Metode procjene genomskog inbridinga

Čavlović, Filip

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

METODE PROCJENE GENOMSKOG INBRIDINGA

DIPLOMSKI RAD

Filip Čavlović

Zagreb, rujan, 2021.

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

Diplomski studij:

Genetika i oplemenjivanje životinja

METODE PROCJENE GENOMSKOG INBRIDINGA

DIPLOMSKI RAD

Filip Čavlović

Mentor:

doc.dr.sc. Maja Ferenčaković

Zagreb, rujan, 2021.

Sveučilište u Zagrebu
Agronomski fakultet

University of Zagreb
Faculty of Agriculture

IZJAVA STUDENTA

O AKADEMSKOJ ČESTITOSTI

Ja, **Filip Čavlović**, JMBAG 01781061895, rođen/a 05.05.1996 u Zagrebu, izjavljujem da sam samostalno izradila/izradio diplomski rad pod naslovom:

METODE PROCJENE GENOMSKOG INBRIDINGA

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 **Filip Čavlović**, JMBAG 01781061895, naslova

METODE PROCJENE GENOMSKOG INBRIDINGA

obranjen je i ocijenjen ocjenom _____ , dana _____ .

Povjerenstvo: potpisi:

1. doc.dr.sc. Maja Ferenčaković mentor _____
2. prof.dr.sc. Ino Čurik član _____
3. izv.prof.dr.sc. Vlatka Čubrić Čurik član _____

Zahvala

Zahvaljujem se profesorima MS studija „Genetika i oplemenjivanje životinja“ na stečenom znanju, a posebice mentorici doc.dr.sc. Maji Ferenčaković na strpljenju, razumijevanju, vodstvu te korisnim savjetima kako kroz pisanje ovog rada, tako i kroz ostatak studija.

SADRŽAJ

Sažetak

Diplomskog rada studenta/ice **Filip Čavlović**, naslova

METODE PROCJENE GENOMSKOG INBRIDINGA

Parenje jedinki u srodstvu, odnosno inbreeding dugi je niz godina predmet istraživanja mnogih radova. Inbreeding se mjeri pomoću koeficijenta inbreedinga do čijeg se izračuna može doći korištenjem podataka iz rodoslovlja ili preko genomskih podataka. Procjena stupnja inbreedinga iz pedigrea ima određene nedostatke, a razvoj molekularnih markera doveo je do razvoja niza metoda procjene koeficijenta inbreedinga iz genomskih podataka. Procjena inbreeding koeficijenta iz podataka SNP-ova i dalje je najčešće korištena metoda no pojava NGS sekvenciranja donosi veću količinu i kvalitetu podataka te ima potencijal postati prvi izbor kod procjene inbreeding koeficijenta. Međutim, s dobrim stranama iskazali su se i određeni problemi kao što su cijena, pohrana, mogućnost analize odnosno nedostatak adekvatnih softvera i pitanje privatnosti. Iako se cijena NGS analiza drastično snizila i dalje je financijski nepristupačna mnogim znanstvenicima, osobito ako se istraživanje bazira na većem broju jedinki. Postojeći problemi sa pohranom podataka WGS analiza su i dalje neriješeni, iako su predložena neka djelomična rješenja, većinom sa sobom donose neka ograničenja ili rizike. U radu sažeto su prikazane neke od metoda procjene inbreedinga iz genomskih podataka kao i trenutne limitacije te moguća buduća rješenja

Ključne riječi: inbreeding, koeficijent inbreedinga, NGS

Summary

Of the master's thesis – student **Filip Čavlović,** entitled

METHODS OF ESTIMATING GENOMIC INBREEDING

Mating of related organisms, also known as inbreeding, has been the topic of many studies. Inbreeding is measured by coefficient of inbreeding which is calculated either by using pedigree data or by the usage of genomic data. Estimation of inbreeding coefficient from pedigree has certain disadvantages and development of molecular markers has led to creation of several methods to estimate inbreeding coefficient from genomic data. Estimation of inbreeding coefficient from SNP data is still the most often used method but development of NGS sequencing brings more data and better quality. However, with beneficial sides of NGS, some problems have also arisen. Problems like price, data storing, data processing and the question of privacy. Although the price has drastically dropped it is still not affordable to most researchers, especially if the research is based on a large number of samples. Existing problems with data storage are still not solved and even though some part solutions have been suggested, they come with certain limitations and risks. This thesis summarizes some of the methods for estimating inbreeding from genomic data as well as current limitations and possible future solutions.

Keywords: inbreeding, inbreeding coefficient, NGS

1. Uvod

Inbreeding, najčešće definiran kao parenje jedinki koje imaju zajedničkog pretka, kroz povijest je privlačio pažnju i polemiku u svim slojevima društva. Široko je korišten u stvaranju pasmina domaćih životinja, uzgoju superiornih jedinki, ali i u medicini. U medicini su posebno poznati inbred miševi koji su doprinijeli istraživanjima raka (npr. *BRAF* gen kod malignog melanoma), pretilosti (hormon leptin) i mnogih drugih bolesti.

Zapadno društvo na brakove između bliskih bioloških srodnika općenito gleda neprihvatljivo te ga brane većina religija i zakoni. S druge strane srodstvo je široko preferencijalno u populaciji Azije i Afrike gdje takvi brakovi trenutno otprilike 20–50%. (Alvarez i sur., 2009.) Dok su kod ljudi štetne posljedice inbreedinga najbolje opisane na slučajevima iz europskih monarhija (Alvarez i sur., 2009.) smatra se da kod drugih vrsta ta praksa nije ni približno toliko opasna. Iako inbriding povećava rizik prijenosa štetnih gena, on također osigurava očuvanje dobrih (korisnih) gena unutar populacije te ponekad prednosti inbreedinga nadmašuju nedostatke.

Ipak, moderna stočarska praksa i praksa uzgoja određenih pasmina domaćih životinja povlače za sobom i pitanje o štetnim posljedicama. Najpoznatiji primjer štetnog učinka inbreedinga je inbreeding depresija. Ona se procjenjuje putem regresije gdje je koeficijent inbreedinga (odnosno mjera kojom iskazujemo stupanj inbreedinga) nezavisna varijabla, a svojstvo koje promatramo je zavisna varijabla. Lako zaključujemo kako je za pravilnu procjenu štetnog efekta inbreedinga nužno što pouzdanije procijeniti koeficijent inbreedinga.

Načini procjene koeficijenta inbreedinga su brojni, a danas ih možemo grubo podijeliti na one koji koriste rodovnik i na one koji koriste informaciju o genomu životinje. Oba načina imaju svoju primjenu, prednosti i nedostatke.

1.1. Inbreeding

Inbreeding je uopćeno definiran kao parenje jedinki koje su u srodstvu preko zajedničkih predaka. (Falconer, 1989.). Parenje u srodstvu, pa tako i sama definicija inbreedinga duže su vrijeme polemika mnogih znanstvenih radova te istraživanja i to primarno zato što je postavljanje praga iznad kojeg se jedinke smatraju srodnima relativan pojam. Iz tog razloga parenje u srodstvu najbolje je definirati kao parenje jedinki čiji je stupanj srodnosti veći od prosjeka stupnja srodnosti unutar promatrane populacije ili vrste. (Curik i sur., 2014.)

Još jedan razlog diskusije oko pojma inbreedinga je što se on koristi za objašnjenje širokog raspona genetskih pojmova kao što su smanjenje genetske raznolikosti konačnih populacija, promjene u efektivnim veličinama populacija, genetskom driftu, promjenama u strukturi populacije, odstupanja od Hardy-Weinberg ravnoteže te smanjenja prosjeka populacije. Inbreeding povećava frekvenciju homozigotnih genotipova te smanjuje frekvenciju heterozigotnih genotipova dok frekvencija alela ostaje nepromijenjena. (Curik i sur, 2014.)

Upravo te promjene u frekvencijama mogu dovesti do redistribucije genetskih varijacija unutar i između populacija (Fernandez i sur., 1995.), smanjenja prosjeka populacije za svojstva povezana sa fitnessom (Charlesworth i sur., 2009.), povećanje homozigotnih recesivnih poremećaja (Alvarez i sur., 2009.) te smanjenje u homeostazi populacije (Lerner, 1954.).

Povijesno inbreeding prati negativna konotacija. Puno prije nego što su provedena znanstvena istraživanja ovaj stav se temeljio na opažanjima prvih uzgajivača domaćih životinja te na činjenici abnormalnog razvoja potomaka koji su inbred. Pogotovo na primjeru ljudi primjećuje se kako potomstvo roditelja koji su u srodstvu povećava mogućnost urođenih poremećaja zato što inbreeding povećava proporciju homozigotnosti, osobito za štetne recesivne alele koji uzrokuju takve poremećaje. (Alvarez i sur., 2009.)

Inbreeding odnosno parenje u srodstvu može dovesti do mnogih pozitivnih i negativnih utjecaja na promatranu populaciju. Parenje u srodstvu se može koristiti i za otkrivanje gena koji uzrokuju abnormalnosti ili su letalni. Takvi geni su obično prisutni

u niskim frekvencijama kod populacija gdje nema parenja u srodstvu, gotovo uvijek su recesivni i njihov učinak ne dolazi do izražaja zbog dominantnih alela stoga do ekspresije štetnih gena ne dolazi kada su naslijeđeni pojedinačno nego su potrebne dvije kopije tog alela pa im se vjerojatnost ekspresije povećava inbreedingom. Otkrivanje letalnih gena omogućuje otklanjanje istih iz populacije preko otklanjanja potomaka koji nose te iste gene kao i njihovih roditelja.

Smanjenje proizvodnosti u korelaciji je sa stupnjem inbreedinga, što je veći inbreeding veće je smanjenje proizvodnosti. Međutim, stvarno smanjenje proizvodnosti nije isto kod svih vrsta i svojstava. Parenje u srodstvu ne utječe puno na svojstvo kao što je kvaliteta mesa, ali ima značajan utjecaj na svojstva fitnesa kao što su životni vijek, plodnost i zdravlje. Moguće je predvidjeti utjecaj inbreedinga na određene osobine te se ta predviđanja temelje na rezultatima dobivenim u eksperimentalnim uvjetima u kojima su postignute različite razine inbreedinga.

Postoje mnogi dokazi da inbreeding ima štetne posljedice te da je povezan sa padom prosječnog fenotipa u mnogim biljnim i životinjskim vrstama. (Charlesworth i sur., 2009.). Ova pojava naziva se inbreeding depresijom. Inbreeding depresija definirana je kao smanjenje prosječne vrijednosti svojstava i smanjenje genetske varijabilnosti uzrokovano parenjem u srodstvu. Pošto je inbreeding depresija negativna posljedica inbreedinga potrebna je točna i precizna procjena koeficijenta inbreedinga kako bi se izbjegla njena pojava. Također do inbreeding depresije dolazi sporije u velikim populacijama odnosno u populacijama s većom razinom genetske varijabilnosti. (Charlesworth i sur., 2009.)

Nadalje, učinci inbreeding depresije se povećavaju u nepovoljnim i stresnim okruženjima. Tako se u studijima ptica pokazalo kako jedinke koje su inbred imaju veću stopu smrtnosti tokom jakog nevremena nego outbred jedinke. Također je na populaciji ovaca prikazano da su inbred jedinke zaražene parazitima imale veću stopu smrtnosti od outbred jedinki (Coltman i sur., 1998.). Na populaciji leptira prikazan je pad uspješnosti parenja jedinki koje su bile u zarobljeništvu s obzirom na divlje jedinke (Joron i sur., 2003.). Navedeni podatci sugeriraju da je inbreeding izazvao veću stopu smrtnosti te da simulirani podatci ili podatci dobiveni kroz opažanje jedinki u kontroliranim uvjetima ne mogu u potpunosti pretpostaviti interakciju okoliša i

inbreedinga odnosno da podatci dobiveni simulacijama mogu prikriti prave utjecaje inbreedinga koji mogu biti jači od očekivanih u divljim populacijama. (Polasek, 2009.)

Raspon pojave inbreeding depresije može se kretati od abnormalnosti kao što su mutirani fenotipovi koji su letalni ili uzrokuju genetske bolesti do manje drastičnih posljedica kao što je spomenuto smanjenje fitnes svojstava inbred individua (Charlesworth i sur., 2009.). Polašek (Polasek, 2009.) u svojoj disertaciji pak navodi kako je inbreeding depresija često je povezana s mnogim negativnim učincima na razini jedinke ili populacije te kako inbreeding depresija utječe na povećanje stope smrtnosti, smanjenje svojstava fitnessa, smanjenje reproduktivne sposobnosti, povećanje podložnosti bolestima te smanjene rezistencije od nametnika i dr. Polašek (Polasek, 2009.) također navodi kako količina i intenzitet inbreeding depresije nemaju isti učinak u svim fazama života jedinke.

Suprotnost inbreedingu je parenje jedinki koje nisu u srodstvu odnosno outbreeding. Zbog promjena u individualnoj heterozigotnosti dolazi do povećanja prosjeka svojstva pa se često se koristi u uzgoju životinja i biljaka. Ova pojava ima nekoliko naziva; heterozis, prednost heterozigota ili heterozigotni vigor (Alam i sur., 2004.). Heterozis možemo promatrati na individualnoj razini i na razini populacije te na razini jednog gena ili cijelog genoma. Na razini pojedinca heterozis može značiti povećanu heterozigotnost u odnosu na populaciju iz koje jedinka dolazi, odnosno poboljšanje fenotipskih svojstava kao posljedice takvog genotipa i/ili na poboljšanu vjerojatnost preživljavanja jedinke zbog poboljšanog fenotipa (David, 1998.). Na razini populacije to može predstavljati usporedbu neke populacije s drugom ili pak s referentnom populacijom, a obzirom na prednosti koje proizlaze iz heterozigotnosti genotipa, a na razini gena ili genoma može značiti regije gena ili genoma koje imaju veću ili manju razinu heterozigotnosti u odnosu na neku drugu regiju ili cijeli genom (Polasek, 2009.).

Postavljene su dvije glavne hipoteze kojima se pokušalo objasniti fenomene inbreeding depresije i heterozisa i to su hipoteza djelomične dominantnosti te hipoteza overdominantnosti. Hipoteza djelomične dominantnosti pretpostavlja da se učinci štetnih recesivnih alela povećavaju u homozigotnim organizmima, a hipoteza overdominantnosti pretpostavlja da je heterozigotna kombinacija alela na nekom

lokusu superiornija od bilo koje homozigotne kombinacije na istom (Falconer i sur. 1996.; Polasek, 2009.).

Glavni problem povezan sa inbreedingom i inbreeding depresijom je njihovo otkrivanje i kvantifikacija. U većini slučajeva procijenjeni inbreeding se mora usporediti s referentnom populacijom koju možemo i ne moramo imati , a također je ponekad teško definirati populaciju od interesa zbog raznih demografskih i populacijskih genetičkih događaja kao što su migracije, preklapanje generacija te nejednakost spolova u populaciji. O svemu ovdje spomenutom bit će više riječi u narednim poglavljima.

Unatoč negativnim utjecajima inbreedinga on je i dalje vrlo koristan u uzgoju životinja te je bitan za razvoj i nasljeđivanje poželjnih svojstava budući da povećava udio sličnih gena. (https://extension.missouri.edu/publications/g2911) Inbreeding je također često korišten kod kreiranja pasmina domaćih životinja zbog veće frekvencije nasljeđivanja poželjnih svojstava te mogućnosti fiksacije gena. Inbreeding ima značajnu ulogu i u medicini, kao što je prije napomenuto, u istraživanju raka i pretilosti a uz to i Alzheimerove bolesti, dijabetesa te raznih patogena i ostalih bolesti. Inbred miševi, odnosno određeni sojevi, pokazali su podložnost infekciji kada su bili izloženi određenim patogenima, dok su neki sojevi pokazali rezistenciju. Neke od bolesti koje uzrokuju ti patogeni su antraks, listerioza, pneumonija i tuberkuloza gdje je patogen bakterija te gripa, herpes (citomegalovirus) te bolest Zapadnog Nila gdje je patogen virus. Genetske studije koje su se bazirale na takvim podatcima omogućile su otkrivanje mnogih gena čija uloga objašnjava zašto su neki domaćini osjetljivi a neki rezistentni na određene patogene, kako kod miševa tako i kod ljudi (https://www.jax.org/news-and-insights/jax-blog/2013/october/inbred-mice-genetic-tools-for-modeling-infectious-diseases). Zato što je inbreeding važan za proučavanje i razumijevanje evolucije biljaka i životinja te za promatranje kvalitete komercijalno važnih pasmina znanstvenici su razvili nekoliko metoda za njegovu procjenu. (Curik i sur. 2014.) U svakoj metodi spominje se pojam identičnost po porijeklu (*engl. Identical by descent*; IDB) stoga ga objašnjavam prije opisa metoda procjene inbreedinga.

1.2. Identičnost po podrijetlu

Aleli se smatraju identičnima po podrijetlu (IBD) ako su naslijeđeni od istog ancestralnog haplotipa odnosno od istog zajedničkog pretka. Pojam IBD prvi je upotrijebio Crow (1954.). Jedinka koja na nekom lokusu ima dva alela koji su IBD naziva se autozigotom. (Falconer, 1989.)

Ono što veže koeficijent inbreedinga (*F*) i IBD kao pojam je definicija po kojoj je on *F* proporcija gena neke individue koji su IBD te se može računati iz rodovnika odnosno pedigrea, ali i procijeniti iz frekvencija alela dobivenih genotipizacijom. Kada je *F* izračunat preko rodovnika IBD status se može tumačiti kroz vjerojatnost ili korelaciju putanje te je dobivena vrijednost uvijek određena strukturom rodovnika. Za takvu procjenu *F* mora se uspostaviti bazna populacija (Curik i sur., 2014.). Pošto se broj predaka u pedigreu po generaciji povećava za 2n, gdje je n broj generacija, dolazi do eksponencijalnog povećanja pedigrea što dovodi do toga da su sve u određenoj generaciji srodne.

Zanimljiv primjer spominje genetičar Adam Rutherford, a vezano za ljudsku populaciju (https://www.scientificamerican.com/article/humans-are-all-more-closely-related-than-we-commonly-think/). On traži čitatelja da broji sve svoje pretke u obiteljskom stablu unatrag. U n-toj generaciji prije sadašnje naše obiteljsko stablo ima 2n grane: dvije za roditelje, četiri za bake i djedove, osam za pradjedove i tako dalje. Broj grana tako raste eksponencijalno. Do 33. generacije - prije otprilike 800 do 1000 godina - imamo ih više od osam milijardi. To je više od broja ljudi koji danas žive, a to je zasigurno mnogo veća brojka od svjetske populacije prije 1000 godina. Iako je na prvi tren paradoksalno, autor navodi kako grane našeg obiteljskog stabla nisu u potpunosti razdvojene i nezavisne i isprepliću se. Mnogi preci mogu zauzimati više grana. Rutherford navodi dalje primjer kako je nečija pra-pra-pra-pra-pra-prabaka mogla je biti ujedno biti i pra-pra-pra-pra-teta, a to sve zajedno dovodi do zaključka da je čovječanstvo prilično srodno i da smo svi srodniji nego što intuitivno mislimo. U navedenom popularnom članku spominje se i 2004. godina kada je putem matematičkog modeliranje i računalne simulacije grupe statističara pod vodstvom Douglasa Rohdea, ustanovljeno kako je najmlađi zajednički predak cijelog današnjeg čovječanstva vjerojatno živio ne prije 1400. godine prije Krista, a moguće i 55. godine

poslije Krista. U vrijeme egipatske kraljice Nefertiti (otprilike 1370. pr. Kr. - otrprilike 1330. pr. Kr.), netko od koga svi potječemo vjerojatno je bio živ negdje u svijetu. Gledajući još više unatrag dolazimo do trenutka kada naša obiteljska stabla dijele ne samo jednog zajedničkog pretka, već svakog zajedničkog pretka, a taj se trenutak naziva genetska izotočka i u njoj obiteljska stabla bilo koje dvije osobe na zemlji, koliko god udaljeno izgledala, vode istoj grupi jedinki. Iako su prvi ljudi napustili Afriku i počeli se širiti svijetom prije najmanje 120.000 godina, genetska izotočka dogodila se mnogo kasnije i to negdje između 5300. i 2200. godine prije Krista (https://www.scientificamerican.com/article/humans-are-all-more-closely-related-than-we-commonly-think/).

Bazna populacija u teoriji bi trebala biti sačinjena od jedinki za čije se roditelje donosi pretpostavka da su nepoznati i da nisu u srodstvu. (Falconer i sur., 1996.) U stvarnosti svi pojedinci za koje se smatra da nisu u srodstvu prema gore navedenom primjeru u srodstvu su preko zajedničkog pretka prije n generacija, odnosno u genetskoj izotočci. Njihovi IBD segmenti zbog zajedničkog pretka prije n generacija imaju IBD segment prosječne duljine 1/(2n) Morgana (M) (Browning i sur., 2010.). Browning (Browning i sur., 2010.) pojašnjavaju kako su rođaci koji su imali zajedničkog pretka pred 50 generacija imaju IBD segment prosječne duljine od 1 centi morgana (cM).

Molekularni pristup procjene ima za cilj utvrđivanje svih IBD segmenata čije detaljno utvrđivanje u stvarnom svijetu nije moguće iz razloga što se ne mogu identificirati svi IBD segmenti za sve zajedničke pretke kroz dugo vremensko razdoblje. (Curik i sur., 2014.) Kada su pak u pitanju rodovnici s velikom količinom informacija, odnosno s mnogo generacija, IBD segmenti mogu postati veoma kratki što može otežavati njihovo otkriće.

U svrhu rješavanja navedenih problema potrebno se osloniti na dodatne informacije koje ili izravno procjenjuju autozigotnost ili su u korelaciji s autozigotnosti. Također podatci o duljini haplotipa mogu dati indormacije o IBD statusu. Prema spomenutoj formuli za očekivanu duljinu IBD segmenta ($L_{EXPECTED}$=1/(2n) M) jasno je da što su dulji homozigotni haplotipovi veća je i vjerojatnost da su IBD (Browning i sur., 2010.).

1.3. Pedigre inbreeding koeficijent

U nizu radova objavljenih između 1913. i 1917. godine Raymond Pearl je napravio prvi pokušaj izračuna inbreedinga na temelju podataka iz rodovnika, odnosno pedigrea. Sewall Wright 1922. godine objavio je pristup izračuna inbreeding koeficijenta iz pedigrea i njegova metoda dobiva široku upotrebu. Njegova metoda izračuna temelji se na načelu da je inbreeding jedinke jednak polovici srodnosti roditelja. Proširenje ovog modela sugerira da se inbreeding može izračunati analizom putanje iz korelacije između proizvoljnih vrijednosti dodijeljenih uniji svih mogućih gameta. Standardna formula za koeficijent inbreedinga, F, glasi:

$$F_X = \sum (1/2)^n \, (1 + F_a)$$

Gdje je *n* broj jedinki u nekom nizu koji povezuje roditelje jedinke x sa zajedničkim pretkom, a F_a je inbreeding zajedničkog pretka.

Međutim, iako je takav koeficijent bio jednostavan za izračun njegovo biološko značenje bilo je teško za protumačiti, osobito za pojedince s nepotpunim ili netočnim pedigreom. (Curik, 2014.)

1948. godine Malecot je definirao inbreeding koeficijent kao vjerojatnost da su dva nasumično odabrana alela nekog lokusa, koji je nasumično odabran između svih lokusa u genomu, identični po podrijetlu. (Malecot, 1948.) U odsustvu selekcije i mutacija donosi se pretpostavka da svi lokusi segregiraju na isti način te se stoga očekuje da imaju isti koeficijent inbreedinga nazvan pedigre inbreeding koeficijent (F_{PED}) (Curik, 2014.). F_{PED} jednak je prosječnoj autozigotnosti genoma jedinke. Procjene inbreedinga temeljene na IBD-u moraju se usporediti sa ancestralnom populacijom u kojoj nijedna jedinka nije srodna odnosno s baznom populacijom za koju je već utvrđeno da je teško odrediva.

Kod izračuna koeficijenta inbreedinga koristeći samo pedigre najčešće se koriste „analiza putanje“ (*engl. path analysis*), matrica aditivne srodnosti (*engl. additive relationship matrix*) (Ballou, 1983.) i stohasticka segregacija gena (*engl. Gene dropping*). Za izračun koeficijenta inbreedinga Wrightov pristup je standard kada se

odnosi na jednu jedinku s jednostavnim pedigreom. Nasuprot tome, matrica aditivne srodnosti pokazala se vrlo učinkovitom za brz izračun koeficijenta inbreedinga za sve individue neke populacije, čak i one sa velikim i kompleksnim pedigreom. (Tier, 1990.; VanRaden, 1992.; Aguliar i sur., 2008.) “Gene dropping” metoda može biti najprikladnija za dobivanje nepristrane procjene, na primjer kod izračuna ancestralnog inbreeding koeficijenta (Suwanlee i sur., 2007.) (Curik i sur., 2014.).

Koeficijenti inbreedinga dobiveni iz pedigrea imaju mnogo nedostataka. Jedan od najčešćih nedostataka je dostupnost i točnost podataka pedigrea. Rodoslovlje je u praksi često nepotpuno, neprecizno i netočno kako za ljude tako i za domaće životinje, a za divlje životinje ne postoje pedigre podatci. Međutim, čak i ako su podatci u pedigreu potpuno pouzdani i točni, F_{PED} ne uzima u obzir stohastičku prirodu nasljeđivanja. (McQuilan i sur., 2008.)

Kako smo već utvrdili, svaki pedigre počinje od neke bazne populacije nesrodnih životinja. Ipak, postoji mogućnost da je došlo do inbreedinga u generacijama čiji podatci nisu prisutni u pedigreu što uzrokuje da svaka procjena koeficijenta inbreedinga iz takvog pedigrea ne može biti i nije potpuno precizna. Također je bitno uzeti u obzir mogućnost lažnog očinstva ili drugih vrsta grešaka u pedigreu koje utječu na procjenu F_{PED} (Bellis i sur., 2005.; Polasek i sur., 2010.; Keller i sur., 2011.)

F_{PED} procjene također pretpostavljaju neutralnost. Curik i sur. (2002.) proveli su simulacijsko istraživanje koje sugerira da procjena koeficijenta inbreedinga iz pedigrea dovodi do pristranih vrijednosti za „ostvarenu“ ili „pravu“ autozigotnost. Pristranost ovisi o intenzitetu selekcije te o genetskom modelu svojstava pod selekcijom. Procjene autozigotnosti koje se temelje na pedigreu niže su od procjene temeljene na lokusima s aditivnom ili djelomičnom dominacijom i više od procjene temeljene na lokusu s overdominacijom. Učinak selekcije na inbreeding koeficijent zanemaren je u predgenomskoj eri zbog pretpostavke da selekcija ne utječe na autozigotnost lokusa, odnosno na nezavisnost segregacije alela u genomu. (Curik i sur. 2014.)

2. Metode procjene inbreedinga iz genomskih podataka

Prvi napredak u korištenju molekularnih informacija za procjenu inbreedinga i individualne multilokusne heterozigotnosti (*engl. Individual multilocus heterozygosity;* MHL) proizašao je iz teorijskih studija na razini populacije od strane Curie-Cohen 1981. godine te Li i Horvitz 1953. godine te iz kompjuterskih simulacija na individualnoj razini koje su izveli Bereskin i sur. 1969. i 1970. godine te Mitton i Pierce 1980. godine.

Mikrosateliti su u neku ruku prvi naširoko korišteni genetski markeri pogodni za mnoge primjene u populacijskoj genetici, a njihovu upotrebu često vidimo i danas. Oni omogućuju procjenu genetske diferencijacije među populacijama (Goldstein i sur., 1995.; Slatkin, 1995.) kao i praćenje toka gena *(engl. Gene flow*) pomoću analize roditelja (Asuka i sur., 2005.). Budući da su mikrosateliti prvenstveno neutralni često se koriste za procjenu parametara bitnih za reprodukciju životinja kao što su inbreeding ili srodnost unutar populacije (Blouin, 2003.). (Chybicki i sur., 2009.)

Izvor potencijalne pogreške kod izračunavanja srodnosti i inbreedinga iz mikrosatelita je prisutnost null alela (*engl. Null allele*). Null aleli su aleli koji se ne amplificiraju što dovodi do toga da jedinke koje imaju jedan ili više null alela na određenom mjestu mogu biti pogrešno definirane kao homozigoti ili se ne uzimati u obzir za to mjesto. Jedan od zahtjeva za točnu procjenu srodnosti iz mikrosatelitskih podataka je zanemarivanje lokusa koji sadrže null alele. (van de Casteele i sur., 2001.)

Međutim, lokusi koji sadrže null alele uobičajeno predstavljaju veliki dio podataka te njihovo uklanjanje može dovesti do značajnog gubitka informacije o određenoj populaciji. (Wagner i sur., 2006.)

Za mjerenje individualnog koeficijenta inbreedinga potrebna je poznata srodnost oba roditelja što zahtjeva podatke o podrijetlu izvan onih dostupnih za većinu divljih populacija (Marshall i sur., 2002.). U tu svrhu razvijeno je nekoliko multilokusnih mjera kako bi se preko mikrosatelitskih podataka došlo do procjene inbreeding depresije u populacijama za koje nisu dostupni koeficijenti inbreedinga (Coltman i sur., 2003.). Najčešće takve mjere su MHL, srednja vrijednost d^2 te unutarnja srodnost (*engl. Internal relatedness;* IR). (Curik i sur., 2014.)

Empirijske studije koje su 2004. godine proveli Slate i sur. (2004.), računalne simulacije Balloux i sur. (2004.) te teorijske analize (DeWoody i sur., 2005.) pokazale su da je broj mikrosatelitskih markera koji se obično koristi u istraživanju (od 15 do 50) prenizak za točnu procjenu heterozigotnosti u cijelom genomu ili za prikaz korelacije s inbreeding koeficijentom. (Curik i sur., 2014.)

Razvoj mikrosatelita pružio je nove mogućnosti za poboljšanje razumijevanja genetskih varijacija u populaciji domaćih životinja (Defaveria i sur., 2013.; Fernandez i sur., 2005.). Procjena koeficijenta inbreedinga iz mikrosatelitskih markera je brza, jeftina i ne zahtjeva podatke o pedigreu. Međutim, procjena inbreedinga iz mikrosatelita objašnjava samo mali dio varijabilnosti od procjena baziranih na pedigreu (Wang, 2016.). Također, usporedba pedigre i mikrosatelitskih inbreeding koeficijenata razmatra dvije vrste različitih homozigotnosti, a to su identičnost po podrijetlu i identičnost po stanju (Hill i sur., 2011.). (Cortes i sur., 2019.)

Temeljem navedenog polimorfizmi jednog nukleotida *(engl. Single nucleotide polymorphisms;* SNP) postaju najčešće korišteni markeri u mnogim studijima. Njih je mnogo u genomu, učinkovito se detektiraju i ispituju i lako analiziraju. (Wakeley, 2001.)

SNP-ovi su zapravo najrasprostranjeniji oblik genetskih varijacija u genomu (za razliku od mikrosatelita) (Seeb i sur., 2011.), i oni mogu pružiti potrebnu rezoluciju za rješavanje mnogih pitanja u ekologiji, evoluciji, biologiji i genetici.

Razvoj platformi za genotipizaciju visoke gustoće dovele su do mogućnosti za povećanje točnosti procijenjenih parametara. Dva najčešća prostupa za genotipizaciju SNP-ova su metoda genotipizacije sekvenciranjem (*engl. Genotyping by sequencing;* GBS) te metode temeljene na nizu (*engl. Array based methods*) u kojima su paneli već poznatih polimorfizama hibridizirani na „chipove“ od strane kompanija kao što su Illumina i Affymetrix. (Humble i sur., 2020.)

GBS metode mogu genotipizirati desetke tisuća SNP-ova, no generiraju velike količine podataka koje zahtijevaju bioinformatičku obradu što može biti vremenski i tehnički izazovno (Shafer i sur., 2017.). Dodatni problem s GBS metodama je što dubina pokrivenosti sekvence nije uvijek dovoljno visoka da se sa sigurnošću utvrde genotipovi što može dovesti do visokih stopa nedostajućih podataka (Huang i sur., 2016.; Benjelloun i sur., 2019.). (Humble i sur., 2020.)

Nasuprot tome, metode temeljene na nizu su brže, zahtijevaju minimalne tehničke napore, imaju nisku stopu grešaka u genotipizaciji te visoku pokrivenost (Shi i sur., 2012.; Thaden i sur., 2020.). (Humble i sur., 2020.)

Razvijeno je nekoliko metoda za procjenu inbreeding koeficijenata baziranih na SNP-ovima. Neki od njih su genomske matrice srodnosti te ROH inbreeding koeficijent.

2.1. Genomske matrice srodnosti

Genomske matrice srodnosti jedan su od načina procjene koeficijenta inbreedinga iz genomskih podataka. To su matrice kovarijance izračunate na temelju SNP informacija pojedinaca odnosno na temelju malog broja alela. Imaju važnu ulogu u mješovitim modelima za analize i predviđanja na području genetike. (Schlather, 2020.)

BLUP (*engl. Best linear unbiased prediction*; BLUP) se s razvojem i sve većom dostupnošću genomskog materijala zamjenjuju sa GBLUP (*engl. Genomic best linear unbiased prediction*; GBLUP). Kako je kod BLUP-a dijagonala numerator matrice srodnosti (*engl. Numerator relationship matrix*; NRM) jednaka 1 + inbreeding koeficijent odgovarajuće jedinke tako je generalno prihvaćeno i kod GBLUP-a gdje se numerator matrica srodnosti zamjenjuje sa jednom od genomskih matrica srodnosti. (Villaneuva i sur., 2021.)

Kako se genomske matrice srodnosti koriste za procjenu genomskog inbreeding koeficijenta potrebno je istražiti koja od matrica daje najbolju procjenu. Villaneuva i sur. (2021.) usporedili su koeficijente inbreedinga dobivene iz procjene dijagonala pet genomskih matrica srodnosti. Pošto ne postoji opća suglasnost nomenklature tih matrica, odabran je naziv prema autoru koji ih je prvi postavio ili formulirao. (Villaneuva i sur., 2021.)

F_{NEJ} je inbreeding koeficijent izračunat iz dijagonalnih elemenata alelne matrice srodnosti prema formuli u radu Nejati-Avaremi i sur (1997.).

$$F_{NEJ} = \frac{\sum_{k=1}^{S}(\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij_k})/2}{S} - 1$$

I_{ij_k} je identitet dva alela, *i* i *j* neke individue na SNP-u *k* te poprima vrijednosti od 0 do 1 gdje je 0 ako aleli nisu identični, a 1 ako jesu. F_{NEJ} je proporcija SNP-ova jedinke koji su homozigotni te ne razlikuje između alela IBD i IBS

$F_{L\&H}$ je inbreeding koeficijent baziran na matrici srodnosti koja opisuje devijacije od Hardy-Weinberg ekvilibrijuma koja je računata prema formuli navedenoj u radu Li i Horwitz (1953.)

$$F_{L\&H} = \frac{SF_{NEJ} - \sum_{k=1}^{S}[1 - 2p_{k(0)}(1 - p_{k(0)})]}{S - \sum_{k=1}^{S}[1 - 2p_{k(0)}(1 - p_{k(0)})]}$$

$p_{k(0)}$ je frekvencija referentnog alela (alel B) SNP-a *k* u referentnoj populaciji. Procjenjuje devijaciju opažene frekvencije homozigota od one očekivane u baznoj populaciji koja je pod utjecajem Hardy-Weinberg ekvilibrijuma. Time ispravlja homozigotnost prisutnu u baznoj populaciji i izražava inbreeding za IBD alele.

F_{VR1} je koeficijent inbreedinga izračunat iz dijagonalnih elemenata genomske matrice srodnosti dobiven metodom 1 navedenom u radu VanRaadena (2008.). F_{VR1} je baziran na varijanci aditivne genetske vrijednosti i daje mjeru relativnu frekvencijama referentnih alela u baznoj populaciji.

$$F_{VR1} = \frac{\sum_{k=1}^{S}(x_k - 2p_{k(0)})^2}{2\sum_{k=1}^{S} p_{k(0)}(1 - p_{k(0)})} - 1$$

Gdje je x_k genotip individue za SNP *k* kodirani kao 0 za genotip AA, 1 za genotip AB te 2 za genotip BB a $p_{k(0)}$ je frekvencija referentnog alela (alel B) SNP-a *k* u referentnoj populaciji. Razlika između F_{VR1} i $F_{L\&H}$ je u tome da F_{VR1} homozigotni genotipovi su dobiveni inverzijom frekvencije njihovih alela te zbog toga rjeđi homozigotni genotipovi pridonose više mjeri inbreedinga nego učestali homozigotni genotipovi.

F_{VR2} je koeficijent inbreedinga izračunat iz dijagonalnih elemenata genomske matrice srodnosti dobiven metodom 2 navedenom u radu VanRaadena (2008.).

$$F_{VR2} = \frac{1}{S}\sum_{k=1}^{S}\frac{(x_k - 2p_{k(0)})^2}{2p_{k(0)}(1 - p_{k(0)})} - 1$$

Gdje su x_k i $p_{k(0)}$ jednaki kao za F_{VR1}. F_{VR2} je sličan F_{VR1} ali je suma kroz markere dobivena drugačije, na način da je utjecaj dan rijetkim alelima još veći. Doprinos svakog SNP-a je podijeljen sa varijancom istog dok je u F_{VR1} zajednički doprinos svih SNP-ova podijeljen sa zajedničkim denominatorom.

F_{YAN} je inbreeding koeficijent izračunat iz dijagonalnih elemenata genomske matrice srodnosti prikazanoj u radu Yang (2010.).

$$F_{YAN} = \frac{1}{S}\sum_{k=1}^{S}\frac{x_k^2 - \left(1 + 2p_{k(0)}\right)x_{k_i} + 2p_{k(0)}^2}{2p_{k(0)}(1 - p_{k(0)})}$$

Gdje su x_k i $p_{k(0)}$ jednaki kao za F_{VR1}. F_{YAN} baziran je na korelaciji između ujedinjujućih gameta te pridaje više važnosti homozigotnosti „minor“ alela odnosno alela čija je frevencija druga u populaciji za određeni SNP, nego homozigotnosti „major“ alela odnosno alela koji su najčešći za određeni SNP u promatranoj populaciji. . (Villaneuva i sur., 2021.)

Na temelju dobivenih rezulata, Villaneuva i sur., (2021.) su zaključili da osim F_{NEJ} (koji se kreće od 0 do 1), vrijednosti za genomske koeficijente koji su istraženi nalaze se izvan raspona Malecotove i Wrightove definicije koeficijenta inbreedinga. Ako se koristi interpretacija inbreedinga kao dobitka ili gubitka varijabilnosti tada $F_{L\&H}$ daje razumne vrijednosti, dok F_{VR1}, F_{VR2} i F_{YAN} ne. Kada se gledaju očekivanja dobivena na razini populacije do izražaja dolaze neke nedosljednosti za F_{VR1}, F_{VR2} i F_{YAN}. Te nedosljednosti uključuju indikaciju da se može izgubiti više varijabilnosti nego što je u početku postojalo (F_{VR1}, F_{VR2} i F_{YAN}), da se varijabilnost smanjila kada se u stvarnosti povećala (F_{VR1}, F_{VR2} i F_{YAN}), da se varijabilnost povećala kada se u stvarnosti smanjila (F_{VR1} i F_{VR2}) i da nije moguće postići varijabilnost veću od one koja je početno postojala (F_{YAN}). (Villaneuva i sur., 2021.)

2.2. ROH segmenti

Većina autora se slaže kako su Broman i Weber 1999. godine prvi koji su u genomu ljudi primijetili duge kromosomske segmente homozigotnih markera. Oni postojanje tih segmenata objašnjavaju kao posljedicu autozigotnosti, odnosno da su rezultat parenja individua koje su u srodstvu te da postoji mogućnost da imaju utjecaj na zdravlje ljudskog organizma. (Broman i sur., 1999.) Lencz i sur. (2007.) ove segmente nazivaju „Runs of Homozygosity“ (ROH) i ovaj je naziv uvriježen i danas.

Prva analiza identifikacije ROH segmenata provedena je 2006. godine. Gibson i sur., (2006.) čiji su rezultati opisali ROH segmente različitih duljina, učestalosti te raspodjele po genomu potaknuli su mnoge ROH analize u humanoj i animalnoj genetici. Kirin i sur. (2010.) provode istraživanje populacijske povijesti i srodstva kod ljudi temeljenih na ROH-u dok su u stočarstvu prve studije o ROH-u izrađene od strane Soelknera i sur., (Soelkner i sur., 2010.) te Ferenčaković i sur. (Ferenčaković i sur., 2011.) na populacijama goveda. U svinjogojstvu prvi radovi na temu ROH-a provedeni su kako bi se istaknuo utjecaj odnosa između populacija, demografske povijesti i utjecaja inbreedinga na frekvenciju homozigotnosti. (Bosse i sur., 2012.; Herrero-Medrano i sur., 2013.). Khanshour (2013b.) te Metzger i sur. (2015.) provode ROH analize u svrhu otkrivanja prisutnosti pozitivne selekcije kod konja. Kod ovaca provedena su istraživanja o populacijskoj strukturi, populacijskoj povijesti te homozigotnosti na temelju ROH segmenata od strane Beynon i sur. (2015.) te Muchadeyi i sur. (2015.). Guangul (2014.) analizira ROH segmente i genomske koeficijente inbreedinga kod koza. (Peripolli i sur., 2017.)

Sve navedene studije definiraju ROH segmente kao dugačke regije kromosomskih segmenata koje se identificiraju molekularnim markerima i homozigotne su te mogu pružiti uvid u evolucijsku povijest i demografsku povijest populacije, procjenu razine inbreedinga, identifikaciju utjecaja selekcije u proizvodnim životinjama i još mnogo toga.

Utvrđeno je također kako distribucija ROH-ova nije slučajna u cijelom genomu (Bosse i sur., 2012.), a genomska područja s najvećom učestalošću ROH segmenata nazivaju se često ROH otoci (engl. *ROH islands*) (Nothangel i sur., 2010.) ili ROH žarišta (engl. *ROH hotspots*) (Pemberton i sur., 2012a.). Suprotno tome, regije s

niskom učestalošću ROH segmenata nazivaju se ROH pustinje (*engl. ROH deserts*) ili ROH „coldspots“ (Curik i sur., 2014.). Obzirom kako selekcija može dodatno smanjiti genetsku raznolikost i povećati razinu homozigotnosti u populaciji vjerojatnije je da će se genomska područja pod selekcijom poklapati sa ROH otocima. (Pemberton i sur., 2012a.) i stoga se analiza ROH otoka obično koristi za otkrivanje genomskih regija pod selekcijom te otkrivanje genetske pozadine ekonomski važnih svojstava u populacijama domaćih životinja. (Fang i sur., 2021.)

Ringbauer i sur. (2021.) proveli su analizu drevne DNA (*engl. Ancient DNA*; aDNA) temeljenu na ROH-u odnosno srodnosti. U radu su identificirali ROH segmente u humanoj aDNA niske pokrivenosti na temelju informacija o haplotipu iz referentnog panela novijih uzoraka. Analizirali su genomske podatke od 1.785 individua koji su živjeli kroz zadnjih 45 000 godina. Otkrili su smanjenje srodnosti s ili ubrzo nakon tranzicije iz nomadskog na sjedilački način života. Takvo otkriće, vjerojatno povezano sa povećanjem veličine lokalnih populacija, opaženo je kroz nekoliko geografskih regija diljem svijeta. (Ringbauer i sur., 2021)

2.2.1. Identifikacija ROH segmenata

Identifikacija ROH segmenata većinom se vrši iz podataka dobivenim iz GWAS (*engl. Genome-wide association study;* GWAS) analiza SNP niza (*engl. SNP array*). (Ku i sur., 2011.; Yang i sur., 2012.; Joshi i sur., 2015.) Ovi podaci su vrlo dostupni, i smatraju standardom s vrlo niskom greškom kod identifikacije ROH segmenata. SNP arrays obično uključuju od 1,2 do 1,5 milijuna SNP-ova tipično s frekvencijama alela većima od 0.05 kako bi se najbolje predstavila struktura haplotipova u odabranoj populaciji. SNP array sa 300 tisuća SNP-ova dovoljni su za identifikaciju ROH segmenata duljih od 1Mb što odgovara pravom ROH-u uzrokovanom autozigotnošću dok s povećanjem duljine željenog segmenta pada potrebna razlučivost SNP niza. (McQuillan i sur., 2008.; Ferenčaković i sur. 2013b.; Ceballos i sur., 2017.).

Brojni faktori utječu na identifikaciju ROH segmenata kao što su gustoća markera, distribucija markera kroz genom, postotak pogrešaka i učestalost rjeđeg alela (*engl. Minor allele frequency;* MAF). (Ferenčaković i sur. 2013b.; Ceballos i sur., 2017.) Unatoč velikom broju faktora o kojima ovisi precizna identifikacija ROH segmenata, još

uvijek ne postoje optimalni uvjeti niti čimbenici koji se moraju zadovoljiti za detekciju ROH segmenata. (Ferenčaković, 2019.)

Uklanjanje SNP-ova s niskom MAF vrijednošću uobičajeno je u GWAS analizama jer su tamo nepotrebni, no inercijom je isto također prihvaćeno kao kontrola kvalitete u ROH analizama. (Ferenčaković i sur. 2013b.; Hillestad i sur., 2017.) Iz razloga što su ROH kontinuirani segmenti važno je zadržati što više genomskih podataka koji upućuju na evolucijske i selekcijske sile poput genetskog drifta ili fiksacije alela, uključujući tako SNP-ove s niskim MAF, kako se ROH-ovi ne bi razdvojili u manje segmente ili izgubili. (Hillestad i sur., 2017.)

Veličina ROH segmenata može upućivati na vrijeme inbreedinga gdje duži ROH segmenti upućuju na nedavno parenje u srodstvu gdje rekombinacija ne skraćuje identične haplotipove naslijeđene od zajedničkog pretka te suprotno tome kraći ROH segmenti upućuju na parenje u srodstvu prije mnogo generacija. Mogućnost da se veličinom ROH segmenata otkriju informacije o starim i nedavnim genetskim promjenama čini ROH korisnim alatom za analizu povijesti populacije, razine inbreedinga i učinka parenja u srodstvu na složene osobine i urođene poremećaje. (Ferenčaković i sur, 2013a.)

Povećanjem gustoće SNP-ova pojavljuju se i novi homozigotni i heterozigotni SNP-ovi u ROH segmentu. Pojava novih SNP-ova uzrokuje povećanje kraćih ROH segmenata i smanjenje dužih iz čega se može zaključiti da se povećanjem gustoće SNP-ova poboljšava rezolucija i smanjuju se greške podjelom dužih ROH segmenata na kraće. Kod detekcije ROH-a od velike je važnosti zadržati što veći broj SNP-ova da bi se dobile bolje informacije o distribuciji homozigotnosti. Veća gustoća SNP-ova pridonosi preciznijoj procjeni ROH-a od niže gustoće. (Hillestad i sur., 2017.)

Kroz razdoblje intenzivne analize ROH segmenata, koje možemo reći da počinje s već spomenutim radom Gibson i sur. (2006), pa sve do danas možemo izdvojiti dvije glavne metode za identifikaciju ROH segmenata: algoritmi opservacijskog prebrojavanja genotipa (*engl. Observational genotype-counting algorithms*) te algoritmi temeljeni na modelu (*engl. Model-based algorithms*). (Ferenčaković, 2019.)

2.2.1.1. “Observational genotype counting”

Kod metoda opservacijskog prebrojavanja razlikujemo „klizeći prozor“ (*engl. Sliding window*) te prebrojavanje. (Ferenčaković, 2019.)

Primjer metode koja koristi sliding window možemo naći unutar programa PLINK (Purcell i sur., 2007.). PLINK je jedan od najčešće korištenih programa za analizu ROH segmenata u ljudskoj i animalnoj populaciji. (Meyermans i sur., 2020.) Generalno, u sliding window metodi prozor određene duljine pomiče se preko podataka, korak po korak te se dobivaju vrijednosti prema podatcima u prozoru. Tako PLINK opisuje ROH na temelju potrebnog broja homozigotnih SNP-ova koji se nalaze u određenom segmentu koristeći metodu kliznog prozora. Prvo se pomiče prozor duž genoma prema broju SNP-ova koje određuje korisnik i odlučuje ako je svaki prozor u skladu sa postavljenim parametrima ili ne (Karimi, 2013.). Parametri uključuju maksimalni broj dopuštenih heterozigota u prozoru da bi se taj prozor smatrao homozigotnim i dopušten broj nedostajućih SNP-ova. Ako broj nedostajućih SNP-ova prelazi postavljen prag, prozor se neće smatrati homozigotnim. Finalno, provjerava se konačna duljina segmenata koji su određeni kao homozigotni, moraju zadovoljiti minimalne vrijednosti postavljene za broj SNP-ova kao i duljinu u Kb da bi se smatrali ROH segmentom. (Meyermans i sur., 2020.)

Primjeri metode prebrojavanja implementirani su u programima SVS (Golden Helix SNP & Variation Suite; SVS v.7.6.8) i cgaTOH (Zhang i sur., 2013.). (Ferenčaković, 2019.)

ROH modul SVS softvera predstavlja algoritam koji kontinuirano kroz cijeli kromosom, ispituje podudaranje svakog mogućeg niza s ulaznim parametrima koje je odredio korisnik. Parametri uključuju minimalni broj SNP-ova u ROH-u, minimalnu duljinu ROH segmenta, minimalnu gustoću, najveću udaljenost te maksimalni broj heterozigota i nedostajućih alela. Algoritam strogo primjenjuje ograničenje na maksimalni dozvoljeni broj heterozigota i nedostajućih alela. Svaki homozigotni SNP smatra potencijalnim početkom novog ROH segmenta. (Curik i sur., 2014.)

Program cgaTOH svaki SNP analizira zasebno te zatim zajedno. Na taj način otkriva TOH (*engl. Tracts of Homozygosity*). Također ima dodatne značajke za klasifikaciju segmenata, kao npr. podudaranje po alelima. SNP-ovi su jedna on

najčešćih varijacija u genomu i mogu se različito prikazati u ljudskim osobinama i osjetljivosti na bolesti među i unutar populacija. SNP-ovi mogu funkcionirati pojedinačno, ali su često u koordinaciji sa drugim SNP-ovima da bi se manifestirala neka bolest što ukazuje na to da se velik dio ljudske raznolikosti može objasniti većim strukturnim razlikama pojedinih genoma koje su mnogo veće od razlika pojedinačnih nukleotida.

U svrhu potpunog iskorištavanja značajki TOH-a te olakšavanja budućih statističkih analiza uvodi se surogat-TOH odnosno regija koja zahvaća skupinu TOH-a sa specifičnim karakteristikama.

cgaTOH softver razvijen je u formatu naredbenog retka i ima mogućnost interaktivne vizualizacije TOH outputa i surogat-TOH-a na razini kromosoma ili na individualnoj razini za pojedinačni surogat-TOH. Također ima mogućnost prikaza NCBI genomskih mapa. cgaTOH je algoritam za identifikaciju različitih skupova SNP-ova koji prikazuju homozigotnost od individualnog TOH-a sve do surogat-TOH-a što omogućuje promatranje raznih aspekata više karakteristika TOH-a i detekciju proširenih homozigotnih segmenata. (Zhang i sur., 2013.)

Karimi 2013. godine provodi istraživanje u kojem je, između ostalog, usporedila rezultate detekcije ROH otoka koristeći 3 različita softvera: PLINK, SVS i cgaTOH. Na temelju dobivenih rezultata ROH otoci detektirani trima softverima uvijek su bili na sličnim mjestima, s neznatnom razlikom u frekvenciji signala. Softver cgaTOH je detektirao veći broj individua u ROH hotspotovima, dok su se rezultati SVS-a i PLINK-a više preklapali. (Karimi, 2013.)

2.2.1.2. “Model based algorithms”

Kod metoda baziranih na modelu razlikujemo Skrivene Markovljeve modele (HMM) koji uzimaju u obzir pozadinske razine LD-a, homozigotno heterogene HMM te HMM Viterbi. (Ferenčaković, 2019.)

Skriveni Markovljevi modeli (*engl. Hidden Markov models;* HMM) ime su dobili po ruskom matematičaru Andreju Andrejeviču Markovu koji je razvio veći dio relevantne statističke teorije. HMM su predstavljeni i proučavani početkom 1970-ih te su primarno bili korišteni u prepoznavanju govora. Krajem 1980-ih uspješno su primijenjeni u analizi bioloških sekvenci. Skriveni Markovljevi modeli su statistički

modeli za prepoznavanje skrivenih informacija iz uočljivih uzastopnih simbola kao što je i nukleotidni niz. Imaju mnoge primjene u analizi sekvenci, posebno za predviđanje egzona (sekvence nukleotida koju kodira gen) i introna (sekvencu nukleotida koja nije kodirajuća), poravnanje dvije sekvence i identifikaciju proteinske domene. HMM imaju široku primjenu u molekularnoj biologiji od profiliranja sekvenci i višestrukog poravnanja proteina do filogenetike. HMM-i su vrsta strojnog učenja koji su sa velikim uspjehom korišteni za dobivanje uvida u „sakrivene“ parametre koji se nalaze u osnovnim podatcima. U slučaju podataka genomskih sekvenci ti sakriveni parametri mogu biti brzina evolucije ili vjerojatnost pojave nukleotida na određenom mjestu. HMM-ovi se mogu proširiti za korištenje distribucije frekvencije alela za mapiranje varijacija temeljnih genetskih parametara u nekom segmentu DNA sekvence ili u cijelom genomu. (Kern i sur., 2010.)

Skriveni Markovljevi modeli se mogu koristiti za opis razvoja opaženih događaja koji ovise o unutarnjim, sakrivenim, čimbenicima koji se ne mogu izravno promatrati. Opaženi događaj nazivamo simbolom a „sakriveni“ čimbenik stanjem. HMM se stoga sastoji od dva stohastička procesa, vidljivog procesa uočljivih simbola te nevidljivog procesa skrivenih stanja zbog čega se naziva i dvostruko-ugrađen stohastički proces (Rabiner, 1989.). Ovakav je pristup također koristan kod modeliranja bioloških sekvenci (DNA, RNA, proteini…). Obično se biološki slijed sastoji od manjih pod struktura sa različitim funkcijama, a različite funkcionalne regije često pokazuju različita statistička svojstva. (Byung-Jun, 2009.)

U pristupima temeljenim na HMM-u općenito se pretpostavlja da je duljina HBD (*engl. Homozygous by descent;* HBD) odnosno segmenata homozigotnih po podrijetlu eksponencijalno distribuirana. Modeliranje jedne eksponencijalne distribucije pretpostavlja da je sva autozigotnost povezana sa predcima prisutnim u istim generacijama. Za populacije s kompleksnom populacijskom povijesti ta pretpostavka može biti previše restriktivna pa su Druet i Gautier (2017.) predložili korištenje različitih eksponencijalnih distribucija za modeliranje grupa HBD segmenata različitih očekivanih duljina. U takvom pristupu grupe HBD-a mogu se promatrati kao skupine predaka prisutnim u različitim generacijama u prošlosti. Takav model bolje objašnjava složene demografske povijesti u kojima različiti predci iz različitih generacija mogu pridonijeti autozigotnosti. Model s više grupa HBD-a također pruža uvid u demografsku

povijest populacije te procjenjuje relativni doprinos prošlih generacija trenutnim razinama inbreedinga. (Druet i Gaurtier, 2021.)

Međutim, kada je doprinos predaka izuzetno visok, model više grupa ima tendenciju podcjenjivati starost segmenata HBD-a prebacivanjem podjele HBD-a prema mlađim grupama iako su ukupne procijenjene razine inbreedinga ostale iste. Iz tog razloga su Druet i Gaurtier (2021.) implementirali ažuriran model više HBD grupa u kojem je HBD grupiranje modelirano u uzastopnim ugniježdenim razinama pri čemu svaka razina odgovara jednom modelu HBD grupe. Prema tom modelu grupe koje nisu HBD su sada modelirane kao mješavina HBD segmenata iz kasnijih generacija te kraćih ne HBD segmenata. Također su u radu dokazali bolja statistička svojstva te bolji performans nove verzije modela sa obzirom na prethodnu verziju.

Primjer za HMM koji uzimaju u obzir pozadinske razine LD-a je BEAGLE. BEAGLE je program za faziranje i imputaciju nedostajućih genotipa. Nasumično nedostajući genotipovi su imputirani tokom faziranja. Ako se koristi referentni panel faziranih genotipova, negenotipizirani markeri u promatranim uzorcima će se imputirati ako su prisutni u referentnom panelu. (Browning, 2018.)

Program Beagle (Browning i Browning, 2018.) koristi HMM koji uključuje LD između SNP-ova i vjerojatnosti haplotipova iz cijelog uzorka kod identifikacije ROH segmenata. Dvije vjerojatnosti koje definira korisnik postavljaju osnovna očekivanja detekcije autozigotnog segmenta u cM SNP podataka. Stopa tranzicije iz HBD u ne HBD je vjerojatnost po cM da SNP koji nije autozigotan postane autozigotan. HBD označava homozigotan po podrijetlu te je konceptualno identičan pojmu autozigotnost. Rezultati sa nižim vrijednostima se tumače da će HBD segmenti biti kraći te obrnuto. Beagle daje matricu vjerojatnosti da je svaki SNP dio autozigotnog segmenta. (Howrigan i sur., 2011.)

Pristup sa HMM također se koristi za podatke WES-a (*engl. Whole exome sequencing;* WES) kao alternativa za otkrivanje varijanti SNP-ova i ROH segmenata male do srednje duljine (Zhuang i sur., 2012.; Mezzavilla i sur., 2015.; Ceballos i sur., 2017.). Međutim, WES analizama nije moguće dobivanje dugih ROH segmenata (Ceballos i sur., 2017.). Iz tog razloga razvijen je specifičan softver H^3M^2 koji je i jedan od primjera za homozigotno heterogene HMM. H^3M^2 je računalni pristup za

identifikaciju ROH segmenata. Algoritam se temelji na HMM koji uključujući udaljenosti između uzastopnih polimorfnih položaja u matricu prijelaznih vjerojatnosti, također može detektirati ROH svake genomske veličine. Ključna značajka algoritma je heterogenost što ga čini vrlo prikladnim za WES podatke. Na temelju rezultata dobivenih u radu, Magi i sur. (2014.) zaključuju da H^3M^2 ima bolji performanse od programa PLINK i GERMLINE na testnim podatcima te naglašuju da je algoritam manje osjetljiv na specifikaciju parametara što osigurava da odabrana konfiguracija parametara nema velik utjecaj na rezultate analize kao u drugim programima. (Magi i sur., 2014.)

Primjer HMM Viterbi algoritma je softver RZooROH. Autozigotni ili segmenti homozigotni po podrijetlu u pojedinim genomima nastaju nasljeđivanjem dvije kopije jednog kromosomskog fragmenta pretka. Zbog rekombinacije, duljina HBD segmenta obrnuto je povezana sa brojem generacija koje povezuju te dvije kopije kromosomskog fragmenta. Razina inbreedinga pojedinca izravno određuje udio njegovog genoma koji je HBD. Identifikacija HBD segmenta ključna je za mnoge primjene u kvantitativnoj i populacijskoj genetici. (Druet i sur., 2019.) Podjela individualnih organizama u HBD i ne-HBD segmente postala je veoma popularna zadnjih godina upravo zbog širokog spektra primjene. (Ceballos i sur., 2018.) Metode bi trebale obrađivati podatke o genotipiziranju i sekvenciranju koji su dostupni i u HBD i ne-HBD segmentima. Osim toga metode bi trebale uzeti u obzir i složene demografske povijesti, frekvenciju alela markera, genetske udaljenosti, vjerojatnosti pogrešaka kod genotipizacije i količinu informacija. (Druet i sur., 2019.)

U tu svrhu Druet i sur. (2019.) razvili su R paket, RzooROH, koji primjenjuje pristup temeljen na HMM-u za skeniranje pojedinačnih genoma za HBD segmente. Metodu su intenzivno testirali na stimuliranom sklopu podataka sa širokim spektrom čimbenika kao što su gustoća markera i stopa pogrešaka te ju usporedili sa drugim metodama. (Druet i sur., 2019.)

Paket radi sa različitim tipovima podataka koji su dobiveni različitim tehnologijama i omogućuje istraživanje i usporedbu različitih specifikacija modela. Paket se oslanja na novi postupak optimizacije proveden u kombinaciji sa ponovnim postavljanjem parametara koji poboljšava procjenu parametara kao i ubrzava izračune.

Paket također nudi nekoliko grafičkih alata za lakše tumačenje rezultata. (Druet i sur., 2019.)

2.1.2. ROH inbreeding koeficijent

McQuillan i sur. 2008. godine su predložili metodu izračuna inbreeding koeficijenta iz podataka ROH-a. F_{ROH} inbreeding koeficijent dobiva se prema formuli

$$F_{ROH} = \sum L_{ROH} / L_{AUTOSOM}$$

Gdje je $\sum L_{ROH}$ ukupna duljina svih ROH-a genoma jedinke iznad određene minimalne duljine a $L_{AUTOSOM}$ predstavlja ukupnu duljinu autosomalnog genoma pokrivenog SNP-ovima. (McQuillan i sur., 2008.)

F_{ROH} ne uzima u izračun ROH na spolnim kromosomima ženskih jedinki pošto one imaju drugačiju distribuciju IBD alela kao ni regije oko centromera pošto su to dulji segmenti genoma bez prisustva SNP-ova te bi mogli uzrokovati pogrešne rezultate. (Curik i sur., 2014.)

ROH inbreeding koeficijent se može podijeliti u vrijednosti za individualne kromosome (na primjer F_{ROH_Ch1} za ROH inbreeding koeficijent prvog kromosoma, F_{ROH_Ch2} za ROH inbreeding koeficijent drugog kromosoma, F_{ROH_Chn} za ROH koeficijent n-tog kromosoma gdje n predstavlja broj kromosoma) ili za određene segmente kromosoma (McQuillan i sur., 2008.). Još jedna prednost F_{ROH} inbreeding koeficijenta je da je referentna populacija poznata i bazirana je na očekivanju da će dvije srodne jedinke dijeliti identične kromosomske segmente određene duljine uz pretpostavku da su IBD. (Curik i sur., 2014.)

Problem kod izračuna F_{ROH} proizlazi kod određivanja broja generacija od zajedničkog pretka što zahtijeva analizu distribucije broja i duljine zajedničkih IBD haplotipova kao funkciju broja generacija do referentne populacije. Ovaj problem se može izbjeći pretpostavkom da očekivana duljina IBD haplotipa ($L_{IBD\text{-}H|gcA}$) prati eksponencijalnu distribuciju čija je srednja vrijednost jednaka 100/(2 gcA) cM gdje je gcA broj generacija od zajedničkog pretka odnosno duljine ROH-a od 16.6, 10.0, 5.0 te 2.5 Mb od zajedničkog pretka prije 3 generacije, 5 generacija, 10 generacija te 20 generacija uz pretpostavku da je 1 cM ≈ 1Mb kako objašnjava Ferenčaković u svom radu (2015.). Izračun je dobiven prema formuli

$$E(L_{IBD\text{-}H}|gcA)=100/(2\ gcA)$$

Pretpostavka da je 1cM ≈ 1Mb je česta, ali odnos između stopa rekombinacije i fizičke udaljenosti varira od vrste do vrste te od kromosoma do kromosoma. Tako je precizniji odnos od 1cM ≈ 0.76 Mb određen na temelju analize 4 populacije svinja (Herrero-Medrano i sur., 2013.) prema čemu bi duljine ROH-a iznosile 21.93, 13.16, 6.58 te 3.29 Mb od zajedničkog pretka prije 3 generacije, 5 generacija, 10 generacija te 20 generacija. Također je određen odnos od 1cM ≈ 1.28 Mb za 29 autosoma kod goveda (Arias i sur., 2009.) prema čemu bi duljine ROH-a iznosile 13.0, 7.8, 3.9, te 1.95 Mb od zajedničkog pretka prije 3, 5, 10 te 20 generacija. (Curik i sur., 2014.)

2.1.3 Procjena ROH segmenata na temelju NGS podataka

Iako je korištenje genetskih markera i dalje najčešće korištena metoda, podatci dobiveni NGS-om (*engl. New generation sequencing;* NGS) rapidno postaju sve popularniji za provedbu analiza. Iako korištenje NGS podataka raste, komparativne studije među navedene dvije metode su oskudne te je znanje o potencijalnim učincima promjene metodologije ograničeno (Jensen i sur., 2021.) WGS pristup (engl. Whole genome sequencing; WGS) analizira svaku varijantu čime se omogućuje pristup genotipizaciji svih baza i za svakog pojedinca se može dobiti više od nekoliko milijuna varijanti.

Sve veća popularnost sekvenciranih podataka daje bolju rezoluciju čak i najkraćeg ROH segmenta. Međutim, stope grešaka u genotipu su mnogo veće nego za podatke SNP arraya što osobito vrijedi za podatke s nižom pokrivenosti koji se, zbog smanjenja troškova, često koriste da bi se mogao povećati broj jedinki u istraživanju. Također sljedovi cijelih egzoma predstavljaju dodatan izazov s obzirom na njihovu veličinu i rijetkost u genomu. Još jedan od problema kod analize sekvenciranih podataka je i nedostatak adekvatnih softvera za procesiranje istih. (Ceballos i sur., 2017.)

WGS s visokom pokrivenošću bio bi najbolja opcija za proučavanje ROH-a, međutim, postoje dvije glavne prepreke za korištenje WGS-a s visokom pokrivenošću. Prvo, nedostatak podataka WGS-a visoke pokrivenosti za populacijske studije te drugo, veliki financijski i računalni trošak za takvu vrstu analize podataka. Suprotno od WGS-a visoke pokrivenosti, podatci WGS-a niske pokrivenosti su dostupniji,

pristupačniji te procesi dobivanja takvih podataka su računalno manje intenzivni no problem je visoka stopa pogrešaka. (Ceballos i sur., 2018.)

Ceballos i sur. (2018.) proveli su istraživanje u kojem su demonstrirali da je moguće postići ekvivalentne rezultate između WGS-a niske pokrivenosti i tehnologije SNP arraya ako se dopuste 3 heterozigotna SNP-a pri radu sa WGS-om niske pokrivenosti. (Ceballos i sur., 2018.)

Dostupnost NGS podataka omogućuje proučavanje genetskih varijanti u pojedinačnim parovima baza bez pristranosti. Međutim, broj pojedinaca sa sekvenciranim podatcima kao što je „1000 bull genome project“ (Daetwyler i sur., 2014.) je još uvijek prilično nizak. (Zhang, 2017.)

Broj individua s većom gustoćom genotipa može se znatno povećati imputacijom koja zaključuje nedostajuće genotipove u markerima niže gustoće prema podatcima iz markera veće gustoće (Ma i sur., 2013., Brondum i sur., 2014.).

Imputacija se oslanja na neravnotežu vezanih gena (*engl. Linkage disequilibrium;* LD) između markera. LD je definiran kao ne slučajna povezanost alela između različitih lokusa u danoj populaciji. (Marchini i sur., 2010.)

Prema Zhang (Zhang, 2017.) kod goveda se varijante niza cijelog genoma obično imputiraju u 2 koraka, na primjer od 50k markera do HD markera te od imputiranih HD markera do varijanti sekvence cijelog genoma. (Brondum i sur., 2014.; Zhang, 2017.) Proces imputacije za učestale varijante je obično vrlo točne rezultate dok je za rijetke varijante to nije slučaj zbog niskog LD-a između uglavnom uobičajenih markera na SNP čipu i rijetkih varijanti na sekvenci. Točnost imputiranja rijetkih varijanti mogla bi se poboljšati za jedinke sa sekvenciranim srodnicima ako se podatci pedigrea koriste za dobivanje informacija za proces imputacije. Kombiniranje sekvenci iz više populacija također može poboljšati preciznost imputacije. Također, određeni softveri imaju bolju izvedbu kod imputacije rijetkih varijanti, no točnost je i dalje relativno niska s obzirom na uobičajene varijante. Iz navedenih razloga točna imputacija rijetkih varijanti je i dalje izazov. (Brondum i sur., 2014.; Zhang, 2017.)

Zaključno, pristup identifikacije ROH segmenata iz WGS-a analizira svaku varijantu tako da sve dostupne baze mogu biti genotipizirane te više od nekoliko milijuna varijanti, od najčešćih do najrjeđih, mogu biti dobivene za svakog pojedinca.

(Nielsen i sur., 2014.; Goodwin i sur., 2016.). Zbog troškova, sekvenciranje niske pokrivenosti se često koristi u svrhu maksimizacije broja sudionika u istraživanju. U tom slučaju rijetke varijante SNP-ova dobivaju se rjeđe, sa većom stopom pogrešaka od češćih varijanti. Stopa pogrešaka za WGS niske pokrivenosti znatno je veća od stope pogrešaka analize iz SNP arraya što dovodi do nepreciznosti kod identifikacije ROH segmenata. Međutim, kako troškovi WGS analiza postaju sve pristupačniji i podatci dostupni otvaraju se nove mogućnosti za detaljno proučavanje ROH-a, replikaciju rezultata dobivenih iz SNP array analiza te za proučavanje odnosa ROH-a u novim populacijama ili svojstvima. (Ceballos i sur., 2017.)

Tablica 1. Sažeti prikaz metoda procjene ROH segmenata

Identifikacija ROH segmenata		
Metoda	Prednosti	Nedostatci
Opservacijsko prebrojavanje genotipa	Mogućnost identifikacije kratkih i srednjih ROH segmenata iz WES analiza Količina podataka u skladu sa računalnim mogućnostima sadašnjice	Nemogućnost identifikacije dugih ROH segmenata iz WES analiza Niža rezolucija od metoda procjene iz sekvenci Performans ovisi o kvaliteti i gustoći SNP podataka
Metode bazirane na modelu	Mogućnost identifikacije kratkih, srednjih i dugih ROH segmenata iz WES analiza (npr. H^3M^2) Bolji performans od metoda prebrojavanja kada je manja količina podataka te kada su podatci varijabilniji	Računalno i vremenski zahtjevno Niža rezolucija od metoda procjene iz sekvenci Performans ovisi o kvaliteti i gustoći SNP podataka
Metode procjene iz sekvence	Bolja rezolucija Procjene iz WGS analiza visoke pokrivenosti je najbolja opcija za proučavanje ROH-a Postoji mogućnost postizanja istih rezultata WGS-a niske pokrivenosti i analiza preko SNP podataka	Visoke stope grešaka (osobito za WGS niske pokrivenosti) Nedostatak adekvatnih softvera Velika količina podataka, računalno zahtjevno za pohranu i analizu

3. Problemi s genomskim podatcima

Sekvenciranje cijelog genoma (WGS) i sekvenciranje cijelog egzona (WES) široko su prihvaćeni u istraživanjima, a u skorije vrijeme i u kliničkoj praksi. (Birney, 2019.). Generirani WGS i WES podatci uključuju ogromne količine informacija od potencijalne važnosti za sadašnje i buduće zdravlje pojedinca, s implikacijama za članove obitelji, ako se mogu prevladati analitičke i interpretacijske prepreke. Široka dostupnost genomskih podataka također nudi mogućnost ponovne uporabe u dodatne istraživačke, zdravstvene ili kliničke svrhe. (Narayanasamy i sur., 2020.)

Od 1990-ih godina cijeli genomi različitih vrsta bili su sekvencirani u različitim projektima. Institut za Genomska Istraživanja je 1955. godine sekvencirao prvi organizam, bakteriju *Haemophilus influenzae*, a 1996. godine potpuno je sekvenciran prvi eukariotski genom, kvasčeva gljivica *Saccharomyces cerevisiae*. Prvi biljni genom je bio sekvenciran 2000. godine, biljka Talijin uročnjak (*lat. Arabidopsis thaliana*) (https://www.nature.com/articles/35048692), a 2003. „Human Genome Project“ (HGP) je završen. (https://www.yourgenome.org/facts/timeline-organisms-that-have-had-their-genomes-sequenced)

Cijena potrebna za sekvenciranje cijelog genoma postupno se smanjivala sve do 2007. godine, kada se sa pojavom NGS tehnologija drastično snizila. Tako se danas cijeli genom može sekvencirati u nekoliko dana, ako ne i sati, za cijenu manju od 1000 američkih dolara, dok bi taj isti proces 2001. godine trajao puno duže i koštao preko 100 milijuna američkih dolara. (Via, 2017.)

Unatoč dobrim stranama tehnologija genotipizacije i masovnog sekvenciranja, zajedno s znanstvenim dostignućima koja omogućuju razmjenu i javni pristup genomskim skupova podataka, postoje i određeni rizici i izazovi koje treba uzeti u obzir. (Via, 2017.) Neki od tih izazova su tehnički zahtjevi i privatnost.

3.1 Tehnički zahtjevi

„Veliki podatci“ (*engl. Big Data*) je slabo definiran pojam koji se primjenjuje na masovne skupove podataka koje je teško obraditi sa standardnim metodama statističkih analiza i upravljanja bazama podataka. (Via, 2017.)

Izazov pri radu s genomskim podatcima dolazi od količine generiranih podataka. Ti podatci su veoma veliki, te kao i drugi veliki podatci zahtijevaju skladištenje na siguran način te alate za učinkovit pristup, analizu i dijeljenje. Procjenjuje se da će do 2025. godine genomika biti najzahtjevnija, ili usporediva sa najzahtjevnijim, u prikupljanju, pohrani, distribuciji i analizi podataka (Via, 2017.). Kako jedan ljudski genom zauzima oko 100 gigabajta prostora za pohranu , a sve je više genoma sekvencirano, potrebe za skladištenjem rastu od gigabajta, do petabajta pa sve do exabajta. Do 2025. će samo za ljudske genomske podatke biti potrebno 40 exabajta skladišnog kapaciteta. (https://medicalfuturist.com/the-genomic-data-challenges-of-the-future/)

Sve veće potrebe za prostorom za pohranu predstavljaju problem iz razloga što su stotine tisuća ljudskih genoma već pohranjene, a najveći centri za sekvenciranje već koriste više od 100 petabajta prostora za pohranu. Pohranjivanje i dijeljenje genomskih podataka također se suočava s istom vrstom potencijalnih opasnosti kao i sve informacije na umreženom poslužitelju, kao što su rušenje servera, gubitak podataka, prekid napajanja, pokušaji hakiranja i brzina prijenosa. Brzina prijenosa posebno je relevantna s podatcima te veličine. Distribucija putem interneta iz središnjih prostora za pohranu može biti veoma spora, osobito kada je riječ o podatcima čija veličina može iznositi i terabajt. U lokalnom kontekstu, ponekad je brže prenijeti podatke s vanjskog pogona za pohranu, no problem s time predstavlja dostupnost, cijena te količina podataka koja se može pohraniti na takvom vanjskom pogonu. (Via, 2017.)

Samo sekvenciranje često nije dovoljno ako se svaki genom ne analizira temeljito da bi se postigli smisleni znanstveni rezultati. Takva analiza genomskih podataka obično generira dodatnih 100 gigabajta podataka po genomu i zahtjeva ogromnu računalnu snagu koju većina osobnih računala ne može pratiti. (Labiotech, 2020.)

Znanstvenici i istraživači koji rade sa velikim skupovima genomskih podataka kontinuirano traže rješenja za trenutne probleme. Dok računala visokih performansa (*engl. High-performance computer;* HPC) većinom mogu podnijeti veliku količinu podataka i njihovu analizu, ekonomski za mnoge nisu pristupačni. (Labiotech, 2020.)

Iz tih razloga u tijeku je provođenje nekoliko strategija za smanjenje podataka generiranih NGS tehnologijama kao što su poboljšanje kompresije podataka i točnosti sekvenciranja. Smanjenje veličine podataka potencijalno bi olakšalo pohranu i dijeljenje. Međutim, takvi će pristupi samo usporiti tempo rasta računalnih potreba.

Alternativa koja postaje sve popularnija je korištenje oblaka (*engl. Cloud*) za spremanje podataka. Usluge cloud-a rješavaju dio problema u pogledu kapaciteta spremanja podataka i računalne snage. (Chow-White i sur., 2015.)

Budući da se mnoge analize genomskih podataka pohranjenih u cloud-u mogu izvesti na daljinu, računalni resursi su optimizirani. Međutim, ti podatci donose i različite etičke izazove. S pravne strane, cloud poslužitelj i genomski podatci pohranjeni u cloud-u mogu se fizički nalaziti na drugom geografskom mjestu, odnosno pod drugačijim regulatornim okvirom, od mjesta samog istraživanja. Također, potrebno je uložiti dodatne sigurnosne napore da bi se osigurala privatnost pojedinca. Oba navedena primjera su uobičajeni problemi pri korištenju usluga upravljana podatcima trećih strana. (Via, 2017.)

Postoji hitna potreba za razvojem računalnih sposobnosti i kapacitetima pohrane. Promjene u bioinformatici i biotehnologiji do kojih su dovele NGS tehnologije su veoma velike. Sekvenciranje i analiza trebali bi biti prioritet, ali zbog tehničkih problema (pohrana, analiza, sigurnost) vrijeme utrošeno na rješavanje istih često je duže je od onog posvećenog prikupljanju i analizi podataka. (Papageorgiu i sur., 2018.)

3.2 Privatnost

Genetski i genomski podatci neke osobe predstavljaju najprivatnije podatke o prošlosti, sadašnjosti i budućnosti pojedinca. S obzirom da dijelimo 50% genoma sa svakim od roditelja te braće i sestara, a 25% sa djedovima, bakama, ujacima, tetama itd. genom ima potencijal otkriti osjetljive podatke i o članovima obitelji. Svako kršenje privatnost podataka na temelju genoma o trenutnom ili budućem zdravstvenom stanju pojedinca potencijalno utječe i na druge članove obitelji. (Chow-White i sur., 2015.; Via, 2017.). Iako su zaštita i povjerljivost ovakvih podataka od iznimne važnosti, složenije je nego sa drugim vrstama podataka. (https://medicalfuturist.com/the-genomic-data-challenges-of-the-future/)

Zaštita privatnosti pojedinca klasična je briga u biomedicinskim istraživanjima. U istraživanjima te kliničkim testiranjima osiguravanje anonimnosti te informirani pristanak smatraju se tradicionalnim zaštitnim mjerama privatnosti. (Chow-White i sur., 2015.). Uz te zaštitne mjere osobni se podatci obično pohranjuju na mjestima sa ograničenim pristupom. Sami prijelaz sa arhiva na digitalne baze podataka donio je mogućnost za analizu velike količine podataka, ali samim time i određeni rizik (Via, 2017.). Neki od tih rizika su nelegalni pristup tim podatcima te prodaja podataka od strane vlasnika.

Međutim, kako bi proveli nekoliko vrsta genomskih izračuna, osobito za fenotipske asocijacije kao GWAS, znanstvenici mogu postići bolju snagu i jačinu signala korištenjem većeg broja podataka, npr. genoma. Stoga dijeljenje i grupiranje velikih količina podataka može rezultirati velikim korisnim informacijama za neku skupinu, iako je privatnost pojedinca ugrožena. (Navarro i sur., 2019.)

Tokom biomedicinskog istraživanja, znanstvenici se mogu susresti s neočekivanim nalazima koji imaju potencijalnu kliničku važnost. Takvi slučajni nalazi (*engl. Incidental findings;* IF) su se povećali s NGS-om. Dok u kliničkom području postoje protokoli kako se nositi i rukovati takvim podatcima, u području genomskih istraživanja nije definirano kako postupati sa istima. To dovodi do problema zato što u nedostatku kliničkih simptoma ili obiteljske povijesti za određenu bolest, istraživanje genoma u potrazi za potencijalno patogenim varijantama može dovesti do lažno pozitivnih rezultata i prekomjerne dijagnoze. (Westbrook i sur., 2013.)

Iako je strogo zakonodavstvo za zaštitnu privatnosti uobičajena praksa u većini zemalja, vrlo mali broj njih donio je sveobuhvatnu politiku za zaštitu i reguliranje genomskih podataka. Tako u SAD-u GINA (*engl. Genetic Information Nondiscrimination Act;* GINA) izričito zabranjuje diskriminaciju temeljenu na genomskim informacijama pri zdravstvenom osiguranju i zapošljavanju. U Europskoj Uniji situacija sa zaštitom genomskih podataka često varira od države do države, ali direktiva EU o zaštiti podataka regulira zaštitu svih podataka, uključujući i zdravstvene. (Via, 2017.)

Privatnost je aspekt većeg pitanja vlasništva nad podatcima i kontrole podataka. Iako se obično misli da pojedinac ili pacijent posjeduju svoje osobne podatke, protivni

trend u nekim biomedicinskim istraživanjima je da ih posjeduje istraživač koji generira skup podataka (Longo i sur., 2016.). Također postoji i notacija da ljudski podatci, osobito zdravstveni, imaju očito medicinsku i komercijalnu vrijednost pa kompanije i nacije često traže kontrolu i vlasništvo nad takvim skupovima podataka. (Navarro i sur., 2019.)

Zaključak

1. Procjena inbreeding koeficijenta točnija je i preciznija iz genomskih podataka nego iz pedigre podataka.

2. Zbog važnosti inbreedinga i dalje se razvijaju nove metode i programi za procjenu istog na temelju dostupnih podataka. Uporaba molekularnih markera dovela je do razvoja niza metoda za procjenu inbreeding koeficijenta

3. Identifikacija ROH segmenata, pa samim time procjena $F_{ROH,}$ iz SNP podataka je i dalje najčešća metoda zbog mnogih benefita koje ima nad ostalim metodama.

4. Pojava NGS tehnologija omogućila je bolju rezoluciju i veću količinu informacija za analizu, međutim iskazali su se i problemi poput cijene, mogućnosti pohrane velikih količina podataka, nedostatak adekvatnih programa za analizu podataka te pitanje privatnosti.

5. S obzirom na sve veću važnost genomike, realno je očekivati razrješenje navedenih problema u skorijoj budućnosti. Međutim, upitno je ako će razvoj rješenja moći pratiti brzi razvoj metoda

Literatura

1 Aguliar, R., Quesada, M., Ashworth, L., Herrarias-Diego, Y., Lobo, J. (2008). Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. Molecular Ecology, 17: 5177-5188. https://doi.org/10.1111/j.1365-294X.2008.03971.x

2 Alam, M. F., Khan, M. R., Nuruzzman, M., Parvez, S., Swaraz, A. M., Alam, I., et al. (2004). Genetic basis of heterosis and inbreeding depression in rice (Oryza sativa L). Journal of Zhejiang University SCIENCE, 5:406-411.

3 Alvarez, G., Ceballos, F. C., Quinteiro, C., (2009). The role of inbreeding in the extinction of a European royal dynasty. PLoS One. 4, e5174.

4 Asuka, Y., Tomaru, N., Munehara, Y.,Tani, N., Tsumura, Y., Yamamoto, S. (2005). Half-sib family structure of Fagus crenata saplings in an old-growth beech–dwarf bamboo forest, Molecular Ecology, vol. 14, 2565-2575.

5 Ballou, J. (1983). Calculating Inbreeding Coefficients from Pedigrees. Genetics and Conservation, Benjamin/Cummings Pub. Co.

6 Ballou, J., (1997). Ancestral inbreeding only minimally affects inbreeding depression in mammalian populations. Journal of Heredity 88, 169 - 178.

7 Balloux, F., Amos, W., Coulson, T., (2004). Does heterozygosity estimate inbreeding in real populations? Molecular Ecology 13, 3021 - 3031.

8 Bellis MA, Hughes K, Hughes S, Ashton JR. (2005). Measuring paternal discrepancy and its public health consequences. Journal of Epidemiology and Community Health, 59(9):749-754.

9 Benjelloun, B., Boyer, F., Streeter, I., Zamani, W., Engelen, S., et al. (2019). An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. Molecular Ecology Resources 1755–0998.13070. 10.1111/1755-0998.13070.

10 Bereskin, B., Shelby, C.E., Hazel, L.N., (1969). Monte Carlo Studies of Selection and Inbreeding in Swine I. Genetic and Phenotypic Trends. Journal of Animal Science, 29, 678-686.

11 Bereskin, B., Shelby, C.E., Hazel, L.N., (1970). Monte Carlo Studies of Selection and Inbreeding in Swine. II. Inbreeding Coefficients. Journal of Animal Science 30, 681-689.

12 Beynon S.E., Slavov G.T., Farre M. et al. (2015) Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. BMC Genetics 16, 65.

13 Birney, E. (2019). The Convergence of Research and Clinical Genomics. The American Journal of Human Genetics. 104. 781-783. 10.1016/j.ajhg.2019.04.003.

14 Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations, Trends Ecol Evol, vol. 18, 503-511.

15 Bosse, M., Megens, H. J., Madsen, O., Paudel, Y., Frantz, L., Schook, L., Crooijmans, R., Groenen, M., (2012). Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genetics 8, e1003100.

16 Broman, K., Weber, J. L., (1999.) Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. American Journal of Human Genetics. 65,1493–500.

17 Brondum, R. F., Guldbrandtsen, B., Sahana, G. *et al.,* (2014). Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics, 15, 728. https://doi.org/10.1186/1471-2164-15-728.

18 Browning, B. L., (2018). Beagle 5.0. University of Washington, Division of Medical Genetics.

19 Browning, S. R., Browning, B. L. (2010). High-resolution detection of identity by descent in unrelated individuals. American Journal of Human Genetics 86, 526–539. doi: 10.1016/j.ajhg.2010.02.021.

20 Byung-Jun, Y. (2009) Hidden Markov Models and their Applications in Biological Sequence Analysis. Current Genomics 10(6): 402-415. doi: 10.2174/138920209789177575. PMID: 20190955; PMCID: PMC2766791.

21 Ceballos F. C., Joshi P. K., Clark D. W., Ramsay M., Wilson J. F. (2018). Runs of homozygosity: windows into population history and trait architecture. National Review of Genetics, 19, 220–234. 10.1038/nrg.2017.109.

22 Charlesworth, D., Willis, J., (2009). The genetics of inbreeding depression. National Review of Genetics 10, 783 - 796.

23 Chow-White, P. A., MacAulay, M., Charters, A., & Chow, P. (2015). From the bench to the bedside in the big data age: ethics and practices of consent and privacy for clinical genomics and personalized medicine. Ethics and Information Technology, 17(3), 189-200.

24 Chybicki, I. J., Burczyk, J. (2009). Simultaneous Estimation of Null Alleles and Inbreeding Coefficients, Journal of Heredity, Volume 100, Issue 1,106–113. https://doi.org/10.1093/jhered/esn088.

25 Coltman, D.W., Bowen, W.D., Wright, J.M., (1998). Birth weight and neonatal survival of harbour seal pups are positively correlated with genetic variation measured by microsatellites. Proceedings: Biological Sciences, 265, 803-809.

26 Coltman. D. W., Slate, J. (2003), Microsatellite measures of inbreeding: a meta-analysis. Evolution, 57, 971-983.

27 Cortes, O., Eusebi, P., Dunner, S., Sevane N, Canon, J. (2019). Comparison of diversity parameters from SNP, microsatellites and pedigree records in the Lidia cattle breed. Livestock science, 219, 80-85.

28 Crow, J. F. (1954). Breeding structure of populations. II. Effective population number. Statistics and Mathematics in Biology. Chapter, 43, 543-556.

29 Curie-Cohen, M., (1981.). Estimates of inbreeding in a natural population: a comparison of sampling properties. Genetics. 100, 339-358.

30 Curik, I., Ferenčaković, M., Sölkner, J. (2014). Inbreeding and runs of homozygosity: A possible solution to an old problem. Livestock science, 166, 26-34.

31 Curik, I., Sölkner, J. and Stipic, N. (2002), Effects of models with finite loci, selection, dominance, epistasis and linkage on inbreeding coefficients based on pedigree and genotypic information. Journal of Animal Breeding and Genetics, 119: 101-115. https://doi.org/10.1046/j.1439-0388.2002.00329.x.

32 Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brandum, R.F., Liao, X., Djari, A., Rodriguez, S., Grohs, C., Jung, S., Esquerre, D., Gollnick, N., Rossignol, M., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P., Coote, D., Chamberlin, A., Van Tassell, C.P., Huggsle, I., Goddard, M.,

Guldbrandsten, B., Lund, M.S., Veerkamp, R., Boichard, D., Fries, R., Hayes, B.J. (2014). The 1000 bull genome project. Nature Genetics, 46(8), 858-865.

33 David, P. (1998). Heterozygosity-fitness correlations: new perspectives on old problems. Heredity, 80, 531-537.

34 Defaveria, J., Viitaniemi, H., Leder, E., Merilä, J., (2013). Charcaterizing genic and nongenic molecular markers: comparison of microsatellites and SNPs. Molecular Ecology Resources, 13, 377–392. https://doi.org/10.1111/1755-0998.12071.

35 DeWoody, Y. D., DeWoody, J. A. (2005). On the Estimation of Genome-wide Heterozygosity Using Molecular Markers, Journal of Heredity, Volume 96, Issue 2, 85–88. https://doi.org/10.1093/jhered/esi017.

36 Druet, T., Bertrand, A. R., Kadri, N. K. (2019). The RZooRoH package. (https://www.semanticscholar.org/paper/The-RZooRoH-package-Druet-Bertrand/05a7404784ca8be53c92fbf641d9a579e59046d8) .

37 Druet, T., Gautier, M. (2017), A whole-genome-based approach for estimation and characterization of individual inbreeding. https://onlinelibrary.wiley.com/doi/10.1111/mec.14324

38 Druet, T., Gautier, M. (2021). An improved hidden Markov model for the characterization of homozygous-by-descent segments in individual genomes. bioRxiv 445246; doi: https://doi.org/10.1101/2021.05.25.445246.

39 Falconer, D.S. (1989). Introduction to Quantitative Genetics. 3rd Edition, Longman Scientific and Technical, New York.

40 Falconer, D.S. (1996). Introduction to Quantitative Genetics. 4th Edition, Longman Scientific and Technical, New York.

41 Fang, Y., Hao, X., Xu, Z., Sun, H., Zhao, Q., Cao, R., Zhang, Z., Ma, P., Sun, Y., Qi, Z., Wei, Q., Wang, Q., & Pan, Y. (2021). Genome-Wide Detection of Runs of Homozygosity in Laiwu Pigs Revealed by Sequencing Data. Frontiers in genetics, 12, 629966. https://doi.org/10.3389/fgene.2021.629966.

42 Ferenčaković, M. (2015), Molecular dissection of inbreeding depression for semen quality traits in cattle. Doctoral thesis.

43 Ferenčaković, M. (2019). Inbreeding and Runs of Homozygosity: Hacks and tricks. International Symposium 27th Animal Science Days- ASD 2019 - book of Abstract. Prag, Češka, 2019. str. 1-1. (plenarno, međunarodna recenzija, sažetak, znanstveni) https://www.bib.irb.hr/1082181.

44 Ferenčaković, M., Hamzic, E., Gredler, B., Curik, I. & Solkner, J. (2011). Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. Agriculturae Conspectus Scientificus. 76, 325-328.

45 Ferenčaković, M., Hamzic, E., Gredler, B., Solberg, T. R., Klemetsdal, G., Curik I., Solkner J. (2013a). Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics. 130, 286-293.

46 Ferenčaković, M., Solkner, J., Curik, I. (2013b): Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics Selection Evolution. 45: 42.

47 Fernández, A., Toro, M., López-Fanjul. (1995) The effect of inbreeding on the redistribution of genetic variance of fecundity and viability in Tribolium castaneum. Heredity 75, 376–381. https://doi.org/10.1038/hdy.1995.149.

48 Fernández, J., Villanueva, B., Pong-Wong, R., Toro, M. A., (2005). Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics 170, 1313–1321. https://doi.org/10.1534/genetics.104.037325.

49 Gibson, J., Newton, E. M., Collins, A. (2006). Extended tracts ofhomozygosity in outbred human populations. Human Molecular Genetics. 15, 789-95.

50 Goldstein, H. (1995). Hierarchical Dana Modeling in the Social Sciences. Journal of educational and behavioral statistics, vol. 20, 201-204.

51 Goodwin, S., McPherson, J., McCombie, W. (2016). Coming of age: ten years of next-generation sequencing technologies. National Review of Genetics 17, 333–351. https://doi.org/10.1038/nrg.2016.49.

52 Guangul, S. A., (2014) Design of community based breeding programs for two indigenous goat breeds of Ethiopia. Doctoral thesis, University of Natural Resources and Life Sciences, Vienna.

53 Herrero-Medrano, J. M., Menges, H-J., Groenen, M., Ramis, G., Bosse, M., Perez-Enciso M. (2013). Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. BMC genetics, 14(1), 106.

54 Hill, W. G., Weir, B. S., (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research 93, 47–64.

55 Hillestad, B., Woolliams, J. A., Boison, S. A., Grove, H., Meuwissen, T., Vage, D. I., Klemetsdal, G. (2017). Detection of runs of homozygosity in Norwegian Red:

Density, criteria and genotyping quality control, Acta Agriculturae Scandinavica, SectionA–Animal Science, 67:34. 107116, DOI: 10.1080/09064702.2018.1501088

56 Howrigan, D. P., Simonson, M. A. & Keller, M. C. (2011).. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. BMC Genomics 12, 460.

57 Huang, H., Knowles, L. L. (2016). Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Systtematic Biology 65: 357–365. 10.1093/sysbio/syu046.

58 Humble, E., Paijmans, A. J., Forcada, J., Hoffman, J. I. (2020). An 85K SNP Array Uncovers Inbreeding and Cryptic Relatedness in an Antarctic Fur Seal Breeding Colony, G3 Genes|Genomes|Genetics, Volume 10, Issue 8, 2787–2799. https://doi.org/10.1534/g3.120.401268.

59 Jensen, A., Lillie, M., Bergström, K. *et al.* (2021). Whole genome sequencing reveals high differentiation, low levels of genetic diversity and short runs of homozygosity among Swedish wels catfish. Heredity 127**,** 79–91. https://doi.org/10.1038/s41437-021-00438-5

60 Joron, M., Brakefield, P. M., (2003). Captivity masks inbreeding effects on male mating success in butterflies. Nature, 424 (6945): 191-194.

61 Joshi, P. K. et al. (2015). Directional dominance on stature and cognition in diverse human populations. Nature 523, 459–462.

62 Karimi, Z. (2013) Runs of Homozygosity patterns in Taurine and Indicine cattle breeds. Major thesis animal breeding and genetics. BOKU.

63 Keller, M., Visscher, P., Goddard, M., (2011). Quantification of inbreeding due to distant ancestors and its detection using dense SNP data. Genetics. 189, 237 - 249.

64 Kern A. D., David Haussler, D. (2010). A Population Genetic Hidden Markov Model for Detecting Genomic Regions Under Selection. Molecular Biology and Evolution, Volume 27, Issue 7, Pages 1673–1685.

65 Khanshour A., Conant E., Juras R. & Cothran E.G. (2013b) Microsatellite analysis of genetic diversity and population structure of Arabian horse populations. Journal of Heredity 104, 386–98.

66 Kirin, M., McQuillan, R., Franklin, C., Campbell, H., McKeigue, P., Wilson, J., (2010). Genomic runs of homozygosity record population history and consanguinity. PLoS One. 5, e13996.

67 Ku, C. S., Naidoo, N., Teo, S. M., Pawitan Y. (2011). Regions ofhomozygosity and their impact on complex diseases and traits. Human Genetics, 129,1–15.

68 Lencz, T., Lambert, C., DeRosse, P., Burdick, K., Morgan, T., Kane, J., Kucherlapati, R., Malhotra, A., (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proceedings of National Academy of Science USA. 104, 19942 - 19947.

69 Lerner, I. M. (1954). Genetic homeostasis. pp. vii+ 134 pp.

70 Li, C., Horvitz, D. G. (1953). Some methods of estimating the inbreeding coeffcient. American Journal of Human Genetics, 5:, 107–117.

71 Longo, D. L., Drazen, J. M. (2016). Data sharing. New England Journal of Medicine, 374: 276-277.

72 Ma, P., Brøndum, R. F., Zhang, Q., Lund, M. S., Su, G. (2013). Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of Dairy Science, 96: 4666-4677.

73 Magi, A. et al. (2014). H3M2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics 30, 2852–2859.

74 Malécot, G., (1948). Les mathématiques de l'hérédité. Masson.

75 Marchini, J., Howie, B. (2010). Genotype imputation for genome-wide association studies. National Review of Genetics 11 499–511. https://doi.org/10.1038/nrg2796

76 Marshall, T. C., D. W. Coltman, J. M. Pemberton, J. Slate, J. A.Spalton, F. E. Guinness, J. A. Smith, J. G. Pilkington, T.H. Clutton-Brock. (2002). Estimating the prevalence of inbreedingfrom incomplete pedigrees. Proceedings of Royal Society London B., 269, 1533-1540.

77 McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., Campbell, H., … Wilson, J. F. (2008). Runs of homozygosity in European populations. American journal of human genetics, 83(3), 359–372.

78 Metzger, J., Karwath, M., Tonda, R., Beltran, S., Agueda, L., Gut, M., Gut, I. G., Distl, O. (2015) Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. BMC Genomics 16, 764.

79 Meyermans, R., Gorssen, W., Buys, N. *et al.* (2020). How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. *BMC Genomics* **21,** 94. https://doi.org/10.1186/s12864-020-6463-x.

80 Mezzavilla, M. et al. (2015). Increased rate of deleterious variants in long runs of homozygosity of an inbred population from Qatar. Human Heredity 79, 14–19.

81 Mitton, J. B., Pierce, B. A., (1980). The distribution of individual heterozygosity in natural populations. Genetics. 95, 1043-1054.

82 Muchadeyi F.C., Malesa M.T., Soma P. & Dzomba E.F. (2015) Runs of homozygosity in Swakara pelt producing sheep: implications on sub-vital performance. Proceedings for Association for the Advancement of Animal Breeding and Genetics 21, 310–3.

83 Narayanasamy, S., Markina, V., Thorogood, A., Blazkova, A., Shabani, M., Knoppers, B. M., ... & Koesters, R. (2020). Genomic sequencing capacity, data retention, and personal access to raw data in Europe. Frontiers in genetics, 11, 303.

84 Navarro, F. C. P., Mohsen, H., Yan, C. et al. (2019) Genomics and data science: an application within an umbrella. Genome Biology 20, 109. https://doi.org/10.1186/s13059-019-1724-1.

85 Nejati-Javaremi, A., Smith, C., Gibson, J. (1997). Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science, 75: 173.

86 Nielsen, T. et al. (2014). Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer 14, 177.

87 Nothnagel, M., Lu, T. T., Kayser, M., Krawczak, M. (2010) Genomic and geographic distribution of SNP defined runs of homozygosity in Europeans. Human Molecular Genetics 19, 2927–35.

88 Papageorgiou, L. & E., Picasi & Raftopoulou, S. & M., Meropi & Megalooikonomou, V. & Vlachakis, D. (2018). Genomic big data hitting the storage bottleneck. EMBnet.journal. 24. 910. 10.14806/ej.24.0.910.

89 Pemberton, T., Absher, D., Feldman, M., Myers, R., Rosenberg, N., Li, J., (2012). Genomic patterns of homozygosity in worldwide human populations. American Journal of Human Genetics, 91, 275 - 292.

90 Peripolli, E., Munari, D. P., Silva, M. V. G. B., Lima, A. L. F., Irgang, R., Baldi, F. (2017). Runs of homozygosity: current knowledge and applications in livestock, Animal genetics, 10.1111/age.12526.

91 Polasek, O. (2009). Investigating the role of human genome-wide heterozygosity as a health risk factor. PhD Thesis, University od Edinburgh.

92 Polasek, O., Hayward, C., Bellenguez, C., Vitart, V., Kolcic, I., McQuillan, R., Saftic, V., Gyllensten, U., Wilson, J., Rudan, I., Wright, A., Campbell, H., Leutenegger, A.-L., (2010). Comparative assessment of methods for estimating individual genome-wide homozygosity by-descent from human genomic data. BMC Genomics. 11, 139.

93 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics, 81(3), 559-575.

94 Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, no. 2,257-286. doi: 10.1109/5.18626.

95 Ringbauer, H., Novembre, J., & Steinrücken, M. (2021). Parental relatedness through time revealed by runs of homozygosity in ancient DNA. Nature Communications, 12(1), 1-11.

96 Schlather, M. (2020). Efficient Calculation of the Genomic Relationship Matrix.

97 Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources, 11: 1-8. https://doi.org/10.1111/j.1755-0998.2010.02979.x.

98 Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A. et al. (2017) Bioinformatic processing of RAD-seq data dramatically impacts downstream

population genetic inference. Methods Ecol. Evol. 8: 907–917. 10.1111/2041-210X.12700.

99 Shi, Y., Zhao, H., Shi, Y., Cao, Y., Yang, D. et al. (2012). Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. Nat. Genet. 44: 1020–1025. 10.1038/ng.2384.

100 Slate, J., David, P., Dodds, K. et al. (2004). Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. Heredity 93, 255–265. https://doi.org/10.1038/sj.hdy.6800485.

101 Slatkin, M. (1995.) A measure of population subdivision based on microsatellite allele frequencies., Genetics, Volume 139, Issue 1, Pages 457–462. https://doi.org/10.1093/genetics/139.1.457

102 Solkner, J., Ferenčaković, M., Gredler, B., Curik, I. (2010). Genomic metrics of individual autozygosity, applied to a cattle population. Proceedings of the 61st Annual Meeting of the European Associationof Animal Production. Heraklion, Greece.

103 Suwanlee, S., Baummung, R., Solkner, J., Curik, I., (2007). Evaluation of ancestral inbreeding coefficients: Ballou's formula versus gene dropping. Conserv. Genet. 8, 489-495.

104 The Arabidopsis Genome Initiative. (2000.) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815. https://doi.org/10.1038/35048692.

105 Tier, B., (1990). Computing Inbreeding Coefficients Quickly. Genet. Sel. Evol. 22, 419-430.

106 Van De Casteele, T., Galbusera, P., Matthysen, E. (2001). A comparison of microsatellite-based pairwise relatedness estimators. Molecular Ecology, 10: 1539-1549. https://doi.org/10.1046/j.1365-294X.2001.01288.x.

107 VanRaden, P.M. (2008). Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science, Volume 91, Issue 11, 4414-4423.

108 VanRaden, P.M., (1992). Accounting for Inbreeding and Crossbreeding in Genetic Evaluation of Large Populations. J. Dairy. Sci. 75, 3136-3144.

109 Via, M. (2017) Big Data in Genomics: Ethical Challenges and Risks. Revista de Bioética y Derecho, no. 41, 33-45.

110 Villanueva, B., Fernández, A., Saura, M., Caballero, A., Fernandez, J., Morales-Gonzales, E., Toro, M. A., Pong-Wong, R. (2021). The value of genomic relationship matrices to estimate levels of inbreeding. Genetics Selection Evolution 53, 42.

111 von Thaden, A., Nowak, C., Tiesmeyer, A., Reiners, T. E., Alves, P. C., Lyons, L. A., ... & Cocchiararo, B. (2020). Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. Molecular ecology resources, 20(3), 662-680.

112 Wagner, A., Creel, S., Kalinowski, S. (2006). Estimating relatedness and relationships using microsatellite loci with null alleles. Heredity 97, 336–345. https://doi.org/10.1038/sj.hdy.6800865.

113 Wakeley, J., Nielsen, R., Liu-Cordero, S. N., & Ardlie, K. (2001). The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. The American Journal of Human Genetics, 69(6), 1332-1347.

114 Wang, J., (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? Theor. Popul. Biol. 107, 4–13. https://doi.org/10.1016/j.tpb. 2015.08.006.

115 Westbrook, M. J., Wright, M. F., Van Driest, S. L., McGregor, T. L., Denny, J. C., Zuvich, R. L., ... & Brothers, K. B. (2013). Mapping the incidentalome: estimating incidental findings generated through clinical pharmacogenomics testing. Genetics in Medicine, 15(5), 325-331.

116 Wright, S. (1922). Coefficients of inbreeding and relationship. The American Naturalist, 56, 330-338.

117 Yang, J., Benyamin, B., McEvoy, B. P., Gordon,S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., Visscher, P.M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature genetics, 42(7), 565–569.

118 Yang, H. C., Chang, L. C., Liang, Y. J., Lin, C. H. & Wang, P. L. (2012). A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human major histocompatibility complex. PLoS ONE 7, e34840.

119 Zhang L., Orloff M.S., Reber S, Li S., Zhao Y. & Eng C. (2013) CGATOH: extended approach for identifying tracts of homozygosity. PLoS One 8, e57772.

120 Zhang Q., (2017) Exploiting whole genome sequence variants in cattle breeding PhD thesis, Aarhus University, Foulum, Denmark and Wageningen University, Wageningen, the Netherlands. DOI: https://doi.org/10.18174/428523.

Životopis

Filip Čavlović rođen je u Zagrebu 05. svibnja 1996. godine. Pohađao je opću gimnaziju Antuna Gustava Matoša u Samoboru te je nakon završetka srednjoškolskog obrazovanja upisao preddiplomski studij Animalne znanosti na Agronomskom fakultetu u Zagrebu. Završetkom preddiplomskog studija upisuje diplomski studij Genetika i oplemenjivanje životinja također na Agronomskom fakultetu u Zagrebu. Engleski jezik uči tokom cjelokupnog obrazovanja dok njemački jezik uči samo u sklopu osnovnoškolskog, a talijanski i latinski jezik samo u sklopu srednjoškolskog obrazovanja.