Procjena DNA ROH segmenata iz simuliranih podataka putem HMM i klasičnih programskih paketa

Strelar, Kristina

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Agriculture / Sveučilište u Zagrebu, Agronomski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:204:413740

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-19**

Repository / Repozitorij:

Repository Faculty of Agriculture University of Zagreb

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

PROCJENA DNA ROH SEGMENATA IZ SIMULIRANIH PODATAKA PUTEM HMM I KLASIČNIH PROGRAMSKIH PAKETA

DIPLOMSKI RAD

Kristina Strelar

Zagreb, rujan, 2020.

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

Diplomski studij:

Genetika i oplemenjivanje životinja

PROCJENA DNA ROH SEGMENATA IZ SIMULIRANIH PODATAKA PUTEM HMM I KLASIČNIH PROGRAMSKIH PAKETA

DIPLOMSKI RAD

Kristina Strelar

Mentor:

Doc. dr. sc. Maja Ferenčaković

Zagreb, rujan, 2020.

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

IZJAVA STUDENTA

O AKADEMSKOJ ČESTITOSTI

Ja, **Kristina Strelar**, JMBAG 0178099939, rođena 21.5.1995. u Zagrebu, izjavljujem da sam

samostalno izradila diplomski rad pod naslovom:

PROCJENA DNA ROH SEGMENATA IZ SIMULIRANIH PODATAKA PUTEM HMM I KLASIČNIH PROGRAMSKIH PAKETA

Svojim potpisom jamčim:

- da sam jedina autorica ovoga diplomskog rada;
- da su svi korišteni izvori literature, kako objavljeni tako i neobjavljeni, adekvatno citirani ili parafrazirani, te popisani u literaturi na kraju rada;
- da ovaj diplomski rad ne sadrži dijelove radova predanih na Agronomskom fakultetu ili drugim ustanovama visokog obrazovanja radi završetka sveučilišnog ili stručnog studija;
- da je elektronička verzija ovoga diplomskog rada identična tiskanoj koju je odobrio mentor;
- da sam upoznata s odredbama Etičkog kodeksa Sveučilišta u Zagrebu (Čl. 19).

U Zagrebu, dana _____ _____

Potpis studenta / studentice

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

IZVJEŠĆE

O OCJENI I OBRANI DIPLOMSKOG RADA

Diplomski rad studentice **Kristine Strelar**, JMBAG 0178099939, naslova

PROCJENA DNA ROH SEGMENATA IZ SIMULIRANIH PODATAKA PUTEM HMM I KLASIČNIH PROGRAMSKIH PAKETA

obranjen je i ocijenjen ocjenom _____ , dana _____ .

Povjerenstvo: potpisi:

1. Doc. dr.sc. Maja Ferenčaković _____
2. Prof. dr. sc. Ino Čurik _____
3. Izv. Prof. dr. sc. Vlatka Čubrić Čurik _____

Zahvala

Ovime zahvaljujem svojoj mentorici, doc. dr. sc. Maja Ferenčaković, na podršci i savjetima tijekom pisanja diplomskog rada. Iznimno sam zahvalna na neumornim odgovaranjima na moja pitanja kao i pomoć u pronalaženju grešaka te ispravljanju istih.

Nadalje, želim se zahvaliti mojim roditeljima Olgici i Mariu te sestri Mateji. Vi ste mi bili podrška ne samo tijekom studija nego tijekom cijelog života. Hvala vam što ste uvijek imali razumijevanja i vjeru u mene. Hvala vam što ste mi pružili mogućnost da bezbrižno studiram i uvijek me gurali da zakoračim stepenicu više.

Veliko hvala i mojim prijateljicama. Marija, hvala ti na zajedničkom učenju i motiviranju. Neke stvari ne bih uspjela da nije bilo tebe. Marija Katarina, hvala ti što si, iako na drugom smjeru, uvijek bila tu kao velika potpora. Bez vas moje studiranje ne bi bilo isto. Ivona, Tajana i Anamarija, hvala vam što ste imale strpljenja i što ste me motivirale, ne samo riječima, već i vlastitim primjerom.

I za kraj, Viktore, hvala ti na nesebičnoj pomoći, trudu i vjeri u mene. Hvala ti što si me stalno podsjećao koliko je i najmanji uspjeh važan. Uz tvoju su podršku i najveći izazovi bili mogući.

Sadržaj

1. Uvod 1

1.1. Cilj istraživanja 1

2. Pregled literature 2

2.1. Runs of Homozigosity 2

2.1.1. Povijesni pregled i primjena 2

2.1.2. Mehanizmi pojavljivanja ROH segmenata u genomu 3

2.1.3. Identificiranje ROH-a i podaci 4

2.2. Programski paketi za obradu podataka 6

2.2.1. AphaSimR 6

2.2.2. PLINK 7

2.2.3. cgaTOH 8

2.2.4. RZooRoH 9

3. Materijali i metode 12

3.1. Simulacija podataka 12

3.2. R paket – RZooRoH 13

3.3. Detekcija ROH segmenata 13

3.4. Statistička obrada i vizualizacija 14

4. Rezultati i rasprava 15

5. Zaključci 30

6. Popis literature 31

7. Životopis 34

8. Prilog 35

8.1. AlphaSimR 35

8.2. PLINK 40

8.3. RZooRoH 40

8.4. cgaTOH 41

Sažetak

Diplomskog rada studentice **Kristine Strelar**, naslova

PROCJENA DNA ROH SEGMENATA IZ SIMULIRANIH PODATAKA PUTEM HMM I KLASIČNIH PROGRAMSKIH PAKETA

Runs of homozygosity (ROH) javljaju se u genomu kao dugi neprekinuti segmenti homozigotnih genotipova koji omogućuju pouzdanu procjenu razine inbreedinga. Danas imaju široku primjenu, pomoću njih otkrivamo i učinke autozigotnosti, no još uvijek postoje polemike oko toga koji je način njihove detekcije optimalan. Cilj ovog rada bio je usporedba triju programskih paketa s različitim algoritmom za detekciju ROH segmenata i to su cgaTOH, PLINK i RZooROH. U tu svrhu je putem programskog paketa AlphaSimR simulirano 60 populacija s različitim efektivnim veličinama (Ne=25 i Ne=300) kroz 50 generacija te su na njima detektirani ROH segmenti sa svakim programskim paketom, a rezultati uspoređeni Tukey-Kramer HSD testom. Rezultati usporedbe programskih paketa pokazuju statistički značajnu razliku ($p < 0,01$) srednjih vrijednosti malih i srednjih segmenata populacija s Ne = 25 dobivenih programskim paketom RZooROH naspram cgaTOH i PLINK dok se oni međusobno ne razlikuju. Kod populacija s Ne = 300 statistički se značajno razlikuju sva tri programska paketa ($p < 0,01$) kod malih segmenata dok kod velikih nema razlike ili je ona prisutna kod RZooROH. Svojim algoritmom RZooROH puno restriktivnije određuje ROH segmente i proglašava ih identičnima po porijeklu, pogotovo kod segmenata veličine do 8MB. Kod velikih segmenata i manje očekivane autozigotnosti te su razlike manje. Grafičkim prikaz statusa homozigotnosti simuliranih genoma naspram procjene putem programskih paketa to također potvrđuje. Svakako su potrebna daljnja istraživanja koja bi uključivala simulaciju segregacije segmenata kroz generacije i potom sposobnost detekcije tih segmenata programskim paketima.

Ključne riječi: ROH, cgaTOH, PLINK, RZooROH, simulacija.

Summary

Of the master's thesis – student **Kristina Strelar,** entitled

EVALUATION OF DNA ROH SEGMENTS FROM SIMULATED DATA THROUGH HMM AND CLASSIC SOFTWARE PACKAGES

Runs of homozygosity (ROH) occur in the genome as long continuous segments of homozygous genotypes that allow a reliable assessment of inbreeding levels. Today, they are widely used, and we can use them to detect the effects of autozygosity, although there is still controversy about which method of their detection is optimal. The goal of this paper was to compare three software packages with different algorithms for detection of ROH segments, namely cgaTOH, PLINK and RZooROH. For this purpose, 60 populations with different effective sizes (Ne=25 and Ne=300) were simulated through the AlphaSimR software package over 50 generations, and ROH segments with each software package were detected on them, and the results were compared with the Tukey-Kramer HSD test. Comparison results of software packages show a statistically significant difference ($p < 0.01$) of the mean values of small and medium segments of populations with Ne=25 obtained by the software package RZooROH versus cgaTOH and PLINK until they differ from each other. In populations with Ne=300, all three software packages are statistically significantly different ($p < 0.01$) in small segments, while in large segments there is no difference, or it is present in RZooROH. With its algorithm, RZooROH determines ROH segments much more restrictively and declares them identical in origin, especially for segments up to 8MB in size. For large segments and less expected autozygosity, these differences are smaller. A graphical representation of the homozygosity status of the simulated genomes versus estimation via software packages also confirms this. Further research is certainly needed that would include simulating the segregation of segments across generations and then the ability to detect those segments by software packages.

Keywords: ROH, cgaTOH, PLINK, RZooROH, simulation.

1. Uvod

Pojava dugih neprekinutih segmenata homozigotnih genotipova (ROH segmenta) u genomu jedinke može se objasniti nasljeđivanjem istog kromosomskog segmenta od oba roditelja, koji su naslijedili taj specifičan segment od zajedničkog pretka. ROH segmente prekidaju rekombinacijski događaji te se putem njihove duljine lako može procijeniti vrijeme do zajedničkog pretka. Pri istraživanjima na ljudskom genomu ROH segmenti često su se koristili za IBD mapiranje, a njihova pojava se, također, povezuje s povećanim rizikom od Alzeheimer-ove bolesti, šizofrenije, autizma, mentalne zaostalosti, raka štitnjače i dojke. Često se koriste i za objašnjavanje demografske povijesti jer se njihova dužina i frekvencija mogu povezati s demografskim procesima.

Kod životinja se ROH segmenti najčešće koriste za procjenu koeficijenta inbreedinga i otkrivanje učinka autozigotnosti. Obzirom da se inbreeding koeficijenti dobiveni putem ROH segmenata (F_{ROH}) smatraju preciznijima od onih procijenjenih iz rodovnika, tako se koriste i kod procjene inbreeding depresije.

Porastom interesa za upotrebu ROH segmenata razvijen je i čitav niz metoda za njihovu detekciju. Takve metode se uglavnom dijele na metode prebrojavanja i na one koje koriste skriveni Markovljev model (HMM). HMM je statistički model koji se može koristiti za opisivanje evolucije vidljivih događaja, te ovisi o internim čimbenicima koje nije moguće izravno promatrati. Preciznost i pouzdanost detekcije ROH segmenata utječu na pouzdanost njihove upotrebe te je dodatan naglasak na pronalaženje najpreciznije metode i programa za njihovu detekciju.

1.1. Cilj istraživanja

Cilj ovog istraživanja je usporedba programskih paketa za detekciju ROH segmenata. Prvi programski paket, RZooRoH, koristi HMM dok drugi, „cgaTOH“, koristi metodu prebrojavanja. Programski paket PLINK koristi pristup „kliznog prozora“ s ciljem definiranja ROH-a kao dužine. Uz samu usporedbu programa napravit će se kratak osvrt na jednostavnost korištenja svih programskih paketa pri dobivanju i izračunu rezultata istraživanja.

2. Pregled literature

2.1. Runs of Homozigosity

2.1.1. Povijesni pregled i primjena

Jedinka je homozigotna za određeni marker ukoliko su oba alela na istom markeru identična. U genomu jedinke se mogu naći homozigotne regije u kojima nema opažene heterozigotnosti. Dugačke homozigotne segmente prvi su uočili Broman i Weber (1999.). Njihova je pretpostavka bila da su pitanju segmenti koji su mogući odraz autozigotnosti te da mogu imati utjecaj na ljudsko zdravlje. Gibson i sur. (2006.) nastavljaju analizirati homozigotne segmente te se osvrću na njihovu duljinu, frekvenciju i distribuciju u HapMap populacijama. Definirali su ih kao neprekidne i kontinuirane segmente DNA sekvence bez heterozigotnosti u diploidnom stanju. Autozigotna je priroda spomenutih segmenata bila pretpostavljena zbog rekombinacijskih događaja koji prekidaju dugačke kromosomske segmente.

Nasljeđivanjem identičnih haplotipova od zajedničkog pretka stvaraju se dugi segmenti homozigotnih genotipova poznatih kao „runs of homozygosity“ (ROH). Nastajanje kratkih segmenata moguće je od daljnjih predaka, ali uz njih se mogu pojaviti i segmenti koji nisu identični po porijeklu (engl. Identical by Descent - IBD). Lencz i sur. (2007.) potvrđuju pretpostavku Bromana i Webera (1999.) ukazujući da se homozigotni segment može sustavno upotrebljavati za mapiranje gena povezanih s bolestima kao što je shizofrenija, sam pojam „Runs of Homozygosity, ROH" također se veže uz njih

Howrigan i sur. (2011.) zaključuju da broj i veličina ROH segmenata identificiranih u genotipskim podacima mogu uvelike ovisiti o specifičnim parametrima i pragovima zadanim tijekom SNP analize. Čišćenje SNP-ova koji pokazuju niske frekvencije minor alela (engl. Minor allele frequency - MAF), onih koji bi mogli odstupati od Hardy Weinberg ekvilibrijuma (HWE), također mogu pokazivati visoki disekvilibrijum vezanosti gena (engl. Linkage Disequilibrium - LD) (Wigginton i sur., 2005.; Albrechsten i sur. 2010.).

McQuillan i sur. (2008.) predstavili su FROH kao genomsku mjeru, odnosno, udio autozonog genoma koji leži u ROH-u određene minimalne duljine u odnosu na ukupni genom u području od interesa. ROH segmenti su se počeli široko prihvaćati u istraživanjima na domaćim životinjama. Sölkner i sur. (2010.) i Ferenčaković i sur. (2011.) su prvi počeli primjenjivati ROH koncept za stoku, a njih su slijedili Purfield i sur. (2012.) i Ferenčaković i sur. (2013 a, b.). Glavni cilj ovih istraživanja bio je usporedba FROH-a i FPED-a s obzirom na dužine ROH-a, dubinu i kvalitetu pedigrea, izračuna algoritama i gustoće markera. Ovi radovi dokazuju da je FROH bolji procjenitelj individualne autozigotnosti nego koeficijent inbreedinga dobiven rodovnikom FPED.

Bjelland i sur. (2013.) te Čurik i sur. (2012.) nastavili su s istraživanjem nadovezujući se na spomenuti rad te počinju primjenjivati ROH koncept za procjenu inbreeding depresije kod goveda i Silio i sur. (2013.) kod svinja. Kim i sur. (2013) su usporedili tri skupine životinja čiji su preci bili izloženi različitim silama selekcije kako bi optimizirali proizvodnju mlijeka Holstein goveda.

Kada se upotrebljavaju podaci sa SNP chip-ova, poznato je da su takvi markeri polimorfni u vrsti od interesa (goveda, ovce, koze, ljudi i dr.). U asocijacijskim istraživanjima cijelog genoma (engl. Genome-Wide association studies - GWAS) izbacivanjem SNP-ova, koji su monomorfni u promatranoj populaciji, dolazi do smanjenja vremena potrebnog za izračunavanje na neinformativnim markerima (neinformativni su jer nema varijance). U ROH analizama izbacivanje fiksiranih SNP-ova predstavlja gubitak informacija. Fiksacija je posljedica inbreedinga i male efektivne veličine populacije. U slučaju autozigotnosti fiksirani SNP-ovi trebaju ostati u podacima dajući realniju sliku stanja populacije od interesa. Isto vrijedi i za LD čišćenje i vezane alele koji proizvode kratke homozigotne segmente.

U većini ROH istraživanja metode za kontrolu kvalitete kopiraju se iz „GWAS" analize u kojima povezani aleli uzrokuju produljeno vrijeme potrebno za izračunavanje te ne pridonose preciznijim i kvalitetnijim rezultatima. U ROH analizama ne uzrokuju toliko problema jer se smatraju lokalnim fenomenom koji proizvodi kratke homozigotne segmente. U slučaju prisutnosti velikih ROH segmenata zbog LD-a, oni su po svojoj prirodi i dalje autozigotni te nema potrebe za njihovim isključivanjem. Devijacija od HWE (eng. Hardy-Weinberg equilibrium) se pojavljuje u populaciji koja je pod utjecajem selekcije. Kada je riječ o domaćim životinjama, većina je bila ili je još uvijek pod velikom selekcijom, stoga su devijacije očekivane (Čurik i sur. 2014.).

Purfield i sur. (2012.) te Ferenčaković i sur. (2013.) su usporedili procjene koje su dobili upotrebljavajući dva SNP-čipa koji se najviše koriste na govedima: „Ilumina BovineSNP50 Genotyping BeadChip" sa 54 001 SNP-ova(50K) i „Illumina BovideHD Genotyping BeadChip" sa 777 972 SNP-ova (HD – High Density). Zaključili su da je 50k čip prikladan samo za identificiranje ROH segmenata dužih od 4Mb (Ferenčaković i sur. 2013.), odnosno 5 Mb (Purfield i sur., 2012.). Analize bazirane na čipovima manje gustoće mogu biti kompromitirane zbog neotkrivenih heterozigotnih SNP genotipova prisutnih u promatranom ROH-u (Ferenčaković i sur. 2013.).

Frekvencije pogrešaka pri genotipiziranju SNP-ova još su jedan faktor koji može utjecati na ROH bazirane procjene autozigotnosti. Identifikacija dugačkih segmenata ROH-a, koji sadrži brojne SNP-ove, može biti pod utjecajem variranja frekvencije pogreške između 0.1% i 0.2%. Svaka pogreška pri genotipiziranju, homozigot u heterozigota ili obrnuto, može utjecati na određivanje ROH segmenata. Potencijalno rješenje je dopustiti određenom broju SNP-ova da budu heterozigoti (Ferenčaković i sur. 2013.).

2.1.2. Mehanizmi pojavljivanja ROH segmenata u genomu

ROH su neprekinuti sljedovi homozigotnih genotipova prisutni u jedinci jer roditelji prenose identične haplotipove na svoje potomstvo. Njihov opseg i učestalost mogu upućivati na porijeklo pojedinca i njegovu populaciju. Što su duži segmenti, to je vjerojatnije da se u rodovniku dogodio nedavni inbreeding (Purfield i sur. 2012.). Distribucija kraćeg ROH-a također može biti informacija o postojanju drevne povezanosti u pedigreu. Pojava ROH segmenata u genomu jedinke može se široko objasniti nasljeđivanjem istog kromosomalnog segmenta od oba roditelja koji su naslijedili taj specifičan segment od zajedničkog pretka (Purfield. i sur. 2012.).

Frekvencije ROH segmenata variraju unutar i između kromosoma, s kromosomima koji pokazuju ROH „hot spot-ove“ odnosno "otoke", kao i „cold spot-ove“ odnosno "pustinje". Razlozi za ovu varijaciju su nejasni i privlače sve veći interes. Sekvenciranje sljedeće generacije može poboljšati naše razumijevanje ROH segmenata i njihovu korisnost kao alat u istraživanju inbreedinga (Čurik i sur. 2014.).

Gibson i sur. (2006.) su predložili dva glavna mehanizma pojave homozigotnih segmenata u genomu potomaka. Prvi govori da roditelji imaju relativno bliskog zajedničkog pretka, stoga je bilo jako malo prilika za rekombinaciju kako bi se segment preinačio. Drugi mehanizam govori da je veza između roditelja daleka, u segmentu nedostaje rekombinacije zbog visokog LD-a, stoga je segment ostao netaknut. Važno je istaknuti da su oba mehanizma smatrana funkcijom srodnosti te su posljedica parenja u srodstvu većeg (nedavnog) ili manjeg (davnog) stupnja. Homozigotni segmenti su rezultat LD-a te su prozvani autozigotnima jer su autozigotni po prirodi.

Broman i Weber (1999.) te Gibson i sur. (2006.) prepoznali su da će LD, kao lokalni fenomen, prouzročiti isključivo kratke homozigotne segmente a oni koji su duži od 1 Mbne mogu biti objašnjeni sa spomenutim mehanizmom. Opseg homozigotnosti u genomu može biti rezultat raznih vrsta jednoroditeljske disomnije (engl. uniparental dysomia - UPD), heterozigotnih delecija i selekcije (Gibson i sur. 2006.). Slučaj disomnije, u kojoj potomak nasljeđuje dvije kopije istog kromosoma (isodisomnija) ili dijelove kromosoma (segmentna isodisomija) od istog roditelja uzrokuje homozigotnost potomaka na svim lokusima kromosoma ili kromosomalnim segmentima.

2.1.3. Identificiranje ROH-a i podaci

Točna identifikacija ROH segmenata ovisi o kontroli nekoliko čimbenika kao što su kvaliteta genotipizacije, minimalna veličina ROH-a i broj dozvoljenih heterozigota, koji mogu ugroziti procjene zbog mogućih pogrešaka u genotipizaciji (Ferenčaković i sur. 2013.). Vrsta čipa koji se koristi za dobivanje podataka također utječe na identifikaciju ROH-a jer široka pokrivenost genoma omogućava identifikaciju većeg broja segmenata istog. Treba napomenuti da su čipovi s gustoćom većom od 50 000 SNP-ova potrebni za precizno otkrivanje ROH-ova manjih od 5,0 Mb (Purfield i sur.2012.; Zhang i sur. 2015.).

Trenutno postoje čipovi visoke gustoće primijenjeni u genotipizaciji za nekoliko proizvodnih vrsta koji su omogućili istraživanja sa svrhom utvrđivanja ROH-a već spomenutih vrsta (Čurik i sur. 2014.). Dobiveni rezultati pokazali su potencijal ovog pristupa u prepoznavanju genskih regija od interesa.

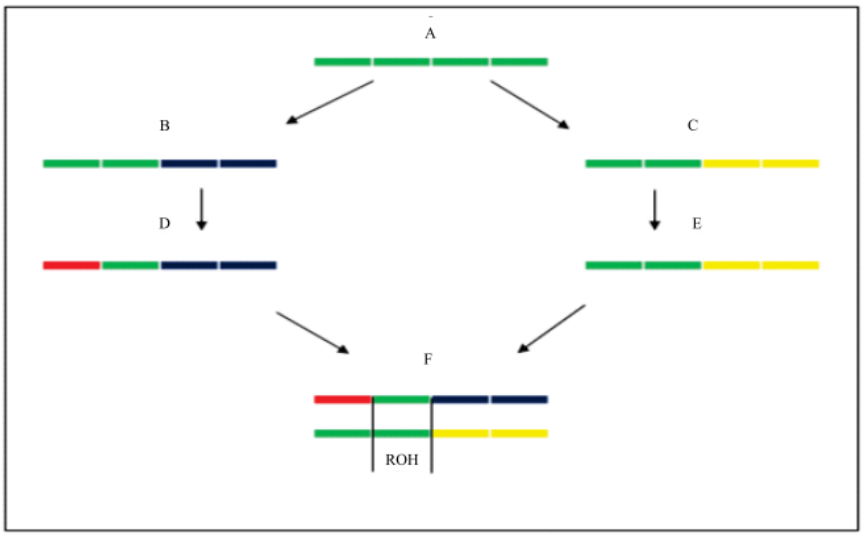
Faktori koji utječu na kvalitetu ROH detekcije su također i gustoća markera, njihova distribucija preko genoma, kvaliteta, stopa pozivanja/grešaka i frekvencija minor alela genotipa. Danas ROH istraživanja najviše upotrebljavaju podatke dobivene putem SNP čipova zbog pristupačnosti podataka i činjenice da su zlatni standard s vrlo niskim greškama kada se gleda stopa grešaka kod pozivanja genotipa (tipično <0.001).

Očekuje se da će dugački ROH segmenti zadržati svoje homozigotno stanje neovisno od SNP pokrivenosti. Ipak, relativna oskudnost SNP-ova na čipovima može značiti da pravi

heterozigotni SNP-ovi između markera mogu biti izostavljeni, pa bi se tako dva bliska ROH segmenata mogla prikazati kao jedan dulji ROH segment (Caballos i sur. 2018.).

Danas su poznate dvije glavne metode identificiranja ROH segmenata: metoda observacijskih algoritama za brojanje genotipova i algoritama na bazi modela. Observacijski algoritam skenira svaki kromosom tako da miče „prozor" fiksirane veličine preko cijele duljine genoma u svrhu pronalaska segmenata kontinuiranih SNP-ova. Ovaj pristup je implementiran u program PLINK v1.07 (Purcell i sur. 2007.) gdje se dani SNP smatra potencijalnim ROH segmentom te izračunava proporciju kompletno homozigotnog „prozora" koji obuhvaća taj SNP. Ukoliko je ova proporcija veća od definiranog praga SNP se smatra dijelom ROH segmenta.

U algoritmu, broj varijable nedostajućih heterozigotnih pozicija ili SNP-ova, mogu se specificirati po „prozoru" kako bi se tolerirale genotipske greške. ROH je detektiran ako je broj kontinuiranih SNP-ova u homozigotnom segmentu prešao zadani prag u smislu SNP broja i/ili pokrivene kromosomalne duljine. Algoritmi koji traže odgovarajuće haplotipove (npr. Germline (Gusev i sur. 2009.)) također se mogu upotrebljavati za izračun IBD-a kako bi identificirali ROH segmente kao poseban slučaj IBD-a unutar jedinke. Pristup s modelom upotrebljava skriveni Markovljev model (engl. Hidden Markov Model -HMM) kako bi se u obzir uzele pozadinske razine LD-a (Browning i sur. 2007.). HMM metode utvrđuju vjerojatnost da je određeni segment identičan po porijeklu (IBD), a zatim procjenjuju starost inbreedinga. Karakterizacija ROH segmenata u različitim populacijama, pasminama ili linijama važna je za dobivanje podataka o evolucijskoj povijesti (Metzger i sur. 2015; Sorbolini i sur. 2015.; Zavarez i sur. 2015), demografiji (Bosse i sur. 2012.), ili povezana sa srodnošću između stanovništva (Marras i sur. 2015.).



Slika 2.1.2.1. Prikaz stvaranja homozigotnih segmenata (ROH-a). Pojedinačni F predstavlja formirani ROH (prikazan zelenom bojom) koji je nastao uparivanjem homozigotnih segmenata homolognog kromosoma zajedničkog pretka A.
Izvor: Basso Robelaro A. i Caetano A. R. 2018.

2.2. Programski paketi za obradu podataka

2.2.1. AphaSimR

AlphaSimR je programski paket za simulaciju programa uzgoja biljaka i životinja. Omogućuje simulaciju više aspekata uzgojnih programa s visokim stupnjem fleksibilnosti. Koristi se za stohastičke simulacije uzgojnih programa do razine DNA sekvence za svakog pojedinca (Faux i sur. 2016.). Sadrži širok spektar funkcija za modeliranje uobičajenih zadataka u uzgojnom programu kao što su selekcija i križanje. Ove funkcije omogućuju izgradnju vrlo složenih uzgojnih programa biljaka i životinja putem skripti u programu R. Također, simulacije mogu koristiti za ocjenu ukupnog učinka i provođenje istraživanja dizajna uzgojnih programa (Faux i sur. 2016.).

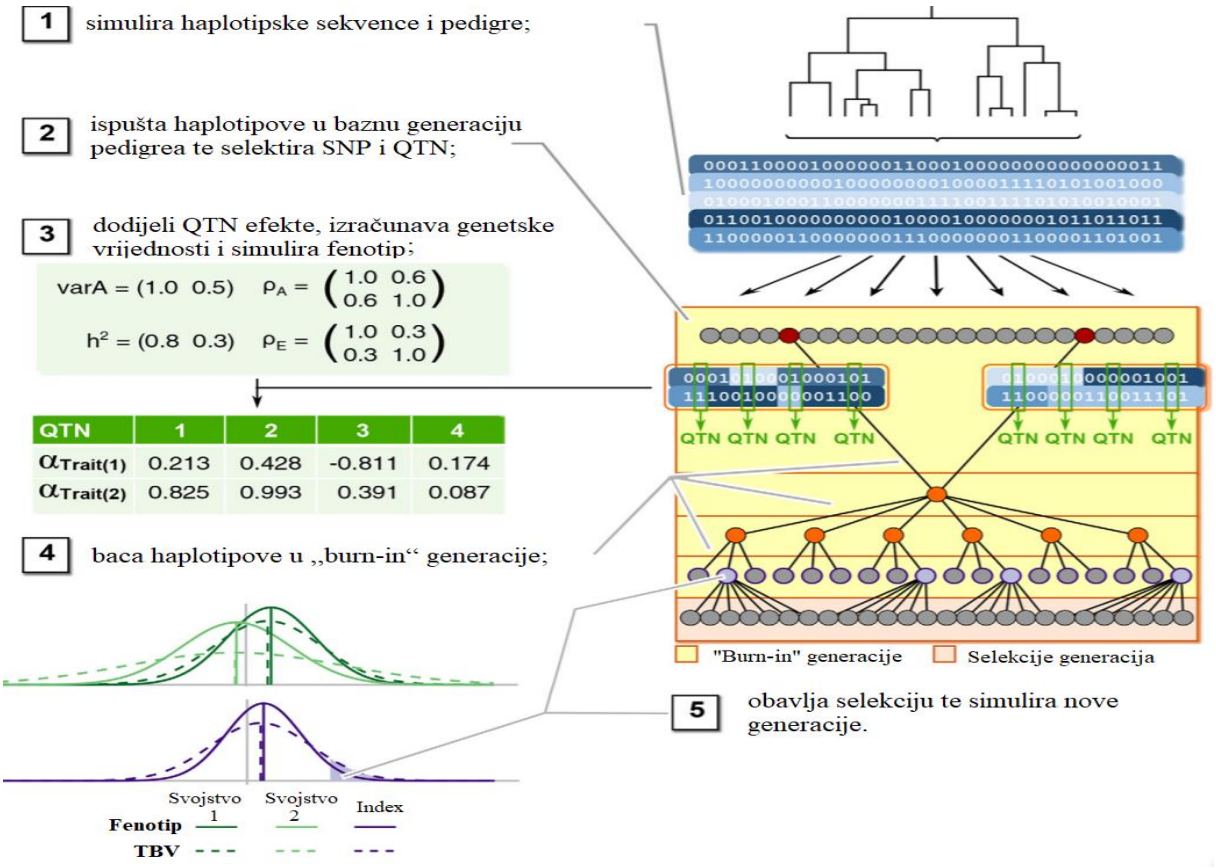
Nadalje, uključen je i „*Markovian coalescent simulator*“ („MaCS“) za brzu simulaciju bi-alelnih nizova prema demografskoj povijesti stanovništva. MaCS proizvodi simulirane podatke koji su gotovo identični podacima simuliranim pod standardnim spajanjem. Potrebno mu je mnogo manje vremena te koristi manje memorije (Faux i sur. 2016.).

AlphaSimR simulira uzgojne programe u sljedećem nizu koraka:

1. simulira haplotipske sekvence i pedigre;
2. ispušta haplotipove u baznu generaciju pedigrea te selektira polimorfizam jednog nukleotida (SNP) i quantitative trait nucleotide (QTN);
3. dodjeljuje QTN efekte, izračunava genetske vrijednosti i simulira fenotipove;
4. baca haplotipove u „burn-in“ generacije; i
5. obavlja selekciju te simulira nove generacije.

Program je fleksibilan s obzirom na povijesnu strukturu stanovništva i raznolikost, nedavnu rodovničku strukturu, arhitekturu svojstava i strategiju odabira. Integrira biotehnologije poput dvostrukih haploida i editiranja gena. Na taj način omogućava korisniku da simulira više svojstava, okruženja, specificira rekombinacijske „hot spot-ove“ i „cold spot-ove“, odnosno „otoke“ i „pustinje“, provodi genomska predviđanja i mnoge druge (Faux i sur. 2016.).

AlphaSimR uključuje funkcionalnosti ponovnog pokretanja koje povećavaju njegovu fleksibilnost. Dopušta pauziranje procesa simulacije tako da se parametri mogu mijenjati ili se može uvesti vanjski rodovnik. Dopušta pokusni dizajn ili rezultate analize prethodno simuliranih podataka. Kombinacijom opcija korisnik može simulirati jednostavne ili složene uzgojne programe s nekoliko generacija, promjenjivu strukturu populacije i varijabilne odluke o uzgoju tijekom vremena (Faux i sur. 2016.).



Slika 2.2.1.1. AlphaSimR ilustrirana simulacija primjenom rodovnika strukturiranog od četiri „burn-in“ generacije i jedne selekcijske generacije za dvije osobine karakterizirane aditivnim genetskim modelom.
Izvor: Faux i sur. 2016.

2.2.2. PLINK

PLINK je dostupan za besplatno preuzimanje na izvornim platformama Linux, MS-DOS, Apple Mac i C/C ++ (Plink…, 2017). Programski alat PLINK dizajnirali su Purcell i sur. (2007.) kako bi se olakšala analiza „whole-genome“ podataka na više načina. Napravljen je kao „open-source“ C/C++ alat.

Koristeći relativno malen broj podataka, naprimjer od 100,000 SNP-ova i 350 jedinki, PLINK-u treba približno 10 sekundi da bi učitao, filtrirao i izveo analizu za sve SNP-ove. Izuzev računalnih izazova, veliki setovi podataka pogoršavaju problem zbunjivanja u genetskim istraživanjima. S povećanom mogućnosti da se otkriju pravi efekti, dolazi i povećani potencijal za pristranost koja može utjecati na rezultate (Purcell i sur. 2007.).

PLINK koristi pristup „kliznog prozora“ s ciljem definiranja ROH-a kao dužine, uključujući minimalan specifičan broj homozigotnih SNP-ova unutar određene kb udaljenosti. Softver podržava samo osnovno otkrivanje ROH-a: naredba „--homozyg“ definira ROH segmente pomoću klizanja „prozora“ koji pretražuje SNP podatke kako bi otkrio homozigotne regije. PLINK prvo određuje može li se određeni SNP nalaziti unutar ROH-a računanjem udjela

potpuno homozigotnih „prozora“ u kojima se pojavljuje spomenuti SNP. Korištenjem unaprijed određene granice „prozora“ 0,05, znači da, ako je 5% ovih „prozora“ potpuno homozigotno, tada je SNP uključen u ROH (Čurik i sur. 2014.).

Također, ROH se može pozvati naredbama prema broju SNP-ova („--homozyg-snp“) ili minimalnoj duljini segmenta („--homozyg-kb“). U oba slučaja, veličina „kliznog prozora“ može se odrediti izračunavanjem proporcija pomoću naredba „--homozyg-window-snp“ ili „--homozyg-window-kb“. Veličina „prozora“ ne smije biti veća od željenog broja SNP-ova, u protivnom program neće otkriti segmente manje od veličine „prozora“ (Čurik i sur. 2014.). PLINK također predviđa određivanje maksimalnog razmaka između dva homozigotna SNP-a (--homozig-gap) i minimalnog praga gustoće za ROH (-homozig-density) (Bjelland i sur. (2013.).

Purcell i sur. (2007.) su dizajnirali alat „PLINK WGAS“ u svrhu ispunjavanja sljedećih zahtjeva: (a) pružanje jednostavnog načina rukovanja velikim WGAS skupinama podataka, (b) pristupa problemu zbunjivanja zbog stratifikacije i ne nasumičnih genotipskih pogrešaka na način da proizvodi široki spektar statistike, (c) za učinkovito provođenje različitih standardnih testova pridruživanja na vrlo velikim skupinama podataka (u populaciji ili obitelji, za bolesti ili kvantitativne ishode, omogućujući haplotip testove, itd.) i (d) pružanje načina za analizu rijetkih varijacija upotrebom zajedničkih SNP panela (Purcell i sur. 2007.).

2.2.3. cgaTOH

TOH je definiran kao kromosomski dio koji mora zadovoljiti kriterij s L uzastopnih homozigotnih SNP-ova ili genetsku udaljenost od minimalno M kb (kb - kilo baza) na samom kromosomu za dani subjekt. L i M predstavljaju parametre definirane od strane korisnika (Reber, 2013.).

Softver cgaTOH pruža intuitivan i interaktivan alat vizualizacije za bolje istraživanje visokopropusnog izlaza s posebnim interaktivnim navigacijskim krugovima. U sustav je uključen i NCBI (National Center for Biotechnology Information) preglednik genomske karte. Identificira razne uzorke SNP-ova koji pokazuju proširenu homozigotnost. Iz tog razloga se mogu primijeti različiti aspekti višestrukih karakteristika TOH-a (Zhang i sur. 2013.).

U svrhu najboljeg kotištenja TOH svojstava te olakšavanja daljnje statističke analize, predstavljena su 3 tipa surogat TOH-a : (a) „cTOH“ – (common TOH) regija koja pokriva „cluster“ uzastopnih SNP-ova koji pripadaju „TOH-u“ u više subjekata, (b) „gTOH“ – (group TOH) regija koja sadržava grupu preklapajućih „TOH“-ova s proksimalnim granicama, (c) „aTOH“ – (allelic TOH) regija koja pokriva TOH-ove s preklapajućm alelno-podudaranim regijama (Zhang i sur. 2013.).

„cgaTOH“ program počinje tako da traži „TOH“-ove i „cTOH“-ove. Kasnije identificira odvojeno „gTOH“ i „aTOH“ te je svaka popraćena interaktivnom regionalnom vizualizacijom. Rezultat toga je program koji je mnogo efektivniji te ima povišenu funkcionalnost i fleksibilnost upotrebljavajući odgovarajuće parametre. Identifikacija TOH-a, „cTOH“-a, „gTOH“-a i „aTOH“-a je implementirana bazirajući se na C++ sa sučeljem za linije naredba s opcijama za izbor parametara te vizualizacijsko sučelje stvoreno od strane Qt za

„cross-platform UI“ (User Interface) (Zhang i sur. 2013.). Ulazni podaci moraju biti obrađeni nakon redovite kontrole kvalitete te u istom formatu kao i ulazne datoteke za PLINK, tj., „.ped“ i „.map“ datoteke. Izlazne datoteke su u tabularnom formatu teksta, koji uključuje podatke o genotipu i fenotipu za sve TOH-ove i surogat-TOH-ove te su spremni za daljnju statističku analizu (Zhang i sur., 2013.).

Zaključno, cgaTOH je algoritam za identificiranje različitih uzorka SNP-a. Oni pokazuju proširenu homozigotnost od pojedinačnih TOH-ova do surogatskih TOH-ova. Na taj način omogućava promatranje različiti aspekata višestrukih karakteristika TOH-a i detektiranje proširene regije homozigotnosti. CgaTOH integrira komponente sa sustavom vizualizacije za čitav genom, omogućujući intuitivnu navigaciju i istraživanje podataka kako bi se pomoglo u daljnjoj analizi (Zhang i sur. 2013.).

2.2.4. RZooRoH

ROH se najčešće tumače kao autozigotni segmenti ili homozigotni po porijeklu („Homozygous-by-Descent“, HBD) , tj. koji se sastoji od parova haplotipova naslijeđenih od zajedničkog pretka bez rekombinacije (i mutacije) u dva različita genetička puta.

Paket RZooRoH nudi funkcije za prepoznavanje segmenata homozigotnog pretka (HBD) i za procjenu pojedinačne autozigotnosti (ili inbreeding koeficijenata). Segmenti HBD-a nastaju kada pojedinac od kopije naslijedi dvije kopije istog segmenta kromosoma. Dvije kopije nasljeđuju se različitim putevima. Jedna se kopija nasljeđuje od majke, a druga od oca. To se događa u slučaju inbreedinga kada su roditelji u srodstvu (imaju zajedničkog pretka). U nedostatku mutacija, to stvara duge segmente homozigotnih genotipova (ROH). Duljina segmenata ovisi o broju generacija od pojedinca do zajedničkog pretka (generacije iz dva različita puta moraju se zbrojiti) ili veličini petlje za križanje. Paket RZooRoH dostupan je za većinu platformi (Linux, MS Windows i MacOS) iz CRAN repozitorija (Druet i sur. 2019.).

Leutenegger i suradnici (2003.) dali su temelj modeliranju IBD procesa duž kromosoma razvijajući HMM za identificiranje HBD segmenata. Takav HMM okvir omogućuje učinkovito korištenje raspoloživih genetskih podataka sadržanih u nizovima homozigotih i heterozigotnih markera kao i mapama vezanosti (linkage maps). Može radit s podacima sekvence čitavog genoma (Narasimhan i sur., 2016), uključujući one dobivene eksperimentima s low-fold sekvenciranjem (Vieira i sur., 2016). Svi pristupi, utemeljeni na modelima oslanjaju se na HMM s obzirom da svaki marker pripada non-HBD ili HBD segmentu. Vjerojatnost prijelaza između ta dva (skrivena) stanja sukcesivnih markera ovisit će o njihovoj genetskoj udaljenosti, parametru koji kontrolira brzinu promjena po jedinici genetske udaljenosti i pojedinačnom inbreending koeficijentu. Uzimajući u obzir samo spomenuta dva stanja (HBD ili non-HBD) program zapravo pretpostavlja da svi segmenti HBD-a unutar određene jedinke imaju istu očekivanu duljinu (Duret i sur. 2017.).

HMM koriste frekvencije alela i stope pogrešaka genotipizacije slično kao ROH temeljen na vjerojatnosti. Iz tog razloga su manje osjetljivi na pristranost markera ili kriterije filtriranja (minimalni MAF). Ponekad se ROH, temeljen na pravilima, provodi s monomorfnim markerima. U HMM-u su ti markeri automatski neinformativni. Pored toga, HMM koristi

podatke iz genetske mape (udaljenost između markera), automatski uzimajući u obzir gustoću markera ili razmaka markera, te postaje robusniji za promjenjive stope rekombinacija duž genoma. Također, HMM automatski istražuje sve moguće duljine HBD segmenata (ne zahtijevaju definiciju optimalne veličine „prozora“) i pruža vjerojatnost HBD-a (Duret i sur. 2017.). HMM omogućuje rad s nepravilnim razmakom markera (egzome, Genotyping-by-sequencing). To omogućava obradu podataka kako egzoma tako i cijelog genoma (Magi i sur., 2014; Narasimhan i sur., 2016; Vieira i sur., 2016).

Pristupi detekcije ROH-ova utemeljeni na HMM-u uspoređeni su u nekoliko simulacijskih istraživanja. Simulirani scenariji često mogu utjecati na rezultate (odabrana gustoća markera, ujednačeni razmak markera, prisutnost pogrešaka u genotipizaciji ili ne). Parametri ROH-a se mogu prilagoditi kako bi se simulacije bolje uklopile. Usporedbe treba tumačiti s oprezom. Narasimhan i sur. (2016) u utvrdili da HMM ima niže lažno pozitivne rezultate kao i lažno negativne rezultate u usporedbi ROH procjene s PLINK-om.

RZooRoH prema zadanim postavkama čita datoteku genotipa u formatu "Oxford GEN" s jednim retkom po markeru i jedinkama u stupcima (datoteka GEN). Za konvertiranje ulaznih podataka u „.gen“ format Druet i Gautier (2017.) preporučuju korištenje programa PLINK pomoću koda: „plink --file myinput --recode oxford --autosome --out myoutput“. Pet prvih stupaca (odvojeni razmakom ili tabulatorima) sadrže podatke o markeru:

1) Broj kromosoma;

2) ID SNP-a ili ime markera (najviše 50 znakova);

3) Fizički položaj markera u bp;

4) Alel prvog markera (najviše 50 znakova);

5) Alel drugog markera (najviše 50 znakova).

Nakon prvih pet stupaca, svaka vrijednost predstavlja genotip kodiran kao broj kopija alela 1 od pojedinca („0“ za homozigote s drugim alelom, „1“ za heterozigote, „2“ za homozigote s prvim alelom i „9“ za missing genotipove). Alternativno, RZooRoH može očitati vjerojatnost genotipa (GP), vjerojatnost genotipa na phred skali (mjera kvalitete identifikacije nukleobaza nastalih automatskim sekvenciranjem DNA) (GL) ili dubinu čitanja za oba alela (AD). U slučaju GP ili GL, daju se tri vrijednosti po pojedincu (tri kolone) što odgovara vjerojatnosti genotipa ili vjerojatnosti phreda za genotipove „11“ (homozigoti s alelom 1), „12“ (heterozigoti) i „22“ (homozigoti s alelom 2). Uz AD, očekuju se dvije kolone po pojedincu, alelna dubina za alel 1 i alelna dubina za alel 2 (Tom Duret 2017.). Naziv datoteke genotipa i parametri su navedeni u datoteci parametara koja se mora nazvati "pharm.txt". Ispred svakog parametra ili opcije mora stajati simbol „#“. Naprimjer: #INBREEDING_INDICATORS određuje koje su klase IBD (1) i non-IBD (0). Jedan redak s vrijednostima „K“ (broj klasa, „1“ za IBD i „0“ za non-IBD) odvojen razmakom (Tom Duret 2017.).

Formati GT, GP, GL i AD su opisani u odjeljku ulaznih datoteka. Nakon odabira formata ulazni podaci moraju biti učitani funkcijom „zoodata“ stvarajući „zooin“ objekt koji sadrži sve informacije za daljnju analizu. Objekt „zooin“ namijenjen je unutarnjoj upotrebi. Sadrži devet dijelova potrebnih za daljnju analizu: genos, bp, chrbound, nind, nsnps, nchr,

zformat, sample_ids. Pomoću funkcije „zoodata“ korisnik može odrediti ime datoteke s podacima, format podataka o genotipu i markerima, minor frekvencije alela (MAF), pravila filtriranja te naziv datoteke s pojedinačnim ID-em ili s prethodno procijenjenim frekvencijama alela (npr., s većim skupom podataka). Funkcija „zoorun“ omogućava procjenu parametara modela, globalne i lokusne procijenjene autozigotnosti, podjelu u različite klase HBD-a i identifikaciju HBD segmenata. „Zoorun“ zahtijeva dva bitna elementa: „zmodel“ i „zoodata“ objekte koji pružaju modele i podatke (Druet i Gautier 2017.).

Moguće vrijednosti outputa su „PartialF“ ili „TotalF“. Prema zadanim postavkama *inbreeding* nije dostupan na svakoj poziciji markera. Korisnik može zahtijevati ukupni *inbreeding* za svaku poziciju markera po pojedincu s opcijom „TotalF“. To generira jednu datoteku koja se zove „TotalF.txt“ s onoliko stupaca koliko ima pojedinaca te isto onoliko redaka koliko ima markera (prva tri stupca daju informacije o markerima). Korisnik može zahtijevati datoteku za svaku IBD klasu nakon čega program generira nekoliko datoteka s istom strukturom (prema broju IBD klasa). Nazivi datoteka su „PartialX.txt“, gdje je „X“ broj IBD klase (jedan za prvu IBD klasu, dva za drugu, itd.) (Tom Duret 2017.).

Rezultati su grupirani u zres objekt s 12 slotova. Njima se može pristupiti simbolom „@“. Neki slotovi opisuju uzorke u analizi:

@ nind predstavlja broj analiziranih pojedinaca u tijeku;

@ ids predstavlja niz s brojevima analiziranih pojedinaca;

@ sampleids predstavlja niz s nazivom uzoraka (ako su dostavljeni) ili s brojem uzoraka;

@ optimerr predstavlja niz koji pokazuje da li je optim pokrenut bez pogreške (za svakog pojedinca) (Druet i Gautier 2017.).

3. Materijali i metode

3.1. Simulacija podataka

Simulacija podataka izvršena je putem programskog paketa AlphaSimR. Simulirano je ukupno 60 populacija, podijeljenih u dvije grupe od 30. Svaka grupa je imala unaprijed određeni Ne (efektivnu veličinu populacije). Prva grupa jedinki imala je 500 generacija i Ne 25. Druga grupa jedinki imala je 1000 generacija i Ne 300. Simulirala se bazna populacija (populacija osnivača) te njezin razvoj kroz 50 generacija. Kod za simulaciju populacija s određenim Ne se nalazi u prilogu. Prvi dio koda se odnosi na pripremanje radne okoline i postavljanje makro naredbi kako bi kod bio lakši za čitanje. U drugom dijelu se postavljaju parametri koje će populacija osnivača imati.

Naredba „Species“ omogućuje odabir životinjske vrste čije genomske podatke želimo simulirati; autorica se odlučila za simulaciju jedinki goveda. „NChr“ postavlja broj kromosoma koji se simuliraju. „nSnpPerChr“ postavlja željeni broj SNP-ova po kromosomu. „nQtlPerChr“ postavlja željeni broj Qtl-ova (engl. Qunatitative trait loci) po kromosomu. „nGen“ postavlja broj generacija kroz koje će populacija proći u simuliranju, u ovom slučaju jedna generacija zato što se simulira populacija osnivača. „nIndPerGen“ postavlja koliko potomaka će se simulirati po generaciji.

Nakon postavljanja parametara slijedi simuliranje populacije postavljeno u „if petlju“. Unutar „if petlje“ naredbom „runMacs“ se vrši simulacija, a unutar te naredbe AlphaSimR-u je potrebno postaviti i ispuniti objekt „SP“ (engl. SimulationParameters) kako bi se simulacija mogla izvršiti. U objekt „SP“ se odabiru parametri poput: dodjeljivanja spola („SP$setGender(gender=“yes_rand“)“), zapisivanja rodovnika („SP$setTrackPed(isTrack=“TRUE“)“), zapisivanje rekombinacija („SP$setTrackRec(isTrackRec = “TRUE“)“).

Nakon simulacije populacije osnivača simulirano je 60 populacija različitih Ne. Kodovi za njihovu simulaciju populacija su istovjetni izuzev parametara za postavljanje Ne i naziva datoteka. Kodovi će biti objašnjeni na primjeru koda korištenog za simulaciju populacije s Ne=25 (kod za simulaciju populacije s Ne=25 se nalazi u prilogu). Prvi dio koda je istovjetan kao i kod simulacije populacije osnivača. Odnosi se na pripremanje radne okoline i postavljanje makro naredbi. Budući da su se simulacije vršile iz iste populacije osnivača potrebno ju je uvesti u radno okruženje pomoću naredbe „load“. Nakon uvlačenja populacije osnivača slijedi postavljanje parametara populacije s ciljem dobivanja populacija s određenim Ne. U naredbi „nSires“ je postavljen željeni broj očeva, a „nDams“ željeni broj majki te se vršila provjera dobivene Ne (Formula 4.)

Odlučeno je simulirati 500 jedinki po generaciji stoga „nIndPerGen=500“. Kako se za svaku populaciju istih kategorija Ne radio novi direktorij unutar petlje, bilo je potrebno razdvojiti simuliranje populacija u tri dijela, radi restrikcija R-a, kako bi se dobilo 30 (12+12+6) populacija. Same simulacije su se odrađivale unutar dvostruke „for petlje“. Prva „for petlja“ označavala je broj simulirane populacije („for (Rep in 1:12)“) a druga broj generacija („for Gen

in 1:50)“). Unutar „for petlje“ za generacije, vršila se nasumična selekcija očeva („Sires=..., use=“rand“,...“) i majki te simulacija križanja pomoću funkcije „randcross2“ koja nasumično odabire roditeljske kombinacije od svih mogućih kombinacija roditelja. Unutar funkcije „randcross2“ odabran je i broj potomaka po roditeljskom križanju („nProgeny=1“) te ne-balansiranje broja potomaka po roditelju („balance=“FALSE““).

Zadnji korak simulacije bio je eksportiranje bazne populacije u PLINK obliku („.ped“ i „.map“ datoteke). Koristila se naredba „writePlink“ gdje je bilo potrebno napisati argumente „pop“ i „baseName“. Kod korišten za simulaciju populacije s Ne=25 nalazi se u prilogu pod brojem 8.1.

3.2. R paket – RZooRoH

Ulazni podaci konvertirani su u Oxford GEN format (SNP-ovi su u redovima, a jedinke po kolonama) u programu PLINK (kod za konvrtiranje u PLINK-u nalazi se u prilogu). Potrebno je preuzeti paket „RZooRoH“ te postaviti radni direktorij. Prvo se koristi naredba „system.file“ kojom se provjerava postojanje navedenih datoteka, tj. naredba pronalazi puna imena datoteka u paketima (prilog 8.2.)

Genotipske informacije posložene su GP formatom tj. formatom koji sadrži tri kolone za tri moguća genotipa AA AB i BB. „Zoodata“ naredbom se čita podatkovna datoteka i pretvara se u format RZooRoH u "zooin" objekt potreban za daljnju analizu. Pri pretvorbi modela u „zoomodel“ definiralo se K= 4, tj. postavilo se 4 klase po bazi 5. Time se dobilo „zmodel“ koji se „zdatom“ može koristiti dalje za „zoorun“ opciju kojom identificiramo HBD segmente pomoću Viterbi algoritma.

Rezultati su dobiveni u zres obliku. Moguće ih je pregledati pomoću simbola @: typ.res@modlik, typ.res@realized, typ.res@hbdseg. Podatke se eksportiralo u csv obliku uz pomoć naredbe „write.csv“.

3.3. Detekcija ROH segmenata

Detekcija ROH segmenata provedena je uz pomoć paketa cgaTOH i PLINK. Za cgaTOH zadana je duljina ROH-ova 1000000 (1Mb) uz naredbu „-min_length 1000000“. Svako posebno detektiranje je imalo maksimalni dozvoljeni razmak između SNP-ova 1000000 postavljeno naredbom „-max_gap 1000000“. U detekciji bio je dopušten samo jedan heterozigot naredbom „-max_hetero 1“. Kod za detekciju ROH segmenata nalazi se u prilogu pod brojem 8.3. Detekcija ROH segmenata u PLINK-u obuhvaćala je minimalnu duljina segmenta postavljenu je na 1000kb i 15 SNP-ova pri čemu je dopušten jedan heterozigot i četiri missing SNP-a. Najveća dopuštena udaljenost između dva SNP-a iznosila je 1Mb.

3.4. Statistička obrada i vizualizacija

Statistička obrada i vizualizacija rezultata, dobivenih putem programskih paketa cgaTOH, PLINK i RZooRoH, izvršena je u programu SAS 9.4 i JMP14. Za usporedbu srednjih vrijednosti između grupa korišten je Tukey-Kramer HSD test.

4. Rezultati i rasprava

Populacija osnivača je simulirana u AlphaSimR-u sa sljedećim parametrima: 500 jedinki, 2000 SNP-ova, 1 kromosom te nasumično dodjeljivanje spola. Koristeći istu populaciju osnivača dalje se simuliralo 60 populacija podijeljenih u dvije grupe gdje je svaka grupa imala točno određeni Ne (Ne=25 i Ne=300). Kako bi se postigao određeni Ne, manipuliralo se brojem očeva i majki koji su sudjelovali u daljnjim generacijama populacije (Formula 4.1.):

Formula 4.1. $$N_e = \frac{4*N_m*N_f}{N_m+N_f}$$

Tablica 4.1. Kriteriji za simulaciju populacija s unaprijed određenom efektivnom veličinom populacije (Ne)

Simulirana Ne	Broj očeva	Broj majki
25	7	49
300	100	300

Sve simulirane populacije za Ne=25 imale su 500 jedinki po generaciji, 50 generacija, jednog potomka po križanju, nebalansirane potomke po roditeljima te diskretne generacije. Nakon simulacije potrebnih populacija svi dobiveni podaci konvertirani su u Oxford GEN format (SNP-ovi su u redovima, a jedinke po kolonama) u programu PLINK. Konvertirane datoteke dalje se koristilo u R paketu RZooRoH, PLINK-u te cgaTOH-u. Dobivene rezultate se uspoređivalo prema podacima prosjeka ROH inbreeding koeficijenata, broja ROH segmenata, prosjeka parcijalnih ROH inbreeding koeficijenata te prosjeka broja ROH segmenata iz različitih ROH kategorija.

Tablica 4.2. prikazuje deskriptivnu statistiku i usporedbu prosjeka ROH inbreeding koeficijenata 30 simuliranih populacija s 500 jedinki efektivne veličine 25 kroz 50 generacija procijenjenih različitim programskim paketima. FROH se može određivati za veće od FROH > 1Mb, FROH>2Mb, ali može i za kategorije FROH 1-2Mb, FROH 2-4 MB, itd.. To se radi da bi se izdvojio inbreeding vezan za neku generaciju. Kada je FROH 1-2 tada se radi o segmentima koji dolaze od zajedničkog pretka, udaljenog između 50 i 25 generacija. Tablica 4.3. prikazuje deskriptivnu statistiku i usporedbu prosjeka broja ROH segmenata iz gore navedene populacije.

Tablica 4.2. Deskriptivna statistika i usporedba prosjeka ROH inbreeding koeficijenata 30 simuliranih populacija s 500 jedinki efektivne veličine 25 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

		FROH > 1	FROH > 2	FROH > 4	FROH > 8	FROH > 16
cgaTOH	prosjek	0,697 A	0,649 A	0,580 A	0,452 A	0,273 A
	standardna devijacija	0,063	0,068	0,080	0,097	0,115
	minimum	0,594	0,543	0,455	0,327	0,145
	maximum	0,913	0,903	0,880	0,825	0,713
PLINK	prosjek	0,689 A	0,645 A	0,576 A	0,444 A	0,269 A
	standardna devijacija	0,063	0,068	0,080	0,097	0,113
	minimum	0,580	0,535	0,445	0,298	0,151
	maximum	0,915	0,897	0,875	0,815	0,702
RZooROH	prosjek	0,435 B	0,430 B	0,407 B	0,342 B	0,224 A
	standardna devijacija	0,040	0,041	0,046	0,057	0,076
	minimum	0,371	0,362	0,335	0,260	0,138
	maximum	0,568	0,568	0,564	0,551	0,512

Tablica 4.3. Deskriptivna statistika i usporedba prosjeka broja ROH segmenata 30 simuliranih populacija s 500 jedinki efektivne veličine 25 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

		nseg > 1	nseg > 2	nseg > 4	nseg > 8	nseg > 16
cgaTOH	prosjek	10,690 A	7,201 A	4,829 A	2,614 A	1,020 A
	standardna devijacija	1,561	0,869	0,579	0,303	0,354
	minimum	5,436	4,902	3,760	2,068	0,510
	maximum	13,836	8,816	6,158	3,304	2,124
PLINK	prosjek	10,454 A	7,293 A	4,894 A	2,589 A	1,021 A
	standardna devijacija	1,440	0,925	0,666	0,314	0,358
	minimum	6,074	4,760	3,746	1,942	0,550
	maximum	13,246	8,918	6,682	3,186	2,062
RZooROH	prosjek	4,076 B	3,750 B	2,981 B	1,859 B	0,814 B
	standardna devijacija	0,607	0,551	0,378	0,183	0,217
	minimum	2,087	2,077	1,948	1,423	0,488
	maximum	4,970	4,595	3,723	2,206	1,382

Tablica 4.2. za FROH>1 prikazuje kako se prosjeci rezultata cgaTOH-a i PLINK-a značajno ne razlikuju, dok RZooRoH ima značajno manji prosjek. Isto se odnosi na FROH>2, FROH>4 i FROH>8. Za FROH>16 nema značajne razlike između rezultata. U Tablici 4.3. prosjeci dobiveni cgaTOH-om i PLINK-om prikazuju da nema značajne razlike. Prosjeci dobiveni RZooRoH-om statistički se značajno razlikuju od prosjeka dobivenih cgaTOH-om i PLINK-om. Možemo zaključiti kako za manje segmente RZooROH rjeđe dodjeljuje IBD status, odnosno homozigotni niz proglašava HBD.

Tablica 4.4. prikazuje deskriptivnu statistiku i usporedbu prosjeka parcijalnih ROH inbreeding koeficijenata 30 simuliranih populacija s 500 jedinki efektivne veličine 25 kroz 50 generacija procijenjenih različitim programskim paketima. Za razliku od tablice 4.2., ovdje se nalazi FROH 1-2, FROH 1-3, itd.. Takav FROH upućuje da se radi o segmentima koji dolaze od zajedničkog pretka. Tablica 4.5. prikazuje deskriptivnu statistiku i usporedbu broja ROH segmenata iz različitih ROH kategorija na opisanoj populaciji.

Tablica 4.4. Deskriptivna statistika i usporedba prosjeka parcijalnih ROH inbreeding koeficijenata 30 simuliranih populacija s 500 jedinki efektivne veličine 25 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

		FROH 1-2	**FROH 2-4**	**FROH 4-8**	**FROH 8-16**
cgaTOH	**prosjek**	0,049 [A]	0,069 [A]	0,129 [A]	0,179 [A]
	standardna devijacija	0,015	0,019	0,035	0,035
	minimum	0,010	0,024	0,058	0,117
	maximum	0,078	0,108	0,205	0,242
PLINK	**prosjek**	0,044 [A]	0,069 [A]	0,133 [A]	0,176 [A]
	standardna devijacija	0,012	0,021	0,039	0,033
	minimum	0,018	0,024	0,062	0,117
	maximum	0,072	0,120	0,217	0,237
RZooROH	**prosjek**	0,005 [B]	0,024 [B]	0,068 [B]	0,123 [B]
	standardna devijacija	0,002	0,008	0,017	0,025
	minimum	0,000	0,007	0,025	0,073
	maximum	0,011	0,036	0,104	0,163

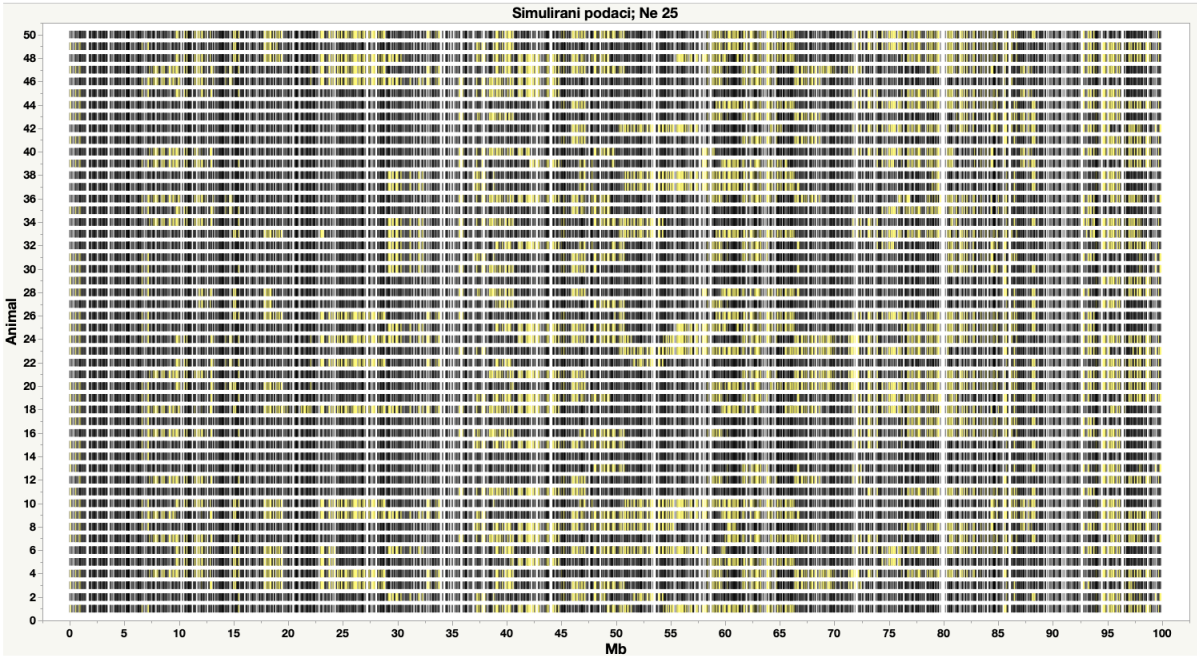
Tablica 4.5. Deskriptivna statistika i usporedba prosjeka broja ROH segmenata iz različitih ROH kategorija 30 simuliranih populacija s 500 jedinki efektivne veličine 25 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

		nseg 1-2	**nseg 2-4**	**nseg 4-8**	**nseg 8-16**
cgaTOH	**prosjek**	3,500 [A]	2,380 [A]	2,223 [A]	1,601 [A]
	standardna devijacija	1,094	0,640	0,601	0,306
	minimum	0,560	0,761	1,115	1,036
	maximum	5,456	3,606	3,648	2,171
PLINK	**prosjek**	3,171 [A]	2,408 [A]	2,314 [A]	1,574 [A]
	standardna devijacija	0,867	0,737	0,694	0,303
	minimum	1,363	0,734	1,085	0,979
	maximum	5,072	4,265	4,082	2,178
RZooROH	**prosjek**	0,338 [B]	0,797 [B]	1,166 [B]	1,089 [B]
	standardna devijacija	0,134	0,243	0,304	0,225
	minimum	0,020	0,242	0,426	0,641
	maximum	0,650	1,246	1,793	1,462

U Tablici 4.4 uočena je značajna statistička razlika za prosjeke dobivene korištenjem programa RZooRoH, dok između prosjeka dobivenih cgaTOH-om i PLINK-om nema statistički značajne razlike i to vrijedi za sve parcijalne koeficijente inbreedinga. U Tablici 4.5. statistički značajna razlika između prosječnog broja segmenata dobivenog korištenjem programa RZooRoH u odnosu na prosjeke dobivene programima cgaTOH i PLINK.

Status SNP-s (homozigot – heterozigot) iz simuliranih podataka su grafički prikazani i uspoređeni s predviđanjem ROH regija cgaTOH-a, PLINK-a i RZooRoH-a. U figurama su heterozigoti označeni žutom, a homozigoti crnom bojom. Na vizualan način prikazuju usporedbu statusa SNP-ova u simuliranim podacima (A) s procijenjenim ROH segmentima (B) putem programskog paketa cgaTOH (crno). Prikazano je uvijek samo prvih 50 jedinki prve simulacije populacije kako za Ne = 25 tako i za Ne=300.

A



B

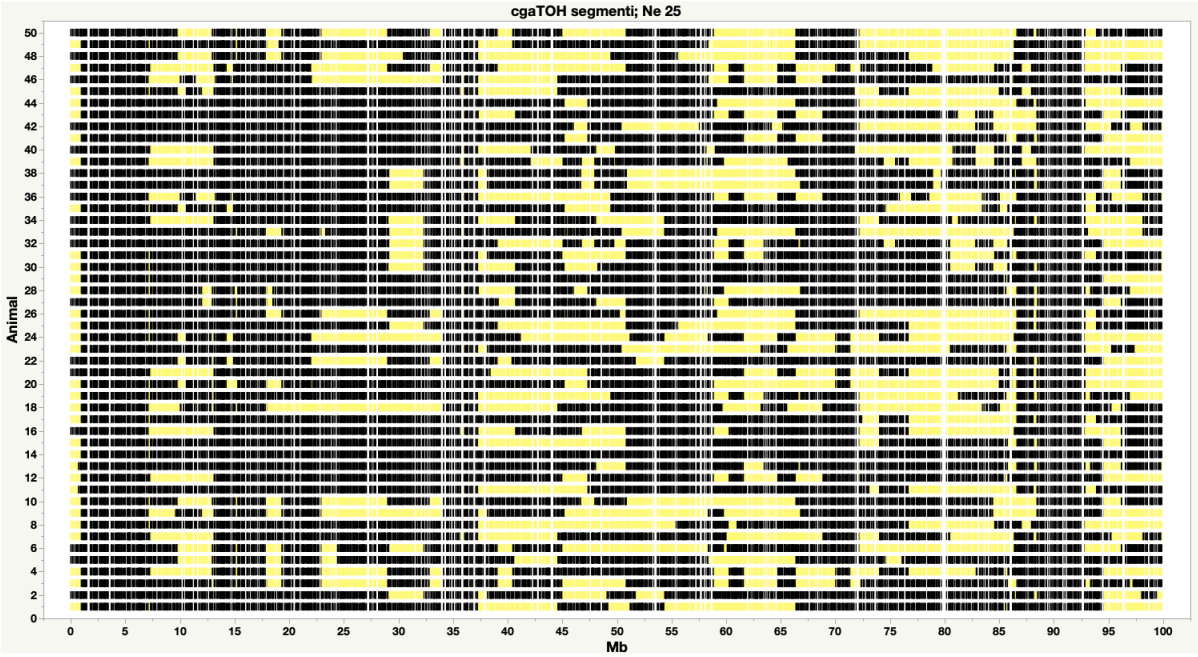
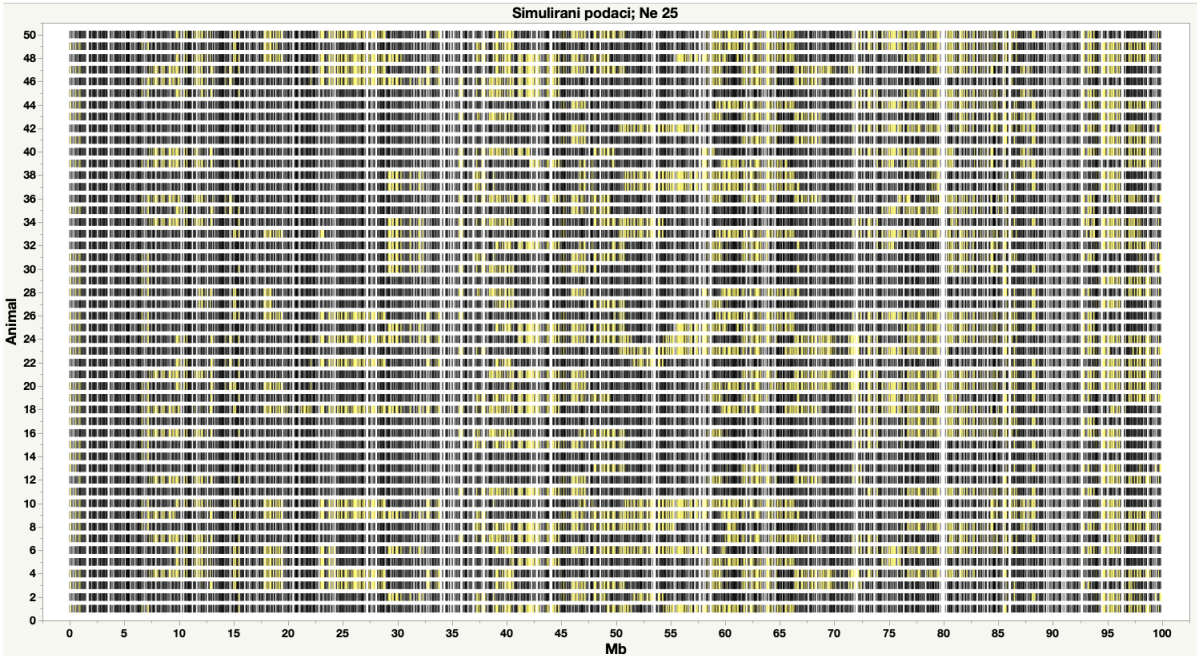


Figura 4.1. Usporedba statusa SNP-ova u simuliranim podacima (homozigot – crno, heterozigot – žuto) (A) s procijenjenim ROH segmentima (B) putem programskog paketa cgaTOH (crno). Prikazano je prvih 50 jedinki prve simulacije populacije s Ne = 25 nakon 50 generacija.

A



B

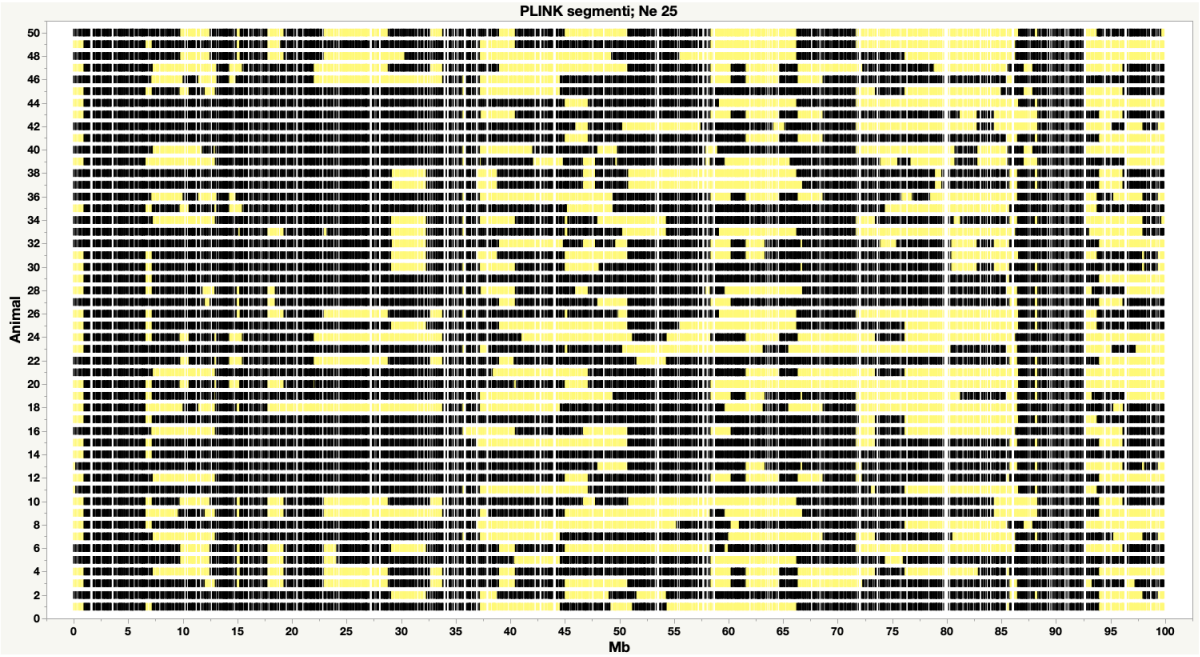
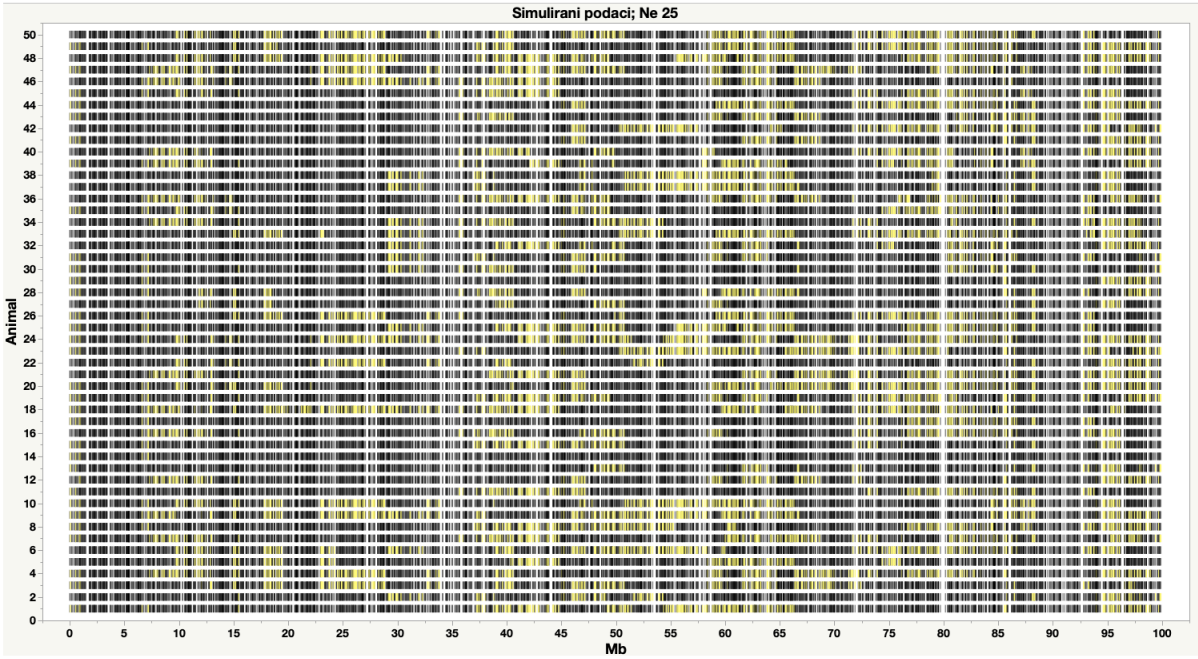


Figura 4.2. Usporedba statusa SNP-ova u simuliranim podacima (homozigot – crno, heterozigot – žuto) (A) s procijenjenim ROH segmentima (B) putem programskog paketa PLINK (crno). Prikazano je prvih 50 jedinki prve simulacije populacije s Ne = 25 nakon 50 generacija.

A



B

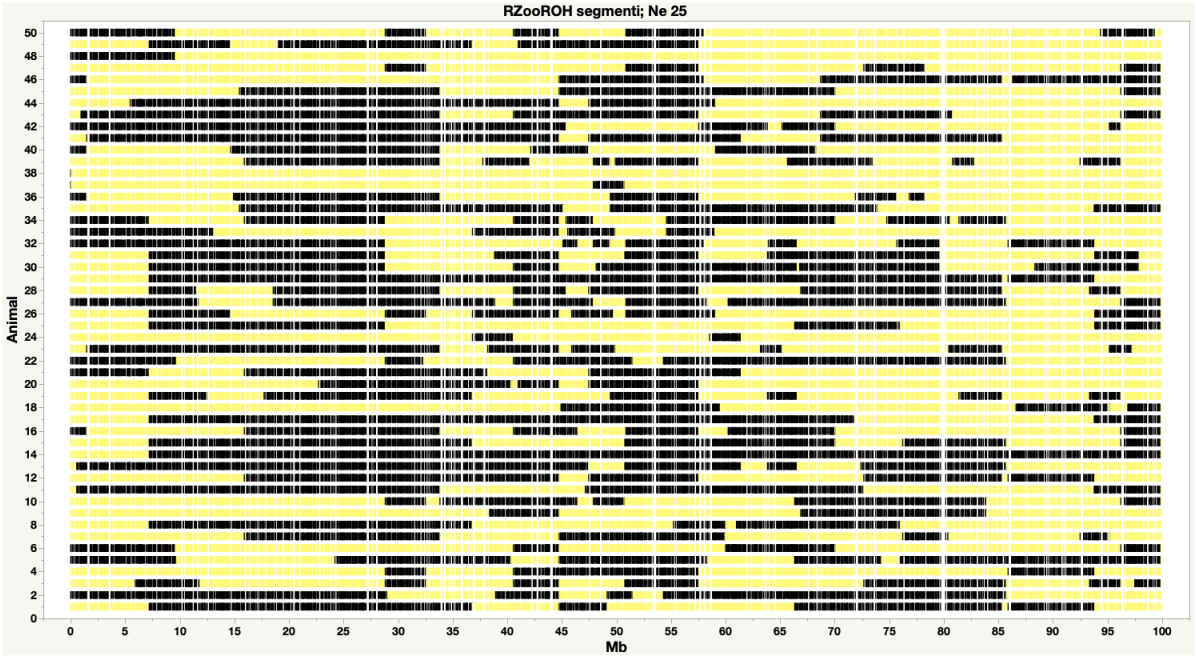


Figura 4.3. Usporedba statusa SNP-ova u simuliranim podacima (homozigot – crno, heterozigot – žuto) (A) s procijenjenim ROH segmentima (B) putem programskog paketa RZooROH(crno). Prikazano je prvih 50 jedinki prve simulacije populacije s Ne = 25 nakon 50 generacija.

Uočena je značajna razlika statusa SNP-ova s procjenjenim ROH segmentima dobivenih programom RZooRoH u odnosu na ostala dva programa. Rezultati RZooRoH-a sadrže znatno manje segmenata nego figure izrađene s druga dva programa. Posebno se izdvaja regija 86 – 94Mb koju cgaTOH i PLINK za gotovo svaku jedinku proglašavaju ROH segmentom ili više njih dok RZooROH dodjeljuje ROH status svega nekoliko jedinki za taj segment. Isto se može uočiti i za regiju 0-7Mb. Ova razlika može biti ključna za detekciju ROH otoka i ROH pustinja.

Tablica 4.6. prikazuje deskriptivnu statistiku i usporedba prosjeka ROH inbreeding koeficijenata 30 simuliranih populacija s 1000 jedinki efektivne veličine 300 kroz 50 generacija procijenjenih različitim programskim paketima, a Tablica 4.7. prikazuje deskriptivnu statistiku i usporedbu prosjeka broja ROH segmenata.

Tablica 4.6. Deskriptivna statistika i usporedba prosjeka ROH inbreeding koeficijenata 30 simuliranih populacija s 1000 jedinki efektivne veličine 300 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

		FROH > 1	**FROH > 2**	**FROH > 4**	**FROH > 8**	**FROH > 16**
cgaTOH	**prosjek**	0,220 [A]	0,120 [A]	0,058 [A]	0,023 [A]	0,009 [A]
	standardna devijacija	0,012	0,011	0,007	0,003	0,002
	minimum	0,194	0,100	0,045	0,018	0,006
	maximum	0,247	0,150	0,080	0,033	0,012
PLINK	**prosjek**	0,203 [B]	0,112 [B]	0,055 [A]	0,022 [A B]	0,008 [A]
	standardna devijacija	0,012	0,011	0,007	0,003	0,002
	minimum	0,012	0,011	0,007	0,003	0,002
	maximum	0,231	0,142	0,077	0,031	0,012
RZooROH	**prosjek**	0,121 [C]	0,090 [C]	0,049 [B]	0,021 [B]	0,008 [A]
	standardna devijacija	0,009	0,009	0,006	0,003	0,002
	minimum	0,009	0,009	0,006	0,003	0,002
	maximum	0,144	0,116	0,068	0,029	0,012

Tablica 4.7. Deskriptivna statistika i usporedba prosjeka broja ROH segmenata 30 simuliranih populacija s 1000 jedinki efektivne veličine 300 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

		nseg > 1	nseg > 2	nseg > 4	nseg > 8	nseg > 16
cgaTOH	prosjek	10,568 [A]	3,116 [A]	0,842 [A]	0,168 [A]	0,034 [A]
	standardna devijacija	0,465	0,248	0,104	0,026	0,006
	minimum	9,536	2,641	0,653	0,122	0,024
	maximum	11,453	3,694	1,174	0,243	0,046
PLINK	prosjek	9,697 [B]	2,870 [B]	0,791 [B]	0,162 [A B]	0,033 [A]
	standardna devijacija	0,414	0,242	0,099	0,024	0,006
	minimum	0,414	0,242	0,099	0,024	0,006
	maximum	10,566	3,457	1,127	0,232	0,046
RZooROH	prosjek	4,283 [C]	2,149 [C]	0,682 [C]	0,153 [B]	0,033 [A]
	standardna devijacija	0,212	0,190	0,086	0,021	0,006
	minimum	0,212	0,190	0,086	0,021	0,006
	maximum	4,602	2,697	0,978	0,211	0,044

U Tablici 4.6. za FROH>1 i FROH>2 postoje značajne statističke razlike između prosjeka za svaki od tri programa. FROH>4 i FROH>8 prikazuju statistički značajnu razliku prosjeka dobivenih cgaTOH-om i PLINK-om u odnosu na prosjek dobiven RZooRoH-om. Za FROH>16 nema statistički značajne razlike među prosjecima. U Tablici 4.7. za nseg>1 i nseg>2 postoje značajne statističke razlike između prosjeka za svaki od tri programa. Nseg>4 i nseg>8 prikazuju statistički značajnu razliku prosjeka dobivenih cgaTOH-om i PLINK-om u odnosu na prosjek dobiven RZooRoH-om. Za nseg>16 nema statistički značajne razlike među prosjecima.

Tablica 4.8. prikazuje deskriptivnu statistiku i usporedbu prosjeka parcijalnih ROH inbreeding koeficijenata 30 simuliranih populacija s 1000 jedinki efektivne veličine 300 kroz 50 generacija procijenjenih različitim programskim paketima dok Tablica 4.9. prikazuje deskriptivnu statistiku i usporedbu prosjeka broja ROH segmenata.

Tablica 4.8. Deskriptivna statistika i usporedba prosjeka parcijalnih ROH inbreeding koeficijenata 30 simuliranih populacija s 1000 jedinki efektivne veličine 300 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti ispitanu Tukey-Kramer HSD testom ($p < 0,01$).

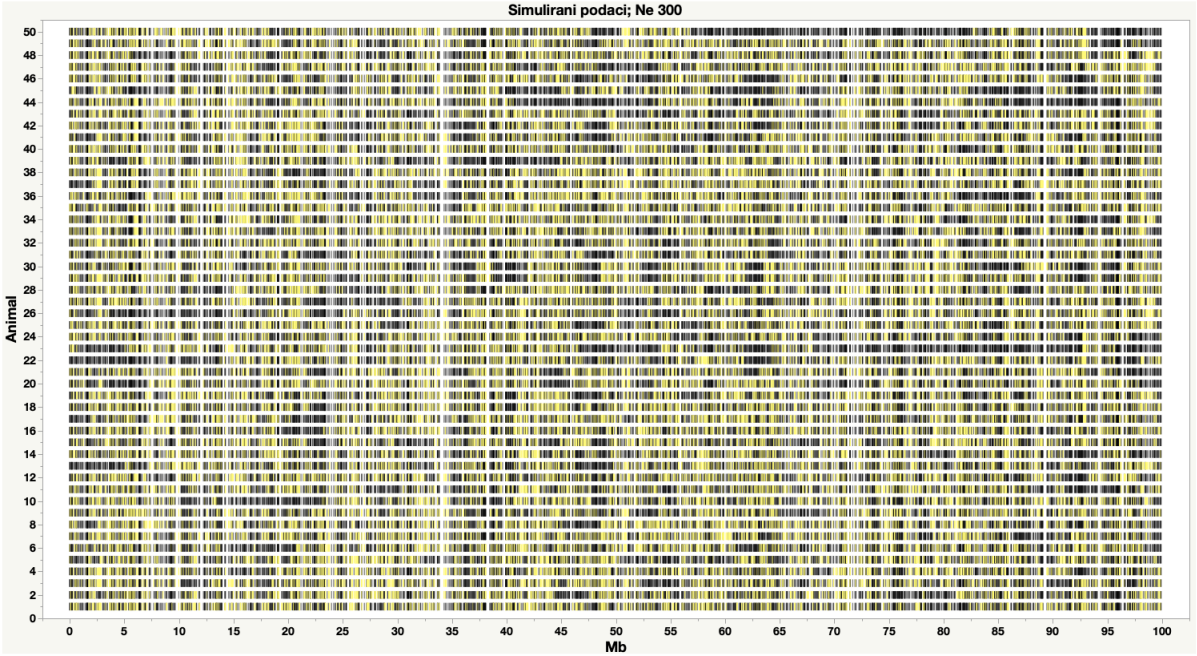
		FROH 1-2	FROH 2-4	FROH 4-8	FROH 8-16
cgaTOH	**prosjek**	0,100 [A]	0,062 [A]	0,036 [A]	0,014 [A]
	standardna devijacija	0,005	0,005	0,005	0,002
	minimum	0,091	0,051	0,027	0,009
	maximum	0,108	0,071	0,051	0,021
PLINK	**prosjek**	0,091 [B]	0,057 [B]	0,033 [B]	0,014 [A B]
	standardna devijacija	0,004	0,005	0,004	0,002
	minimum	0,084	0,048	0,025	0,009
	maximum	0,099	0,066	0,049	0,020
RZooROH	**prosjek**	0,031 [C]	0,041 [C]	0,028 [C]	0,013 [B]
	standardna devijacija	0,002	0,004	0,004	0,002
	minimum	0,026	0,033	0,021	0,009
	maximum	0,035	0,049	0,042	0,017

Tablica 4.9. Deskriptivna statistika i usporedba prosjeka broja ROH segmenata iz različitih ROH kategorija 30 simuliranih populacija s 1000 jedinki efektivne veličine 300 kroz 50 generacija procijenjenih različitim programskim paketima. Različita slova u koloni označuju statističku značajnost između srednjih vrijednosti.

		nseg 1-2	nseg 2-4	nseg 4-8	nseg 8-16
cgaTOH	**prosjek**	7,453 [A]	2,274 [A]	0,674 [A]	0,135 [A]
	standardna devijacija	0,358	0,184	0,084	0,024
	minimum	6,852	1,872	0,513	0,087
	maximum	8,165	2,599	0,952	0,200
PLINK	**prosjek**	6,828 [B]	2,080 [B]	0,629 [B]	0,129 [A B]
	standardna devijacija	0,311	0,179	0,080	0,023
	minimum	6,320	1,747	0,483	0,085
	maximum	7,471	2,399	0,909	0,189
RZooROH	**prosjek**	2,139 [C]	1,470 [C]	0,531 [C]	0,120 [B]
	standardna devijacija	0,148	0,130	0,071	0,019
	minimum	1,814	1,189	0,404	0,083
	maximum	2,421	1,724	0,779	0,172

U Tablici 4.8. za FROH 1-2, FROH 2-4 i FROH 4-8 postoji statistički značajna razlika prosjeka za svaki od navedenih programa. Za FROH 8-16 razlika postoji samo za prosjek dobiven programom RZooRoH. U Tablici 4.9. za nseg 1-2, nseg 2-4 i nseg 4-8 postoji statistički značajna razlika prosjeka za svaki od navedenih programa. Za nseg 8-16 javila se statistički značajna razlika samo za prosjek dobiven programom RZooRoH. U nastavku se nalaze figure za Ne=300 nakon 50 generacija.

A



B

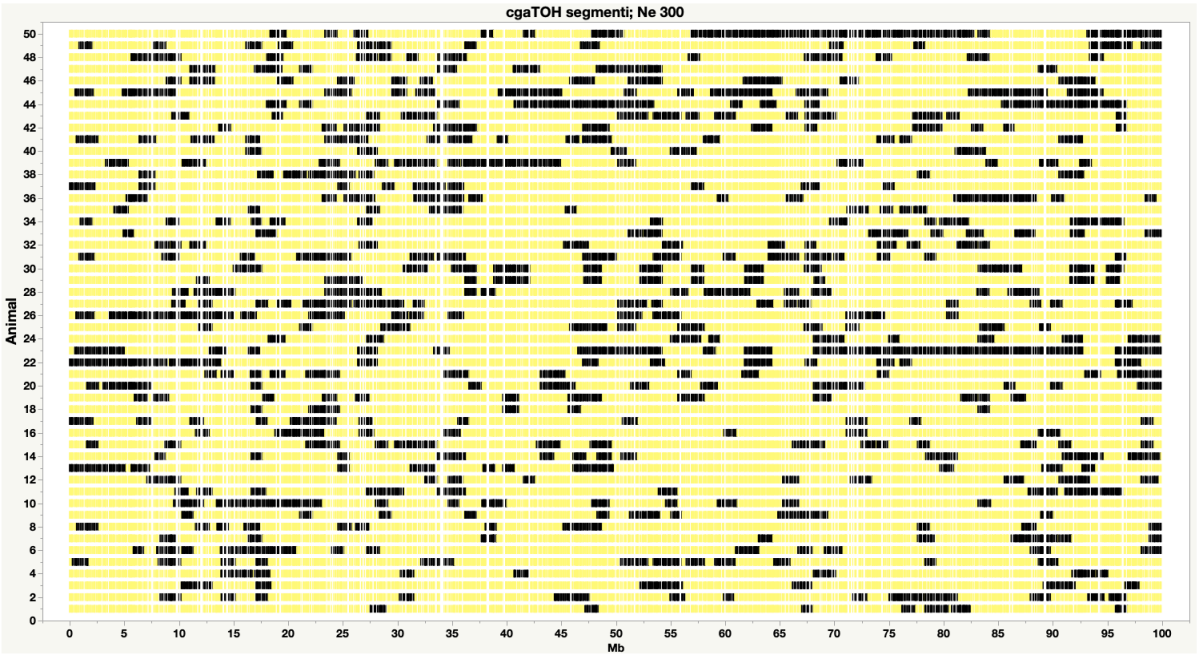
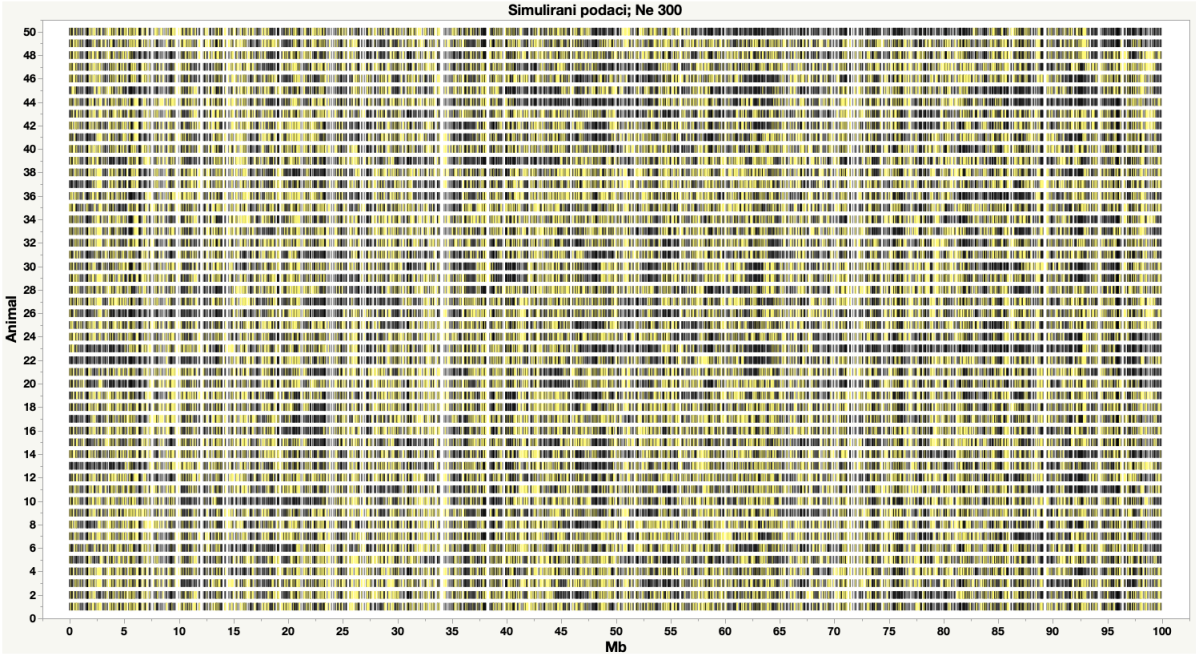


Figura 4.4. Usporedba statusa SNP-ova u simuliranim podacima (homozigot – crno, heterozigot – žuto) (A) s procijenjenim ROH segmentima (B) putem programskog paketa cgaTOH (crno). Prikazano je prvih 50 jedinki prve simulacije populacije s Ne = 300 nakon 50 generacija.

A



B

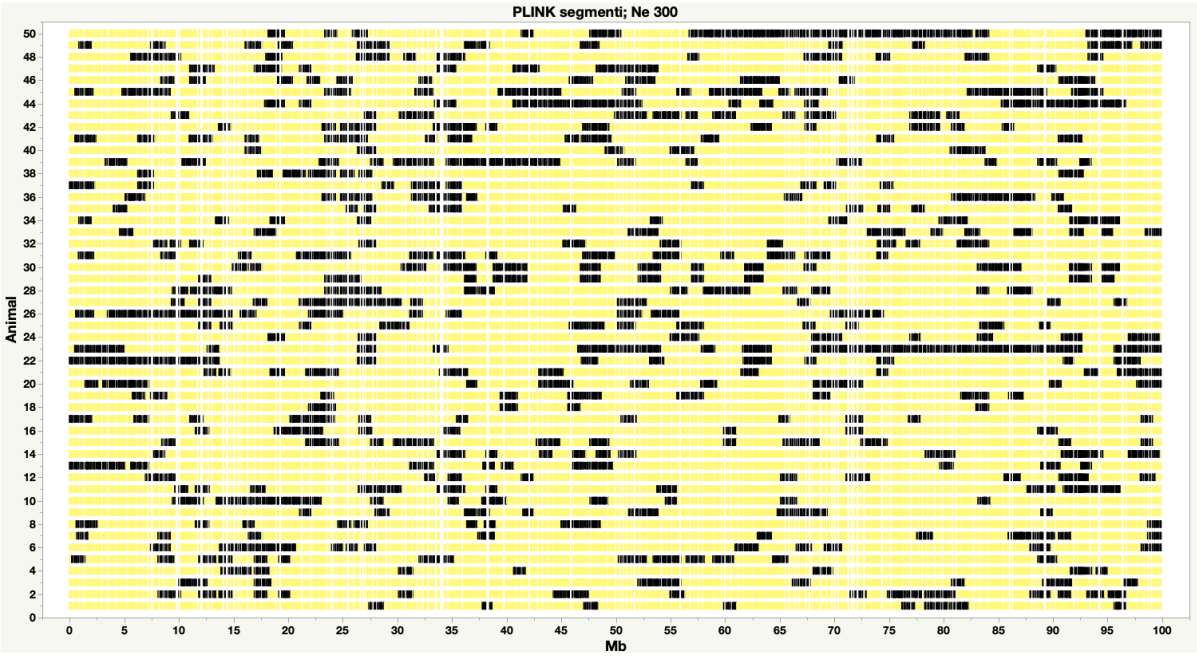
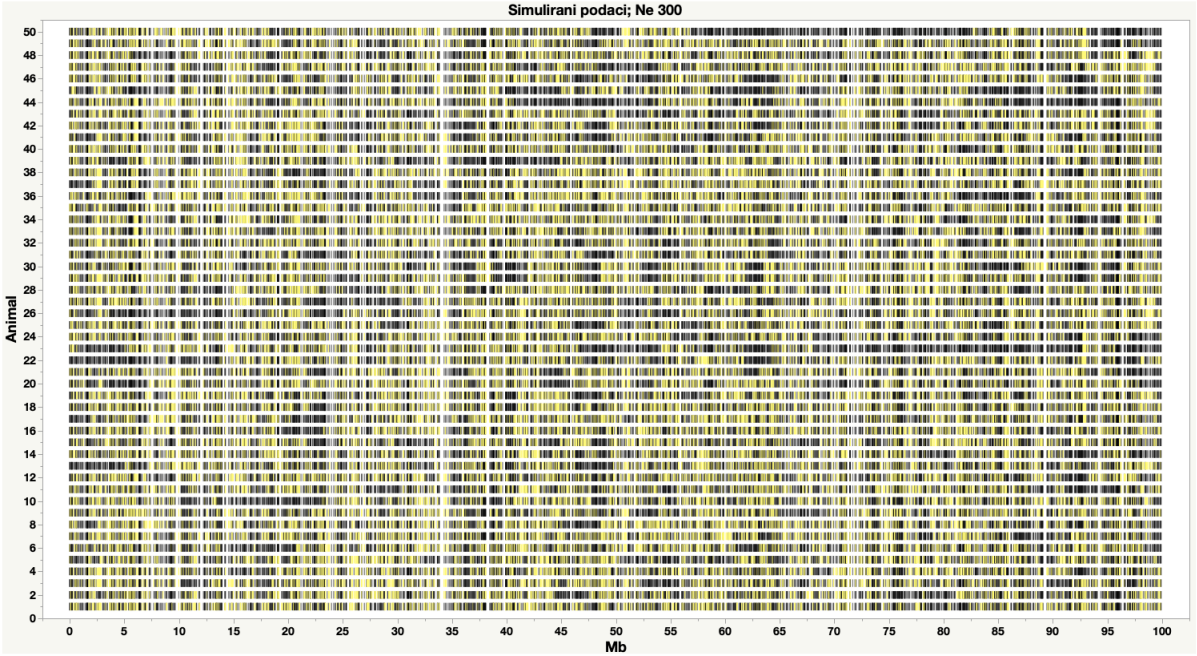


Figura 4.5. Usporedba statusa SNP-ova u simuliranim podacima (homozigot – crno, heterozigot – žuto) (A) s procijenjenim ROH segmentima (B) putem programskog paketa PLINK (crno). Prikazano je prvih 50 jedinki prve simulacije populacije s Ne = 300 nakon 50 generacija.

A



B

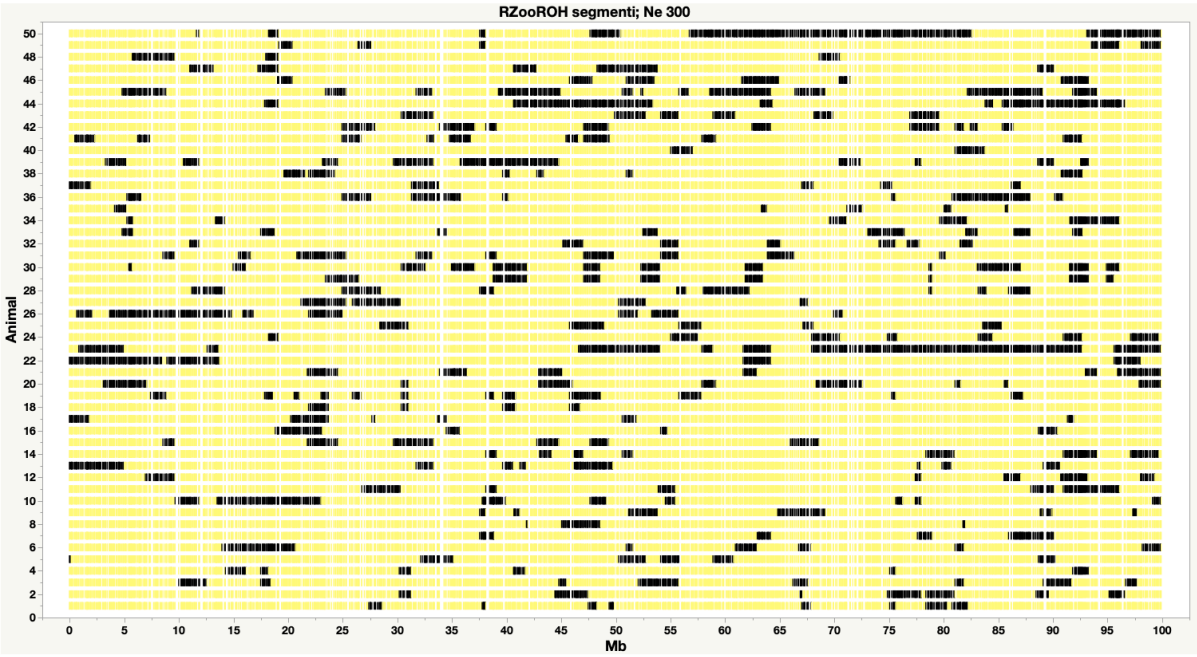


Figura 4.6. Usporedba statusa SNP-ova u simuliranim podacima (homozigot – crno, heterozigot – žuto) (A) s procijenjenim ROH segmentima (B) putem programskog paketa RZooROH(crno). Prikazano je prvih 50 jedinki prve simulacije populacije s Ne = 300 nakon 50 generacija.

U figurama za Ne=300 primjetno je manji broj homozigotnih dijelova (crno) što je i očekivano jer se očekuje manji udio autozigotnosti. Najmanje je vidljivih segmenata prikazano figurom 4.6. izrađenom korištenjem programa RZooRoh. Moguće je primijetiti kako RZooROH samo regije bez ubačenog heterozigota proglašava HBD dok cgaTOH i PLINK

regiju koja je prekinuta heterozigotom samo podijele na dva manja segmenta. To možemo povezati i s usporedbom prosjeka u tablicama. Prosjek dobivem programom RZooRoH se, u većini slučajeva, statistički značajno razlikovao od prosjeka dobivenih programima cgaTOH i PLINK.

Po zahtjevnosti upotrebe cgaTOH i PLINK su jednostavniji i puno brži u analizama. Kod RZooROH-a potrebno je imati određeno poznavanje programskog paketa „R“, pogotovo kada se koriste vlastiti podaci. Također je sam koncept HBD klasa koje treba unaprijed odrediti puno kompliciraniji od onog u cgaTOH-u i PLINKU, a uveden je i pojam „k-rates“ (R_k) koji služi za eksponencijalnu raspodjelu duljine HBD segmenata klase k pri čemu je prosjek 1/k. Klase koje imaju niži R_k zapravo su HBD-ovi veće duljine i dolaze od bližeg zajedničkog pretka i obrnuto. Time i RZooROH može odrediti izvor i starost ROH segmenta. Može se aproksimirati da je broj generacija do zajedničkog pretka $R_k/2$. Kod cgaTOH-a i PLINK-a određuje se samo duljina segmenta, a broj generacija se povezuje s očekivanom duljinom segmenta nakon određenog broja generacija. RZooROH ignorira monomorfne SNP-ove, a to prema Ferenčaković i sur. (2013) može negativno utjecati na detekciju segmenata nastalih zbog selekcije i/ili genetskog drifta.

Stoga je potrebno proširiti ovo istraživanje simulacijama u kojima bi se pratili IBD segmenti kroz generacije pod određenim evolucijskim silama i pod selekcijom i točno utvrdili. Potom bi bilo potrebno procijeniti mogućnosti RZooROH-a i drugih programskih paketa za detekciju tih segmenata.

5. Zaključci

1. Na 30 populacija s 500 jedinki dobivenih nakon 50 generacija i efektive veličine Ne = 25, uspješno su detektirani ROH segmenti sa sva tri programska paketa.

2. Usporedbe srednjih vrijednosti dobivenih inbreeding koeficijenata (FROH) populacija s Ne=25 pokazuju kako postoji statistički značajna razlika za vrijednosti dobivene putem RZooROH-a dok se cgaTOH i PLINK ne razlikuju statistički značajno i to za FROH>1, FROH>2, FROH>4 i FROH>8. Za FROH>16 nema razlike između programskih paketa.

3. Usporedbe srednjih vrijednosti broja ROH segmenata (nseg) populacija s Ne=25 pokazuju kako postoji statistički značajna razlika za sve vrijednosti dobivene putem RZooROH-a dok se cgaTOH i PLINK ne razlikuju statistički značajno.

4. Usporedbe prosjeka parcijalnih inbreeding koeficijenata populacija s Ne=25 pokazuju kako postoji statistički značajna razlika za sve vrijednosti dobivene putem RZooROH-a dok se cgaTOH i PLINK ne razlikuju statistički značajno. Isto vrijedi i za prosjeke brojeva segmenata iz parcijalnih koeficijenata.

5. Na 30 populacija s 1000 jedinki dobivenih nakon 50 generacija i efektive veličine Ne = 300, uspješno su detektirani ROH segmenti sa sva tri programska paketa.

6. Usporedbe srednjih vrijednosti dobivenih inbreeding koeficijenata (FROH) populacija s Ne=300 pokazuju kako postoji statistički značajna razlika između sva tri programska paketa i to za FROH>1 i FROH>2. Kod FROH>4 statistički značajno su različiti rezultati RZooROH-a. Kod FROH>8 cgaTOH se značajno razlikuje od RZooROH-a ali ne i od PLINK-a, dok se PLINK i RZooROH ne razlikuju statistički značajno.Za FROH>16 nema razlike između programskih paketa. Isto vrijedi i za prosjeke broja segmenata.

7. Usporedbe prosjeka parcijalnih inbreeding koeficijenata populacija s Ne=300 pokazuju kako postoji statistički značajna razlika za sve programske pakete za FROH 1-2, FROH 2-4 i FROH 4-8. Kod FROH 8-16 cgaTOH se značajno razlikuje od RZooROH-a ali ne i od PLINK-a, dok se PLINK i RZooROH ne razlikuju statistički značajno. Isto vrijedi i za prosjeke brojeva segmenata iz parcijalnih koeficijenata.

8. Potrebno proširiti ovo istraživanje simulacijama u kojima bi se pratili IBD segmenti kroz generacije pod određenim evolucijskim silama i pod selekcijom kako bi se točno utvrdili te onda ispitali programi za detekciju.

6. Popis literature

1. Albrechtsen A., Nielsen F. C. i Nielsen R. (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution. Oxford University Press, 27(11), pp. 2534–2547. doi: 10.1093/molbev/msq148.

2. Bjelland D., Weigel K., Vukasinovic N. & Nkrumah J. (2013). Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. J Dairy Sci 96, 4697-706. doi: 10.3168/jds.2012-6435.

3. Bosse M., Megens H.-J., Madsen O., Paudel Y., Frantz L., Schook L., Crooijmans R., Groenen M. (2012). Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. doi: 10.1371/journal.pgen.1003100

4. Broman K. W. i Weber J. L. (1999). Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. The American Journal of Human Genetics. Cell Press, 65(6), pp. 1493–1500. doi: 10.1086/302661.

5. Caetano A. R., Basso A. R., (2018). Runs of homozygosity for autozygosity estimation and genomic analysis in production animals. Pesq. agropec. bras., Brazil, v.53, n.9, p.975-984, doi: 10.1590/S0100-204X2018000900001

6. Ceballos F. C., Hazelhurst S., Ramsay M. (2018). Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage dana. BMC Genomics, Article number: 106 (2018) https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4489-0 pristupljeno: 18.4.2020.

7. Čurik I., Ferenčaković M. i Sölkner J. (2014). Inbreeding and runs of homozygosity: A possible solution to an old problem. Livestock Science, 166(1). doi: 10.1016/j.livsci.2014.05.034.

8. Čurik I., Ferenčaković M., Sölkner J. (2012). Modeling perspectives in the estimation of inbreeding depression based on genomic information: lessons from the bull fertility. In: Genetika 2012: book of abstracts/ 6th Congress of the Genetic Society of Slovenia [and] 6th Meeting of the Slovenian Society for Human Genetics with International Participation/Uroš Potočnik (Ed.). Ljubljana: Genetic Society of Slovenia, 2012. 82- 82.

9. Druet T., Bertrand A., Kadri N., Gautier M. (2019). The RZooRoH package. doi: 2019-03-21

10. Druet, T., i Gautier, M. (2017). A model-based approach to characterize individual inbreeding at both global and local genomic scales. Molecular ecology. 26(20), 5820-5841. doi: 10.1111/mec.14324.

11. Druet, T., i Gautier, M. (2017). A whole-genome-based approach for estimation and characterization of individual inbreeding. bioRxiv, 106765. doi: https://doi.org/10.1101/106765 Pristupljeno 21.5.2020.

12. Duret T. 2017. ZooROH_Manual. https://github.com/tdruet/ZooRoH/blob/master/ZooRoH_manual.pdf proistupljeno: 21.5.2020.

13. Faux A.-M., Gorjanc G., Gaynor R. C., Battagin M., Edwards S. M., Wilson D. L., Hickey J. M. (2016). AlphaSim: Software for Breeding Program Simulation. The Plant Genome, 9(3), 0. doi:10.3835/plantgenome2016.02.0013

14. Ferenčaković M., Hamzic E., Gredler B., Čurik I, Sölkner J. (2011). Runs of Homozygosity Reveal Genomewide Autozygosity in the Austrian Fleckvieh Cattle. Agriculturae Conspectus Scientificus, Vol. 76 No. 4, str. 325-329. https://hrcak.srce.hr/72108 pristupljeno: 20.4.2020.

15. Ferenčaković M., Sölkner J., Čurik I. (2013). Estimating autozygosity from high-throughput information: Effect of SNP density and genotyping errors. Genet. Sel. Evol. 45:42. https://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-45-42 pristupljeno: 18.4.2020.

16. Ferenčaković M., Sölkner j., Kapš M., Čurik I. (2017). Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. J. Dairy Sci. 100:4721–4730 https://www.journalofdairyscience.org/article/S0022-0302(17)30323-5/fulltext Pristupljeno: 18.4.2020.

17. Ferenčaković, M., Hamzić, E., Gredler, B., Solberg, T.R., Klemetsdal, G., Curik, I., Sölkner, J. (2013a): Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics, 130: 286-293. doi: http://dx.doi.org/10.1111/jbg.12012 6.

18. Ferenčaković, M., Solkner, J., Curik, I. (2013b): Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics Selection Evolution, 45: 42. doi: http://dx.doi. org/10.1186/1297-9686-45-42.

19. Gibson, J., Morton, N. E. i Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. Human Molecular Genetics. 15(5), pp. 789–795. doi: 10.1093/hmg/ddi493.

20. Gusev A., Lowe J. K., Stoffel M., i sur. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19: 318-326 originally published online October 29, 2008, doi:10.1101/gr.081398.108

21. Howrigan D. P. (2015). Genome-wide autozygosity is associated with lower general cognitive ability. Molecular Psychiatry (April), pp. 1–7. doi: 10.1038/mp.2015.120.

22. Kim E.-S., Cole J.B., Huson H., Wiggans G.R., Van Tassell C.P., Crooker B.A., Liu G., Da Y., Sonstegard T.S., (2013). Effect of Artificial Selection on Runs of Homozygosity in U.S. Holstein Cattle. PLoS ONE. 8, e80813. doi: 10.1371/journal.pone.0080813

23. Lencz T., Lambert C., DeRosse P., Burdick K. E., Morgan T. V., Kane J. M., Kucherlapati R. i Malhotra A. K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. 104(50), pp. 19942–7. doi: 10.1073/pnas.0710021104.

24. Leutenegger A. L., Prum B., Génin E., Verny C., Lemainque A., Clerget-Darpoux F., i Thompson E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. The American Journal of Human Genetics, 73(3), 516-523. doi: 10.1086/378207.

25. Magi A., Tattini L., Palombo F., Benelli M., Gialluisi A., Giusti B., Abbate R., Seri M., Gensini G. F., Romeo G. i sur. (2014). H 3 m 2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics, doi: 30(20):2852–2859.

26. McQuillan R., Leutenegger A., Abdel-Rahman R., Franklin C., Pericic M., Barac-Lauc L., Smolej-Narancic N., Janicijevic B., Polasek O., Tenesa A., Macleod A., Farrington S., Rudan P., Hayward C., Vitart V., Rudan I., Wild S., Dunlop M., Wright A., Campbell H., Wilson,J., 2008. Runs of homozygosity in European populations. Am. J. Hum. Genet. 83, 359 - 372. https://www.sciencedirect.com/science/article/pii/S000292970800445X pristupljeno: 20.4.2020.

27. Narasimhan V., Danecek P., Scally A., Xue Y., Tyler-Smith C., Durbin R. (2016). Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics, doi: 32(11):1749–1751.

28. Purfield D. C., Berry D. P., McParland S., Bradley D. G. (2012). Runs of homozygosity and population history in cattle. BMC Genetics. 13(1), 70. doi:10.1186/1471-2156-13-70

29. Silió L., Rodríguez M.C., Fernández A., Barragán C., Benítez R., Óvilo C., Fernández A.I., (2013). Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. J. Anim. Breed. Genet .170, 349-360. doi: 10.1111/jbg.12031.

30. Sölkner J., Ferenčaković M, Gredler B., Čurik I. (2010). Genomic metrics of individual autozygosity, applied to a cattle population.. Page 306 in 61st Annual Meeting of the European Association of Animal Production, Heraklion. Wageningen Academic Publishers, Wageningen, the Netherlands

31. Vieira F. G., Albrechtsen A., Nielsen R. (2016). Estimating ibd tracts from low coverage ngs data. Bioinformatics, doi: 32(14):2096–2102.

32. Wigginton J. E., Cutler D. J. i Abecasis G. R. (2005). A Note on Exact Tests of HardyWeinberg Equilibrium. The American Journal of Human Genetics. Cell Press, 76(5), pp. 887– 893. doi: 10.1086/429864.

33. Zhang L., Orloff M. S., Reber S., Li S., Zhao Y., Eng, C. (2013). cgaTOH: Extended Approach for Identifying Tracts of Homozygosity. PLoS ONE, 8(3), e57772. doi:10.1371/journal.pone.0057772

7. Životopis

Kristina Strelar rođena je 21.5.1995. u Zagrebu. Svoje obrazovanje započela je u OŠ Sesvetsk Kraljevec. Pohađala je opću gimnaziju u SŠ Dugo Selo. Oduvijek je imala veliki interes prema životinjama što ju je potaknulo da, po završetku srednjoškolskog obrazovanja, upiše Agronomski fakultet Sveučilišta u Zagrebu. Stekla je titulu baccalaurea Animalnih znanosti nakon čega upisuje diplomski studij, na istom fakultetu, smjer Genetika i oplemenjivanje životinja. Za vrijeme studija stekla je certifikat *Contribution of Animal Breeding to Global Food Security* (*University of Natural Resources and Life* Sciences, Beč, Austria). Pohađala je Modul "Upravljanje ljudskim resursima", (Akademija za industrijski razvoj) te stekla još jedan certifikat. Kristina trenutno radi praksu u konzultantskoj firmi Apsolon.

8. Prilog

8.1. AlphaSimR

```
#######################################
 1. POSTAVLJANJE RADNOG DIREKTORIJA
#######################################

setwd("C:/Users/krist/Desktop/Ne24,5")
getwd()

##################################
2. UĆITAVANJE POTREBNIH PAKETA
##################################
library("R6")
library("AlphaSimR")
library("tidyverse")

##################################################
3. UCITAVANJE POTREBNIH FUNKCIJA ZA SIMULACIJU
##################################################

CreateAndSetScenarioFolder <- function(x) {
  # Create folder and set it as a working directory
  # x - character, folder name
  if (dir.exists(paths = x)) {
    unlink(x = x, recursive = TRUE)
  }
  dir.create(path = x)
  setwd(dir = x)
}

DetermineGenerationFromBase = function(x, Unknown = "0") {
  # Determine how many generations is an individual removed
from the base of a pedigree
  # x - pedigree table (tibble or data.frame) with individual,
father, and mother columns;
  #     with 0 as unknown parent; the pedigree is assumed
sorted such that parents
  #     preceede progeny and that pedigree is extended = that
all parents are listed
  #     as individuals)
  # unknown - a value used to denote unknown parent
  if (is.matrix(x)) {
    stop("x must be tibble or data.frame!")
  }
  nInd = nrow(x)
  Generation = rep(x = NA, times = nInd)
```

```
  FIDVec = match(x = x[[2L]], table = x[[1L]], nomatch = 0L)
  MIDVec = match(x = x[[3L]], table = x[[1L]], nomatch = 0L)
  FatherIsUnknownVec = x[[2L]] %in% Unknown
  MotherIsUnknownVec = x[[3L]] %in% Unknown
  for (Ind in 1L:nInd) {
    # Ind = 1L
    FID = FIDVec[Ind]
    MID = MIDVec[Ind]
    FatherIsUnknown = FatherIsUnknownVec[Ind]
    MotherIsUnknown = MotherIsUnknownVec[Ind]
    if (FatherIsUnknown & MotherIsUnknown) {
      Generation[Ind] = 0.0
    } else if (!FatherIsUnknown & MotherIsUnknown) {
      Generation[Ind] = (Generation[FID] + 1.0) / 2.0
    } else if (FatherIsUnknown & !MotherIsUnknown) {
      Generation[Ind] = (Generation[MID] + 1.0) / 2.0
    } else {
      Generation[Ind] = ((Generation[FID] + 1.0) +
(Generation[MID] + 1.0)) / 2.0
    }
  }
  Generation
}

IbdInbreeding = function(x) {

  Names = rownames(x) %>%
    strsplit(split = "_") %>%
    sapply(FUN = function(z) z[[1L]]) %>%
    as.integer()
  nInd = nrow(x) / 2L
  nLoc = ncol(x)
  Ret = data_frame(IId = rep(x = 0L, times = nInd),
                   IbdInb = 0.0)
  for (Ind in 1L:nInd) {

    Ret$IId[Ind] = Names[2L * Ind - 2L + 1L]

    Ret$IbdInb[Ind] = sum(x[2L * Ind - 2L + 1L,  ] == x[2L *
Ind - 2L + 2L,  ]) / nLoc
  }
  Ret
}

RohInbreeding = function(x, RohLengthThreshold, GenomeLength)
{

  x %>%
    subset(lengthBps >= RohLengthThreshold) %>%
    group_by(id) %>%
    summarise(RohInb = sum(lengthBps) / GenomeLength)
```

```
}

###############
  ARGUMENTI
###############
Args = commandArgs(trailingOnly = TRUE)

Args = c(1, "Base", "SceA", "SceB")
if (length(Args) < 2) {
  stop("Must provide replicate number (1, 2, ...) as the first
argument and\n
       scenario (Base, SceA, SceB, SceC, and SceD) as the
second argument!")
}
Rep = Args[1]
Sce = Args[-1]


#########################################
 4. POSTAVLJANJE PARAMETARA SIMULACIJE
#########################################

Species = "CATTLE"
nChr = 1
nSnpPerChr = 2000
nQtlPerChr = 100
nGen       =   50
nIndPerGen = 500

nSires=7
nDams=49


#########################################################
 5. PROVJERA EFEKTIVNE VELICINE POP ZA POSTAVLJENE PARAMETRE
#########################################################

4 * nSires * nDams / (nSires + nDams)

MeanGInitial = 0
VarGInitial = 1
VarE = 3


#########################################################
6. SIMULACIJA FOUNDER (FounderPop) I BAZNE (BasePop)
POPULACIJE
#########################################################

if ("Base" %in% Sce) {
  ScenarioName = paste("Rep_", Rep, sep = "")
```

```
  CreateAndSetScenarioFolder(x = ScenarioName)

  FounderPop = runMacs(nInd = nIndPerGen,
                       nChr = nChr,
                         segSites = nSnpPerChr + nQtlPerChr,
                      species = Species)
  SP = SimParam$new(founderPop = FounderPop)
  SP$setGender(gender = "yes_rand")
  SP$addTraitA(nQtlPerChr = nQtlPerChr, mean = MeanGInitial,
var = VarGInitial)
  SP$addSnpChip(nSnpPerChr = nSnpPerChr)
  SP$setTrackPed(isTrackPed = TRUE)
  SP$setTrackRec(isTrackRec = TRUE)
  BasePop = newPop(rawPop = FounderPop)
  BasePop = setPheno(pop = BasePop, varE = VarE)

  save.image(file = "Data.RData")

  setwd(dir = "..")
  cat("DONE with", Sce, "\n")
}


#########################################################
```

7. SIMULACIJA POPULACIJA

```
 - podijeljena je u 3 (12+12+6) dijela zato što widows ne može
napraviti više od 12 poddatoteka, #
ovdje radim 30 populacija s 50 generacija u svakoj populaciji,
zato gen in 1:50
#########################################################

Pop = BasePop

for(Rep in 1:12) {

  ScenarioName = paste("Rep_", Rep, "_Scenario_A", sep = "")
  CreateAndSetScenarioFolder(x = ScenarioName)

  for (Gen in 1:50) {

    Sires = selectInd(pop = Pop, nInd = nSires, use = "rand",
gender = "M")
    Dams  = selectInd(pop = Pop, nInd = nDams,  use = "rand",
gender = "F")
    Pop = randCross2(females = Dams, males = Sires, nCrosses =
nIndPerGen,
                     nProgeny = 1, balance = FALSE)
  }
  writePlink(pop = Pop, baseName = paste("Rep_", Rep, "_50",
sep = ""))
```

```
  Pop=BasePop

}

save(Pop, file = "C:/Users/krist/Desktop/Ne24,5/pop50")

setwd("C:/Users/krist/Desktop/Ne24,5")


for(Rep in 13:24) {

  ScenarioName = paste("Rep_", Rep, "_Scenario_B", sep = "")
  CreateAndSetScenarioFolder(x = ScenarioName)

  for (Gen in 1:50) {

    Sires = selectInd(pop = Pop, nInd = nSires, use = "rand",
gender = "M")
    Dams  = selectInd(pop = Pop, nInd = nDams,  use = "rand",
gender = "F")
    Pop = randCross2(females = Dams, males = Sires, nCrosses =
nIndPerGen,
                     nProgeny = 1, balance = FALSE)
  }
  writePlink(pop = Pop, baseName = paste("Rep_", Rep, "_50",
sep = ""))

  Pop=BasePop
}


setwd("C:/Users/krist/Desktop/Ne24,5")


for(Rep in 25:30) {


  ScenarioName = paste("Rep_", Rep, "_Scenario_C", sep = "")
  CreateAndSetScenarioFolder(x = ScenarioName)

  for (Gen in 1:50) {

    Sires = selectInd(pop = Pop, nInd = nSires, use = "rand",
gender = "M")
    Dams  = selectInd(pop = Pop, nInd = nDams,  use = "rand",
gender = "F")
    Pop = randCross2(females = Dams, males = Sires, nCrosses =
nIndPerGen,
                     nProgeny = 1, balance = FALSE)
  }
```

```
  writePlink(pop = Pop, baseName = paste("Rep_", Rep, "_50",
sep = ""))

  Pop=BasePop
}


#######################################################
 8. EKSPORTIRANJE BAZNE POPULACIJE U PLINK OBLIKU (ped i map
datoteke)
#######################################################


Gen = 0
writePlink(pop = BasePop,
           baseName = paste0("Data",
                             formatC(x = Gen, flag = "0",
width = nchar(nGen))))
```

8.2. PLINK

```
./plink --chr-set 1  --homozyg-density 100 --homozyg-gap 1000
--homozyg-kb 1000 --homozyg-snp 15 --homozyg-window-het 1 --
homozyg-window-missing 2  --homozyg-window-snp 15 --homozyg-
het 1 --file Rep_2_50 --out R_2_50
```

8.3. RZooRoH

```
library(RZooRoH)
setwd("C:/Users/krist/Desktop/Ne300")
getwd()
file_1_ <-
system.file("exdata","Rep_1_50.gen",package="RZooRoH")
R_1_50 <- zoodata(genofile = file_1_, zformat = "gp")
model <- zoomodel(predefined=FALSE,K=4,base=5)
R_1_50_m <- zoorun(model, R_1_50)
write.csv(R_1_50_m@hbdseg,"C:/Users/krist/Desktop/Ne300/R_1_50
_r.csv",row.names = FALSE)
```

8.4. cgaTOH

```
./TOH_ClusteringSuite_v1_0 -map Rep_1_50 -p Rep_1_50 -o
R_1_50_c -l 15 -min_length 1000000 -max_gap 1000000 -
max_missing 2 -max_hetero 1 -skip_clustering -force_proceed
```