Procjena efektivne veličine populacije iz vezanih i nevezanih gena

Došen, Valentina

Sveučilište u Zagrebu
Agronomski fakultet
University of Zagreb
Faculty of Agriculture

PROCJENA EFEKTIVNE VELIČINE POPULACIJE IZ VEZANIH I NEVEZANIH GENA

DIPLOMSKI RAD

Valentina Došen

Zagreb, rujan, 2019.

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

Diplomski studij:

Genetika i oplemenjivanje životinja

PROCJENA EFEKTIVNE VELIČINE POPULACIJE IZ VEZANIH I NEVEZANIH GENA

DIPLOMSKI RAD

Valentina Došen

Mentor: doc. dr. sc. Maja Ferenčaković

Neposredni voditelj: dr. sc. Vladimir Brajković

Zagreb, rujan, 2019.

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

IZJAVA STUDENTA

O AKADEMSKOJ ČESTITOSTI

Ja, **Valentina Došen**, JMBAG 0178093181, rođen/a 19.2.1994. u Zagrebu, izjavljujem da sam samostalno izradila diplomski rad pod naslovom:

PROCJENA EFEKTIVNE VELIČINE POPULACIJE IZ VEZANIH I NEVEZANIH GENA

Svojim potpisom jamčim:

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

U Zagrebu, dana _____

_____
Potpis studenta / studentice

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

IZVJEŠĆE

O OCJENI I OBRANI DIPLOMSKOG RADA

Diplomski rad studenta/ice **Valentina Došen**, JMBAG 0178093181, naslova

PROCJENA EFEKTIVNE VELIČINE POPULACIJE IZ VEZANIH I NEVEZANIH GENA

obranjen je i ocijenjen ocjenom _____ , dana _____ .

Povjerenstvo:			potpisi:
1.	Doc.dr.sc. Maja Ferenčaković	mentor	_____
2.	Dr. sc. Vladimir Brajković	neposredni voditelj	_____
3.	Prof. dr. sc. Ino Čurik	član	_____
4.	Izv. prof. dr. sc. Vlatka Čubrić Čurik	član	_____

Zahvala

Veliku zahvalnost, u prvom redu, dugujem svojoj mentorici doc. dr. sc. Maji Ferenčaković i neposrednom voditelju dr. sc. Vladimiru Brajkoviću koji su mi omogućili svu potrebnu literaturu te pomogli svojim savjetima pri izradi ovog diplomskog rada. Također, zahvaljujem se svojim prijateljima i prijateljicama, koji su uvijek bili uz mene i bez kojih cijeli ovaj tijek mog studiranja ne bi prošao tako lako i zabavno. Posebnu zahvalnost iskazujem cijeloj svojoj obitelji koja me je uvijek podržavala i upućivala na pravi put.

Velika HVALA svima!

Sadržaj

Sažetak

Diplomskog rada studenta/ice **Valentina Došen**, naslova

PROCJENA EFEKTIVNE VELIČINE POPULACIJE IZ VEZANIH I NEVEZANIH GENA

Efektivna veličina populacije (*Ne*) predstavlja ključni parametar u populacijskoj genetici koji opisuje količinu genetskog drifta u populaciji. Drugim riječima, *Ne* je definirana kao veličina hipotetski idealne populacije (*N*) koja bi doživjela istu količinu genetske promjene, tj. isti iznos genetskog drifta, kao realna populacija koju razmatramo. Procjena *Ne* putem metode disekvilibrijuma vezanih gena (eng. "*linkage disequilibrium*" - LD) sve se češće koristi u procjeni parametara populacije. Međutim, ova procjena, zbog specifičnosti formule, uključuje i markere koji nisu vezani, već se samo nalaze u korelaciji (eng. "*gametic disequilibrium*" - GD). Procjena *Ne* putem metode disekvilibrijuma vezanih markera (*NeLD*) često se koristi u populacijskoj genetici ljudi i domaćih životinja. Suprotno, procjena *Ne* iz informacije disekvilibrijuma nevezanih markera (*NeGD*) često se koristi u populacijskoj genetici divljih životinja.Ovim radom nastojat će se odvojiti GD i LD i na taj način usporediti procjene *Ne* s i bez GD-a.

Cilj ovog rada je procijeniti *Ne* programskim paketima NeEstimator i SNeP. Procjene će se potom usporediti budući da ovi programski paketi djeluju na različitim algoritmima. U radu se koriste podaci šest populacija ovaca s ukupnim brojem od 354 jedinke.

Ključne riječi: efektivna veličina populacije (*Ne*), genetski drift, LD („linkage disequilibrium“), GD („gametic disequilibrium“), NeEstimator, SNeP

Summary

Of the master's thesis – student **Valentina Došen,** entitled

ESTIMATION OF EFFECTIVE POPULATION'S SIZE BY LINKED UND UNLINKED GENES

Effective population size (*Ne*) is a key population genetic parametar that describes the amount of genetic drift in a population. In other words, Ne is defined as the size of a hypothetically ideal population (*N*) that would experience the same amount of genetic change (genetic drift) as the real population we are considering. Estimation of effective population size (*Ne*) by the linkage disequilibrium (LD) method is increasingly used to estimation of population's parameters. However, due to the specificity of the formula, this assessment also includes markers that are not linked but only correlated (gametic disequilibrium – GD). Estimation of *N*e by disequilibrium linked marker (*NeLD*) method is often used in population genetics of humans and domestic animals. On the contrary, the estimation of *Ne* from disequiliubrium unlinked markers information (*NeGD*) is often used in genetics of wild animals. This paper will seek to separate GD and LD and thus compare the estimates *Ne* with and without GD.

The aim of this paper is to estimate *Ne* by programs NeEstimator and SNeP. The estimates will then be compared because programs run on different algorithms. The paper uses data from six sheep populations with a total of 354 individuals.

Key words: effective population size (*Ne*), genetic drift, linkage disequilibrium (LD), gametic disequilibrium (GD), NeEstimator, SNeP

.

1. Uvod

Domestikacija domaćih životinja označila je bitan korak u čovjekovom demografskom i kulturnom razvoju. Evolucijske sile, poput mutacije, selekcijskog uzgoja, adaptacije i genetskog drifta prouzročile su stvaranje velikog spektra različitih pasmina. Iznimno je važno očuvati varijabilnost svih pasmina životinja jer je upravo biološka raznolikost ključ njihovog održavanja i opstanka. Poželjna svojstva mogu se selekcijom poboljšati samo ako postoji genetska varijabilnost u populaciji. Genetski napredak za ta svojstva ovisi o mnogim čimbenicima, uključujući genetske korelacije, intenzitet selekcije, koeficijent srodstva, mutacije i genetski drift (Groeneveld, 2010.).

Nekoliko je vrlo važnih čimbenika koji određuju brzinu evolucijskih procesa, a veličina populacije je svakako jedna od njih. Samo znanje o ukupnom broju jedinki (*N*) u populaciji ipak nije dovoljno za točno poznavanje i razumijevanje tih evolucijskih procesa. Ako u određenom slučaju imamo dvije populacije jednakih veličina, one mogu imati vrlo različite stope genetske promjene. Tako je Wright (1931.) razvio koncept efektivne veličine populacije (*Ne*) kao način sažimanja demografskih podataka putem kojih je moguće predvidjeti evolucijske posljedice konačne veličine populacije. Dakle, izračunom *Ne* moguće je izračunati stopu gubitka genetske varijabilnosti.

Gotovo sav uspjeh u očuvanju vrsta brojnih populacija danas možemo zahvaliti napretku molekularne genetike, ali i paralelnom razvoju nekih multidisciplinarnih grana poput ekološke genetike te populacijske i konzervacijske genetike. Molekularna ekologija putem molekularne genetike bavi se ekologijom i divljim životinjama. Konzervacijska genetika bavi se čimbenicima koji utječu na izumiranje vrsta, a populacijska genetika odnosi se na genetsku osnove evolucije te istražuje frekvenciju i preživljavanje genotipova prirodnih populacija. Sve ove grane genetike pomažu shvaćanju strukture populacija, određivanja filogenetskih odnosa, rješavanju taksonomskih pitanja te doprinose planiranju zaštite ugroženih vrsta, a sve to sa ciljem očuvanja genske raznolikosti.

Genska raznolikost je, po definiciji, razlika u slijedu nukleotida u molekuli DNA. Te razlike rezultiraju različitim slijedom aminokiselina u proteinima koje ti geni kodiraju, a kasnije ta varijacija u slijedu proizvodi morfološke nejednakosti i utječe na preživljavanje jedinki u populaciji. Ako na lokusu u skupini koju istražujemo postoji više od jednog alela onda se lokus smatra polimorfnim, a ako su sve jedinke u skupini homozigoti onda se lokus smatra fiksiranim (Mank i sur., 2009.). Genska raznolikost nije samo ovisna o broju mutacija, nego i o *Ne*, odnosno broju jedinki koje se razmnožavaju u populaciji. Prema tome, populacije

koje imaju malu *Ne* obično imaju nisku genetsku raznolikost. Stoga, genska raznolikost čini snažan evolucijski potencijal vrste i usko je povezana s obilježjima vezanim za sposobnost preživljavanja, kao što su rast, razvoj, otpornost na bolesti i plodnost.

Pad genske raznolikosti najčešće pratimo kod malih i izoliranih populacija. Upravo u slučaju inbridinga, odnosno parenja u srodstvu, populacije gube brojnost. Većina ugroženih populacija ima nižu gensku raznolikost od populacija koje nisu ugrožene.

Koncept *Ne* uveo je Sewall Wright već tridesetih godina. Efektivna veličina populacije predstavlja broj jedinki koji doprinose svojim gametama sljedećoj generaciji ili jednostavnije rečeno, to je upravo broj jedinki koje sudjeluju u stvaranju potomstva (Charlesworth, 2009.).

Iz navedenog slijedi logičan zaključak da *Ne* zbilja predstavlja osnovni parametar u populacijskoj genetici koji je bitan za formiranje adekvatnog programa očuvanja ugroženih vrsta. U populacijskoj genetici za procjenu parametara populacije sve se češće koristi procjena *Ne* metodom disekvilibrijuma vezanih gena („*linkage disequilibrium*“, tj. LD). Metoda za procjenu *Ne* putem LD-a razvijena je otprilike prije 40 godina, ali u velikoj mjeri ovisi o dostupnosti velike količine podataka o genetskim markerima (Barbato, 2015.). Problem je što ta procjena, zbog specifičnosti formule, uključuje i markere koji nisu vezani, nego se samo nalaze u korelaciji, a to onda nazivamo „*gametic disequilibrium*“, tj. GD.

1.1. Cilj rada

Cilj ovog diplomskog rada je usporediti procjene efektivne veličine populacije putem dva programska paketa, a to su SNeP i NeEstimator. SNeP vrši procjenu *Ne* putem LD-a dok NeEstimator to čini putem GD-a. Time će se usporediti trenutna efektivna veličina (*NeGD*) s povijesnom efektivnom veličinom (*NeLD*).

2. Razrada literature

2.1. „Linkage“ i „gametic“ disekvilibrijum

„Linkage disequilibrium“, odnosno neravnoteža vezanih gena definira se kao neslučajna povezanost između alela na različim lokusima u zadanoj populaciji (Waples, 2010.). Može se, u teoriji, proizvesti pomoću nekoliko faktora koji djeluju odvojeno ili zajedno, a to su epistatska selekcija, migracija ili slučajni drift u konačnoj populaciji. Mnoge procjene neravnoteže veze odvijale su se, kako na laboratorijskim, tako i na prirodnim populacijama. Dodatni podatci o neravnoteži veze vjerojatno će doći izravnim istraživanjima DNK, zatim s mjesta restrikcijskih enzima ili direktno iz DNK sekvence (Hill, 1981.).

LD predstavlja temeljnnu ulogu u mapiranju gena, identificiranju kromosomskih regija koje su pod selekcijom te upravljanju genetskim resursima i raznolikošću. Također je zanimljiv zbog činjenice da može puno toga otkriti o povijesti i podrijetlu populacije jer je raspodjela LD-a dijelom određena poviješću populacije. Definira se kao neslučajna povezanost alela na različitim lokusima. Ti se lokusi uvijek zajedno nasljeđuju, a budući da se nalaze međusobno jako blizu „crossing over“ se na takvim mjestima ne događa.

Iako se očekuje da će markeri u LD-u biti čvrsto povezani, precizan opseg zajedničkih haplotipova ovisi o stohastičkom procesu koji uključuje mnoštvo faktora uključujući migraciju, selekciju, mutaciju i genomske obrasce rekombinacije (Pritchard, 2001.). Očekuje se da će LD varirati, ne samo između populacije, već i između lokusa u istoj populaciji. Stupanj LD-a na određenim lokusima i populaciji diktira odabir SNP markera koji će biti genotipizirani te veličinu uzorka potrebnog za analizu.

Dok su kod LD-a aleli na različitim lokusima povezani isključivo zbog same veze (eng. „linkage), kod gametskog disekvilibrijuma (GD) aleli mogu biti povezani kao posljedica selekcije, genetskog drifta, mutacije i inbridinga (predavanje iz Populacijske genetike, 2017., prof. dr. sc. I. Čurik). Selekcija, kao faktor evolucije, omogućuje da određene jedinke imaju tedenciju da budu reprodukcijski uspješnije, što znači da ostavljaju više potomaka i svojih gena u narednim generacijama. Naravno, dugotrajna pozitivna selekcija jednog alela u konačnici vodi k smanjenju učestalosti ostalih alela u populaciji. Gubitak alela ili njegova fiksacija u populaciji čistom slučajnosšću i nevezano sa selekcijom nazivamo genetski drift. Najvažniji faktor za drift je veličina populacije; što je populacija manja veći je potencijal za genetski drift. Nadalje, mutaciju definiramo kao iznenadnu nasljednu promjenu genetskog materijala. Sve nasljedne varijabilnosti kod živih bića, i dobre i loše, imaju izvor u mutacijama genetskog materijala. Zatim, inbriding najjednostavnije definiramo kao parenje u

srodstvu. Stoga, kao rezultat parenja povezanih i, posljedično, sličnih životinja, njihovi potomci akumuliraju alele i genotipove tog pretka koji je zajednički srodnim pojedincima. Inbreeding dovodi do povećanja učestalosti homozigotnih genotipova u potomcima i smanjenja učestalosti heterozigotnih genotipova uz zadržavanje frekvencija alela. Srodno parenje prati smanjenje genetske varijabilnosti

Dakle, LD će pružiti informacije iz disekvilibrijuma vezanih markera, dok će GD to učiniti iz disekvilibrijuma nevezanih markera te se kod takvih markera ne govori o vezi, nego o koleraciji.

.

2.2. Važnost efektivne veličine populacije

Efektivnu veličinu populacije, prema Wright-Fischeru, definiramo kao veličinu hipotetski idealne populacije (*N*) koja bi doživjela istu količinu genetske promjene, tj. isti iznos genetskog drifta, kao realna populacija koju razmatramo (predavanje iz Konzervacijske genetike, 2017., prof. dr. sc. I. Čurik). Idealna populacija zasnovana je na nerealnim uvjetima; parenje je slučajno, jednak je omjer muških i ženskih jedinki, konstantan je broj jedinki u obitelji, kao što je i konstantan broj potomaka. Svrha koncepta *Ne* je izračunati stopu promjene evolucijskih procesa uzrokovanu slučajnim uzorkovanjem frekvencije alela u konačnoj populaciji (genetski drift).

Ne također može ukazati na to koliko je neka populacija „ranjiva“ te kako neravnomjeran broj ženskih i muških jedinki može mijenjati varijabilnost genetskog drifta i inbridinga. Stoga, mala vrijednost *Ne* može imati velike posljedice za samu populaciju. Smanjenje *Ne* upućuje na povećanje inbridinga, pojavu genetskog drifta i smanjenje genetske varijabilnosti u populaciji. Povećanje inbridinga za posljedicu će imati promjenu u frekvenciji genotipova, odnosno povećanje frekvencije homozigotnih genotipova uz istovremeno smanjenje heterozigota. U velikoj populaciji u kojoj pratimo slučajno sparivanje jedinki genetski drift neće imati značajan utjecaj na populaciju, ali ako je *Ne* mali drift će snažno utjecati na frekvenciju alela. U tom slučaju, genetski drift može dovesti do fiksacije štetnog alela.

Stopa mutacije i *Ne* određuju razinu getekske varijabilnosti u neutralnoj ili slabo selektiranoj populaciji. Nadalje, učinkovitost eliminacije štetne mutacije ili pak širenja povoljne mutacije kontrolirana je vrjednošću *Ne* i intenzitetom selekcije. Vrijednost *Ne* uvelike je pod utjecajem varijabilnosti DNA sekvence, kao i stopom evolucije DNK sekvence.

Važnost *Ne*, kao evolucijskog čimbenika, istaknuta je saznanjima da je vrijednost *Ne* često znatno niža od cenzusnog broja određene populcije. Vrste s povijesno niskom vrijednosti *Ne*, poput ljudi, dokazuju smanjenje varijabilnosti i učinkovitosti selekcije u usporedbi s drugim vrstama. *Ne* također može varirati kroz različita mjesta na genomu, a razlog je vrlo vjerojatno razlika u načinu prijenosa različitih komponenti genoma (na primjer, X kromosomi s jedne te autosomi s druge strane). Također, takvo variranje se može pojaviti kao posljedica učinka selekcije na nekom mjestu na genomu. Važna posljedica selekcije je smanjenje *Ne* u genomskim regijama s niskom razinom genetske rekombinacije, s učincima koji su vidljivi na razini molekularne sekvence.

2.1.1. Čimbenici koji utječu na efektivnu veličinu populacije

- podjela na dva spola: mali broj jedinki jednog spola može snatno smanjiti *Ne*
- varijacija u broju potomaka: veća odstupanja u broju potomaka od očekivanoga
- inbreeding: korelacija između majčinih i očevih alela uzrokovana zajedničkim pretkom
- način nasljeđivanja: *Ne* ovisi o načinu na koji se događa prijenos alela, odnosno radi li se o autosomalnom, X-vezanom, Y-vezanom prijenosu ili preko organela
- dobna struktura: u dobno strukturiranoj populaciji, *Ne* je znatno niži od cenzusnog broja (*N*)
- promjene u veličini populacije: periodi s malom veličinom populacije, odnosno „bottle-neck“ imaju nesrazmjerni učinak na ukupnu vrijednost *Ne*
- prostorna struktura: *Ne* koja određuje srednju vrijednost neutralne varijabilnosti unutar lokalne populacije je često neovisna o detaljima migracijskog procesa koji povezuje populacije. Ograničena migracija između populacija znatno povećava *Ne* za cijelu populaciju, gdje visoka razina lokalnog izumiranja ima suprotan učinak
- genetska struktura: direktna selekcija uzrokuje smanjenje *Ne* na povezanim alelima, dok uravnotežena selekcija povećava *Ne* na mjestima gdje su aleli usko povezani (Hill-Robertsonov učinak)

2.3. Određivanje efektivne veličine populacije

Tri su glavna načina prema kojima genetski drift može biti modeliran u populaciji. Ovi teorijski modeli vode do općeg pristupa koji se može primjeniti u situacijama većeg biološkog interesa, što pak otkriva korisnost koncepta *Ne*.

Wright-Fischer populacija

Kako bismo shvatili važnost računanja *Ne*, potrebno je razumjeti kako genetski drift može biti modeliran u jednostavnim slučajevima Wright-Fisher populacije. To je populacija u kojoj je parenje slučajno, jednak je broj muških i ženskih jedinki, konstantan je broj jedinki u obitelji te svi roditelji moraju imati jednaku vjerojatnost da budu roditelji. Ukoliko je *Ne* razmjerno velik onda to podrazumijeva da je veličina obitelji definirana Poisson distribucijom. Poissonova distribucija je raspodjela vrlo rijetkih slučajnih događaja (kod kojih je vjerojatnost pojavljivanja vrlo mala), a definirana je aritmetičkom sredinom jer je njena varijanca jednaka aritmetičkoj sredini. Kada je *N* vrlo velik, Poissonova distribucija je približava binomnoj, no razlika je u tome što kod binomne raspodjele znamo koliko se puta neki događaj pojavio, ali i koliko se puta nije pojavio, dok kod Poissonove raspodjele znamo samo koliko se puta neki događaj pojavio (http://www.unizd.hr/Portals/13/NASTAVNI_MATERIJALI/04%20-%20Distribucije.pdf). Populacije morskih organizama, koje polažu veliki broj jajašaca i sperme i nasumično tvore nove zigote, najsličnije su idealnoj populaciji (Charlesworth, 2009.).

Alternativni pristup, koji ima središnju ulogu u suvremenoj interpretaciji varijacije DNK objašnjen je teorijom koalescentnog procesa (eng. „*coalescent teory*“). Koalescentna teorija definirana je kao metoda rekonstrukcije povijesti uzorka alela iz populacije prateći njihovo rodoslovlje do najnovijeg zajedničkog pretka. Umjesto da se pogledaju svojstva stanovništva u cjelini, razmatraju se određeni aleli na genskom lokusu koji su uzorkovani iz neke populacije. Ukoliko pratimo njigove pretke u prošlosti, pokazat će se da potječu od istog ancestralnog alela, odnosno podvrgnuti su koalescenciji (Charlesworth, 2009.).

Realniji modeli genetskog drifta

Pretpostavke Wright-Fisher populacije ne vrijede kod većine populacija od biološkog interesa: gotovo sve vrste imaju dva spola, mogu postojati varijacije u reproduktivnom uspjehu, parenje nije slučajno, generacije se preklapaju umjesto da su diskretne, veličina populacije varira kroz vrijeme te vrste mogu biti podijeljene na lokalnu populaciju ili na diskretne genotipove. Nadalje, potrebno je analizirati učinke evolucijskih sila, poput selekcije i rekombinacije, ali i drifta.

Ne opisuje vremenski okvir genetskog drifta u ovim složenijim situacijama: zamijenimo 2*N* s 2*Ne*, gdje je *Ne* dobiven iz formule koja uzima u obzir relevantne biološke detalje. To se dobije računanjem na temelju varijance ili koeficijenta inbreedinga, no odnedavno se

primjenjuje koalescentna teorija. Općenito, upotreba *Ne* samo daje aproksimaciju stope genetskog drifta za znatno veliku veličinu populacije. Točni proračuni promjene varijance ili frekvencije alela ili koeficijenta inbreedinga su često potrebni u aplikacijama u kojima je veličina populacije jako mala ili je vremenaki okvir kratak.

Određivanje efektivne veličine populacije: opća metoda

Koalescentna teorija pruža fleksibilnu i moćnu metodu za dobivanje formule za *Ne*. Ovaj pristup uključuje strukturirani koalescentni proces u kojem je nekoliko odjeljaka (poput dobi ili spola) u populaciji iz kojih aleli mogu biti uzorkovani. Aleli su prvo uzorkovani iz jenog ili više odjeljka i vjerojatnost pomicanja alela u druge odjeljke određena je pravilima nasljeđivanja. Pretpostavka je da aleli među odjeljcima protječu puno brže nego koalescencija alela, a takva pojava se naziva brza aproksimacija vremenske skale (Charlesworth, 2009.). To znači da možemo uzorkovane alele tretirati kao da su iz stanja ekvilibrijuma. To daje opću formulu za brzinu koalescencije koju je lako primjeniti na individualne slučajeve.

2.4. Genetske metode izračuna efektivne veličine populacije

Poznavanje *Ne* posebno je važno kada pokušavamo sačuvati malu populaciju neke vrste. Kada je *Ne* mali, očuvana populacija može brzo izgubiti genetsku varijabilnost. *Ne* nema fiksnu vezu s veličinom populacije te je precizna procjena dosta složena. Dva su načina kako možemo pristupiti procjeni *Ne*: genetski i ekološki. Genetske metode izravno ukazuju na genetske posljedice *Ne*, dok su ekološke metode neizravne i ovise o mjerenju ekoloških parametara za koje je dokazano da teoretski utječu na *Ne*.

Korištenje genetskih metoda na prvi će se pogled pokazati kao najbolji način procjene *Ne*, no postoje problemi koje je potrebno prevladati kako bi taj način bio praktičan. Jedan od njih je potreba za dobivanjem velikih količina genetskih informacija. Ovaj problem danas nije toliko ograničavajuć zbog tehnološkog napretka. Također, problem predstavlja potreba da se uklone zbunjujući efekti imigracije i podjele stanovništva te mogućnost da odabir djeluje na markerima ili lokusima povezanim s markerima.

Dugoročne genetske metode uključuju praćenje genetske promjene populacija tijekom vremena. Jedna takva metoda uključuje mjerenje genetske divergencije replicirane populacije. Metoda se oslanja na očekivanje da se, u odsutnosti selekcije, heterozigotnost s vremenom smanjuje jer se replike razilaze u frekvenciji gena zbog inbreedinga. S obzirom na oslanjanje na replicirajuću populaciju, ova metoda se može primjeniti samo u prirodnim populacijama pod posebnim okolnostima.

Druga metoda, koja je dosta općenitija, je „temporalna“ metoda. Ona uključuje praćenje promjena frekvencije gena unutar jedne populacije mjerenjem frekvencije više polimorfnih markera u početnom uzorku, a zatim ponavljanjem vježbe jednu ili više generacija kasnije. Takvi podaci omogućuju procjenu veličine populacije koja utječe na varijancu. Praćenje

određenog broja povezanih grupa također može pružiti informacije o tome ponaša li se bilo koja regija genoma atipično zbog jake selekcije. Nedostatak metode leži u tome što postaje učinkovita tek nakon što prođe niz generacija, ali ovo je samo prolazna poteškoća dok se ne skupi dovoljno povijesnih podataka.

Kratkoročne genetske metode koriste podatke iz jednog uzorka. Jedna takva metoda bazira se na uspoređivanju letalnih alela između i unutar populacije, ali budući da zahtijeva procjenu smrtonosnih frekvencija, metoda je općenito nepraktična. Predložen je srodni pristup korištenjem analize kromosomskih prepravki. Ova metoda je bila korisna u procjeni *Ne*.

Druga vrsta kratkotrajne genetske metode temelji se na odnosu između LD-a i *Ne*. Međutim, upotreba povezanih lokusa stvara praktični problem procjene odnosa veze i interpretacija rezultata je složena jer povijesni događaji mogu imati veliki utjecaj na rezultate. Ovaj pristup ima brojne prednosti pod uvjetom da se pretpostavke mogu potvrditi. Te pretpostavke uključuju saznanje da su korišteni lokusi nepovezani i potvrđuju da populacija nije podijeljena niti je podložna značajnoj imigraciji (Nunney, 1994.).

2.5. Efektivna veličina populacije i „linkage disequilibrium“

Neravnoteža veze (LD) može se koristiti za procjenu vrijednosti Ne ukoliko je poznata brzina rekombinacije. Korištenje LD-a ima prednost u tome što je stopa rekombinacije više kontrolirana od mutacijske stope, a novije vrijednosti *Ne* mogu se procijeniti jer rekombinacijska stopa može biti veća od mutacijske stope (Hayes, 2003).

Jačina disekvilibrijuma (D) između alela na dva gen lokusa je definirana kao razlika između promatrane frekvencije dviju gameta i njezine očekivane frekvencije, temeljene na frekvenciji populacijskih alela. D se može procijeniti izravno iz gametskih frekvencija. Za većinu nemodelskih vrsta, međutim, dostupni su samo genotipski podaci pa se ne mogu sa sigurnošću rekonstruirati gametske frekvencije zbog dvosmislenosti koja je povezana s dvostrukim heterozigotima. U tom se slučaju za procjenu parametra D najčešće koristi Burrowsova kompozitna delta metoda (Δ), koja je jednostavna za izračunavanje i ne ovisi o pretpostavci o slučajnom parenju. Budući da su D i Δ osjetljivi na frekvenciju alela, često se koristi standardizirani oblik nejednakosti veze (r) koji se može protumačiti kao koeficijent korelacije alela na različitim genima. I D i r mogu biti pozitivni ili negativni pa se kvadratni izrazi D^2 i r^2 često upotrebljavaju kada nas zanima razmjernost neravnoteže veze.

Pretpostavka LD metode je da je razmjer slučajne asocijacije alela kod različitih lokusa gena određen s tri varijable, a to su *Ne*, broj uzorkovanih jedinki (S) i stopa rekombinacije između lokusa (c). Za većinu prirodnih populacija rekombinacijska frakcija neće biti poznata.

Međutim, ukoliko broj markera nije veliki ili je broj kromosoma mali, bilo bi razumno pretpostaviti da su lokusi nepovezani (c=0.5). Mnoge prirodne populacije su se migracijama povezale s drugim populacijama. U bilo kojem trenutku, populacija od interesa može sadržavati jedinke proizašle više od jednog „*gene pool*"-a (Waples, 2011.). Takva mješavina stvara dobro poznati Wahlundov efekt koji se očituje kao nedostatak heterozigota u usporedbi s očekivanom frekvencijom jednog lokusa u Hardy-Weinbergu. Smjese također stvaraju svojevrsni dvostruki Wahlund efekt koji je prepoznatljiv kao neravnoteža veze (Wang, 2005.).

2.6. Korekcija pristranosti za procjenu efektivne veličine populacije putem metode „linkage disequilibrium"-a

Iako je *Ne* od središnje važnosti za evolucijsku i konzervacijsku biologiju, pokazalo se nedostižnim dobiti pouzdane procjene ovog ključnog parametra. Znanstvenici još uvijek vode raspravu o tome je li omjer *Ne* naspram cenzusne veličine populacije ograničen u određenim slučajevima ili granice mogu biti manje za neke vrste. Zbog poteškoća u dobivanju odgovarajućih demografskim podataka za izračunavanje *Ne*, razvijeno je nekoliko genetskih metoda za izračun *Ne*. Metode koje će biti spomenute u ovome radu su temporalna metoda i metoda disekvilibrijuma vezanih gena. Temporalna metoda koristi podatke o stopi promjene frekvencije alela između uzoraka uzetih u različito vrijeme.

Metoda putem LD-a ,koji predtavlja neslučajnu povezanost alela u različitim genskim lokusima, znatno rjeđe se koristi u istraživanjima. Prednost ove metode je u tome što zahtijeva samo jedan uzorak iz populacije, dok temporalna metoda zahtijeva najmanje dva uzorka iz populacije. Hill (1981) je pokazao da ova metoda ima malu preciznost ukoliko se ne koriste čvrsto povezani lokusi i zaključio da metoda ima ograničenu korisnost za procjenu *Ne*. S druge strane, Waples (1991.) je zaključio da metoda ima veću korist kada je Ne mala te stoga može biti od iznimnog značaja za evolucijsku i konzervacijsku biologiju koje se upravo bave malim vrijednostima *Ne*. Waples je također sugerirao da, ukoliko postoje, podaci za određeni broj nepovezanih lokusa mogu pružiti odgovarajuću preciznost kako bi metoda bila korisna.

2.7. Procjena „linkage disequilibrium"-a i efektivne veličine populacije

Neravnoteža veze (D) između alela na dva gen lokusa definirana je kao razlika između promatrane frekvencije gameta na dva lokusa i njezine očekivane frekvencije temeljene na slučajnoj povezanosti i frekvenciji alela u populaciji. D se može procijeniti izravno iz gametskih frekvencija, međutim, za većinu prirodnih populacija dostupni su samo genotipski

podaci, što znači da se frekvencije gameta ne mogu sa sigurnošću rekonstruirati zbog dvosmislenosti glede gameta koje se ujedinjuju u dvostruke heterozigote. U tom slučaju, najčešće korištena metoda za procjenu disekvilibrijuma jest Burrowsov Δ jer je jednostavan za izračun i ne ovisi o slučajnom parenju.

3. Programski paketi za izračun efektivne veličine populacije

3.1. NeEstimator

Tijekom posljednjeg desetljeća iznimno raste zanimanje za korištenjem genetskih metoda u procjeni *Ne*. Prvenstveno, to možemo zahvaliti napretku u razvoju molekularnih markera i metoda ekstrakcije DNA iz prirodnih populacija. Donedavno se pri procjeni *Ne* koristila temporalna metoda, što zahtijeva najmanje dva uzorka iz iste populacije.

S obzirom na raznolikost dostupnih metoda, nastojalo se razviti softver koji na isti skup podataka može primjeniti više metoda. To se i postiglo prvotnom verzijom navedenog programskog paketa koji je nosio naziv NeEstimator v1.4. Međutim, softveru su prethodila najnovija dostignuća u metodama s jednim uzorkom, što je ograničavalo njegovo korištenje. Uslijedila je obnovljena verzija softvera pod nazivom NeEstimator v.2.0., a upravo će ta verzija biti opisana i korištena u radu.

NeEstimator je programski paket za procjenu suvremene *Ne* iz genetskih podataka. Pri procjeni ovaj programski paket koristi nekoliko različitih metoda i jednu ulaznu datoteku. Uključuje: a) metodu jednog uzorka (metoda temeljena na LD-u); b) metoda temeljena na heterozigotima („heterozygote excess“); c) metoda koja se temelji na molekularnoj razini i prati pretka („molecular coancestry“); d) metoda na dva uzorka (temporalna metoda).

Ta nova verzija ima fleksibilno grafičko sučelje i pogodno je za simulirane skupove podataka koji sadrže različiti broj genotipova s dva ili više lokusa te imaju dva ili više alela po lokusu. Genotipovi mogu predstavljati jednu do više populacija koje mogu biti uzorkovane jednom ili više puta. NeEstimator v.2.0 također ima verzije za operativne sustave Windows, MacOS i Linux.

Prednosti programskog paketa NeEstimator v.2.0 nad prijašnjom verzijom su sljedeće:

1) poboljšane su metode računanja nedostajućih vrijednosti
2) mogućnost uklanjanja rijetkih alela
3) intervali pouzdanosti za sve metode
4) mogućnost analize skupova podata s velikim brojem genetskih markera (10 000 i više)
5) mogućnost serijske obrade velikog broja različitih skupova podataka, a to će olakšati usporedbu unakrsnih metoda koristeći simulirane podatke
6) korekcija za procjene vremenske metode kada uzorkovane jedinke nisu maknute iz populacije.

Korisniku se, svakako, omogućuje značajna kontrola nad ulaznim podacima i sastavom, ali i formatom ispisa.

3.1.1. Unos podataka

NeEstimator v.2.0 prihvaća uobičajene ulazne formate, odnosno GENEPOP ili FSTAT. Korisnik odabire direktorij za odabir odgovarajuće ulazne datoteke. Datoteke su u prihvatljivim formatima, a to su .TXT, .GEN i .DAT. Jedna ili više metoda za izračun *Ne* mogu se izvesti istovremeno, a obično se izvode na jednoj datoteci za unos podataka. Postoji nekoliko opcija za obradu mnogih zasebnih datoteka. Ulazne datoteke mogu uključiti veliki broj uzoraka. Za metode jednog uzorka (LD, „*heterozygote excess*“ i „*molecular coancestry*“) svaki se uzorak tretira kao zasebna „populacija“. Za temporalne metode, svaka populacija je predstavljena s dva ili više uzorka uzorkovanih u različito vrijeme i odvojeni poznatim brojem generacija. Generacije za svaki uzorak mogu biti definirane kao cijeli ili frakcijski brojevi. U najjednostavnijim okolnostima, ulazna datoteka za temporalnu metodu sadržavala bi dva uzorka, odvojena brojem generacija koji definira korisnik. Temporalna metoda proizvela bi jednu procjenu *Ne* na broj generacija između uzoraka, dok bi metoda jednog uzorka proizvela odvojene procjene na svaku uzorkovanu generaciju. Također, mogu se primjeniti naprednije strategije uzorkovanja. Na primjer, korisnik može odrediti da je prva populacija uzorkovana iz generacije nula i dva, druga populacija iz generacije tri i pet, a da su preostale tri populacije uzorkovane iz generacije nula, četiri i pet. U ovakvom slučaju, datoteka ulaznih podataka sadržavala bi 13 ukupnih uzoraka, analiziranih kao pet zasebnih populacija.

Programski paket također omogućuje korisniku fleksibilnost u definiranju parametara za analize. Za sve metode, korisnik može odlučiti ukloniti rijetke alele s frekvencijom ispod kriterija kojeg određuje korisnik, a naziva se *Pcrit*. Nadalje, podskupine ulaznih podataka mogu se odabrati za analizu u grafičkom korisničkom sučelju. Na primjer, ako datoteka podataka sadrži deset populacija, programski paket se može usmjeriti samo na analizu prva dva. Korisnik, također, može ograničiti broj analiziranih jedinki (npr. na 10 ili 20) u svakom uzorku. Uz to, lokusi se mogu selektivno isključiti određivanjem raspona (npr. lokus 1-5 i 10-15) ili navođenjem lokusa jedinki (npr. lokus 2, 4, 6). Za LD metodu korisnik može birati između pretpostavki o slučajnom ili monogamnom parenju. Kada ulazna datoteka sadrži tisuće lokusa ili veliki broj jedinki po populaciji, metode LD-a i metode praćenja pretka mogu potrajati satima ili danima. Sučelje će otprilike procijeniti tu mogućnost i po potrebi će staviti dijaloški okvir s upozorenjem te korisnik tada može odlučiti hoće li nastaviti ili koristiti neke opcije dostupne na sučelju za ograničavanje podataka. Ako korisnik odluči pokrenuti, ekran terminala će ispisati napredak u određenim ciljevima pa korisnik može imati u vidu kada je pokrenuti postupak završen.

3.1.2. Datoteke izlaznih podataka

Jedan od potencijalnih nedostataka ovog programskog paketa koji nudi višestruke analize su velike datoteke izlaznih podataka. NeEstimator to prevladava tako što stvara jednostavnu zadanu izlaznu datoteku koja opisuje procijenjene parametre populacije za svaku odabranu metodu analize i korisniku daje izbor odabira dodatnih i detaljnih izlaznih datoteka. Na primjer, korisnik može odabrati da se rezultati svake metode ispisuju u zasebnoj datoteci koja je organizirana u pojednostavljenom formatu koji je lako analizirati i uvesti u drugi softver. Ostale opcije uključuju podatke o frekvenciji na svakom lokusu i rezultate za svaki par lokusa u LD metodi.

3.1.3. Intervali pouzdanosti

NeEstimator osigurava intervale pouzdanosti za sve metode i u nekoliko slučajeva implementira nove i poboljšane rutine. Potencijalne pristranosti povezane sa standardnim parametrijskim intervalima pouzdanosti za LD metodu smanjuju se primjenom metode „*jackknife*“. Omogućuje korisniku utvrđivanje relevantnosti jednog ili oba intervala za analize. Za metodu temeljenu na heterozigotima implementacija programskog paketa ispravlja pogrešku u intervalu pouzdanosti kod metode koju su predložili Zhdanova i Pudovkin (2008.). Nomura (2008.) nije predložio metodu za izgradnju intervala pouzdanosti za metodu temeljenu na pretku (eng. *molecular coancestry method*) te programski paket implementira novu metodu „*jackknife*“-a razvijenu posebno za tu svrhu. Važno upozorenje je da učinkovitost novih metoda za intervale pouzdanosti implementirane u NeEstimator nije ocijenjena. Konkretno, uporaba velikog broja (100 ili 1000) SNP lokusa, od kojih će mnogi biti neizbježno povezani, uvodi važna pitanja koja se odnose na pseudoreplikaciju. Preciznost procjena na temelju velikog broja lokusa može biti znatno manja nego što je to preporučeno tradicionalnim metodama za računanje intervala pouzdanosti.

3.1.4. Negativne ili beskonačne procjene efektivne veličine populacije

Sve razmatrane metode temelje se na genetskom indeksu koji ima dvije komponente: jedna zbog genetskog drifta, a jedna zbog uzorkovanja ograničenog broja jedinki. Nepristrani procjenitelji ovise o poznavanju veličine uzorka (tako da očekivana pogreška uzorkovanja može biti izračunata). Stvarna količina pogreške uzorkovanja može biti veća od očekivane pa je u tom slučaju moguće da procjena *Ne* bude negativna. Uobičajena interpretacija u ovom slučaju je da je procjena *Ne* beskonačna, tj. nema dokaza za varijacije u genetskim

karakteristikama prouzrokovane ograničenim brojem roditelja; sve se to može objasniti pogreškom uzorkovanja. Ekvivalentni fenomen može se javiti i kod nepristranim procjenitelja genetske udaljenosti ili FST-a.

U NeEstimatoru (v2.0) navode se negativne procjene i intervali pouzdanosti su izneseni kao „beskonačnost“ u glavnoj izlaznoj datoteci. Međutim, u dodatnim izlaznim datotekama stvarne negativne vrijednosti su iznesene jer negativne procjene *Ne* sadrže vrijedne informacije kada je uključena harmonijska srednja vrijednost kojom se dobije ukupna procjena *Ne* (npr. ako postoji nekoliko ponovljenih uzoraka iz iste populacije).

3.1.5. Rijetki aleli

NeEstimator (v2.0) pruža opcije za izbacivanje rijetkih alela za sve metode osim metode molekularnog pretka, eng. *molecular coancestry* (za koje frekvencija alela nije problem), koristeći iste protokole kao *LDNe*. Softver već prema zadanim postavkama provodi i izvještava rezultate za odvojene analize koje koriste sve alele ili koji izbacuju alele s frekvencijom ispod *Pcript* vrijednosti od 0.01, 0.02 i 0.05. Korisnik može promijeniti ove zadane postavke za implementaciju bilo koje željene vrijednosti *PCrit*. Korisnik također ima mogućnost izbora dodatne izlazne datoteke koja sadrži frekvencije alela za svaki lokus za svaku populaciju i izvještava o broju alela po lokusu koji su uklonjeni jer su bili ispod vrijednosti koju je korisnik odredio.

3.2. SNeP

Ne može se procijeniti koristeći tri metodološke kategorije, a to su demografska kategorija, kategorija temeljena na rodovniku i kategorija temeljena na markerima. Podaci o rodovniku tradicionalno se koriste pri procjeni *Ne* stoke. Međutim, pouzdane procjene *Ne* ovise o kompletnosti podataka u rodovniku. To je izvedivo kod nekih domaćih populacija, čiji su demografski parametri precizno promatrani za dovoljno velik broj generacija. Ipak, u praksi, primjena ovog pristupa ostaje ograničena na nekoliko slučajeva koji uključuju visoko upravljane pasmine.

Jedno od rješenja prevladavanja ograničenja koje uzrokuje nepotpuni rodovnik je procijeniti *Ne* koristeći genomske podatke. Nekoliko autora prepoznalo je da se *Ne* može procijeniti iz informacija o LD-u. LD opisuje neslučajnu povezanost alela u različitim lokusima kao funkciju rekombinacijske stope između fizičkih položaja lokusa u genomu. Međutim, potpisi LD-a mogu također proizlaziti iz demografskih procesa kao što su

„*admixture*“ i genetski drift ili putem procesa „*hitchhiking*“. U takvim scenarijima aleli na različitim lokusima postaju povezani, ovisno o njihovoj blizini u genomu. Pod pretpostavnom da je populacija zatvorena i panmiktična, vrijednost LD-a izračunata između neutralnih nepovezanih lokusa ovisi isključivo o genetskom driftu. Ta se pojava može upotrijebiti za predviđanje *Ne* zbog poznate veze između varijance u LD-u (izračunate koristeći frekvenciju alela) i *Ne*.

Najnoviji napretci u tehnologiji genotipizacije (korištenje SNP-ova s desetcima tisuća DNK sondi) omogućili su ogromne količine podataka o povezanosti gena u genomu, što je važno za procjenu *Ne*, kako kod stoke tako i kod ljudi. Međutim, nedostaju softverski alati koji omogućuju procjenu *Ne* putem LD-a.

SNeP predstavlja alat za procjenu *Ne* koristeći SNP podatke u cijelom genomu. Metoda koju SNeP koristi za izračunavanje LD-a ovisi o dostupnosti faznih podataka. Ovaj program je razvijen u C++ programu i namijenjem operacijskim sustavima Windows, OSX i Linux. Ovaj programski paket omogućuje procjenu *Ne* tijekom generacija koristeći SNP podatke koji su ispravljeni za veličinu uzorka, fazu i rekombinacijsku stopu (Barbato, 2015.).

SNeP proizvodi tri izlazne datoteke. U izlaznoj datoteci NeAll (imedatoteke.NeAll) je predstavljeno nekoliko kolona koje su korištene za procjenu *Ne*, a to su GeneAgo (broj generacija u prošlosti koje bi odgovarale vrijednosti *Ne*), *Ne* (vrijednost efektivne veličine populacije), dist (prosječna udaljenost između lokusa), r^2 (LD prosjek), r^2SD (LD standardna devijacija) te items (broj usporedbi koje su pridonijeli procjeni). Takva izlazna datoteka može se lako uvesti u Microsoft Excel, R ili drugi softver koji je u mogućnosti prikazati rezultate. Druga izlazna datoteka sadrži nastavak NeChr (imedatoteke.NeChr), a pohranjuje podatke o LD-u za svaku parnu usporedbu. Sadrži zaglavlje s kolonama: CHR (kromosom na koje se nalazi SNP), dist (udaljenost u baznim parovima između SNP-ova) te r^2 (vrijednost LD-a).
Izlazna datoteka log (imedatotekeSNeP.log) služi programu kao podsjetnik na postavke opcija koje se koriste za određenu analizu.

Format koji je potreban za ulazne datoteke je standardni PLINK format, odnosno datoteke ped i map. Softversko sučelje omogućuje korisniku da kontrolira sve parametre analize, na primjer raspon udaljenosti između SNP-ova. Pored toga, SNeP uključuje opciju izbora MAF praga (zadano je 0.05).

4. Materijali i metode

U radu su korišteni podaci različitih pasmina ovaca genotipiziranih putem Illumina Ovine SNP50v1 SNP čipa. U analizi su obuhvaćene ukupno 354 jedinke iz šest populacija, točnije Merinos de Rambouillet, Australian Pool Dorset, Altamurana, Australian Merino, Santalnes i Castellana. Populacije su grupirane u tri skupine, ovisno o broju jedinki, a svaka skupina je sadržavala dvije pasmine. Tako se procjena *Ne* radila posebno na skupini koja je sadržavala oko 100 jedinki, pa na skupini koja je sadržavala oko 50 jedinki i naposlijetku na grupi koja je sadržavala oko 25 jedinki. Tablica 1. prikazuje točan broj jedinki i pasminu. Procjena *Ne* vršena je putem programskih paketa SNeP i NeEstimator.

Tablica 1. Pasmine ovaca koje su korištene u procjeni Ne

OKVIRNA VELIČINA GRUPE	PASMINA	TOČAN BROJ JEDINKI
	Australian Poll Dorset	108
100	Merinos de Rambouillet	102
	Australian Merino	50
50	Santalnes	47
	Altamurana	24
25	Castellana	23

4.1. WIDDE

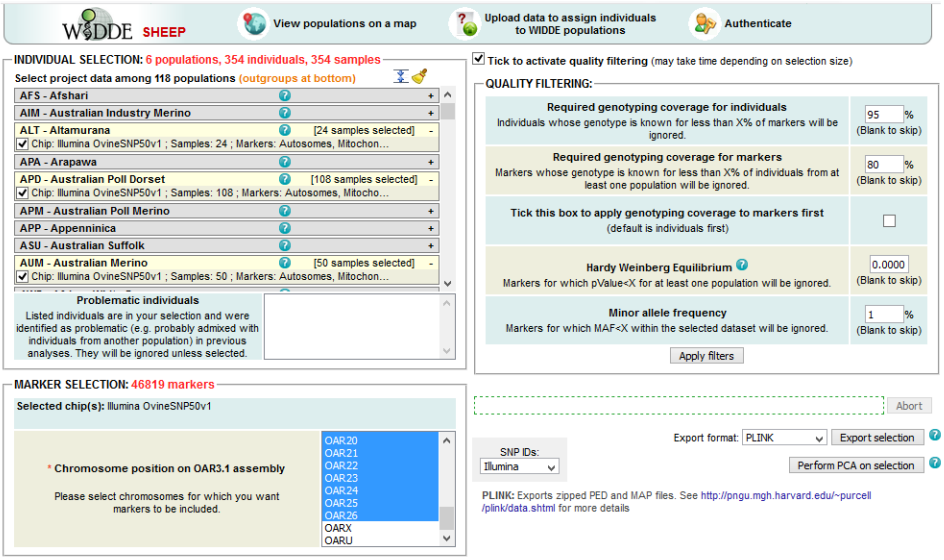
Svi podaci preuzeti su na internetu s baze podataka pod nazivom WIDDE. Ova baza podataka skladišti i upravlja s genotipiziranim podacima. WIDDE sadrži javne (slobodno dostupne) i privatne (zahtijevaju prijavu i lozinku) podatke, a za sada raspolaže podacima za ovce i goveda.

Prije samog preuzimanja genotipiziranih podataka od korisnika se traži specifikacija u tri koraka. Prvi korak omogućuje odabir pasmina koje nas interesiraju. Klikom na njihov naziv pružaju se informacije o SNP čipu koji je korišten za genotipizaciju te broj jedinki (Slika1).

Nakon valjanog odabira pasmina slijedi odabir kromosoma. Spolni kromosomi mogu se izostaviti, kako je i učinjeno u ovoj analizi. Dakle, za potrebe procjene *Ne* odabrani su samo autosomi (Slika 1.).

U trećem koraku nalazi se filter kvalitete, eng. *quality filtering,* gdje je omogućena primjena filtera kvalitete na pasmine koje smo odabrali. Filter uključuje 1) pokrivenost

genotipizacije za jedinke 2) pokrivenost genotipizacije za markere 3) Hardy Weinbergov test 4) prag frekvencije alela koji su slabo zastupljeni. Kao što je također prikazano na slici 1., pokrivenost genotipizacije za jedinke postavljena je na 95%, a to znači da se izbacuje jedinka kojoj nedostaje više od 5% markera. Pokrivenost genotipizacije za markere postavljena je na 80%, drugim riječima, izbacuje se SNP koji je prisutan u manje od 80% jedinki. Opcija Hardy Weinberg Equilibrium iznosi 0.00001, a frekvencija malih alela je 1%. Naposlijetku, slijedi izvoz podataka koje preuzimamo u ped i map formatu.



Slika 1. Odabir pasmina u bazi podataka WIDDE
Izvor: http://widde.toulouse.inra.fr/widde/widde/main.do?module=sheep

4.1. Program SNeP

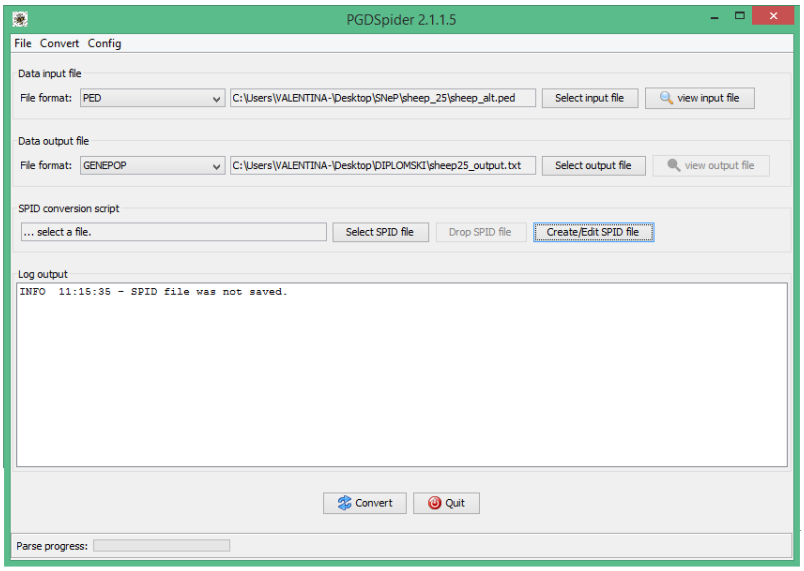
Nakon preuzimanja, ped i map datoteke su spremne za izravan unos u SNeP programskom paketu. Programski paket se pokreće putem naredbenog redka pa se, nakon pozicioniranja u radnom direktoriju, unose podaci u ped i map obliku. Sintaksa poretanja programskog paketa glasi:

```
SnePv1.1.exe -ped imedatoteke.ped -map imedatoteke.map -svedf
-out.
```

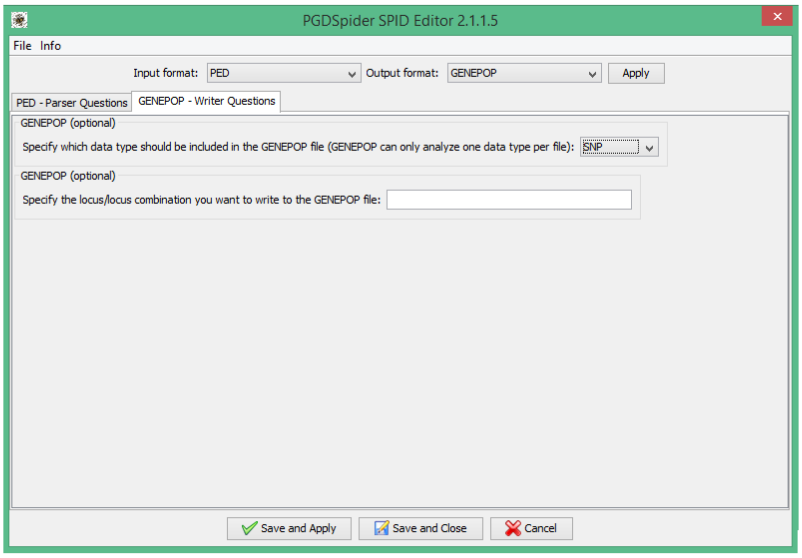
Programski paket nije bio u mogućnosti prepoznati da se u ped datoteci nalaze dvije populacije pa ih je tretirao kao jednu. Iz tog razloga, bilo je potrebno napraviti dvije ped datoteke od kojih je svaka sadržavala jednu populaciju i odvojeno raditi procjenu *Ne* za svaku tu populaciju.

4.2. PGDSpider

NeEstimator je programski paket koji ne prihvaća ped i map format datoteke, nego zahtijeva pretvorbu u GENEPOP format. Pretvorba formata postignuta je programskim paketom PGDSpider. PGDSpider je vrlo moćan alat za automatsku pretvorbu podataka za populacijsku genetiku i genomske programe. Osim konvencionalnih populacijskih genetičkih formata, PGDSpider integrira formate u genomici koji se obično koriste za pohranjivanje i rukovanje sekvenciranih podataka sljedeće generacije, eng. *next generation sequencing*. Napisan je u Javi i zbog toga je neovisan o toj platformi (Lischer, 2012.). Vrlo je jednostavan za instalaciju i omogućuje korisniku odabir željenih postavki. Slika 2. sadrži prikaz odabira ulazne datoteke (ped format) te odabiremo naziv izlazne datoteke koja će biti u GENEPOP formatu. Također, potrebno je odabrati „Create/Edit SPID file“. Zatim, slijedi učitavanje map datoteke te odabit određenih stavki koje prikazuje Slika 3. Zadnji korak prikazan je na Slici 4., a potrebno je specificirati o kojem se tipu podataka radi, odnosno odabrati SNP-ove.



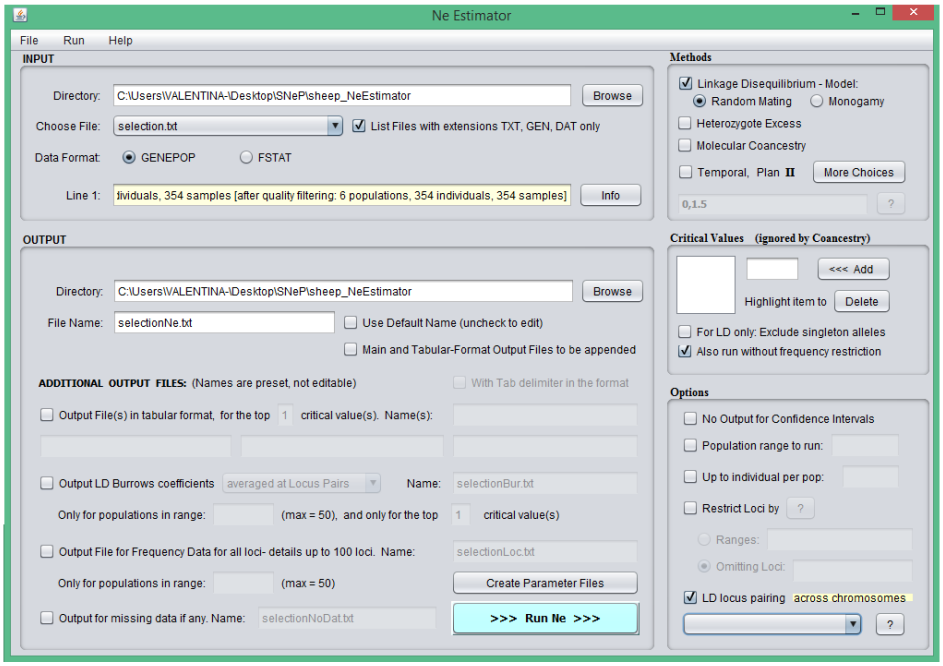
Slika 2. Izgled programa PGDSpider
Izvor: arhiva autorice



Slika 3. Izgled programa PGDSpider - treći korak
Izvor: arhiva autorice

4.1. Programski paket NeEstimator

Ped i map podaci su, nakon pretvorbe u GENEPOP format, spremni za unos u programski paket NeEstimator. Programski paket je neovisan o platformi Java te je jednostavan za pokretanje. Slika 5. pokazuje sve opcije koje korisnik može odabrati za procjenu *Ne*. U „input“ dijelu programa odabire se željeni direktorij, a zatim datoteka koja je programskim paketom PGDSpider pretvorena u GENEPOP format. U „output“ dijelu potrebno je imenovati naziv izlazne datoteke te direktorij u kojem će ona biti spremljena. Prema potrebi i temi ovoga rada izabran je izračun *Ne* putem metode LD-a. Kritična vrijednost nije odabrana, kao što je i vidljivo na slici.



Slika 4. Prikaz programa NeEstimator
Izvor: arhiva autorice

4.2. SAS

U programskom paketu SAS 9.4. uspoređeni su i grafički prikazani dobiveni rezultati procjene *Ne* putem programskog paketa SNeP. Za svih šest populacija napravljena je regresija te se procjena *Ne* mogla usporediti između procjene dobivene programskim paketom SNeP i one dobivene programskim paketom NeEstimator. Sintaksa za izračun regresije za pojedinu populaciju glasi:

```
data sheep_25_alt;
infile "C:\Users\VALENTINA-
\Desktop\SNeP\populacija_25\rezultat_alt\sheep_alt.NeAll"
dlm='09'x dsd firstobs=2;
input GenAgo Ne dist r2 r2SD items;
if GenAgo > 100 then delete;
run;

proc reg data = sheep_25_alt outest=rsheep_25_alt tableout
alpha=0.05;
model ne = genago;
run;
quit;
```

Kao što je vidljivo u kodu, podaci su se filtrirali, odnosno izbačena su sva rješenja za koje je u izlaznoj datoteci GenAgo iznosio više od 100. Time je kod svake populacije izbačeno 14 rješenja, a zadržano 13.

Parametri procjene i statistički parametri iščitavaju se iz dobivene SAS tablice iz koje se iščitava vrijednost *Ne*, kao i interval pouzdanosti.

5. Rezultati i rasprava

Nakon grupiranja populacija u tri grupe izvšena je procjena *Ne* za svaku populaciju u svakom od programskih paketa. Dobiveni rezultati prikazani su u Tablici 1. Iz tablice je vidljivo kako je vrijednost *Ne* za svaku populaciju veća od njezine cenzusne veličine, osim u slučaju manjih populacija. Dakle, jedino populacije čiji je okviran broj jedinki 25 imaju vrijednost *Ne* manju od cenzusnog broja jedinki Isto tako, iz tablice je moguće iščitati intervale pouzdanosti. Interval pouzdanosti za SNeP programski paket dobiven je procjenom regresije u SAS-u, dok je interval pouzdanosti za NeEstimator vidljiv u njegovoj izlaznoj datoteci. Osim procjene *Ne* i intervala pouzdanosti, izlazna datoteka programskog paketa NeEstimator korisniku daje informacije o harmonijskoj srednjoj vrijednosti uzorka, vrijednosti neovisnih usporedbi, zatim jednu vrstu intervala pouzdanosti koji se zove „*jack-knife*“ te dobivenu i očekivanu korelaciju alela na različitim lokusima.

Tablica 2. Rezultati procjene *Ne* putem programa NeEstimator i SNeP

Okvirna veličina grupe	Pasmina	Točan broj jedinki	Procjena Ne putem NeEstimator	IC „jack-knife“ kod NeEstimator *	Procjena Ne putem SNeP	Interval pouzdanosti kod SNeP
100	Australian Poll Dorset	108	79.8	63.2-103.9	129.6	272.1-300.2
	Merinos de Rambouillet	102	239.6	174.7-366.2	286.2	126.9-132.3
50	Australian Merino	50	76.0	46.8-160.3	131.1	112.8-149.4
	Santalnes	47	67.5	42.5-134.1	109.7	96.4-123.0
25	Altamurana	24	31.5	20.7-56.0	22.4	19.2-25.6
	Castellana	23	31.4	19.1-65.7	17.9	14.3-21.5

*IC „jack-knife“ – interval pouzdanosti

Razlike u procjeni *Ne* između programskih paketa su već na prvi pogled očite. Razlike su potpuno logične budući da programski paketi djeluju na različitim algoritmima.

Softver koji procjenjuje *Ne* putem LD-a je SNeP. Programski paket SNeP procjenjuje *Ne* putem metode disekvilibrijuma vezanih markera (*NeLD*) te će računati povijesnu vrijednost *Ne*. Takve se procjene danas često kosriste u populacijskoj genetici ljudi i domaćih životinja.

SNeP računa povijesnu Ne baziranu na vezi između r^2 (prosječna udaljenost svakog para SNP-a), Ne i c (rekombinacijska stopa). Prilikom rukovanja ovim softverom uvjerili smo se da je jedini nedostatak softvera nemogućnost prepoznavanja populacija u ped datoteci, odnosno dvije prisutne populacije program je tretirao kao jednu. Iz tog razloga, bilo je potrebno fizički odvojiti populacije.

S druge strane, softver koji procjenjuje *Ne* putem GD-a je NeEstimator. Programski paket NeEstimator procjenjuje *Ne* iz informacije disekvilibrijuma nevezanih markera (*NeGD*), a takva metoda procjenjuje trenutnu *Ne* te se danas učestalo koristi u populacijskoj genetici divljih životinja. NeEstimator predstavlja nadograđenu verziju programskog paketa LDNE, koja je spomenuta u radu. Omogućuje analizu velikog skupa podataka i za cilj ima pružiti korisniku suvremene nepristrane procjene *Ne*. Procjene *Ne*, posebno one koje su vezane za generacije u daljoj prošlosti, snažno su pod utjecajem manipulacijskih faktora, kao što je MAF. Proučavajući sve detaljnije NeEstimator, nije se bilo teško uvjeriti kako ovaj softver raspolaže s raznim mogućnostima. Uključuje čak četiri metode procjene *Ne*, a za rješavanje problematike ovog rada korištena je metoda procjene *Ne* putem LD-a. Ovaj programski paket je bio vrlo koristan i praktičan u izračunu. Zasigurno treba naglasiti njegovu mogućnost prepoznavanja i razdvajanja populacija. Upravo iz tog razloga, bilo je dovoljno uvesti jednu ulaznu datoteku sa svih šest populacija koje je softver prepoznao i za svaku od njih izračunao *Ne*. NeEstimator neće prihvatiti map i ped format ulazne datoteke, što čak niti ne stvara problem jer se jednostavna pretvorba formata izvršila preko programskog paketa PGDSpider. Format koji NeEstimator prihvaća je GENEPOP ili FSTAT format pa je za analizu korišten GENEPOP. Korisnik, također, može izbrisati sve rijetke alele s frekvencijom koju po izboru zadaje, no u slučaju ove procjene nije korišten niti jedan prag frekvencije rijetkih alela. Metoda procjene putem LD-a za svih šest populacija trajala je gotovo cijeli dan, no izgled i jednostavnost izlazne datoteke pokazao se kao prednost. Izlazna datoteka softvera pruža informacije o procjeni *Ne* za svaku populaciju. Osim efektivne veličine populacije, sadržana je i informacija o intervalu pouzdanosti.

Oba programska paketa su se pokazala iznimno korisnim u analizi ovoga rada. Dobivene razlike u procjeni *Ne* su posljedica djelovanja na različim algoritmima procjene. SNeP će procjenjivati *Ne* preko vezanih markera, dok će NeEstimator procjenjivati iz informacije nevezanih markera. Nakon dobivenih rezultata donešeno je mišljenje kako niti jedan od programa ne prati pravilnost u procjeni *Ne.* Jednostavnije rečeno, kod velikih populacija veće su razlike u procjeni *Ne* dok kod malih populacija pratimo malu razliku u vrijednosti Ne

unutar programa. Dakle, ovisno o tome koja je namjera procjene *Ne*; potrebno je odlučiti se između ova dva programska paketa.

6. Zaključak

- *Ne* definirana je kao veličina „idealne“ Wright-Fischer populacije koja bi dala istu vrijednost određenih genetskih svojstava kao razmatrana populacija
- *Ne* može predvidjeti gubitak i raspodjelu genetske varijacije, vjerojatnost fiksacije korisnih ili štetnih alela te kondiciju i preživljavanje malih populacija
- izračun *Ne* putem metode LD-a prvi je puta korištena prije otprilike četrdeset godina te se od tada primjenjuje, razvija i poboljšava
- kvantitativna vrijednost procjene jako ovisi o veličini uzorka, vrsti LD procjene, dok kvalitativna vrijednost više ovisi o genetskim informacijama nego o manipulaciji podataka
- dostupni programski paket koji procjenjuje *Ne* putem LD-a je SNeP, a putem GD-a je NeEstimator
- SNeP se fokusira na procjenu povijesnih trendova *Ne*, dok je cilj programskog paketa NeEstimator stvaranje suvremenih nepristranih procjena *Ne*.
- *Ne* je moguće izračunati iz slabo povezanih ili nepovezanih lokusa i iz čvrsto povezanih lokusa
- oba programska paketa korištena u analizi pokazala su se kao izvrsni alati u procjeni *Ne*
- moguća komplementarna analiza obje metode može unaprijediti procjenu *Ne,* kako NeLD tako i NeGD

7. Popis literature

1. Barbato, M., Orozco-terWengel, P., Tapio, M., i Bruford, M. W. (2015). SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Frontiers in genetics*, *6*, 109.
2. Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195.
3. Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., Negrini, R., Finlay, E. K., Jianlin, H., Groeneveld, E., i Weigend, S. (2010). Genetic diversity in farm animals–a review. *Animal genetics*, *41*, 6-31.
4. Hayes, B. J., Visscher, P. M., McPartlan, H. C., i Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome research*, *13*(4), 635-643.
5. Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics Research*, *38*(3), 209-216.
6. Lischer HEL and Excoffier L (2012) PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28: 298-299.
7. Mank, J. E., Nam, K., i Ellegren, H. (2009). Faster-Z evolution is predominantly due to genetic drift. *Molecular biology and evolution*, *27*(3), 661-670.
8. Nunney, L., i Elam, D. R. (1994). Estimating the effective population size of conserved populations. *Conservation Biology*, *8*(1), 175-184.
9. Pritchard, J. K., i Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, *69*(1), 1-14.
10. Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1459), 1395-1409.
11. Waples, R. S. (1991). Genetic methods for estimating the effective size of cetacean populations. *Report of the International Whaling Commission (special issue)*, *13*, 279-300.
12. Waples, R. S., i Do, C. H. I. (2008). LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular ecology resources*, *8*(4), 753-756.

13. Waples, R. S., i Do, C. H. I. (2010). Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, *3*(3), 244-262.
14. Waples, R. S., i England, P. R. (2011). Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. *Genetics*, *189*(2), 633-644.
15. Materijali s predavanja iz Populacijske i Konzervacijske genetike, 2017., prof. dr. sc. I. Čurik)
16. http://www.unizd.hr/Portals/13/NASTAVNI_MATERIJALI/04%20%20Distribucije.pdf

Životopis

Rođena sam 19.02.1994. u Zagrebu. Od rođenja živim u malom gradu Hrvatska Kostajnica, odakle su mi roditelji. U Hrvatskoj Kostajnici upisujem OŠ „Davorin Trstenjak“, nakon koje se opredjeljujem za opću gimnaziju. Godine 2012. upisujem Agronomski fakultet u Zagrebu, smjer Animalne znanosti te stječem diplomu prvostupnice inženjerke animalnih znanosti. Na jesen, 2016. godine upisujem diplomski studij Genetika i oplemenjivanje životinja. Ljubiteljica sam rekreativnog trčanja, gledanja filmova te kuhanja. Poznavanje engleskog jezika procijenila bih B razinom.