



UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I
INFORMATIKU



Igor Stranjanac

Regresione metode u QTL analizi diploidnih i tetraploidnih ruža

Master rad

Mentor:

prof. dr Zagorka Lozanov-Crvenković

Novi Sad, 2015

Sadržaj

Predgovor	2
1 Osnovni pojmovi	4
1.1 Uvod	4
1.2 Pregled pojmoveva iz genetike	5
1.2.1 Osnovni elementi genetičke analize	5
1.2.2 Genetičke mape	12
1.3 Pregled statističkih pojmoveva	14
1.3.1 Osnovni pojmovi	14
1.3.2 Testiranje hipoteza	18
1.3.3 Linearni modeli	19
1.3.4 Metoda maksimalne verodostojnosti	20
1.4 Molekularni markeri	21
1.4.1 SNP (Single Nucleotid Polymorphism) markeri	23
1.5 Poliploidni	23
1.6 Obrasci nasleđivanja tetraploidnih organizama	25
1.7 Oplemenjivanje ruža	26
1.8 Morfologija ruža	27
1.9 Molekularni makeri u oplemenjivanju	28
2 Statistički modeli QTL analize	30
2.1 Pronalaženje QTL-a	30
2.2 Metode za detekciju QTL-ova	32
2.2.1 Model sa jednim QTL-om	33
2.2.2 Analiza pojedinačnog markera	36

2.2.3	Intervalno mapiranje	41
2.3	Permutacioni test	53
3	Numerički rezultati	56
3.1	Implementacija modela	56
3.1.1	Model 1	56
3.1.2	Model 2	61
3.1.3	Model 3	63
3.2	Modeli interakcije - primena višestruke regresije	65
4	Zaključak	67
	Literatura	69
	Biografija	71

Predgovor

QTL (eng. *Quantitative trait loci*) analiza predstavlja metod identifikacije gena koji se koristi u genetskoj epidemiologiji i oplemenjivanju biljaka i životinja. Primenom navedenog metoda uzima se uzorak DNK-a iz koga se čitaju genetičke faktori koristeći molekularne markere. Ove analize imaju za cilj identifikovanje markera, skupa markera ili regiona iz mnoštva zajedničkih genetičkih faktora različitih individua koji su smatraju odgovornim za ispoljavanje osobine od interesa. Na taj način se dobijaju markeri koji se nalaze u blizini određenih gena koji se dalje smatraju uticajnim za ispoljavanje fenotipske karakteristike. QTL analiza može biti usmerena na utvrđivanje pojedinačnih markera ili čitavih regiona hromozoma.[2]

U prvoj glavi rada definisće se osnovni pojmovi iz oblasti statističke i genetičke analize. U drugoj glavi objasniće se najjednostavniji oblik QTL analize, a zatim će se obraditi moguće modifikacije metode. Osobina od interesa koja će se obraditi u ovom radu je otpornost na niske temperature.

Treća glava rezervisana je za poređenje rezultata koji su dobijeni primenom različitih metoda. Zatim se rezultati analiziraju i proverava se njihova validnost. Primenom višestruke regresije i grafičkim prikazom rezultata proveravaju se međusobni efekti dobijenih regiona na hromozomu, kako bi se ustanovili efekti na fenotipsku osobinu.

U četvrtoj glavi sumiraće se rezultati analize. U zaključku će se sumirati rezultati rada i dati predlozi značajnosti za dalji razvoj i modifikaciju modela.

Zahvaljujem se mentoru prof. dr Zagorki Lozanov-Crvenković na svim sugestijama i stručnom usmeravanju, kao i razumevanju prilikom izrade ovog rada. Takođe, zahvaljujem se članovima komisije, prof. dr Ivani Štajner-Papuga i prof. dr Nataši Krejić na korekcijama i savetima vezanim za poboljšanje rada, kao i svim ostalim profesorima

i asistentima sa kojima sam sarađivao tokom osnovnih i master akademskih studija. Želeo bih da se zahvalim i svim kolegama iz Pheno Geno tima na čelu sa direktorkom Biljanom Božanić Tanjga i Piterom Koksom, koji su mi, prilikom mog praktičnog rada u njihovoj firmi, približili ovu temu, a posebno bih izdvojio dr Mirjanu Vukosavljev koja mi je, svojim savetima i sugestijama, mnogo pomogla u izboru i razumevanju ove teme.

Posebnu i najveću zahvalnost dugujem svojim roditeljima, Miroslavu i Eriki, i bratu Miloradu na beskrajnoj podršci i razumevanju u toku školovanja i života.

1 Osnovni pojmovi

1.1. Uvod

Od početka XX veka, matematičko modeliranje je odigralo odlučujuću ulogu u stvaranju i razvoju genetike kao naučne discipline. Začetnikom klasične genetike, koja se bavi mehanizmima prenošenja naslednih faktora sa roditelja na potomstvo, smatra se austrijski biolog, botaničar i matematičar Gregor Johan Mendel (1822-1884), koji je svojim radovima i zaključcima, koji su inače bili daleko ispred svog vremena, postavio temelje savremene genetike. Tek su sa početkom XX veka genetičari uspeli da povežu Mendelova istraživanja (zakonitosti nasleđivanja) sa Darwinovom teorijom evolucije. Iako Mendelovi zaključci predstavljaju osnovne zakone nasleđivanja, oni su svedeni na osobine koje zavise samo od uticaja jednog gena. Za mnoge osobine od interesa koje se javljaju prilikom oplemenjivanja raznih vrsta poljoprivrednih dobara je ustanovljeno da zavise od uticaja više različitih gena. Usled razvoja raznih molekularnih tehnika koje obezbeđuju podatke o genetskim varijacijama, dolazimo i do razvoja novih statističkih modela i algoritama na osnovu kojih se razne kvantitativne osobine povezuju sa odgovarajućim skupom markera (gena).

Kao što je poznato, za razne bolesti i zdravstvene probleme koji se javljaju u svetu smatra se da su nasledni. Takođe, jasno je da neke bolesti nisu povezane samo sa jednim genom, već zavise od kombinacije i međusobne interakcije istih. Kao i bolesti, razne osobine poput visine i težine, zavise od uticaja više gena i takve osobine se nazivaju *poligenske*.

Dakle, jedan od razloga zbog čega je poželjno pronaći gene koji utiču na određenu osobinu su povećanje prinosa poljoprivrednih proizvoda izborom odgovarajućih biljnih i

životnjskih vrsta, kao i lečenje raznih bolesti kod živih bića. Međutim, zbog kompleksne interakcije više gena i životne sredine, problem pronalaženja gena čini teškim za rešavanje.

Efekti kombinacije određenih gena se ne mogu uočiti gledajući individualni rezultat istih; kao što se ni kvalitet fudbalske ekipe ne može oceniti posmatrajući samo individualne kvalitete igrača. Uticaj životne sredine se može kontrolisati prilikom pronalaženja odgovornih gena kod biljaka i životinja, ali kod ljudi taj uticaj nije moguće redukovati, te se stoga eksperimenti vrše na miševima i pacovima, iz razloga što postoji veliki broj gena u ljudskom genomu koji je zastupljen i kod miševa i pacova.

1.2. Pregled pojmove iz genetike

1.2.1. Osnovni elementi genetičke analize

Radi lakšeg razumevanja, u ovom delu objasnićemo neke osnovne elemente genetičke analize koji će se koristiti u daljem radu.

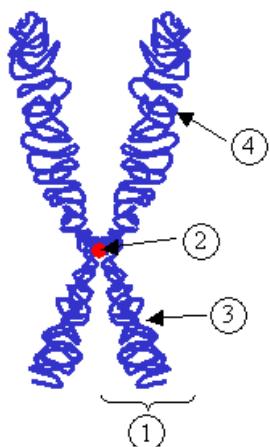
Živa bića se sastoje od ćelija, a ćelije predstavljaju osnovnu jedinicu života. Dezok-siribonukleinska kiselina (DNK) je nukleinska kiselina koja sadrži uputstva za razvoj i pravilno funkcionisanje svih živih organizama. Sva živa bića svoj genetički materijal nose u obliku DNK. Ćelije sadrže DNK organizovan u duge strukture koje se zovu *hromozomi*. Dakle, hromozom predstavlja dva niza DNK sekvene koji imaju oblik dvostrukog heliksa. *Hromatida* je uzdužna polovina hromozoma, što se može videti na grafiku 1.1. Hromatide mogu biti:

- sestrinske - to su hromatide jednog hromozoma gde su njegove uzdužne polovine spojene centromerom;
- nesestrinske - to su hromatide para homologih hromozoma.

Svaki organizam ima drugačiji set hromozoma. Set hromozoma u organizam naziva se *genom*. Za primer se može uzeti ljudski genom, koji se sastoji od 46 hromozoma: 22 para homologih hromozoma i jednog para polnih hromozoma koji nisu homologi.

Homolog predstavlja par hromozoma koji sadrže identičan niz gena, na istim pozicijama na hromozomu, svaki dobijen od jednog roditelja.

Biljke, životinje i ljudi su najčešće diploidni organizmi, što znači da se svaki tip hromozoma pojavljuje u paru koji se naziva homolog, gde je jedan deo para nasleđen od muškog roditelja, a drugi od ženskog roditelja. Prilikom prenosa genetskog materijala na potomstvo, članovi homologa segregiraju i prenose genetski materijal sa jednakom verovatnoćom, u skladu sa Mendelovim zakonom segregacije, o kom će kasnije biti reči.



Grafik 1.1: Hromozom sa svojim elementima: 1. hromatida, 2. centromera, 3. kratki krak, 4. dugi krak

Genom predstavlja skup svih gena u organizmu. *Geni* predstavljaju osnovnu jedinicu nasleđivanja, nanizanu duž DNK. Praktično gledano, gen predstavlja sekvencu nukleotida DNK koja prenosi naslednu informaciju. Kao što hromozomi idu u paru, isto tako i gen ima svoja dva homologna oblika, ukoliko pričamo o diploidnim organizmima. Homologni oblik gena naziva se *alel*. Kao što nasleđujemo po jedan hromozom od oba roditelja, tako se i po jedan alel nasleđuje od svakog roditelja. Gen može da ima dva ili više alela, koji se nalaze na paru homolognih hromozoma. U užem smislu, gen se sastoji od tačno dva alela, s time da svaki gen ima svoj određenu poziciju na hromozomu. Pozicija gena na hromozomu naziva se *genski lokus*.

Kombinacija dva alela nekog gena naziva se *genotip* i predstavlja definiciju genotipa u užem smislu. Genotip u širem smislu predstavlja gensku konstituciju nekog organizma koja se odnosi na celovitu naslednu osnovu, tj. sve gene koje taj organizam pose-

duje. Aleli u genotipu mogu da budu isti (*homozigotni genotip*) i različiti (*heterozigotni genotip*).

Kao i genotip, fenotip se može posmatrati u užem i širem smislu. U užem smislu, *fenotip* predstavlja jednu, vidljivu osobinu organizma koja nastaje usled interakcije genotipa i uslova životne sredine. Fenotip u širem smislu predstavlja skup svih osobina jednog organizma koje su nastale zajedničkim delovanjem genotipa i uslova sredine u kojima se dati organizam razvija.

Fenotipske osobine se najgrublje mogu podeliti na *kvalitativne* (oblik cveta, boja cveta, tip ruže) i *kvantitativne* (visina biljke, težina latica, prečnik cveta). Sve što se dešava u okolini organizma nakon prve deobe ćelija nazivamo uslovima životne sredine. Treba napomenuti da jedinke koje imaju identičan genotip mogu imati totalno drugačije fenotipove, dok sa druge strane, što su uslovi sredine sličniji, to će fenotipovi biti sličniji.

Kvantitativne osobine su osobine organizma koja mogu da se mere na neprekidnoj, kontinualnoj skali, kao što su visina, težina, vreme cvetanja. To su one osobine koje variraju neprekidno u populaciji, i mnoge kvantitativne osobine prate normalnu raspodelu. Diskretne (kvalitativne) osobine mogu biti sortirane po utvrđenim pravilima i najčešće zavise samo od jednog gena, dok kvantitativne osobine zavise kako od više gena, tako i od njihove međusobne interakcije.[11]

Ako uticaj sredine ima mali efekat na osobinu od interesa, u odnosu na genotip, potrebno je više gena da bismo imali neprekidnu raspodelu, dok u slučaju kada je uticaj okruženja visok, tada osobine mogu da budu neprekidnog tipa, čak i ako mali broj gena utiče na ispoljavanje osobine.[11]

Na fenotype utiču aleli, i to na sledeće načine:

- dominantno,
- recesivno,
- kodominantno,
- izrazito dominantno.

Alel u genotipu ne mora nužno da se manifestuje kroz fenotip. Alel koji maskira postojanje drugog se naziva *dominantan* u odnosu na taj drugi alel, koji se naziva *recesivnim* u odnosu na dominantni alel. Ukoliko se oba alela ogledaju u fenotipu, tada za njih kažemo da su *kodominantni* jedno drugom. Ukoliko je uticaj heterozigotnog genotipa na fenotip veći nego uticaj homozigotnog genotipa, tada je homozigotni genotip određen aleлом čiji je uticaj veći od homozigota sa drugim aleлом, *izrazito dominantan* u odnosu na drugi alel. Možemo definisati pojam *doza alela* koji se odnosi na broj dominantnih ili recesivnih alela u genu. Narednim primerima će se čitaocu pokušati približiti načini uticaja alela na fenotip.

Primer 1.1. Krvna grupa kod ljudi predstavlja dobar primer uticaja alela na fenotip. U genu postoje tri alela: A, B i O koji utiču na antigen u crvenim krvnim ćelijama. Neka su antigeni označeni sa A,B i O. Ako individua poseduje genotip AO ili BO, tada samo antigen A ili B mogu da se prepoznaju u krvi. Postojanje antiga A ili B maskira postojanje O, što dovodi do zaključka da su A i B dominantni u odnosu na O, odnosno O je recesivan u odnosu na A i B. Ukoliko individua poseduje genotip AB, tada oba antiga (i A i B) mogu da se prepoznaju krvi. Stoga su antigeni A i B kodominantni jednog drugom.

Primer 1.2. Recimo da je boja cveta određena genotipom na jednom lokusu. Ako pretpostavimo da genotip *aa* daje bele cvetove i genotip *AA* daje crvene cvetove i genotip *Aa* daje crvene cvetove, tada kažemo da je alel *A* u potpunosti dominantan u odnosu na alel *a*.

U daljem toku rada dominantni aleli će se označavati sa *A*, dok će se recesivni označavati sa *a*.

Genska interakcija predstavlja situaciju u kojoj više gena utiče na fenotip. *Epistaza* predstavlja zavisnost (interakciju) alela (gena) u cilju povećanja efekta na fenotip. Epistazom nazivamo devijaciju aditivnosti između različitih lokusa, tj. epistaza se javlja kada je ukupan uticaj više lokusa različit od zbira individualnih efekata. Ona je veoma česta i bitna, jer geni predstavljaju kodirane supstance koje međusobno interakuju i promene u jednoj supstanci mogu da utiču na ceo lanac reakcija.

Mendelova istraživanja su imala ogroman uticaj na čitav razvoj genetike, te zbog toga navodimo njegov prvi zakon nasleđivanja:

Definicija 1.1. Mendelov prvi zakon nasleđivanja ili zakon segregacije: Paran broj alela za neko svojstvo razdvaja se ili segregira tokom stvaranja gameta tako da svaki gamet dobije po polovinu početnog broja alela. Fizička osnova segregacije alela je razdvajanje homolognih hromozoma.[13]

Primer 1.3. U slučaju diploidnih organizama, neka su oba roditelja heterozigoti na određenom markeru M (genotip je oblika Aa). Prema Mendelovom zakonu segregacije, genotip potomstva na markeru M može biti AA , Aa ili aa . Verovatnoća da je genotip potomstva AA je:

$$p(AA) = \frac{\binom{1}{1}}{\binom{2}{1}} \cdot \frac{\binom{1}{1}}{\binom{2}{1}} = \frac{1}{4}.$$

Dakle, od oba roditelja biramo dominantne alele. Analogno, verovatnoća da je genotip potomstva aa je verovatnoća da od oba roditelja biramo recesivne alele i on je $p(aa) = \frac{1}{4}$. Verovatnoća da se pojavi genotip potomstva oblika Aa dobija se tako što od jednog roditelja se prenese dominantni alel, a od drugog recesivni. Dakle,

$$p(Aa) = \frac{\binom{1}{1}}{\binom{2}{1}} \cdot \frac{\binom{1}{1}}{\binom{2}{1}} + \frac{\binom{1}{1}}{\binom{2}{1}} \cdot \frac{\binom{1}{1}}{\binom{2}{1}} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Osnovu mapiranja gena i QTL analize predstavlja odnos gena i osobina koji otkriva da varijacije u genotipu jednog gena utiču na varijaciju fenotipa istog. Mapiranje lokusa za kvantitativne osobine u suštini predstavlja pretragu gena čija varijacija u genotipu može da objasni varijaciju u kvantitativnoj osobini od interesa.

Lokus za kvantitativna osobine (engl. *Quantitative Trait Locus - QTL*) predstavlja lokus čiji genotip utiče na ispoljavanje kvantitativne osobine. U daljem radu, lokuse za kvantitativna svojstva nazivaćemo QTL-ovima.

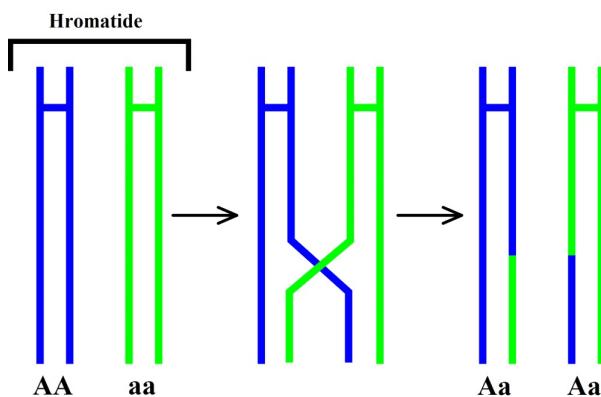
U slučaju diploidnih organizama, roditelji prenose na potomstvo samo jedan homologni hromozom. Isto tako roditelji nasleđuju po jedan homologni hromozom od svojih roditelja. Dakle, ne zna se od kog pretka je potomstvo navelio navedeni hromozom, što nas dovodi do pojma rekombinacije gena.

Uzrok rekombinacije je događaj pod nazivom **krossover** (eng. *crossover*), koji predstavlja razmenu genetičkog materijala između nesestrinskih hromatida homolognih hromozoma. Drugim rečima, krossover je događaj koji se dešava između lokusa prilikom

deljenja ćelija kako bi se dobili gameti. *Gameti* su polne ćelije koje se stvaraju prilikom deobe ćelije i one sadrže *haploidan* broj hromozoma. Pojam haploidan podrazumeva upola manji broj hromozoma od broja hromozoma roditelja. Dakle, hromozomi se u gametima ne nalaze u paru. Kod vrsta koje nisu poliploidne, haploidan broj hromozoma predstavlja jednu garnituru hromozoma i istovremeno je upola manji od diploidnog broja. Kada govorimo organizmima koje imaju veći nivo ploidnosti, tada svaki gamet takođe prenosi haploidan broj gameta.

Odigravanja krosovera između dva gena na istom hromozomu zavisi od njihovog međusobnog rastojanja. Što je to rastojanje veće i verovatnoća da će doći do krosovera je veća i obratno.

Rekombinacija može da se desi između bilo koja dva lokusa (gena) u gametu. Ako aleli dva lokusa ne potiču od istog roditelja, tada se kaže da se desila rekombinacija između ta dva lokusa. Gameti koji sadrže hromatide koje su razmenjivale delove nazivaju se *rekombinantni gameti*. Oni se razlikuju od haploidnih gameta koji su činili deo diploidnog ili poliploidnog organizma. Jedinke koje nastaju od takvih gameta nazivaju se *rekombinanti*. *Rekombinacija gena* predstavlja čitav proces koji se javlja kao rezultat krosovera.



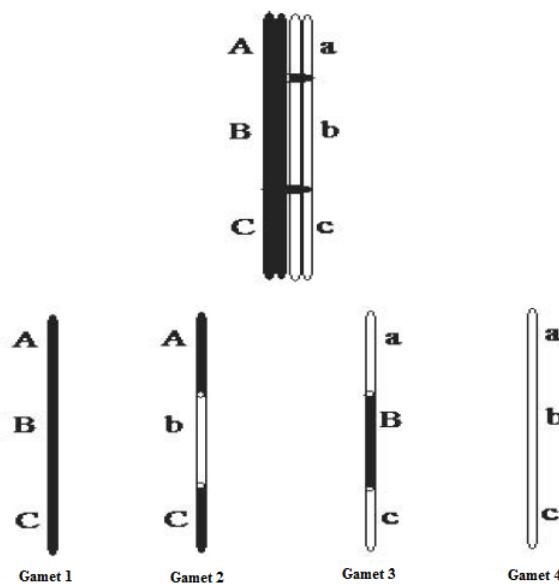
Grafik 1.2: Grafik krosovera između hromatida u slučaju diploidnih homozigotnih jedinki

Na grafiku 1.2 je prikazan primer rekombinacije između nesestrinskih hromatida, u slučaju diploidnih homozigotnih jedinki.

Stopa rekombinacije predstavlja stopu pojave rekombinantnih gameta u odnosu na

ukupan broj gameta. Stopa rekombinacije predstavlja verovatnoću da je gamet rekombinantan. *Hijazmate* predstavljaju mesta na nesestrinskim hromatidama na kojima se vrši razmena genetskog materijala.

Primer 1.4. Neka su data četiri homologne sestrinske hromatide sa dve hijazmate i njena četiri gameta. Hromatide koje su obojene crnom bojom potiču od hromozoma koji potiče od oca, dok bele hromatide potiču od majčinskih hromozoma. Neka su aleli koji potiču sa hromozoma od oca označeni velikim slovom, a aleli koji potiču sa hromozoma od majke malim slovima. Kod dve hijazmate, iste hromatide su učestvovalle u krosoveru, što je rezultiralo gametima 2 i 3. Iste ti gameti su formirani od materijala koji potiče od oba roditelja. Gameti 1 i 4 su čitav genetski materijala nasledili od oca i majke, respektivno. Na gametu 2, *A* i *b*, *b* i *C* su rekombinantni, dok su na gametu 3 aleli *a* i *B*, *B* i *c* rekombinantni. Ali *A* i *C* na gametu 2 i *a* i *c* na gametu 3 nisu rekombinantni. Još jedan način za određivanje da li su dva lokusa rekombinantni jeste da se posmatra broj krosovera između njih. Ako je broj krosovera između dva lokusa neparan, tada su oni rekombinantni, inače oni nisu rekombinantni.



Grafik 1.3: Rekombinacija između hromatida u slučaju diploidnih homozigotnih jedinki iz primera 1.4.

Dva ili više gena su povezani ako postoji tendencija da se nasleđuju zajedno. Dva

gena koji su dovoljno blizu jednog drugom na istom hromozomu imaju tendenciju da se zajedno nasleđuju i takve gene nazivamo *haplotipovi*. Geni na različitim hromozomima su uvek nepovezani. Ako su lokusi blizu jedno drugom na istom hromozomu, za njih se kaže da su povezani, i oni se često nasleđuju zajedno, ali to nije pravilo: povremeno može doći do rekombinacije između lokusa. Ako se rekombinacija desi, dva lokusa će se naći na različitim homolozima ili čak hromozomima. Što su lokusi više udaljeni, veća je verovatnoća da će se rekombinacija desiti i ako su oni veoma udaljeni jedno od drugog, šansa da oni budu razdvojeni tokom mejoze jednaka je verovatnoći da oni ostanu na istim hromozomima i lokusi u tom slučaju su ponovo nepovezani.

1.2.2. Genetičke mape

Kreiranje genetičke mape predstavljaju veoma bitan, i uz to kompleksan korak i problem u QTL analizi. Međutim, u ovom radu, one će biti samo navedene i autor se neće baviti problemom genetičkih mapa.

Postoje tri tipa genetičkih mapa:

- citogenetičke mape,
- fizičke mape,
- molekularne mape.

Najbitnije mape koje se koriste za QTL mapiranje jesu molekularne mape. One predstavljaju genetičke markere ili gene i njihovu tačnu poziciju na hromozomu, merenu u Morganima.

Genetički marker predstavlja niz baza koji se nalaze na određenoj lokaciji u genomu, koja varira između individua. Drugim rečima, marker je sekvenca DNK koje je identifikovana u smislu njegove lokacije na hromozomu. *Morgan* predstavlja očekivani broj krosovera na jednom gametu između dva markera. Najčešće se kao jedinica rastojanja koristi *centimorgan* (cM) koji se procenjuje na osnovu broja krosovera gameta na 100 gameta (u %). Dakle, 1% krosovera predstavlja jedinicu rastojanja ili centimorgan (cM), što znači da je $1 \text{ cM} = 1\%$ krosovera.

Genetički marker ne mora da predstavlja gen. Međutim, termini koji se koriste za gene kao što su aleli i genotipovi se takođe koriste za markere, iako markeri nisu geni. U suštini, markeri se nalaze u genima ili blizu gena, i ukoliko su blizu gena, to znači da se nije desila rekombinacija između markera i gena, te se genetički materijal nasleđuje zajedno.

Kako broj krosovera najčešće nije lako primetljiv, genetička distanca se ne može direktno izmeriti. Međutim, kako krossover uzrokuje rekombinaciju, a stopa rekombinacije između dva markera može da se izračuna, tada se genetička distanca može proceniti na osnovu stope rekombinacije. Veza između genetičke distance i stope rekombinacije je aproksimirana *funkcijom mapiranja*. Funkcija mapiranja je monotona funkcija oblika $r = M(d)$, koja povezuje stopu rekombinacije r i genetičku distancu d izraženu u Morganima i najčešće zavisi od određenih prepostavki.

Najčešće korišćene funkcije mapiranja su:

- **Haldanova funkcija mapiranja (detaljnije u [10])**

$$r = \frac{1}{2}(1 - e^{-2d}),$$

- **Kosambi funkcija mapiranja**

$$r = \frac{1}{2}(\tanh(2d)) = \frac{1}{2} \frac{e^{4d} - 1}{e^{4d} + 1},$$

- **Karter-Falkoner funkcija mapiranja**

$$r_i = \frac{1}{2}\tanh[4d - \tan^{-1}(2r_{i-1})], \quad i = 1, 2, \dots$$

Kao što je prethodno navedeno, mapiranje markera, tj. razviće molekularnih (DNK) mapa predstavlja bitan korak u QTL analizi. Na osnovu njih se utvrđuje pozicija i relativna genetička distanca između markera duž hromozoma. U cilju što preciznije identifikacije QTL-ova, potrebno je postići odgovarajuću pokrivenost genoma u smislu gustine molekularnih markera. Za većinu vrsta, dovoljno je pronaći 100-150 markera, ravnomerno raspoređenih duž hromozoma.[2]

Mape velike gustine (idealno 1 cM između svaka dva markera) su potrebne radi lokalizacije i mapiranja lokusa odgovornih za ispoljavanje osobine od interesa. Da bi se povećala tačnost sa kojom se detektuju lokusi, isplativije je povećati broj biljaka u populaciji nego broj markera. Samim tim se poboljšava gustina mape, jer se nalaze nove rekombinacije (r u Haldanovoj funkciji).[2]

Koristeći markere, šablon nasleđivanja se može pratiti kroz porodično stablo. Na primer, analizirajući markere povezane sa bojom očiju kroz više generacija, moguće je odrediti od koga je dete nasledilo alele za boju očiju. Što je važnije, pronalaženje markera povezanih sa nekom bolešću dovodi do toga da se locira odgovorni gen za tu bolest koji je veoma bitan u smislu pronađaska leka.

1.3. Pregled statističkih pojmoveva

1.3.1. Osnovni pojmovi

Uvedimo sada neke osnovne elemente statistike i verovatnoće koje će se koristiti u daljem radu.

Statistička populacija se razlikuje od prirodne populacije u tom smislu da je ona uža i usredstvena je na određeni aspekt prirodne populacije. Statistička populacija se čak razlikuje u različitim studijama, čak i ako studije obuhvataju iste populacije, što će se ilustrovati narednim primerom.

Primer 1.5. : Plantaža kalifornijskog bora u južnoj Australiji je primer prirodne populacije. Neka su u studiji obuhvaćene sledeće osobine: broj šišarki na godišnjem nivou, prečnik stabla i kvalitet grana. Kada se ove tri osobine posmatraju odvojeno, tada svaka od tih osobina daju statističku populaciju koju čini kolekcija svih vrednosti navedenih osobina u prirodnoj populaciji. Dakle, postoje tri statističke populacije koje su asocijirane u prirodnoj populaciji. Naravno, navedene tri osobine se mogu posmatrati zajedno, u istoj studiji, odakle se dobija nova statistička populacija koja predstavlja skup svih trojki vrednosti osobina iz prirodne populacije.

U daljem radu, umesto termina statističke populacije, koristiće se termin populacija.

Osnovni model u teoriji verovatnoće jeste *eksperiment* (pojava, opit) kod koga ostvarivanje određenih uslova ne dovodi do jednoznačnog rezultata. Skup svih mogućih ishoda nekog eksperimenta označavaćemo sa Ω , dok elemente skupa Ω ćemo nazivati *elementarnim događajima*, u oznaci ω . Svakom elementarnom događaju korespondiramo neku njenu brojnu karakteristiku.[7]

Populacija može da se opiše raspodelom, koja opisuje proporciju određenog seta numeričkih vrednosti u populaciji. Tačnije rečeno, proporcija određenog seta brojnih vrednosti jeste verovatnoća pojave istog seta prilikom slučajnog uzorkovanja. Promenljiva veličina koja brojne vrednosti prima sa određenim verovatnoćama se naziva *slučajna promenljiva*. Pod slučajnim uzorkovanjem podrazumevamo situaciju u kojoj svaka numerička vrednost iz populacije ima jednaku verovatnoću da bude uzorkovana.[7]

Zakon raspodele predstavlja skup vrednosti slučajne promenljive i njene odgovarajuće verovatnoće. U realnim statističkim populacijama, raspodela je retko poznata. Za bilo kakvo statističko zaključivanje, ona se aproksimira matematički jednostavnijim raspodelama. Najčešće raspodele koje se koriste prilikom mapiranja lokusa za kvantitativna svojstva su:

Normalna raspodela

Normalnu (Gausovu) raspodelu karakteriše sledeća funkcija gustine:

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty \leq x \leq \infty,$$

gde je μ očekivana vrednost, a σ varijansa navedene raspodele, i označava se $N(\mu, \sigma^2)$. Normalna raspodela se najčešće koristi za aproksimaciju raspodele ispitivane kvantitativne osobine čiji su uzroci varijanse nepoznati i nemaju dominantan uticaj na ukupnu varijaciju.

Bernulijeva raspodela

U ovu grupu raspodela spadaju populacije koje se mogu podeliti na tačno dve grupe, dve potpopulacije. Neka se u eksperimentu koji se vrši u jednom ponavljanju posmatra događaj A , gde je skup svih mogućih ishoda $\Omega = \{A, \bar{A}\}$, tj. da li se događaj A dogodio ili nije. Neka je p verovatnoća da se događaj A desio. Tada je raspodela verovatnoća

Bernulijeve raspodele jednaka

$$P(X = k) = p^k(1 - p)^{n-k}, \quad k = 0, 1, \dots, n.$$

Za primer se može uzeti populacija ruža i dihotomna osobina kao što je kontinuirano cvetanje biljke u toku godine gde se razlikuju biljke koje cvetaju kontinuirano u toku godine i one koje ne cvetaju. Jedno izvođenje eksperimenta bi predstavljalo posmatranje navedene osobine za samo jedan cvet iz čitave populacije.

Binomna raspodela

Neka su X_i , $i = 1, \dots, n$, nezavisne Bernulijeve slučajne promenljive i neka se n puta ponavlja eksperiment. Broj pojavljivanja događaja A je $0, 1, \dots, n$. Neka je S_n predstavlja broj realizacija događaja A. Tada se distribucija $S_n = \sum_{i=1}^n X_i$ naziva binomna raspodela, a raspodela verovatnoća je data sa

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

Politomna raspodela

U slučaju kada populacije sadrže $k > 2$ različitih kategorija, one se tada opisuju sa politomnom raspodelom. Svaka individua u populaciji se može prikazati kao vektor $\mathbf{x} = (x_1, \dots, x_{k-1})$, gde je $x_j = 0$ ili 1. Ukoliko individua spada u kategoriju j , tada je $x_j = 1$ i inače je 0. Individua koja pripada k -toj kategoriji se opisuje nula vektorom $\mathbf{x} = (0, \dots, 0)$, koji je dužine $k-1$. Neka je p_j verovatnoća da individua pripada j -oj kategoriji. Tada je raspodela verovatnoća politomne raspodele

$$f(\mathbf{x}) = p_1^{x_1} \cdots p_{k-1}^{x_{k-1}} (1 - \sum_{j=1}^{k-1} p_j)^{1 - \sum_{j=1}^{k-1} x_j}, \quad x_j = 0 \text{ ili } 1, \quad \sum_{j=1}^{k-1} x_j \leq 1.$$

Multinomna raspodela

Kao što se Bernulijeva raspodela može predstaviti kao binomna raspodela u slučaju jednog izvođenja eksperimenta, isto tako se politomna raspodela može predstaviti kao multinomna raspodelu u slučaju jednog izvođenja eksperimenta. Neka je $X_i, i = 1, \dots, n$

skup slučajnih vektora koji prate politomnu raspodelu. Tada slučajna promenljiva $N = \sum_{i=1}^n X_i = (n_1, \dots, n_{k-1})^t$ prati *multinomnu raspodelu* sa indeksom n i parametrima $\mathbf{p} = (p_1, \dots, p_{k-1})^t$.

Odgovarajuća raspodela slučajne promenljive N je data sa

$$P((n_1, \dots, n_{k-1})^t) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} \cdots p_{k-1}^{n_{k-1}} p_k^{n_k}, \quad 0 \leq n_j \leq n, \quad \sum_{j=1}^{k-1} n_j \leq n,$$

gde je $n_k = n - \sum_{j=1}^{k-1} n_j$ i $p_k = 1 - \sum_{j=1}^{k-1} p_j$.

Multinomna raspodela je uopštenje binomne raspodele gde N predstavlja broj nezavisno izvedenih eksperimenata. Svaki eksperiment ima k mogućih ishoda sa odgovarajućim verovatnoćama p_1, p_2, \dots, p_k . Komponenta n_j predstavlja broj pojavljivanja j -tog ishoda (događaja).

Poasonova raspodela

Poasonova raspodela spada u diskretne raspodele čiji je zakon raspodela data sa

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Očekivana vrednost i varijansa slučajne promenljive koja je opisana sa Poasonovom raspodelom su jednake parametru λ . U praksi, Poasonova raspodela se najčešće koristi za opisivanje promenljivih koje broje pojavljivanje određenih događaja i predstavlja aproksimaciju binomne raspodela kada je n veliko i $np \leq 10$. Tada uzimamo $\lambda = np$ i prelazimo na Poasonovu raspodelu $\mathcal{P}(\lambda)$.

Raspodele iz familije eksponencijalnih distribucija

Sve navedene raspodela spadaju u familiju eksponencijalnih distribucija koja se opisuje sledećom raspodelom verovatnoća:

$$f(x) = \exp\left\{\frac{\theta' T(x) - b(\theta)}{\varphi} + c(x, \theta)\right\} I_A(x)$$

gde je $\varphi > 0$ disperzionalni parametar, $a(\cdot)$, $b(\cdot)$ i $c(\cdot, \cdot)$ poznate funkcije, θ prirodni para-

metar odgovarajuće raspodele i $I_A(x)$ indikator funkcija seta događaja A . U familiju eksponencijalnih raspodela spadaju i eksponencijalna, gama i log-normalna raspodela.,

1.3.2. Testiranje hipoteza

Testiranje hipoteza je oblast statističke analize koja je u širokoj upotrebi, jer omogućuje sistematično donošenje odluka o problemima koji u sebi sadrže neodređenost. Ono omogućuje da se podaci dobijeni iz uzorka mogu kombinovati sa statističkom teorijom i na taj način izvode zaključci o celoj populaciji. Takođe, ono otklanja uticaj subjektivnosti pojedinca, i na taj način dovodi do racionalnijeg i tačnijeg zaključivanja.

Definicija 1.2. Statistička hipoteza je svaka pretpostavka o tome da obeležje X ima raspodelu koja pripada nekom podskupu skupa dopustivih raspodela.

Na primer, za dati lokus u genomu određene biljke ili životinje, želimo da saznamo da li je taj lokus odgovoran za ispoljavanje kvantitativne osobine. Prepostavimo da se radi o diploidnom organizmu, za koji znamo da postoje tri moguće kombinacije genotipa AA , Aa i aa . Neka μ_{AA} , μ_{Aa} i μ_{aa} predstavljaju srednje vrednosti svih individua koje imaju genotip AA , Aa i aa , respektivno. Da bismo saznali da li je lokus zapravo QTL, testiramo hipotezu $\mu_{AA} = \mu_{Aa} = \mu_{aa}$.

Neka je data familija distribucija $\{P(x, \theta) : \theta \in \Theta\}$, gde je Θ skup svih mogućih vrednosti parametra θ . Skup Θ se deli na dva međusobno disjunktna podskupa, Θ_0 i Θ_1 . Polazi se od nulte hipoteze $H_0(\theta \in \Theta_0)$, nakon koje se definiše alternativna hipoteza $H_1(\theta \in \Theta_1)$. Postupak odbacivanja statističkih hipoteza na osnovu realizovanog uzorka naziva se statistički test. Neka je T statistički test koji se sprovodi na sledeći način: ako je $T > c$, gde je c određena specifična konstanta, tada se odbija nulta hipoteza, u suprotnom se ona prihvata.

Moguće je da se pojave dve vrste greške: prva vrsta se odnosi na situaciju kada se, zbog slučajnosti uzorka, odbaci hipoteza H_0 , koja je zapravo tačna. Ova vrsta greške se naziva *greška prve vrste*. Druga vrsta greške se naziva *greškom druge vrste* i javlja se u situaciji kada se ne odbaci hipoteza H_0 , iako je hipoteza H_1 tačna.

Prilikom QTL analize, nulta hipoteza je uvek ista i ona se odnosi na situaciju kada

lokus nije QTL. Ako se lokus proglaši QTL-om, a on to nije, čini se greška prve vrste. Greška druge vrste se čini ukoliko se lokus ne proglaši QTL-om, a on to nije.

1.3.3. Linearni modeli

Pronalaženjem veza između pojava bavi se regresiona analiza. Regresiona analiza je od velikog značaja, kako u ekonomiji i privredi, tako i u prirodnim naukama, kao što su: hemija, fizika, biologija, farmakologija, toksikologija, biohemija, medicina i druge. Problem opisivanja ovakvih veza svodi se na pronalaženje modela koji povezuje jednu ili više zavisnih promenljivih sa jednom ili više nezavisnih, objašnjavajućih, promenljivih pomoću neke funkcionalne zavisnosti. Oblik ove funkcionalne zavisnosti je najčešće nepoznat, pa ostaje na istraživaču da izabere onu koja je po nekom kriterijumu najbolja. Veoma često se koriste polinomne funkcije, ali isto tako i eksponencijalne ili neke druge funkcije. Opšti problem nalaženja funkcije koja dobro aproksimira posmatrani skup podataka, često se naziva “fitovanje” krive ili određivanje regresione linije.

Linearni modeli se koriste za opisivanje veze između zavisne promenljive y i nekog broja kovarijata (nezavisnih promenljivih) x_j . U QTL analizi, izmerene fenotipske vrednosti kvantitativne osobine su zavisne promenljive, a genotipovi makera se uzimaju kao kovarijate, te se linearni model koristi kao instrument kojim opisuјemo vezu između osobine i markera.

Neka $\{y_i, (x_{i1}, \dots, x_{ip}) : i = 1, \dots, n\}$ predstavlja vrednosti zavisne promenljive i kovarijata izmerene za n individua. Linearni model je

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

gde β_j predstavlja nepoznate parametre i ε_i predstavlja slučajne promenljive čija je očekivana vrednost nula.

Linearni model se koristi za detekciju uzročnih kovarijata, predviđanje buduće vrednosti zavisne promenljive i kontrolu procesa koji utiču na model. Prepostavke koje se najčešće čine su:

- ε_i predstavlja nezavisne slučajne promenljive sa varijansom σ_i^2 , ili
- ε_i predstavlja nezavisne slučajne promenljive iz iste raspodele čija varijansa iznosi σ^2 , ili
- ε_i predstavlja nezavisne slučajne promenljive koje prate normalnu raspodelu čija je varijansa σ^2 .

Linearni model se može izraziti u matričnom obliku. Neka je $\mathbf{y} = (y_1, \dots, y_n)^t$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^t$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^t$ i $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_n)$, gde $\mathbf{1}$ predstavlja jedinični vektor, i $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^t$. Tada se model može zapisati na sledeći način:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Postoje razne metode koje se koriste za ocenjivanje nepoznatih parametara $\boldsymbol{\beta}$, kao što su metoda maskimalne verodostojnosti, metoda momenta, Bajesovsko ocenjivanje. U radu će se koristiti metoda maksimalne verodostojnosti.

1.3.4. Metoda maksimalne verodostojnosti

Osnovna ideja ove metode je da se pomoću izabranog uzorka (x_1, x_2, \dots, x_n) odabere ona vrednost nepoznatog parametra θ koja daje najveću verovatnoću da baš taj uzorak bude odabran. Da bismo formulisali metodu maksimalne verodostojnosti, potrebno je definisati funkciju verodostojnosti.

Definicija 1.3. Funkcija verodostojnosti za slučajni uzorak (X_1, X_2, \dots, X_n) na osnovu realizovanog uzorka (x_1, x_2, \dots, x_n) obima n je

$$L(\theta, x_1, x_2, \dots, x_n) = \begin{cases} p(x_1, \theta)p(x_2, \theta)\dots p(x_n, \theta), & \text{za diskretnu raspodelu} \\ \phi(x_1, \theta)\phi(x_2, \theta)\dots\phi(x_n, \theta), & \text{za neprekidnu raspodelu} \end{cases}$$

Dalje ćemo prilikom određivanja funkcije verodostojnosti koristiti oznaku

$$L(\theta, x_1, x_2, \dots, x_n) = f(x_1, \theta)f(x_2, \theta)\dots f(x_n, \theta)$$

i za diskretnu i neprekidnu raspodelu.

Definicija 1.4. Neka je $L(\theta, x_1, x_2, \dots, x_n)$ funkcija verodostojnosti za prost slučajni uzorak (X_1, X_2, \dots, X_n) na osnovu realizovanog uzorka (x_1, x_2, \dots, x_n) .

Ako je $\hat{\theta} = u(x_1, x_2, \dots, x_n)$ u skupu Θ i maksimizira $L(\theta)$, tada je statistika $\hat{\theta} = u(X_1, X_2, \dots, X_n)$ ocena nepoznatog parametra θ dobijena metodom maksimalne verodostojnosti. Realizovana vrednost $\hat{\theta} = u(x_1, x_2, \dots, x_n)$ je ocena maksimalne verodostojnosti za θ na osnovu uzorka (x_1, x_2, \dots, x_n) . Dalje ćemo i za statistiku i za njenu realizovanu vrednost koristiti istu oznaku $\hat{\theta}$, bez opasnosti od zabune.

U velikom broju slučajeva funkcija verodostojnosti zadovoljava uslove regularnosti, pa se njen maksimum može odrediti rešavanjem jednačine

$$\frac{dL(\theta)}{d\theta} = 0.$$

Kako funkcija $L(\theta)$ i $\ln L(\theta)$ postižu maksimum za istu vrednost θ , često je, zbog oblika raspodele, lakše naći maksimum prirodnog logaritma funkcije verodostojnosti.

1.4. Molekularni markeri

Ideja o korišćenju genetičkih markera u cilju identifikacije odgovornih lokusa za kvantitativna svojstva javila se početkom dvadesetog veka. Morfološki markeri se nisu pokazali kao najbolje rešenje za postizanje postavljenog cilja, iako su korišćeni i u drugim istraživanjima u cilju proučavanja genetičke osnove kvantitativnih osobina.[2] Genetički markeri su veoma bitni sa stanovišta QTL analize, jer se na osnovu njih procenjuju lokusi koji utiču na osobine, te zbog toga će se sada malo detaljnije obraditi. Kao što je ranije objašnjeno, markeri predstavljaju delove genoma koji oslikavaju genetičke razlike između različitih organizama ili vrsta, a na osnovu njih se indirektno dobija informacija o genima (ili delovima genoma) koji se odnose na određene osobine koje se ispituju. Molekularni markeri se koriste i za mapiranje gena koji kontrolišu ekspresiju različitih ekonomski značajnih agronomskih svojstava. Mogu se svrstati u tri velike grupe:

- morfološki markeri - zasnovani su na osobinama koje se mogu vizuelno uočiti,
- biohemijski markeri - obuhvataju proteine (strukturne, rezervne), kao i izoenzime (različiti alelni vidovi jednog enzima),
- molekularni markeri - čine fragmenti DNK molekula, koji mogu biti deo gena ili nekodirajućih delova genoma.

Morfološki markeri su obično vizuelno ocenjene fenotipske karakteristike kao što su npr. boja cveta, tip cveta, broj latica, i slično. Nedostatak ovih markera su ograničena dostupnost, nemogućnost razlikovanja heterozigota od homozigota, asocijacija sa štetnim fenotipskim efektima. Glavni nedostatak biohemijskih markera je nedovoljan broj. Molekularni markeri su fragmenti DNK molekula, koji mogu biti deo gena ili nekodirajućih delova genoma. Osnovna karakteristika ovih markera je visok stepen polimorfizma, tačnije varijacija u određenoj DNK sekvenci prisutnih u populaciji. Većina takvih varijacija prisutna je u nekodirajućim delovima genoma i oni imaju veoma mali ili nikakav efekat na fenotip i funkcije organizma, a mogu se lako detektovati na nivou DNK. Ovi markeri najviše ispunjavaju zahteve idealnog genetičkog markera. Kada bismo trebali da definišemo idealan marker, tada bi on trebao da ispunjava sledeće kriterijume: visok stepen polimorfizma, ravnomerna raspoređenost po genomu, kodominantno nasleđivanje, odsustvo uticaja na ispoljavanje osobina koje se ispituju, odsustvo uticaja faktora spoljašnje sredine na njihovu detekciju, mogućnost da se lokus lako i brzo uoči u ranim fazama razvića.[2] Neki od najčešće korišćenih molekularnih markera su:

1. RFLP markeri - markeri zasnovani na hibridizaciji
2. SSR markeri - markeri zasnovani na lančanoj reakciji polimeraze
3. AFLP markeri - markeri zasnovani na PCR-u i restrikcionom „sečenju“
4. SNP - polimorfizam pojedinačnih nukleotida - markeri zasnovani na tačkastim mutacijama u DNK molekulu

Primena molekularnih markera danas predstavlja nezamenljivi deo selekcije i oplemenjivanja raznih kultura biljaka i životinja. Njihovo korišćenje omogućava određivanje stepena genetičkog diverziteta između različitih useva, unutar populacija, između

srodnih vrsta, itd. Kada su vrlo blisko vezani za gen od interesa, mogu se koristiti za indirektnu selekciju svojstva od interesa, i ovaj postupak predstavlja najjednostavniji vid selekcije pomoću molekularnih markera (eng. *Marker Assisted Selection*). Selekcija pomoću molekularnih markera će detaljnije biti obrađena u daljem tekstu.

Zahvaljujući molekularnim markerima, poligena svojstva mogu biti razložena na pojedinačne komponente (QTL-ove), što doprinosi boljem razumevanju nasleđivanja ovih svojstava i omogućava korišćenje marker asistirane selekcije zajedno sa metodama klasične selekcije.

1.4.1. SNP (Single Nucleotid Polymorphism) markeri

SNP markeri predstavljaju varijacije u sekvenci DNK koje nastaju usled izmene samo jednog nukleotida u genomu. Navedeni markeri u genomu formiraju različite blokove DNK (haplotipove). SNP aleli u haplotipovima se zajedno nasleđuju i identifikacija nekoliko pažljivo odabralih SNP-ova u ciljanim regionima definisana određenim haplotipovima obezbeđuje dovoljno informacija da se može predvideti konstitucija ostatka SNP-ova u regionu.

Iako se hiljade SNP-ova koriste u analizi humanog i životinjskih genoma, njihovo korišćenje u analizi biljnih genoma je još uvek u začetku. Glavna prednost u odnosu na ostale sisteme im je što su prisutni u praktično neograničenom broju u genomu (približno 68 000 SNP-ova kod ruža), kao i dostupnost pouzdanih testova za karakterizaciju velikog broja biljaka.[14]

1.5. Poliploidi

Ploidnost je jedna od onih tema koja se ne pominju u svakodnevnoj konverzaciji, ali gledajući iz ugla biologije ono je veoma bitan pojam. Reč ploidnost potiče verovatno od često korišćenih izraza haploid i diploid, koje se odnose na postojanje jednog i dva seta hromozoma, respektivno. Ploidnost se definiše kao broj kopija jednog seta hromozoma u ćeliji.[15]

Kao primer ploidnosti možemo uzeti diploidni ljudski genom koji sadrži 23 para hromozoma, odnosno ukupno 46 hromozoma. Dakle, ljudski genom sadrži dve kopije jednog seta hromozoma.

Poliploidni organizam je onaj koji sadrži više od dve kopije jednog seta hromozoma u svojim ćelijama (detaljnije u [15]). U ovom radu će pažnja biti usmerena na tetraploidne organizme (ruže), koje sadrže četiri kopije ($n = 4$), kao i diploidne koje sadrže dve kopije ($n = 2$).

Poliploidi se često javljaju kod uzgajanih biljaka koje se koriste:

- u ishrani, kao što su krompir, kafa, šećerna trska,
- u tekstilnoj industriji (npr. pamuk),
- ukrasne biljke (npr. ruže).

Kada govorimo o QTL analizi kod poliploida, nailazimo na mnogobrojne prepreke: većina dosadašnjih istraživanja i metode su razvijene za diploidne vrste. Činjenica da su ljudi diploidni organizmi je verovatno doprinela tome da ova oblast nije dovoljno istražena. Veći broj mogućih genotipova na markeru ili lokusu otežava problem QTL analize, jer potomak dva tetraploidna roditelja može da ima ukupno 36 mogućih kombinacija roditeljskih alela, tj. jedan od 8 mogućih alela koji se nasleđuju, pod pretpostavkom da nema dvostrukе redukcije (u tom slučaju postoji 100 mogućih kombinacija alela), dok, sa druge strane, kod diploida imamo situaciju gde postoje svega četiri moguće kombinacije roditeljskih alela. *Dvostruka redukcija* (eng. *double reduction*) je proces prilikom kog se od heterozigotne jedinke dobije homozigotna gamet. Dvostruka redukcija je pojava koja prouzrokuje naizgled nemoguću situaciju kod poliploida: da se kod potomstva javi genotip $AAaa$, a segregacija roditelja je $Aaaa \times aaaa$. Ako se dvostrukе redukcije ne identifikuju, konstrukcija genetičke mape i QTL analiza mogu da daju pogrešne rezultate. Srećom, stopa pojave dvostrukе redukcije je najčešće mala. Maksimalnom stopom se smatra 0,25, dok neki autori koriste granicu od 1/6.[15]

Dakle, zbog navedenih razloga, heterozigotnost kod poliploida je veća nego kod diploida. Utvrđivanje doze alela na markerima takođe predstavlja poteškoću, koja predstavlja problem prilikom rekonstrukcije haplotipova i molekularnih mapa.[15]

1.6. Obrasci nasleđivanja tetraploidnih organizama

Kao što je rečeno u prvom Mendelovom zakonu nasleđivanja, *segregacija* je odvajanje homolognih hromozoma u različite ćelije prilikom procesa razdvajanja ćelija. Segregacija genotipa potomstva tetraploidnih organizama je u skladu sa Mendelovim prvim zakonom nasleđivanja, ali je kompleksnija u odnosu na segregaciju genotipa kod diploidnih vrsta. Radi lakšeg razumevanja segragacije uzmimo jedan primer:

Primer 1.6. Neka je genotip roditelja X na određenom markeru M oblika *AAAA* (dakle imamo 4 dominantna alela), a genotip roditelja Y na markeru M *AAa* (uočimo postojanje jednog recesivnog alela). Primenjujući zakon segregacije, genotip potomstva na markeru M može biti *AAAA* ili *AAa*. Verovatnoća da je genotip potomstva *AAAA* dobijamo na sledeći način: od roditelja X svakako se prenose dva dominantna alela, dok od roditelja Y od tri moguća dominantna alela biramo dva:

$$p(\text{AAAA}) = \frac{\binom{4}{2}}{\binom{4}{2}} \cdot \frac{\binom{3}{2}}{\binom{4}{2}} = \frac{1}{2}.$$

Verovatnoća da se u potomstvu dobije genotip *AAa* takođe zavisi samo od roditelja Y, gde pri segregaciji biramo jedan od tri moguća dominantna alela i jedini recesivni alel:

$$p(\text{AAa}) = \frac{\binom{4}{2}}{\binom{4}{2}} \cdot \frac{\binom{3}{1} \cdot \binom{1}{1}}{\binom{4}{2}} = \frac{1}{2}.$$

Raspodela pri segregaciji roditelja sa *AAAA* i *AAa* genotipom na određenom markeru su:

Roditelj X	Roditelj Y	<i>AAAA</i>	<i>AAa</i>
<i>AAAA</i>	<i>AAa</i>	1/2	1/2

U tabeli 1.1 su predstavljane raspodele genotipa potomstva, u skladu sa odgovarajućim genotipom roditelja:

Roditelj <i>X</i>	Roditelj <i>Y</i>	<i>AAAA</i>	<i>AAAa</i>	<i>Aaaa</i>	<i>Aaaa</i>	<i>aaaa</i>
<i>AAAA</i>	<i>AAAa</i>	1/2	1/2			
<i>AAAA</i>	<i>AAaa</i>	1/6	2/3	1/6		
<i>AAAA</i>	<i>Aaaa</i>		1/2	1/2		
<i>AAAa</i>	<i>AAAa</i>	1/4	1/2	1/4		
<i>AAAa</i>	<i>AAaa</i>	1/12	5/12	5/12	1/12	
<i>AAAa</i>	<i>Aaaa</i>		1/4	1/2	1/4	
<i>AAAa</i>	<i>aaaa</i>			1/2	1/2	
<i>AAaa</i>	<i>AAaa</i>	1/36	2/9	1/2	2/9	1/36
<i>AAaa</i>	<i>Aaaa</i>		1/12	5/12	5/12	1/12
<i>Aaaa</i>	<i>Aaaa</i>			1/4	1/2	1/4
<i>Aaaa</i>	<i>aaaa</i>				1/2	1/2
<i>aaaa</i>	<i>AAAa</i>			1/2	1/2	
<i>aaaa</i>	<i>AAaa</i>			1/6	2/3	1/6
<i>aaaa</i>	<i>Aaaa</i>				1/2	1/2

Tabela 1.1: Distribucije genotipa potomstva u zavisnosti od genotipa roditelja

1.7. Oplemenjivanje ruža

Oplemenjivanje baštenskih ruža predstavlja dugotrajan proces (8 godina) koji počinje sa ukrštanja dveju sorta, a završava se sa plasiranjem nove sorte na tržište. Proces oplemenjivanja se sastoji od sledećih koraka:

- selekcije potomaka dveju ruža u staklenicima (prve dve godine),
- testiranja osobina na poljima i selekcija potomaka na polju (od treće do osme godine).

Najbitniji razlog zbog čega nije praktično koristiti marker asistiranu selekciju u oplemenjivanju ruža je taj što su one izrazito heterozigotne biljke, najčešće tetraploidi, te

se i najbitnije komercijalne osobine kao što su oblik cveta, boja, visina se mogu videti golim okom. Međutim, osobine kao što su otpornost na bolesti, na nisku/visoku temperaturu i miris, zahtevaju genetski bazirana istraživanja. Treba napomenuti da su većina istraživanja u ovoj oblasti rađena na divljim, diploidnim ružama (genom se sastoji od $2n$ hromozoma).[14]

Baštenske ruže su ekonomski veoma značajne biljke, jer oplemenjivanje, koje datira od 3 000 godina pre nove ere, nije usmereno samo na njihovo korišćenje kao ukras, već i na njihovu upotrebu u farmaciji, kozmetičkoj industriji i ishrani.[14]

1.8. Morfologija ruža

Prepostavimo da govorimo o tetraploidnoj populaciji ruža. Na svakom lokusu, dva od četiri alela po roditelju se prenose na potomstvo, te postoji 6 mogućih kombinacija dva alela po roditelju, što ukupno daje 36 mogućih kombinacija alela. U zavisnosti od doze alela, razlikuju se 5 tipova markera kada se posmatraju genotipovi tetraploidnih roditelja i u zavisnosti od doze alela na određenom markeru ih delimo na **nulipleks**, **simpleks**, **oupleks**, **triplex** i **kvadriplex**. U daljem radu, označavaćemo navedene grupe sa 0, 1, 2, 3, 4, respektivno. Segregacijom svih mogućih kombinacija tipova markera kod potomstva, primećuje se postojanje mnogo simetričnih tipova markera roditelja. Kao primer možemo uzeti situaciju kada je jedan roditelj koji je na određenom markeru tripleks za dominantni alel i segregira sa roditeljem koji je na istom markeru kvadriplex za dominantni alel (u oznaci 3×4). Tada je moguće zapisati navedene genotipe kao simpleks za prvog roditelja i nulipleks za drugog roditelja, uz odgovarajući, recessivan alel. Dakle, određen tip doze alela svakog markera se može prevesti na drugi tip doze koristeći sledeće pravilo: kvadriplex jednog alela = nulipleks drugog alela, tripleks jednog alela = simpleks drugog alela. Geni koji utiču na određene osobine se nalaze na parovima hromozomima u ćelijama. Ruže sadrže set od 7 kopija svakog hromozoma i imajući u vidu ploidnost ($4n$, tetraploidi), vidimo da ruže sadrže 28 hromozoma.[15]

1.9. Molekularni makeri u oplemenjivanju

Jedna od najočiglednijih primena QTL analize je indirektna selekcija pomoću molekularnih markera (MAS). Kada su markeri vrlo blisko vezani za gen od interesa, mogu se korisiti za selekciju svojstva od interesa, i ovaj postupak predstavlja najjednostavniji vid selekcije pomoću molekularnih markera. Ovo je metoda koja podrazumeva selekciju poželjnih svojstava pomoću molekularnih markera blisko vezanih za gen(e) koji determinišu dato svojstvo. Dobijeni podaci se mogu koristiti za "otisak" genotipova, u cilju njihove identifikacije i zaštite, razumevanja veza između jedinki koje se ispituju, efikasnog upravljanja genetičkim resursima, olakšavanja unošenja hromozomskih segmentata iz različitih vrsta, čak i u cilju obeležavanja specifičnih gena.[2]

Odabiranje biljaka sa odgovarajućim kombinacijama gena je suštinska komponenta oplemenjivanja. Selekcija na nivou genotipa prevazilazi nedostatke metoda klasične selekcije kao što su nepouzdanost, nemogućnost izvođenja, složenost i cena fenotipskih testova, te se na ovaj način može značajno povećati uspešnost u opomenjivanju biljaka u poređenju sa konvencionalnim metodama. Prednosti MAS-a su:

- ušteda vremena zamenom složenih ogleda u polju (koje je moguće izvesti samo u određenom periodu u toku godine ili na specifičnim lokacijama ili su tehnički složeni) marker testovima,
- eliminacija neopouzdanih fenotipskih procena vezanih za poljske oglede koji su nastali usled uticaja sredine,
- selekcija genotipova u ranim fazama razvića,
- istovremena selekcija nekoliko osobina ili nekoliko gena jedne osobine,
- sprečavanje unošenja nepoželjnih gena (engl. *linkage drag*),
- testiranje za specifična svojstva u slučajevima kada fenotipsko ocenjivanje nije izvodljivo.

Postoji veliki broj studija na temu mapiranja QTL-ova, za različite osobine kod velikog broja biljnih vrsta, ali svega nekoliko je pokazalo mogućnost korišćenja molekularnih markera za poboljšanje kvantitativnih svojstva.[2] Mnogi markeri pokazali su se

kao nepouzdani u predviđanju željenog fenotipa. U velikom broju slučajeva, razlog za ovo je nedovoljna preciznost u mapiranju QTL-ova.[2] Metode i marker sistemi koji se koriste u ovim istraživanjima su skupi i to je još jedan od ograničavajućih faktora u primeni MAS-a.

2 Statistički modeli QTL analize

2.1. Pronalaženje QTL-a

Većina osobina koje posmatramo kod biljaka, životinja i ljudi su kvantitativne i stoga je poželjno pronaći najznačajnije QTL-ove.

Kada dodamo pretpostavku da postoji epistaza, koja je predstavlja interakciju gena u cilju povećanja efekta na fenotip, problem pronalaženja odgovarajućih lokusa postaje veoma kompleksan. Drugim rečima, isključujući epistazu dobili bismo situaciju sličnu formiranju fudbalskog tima na osnovu igrača koji dobro izvode penale. To nije irelevantan test, ali najznačajniji deo, a to je kako igrači međusobno sarađuju, je ispušten iz vida. Kao uvod u postupak mapiranja, postavimo ciljeve analize:

1. ocena fenotipa osobine od interesa i određivanje genotipa,
2. određivanje regiona u genomu u kojima se nalaze potencijalni lokusi,
3. određivanje efekta QTL-a na kvantitativnu osobinu od interesta,
4. identifikovanje gena koji utiče na ciljanu osobinu.

Mapiranje QTL-ova možemo podeliti na tri koraka:

1. detekcija genetičkih faktora (alela),
2. pronalazak njihove pozicije u skladu sa datim markerima,
3. procena njihovog pojedinačnog i međusobnog efekta na ispoljavanje kvantitativne osobine.

QTL analiza se u suštini svodi na detekciju statistički značajnih asocijacija između fenotipa osobine i markera (ili intervala) na mapi. Hipoteza koja se testira je ista ona koja je spomenuta u sekciji **1.3.2**, dakle :

$$H_0 : \text{ne postoji QTL u genomu,}$$

$$H_1 : \text{postoji bar 1 QTL u genomu.}$$

Upoređujući srednje vrednosti između marker grupa u odnosu na moguće kombinacije alela u dатој populацији може се utvrditi да ли постоји statistički značajna razlika između srednjih vrednosti različitih grupa markera. Testiranje razlika između srednjih vrednosti се може istražiti korišćenjem analize varijanse или jednostrukog regresije, где се више пажње posvećује објашњеној варијанси фенотипских података. Маркери који објашњавају највећи део варијансе су највероватније више повезани са QTL-ом него остали маркери. Исто тако, када је дисперзија између различитих група већа у односу на варијансу унутра група, добијамо основу за коришћење F-теста значajности. Такође, може се применити t-тест на основу једноstrukог регресије. У суštini, логаритам са основом 10 p-vrednosti (ЛОР вредност) се користи као мера асocijације позиције маркера и његовог ефекта на фенотип. Wald тест се може користи као тести статистика уместо ЛОР вредности.[15] Још један приступ QTL мапирању је метода максималне веродостојности. На основу оцењивања добијених наведеном методом, конструише се количник између оцењивања добијених под алтернативном и nullом хипотезом присуства локуса.[15]

Подаци за анализу квантитативних особина сastoje сe, sa jedne strane, od seta molekularnih markera i njihovih genotipova, koji su izmereni za svaku jedinku iz populacije i sa druge strane, фенотипских мерења одређене особине за сваку јединку populacije. Da bi сe детектовао QTL u genomu, потребно је izvršiti statističke testove за сваку локацију на молекуларној мapi. Под локацијом можемо подразумевати локацију маркера (eng. *single marker analysis*), као и локацију локуса унутар интервала између два суседна маркера (eng. *interval mapping*). Ако постоји QTL на одређеној локацији, можемо применjujući анализу варијансе (*F-test*) да откријемо везу, односно меру асocijације између резултата фенотипских мерења и локације маркера или интервала између два маркера.[15]

Permutacioni test je метод оценjivanja praga зnačajnosti između fenotipa i markera i може се применити приликом сваке методе мапирања QTL-а. Ако подаци ukazuju на постојање QTL-а, тада везу између фенотипа и генотипа можемо прекинuti једnostavnom

permutacijom fenotipskih rezultata svih individua, pri tom ne menjajući genotip, tj. molekularnu mapu. Ako nema indikacije da postoji QTL u genomu, permutujući rezultate fenotipskih merenja svih n jedinki populacije, neće se promeniti rezultati test statistike koja se koristi. Permutacioni test će detaljnije biti obrađen u daljem toku rada.[9]

2.2. Metode za detekciju QTL-ova

Kao što je rečeno u prvom poglavlju, kvantitativne osobine su one osobine čije je ispoljavanje kontrolisano od strane većeg broja gena. Kvalitativne osobine su osobine čiji fenotip je diskretna slučajna promenljiva i najčešće se nalazi pod uticajem jednog gена, dok je uticaj sredine veoma mali. QTL analiza se može primeniti i na kvalitativnim osobinama, i to posmatrajući samo prisustvo ili odsustvo date osobine (na primer, prisustvo ili odsustvo crvene boje latica). Segregacija gena, uticaj spoljašnje sredine i njihova međusobna interakcija uzrokuje neprekidno variranje svojstva. Veliki broj ekonomski značajnih osobina gajenih biljaka (kao što su prinos, komponente prinosa, tolerantnost na sušu, komponente kvaliteta, itd.) klasifikovane su kao kvantitativne osobine. Poboljšanje ovih osobina dugo je bilo zasnovano samo na metodama klasične selekcije (fenotipska ekspresija) i s obzirom na to da ove osobine kontroliše veći broj gena i da na fenotipsko ispoljavanje utiče sredina, za izučavanje njihove genetičke pozadine nije moguće primenjivati jednostavno predviđanje prema mendelovskim očekivanim verovatnoćama, koja se sa uspehom primenjuju u analizi kvalitativnih osobina.

Glavni napredak u karakterizaciji QTL-ova započeo je razvićem DNK markera krajem prošlog veka, o kojima je bilo reči prethodnom poglavlju. Osnovna ideja u ovom pristupu je praćenje korelacije između pojave ispitivane osobine i markera. Ukoliko se utvrdi korelacija, to znači da su regioni genoma gde se markeri nalaze uključeni u ispoljavanje ispitivane osobine. Sledeći korak u identifikaciji QTL-ova podrazumeva povezivanje podataka iz odgovarajuće mape i podataka o varijabilnosti ispitivanog kvantitativnog svojstva. Ovaj korak predstavlja povezivanje genotipa i fenotipa odgovarajućeg markera. Posle razvrstavanja ispitivanih jedinki populacije u odgovarajuće klase molekularnih markera, na osnovu prisustva/odsustva određenog alela, ideja je da se ispituje značajnost razlika između ovih klasa uzimajući u obzir izmerene vrednosti pos-

matranog svojstva. Ukoliko se za neki marker detektuje značajna razlika između srednjih vrednosti fenotipskog svojstva odgovarajućih parova marker klasa, onda taj marker utiče na ispoljavanje tog svojstva. Suština je u tome da se blisko vezani marker i QTL neće razdvojiti rekombinacijom (manja je verovatnoća rekombinacije), pa će srednja vrednost fenotipskog svojstva kod marker klase blisko vezane za QTL biti značajno različita od srednje vrednosti klase koja nije vezana za marker.

U poslednje dve decenije statističke metode za detekciju QTL-ova razvile su se od jednostavnih do vrlo sofisticiranih. Metode za detekciju QTL-ova (utvrđivanje broja, pozicije i efekta) koje su u širokoj upotrebi možemo podeliti na:

- analize zasnovane na pojedinačnim markerima (engl. *single marker analysis*),
- jednostavno intervalno mapiranje (engl. *simple interval mapping analysis - SIM*),
- kompozitno intervalno mapiranje (engl. *composite interval mapping - CIM*),
- višestruko QTL mapiranje (eng. *multiple QTL mapping*).

Prepostavke koje ćemo postaviti su:

- nema dvostrukе redukcije,
- geni su nezavisni jednih od drugih,
- ne postoji interakcija gena i prirodnog okruženja,

2.2.1. Model sa jednim QTL-om

Genetska konstitucija populacije je karakterisana frekvencijom genotipa i frekvencijom alela. Ukoliko se radi o jednom lokusu koji sadrži dva alela (dakle, govorimo o diploidnoj populaciji), tada su mogući genotipovi AA , Aa i aa . Frekvencija genotipa predstavlja procenat određenog genotipa u populaciji (na primer 16 % jedinki u populaciji poseduje genotip AA , te je stoga frekvencija genotipa AA 0.16). Frekvencija alela se odnosi na procenat učešća određenog alela u populaciji. (npr. 40 % jedinki u populaciji poseduje alel A , te je stoga frekvencija alela A 0.4).

Frekvencija alela se može sračunati na osnovu frekvencije genotipa. Označimo

frekvencije genotipova AA , Aa i aa sa P , Q i R , respektivno. Neka je N veličina populacije. Tada imamo N genotipova i $2N$ alela u populaciji. PN jedinki ima genotip AA , odnosno $2PN$ dominantnih alela A . Isto tako, QN jedinki ima genotip oblika Aa , odnosno sadrži QN dominantnih alela. Dakle, frekvencija alela A se dobija kao $p = (2PN + QN)/(2N) = (2P + Q)/2$. Slično, frekvencija alela a je $q = (Q + 2R)/2$.

Imajući u vidu da na fenotip utiče genetska konstitucija jedinke i sredina u kojoj se individua razvija, fenotip možemo rastaviti na sledeći način:

$$Y = G + E,$$

gde Y predstavlja fenotipsku vrednost, G efekat genotipa, a E uticaj sredine.

Prepostavimo da postoje samo dva alela u QTL populaciji. Situacija sa više od dva alela se može posmatrati slično. Dakle, neka su A i a dominantni i recesivni alela, a p i q njihove frekvencije, respektivno. Bez gubitka opštosti, genotip može da se definiše na sledeći način:

$$G = \begin{cases} c, & \text{ako } AA, \\ d, & \text{ako } Aa, \\ -c, & \text{ako } aa. \end{cases}$$

Naš cilj je da pronađemo promenljive c i d , takve da modeliramo različiti uticaj genotipa na fenotip. Prepostavimo sada da je genotip funkcija dva alela, x_1 i x_2 , koji predstavljaju alele dobijene od oca i majke, respektivno. Funkcija genotipa je sada oblika:

$$G(x_1, x_2) = \begin{cases} c, & \text{ako } (x_1, x_2) = (A, A), \\ d, & \text{ako } (x_1, x_2) = (A, a) \text{ ili } (x_1, x_2) = (a, A), \\ -c, & \text{ako } (x_1, x_2) = (a, a). \end{cases}$$

Neka je $H(x_1, x_2) = G(x_1, x_2) - \mu$, gde je $\mu = E(G(x_1, x_2)) = c(p - q) + 2pqd$.

Tražimo funkcije $h_1(x_1)$ i $h_2(x_2)$ takve da minimizuju

$$E(H(x_1, x_2) - \tilde{h}_1(x_1) - \tilde{h}_2(x_2))^2,$$

gde za funkcije $\tilde{h}_1(x_1)$ i $\tilde{h}_2(x_2)$ važi $E(\tilde{h}_1(x_1)) = E(\tilde{h}_2(x_2)) = 0$. Neka je $h_1(x_1) = E(H(x_1, x_2)|x_1)$, $h_2(x_2) = E(H(x_1, x_2)|x_2)$. Kako su x_1 i x_2 nezavisni, dobija se sledeće jednačina

$$\begin{aligned} E[(H(x_1, x_2) - \tilde{h}_1(x_1) - \tilde{h}_2(x_2))^2] &= E[(H(x_1, x_2) - h_1(x_1) - h_2(x_2))^2] \\ &+ 2E[(H(x_1, x_2) - h_1(x_1) - h_2(x_2))(h_1(x_1) - \tilde{h}_1(x_1) + h_2(x_2) - \tilde{h}_2(x_2))] \\ &+ E[(h_1(x_1) - \tilde{h}_1(x_1) + h_2(x_2) - \tilde{h}_2(x_2))^2] \\ &= E[(H(x_1, x_2) - h_1(x_1) - h_2(x_2))^2] \\ &+ 2E[(H(x_1, x_2) - h_1(x_1) - h_2(x_2))(h_1(x_1) - \tilde{h}_1(x_1) + h_2(x_2) - \tilde{h}_2(x_2))] \\ &+ E[(h_1(x_1) - \tilde{h}_1(x_1))^2] + 2E[(h_1(x_1) - \tilde{h}_1(x_1) + h_2(x_2) - \tilde{h}_2(x_2))] \\ &+ E[(h_1(x_1) - \tilde{h}_2(x_2))^2] \\ &= E[(H(x_1, x_2) - h_1(x_1) - h_2(x_2))^2] + E[(h_1(x_1) - \tilde{h}_1(x_1))^2] + E[(h_2(x_2) - \tilde{h}_2(x_2))^2] \end{aligned}$$

Zaključujemo da $h_1(x_1)$ i $h_2(x_2)$ predstavljaju rešenje početnog problema. Zbog simetrije funkcije $H(x_1, x_2)$ i nezavisnosti x_1 i x_2 , dobija se

$$h_1(x) = h_2(x) \equiv h(x)$$

$$h(A) = pc + qd - \mu = q[c + d(q - p)] \equiv \alpha_1,$$

$$h(a) = pd - qc - \mu = -p[c + d(q - p)] \equiv \alpha_2.$$

U literaturi [17], α_1 predstavlja prosečan efekat alela A , α_2 prosečan efekat alela a , dok razlika $\alpha = \alpha_1 - \alpha_2$ predstavlja prosečan efekat substitucije gena. Iz dobijenih jednakosti dobijamo da je $\alpha = c + d(q - p)$. Neka je sada $D(x_1, x_2) = H(x_1, x_2) - h(x_1) - h(x_2)$. Vidimo sada da genetska komponenta ima oblik

$$G(x_1, x_2) = \mu + h(x_1) + h(x_2) + D(x_1, x_2),$$

gde $h_1(x_1) + h_2(x_2)$ predstavlja aditivni efekat lokusa, a komponenta $D(x_1, x_2)$ dominantni efekat.

Dakle, sada možemo zapisati

$$D(x_1, x_2) = \begin{cases} -2q^2d, & \text{ako } (x_1, x_2) = (A, A), \\ 2pqd, & \text{ako } (x_1, x_2) = (Aa) \text{ ili } (aA), \\ -2p^2d, & \text{ako } (x_1, x_2) = (a, a), \end{cases}$$

i

$$G(x_1, x_2) = \begin{cases} \mu + 2q\alpha - 2q^2d, & \text{ako } (x_1, x_2) = (A, A), \\ \mu + (q-p)\alpha + 2pqd, & \text{ako } (x_1, x_2) = (Aa) \text{ ili } (aA), \\ \mu - 2p\alpha - 2p^2d, & \text{ako } (x_1, x_2) = (a, a). \end{cases}$$

$$\equiv \mu + A(x_1, x_2) + D(x_1, x_2).$$

gde $A(x_1, x_2)$ predstavlja aditivni efekat. Treba napomenuti i da je $D(x_1, x_2) = 0$ ako i samo ako je $d = 0$. Drugim rečima, ukoliko postoji dominantni alel u genotipu, tada dolazi i do ispoljavanja dominantnog efekta.

2.2.2. Analiza pojedinačnog markera

Neka A i a predstavljaju alele potencijalnog QTL-a, dok p i q predstavljaju njihove frekvencije. Pretpostavimo da su frekvencije alela nepoznate. Označimo sa y fenotipske vrednosti ispitivane osobine i neka x_1, x_2 predstavljaju odgovarajuće alele na lokusu dobijene od roditelja. Definišimo prvo slučajnu promenljivu

$$I(x) = \begin{cases} q, & \text{ako } x = A, \\ -p, & \text{ako } x = a. \end{cases}$$

Neka je $E(I(x)) = 0$ i $Var(I(x)) = pq$. Tada se kvantitativna osobina može modeli-

rati kao

$$y = \mu + \alpha[I(x_1) + I(x_2)] - 2dI(x_1)I(x_2) + \varepsilon, \quad (2.1)$$

gde je μ očekivana vrednost populacije, α prosečan efekat substitucije gena, d dominantni efekat i ε slučajna promenljiva uzrokovana uticajem sredine. Varijansa aditivnog i dominantnog efekta genotipa se dobija kao

$$\begin{aligned} V_a &= \text{Var}(\alpha[I(x_1) + I(x_2)]) = 2\alpha^2 \text{Var}(I(x_1)) = 2pq\alpha^2, \\ V_d &= \text{Var}(-2dI(x_1)I(x_2)) = 4d^2 E(I^2(x_1)I^2(x_2)) = 4(pq)^2 d^2. \end{aligned}$$

Neka su alel frekvencije $p = q = \frac{1}{2}$. Tada model dobija oblika

$$y = \begin{cases} \mu + \alpha - \frac{1}{2}d + \varepsilon, & \text{ako } (x_1, x_2) = (A, A) \\ \mu + \frac{1}{2}d + \varepsilon, & \text{ako } (x_1, x_2) = (A, a) \text{ ili } (a, A) \\ \mu - \alpha - \frac{1}{2}d + \varepsilon, & \text{ako } (x_1, x_2) = (a, a) \end{cases} \quad (2.2)$$

Primetimo da kada je $p = q$, $\alpha = c + (q - p)d = c$. Neka je $\beta_0 = \mu - c - \frac{1}{2}$. Tada je

$$y = \begin{cases} \beta_0 + 2c + \varepsilon, & \text{ako } (x_1, x_2) = (A, A) \\ \beta_0 + c + d + \varepsilon, & \text{ako } (x_1, x_2) = (A, a) \text{ ili } (a, A) \\ \beta_0 + \varepsilon, & \text{ako } (x_1, x_2) = (a, a) \end{cases} \quad (2.3)$$

Uvedimo sada dve slučajne promenljive:

$$\delta_1 = \begin{cases} 2, & \text{ako je genotip } AA, \\ 1, & \text{ako je genotip } Aa, \\ 0, & \text{ako je genotip } aa; \end{cases}$$

$$\delta_2 = \begin{cases} 1, & \text{ako je genotip } Aa, \\ 0, & \text{inače.} \end{cases}$$

Naš model dobija onda oblik

$$y = \beta_0 + \beta\delta_1 + \gamma\delta_2 + \varepsilon, \quad (2.4)$$

gde je $\beta = c$ i $\gamma = d$.

QTL mapiranje koje sprovodi test za svaki marker naziva se *analiza pojedinačnog markera*. Ispituje se da li je marker QTL ili je povezan sa QTL-om. Neka y predstavlja vrednosti ispitivane osobine i neka su dominantni i recessivni alel označeni sa A i a . Prepostavimo zatim da je svaki marker potencijalni QTL.

Analiza pojedinačnog markera testira za svaki marker hipotezu $H_0 : \beta = \gamma = 0$. Neka se uzorak sastoji od n jedinki. Tada se model može predstaviti u matričnom obliku kao

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

gde je $\mathbf{y} = (y_1, \dots, y_n)^T$ vektor vrednosti osobine, $\mathbf{X} = (\mathbf{1}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$, $\boldsymbol{\beta} = (\beta_0, \beta, \gamma)$.

Prepostavimo da postoji ukupno m markera u istraživanju. Tada se odgovarajući testovi primenjuju za svih m markera. Da bi se odbacila nulta hipoteza za određeni marker, prag značajnosti se određuje na svih m testova kako bi se kontrolisala ukupna greška prve vrste (odbacivanje hipoteza H_0 , koja je zapravo tačna). Ukoliko gustina markera u genomu nije dovoljna, tada ova vrsta analize nije mnogo korisna (detaljnije u [2], [15], [17]).

Metod 1 - Analiza varijanse

Najjednostavniji metod koji se koristi u QTL analizi zasniva se na primeni analize varijanse za svaki pojedinačni marker.

Analiza varijanse (disperziona analiza) je jedan od najčešće korišćenih statističkih metoda. Koristi se u ispitivanjima u kojima se posmatra kako jedan ili više kontrolisanih faktora utiče na formiranje vrednosti obeležja.[18]

Pravi način za testiranje hipoteze o jednakosti više očekivanja je analiza varijanse. Važno je uočiti da se testiraju hipoteze o matematičkim očekivanjima, a ne o disperzi-

jama. Naziv analiza varijanse potiče od toga što se zaključci o matematičkim očekivanjima obeležja izvode na osnovu analiziranja varijanse obeležja.

Analiza varijanse pruža mogućnost da detektujemo razlike između više genotipskih oblika molekularnih markera i da testiramo značajnost razlike između svih genotipova. Posmatra se dejstvo nekog faktora A na vrednosti obeležja X . Faktor A deluje preko $k \geq 2$ svojih nivoa ili stanja koji se obeležavaju sa A_1, A_2, \dots, A_n . U slučaju QTL analize, faktori predstavljaju marker klase. Marker klase se odnose na moguće varijacije genotipa markera. Kod diploida imamo klase AA, Aa i aa , dok kod tetraploida $AAAA, AAAa, AAaa, Aaaa$ i $aaaa$. Prilikom izvođenja eksperimenta cela populacija E na kojoj se posmatra obeležje X deli se na potpopulacije $1, 2, \dots, k$ i na elemente svake od potpopulacija deli je samo jedan od nivoa faktora A .

Prilikom primene analize varijanse testira se hipoteza $H_0 : (m_1 = m_2 = \dots = m_k = m)$ protiv alternativne $H_1 : (m_i \neq m_j, \text{ za bar } 1 \text{ par } i, j)$.

Nulta hipoteza se odnosi na situaciju kada nema statistički značajnih razlika između srednjih vrednosti, tj. kada ispitivani marker ne predstavlja QTL. Za testiranje hipoteze H_0 koristićemo F-test statistiku, tačnije p-vrednost F-test statistike.

Statistički test se može sprovesti korišćenjem p-vrednosti, gde je p-vrednost veličina kritične oblasti čija je granica registrovana vrednost t_{reg} test statistike.[18] Na primer, ako je oblast odbacivanja nulte hipoteze oblika $T > c$, tada se verovatnoća

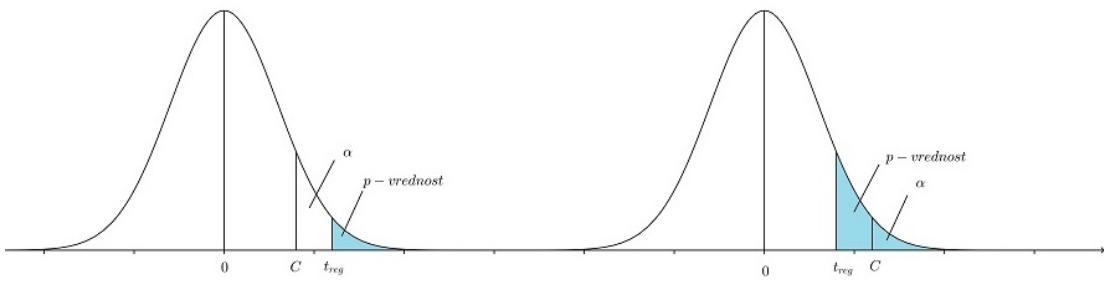
$$p = P_{H_0}\{T > t_{reg}\}$$

naziva p-vrednost. Odluka se donosi na sledeći način:

- **Ako je $p < \alpha$, H_0 se odbacuje;**
- **Ako je $p > \alpha$, H_0 se ne odbacuje.**

Nultu hipotezu odbacujem ako je p-vrednost ispod određene granice, ili ako je negativni logaritam p-vrednosti prevaziđa određeni prag značajnosti. Vrednost $-\log_{10}p$ nazivamo **LOP vrednošću** (negativni logaritam p-vrednosti) i nju ćemo koristiti kao test statistiku prilikom primene ovog modela.

Ovakav princip pronalaženja odgovarajućih lokusa prepostavlja da je svaki marker



Grafik 2.1: Grafik p-vrednosti

potencijalni QTL, i za svaki marker se prilikom analize varijanse posmatra razlike između srednjih vrednosti fenotipa različitih marker klasa.

Dakle, prilikom primene analize varijanse QTL analiza se sastoji od dve faze:

1. Određivanje LOP vrednosti koja predstavlja meru asocijacije lokacije markera i efekta na fenotipsku osobinu od interesa.
2. Određivanje praga odsecanja LOP vrednosti na osnovu kojeg dobijamo potencijalne QTL-ove. Prag određujemo vršenjem permutacija fenotipskih osobina po markeru.

Visoke LOP vrednosti, koje nadmašuju prag značajnosti ukazuju na veću verovatnoću tačnosti alternativne hipoteze.

Prednosti analize varijanse se ogledaju u situacijama kada imamo gustu molekularnu mapu, sa markerima koji su ravnomerno i dovoljno blizu raspoređeni po čitavom genomu. Mane analize zasnovane na analizi varijanse, a i generalno na analizama pojedinačnog markera, su:

- jedinke čije genotipske informacije nedostaju, odnosno kojima genotip na nekom markeru nije određen, se odbacuju,
- kada su markeri proređeni, lokusi mogu biti daleko od svih markera, što umanjuje mogućnost detekcije QTL-a,
- ne dobijaju se odvojene ocene lokacije i efekta QTL-a.

Dakle, glavni nedostatak ove metode je taj da što je QTL udaljeniji od markera

to ga je teže detektovati, zato što može da dođe do rekombinacije između markera i QTL-a. Ovo ukazuje na mogućnost precenjivanja, odnosno potcenjivanja efekta QTL-a identifikovanog primenom ove metode.[12]

Asocijacija između markera i osobina se testirati na nekoliko različitih nivoa:

- ANOVA u odnosu na dozu alela,
- ANOVA u odnosu na prisustvo ili odsustvo alela,
- primena t-testa u odnosu na prisustvo ili odsustvo alela.

Prvi korak u primeni analize varijanse, bilo da je ono rađeno u odnosu na dozu alela ili na prisustvo ili odsustvo alela, je izbacivanje svih jedinki kojima nedostaje fenotipsko merenje i svih jedinki za koje nije određen genotip markera.

Nakon toga se za svaki marker proverava da li postoje statistički značajne razlike između srednjih vrednosti fenotipskih merenja marker klase.

U slučaju pojave dva ili više QTL-ova koji prelaze prag odsecanja, tada se primenom višestruke regresije ispituje njihov zajednički uticaj na ispitivanu osobinu od interesa. Tačnije, ispituje se da li zajedničko nasleđivanje specifičnih markera utiče statistički značajno na manifestaciju fenotipa.[14]

2.2.3. Intervalno mapiranje

Mešavina Gausovih raspodela

Prilikom mapiranja značajnih lokusa, lokacije i genotip mnogih od njih su nepoznati, jer je poznat samo genotip markera. Međutim, znajući genotip svakog makera, moguće je na osnovu uslovnih verovatnoća odrediti genotipove lokusa koji nedostaju. Navedene uslovne verovatnoće imaju veoma bitnu ulogu prilikom intervalnog mapiranja.

Neka je Q potencijalni QTL, dok su \mathcal{A} i \mathcal{B} susedni markeri koji se nalaze levo i desno od njega. Neka su odgovarajući aleli za navedene markere Q, q, A, a i B, b respektivno. Neka je p_{ij} , $i, j = 0, 1$ zajednički zakon raspodele ukoliko se desilo i rekombinacija između \mathcal{A} i Q i j rekombinacija između Q i \mathcal{B} na jednom hromozomu. Neka je P_1

verovatnoća da se desila rekombinacija između \mathcal{A} i \mathcal{B} na jednom hromozomu, a $P_0 = 1 - P_1$. Označimo homozigotni genotip sa dominantnim alelima sa 2, heterozigotni genotip sa 1 i homozigotni genotip sa recessivnim alelima sa 0. Dakle, indeksi 222, 212 i 202 predstavljaju genotipove $AAQQBB$, $AAQqBB$ i $AAqqBB$ respektivno.

U slučaju kada imamo indeks 222, marginalna verovatnoća se određuje na sledeći način: verovatnoća pojave homozigotnog genotipa sa dva dominantna alela na markeru \mathcal{A} je $\frac{1}{4}$. Kako nema rekombinacije između \mathcal{A} i \mathcal{Q} , niti između \mathcal{A} i \mathcal{B} na oba hromozoma, verovatnoća da nema rekombinacije je p_{00}^2 . Dakle, marginalna verovatnoća za indeks 222 je $\frac{1}{4}p_{00}^2$. Frekvencija genotipa susednih markera se dobija na taj način što posmatramo da li je došlo do rekombinacije između susednih markera. Kako u slučaju indeksa 22 nije došlo do rekombinacije, vidimo da frekvencija genotipa iznosi $\frac{P_0^2}{4}$. Uslovnu verovatnoću za genotip nepoznatog lokusa računamo uz pomoć definicije uslovne verovatnoće, i delimo marginalnu verovatnoću indeksa 222 sa frekvencijom genotipa susednih markera, dobijajući uslovnu verovatnoću pojave genotipa 2 na lokusu koji se nalazi između dva susedna markera čiji su genotipovi poznati (oba poseduju genotip 2). U slučaju indeksa 112, tada se marginalna verovatnoća dobija slično: verovatnoća pojave heterozigotnog genotipa kod markera \mathcal{A} koja iznosi $\frac{1}{4}$, dok verovatnoća da se desi rekombinacija između susednih markera i lokusa koji se nalazi između njih se dobija na sledeći način: prva situacija se odnosi na ne postojanje rekombinacije između markera i lokusa na prvom hromozomu, dok na drugom hromozomu se dolazi do rekombinacije između lokusa \mathcal{Q} i markera \mathcal{B} . Dakle, verovatnoća je $p_{00}p_{01}$. Druga situacija se odnosi na postojanje rekombinacije između oba markera i lokusa na jednom hromozomu i postojanja rekombinacije između markera \mathcal{A} i lokusa \mathcal{Q} . Dakle, verovatnoća je $p_{11}p_{10}$, odnosno marginalna verovatnoća za indeks 112 je $\frac{1}{2}(p_{00}p_{01} + p_{11}p_{10})$. Analogno kao i prethodni primer sa indeksom 222, dobijamo uslovnu verovatnoću za indeks 112. U tabeli 2.1 su predstavljene uslovne verovatnoće svih mogućih kombinacija indeksa.

Povežimo sada zajedničke verovatnoće p_{11}, p_{10}, p_{01} i p_{00} sa stopama rekombinacije. Neka su r, s i γ stope rekombinacije između \mathcal{A} i \mathcal{Q} , \mathcal{Q} i \mathcal{B} i \mathcal{A} i \mathcal{B} . Stope rekombinacije su u suštini marginalne verovatnoće da se desi rekombinacija između dva lokusa. Neka E_1, E_2 i E_3 predstavljaju događaje da se dese rekombinacije između \mathcal{A} i \mathcal{Q} , \mathcal{Q} i \mathcal{B} i \mathcal{A} i \mathcal{B} . Dakle, vidimo da je $r = P(E_1), s = P(E_2)$ i $\gamma = P(E_3)$. Neka $\bar{E}_j, j = 1, 2, 3$ predstavlja

Genotip markera	Frekvencija genotipa	Genotip lokusa		
		2	1	0
22	$\frac{P_0^2}{4}$	$\frac{p_{00}^2}{P_0^2}$	$\frac{2p_{00}p_{11}}{P_0^2}$	$\frac{p_{11}^2}{P_0^2}$
21	$\frac{P_0P_1}{2}$	$\frac{p_{00}p_{11}}{P_0P_1}$	$\frac{p_{00}p_{10}+p_{11}p_{01}}{P_0P_1}$	$\frac{p_{11}p_{10}}{P_0P_1}$
20	$\frac{P_1^2}{4}$	$\frac{p_{01}^2}{P_1^2}$	$\frac{2p_{01}p_{10}}{P_1^2}$	$\frac{p_{10}^2}{P_1^2}$
12	$\frac{P_0P_1}{2}$	$\frac{p_{00}p_{10}}{P_0P_1}$	$\frac{p_{00}p_{01}+p_{11}p_{10}}{P_0P_1}$	$\frac{p_{11}p_{01}}{P_0P_1}$
11	$\frac{P_0^2+P_1^2}{2}$	$\frac{p_{00}p_{11}+p_{10}p_{01}}{P_0^2+P_1^2}$	$\frac{p_{00}^2+p_{11}^2+p_{10}^2+p_{01}^2}{P_0^2+P_1^2}$	$\frac{p_{00}p_{11}+p_{10}p_{01}}{P_0^2+P_1^2}$
10	$\frac{P_0P_1}{2}$	$\frac{p_{11}p_{01}}{P_0P_1}$	$\frac{p_{00}p_{01}+p_{11}p_{10}}{P_0P_1}$	$\frac{p_{00}p_{10}}{P_0P_1}$
02	$\frac{P_1^2}{4}$	$\frac{p_{10}^2}{P_1^2}$	$\frac{2p_{01}p_{10}}{P_1^2}$	$\frac{p_{01}^2}{P_1^2}$
01	$\frac{P_0P_1}{2}$	$\frac{p_{11}p_{10}}{P_0P_1}$	$\frac{p_{00}p_{10}+p_{11}p_{01}}{P_0P_1}$	$\frac{p_{00}p_{11}}{P_0P_1}$
00	$\frac{P_0^2}{4}$	$\frac{p_{11}^2}{P_0^2}$	$\frac{2p_{00}p_{11}}{P_0^2}$	$\frac{p_{00}^2}{P_0^2}$

Tabela 2.1: Uslovne verovatnoće lokusa kod kojih je genotip nepoznat.

komplement događaja $E_j, j = 1, 2, 3$.

Sada zajedničke verovatnoće možemo predstaviti kao

$$\begin{aligned} p_{11} &= P(E_1 \cap E_2), & p_{00} &= P(E_1^c \cap E_2^c) \\ p_{10} &= P(E_1 \cap E_2^c), & p_{01} &= P(E_1^c \cap E_2) \\ p_{10} + p_{01} &= P(E_3), & p_{00} + p_{11} &= P(E_3^c) \end{aligned}$$

Primetimo sledeće veze:

$$\begin{aligned} p_{10} &= P(E_1 \cap E_2^c) = 1 - P(E_1^c \cup E_2) \\ &= 1 - P(E_1^c) - P(E_2) + P(E_1^c \cap E_2) \\ &= P(E_1) - P(E_2) + p_{01}, \end{aligned}$$

$$p_{01} = P(E_3) - p_{10} = P(E_3) - P(E_1) + P(E_2) - p_{01},$$

$$\begin{aligned} p_{11} &= 1 - P(E_3) - p_{00} \\ &= 1 - P(E_3) - [1 - P(E_1) - P(E_2) + p_{11}] \\ &= P(E_1) + P(E_2) - P(E_3) + p_{11}, \end{aligned}$$

$$\begin{aligned} p_{00} &= P(E_1^c \cap E_2^c) = 1 - P(E_1 \cup E_2) \\ &= 1 - P(E_1) - P(E_2) + P(E_1 \cap E_2) \\ &= 1 - P(E_1) - P(E_2) + p_{11}. \end{aligned}$$

Na kraju, dobijamo vezu između stope rekombinacije i zajedničkih verovatnoća p_{ij} , $i, j = 0, 1, 2$:

$$\begin{aligned} p_{11} &= \frac{1}{2}[P(E_1) + P(E_2) - P(E_3)] = \frac{1}{2}(r + s - \gamma) \\ p_{10} &= \frac{1}{2}[P(E_1) - P(E_2) + P(E_3)] = \frac{1}{2}(r - s + \gamma) \\ p_{01} &= \frac{1}{2}[P(E_2) - P(E_1) + P(E_3)] = \frac{1}{2}(s - r + \gamma) \\ p_{00} &= 1 - \frac{1}{2}[P(E_1) + P(E_2) - P(E_3)] = 1 - \frac{1}{2}(r + s - \gamma) \end{aligned}$$

Ako prepostavimo da su rekombinacije na svim hromozomima nezavisne, odnosno da krosoveri u jednom regionu hromozoma ne utiču na verovatnoću pojave krosovera u susednom regionu. Tada važi da je $\gamma = r + s - 2rs$, odakle tabela uslovnih verovatnoća dobija sledeći oblik:

Genotip markera	Frekvencija genotipa	Genotip lokusa		
		2	1	0
22	$\frac{(1-\gamma)^2}{4}$	$\frac{(1-r)^2(1-\gamma)^2}{(1-\gamma)^2}$	$\frac{2rs(1-r)(1-s)}{(1-\gamma)^2}$	$\frac{r^2s^2}{(1-\gamma)^2}$
21	$\frac{\gamma(1-\gamma)}{2}$	$\frac{s(1-r)^2(1-s)}{\gamma(1-\gamma)}$	$\frac{r(1-r)(1-2s+2s^2)}{\gamma(1-\gamma)}$	$\frac{r^2s(1-s)}{\gamma(1-\gamma)}$
20	$\frac{\gamma^2}{4}$	$\frac{(1-r)^2s^2}{\gamma^2}$	$\frac{2rs(1-r)(1-s)}{\gamma^2}$	$\frac{r^2(1-s)^2}{\gamma^2}$
12	$\frac{\gamma(1-\gamma)}{2}$	$\frac{r(1-s)^2(1-r)}{\gamma(1-\gamma)}$	$\frac{s(1-s)(1-2r+2r^2)}{\gamma(1-\gamma)}$	$\frac{s^2r(1-r)}{\gamma(1-\gamma)}$
11	$\frac{(1-\gamma)^2+\gamma^2}{2}$	$\frac{2rs(1-r)(1-s)}{(1-\gamma)^2+\gamma^2}$	$\frac{(1-2r+2r^2)(1-2s+2s^2)}{(1-\gamma)^2+\gamma^2}$	$\frac{2rs(1-r)(1-s)}{(1-\gamma)^2+\gamma^2}$
10	$\frac{\gamma(1-\gamma)}{2}$	$\frac{s^2r(1-r)}{\gamma(1-\gamma)}$	$\frac{s(1-s)(1-2r+2r^2)}{\gamma(1-\gamma)}$	$\frac{r(1-s)^2(1-r)}{\gamma(1-\gamma)}$
02	$\frac{\gamma^2}{4}$	$\frac{r^2(1-s)^2}{\gamma^2}$	$\frac{2rs(1-r)(1-s)}{\gamma^2}$	$\frac{(1-r)^2s^2}{\gamma^2}$
01	$\frac{\gamma(1-\gamma)}{2}$	$\frac{r^2s(1-s)}{\gamma(1-\gamma)}$	$\frac{r(1-r)(1-2s+2s^2)}{\gamma(1-\gamma)}$	$\frac{s(1-r)^2(1-s)}{\gamma(1-\gamma)}$
00	$\frac{P_0^2}{4}$	$\frac{r^2s^2}{(1-\gamma)^2}$	$\frac{2rs(1-r)(1-s)}{(1-\gamma)^2}$	$\frac{(1-r)^2(1-\gamma)^2}{(1-\gamma)^2}$

Tabela 2.2: Uslovne verovatnoće lokusa čiji je genotip nepoznat, uz prepostavku nezavisnosti krosovera.

Metod 2 - Lander-Botstein

Intervalno mapiranje lokusa za kvantitativna svojstva koje su razvili Lander i Botstein prepostavlja da u čitavom genomu postoji najviše jedan QTL. Prepostavimo da je lokus uistinu odgovoran za ispoljavanje kvantitativne osobine. Tada genotipovi istog dele čitavu populaciju na više potpopulacija i kvantitativna osobina prati različite raspodele unutar tih potpopulacija. Kada je individua izabrana na slučajan način iz populacije, bez znanja kojoj potpopulaciji pripada, slučajno izabrana osobina poseduje mešavinu Gausovih raspodela (detaljnije u [8], [17]). Mešavina normalnih (Gausovih) raspodela ima oblik

$$F(y; \boldsymbol{\theta}, \mathbf{p}) = \sum_{j=1}^k p_j F_j(y; \boldsymbol{\theta}_j),$$

gde F_j predstavlja diskretnu slučajnu promenljivu sa raspodelom verovatnoća za koju važi $p_j = 0$, $\sum_{j=1}^k p_j = 1$. U slučaju neprekidne slučajne promenljive, tada mešavina Gausovih raspodela ima funkciju gustine:

$$f(y; \boldsymbol{\theta}, \mathbf{p}) = \sum_{j=1}^k p_j f_j(y; \boldsymbol{\theta}_j).$$

Označimo sa (y_1, \dots, y_n) slučajan uzorak iz populacije. Logaritam funkcije verodostojnosti na osnovu slučajnog uzorka je

$$l(\boldsymbol{\theta}, \mathbf{p}) = \sum_{j=1}^k \ln \left(\sum_{j=1}^k p_j f_j(y_i; \boldsymbol{\theta}_j) \right).$$

Prilikom ocenjivanja nepoznatih paramatera u slučaju modela sa mešavinom Gausovih raspodela može se primeniti metoda maksimalne verodostojnosti, u kombinaciji sa takozvanim EM algoritmom (eng. *Expectation-Maximization algorithm*). U ovakvim modelima, članovi potpopulacije su nepoznati.

Takođe, svaka individua se može predstaviti preko vektora $\mathbf{z} = (z_1, \dots, z_{k-1})^t$ koji je definisan sa

$$z_j = \begin{cases} 1, & \text{ako individua pripada } j-\text{toj potpopulaciji,} \\ 0, & \text{inače,} \end{cases} \quad j = 1, \dots, k-1$$

Vektor \mathbf{z} prati politomnu raspodelu iz sekcije 1.3.1,

$$p(\mathbf{z}) = \prod_{j=1}^k p_j^{z_j}, \quad z_k = 1 - \sum_{j=1}^{k-1} z_j, \quad p_k = 1 - \sum_{j=1}^k p_j.$$

Za jedinku i , vektor je oblika $\mathbf{z}_i = (z_{i1}, \dots, z_{ik-1})^t$. Ako su \mathbf{z}_i -ovi poznati, tada imamo "kompletne" podatke $\{(y_i, \mathbf{z}_i), i = 1, \dots, n\}$, dok inače kada nedostaju \mathbf{z}_i -ovi, podatke $\{(y_i), i = 1, \dots, n\}$ nazivamo "nekompletnim" podacima. Zajednička raspodela (y_i, \mathbf{z}_i)

ima gustinu

$$\prod_{j=1}^k [p_j f_j(y_i; \boldsymbol{\theta}_j)]_{ij}^z.$$

Stoga, logaritam funkcije verodostojnosti koji obuhvata "kompletne" podatke je oblika

$$l_c(\boldsymbol{\beta}, \mathbf{p}) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \ln(p_j) + \sum_{i=1}^n \sum_{j=1}^k z_{ij} \ln(f_j(y_i; \boldsymbol{\theta}_j))$$

Sada ćemo malo pojasniti EM algoritam. On se sastoji od dva koraka: u prvom koraku se računa očekivanje logaritma funkcije verodostojnosti na osnovnu kompletne podatke i izračunavaju se nepoznati parametri. U drugom koraku se maksimiziraju uslovna očekivanja u skladu sa dobijenim parametrima. Navedeni koraci se ponavljaju dok nepoznati parametri ne konvergiraj (detaljnije u [1], [3], [16]).

Prilikom primene intervalnog mapiranja, *LOD skor* se koristi kao test statistika. LOD skor predstavlja logaritam sa osnovom 10 količnika funkcija verodostojnosti pod nultom i alternativnom hipotezom. Pretpostavka je koja se koristi je $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$.

Dalje, neka je ocena parametara $\hat{\boldsymbol{\beta}}$ i $\hat{\sigma}^2$ dobijena maksimalnom verodostojnosti pod alternativnom hipotezom, a parametri $\hat{\beta}_0$ i $\hat{\sigma}_0^2$ pod nultom hipotezom. Funkcije maksimalne verodostojnosti pod alternativno i nultom hipotezom $L_1(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ i $L_0(\hat{\beta}_0, \hat{\sigma}_0^2)$. LOD skor test statistika se dobija na sledeći način

$$LOD = \log_{10} \left(\frac{L_0}{L_1} \right).$$

Visoke vrednosti LOD skora nam ukazuju na veću verovatnoću da je alternativna hipoteza tačna. Ukoliko LOD skor prelazi određeni prag značajnosti, tada marker proglašavamo QTL-om ili da je povezan sa QTL-om.[17]

Prilikom primene intervalnog mapiranja u QTL analizi, markeri razdvajaju čitav genom na intervale između dva susedna markera. Na svakom lokusu u intervalu, njegov genotip se može izvesti na osnovu genotipa susednih markera.[8]

Neka Q i q predstavljaju alele potencijalnog QTL-a i A , a i B , B predstavljaju alele susednih markera sa leve i desne strane, respektivno. Neka je x promenljiva koja pred-

stavlja kombinaciju genotipa susednih markera, definisana na sledeći način:

$$x = \begin{cases} 1, & \text{za } AABB, \\ 2, & \text{za } AABb, \\ 3, & \text{za } AAbb, \\ 4, & \text{za } AaBB, \\ 5, & \text{za } AaBb, \\ 6, & \text{za } Aabb, \\ 7, & \text{za } aaBB, \\ 8, & \text{za } aaBb, \\ 9, & \text{za } aabb. \end{cases}$$

Uvedimo sada dve slučajne promenljive koje predstavljaju indikatore događaja:

$$\delta_1 = \begin{cases} 1, & \text{ako je genotip } Qq, \\ 0, & \text{inače;} \end{cases}$$

$$\delta_2 = \begin{cases} 1, & \text{ako je genotip } QQ, \\ 0, & \text{inače.} \end{cases}$$

U narednom koraku izvodimo model koji uključuje mešavinu normalnih raspodela. Prepostavimo da fenotipska vrednost osobine, koja zavisi od genotipa potencijalnog QTL-a, prati normalnu raspodelu

$$y | (\delta_1, \delta_2) \sim \varphi(y; \beta_0 + \beta_1 \delta_1 + \beta_2 \delta_2, \sigma^2),$$

gde $\varphi(y; \mu, \sigma^2)$ predstavlja funkciju gustine normalne raspodele sa parametrima μ i σ^2 . Prema definiciji δ_1 i δ_2 , očekivane fenotipske vrednosti individua sa genotipom Qq , QQ i qq na datom markeru je $\beta_0 + \beta_1$, $\beta_0 + \beta_2$ i β_0 , respektivno. Par (δ_1, δ_2) predstavlja vektor koji prati trinomnu distribuciju. Posmatrajmo interval između dva markera. Neka je r stopa rekombinacije između markera sa leve strane potencijalnog

QTL-a i njega samog, i neka je γ stopa rekombinacije između dva markera koji određuju interval. Stope rekombinacije dobijamo na osnovu genetičkih distanca između markera, koristeći Haldanovu funkciju mapiranja.

Označimo sa $p_1(x, r, \gamma)$ i $p_2(x, r, \gamma)$ uslovne verovatnoće da potencijalni QTL iz intervala poseduje genotip Qq i QQ , respektivno. Funkcija gustine (δ_1, δ_2) je data sa

$$(\delta_1, \delta_2) \sim p_1(x, r, \gamma)^{\delta_1} p_2(x, r, \gamma)^{\delta_2} p_0(x, r, \gamma)^{1-\delta_1-\delta_2},$$

gde $p_0(x, r, \gamma) = 1 - p_1(x, r, \gamma) - p_2(x, r, \gamma)$.

Zakon raspodele promenljivih y i (δ_1, δ_2) je

$$f(y, \delta_1, \delta_2) = p_1(x, r, \gamma)^{\delta_1} p_2(x, r, \gamma)^{\delta_2} p_0(x, r, \gamma)^{1-\delta_1-\delta_2} \varphi(y; \beta_0 + \beta_1 \delta_1 + \beta_2 \delta_2, \sigma^2). \quad (2.5)$$

Marginalna funkcija gustine za y je tada

$$\begin{aligned} f(y; x, \boldsymbol{\beta}, \sigma^2) &= \sum_{(\delta_1, \delta_2)} p_1(x, r, \gamma)^{\delta_1} p_2(x, r, \gamma)^{\delta_2} p_0(x, r, \gamma)^{1-\delta_1-\delta_2} \\ &\quad \times \varphi(y; \beta_0 + \beta_1 \delta_1 + \beta_2 \delta_2, \sigma^2) \\ &= p_1(x, r, \gamma) \varphi(y; \beta_0 + \beta_1, \sigma^2) + p_2(x, r, \gamma) \varphi(y; \beta_0 + \beta_2, \sigma^2) \\ &\quad + p_0(x, r, \gamma) \varphi(y; \beta_0, \sigma^2), \end{aligned} \quad (2.6)$$

gde je $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^t$. Dobijena raspodela predstavlja mešavinu Gausovih raspodela.

Primenimo sada gore navedeni model na određen set podataka. Za svaki interval, imamo podatke oblika $\{(y_i, x_i) : i = 1, \dots, n\}$. Za svaku lokaciju unutar intervala, znajući vrednost r definišemo funkciju verodostojnosti

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n f(y_i, x_i, \boldsymbol{\beta}, \sigma^2). \quad (2.7)$$

LOD skor se tada računa za svaku lokaciju

$$LOD = \log_{10} \frac{L(\hat{\boldsymbol{\beta}}_0, \hat{\sigma}_0^2)}{L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)},$$

gde su $\hat{\beta}$, σ^2 ocene nepoznatih parametara koje su dobijene metodom maskimalne verodostojnosti, dok su $\hat{\beta}_0$, $\hat{\sigma}_0^2$ ocene dobijene metodom maksimalne verodostojnosti pod hipotezom $\beta_1 = \beta_2 = 0$.

U tom slučaju,

$$L(\hat{\beta}_0, \hat{\sigma}_0^2) = \left[\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}} \right]^n \prod_{i=1}^n e^{-\frac{(y_i - \hat{\beta}_0)^2}{2\hat{\sigma}_0^2}},$$

gde je $\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2$, a $\hat{\beta}_0 = \frac{1}{2} \sum_{i=1}^n y_i$.

Dakle, ocene parametara funkcije verodostojnosti ($\hat{\beta}$ i $\hat{\sigma}^2$) se dobijaju primenom EM algoritma. "Kompletne" podatke čine $\{(y_i, \delta_{1i}, \delta_{2i}) : i = 1, \dots, n\}$ i funkcija verodostojnosti na osnovu kompletnih podataka je oblika

$$\begin{aligned} f(y; x, \boldsymbol{\beta}, \sigma^2) &= \sum_{(\delta_1, \delta_2)} p_1(x, r, \gamma)^{\delta_1} p_2(x, r, \gamma)^{\delta_2} p_0(x, r, \gamma)^{1-\delta_1-\delta_2} \\ &\quad \times \varphi(y; \beta_0 + \beta_1 \delta_1 + \beta_2 \delta_2, \sigma^2). \end{aligned} \tag{2.8}$$

Neka je $\boldsymbol{\delta}_k = (\delta_{k1}, \dots, \delta_{kn})^t$, $k = 1, 2$. Znajući da $\delta_{1i}\delta_{2i} = 0$ i $\delta_{ki}^2 = \delta_{ki}$, nakon logaritmovanja funkcije verodostojnosti dobija se

$$l(\beta, \sigma^2; \delta_1, \delta_2) = l_1(\beta, \sigma^2; \delta_1, \delta_2) + l_2(\beta, \sigma^2; \delta_1, \delta_2),$$

gde je

$$l_1(\boldsymbol{\beta}, \sigma^2; \delta_1, \delta_2) = \sum_{i=1}^n \left[\delta_{1i} \ln \frac{p_1(x_i, r, \gamma)}{p_0(x_i, r, \gamma)} + \delta_{2i} \ln \frac{p_2(x_i, r, \gamma)}{p_0(x_i, r, \gamma)} + p_0(x_i, r, \gamma) \right]$$

i

$$\begin{aligned}
l_2(\boldsymbol{\beta}, \sigma^2; \delta_1, \delta_2) &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \delta_{1i} - \beta_2 \delta_{2i})^2 \\
&= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [(y_i - \beta_0)^2 - 2(y_i - \beta_0)(\beta_1 \delta_{1i} + \beta_2 \delta_{2i}) \\
&\quad + \beta_1^2 \delta_{1i}^2 + \beta_2^2 \delta_{2i}^2].
\end{aligned}$$

Prvi korak u EM algoritmu uključuje određivanje očekivane uslovne verovatnoće vrednosti parametara δ_{ki} , $k = 1, 2$ (E korak). Kako je funkcija l linearna po δ_{ki} , uslovne verovatnoće se dobijaju kao

$$\delta_{ki}^{(s)} = \frac{p_k(x_i, r, \gamma) \varphi(y_i; \mu_k^{(0)}, \sigma^{2(0)})}{\sum_{l=0}^k p_k(x_i, r, \gamma) \varphi(y_i; \mu_k^{(0)}, \sigma^{2(0)})}, \quad s = 1, 2, \dots$$

gde je $\mu_k^{(0)}$, $k = 1, 2$ i $\sigma^{2(0)}$ predstavljaju početne proizvoljne vrednosti nepoznatih parametara, a $\mu_0^{(0)} = \beta_0^{(0)}$, $\mu_1^{(0)} = \beta_0^{(0)} + \beta_1^{(0)}$, $\mu_2^{(0)} = \beta_0^{(0)} + \beta_2^{(0)}$.

Kako l_1 ne sadrži nepoznate parametre, u M koraku traži se maksimum samo za funkciju l_2 , gde δ_{ki} ažuriramo sa $\delta_{ki}^{(s)}$, $s = 1, 2, \dots$

Dakle, nakon ažuriranja funkcije l_2 sa novim $\delta_{ki}^{(s)}$, dobijamo

$$\begin{aligned}
l_2(\boldsymbol{\beta}, \sigma^2; \delta_1, \delta_2) &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 \delta_{1i}^{(s)} - \beta_2 \delta_{2i}^{(s)})^2 \\
&= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [(y_i - \beta_0)^2 - 2(y_i - \beta_0)(\beta_1 \delta_{1i}^{(s)} + \beta_2 \delta_{2i}^{(s)}) \\
&\quad + \beta_1^2 \delta_{1i}^{(s)} + \beta_2^2 \delta_{2i}^{(s)}], \quad s = 1, 2, \dots
\end{aligned}$$

Radeći prvi izvod po nepoznatim parametrima i izjednačavajući ih sa nulom, a zatim i rešavajući sistem jednačina, dobijamo sledeće ocene parametara:

$$\begin{aligned}\beta_0^{(s+1)} &= \frac{\bar{y} - \bar{\delta}_1 \bar{y}_{\delta_1} - \bar{\delta}_2 \bar{y}_{\delta_2}}{1 - \bar{\delta}_1 - \bar{\delta}_2} \\ \beta_1^{(s+1)} &= \bar{y}_{\delta_1} - \beta_0^{(s+1)} \\ \beta_2^{(s+1)} &= \bar{y}_{\delta_2} - \beta_0^{(s+1)} \\ \sigma^{2(s+1)} &= \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0^{(s+1)} - \beta_1^{(s+1)} \delta_{1i}^{(s)} - \beta_2^{(s+1)} \delta_{2i}^{(s)})^2,\end{aligned}$$

$$\text{gde } \bar{\delta}_k = \sum_{i=1}^n \delta_{ki}, \text{ a } \bar{y}_{\delta_k} = \frac{\sum_{i=1}^n y_i \delta_{ki}^{(s+1)}}{\sum_{i=1}^n \delta_{ki}^{(s+1)}}, k = 1, 2.$$

Metod 3 - Haley-Knott regresija

Prethodno obrađen model mešavine raspodela prepostavlja normalnost kvantitativne osobine. Primena Haley-Knott regresije ne prepostavlja normalnost raspodele. Isto tako, vreme potrebno za primenu Haley-Knott regresije je manje zahtevno od metode mešavine raspodela.[6] Dakle na osnovu modela (2.3) iz poglavlja **2.2.2**

$$y = \begin{cases} \beta_0 + 2c + \varepsilon, & \text{ako je genotip markera } QQ, \\ \beta_0 + c + d + \varepsilon, & \text{ako je genotip markera } Qq, \\ \beta_0 + \varepsilon, & \text{ako je genotip markera } qq. \end{cases}$$

gde su c i d aditivni i dominantni efekat QTL respektivno, i ε prati raspodelu sa očekivanjem 0 i varijansom σ^2 . Dakle, ε ne mora nužno da prati normalnu raspodelu. Tada se očekivanje promenljive y_i računa kao

$$\begin{aligned}E(y_i) &= \beta_0 p_0(x_i, r, \gamma) + (\beta_0 + c + d) p_1(x_i, r, \gamma) + (\beta_0 + 2c) p_2(x_i, r, \gamma) \\ &= \beta_0 + cp_1(x_i, r, \gamma) + 2p_2(x_i, r, \gamma) + dp_1(x_i, r, \gamma).\end{aligned}$$

Neka je sada $z_{i1} = p_1(x_i, r, \gamma) + 2p_2(x_i, r, \gamma)$, $z_{i2} = p_1(x_i, r, \gamma)$, $c = \beta_1$ i $d = \beta_2$. Tada

linearni model poprima sledeći oblik:

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \varepsilon_i, i = 1, \dots, n. \quad (2.9)$$

Primetimo da kovarijate u modelu zavise od r . Prvi korak prilikom implementacije Haley-Knott regresije je da se izračunaju kovarijate z_{i1}, z_{i2} na osnovu uslovnih verovatnoća dobijenih od susednih markera koji određuju intervale. Haley-Knott regresija tretira model za svaku regresiju kao običan regresioni model i testira nullu hipotezu $H_0 : \beta_1 = \beta_2 = 0$. Test statistika koja se koristi je F-test statistika, koja prati Fišerovu raspodelu. Strategija koja se koristi je ista kao i kod prethodnog modela: lokusi čije F vrednosti koje prelaze određeni prag značajnosti se proglašavaju lokusom koji utiče na ispoljavanje osobine.

2.3. Permutacioni test

Postavljajući proizvoljne pragove značajnosti dovodi do mogućnosti pojave greške prve vrste (kada ne postoji potencijalni QTL, ali je ipak pronađen marker statistički značajan). Permutacioni test se koristi prilikom određivanja praga značajnosti LOP vrednosti kvantitativnih lokusa. Lokusi čija vrednost test statistike pređe prag značajnosti proglašava se potencijalnim QTL-om (detaljnije u [9], [15], [17]).

Znamo da se prilikom testiranja hipoteza obično koristi test statistika i neka je ona oblika $T = f(X_1, X_2, \dots, X_n)$. Statistički test se sprovodi tako što se odredi odgovarajuća kritična oblast, (oblast odbacivanja nulte hipoteze) $C(T)$, vezana sa statistiku T i odluka se donosi na sledeći način:

- **Ako registrirana vrednost t_{reg} statistike T ne pripada $C(T)$, nemamo razloga da odbacimo hipotezu H_0 .**
- **Ako registrirana vrednost t_{reg} statistike T pripada $C(T)$, hipoteza H_0 se odbacuje.**

Permutacioni test obuhvata permutaciju rezultata fenotipskih merenja svih jedinki iz populacije. Nakon permutacije, primenjujemo identičan statistički metod, koji je rađen

na originalnom setu podataka. Navedeni proces se ponavlja određen broj puta, uzimajući iz svake dobijene permutacije maksimalne LOP vrednosti. Na taj način dobijamo vektor dužine n koji se sastoji od maksimalnih LOP vrednosti. Distribucija ovih vrednosti nam daje ocenu praga značajnosti, iznad koje proglašavamo lokus značajnim.

Dakle, permutacioni test predstavlja metodu determinacije praga značajnosti. Primena navedenog metoda ne zahteva dodatne pretpostavke. Ideja permutacionog testa je da se izmešaju rezultati fenotipskih merenja za svih n jedinki iz populacije, a da pritom genetska mapa koja se sastoji od markera i njihovih pozicija ostane nepromenjena i da se nakon toga primeni odgovarajuća statistička metoda na nove, permutovane podatke. [15]

Algoritam za permutacioni test

Pretpostavimo da se QTL analiza vrši nad n jedinki, tj. da se populacija sastoji od n jedinki. Definišimo vektor $s(K) = (s_1, \dots, s_K)$ koji predstavlja permutovani vektor fenotipskih merenja osobine, gde K predstavlja broj permutacija. Označimo originalni set podataka $\{(y_i, x_{i1}, \dots, x_{iM}) : i = 1, \dots, n\}$, gde y_i predstavlja fenotipske vrednosti posmatrane osobine od interesa, x_{i1}, \dots, x_{iM} genotipske podatke i -te individue, a M broj markera koji se koristi u analizi. Definišimo promenljivu $T(y)$ koja predstavlja test statistiku dobijenu na osnovu originalnih podataka, npr. test količnika funkcija verodostojnosti, koji se može koristiti u testiranju hipoteze da nema tačke promene protiv alternativne hipoteze da postoji promena.

Neka $\{(y_{si}, x_{i1}, \dots, x_{iM}) : i = 1, \dots, n\}$ predstavlja permutovane podatke i označimo sa $T_{max}(\mathbf{y}_s)$ maksimum test statistike dobijene na osnovu permutovanih podataka.

Dakle, algoritam za permutacioni test je sledeći: permutovati fenotipske vrednosti individua i izračunati test statistiku na osnovu permutovanih podataka. Neka maksimum test statistike dobijene u k -toj permutaciji zapišemo sa $T_{max}(\mathbf{y}_s^{(k)})$. Tada dobijena distribucija $\{T_{max}(\mathbf{y}_s^{(k)}) : k = 1, \dots, K\}$ je aproksimacija raspodele $T_{max}(\mathbf{y})$. Prag značajnosti na nivou α je aproksimiran sa $(1 - \alpha)$ percentilom empirijske distribucije $\{T_{max}(\mathbf{y}_s^{(k)}) : k = 1, \dots, K\}$.

Permutacijom fenotipskih vrednosti osobine između svih jedinki u populaciji prekidamo vezu između osobine sa određenim lokusom bilo da je QTL ili jednostavno neki

nepovezan marker. Ako bi postojali QTL-ovi u genomu, tada bi permutovane vrednosti fenotipskih merenja imali distribuciju koja je mešavina više raspodelu, i to takvu koja nije povezana ni sa jednim lokusom. Ako ne postoje QTL-ovi, permutovani podaci bi se slagali sa raspodelom koja ne predstavlja mešavinu više raspodela.

Da bi se izveo permutacioni test, potrebno je odrediti koliki je broj K , tj. odrediti potreban broj permutacija. Što veći broj permutacija K , to su bolje aproksimacije praga značajnosti iznad kog proglašavamo lokus značajnim. Sa druge strane, veći broj permutacija K je vremenski zahtevna. Preporuka je da $K \geq 1000$ za $\alpha = 0.05$ i $K \geq 10000$ za $\alpha = 0.01$. [9]

Permutacioni test je univerzalni metod, kome je potrebno više vremena prilikom izvođenja simulacija, u odnosu na druge simulacione metode.

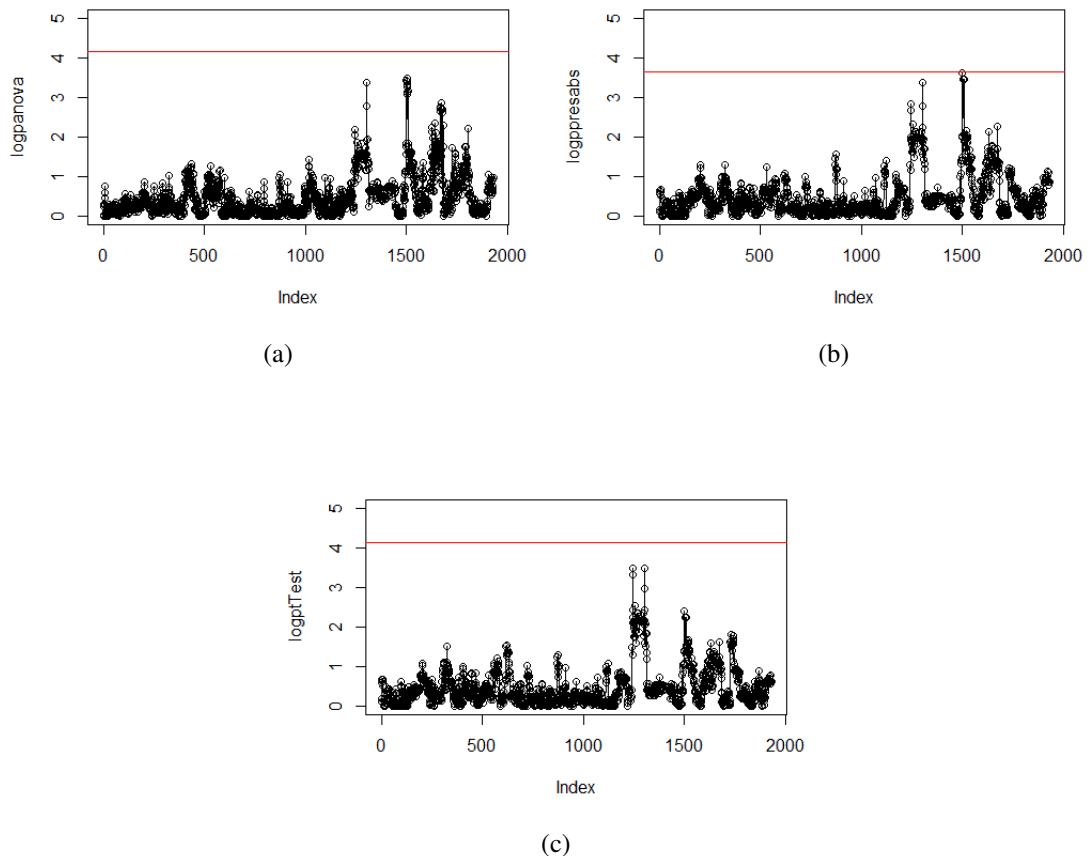
3 Numerički rezultati

3.1. Implementacija modela

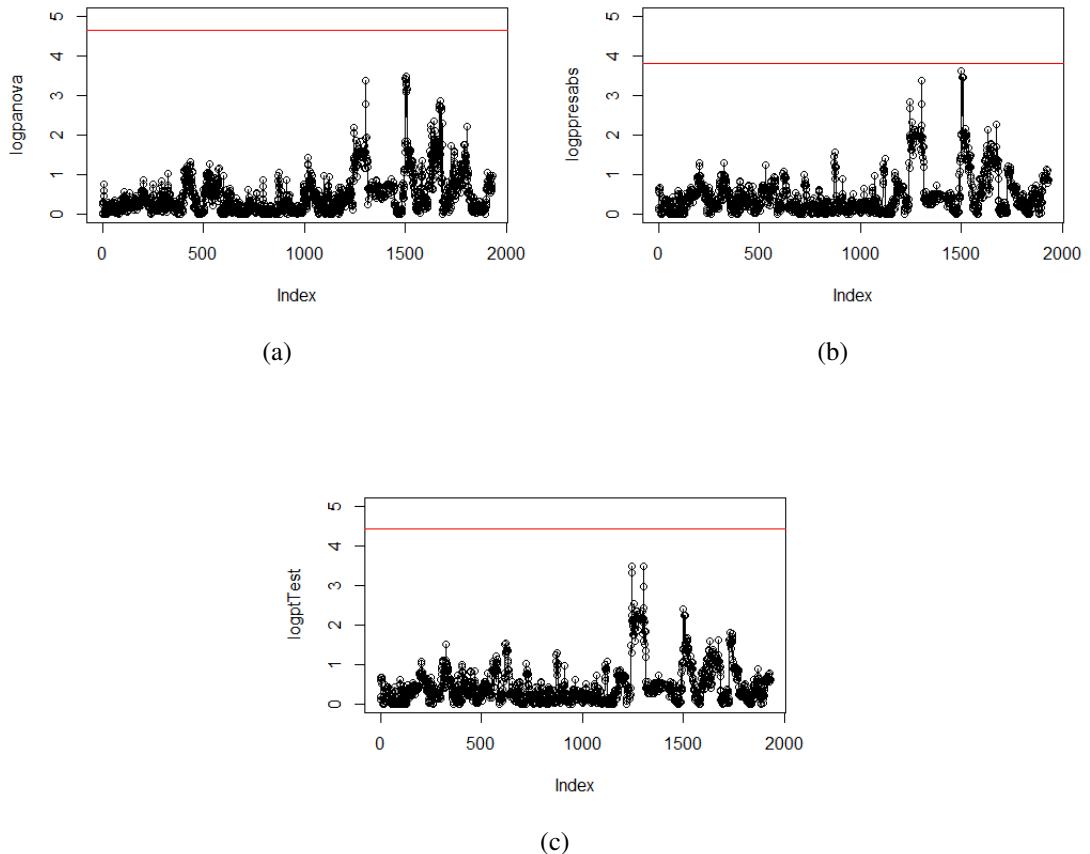
Modeli koji su predstavljeni u prethodnom poglavlju će biti prikazani u ovom delu rada, kao i njihova implementacija u softverskom paketu R, verziji 0.98.507. Posmatrajući raznovrsnosti u osobinama ruža, autor se odlučio da u ovom radu obradi otpornost na niske temperature, koristeći podatke dobijene iz firme Pheno Geno Roses iz Novog Sada. Ovi podaci uključuju populaciju koja sadrži 103 jedinke, koje su proizvod ukrštanja dve različite populacije. Prvi model će biti primenjen na tetraploidnu, a drugi i treći će se primeniti na hipotetičku diploidnu populaciju. Fenotipski rezultati vezani za otpornost ruže na niske temperature su dobijeni na taj način što je za svaku granu meren procenat izmrzlog dela grane u odnosu na čitavu dužinu grane, i onda je uzeta srednja vrednost za sve grane i-te individue. Sledeći deo rada sadrži rezultate za svaki od predstavljenih modela, kao i pojedinačnu analizu rezultata.

3.1.1. Model 1

Prvi model, koji prepostavlja da je svaki marker potencijalni QTL, primeniće se na tetraploidnu populaciju. Ovaj model se zasniva na primeni analize varijanse za svaki marker, i to u odnosu na dozu alela sa jedne strane, i u odnosu na prisustvo ili odsustvo alela sa druge strane. Pored analize varijanse, primenjen je i Studentov t-test na podatke kod kojih se posmatra prisustvo ili odsustvo alela. Prilikom primene navedenih testova, pragovi značajnosti su određeni na nivoima značajnosti $\alpha = 0.1$ i $\alpha = 0.05$. Grafici 3.1. i 3.2. predstavljaju rezultate primene prvog modela na tetraploidnu populaciju.



Grafik 3.1: Rezultati primene analize varijanse u odnosu na dozu alela (a) i u odnosu na prisustvo ili odsustvo alela (b), kao i t-testa (c) na nivou značajnosti $\alpha = 0.1$. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija.



Grafik 3.2: Rezultati primene analize varijanse u odnosu na dozu alela (a) i u odnosu na prisustvo ili odsustvo alela (b), kao i t-testa (c) na nivou značajnosti $\alpha = 0.05$. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija.

Posmatrajući grafike kod kojih je uzet nivo značajnosti $\alpha = 0.1$, možemo izdvojiti markere *RhK5_338_425R* i *Rh12GR_51628_738R* mogu biti potencijalni lokusi koji utiču na otpornost biljke na niske temperature, iako njihove LOP vrednosti ne prelaze prag značajnosti koji iznosi 4.154751874, 3.651072051 i 4.141287094 u slučaju primene analize varijanse u odnosu na dozu alela, u odnosu na prisustvo/odsustvo alela i t testa, respektivno. U slučaju primene t testa, vidimo da LOP vrednost markera *Rh12GR_51628_738R* nije najveća, ali se može primetiti da odstupa od LOP vrednosti većine markera. Ono što se takođe može primetiti na graficima 3.1. i 3.2. je da su, prilikom primene analize varijanse, najviše LOP vrednosti dobijene za lokuse na hro-

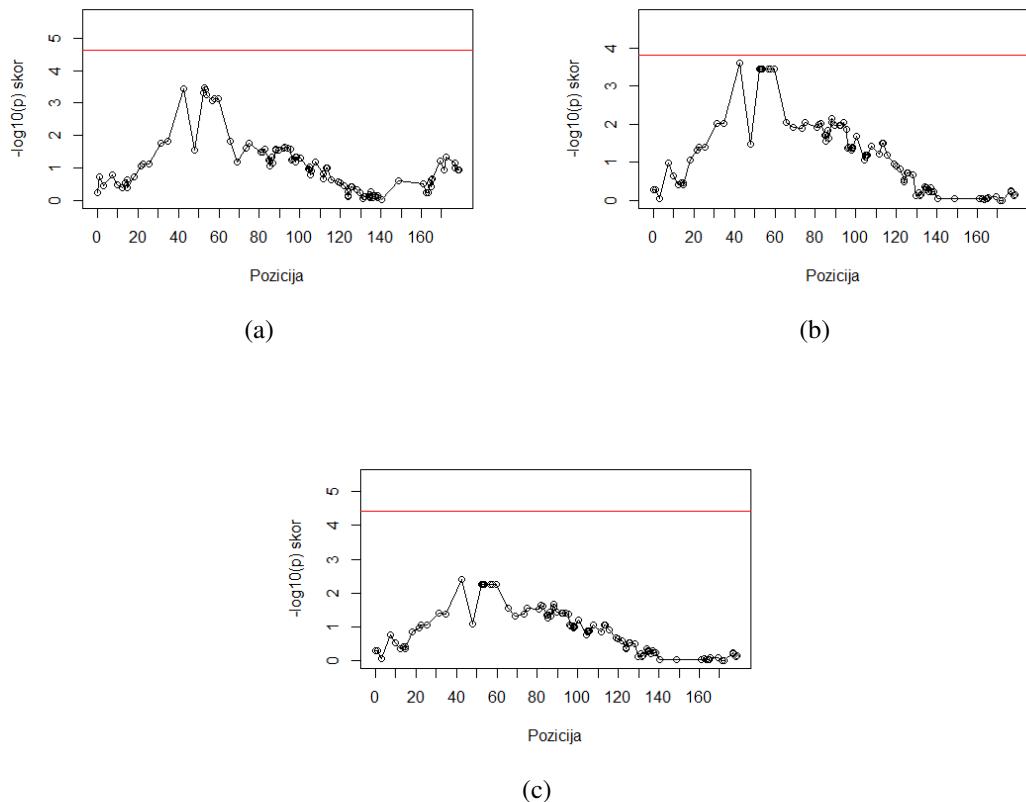
mozomu 1b, tačnije u regionu od 52.324 cM do 59.924 cM, a zatim i na hromozomu 1a. Ako se pogleda grafik 3.2., koji se odnosi na nivo značajnosti od $\alpha = 0.05$, može se primetiti da takođe ne postoji LOP vrednost koja prelazi pragove značajnosti ni za jedan od tri testa. Ali identičan region na hromozomu 1b se može izdvojiti. LOP vrednosti potencijalno značajnih markera koji se nalaze u navedenom regionu, kao i pragovi značajnosti, su predstavljeni u tabeli 3.1. Što se tiče hromozoma 1a, premda se mogu primetiti dva markera koji se izdvajaju i koji su uz to jedan drugom susedni, s obzirom da ne postoji više značajnih markera u njihovoј okolini, možemo zaključiti da navedena dva markera nisu potencijalni odgovorni lokusi.

Markeri	Hromozom	Pozicija	ANOVA (doza alela)	ANOVA (prisustvo/odsustvo)	t test
RhK5_338_425R	1b	52.815	3.484966	3.456969	2.246357
Rh12GR_51628_738R	1b	42.63	3.428384	3.603424	2.408971
Rh12GR_12827_735R	1b	53.305	3.40384	3.456969	2.246357
RhK5_13686_1038R	1b	52.324	3.325494	3.367405	3.490243
RhK5_16002_554R	1b	53.795	3.25246	3.456969	2.246357
Rh12GR_53166_2167R	1b	59.924	3.149556	3.456969	2.246357
Rh88_26067_604R	1b	57.893	3.122941	3.456969	2.246357
Rh12GR_86975_155R	1b	56.902	3.085791	3.456969	2.246357

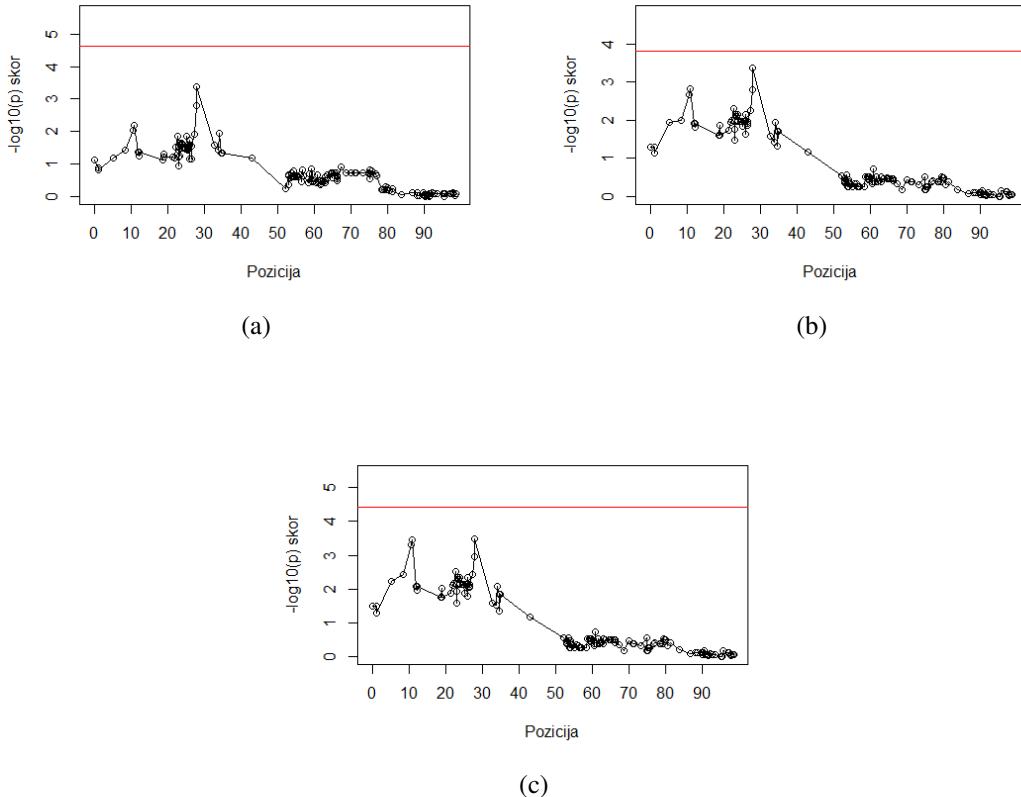
Tabela 3.1: Rezultati test statistika dobijeni primenom modela 1 na tetraploidnu populaciju.

Hromozomi 1b i 1a su respektivno predstavljeni graficima 3.3. i 3.4., gde smo uzeli u obzir nivo značajnosti od 0.05. Treba primetiti i da se marker *Rh12GR_51628_738R* izdvaja od susednih markera, tačnije susedni markeri nemaju približne LOP vrednosti, što možemo videti na grafiku 3.3. Znajući da je efekat markera koji su susedni nekom značajnom lokusu približno značajan kao i efekat odgovornog makera, zaključili bismo da *Rh12GR_51628_738R* nije QTL i odbacili bi ga iz dalje analize. Međutim, ne želimo da ga odbacimo sa sigurnošću iz više razloga. Prvi razlog je taj što između navedenog markera i prvih susednih markera sa leve i desne strane postoji razlika u rastojanju od čak 7 cM i 8 cM, što nam ukazuje na činjenicu da neki markeri nedostaju sa mape. Drugi

razlog zbog čega ne bi trebalo da odbacimo marker *Rh12GR_51628_738R* je taj što se između njega i potencijalno značajnog regiona koji smo prethodno naveli nalazi samo jedan marker (*Rh12GR_28725_1492R* na poziciji 48.041 cM), čija LOP vrednost veoma odstupa od markera koji se smatraju potencijalnim QTL-om, što se može primetiti na grafiku 3.3. Ova činjenica nas navodi na zaključak da je došlo do dvostrukе redukcije, zbog koje je moguće da su markeri u ovom regionu od 42 do 52 cM odbačeni, jer nisu bili dovoljno informativni, dok činjenica da jedini marker koji se nalazi u navedenoj oblasti i čija LOP vrednost odskače od svojih najbližih suseda implicira da, takođe usled dvostrukе redukcije, doza istog nije tačno određena, te da zbog toga dolazi do drastičnog odstupanja.



Grafik 3.3: Rezultati primene analize varijanse u odnosu na dozu alela (a) i u odnosu na prisustvo ili odsustvo alela (b), kao i t-testa (c) na nivou značajnosti $\alpha = 0.05$ na hromozomu 1b. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija



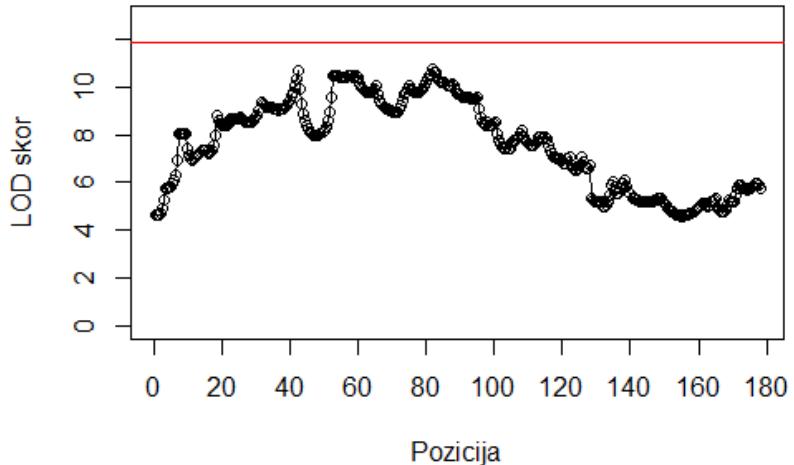
Grafik 3.4: Rezultati primene analize varijanse u odnosu na dozu alela (a) i u odnosu na prisustvo ili odsustvo alela (b), kao i t-testa (c) na nivou značajnosti $\alpha = 0.05$ na hromozomu 1a. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija

Na osnovu svega navedenog, možemo zaključiti da se u regionu od 52.324 cM do 59.924 na hromozomu 1b nalaze potencijalni lokusi koji su odgovorni za ispoljavanje otpornosti na niske temperature, s tim da marker *Rh12GR_51628_738R* na poziciji 42.63 ne odbacujemo i smatramo da je takođe i on potencijalni lokus.

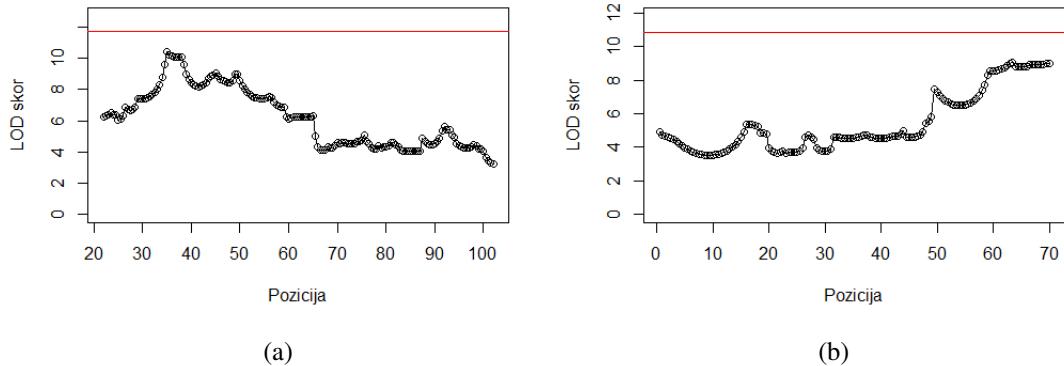
3.1.2. Model 2

Kako bi se primenio model intervalnog mapiranja, korišćena je hipotetička diploidna populacija, s obzirom da je primena na tetraploidnu populaciju zahtevala kompleksno

računanje uslovnih verovatnoća iz poglavlja 2.2.3. Model intervalnog mapiranja je vremenski daleko najzahtevniji i prosečno najviše vremena je bilo potrebno za obradu podataka. Što se tiče EM algoritma, primećeno je da nepoznati parametri konvergiraju nakon 30 iteracija, te je $n = 30$ uzet u obzir prilikom primene ovog modela, gde je n broj iteracija EM algoritma. Kako nijedan lokus nije prešao prag značajnosti, može se ipak izdvojiti jedan region na hromozomu 1b za koji možemo reći da utiče na ispoljavanje osobine. Međutim, možemo primetiti regije na hromozomima 21 i 24, koji se takođe nalaze relativno blizu praga značajnosti, i koji se mogu izdvojiti kao potencijalno značajni regioni. Posmatrajući hromozom 24, postoji mogućnost da je došlo do dvostrukе redukcije, kao i da nedostaje deo markera na kraju hromozoma, te da upravo iz tog razloga nemamo oblik zvona što je karakteristično prilikom mapiranja QTL-ova. Činjenica da nedostaje deo markera jeste upravo zbog dvostrukе redukcije, na osnovu koje su potencijalni markeri na navedenom hromozomu odbačeni, jer nisu dovoljno informativni. Rezultati za hromozome 1b, 21 i 24 su predstavljeni graficima 3.5. i 3.6.



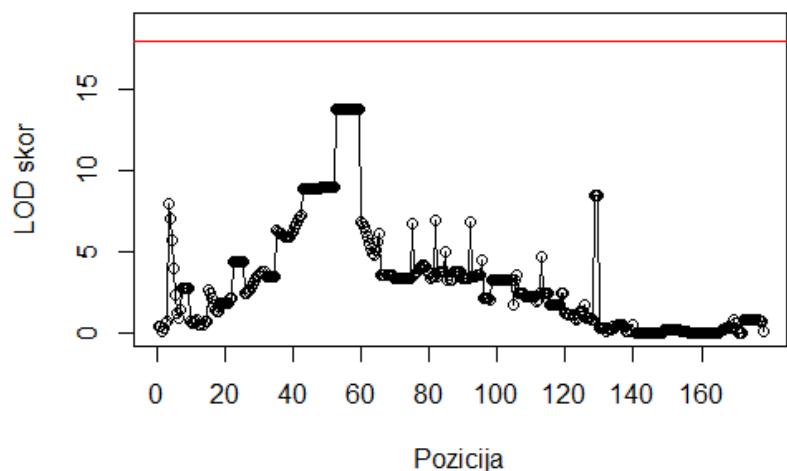
Grafik 3.5: Rezultati nakon primene modela 2 na hromozomu 1b na nivou značajnosti $\alpha = 0.1$. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija.



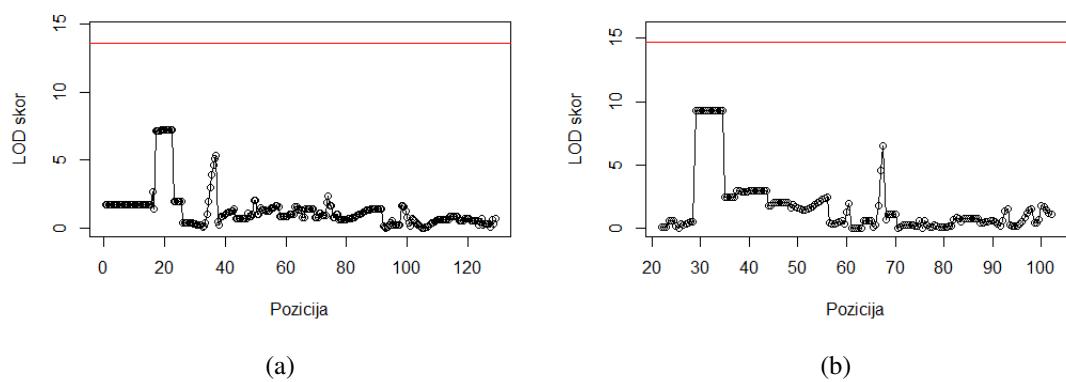
Grafik 3.6: Rezultati primene modela 2 na hromozomima 21 i 24 na nivou značajnosti $\alpha = 0.1$. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija.

3.1.3. Model 3

Treći model, koji se odnosi na Haley-Knott regresiju, je primenjen na diploidnu populaciju, iz istog razloga kao i model 2, zbog komplikovanog izvođenja uslovnih verovatnoća. Što se tiče vremena potrebnog za obradu podataka, model koji su izveli Haley i Knott nije toliko zahtevan kao model 2, iako je i on bio vremenski zahtevan. Na osnovu obrađenih podataka, premda nije bilo lokusa koji su prešli prag značajnosti, vidimo da se na hromozomu 1b rezultati grupišu oko jednog regionala, kao i u slučaju primene metode maksimalne verodostojnosti iz intervalnog mapiranja Lander-a i Botstein-a. Dakle, nije u pitanju jedan lokus koji se izdvojio, već više susednih lokusa, što nam ukazuje da postoji mogućnost da su isti odgovorni za ispoljavanje osobine, s obzirom da se geni najčešće nasleđuju zajedno, tzv. haplotipovi, o kojima je bilo reči u prvom poglavlju. Dodatno, na hromozomima 13b i 21 mogu se primetiti potencijalno značajni regionali. Posmatrajući prethodni model i upoređujući ga sa modelom koji su izveli Haley i Knott, vidimo da su se izdvojila dva regionala koja su potencijalno odgovorni za ispoljavanje osobine, a to su hromozomi 1b i 21. Rezultati hromozoma 1b, 13b i 21 su predstavljeni graficima 3.7. i 3.8.



Grafik 3.7: Rezultati nakon primene Haley-Knott regresije na hromozomu 1b na nivou značajnosti $\alpha = 0.1$. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija.



Grafik 3.8: Rezultati primene Haley-Knott regresije na hromozomima 13b i 21 na nivou značajnosti $\alpha = 0.1$. Prag značajnosti je označen crvenom bojom, dobijen na osnovu 500 permutacija.

3.2. Modeli interakcije - primena višestruke regresije

Kao što je spomenuto u prvom poglavlju, moguća je pojava više lokusa koji su odgovorni za ispoljanje kvantitativne osobine. U tom slučaju, potrebno je odrediti efekte interakcije datih lokusa, odnosno efekte epistaze. Postojanje interakcije između lokusa ukazuje na situaciju u kojoj efekat jednog lokusa je drugaćiji za različit oblik genotipa drugog lokusa.

U slučaju postojanja interakcije između dva lokusa, modeli koji se koristi je

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2 + \varepsilon,$$

gde je x_j promenljiva definisana kao

$$x_j = \begin{cases} 1, & \text{ako je lokus } j \text{ heterozigot,} \\ 0, & \text{ako je lokus } j \text{ homozigot,} \end{cases} \quad j = 1, 2.$$

Ukoliko bismo generalizovali model na m lokusa, tada model koji ispituje epistazu poprima oblik

$$y_i = \alpha + \sum_{j=1}^m \beta_j x_j + \sum_{\substack{k=1 \\ j \neq k}}^p \gamma_{jk} x_j x_k + \varepsilon_i, \quad i = 1, \dots, n.$$

Kako su prilikom primene intervalnog mapiranja dobijeni regioni (intervali na hromozomima), a ne specifični markeri, prilikom provere epistaze korišćeni su rezultati primene modela broj 1 na tetraploidnu populaciju. Dakle, markeri koji su uzeti u obzir su $Rh12GR_51628_738R$ i $RhK5_338_425R$ i oni se nalaze na hromozomu 1b. Označimo ih x_1 i x_2 , respektivno. Rezultati nakon primene višestruke regresije kao modela interakcije su dati u sledećoj tabeli:

Markeri	p vrednost	LOP vrednost
$y = x_1$	0.000327	3.484966
$y = x_2$	0.000373	3.428384
$y = x_1 + x_2$	0.0002695	3.569441
$y = x_1 + x_2 + x_1 \cdot x_2$	0.0002868	3.542421

Tabela 3.2: LOP vrednosti za markere $Rh12GR_51628_738R$ i $RhK5_338_425R$, gde je $Rh12GR_51628_738R$ označen sa x_1 , $RhK5_338_425R$ označen sa x_2 , aditivni model sa dva markera označen je sa $y = x_1 + x_2$, dok je interakcija označena sa $y = x_1 + x_2 + x_1 \cdot x_2$.

Posmatrajući i upoređujući LOP vrednosti dobijenih pojedinačno od svakog markera i LOP vrednosti dobijene na osnovu modela sa i bez njihove interakcije, dolazimo do sledećeg zaključka: nakon uključivanja interakcije između dva markera, LOP vrednost se neznatno smanjila u odnosu na aditivni model. Kako opet aditivni model sa navedena dva markera i model interakcije imaju veće LOP vrednosti u odnosu na pojedinačne efekte istih, implicira da markeri zajedno imaju veći efekat na otpornost biljke na niske temperature nego kad se posmatraju pojedinačno.

4 Zaključak

Ideja ovog rada je bio da ukaže kolika je značajna i prisna veza između statistike i genetike kao naučnih disciplina. Ta prisnost se može videti i u činjenici da je, kao što je i rečeno na početku rada, Mendel koji se smatra začetnikom genetike, po profesiji bio diplomirani matematičar. QTL analiza se danas umnogome koristi kao osnovni alat za određivanje markera koji su povezani sa fenotipskom osobinom. Nedostatak ovog pristupa je što dobijeni rezultati se odnose samo za datu populaciju i ne mogu se bez prethodne validacije primeniti na neku drugu populaciju, kao i to da se čak u dobijenim regionima mogu nalaziti stotine gena koji su potencijalno bitni lokusi.

U radu su obrađena tri modela, koji su implementirani uz pomoć softverskog paketa R. Rezultati ovog rada se nalaze u poglavljju broj 3 i ukazuju na to da, iako nijedan lokus prilikom primene analize pojedinačnog markera i intervalnog mapiranja nije prešao prag značajnosti, mogu se primetiti određeni regioni u genomu koji se ističu. Pragovi značajnosti su dobijeni izvođenjem permutacionih testova za svaku metodu mapiranja na nivoima značajnosti od $\alpha = 0.1$ i $\alpha = 0.05$.

Postoje još mnogi modeli koji uključuju korišćenje Bejzove statistike, uz primenu skrivenih lanaca Markova (detaljnije u [4] i [5]), kao i mnoge modifikacije već postojećih modela. U ovom radu je obrađen model interakcije između markera koji su pokazali visoke LOP vrednosti. U tom smeru, pažnja se može usmeriti i na modele koji uzimaju u obzir interakciju između više markera, kao i one modele interakcije koji uzimaju u obzir i markere koji ne poseduju visoke LOP vrednosti, ali čiji efekat interakcije je potencijalno statistički značajan. Iako možda ovakav vid pretrage nije najisplativiji, moguće je identifikovati veći broj značajnih regiona u genomu.

U radu je pretpostavljeno da su krosoveri nezavisni jedni od drugih, što u realnim

situacijama nije slučaj, pa bi bilo interesantno ispitati moguće modifikacije modela.

Dakle, ono što je budućnost, moguće je proširenje modela na poliploidne organizme, i bilo bi veoma interesantno pronaći odgovarajuće uslovne verovatnoće za poliploide i implementirati ih u postojeće modele intervalnog mapiranja.

Literatura

- [1] A.P. Dempster, N.M. Laird, D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, B, 39 (1), 1–38, 1977.
- [2] A.S. Nikolić, *Identifikacija genskih lokusa (QTL) uključenih u kontrolu odgovora kukuruza (Zea Mays L.) na stres suše*, Univerzitet u Beogradu, Biološki fakultet, 2012.
- [3] C. Do, S. Batzoglou, *What is the expectation maximization algorithm?*, Nature Publishing Group, 2008.
- [4] C.A. Hackett, J.E. Bradshaw, G.J. Bryan, *QTL mapping in autotetraploids using SNP dosage information*, Theor Appl Genet 127, 1885–1904, 2014.
- [5] C.A. Hackett, J.E. Bradshaw, J.W. McNicol, *Interval Mapping of Quantitative Trait Loci in Autotetraploid Species*, Genetics Society of America, 2001.
- [6] C.S. Haley, S.A. Knott, *A simple regression method for mapping quantitative trait loci in line crosses using flanking markers*, Heredity 69, 315–324, 1992.
- [7] D. Rajter-Ćirić, *Verovatnoća*, Univerzitet u Novom Sadu, Prirodno-matematički fakultet, 2009.
- [8] E.S. Lander, D. Botstein, *Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps*, Genetics Society of America, 1989.
- [9] G.A. Churchill, R.W. Doerge, *Empirical threshold values for quantitative trait mapping*, Genetics, 138, 963–971, 1994.

- [10] J.B.S. Haldane, *The combination of linkage values, and the calculation of distances between the loci of linked factors*, J. Genet., 8, 299–309, 1919.
- [11] K. Ljungberg, *Numerical Methods for Mapping of Multiple QTL*, Uppsala Universitet, Uppsala, Švedska, 2013.
- [12] K.W. Broman, *Review of statistical methods for QTL mapping in experimental crosses*, Lab Animal 30(7), 44–52, 2001.
- [13] M. Pavlica, V. Zoldoš, T. Horvat, N. Malenica, J. Mlinarec, V. Vičić, *Praktikum iz genetike - skripta*, Sveučilište u Zagrebu, Prirodno-matematički fakultet, Biološki odsjek, 2013.
- [14] M. Vukosavljev, *Towards marker assisted breeding in garden roses: from marker development to QTL detection*, Wageningen University, Wageningen, 2014.
- [15] P. Bourke, *QTL analysis in polyploids Model testing and power calculations for an autotetraploid*, Laboratory of Plant Breeding, Wageningen University and Research Centre, 2014.
- [16] R.J. Kerr, G.M. McLachlan, J.M. Henshall, *Use of the EM algorithm to detect QTL affecting multiple-trait in an across half-sib family analysis*, Genet. Sel. Evol. 37, 83–103, 2005.
- [17] Z. Chen, *Statistical Methods for QTL Mapping*, National University of Singapore, Singapore, 2013.
- [18] Z. Lozanov-Crvenković, *Statistika*, Novi Sad, 2012.

Biografija



Igor Stranjanac je rođen 07.02.1990. u Novom Sadu. Osnovnu školu „Dositej Obradović“ u Novom Sadu završio je 2005. godine. Upisuje Srednju ekonomsku školu „Svetozar Miletić“ u Novom Sadu koju završava 2009. godine, kao nosilac Vukove diplome.

Iste godine upisao je osnovne akademske studije Primenjene matematike na Prirodno-matematičkom fakultetu u Novom Sadu, modul Matematika finansija. Zaključno sa oktobarskim rokom 2012. godine stiče zvanje Matematičar primjenjene matematike i iste godine upisuje master studije na istom usmerenju. Zaključno sa septembarskim ispitnim rokom 2014. godine, položio je sve ispite predviđene nastavnim planom i programom i time stekao uslov za odbranu master rada. Od februara 2014. godine radi u firmi Pheno Geno Roses iz Novog Sada na poslovima vezanim za rad, razvoj i kreiranje baze podataka.

Novi Sad, decembar 2015. godine

Igor Stranjanac

UNIVERZITET U NOVOM SADU
PRIRODNO - MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: Monografska dokumentacija

TD

Tip zapisa: Tekstualni štampani materijal

TZ

Vrsta rada: Master rad

VR

Autor: Igor Stranjanac

AU

Mentor: prof. dr Zagorka Lozanov-Crvenković

MN

Naslov rada: Regresione metode u QTL analizi diploidnih i tetraploidnih ruža

NR

Jezik publikacije: Srpski (latinica)

JP

Jezik izvoda: s / e

JI

Zemlja publikovanja: Republika Srbija

ZP

Uže geografsko područje: Vojvodina

UGP

Godina: 2015.

GO

Izdavač: Autorski reprint

IZ

Mesto i adresa: Departman za matematiku i informatiku, Prirodno-matematički fakultet,
Univerzitet u Novom Sadu, Trg Dositeja Obradovića 4, Novi Sad,

MA

Fizički opis rada: (4/78/18/4/4/4/0)

FO

Naučna oblast: Matematika

NO

Naučna disciplina: Statistika

ND

Ključne reči: statistika, regresija, analiza varijanse

PO

UDK

Čuva se: Biblioteka Departmana za matematiku i informatiku Prirodno-matematičkog
fakulteta Univerziteta u Novom Sadu

ČU

Važna napomena:

VN

Izvod: Ovaj rad se bavi regresionim modelima u QTL analizi. U prvom poglavlju rada definisani su osnovni pojmovi iz oblasti genetičke analize, kao što su DNK, molekularni markeri, fenotip, genotip. Obrađeni su i pojmovi iz oblasti statistike, kao što su linearni modeli i metoda maksimalne verodostojnosti. Objasnjeni su obrasci nasleđivanja diploidnih i poliploidnih populacija, sa kratkim osvrtom na oplemenjivanje, morfologiju

ruža i primenu molekularnih markera u oplemenjivanju ruža.

U drugom poglavlju izvedena i predstavljena su tri modela koji se koriste prilikom QTL analize: model koji koristi pojedinačne markere i intervalno mapiranje.

U trećem poglavlju predstavljena je implementacija modela iz prethodnog poglavlja na tetraploidnu i diploidnu populaciju. Rezultati su obrađeni i grafički predstavljeni pomoću programskog paketa R. Rezultati su analizirani i primenjen je model interakcije na tetraploidnu populaciju.

Četvrtog poglavlje predstavlja sumiranje rezultata čitave analize. U zaključku su dati predlozi značajnosti za dalji razvoj i modifikaciju modela.

IZ

Datum prihvatanja teme od strane NN veća: 20.01.2015.

DP

Datum odbrane: Decembar 2015.

DO

Članovi komisije:

KO

Predsednik: dr Nataša Krejić, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Mentor: dr Zagorka Lozanov-Crvenković, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Ivana Štajner-Papuga, vanredni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE KEY
WORDS DOCUMENTATION

Accession number:

ANO

Identification umber:

INO

Document type: Monograph type

DT

Type of record: Textual printed material

TR

Contents Code: Master thesis

CC

Author: Igor Stranjanac

AU

Mentor: Zagorka Lozanov-Crvenković, Ph.D.

MN

Title: Regression methods in QTL analysis of diploid and tetraploid roses

TI

Language of text: Serbian (Latin)

LT

Language of abstract: en / s

LA

Country of publication: Republic of Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2015.

PY

Publisher: Author's reprint

PU

Publ. place: Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, Novi Sad

PP

Physical description: (4/78/18/4/4/4/0)

PD

Scientific field: Mathematics

SF

Scientific discipline: Statistics

SD

Subject / Key words: statistics, regression, analysis of variance

SKW

UC

Holding data: Library of the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad

HD

Note:

N

Abstract: This thesis explores regression models in QTL analysis. The first chapter provides the basic concepts and terms which are used in the area of genetics, such as DNA, molecular markers, phenotype, genotype. Also, basic statistic terms are defined in this part, such as linear models and maximum likelihood. Patterns of inheritance of diploid and polyploid populations were explained, as well as rose breeding, rose morphology

and application of molecular markers in rose breeding. In the second chapter three models are provided, which are mostly used in QTL analysis: single marker analysis and interval mapping. Implementation of three models from the previous chapter on a diploid and tetraploid population is given in the third chapter. In this chapter the data is processed and results are interpreted and graphicly presented by using R software. The interaction model is applied to the tetraploid population and the results are briefly analyzed. The fourth chapter provides a conclusion of the analysis. Suggestions for further development and modifications are given in the conclusion.

AB

Accepted by the Scientific Board on: 20.01.2015.

ASB

Defended: December 2015.

DE

Thesis defend board:

President: Nataša Krejić, Ph.D., full professor, Faculty of Sciences, University of Novi Sad

Mentor: Zagorka Lozanov-Crvenković, Ph.D., full professor, Faculty of Sciences, University of Novi Sad

Member: Ivana Štajner-Papuga, Ph.D., associate professor, Faculty of Sciences, University of Novi Sad