti BWA-MEM nebo Bowtie2 čas přibližně poloviční, nicméně toto bylo kompenzováno vysokou paměťovou náročností, která je ze všech používaných nástrojů jednoznačně nejvyšší.

U měření přesnosti pomocí Damerau-Levenshteinovy vzdálenosti má velikost vstupních dat jen minimální nebo žádný vliv na výslednou chybovost nástroje. Největší chybovostí z hlediska přesnosti se i zde vyznačuje nástroj BWA-SW (z důvodu uvedeného výše) a dále nástroj Kart. U nástroje Kart je extrémně nízká časová a paměťová náročnost kompenzována právě větší chybovostí.

Na závěr je třeba vyzdvihnout nástroje Bowtie2 a BWA-MEM, které jako jediné dokázaly s minimální chybovostí zpracovat data, ve kterých byly jen malé rozdíly (varianty jednoho genu), jak dokazuje obrázek 27. Zbytek nástrojů zarovnal všechny ready k několika málo genům s vysokým pokrytím, které neodpovídá požadovanému výsledku.

Obecně se tedy dá konstatovat, že v aplikaci, kde nám nevadí nižší preciznost nástroje a kde pracujeme s velkými objemy vstupních dat, je možné doporučit hybridní nástroj Kart nebo paměťově méně efektivní GEM. Jako ”zlatá střední cesta”se jeví nástroje Bowtie2 a BWA-MEM, které zároveň lze doporučit k použití např. při rozeznávání variant téhož genu.

7 Závěr

DNA data pocházející ze sekvenačních systémů nové generace a jejich zpracování je v současné době stále nedílnou součástí komplexních výzkumných postupů. Díky stále se rozvíjejícímu technologickému i programovému zázemí se rozšiřuje i spektrum možností jak zvýšit úspěšnost transplantací a zmírnit, či eliminovat posttransplantační komplikace příjemce.

Cílem této práce bylo vytvořit nástroj, který bude automaticky hodnotit NGS zarovnávací nástroje na základě předem definovaných metrik. Metriky byly definovány s ohledem na konkrétní účely. Výstup této práce by měl přispět k rozhodování o použití konkrétního zvoleného nástroje jako součást komplexního postupu při práci s DNA daty.

Provedenou rešerší dostupného softwaru jsem zjistil, že vzhledem k rozdílným parametrům a rozdílnému určení nelze hodnotit všechny nástroje na základě stejných pravidel. Vybrané nástroje dosahovaly předpokládaných výsledků a byla demonstrována důležitost výběru vhodného nástroje pro konkrétní data, pro která je daný nástroj určen. Na základě uvedeného jsem došel k závěru, že nástroje Bowtie2 a BWA-MEM jsou téměř univerzální nástroje a jsou kompromisem při výběru s ohledem na téměř všechny testované metriky. Nástroje GEM3 a Kart mají určité přednosti (zejména časovou efektivitu), ale ty jsou vykoupeny nedostatky v oblasti paměťové náročnosti či chybovosti. Jejich užití je tedy třeba předem pečlivě zvážit s ohledem na charakter vstupních dat (velikost, potřeba výsledné přesnosti apod.).

Tuto práci jsem napsal jako základ pro další výzkum a lze na ni navázat buď rozšířením oblasti zkoumání dalších nástrojů, jejich parametrů a vhodnosti použití, nebo přechodem na další krok ve zpracování DNA dat jako např. de novo assembly, identifikace variant atd. Další možnosti rozšíření práce by spočívaly v přidání dalších metrik přesnosti zarovnání nebo např. v testování v poslední době se rozšiřujících cloudových služeb.

Odkazy

1. TEPLÁ, Milada. *Biochemie - základní kapitoly*. 2013. Dostupné také z: http://www.studiumbiochemie.cz/na2.html.
2. ELZANOWSKI, Andrzej; OSTELL, Jim. *The Genetic Codes*. 2019. Dostupné také z: https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi.
3. KOLÍSKO, Martin. *Moderní metody sekvenování DNA*. 2017. Dostupné také z: https://ziva.avcr.cz/files/ziva/pdf/moderni-metody-sekvenovani-dna.pdf.
4. *Sanger Sequencing*. 2022. Dostupné také z: https://eurofinsgenomics.eu/en/eurofins-genomics/material-and-methods/sanger-sequencing/.
5. *DNA Sequencing Core Facility*. Dostupné také z: https://biology.unt.edu/~jajohnson/Chromatogram_Interpretation.
6. BEHJATI, Sam; TARPEY, Patrick. What is next generation sequencing? *Arch Dis Child Educ Pract Ed.* 10/2013. Dostupné z DOI: 10.1136/archdischild-2013-304340.
7. CROSSLEY, Beate; JIANFA, Bai; GLASER, Amy; MAES, Roger; PORTER, Elisabeth; KILLIAN, Mary; CLEMENT, Travis; KURT-TOOHAY, Kathy. Guidelines for Sanger sequencing and molecular assay monitoring. *Journal of Veterinary Diagnostic Investigation*. 2020, roč. 32, s. 767–775. Dostupné z DOI: d0oi..1o1rg7/710/1.10147076/130847062308970250890353.
8. JOSHI, Mohini; DESHPANDE, Jayant. Polymerase chain reaction: methods, principles and application. *International Journal of Biomedical Research*. 2011, roč. 2, č. 1, s. 81–97. Dostupné z DOI: https://doi.org/10.7439/ijbr.v2i1.83.
9. *PCR(polymerázová řetězová reakce)*. Dostupné také z: https://cit.vfu.cz/opvk2011/?title=popis_metod-pcr&lang=cz.

10. GENOMICS, Functional. *Ion Torrent: Proton / PGM sequencing*. Dostupné také z: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/ion-torrent-proton-pgm-sequencing/.

11. GENOMICS, Functional. *454 sequencing*. Dostupné také z: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/.

12. SCHADT, Eric; TURNER, Steve; KASARSKIS, Andrew. A window into third-generation sequencing. *Human Molecular Genetics*. 2010, roč. 20, s. 227–240. Dostupné z DOI: https://doi.org/10.1093/hmg/ddq416.

13. *DNA Sequence formats*. Dostupné také z: https://www.genomatix.de/online_help/help/sequence_formats.html.

14. *DNA Sequence formats*. Dostupné také z: https://www.animalgenome.org/bioinfo/resources/manuals/seqformats.

15. *Sequence Alignment/Map Format Specification*. 2022. Dostupné také z: https://samtools.github.io/hts-specs/SAMv1.pdf.

16. *Intro to NGS Data Analysis Workflow*. 2020. Dostupné také z: https://diagnostech.co.za/intro-to-ngs-data-analysis-workflow/.

17. REINERT, Knut; LANGMEAD, Ben; WEESE, David; EVERS, Dirk. Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*. 2015, roč. 16, s. 133–151. Dostupné z DOI: 10.1146/annurev-genom-090413-025358.

18. LI, Heng; DURBIN, Richard. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009, roč. 25, č. 14, s. 1754–1760. Dostupné z DOI: https://doi.org/10.1093/bioinformatics/btp324.

19. *Hashovací tabulka*. Dostupné také z: https://www.algoritmy.net/article/32077/Hashovaci-tabulka.

20. *BOWTIE 2*. Dostupné také z: https://bowtie-bio.sourceforge.net/bowtie2/index.shtml.

21. LANGMEAD, Ben; SALZBERG, Steven. Fast gapped-read alignment with Bowtie2. *Nat Methods*. 2012, roč. 9, č. 4, s. 357–359. Dostupné z DOI: 10.1038/nmeth.1923.

22. *Burrows-Wheeler Aligner*. 2010. Dostupné také z: https://bio-bwa.sourceforge.net/.

23. ALKAN, Can; KIDD, Jeffrey; MARQUES-BONET, Tomas; AKSAY, Gozde; ANTONACCI, Francesca; HORMOZDIARI, Fereydoun; KITZMAN, Jacob; BAKER, Carl; MALIG, Maika; MUTLU, Onur; SAHINALP, Cenk; GIBBS, Richard; EICHLER, Evan. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet.* 2009, roč. 41, č. 10, s. 1061–1067. Dostupné z DOI: 10.1038/ng.437.

24. LI, Heng; DURBIN, Richard. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008, roč. 18, č. 19, s. 1851–1858. Dostupné z DOI: 10.1101/gr.078212.108.

25. RUMBLE, Stephen; LACROUTE, Phil; DALCA, Adrian; FIUME, Marc; SIDOW, Arend; BRUDNO, Michael. SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Comput Biol*. 2009, roč. 5, č. 5. Dostupné z DOI: https://doi.org/10.1371/journal.pcbi.1000386.

26. NOVOCRAFT. *NovoAlign*. 2020. Dostupné také z: https://www.novocraft.com/products/novoalign/.

27. SANTIAGO, Marco-Sola. *GEM-Mapper v3*. 2019. Dostupné také z: https://github.com/smarco/gem3-mapper.

28. LIN, Hsin-Nan; HSU, Wen-Lian. Kart: a divide-and-conquer algorithm for NGS read alignment. *Bioinformatics*. 2017, roč. 33, č. 15, s. 2281–2287. Dostupné z DOI: https://doi.org/10.1093/bioinformatics/btx189.

29. LIN, Hsin-Nan; HSU, Wen-Lian. Comprehensive comparison of cloud-based NGS data analysis and alignment tools. *Informatics in Medicine Unlocked*. 2020, roč. 18, s. 100–296. ISSN 2352-9148. Dostupné z DOI: https://doi.org/10.1016/j.imu.2020.100296.

30. RUFFALO, Matthew; LAFRAMBOISE, Thomas; KOYUTÜRK, Mehmet. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*. 2011, roč. 27, č. 20, s. 2790–2796. ISSN 1367-4803. Dostupné z DOI: 10.1093/bioinformatics/btr477.

31. SCHILBERT, Hanna Marie; REMPEL, Andreas; PUCKER, Boas. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. *Plants*. 2020, roč. 9, č. 4. ISSN 2223-7747. Dostupné z DOI: 10.3390/plants9040439.

32. WANG, Wei-An; WU, Chin-Ting; LU, Tzu-Pin; TSAI, Mong-Hsun; LAI, Liang-Chuan; CHUANG, Eric Y. *2014 International Conference on Electrical Engineering and Computer Science (ICEECS)*. Comparisons and performance evaluations of RNA-seq alignment tools. 2014. Dostupné z DOI: 10.1109/ICEECS.2014.7045249.

33. BLACK, Paul. *Hamming distance*. 2006. Dostupné také z: https://www.nist.gov/dads/HTML/HammingDistance.html.

34. *LEVENSHTEIN DISTANCE*. Dostupné také z: https://devopedia.org/levenshtein-distance.

35. SINGH, Prabhjot; DHAWAN, Sumit; AGARWAL, Shubham; THAKUR, Dr. Narina. Implementation of an efficient Fuzzy Logic based Information Retrieval System. *EAI Endorsed Transactions on Scalable Information Systems*. 2015, roč. 2. Dostupné z DOI: 10.4108/sis.2.5.e5.

36. *Computational complexity classes comparison*. Dostupné také z: https://cs.wikipedia.org/wiki/Asymptotick%C3%A1_slo%C5%BEitost#/media/Soubor:Comparison_computational_complexity.svg.

37. OTTO, Christian; STADLER, Peter F.; HOFFMANN, Steve. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics*. 2014, roč. 30, č. 13, s. 1837–1843. ISSN 1367-4803. Dostupné z DOI: 10.1093/bioinformatics/btu146.

38. ROBINSON, J; HALLIWELL, JA; MCWILLIAM, H; LOPEZ, R; MARSH, SGE. IPD - the Immuno Polymorphism Database. *Nucleic Acid Research*. 2013. Dostupné také z: https://www.ebi.ac.uk/ipd/.

39. CLARK, K; KARSCH-MIZRACHI, I; LIPMAN, DJ; OSTELL, J; SAYERS, EW. *Nucleic Acids Res.* 2016, roč. 44, s. 67–72. Dostupné z DOI: 10.1093/nar/gkv1276.

40. VAN ROSSUM, Guido; DRAKE, Fred L. Python 3 Reference Manual. 2009. ISBN 1441412697. Dostupné také z: https://www.python.org/downloads/.

41. Anaconda Software Distribution. *Anaconda Documentation*. 2020. Dostupné také z: https://docs.conda.io/en/latest/.

42. WEICHUNG, Huang; LEPING, Li; MYERS, JR; GABOR, TM. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012, roč. 28, s. 593–594. Dostupné také z: https://www.niehs.nih.gov/research/resources/software/biostatistics/art/index.cfm.

43. COCK. *Biopython*. 2009. Dostupné také z: https://biopython.org/.

44. *Pysam*. Dostupné také z: https://github.com/pysam-developers/pysam.

45. *Python subprocess.run*. Dostupné také z: https://docs.python.org/3/library/subprocess.html.

46. *Python time*. Dostupné také z: https://docs.python.org/3/library/time.html.

47. *Python numpy*. Dostupné také z: https://numpy.org/.

48. *Python matplotlib*. Dostupné také z: https://matplotlib.org/.

49. *Python Levenshtein*. Dostupné také z: https://pypi.org/project/python-Levenshtein/.

50. *Python FastDamerauLevenshtein*. Dostupné také z: https://pypi.org/project/fastDamerauLevenshtein/.

51. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. [B.r.], roč. 30, s. 2114–2120. Dostupné z DOI: https://doi.org/10.1093/bioinformatics/btu170.

52. *Proceedings of ACM SIGPLAN 2007 Conference on Programming Language Design and Implementation (PLDI 2007)*. Valgrind: A Framework for Heavyweight Dynamic Binary Instrumentation. SanDiego, USA: Nethercote, Nicholas a Seward Julian, 2007. Dostupné také z: https://valgrind.org/.

53. *MSParser*. Dostupné také z: https://pypi.org/project/msparser/.

54. *ConfigParser*. Dostupné také z: https://docs.python.org/3/library/configparser.html.

55. *CSV*. Dostupné také z: https://docs.python.org/3/library/csv.html.

56. *sys*. Dostupné také z: https://docs.python.org/3/library/sys.html.

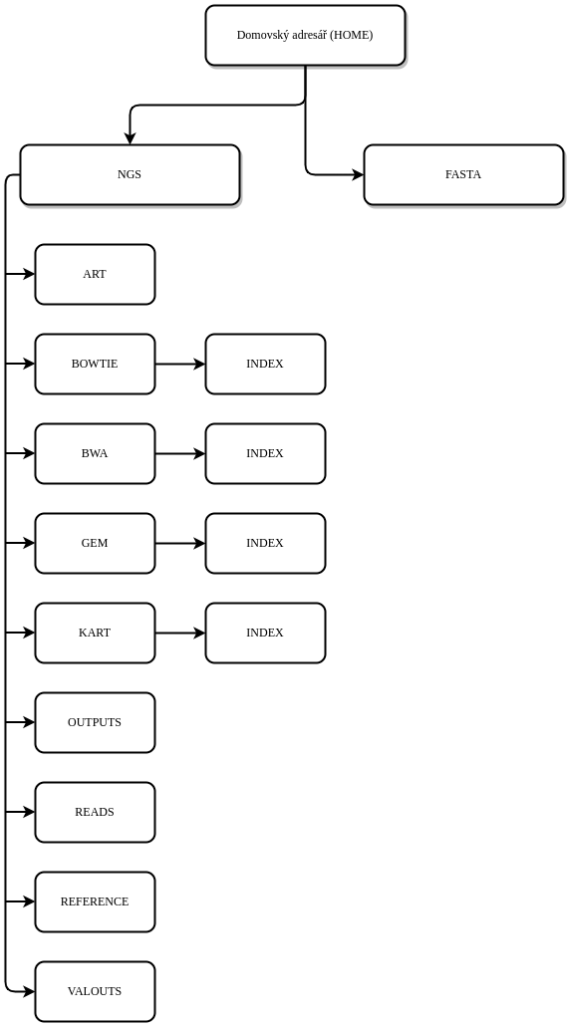
Seznam obrázků

Seznam tabulek

A Uživatelská dokumentace

A.1 Adresářová struktura

Pro správnou funkci nástroje je potřebná specifická adresářová struktura, jejíž vytvoření poskytuje spuštení skriptu makeFolders.sh v domovském adresáři. Složka NGS obsahuje všechny složky potřebné pro funkci nástroje a složka FASTA obsahuje referenční sekvence.



Obrázek 26: Schéma potřebné adresářové struktury

A.2 Instalace

K funkci nástroje je nutné mít instalováno následující:

- Python 3.6, nebo vyšší [40]
- Miniconda, nebo Anaconda [41]
- Art MountRainier [42]
- Bowtie2 [21]
- BWA [18]
- GEM3 [27]
- Kart [28]
- Biopython [43]
- Pysam [44]
- Subprocess.run [45]
- Time [46]
- Numpy [47]
- Matplotlib [48]
- Levenshtein [49]
- FastDamerauLevenshtein [50]
- Trimmomatic [51]
- Valgrind [52]
- MSParser [53]

- ConfigParser [54]
- csv [55]
- sys [56]

K instalaci lze využít odkazů uvedených u každého potřebného softwaru. Konkrétnější informace o balíčcích a instalaci lze nalézt tamtéž. Zarovnávací nástroje je doporučeno instalovat prostřednictvím Condy. Při manuální instalaci nástroje je třeba přidat cestu ke spouštěcímu skriptu do systémové proměnné PATH.

A.3 Spuštění nástroje

Nástroj se spouští pomocí příkazové řádky (Terminálu) spuštěním hlavního souboru ngsAlignersComparator.py, který musí být umístěn v domovském adresáři. Vedle hlavního souboru existuje ještě konfigurační soubor config.ini, kde lze nastavit následující parametry:

Parametr	**Popis parametru**	**Výchozí hodnota**
refSeq	Referenční sekvence	sequenceMedium.fasta
l	Parametr ARTu	250
ss	Parametr ARTu	MSv1
f	Parametr ARTu	100
m	Parametr ARTu	300
s	Parametr ARTu	50
art	Generování nových syntetických dat	true
measureTime	Zapnout měření časových metrik	true
measureMemory	Zapnout měření pamětových metrik	true
measureDistance	Zapnout měření přesnosti	true
trimmomatic	Předzpracování vstupních dat	true
makeIndex	Generování indexů	true
bowtie2	Připojení zarovnávacího nástroje	true
bwasw	Připojení zarovnávacího nástroje	true
bwamem	Připojení zarovnávacího nástroje	true
gem	Připojení zarovnávacího nástroje	true
kart	Připojení zarovnávacího nástroje	false
novoalign	Připojení zarovnávacího nástroje	false

Tabulka 14: Parametry konfigurace nástroje

A.4 Čištění adresářů

Pro potřebu uvedení adresářů do výchozího stavu je k dispozici skript clearFolder.sh, který smaže všechna data, která byla vytvořena za běhu hlavního programu.

A.5 Výstupy

Výstupy programu lze nalézt s časovou známkou v souboru results.txt, který se vytvoří ve složce NGS po spuštění nástroje. Výstupní obrázky lze nalézt v téže složce. Při opětovném spuštění nástroje se obrázky přemažou novými, textové výstupy se ukládají do téhož souboru s aktuální časovou známkou. Při použití skriptu pro čištění adresářů dojde ke smazání tohoto souboru s výsledky.

B Detailní popis struktury vstupních dat experimentů 1.1.N - 3.2.N

B.1 1.1.N

Původ dat	KIR:KIR2DL1*007
Délka ref. sekvence	14749
Počet vygenerovaných readů - 1. z páru	2900 (725000 bp)
Počet vygenerovaných readů - 2. z páru	2900 (725000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	nebylo provedeno
Počet bp po ořezu nekvalitních bází - 2. z páru	nebylo provedeno

Tabulka 15: Charakteristika vstupních dat

B.2 1.2.T

Původ dat	KIR:KIR2DL1*007
Délka ref. sekvence	14749
Počet vygenerovaných readů - 1. z páru	2900 (725000 bp)
Počet vygenerovaných readů - 2. z páru	2900 (725000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	706300 bp
Počet bp po ořezu nekvalitních bází - 2. z páru	507600 bp

Tabulka 16: Charakteristika vstupních dat

B.3 1.2.N

Původ dat	KIR:KIR2DL1*007
Délka ref. sekvence	14749
Počet vygenerovaných readů - 1. z páru	2900 (725000 bp)
Počet vygenerovaných readů - 2. z páru	2900 (725000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	nebylo provedeno
Počet bp po ořezu nekvalitních bází - 2. z páru	nebylo provedeno

Tabulka 17: Charakteristika vstupních dat

B.4 2.1.T

Původ dat	KIR2DL1*007 KIR2DL2*0010101 KIR2DL3*006 KIR2DL4*010 KIR2DL5A*025 KIR2DL5B*003 KIR2DL5A*025 KIR2DP1*004 KIR2DS1*006 KIR2DS2*010 KIR2DS3*009 KIR2DS4*010 KIR2DS5*010 KIR3DL1*019 KIR3DL2*018 KIR3DL3*005 KIR3DP1*001 KIR3DS1*055
Délka ref. sekvence	205140
Poqčet vygenerovaných readů - 1. z páru	41000 (10200000 bp)
Počet vygenerovaných readů - 2. z páru	41000 (10200000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	9900000
Počet bp po ořezu nekvalitních bází - 2. z páru	7100000

Tabulka 18: Charakteristika vstupních dat

B.5 2.1.N

Původ dat	KIR2DL1*007 KIR2DL2*0010101 KIR2DL3*006 KIR2DL4*010 KIR2DL5A*025 KIR2DL5B*003 KIR2DL5A*025 KIR2DP1*004 KIR2DS1*006 KIR2DS2*010 KIR2DS3*009 KIR2DS4*010 KIR2DS5*010 KIR3DL1*019 KIR3DL2*018 KIR3DL3*005 KIR3DP1*001 KIR3DS1*055
Délka ref. sekvence	205140
Počet vygenerovaných readů - 1. z páru	41000 (10200000 bp)
Počet vygenerovaných readů - 2. z páru	41000 (10200000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	nebylo provedeno
Počet bp po ořezu nekvalitních bází - 2. z páru	nebylo provedeno

Tabulka 19: Charakteristika vstupních dat

B.6 2.2.T

Původ dat	KIR2DL1*007 KIR2DL2*0010101 KIR2DL3*006 KIR2DL4*010 KIR2DL5A*025 KIR2DL5B*003 KIR2DL5A*025 KIR2DP1*004 KIR2DS1*006 KIR2DS2*010 KIR2DS3*009 KIR2DS4*010 KIR2DS5*010 KIR3DL1*019 KIR3DL2*018 KIR3DL3*005 KIR3DP1*001 KIR3DS1*055
Délka ref. sekvence	205140
Počet vygenerovaných readů - 1. z páru	41000 (10200000 bp)
Počet vygenerovaných readů - 2. z páru	41000 (10200000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	9900000
Počet bp po ořezu nekvalitních bází - 2. z páru	7100000

Tabulka 20: Charakteristika vstupních dat

B.7 2.2.N

Původ dat	KIR2DL1*007 KIR2DL2*0010101 KIR2DL3*006 KIR2DL4*010 KIR2DL5A*025 KIR2DL5B*003 KIR2DL5A*025 KIR2DP1*004 KIR2DS1*006 KIR2DS2*010 KIR2DS3*009 KIR2DS4*010 KIR2DS5*010 KIR3DL1*019 KIR3DL2*018 KIR3DL3*005 KIR3DP1*001 KIR3DS1*055
Délka ref. sekvence	205140
Počet vygenerovaných readů - 1. z páru	41000 (10200000 bp)
Počet vygenerovaných readů - 2. z páru	41000 (10200000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	nebylo provedeno
Počet bp po ořezu nekvalitních bází - 2. z páru	nebylo provedeno

Tabulka 21: Charakteristika vstupních dat

B.8 3.1.T

Původ dat	Prvních 1145798 bází chromosomu 6 genomu hg19 (bez N)
Délka ref. sekvence	1145798 bp
Počet vygenerovaných readů - 1. z páru	225900 (56400000 bp)
Počet vygenerovaných readů - 2. z páru	225900 (56400000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	54800000
Počet bp po ořezu nekvalitních bází - 2. z páru	39500000

Tabulka 22: Charakteristika vstupních dat

B.9 3.1.N

Původ dat	Prvních 1145798 bází chromosomu 6 genomu hg19 (bez N)
Délka ref. sekvence	1145798 bp
Počet vygenerovaných readů - 1. z páru	225900 (56400000 bp)
Počet vygenerovaných readů - 2. z páru	225900 (56400000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	nebylo provedeno
Počet bp po ořezu nekvalitních bází - 2. z páru	nebylo provedeno

Tabulka 23: Charakteristika vstupních dat

B.10 3.2.T

Původ dat	Prvních 1145798 bází chromosomu 6 genomu hg19 (bez N)
Délka ref. sekvence	1145798 bp
Počet vygenerovaných readů - 1. z páru	225900 (56400000 bp)
Počet vygenerovaných readů - 2. z páru	225900 (56400000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	54800000
Počet bp po ořezu nekvalitních bází - 2. z páru	39500000

Tabulka 24: Charakteristika vstupních dat

B.11 3.2.N

Původ dat	Prvních 1145798 bází chromosomu 6 genomu hg19 (bez N)
Délka ref. sekvence	1145798 bp
Počet vygenerovaných readů - 1. z páru	225900 (56400000 bp)
Počet vygenerovaných readů - 2. z páru	225900 (56400000 bp)
Parametr -ss	MSv1
Parametr -m	300
Parametr -f	100
Parametr -s	50
Typ readů	paired-end
Počet bp po ořezu nekvalitních bází - 1. z páru	nebylo provedeno
Počet bp po ořezu nekvalitních bází - 2. z páru	nebylo provedeno

Tabulka 25: Charakteristika vstupních dat

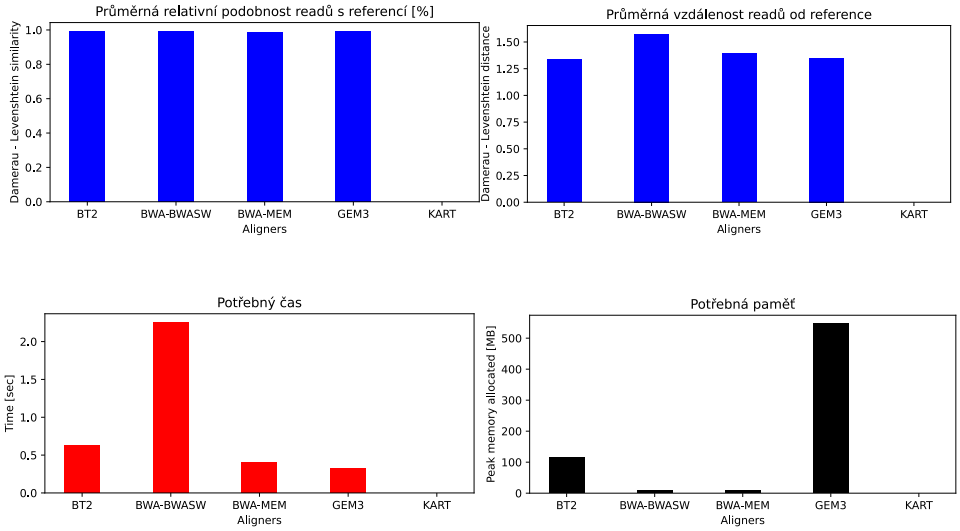
C Výsledky opakovaných experimentů

C.1 Experiment 1.2.T

Nástroj	**Čas běhu [s]**	**Potřebná paměť [MB]**	**Průměrná vzdálenost readů od reference**	**Průměrná vzdálenost readů od reference [%]**
Bowtie2	0.63	115.305	1.334	0.993
BWA-SW	2.225	8.490	1.390	0.987
BWA-MEM	0.403	7.784	1.391	0.987
Kart	-	-	-	-
GEM3	0.322	546.635	1.351	0.993

Tabulka 26: Porovnání získaných hodnot pro malá vstupní data

Malá vstupní data reprezentující jeden gen skupiny KIR, předzpracováno. Z hlediska času vykazuje nejlepší výsledky GEM3, nejpřesnější je nástroj Bowtie2. Z hlediska nároků na paměť jsou nejúspornější oba nástroje ze skupiny BWA. Kart nebyl testován.



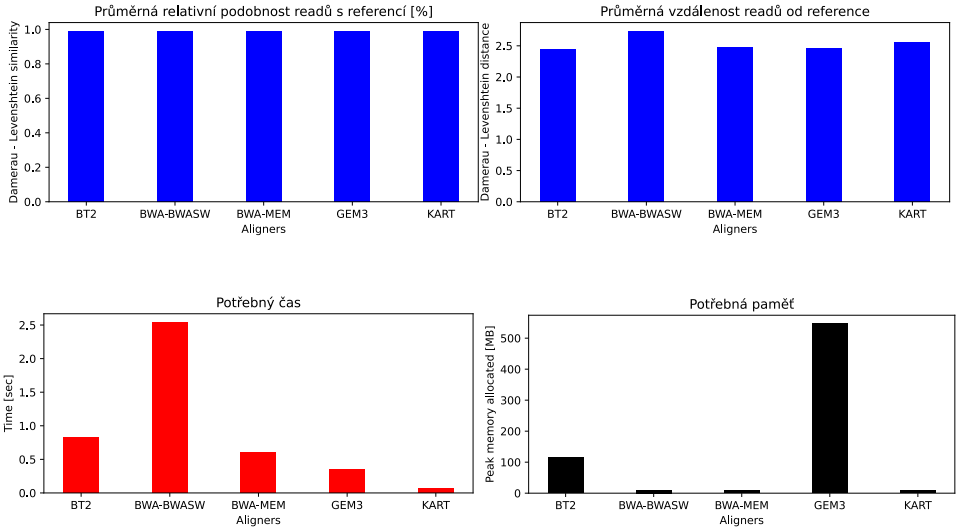
Obrázek 27: Vizualizace výsledků experimentu 1.2.T

C.2 Experiment 1.2.N

Nástroj	**Čas běhu [s]**	**Potřebná paměť [MB]**	**Průměrná vzdálenost readů od reference**	**Průměrná vzdálenost readů od reference [%]**
Bowtie2	0.824	114.458	2.444	0.990
BWA-SW	2.541	9.708	2.736	0.989
BWA-MEM	0.600	8.349	2.485	0.990
Kart	0.070	8.277	2.550	0.989
GEM3	0.351	546.634	2.468	0.990

Tabulka 27: Porovnání získaných hodnot pro malá vstupní data

Malá vstupní data reprezentující jeden gen skupiny KIR, nepředzpracováno. Z hlediska času a nároků na paměť vykazuje nejlepší výsledky Kart, z hlediska přesnosti je nejlepší nástroj Bowtie2.



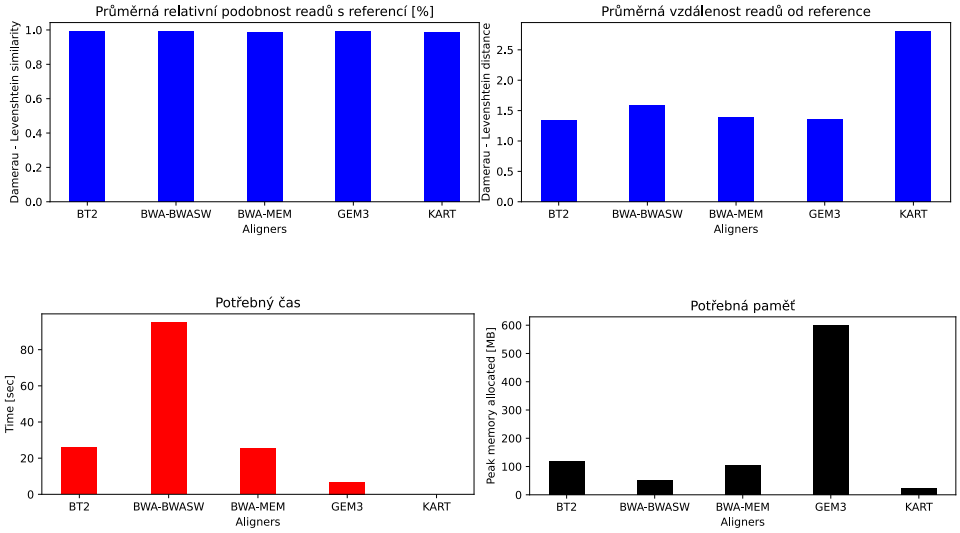
Obrázek 28: Vizualizace výsledků experimentu 1.2.N

C.3 Experiment 2.2.T

Nástroj	**Čas běhu [s]**	**Potřebná paměť [MB]**	**Průměrná vzdálenost readů od reference**	**Průměrná vzdálenost readů od reference [%]**
Bowtie2	25.916	118.697	1.341	0.993
BWA-SW	95.059	51.994	1.583	0.992
BWA-MEM	25.352	102.728	1.397	0.988
Kart	-	-	-	-
GEM3	6.729	599.573	1.353	0.993

Tabulka 28: Porovnání získaných hodnot pro střední vstupní data

Středně velká vstupní data reprezentující 18 genů skupiny KIR, předzpracováno. Z hlediska času vykazuje nejlepší výsledky GEM3, nejpřesnější je nástroj Bowtie2. Z hlediska nároků na paměť je nejúspornější nástroj BWA-SW. Kart nebyl testován.



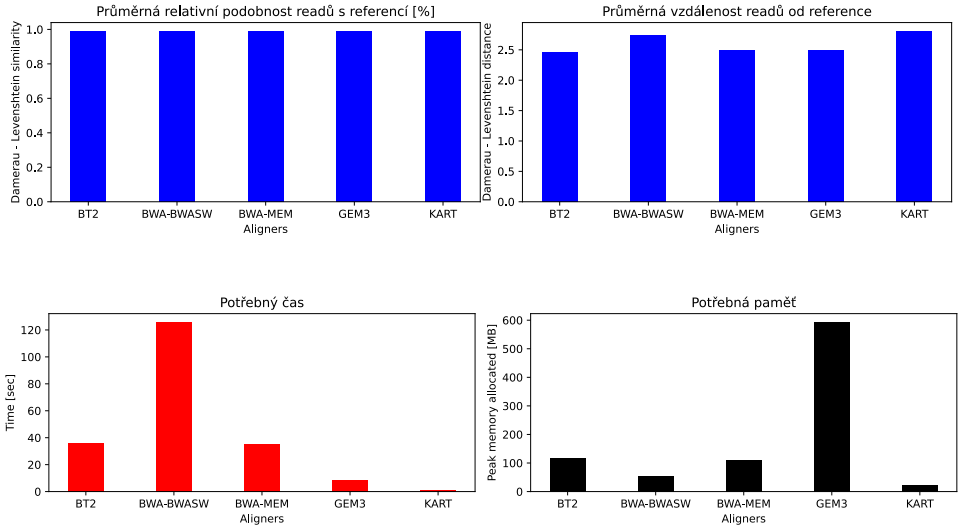
Obrázek 29: Vizualizace výsledků experimentu 2.2.T

C.4 Experiment 2.2.N

Nástroj	**Čas běhu [s]**	**Potřebná paměť [MB]**	**Průměrná vzdálenost readů od reference**	**Průměrná vzdálenost readů od reference [%]**
Bowtie2	36.242	117.558	2.457	0.990
BWA-SW	125.824	53.051	2.734	0.989
BWA-MEM	35.565	111.337	2.496	0.990
Kart	0.847	22.153	2.808	0.988
GEM3	8.840	593.320	2.488	0.990

Tabulka 29: Porovnání získaných hodnot pro střední vstupní data

Střední vstupní data reprezentující 18 genů skupiny KIR, nepředzpracováno. Z hlediska času a nároků na paměť vykazuje nejlepší výsledky Kart, z hlediska přesnosti je nejlepší nástroj Bowtie2.



Obrázek 30: Vizualizace výsledků experimentu 2.2.N