УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Надежда М. Богдановић

РАЗВОЈ МЕТОДЕ ЗА ДЕТЕКТОВАЊЕ ПРОМЕНА БРОЈА КОПИЈА У ГЕНОМСКИМ СЕКВЕНЦАМА

мастер рад

Београд, 2020.

Ментор:

др Јована Ковачевић, доцент
Универзитет у Београду, Математички факултет

Чланови комисије:

др Невена Вељковић, научни саветник
Универзитет у Београду, ИНН Винча

др Саша Малков, ванредни професор
Универзитет у Београду, Математички факултет

Датум одбране: 09.09.2020.

„У Њему је једино моћућа разумност човечијеī бића, у Њему је једино моћуће оūравдање човечијеī ūостојања. У ūу истину се стичу све ūајне неба и земље, све вредности свих светова о којима човек може мислити, све радости свих савршенстава која човек може ūостизаūи.”

— *Ава Јустин Поūовић*

Наслов мастер рада: Развој методе за детектовање промена броја копија у геномским секвенцама

Резиме: Сазнања о промени броја копија геномских секвенци се последњих година све чешће користе у различитим биомедицинским истраживањима, а нарочито у истраживањима везаним за ретке генетске болести и канцер. У таквим истраживањима је од посебног значаја испратити промене броја копија на нивоу сваке ћелије појединачно, што се постиже методама секвенцирања једне ћелије. Комбиновањем постојећих метода секвенцирања једне ћелије и постојећих рачунарских метода није могуће прецизно одредити број сваке од копија геномских секвенци, нити са ког хромозома оригинално потичу. Овај рад има за циљ да представи нову методу секвенцирања једне ћелије - баркодирање генома, као и развој рачунарске методе за детекцију броја копија, која се посебно развија за ову методу секвенцирања. Истраживање о развоју методе детектовања варијација у броју копија је рађено у оквиру компаније *Digenomix*.

Кључне речи: детектовање, варијација у броју копија, баркодирање, геном, секвенцирање

Садржај

1 Увод **1**

2 Секвенцирање једне ћелије **5**

2.1 Постојеће методе секвенцирања једне ћелије 7

2.2 Баркодирање генома 9

3 Детектовање варијација у броју копија геномских секвенци **20**

3.1 Постојеће методе за детекцију варијација у броју копија 22

3.2 Архипелази 24

3.3 Процес детекције варијација у броју копија на основу формираних острва и архипелага 30

4 Имплементација методе за детекцију варијација у броју копија **35**

4.1 Подршка у виду постојећих биоинформатичких алата 35

4.2 Организација модула 39

5 Провера тачности методе **57**

5.1 Први ниво провере 57

5.2 Други ниво провере 58

5.3 Поузданост методе 59

6 Закључак **61**

Библиографија **63**

Глава 1

Увод

Свака здрава ћелија има своју предефинисану гарнитуру хромозома. Током свог развоја она може проћи кроз низ стресних услова попут нагле промене средине у којој се налази (велика промена pH вредности, нагло повећање или смањење количине кисеоника), напада других организама на ћелију или озбиљних поремећаја метаболичких процеса у самој ћелији.

Промена броја копија (енгл. *copy number variation (CNV)*) представља појављивање неких сегмената ДНК у једном организму у већем или мањем броју него у референтном геному. До ове појаве, као и до разних других промена у геномској секвенци (нпр. SNP и слично), може доћи у току ћелијске деобе. Ћелија код које је дошло до промена броја копија може одмах угинути или опстати, у зависности од услова у којима се налази. Ако су услови стресни и ћелија у њима успе да преживи, тада се промена броја копија може посматрати као еволутивно решење за опстанак у оваквим условима. Промена броја копија се чешће назива **варијацијом у броју копија**, па ће и овај назив бити коришћен у наставку рада. Процес детекције оваквих варијација ће бити детаљно описан у глави 3.

Када ћелија са поремећеним бројем копија делова геномске секвенце преживи и пренесе свој генетски склоп на потомство, она тиме започиње читаву нову популацију мутираних ћелија. Овај процес често претходи процесу канцерогенезе, па је детектовање варијација у броју копија један од метода за праћење развоја канцерогенезе у ткивима.

Овакве патогене ћелије често угину саме од себе или не успевају да се поделе. Уз то, у вишећелијским организмима постоје и ефикасни механизми за њихово уклањање, због чега их често није могуће детектовати и анализи-

рати. За разлику од њих, једноћелијски организми, попут пивског квасца, немају тако добро развијене механизме за детекцију и одстрањивање патогених варијетета, те их је самим тим много лакше анализирати. Сходно томе, ово истраживање се фокусира на геном пивског квасца (лат. *Saccharomyces cerevisiae*), чији је референтни геном већ познат научним круговима, у којима се користи за многобројна истраживања.

У истраживању су коришћени подаци добијени од девет различитих ћелија квасца означених идентификаторима 6Y5b_S33, 6Y5d_S34, 6Y5j_S37, 6Y5l_S38, 6Y5n_S39, 6Y5p_S40, 6Y6b_S41, 6Y6d_S42, 6Y6f_S43 и 6Y6h_S44. Истраживање ће бити представљено кроз узорак 6Y5j_S37, док су остали узорци коришћени за тестирање. Резултати за узорак 6Y5j_S37 и јавно су доступни на *github* репозиторијуму [1].

Како би се геномски садржај ћелије могао анализирати, мора му се одредити секвенца, односно редослед азотних база A, C, T, G, за сваки хромозом посебно. Ово се постиже процесом **секвенцирања** који спроводи машина секвенцер. За потребе овог истраживању коришћен је *Illumina MiSeq* секвенцер упарених кратких очитавања једне ћелије, што ће детаљно бити описано у глави 2.

Детектовање варијација у броју копија из података добијених коришћењем постојећих метода секвенцирања врши се рачунањем одређених статистика, што је објашњено у глави 3. Методе које су у широкој употреби нису довољно тачне и прецизне и имају велики број лажно позитивних и лажно негативних погодака. Додатно, на овај начин није могуће јасно одредити која копија гена долази са ког конкретног хромозома. Лабораторија *Digenomix* развила је нову методу секвенцирања једне ћелије - баркодирање генома (BIG, **B**arcode **I**n **G**enome) [2]. Метода се заснива на фрагментисању генома и додељивању јединственог идентификатора сваком од фрагмената, што омогућава њихово касније разазнавање приликом детекције варијација у броју копија. Истраживање о развоју методе детектовања варијација у броју копија је рађено у оквиру компаније *Digenomix*.

У данашње време потпуно је незамисливо било какво биолошко, хемијско или медицинско истраживање без подршке рачунарских метода. Оне омогућавају прецизну и брзу анализу података који се у истраживању користе. За различите проблеме развијане су различите методе, али за сваку од њих постоје два кључна захтева како би биле прихваћене од стране својих кори-

сника: **једноставност при коришћењу** и **ефикасност**. Метода детекције варијација у броју копија се заснива на два кључна корака: **новој методи секвенцирања ћелије (баркодирање генома)**, са биоинформатичком репродукцијом инсерата и контига на рачунарском нивоу (што ће ближе бити објашњено у глави 2.2) и **биоинформатичкој анализи** репродукованих података. Овај рад има за циљ да изложи и објасни поменуте биоинформатичке процесе за рад са техником баркодирања генома. Биоинформатички процеси описани су у програмском језику *Python*, уз коришћење стандардних биоинформатичких алата као што су *IGV*, *samtools* и *bedtools*.

Како је сама метода баркодирања сасвим нова и још увек у развоју, прати је и динамичан развој рачунарских метода. Овај рад ће представити основне претпоставке везане за детекцију варијација у броју копија, које се заснивају на резултатима добијеним из протокола баркодирања.

У глави 1 овог рада је објашњено шта су варијације у броју копија геномских секвенци и наведени су неки биолошки узроци ових варијација у броју копија геномских секвенци. Такође, представљен је сам приступ истраживању, као и скуп узорака над којима је истраживање вршено.

У глави 2 овог рада је извршено упоређивање технике баркодирања са осталим техникама секвенцирања и објашњено је како се изводи биоинформатичка анализа података добијених методом баркодирања. Како би варијације могле бити детектоване, најпре је представљен поступак реконструкције инсерата, а самим тим и хромозома на које се инсерти мапирају.

Глава 3 представља протокол који користе неке постојеће методе детектовања варијација у броју копија и њихове недостатке. Затим се представља нова идеја о детектовању варијација у броју копија, путем бојења инсерата који припадају различитим варијацијама. Обојени инсерти су представљени у неком од алата за визуелизацију геномских секвенци. Приступ се заснива на чињеници да је захваљујући баркодирању фрагмената могуће преклапање поравнатих инсерата једино у случају да припадају различитим варијацијама.

У глави 4 је представљена имплементација реконструкције инсерата и детектовања варијација у броју копија на рачунарском нивоу. Такође, приказани су и коришћени алати за рад са стандардним форматима биоинформатичких датотека.

У глави 5 је дискутовано како се може проверити да ли развијена метода може дати коректне резултате. Провера се врши на два нивоа: провера

исправности рачунарске имплементације методе и провера исправности детектованих варијација у броју копија геномског материјала.

У глави 6 је резимирано представљено истраживање. Такође, продискутовани су потенцијални начини примене резултата детектовања варијација у броју копија.

Глава 2

Секвенцирање једне ћелије

За боље праћење и разумевање даљег текста, неопходно је увести неколико термина који су саставни део стандардног биоинформатичког речника:

- **узорак**: геном ћелије која се анализира
- **фрагмент**: део молекула ДНК на који су прикачени адаптери
- **секвенцирање**: утврђивање редоследа азотних база (секвенци) на фрагментима
- **очитавање**: резултат пропуштања крајева фрагмента кроз секвенцер; очитавања су рачунарски записи секвенце азотних база крајева фрагмената
- **инсерт**: фрагмент без адаптера;
- **амплификација**: умножавање фрагмената;
- **варијација у броју копија**: варијација у броју копија геномских секвенци у ћелији [3]; овај термин ће бити детаљно обрађен у глави; 3
- **мапирање**: поравнање очитавања или инсерата у односу на референтни геном;
- **библиотека**: уопштено - колекција молекула припремљених у одређеном хемијском раствору; у овом раду односи се на колекцију фрагмената;

- **GC садржај**: укупан број молекула гуанина и цитозина у ДНК; аналогно се користи и термин GC%, који означава проценат ових азотних база у неком сегменту ДНК;

- **просечна дужина инсерта**: у овом истраживању је просечна дужина инсерта 200bp, а највећа прихваћена дужина је 1000bp[1]

Неки од ових термина ће бити детаљније објашњени у поглављима која следе.

Постоје два приступа секвенцирању на основу тога колико је ћелија укључено у сам процес: **секвенцирање скупа ћелија** и **секвенцирање једне ћелије**. Приступ секвенцирања скупа ћелија подразумева да су под нормалним околностима геноми скупа ћелија које се заједно секвенцирају идентични. Ћелије из скупа се углавном узимају из истог ткива, или из исте културе. Овакав приступ подразумева да су под нормалним околностима геноми скупа ћелија које се заједно секвенцирају идентични. Ово није случај код канцерогених или измењених нервних ћелија. Стечене варијанте се појављују у релативно малом броју ћелија и зато их је тешко детектовати. Код секвенцирања једне ћелије, циљ је испитивање разлика међу ћелијама и откривање функционално значајних варијанти у ћелији што је могуће раније. Кључна разлика у раду са подацима између ова два приступа лежи у њиховом односу према губитку, интерпретацији и употреби података [4, 5, 2].

Анализа геномских података добијених секвенцирањем једне ћелије омогућава разумевање функционисања једне ћелије у контексту њене микрооколине. Једна од предности таквог приступа јесте изучавање канцерогенезе и субклоналних популација канцера [6], као и испитивање одговора ћелије на стресне животне услове.

Наредна поглавља ће представити неке од постојећих метода секвенцирања једне ћелије, као и њихове предности и недостатке. Након тога биће описана најновија технологија секвенцирања једне ћелије -**технологија баркодирања генома** [2], која има за циљ да надомести недостатке постојећих метода секвенцирања. Након тога ће бити изведена анализа резултата, чији је циљ да ближе објасни саму технологију.

[1]Ово ограничење је последица засићења раствора транспозазе - што је концентрација транспозазе већа, то инсерти по правилу морају бити краћи. Горња граница од 1000bp је експериментално потврђена у лабораторији.

2.1 Постојеће методе секвенцирања једне ћелије

Како се последњих година све више открива о могућностима примене и значају података добијених секвенцирањем једне ћелије, сасвим је очекиван и пораст броја нових метода у овој области. У наставку ће бити описане неке од метода које се најчешће користе.

QPCR

Quantitative **P**olymerase **C**hain **R**eaction [7] се често примењује у специфичним доменима молекуларне биологије, као и у дијагностици приликом квантификације различитих биомаркера[2]. Ова метода се заснива на количинским мерењима информационе РНК у једној ћелији, с обзиром на то да је количина информационе РНК слика експресије гена у тој ћелији.

Ова метода успешно анализира ДНК и РНК молекуле, као и многе протеине, па чак и многе њихове комбинације. Нажалост, могуће је таргетирати ограничен број молекула - највише 96, чиме се постиже фаворизација одређених делова ДНК или РНК молекула, што ову технологију чини слабо применљивом у детекцији варијација у броју копија.

MDA

Multiple **D**isplacement **A**mplification [8] је метода која почива на умножавању различитих фемтограма[3] ДНК молекула. Како би се овај процес остварио, користи се ДНК полимераза из бактериофага[4] *phi*29, чија је улога управо синтетисање великог броја копија фемтограма.

Ова метода покрива већи проценат генома за разлику од PCR метода. MDA методом није могуће детектовати варијације у броју копија због експоненцијалне мултипликације ДНК фемтограма - не може се разазнати да

[2]Биолошке карактеристике које се објективно могу мерити и помоћу којих се детектују патолошки процеси у организму

[3]Фемтограм је мерна јединица за масу која износи $0.000000000000001(10^{-15})$ грама. Овде се под појмом фемтограм мисли на сегмент молекула ДНК поменутог реда величине.

[4]Вирус који напада бактерије. Овакви вируси су често богати полимеразама, чија је улога синтетисање ДНК или РНК ниски домаћина.

ли је варијација настала као артефакт секвенцирања или је заиста стечена варијација.

MALBAC

Multiple **A**nnealing and **L**ooping **B**ased **A**mplification **C**ycles [8] се сматра псеудо-линеарном методом амплификације која је у стању да избегне експоненцијални раст дупликата. То се постиже коришћењем специјалних основа које допуштају ампликонима[5] да поседују комплементарне крајеве. Резултат је итеративно умножавање уместо експоненцијалног.

Осим што резултује смањеним амплификационим шумом, ова метода је у стању да секвенцира обимне узорке и да умножава сегменте са обогаћеним GC садржајем, што се сматра проблематичним за остале методе. Ипак, метода је много осетљивија на контаминацију ДНК молекула услед лоше припреме за секвенцирање од осталих метода, или на различито припремљене библиотека за секвенцирање, што води до акумулације и пропагације грешке у случају њеног настанка у току процеса амплификације, а самим тим и до лажно позитивних варијација у броју копија.

LIANTI

Linear **A**mplification via **T**ransposon **I**nsertion [9] је метода секвецирања једне ћелије заснована на линеарној амплификацији, што значи да у процесу амплификације не учествују апликони који садрже неку грешку. На места у геномској ДНК се на случајно одабраним позицијама умећу промотерске секвенце[6], на основу којих се фрагменти ДНК амплификују у РНК молекуле. Након тога се врши инверзна транскрипција[7], чиме се завршава припрема библиотеке за секвенцирање.

Од свих горепоменутих метода, LIANTI постиже најбољу покривеност генома и у стању је да детектује варијације у броју копија много боље од осталих метода. Нажалост, начин изолације молекула ДНК из језгра при овој методи

[5]Ампликон је мултипликативни шаблон, односно шаблон за процес амплификације. Обично се састоји од етикете (енгл. *tag*), предње основе (енгл. *forward primer*), циљне секвенце (енгл. *target sequence*), инвертоване основе (енгл. *reverse primer*), инвертоване етикете и адаптера (енгл. *adapter*).

[6]Места на ДНК за која се везују протеини који иницирају транскрипцију - препис молекула ДНК у информациони молекул РНК.

[7]Начин добијања молекула ДНК од молекула информационе РНК

узрокује његову хемијску нестабилност, те ова метода пријављује известан број лажно позитивних и негативних варијација у броју копија, због чега је неопходно и секвенцирање сродних ћелија које служе као нека врста контроле, чиме су изгубљене суптилне промене на геному.

2.2 Баркодирање генома

Иако су се показале веома корисним, технике секвенцирања једне ћелије о којима је било речи у претходном поглављу су веома скупе у погледу временских и људских ресурса и имају тенденцију да при амплификацији фаворизују неке делове генома више од других [2]. Како би се ти недостаци превазишли, последњих пар година је у развоју сасвим нова техника секвенцирања једне ћелије - **баркодирање генома** [2].

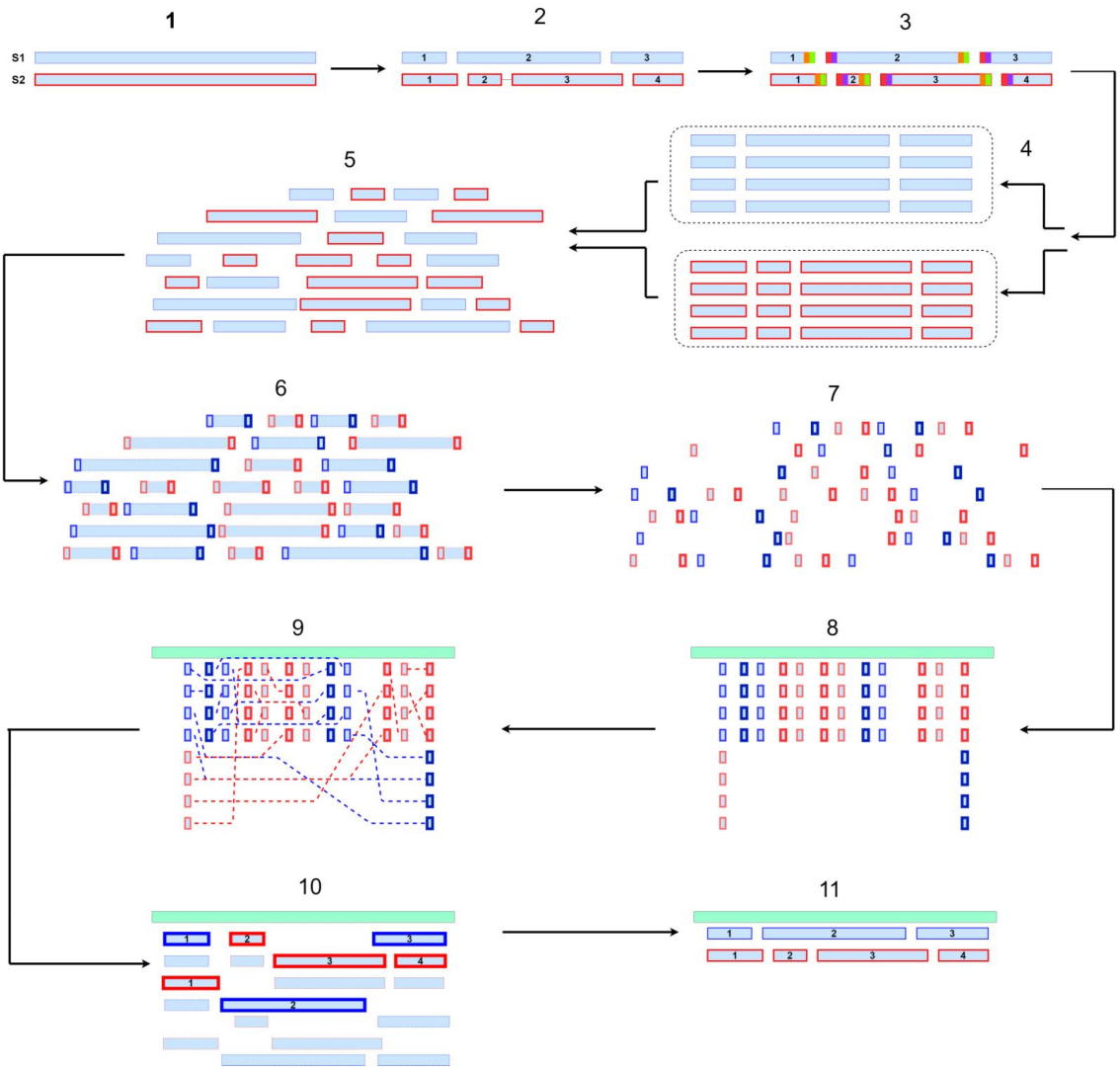
Техника се заснива на генерисању **баркода**, јединственог идентификатора за сваки *фрагмент*, *очитавање*, а касније и за сваки *инсерт*. Баркод се генерише на основу **дужине фрагмента** (или очитавања или инсерта) и његове **почетне позиције** у односу на референтни геном. На основу овако генерисаног баркода, могуће је увек знати који фрагмент је дошао са ког молекула ДНК, што у бити води детекцији варијација у броју копија.

Протокол технике баркодирања [2], приказан је на слици 2.1 и састоји од пет корака, који ће бити детаљно описани у секцијама које следе:

1. **Прирема ДНК молекула за секвенцирање**;
2. **Упарено секвенцирање**;
3. **Упарено мапирање**;
4. **Реконструкција инсерата**;
5. **Дедупликација инсерата**;

Припрема ДНК молекула за секвенцирање

Процес припреме ДНК молекула за секвенцирање састоји се од три корака:



Слика 2.1: Протокол баркодирања. У корацима 1-5 се припрема ДНК молекул за секвенцирање: у кораку 1 се врши екстракција ДНК из језгра и пречишћавање, у кораку 2 се врши фрагментација, у кораку 3 се врши тагментација и баркодирање, у кораку 4 се врши амплификација, у кораку 5 су представљени фрагменти који се прослеђују секвенцеру; у кораку 6 се врши секвенцирање крајева фрагмената; у кораку 7 је приказан резултат секвенцирања; у кораку 8 се мапирају упарена очитавања; кораци 9 и 10 представљају реконструкцију инсерата; а корак 11 дедупликацију инсерата

1. **Екстракција ДНК из језгра**: У овом кораку се молекул ДНК издваја из језгра и припрема за корак пречишћавања (приказано у кораку 1 на слици 2.1);

2. **Пречишћавање ДНК молекула**: У овом кораку се ДНК пречи-

шћава, односно ослобађа од хистона и других ензима, који су заслужни за компресију ДНК молекула и његову просторну структуру. Када се ензими одвоје од ДНК молекула, могуће је приступити одређеним сегментима и извршити фрагментацију молекула (приказано у кораку 1 на слици 2.1);

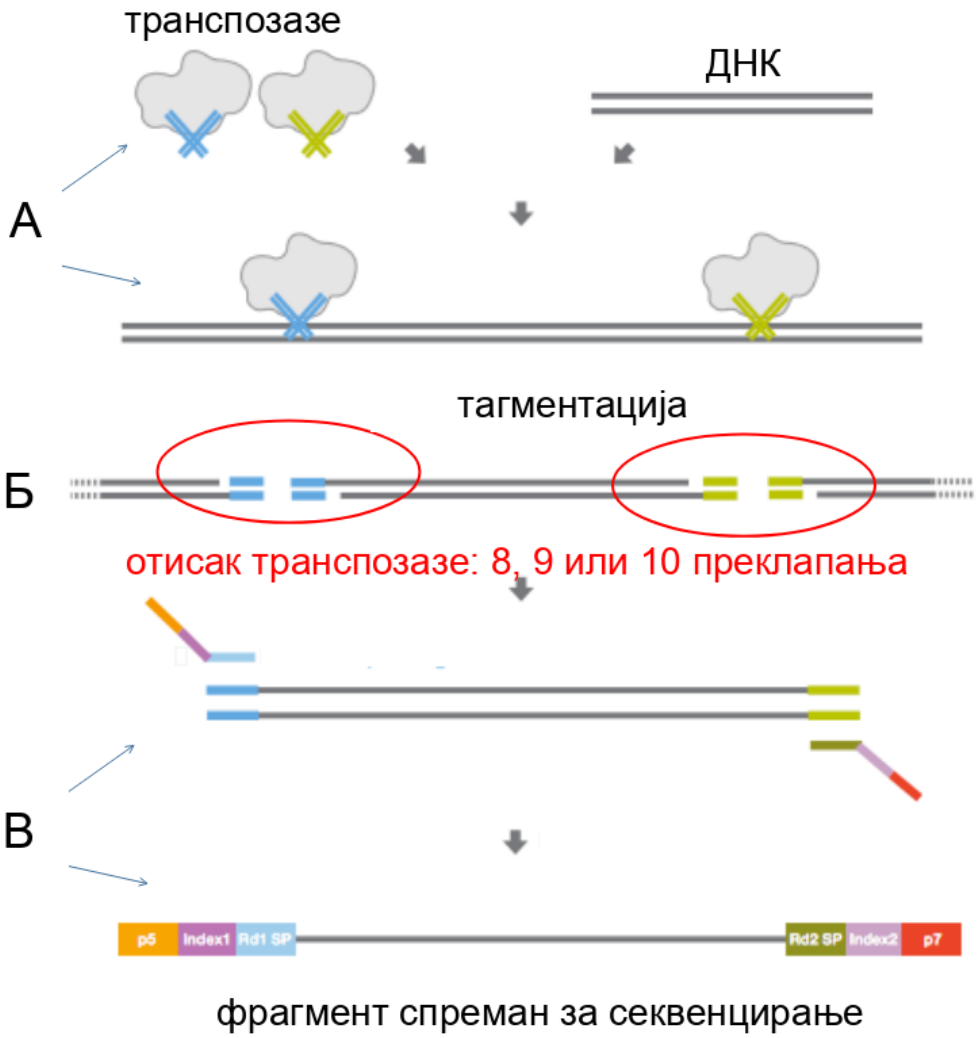
3. **Фрагментација и Тагментација**: Ово је уједно и најбитнији корак овог дела протокола у којем се изводи фрагментација, односно исецање молекула ДНК, коришћењем ензима *Tn5* [10]. Ово је ензим транспозазе екстрахован из бактеријских ћелија, чија активна места везивања[8] садрже два метална јона магнезијума и мангана. Њена улога је исецање молекула ДНК на произвољним местима у односу на референтни геном, тако да настају фрагменти различитих дужина. Комбинација фрагмената произвољних дужина и њихових произвољних позиција даје јединствени идентификатор за сваки фрагмент - баркод. Транспозаза се везује за молекул ДНК дужином од осам, девет, или десет базних парова[9]. Након пресека, на оба ланца остају делови који „штрче" у дужини од 8, 9, или 10bp, као што се види на слици 2.2. Касније, приликом поравнања на референтни геном, инсерти који су припадали истом молекулу ДНК ће се такође преклапати на 8, 9, или 10bp, односно на местима која „штрче". Такође, у овом кораку се изводи и баркодирање на молекулском нивоу. На сваки фрагмент се поред PCR адаптерских секвенци[10] „лепе" и секвенце баркода за сваки фрагмент [12] (приказано у корацима 2 и 3 на слици 2.1);

4. **Амплификација**: Фрагменти формирани у претходном кораку пролазе кроз процес амплификације. Такав скуп фрагмената се прослеђује секвенцеру. Да бисмо резултате секвенцирања могли да користимо за бројање копија, амплификација мора бити **равномерна**, тј. да не производи варијације у броју очитавања. Ово се постиже засићењем раствора транспозазе (енгл. *saturated transposition*) и баркодирањем.

[8]Места на којима се ензим везује за молекул ДНК.

[9]Дужина од девет базних парова се сматра канонском, дужина од десет базних парова је откривена пре 30 година [11] и на њу се није реферисало све до сад, а дужина од осам базних парова је откривена у процесу развоја технологије баркодирања и развоја методе за детекцију варијација у броју копија

[10]Кратке секвенце нуклеотидних база које служе за означавање 5' и 3' краја фрагмента који ће се умножавати.



Слика 2.2: Тагментација [13]: **А**: Везивање транспозазе са парцијалним адаптерима (плаво и зелено) за ДНК молекул **Б**: Тагментација са циљем фрагментисања и лепљења парцијалних адаптера на фрагментисане молекуле. Као што се може приметити у регијама означеним црвеном бојом, транспозаза сече молекул тако да оставља регије за које делује да „штрче" у односу на остале. Та места називамо отисцима транспозазе. Приликом поравнања инсерата на референтни геном, горњи и доњи инсерт (овде горњи и доњи фрагмент) ће се преклапати на 8, 9 или 10bp управо због делова који штрче; **В**: Лепљење PCR адаптера.

PCR библиотеке за амплификацију су осетљиве на GC садржај фрагмената [14]. У кораку тагментације се на сваки фрагмент лепи баркод (универзална секвенца), који уједначава GC садржај фрагмента. Тиме сви фрагменти имају једнаку шансу да буду умножени приликом процеса

амплификације [2], без обзира на GC садржај сваког фрагмента. Такође, PCR библиотеке за амплификацију су осетљиве на дужину фрагмената. То значи да у случају неуједначених дужина фрагмената, дужи фрагменти имају већу шансу да буду умножени него они краћи, што доводи до неравномерне амплификације. Овај проблем се решава засићењем раствора транспозазе - већа концентрација транспозазе резултује мањим фрагментима уједначене дужине [2] (приказано у кораку 4 на слици 2.1);

Специфичности припреме молекула ДНК за потребе овог истраживања

Пошто је пивски квасац који је коришћен приликом овог истраживања хаплоидан организам,[11], за потребе овог истраживања било је потребно симулирати на неки начин варијације у броју копија. Ово се постиже вештачким додавањем сегмената ДНК молекула на произвољним местима у односу на референтни геном. Ова процедура се разликује од стандардне [2], где нема вештачког додавања сегмената ДНК молекула.

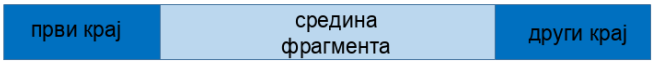
Такође, за потребе овог истраживања фрагментација се одвија на два нивоа. Први ниво подразумева одсецање делова генома дужине 2000bp до 10000bp (ово је оквирна горња граница, може бити и више). Само ови фрагменти (у глави 3.2 они одговарају архипелазима) ће бити испитивани, уместо целог генома, како би било лакше извести основна правила о детектовању броја копија. Селекција регија које ће бити секвенциране врши се у кораку припреме за секвенцирање на случајан начин, лепљењем одговарајућих адаптера. Случајни одабир регија се постиже случајно одабраним адаптерским секвенцама. Овакав приступ одступа од стандардног протокола [2] у којем се секвенцира цео геном, али је неопходан у раним фазама истраживања детекције варијација у броју копија. Други ниво фрагментације подразумева исецање горепоменутих регија на случајно одабраним местима. Ова процедура се не разликује од стандардне, осим што се примењује на регије генома, уместо на цео геном.

[11] Сваки хромозом у ћелији има једну копију

Упарено секвенцирање

Умножени и хемијски припремљени фрагменти се прослеђују секвенцеру, како би се извршило очитавање секвенци. Приликом истраживања, фрагменти су очитавани коришћењем *Illumina MiSeq* секвенцера [15] и технике секвенцирања упарених очитавања, која подразумева да се секвенцеру представљају почетни и крајњи делови фрагмента, као што је приказано на слици 2.3. Истраживање [2] је показало да је на основу оптималне дужине крајева фрагмента у интервалу (50, 70) базних парова могуће једнозначно идентификовати сваки фрагмент. Процедура је приказана у кораку 6 на слици 2.1.

У овом истраживању, секвенцирање је *псеудотаргетирано*, што значи да се не секвенцира цео геном, већ само неки његови случајно одабрани делови, као што је објашњено у претходном поглављу. Ова процедура се разликује од стандардне процедуре, приликом које се секвенцира геном у целости.



Слика 2.3: Секвенцеру се прослеђује цео фрагмент, а он скенира само крајеве тог фрагмента

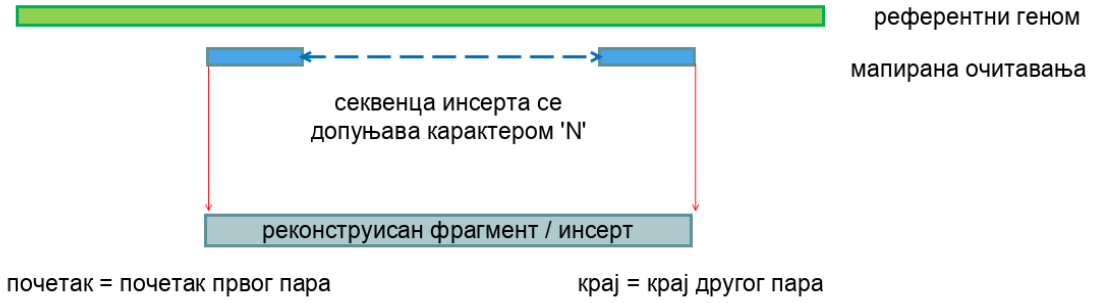
Упарено мапирање

Упарена очитавања се представљају софтверу за мапирање (маперу) који у односу на референтни геном поравнава ниске из очитавања и тиме позиционира сама очитавања у односу на референтни геном. У те сврхе је коришћен је *bowtie2* [16] мапер упарених очитавања. Процедура је приказана у кораку 8 на слици 2.1.

Реконструкција инсерата

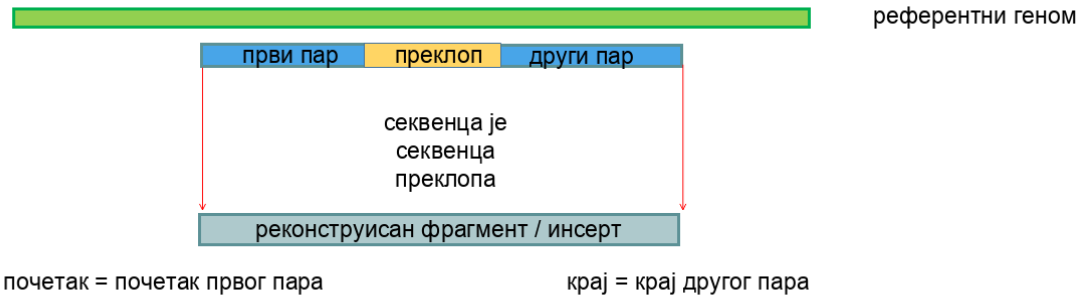
Након секвенцирања су изгубљени сви подаци о инсерту, осим позиција и секвенци упарених очитавања и потребно их је реконструисати. Потребно је да реконструисати читаву дужину и секвенцу инсерта. Процедура је приказана у корацима 9 и 10 на слици 2.1.

Процес реконструкције дужине инсерта је приказан на слици 2.4. Почетак инсерта представља мању од почетних вредности упарених очитавања,



Слика 2.4: Реконструкција инсерата када се упарена очитавања не преклапају

а крај представља већу од крајњих вредности упарених очитавања. Пошто средишњи део фрагмента никад није био секвенциран, не може се са сигурношћу реконструисати секвенца средишњег дела инсерта. У биоинформатици постоји конвенција за описивање таквих непознатих секвенци: средишњи део секвенце биће испуњен са онолико карактера 'N'[12], колика је и дужина средишњег региона мерена у базним паровима. Уколико се два упарена очитавања преклапају, онда се средишњи део секвенце не допуњава 'N' карактерима, већ је ниска на месту преклопа уједно и ниска средишње секвенце, што је приказано на слици 2.5



Слика 2.5: Реконструкција инсерата када средишњи део инсерта представља преклоп између упарених очитавања

У кораку који непосредно претходи реконструкцији се такође ради и фил-

[12]Значи да се на том месту може наћи било која од база: аденин, тимин, гуанин, или цитозин

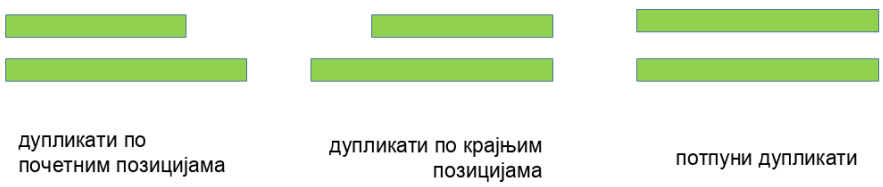
трирање упарених очитавања са циљем одстрањивања података лошег квалитета:

- Ако један од елемената (енгл. *mate*) пара очитавања није мапиран, не може се извршити реконструкција, па се одстрањују оба елемента.
- Ако су елементи пара превише удаљени (ако су поравнати на различитим хромозомима, или су удаљени више од 1000bp) онда се оба елемента (цео пар) одстрањују и не врши се реконструкција;
- Ако збир скорова квалитета мапирања очитавања није већи од 0, онда се не врши реконструкција и оба елемента се одстрањују, јер се скоро никакви логички закључци не могу извести за очитавања која имају изузетно лош индикатор квалитета мапирања;

Дедупликација инсерата

Инсерти могу имати своје дупликате у односу на позиције мапирања на референтном геному. Делимо их у три групе, као што је приказано на слици 2.6:

- Дупликати по почетним позицијама;
- Дупликати по крајњим позицијама;
- Дупликати и по почетним и по крајњим позицијама;



Слика 2.6: Дедупликација инсерата.

Дупликати настају као последица процеса амплификације. Полимераза, чији је примарна улога формирање умножака нуклеинских киселина на основу шаблона, има ограничен животни век. Може се десити да у току формирања једног полимера она просто откаже, не формиравши притом полимер

до краја. У зависности од тога да ли се полимер формира на основу примарног шаблона, или његовог комплемента, настају дупликати по почетним, или по крајњим позицијама.

Постоји могућност да транспозаза која сече молекул ДНК има неки дефект и да пресече само један ланац ДНК молекула уместо оба. У том случају настају дупликати који припадају првој или другој групи дупликата. Дупликати из треће групе веома личе на додатне копије геномског материјала, али су заправо последица процеса амплификације фрагмената. Овај закључак произилази из сазнања да транспозаза сече молекуле ДНК на случајно одабраним позицијама. То значи да, када би постојала додатна копија, била би мала вероватноћа да буде исечена на потпуно истим позицијама и да резултује фрагментима потпуно исте дужине.

Дупликати дају лажну представу о варијацијама у броју копија и потревно их је елеминисати, као што је приказано на слици 2.1 у кораку 11. Овај проблем се решава тако што се задржава најдужи дупликат. Ако су дупликати исте дужине, бира се онај са већим квалитетом мапирања.

Анализа резултата добијених применом протокола баркодирања

Резултат протокола баркодирања смешта се у датотеке са називима облика *imeuzorka_inserts_d.bam* датотеке, што означава да је за дати узорак извршена реконструкција инсерата и њихова дедупликација. Садржај тако добијене *.bam* датотеке[13] представљен је у колонама:

```
M02294:61:000000000-CM7ND:1:1104:16772:16728    0       NC_001133.9     120196  255     173M
```

Слика 2.7: Првих пет колона *.bam* датотеке: име инсерта, заставица, име хромозома, почетна позиција поравнања, индикатор квалитета мапирања и CIGAR ниска

1. **Име инсерта** - исто је као и име упарених очитавања од којих је реконструисан;
2. **Заставица** - целобројна маска која носи додатне информације о мапирању (да ли је очитавање упарено, да ли је пар очитавања добро

[13]Ова датотека је бинарна компресована верзија *.sam* датотеке која се користи за чување информација о поравнањима на референтни геном

мапиран, да ли постоје секундарна места мапирања, ...) У датотеци са инсертима, ова заставица увек има вредност 0, што означава да је у питању инсерт који нема свог пара (јер инсерти немају парове), да су парови очитавања од којих је реконструисан били правилно упарени и коначно, да је мапирање оваквог инсерта у односу на референтни геном једнозначно[14];

3. **Име хромозома** на који се инсерт мапира[15];
4. **Почетна позиција** мапирања инсерта у односу на референтни геном;
5. **Индикатор квалитета мапирања**;
6. **CIGAR ниска** [17] - Секвенца мапираног очитавања може имати додатне очитане базе које не постоје на референтном геному - *уметања* (енгл. *insertions*), или јој пак могу недостајати неке базе које се налазе на референтном геному - *губици* (енгл. *deletions*). CIGAR ниска је индикатор таквих и још неких суптилних промена очитавања у односу на референтни геном. Како овај рад има за циљ детекцију варијација у броју копија у општем случају, детаљи таквог типа неће бити разматрани и сва уметања и губици у упареним очитавањима биће занемарени, па ће се сматрати да је реконструисани инсерт у потпуном редоследу азотних база мапиран на референтни геном;
7. **Име пара** - Како инсерти немају парове, користи се '*' као ознака за недостајућу информацију;
8. **Дужина инсерта**;
9. **Секвенца инсерта**;
10. **Ниска појединачних квалитета мапирања (енгл. *Phred Score*)** - Сваки карактер ове ниске одговара једној бази из секвенце инсерта и означава њен појединачни квалитет поравнања на референтни геном;

[14]Протокол обезбеђује и могућност рада са инсертима који немају једнозначно мапирање у односу на референтни геном, али тај правац анализе података неће бити обрађен у овом раду

[15]У општем случају, то је име референтне секвенце у односу на коју се врши поравнање, која не мора бити нужно хромозом. Међутим, у овом истраживању се анализирају само поравнања на хромозоме.

11. **Етикете (енгл. *tag*)** - Наредне колоне означавају додатне информације везане за мапирање:

    - XG: Укупан број уметања и губитака;
    - NM: Број уметања сабран са бројем база које се не поклапају са референтним геномом;
    - XM: Број база које се не поклапају са референтним геномом;
    - XN: Број карактера на референтном геному различитих од 'A', 'C', 'T', 'G';
    - XO: Број различитих острва [16] уметања и губитака у једном очитавању или инсерту;
    - BC: Баркод;
    - YT: Носи информацију за сагласност упарених очитавања. Пошто инсерт нема пара, етикета за реконструисани инсерт ће увек бити „UU", што значи неупарен;
    - RG: Скривена етикета, која означава да је прослеђена YC етикета;
    - YC: Носи информацију о боји у виду уређене тројке која одговара компонентама RGB система боја;

    Имена ових етикета су унапред одређена стандардом за креирање *sam* датотека. Етикете које су коришћене у истраживању генерише сам мапер, а постоје и кориснички дефинисане етикете, као што је већ поменута етикета YC.

Ова датотека је сортирана помоћу алата *samtools* најпре по почетној позицији мапирања у односу на референтни геном, а затим и по квалитету мапирања.

[16] Једно острво инсерција и делеција (острво индела) чини више од једног узастопног уметања или губитка.

Глава 3

Детектовање варијација у броју копија геномских секвенци

Биће уведени неки битни термини дефинисани у односу на произвољни референтни геном и произвољни скуп узорака:

- **дубина поравнања**: број који означава колико је инсерата поравнато на дату базу на референтном геному;
- **острво**: група поравнатих инсерата, таква да у њиховој пројекцији на референтни геном нема непокривених база;
- **архипелаг**: групација острва; кластер инсерата мапираних у односу на референтни геном;
- **регија**: интервал на референтном геному; дефинише се уређеном тројком (x, y, z)[1];
- **регија лошег мапирања**: регија на референтном геному на којој се нагомилавају мапирани инсерти;
- **покривена регија**: регија на референтном геному где постоје поравнати инсерти из било ког узорка;
- **непокривена регија**: (енгл. *gap*), супротно од покривене регије - регија на архипелагу где нема поравнатих инсерата. Користе се и алтернативни термини: размак или рупа (енгл. *gap*);

[1](хромозом поравнања, почетна позиција поравнања, завршна позиција поравнања)

Варијације у броју копија су делови ДНК, чија дужина варира од 50bp do 1Mb (или чак више). Појављују се у једном геному у различитом броју копија у односу на референтни геном. По својој природи могу бити:

- дупликације (умношци) - могу бити тандемске[2] или мултиалелске дупликације[3];
- делеције (губици);
- инсерције (уметања између већ постојећих геномских секвенци);
- комплексни реаранжмани који обухватају већи број хромозома[4];

Детекцијом оваквих промена, односно детекцијом **варијација у броју копија** геномских секвенци у ћелији, могу се поредити различите ћелије и открити узроци патогености у ткиву или чак у организму. Један од већих изазова у савременим научним истраживањима представља поређење канцерогених ћелија, са циљем одређивања субклоналних популација[5], што је битно за различите предиктивне дијагностике и медицинске третмане [18]. Ове ћелије имају често потпуно поремећену гарнитуру хромозома у односу на здраве ћелије, па се не може закључити која ћелија је ћерка мутант, а која је референтна ћелија, а самим тим ни у ком правцу се развија болест[6]. Стога је веома битно да се у агресивним ситуацијама, као што је канцерогенеза, одреди колико укупно има копија геномских секвенци и са ког хромозома свака од копија потиче.

Секција 3.1 има за циљ да представи неке постојеће методе детекције варијација у броју копија. Осим што су неинтуитивне, ове методе користе велики број апроксимација различитих величина, што често доводи до лажно позитивних и лажно негативних резултата. Такође, нису у стању да одреде за

[2]Дупликације одређеног сегмента молекула ДНК који се у свим анализираним геномима дуплицирају на исти начин

[3]Дупликације чији се број дуплицираних сегмената молекула ДНК разликује између анализираних генома

[4]Сегмент молекула ДНК у једном геному може бити дуплициран а у другом делетиран

[5]Када једна канцерогена ћелија мутира и тако мутирана ћелија се размножи, читаво њено потомство се назива субклоналном популацијом те ћелије.

[6]Правац развитка болести се може открити на нешто вишем степену, путем дијагностиковања биомаркерима. Нажалост, често се у оваквим ситуацијама може само реаговати на дату медицинску слику неким третманом, али је већ касно за предупређивање болести, нарочито код агресивних типова канцера.

произвољно поравнато очитавање[7] са којег хромозома оно оригинално потиче, што значи да се њиховим коришћењем не могу детектовати узастопне копије неког гена у оквиру истог хромозома.

У поглављима која следе биће приказано да за детекцију копија нису неопходне апроксимације које се користе у постојећим методама, а поврх тога и да је могуће одредити и порекло сваке од копија. За то ће бити коришћен протокол детекције варијација у броју копија који се састоји из две целине:

1. Формирање и анализа архипелага, описани у поглављу; 3.2
2. Детектовање варијација у броју копија на основу формираних острва архипелага, описана у поглављу 3.3;

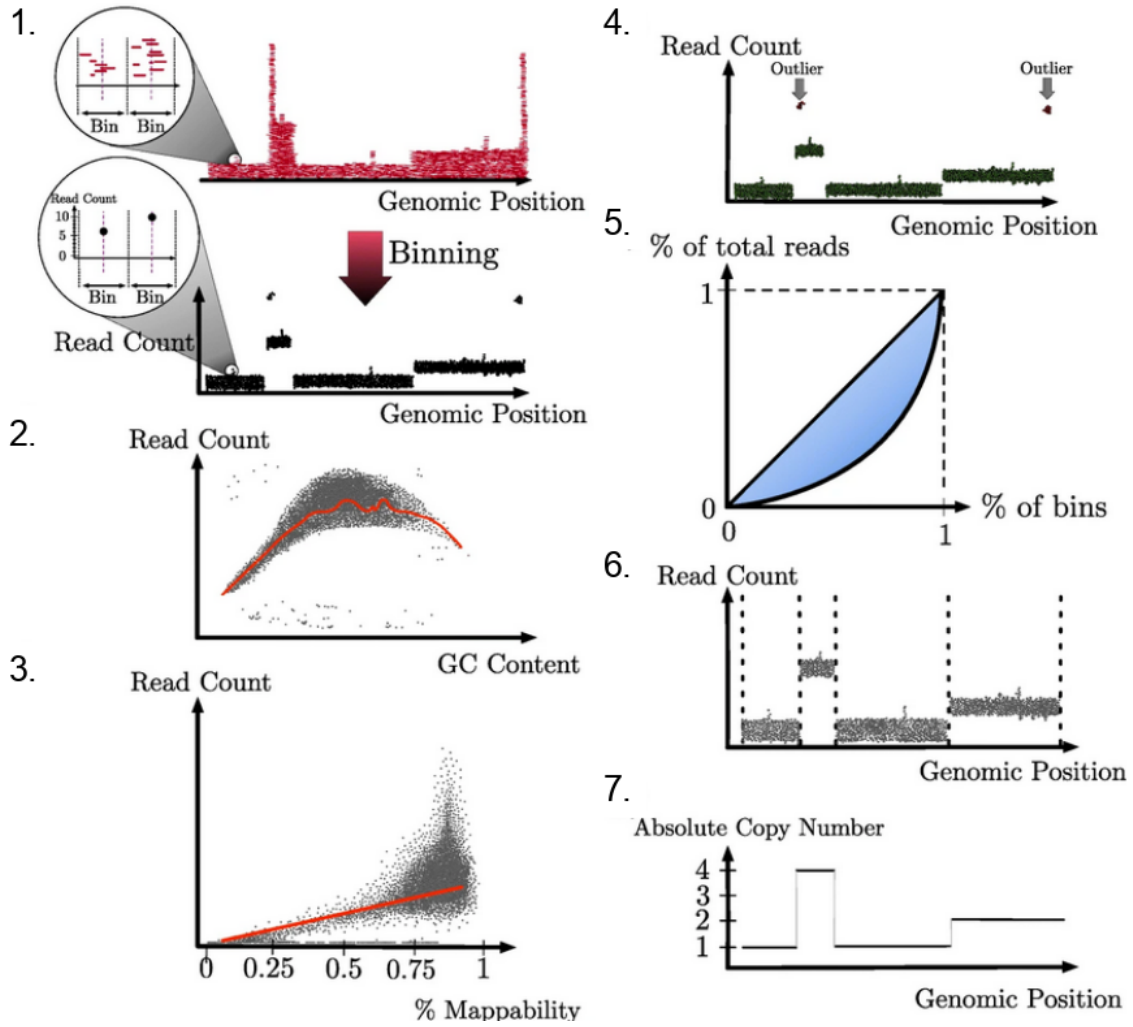
У овом и у наредном поглављу биће приказане слике добијене алатом *IGV* [19]. *IGV* је алат за визуелизацију геномских секвенци. У овом истраживању ће бити коришћен за визуелизацију мапирања у односу на референтни геном и обележавање регија од интереса на референтном геному. Алат пружа могућност за означавање мапирања бојама, као и њихово груписање. Овај алат је један од најзначајнијих алата коришћених у овом истраживању.

3.1 Постојеће методе за детекцију варијација у броју копија

Постојеће методе користе различите апроксимације дубина поравнања сваке базе. Већина њих прати протокол приказан на слици 3.1 са корацима:

1. Смештање очитавања у контејнере (енгл. *bin*) ширине 100bp. Број очитавања у сваком контејнеру се рачуна на основу дубине покривености на регији коју заузима контејнер;
2. Креирање тачкастог дијаграма где је на хоризонталној оси приказан GC садржај сваког контејнера, док је на вертикалној оси приказан број очитавања из сваког контејнера. Црвеном кривом је представљена регресија;

[7]Постојеће методе не користе инсерте, нити их реконструишу. То значи да простор између два упарена очитавања морају некако да апроксимирају. Ово није случај са методом баркодирања, јер она реконструише инсерте и нема потребе за апроксимацијама покривености између два упарена очитавања.



Слика 3.1: Протокол детекције варијација у броју копија који користе постојеће методе [20].

3. Апроксимирање мапирања. Потребно је исправити црвену линију са претходног графика како би се смањио број лажно позитивних и негативних очитавања;

4. Креирање тачкастог дијаграма где су на хоризонталној оси приказане геномске позиције, а на вертикалној број очитавања по контејнеру. Циљ је детектовање елемената ван границе који се елиминишу у следећем кораку;

5. Елиминација елемената ван границе. Крива приказана на графику је Лоренцова крива којом се представља укупан број очитавања у свим

контејнерима. За сваки елемент из претходног корака се рачуна Гинијев индекс. Гинијев индекс свих елемената би визуелно био приказан као површина два пута већа од плаве површине приказане на графику. Што је већи индекс елемента, то је већа вероватноћа да је елемент ван границе;

6. Сегемнтација. Поново се генерише дијаграм као из корака 4, само што се у овом кораку дефинишу јасне границе између различитх сегмената (скупова тачака). Сегментне границе су приказане испрекиданом линијом;

7. Одређивање броја копија. Сваком сегменту се додељује позитиван цео број на основу висине самог сегмента. У случају да је овај број рационалан, заокружује се на већу целобројну вредност;

Две најчешће коришћене методе за детекцију варијација у броју копија су:

- **HMMcopy** [20]: У кораку сегментације се користи HMM (**H**idden **M**arkov **M**odel) приступ. Стања модела одговарају бројевима копија[8], а прелази између стања одговарају сегментним границама. Локације се реконструишу пратећи путању дијаграма. Мана овог приступа је што захтева интервенцију човека приликом подешавања одређених параметара и што није у стању да оцени број хромозома у секвенцираној ћелији;

- **Ginko** [20]: Користи унапред задату листу лоших контејнера како би сваки контејнер који генерише могао да упореди са листом и евентуално га поправи. Његова највећа предност је уједно и његова највећа мана - предефинисана листа лоших контејнера. Такође, ограничен је само на одређене верзије референтног генома;

3.2 Архипелази

Приликом поравнања у односу на референтни геном, инсерту се придружује и његова **пројекција**, односно регија на референтом геному коју инсерт покрива својим поравнањем. Таква пројекција представља низ база на референтном геному које тај инсерт покрива. На пример, ако је инсерт поравнат

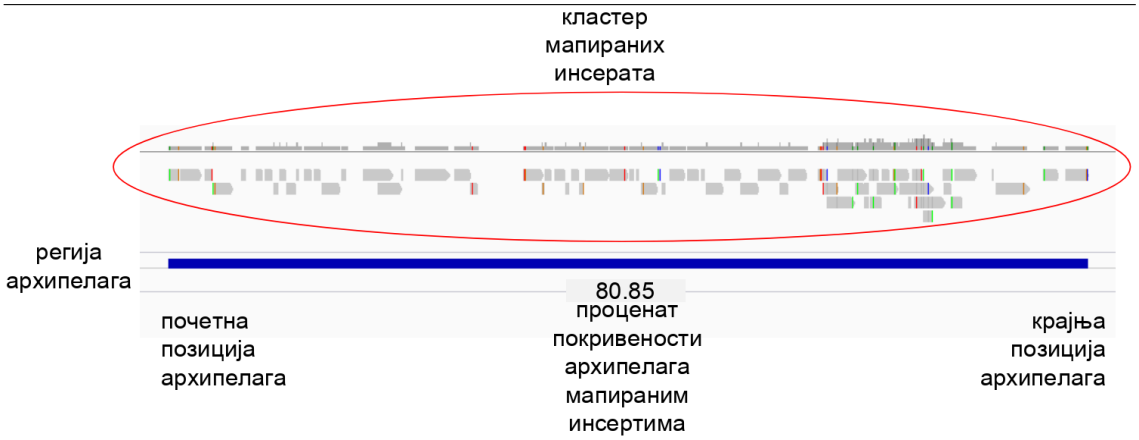
[8]За једну копију стање се означава као 1, за две копије стање се означава као 2, итд.

на референтом геному на хромозому NC_001133.9, са почетном позицијом 119.441 и крајњом позицијом 119.679, онда је његова пројекција низ база са редним бројевима у интервалу [119.441, 119.679] на NC_001133.9 хромозому. Инсерти својим поравнањима могу да формирају различите групе у односу на позиције на референтном геному. **Острво** је група инсерата чија пројекција не садржи непокривене базе, што је приказано на слици 3.2.

Слика 3.2: Пример групе инсерата која чини острво и пример групе инсерата која не чини острво. Део слике означен црвеним правоугаоником представља базе референтног генома и уједно степен покривености сваке од база инсертима. Обојене вертикалне линије означавају места мутација на инсертима у односу на референтни геном. Приказано у *IGV* алату за визуелизацију мапирања [19], на подацима добијеним из 6Y5j_S37_L001 узорка

Архипелази су групе острва. У својој бити, они су концентрисане групе поравнатих инсерата, односно њихови кластери, као што је показано на слици 3.3. Са слике се може приметити да и архипелаг, као и инсерт и острво, има своју пројекцију на референтни геном, што је приказано дугим испуњеним плавим правоугаоником. Број испод архипелага означава проценат покривености ове регије референтног генома, односно архипелага, поравнатим инсертима. У даљем тексту ће се термин *архипелаг* односити на његову пројекцију. Исто важи и за термин *острво*.

Пројекција архипелага се може представити и уређеном тројком (хромозом, почетна позиција, крајња позиција), која представља координате архипелага на референтном геному. Процес дефинисања оваквих тројки је описан следећим корацима:



Слика 3.3: Пројекција архипелага на референтни геном приказана је плавим правоугаоником. Сваки архипелаг има своју почетну и крајњу позицију на референтном геному, као и проценат покривености. Приказано у *IGV* алату за визуелизацију мапирања, на подацима добијеним из 6Y5j_S37_L001 узорка.

1. Рачунање координата архипелага се локализује на појединачне хромозоме. Ово значајно убрзава сам процес;

2. На сваком хромозому се одреде **непокривене регије** генома, који представљају неку врсту негатива мапираним инсертима (места са мапираним инсертима представљају позитиве);

3. Први инсерт на хромозому представља почетак првог архипелага. Нови архипелаг настаје, а тренутни се завршава, када се наиђе на непокривену регију дужине веће од 2000bp[9]. Другим речима, размак између два архипелага је минимално 2000bp;

4. Одсецају се делови архипелага који имају преклапања са регијама лошег мапирања у више од 10%[9];

5. Одбацују се архипелази дужине мање од 1000bp, пошто је то горња граница за дужину инсерта[9]. Задржавање таквих архипелага би значило да је теоретски могуће имати архипелаг од само једног инсерта, што је у супротности са дефиницијом архипелага;

6. Одбацују се архипелази чија је покривеност инсертима мања од 55%[9];

[9]Овај параметар се експериментално потврђује у лабораторији

Као што је већ поменуто у секцији 2.2, протокол баркодирања секвенцира само таргетиране делове генома. Сваки архипелаг одговара одређеном сегменту молекула ДНК који је таргетирано секвенциран, те тиме мора имати исте основне особине као и референтни геном, под условом да је протокол баркоирања обављен на исправан начин. Особине које користимо за валидацију протокола баркодирања су:

- **проценат G и C база**: Проценат G и C азотних база у ДНК молекулу (скраћено GC%) указује на стабилност самог молекула ДНК, а код генома пивског квасца износи око 38%. Анализом података добијених из 6Y5j_S37_L001 узорка долази се до следећих резултата:
  - **просек GC% покривених регија за све архипелаге**: 38.65% - рачуна се тако што се издвоји секвенца за сваки инсерт и обједини у заједничку секвенцу која представља покривене регије архипелага. За тако обједињену секвенцу се израчуна GC%;
  - **просек свих просечних GC% за све архипелаге на покривеним регијама**: 38.98% - рачуна се тако што се за сваки инсерт у архипелагу израчуна његов GC%, а онда се израчуна просек таквих добијених вредности;
  - **просек GC% непокривених регија за све архипелаге**: 37.64% - процес рачунања је исти, с тим што се вредности рачунају за рупе уместо за инсерте;
  - **просек свих просечних GC% за све архипелаге на непокривеним регијама**: 38.34% - процес рачунања је исти, с тим што се вредности рачунају за рупе уместо за инсерте;

  Просечна вредност ове четири вредности износи 38.04%, што је веома блиско вредности референтног генома.

- **проценат покривености мапираним инсертима**: Висок проценат покривености у архипелагу указује на то да ће приликом секвенцирања целог генома највећи број очитавања, а затим и инсерата бити исправно мапиран. За узорак 6Y5j_S37_L001, тај проценат износи 76.64%[10]. Додатно, ако је библиотека за секвенцирање припремљена на исправан

[10]У лабораторији је експерименталним путем утврђено да је 70% доња граница за овај параметар.

начин, проценат инсерата ван архипелага треба да буде веома мали. За узорак 6Y5j_S37_L001, тај проценат износи 1.83%, што је 71 инсерт од укупно 4905 инсерата;

Ако анализа резултата добијених протоколом баркодирања указује на то да је GC% у архипелазима приближно исти као код референтног генома а да је притом сваки архипелаг у великој мери покривен мапираним инсертима, може се стећи висок степен поверења у технологију баркодирања генома. Ово значи да је анализирајући архипелаге могуће валидирати протокол баркодирања, а самим тим је и дата мотивација за прелазак на следећу фазу истраживања: секвенцирање целог генома.

Регије лошег мапирања

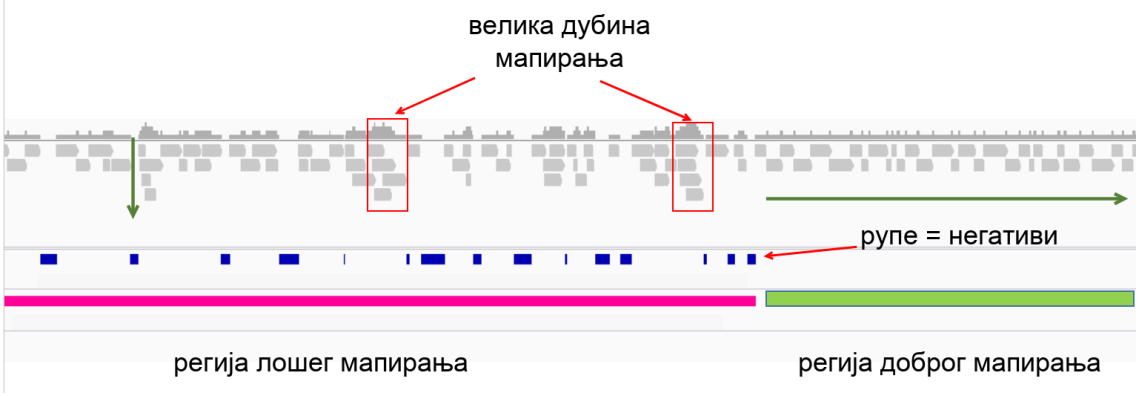
Када би се фрагментисао референтни геном на произвољним позицијама и када би се урадило мапирање таквих фрагмената[11] у односу на исти тај референтни геном, очекивало би се да такво поравнање буде савршено: праволинијско и без непокривених регија. Међутим, када се тако описан поступак спроведе у дело, откривају се неке регије на којима поравнање није тако савршено као што се очекивало. Такве регије називамо **регије лошег мапирања**. На њима дубина мапирања горепоменутих фрагмената значајно одступа од дубине мапирања на осталим регионима. Уочена је учестала појава понављајућих секвенци у овим регијама. Регије лошег мапирања су тиме особина самог референтног генома, што значи да не зависе од конкретног узорка. Једна од њих је приказана на слици 3.4.

Распоред непокривених регија у оквиру регије лошег мапирања оставља карактеристичан визуелни отисак, попут оног на слици 3.5.

И регије лошег и регије доброг мапирања се дефинишу уређеном тројком (хромозом, почетна позиција, крајња позиција), где се координате задају у односу на референтни геном. Процес дефинисања регија лошег мапирања у овим терминима је врло једноставан и аналоган је процесу дефинисања архипелага.

За инсерте мапиране на ове регије се не може закључити да ли носе варијације у броју копија, с обзиром на то да је на регијама лошег мапирања дубина

[11]Овде се под термином фрагмент мисли на малу целину, а не на физички молекул, као што је описано у глави 3.

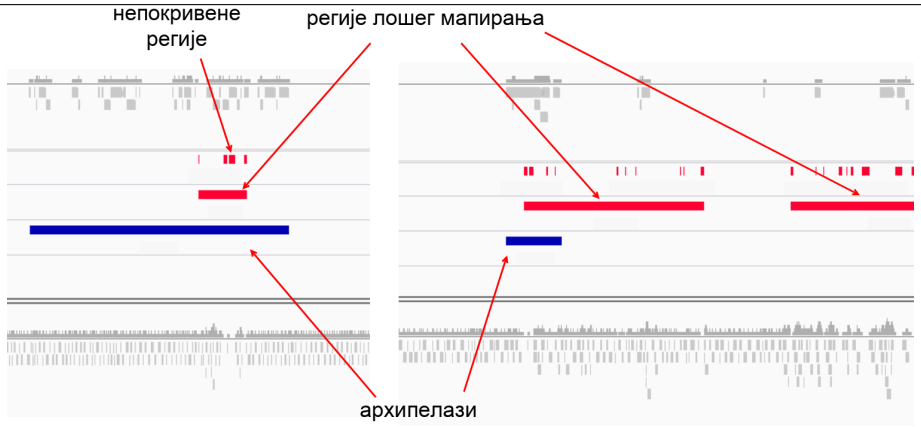


Слика 3.4: Регија лошег мапирања (розе) наспрам регије доброг мапирања (зелено). На регији лошег мапирања се може приметити велика дубина мапирања, као и појава непокривених регија - рупа. Зелене стрелице показују тенденцију да се мапирање у регијама лошег мапирања простире у дубину, а на регијама доброг мапирања у ширину. На овој слици су очитавања добијена фрагментисањем референтног генома, а затим поравната у односу на референтни геном. Плави правоугаоници представљају отисак регије лошег мапирања. Приказано у *IGV* алату за визуелизацију мапирања.

Слика 3.5: Отисак регије лошег мапирања. Бројеви означавају дужину сваке од непокривених регија која учествује у отиску. На слици делује као да неке регије немају придружен број, што је последица начина приказивања елемената у *IGV* алату за визуелизацију мапирања.

мапирања већа него на осталим регијама[12]. Сходно томе, архипелазима који се преклапају регијама лошег мапирања на више од 10% своје дужине, биће одсецани делови преклопа са оваквим регијама, као што је приказано на слици 3.6.

[12]О самој вези између дубине и варијација у броју копија биће речи у поглављу 3.3.



Слика 3.6: Преклапање са регијама лошег мапирања. Архипелаг лево неће бити одсечен, а архипелаг десно хоће. Приказано у *IGV* алату за визуелизацију мапирања.

3.3 Процес детекције варијација у броју копија на основу формираних острва и архипелага

Пивски квасац је хаплоидан организам[13] - има једну копију сваког хромозома. То значи да је очекивана дубина покривања сваке базе једнака један. Свако место наслаганих инсерата, односно место са дубином покривености веће од један, као што је предстаљено на слици 3.4, представља место варијација у броју копија. Овакав закључак представља и основ идеје о детекцији: траже се сви инсерти који се преклапају, те тиме повећавају дубину покривености одређених база преко један.

Смисао оваквог приступа проистиче из самог начина фрагментисања и таргетирања генома. Како је геном секвенциран таргетирано, постоје тачно два случаја када се два мапирана инсерта могу преклапати (с обзиром на то да су дупликати и очитавања са лошим квалитетом мапирања одстрањена у процесу реконструкције инсерата):

- инсерти припадају истој копији - дужина преклопа је у том случају 8, 9, или 10 базних парова;

[13]У овом истраживању, додатне копије делова хромозома су додаване вештачким путем приликом припреме за секвенцирање, како би се симулирала патогеност ћелије.

- инсерти припадају различитим копијама - дужина преклопа је различита од отиска транспозазе (свака копија геномских секвенци има очекивано исту секвенцу, тако да ће се све копије мапирати на исто место у односу на референтни геном; на регији на референтном геному која одговара гену са више копија биће примећено више наслаганих инсерата који се преклапају);

Једна од карактеристика добре методе за детекцију јесте једноставност и интуитивност при коришћењу. Најједноставније би било када бисмо могли некако да визуелно симулирамо хромозоме и копије гена на њима и да копије геномских секвенци представимо неким целим бројем и укажемо на њих. Сличан ефекат се постиже коришћењем линеарних алата за визуелизацију геномских секвенци, као што је *IGV*. Поред многобројних опција које пружа својим корисницима, овај алат обезбеђује две врло важне могућности: **бојење мапираних инсерата** на основу етикете која им се додели и **распоређивање инсерата у различите групе** на основу боје која им је додељена. Различите боје би представљале различите копије геномских секвенци. На тај начин је за детекцију копија довољно сваком инсерту доделити одређену етикету која представља боју, па да процес детекотвања буде комплетан.

Како би детектовање варијација у броју копија била успешно, неопходно је поставити неколико правила за бојење инсерата, која су наведена по приоритету примене:

1. Додела етикета са бојама инсертима, у даљем тексту бојење, се врши у сваком острву посебно. Већ је речено да тражимо инсерте који се преклапају, а инсерти се могу преклапати само у оквиру острва. Такође, потребно је најпре реконструисати примарну копију неке геномске секвенце, а тек онда реконструисати и остале варијације у броју копија. Ово се постиже управо доделом увек исте боје првом инсерту у острву.;

2. Као што је већ објашњено, два инсерта који се преклапају припадају различитим молекулима, па се и различито боје. Овде је битно пазити на инсерте који се преклапају на 8, 9 и 10bp. Они представљају такозвани **отисак транспозазе**, односно реферишу на места на молекулу ДНК за која се транспозаза везала, а затим и пресекла сам молекул. То значи да инсерти који имају преклоп на 8, 9 или 10bp припадају истом молекулу, односно истој копији, па такве инсерте бојимо истом бојом.

Овај механизам рада транспозазе је већ објашњен у одељку 2.2, у кораку тагментације;

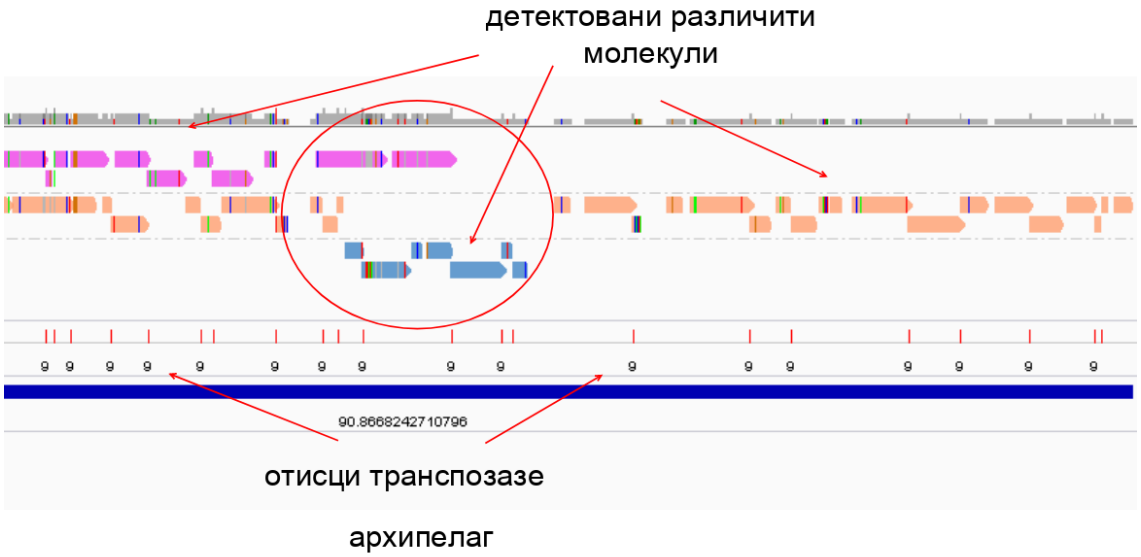
3. Бојење дуж хромозома је **похлепно**. То значи да покушавамо што више инсерата да доделимо већ постојећим молекулима. То има смисла ако се узме у обзир да се додатна копија не може назвати варијацијом, ако се најпре не зна која је примарна копија. Ово је упоредиво са постојећим методама којима су за детекцију потребни подаци из додатне валидационе ћелије како би утврдили шта јесте варијација у броју копија, а шта није. За разлику од тих метода, овде није потребно позивати се на другу ћелију. Ово правило се користи када је потребно одлучити да ли је неопходно генерисати нову боју за бојење текућег инсерта, или се он може обојити неком од већ постојећих боја;

4. Бојење почива на **принципу блискости** између инсерата: што су два инсерта ближе поравната један другом у односу на референтни геном, то је већа вероватноћа да потичу с истог молекула. Овај принцип проистиче из начина припреме библиотеке за секвенцирање и експериментално је потврђен у лабораторији. Ово правило је уједно и природна допуна претходног правила. Ако је применом претходног правила донета одлука о томе да се инсерт бојим неком већ постојећом бојом, неопходно је одлучити којом. Ту на снагу ступа принцип блискости који каже да се инсерт боји бојом једног од његових претходника који му је најближи, а с којим се не преклапа;

Исправност оваквог приступа бојењу лежи у чињеници да бојење различитом бојом почива на принципу детектовања промене дубине на острву. Све тренутне одобрене методе почивају на овом приступу, те њега није потребно посебно доказивати, већ је довољно позвати са на доказе већ поменутих метода.

На слици 3.7 је приказан један пример бојења, односно детекције различитих копија. Са слике се јасно види смисао принципа похлепног бојења - најбројнији су инсерти који су обојени наранџастом бојом. Тиме је најпре формирана примарна копија, а остале се сматрају секундарним у односу на њу. Може се приметити да у средишњем делу слике постоји размак између инсерата обојених наранџастом бојом и да би се инсерти обојени плавом бојом уклопили лепо у ту празнину када би били обојени наранџастом бојом.

Међутим, са слике 3.8 се види да се инсерти 4 и 5 преклапају на броју база различитом од 8, 9, 10bp. У лабораторији је потврђено да су ове ситуације последица дубине секвенцирања[14] и да би се празнине попут ове попуниле инсертима када би дубина секвенцирања била већа.

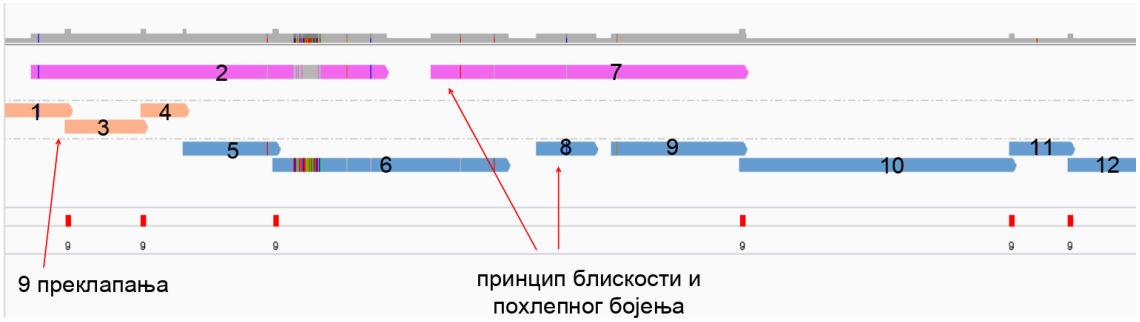


Слика 3.7: Визуелна детекција варијација у броју копија. Примарна копија је означена наранџастом бојом, а остале боје представљају додатне копије. Приказано у *IGV* алату за визуелизацију мапирања.

Слика 3.8 представља увеличану регију означену на слици 3.7. Пошто је референтни молекул конвенцијом овог истраживања означен наранџастом бојом, први инсерт у сваком острву бојимо том бојом. Ово је директна примена правила 1, а имплицитна примена принципа похлепности(правило 3). Следећи поравнат на референтни геном је инсерт са редним бројем 2, пошто су инсерти сортирани по почетној позицији поравнања у односу на референтни геном. Он се преклапа са инсертом 1. на броју базних парова који је већи од дужине отиска транспозазе, па се боји другом бојом. Затим, следе инсерти 3. и 4. који се боје истом бојом као 1. јер се преклапају на 9bp, што је директна примена правила 2. Инсерт 5. се преклапа и са 4. и са 2. на броју базних парова различитих од 8, 9, 10 па се боји новом бојом. Инсерт 6. се преклапа на 9bp са 5. па се и он боји плавом бојом. Инсерт 7. се преклапа са 6. на броју базних парова различитих од 8, 9, 10, па се не може обојити

[14] Дубина секвенцирања се израчунава као d = LN/G, где је L просечна дужина очитавања, N је укупан број очитавања, а G је дужина генома [21].

плавом бојом. Дакле, треба га обојити или наранџастом, или розе бојом или увести нову. Због принципа похлепности нећемо уводити нову боју, већ ћемо одабрати наранџасту или розе, јер се 7. не преклапа на броју базних парова различитих од 8, 9, 10 ни са једним инсертом обојеним од те две боје. Ту на снагу ступа принцип блискости, који каже да се у случају одабира већ постојеће боје инсерт боји бојом оног инсерта који му је најближи а с којим се не преклапа. У овом случају то је розе, па се инсерт 7. боји розе бојом. Даље се бојење врши по претходно објашњеној процедури.



Слика 3.8: Принципи бојења. Инсерти су означени у редоследу којим се поравнавају у односу на референтни геном. Црвени правоугаоници и бројеви испод њих означавају места преклапања инсерата на 9bp, односно отисак транспозазе. Приказано у *IGV* алату за визуелизацију мапирања.

Слике 3.7 и 3.8 су добијена учитавањем датотека *6Y5j_inserts_d_cnv_visual_sorted.bam*, *6Y5j_inserts_d_fltered_overlaps_8_9_10.bed* и *6Y5j_inserts_d_archipelago_call_2000_coverage.bed* у *IGV* алат. Датотеке су јавно доступне на *github* репозиторијуму [1].

Овим је показано да је могуће детектовати варијације у броју копија у подацим добијеним техником баркодирања генома. У наредној глави биће описан развој ове методе у програмском језику *Python*.

Глава 4

Имплементација методе за детекцију варијација у броју копија

Циљ развоја методе за детекцију варијација у броју копија је да пружи подршку биохемичарима и дијагностичарима у лабораторији који се баве дијагностиком или даљим развојем технике баркодирања. Сходно томе, она мора бити довољно **једноставна** за разумевање и употребу, али и **ефикасна**, јер је предвиђена да барата са великим количинама података. Такође, веома је битно да је **лако доступна** и **униформна** за све кориснике. У даљем тексту биће приказано како ова метода испуњава поменуте захтеве, најпре кроз опис коришћених алата и библиотека програмског језика *Python*, а затим и кроз увид у програмске модуле методе.

4.1 Подршка у виду постојећих биоинформатичких алата

Ради бољег разумевања програмских модула за подршку методе детекције варијација у броју копија, најпре је потребно објаснити који су алати коришћени за развој методе и пратећих модула. Метода је развијана на *Ubuntu* оперативном систему, а програмски модули су компатибилни са оперативним системима из фамилије *UNIX* система. За програмски језик је изабран програмски језик *Python* пре свега због своје широке распрострањености у

биохемијским и биоинформатичким круговима, што методу чини лако разумљивом и доступном свима. Такође, има велику подршку у библиотекама за рад са специфичним биоинформатичким подацима. Неке од тих библиотека су и нестандардне *Python* библиотеке, *pysam* и *pyfadix*, за подршку у раду са одређеним форматима стандардних биоинформатичких датотека. Те библиотеке имају и свој пандан у виду алата командне линије *samtools* и *bedtools*, за рад са .sam и .bed датотекама. Као алат кључан за развој методе, коришћен је алат за визуелизацију мапирања - *IGV*.

Стандардни формати биоинформатичких датотека

Ради разумевања сврхе коришћења одређених алата, потребно је најпре појаснити неке стандардне **формате** биоинформатичких датотека:

- **fasta** [22]: Датотека за представљање нуклеотидних или пептидних секвенци[1]. Датотека је подељена на геномске секвенце, а почетак нове геномске секвенце означава се знаком >, након чега следи име секвенце, затим информације о типу организма и коначно, низ карактера. Садржај датотеке приказан је на слици 4.1;

```
>NC_001133.9 Saccharomyces cerevisiae S288c chromosome I, complete sequence
ccacaccacacccacacacccacacaccacaccacacaccacaccacacccacacacacatCCTAACACTACCCTAAC
ACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCAT
TCAACCATACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATACTGTTCTTCTACCCACCATAT
TGAAACGCTAACAAATGATCGTAAATAACACACACGTGCTTACCCTACCACTTTATACCACCACCACATGCCATACTCAC
CCTCACTTGTATACTGATTTTACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTC
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATGCACGGCACTTGCCTCAGCGG
TCTATACCCTGTGCCATTTACCCATAACGCCCATCATTATCCACATTTTGATATCTATATCTCATTCGGCGGTcccaaat
attgtataaCTGCCCTTAATACATACGTTATACCACTTTTGCACCATATACTTACCACTCCATTTATATACACTTATGTC
AATATTACAGAAAAATCCCCACAAAAATCacctaaacataaaaatattctacttttcaacaataataCATAAACATATTG
GCTTGTGGTAGCAACACTATCATGGTATCACTAACGTAAAAGTTCCTCAATATTGCAATTTGCTTGAACGGATGCTATTT
CAGAATATTTCGTACTTACACAGGCCATACATTAGAATAATATGTCACATCACTGTCGTAACACTCTTTATTCACCGAGC
AATAATACGGTAGTGGCTCAAACTCATGCGGGTGCTATGATACAATTATATCTTATTTCCATTCCCATATGCTAACCGCA
ATATCCTAAAAGCATAACTGATGCATCTTTAATCTTGTATGTGACACTACTCATACGAAGGGACTATATCTAGTCAAGAC
```

Слика 4.1: Приказ садржаја *fasta* датотеке са екстензијом *.fna*.

- **fastq** [23]: Датотека која садржи информације о очитавањима, веома слична *fasta* датотеци, осим што за сваку секвенцу садржи оцену квалитета очитавања. Ако је секвенцирање било типа „упарених очитавања”,

[1]Низови аминокиселина. Свака аминокиселина се може представити нуклеотидном секвенцом дужине 3bp.

онда се у анализи података користе две упарене датотеке уместо једне. Ознака за почетак очитавања је карактер @, иза којег следи идентификатор очитавања. Ако су очитавања упарена, онда је ово име исто за оба упарена очитавања у обе упарене датотеке. Затим следи знак '+', који је ознака да се у следећем реду налази низ оцена квалитета секвенцирања за сваку од база секвенце очитавања. Садржај датотеке је приказан на слици 4.2;

```
@M02294:61:000000000-CM7ND:1:1101:16527:1732 1:N:0:37
TACCTGTCACGTTCACGCCCATTGCAGCGTCACAGTCACGCAGACGTTCTGCCGCCACTTTGCTGTTTAATACCTGAACA
+
CCCCCFGGCF7EFG8E,EC@@EF9<,,C+,8,8,,;C,C7E7++;+C,;,,6,++++8:CC,9E,<C,,,C,<69,,69,
```

Слика 4.2: Приказ садржаја *fastq* датотеке.

- **sam** и **bam** [24]: Датотеке које садрже информације о поравнањима очитавања или инсерата на референтни геном. Осим информације о позицијама поравнања, ове датотеке носе и додатне информације попут квалитета поравнања, броја успешних погодака у односу на референтни геном, као и о броју индела. Ови типови датотека су детаљно објашњени у поглављу 2.2;
- **bed** [25]: Датотека која садржи информације о пројекцијама на референтни геном. Пројекције су детљно објашњене у поглављу 3.2. Користи се за праћење позиција мапираних података у односу на референтни геном. Формат је веома користан када се раде филтрирања позиција података, или када је потребно приказати места поравнања податка који носе одређену особину. Обавезне колоне редом представљају: секвенцу на референтном геному на коју се врши поравнање, почетну позицију поравнања и крајњу позицију поравнања. Остале колоне су опционе и носе додатне информације о координатама које су представљене у прве три колоне. На овој слици су представљене координате архипелага са додатном колоном која носи информацију о покривености архипелага. Формат је приказан на слици 4.3;

Алати и библиотеке

За сада су модули за детекцију варијација у броју копија доступни само за кориснике оперативних система из фамилије *UNIX* система. Такође, кори-

```
NC_001137.3 7040    19293   81.99624581735085
NC_001137.3 23118   27230   82.85505836575875
NC_001137.3 126265  135603  82.53373313343329
NC_001137.3 192876  196103  73.34986055159591
NC_001136.10    19834   31375   82.54917251537995
NC_001136.10    236840  242703  90.60208084598328
NC_001136.10    320733  331536  65.79653799870407
NC_001136.10    373319  387988  73.04519735496625
NC_001136.10    540257  542733  87.52019386106623
NC_001136.10    626302  632905  93.27578373466606
NC_001136.10    695389  702380  91.64640251752253
NC_001136.10    743680  755666  66.37744034707158
NC_001136.10    763129  775659  63.91859537110934
NC_001136.10    798487  802756  63.457484188334504
```

Слика 4.3: Приказ садржаја *bed* датотеке.

сник мора имати инсталирану барем верзију 3.6 програмског језика *Python* и барем *Java8*, као и следеће алате и библиотеке:

1. **bowtie2** [16]: Алат за поравнање очитавања на референтни геном. Носи епитет веома брзог и ефикасног алата, који је у стању да подржи неколико глобалних и локалних метода поравнања и кратких и дугих очитавања. Најбоље резултате даје приликом поравнања секвенци 50 до 100bp на велике геноме. Приликом развоја методе коришћена је верзија 2.3.4.1. Алат је доступан и за *Windows* оперативне системе;

2. **samtools** [26]: Алат за рад са датотекама које чувају информације о поравнањима на референтни геном. Кориснику пружа мноштво опција за претрагу, обједињавање, индексирање и анализу поменутих датотека. Како би у раду са великом количином података био ефикасан, писан је у програмском језику *C*. Приликом развоја методе коришћена је верзија 1.7. Алат је доступан само за оперативне системе који припадају фамилији *UNIX* оперативних система;

3. **bedtools** [27]: Алат за рад са датотекама које чувају информације о геномским интервалима (*bed* датотеке). Често га називају биоинформатичким „швајцарским ножем геномске аритметике", јер омогућава корисницима да раде сортирање, пресек, обједињавање, пребројавање и рачунање статистика геномских интервала, али и још много тога. Писан је већином у програмским језицима *C* и *C++*. Приликом развоја

методе коришћена је верзија v2.26.0. Алат је доступан само за оперативне системе који припадају фамилији *UNIX* оперативних система;

4. **pysam** [28]: Модул програмског језика *Python*, који се користи за рад са *sam* датотекама, које садрже поравнања кратких очитавања. Модул је креиран као омотач око алата *samtools*, чему дугује своју ефикасност. Приликом развоја методе коришћена је верзија v0.16.01. Модул је доступан само за оперативне системе који припадају фамилији *UNIX* оперативних система;

5. **pyfaidx** [29]: Модул програмског језика *Python* који се користи за рад са геномским датотекама. Почива на раду *samtools*-ове функције *faidx*, чији је задатак да индексира геномску датотеку. Индекси су показивачи на различита места у геномској датотеци и значајно убрзавају операције за рад са геномским подацима. Приликом развоја методе коришћена је верзија v0.5.9.1. Модул је доступан и за *Windows* оперативне системе;

6. **I**ntegrative **G**enomics **V**iewer [19]: Ефикасан, интерактиван алат за визуелизацију и истраживање геномских података. Једноставан је за коришћење и подржава флексибилну интеграцију геномских података. Доступан је у виду десктоп и веб апликација, а подржава више платформи, с обзиром на то да је писан у програмском језику *Java*. Да би се могао користити, на оперативном систему мора бити инсталирана барем *Java8*. Приликом развоја методе коришћена је верзија 2.7. Алат је доступан и за *Windows* оперативне системе;

4.2 Организација модула

Постоје два главна модула који се могу покретати независно један од другог: модул за **реконструкцију инсерата** и модул за **детекцију варијација у броју копија**, као што је приказано на слици 4.4. Модулу за детекцију броја копија придодат је помоћни модул за генерисање регија лошег мапирања. Имплементација препознавања регија лошег мапирања издвојена је у засебан помоћни модул зато што су регије лошег мапирања директно у вези са референтним геномом и независне су од самог узорка који се испитује. Због тога је помоћни модул потребно покренути само једном, независно од узорка. Оба главна модула садрже конфигурациону *jason* датотеку, помоћу које се

Слика 4.4: Организација модула. Пуне стрелице приказују редослед корака, а испрекидане ток података.

главној скрипти у модулу прослеђују директоријуми који садрже податке о различитим узорцима. Главне скрипте оркестрирају даљи редослед позива одређених модула и токове података. Рачунарска имплементација методе је доступна на захтев.

Модул за реконструкцију инсерата

Модул се покреће позивом из командне линије:

```
$ ./reconstruction.py config.json
```

Модул прати основне кораке неопходне за реконструкцију инсерата, објашњене у глави 2.2:

1. мапирање упарених очитавања;
2. тримовање;
3. реконструкција инсерата;
4. дедупликација инсерата;

Конфигурациона датотека, осим што прослеђује путање до података о секвенцираним узорцима, такође обезбеђује и подешавање опционих параметара. На пример, у случају да корисник већ има датотеке са поравнатим очитавањима за које жели да реконструише инсерте и није му потребно ново поравнање, он покретање процеса поравнања може прескочити подешавањем одговарајућег конфигурационог параметра. Пример конфигурационе датотеке приказан је на слици 4.5.

Мапирање упарених очитавања

За мапирање очитавања коришћен је алат *bowtie2*, описан у поглављу 4.1. За сваки алат који ради мапирање, па и за *bowtie2*, неопходно је индексирати геном. Индексирање је поступак креирања неколицине индекс датотека које садрже показиваче на различита места на референтном геному. Приликом мапирања очитавања, мапер не користи датотеку која садржи референтни геном, већ само индекс датотеке, како би сам процес поравнања био што ефикаснији. То се постиже следећим позивом из командне линије, којем се

```
{"samples":
»   [
»   »   {"6Y5j":{"alignment_mode":"ON",
»   »   »   "trimming_mode":"ON",
»   »   »   "alignment_command_line" : "bowtie2-build reference/Yeast.fna reference/
            bowtie2_index
                                    | bowtie2 -X 1000 -p 8 -x reference/bowtie2_index
                                       -1 6Y5j/6Y5j_S37_L001_R1_001.fastq.gz -2 6Y5j/
                                      6Y5j_S37_L001_R2_001.fastq.gz -a",
»   »   »   "initial_bam": "6Y5j/6Y5j_S37_L001.bam"}}
»   ],
"reference_genome": "reference/Yeast.fna"}
```

Слика 4.5: Пример конфигурационе датотеке. Реконструкција се врши над подацима добијеним секвенцирањем узорка 6Y5j_S37. Опционе функционалности се подешавају постављањем одређеног параметра на ON/OFF

прослеђује датотека генома који се индексира и путања до директоријума у којем ће ти индекси бити сачувани:

```
$ bowtie2-build reference/Yeast.fna reference/bowtie2_index
```

Затим се покреће сам алат који врши поравнање командом[2]:

```
$ bowtie2 -X 1000 \\ максимална величина инсерта
          -p 8 \\ број нити
           \\ директоријум са индексима генома
          -x reference/bowtie2_index
          -a \\ хватање секундарних места поравнања
           \\ парови очитавања
           -1 6Y5j/6Y5j_S37_L001_R1_001.fastq.gz
           -2 6Y5j/6Y5j_S37_L001_R2_001.fastq.gz
```

Резултат овог корака смешта се у датотеку са називом *imeuzorka.bam*, која садржи поравната очитавања у односу на референтни геном.

Тримовање

Алат који изводи процес поравнања је у стању да препозна секвенце PCR адаптера које су прикачене у процесу тагментације на фрагмент. Међутим, некад може да се деси да алат пропусти да детектује такве адаптере. Корак

[2]Прослеђивањем параметра -a, омогућује се хватање секундарних места мапирања за очитавања који се могу мапирати на више места. Иако у овом истраживању нису коришћена секударна поравнања, остављена је и та могућност.

тримовања обезбеђује њихову детекцију и одстрањивање. Техника баркодирања користи секвенцу AAGAGACAG на крају 5' фрагмента, а CTGTCTCTT на крају 3' фрагмента. Експериментално је утврђено да у највећем броју случајева остану неодстрањене само CAG секвенце на 5' крају очитавања и CTG секвенце на 3' крају очитавања, као на слици 4.6. Овакве делове секвенце PCR адаптера прикачене на очитавања називамо **PCR ожиљци**. Детектују се тако што се из секвенци очитавања изолују прве (последње) три базе. Ако се таква секвенца, која одговара PCR ожиљку, разликује у односу на секвенцу на референтном геному, одстрањује се из очитавања. Уколико ожиљци не би били одстрањени, могли би да утичу на дужину преклапања између два инсерта и тиме би долазило до грешака приликом детекције варијација у броју копија. Ово се постиже коришћењем библиотеке *pysam* за приступање очитавањима и креирање нових очитавања и библиотеке *pyfaidx* за приступање секвенцама на геному са унапред задатим геномским интервалима. Резултат овог корака је датотека с називом *imeuzorka_trimmed.bam* са тримованим, упареним очитавањима.

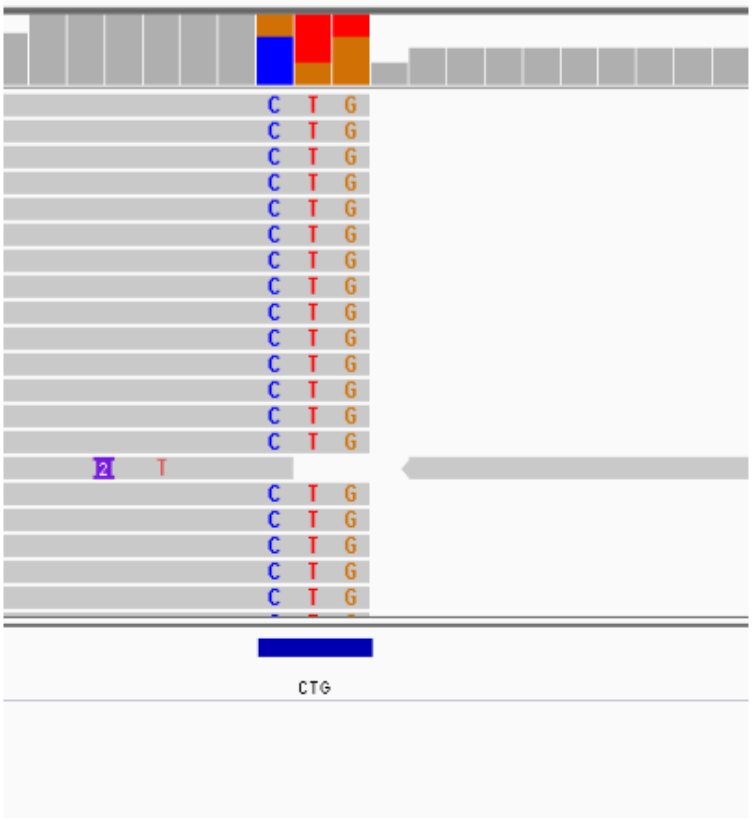
Реконструкција инсерата

У овом кораку, по којем читав модул носи име, одвија се реконструкција инсерата на основу приниципа објашњених у поглављу 2.2. Корак се извршава коришћењем библиотеке *pysam* за приступ очитавањима. Резултат корака реконструкције се чува у датотеци *imeuzorka_inserts.bed*, уместо у датотеци *bam* формата, како би се ефикасно извео корак дедупликације.

Дедупликација реконструисаних инсерата

Пошто је неопходно уклонити не само апсолутне, него и дупликате по почетној и крајњој позицији, није коришћен ниједан постојећи алат за дедупликацију, већ се процес извршава позивом *sort* алата командне линије. Како би се дедупликација над инсертима из претходног корака обавила у складу са принципима описаним у поглављу 2.2, примењује се команда:

```
$ sort -k6,6rn                \\ 1
| sort -k2,2 -k3,3n -k4,4n -u \\ 2
| sort -k1,1nr                \\ 3
| sort -k2,2 -k3,3n -u        \\ 4
```



Слика 4.6: PCR ожиљак на 3' крају очитавања. Приказано у *IGV* алату за визуелизацију мапирања

```
| sort -k1,1nr                  \\ 5
| sort -k2,2 -k4,4n -u          \\ 6
| sort -k2,2 -k3,3n             \\ 7
```

Команда прима датотеку са реконструисаним инсертима у коју је на почетку уметнута колона која означава дужину сваког инсерта. Први ред означава да се датотека сортира по скору квалитета мапирања, чиме се постиже да се у случају дупликата једнаких по почетној и крајњој позицији задржава онај са већим квалитетом мапирања. Други, четврти и шести ред обезбеђују уклањање дупликата, а трећи и пети да се између два дупликата одаберe дужи. Седми ред обезбеђује сортирање у односу на почетну позицију поравнања, а затим и односу на хромозом.

Резултат овог корака смешта се у датотеку с називом *imeuzorka_ inserts_ d.bed*. Коначна резултујућа *bam* датотека се креира коришћењем *pysam* библиотеке, која сваки инсерт посматра као објекат којем се прослеђују атрибути из да-

тотеке *imeuzorka_ inserts_ d.bed*.

Модул за детектовање варијација у броју копија

Модул се позива командом

```
$ ./cnv_detection.py config.json
```

и састоји из следећих логичких целина, у складу са описом методе детекције у поглављу 4.2:

1. детектовање архипелага;
2. пресек са регијама лошег мапирања;
3. филтрирање архипелага и инсерата;
4. детектовање острва;
5. бојење инсерата;

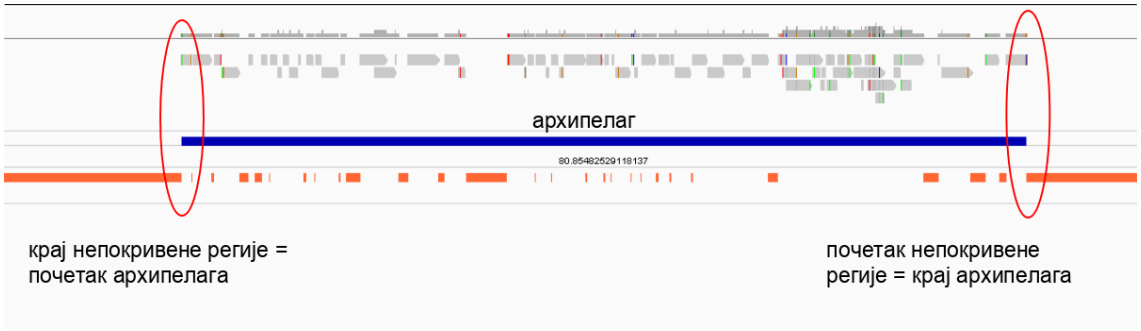
Детектовање архипелага

За детектовање архипелага, потребно је најпре креирати непокривене регије што се постиже командом:

```
$ bedtools complement
 -i imeuzorka_inserts_d.bed
 -g indeks_fasta_datoteke
 > imeuzorka_inserts_d_uncovered_regions.bed
```

Почетком архипелага сматра се почетак првог инсерта у том архипелагу, а крајем архипелага крај последњег инсерта у том архипелагу. Како се непокривене регије сматрају негативима инсерта, архипелази се могу наћи међусобним поређењем непокривених регија. Тиме би почетак архипелага био крај непокривене регије која му претходи, а крај архипелага би био почетак непокривене регије која му следи, што је приказано на слици 4.7.

Крећући се дуж референтног генома од непокривене до непокривене регије, наилазак на регију једнаку или већу од 2000bp означава почетак новог архипелага, што је приказано следећим кодом:



Слика 4.7: Почетак и крај архипелага у односу на позиције непокривених регија које му претходе и следе. Приказано у *IGV* алату за визуелизацију мапирања

```
def create_archipelagos():
    for contig in gap_map.keys():
        # we want to focus on one contig at the time
        gaps_on_contig = gap_map[contig]

        arch_start = -1
        arch_stop = -1

        # find starting gap
        gap_start = 0

        arch_start = gaps_on_contig[gap_start].stop
        arch_stop = gaps_on_contig[gap_start+1].start

        for i in range(gap_start+1, len(gaps_on_contig)-1):
            # if current gap is >= 2000bp
            # we can begin the new archipelago
            if gaps_on_contig[i].length >= MAX_GAP:
                arch_stop = gaps_on_contig[i].start

                archipelago = Interval(
                    contig, arch_start, arch_stop)
                archipelago_array.append(archipelago)
```

```
                arch_start = gaps_on_contig[i].stop
            else:
                arch_stop = gaps_on_contig[i+1].start
        # for the last one
        last_gap = gaps_on_contig[-1]
        arch_stop = last_gap.start
        archipelago = Interval(contig, arch_start, arch_stop)
        archipelago_array.append(archipelago)

    return
```

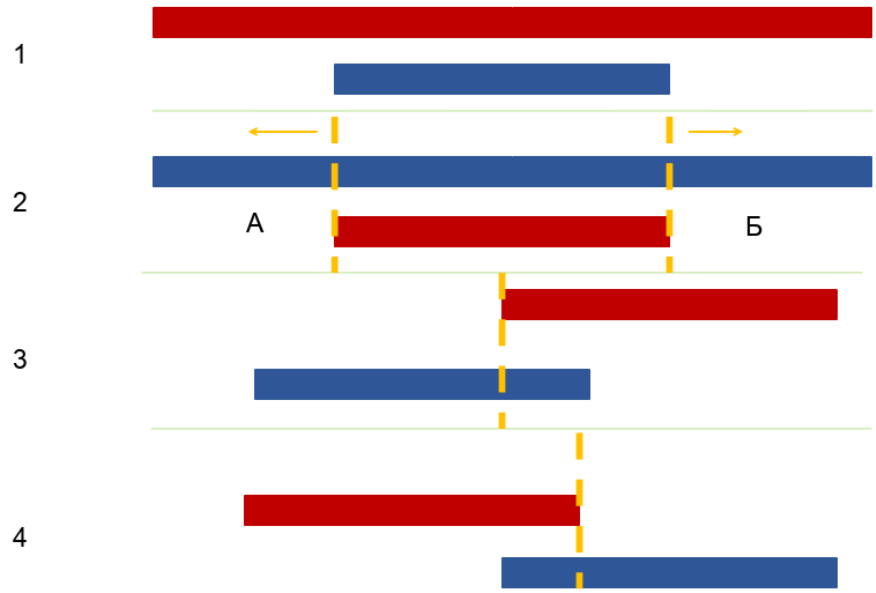
Резултат корака детектовања архипелага се чува у датотеци под називом *imeuzorka inserts_ potential_ archipelagos_ 2000.bed*. Ово још увек није коначан облик архипелага садржаних у поменутој датотеци. Потребно је филтрирати их у односу на регије лошег мапирања и покривеност.

Пресек са регијама лошег мапирања

И овај корак као и претходни користи информације о непокривеним регијама на референтном геному. Четири могућа случаја преклапања архипелага са регијама лошег мапирања су приказана на слици 4.8. У првом случају, архипелаг се сасвим одбацује јер потпуно упада у регију лошег мапирања. У осталим случајевима, одсецање архипелага се врши ако се преклапа са регијом лошег мапирања на више од 10%.

У другом случају са слике, потребно је поделити га на архипелаге А и Б. Почетак архипелага А остаје исти као почетак старог архипелага, а крај архипелага Б остаје исти као крај старог архипелага. Мора се водити рачуна да се не одсече архипелаг на месту где је поравнат неки инсерт (односно по средини инсерта). Пошто се користимо информацијама о непокривеним регијама уместо о инсертима, потребно је пронаћи први размак који претходи инсерту који се налази на месту пресека. Почетак тог размака ће бити крај архипелага А. Ситуација за архипелаг Б се решава супротно: потребно је пронаћи први размак након инсерта који се налази на месту пресека и његов крај. Крај тог размака ће бити почетак архипелага Б. Како би се овај процес обавио што ефикасније, смештање непокривених регија у оквире одговарајућег архипелага се постиже бинарном претрагом, која је имплементирана у *bisect*

модулу програмског језика *Python*. Трећи и четврти случај су подслучајеви другог, па се решавају на аналогно.



Слика 4.8: Случајеви преклапања регије лошег мапирања и архипелага.

Филтрирање архипелага и инсерата

Ови процеси су описани детаљно у поглављу 3.3. Проценат покривености архипелага се рачуна као $100 \cdot (N_{nepokrivene}/N_{arhipelag})$, где је $N\_nepokrivene$ укупна дужина свих непокривених регија на архипелагу, а $N\_arhipelag$ дужина читавог архипелага. Укупна дужина свих непокривених регија се рачуна тако што се саберу све дужине непокривених регија које упадају у интервал архипелага. Смештање непокривених регија у оквире одговарајућег архипелага се постиже бинарном претрагом, која је имплементирана у *bisect* модулу програмског језика *Python*. Задржавају се само они архипелази који имају покривеност барем 55%[3]. Датотека са коначним скупом архипелага који настају као резултат овог корака назива се *imeuzorka_inserts_d_archipelago call_2000.bed*.

[3]Иако је у секцији 3.2 представљено да је пожељно да сви архипелази имају покривеност већу од 70%, у лабораторији је одлучено да се приликом истраживања задрже и они чија је покривеност између 55% и 70%. Подаци добијени из оваквих архипелага служе лабораторији као повратна информација о томе како би требало унапредити библиотеку за секвенцирање.

Филтрирање инсерата подразумева издвајање само оних инсерата који припадају неком архипелагу. Смисао овог корака лежи у томе што се боје само инсерти који припадају неком архипелагу, односно регији за коју може да се закључи да ли постоји варијација у броју копија или не. Резултат се чува у датотеци с називом *imeuzorka_ inserts_ d_ filtered.bed*.

Детектовање острва

Архипелази су у глави 3.2 представљени као групе острва инсерата, па би према томе било логично да је за детектовање архипелага неопходно најпре детектовати и острва. Међутим, по својој дефиницији они имају заједничке особине: они су интервали на референтном геному (пројекције) и оба зависе од размака између инсерата. Једина разлика је у томе што за детектовање острва не постоји захтев за минималном раздаљином од 2000bp (као код архипелага). Сасвим је довољно детектовати непокривену регију произвољне дужине, односно регију на којима је дубина покривености базе 0. Резултат се чува у датотеци с називом *imeuzorka_ inserts_ d_ islands.bed*.

Бојење инсерата

Бојење инсерата одвија се у складу са принципима из поглавља 3.3. Поступак је приказан следећим псеудо-кодом:

```
begin
  // mapa koja cuva koordinate poslednjeg inserta
  // koji je obojen tom bojom
  color_map = {
    color : last_insert_colored_with_that_color=null
  }
  // inserti moraju biti sortirani po
  // koordinati pocetka poravnanja
  // pocinjemo sa primarnom kopijom
  molecules_in_island = 1
  for insert in island of sorted inserts:
    // pohlepni prinicip:
    // prvi insert u ostrvu se uvek boji prvom bojom
    if first_in_island == True:
```

```
 insert.color(1)
 color_map[1] = insert
else:
 // ako se insert i njegov prethodnik
 // preklapaju na 8, 9, ili 10 baza
 // bojimo ih istom bojom
 if insert.overlaps(previous) on (8, 9, 10)
    insert.color(previous.color)
    color_map[previous.color] = insert

 // ako se ne preklapaju istom bojom
 // treba naci neki od prethodnika
 // s kojim se preklapa na 8, 9, 10
 // ili onaj koji mu je pozicijom poravnanja najblizi
 // i obojiti ih istom bojom
 elif insert.overlaps(previous) on !(8, 9, 10) \
 or there is gap between them:
    min_distance = 'inf'
    nearest_color = null
    for color in color_map:

     // ako se preklapaju na 8, 9, 10
     // oboj ih istom bojom
     if insert.overlaps(color.insert) on (8, 9, 10):
        insert.color(color.insert)
        color_map[color] = insert
     // ako nema preklapanja
     // mogu se obojiti istom bojom
     else if !insert.overlaps(color.insert):
        if distance(insert, color.insert) < min_distance:
         min_distance = distnce
         nearest_color = color
         non_overlap_condition = True
   end for
```

```
        // ako se naislo na insert s kojim se trenutni ne preklapa
        // i koji mu je najblizi od svih inserata
        // obojenih istom tom bojom
        // oboj ih istom bojom
        if non_overlap_condition == True:
            insert.color(nearest_color)
            color_map[nearest_color] = insert
        // inace se mora obojiti novom bojom
        // ovo je mesto detekcije nove kopije
        else:
            // detektovali smo novi molekul
            molecules_in_island += 1
            new_color = generate_new_color
            insert.color(new_color)
            color_map[new_color] = insert


  end for
end
```

Пре почетка бојења неопходно је сортирати инсерте по почетној позицији поравнања у односу на референтни геном. Први инсерт у острву се увек боји истом бојом, чиме се поштује принцип похлепности. Затим се пореде међусобно инсерти. У случају да се преклапају на 8, 9, или 10bp, боје се истом бојом, пошто припадају истом молекулу. Ако се преклапају на броју базних парова који се разликује од отиска транспозазе (или ако се не преклапају уопште[4]), потребно је проверити да ли постоји неки од претходних инсерата с којим се не преклапа (или преклапа на дужини отиска транспозазе). Тада је потребно обојити их истом бојом. Овим се поштује приницип блискости. У случају да такав инсерт не постоји, значи да је детектована нова копија. За такав инсерт, који припада новој копији, потребно је генерисати нову боју и обојити га том бојом. Резултат овог корака се чува у датотеци с називом

[4]Инсерти се обрађују у редоследу по ком су сортирани и увек се пореди текући са претходним. У ситуацији у којој се претходни и текући не преклапају, могло би се закључити да је овде потребно применити принцип блискости. Али, могуће је да постоји неки од претходних инсерата који је дужи од директног претходника текућег инсерта, који се са текућим преклапа на 8, 9, 10bp. Тиме инсерти који припадају истом молекулу не би били обојени истом бојом.

imeuzorka_inserts_d_cnv_visual_sorted.bam. У датотеци *imeuzorka_color.bed* се чува податак о томе колико у свако острву има детектованих молекула.

Ефикасност овог поступка лежи у два кључна одсецања простора претраге:

1. **инсерти су сортирани** и довољно је поредити само суседе, уместо сваког са сваким;

2. **користи се мапа бојења**: она чува редни број за сваку боју и координате последњег инсерта у текућем острву који је бојен том бојом. Редни број боје се додељује приликом креирања сваке боје. Боје се генеришу на случајан начин, осим прве боје у структури. Она је увек наранџаста, по конвенцији. То значи да се примарној копији увек додељује прва боја из структуре, а последња боја у оваквој структури одговара последњој[5] детектованој копији у острву. У случајевима када је потребно вратити се уназад и пронаћи најближи инсерт који се не преклапа са тренутним, није потребно пролазити кроз читаво острво уназад и проверавати све инсерте. Довољно је проћи кроз мапу и покупити последње обојене инсерте;

Детектовање регија лошег мапирања

Пошто су регије лошег мапирања директно у вези са референтним геномом, а не са узорком који се обрађује, овај модул је издвојен од осталих као помоћни. Резултат се чува у датотеци *bad_mapping_archipelagos.bed*. Процес се састоји из следећих корака:

1. фрагментисање референтног генома на случајно одабраним местима[6] и креирање упарених очитавања;

2. покретање процеса мапирања над фрагментима добијеним из претходног корака;

3. реконструкција инсерата на основу мапираних упарених очитавања из претходног корака;

[5]Последња у редоследу акција доделе у острву, а не последња у смислу позиције поравнања на референтни геном.

[6]На захтев лабораторије, приликом фрагментисања се на почетак сваког фрагмента 5' додаје секвенца од девет базних парова која је иста као секвенца краја претходног фрагмента 3', како би се симулирао рад транспозазе.

4. детектовање непокривених регија на основу резултата мапирања из претходног корака;

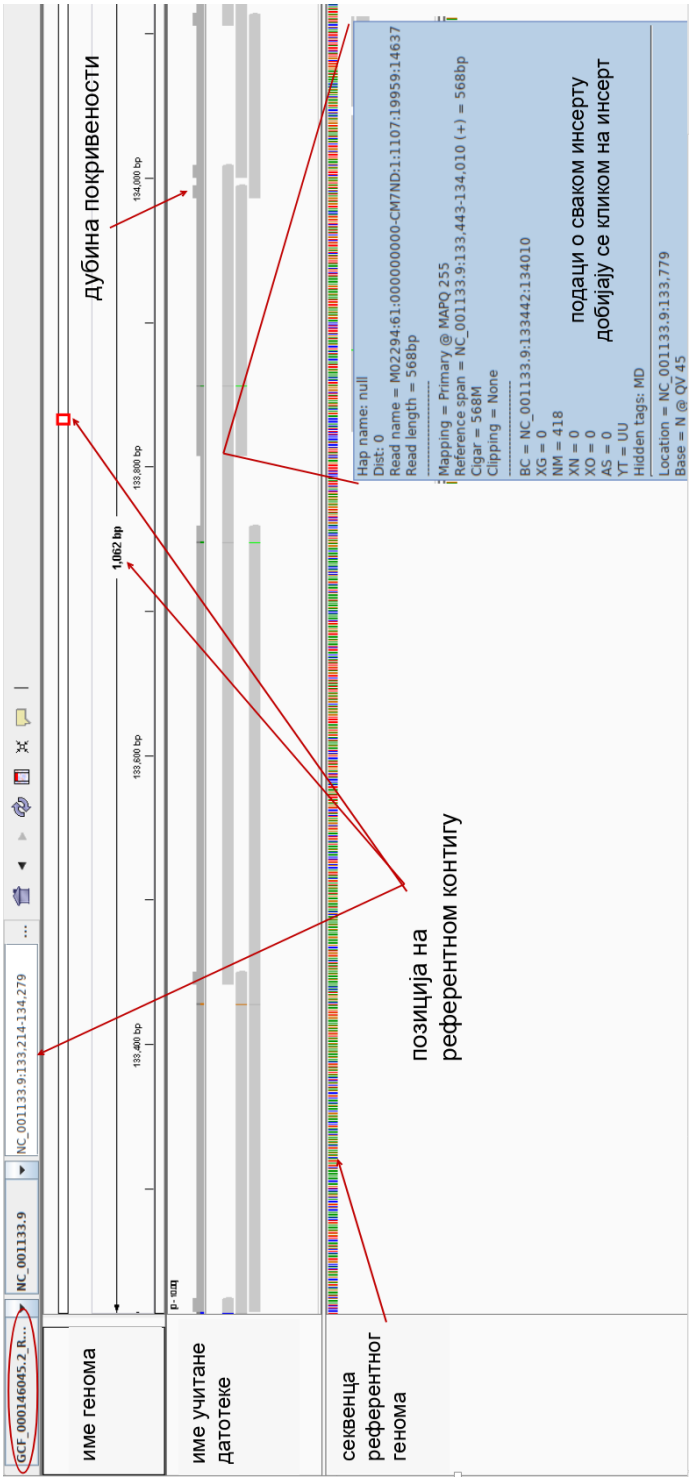
5. детектовање архипелага на основу непокривених регија детектованих у претходном кораку;

Коришћење IGV алата у процесу детектовања варијација у броју копија

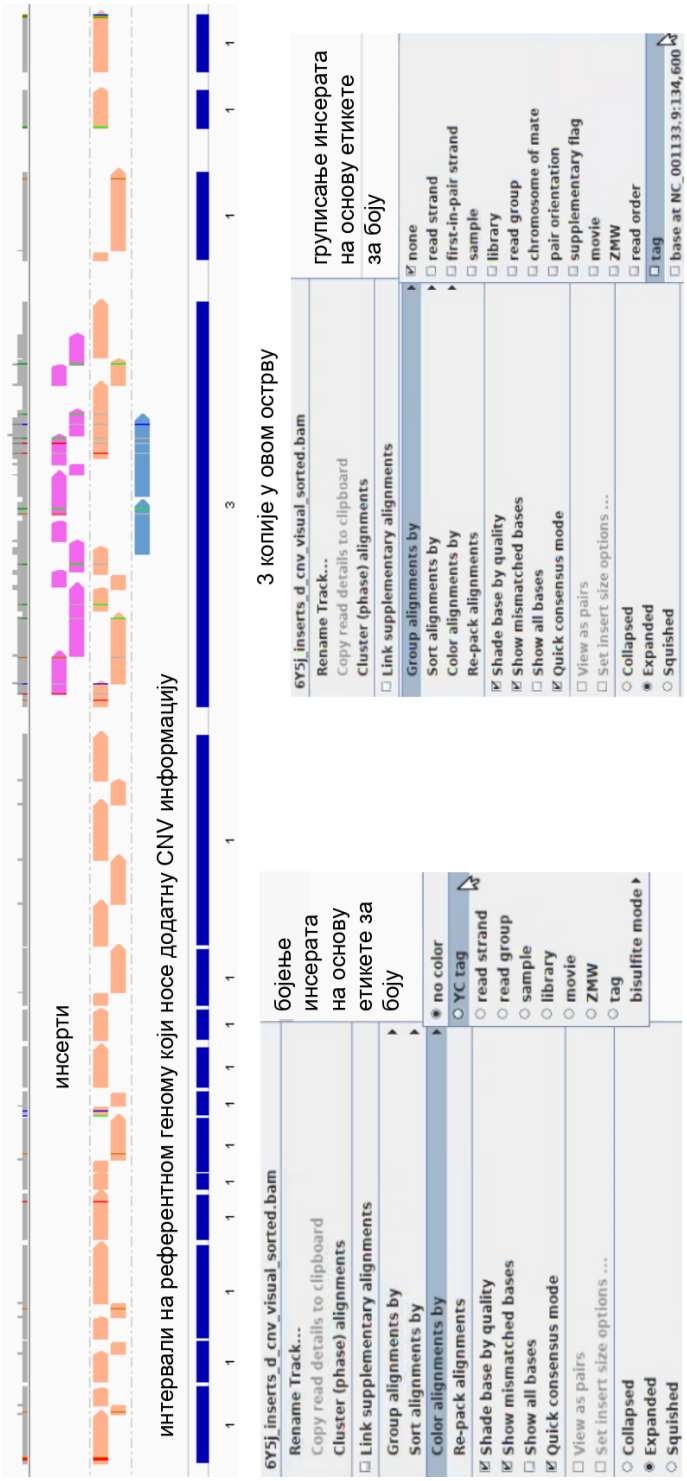
На слици 4.9 је приказан стандардни изглед IGV интерфејса приликом учитвања датотеке са поравнањима. Интерфејс алата приказује учитани геном у виду секвенце и обележене позиције за сваку базу. Поред позиције и дубине поравнања за сваки инсерт, могуће је приказати и податке о квалитету мапирања инсерата, CIGAR стрингу и баркоду.

Иако је етикета додељена сваком инсерту у *bam* датотеци, потребно је подесити одређене параметре у IGV-у, како би молекули били јасно видљиви, као што је приказано на слици 4.10. Прва акција приликом подешавања параметра за боју је клик десног миша у пределу поравнања инсерата. У падајућем менију се затим одабира опција *Color alignments by*, а затим и опција *tag*, након чега се уноси име етикете на основу које се врши бојење. Етикета која носи информацију о боји је *YC* (може постојати као већ понуђена опција у падајућем менију). Након наведених корака, IGV сваки инсерт боји бојом која му је додељена. IGV пружа још једну могућност - груписање инсерата на основу одређеног критеријума, што ће помоћи у коначној формулацији молекула. Ово се постиже десним кликом на област са поравнатим инсертима и кликом на опцију *Group alignments by*. Затим се кликне на опцију *tag*, након чега се уноси име етикете на основу које се врши груписање. У овом случају, то је поново етикета која носи информацију о боји. Резултат су обојени инсерти груписани у различите линије (енгл. *swimming line*).

У случају да неки од алата не подржава бојење инсерата, информацију о броју молекула је ипак могуће испратити. Број инсерата у сваком острву се бележи у посебну *bed* датотеку. Прве три колоне те датотеке носе информацију о регији на референтном геному за свако острво. Регије се у IGV-у приказују дугим плавим правоугаоницима, као што је приказано на слици 4.10. Четврта колона носи информацију о броју детектованих молекула у том острву и у IGV-у се приказује као број испод сваког правоугаоника. Према



Слика 4.9: Приказ IGV интерфејса приликом учитавања датотеке са поравнањима.



Слика 4.10: Подешавање видљивости боја за инсерте.

томе, не само да је могуће јасно разликовати копије и означити их различитим бојама, већ је могуће и забележити регије на референтном геному где долази до варијације у броју копија.

Глава 5

Провера тачности методе

Након генерисања резултата потребно је проверити их. Тестирање постојећих метода [20] за детектовање варијација у броју копија се може вршити на два начина. Први начин је бојење варијација флуоресцентним бојама у лабораторији, након чега се резултати методе пореде са обојеним варијацијама. Други начин је упоређивање са валидираним и објављеним скуповима података који садрже већ детектоване варијације у броју копија, попут скупа *ICR96 Exon Validation series* [30]. Оваква поређења нису била могућа, због недоступности података секвенцираних другим методама секвенцирања једне ћелије и због недоступности података о вештачки додатим варијацијама приликом припреме за секвенцирање. Стога је извршена алтернативна провера тачности методе која обухвата два нивоа:

1. Први ниво провере подразумева испитивање исправности кода, односно да ли је предложена метода коректно имплементирана;

2. Други ниво провере подразумева испитивање да ли је могуће детектовати варијације у броју копија применом ове методе, а ако је могуће, да ли детектоване копије одговарају вештачки додатим копијама у лабораторији;

5.1 Први ниво провере

Да би се потврдило да приказана имплементација даје очекиване резултате, извршено је тестирање засебно за сваку јединицу кода (енгл. *unit test*) за сваки модул посебно. Свака јединица је тестирана над вештачки креираним

подацима и над реалним подацима из генома узорака 6Y5b_S33, 6Y5d_S34, 6Y5j_S37, 6Y5l_S38, 6Y5n_S39, 6Y5p_S40, 6Y6b_S41, 6Y6d_S42, 6Y6f_S43 и 6Y6h_S44. Тестирање је извршавано тако што је покретан рачунарски код за сваку јединицу, а резултати тог покретања су упоређивани са очекиваним резултатима. Када су очекивани и добијени резултати исти, значи да је код јединице прошао тест пример и да расте степен поверења у његову тачност. У случају да код не пролази тест пример, на основу самог тест примера се лоцира грешка у коду, она се исправља, а затим се код покреће поново за све тест примере (и онај на којем је пао, али и они које је прошао).

Тест примери се генеришу за уобичајене случајеве које је могуће сусрести у раду са одређеним подацима. На пример, приликом реконструкције инсерата тест примери су парови очитавања који се не преклапају и парови очитавања који се преклапају. Такође, тестирају се ситуације у којима постоји неколико могућих прилика за доношење одлуке. На пример, приликом одсецања архипелага неопходно је тестирати сва четири случаја приказана на слици 3.6. Представљена имплементација је дала очекиване резултате за све тест примере који су јој прослеђени.

5.2 Други ниво провере

Други ниво провере подразумева испитивање да ли је могуће детектовати варијације у броју копија применом ове методе. Како би се ово испитало да ли је метода у стању да детектује варијације у броју копија, било је неопходно покренути методу над узорком који садржи варијације у броју копија и над узорком који не садржи варијације у броју копија.

За први случај, у лабораторији је припремљен посебан узорак генома квасца *6Y5j* (остали узорци су коришћени за тестирање рада транспозазе у лабораторији и за тестирање првог нивоа рачунарске имплементације), тако што су му додате вештачке копије његовог геномског материјала. Овако биолошки модификован узорак је секвенциран и добијени су резултати који представљају улаз у методу. На добијена очитавања примењен је најпре модул за реконструкцију инсерата а затим и модул за детектовање варијација у броју копија. Исправно детектовање варијација у броју копија би значило детектовање тачно оних варијација које су вештачки додаване у лабораторији пре самог секвенцирања. На пример, ако су за геномску секвенцу са контига

NC_001133.9 на интервалу 133023-136327 на геному узорка додате вештачке копије, исправно детектовање варијација у броју копија би значило да су варијације детековане баш на контигу NC_001133.9 на интервалу 133023-136327. Директно поређење овог типа није било могуће, јер лабораторија није доставила податке о томе где су тачно додате вештачке копије. Резултати методе су прослеђени лабораторији на испитивање исправности, где је утврђено да се резултати добијени методом поклапају са оригиналном поставком варијација у броју копија.

Како лабораторија није обезбедила узорак за који се сигурно зна да не садржи варијације у броју копија, било је неопходно вештачки креирати тај узорак. Стога је модул за детектовање варијација у броју копија примењен над инсертима добијеним реконструкцијом фрагмената референтног генома квасца. Разлика у односу на процес фрагментације који се врши за детектовање регија лошег мапирања је у томе што се не симулира рад транспозазе додавњем вештачких секвенци на крајеве фрагмената. Утврђено је да на местима на референтном геному на којима нема варијација у броју копија, ни сама метода за детектовање варијација у броју копија није детектовала варијације. На местима на референтном геному где су дубине мапирања различите од један, детектоване су варијације у броју копија. Већина тих регија се поклапа са регијама лошег мапирања. Испитивањем сваке овакве регије појединачно, дошло се до закључка да је регије лошег мапирања могуће прецизније детектовати покретањем модула детектовања броја копија над фрагментима референтног генома. Ова чињеница ће бити значајна приликом детаљног истраживања сваке од регија лошег мапирања и њихове везе са инсертима који имају вишеструка места поравнања, што је следећи корак у развоју методе за детектовање варијација у броју копија.

5.3 Поузданост методе

У овом истраживању није било могуће директно упоредити добијене резултате са тачним резултатима над реалним подацима због недоступности ових резултата од стране лабораторије. Ипак, можемо претпоставити да поузданост развијене методе произилази из следећих чињеница:

- Могућност праћења инсерата заснива се на чињеници да је свакоме од њих додељен баркод;

- Једнозначност баркода је последица исецања молекула ДНК на случајно одабраним местима.

- Равномерна амплификација, која се постиже довољним засићењем раствора транспозазе, омогућава довољно добру покривеност приликом мапирања, тако да је могуће детектовати додатне копије геномског материјала;

- Задовољавајућа покривеност генома се постиже реконструкцијом инсерата, захваљујући којој није потребно апроксимирати покривеност сваке од база које припадају регијама између два поравната очитавања;

- Пошто је амплификација равномерна и пошто је показано да су дупликати последица процеса амплификације, долази се до закључка да је за два поравната инсерта могуће да се преклапају на месту поравнања на референтном геному једино ако припадају различитим варијацијама;

Из наведеног произилази неколико закључака о условима које је неопходно испунити како би метода била успешна:

- Метода је тренутно примењива само на податке добијене псеудотаргетираним секвенцирањем и то само једне ћелије;

- Раствор транспозазе мора бити засићен како би се постигли довољно кратки и уједначени фрагменти, с обзиром на то да је PCR амплификација осетљива на дужину фрагмената;

С обзиром на то да је већина дектованих архипелага имала покривеност већу од 70% и има сличне особине као и сам референтни геном, очекује се да ће метода бити успешна и приликом секвенцирања целог генома.

Глава 6

Закључак

У данашње време су варијације у броју копија геномских секвенци ћелије у фокусу великог броја научних истраживања, посебно у контексту микрооколине појединачних ћелија. Постоје различите методе секвенцирања једне ћелије и различите методе детектовања варијација у броју копија у оквиру једне ћелије. Међутим, у истраживањима која примењују различите комбинације постојећих метода секвенцирања и метода детектовања варијација у броју копија није потврђено да је могуће прецизно детектовати узастопне копије делова геномске секвенце, нити је могуће утврдити порекло сваке од тих копија. Научници из лабораторије *Digenomix* су развили нову методу секвенцирања једне ћелије - баркодирање генома, за коју се верује да ће у свом коначном облику бити у стању да превазиђе друге методе секвенцирања једне ћелије и надомести њихове недостатке. Иновација ове методе је генерисање јединственог идентификатора - баркода, захваљујући којем је могуће равномерно умножити сваки од фрагмената и у великој мери смањити грешке приликом секвенцирања. Како би та метода била комплетна и потврђена, неопходно је извршити биоинформатичку анализу резултата секвенцирања и развити рачунарски приступ за детектовање варијација у броју копија. Истраживање о развоју методе детектовања варијација у броју копија је рађено у оквиру компаније *Digenomix*.

Варијација у броју копија је тип структурне варијације (варијација у структури хромозома), посматран код јединки исте врсте, код којих неки сегменти ДНК могу бити дупликати (или трипликати, или мултипликати). Број копија се може разликовати од јединке до јединке, а може укључивати и гене или више гена. Научници су временом показали да је варијација у

броју копија прилично честа појава, као и то да ове варијације могу захватити функционалне или регулаторне гене, што често резултује различитим болестима.

Сазнања о варијацијама у броју копија доприносе различитим биомедицинским истраживањима о: настанку и преносу ретких генетских обољења на потомство [31], узроку различитих хромозомалних поремећаја [32], експресији гена [33] и разликовању типова и подтипова канцера [34]. Према томе, резултати развоја методе детектовања варијација у броју копија применљиви су у различитим научним истраживањима и дијагностичким поступцима.

Како је сама метода баркодирања генома још увек у развоју, тако је и имплементацију рачунарске методе која је прати неопходно још усавршавати у складу са новим открићима. У даљем раду би било неопходно усавршити процес тримовања очитавања и истражити како индели утичу на детектовање варијација у броју копија. Такође, било би значајно испитати и утицај инсерата са вишеструким местима мапирања на детектовање варијација у броју копија.

Библиографија

[1] Digenomix, "6Y5j rezultati." Online at: `https://github.com/Nacili/MASTER`.

[2] L. Xi, A. Belyaev, S. Spurgeon, X. Wang, H. Gong, R. Aboukhalil, and R. Fekete, "New library construction method for single-cell genomes," *PLOS ONE*, vol. 12, pp. 1–14, 07 2017.

[3] GeneQuantification, "Gene Quantification." Online at: `http://cnv.gene-quantification.info/`.

[4] B. T. K. J. E James, S Jai-Yoon, "The promise of single-cell sequencing," *Nature Methods*, vol. 11, pp. 25–27, 2014.

[5] B. N. Olsen TK, "Introduction to single-cell rna sequencing," *Current Protocols In Mollecular Biology*, vol. 122, no. 1, p. e57, 2018.

[6] G. Macintyre, P. Van Loo, N. M. Corcoran, D. C. Wedge, F. Markowetz, and C. M. Hovens, "How subclonal modeling is changing the metastatic paradigm," *Clinical Cancer Research*, vol. 23, no. 3, pp. 630–635, 2017.

[7] BioSistemika, "qPCR, Microarrays or RNA-sequencing: When To Choose One Over the Other?." Online at: `https://biosistemika.com/blog/qpcr-microarrays-rna-sequencing-choose-one/`.

[8] M. Chen, P. Song, D. Zou, X. Hu, S. Zhao, S. Gao, and F. Ling, "Comparison of multiple displacement amplification (mda) and multiple annealing and looping-based amplification cycles (malbac) in single-cell sequencing," *PLOS ONE*, vol. 9, pp. 1–12, 12 2014.

[9] C. Chen, D. Xing, L. Tan, H. Li, G. Zhou, L. Huang, and X. S. Xie, "Single-cell whole-genome analyses by linear amplification via transposon insertion (lianti)," *Science*, vol. 356, no. 6334, pp. 189–194, 2017.

[10] M. J. Levin HL, "Dynamic interactions between transposable elements and their hosts," *Nature Reviews Genetics*, vol. 12, no. 9, pp. 615–627, 2011.

[11] C. A. Chu CC, "A 10- rather than 9-bp duplication associated with insertion of Tn5 in Escherichia coli K-12. Plasmid," *PubMed*, vol. 22, no. 3, p. e0181163, 1989.

[12] H. Willems, "Adaptor pcr for the specific amplification of unknown dna fragments," *BioTechniques*, vol. 24, no. 1, pp. 26–28, 1998. PMID: 9454945.

[13] Illumina, "Nextera XT LibraryPrep: Tipsand Troubleshooting." Online at: `file:///tmp/mozilla_naca0/nextera-xt-troubleshooting-technical-note.pdf`.

[14] A. e. a. Adey A, Morrison H.G., "Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.," *Genome Biology*, vol. 11, no. 119, 2010.

[15] Illumina, "llumina MiSeq." Online at: `http://www.illumina.com/systems/sequencing-platforms/miseq.html`.

[16] J. H. University, "Samtools." Online at: `http://bowtie-bio.sourceforge.net/bowtie2/index.shtml`.

[17] J. H. University, "CIGAR Strings For Dummies." Online at: `https://jef.works/blog/2017/03/28/CIGAR-strings-for-dummies/`.

[18] S. A. B Pereira, O Rueda, "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes," *Briefings in Bioinformatics*, vol. 7, no. 1, 2016.

[19] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, pp. 178–192, 04 2012.

[20] N. N. Mallory XF, Edrisi M, " Methods for copy number aberration detection from single-cell DNA-sequencing data," *Genome Biology*, vol. 21, no. 208, 2020.

[21] I. N. Sims D, Sudbery I, " Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, pp. 121–132, 2014.

[22] U. of Mitchigan, "FASTA format." Online at: `https://zhanglab.ccmb.med.umich.edu/FASTA/`.

[23] Illumina, "FASTQ files explained." Online at: `https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html`.

[24] Metagenomics, "SAM file format." Online at: `http://www.metagenomics.wiki/tools/samtools/bam-sam-file-format`.

[25] Ensembl, "BED file format." Online at: `https://m.ensembl.org/info/website/upload/bed.html`.

[26] G. R. Limited, "Samtools." Online at: `https://github.com/samtools/samtools`.

[27] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, pp. 841–842, 01 2010.

[28] A. Heger, "pysam - An interface for reading and writing SAM files." Online at: `https://pysam.readthedocs.io/en/latest/api.html`.

[29] M. Shirley, "Welcome to pyfaidx's documentation!." Online at: `https://pythonhosted.org/pyfaidx/`.

[30] Y. S. e. a. Mahamdallie S, Ruark E, "The icr96 exon cnv validation series: a resource for orthogonal assessment of exon cnv calling in ngs data.," *Wellcome Open Research*, vol. 2, no. 35.

[31] R. V. e. a. SGross A.M., Ajay S.S., "Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease," *Genetics in Medicine*, vol. 21, p. 1121–1130, 2019.

[32] R. R. Le Caignec, C., "Copy number variation goes clinical.," *Genome Biology*, vol. 10, p. 301, 2009.

[33] L. J. e. a. Shao X., Lv N., "Copy number variation is highly correlated with differential gene expression: a pan-cancer study.," *BMC Medical Genetics*, vol. 20, p. 175, 2019.

[34] C. Park, K.-A. Yoon, J. Kim, I. H. Park, S. J. Park, M. K. Kim, W. Jang, S. Y. Cho, B. Park, S.-Y. Kong, and E. S. Lee, "Integrative molecular profiling

identifies a novel cluster of estrogen receptor-positive breast cancer in very young women," *Cancer Science*, vol. 110, no. 5, pp. 1760–1770, 2019.

Биографија аутора

Надежда Богдановић, рођена 1995. године у Београду. Дипломирала 2018. године као студент Математичког факултета Универзитета у Београду. Тренутно запослена у *APIS Assay Technologies Ltd.* молекуларном дијагностичком центру као Биоинформатички Инжењер.