UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULETET

Sanja Mijalković

Pronalaženje mesta integracije HIV virusa u ljudskom genomu

Master rad

Beograd, 2017

Podaci o mentoru i članovima komisije

Mentor

prof. dr Nenad Mitić, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

Članovi komisije

prof. dr Nenad Mitić, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

prof. dr Saša Malkov, vanredni profesor, Matematički fakultet, Univerzitet u Beogradu

dr Miloš Beljanski, naučni savetnik, Institut za opštu i fizičku hemiju

Datum odbrane:

Sadržaj

Spisak slika

Spisak tabela

1 Uvod

Sekvenciranje, tj. digitalizacija genoma je uvela mnoge novine u genomiku i omogućila detaljnije analize u ovoj oblasti. Međutim, sa sobom donosi i mnogo izazova. Naime, genomski podaci su veličine i do nekoliko stotina gigabaza, pa ih je nemoguće obraditi ručno. Otuda se javila potreba za primenom računarskih analiza u svrhu obrade ovog tipa podataka, što dovodi do razvoja bioinformatike kao nove grane informatike.

Bioinformatika je oblast u razvoju i potreba za novim bioinformatičkim rešenjima svakodnevno raste sa pojavom novih genetičkih otkrića. Primenjuje se npr. u svrhu pronalaženja korelacija između genoma ispitivanog organizma i njegovih karakteristika, ali i naslednih ili stečenih bolesti. Kao takva bioinformatika je pored primene pri analizi bolesti sa visokom stopom smrtnosti kao što je rak, našla svoju primenu i u analizi nekih retkih genetskih oboljenja. Moguća je i primena bioinformatike u analizi HIV virusa, što je i centralna tema ovog rada.

HIV virus je danas, nažalost, široko zastupljen i poznato je da je AIDS, bolest koju on izaziva, neizlečiva. Upravo zbog toga privlači veliku pažnju naučnika. Pored toga, još uvek ima dosta nepoznanica o tome kako HIV izaziva AIDS, tj. o HIV integraciji i najčešćim mestima ugradnje u ljudski genom. Dodatni problem je nedostatak alata koji omogućavaju precizne analize takvih mesta.

U radu je opisan nov algoritam za rešavanje ovog problema, nazvan HIVSeeker, koji je detaljno opisan u narednim poglavljima.

1.1 Nukleinske kiseline

Nukleinske kiseline su dezoksiribonukleinska kiseline (u daljem tekstu DNK) i ribonukleinska kiselina (u daljem tekstu kao RNK). DNK i RNK su makromolekuli. Dok je RNK jednolančani makromolekul, DNK je dvolančani, što znači da se sastoji od dva duga polinukleotidna niza, poznatijih kao DNK lanci, čije su gradivne jedinice 4 različita nukleotida [3]. Svaki nuklotid se sastoji od:

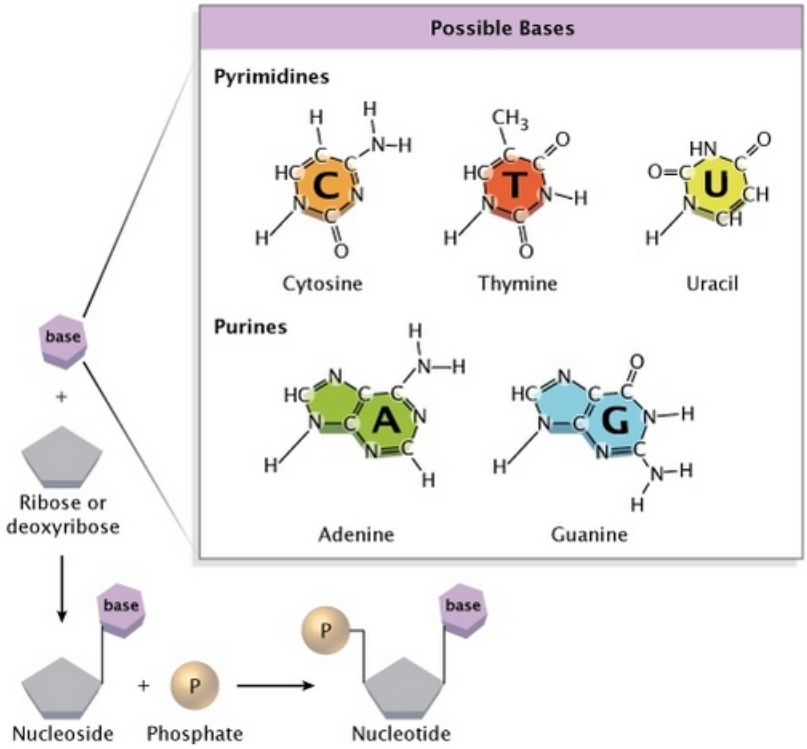
- Azotne baze koja može biti:
  - Purinska - azotne baze koje se sastoje od dva prstena. U ovu grupu spadaju Adenin (A) i Guanin (G)

  - Pirimidinska - azotne baze koje se sastoje od samo jednog prsten, a ovoj grupi pripdaju Citozin (C), Timin (T) i Uracil (U).

- Šećera koji može biti
  - Dezoksiriboza, u slučaju DNK
  - Riboza, u slučaju RNK

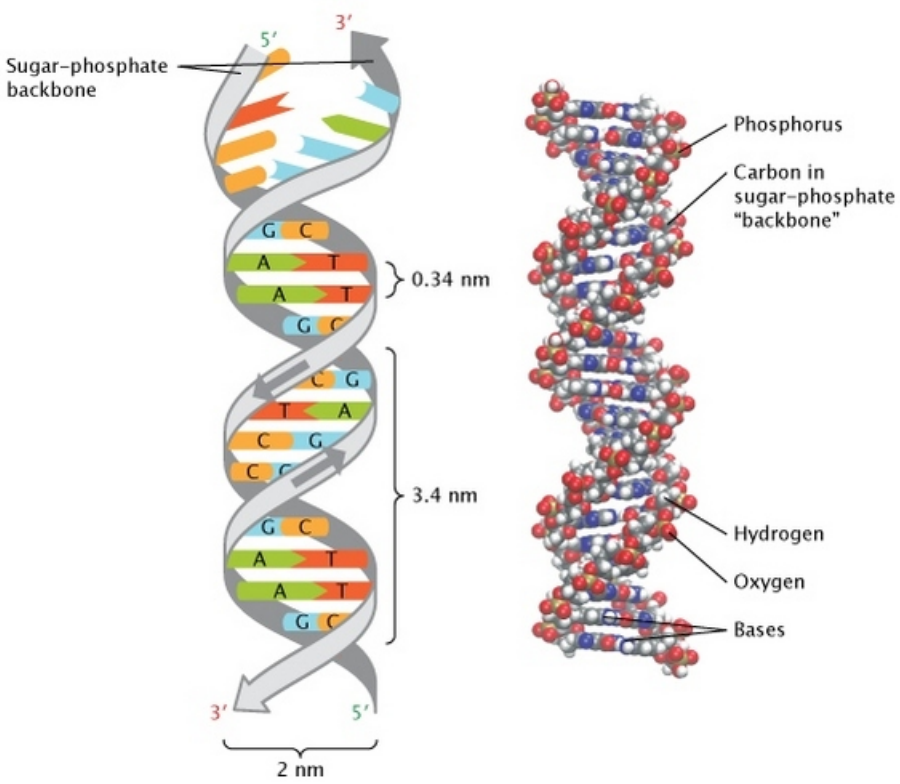
- Fosfatne grupe

A,T,C i G nukleotidi su prisutni u DNK sekvenci, dok je u RNK sekvenci Timin zamenjen Uracilom. Struktura nukleotida je predstavljena na slici 1.



Slika 1: *Strukture nukleotida*

Susedni nukleotidi u okviru jednog lanca su povezani fosfodiestarskim vezama. Dva DNK lanca su povezana vodoničnim vezama između komplementarnih susednih nukleotida. U DNK strukturi, A se uvek vezuje dvostrukom vodoničnom vezom za T, dok se C uvek vezuje za G trostrukom vodoničnom vezom, i obrnuto, formirajući dvostruku spiralu DNK (eng. *double-helical structure of DNA*), slika 2 [4].



Slika 2: *Trodimenzionalna struktura DNK*

U DNK su zapisane informacije pomoću redosleda nukleotida u DNK lancima. Biološke poruke koje se prenose između generacija zapisane u četvoroslovnoj azbuci, A,T,C i G gde svako slovo predstavlja jedno azotnu bazu. Iako različiti organizmi imaju istu strukturu DNK, redosled nukleotida i dužine DNK sekvenci su drugačije pa su samim tim i njihove biološke poruke različite, što omogućava raznolikost vrsta.

Segmenti DNK koji nose informacije se zovu **geni**. Na osnovu gena se procesom transkripcije dobija RNK molekul, a procesom translacije se od RNK molekula dobija protein. Prenošenje informacija na ovaj način je poznato kao **centralna dogma molekularne biologije** [5]. U genomu se nalaze recepti svih proteina koje taj organizam moze da sintetiše [3], zato što su svi geni jednog organizma zapisani u genomu.

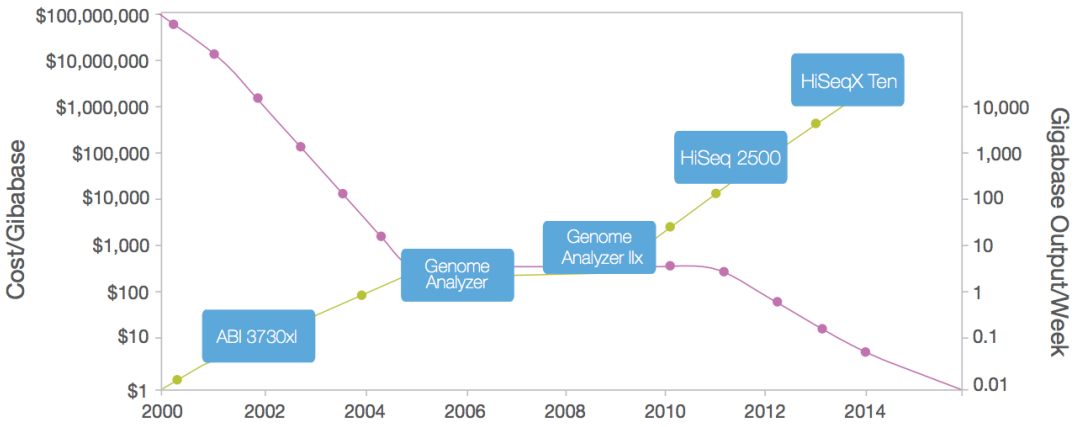
Pored transkripcije i translacije, bitan je mehanizam replikacije DNK. Procesom replikacije se udvaja molekul DNK, a samim tim i sav genetski materijal ćelije. Proces replikacije omogućava ćelijsku deobu.

1.2 Sekvenciranje genoma

Već je spomenuto da sekvence gena, pored ostalih faktora, putem centralne dogme direktno utiču na proteine koji se proizvode u jednoj ćeliji. Kako su geni zapisani u DNK sekvenci, određivanje redosleda baza DNK lanca je jedan od najvažnijih koraka u izučavanju genoma nekog organizma. Sekvenciranje genoma je postupak određivanja redosleda nukleotida u DNK sekvenci.

Sangerova metoda sekvenciranja [8] se pojavila 1977. godine i omogućila je naučnicima da precizno i ponovljivo mogu odrediti sekvencu DNK lanca. Mana Sangerove metode je mala brzina, jer je sekvenciranje jednog celog genoma trajalo nekoliko godina. Nekoliko različitih metoda sekvenciranja se pojavilo u narednom periodu, međutim nedostatak brzine nije prevaziđen. Takve metode se mogu svrstati u **metode prve generacije sekvenciranja** (eng. *First generation sequencing*).

Pojava novih metoda, koje omogućuju paralelna sekvenciranja celog genoma se jednim imenom nazivaju **metode sekvenciranja nove generacije**, u daljem tekstu NGS (eng. *Next Generation Sequencing*). Metode sekvenciranja nove generacije se pojavljuju 2005. godine i pomeraju granice sekvenciranja sa tadašnjih 84 kilobaze (KB) na jednu gigabazu (GB) po pokretanju mašine. NGS uvodi fundamentalno novi pristup koji je napravio revoluciju u sekvenciranju, kako po pitanju kapaciteta, tako i po pitanju brzine i cene istog. Od tog trenutka pa do danas kriva količine sekvenciranih podataka prevazilazi krivu Murovog zakona. Na slici 3 je predstavljen rast količine podataka(zelena linija) tokom godina a u isto vreme i opadanja cene(ljubičasta linija).



Slika 3: *Cena i količina sekvenciranih podataka od 2000. godine*

Dolazak NGS-a je omogućio sekvenciranja celih genoma, što je izazvalo promenu pri-

stupa naučnika prema analizi. Od tog trenutka pa na dalje je sve više podataka postajalo pristupačno, pa se samim tim i spektar mogućih analiza drastično povećao. Već 2014. godine se ograničenje količine sekvenciranja pri jednom pokretanju mašine pomerilo na 1.8 terabaza (TB), što je 1000 puta više podataka od inicijalne pojave NGS.-a. Tada je postalo moguće sekvencirati i do 45 genoma dnevno na jednoj mašini po ceni od do tada neverovatnih 1000$ po uzorku.

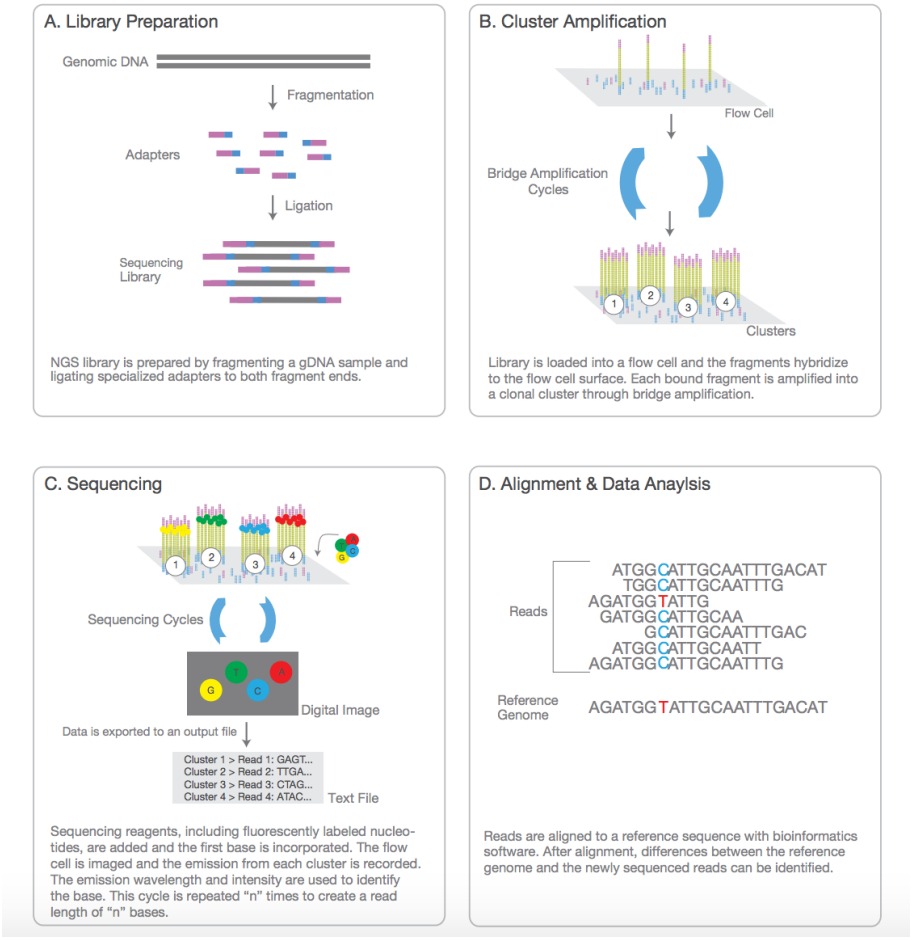
Illumina je jedna od najpoznatijih firmi koje se bave proizvodnjom sekvencera nove generacije. Algoritam po kome se vrši sekvenciranje se sastoji od četiri glavne faze i to su:

- Priprema biblioteka (eng. *Library Preparation*)
- Generisanje klastera (eng. *Cluster Generation*)
- Sekvenciranje
- Analiza podataka

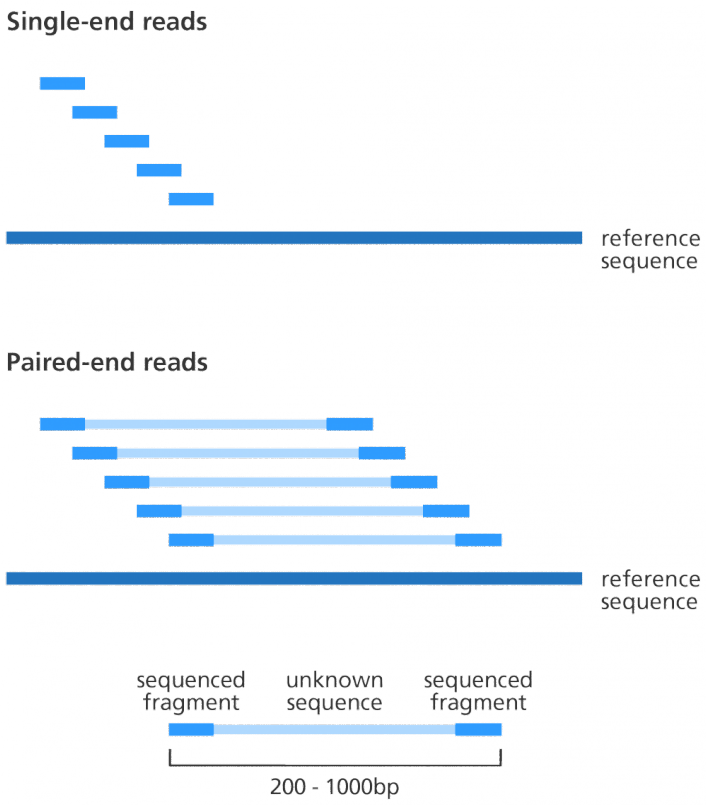
Detaljnije informacije o svim koracima algoritma su dostupne u literaturi [6] a grafička reprezentacija procesa je predstavljena na slici 4.

Problem kod očitavanja sekvence genoma je što korišćenjem Illumina mašine za sekvenciranje nije moguće očitati celu sekvencu genoma odjednom. Zapravo, genom mora da se isecka na sitne komade (fregmente) čija dužina najčešće varira od 300 do 500 baznih parova. Nakon toga se delovi tih fragmenata očitavaju, određen broj baza samo sa jedne ili sa obe strane kao sto je prikazano na slici 5. Rezultat sekvenciranja ovakvim metodama je veliki broj očitavanja čija dužina varira od verzije mašine korišćene pri sekvenciranju, ali se kreće između 40 do 200 baznih parova očitanih sa delova DNK fregmenta. Svakoj očitanoj bazi je pridružen kvalitet očitavanja, odnosno sigurnost da je očitana baza tačna. Sekvence su najčešće zapisane u datoteci u FASTQ formatu, o čemu će biti reči kasnije u ovom poglavlju.

Očitavanja mogu da se čitaju samo sa jedne strane DNK fregmenta, kao što je bio slučaj ranije. Takvo sekvenciranje se zove očitavanje sa jednog kraja (eng. *Single end*). Naknadno je metoda unapređena tako da se očitavanja čitaju sa obe strane, kako bi se sačuvalo više informacija o jednom fregmentu DNK. Takva metoda sekvenciranja se zove upareno sekvenciranje (eng. *Paired end*) i prikazano je na slici 5. Informacija o uparenim očitavanjima se čuva uz identifikator očitavanja, tako da u svakom trenutku može da se prati i uparenje datog očitavanja.



Slika 4: *Četiri faze Ilumina algoritma za očitavanje i tumačenje genomskih podataka. Slika A: Priprema biblioteka i hemije, slika B: Generisanje klastera, slika C: Sekvenciranje, slika D: Analiza podataka*



Slika 5: *Sekvenciranje očitavanja samo sa jednog kraja i upareno sekvenciranje*

1.3 Ljudski genom

Napredak u razumevanju ljudskog genoma tokom dvadesetog veka se može grubo prirodno podeliti u četiri glavne faze. Prvo veliko otkriće je postojanja hromozoma. Hromozomi predstavljaju osnovu ćelijskog nasleđivanja. Tokom druge faze otkrivan je dvolančani molekul dezoksiribonukleinske kiseline (u daljem tekstu DNK). Tokom treće faze, koja odgovara trećoj četvrtini XXog veka, se saznaje više o osnovnim konceptima nasleđivanja, kao i o biološkim mehanizmima ćelije za čitanje informacije zapisanih u genima.

Poslednja četvrtina veka je posvećena pronalazenju sekvence celog genoma i otkrivanju mape gena. Upravo u tom periodu je pokrenut i projekat ljudskog genoma (eng. The Human Genome Project). Započet je 1993. godine a završio se 2003. Cilj projekta je bio da se odredi sekvenca nukleotida koji cine ljudski genom i da se napravi referentni ljudski genom, kao i da se poboljša razumevanje sekvence genoma i napravi mapa gena koja će olakšati dalja istraživanja. Inicijalna verzija ljudskog genoma je pokrivala 94% celog genoma. Reference je sadržala preko 3 milijarde nukleotidnih baza u tačno određenom redosledu.

U projekat ljudskog genoma bilo je uključeno 20 velikih grupa iz Sjedinjenih Američkih država, Velike Britanije, Japana, Francuske, Nemačke i Kine [10].

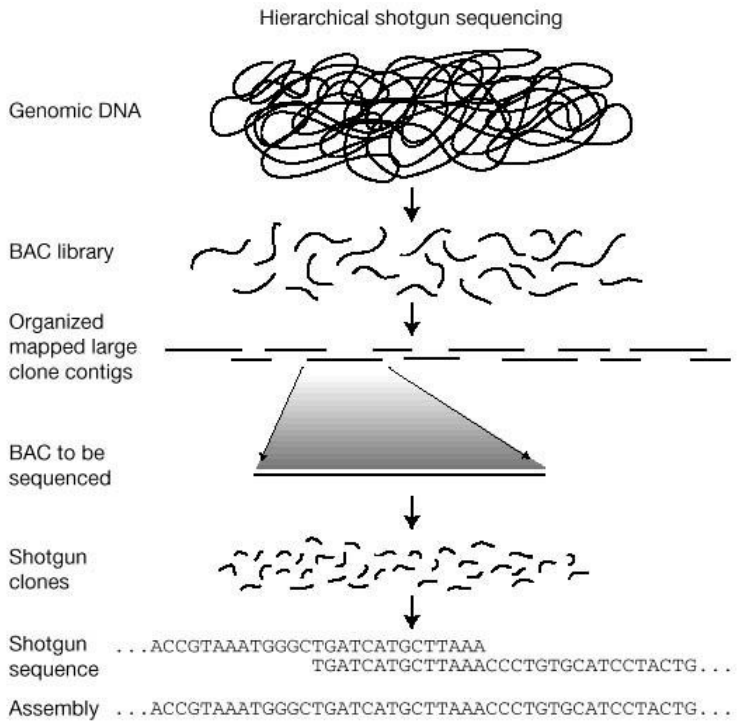
Verovalo se da prva verzija celokupnog ljudskog genetskog materijala treba da bude kompozicija genoma nekoliko osoba koje su dovoljno raznovrsnog etničkog porekla kako bi referentni genom pružao što reprezentativniju sliku o prirodi ljudskog genoma svih ljudi na planeti. Svi davaoci su se prijavljivali na dobrovoljnoj bazi, od kojih je ukupno odrabrano 21. Samo su tri informacije zapamćene i šifrovane za svaki uzorak:

- Godine
- Pol
- Etnička grupa kojoj pripadaju i za koju su se sami opredelili

Nakon analize pomenutih donora, na osnovu kombinacije nekoliko faktora od kojih je ključan bio postizanje etničke raznolikosti kao i nekoliko tehničkih faktora, odabrano je samo njih pet čiji su genomi sekvencirani i ušli u sekvencu prve verzije ljudskog genoma. Pet osoba (dva muskarca i tri žene) jedna Afričko-Američkog, jedna Azijsko-Kineskog, jedna Hispanskog-Meksičkog i dve osobe Kavkaskog porekla [11].

Skica sekvenciranja metodom nove generacije korišćene u projektu ljudskog genoma je

prikazana na slici 6 [10].



Slika 6: *Skica sekvenciranja u projektu Ljudskog genoma*

Kada je sekvenciranje završeno, pažnja je usmerena na sastavljanje referentnog genoma (eng. *Genome assembly*) na osnovu očitavanja sekvenciranih genoma davalaca. Korišćeni algoritmi su predstavljali unapređenja algoritama upotrebljenih za pronalaženje sekvence genoma Drosofile (eng. *Drosophila genome*) koji je do detalja opisan u [12]. Algoritam za sastavljanje genoma se sastoji od pet glavnih faza, izvršavanih sledećim redosledom:

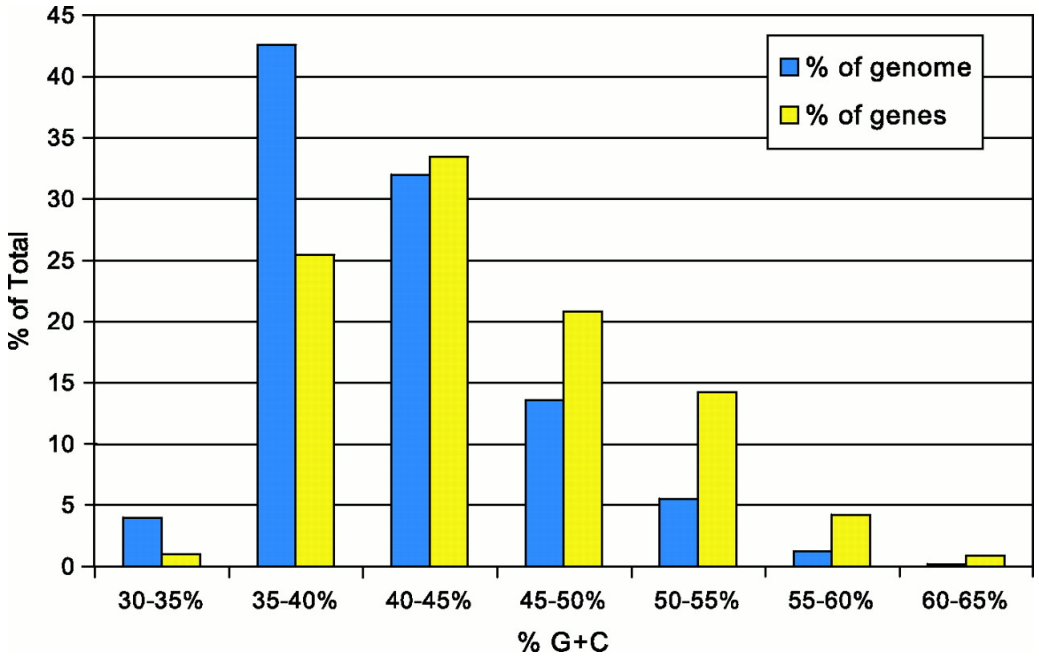
- (eng. *Screener*) Skeniranje svih sekvenci i razdvajanje na dve podgrupe
- (eng. *Overlapper*) Poređenje svakog očitavanja sa svakim drugim očitavanjem u potrazi za preklapanjem minimalnih 40 baznih parova, sa ne više od 6% nepoklapanja u istih 40 baza. Od ovakvih preklapanja se pravi graf
- (eng. *Unitigger*) - Izdvajanje jedinstvenih putanja kroz graf koje predstavljaju nedvosmislene celine genoma
- (eng. *Scaffolder*) Pravljenje Skafolda - većih celina pomoću uparenih očitavanja na velikim udaljenostima

- (eng. *Repeat Resolver*) Razrešavanje ponavljanja u genomu

Detaljnije informacije o algorimu sastavljanja genoma su dostupne u [11].

Na ovaj način je dobijena prva reprezentativna sekvenca ljudskog genoma. To je bio najveći genom koji je tako temeljno sekvenciran do tada, zapravo 25 puta duži od bilo kog prethodno sekvenciranog genoma. To je i prvi genom sisara koji je ikada sastavljen. Izuzetan značaj ovog projekta je i u tome sto je sekvenciran baš genom naše vrste, ljudi [10].

Pored pronalaženja same sekvence ljudskog genoma, bilo je potrebno i anotirati ga. Tokom ovog projekta otkriveno je i analizirano nekoliko karakteristika genoma. Otkrivene su različite distibucije nekoliko bitnih odlika sekvence, kao što su distibucije gena, prenosivih elemenata (eng. *transposable elements*), sadržaj GC baza kao i CpG ostrva u genomu. Na osnovu ovih informacija je bilo lakše izvesti zaključke o funkcijama određenih delova sekvence. Pokazano je i da je broj gena u uskoj vezi sa GC sadržajem u sekvenci genoma, što se može videti na slici 7 [11].



Slika 7: *Veza između GC sadrzaja i gustine gena. Plavi pravougaonici pokazuju procenat genoma koji (sa prozorom 50 kilobaza) sa GC sadržajem. Procentat totalnog genskog sadržaja koji odgovara svakom GC pravougaoniku je predstavljen žutim pravougaonicima. Na ovom grafiku se vidi da oko 5% genoma ima GC sadržaj između 50 i 55%, a da taj deo sadrži blizu 15% svih gena.*

Pretpostavljalo se da postoji 30000 - 40000 gena koji kodiraju proteine, što je bilo neočekivano malo, samo duplo više nego kod muve mada dosta složenije strukture. Ot-

kriveno je i da su na stotine gena zapravo verovatno nastale horizontalnim transferom sa bakterija u nekom trenutku pri evoluciji u lancu sisara. Pronađeno je i 1.4 miliona zamena pojedinačnih baza, u daljem tekstu SNP (eng. *Single Nucleotide Polymorphism*). Sva ova otkrića su pretočena u baze podataka koje su kasnije unapređivane kao i sama sekvenca genoma. Rad na referentnom genomu nikada nije zaustavljen i traje i dan danas.

1.4 Formati podataka

Tokom analize pronalazenja mesta integracije HIV virusa u ljudskom genomu korišćeno je više različitih formata datoteka a najznacajniji su FASTA, FASTQ i BAM datoteke. Pomenuti formati datoteka imaju široku primenu u bioinformatici.

1.4.1 FASTA

FASTA format datoteke je pogodan za zapisivanje nukleotidnih i proteinskih sekvenci. U ovom formatu su obično zapisane sekvence genoma. U upotrebi je više različitih vrsta fasta zaglavlja. Nije redak slučaj da svaka od (velikih) baza podataka koristi svoj poseban oblik fasta zaglavlja (videti npr. [1]). Bez obzira na specifičan oblik zaglavlja koje se koristi, za zapis u fasta formatu važe seledeća opšta pravila:

- Na početku zapisa se nalazi zaglavlje koje sadrži identifikaciju sekvence koja se prikazuje.
- Sekvenca se zapisuje iza zaglavlja.
- Zapis identifikacije se razlikuje od sekvence pomoću znaka veće (">") koji se zapisuje u prvoj koloni.

Primer zapisa dela genomske sekvence ljudskog hromozoma jedan:

```
>chr1 human-genome-19
AGCCGGCCCGCCCGCCCGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGA
CACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGACACATG
CTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAG
CGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGA
GCAAAGTCGCACGGCGCCGGGCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGG
CGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTAGAGTGGTGGC
CAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCT
AGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCC
TCTTGCTCCAACAGTACTGGCGGATTATAGGGAAACACCCGGAGCATATGCTGTTTGGTCTCAGTAGACTCCTAAATATG
```

1.4.2 FASTQ

FASTQ format je tekstualni format datoteke u kome mogu biti zapisane nukleotidne sekvence očitavanja (eng. *reads*) i odgovarajući kvaliteti očitanih baza. Ovaj format predstavlja nadgradnju FASTA formata, kome nedostaje mogućnost čuvanja kvaliteta zapisa nukleotidne sekvence. Postao je široko korišćen i često je standardni format datoteke za razmenu sekvenciranih podataka, kao i za format ulaznih podataka u bioinformatičkim analizama [2]. Svako čitanje je opisano pomoću četiri uzastopne linije u FASTQ datoteci u sledećem redosledu:

- Identifikator čitanja, uvek pocinje karakterom "@", praćen sekvencom koja jednoznačno opisuje samo čitanje, nakon čega opciono može slediti dodatni opis
- Sekvenca očitanih nukleotida
- Separator koji počinje znakom "+" nakon kog može da se navede komentar
- Sekvenca karaktera koji predstavljaju kvalitete očitavanja odgovarajućih nukleotida iz nukleotidne sekvence. Ova sekvenca mora biti iste dužine kao nukleotidna sekvenca

Ukoliko se radi o sekvenciranim podacima sa uparenim očitavanjima (eng. *paired-end reads*) onda se parovi najčešće smeštaju u dve odvojene datoteke, jedna za uparena čitanja sa jedne strane, druga za uparena čitanja sa druge strane. Najčešće se to može videti u imenu datoteke, a na kraju sekvence identifikatora očitavanja se nalazi "/1" ili "/2".

Prikazana su tri primera čitanja delova ljudskog genoma u FASTQ formatu. Prvi primer očitavanja ima identifikator "C19C7ACXX121128:4:1305:10471:4944", dok se na kraj identifikatora dodaje "/1" što označava da je to prvo očitavanje u paru uparenih očitavanja. Drugi red predstavlja niz očitanih baza dok se u četvrtom redu nalaze pridruženi kvaliteti očitavanja. To znaci da je "B" kvalitet prve očitane baze, u ovom slučaju baze "A", kvalitet druge baze "A" je "C", kvalitet treće baze "C" je "@" i tako dalje. Kvaliteti očitanih baza su zapisani u posebnom formati i vise informacija je dostupno u [9]. Isto važi i za drugi i treći primer.

Prvi primer:

```
@C19C7ACXX121128:4:1305:10471:4944/1
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCG
+
BC@FFFFDHHHHHGHGGIIEHHJIAHIGGGI@EGFGHIFHFGIJIIIJIIIGHI<GEGGHGCHGJ>C8C@@FC:CEGIG
```

Drugi primer:

```
@C19C7ACXX121128:4:1305:10471:4945/1
TCTTGCTCCAACAGTACTGGCGGATTATAGGGAAACACCCGGAGCATATGCTGTTTGGTCTCAGTAGACTCCTAAATATG
+
CCCFFFFFGHDHHJIJIIJEGGIJJJJJJJIJJJJJJJIGHGIIIGHHJIJIJHIJJJIIJJJIJJIJGEHGHFFFFEEC
```

Treći primer:

```
@C19C7ACXX121128:4:1305:10471:4946/1
CTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAG
+
CEECCCDCADDDDDDDDBCD#FFDHHHHHGHGGIIEHHJIAHIGGGI@EGFIJII<GEGGHGCHGJ>C8C@@FC:CEGIG
```

1.4.3 BAM/SAM

SAM format datoteke (eng. *Sequence Alignment/Map format*) je tekstualni format u kome su podaci u redu razdvojeni tabulatorom kao graničnikom. Uglavnom se koristi za zapis poravnatih sekvenci, tj da opiše kako su sekvence poravnate u odnosu na referentni genom. Sastoji se iz dve sekcije, sekcije zaglavlja koja nije obavezna i sekcije poravnatih očitavanja.

Sekcija zaglavlja sadrži informacije o zapisima poravnanja, tj. detalje o samom algoritmu poravnanja. Redovi zaglavlje počinju simbolom "@" i po tome se razlikuju od redova koji pripadaju sekciji poravnatih očitavanja. Svaka linija je tabelarnog formata i više različitih tipova informacija je zapisano u linijama zaglavlja, a tip podataka u svakom redu se zapisuje kodom odmah nakon simbola "@". Nakon koda dozvoljeno je uneti polja te linije zaglavlja u formatu zapisa *TAG:VALUE* gde je *TAG* predstavljen sa dva karaktera koji definišu format dok se sadržaj zapisuje u *VALUE* formatu.

Sekcija poravnatih očitavanja sadrži poravnata očitavanja sa svim detaljnim informacijama o poravnanjima. Svako očitavanje je zapisano jednim redom u tekstualnoj datoteci pomoću nekoliko obaveznih i nekoliko neobaveznih polja. Postoji 11 obaveznih polja koja moraju biti popunjena za svako poravnato očitavanje i pružaju osnovni skup informacija o njemu [13]. Obavezna polja su:

- QNAME: Identifikator očitavanja
- FLAG: Zastavica čija vrednost se kreće od 0 do 65535 a binarni zapis tog broja je kombinacije različitih zastavica, od kojih svaka ima svoj binarni broj.
- RNAME: Ime hromozoma u referenci na koji je očitavanje poravnato ili "*"ukoliko je ta informacija nedostupna
- POS: Krajnje leva pozicija na genomu kojom počinje poravnanje očitavanja ili "0" za očitavanje čije poravnanje nije pronađeno

- MAPQ: Kvalitet poravnanja
- CIGAR: CIGAR niska opisuje poravnanje očitavanja (eng. *Concise Idiosyncratic Gapped Alignment Report*)
- RNEXT: Ime hromozoma u referenci na koji je upareno očitavanje poravnato
- PNEXT: Krajnje leva pozicija na genomu kojom počinje poravnanje uparenog očitavanja
- TLEN: Dužina fragmenta
- SEQ: Nukleotidna sekvenca očitavanja ili "@"ukoliko ova informacija nije dostupna
- QUAL: Sekvenca karaktera koji predstavljaju kvalitete očitavanja odgovarajućih nukleotida iz nukleotidne sekvence. Ova sekvenca mora biti iste dužine kao nukleotidna sekvenca (SEQ)

Postoje dodatna opciona polja koja se popunjavaju po potrebi i različiti algoritmi poravnanja unose različit skup opcionih polja. Sva opciona polja su formata *TAG:TYPE:VALUE*. *TAG* je predstavljen sa dva slova koji su naziv tog opcionog polja, *TYPE* je tip polja a nakon toga sledi vrednost.

BAM datoteka predstavlja binarni format SAM datoteke. Za BAM datoteku postoji index datoteka koja je često neophodna kako bi njen sadržaj mogao biti obrađen.

Više informacija o SAM/BAM formatu datoteke je dostupno u [13].

Primer očitavanja poravnatih u odnosu na referencu praćen primerom zapisa tih očitavanja u BAM datoteci može se videti na slikama 8 i 9.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                  ATAGCT..............TCAGC
-r003                        ttagctTAGGC
-r001/2                                    CAGCGGCAT
```

Slika 8: *Zapis očitavanja u BAM datoteci*

U algoritmu HIVSeeker, koji je centralna tema ovog rada, će, pored obaveznih polja, biti korišćena neka opciona polja, od kojih je najznačajnije "SA" polje koje govori o tome

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Slika 9: *Očitavanja poravnata u odnosu na referencu*

da očitavanje ima dodatno poravnanje na nekoj drugoj poziciji u genomu. U ovom polju su upisane sve informacije o tome gde se dodatno poravnanje nalazi. Bez korišćenja ove zastavice i funkcionalnosti algoritma BWA-MEM da upiše tu informaciju, ne bi bilo moguće upotrebiti algoritam HIVSeeker.

2 HIV virus

Hiv virus pripada porodici Retrovirusa (Retroviridae) koja poseduje nekoliko jedinstvenih karakteristika po kojima zavređuje posebno mesto u taksonomiji i virusologiji. Ovi virusi imaju neobičan životni ciklus koji ih izdvaja zbog kršenja centralne dogme molekularne biologije po kojoj se informacija od RNK kreće ka proteinima i nikada unazad ka DNK. Ovo nije jedini takav izuzetak u prirodi i kasnije su nova otkrića pokazala da je protok informacija u nukleinskim kiselinama mnogo složeniji nego što se u početku smatralo. Ipak, njihovo otkriće je napravilo svojevrstan preokret u medicini i molekularnoj biologiji.

Iako potpuno razumevanje ovih virusa i njihov značaj u opštoj populaciji dobijaju na značaju tek krajem dvadesetog veka, prvi članovi ove porodice su identifikovani još davne 1908 [15]. Danski tim naučnika je pokazao da je pileća leukoza, bolest iz porodice malignih krvih poremećaja, izazvana virusom koji su nazvali *avian leukosis virus*. Na samom početku je pokazano da su ovi virusi značajni u nastanku malignih oboljenja, iako mehanizam nastanka bolesti nije bio jasan. To je dodatno motivisalo naučnike da pronalaze nove tumore izazvane virusima kod ptica, i taj broj je do 1930. prevazišao 20. U tom periodu započinje proširivanje ovog biološkoh modela na sisare. Decenije su prošle u istraživanju uticaja retrovirusa na onkogenezu nižih sisara bez odgovora na pitanje - da li postoje ljudski retrovirusi od kliničkog značaja? Odgovor dolazi tek 1980. godine kada je otkriven HTLV-1, retrovirus koji izaziva vrstu agresivne T-ćelijske vlasaste leukemije. U tom periodu svetom se širi neobična nova bolest koja se manifestuje padom imuniteta sa raznorodnom simptomatologijom. Ovo su prvi registrovani slučajevi virusom izazvane humane imunodeficijencije, a pravi uzročnik ostaje nedovoljno istražen, čemu je posvećeno više pažnje u narednom periodu.

Jedan od naučnika najzaslužnijih za izolaciju i opis HIV-a, američki biohemičar Robert Galo u svom revijalnom radu iz 2005. godine navodi da je deo interesovanja za retroviruse bio u njihovoj sposobnosti da se prenose između vrsta - ova osobina nije nečuvena ali nije ni česta, upravo zbog specifičnosti biohemijskog aparata koju svaki virus obično zahteva [16]. Galo ističe da su svi dotadašnji slučajevi međuvrsnog prenosa bili stari događaji otkrivani kao evolutivne promene u genomu, dok je virus gibonske leukemije prešao sa gibona na vunastog majmuna, iako ove dve vrste majmuna nisu filogenetski bliske, što je potvrdilo prenos između dveju vrsta u toku života. To je probudilo njegov interes i usmerilo ga ka istraživanju koje dovodi do rada koji prvi opisuje HIV kao uzročnika sindroma stečene humane imunodeficijencije (eng. *Acquired ImmunoDeficiency Syndrome*

- *AIDS*). Iste godine dva različita instituta su objavila relevantne radove - u aprilu 1984. Lancet je objavio rad Luka Montanjea sa Paster instituta koji dokumentuje izolaciju HIV-a; u maju izlazi Galov rad koji daje HIV-u centralnu ulogu u AIDS-u [17]. Patent na otkriće i kasnije testiranje je pripao američkom timu, nakon čega je usledila borba dve države. Preko dve decenije tužbi, presuda, dogovora, pomirenja i saradnje je konačno razrešeno dodelom Nobelove nagrade za medicinu koju je dobio Montanje.

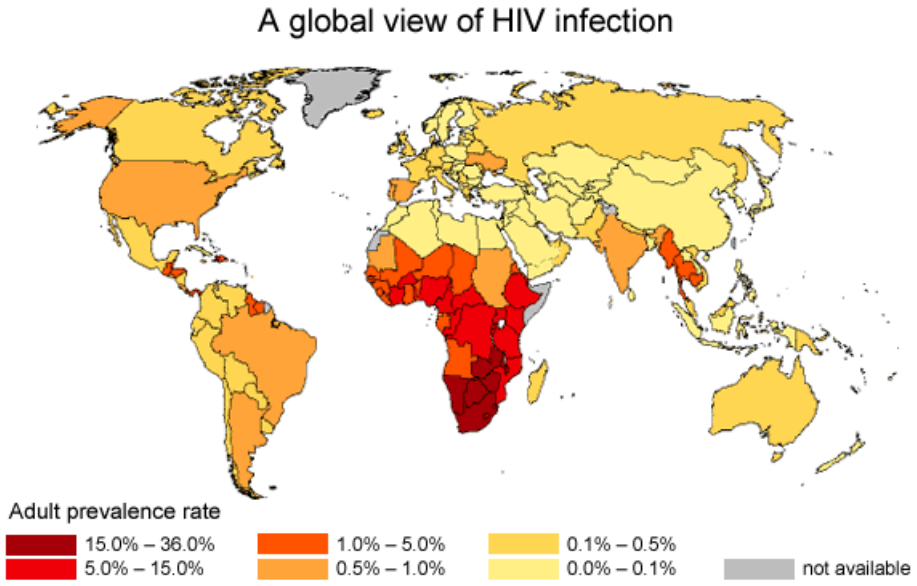
2.1 Epidemiologija

Virus koji je identifikovan od strane Montanjea i Galoa se danas naziva HIV-1 virus i on je glavni svetski patogen koji dovodi do AIDS-a. HIV-2 je otkriven nesto kasnije, manje je proučavan i uglavnom je odgovoran za infekcije na Afričkom kontinentu. U daljem tekstu, ukoliko nije naglašeno, pretpostavlja se da se informacija odnosi na HIV-1.

Već pri pojavljivanju prvih slučajeva stečene imunodeficijencije u Atlanti u SAD postalo je jasno da su izvesne socijalne grupacije izloženije i sklonije virusu - mladi homoseksualci bez urođenih sklonosti ka imunodeficijenciji i intravenski zavisnici od narkotika. Krvno-seksualni prenos je dosta rano ustanovljen kao način prenosa virusa. Od opisivanja i izolacije virusa do razvijanja prvih testova na antitela u krvi prošlo je samo nekoliko godina i prvi testovi su bili dostupni već 1985. godine. Kasnije se distribucija inficiranih pojedinaca menja i zahvata širu i nespecifičniju populaciju. Danas je distribucija geografski dosta uslovljena - najteže je pogođena Afrika u kojoj se po procenama nalazi 25 miliona inficiranih pojedinaca dok je 3 miliona u uznapredovalom stadijumu [18]. Procenjuje se da je ukupno 60 miliona pojedinaca zaraženo virusom od njegovog pojavljivanja, sa 5 miliona novih slučajeva godišnje. Najviše novih slučajeva se registruje u istočnoj Evropi, Kini i jugoistočnoj Aziji, te bi ova žarišta mogla po broju da nadmaše i samu Afriku. Svetska distribucija HIV-a je prikazana na slici 10 [19]. Poreklo virusa nije poznato.

2.2 Kliničke manifestacije

HIV virus pripada rodu Lentivirus, od latinske reči *lentus, lenti m.* što znači spor [18]. Ovakvo ime pokazuje da je relativno ranu ustanovljeno da virus ima dugu asimptomatsku fazu, što znači da nakon primarne infekcije koja može nastati krvnim ili seksualnim putem posle dve nedelje nastupa nespecifična akutna virusna simptomatologija - temperatura, upala ždrela, osipi, malaksalost, gubitak telesne težine. Posle nekoliko nedelja bolest ulazi u fazu kliničkog mirovanja koja ima promenljivo trajanje, ali samo kod manje od 1%



Slika 10: *Globalna mapa stepena zaraženosti HIV virusom*

pacijenata traje manje od 2 godine. Nakon dugog mirovanja dolazi do postepenog sloma mnogih sistema usled odsustva imunog odgovora. Infekcija virusom se odlikuje teškim i upornim infekcijama od kojih su neke relativno specifične za AIDS: respiratorna infekcija sa *Pneumocystis Carinii*, sistemske i lokalne gljivične infekcije, infekcije mozga i pojava Kapošijevog sarkoma. Interesantno je napomenuti da postoji nekoliko maligniteta koji se pojavljuju kao posledica pada imuniteta, ali nijedan od njih nije direktno izazvan HIV virusom, iako je u ostatku porodice Retroviridae to česta pojava da integracija virusa dovede do maligninteta. Ovo se objašnjava mehanizmom onkogeneze u kojoj se virusni geni zamenjuju važnim onkogenima domaćina - dakle, prepisivanje koje dovodi do karcinoma ne dovodi do virusne replikacije, pa su u ovom slučaju kancerogeneza i akutna virusna infekcija ekskluzivno disjunktne. Iako direktnim dejstvom HIV ne izaziva kancer, kao sekundarni kofaktor ima velikog uticaja na stvaranje maligniteta koji se pojavljuju u 4-10% nelečenih pacijenata [20]. Razumevanje ovih razlika u konačnom efektu virusne infekcije ima implikacije na kasniju primenu metodologije predložene u ovom radu.

Paralelno sa opisanim makroskopski primetnim kliničkim događajima, na ćelijskom nivou se odvija životni ciklus virusa. Primarna infekcija počinje ulaskom virusa u krv preko sluznice ili krvi (seksulani kontakt, krvni kontakt ili vertikalni prenos sa majke). Virusu su potrebni receptori CD4 i CCR5 da bi ušao u ćeliju - ovi membranski proteini se nalaze na posebnoj vrsti T-limfocita koji se nazivaju CD4 ili CD4+ limfociti, a često

se nazivaju i T-helper ćelijama zbog njihove centralne uloge u pomaganju i koordinisanju efektivnog imunog odgovora. Nakon ulaska u ćeliju dolazi do ugradnje virusa u genom domaćina i njegove aktivne replikacije i transkripcije što dovodi do stvaranja novih virusa. Zreli virusi se formiraju, uništavaju CD4 T-helper ćeliju domaćina i inficiraju nove CD4 ćelije. Ova faza koincidira sa akutnom fazom HIV infekcije, i tokom nje dolazi do uništenja velikog dela ovih esencijalnih ćelija, naročito podvrste čije se ćelije nazivaju memorijske T-helper ćelije. Nakon ove intenzivne faze replikacije, dolazi do smanjenja količine virusa u krvi i organizmu, a jedan deo virusa ostaje u stanju mirovanja u DNK materijalu živih CD4 ćelija. Virusi su u većini u latentnom stanju, što znači da se ne prepisuju, niti pojavljuju u citoplazmi zaraženih ćelija. Ova faza odgovara fazi kliničkog mirovanja, ali činjenica da sa vremenom dolazi do terminalne faze bolesti (AIDS) govori da postoji tiha replikacija jednog dela umetnutog virusa koja dovodi do stabilnog pada CD4 limfocita. Kada broj ovih ćelija padne ispod kritične granice, dolazi do manifestnog AIDS-a.

3 Problem pronalaženja mesta integracije HIV virusa u ljudskom genomu

3.1 Opis problema

HIV virus, kao jedan od najopasnijih virusa današnjice, za koji ne postoji zvaničan lek privlači pažnju kako naučnika tako i velikih farmaceutskih kuća. Pored toga što je virus već, uslovno rečeno, davno otkriven još uvek njegove metode integracije i širenja infekcije nisu potpuno jasne.

Takođe se pokazalo da, iako je poznato da postoji latentno stanje inficiranih ćelija HIV virusom, teško je izmeriti količinu ćelija u latentnoj fazi. Ukoliko se ova količina biti procenjuje na osnovu broja virusnih infektivnih čestica, onda će količina podcenjena jer se u latentnoj fazi virus, bar koliko je do sada poznato, ne replikuje. Sa druge strane, ukoliko se pokuša proceniti količina ćelija inficiranih HIV virusom u latentnoj fazi na osnovu broja ugradnja HIV virusa u ljudski genom, onda će ona biti precenjena iz razloga što su mnoge integracije defektne i nefunkcionalne, samim tim ne dovode ćeliju do latentnog stanja [28].

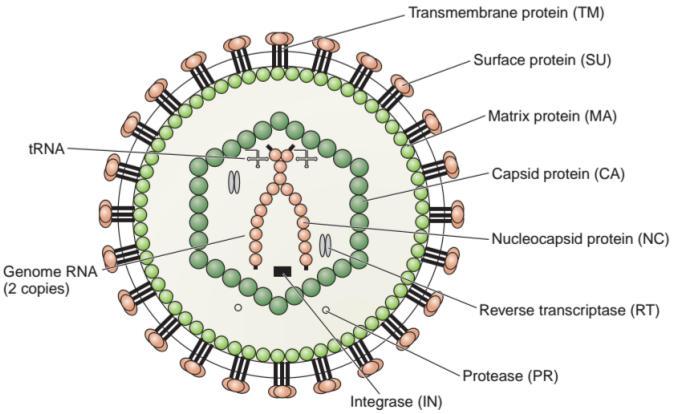
Procenjivanje rezervoara latentnih ćelija je još uvek nedefinisano, poznato je da bi inficirana ćelija ušla u latentno stanje mirovanja mora se desiti kompletna integracija HIV virusa sa netaknutom sekvencom i da su se posledično prekinule, tj. pauzirale replikacije i aktivnosti svih gena integrisanog HIV virusa [29]. Da li lokacija ugradnje virusa utiče na predispoziciju da uđe u latentno stanje je diskutovano u nekoliko ranijih istraživanja [30], [31] i [32] ali nije izveden konačan zaključak.

Već duži niz godina se analizira metod ugradnje virusa, ali mesta integracije nisu sasvim poznata. Postoje različita istraživanja ne temu pronalaženja verovatnih mesta ugradnje i takozvanih vrućih tačaka (eng. *hotspots*) u ljudskom genomu, kao što je [33]. Međutim nikada nije razjašnjen uzrok integracije sekvence baš na ta mesta niti priroda odabira mesta ugradnje.

Iz prethodno navedenog se vidi da je potrebno detaljnije istraživanje mesta integracija HIV virusa u ljudskom genomu. Upravo iz tog razloga je ovaj rad posvećen predstavljanju i opisivanju toka analize i algoritma za pronalaženje mesta integracije. Ne postoji široko prihvaćen alat u bioinformatičkoj zajednici korišćen za ove svrhe, tako da ne postoji drugi alat kojim je bilo moguće uporediti rezultate, ali je potvrda ispravnosti rezultata urađena drugim putem koji je opisan u poglavlju 5.

3.2 Molekularni mehanizmi infekcije i replikacije retrovirusa

Virusna infektivna čestica-partikula se sastoji od genetskog materijala sa proteinskim omotačem koji se zajedno nazivaju nukleokapsid i virusnog omotača koji se gubi pri ulasku u ćeliju ali ima funkcionalnu ulogu u ekstracelularnoj sredini i pri infekciji ćelije. Sam omotač se sastoji iz fosfolipidnog dvosloja poput ćelijske membrane eukariotskih ćelija i omogućava laku fuziju sa njom putem hidrofobnih interakcija. U fosfolipidni dvosloj su, baš kao i kod eukariota, umetnuti površinski i transmembranski proteini. Uz kapsid su povezana tri već sintetisana enzima koji imaju ulogu u obavljanju ključnih koraka virusnog replikativnog ciklusa: proteaza (PR), integraza (IN) i reverzna transkriptaza (RT) [21]. U samoj sredini se nalaze dva lanca RNK kao na slici 11.

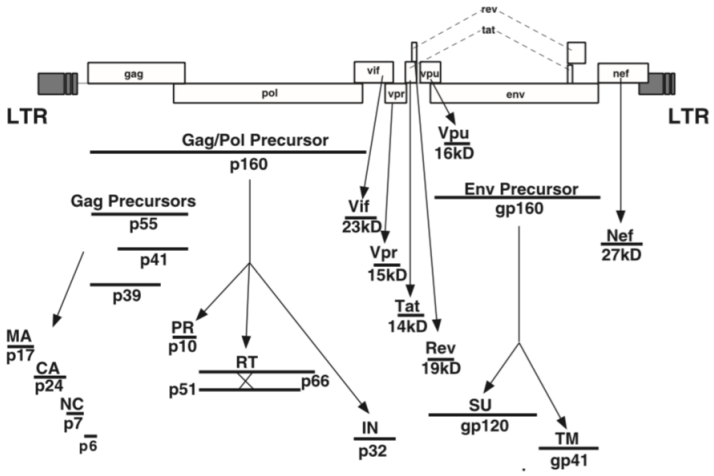


Slika 11: *Virusna infektivna čestica*

3.2.1 Molekularna struktura restrovirusa

Virusni genom ima samo tri gena koja su zajednička za celu virusnu familiju, i nazivaju se Gag, Pol i Env. Svi proteini prisutni u virusnoj čestici su kodirani sa ova tri gena. Uz to, pokazano je da HIV, pored obavezna tri gena, ima još jednu jedinstvenost u odnosu na svoje bliske rođake. Dokazano je da postoje dodatni višestruki i preklapajući otvoreni okviri čitanja koji omogućuju kodiranje više proteina. Novija istraživanja su pokazala da HIV-1 i HIV-2 imaju svoje unikatne gene koji ih razlikuju od ostalih retrovirusa - HIV-1 poseduje Vpu gen, dok HIV-2 poseduje Vpx gen. Prepisivanjem genoma se dobijaju poliproteinski prekursori - prelazni produkti koji se dalje obrađuju do finalnih funkcionalnih i strukturnih proteina kao sto je prikazano na slici 12. Na krajevima genoma se nalaze identične LTR sekvence (eng. *Long Terminal Repeat*) i one igraju ulogu pri integraciji.

Sa svakom integracijom LTR region raste, tako da se na osnovu broja stečenih mutacija u LTR regionima može proceniti koliko je prošlo vremena od prve integracije virusa.

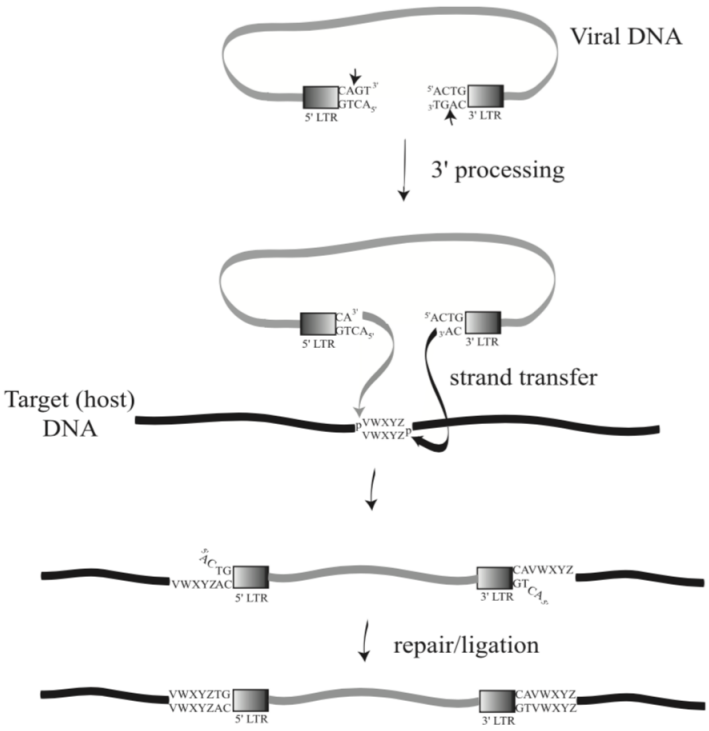


Slika 12: *Proteini koje kodira HIV virus*

Funkcionalni enzimi potiču sa Pol segmenta, Gag kodira strukturne proteine a Env kodira proteine omotača (eng. *envelope*). RNK koja kodira ove proteine se po ulasku u ćeliju prepisuje uz pomoć reverzne transkriptaze relativno složenim mehanizmom koji uključuje dva nedovoljno objašnjena skoka sa kraja na kraj RNK obrasca. Kao rezultat dobija se dvolančana DNK, takozvana komplementarna DNK, koja čini preintegracioni kompleks. Pri reverznom prepisivanju izostaje kontrolna aktivnost polimeraze koju poseduje većina drugih poznatih polimeraza, pa se u ovom procesu unosi dosta pojedinačnih grešaka, i do 10 nukleotida. Zbog ove velike varijabilnosti, ponekada se retrovirusi ne smatraju uniformnom klasom, već takozvanom kvaziklasom [21].

3.2.2 Integracija retrovirusa

Integracija počinje kada se iseku dva nukleotida sa 3' kraja oba lanca provirusne DNK, odmah iza dobro konzervirane sekvence CA. Lepljivi krajevi koji se dobijaju uz pomoć enzima integraze napadaju dve tačke u razmaku od 4-6 baznih parova. Ovim se iseca domaćinska DNK takođe praveći lepljive krajeve koji "štrče" spomenutih 4-6 baza. Interesantno je primetiti da se dva lepljiva kraja ne podudaraju i nisu komplementarni, te da dvonukleotidni fragment sa virusnog 5' kraja ostaje da slobodno "visi". Mehanizmi DNK popravke domaćina će popuniti prazninu od 4-6 baza sa obe strane, čime je napravljen artefakt integracije - mala duplikacija koja ograničava virus. Na slici 13 je šematski prikazan proces ugradnje.



Slika 13: *Skica procesa ugradnje HIV virusa*

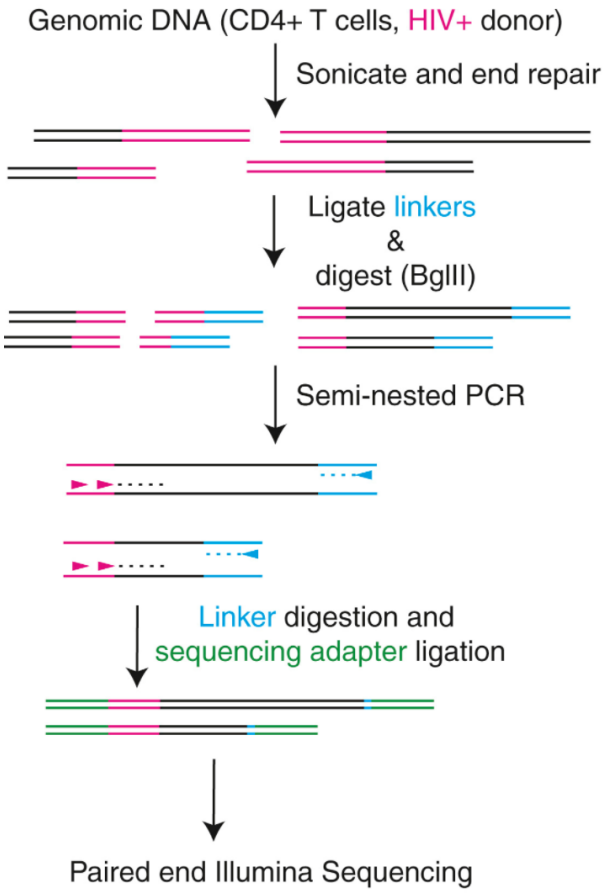
S obzirom na način odvijanja procesa ugradnje, bitno je naglasiti da je moguća integracija i u jedan i u drugi lanac DNK, što isto mora biti uzeto u obzir pri pisanju algoritma za pronalaženje mesta ugradnje.

3.2.3 Priroda podataka korišćenih za razvoj algoritma

Podaci za testiranje i analizu preuzeti su iz rada [34]. Sekvencirani su genomi osoba inficiranih HIV virusom, kao i nekoliko zdravih osoba radi kontrole. U svrhe testiranja analize, korišćeni su uzorci zaraženih osoba. Zbog složenosti procesa sekvenciranja, u radu je prikazana samo šema procesa sekvenciranja na slici 14. Sekvencirane su ćelije imunog sistema, CD4+ T ćelije, pošto su one inficirane HIV virusom.

Prosečna pokrivenost na mestima integracije je u proseku 3000 očitavanja, ali ide i do 290000.

Bitno je naglasiti i da se ne garantuje sekvenciranje ugradnje oba kraja HIV virusa, jer je pitanje kako će se hromozomi iscepkati na fregmente. Zbog toga, u algoritmu HIVSeeker nije implementirano uparivanje mesta integracije već je svaka integracija prijavljena nezavisno u tabeli, što bi značilo da u idealnom slučaju je za svako mesto integracije moguće



Slika 14: *Skica procesa sekvenciranja*

pronaći po 2 reda u tabeli, jedno tipa *START* a drugo tipa *END*.

Zbog složenosti procedure, svi detalji o pripremi materijala za sekvenciranje i samom sekvenciranju su dostupni u radu [34].

4 Tok analize za pronalaženja mesta integracije HIV virusa u ljudskom genomu

Problem pronalaženja mesta ugradnje virusa je složen problem i sa biološke i sa računarske strane. Kako bi pronalaženje mesta integracije HIV virusa u ljudskom genomu bilo moguće potrebno je uraditi niz operacija nad ulaznim podacima koje će biti detaljno opisane u ovom poglavlju. Sama analiza pronalaženja mesta integracije se odvija u dva glavna koraka:

1. Priprema podataka se sastoji od poravnanja očitavanja na specifičan način, povezivanjem nekoliko već postojećih alata i njihovom podešavanjem kako bi rezultujuća BAM datoteka bila pogodana za ulaz u drugi deo analize, algoritma za pronalazenje mesta integracije HIV virusa.

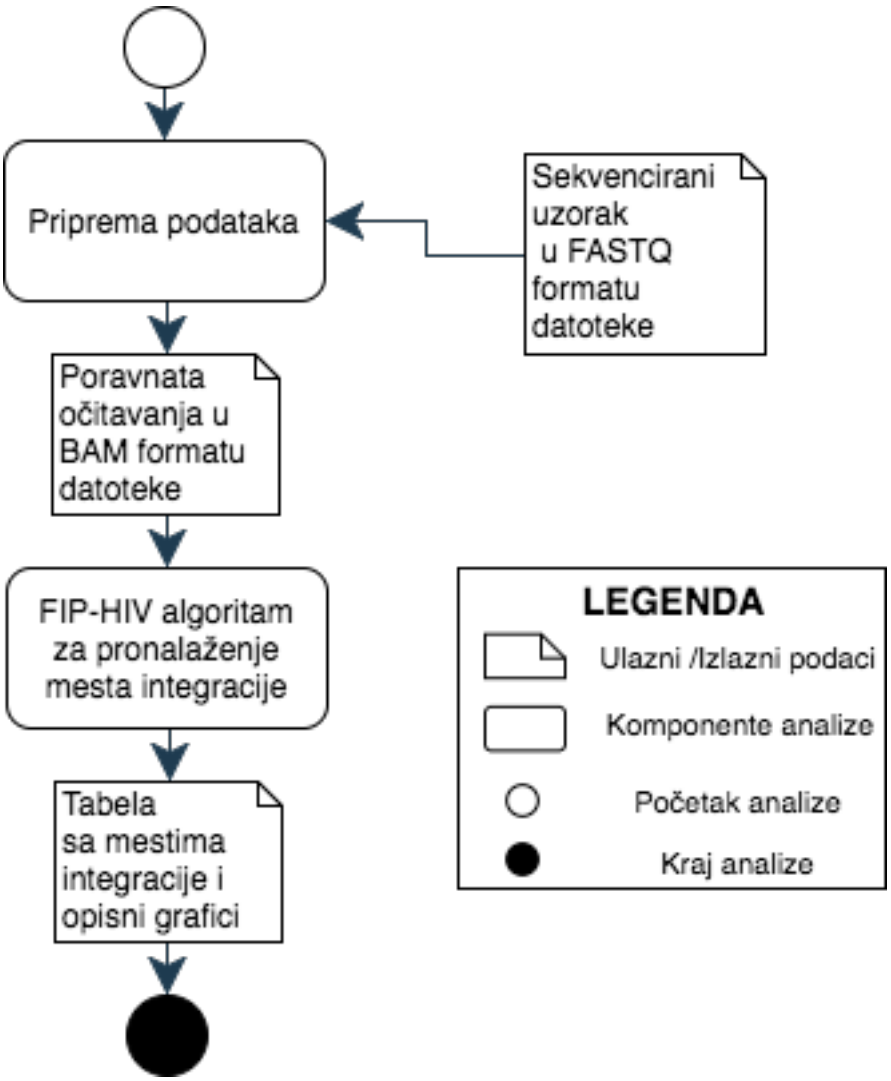
2. Primena algoritma HIVSeeker koji služi za pronalaženje mesta integracije kao i pouzdanosti u izračunatu poziciju integracije. Program je napisan u python programskom jeziku, korišćenjem biblioteka pysam i pandas.

Dijagram toka analize je prikazan na slici 15.

4.1 Ciljevi analize

Ciljevi analize su da pronađu pozicije mesta integracije HIV sekvence u ljudskom genomu. Željeni format izlaza je tabela koja sadrži informacije o mestima integracije. Svaki red tabele opisuje jedno mesto integracije. Kolone koje jednoznačno i detaljno opisuju mesto integracije su:

- Hromozom
- Pozicija na hromozomu
- Vrsta integracije - da li je početak ili kraj ugradnje
- Orijentacija intergacije
- Pozicija ugradnje na HIV sekvenci
- Broj očitavanja koji podrzavaju mesto ugradnje



Slika 15: *Skica analize*

4.2 Priprema podataka

Priprema podataka je prvi korak u pronalaženju mesta integracije HIV virusa. Rezultat ovog koraka je BAM datoteka koja se dobija obradom ulaznih očitavanja datih u FASTQ formatu. Priprema podataka se zasniva na korišćenju algoritma za poravnanje nukleotidnih sekvenci u odnosu na zadati genom.

4.2.1 Poravnanje očitavanja

Poravnanje sekvenci je algoritam koji se koristi za pronalazenje sličnosti između dve ili više sekvenci. U slučaju koji je opisan u ovom radu poravnanje sekvenci znači pronalaženje mesta na referentnom genomu odakle najverovatnije potiču ulazna očitavanja. Rezultat poravnanja je BAM datoteka koja za svaki očitavanje sadrži temeljne informacije o poravnanju na referentni genom sa najboljim kvalitetom ili informaciju da očitavanje nije nigde poravnato na datom genomu zbog nedovoljne sličnosti. Jedan od najpoznatijih programa koji se koristi za pronalaženje poravnanja je BWA-MEM [14] koji je je korišćen u radu. Ulaz u BWA-MEM algoritam su referentni genom i sekvencirani uzorak. U ovom radu je referentni genom odabran na poseban način i to će biti opisano u narednom poglavlju.

4.2.2 Odabir genoma

Kod postupka poravnavanja očitavanja, osnovno pitanje je u odnosu na koji referentni genom raditi poravnanja. U slucaju kada je sekvenciran jedan uzorak znamo kog je porekla i u odnosu na koji genom treba poravnati očitavanja. Međutim, u situaciji kada je sekvenciran genom ljudske osobe koja je zaražena HIV virusom to znači da DNK te osobe sadrži nukleotide koji poticu i sa HIV sekvence i sa ljudskog genoma. Postavlja se pitanje da li treba poravnati očitavanja u odnosu na HIV ili na ljudski referentni genom. Ukoliko se poravnanje radi u odnosu na ljudski genom dobijaju se očitanja koja se odnose samo na njega, a informacija o očitavanjima koji su delom na ljudskom genomu a delom na HIV virusu bi bila izgubljena. Sa druge strane, ako bi očitavanja bila poravnata u odnosu na HIV sekvencu, bilo bi obrnuto, samo očitavanja koji se mapiraju na HIV bi ostala, a opet bi bila izgubljena ona očitavanja koja u sebi sadrže i deo referentnog genoma i deo HIV virusa. Ovakve slučajeve se obeležavaju kao podeljena očitanja. Upravo su takvi slučajevi relevantni za odrađivanje mesta integracije i oni treba da budu sačuvani kao ulazni podaci u drugi deo analize.

Rešenje ovakvog problema je da kao genom budu korišćene i HIV sekvenca i sve sekvence

ljudskog referentnog genoma, kako bi algoritam poravnanja mogao istovremeno da radi nad oba genoma. Stoga, algoritmu poravnanja će biti kao ulaz biti dat referentni genom u FASTA formatu datoteke koja pored svih hromozoma, kao jedan dodatni hromozom sadrži i sekvencu HIV virusa - kao dodatni deo ljuskog genoma, sto u slučaju osobe zaražene HIVom i jeste slučaj.

Sekvenca HIV virusa je preuzeta sa zvanične veb stranice NCBI, dodatne informacije dostupne su u [35]. Ljudski referentni genom je verzija hg37.

4.2.3 Podešavanje i pokretanje BWA-MEM algoritma

Za pripremu podataka, kao sto je već pomenuto, korišćen je BWA-MEM algoritam. U izobilju programa za poravnanje odabran je baš ovaj iz više razloga, prvenstveno zato sto nudi opcije koje su neophodne za sprovođenje dalje analize. Jedna od najbitnijih opcija u slučaju ugradnje HIV virusa je da BWA-MEM može da radi i podeljena poravnanja očitavanja (eng. *split read alignment*). To su očitavanja koji se različitim delovima poravnavaju na različita mesta u genomu, tako da svako drugo moguće jedinstveno poravnanje ima drastično manji kvalitet poravnanja. Ovo naknadno omogućava pronalaženje podeljeno poravnatih očitavanja jer oni potencijalno leže na mestima integracije. BWA-MEM algoritam uključuje parametar (-M) kojim se alat podešava tako da sačuva i takva poravnanja u BAM datoteci. Ovakva poravnanja se beleže u dva ili više redova, gde svaki red predstavlja po jedno moguće parcijalno poravnanje. Podeljena poravnanja mogu biti i na udaljenim mestima u genomu i obeleženi su zastavicom "SA", što je zastavica za dodatno poravnanje (eng. *supplementary alignment*). Ako se ne podesi parametar "-M" pri pokretanju BWA-MEM algoritma tada se sva podeljena poravnanja prebacuju u druga poravnanja (eng. *secondary alignments*) koja se zatim filtriraju.

Proizvod BWA-MEM programa je datoteka u SAM formatu. Ulaz u dalju analizu je BAM datoteka koja je sortirana i indeksirana, tako da je neophodno izvršiti konverziju SAM u BAM datoteku i njeno sortiranje i indeksiranje. Korišćeni alati za ovu konverziju su:

- Sambamba view
- Sambamba sort

Kako bi vreme izvršavanja programa bilo smanjeno, da bi se izbeglo pisanje rezultata na disk, preko komandne linije je direktno preusmeren izlaz iz BWA-MEM programa u

Sambamba View alat koji radi konverziju iz SAM formata u BAM format, a zatim je izlaz direktno preusmeren u Sambamba Sort alat koji sortira dobijenu BAM datoteku i automatski proizvodi indeks datoteku.

4.3 HIVSeeker algoritam

Algoritam HIVSeeker je razvijen sa idejom da bude javno dostupni alat koji pronalazi mesta integracije HIV virusa u ljudskom genomu iz NGS sekvenciranih podataka, i kao takav predstavlja centralnu temu ovog rada. Detalji implementacije algoritma su napisani u produžetku ovog poglavlja dok se više o performansama algoritma i samim rezultatima moze naći u narednom poglavlju. Kod je priložen kao dodatak uz elektronsku veziju rada. Algoritam je pisan u python programskom jeziku i podeljen je u nekoliko glavnih celina koje se moraju izvršavati sekvencijalno.

4.3.1 Zahtevi algoritma HIVSeeker

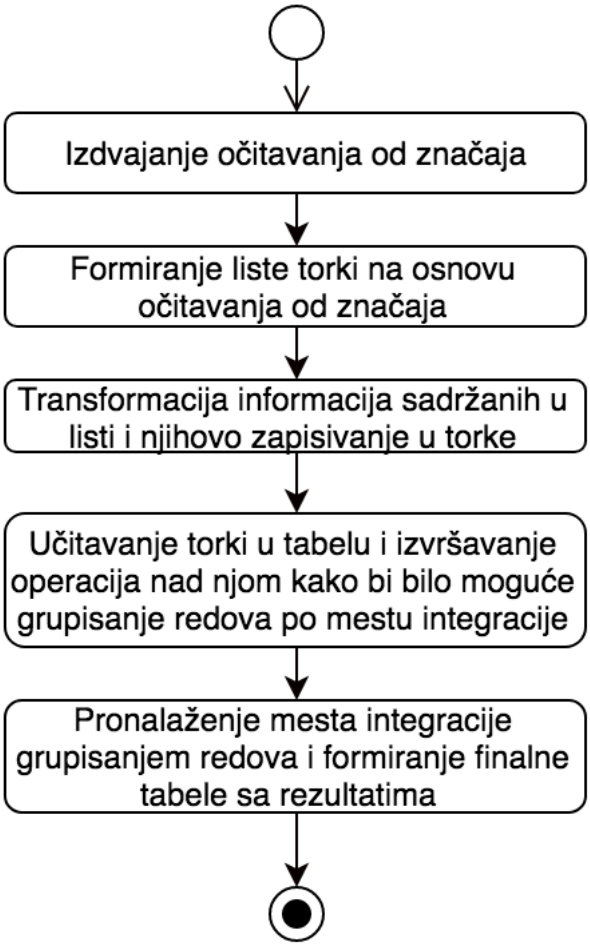
Zahtevi za pokretanje algoritma su:

- Ulazni podaci moraju biti u BAM formatu datoteke
- BAM datoteka mora biti dobijena korišćenjem BWA-MEM algoritma, ili nekog srodnog algoritma za poravnanje koji nudi opciju čuvanja dodatnih poravnanja, kao i upisivanje detalja o tome u već pomenutu "SA" zastavicu. Poželjan format ulaza sa kojim je algoritam najviše testiran se može dobiti praćenjem instrukcija u prethodnom poglavlju u kome je detaljno opisana priprema podataka.
- Neophodne instalirane biblioteke su pandas i pysam, kao i programski jezik python, verzija 2.7 ili kasnije.

4.3.2 Implementacija algoritma HIVSeeker

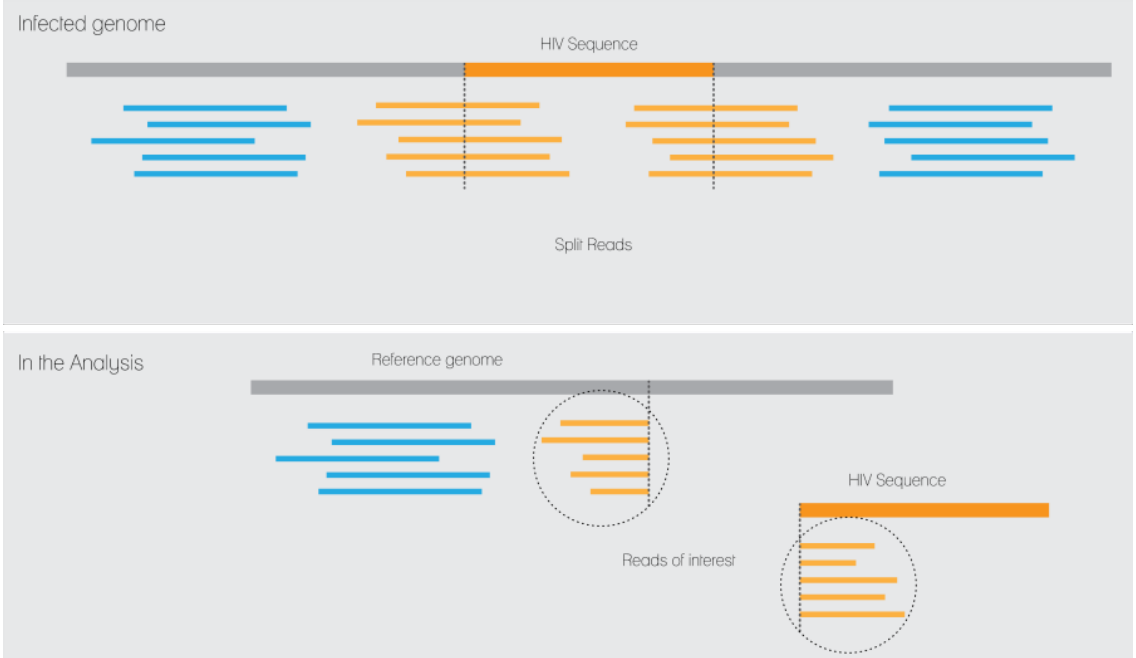
Postupak pronalazenja mesta integracije ima nekoliko faza. Algoritam počinje izdvajanjem očitavanja koja imaju podeljena poravnanja, sa uslovom da je bar jedno podeljeno poravnanje sa HIV sekvence, a drugi deo je negde na ljudskom genomu. Dijagram aktivnosti je prikazan na slici 16.

Na slici 17 su prikazana moguća očitavanja, dok su narandžastom bojom obeležena ona koja će biti korišćena u pronalaženju mesta integracije. Plava očitavanja se mapiraju samo na ljudski genom, dok su narandžasta delom na HIV sekvenci delom na ljudskom genomu.



Slika 16: *HIVSeeker algoritam*

Na prvoj slici se vidi kako u inficiranom genomu izgledaju poravnata očitavanja, dok je na drugoj slici prikazano kako ta očitavanja izgledaju u procesu analize i kako mogu biti prepoznata.



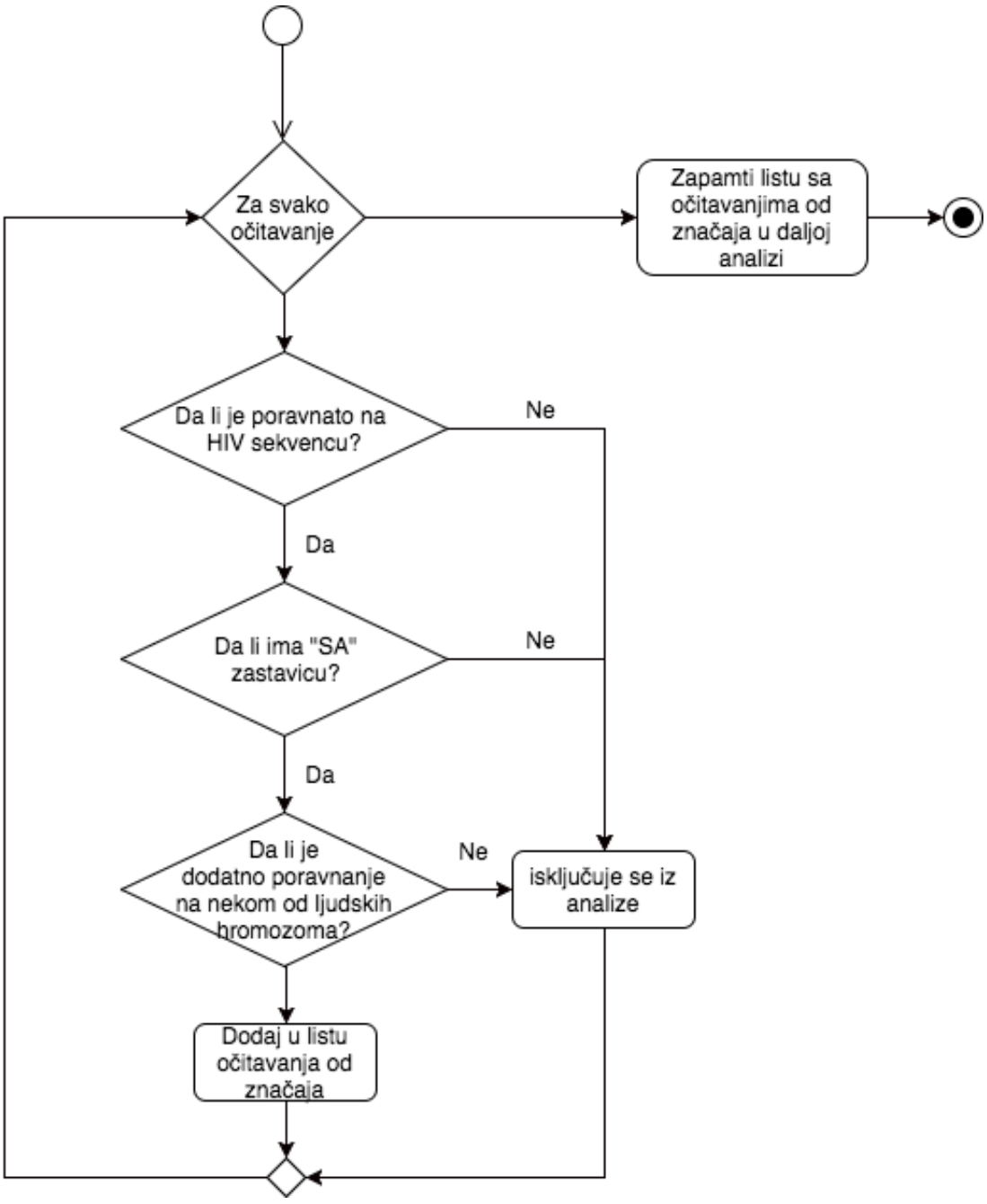
Slika 17: *Opisna slika podeljenih očitavanja kod osobe u koju se HIV integrisao u stvarnosti i kako se to vidi u toku analize. Zaokružena su očitavanja od interesta pri pronalaženju mesta integracije.*

Pronalaženje očitavanja od značaja se sprovodi sledećim algoritmom:

1. Prolazi se kroz sva očitavanja u BAM datoteci i filtriraju se očitavanja od znacaja. Pri postupku filtracije proverava se:

   - Da li je očitavanje poravnato na HIV sekvencu. Ako nije, ne koristi se u daljoj analizi.
   - Da li očitavanje ima postavljenu "SA" zastavicu, što, kao sto je već navedeno u tekstu, znači da ima dodatna poravnanja. Ukoliko postoje dodatna poravnanja, analizira se zastavica SA i proverava se koja su dodatna poravnanja. Ako dodatna poravnanja ne postoje očitavanje se preskace u daljoj analizi.
   - Da li je dodatno poravnanje takođe poravnato na HIV sekvencu. Ukoliko jeste, to očitavanje se preskače zato što ne može pomoći pri pronalaženju mesta integracije u ljudski genom. Ukoliko je dodatno poravnanje na nekom od ljudskih hromozoma zaključujemo da je relevantno za dalju analizu.

Dijagram aktivnosti procesa filtriranja očitavanja je dat na slici 18.



Slika 18: *Postupak filtriranja očitavanja*

2. Sva izdvojena očitavanja se analiziraju i formira se lista torki sa informacijama potencijalno relevantinim za mesto integracije koje će biti korišćene u daljoj analizi:

   - Za svako očitavanje se formira torka sa sledećim poljima:
     - Jedinstveni identifikator očitavanja (*Id očitavanja*)

- Jedinstveni identifikator hromozoma na referenci na kome se nalazi dodatno poravnanje (*Id hromozoma*)
- Tačna pozicija na hromozomu od koje počinje dodatno poravnanje (*Hromozom pozicija*)
- Orijentacija poravnanja podeljenih ridova
- Opis poravnanja za dodatni deo očitavanja na referentno genomu (*CIGAR niska*)
- Pozicija početka poravnanja na HIVu
- Pozicija kraja poravnanja na HIVu
- Opis poravnanja na HIV sekvenci (*CIGAR niska*)
- Hromozom na koji se poravnava upareno očitavanje (eng. *Paired end read*)
- Tačna pozicija na hromozomu od koje počinje poravnanje uparenog očitavanja

• Orijentacija poravnanja podeljenih očitavanja određuje se pomoću informacije iz opisa poravnanja, koja govori o tome na koji se DNK lanac poravnalo očitavanje. Ukoliko oba podeljena poravnanja dolaze sa istog lanca to znači da se HIV virus ugradio u istom smeru kao sto je i referentni genom predstavljen, odnosno u 5' -> 3' smeru i to se obeležava kao pozitivnu orijentaciju ("+"). Inače se HIV virus ugradio u suprotnom smeru, dakle suprotne je orijentacije i onda se to obeležava negativnom orijentacijom ("-"). Ova informacija je značajna pri grupisanju očitavanja pri određivanju mesta integracije. Samo očitavanja sa istom orijentacijom se broje kao podrška jednog mesta integracije.

3. Kada su formirane torke opisane pod 2., svakoj torci se pridružuje informacija o poziciji završetka poravnanja na ljudskom hromozomu kako bi svi potrebni podaci za pronalaženje mesta integracije bili grupisani. Za pronalaženje tih informacija potrebno je sprovesti sledeći algoritam:

   • Ponovo se prolazi kroz sva očitavanja u BAM datoteci i pronalaze se sva koja imaju SA zastavicu ali da ovaj put nisu poravnati na HIV sekvencu.

   • Formira se rečnik od odabranih očitavanja:

     – Ključ je id očitavanja na koji je nadovezana pozicija početka poravnanja na referenci

  - Vrednost je torka (eng. *tuple*) (pozicija kraju poravnanja na referenci, torka celog očitavanja)

- Prolazi se kroz listu torki napravljenih u koraku 2. i za svaku se proverava da li se nalazi u napravljenom rečniku. Ukoliko se nalazi, torka se proširuje informacijom o poziciji kraja poravnanja na referentnom genomu.
- Svaka torka sadrži sedeće informacije, a u zagradama se nalaze pridružena imena polja iz koda:
  - Jedinstveni identifikator očitavanja (*ReadName*)
  - Jedinstveni identifikator hromozoma na referenci na kom se nalazi dodatno poravnanje (*RefName*)
  - Tačna pozicija na hromozomu od koje počinje dodatno poravnanje (*RefStart*)
  - Orijentacija poravnanja podeljenih ridova (*Ori*)
  - Opis poravnanja za dodatni deo očitavanja na referentno genomu (*RefCigar*)
  - Pozicija početka poravnanja na HIVu (*HivStart*)
  - Pozicija kraja poravnanja na HIVu (*HivEnd*)
  - Opis poravnanja na HIV sekvenci (*HivCigar*)
  - Hromozom na koji se poravnava upareni rid (*MateRefName*)
  - Tačna pozicija na hromozomu od koje počinje poravnanje uparenog očitavanja (*MateRefPos*)
  - Tačna pozicija na hromozomu na kojoj se završava poravnanje (*RefEnd*)

4. Sada kada su izdvojene sve potrebne informacije iz očitavanja od značaja, prelazi se na analizu torki i pronalaženje mesta integracije. Koristiće se python biblioteka pandas za analizu torki. Sprovodi se sledeći algoritam:
   - Učitavaju se torke u pandas tabelu (eng. *pandas dataframe*)
   - Za svaki red u pandas tabeli transformiše se kolona MateRefName kako bi u njoj bilo zapisano ime hromozoma na koji se poravnava a ne identifikator hromozoma
   - Cilj algoritma je da pronađe tačna mesta integracije na hromozomu RefName, kao i tip integracije, odnosno da li se mesto odnosi na početak ili kraj integri-

sane HIV sekvence. Pored toga, cilj algoritma je i da pronađe mesto početka poravnanja na HIV sekvenci. Kako bi ovakve informacije bile pronađene, potrebno je da pored svih informacija koje torke već sadrže izvučemo dodatne podatke. Parsiranjem CIGAR niski (str. 14) se dobijaju detaljnije informacije o tome koju poziciju integracije red u tabeli podržava. Parsiranje se izvodi na sledeći način:

- U zavisnosti od CIGAR niske na referenci (RefCigar), koja može biti oblika xSyM ili xMyS, možemo zaključiti da li očitavanje podržava mesto kraja integracije HIV virusa u prvoj situaciji (xSyM) a početak ukoliko je drugi tip CIGAR niske (xMyS)
- U zavisnosti od toga se uzima pozicija početka poravnanja za mesto integracije, ili pozicija kraja poravnanja za mesto integracije HIV sekvence
- U zavisnosti od CIGAR niske na HIV sekvenci (HivCigar), analogno referenci, moze se iznačunati pozicija na HIV sekvenci
- Dobijene informacije se zapisuju za svaki red u tabeli dodavanjem tri kolone:
  - Pozicija mesta integracije (*InsertionSite*)
  - Tip integracije (*Kind*)
  - Pozicija integracije na HIV sekvenci, tj pozicija od koje se HIV sekvenca ugradila (*BreakPoint*)
- Brišu se duplikati iz tabele po ključu: (ReadName, RefName, InsertionSite, Kind, Ori, BreakPoint)
- Svakom redu se dodaje kolona u kojoj je zapisan broj redova koji podrzava isto mesto ugradnje HIV virusa. Na početku je inicijalizovano na 1, pošto su inicijalno sva očitavanja pojedinačni redovi u tabeli. (*num*)

5. Kada je formirana tabela sa svim navedenim informacijama, prelazi se na grupisanje redova koja podržavaju isto mesto integracije.

   - Grupišu se redovi na osnovu vrednosti u sledećim kolonama: RefName, InsertionSite, Kind, Ori, BreakPoint. Cilj je pronalaženje mesta integracije na referentnom genomu, mesto odakle se desila ugradnja na HIV sekvenci, orijentacija očitavanja koji podržavaju mesto integracije i da vrsta integracije bude

ista, jer su očitavanja sa jednakim vrednostima ovih polja podrška za jedinstveno mesto ugradnje.

- Prebrojava se koliko redova ima jednake vrednostima ovih svih 5 kolona i taj broj se upisuje u kolonu "num"
- Odbacuju se sve kolone koje više nisu od značaja: RefStart, RefEnd, HivStart, HivEnd, MateRefName.
- Dobijena tabela se sortira po koloni "num"
- Eliminišu se sva mesta integracije gde podrška nije zadovoljavajća. Minimalni broj očitavanja je moguće menjati.

Rezultat algoritma je tabela koja sadrži informacije o mestima integracije. Ne mora svako mesto integracije biti pronađeno sa obe strane, već je moguće da samo jedna od dve pozicije budu uključene.

5 Provera rada HIVSeeker algoritma i rezultati

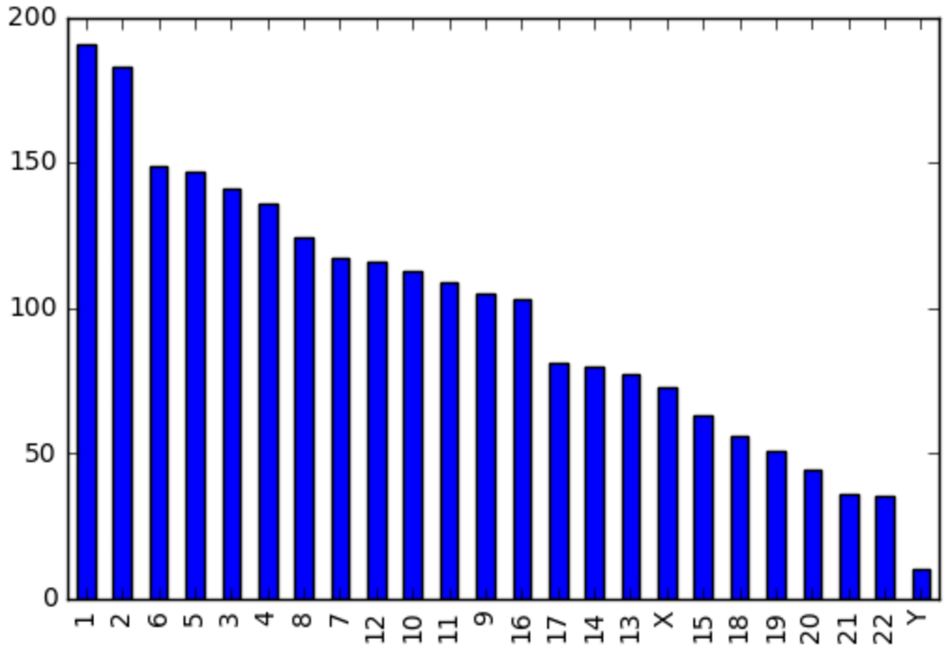
Za proveru rada programa HIVSeeker je korišćen javno dostupan uzorak CD4+ T ćelija inficiranih HIV-om, gde je materijal za sekvenciranje pripremljen na poseban način tako da detekcija lokacija integracija iz NGS sekvenciranih podataka bude moguća. Međutim, za ovaj realan uzorak ne postoje unapred definisana mesta integracije HIV virusa, tako da je jedino moguće koristiti ga za demonstraciju funkcionalnosti algoritma, ali nije moguće koristiti ga i za proveru ispravnosti rezultata. Upravo zbog toga je generisan i sintetički uzorak sa unapred poznatim mestima itegracije kako bi se proverila i ispravnost algoritma. U ovom poglavlju su detaljno opisani dobijeni rezultati pokretanja HIVSeeker algoritma nad oba uzorka. Rezultat analize je tabela u kojoj su upisane informacije o mestima integracije.

5.1 Provera funkcionalnosti programa HIVSeeker nad realnim uzorkom

Provera funkcionalnosti programa je, kao što je već pomenuto, rađena nad javno dostupnim podacima iz rada [34]. Korišćen uzorak ima oznaku "SRR1560523 ", dok je veličina ulaznih podataka u kompresovanom FASTQ formatu u zbiru približno 500 megabajta. Veličina rezultujuće BAM datoteke iz pripreme podataka je 430 megabajta, što odgovara očekivanjima s obzirom na veličinu ulaznih podataka. Što se tiče performansi izvršavanja celokupne analize, korak priprema podataka nad ovim podacima je trajalo 41 minut na mašini sa 8 procesora i 15 gigabajta RAM memorije. Nakon toga, izvršavanje HIVSeeker programa je trajala približno 15 minuta. To pokazuje da je program upotrebljiv u realnom vremenu nad realnim podacima.

5.1.1 Rezultati nad realnim uzorkom

Broj mesta integracije je promenljiv u zavisnosti od podešenih parametara pri pokretanju analize. Ukoliko je poželjno da analiza ima veću osetljivost (eng. *sensitivity*) onda je dobra praksa smanjiti parametar minimalne podrške očitavanja na mestima integracije pri pokretanju alata HIVSeeker. Radi bolje interpretacije, rezultati su uglavnom predstavljeni pomoću grafika dok je tabela sa svim rezultatima dostavljena kao dodatak. Kada je granica minimalne podrške podešena na vrednost 30, ukupan broj mesta integracije je 2663. Na slici 19 je prikazan broj mesta integracije po hromozomu, pri čemu se na Y osi nalazi broj integracija, dok su na X osi prikazani hromozomi.



Slika 19: *Broj mesta integracije po hromozomu, sa minimalnom granicom podrške podešenom na vrednost 30. Y osa na grafiku označava broj mesta integracije, dok se na X osi nalaze hromozomi*

Pored broja mesta ugradnje po hromozomu, slika 20 predstavlja distribuciju ovih mesta integracije po hromozomima. Kao što se vidi na slici, hromozomi nisu iste dužine i cela slika je skalirana prema dužini najdužeg hromozomu 1. Svaka uspravna crta predstavlja mesto integracije. Beli delovi predstavljaju odsustvo dokaza da se ugradnja tu dogodila. Regioni sa gušćim linijama se mogu smatrati vrućim tačkama integracije i mestima privlačenja HIV virusa na ljudski genom.

Sa druge strane, ukoliko je od velikog značaja preciznost (eng. *precision*), onda bi trebalo podići vrednost parametra minimalne podrške. Nije moguće diskutovati o tome kolika je minimalna vrednost podrške potrebna jer to zavisi isključivo od metode sekvenciranja uzorka i prosečne dubine pokrivenosti pozicija u genomu. Pošto je za parametar minimalne podrške podešen na vrednost 30 dobijeno 2663 mesta integracije, što je veliki broj rezultata, a prosečna pokrivenost genoma u ovom uzorku je daleko veća od 30 minimanlih očitavanja, moguće je podići donju granicu minimalne pokrivenosti mesta ugradnje i posmatrati kakve promene u graficima će nastati.

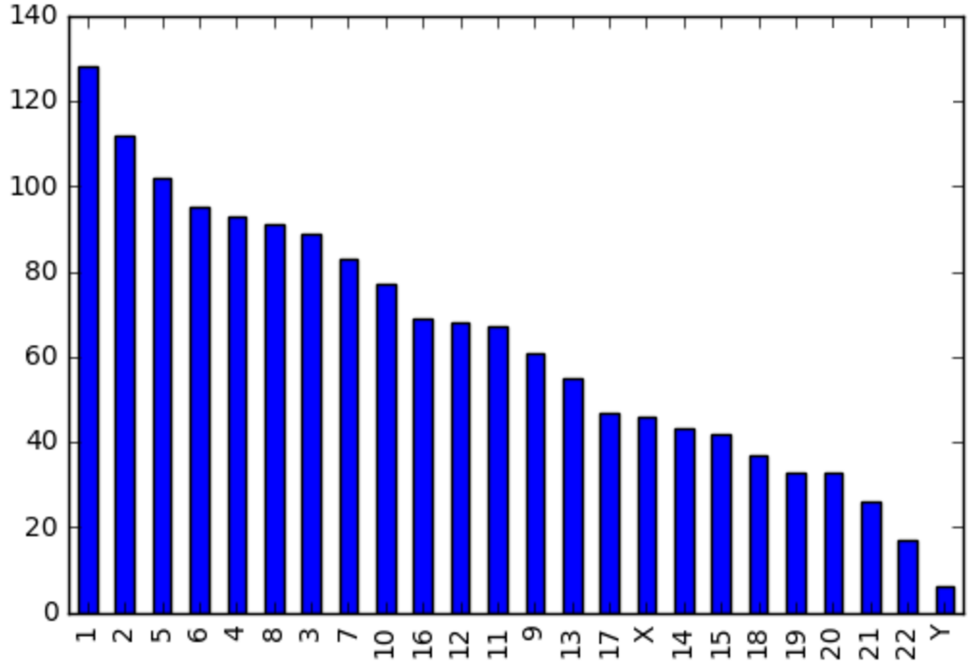
Kada se podigne vrednost minimalne podrške na 50 broj pronađenih mesta ugradnje je 1544. Na slici 21 je prikazan broj mesta integracije po hromozomu, sa granicom podrške podešenom na vrednost 50. Kada se uporedi broj mesta integracije sa različito podešenim

Slika 20: *Distribuciju mesta integracije po hromozomima, sa minimalnom granicom podrške podešenom na vrednost 30*

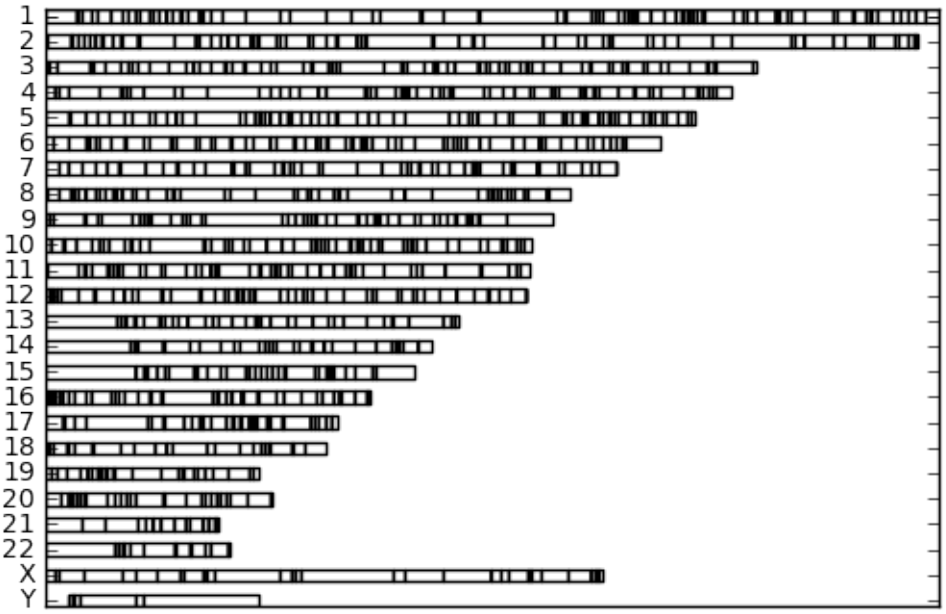
parametrom minimalne podrške, na slici se može videti da je odnos broja integracija po različitim hromozomima promenjen, ali neznatno s obzirom na međusobno malu razliku u broju integracija, tako da se može reći da se broj integracija po hromozomima ravnomerno smanjio, što je očekivano ponašanje upotrebljenog filtera.

Na slici 22 je predstavljena distribucija mesta integracije po hromozomima sa minimalnom podrškom očitavanja podešenom na vrednost 50. Kada se uporede slike 20 i 22 primećuje se kako se gustina mesta integracije ravnomerno smanjila što je takođe bilo očekivano ponašanje.

Ukoliko je preciznost od izuzetno velikog značaja, implementiran je jos jedan filter. Ovaj filter je nadgradnja algoritma kako bi se podigla sigurnost u mesta integracije. Ideja je da se zadrže samo ona očitavanja koja su poravnata ili na hromozom ili na HIV sekvencu najmanje sa unapred predefinisanim brojem baza iz sekvence očitavanja. Parametar koji ovo govori je minmalna dužina poravnanja. Na ovaj način će biti odbačena sva mesta integracije koja nisu podržana ovako opisanim poravnatim očitavanjima. Kada se takav metod filtriranja primeni a minimalna dužina poravnanja se podesi na 40, broj preostalih mesta integracije je 57. Odabrana je vrednost 40 nakon niza testova sa različitim vrednostima, gde se 40 pokazala kao optimalna. Pouzdanost u ova mesta integracije je znatno veća, a distribucija po hromozomima je prikazana na slici 23, dok je broj integracija po



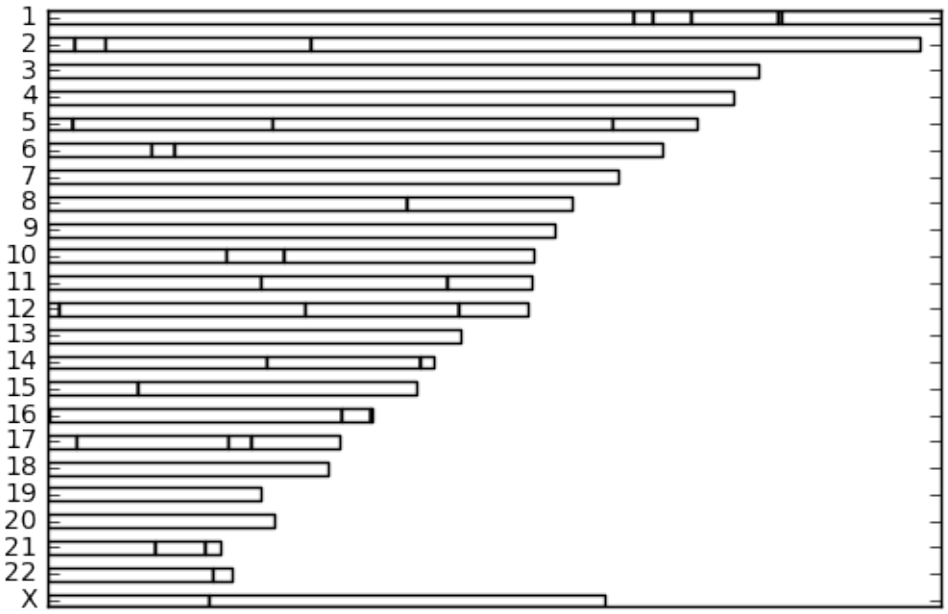
Slika 21: *Broj mesta integracije po hromozomu, sa minimalnom granicom podrške podešenom na vrednost 50. Y osa na grafiku označava broj mesta integracije, dok se na X osi nalaze hromozomi*



Slika 22: *Distribuciju mesta integracije po hromozomima, sa minimalnom granicom podrške podešenom na vrednost 50*

hromozomima prikazana na slici 24.

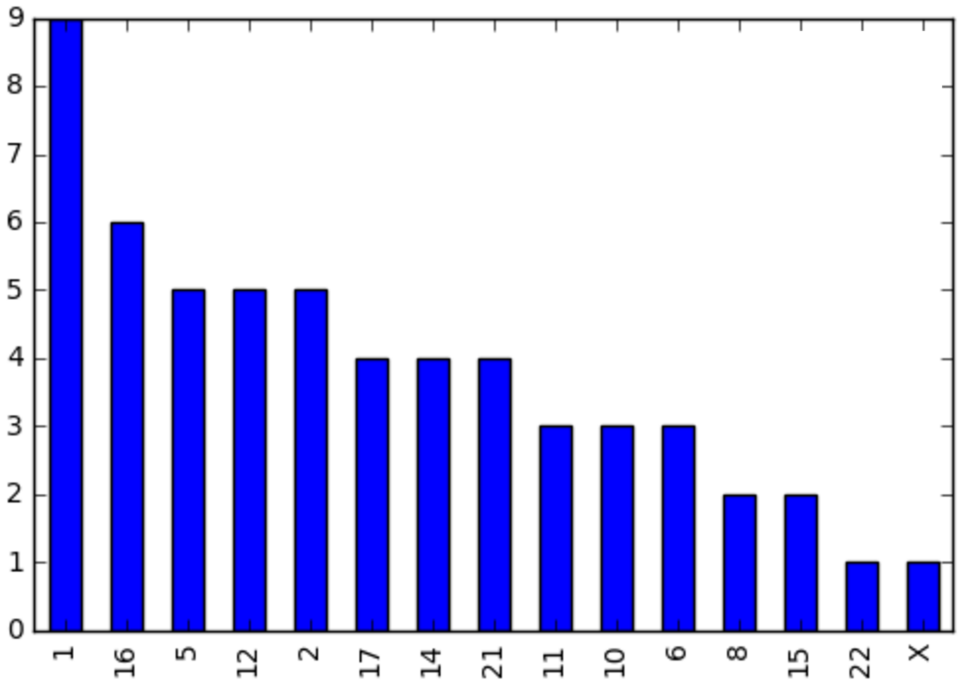
Pored tog filtera, omogućen je i filter očitavanja na osnovu broja nepoklapajućih baza u jednom očitavanju koje podržava mesto integracije. Nad finalnim rezultatima je pokrenut filter maksimalnog broja nepoklapajućih baza podešenim na 3. Ovaj filter nije promenio krajnje rezultate, osim malo u broju očitavanja koji podržavaju mesto integracije. Pored toga, i ovaj filter je omogućen pri pokretanju programa, zato što bi na drugim uzorcima možda imao veći značaj, jer ovaj filter pomaže kada je uzorak sekvenciran sa niskim kvalitetom.



Slika 23: *Distribuciju mesta integracije po hromozomima, sa minimalnom granicom podrške podešenom na vrednost 30 i sa dodatnim filterom minimalne dužine poravnanja podešenom na vrednost 40*

Bitno je naglasiti da zbog prirode sekvenciranja nije moguće očekivati da će biti uhvaćena oba mesta integracije jedne sekvence HIV virusa, odnosno ne garantuje se da će metodom sekvenciranja biti sekvencirana i levu i desnu pozicija ugradnje. Zbog toga mesta integracije nisu uparena, tako da se neka mesta integracije pojavljuju na dva mesta u tabeli. Još jedna bitna informacija je da su levi i desni deo HIV sekvence takozvani *LTR* (str. 22) delovi genoma i da je redosled baza u tim delovima veoma sličan do identičan.

Takođe je bitno pomenuti da je radi provere ispravnosti metode urađena provera sličnosti između ljudskog referentnog genoma i genoma HIV virusa. Jedan od najpoznatijih algoritama za promalaženje sličnosti između genomskih sekvenci, BLAST program [36], je pokazao da ne postoje poklapanja.

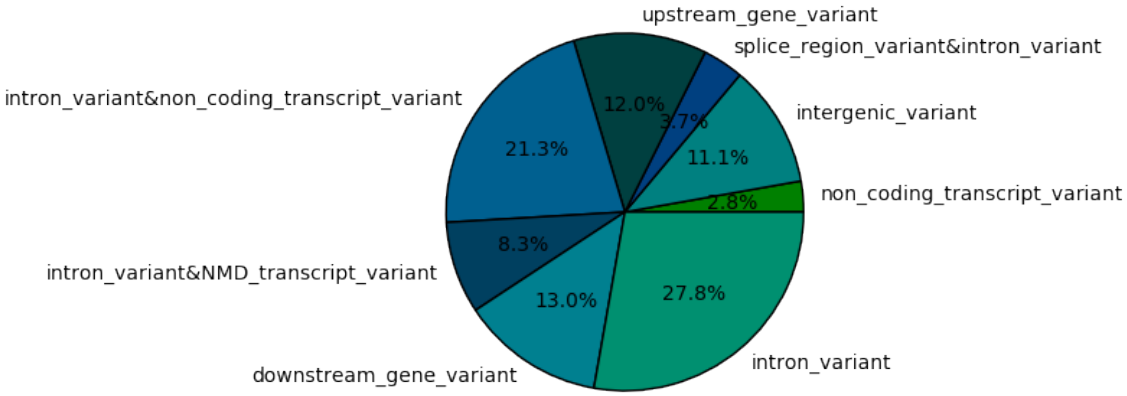


Slika 24: *Broj mesta integracije po hromozomu, sa minimalnom granicom podrške podešenom na vrednost 30 ali sa primenjenim filterom minimalne dužine poravnanja podešenom na vrednost 40. Y osa na grafiku označava broj mesta integracije, dok se na X osi nalaze hromozomi*

5.1.2 Analiza dobijenih mesta integracije

Da bi bolje razumeli pronađena mesta integracije, potrebno je proveriti funkciju i značaj tih pozicija na ljudskom genomu. Pošto je baza gena poznata, sa precizno datim koordinatama početka gena i kraja gena, moguće je anotirati postojeće rezultate. Anotacija je izvršena nad krajnjim rezultatom mesta integracije (ukupno 57), sa minimalnom granicom podrške podešenom na vrednost 30 i sa primenjenim filterom minimalne dužine poravnanja podešenom na vrednost 40. Na slici 25 je prikazan pita dijagram koji opisuje prirodu pronađenih pozicija ugradnje u odnosu na poznate gene. Nomenklatura za lokaciju mesta integracije je preuzeta od VEP-a (eng. *Variant Effect Predictor*), poznatog alata za anotaciju [37]. Bitno je naglasiti i da jedno mesto na genomu može pripadati više različitih kategorija. Upravo zbog toga, uz pita dijagram, prikazana je i tabela sa tačnim brojem mesta integracija koja pripadaju jednoj kategoriji prirode lokacije na genomu.

Pored 12 mesta integracije koji su pronađeni u integrenskim regijama pa samim tim nisu povezani na sa jednim genom, geni sa kojima su povezana ostala mesta integracije, kao i broj odgovarajućih mesta integracije su:



Slika 25: *Pita dijagram koji opisuje raspodelu prirode mesta integracija sa minimalnom granicom podrške podešenom na vrednost 30 ali sa primenjenim filterom minimalne dužine poravnanja podešenom na vrednost 40*

Priroda pozicija	Broj mesta integracije
downstream_gene_variant	14
intergenic_variant	12
intron_variant	30
intron_variant&NMD_transcript_variant	9
intron_variant & non_coding_transcript_variant	23
non_coding_transcript_variant	3
splice_region_variant&intron_variant	4
upstream_gene_variant	13

Tabela 1: *Tabela sa brojem mesta integracija koja odgovaraju određenoj kategoriji prirode lokacije na genomu*

- AC008271.1: 1
- AF131217.1: 2
- ALOXE3: 1
- AP000442.4: 1
- AXDND1: 2
- CA10: 1
- CMIP: 2
- CTD-2044J15.1: 2
- FANCA: 2
- HES7: 1
- KDM6A: 1
- MAPK8: 1
- MARK3: 1
- MRPL28: 2
- NPHS2: 2
- POU2AF1: 2
- PPARD: 2
- RBM19: 2
- RNU6-487P: 1
- RNU7-58P: 1
- RP11-1042B17.3: 1
- RP11-112H10.4: 2
- RP11-122F24.1: 1
- RP11-170M17.1: 2
- RP11-545A16.1: 2
- RP11-736N17.8: 2
- SEPT4: 2
- SFXN5: 2
- SLC37A1: 2
- SNRPN: 2
- SNURF: 2
- TMEM8A: 2
- TSPAN8: 1
- TSPAN9: 2
- VPS13B: 2

To što je većina pronađenih mesta integracije u blizini, ili se nalazi unutar jednog ili više gena je u saglasnosti sa [38], gde tvrde da se HIV virus češće ugrađuje u regione bogate genima, zato što to omogućava efikasniju ekspresiju genoma virusa.

U nastavku je prikazana tabela sa svih 57 anotiranih rezultata, soritiranih po koloni "Podrška", gde je zapisan broj očitavanja koji podržava to mesto integracije.

Hromozom	Pozicija	Vrsta	Orijentacija	HivPozicija	Podrška	Priroda pozicije
8	100290201	End	-	180	102	VPS13B: intron_variant
8	100290201	End	-	62	75	VPS13B: intron_variant
12	114385456	End	+	90	93	RBM19: intron_variant
12	114385456	End	+	9175	83	RBM19: intron_variant
17	56611449	End	+	9175	77	RP11-112H10.4: intron_variant & non_coding_transcript_variant, SEPT4: upstream_gene_variant
17	56611449	End	+	90	73	RP11-112H10.4: intron_variant & non_coding_transcript_variant, SEPT4: upstream_gene_variant

nastavak na sledećoj strani

Hromozom	Pozicija	Vrsta	Orijentacija	HivPozicija	Podrška	Priroda pozicije
6	35314360	End	+	90	76	PPARD: intron_variant
6	35314360	End	+	9175	75	PPARD: intron_variant
1	204475727	Start	-	199	72	intergenic_variant
21	43948434	Start	-	9176	72	SLC37A1: downstream_gene_variant
21	43948434	Start	-	91	57	SLC37A1: downstream_gene_variant
6	28918878	Start	-	199	69	intergenic_variant
1	204476230	Start	+	199	63	intergenic_variant
16	81596855	End	+	9175	63	CMIP: intron_variant
16	81596855	End	+	90	56	CMIP: intron_variant
17	8022545	Start	-	199	61	ALOXE3: upstream_gene_variant, HES7: downstream_gene_variant
12	3214695	End	+	90	60	TSPAN9: intron_variant & NMD_transcript_variant
12	3214695	End	+	9175	58	TSPAN9: intron_variant & NMD_transcript_variant
11	59327880	Start	+	199	54	RNU7-58P: downstream_gene_variant, AP000442.4: downstream_gene_variant
X	44797648	Start	+	62	52	KDM6A: intron_variant
21	29846328	Start	-	9175	50	AF131217.1: intron_variant & non_coding_transcript_variant
21	29846328	Start	-	90	45	AF131217.1: intron_variant & non_coding_transcript_variant
15	25212010	Start	-	90	49	SNRPN:intron_variant, SNURF: intron_variant & NMD_transcript_variant
15	25212010	Start	-	9175	43	SNRPN: intron_variant, SNURF: intron_variant & NMD_transcript_variant
14	103556820	Start	-	9175	47	RP11-736N17.8: upstream_gene_variant
14	103556820	Start	-	90	42	RP11-736N17.8: upstream_gene_variant
2	73193296	End	+	9177	47	SFXN5: intron_variant & non_coding_transcript_variant

nastavak na sledećoj strani

Hromozom	Pozicija	Vrsta	Orijentacija	HivPozicija	Podrška	Priroda pozicije
2	73193296	End	+	92	44	SFXN5: intron_variant & non_coding_transcript_variant
5	6683516	End	+	9175	45	CTD-2044J15.1: downstream_gene_variant
5	6683516	End	+	90	40	CTD-2044J15.1: downstream_gene_variant
16	89849407	End	+	9175	44	FANCA: upstream_gene_variant
16	89849407	End	+	90	39	FANCA: upstream_gene_variant
14	60994730	End	+	187	44	RP11-1042B17.3: intron_variant & non_coding_transcript_variant
12	71555944	Start	+	194	43	TSPAN8: upstream_gene_variant
1	179516510	Start	-	90	42	NPHS2: downstream_gene_variant, RP11-545A16.1: intron_variant & non_coding_transcript_variant, AXDND1: intron_variant
1	179516510	Start	-	9175	42	NPHS2: downstream_gene_variant, RP11-545A16.1: intron_variant & non_coding_transcript_variant, AXDND1: intron_variant
1	203289949	Start	-	92	42	RNU6-487P: upstream_gene_variant
11	111273454	Start	-	9175	41	POU2AF1: intron_variant
11	111273454	Start	-	90	39	POU2AF1: intron_variant
2	7414794	End	+	90	41	intergenic_variant
2	7414794	End	+	9175	36	intergenic_variant
5	62647064	Start	-	62	38	intergenic_variant
10	65682587	Start	-	9175	35	RP11-170M17.1: intron_variant & non_coding_transcript_variant
10	65682587	Start	-	90	32	RP11-170M17.1: intron_variant & non_coding_transcript_variant
14	103907887	Start	-	199	34	MARK3: intron_variant & non_coding_transcript_variant
2	15871348	End	+	90	34	AC008271.1: intron_variant
22	46032368	Start	-	93	34	intergenic_variant
1	168583488	End	+	9175	34	intergenic_variant

nastavak na sledećoj strani

Hromozom	Pozicija	Vrsta	Orijentacija	HivPozicija	Podrška	Priroda pozicije
1	168583488	End	+	90	32	intergenic_variant
10	49603226	End	+	220	33	MAPK8: intron_variant & non_coding_transcript_variant
5	157448180	End	+	90	33	intergenic_variant
17	50095997	End	+	9177	32	CA10: intron_variant
16	419061	End	+	90	32	TMEM8A: downstream_gene_variant, MRPL28: non_coding_transcript_exon_variant & non_coding_transcript_variant
16	419061	End	+	9175	32	TMEM8A: downstream_gene_variant, MRPL28: non_coding_transcript_exon_variant & non_coding_transcript_variant
1	163384810	Start	-	9177	32	intergenic_variant
1	163384810	Start	-	92	31	intergenic_variant
5	7000445	End	+	90	30	RP11-122F24.1: intron_variant & non_coding_transcript_variant

Tabela 2: *Tabela koja sadrži detalje o pronađenim mestima integracije*

5.2 Provera ispravnosti HIVSeeker programa nad sintetički generisanim uzorkom

Usled nedostatka postojećih uzoraka sa potvrđenim mestima integracije, provera ispravnosti analize pronalaženja mesta integracije opisane u ovom radu, kao i algoritma HIVSeeker je uradjena korišćenjem generisanih sintetičkih uzoraka. Postupak pravljenja uzoka će biti detaljno opisan u ovom poglavlju, kao i rezultati dobijeni pri testiranju algoritma.

5.2.1 Postupak generisanja sintetičkog uzorka

Kako bi se izvršila provera rezultata, napravljen je sintetički uzorak koji simulira uzorak čoveka zaraženog HIV virusom. Najpribližniji sintetički generisan reprezentativni uzorak uzorka nad kojim je razvijan algoritam je sekvenciranje celog egzoma (eng. *Whole exome sequencing*). Za simulaciju je korišćen alat BEDTools GetFasta [39] kako bi od ljudskog

referentnog genoma bio napravljen ceo egzom. Nakon toga je ručno dodata sekvenca HIV virusa u nekoliko mesta u ljudskom genomu. Kako bi malo zakomplikovali testiranje, nije uvek dodavana cela sekvenca HIV virusa, već su i isecani delovi sekvence. Mesta na koja je ubacena sekvenca su:

- chr1 152278081 - Hiv sekvenca je dodata od početka do kraja, dužina 9181
- chr7 100638803 - Hiv sekvenca je dodata od pozicije 66 do pozicije 8680
- chr10 857527 - dodata je cela HIV sekvenca
- chr18 55148133 - dužina dodatog isečka je 1330, od pozicije 630 do 1960.

Nakon tako napravljene reference celog egzoma, simulacija sekvenciranja je rađena ART alatom [40]. Prosečna pokrivenost očitavanjima svake baze je približno 30.

5.2.2 Rezultati nad sintetičkim uzorkom

Nakon pokretanja algoritma HIVSeeker nad generisanim uzorkom, usledila je analiza dobijenih rezultata i ujedno i provera ispravnosti algoritma.

Tokom analize dobijeni broj očitavanja koji potencijalno pokrivaju mesta integracije je 257. Parametar donje granice broja očitavanja koji podržavaju mesto ugradnje je podešen na minimalnih 13 očitavanja. Ovakva granica je podešena u skladu poznavanja sintetičkih podataka i očekivane podrške referentnih baza.

Algoritam je uspeo uspešno da pronađe sva 4 mesta integracije na ljudskom genomu, po dva za svako mesto ugradnje. Što se tice ugrađivanja u sedmi, deseti i osamnaesti hromozom, mesta integracije su potpuno precizno pronađena, do na nekoliko baza zbog postojanja kratkih homologih sekvenci između mesta isecanja HIV virusa i mesta ugradnje u ljudskom genomu. Što se tiče ugradnje u prvom hromozomu, pozicija na ljudskom referentnom genomu je dobro određena, ali pozicija HIV sekvence nije uspešno otkrivena. Vrlo je verovatno da se ovakva situacija desila upravo zbog homologih baza na kraju i početku HIV sekvence virusa, i da su zbog toga početak i kraj ugradnje zamenjenio.

Izlaz iz algoritma je tabela sa potencijalnim mestima integracije. Svaki red u tabeli predstavlja jedno mesto integracije, dok su kolone koje ga opisuju:

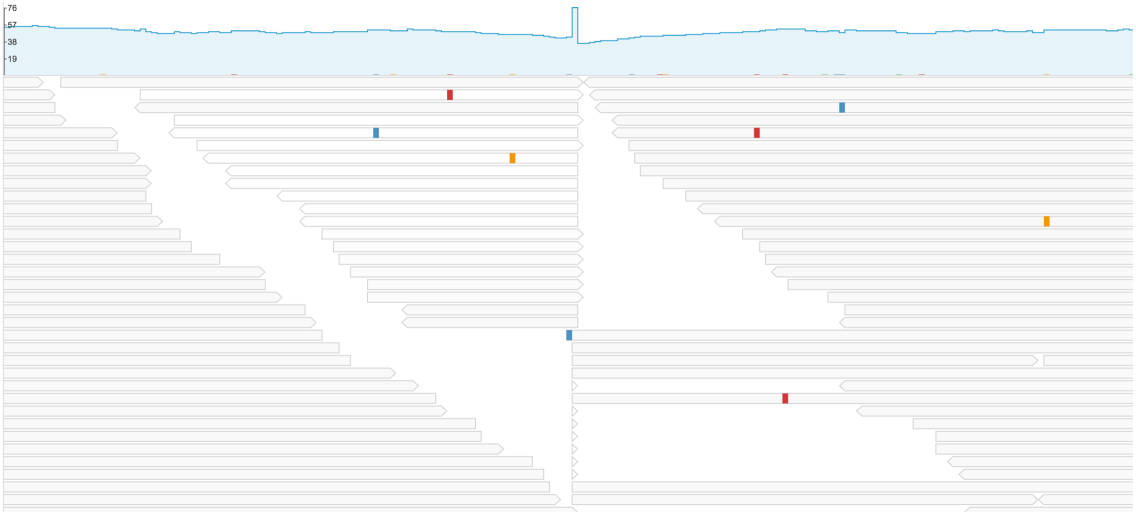
- Hromozom
- Pozicija na hromozomu

RefName	InsertionSite	Kind	Orientation	HivBreakPoint	NumberOfSupportingReads
1	152278081	Start	+	9083	22
1	152278081	End	+	96	14
10	857527	Start	+	0	20
10	857526	End	+	9181	15
18	55148132	End	+	1960	29
18	55148133	Start	+	630	24
7	100638806	Start	+	66	30
7	100638803	End	+	8750	26

Tabela 3: *Tabela sa rezultatioma nad sintetički generisanim uzorkom.*

- Vrsta integracije - da li je početak ili kraj ugradnje
- Orijentacija intergacije
- Pozicija ugradnje na HIV sekvenci
- Broj očitavanja koji podržavaju mesto ugradnje

Vizualni prikaz mesta integracije takođe potvrđuje mesta integracije. Na slici 26 se može videti kako vizualno izgleda mesto integracije na hromozomu 10. Horizontalne linije su očitavanja koja se poravnavaju na tom delu referentnog genoma, dok je na vrhu slike prikazan broj očitavanja koji podržavaju referentnu bazu. Mesto integracije je prikazano na sredini slike, gde se moze videti da su očitavanja naglo isečena na tom mestu na istoj poziciji. To su upravo očitavanja koja se prostiru preko mesta integracije, a deo koji je isečen se poravnava na hiv sekvencu. Alat koji je korišćen za vizualizaciju je alat za interaktivno gledanje genoma (eng. *Interactive genomics viewer*) [41] .



Slika 26: *Slikoviti prikaz mesta integracije*

6 Zaključak i dalji rad

Centralna tema ovog rada je progam HIVSeeker, koji služi za pronalaženje mesta integracije HIV virusa u ljudskom genomu. Algoritam na osnovu kog je program napisan, kao i sam postupak za pripremu podataka je detaljno bio opisan u prethodnim poglavljima.

Program je pokrenut nad dva različita uzorka, jedan zbog demonstriranja funkcionalnosti algoritma a drugi zbog provere ispravnosti rezultata:

- Za demonstraciju funkcionalnosti korišćen je javno dostupan uzorak CD4+ T ćelija inficiranih HIV-om, gde je materijal za sekvenciranje pripremljen na poseban način tako da detekcija lokacija integracija iz NGS sekvenciranih podataka bude moguća. Nedostatak korišćenja ovog uzorka je što za njega nisu unapred poznata mesta integracije pa se stoga ne može koristiti i za proveru ispravnosti

- Radi provere ispravnosti rezultata je generisan sintetički uzorak sa unapred poznatim mestima integracije. Nad tim podacima, korišćenjem HIVSeeker programa tačno su pronadjena su sva mesta integracije, pa je na taj način potvrđena ispravnost dobijenih rezultata

Rezultati dobijeni nad oba uzorka su detaljno opisani u prethodnom poglavlju.

Pomoću parametara HIVSeeker programa, moguće je kontolisati broj pronađenih mesta integracije, u zavisnosti od željenog ishoda. Ukoliko bi rezultati programa prolazili kroz neku dodatnu proveru ispravnosti pozicija integracije, npr. duboko sekvenciranje uzorka samo na pozicijama gde su pronađena mesta integracije, poželjno je da rezultata bude što više kako bi odziv pravih pronađenih mesta integracije bio što veći, a preciznost bi bila postignuta naknadnim filtriranjem. Ukoliko je izlaz iz HIVSeeker programa konačan rezultat analize bez naknadne filtracije, bolja praksa je podešavanja ovih parametara na više vrednosti kako bi preciznost dobijenih rezultata bila što veća, možda po cenu odziva.

Takođe je primećeno da su integracije zastupljene na skoro svim hromozomima ali ne sa istom učestalošću. U finalnim rezultatima se vidi da je prvi hromozom najugroženiji HIV integracijama. Iz tabele sa anotiranim mestima integracije se može videti da od 9 lokacija na prvom hromozomu 6 pripada intergenskim regionima dok se preostale 3 nalaze na genima. Ako se malo bolje pogleda, vidi se i da se ovih 9 mesta integracije može tumačiti kao 6, zato što su tri mesta integracije zapisana dva puta sa jedinom razlikom u poziciji integracije na hiv sekvenci. Tako uparena mesta integracije imaju jedine različite vrednosti 9175 i 90 kao poziciju početka ugradnje HIV virusa, i to je slučaj u 2 od 3 para,

dok treći par ima vrednosti 9177 i 92, koje su takođe jako blizu. Ovo se objašnjava upravo već pomenutom homologijom sa početka i kraja HIV sekvence genoma. Nakon grupisanja ovih mesta integracije, podrška za ta mesta bi bila jos veća, jer bi brojevi podržavajućih očitavanja bili sabrani.

Većina pronađenih mesta integracije se nalazi u intronima gena ili u regulatornim delovima gena, dok se mali procenat integracija nalazi u intergenskim regionima.

6.1 Dalji rad

HIVSeeker algoritam omogućava pronalaženje mesta integracije HIV virusa u ljudskom genomu iz sekvenciranih inficiranih genoma NGS tehnologijom. U nedostatku podataka sa utvrđenim mestima integracije, nije bila moguća provera ispravnosti nad realnim podacima, već samo nad sintetičkim. Predlozi za unapređenje programa i dalji rad na njemu su:

1. Automatsko anotiranje rezultata, pronalaženje značaja pozicije u genomu za sve dobijene rezultate može biti omogućena

2. Uparivanje početka i kraja integracije HIV sekvence ukoliko su pri sekvenciranju očitana oba mesta integracije

3. Dodavanje različitih funkcija i parametara za filtriranje očitavanja koja pružaju podršku mestima integracije, na primer filtriranje očitavanja po kvalitetu poravnanja, ili po srednjoj vrednosti kvaliteta očitanih baza

4. Uvođenje mogućnosti otkrivanja novih mesta integracije i provera pronađenih pomoću uparenih očitavanja. Trenutno se informacije o uparenim očitavanjima nalaze u tabeli torki, ključevi *MateRefName* i *MateRefPos* ali se ne koriste u daljoj analizi

5. Provera koji delovi HIV sekvence su ugrađeni i da li je ta ugradnja funkcionalna

6. Analiza sekvenci na referentnom genomu u koje se HIV integrisao i njihova klasterizacija

Literatura

[1] Zaglavlja FASTA formata: http://www.uniprot.org/help/fasta-headers

[2] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, Peter M. Rice; The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 2010; 38 (6): 1767-1771. doi: 10.1093/nar/gkp1137

[3] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. The Structure and Function of DNA. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26821/

[4] Pray, L. (2008) Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100

[5] Crick F. Central dogma of molecular biology. Nature. 1970;227:561–563

[6] Vodič kroz sekvenciranje: https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

[7] Jay E, Bambara R, Padmanabhan R, Wu R. DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping*. Nucleic Acids Research. 1974;1(3):331-353.

[8] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-5467.

[9] Objašnjenje zapisa kvaliteta u fastq formatu: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

[10] Initial sequencing and analysis of the human genome, International Human Genome Sequencing Consortium Eric S. Lander et al , Nature 409, 860-921 (15 February 2001), doi:10.1038/35057062; Received 7 December 2000; Accepted 9 January 2001

[11] The Sequence of the Human Genome, J. Craig Venter et al, Science, 16 Feb 2001, Vol. 291, Issue 5507, pp. 1304-1351, DOI: 10.1126/science.1058040

[12] E. W. Myers et al., Science 287, 2196 (2000).

[13] Dokumentacija o SAM/BAM formatu datoteke: https://samtools.github.io/hts-specs/SAMv1.pdf

[14] Heng Li: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

[15] Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. A Brief Chronicle of Retrovirology. Available from: https://www.ncbi.nlm.nih.gov/books/NBK19403/

[16] Gallo RC. The discovery of the first human retrovirus: HTLV-1 and HTLV-2. Retrovirology. 2005;2:17. doi:10.1186/1742-4690-2-17.

[17] Vahlne A. A historical reflection on the discovery of human retroviruses. Retrovirology. 2009;6:40. doi:10.1186/1742-4690-6-40.

[18] Fields Virology, David M. Knipe and Peter Howley, ISBN: 9781451105636, Publication Month: June 2013, Edition: 6, 2-volume set

[19] Veb stranica: https://www.accesscontinuingeducation.com/ACE4010/c1/

[20] Mesri EA, Feitelson M, Munger K. HUMAN VIRAL ONCOGENESIS: A CANCER HALLMARKS ANALYSIS. Cell host & microbe. 2014;15(3):266-282. doi:10.1016/j.chom.2014.02.011.

[21] Nicholas H. Acheson: Fundamentals of Molecular Virology, 2nd Edition. ISBN : 978-1-118-21489-3

[22] Peng K., Radivojac P., Vucetic S., Dunker A.K., Obradovic Z.: Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics, 7:208, 1-17. (2006)

[23] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK.: DisProt: the Database of Disordered Proteins. Nucleic Acids Res, 35(Database issue):D786-793. (2007)

[24] Lobanov, M., Garbuzynskiy S., Galzitskaya O.: Statistical Analysis of Unstructured Amino Acid Residues in Protein Structures. Biokhimiya, Vol. 75, 236-246 (2009)

[25] Kyte, J. and Doolittle, R.: A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157 (1) 105-132 (1982)

[26] Campen A., Williams R., Brown C., Meng J., Uversky V., Dunker A.K.: TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. Protein and Pept Lett, 15(9):956-963. (2008)

[27] Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.: AAIndex: amino acid index database, progress report, Nucleic Acids Research, Volume: 36, Database issue, 202-205 (2008)

[28] Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure, Ya-Chi Ho et al, Cell

[29] HIV latency. Siliciano, R.F. and Greene, W.C. Cold Spring Harbor perspectives in medicine. 2011; 1: a007096

[30] HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. Jordan, A., Bisgrove, D., and Verdin, E. EMBO J. 2003; 22: 1868–1877

[31] The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. Jordan, A., Defechereux, P., and Verdin, E.

[32] HIV latency and integration site placement in five cell-based models. Sherrill-Mix, S., Lewinski, M.K., Famiglietti, M., Bosque, A., Malani, N., Ocwieja, K.E., Berry, C.C., Looney, D., Shan, L., Agosto, L.M. et al. Retrovirology. 2013; 10: 90

[33] HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., Spindler, J., Ferris, A.L., Mellors, J.W., Kearney, M.F. et al. Science. 2014; 345: 179–183

[34] HIV-1 Integration Landscape during Latent and Active Infection, Lillian B. Cohn et al., Cell Link: http://www.cell.com/cell/fulltext/S0092-8674(15)00063-X

[35] HIV fasta https://www.ncbi.nlm.nih.gov/nuccore/4558520?report=fasta AF033819 9181 bp RNA linear VRL 26-JUL-2016 DEFINITION HIV-1, complete genome. ACCESSION AF033819 VERSION AF033819.3 GI:4558520

[36] Link: https://blast.ncbi.nlm.nih.gov/Blast.cgi

[37] Zvanična veb stranica VEP programa:

http://www.ensembl.org/info/docs/tools/vep/index.html

[38] Gonçalves J, Moreira E, Sequeira IJ, Rodrigues AS, Rueff J, Brás A. Integration of HIV in the Human Genome: Which Sites Are Preferential? A Genetic and Statistical Assessment. International Journal of Genomics. 2016;2016:2168590. doi:10.1155/2016/2168590

[39] http://bedtools.readthedocs.io/en/latest/content/tools/getfasta.html

[40] Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593-594. doi:10.1093/bioinformatics/btr708.

[41] Interactive genomics viewer http://www.igv.org/