Oznaka poročila: ARRS-CRP-ZP-2012-05/48

ZAKLJUČNO POROČILO O REZULTATIH CILJNEGA RAZISKOVALNEGA PROJEKTA

A. PODATKI O RAZISKOVALNEM PROJEKTU

1.Osnovni podatki o raziskovalnem projektu

Šifra projekta	V4-1084
Naslov projekta	Uvedba genomske selekcije v slovensko govedorejo na primeru rjave pasme
Vodja projekta	24769 Gregor Gorjanc
Naziv težišča v okviru CRP	5.09.12 Uvajanje genomske selekcije v rejsko delo v govedoreji
Obseg raziskovalnih ur	1122
Cenovni razred	C
Trajanje projekta	10.2010 - 09.2012
Nosilna raziskovalna organizacija	481 Univerza v Ljubljani, Biotehniška fakulteta
Raziskovalne organizacije - soizvajalke	
Raziskovalno področje po šifrantu ARRS	4 BIOTEHNIKA 4.02 Živalska produkcija in predelava 4.02.01 Genetika in selekcija
Družbeno-ekonomski cilj	08. Kmetijstvo

2.Raziskovalno področje po šifrantu FOS[1]

Šifra	4.02
- Veda	4 Kmetijske vede
- Področje	4.02 Znanosti o živalih in mlekarstvu

3.Sofinancerji[2]

	Sofinancerji		
1.	Naziv		
	Naslov		

B. REZULTATI IN DOSEŽKI RAZISKOVALNEGA PROJEKTA

4.Povzetek projekta[3]

SLO

Selekcija domačih živali je dolgo temeljila na fenotipskih opažanjih rejcev. V zadnjih desetletjih so metode selekcije napredovale, a je selekcija še vedno temeljila na fenotipskih podatkih. Razvoj tehnik molekularne genetike je v zadnjem desetletju omogočil vpogled v genom. Medtem, ko je iskanje genov zamuden proces, lahko poznane variabilne dele genoma uporabimo kot označevalce regij, kjer se lahko nahajajo geni za gospodarsko pomembne lastnosti. Vključitev tega dodatnega vira informacije v sistem ocenjevanja plemenskih vrednosti živali vodi do t.i. genomske selekcije. Prednost tega pristopa pri selekciji goveda je v možnosti ocene plemenskih vrednosti že ob rojstvu, kar lahko poveča sedanji letni genetski napredek tudi za 100%. V projektu smo na primeru rjave pasme goveda razvili potrebno delovno okolje za delo z velikim število genetskih označevalcev in nadgradili programsko opremo za ocenjevanje komponent variabilnosti in plemenskih vrednosti (genetsko vrednotenje) v Sloveniji. Razvoj je temeljil na podatkih za 191 bikov rjave pasme, katerih seme je bilo genotipizirano z Illumina BovineSNP50K čipom s 54.001 označevalcev SNP. Začetne analize nacionalnih podatkov so pokazale, da vključitev genomske informacije v sistem genetskega vrednotenja na osnovi genomsko nadgrajenega mešanega statističnega modela z rodovniki ne prinaša pričakovanih izboljšav. Točnosti (kot korelacija) plemenskih vrednosti za mlade bike so bile v primeru klasičnega pristopa 0,58, 0,51 in 0,58 za količino mleka, mlečnih maščob in beljakovin v laktaciji, medtem ko so bile točnosti z vključitvijo genomske informacije 0,61, 0,60, 0,61. Točnosti so bile nekoliko večje z vključitvijo genomske informacije, a je razlika premajhna v primerjavi z večjimi študijami v tujini. Razlog za neznatno povečanje točnosti je v premajhnem številu genotipiziranih bikov v Sloveniji. Zaradi majhnosti naših populacij smo preučili alternativno metodo, kjer v nacionalni obračun vključimo rezultate ocenjevanja povezav med označevalci SNP in fenotipskimi vrednostmi iz mednarodnih konzorcijev (npr. InterGenomics pri rjavi pasmi) kot neodvisno a povezano spremenljivko. Rezultati te analize so pokazali, da je na takšen način možno v nacionalni sistem vključiti genomsko informacijo s pričakovanim povečanjem točnosti za mlade živali. Za mlade bike se je točnost povečala na 0,79, 0,77, 0,74 za prej omenjene lastnosti. S tem smo pokazali, da je možno genomsko selekcijo uvesti tudi v majhne populacije, pri čemer je nujno potrebno sodelovanje z rejskimi programi iz večjih držav ali konzorciji več držav na tem področju. V Sloveniji lahko genomsko selekcijo za omenjene lastnosti mlečnosti uvedemo takoj, medtem ko je potrebno za večji nabor lastnosti (konformacija, zdravje vimena, plodnost, dolgoživost, itd.) najprej zagotoviti sodelovanje za več lastnosti v InterGenomics konzorciju preko InterBull centra na Švedskem.

ANG

Selection of domestic animals has been based on observations of phenotype for a long time. In last decades several methods of selection have been introduced, however all of them utilized only phenotypic information. Development of techniques of molecular genetics in the last decade has enabled cost effective screening of genomes. While gene discovery is still very slow process we can use polymorphic sites in genome as markers of regions with potential gene for economically important traits. Inclusion of this information into the system of genetic evaluation leads to so called genomic selection. The advantage of this approach to selection in cattle is that we can evaluate breeding values at the birth, which can lead up to 100% greater genetic gain per year. In this project working environment for handling data of many genotype markers has been developed and merged with the existing infrastructure for variance component estimation and inference of breeding values (genetic evaluation) in Slovenia. Development was based on the data for 191 bulls of brown breed whose semen was genotyped using Illumina BovineSNP50K chip for 54,001 SNP markers. Initial analyses of national data showed that introduction of genomic information into genetic evaluation based on the single-step methodology did not provide the expected improvements. Accuracies of breeding values for young bulls were 0.58, 0.51, and 0.58 for milk, fat, and protein yield in lactation using the classical approach, while with the single-step method the accuracies were 0.61, 0.60, and 0.61. Accuracies were slightly higher, but too small in comparison to larger foreign studies. The reason for small improvements in accuracies for young bulls is in too small number of genotyped animals in Slovenia. Due to the small size of our population an alternative method was tested where associations between SNP marker genotypes and phenotypic values from international consortia (such as InterGenomics in brown breed) is introduced as independent but correlated trait. Results from this analysis showed that this is a viable alternative to incorporate genomic information into the national system with the expected improvements in the accuracies for young animals. For young bulls the accuracies were 0.79, 0.77, and 0.74 for

analysed milk traits. This shows that genomic selection can be introduced in small populations conditional on collaboration with breeding programmes from other larger countries or consortia of several countries in this area. Introduction of genomic selection for milk traits in Slovenia can proceed, while for other traits (conformation, udder health, fertility, longevity, etc.) expansion of collaboration with the InterGenomics consortia via the InterBull centre in Sweden is needed.

5.Poročilo o realizaciji predloženega programa dela na raziskovalnem projektu[4]

Uveljavljeni preizkus bikov na osnovi fenotipskih vrednosti potomcev (progeni test) omogoča vrednotenje plemenskih vrednosti s točnostjo (merjeno s korelacijo) več kot 90%. Takšen preizkus daje zanesljive rezultate, a traja dolgo, tudi do šest let v primeru spremljanja mlečnosti. Zaradi visoke točnosti in dolgega generacijskega intervala selekcija bikov na osnovi preizkusa na potomcih omogoča dolgoročno stabilen a zmeren genetski napredek. Vložek v takšen sistem je znaten, saj moramo zbirati podatke po določenem biku dolgo časa. Razlog za dolgotrajen postopek je v dejstvu, da spremljamo izražanje fenotipa pri potomcih. V kolikor bi selekcijo izvajali samo na osnovi informacij o starših bika, bi bila točnost manjša (60%), a bi hkrati skrajšali generacijski interval (na okoli dve leti) in posledično ravno tako dosegli napredek, ki pa je lahko bolj variabilen/nestabilen. Ta sistem se v praksi redko uporablja, saj je poznano, da so bikovske matere pogosto bolje oskrbovane kot ostale krave v populaciji. Posledično so rezultati prireje pri teh kravah pristrani in manj uporabni za selekcijsko delo.

Pri preizkusu na potomcih ocenjujemo učinek alelov, ki jih bik prenese na potomce. V kolikor bi lahko določili kakšne alele nosi določen bik, bi lahko te informacije uporabili za selekcijo bistveno prej in tako skrajšali generacijski interval. S tem namenom raziskovalci na področju genetike in selekcije že nekaj desetletjih skušajo odkriti gene, ki vplivajo na gospodarsko pomembne lastnosti. Iskanje takšnih genov je zamuden proces in do sedaj je poznanih le nekaj primerov in njihove uporaba v praksi ne poveča znatneje točnosti plemenskih vrednosti mladih bikov. Tako njihova uporaba ni ekonomsko zanimiva za selekcijo na gospodarsko pomembne lastnosti. Razvoj tehnik na področju molekularne genetike v zadnjem desetletju je vendarle privedel do cenovno dostopnih metod za vpogled v genom. Medtem ko je iskanje genov še vedno dolgotrajen proces, lahko t.i. genomsko informacijo kljub temu uporabimo za selekcijsko delo kar pogosto poimenujemo kot genomska selekcija. Genomska informacija predstavljena z velikim številom variabilnih mest v genomu, izmed katerih so najbolj pogoste točkovne razlike imenovane označevalci SNP. Uporaba teh informacij omogoča vrednotenje plemenskih vrednosti tudi pri mladih živalih z zglednimi točnostmi (80%), ki so sicer manjše kot pri preizkusu na potomcih, a je znatno krajši tudi generacijski interval in posledično je lahko genetski napredek na leto celo večji in znatno cenejši. Z istim pristopom je možno vrednotiti tudi ženski del populacije kar povečuje možnost napredka v celotni populaciji.

Z namenom preučitve vpeljave genomske selekcije v rejsko delo pri govedu v Sloveniji smo postopke preučili na primeru 191 genotipiziranih bikov rjave pasme. Seme teh bikov smo v okviru rednih nalog rejskega programa v govedoreji poslali v genetski laboratorij na izolacijo DNK in genotipizacijo z Illumina BovineSNP50K čipom za 54,001 označevalcev SNP. Na primeru teh podatkov smo razvili set programov, ki nam omogočajo polavtomatski zajem in kontrolo ter čiščenje podatkov glede na številne kriterije (uspešnost genotipizacije po živali in po označevalcih, frekvenci alelov in genotipov ter odstopanju od Hardy-Weinbergovega ravnotežja). Po vseh kontrolah smo na koncu obdržali 34,450 označevalcev in 183 bikov (podrobnosti so prikazane v četrtem delu priloge - poglavje material in metode). Število zavrženih označevalcev in živali je znatno, a ne bistveno manjše kot pri podobnih raziskavah v tujini. Osem bikov smo izločiti, saj je bil uspeh genotipizacije zaradi starosti semena slab (pod 80%). Razviti set programov hkrati omogoča tudi avtomatsko kontrolo rodovnikov. V obstoječih podatkih nismo zasledili neskladij med zabeleženimi rodovniki in označevalci SNP bikov (sinov in očetov). Ker je takšen pristop k preverjanju porekla znatno bolj natančen kot uveljavljeni postopek z mikrosatelitnimi označevalci, bi kazalo v prihodnosti preiti na sistem z označevalci SNP.

Da bi se bolje spoznali z metodami genomske selekcije smo v sodelovanju s tujimi partnerji izvedli serijo teoretskih študij. Za to smo izdelali program za simulacijo podatkov imenovan AlphaDrop, ki smo ga objavili v mednarodni reviji, kjer je prosto dostopnih tudi deset neodvisno simuliranih setov podatkov. To delo je bilo sprejeto pozitivno in že smo dobili prve citate. Podrobnosti o programu so predstavljene v prvem delu priloge. Na teh simuliranih podatkih smo preučevali različne pristope k selekciji z uporabo različnih virov genetskih informacij (rodovniki, označevalci SNP, haplotipi ali geni – QTL). Rezultati so pokazali, da vključitev genomskih informacij omogoča povečanje točnosti plemenskih vrednosti, še posebej pri mladih živalih. Razvili smo metodo za vključitev poljubno dolgih haplotipov (kombinacije označevalcev SNP na gametah) v vrednotenje, kar bi na podlagi teoretskih pričakovanj omogočilo še nadaljnje povečanje točnosti. Pri tem se je izkazalo, da daje uporaba označevalcev SNP ali haplotipov primerljive točnosti. Pokazali smo tudi, da četudi bi poznali genotip vseh genov (QTL), ki vplivajo na gospodarsko pomembne lastnosti, ne moremo oceniti plemenskih vrednosti s 100% točnostjo zaradi statističnega ocenjevanja na vzorcu podatkov. Prav tako smo pokazali, da točnosti napovedi iz generacije v generacijo padajo. Slednje potrjuje pomen dolgoročnega spremljanja prireje in beleženja podatkov, saj bomo lahko la na osnovi teh zagotavljali ustrezno točnost sistema v prihodnosti. Podrobnosti te študije so predstavljene v drugem delu priloge.

Opremljeni s teoretskimi osnovami o genomski selekciji smo izvedli še eno teoretično študijo s katero smo pokazali, da je v primeru selekcije točnost plemenskih vrednosti izvrednotena s klasičnim pristopom (fenotipske vrednosti + rodovniki) znatno zmanjšana, še posebej pri mladih živalih. Slednje potrjuje praktična opažanja, da je točnost pri mladih bikih v realnosti bistveno manjša kot bi pričakovali na podlagi teoretskih izračunov na kar očitno vpliva zelo ostra selekcija bikovskih mater. Pri tem smo pokazali, da so ocene plemenske vrednosti z uporabo genomskih informacij znatno manj podvržene vplivu selekcije. Posledično je točnost takšnih ocen primerjalno na točnost klasičnih plemenskih vrednosti še bistveno večja kot smo pričakovali. Podrobnosti te študije so predstavljene v tretjem delu priloge.

Opremljeni s teoretičnim znanjem o genomski selekciji smo preučili možnost vključitve genomske informacije v oceno plemenskih vrednosti na slovenskih podatkih. Na voljo smo imeli 1,342,134 fenotipskih vrednosti za količino mleka, mlečnih maščob in beljakovin v laktaciji od 57,670 krav rjave pasme med leti 1997 in 2011. Za te živali smo sestavili rodovnik, ki je vseboval 79,573 živali. Poleg tega smo v analize vključili še genotipe bikov za označevalcev SNP. Ob zasnovi projekta smo na podlagi dotedanjih spoznanj zastavili delo z metodo, ki hkrati vključuje vse tri vire informacij: fenotipske vrednosti, rodovnike in označevalce SNP. Uporaba te metode na naših podatkih ni prinesla pričakovanih rezultatov. Točnosti (kot korelacija) plemenskih vrednosti za mlade bike so bile v primeru klasičnega pristopa 0,58, 0,51 in 0,58 za količino mleka, mlečnih maščob in beljakovin v laktaciji, medtem ko so bile točnosti z vključitvijo genomskih informacij 0,61, 0,60, 0,61. Pri tem je zaznati določeno povečanje točnosti (še posebej pri količini maščobe, kar je pričakovano zaradi gena DGAT z izrazito velikim učinkom na vsebnost maščobe v mleku), a je razlika manjša v primerjavi z večjimi študijami v tujini. Razlog za neznatno povečanje točnosti je v premajhnem številu genotipiziranih bikov v Sloveniji. Četudi bi genotipizirali vse bike rjave pasme pri nas, se točnosti ne bi znatneje povečale, saj je število vseh bikov majhno. Od 736 bikov, ki se pojavljajo v naših rodovnikih ima le 399 bikov znatnejše število potomcev, da so še lahko vključeni v mednarodno primerjavo pri InterBull centru. Pri tem je potrebno poudariti, da je od teh 399 bikov le 191 takšnih, ki smo jih vzredili in uporabljali pretežno le v Sloveniji, medtem ko so ostali biki tuji. Zaradi majhnosti naših populacij smo preučili metodo, kjer v nacionalni obračun vključimo rezultate ocenjevanja povezav med označevalci SNP in fenotipskimi vrednostmi iz mednarodnih konzorcijev (npr. InterGenomics pri rjavi pasmi) kot neodvisno a povezano spremenljivko. Rezultati so pokazali, da je na takšen način možno v nacionalni sistem vključiti genomsko informacijo s pričakovanim povečanjem točnosti za mlade živali. Za mlade bike se je točnost povečala na 0,79, 0,77, 0,74 za prej omenjene lastnosti. Ti rezultati so primerljivi rezultatom iz večjih populacij. S tem smo pokazali, da je možno genomsko selekcijo uvesti tudi v majhne populacije, pri čemer je nujno potrebno sodelovanje z rejskimi programi iz večjih držav ali konzorciji več držav. V Sloveniji lahko genomsko selekcijo za omenjene lastnosti mlečnosti uvedemo takoj, medtem ko je potrebno za večji nabor lastnosti (konformacija, zdravje vimena, plodnost, dolgoživost, itd.) najprej zagotoviti sodelovanje z več lastnostmi v InterGenomics konzorciju pri InterBull centru. Podrobnosti tega dela študij so predstavljene v četrtem delu priloge.

Uvedba genomske selekcije brez spremembe rejskega programa ni smiselna. Povečanje točnosti plemenskih vrednosti za mlade živali odpira zelo veliko novih možnosti za poenostavitev in s tem tudi pocenitev selekcijskega dela. Formalne ekonomske analize nismo izvedli, saj lahko ob točnosti okoli 80% v skladu z Evropsko zakonodajo vključimo mlada moška teleta v reprodukcijo praktično takoj ob nastopu spolne zrelosti. Za Slovenske razmere bi kazalo uvesti sistem, kjer bi vzorčenje, genotipizacijo in izračun plemenskih vrednosti mladih živali 1) pokrili rejci sami ali 2) vzrejališča ali osemenjevalni centri v dogovoru z rejci pokrijejo stroške za vnaprej izbrane živali po ustaljenih kriterijih, pri čemer se obe strani zavežeta za morebiten nakup/prodajo v skladu z uveljavljenim cenikom. Takšni sistemu so vpeljani tudi v tujini. Vzorčenje tkiv (sluznica ali del ušesa tekom označevanja ali rovašenja) bi lahko opravljali kontrolorji v sklopu rednih delovnih obveznosti na kmetijah npr. ob označevanju novorojenih telet. Vnaprej sestavljen seznam živali v okviru rejskega programa bi lahko osnovali na osnovi povprečja plemenskih vrednosti staršev pri čemer bi v poštev za vzorčenje prišle živali s PV12 nad določenim pragom. Pragove bo potrebno v prihodnje določiti ob pregledu variabilnosti pri posameznih pasmah in številčnosti populacij. Podoben sistem bi kazalo uvesti tudi za pred-odbiro bikovskih mater pri čemer bi se lahko genotipizacija pokrila iz sredstev rejskega programa ob pogoju vzpostavitve predkupne pravice za moško tele s strani vzrejališča ali osemenjevalnega centra. Osnovna cena takšnih živali bi recimo bila 5000 EUR, pri čemer bi vsaka nadaljnja točka za PV12 prinesla še dodatnih 1000 EUR do določenega maksimuma. Te cene so podane zgolj kot primer in so stvar dogovorov v rejski organizaciji. Cena se mora vsekakor nanašati na plemensko vrednost z upoštevanjem dejstva, da je takšna žival on nastopu spolne zrelosti že lahko primerna za vključitev v postopke za preverjanje reprodukcijske sposobnosti na osemenjevalni centrih in v primeru pozitivnih rezultatov zelo kmalu vključena v zbiranje in distribucijo semena. Pri tem bi kazalo uvesti kratkotrajni preizkus v okviru biološkega testa, da se z nekaj osemenitvami preveri ali ni slučajno odbrani bik nosilec nezaželenih mutacij in prenašalec težkih telitev. Takšen preizkus je možno opraviti zelo hitro in ne bo bistveno zaviral genetskega napredka. Za praktično uvedbo genomske selekcije je tako potrebno doseči dogovor vseh vpletenih deležnikov, medtem ko smo v okviru tega projekta že začeli obveščati rejce o možnostih, ki jih prinaša genomska selekcija (peti del priloge).

6.Ocena stopnje realizacije programa dela na raziskovalnem in zastavljenih raziskovalnih ciljev[5]

Cilj predlaganega projekta je bil preučiti možnosti uvedbe genomske selekcije v slovensko govedorejo na primeru rjave pasme. Pri tem smo dosegli vse zastavljene cilje, ki so podrobneje predstavljeni spodaj.

1) Pri delu z obstoječimi podatki za 54.001 označevalcev SNP za 191 bikov rjave pasme smo razvili delovno okolje s programom SAS za delo z velikim številom genetskih označevalcev. Obstoječa programska oprema hkrati omogoča tudi avtomatsko preverjanje porekla za genotipizirane živali.

2) Razvili smo programsko okolje za oceno povezav med velikim številom genetskih označevalcev in fenotipskimi vrednostmi ter kombinacijo teh informacij s fenotipskimi vrednosti in poreklom z namenom, da čim bolj natančno ocenimo plemenske vrednosti živali, še posebej mladih genotipiziranih živali. Programska oprema je pripravljena za vključitev v rutinski obračun plemenskih vrednosti.

3) Razvito programsko opremo smo preizkusili na lastnosti mlečnostih, kjer smo imeli na voljo nacionalne podatke o mlečnosti krav in njihovih rodovnikih, genotipih za 54.001 označevalcev SNP za 191 slovenskih bikov ter rezultate ocen povezav med označevalci SNP in fenotipskimi vrednostmi iz konzorcija InterGenomics. Rezultati analiz so pokazali, da lahko v Sloveniji uspešno uvedemo genomsko selekcijo (točnosti plemenskih vrednosti pri mladih živali so zadovoljivo visoke) v primeru sodelovanja z mednarodnimi partnerji (velike države ali konzorciji kot je npr. InterGenomics pri rjavi pasmi) pri ocenjevanju povezav med označevalci SNP in fenotipskimi vrednostmi.

4) Na osnovi prikazanega povečanja točnosti plemenskih vrednosti za mlade živali na osnovi genomskih informacij smo predlagali spremembe rejskega programa pri rjavi pasmi goveda, medtem ko bo za uvedbo genomske selekcije pri ostalih pasmah goveda (lisasta in črno-bela) najprej potrebno poiskati mednarodne partnerje za sodelovanje pri ocenitvi povezav med označevalci SNP in fenotipskimi vrednosti.

5) Rezultate številnih analiz smo dokumentirali z objavami in predstavitvami v znanstveni in strokovni javnosti kot smo opisali v drugih točkah tega poročila. Tekom izvajanja del na projektu smo vzpostavili celo serijo povezav s tujimi raziskovalnimi skupinami, kar bo še pospešilo nadajljni razvoj na tem področju.

7.Utemeljitev morebitnih sprememb programa raziskovalnega projekta oziroma sprememb, povečanja ali zmanjšanja sestave projektne skupine[6]

Pri projektu nismo izvedli sprememb programa ali sestave projektne skupine.

8.Najpomembnejši znanstveni rezultati projektne skupine[7]

	Znanstveni dosežek

1.	COBISS ID		3034248	Vir: COBISS.SI
	Naslov	SLO	/	
		ANG	Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods	
	Opis	SLO	/	
		ANG	An approach is described for simulating data sequence, genotype, and phenotype data to study genomic selection and genome-wide association studies (GWAS). The simulation method, implemented in a software package called AlphaDrop, can be used to simulate genomic data and phenotypes with flexibility in terms of the historical population structure, recent pedigree structure, distribution of quantitative trait loci effect, and with sequence and single nucleotide polymorphism-phased alleles and genotypes. Ten replicates of representative scenario used to study genomic selection in livestock were generated and have been made publically available. The simulated data sets were structured to encompass a spectrum of additive quantitative trait loci effect distributions, relastionship structures, and single nucleotide polymorpism chip densities.	
	Objavljeno v		Genetics Society of America; G3; 2012; Vol. 2, no. 4; str. 425-427; Avtorji / Authors: Hickey John M., Gorjanc Gregor	
	Tipologija		1.03 Kratki znanstveni prispevek	
2.	COBISS ID		3028872	Vir: COBISS.SI
	Naslov	SLO	/	
		ANG	An imputation strategy which results in an alternative parameterization of the single stage genomic evaluation	
	Opis	SLO	/	
		ANG	A method is presented for imputing genotypes in pedigreed populations based on long-range phasing, haplotype libraries, recombination modelling and segragation analysis. In two very different data sets, one a pig data set comparising animals from a single line and the other a multiple breed cattle data set, imputation accuracy was high and was always higher than that of Impute2 a widely used alternative. Accuracy was highest for animals which had both parents genotyped at high-density, however some animals with neither parent genotyped at high-density also had high imputation accuracy. The method imputes genotypes or sum of the allele probabilities for all animals in the pedigree and thus facilitates single stage genomic evaluations combining all available pedigree, genomic, and phenotypic information in a single step. This was expored using bith simulated and real data with favourable results.	
	Objavljeno v		Interbull Centre; Proceedings of the 2011 INTERBULL meeting, Stavanger, Norway, August 27-29 2011; 2011; Str. 38-41; Avtorji / Authors: Hickey J.M., Gorjanc Gregor, Kinghorn B.P., Tier B., Werf J.H.J. van der, Cleveland Michael A.	
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci	
3.	COBISS ID		3093384	Vir: vpis v poročilo
	Naslov	SLO	/	
		ANG	Accuracy of genomic prediction for protein yield using different models in Slovenian Brown bulls	
	Opis	SLO	/	
			Use of genomic (SNP) information enables more accurate inference of breeding values (BV), especially for young animals. The objective of this study was to compare the correlation between progeny based and genomically based evaluation for protein yield of Slovenian Brown bulls	

		ANG	based on different sources of information and scenarios. Emphasis was to integrate genomic information into national evaluation. Four data sets were constructed as follows: 1) DS1 - phenotypic data (PD) from national genetic evaluation (1.342.134 test-day records of 57.670 cows recorded between years 1997 and 2011); 2) DS2 - DS1 + 50K Illumina SNP genotypes for 183 bulls; 3) DS3 - DS1 + direct genomic value (DGV) for 183 Slovenian bulls evaluated at InterBull (InterGenomics) treated as correlated trait; 4) DS4 - DS1 + DGV for 399 bulls in the national pedigree. Two scenarios were evaluated. In the first scenario, all PD was used in analysis, while in the second PD was removed for years 2008 to 2011 to exclude daughter information of 35 genotyped bulls. Animal model was used for the analysis of DS1, while the joint pedigree and genomic relationship model was used for the analysis of DS2. For the analysis of DS3 and DS4, bivariate animal model was used. Both theoretical and validation based accuracies were evaluated. Application of joint pedigree and genomic relationship model did not increase validation accuracies due to the limited number of genotyped animals. Theoretical accuracies were 0.58 (DS1), 0.61 (DS2), 0.84 (DS3), and 0.74 (DS4), while validation accuracies were 0.56 (DS1), 0.54 (DS2) 0.86 (DS3), and 0.72 (DS4). For comparison, theoretical accuracy of genomically enhanced breeding value at InterGenomics for validation bulls is 0.90. Results show that integration of genomic information into national evaluation was successful.	
	Objavljeno v		Book of Abstracts of the 63rd Annual Meeting of the European Federation of Animal Science, Bratislava, Slovakia, 27-31 August 2012.- Str. 134; Avtorji / Authors: Špehar, M., Potočnik, K. and Gorjanc Gregor	
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci	
4.	COBISS ID		Še ni vpisano1	Vir: vpis v poročilo
	Naslov	SLO	/	
		ANG	Reliability of breeding values in selected populations	
	Opis	SLO	/	
		ANG	Selection reduces genetic variance in population. However, this is not taken into account when reliabilities are computed from prediction error variance (PEV) and base population additive genetic variance. Results of simulations confirmed that when selection is present PEV based reliabilities are too high and do not reflect the true uncertainty of EBV. Drop in reliability is substantial for parent average, while EBV for progeny tested or genomically evaluated animals is reduced only slightly. This implies that genomic EBV are in comparison to pedigree EBV even more reliable than anticipated from the comparison of PEV based reliabilities.	
	Objavljeno v		Interbull Centre; Proceedings of the 2012 INTERBULL meeting, Cork, Ireland, May 28 - June 01 2012; 2012; ; Avtorji / Authors: Gorjanc Gregor, Hickey J.M., Bijma P. http://www.icar2012.ie/presentations/videos/IBOM2_6_29_5_2012_Gregor	
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci	
5.	COBISS ID		Še ni vpisano2	Vir: vpis v poročilo
	Naslov	SLO	/	
		ANG	Accuracy of genomic prediction for milk traits with different approaches in Slovenian Brown bulls	
	Opis	SLO	/	
			Use of genome-wide single nucleotide polymorphism (SNP) marker information enables more accurate inference of breeding values (EBV), especially for young animals. The objective of this study was to compare the accuracy of genomic and progeny based evaluation for milk traits in a small population of Slovenian Brown bulls using different approaches to	

		ANG	utilize genomic information. Four approaches were considered: 1) NAT - phenotypic and pedigree data from national genetic evaluation used in univariate repeatability test-day model, 2) NATss - NAT approach with the inclusion of genome-wide SNP genotypes for 183 Slovenian Brown bulls via the improved relationship matrix of single-step methodology, 3) MT1 - NAT approach with direct genomic values (DGV) as correlated trait for 183 Slovenian Brown bulls available externally from the InterGenomics consortium, and 4) MT2 - the same as MT1 but with using DGV for 399 bulls in the national pedigree. Performance of different approaches was assessed with the analysis of theoretical and validation accuracies. For validation bulls in juvenile stage (reduced dataset) increase in accuracy with the NATss approach did was negligible in comparison to the approach NAT due to the small reference population in Slovenia. With the MT1 and MT2 approach the average theoretical accuracies were 0.90 (MT1) and 0.79 (MT2) for milk, 0.86 (MT1) and 0.77 (MT2) for fat, and 0.85 (MT1) and 0.74 (MT2) for protein yield. These results confirm the expected increase in accuracy due to the inclusion of genomic information (via DGV) in the national evaluation system. However, accuracies with the MT1 approach were unrealistically high. Validation accuracies were lower in comparison to the average theoretical accuracies, especially for the NAT and NATss approaches. With the MT1 and MT2 validation accuracies were 0.92 (MT1) and 0.74 (MT2) for milk, 0.91 (MT1) and 0.81 (MT2) for fat, and 0.87 (MT1) and 0.72 (MT2) for protein yield. Use of larger number of animals with DGV information (national and foreign bulls used in Slovenia) as in the MT2 approach resulted in realistic accuracy of genetic evaluation. These results show that the integration of genomic information into national evaluation was successful. Further research is needed to quantify the effect of potential double counting of available information.
	Objavljeno v		Špehar M., Potočnik K., Gorjanc Gregor - poslano v objavo v revijo Czech Journal of Animal Science
	Tipologija		1.01 Izvirni znanstveni članek

9.Najpomembnejši družbeno-ekonomsko relevantni rezultati projektne skupine[8]

	Družbenoekonomsko relevantni dosežki			
1.	COBISS ID		3116168	Vir: vpis v poročilo
	Naslov	SLO	/	
		ANG	Whole-genome evaluation of complex traits using SNP, haplotype, or QTL information	
	Opis	SLO	/	
			Whole-genome technologies provide rich data for dissection of complex traits. While gene discovery is still largely limited, the data at hand can be successfully used for evaluation of genetic merit. The aim of this work was to demonstrate the value of different sources of information (pedigrees, Single Nucleotide Polymorphisms – SNP, haplotypes, or Quantitative Trait Loci – QTL) for genetic evaluation of non-phenotyped individuals in a typical animal breeding scenario via simulation. In the first step a coalescent simulation was used to create a base population with structured chromosomes that were in the second step dropped and recombined through the pedigree of 10 generations with 50 sires per generation, 10 dams per sire, and 2 offspring per dam. Phenotypic values were simulated with different genetic architectures (QTL effects were sampled from Gaussian or gamma distribution and minor allele frequency less than 0.3) and heritability of 0.25. Genotypic data was available for all individuals from generation 4 onwards, while phenotypic data was available for	

		ANG	individuals in generations 4 and 5. Genetic evaluation was based on linear mixed models with relationship matrix between individuals. This matrix was built using pedigree, SNP, haplotype, or QTL data. Haplotypes of different length were considered (from 5 to all the way up to 2000 SNP) with an option to account for similarities between haplotypes while building relationship matrices. The accuracy of different methods was assessed by correlation between true and evaluated additive genetic values for individuals in generations 6, 8 and 10. Average accuracy over ten replications for Gaussian trait over generations was between 0.45 to 0.10 for pedigree data, 0.50 to 0.35 for SNP and haplotype data and 0.6 to 0.4 for QTL data. In the case of long haplotypes accuracies dropped considerably, but accounting for similarities between haplotypes prevented this drop. In the case of gamma trait accuracies were slightly higher in generation 6 and dropped faster in the later generations in the case of pedigree, SNP, and haplotype data due to recombinations. On the other hand accuracies were substantially higher with QTL data and quite stable over generations (from 0.75 to 0.65) though still far from perfect (even though QTL genotypes are known), due to estimation errors. Results demonstrate the value and limitations of genotypic information for the evaluation of additive genetic merit in animal populations.	
	Šifra		B.03 Referat na mednarodni znanstveni konferenci	
	Objavljeno v		Genetika 2012 : book of abstracts; Ljubljana: Slovensko genetsko društvo; 2012; str. 84.; Avtorji / Authors: Gorjanc Gregor, Hickey J.M.	
	Tipologija		1.08 Objavljeni znanstveni prispevek na konferenci	
2.	COBISS ID		3062664	Vir: vpis v poročilo
	Naslov	SLO	Application of latent Gaussian models in genetics	
		ANG	Application of latent Gaussian models in genetics	
	Opis	SLO	Latent Gaussian models represent an important class of models used in genetics, especially for the analysis of traits having complex genetic architecture and being under influence of environment. These complexities severely limit the discovery of genes and their interactions in such traits. Gaussian approximation of joint effects of all genes provides a way to avoid these complexities by introducing conceptual latent variables representing different modes of gene effects: additive, dominance, epistasis, and imprinting. Introduction of latent variables enables conceptual treatment of genetic effects, where even pedigrees alone can be used as the only source of genetic information, though with lost resolution on individual gene effects. Latent Gaussian models are used heavily in animal and plant breeding, but also in evolutionary biology and human genetics. In all fields latent Gaussian models are used to dissect phenotypic variation into genetic and environmental components to better understand the state of nature. In animal and plant breeding the primal goal is prediction for selective breeding. Recent advancements in molecular genetics lead to the generation of massive amount of information via genetic markers. These can be used in the same framework by replacing or upgrading pedigree information.	
		ANG	Latent Gaussian models represent an important class of models used in genetics, especially for the analysis of traits having complex genetic architecture and being under influence of environment. These complexities severely limit the discovery of genes and their interactions in such traits. Gaussian approximation of joint effects of all genes provides a way to avoid these complexities by introducing conceptual latent variables representing different modes of gene effects: additive, dominance, epistasis, and imprinting. Introduction of latent variables enables conceptual treatment of genetic effects, where even pedigrees alone can be used as the only source of genetic information, though with lost resolution on individual gene	

			effects. Latent Gaussian models are used heavily in animal and plant breeding, but also in evolutionary biology and human genetics. In all fields latent Gaussian models are used to dissect phenotypic variation into genetic and environmental components to better understand the state of nature. In animal and plant breeding the primal goal is prediction for selective breeding. Recent advancements in molecular genetics lead to the generation of massive amount of information via genetic markers. These can be used in the same framework by replacing or upgrading pedigree information.	
	Šifra		B.04 Vabljeno predavanje	
	Objavljeno v		Second Workshop on Bayesian Inference for Latent Gaussian Models with Applications, Norwegian University of science and Technology, [Trondheim], 30 May - 1 June, 2012. [S.l.: s.n.], 2012, str. 27.; Avtorji / Authors: Gorjanc Gregor	
	Tipologija		1.06 Objavljeni znanstveni prispevek na konferenci (vabljeno predavanje)	
3.	COBISS ID		2931336	Vir: vpis v poročilo
	Naslov	SLO	Študijsko gradivo za predmet Genomika domačih živali	
		ANG	/	
	Opis	SLO	Sodelovanje v tem projektu je omogočilo nadgradnjo predmeta Genomika domačih živali na drugi stopnji študija kmetijstvo zootehnika (MSc) z vsebinami s področja genomskih asociacijskih študij in genomske selekcije.	
		ANG	/	
	Šifra		D.10 Pedagoško delo	
	Objavljeno v		Avtorji / Authors: Kunej, T., Gorjanc, Gregor, Horvat, S. Študijsko gradivo za predmet Genomika domačih živali, 2010-2011 : drugostopenjski M.sc. študijski program "Kmetijstvo-zootehnika". [Elektronska izd.]. Domžale: Biotehniška fakulteta, Oddelek za zootehniko, 2011. 1 CD-ROM.	
	Tipologija		2.05 Drugo učno gradivo	
4.	COBISS ID		Še ni vpisano3	Vir: vpis v poročilo
	Naslov	SLO	Mentorstvo doktorandom Marija Špehar in Vesna Mrak	
		ANG	/	
	Opis	SLO	Sodelovanje v tem projektu je omogočilo raziskovalno delo pri čemer nastajata dve doktorski nalogi v sodelovanju z doktorandkama Marija Špehar (doktorski študij zootehnike) in Vesna Mrak (doktorski študij statitike). Obe doktorski nalogi sta še v teku.	
		ANG	/	
	Šifra		D.09 Mentorstvo doktorandom	
	Objavljeno v		/	
	Tipologija		2.13 Elaborat, predštudija, študija	
5.	COBISS ID		Še ni vpisano4	Vir: vpis v poročilo
	Naslov	SLO	Razvoj programske opreme za rutinsko genomsko vrednotenje domačih živali	
		ANG	/	
	Opis	SLO	Rezultat tega projekta je razvita programska oprema za rutinsko genomsko vrednotenje domačih živali, ki zajema delo s fenotipskimi in rodovniškimi podatki kakor tudi velik številom genomskih (SNP) označevalcev.	
		ANG	/	
			F.23 Razvoj novih sistemskih, normativnih, programskih in	

	Šifra	metodoloških rešitev
	Objavljeno v	/
	Tipologija	2.21 Programska oprema

10.Drugi pomembni rezultati projektne skupine[9]

Gorjanc G. Predavanje na temo genomske selekcije na letni skupščini Zveze rejcev rjavega goveda. Kmetijska šola Grm Novo mesto, 2011.

Gorjanc, G. Predstavitev rezultatov projekta na delovnem sestanku vodij selekcije v govedoreji. Kmetijsko gozdarski zavod Murska Sobota, 2012.

Gorjanc G. Genomska selekcija, Kmečki glas, 2012.

11.Pomen raziskovalnih rezultatov projektne skupine[10]

11.1.Pomen za razvoj znanosti[11]

SLO

Rezultati našega dela so številni in segajo na več področij znanosti o reji domačih živali kakor tudi genetike drugih vrst.

1) Izdelan in dokumentiran program (AlphaDrop) za simulacijo genomskih asociacijskih študij in genomske selekcije za potrebe testiranja novih metod genomske selekcije (Hickey in Gorjanc, 2012). Program AlphaDrop in simulirani seti podatkov so prosto dostopni na spletu (http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.111.001297/-/DC1).

2) Preučitev uporabe različnih virov genetskih informacij (rodovniki, označevalci SNP, haplotipi in geni - QTL) za genetsko vrednotenje kompleksnih lastnosti (Hickey in sod., 2012). S pomočjo simulacij smo pokazali praktične omejitve genetskega vrednotenja glede na različne vire informacij. Četudi bi poznali vse gene, ki so vplivajo na kompleksne lastnosti, je v sistemu genetskega vrednotenja prisotna napaka, saj je potrebno vpliv genov oceniti za vsako populacijo ločeno. Slednje v sistem vnaša statistično napako, ki pa je manjša kot pri drugih virih genetskih informacij (rodovniki > označevalci SNP = haplotipi). Pri tem delu smo razvili metodo, ki omogoča uporabo haplotipov (poljubne dolžine) namesto označevalcev SNP za izgradnjo genomske matrike sorodstva pri genetskem vrednotenju.

3) Preučitev vpliva selekcije na točnost genetskega vrednotenja in primerjava med klasičnim in genomskim pristopom (Gorjanc in sod., 2012). Rezultati so pokazali, da je v primeru selekcioniranih populacij točnost klasičnega genetskega vrednotenja znatno zmanjšana pri mladih živalih brez fenotipskih podatkov, medtem ko selekcija nima tako izrazitega vpliva na točnost v primeru uporabe genomske informacije. To pomeni, da je genomski pristop v selekcioniranih populacijah znatno bolj točen kot je bilo znano do sedaj.

4) Preučitev različnih možnosti uporabe genomskih informacij za genetsko vrednotenje živali v majhnih populacijah (Špehar in sod., 2012). Pokazali smo, da je možno genomsko selekcijo uvesti tudi v majhnih populacijah, pri čemer je nujno potrebno sodelovanje z večjimi državami ali konzorciji.

ANG

1) Developed and documented program (AlphaDrop) for simulation of data for genome-wide association studies and genomic selection studies with the aim to test and develop new methods in this area (Hickey and Gorjanc, 2012). Program AlphaDrop and simulated data sets are freely available (http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.111.001297/-/DC1).

2) The study of different sources of genetic information (pedigrees, SNP markers, haplotypes and genes - QTL) for genetic evaluation of complex traits (Hickey et al., 2012). Practical

limitation of genetic evaluation using different source of genetic information have been demonstrated using simulation approach. Even if all genes for complex traits would be known the accuracy of genetic evaluation will not be perfect, due to the estimation errors introduced via the estimation of gene effects. However, in some cases errors are small than with the use of other source of genetic information (pedigree > SNP markers = haplotypes). A unique method has been developed which allows the use of haplotypes of any length to be used instead of SNP markers for the construction of genomic relationship matrix for genetic evaluation.

3) The study of effect of selection on the accuracy of genetic evaluation comparing classical and genomic approach (Gorjanc et al., 2012). Results show that selection reduces accuracy of genetic evaluation especially in young non-phenotyped animals, while this effect is much smaller when genomic evaluation is being used. This implies that genomic selection is even more accurate in young animals than anticipated until now.

4) The study of different ways to integrate genomic information into genetic evaluation in small populations (Špehar et al., 2012). Results show that small populations can introduce genomic selection conditional that there is collaboration in estimating associations between SNP marker genotypes and phenotypic values with breeding programmes from larger countries of consortia of many countries.

11.2.Pomen za razvoj Slovenije[12]

SLO

Rezultati našega dela uvajajo v Slovensko živinorejo uporabo genomskih informacij za potrebe selekcije. S tem se postavljamo ob bok velikim in razvitim živinorejskih državam v Evropi in drugod po svetu. Uporaba genomskih informacij pri selekciji kakor tudi za druge namene (določitev pasme, spremljanje rodovnikov, odkrivanje genov in drugih genetskih zakonitosti, uravnavanje prehrane glede na genotip živali, ipd.) predstavlja novo obdobje v živinoreji in celotnem kmetijstvu, kjer se bodo vsakdanje odločitve sprejemale na osnovi bogatih virov podatkov (genomika, geografski informacijski sistemi (GIS), ipd.).

ANG

Results of this project introduce genomic information into the breeding activities of Slovenian animal production. This puts Slovenia aside bigger and more developed countries in this area in Europe and in other parts of the world. Use of genomic information in selection as well as other activities (breed classification, pedigree checks, discovery of genes and other genetic principles, management of nutrition according to the animal’s genotype, etc.) introduces a new era in animal production as well as whole agriculture in Slovenia where everyday decisions will be based on rich data sources (genomics, geographic information systems (GIS), etc.).

12.Vpetost raziskovalnih rezultatov projektne skupine.

12.1.Vpetost raziskave v domače okolje

Kje obstaja verjetnost, da bodo vaša znanstvena spoznanja deležna zaznavnega odziva?

☑ v domačih znanstvenih krogih

☑ pri domačih uporabnikih

Kdo (poleg sofinancerjev) že izraža interes po vaših spoznanjih oziroma rezultatih?[13]

Interes po naši spoznanjih pridobljenih tekom projekta izražajo raziskovalne skupine, ki se s podobno tematiko (uporaba genomskih informacij za napovedovanje fenotipskih vrednosti) srečujejo na področju genetike laboratorijskih živali in ljudi.

12.2.Vpetost raziskave v tuje okolje

Kje obstaja verjetnost, da bodo vaša znanstvena spoznanja deležna zaznavnega odziva?

☑ v mednarodnih znanstvenih krogih

☑ pri mednarodnih uporabnikih

Navedite število in obliko formalnega raziskovalnega sodelovanja s tujini raziskovalnimi inštitucijami:[14]

Tekom dela na projekta smo navezali stike s kar nekaj raziskovalnimi inštitucijami po svetu vendar še nimamo formalnih oblik sodelovanja:
1) University of New England, New South Wales, Armidale, Australia (prof. dr. Julius van der Werf)

2) Animal Genetics and Breeding Unit (AGBU), Armidale (dr. Bruce Tier)

3) University of Wagenignen, Wagenignen, Netherlands (prof. dr. Piter Bijma)

4) International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico, (dr. John Hickey)

Kateri so rezultati tovrstnega sodelovanja:[15]

Nekatere objave navedene pod številkami 1., 2. in 4.

C. IZJAVE

Podpisani izjavljam/o, da:
- so vsi podatki, ki jih navajamo v poročilu, resnični in točni
- se strinjamo z obdelavo podatkov v skladu z zakonodajo o varstvu osebnih podatkov za potrebe ocenjevanja in obdelavo teh podatkov za evidence ARRS
- so vsi podatki v obrazcu v elektronski obliki identični podatkom v obrazcu v pisni obliki
- so z vsebino letnega poročila seznanjeni in se strinjajo vsi soizvajalci projekta
- bomo sofinancerjem istočasno z zaključnim poročilom predložili tudi študijo ali elaborat, skladno z zahtevami sofinancerjev

Podpisi:

zastopnik oz. pooblaščena oseba raziskovalne organizacije:	in	*vodja raziskovalnega projekta:*
Univerza v Ljubljani, Biotehniška fakulteta		Gregor Gorjanc

ŽIG

Kraj in datum: Ljubljana 10.10.2012

Oznaka prijave: ARRS-CRP-ZP-2012-05/48

[1] Zaradi spremembe klasifikacije je potrebno v poročilu opredeliti raziskovalno področje po novi klasifikaciji FOS 2007 (Fields of Science). Prevajalna tabela med raziskovalnimi področji po klasifikaciji ARRS ter po klasifikaciji FOS 2007 (Fields of Science) s kategorijami WOS (Web of Science) kot podpodročji je dostopna na spletni strani agencije (http://www.arrs.gov.si/sl/gradivo/sifranti/preslik-vpp-fos-wos.asp). Nazaj

[2] Podpisano izjavo sofinancerja/sofinancerjev, s katero potrjuje/jo, da delo na projektu potekalo skladno s programom, skupaj z vsebinsko obrazložitvijo o potencialnih učinkih rezultatov projekta obvezno priložite obrazcu kot priponko (v skeniranem PDF formatu) in jo v primeru, da poročilo ni polno digitalno podpisano, pošljite po pošti na Javno agencijo za raziskovalno dejavnost RS. Nazaj

[3] Napišite povzetek raziskovalnega projekta (največ 3.000 znakov v slovenskem in angleškem jeziku) Nazaj

[4] Napišite kratko vsebinsko poročilo, kjer boste predstavili raziskovalno hipotezo in opis raziskovanja. Navedite ključne ugotovitve, znanstvena spoznanja, rezultate in učinke raziskovalnega projekta in njihovo uporabo ter sodelovanje s

tujimi partnerji. Največ 12.000 znakov vključno s presledki (približno dve strani, velikosti pisave 11). Nazaj

[5] Realizacija raziskovalne hipoteze. Največ 3.000 znakov vključno s presledki (približno pol strani, velikosti pisave 11) Nazaj

[6] V primeru bistvenih odstopanj in sprememb od predvidenega programa raziskovalnega projekta, kot je bil zapisan v predlogu raziskovalnega projekta oziroma v primeru sprememb, povečanja ali zmanjšanja sestave projektne skupine v zadnjem letu izvajanja projekta (obrazložitev). V primeru, da sprememb ni bilo, to navedite. Največ 6.000 znakov vključno s presledki (približno ena stran, velikosti pisave 11). Nazaj

[7] Znanstveni in družbeno-ekonomski dosežki v programu in projektu so lahko enaki, saj se projekna vsebina praviloma nanaša na širšo problematiko raziskovalnega programa, zato pričakujemo, da bo večina izjemnih dosežkov raziskovalnih programov dokumentirana tudi med izjemnimi dosežki različnih raziskovalnih projektov.

Raziskovalni dosežek iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A'' ali A'. Nazaj

[8] Znanstveni in družbeno-ekonomski dosežki v programu in projektu so lahko enaki, saj se projekna vsebina praviloma nanaša na širšo problematiko raziskovalnega programa, zato pričakujemo, da bo večina izjemnih dosežkov raziskovalnih programov dokumentirana tudi med izjemnimi dosežki različnih raziskovalnih projektov.

Družbeno-ekonomski rezultat iz obdobja izvajanja projekta (do oddaje zaključnega poročila) vpišete tako, da izpolnite COBISS kodo dosežka – sistem nato sam izpolni naslov objave, naziv, IF in srednjo vrednost revije, naziv FOS področja ter podatek, ali je dosežek uvrščen v A'' ali A'.

Družbenoekonomski dosežek je po svoji strukturi drugačen, kot znanstveni dosežek. Povzetek znanstvenega dosežka je praviloma povzetek bibliografske enote (članka, knjige), v kateri je dosežek objavljen.

Povzetek družbeno ekonomsko relevantnega dosežka praviloma ni povzetek bibliografske enote, ki ta dosežek dokumentira, ker je dosežek sklop več rezultatov raziskovanja, ki je lahko dokumentiran v različnih bibliografskih enotah. COBISS ID zato ni enoznačen izjemoma pa ga lahko tudi ni (npr. v preteklem letu vodja meni, da je izjemen dosežek to, da sta se dva mlajša sodelavca zaposlila v gospodarstvu na pomembnih raziskovalnih nalogah, ali ustanovila svoje podjetje, ki je rezultat prejšnjega dela … - v obeh primerih ni COBISS ID). Nazaj

[9] Navedite rezultate raziskovalnega projekta iz obdobja izvajanja projekta (do oddaje zaključnega poročila) v primeru, da katerega od rezultatov ni mogoče navesti v točkah 7 in 8 (npr. ker se ga v sistemu COBISS ne vodi). Največ 2.000 znakov vključno s presledki. Nazaj

[10] Pomen raziskovalnih rezultatov za razvoj znanosti in za razvoj Slovenije bo objavljen na spletni strani: http://sicris.izum.si/ za posamezen projekt, ki je predmet poročanja Nazaj

[11] Največ 4.000 znakov vključno s presledki Nazaj

[12] Največ 4.000 znakov vključno s presledki Nazaj

[13] Največ 500 znakov vključno s presledki (velikosti pisave 11) Nazaj

[14] Največ 500 znakov vključno s presledki (velikosti pisave 11) Nazaj

[15] Največ 1.000 znakov vključno s presledki (velikosti pisave 11) Nazaj

Obrazec: ARRS-CRP-ZP/2012-05 v1.00c
04-39-81-67-BE-FD-E7-9A-21-9D-9E-E1-06-56-EC-43-15-C5-8E-3B

Priloga k zaključnemu poročilu

CRP V4-1084
»Uvedba genomske selekcije v slovensko govedorejo na primeru rjave pasme«

Doc. dr. Gregor Gorjanc

Ljubljana, Oktober 2012

KAZALO

1. Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods, G3, 2:425-427

Izdelava in dokumentacija programa (AlphaDrop) za simulacijo genomskih asociacijskih študij in genomske selekcije za potrebe testiranja novih metod genomske selekcije (Hickey in Gorjanc, 2012). Program AlphaDrop in simulirani seti podatkov so prosto dostopni na spletu (http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.111.001297/-/DC1

Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods

John M. Hickey*,[1] **and Gregor Gorjanc**†
*School of Environmental and Rural Science, University of New England, Armidale, 2351 New South Wales, Australia, and
†Department of Animal Science, Biotechnical Faculty, University of Ljubljana, 1230 Domžale, Slovenia

ABSTRACT An approach is described for simulating data sequence, genotype, and phenotype data to study genomic selection and genome-wide association studies (GWAS). The simulation method, implemented in a software package called AlphaDrop, can be used to simulate genomic data and phenotypes with flexibility in terms of the historical population structure, recent pedigree structure, distribution of quantitative trait loci effects, and with sequence and single nucleotide polymorphism-phased alleles and genotypes. Ten replicates of a representative scenario used to study genomic selection in livestock were generated and have been made publically available. The simulated data sets were structured to encompass a spectrum of additive quantitative trait loci effect distributions, relationship structures, and single nucleotide polymorphism chip densities.

KEYWORDS
genome-wide association studies (GWAS)
shared data resources
GenPred
simulation method
quantitative trait loci (QTL)
pedigrees

Simulation studies have made important contributions to the advancement of animal and plant breeding. With many breeding programs now incorporating genomic information at great expense, simulation is both useful and necessary to compare, at low cost, the potential that different analysis methods have to increase the accuracy of estimating breeding values and to compare the alternative structures of breeding programs. Furthermore, simulation can be used to test and benchmark software packages. Recently, many alternative strategies for simulation have been applied within the context of livestock. These strategies use different ways to simulate data, have distributions of quantitative trait loci (QTL) effects, and have different relationship structures. This complicates the comparison of the results and conclusions drawn from the different studies. The first objective of this note was to describe a simple simulation method that can be used to simulate animal or plant genomic data and phenotypes with flexibility in terms of historical population structure, recent pedigree structure, distribution of QTL effects, and with sequence and single nucleotide polymorphism (SNP)-phased alleles and genotypes. The second objective of this note was to provide a set of publically available simulated data sets that cover a spectrum of QTL distributions, relationship structures, and SNP densities. The data were simulated to represent a livestock population and mimic some of the scenarios in which genomic selection is applied.

MATERIALS AND METHODS

Method of simulation

A system to simulate sequence, SNP, and QTL data using a combination of coalescent and gene drop methods was developed. The system is packaged in a Fortran 95 program called AlphaDrop, which calls the Markovian Coalescence Simulator (MaCS) (Chen *et al.* 2009). AlphaDrop has full flexibility in terms of number of chromosomes, QTL, and SNP chips and their density, pedigree structure, and whether the underlying sequence data are outputted. Through the use of MaCS, full flexibility is available in terms of the structure

doi: 10.1534/g3.111.001297
Manuscript received September 30, 2011; accepted for publication November 9, 2011

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.111.001297/-/DC1
[1]Corresponding author: School of Environmental and Rural Science, University of New England, Armidale, 2351 New South Wales, Australia E-mail: john.hickey@une.edu.au

and size of the ancestral population. QTL effects are restricted to being additive and sampled from normal or gamma distributions. MaCS and AlphaDrop are each controlled by a single specification file, examples of which are given in the supporting information, File S1.

Briefly, AlphaDrop starts by setting up the data structures in terms of SNP chips and pedigree. It then calls MaCS, which simulates a sample of haplotypes with sequence information for each chromosome according to the specified ancestral population and mutation and recombination rates. AlphaDrop then drops these haplotypes through a pedigree with a recombination rate assuming 1 recombination event every 100 centimorgans (cM) but no mutation. Internally or externally generated pedigrees can be used. Currently the internal pedigrees are restricted to mammalian species. To simulate data for other species, such as plant species, an externally created pedigree needs to be supplied. The base generation of the pedigree is the most recent generation of the ancestral population simulated using MaCS. Next, the segregating sites are sampled at random to become SNP markers, and a number of SNP chips of different density are provided. The user has full control over the number and density of these chips. The full sequence and phased data can also be outputted if required.

AlphaDrop then selects two samples of segregating sites to possibly become QTL. These are called candidate QTL. The first set comprises a user-specified number of candidate QTL selected at random from across the genome. The second set comprises a user-specified number of candidate QTL selected at random from across the genome with the restriction that the minor allele frequency must be less than a certain threshold. This restriction was designed to facilitate the possibility that QTL have lower minor allele frequency than SNP. Four different traits are then generated assuming an additive genetic model. The first pair of traits is generated using the unrestricted candidate QTL loci. For the first trait (PolyUnres), the allele substitution effect at each QTL locus is sampled from a normal distribution with a mean of zero and standard deviation of one unit. For the second trait (GammaUnres), a random subset of the unrestricted set of candidate QTL loci are selected and the allele substitution effect at each QTL locus is sampled from a gamma distribution with a user-specified shape and scale parameter and a 50% chance of being positive or negative. The second pair of traits (PolyRes and GammaRes) is generated in the same way as the first pair except that the candidate QTL comprise a set with the restriction that their minor allele frequency could not exceed a user specified threshold.

Phenotypes with user-defined heritability are generated for each trait. To ensure that the heritability of the four traits remains constant, the residual variance is scaled relative to the variance of the breeding values of individuals in the base generation of the pedigree, which was given by $\mathbf{a}'\mathbf{a}/(n - 1)$, where $\mathbf{a}$ is a vector of breeding value of individuals in the base generation and n is the number of individuals in that generation.

AlphaDrop efficiently stores sequence information, and this makes the simulation of sequence data in large pedigrees computationally feasible. Gametes comprise strings of 0s and 1s, representing SNP alleles. Gametes can therefore be thought of as large binary numbers and represented as integers. AlphaDrop breaks gametes into haplotypes of a certain length. Each haplotype can be represented as long integer, and these long integers are only decompressed into their binary numbers where a recombination occurs.

Simulated data sets

Ten replicates of a livestock data structure were simulated. The structure was designed to cover a spectrum of QTL distributions, relationship structures, and SNP chip densities and to mimic some of the scenarios in which genomic selection is applied. In each replicate sequence data for 4000 base haplotypes for each of 30 chromosomes was simulated using the MaCS (Chen *et al.* 2009). The 30 chromosomes were each 100 cM in length comprising approximately 10^8 base pairs and were simulated using a per site mutation rate of $2.5{*}10^{-8}$ and an effective population size (Ne) of 100 in the final generation of the sequence simulation. The reduction of Ne in the preceding generations was modeled with a Ne 1000 years ago of 1256, a Ne 10,000 years ago of 4350, and a Ne 100,000 years ago of 43,500 with linear changes in between. This reflects estimates by Villa-Angulo *et al.* (2009) for the Holstein population.

A pedigree was simulated comprising 10 generations of individuals, with 50 sires per generation, 10 dams per sire, and 2 offspring per dam. Base individuals in the pedigree had their gametes randomly sampled from the 4000 haplotypes of the sequence simulation allowing for recombination according to the genetic distance using 1% probability of a recombination event per cM. Subsequent generations in the pedigree had their gametes generated through Mendelian

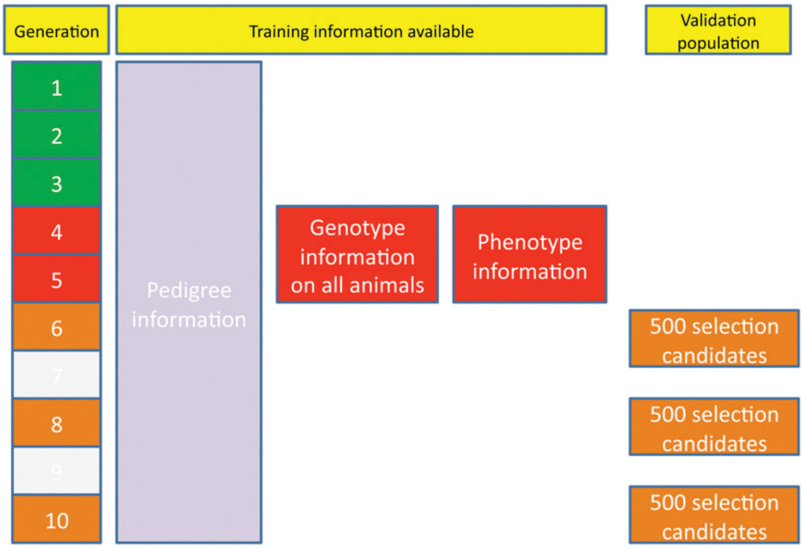

Figure 1 Structure of training and testing data sets.

inheritance with recombination. The total number of segregating sites across the resulting genome was approximately 1,670,000. A set of 9000 segregating sites were randomly selected from the sequence to be used as candidate QTL loci in two different ways, one a randomly sampled set and the other being a randomly sampled set with the restriction that their minor allele frequency could not exceed 0.30. In addition, a random samples of 60,000 and 300,000 segregating sites was selected from the sequence to be used as SNP on two different SNP chips.

Four different traits were simulated assuming an additive genetic model. The first pair of traits was generated using the 9000 unrestricted candidate QTL loci. For the first trait (PolyUnres), the allele substitution effect at each QTL locus was sampled from a normal distribution with a mean of zero and standard deviation of one unit. For the second trait (GammaUnres), a random subset of 900 of the candidate QTL loci were selected and their allele substitution effects at each QTL locus were sampled from a gamma distribution with a shape of 0.4 and scale of 1.66 (Meuwissen *et al.* 2001) and a 50% chance of being positive or negative. The second pair of traits (PolyRes and GammaRes) was generated in the same way as the first pair except that the candidate QTL loci comprised the 9000 with the restriction that their minor allele frequency could not exceed 0.30. Phenotypes with a heritability of 0.25 were generated for each trait.

Training and validation data sets

Subsets of the data were extracted for training and validation. The training set comprised the 2000 individuals in generations 4 and 5 (*i.e.* 1000 animals in each generation). Three validation sets were extracted, consisting of 1500 animals, with 500 animals sampled at random from each of generations 6, 8, and 10. The structure of the training and testing data sets are illustrated in Figure 1.

DISCUSSION

A system to simulate data for the study of genomic selection in livestock and plants was developed. The system, which combines coalescent and gene drop methods, was designed to be simple and flexible. It makes routine simulation of sequence data for large pedigrees possible. Other genome simulation packages are publically available, such as Fregene (Chadeau-Hyam *et al.* 2008), HaploSim (Coster *et al.* 2010), and QMSim (Sargolzaei and Schenkel 2009). However, given that these packages are based on gene dropping approaches they are less computationally efficient in comparison with the combination of coalescent and gene drop approaches presented here. There are important questions relating to the simulation of genomic data that remain to be resolved. It is not clear whether coalescent or gene drop methods generate realistic genomic data and whether simple additive genetic models are sufficient. Like the simulated data from all other packages, the data simulated by AlphaDrop may not fully reflect the structure of real data. However, the presented approach uses realistic mutation rates, recombination rates, evolution of historical effective population sizes, and numbers of nucleotide base pairs to reflect whole genome level sequence. Simulated data would benefit from having standardized methods to validate its quality. Further development of AlphaDrop is ongoing.

ACKNOWLEDGMENTS

We thank Brian Kinghorn, Hans Daetwyler, Julius van der Werf, Sam Clark, and Gustavo de los Campos for their useful discussions. This research was funded by the Australian Research Council project LP100100880 of which Genus Pty, Aviagen LTD, and Pfizer are cofunders. The simulated data sets are available in the supporting information. AlphaDrop, and its source code, is available from http://sites.google.com/site/hickeyjohn/workstuff.
The authors declare that they have no competing interests.

LITERATURE CITED

Chadeau-Hyam, M, C. J. Hoggart, P. F. O'Reilly, J. C. Whittaker, M. De Lorio *et al.*, 2008 Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. BMC Bioinformatics 9: 364.

Chen, G. K., P. Marjoram, and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. 19: 136–142.

Coster, A. J., W. M. Bastiaansen, M. P. L. Calus, J. A. M. Van Arendonk, and H. Bovenhuis, 2010 Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet. Sel. Evol. 42: 9.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Villa-Angulo, R., L. K. Matukumalli, C. A. Gill, J. Choi, C. P. Van Tassell *et al.*, 2009 High-resolution haplotype block structure in the cattle genome. BMC Genet. 10: 19.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: A large scale genome simulator for livestock. Bioinformatics 25: 680–681.

Edited by Dirk-Jan de Koning and Lauren M. McIntyre

2. Genomic evaluations using similarity between haplotypes, Journal of Animal Breeding and Genetics (sprejeto v objavo)

Preučitev uporabe različnih virov genetskih informacij (rodovniki, označevalci SNP, haplotipi in vzorčni geni - QTL) za genetsko vrednotenje kompleksnih lastnosti (Hickey in sod., 2012). S pomočjo simulacij smo pokazali praktične omejitve genetskega vrednotenja glede na različne vire informacij. Četudi bi poznali vse gene, ki so vplivajo na kompleksne lastnosti, je v sistemu genetskega vrednotenja prisotna napaka, saj je potrebno vpliv genov oceniti za vsako populacijo ločeno. Slednje v sistem vnaša statistično napako, ki pa je manjša kot pri drugih virih genetskih informacij (rodovniki, označevalci SNP in haplotipi). Pri tem delu smo razvili metodo, ki omogoča uporabo haplotipov (poljubne dolžine) namesto označevalcev SNP za izgradnjo genomske matrike sorodstva pri genetskem vrednotenju.

Genomic evaluations using similarity between haplotypes

J. M. Hickey[1], B. P. Kinghorn[1], B. Tier[2], S. A. Clark[1,3], J. H. J. van der Werf[1,3] & G. Gorjanc[4]

1 School of Environmental and Rural Science, University of New England, Armidale, Australia

2 Animal Genetics and Breeding Unit, University of New England, Armidale, Australia

3 Cooperative Research Centre for Sheep Industry Innovation, Armidale, Australia

4 Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Domžale, Slovenia

Correspondence: J. M. Hickey, School of Environmental and Rural Science, University of New England, Armidale, Australia; E-mail: john.hickey@une.edu.au

Long-range phasing and haplotype library imputation methodologies are accurate and efficient methods to provide haplotype information that could be used in prediction of breeding value or phenotype. Modeling long haplotypes as independent effects in genomic prediction would be inefficient due to the many effects that need to be estimated and phasing errors, even if relatively low in frequency, exacerbate this problem. One approach to overcome this is to use similarity between haplotypes to model covariance of genomic effects by region or of animal breeding values. We developed a simple method to do this and tested impact on genomic prediction by simulation. Results show that the diagonal and off-diagonal elements of a genomic relationship matrix constructed using the haplotype similarity method had higher correlations with the true relationship between pairs of individuals than genomic relationship matrices built using unphased genotypes or assumed unrelated haplotypes. However, the prediction accuracy of such haplotype based prediction methods was not higher than those based on unphased genotype information.

Keywords: genomic selection, haplotypes, similarity

INTRODUCTION

Long-range phasing (Kong *et al.* 2008) and haplotype library imputation methods provide an accurate framework for both genome-wide phasing (AlphaPhase - Hickey *et al.* 2011) and imputation (AlphaImpute - Hickey *et al.* 2012), which is computationally feasible even for large data sets. This could facilitate the use of haplotype information routinely in livestock breeding programs. Using haplotypes and their effects is appealing as breeding programs can, in some sense, be thought of as systems to construct individuals in the next generation from favourable haplotypes and remove the unfavourable haplotypes from the population. Haplotypes provide more accurate information about genomic regions being identical by descent (IBD) than SNP, and long haplotypes have greater ability to do this compared to short ones. Furthermore, haplotype effects may represent the effects of multiple QTL that maybe hard to separate and these could include .local epistatic effects among genes within the segment.

AlphaPhase has higher phasing accuracy (i.e. >97% of all alleles correctly phased) than statistically based phasing methods such as fastPHASE (Scheet & Stephens 2006) which uses a Hidden Markov Model (Hickey *et al.* 2011). However, estimating the effects of the resulting haplotypes directly is hampered by a number of issues generic to the use of haplotypes that are exacerbated by the use of long haplotypes (which are necessary for long-range phasing) and phasing errors. Firstly, treating haplotype effects as independent greatly increases the number of effects to be estimated for a given data set. Secondly, rare haplotypes have very few phenotypic records, which results in unreliable estimates of their effects. The use of long haplotypes and the occurrence of phasing errors increase this problem. Moreover, treating haplotype effects as independent ignores the coalescent process that generated the existing haplotypes. Closely related haplotypes are expected to have similar effects due to sharing regions that are IBD. Specifically, long-range phasing is generally based on the haplotypes of 10 to 20 centimorgans in length. Haplotypes of this length are on average shared with individuals separated by no more than 4 meiotic events. This severely reduces both the individuals who can contribute phenotypic information for genomic prediction and also the distance (in terms of relationships) for which genomic predictions can be used, unless a method that models the relationship between such haplotypes is used.

Some studies exploring the benefit of using haplotype information for genomic prediction have treated haplotypes as independent effects (Hayes *et al.* 2007; Villumsen *et al.* 2009). This approach would be inefficient when considering long haplotypes because treating these haplotypes as being independent would lose a lot of information. Alternatively, approaches were used for modelling IBD relationship between haplotypes (Meuwisen & Goddard 2001; Calus *et al.* 2008; Druet & Georges 2008). These methods are computationally intensive and they are not easily compatible with the heuristic nature of the long-range phasing and haplotype library imputation framework. For example, the method of Druet & Georges (2008) is based on a Hidden Markov model, which is similar to that used in fastPHASE (Scheet & Stephens 2006), and is a probabilistic rather than heuristic model.

The objective of this research was to develop a simple and computationally inexpensive method that approximates the similarity between long haplotypes generated by long-range phasing with the aim to increase the accuracy of genomic prediction. The accuracy of such a method was compared with other widely used genomic prediction methods that ignore the haplotype information.

MATERIALS AND METHODS

Haplotype similarity

The haplotype similarity method is based on a genomic relationship matrix ($\mathbf{G}_{\mathrm{HAP}}$) amongst individuals that is built using haplotypes. To build this matrix the genome is divided into *k* segments of *n* SNP in length. For each segment *k* a $\mathbf{G}_{\mathrm{HAP,k}}$ is built. The genome wide $\mathbf{G}_{\mathrm{HAP}}$ is then constructed as the mean of all $\mathbf{G}_{\mathrm{HAP,k}}$ across the *k* genomic segments (Van Arendonk *et al.* 1994). Each element of $\mathbf{G}_{\mathrm{HAP,k}}$ is filled as the sum of the elements of a matrix $\mathbf{W}_{ij}$ divided by 2. $\mathbf{W}_{ij}$ is a 2×2 matrix with elements measuring the haplotype similarity of individual *i* and *j*, and is a haplotype version of the classical gametic relationship matrix between a pair of individuals (Jamrozik & Schaeffer 1991). The elements of $\mathbf{W}_{ij}$ are based on the elements of **H**, a haplotype similarity matrix at the genome segment *k*. An example of how $\mathbf{G}_{\mathrm{HAP}}$ is constructed is given in detail in Appendix A.

The similarity matrix **H** can be defined in different ways. In this study three different definitions of **H** are reported. $\mathbf{H}_1$ is based simply on the proportion of matching SNP alleles for each pair of haplotypes. $\mathbf{H}_2$ rewards segments of consecutive matching SNP alleles between haplotypes in the definition of their similarity, with longer segments getting greater rewards. $\mathbf{H}_3$ is constructed in the same way as $\mathbf{H}_2$ but the similarity measure also accounts for the allele frequencies of the matching SNP alleles. Accounting for allele frequency was an attempt to penalize matching segments of SNP alleles in pairs of haplotypes that may arise due to SNP alleles being common, therefore being IBS rather than IBD. Genomic relationship matrices using the different definitions of **H** are denoted as $\mathbf{G}_{\mathrm{HAP1}}$, $\mathbf{G}_{\mathrm{HAP2}}$, and

$\mathbf{G}_{HAP3}$. Further details and a worked example are given in Appendix A and prototype R code is given in supplementary material.

Comparison to other methods

The haplotype similarity method was compared to building the genomic relationship matrix ($\mathbf{G}_{HAPIdentity}$) using an identity matrix instead of **H**. This implies that the effects of haplotypes are independent and is a GBLUP version of the methods of Hayes *et al.* (2007) and Villumsen *et al.* (2009). Comparisons were also made to a standard pedigree based relationship matrix (**A**), two genomic relationship matrices based on individual SNP genotypes ($\mathbf{G}_{SNPV}$ – VanRaden (2008) and $\mathbf{G}_{SNPY}$ - Yang *et al.* (2010)), and a genomic relationship matrix built using the true underlying quantitative trait loci (**QTL**) genotypes and their effects ($\mathbf{G}_{QTL}$). For the purpose of comparison of variance component estimates two versions of $\mathbf{G}_{QTL}$ were constructed: without ($\mathbf{G}_{QTL}$) and with ($\mathbf{G}_{QTLS}$) allele frequency scaling (VanRaden 2008). The method of (VanRaden 2008) scales the genomic relationship matrix according to the allele frequencies in the current population. $\mathbf{G}_{SNPV}$, $\mathbf{G}_{QTL}$, and $\mathbf{G}_{QTLS}$ had a small number (0.01) added to their diagonals to ensure that they were of full rank.

The various relationship matrices were evaluated by comparing their elements to the elements of $\mathbf{G}_{QTL}$, which were assumed to reflect the true relationship at QTL loci between individuals averaged over all the QTL. The correlation between estimated breeding values and true breeding values was calculated in order to assess the accuracy of estimated breeding values.

Simulations

Sequence data for 4000 base haplotypes for each of 30 chromosomes was simulated using the Markovian Coalescence Simulator (MaCS) (Chen *et al.* 2009) and AlphaDrop (Hickey & Gorjanc 2012). The chromosomes were each 100 cM in length comprising approximately 10^8 base pairs and were simulated using a per site mutation rate of 2.5×10^{-8} and an effective population size (N_e) of 100 in the final generation of the sequence simulation. The reduction of N_e in the preceding generations was modelled with a N_e=1,256 1,000 years ago, a N_e=4,350 10,000 years ago, and a N_e=43,500 100,000 years ago with linear changes in between. This reflects estimates by Villa-Angulo *et al.* (2009) for the Holstein cattle population.

After sequence simulation a pedigree was simulated comprising 10 generations of individuals, with 50 sires per generation, 10 dams per sire, and 2 offspring per dam. Base individuals in the pedigree had their gametes randomly sampled from the 4000 haplotypes of the sequence simulation allowing for recombination according to the genetic distance using 1% probability of a crossover event per cM. Subsequent generations in the pedigree had their gametes generated through Mendelian inheritance with recombination. The total number of segregating sites across the resulting genome was approximately 1,670,000. A random sample of 60,000 segregating sites was selected, with a restriction that 2,000 markers are sampled from each of the 30 chromosomes, from the sequence to be used as SNP on a 60,000 SNP array. In addition a set of 9,000 segregating sites were randomly selected, with a restriction that 300 are sampled from each of the 30 chromosomes, from the sequence to be used as candidate QTL loci in two different ways: (i) a randomly sampled set of loci; and (ii) a randomly sampled set of loci with the restriction that minor allele frequency could not exceed 0.30.

Four different traits were simulated assuming an additive genetic model. The first pair of traits was generated using the 9,000 candidate QTL loci that were sampled without restriction. For the first trait (GaussUnres) the allele substitution effect at each QTL locus was sampled from a normal distribution with a mean of zero and standard deviation of 1.0 divided by the square root of the number of QTL. For the second trait (GammaUnres) a random subset of 900 of the candidate QTL loci were selected and their allele substitution effects were sampled from a gamma distribution with a shape of 0.4 and scale of 1.66 (Meuwissen *et al.* 2001) divided by the square root of the number of QTL and a 50% chance of the effect being either positive or negative. The second pair of traits (GaussRes and GammaRes) was generated in the same way as the first pair except that the candidate QTL loci comprised the 9,000 with the restriction on minor allele frequency. From these data the true breeding value of an individual was obtained as a sum of all QTL substitution effects carried by that individual.

The way in which the data were simulated gives two reference values for the base population genetic variance. The first is the genetic variance in a conceptual infinitely-sized population of unrelated gametes, denoted as SIM_{CP} (simulated genetic variance in the conceptual population). The second is the actual genetic variance among base animals, denoted as SIM_{PB} (genetic variance in the pedigree base population). Phenotypes with a heritability of 0.25 at the base generation of pedigree were generated for each trait by scaling the residual variance relative to the variance of the true breeding values of individuals in the base generation of the pedigree, which was given by **a'a**/(n-1), where **a** is a vector of true breeding values of individuals in the base generation of the pedigree and n is the number of individuals in that generation. Ten replicates of each scenario were simulated.

Haplotype lengths for analysis of simulated data

The simulated data were analysed using eight different haplotype lengths. These were 5, 10, 20, 50, 100, 200, 400, and 2000 SNP long resulting in 12000, 6000, 3000, 1200, 600, 300, 150, and 30 genomic segments, respectively. Since a 60,000 SNP density was used the 2000 SNP long haplotypes spanned entire chromosomes (100 cM). Of the remaining haplotype definitions the longest (400 SNP) spanned 20 cM while the shortest (5 SNP) spanned 0.25 cM. Phase was assumed to be known and the resulting haplotypes were used to build the different **H** matrices. Summary statistics relating to the numbers of haplotypes per genomic segment, total number of haplotypes per genome, and the distribution of haplotype frequencies were obtained from the first replicate averaged across the 30 chromosomes.

Training and validation data sets

To test the accuracy of estimation of breeding values subsets of the data were extracted for training and validation. The training set comprised the 2000 individuals in generations 4 and 5. Three validation sets were extracted. The first (Gen6), the second (Gen8), and the third (Gen10) set each comprised of 500 individuals sampled at random from generation 6, 8, and 10, respectively. This allowed the accuracy to be tested in very closely related individuals to the training population as well as in individuals who were less related.

Models

The analysis models were fitted using a mixed model which can be generically written as:

$$\mathbf{y}=\mathbf{Xb}+\mathbf{Zu}+\mathbf{e},$$

where **y** is a vector of phenotype records, **b** is a vector of fixed effects (intercept only), **u** is a vector of breeding values, and **e** is a vector of residuals, while **X** and **Z** are incidence matrices linking the phenotype records to location parameters. The analysis models differed in the specification of the covariance structure for **u**, e.g., $Var(\mathbf{u}) = \mathbf{G}\sigma_g^2$. The different definitions of **G** are described above. All analyses were carried out using ASReml (Gilmour *et al.* 2006).

RESULTS

There was an exponential increase in the average number of distinct haplotypes per genome segment with the increasing haplotype length (Table 1). The total number of distinct haplotypes per genome also increased with the increasing haplotype length up to a length between 200 and 400 SNP (10 and 20 cM). For longer haplotypes the total number of distinct haplotypes decreased which could be attributed to the limited number of individuals in the data set.

Table 1 Summary statistics of simulated haplotypes

Haplotype length (SNP)	5	10	20	50	100	200	400	2000
Haplotype length (cM)	0.25	0.5	1	2.5	5	10	20	100
Number of genomic segments	12000	6000	3000	1200	600	300	150	30
Average number of haplotypes per segment	7	17	43	154	375	827	1653	5133
Total number of haplotypes per genome (×1000)	84	100	130	185	225	248	248	154

The correlations between the diagonal and off-diagonal elements of $\mathbf{G}_{\mathrm{QTL},}$ (built using the QTL for the GaussUnres trait) with **A**, $\mathbf{G}_{\mathrm{SNPV}}$, $\mathbf{G}_{\mathrm{SNPY}}$, $\mathbf{G}_{\mathrm{HAPIdentity}}$, and different $\mathbf{G}_{\mathrm{HAP}}$ are shown in Figure 1. Variation in correlations between replicates was very small. Results for GaussRes, GammaUnres, and GammaRes traits were similar and are shown in the supplementary material. The low correlation between $\mathbf{G}_{\mathrm{QTL}}$ and **A** is attributable to many of the elements of **A** being zero. $\mathbf{G}_{\mathrm{HAP1}}$ had the highest correlation for both diagonal and off-diagonal elements and it was invariant to haplotype length. $\mathbf{G}_{\mathrm{HAPIdentity}}$ and $\mathbf{G}_{\mathrm{HAP2}}$ had lower correlations as haplotype length increased, and correlations were lowest for $\mathbf{G}_{\mathrm{HAPIdentity}}$. $\mathbf{G}_{\mathrm{HAP3}}$ had lower correlations for the diagonal elements compared to $\mathbf{G}_{\mathrm{HAP1}}$, $\mathbf{G}_{\mathrm{HAP2}}$, and $\mathbf{G}_{\mathrm{SNPY}}$ and was worse than $\mathbf{G}_{\mathrm{SNPV}}$ for off-diagonal elements. $\mathbf{G}_{\mathrm{SNPY}}$ had higher correlations compared to $\mathbf{G}_{\mathrm{SNPV}}$ for the diagonal elements, while for the off-diagonal elements little difference was observed. For all haplotype lengths $\mathbf{G}_{\mathrm{HAP3}}$ was not positive definite while the other relationship matrices were positive definite.

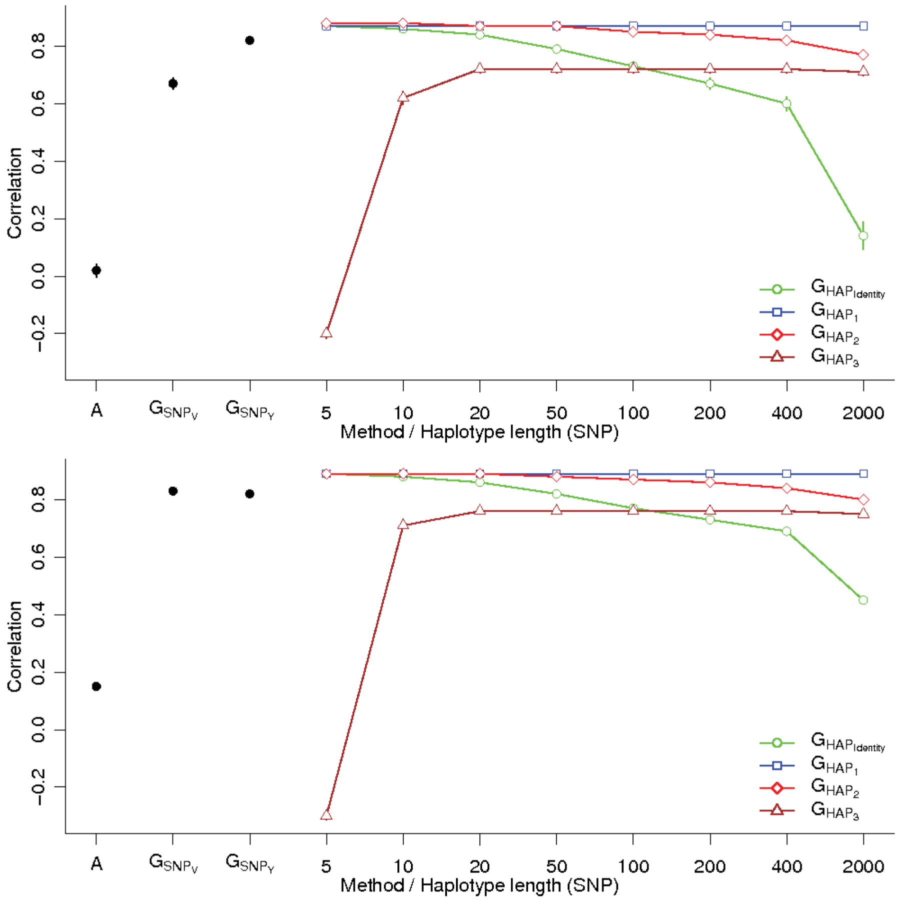

Figure 1 Average correlation (± standard deviation) over replicates between diagonal (top) and off-diagonal (bottom) elements of $\mathbf{G}_{\mathrm{QTL}}$ and other matrices for the GaussUnres trait

Variance components for the different relationship matrices for the GaussUnres trait are given in Figure 2. For comparison $\mathrm{SIM}_{\mathrm{CP}}$ and $\mathrm{SIM}_{\mathrm{PB}}$, the conceptual population and pedigree base genetic variances, are also shown. The different methods partitioned the genetic and residual variance differently and estimated the genetic variance for different base populations. The genetic variance estimated using, $\mathbf{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$ was similar to $\mathrm{SIM}_{\mathrm{PB}}$ and to the variance estimated by **A**. $\mathbf{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$ appeared to reflect variance in a recent base, most likely the population at the start of the simulated pedigree, or the current population of the genotyped animals with phenotypes. The genetic variance estimated using $\mathbf{G}_{\mathrm{HAP1}}$ and $\mathbf{G}_{\mathrm{QTL}}$ was similar to $\mathrm{SIM}_{\mathrm{CP}}$, which was much higher than $\mathrm{SIM}_{\mathrm{PB}}$. In contrast, $\mathbf{G}_{\mathrm{QTLS}}$ estimated a variance in a recent base population, possibly the current population, which is to be expected because the scaling is based on the allele frequencies in the current population. Results obtained with $\mathbf{G}_{\mathrm{HAPIdentity}}$ were harder to interpret. When constructed using haplotypes of intermediate length estimates appeared to recover the genetic variance in a recent base generation. However, when constructed using long haplotypes the separation of genetic and residual variance became an issue, which was to be expected given that extremely long haplotypes are likely to be unique to each animal and thus confounded with the residual term. Genetic variance obtained with $\mathbf{G}_{\mathrm{HAP1}}$ was not sensitive to the haplotype length. Genetic variance obtained with $\mathbf{G}_{\mathrm{HAP2}}$ was sensitive to haplotype length. With short haplotypes the estimates of genetic variance were similar to the value in a conceptual population of unrelated gametes ($\mathrm{SIM}_{\mathrm{CP}}$), but the estimates decreased almost to the level of more recent base ($\mathrm{SIM}_{\mathrm{PB}}$) as haplotype lengths increased.

The accuracy of the estimated breeding values for the GaussUnres trait and the GammaUnres trait are shown in Figures 4 for all methods except $\mathbf{G}_{\mathrm{HAP3}}$. Results for the other traits are shown in the supplementary material. $\mathbf{G}_{\mathrm{QTL}}$ was always the best performing method and more differences between methods were seen for the traits with a gamma distribution of QTL substitution effects. The ranking of $\mathbf{G}_{\mathrm{SNPV}}$, $\mathbf{G}_{\mathrm{SNPY}}$, $\mathbf{G}_{\mathrm{HAPIdentity}}$, and other $\mathbf{G}_{\mathrm{HAP}}$ matrices in terms of accuracy was not greatly affected by the degree of relationship between the training and validation individuals, as measured by the decay in accuracy with increasing numbers of generations between the training and validation populations. The ranking of methods was also not affected by the underlying genetic model. The accuracy of estimated breeding values using genomic information outperformed pedigree information. There was little difference in the performances of $\mathbf{G}_{\mathrm{SNPV}}$ and $\mathbf{G}_{\mathrm{SNPY}}$. With short haplotypes the haplotype based methods performed as well as $\mathbf{G}_{\mathrm{SNPV}}$ or $\mathbf{G}_{\mathrm{SNPY}}$. With longer haplotypes $\mathbf{G}_{\mathrm{HAP}}$ performed as well as $\mathbf{G}_{\mathrm{SNPV}}$ or $\mathbf{G}_{\mathrm{SNPY}}$, while $\mathbf{G}_{\mathrm{HAPIdentity}}$ had significantly poorer performance.

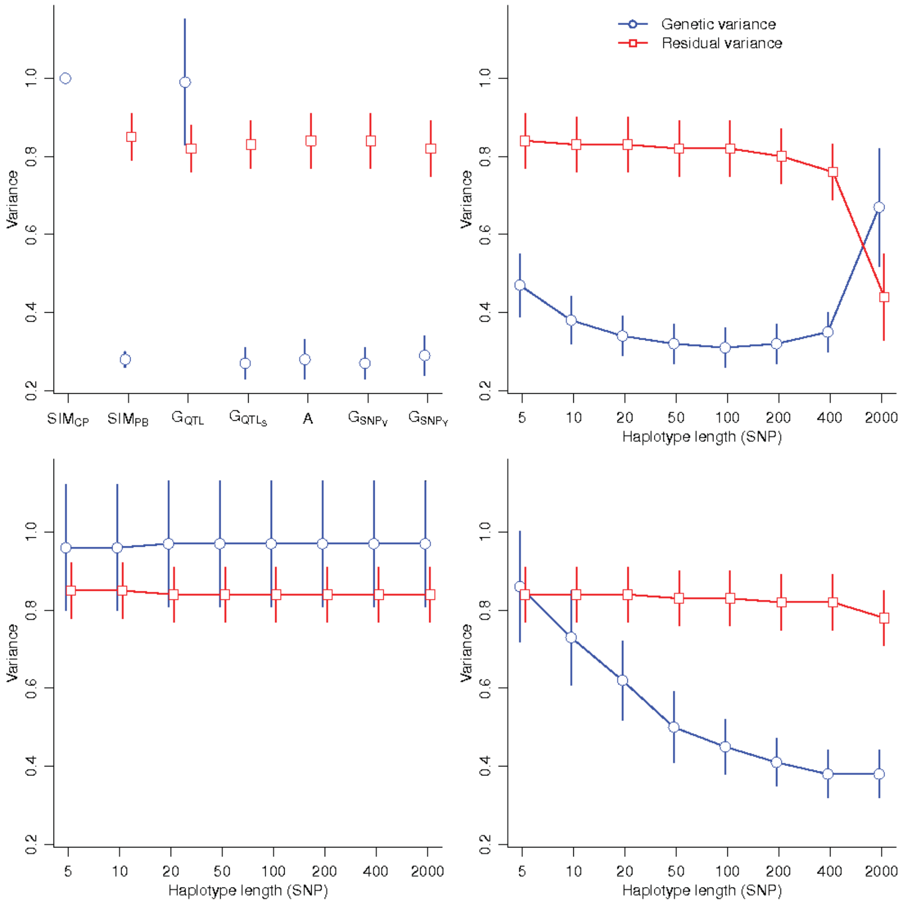

Figure 2 True and estimated genetic and residual variance components (average ± standard deviation over replicates) for the GaussUnres trait estimated using the different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathrm{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathrm{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathrm{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

The simple method of calculating haplotype similarity ($\mathbf{G}_{\mathrm{HAP1}}$) had the same or higher accuracy of estimated breeding values than $\mathbf{G}_{\mathrm{HAP2}}$. Given this result the simple haplotype similarity matrix $\mathbf{H}_1$ has been included in the standard output of the software package AlphaPhase (Hickey *et al.* 2011), and a wrapper has been written to produce the resulting $\mathbf{G}_{\mathrm{HAP1}}$ matrix.

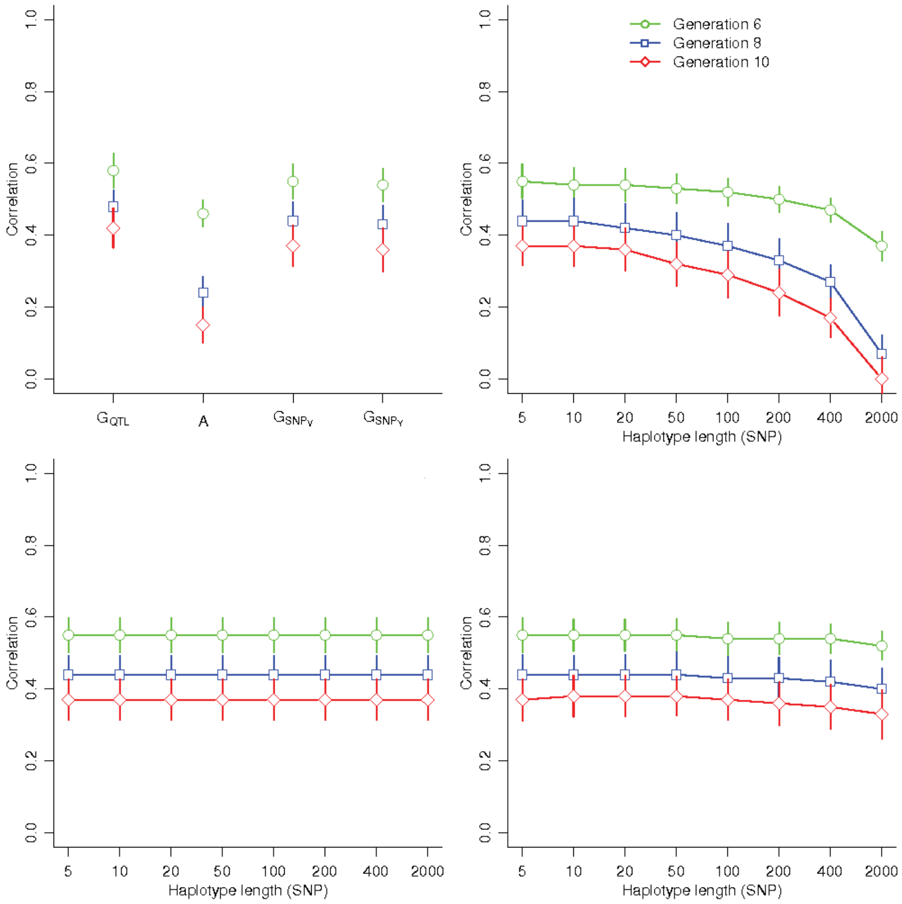

Figure 3 Average accuracy (± standard deviation) over replicates of estimated breeding values for the GaussUnres trait when estimated using different relationship matrices: (top-left) $\mathbf{G}_{\text{QTL}}$, pedigree (**A**), $\mathbf{G}_{\text{SNPV}}$, and $\mathbf{G}_{\text{SNPY}}$, (top-right) $\mathbf{G}_{\text{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\text{HAP1}}$, and (bottom-right) $\mathbf{G}_{\text{HAP2}}$ using 8 different haplotype lengths

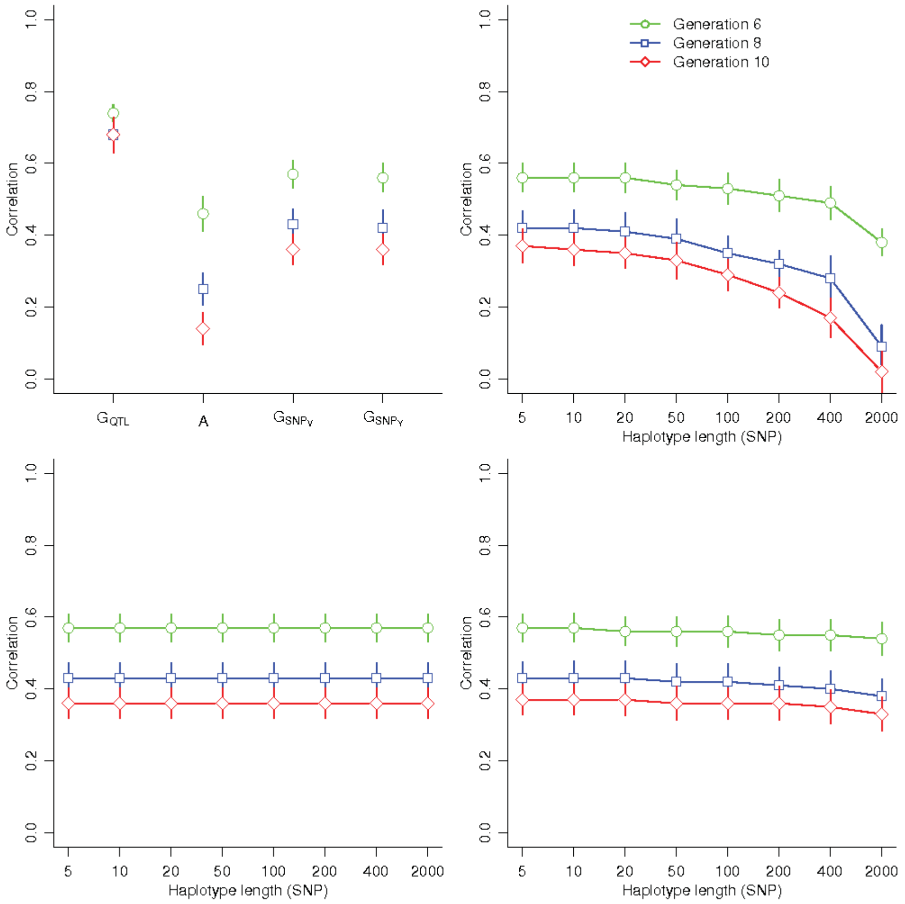

Figure 4 Average accuracy (± standard deviation) over replicates of estimated breeding values for the GammaUnres trait when estimated using different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathbf{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathbf{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathbf{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

DISCUSSION

The haplotype similarity method allows the haplotypes generated by haplotype library imputation to be used directly for genomic prediction. It is a simple and computationally inexpensive approach that fits within the framework of the long-range phasing and haplotype library imputation method. While the method did not result in more accurate estimated breeding values, it resulted in genomic relationship matrices that had higher correlations with the true relationship at the QTL loci than matrices built using SNP (e.g. the method of

VanRaden 2008) indicating that it is a more accurate way of parameterising genomic relationships than a method based on accumulating information from single SNP genotypes. Constructing genomic relationship matrices using haplotype similarity was better than treating haplotypes as independent effects (e.g. Villumsen *et al.* 2009) and this was especially evident for longer haplotypes. Using haplotype similarity would also enable rare haplotypes or unique haplotypes that could be created by small proportions of phasing errors to borrow information from more frequent haplotypes that are more likely to be IBD for some sub-segment. Longer haplotypes could capture genetic variance controlled by complex mechanisms such as local epistasis that may be present in real data. There is considerable biological evidence for the importance of local epistasis (Clark 2004).

The definition of haplotype similarity needs to be refined as it parameterises IBS relationships between haplotypes under the assumption that these are correlated with the underlying IBD relationships. $\mathbf{G}_{\mathrm{QTL}}$ always outperformed $\mathbf{G}_{\mathrm{HAP}}$ and $\mathbf{G}_{\mathrm{SNP}}$ in terms of accuracy of prediction, suggesting that the correlation between IBS and IBD in both $\mathbf{G}_{\mathrm{HAP}}$ and $\mathbf{G}_{\mathrm{SNP}}$ is not optimal for the maximisation of prediction accuracy. The simplest method of describing haplotype similarity (which did not account for segments of consecutive matches) gave the best results, suggesting that further research is needed to refine the use of segments of consecutive matching alleles and the allele frequency at these loci in the definition of haplotype similarity. In addition the haplotype similarity method required the haplotype lengths to be set. Ideally, a method would use variable haplotype length to better capture IBD information from regions with varying haplotype diversity. Routes to doing this will consider chromosome-wide information for IBD inference at each locus but will require whole chromosome alignment of haplotypes at adjacent genomic regions, which is not a trivial task within the long-range phasing and haplotype library imputation framework (Hickey *et al.* 2011), because this framework phases chromosome segments independently.

The correlations between the elements of the different relationship matrices with the true relationships at the QTL suggest that the haplotype similarity method is better at capturing IBD information than the commonly used SNP based genomic relationship matrices (e.g. VanRaden 2008). This was especially evident with the diagonal elements, meaning that inbreeding at the QTL is better captured by the haplotype similarity. This greater precision was facilitated by the use of between 150,000 and 250,000 haplotypes as opposed to 60,000 SNP. Despite this there was no advantage for the IBD method in genomic prediction compared to $\mathbf{G}_{\mathrm{SNP}}$. However both methods gave similar accuracies as $\mathbf{G}_{\mathrm{QTL}}$, suggesting they are close to the asymptotic performance for a data set of this size. It may have been unrealistic to expect that the haplotype similarity method would outperform individual SNP based methods under the genetic model simulated in this study which involved QTL that were IBS, therefore having the same additive effect regardless of genetic background, family, and sub-population. Further study is required using simulated genetic models involving epistatic interactions, epigenetic factors, or other causes of family specific QTL effects that may favour IBD methods. When local epistasis is an important contributor to genetic variation it can be expected that haplotypes can improve the accuracy of genomic prediction. There is a least some evidence that local epistasis is important, for example non linear prediction models, such as Reproducing Kernel Hilbert Space models (Gianola & van Kaam 2008), which weight information on close relatives more strongly than that from distant relatives are sometimes more accurate than the standard linear prediction models (de los Campos *et al.* 2009).

With $\mathbf{G}_{\mathrm{QTL}}$ the accuracies were not very high for the GaussUnres and GaussRes trait, while higher values were observed for the GammaUnres and GammaRes trait. The later is due to the smaller number of QTL with larger substitution effects. However, in all the scenarios $\mathbf{G}_{\mathrm{QTL}}$ did not produce estimates of perfect accuracy as there is still the need to infer all QTL substitution effects from obtained data even though QTL allele/genotype information is available. During the estimation process accumulated noise over so many polymorphisms is propagated to breeding values and their lower accuracy.

The estimates of genetic variance from $\mathbf{G}_{\mathrm{HAP1}}$ and $\mathbf{G}_{\mathrm{QTL}}$ were much higher than for the other methods. In some sense the haplotype similarity method is the same as estimating the genetic variance using a much deeper ancestral pedigree and hence it is not surprising that the variance recovered is much higher. However the estimated values of this variance were very close to the actual variance simulated for the QTL effects. This suggests that the haplotype similarity method recovers genetic variance that is present in a conceptual population of unrelated gametes. It suggests that a large proportion of the genetic variance is locked up in the limited permutations of QTL loci that all of the individuals in a population carry (i.e., no gamete carries the favourable allele at all QTL and no gamete carries the unfavourable allele at all QTL). Breeding programs may need to be designed to exploit this. Animals could be mated based on the likely QTL alleles that they carry in order to help reveal the hidden variance. However, for many QTL involved, much recombination will be required to unlock most of this variation, and that will take many generations under any biologically conventional breeding system. Estimating this variance is also possible using the IBS SNP based methods that do not scale the genomic relationship matrices for SNP allele frequency.

CONCLUSION

A simple method to account for haplotype similarity was developed, which fits within the long-range phasing and haplotype library imputation framework. The method can be used to set up a genomic relationship matrix that parameterizes IBD more precisely than methods that ignore haplotype similarity or use SNP individually. With this method haplotypes of any size and therefore also any number can be fitted without decrease in predictive accuracy that is observed with long haplotypes. The haplotype similarity method models the genetic variance in a conceptual population of unrelated gametes of which the current population is a subset and consequently reveals more genetic variance that can be used for selection if breeding programs are appropriately designed.

AVAILABILITY

A computer program, written in Fortran 95, which implements haplotype similarity method as a post processing of AlphaPhase output is freely available from http://sites.google.com/site/hickeyjohn.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

The research was funded by the Australian Research Council project LP100100880 of which Genus Pty, Aviagen LTD, and Pfizer are co-funders. Matt Cleveland is acknowledged for useful discussions.

REFERENCES

Boichard D., Guillaume F., Baur A., Croiseau P., Rossignol M.N., Boscher M.Y., Druet T., Genestout L., Eggen A., Journaux L., Ducrocq V., Fritz S. (2010) Genomic Selection In French Dairy Cattle. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production: 1-6 August 2010; Leipzig*. pdf 07-16.

Browning S.R., Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81,** 1084-1097.

Calus M.P., Meuwissen T.H., de Roos A.P., Veerkamp R.F.(2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, **178**, 553-561.

Chen G.K., Marjoram P., Wall J.D. (2009) Fast and flexible simulation of DNA sequence data. *Genome Res.*, **19**, 136-142.

Clark A.G. (2004) The role of haplotypes in candidate gene studies. *Genet. Epidemiol.*, **27**, 321-333.

de los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K. A., Cotes J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, **182**, 375–385.

Druet T., Georges M. (2010) A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, **184**, 789–798.

Fernando R.L., Grossman M. (1989) Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.*, **21**, 467-477.

Gianola D., van Kaam, J.B.C.H.M. (2008) Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics*, **178**, 2289-2303.

Gilmour A.R., Gogel B.J., Cullis B.R., Thompson R. (2006) *ASReml User Guide Release 2.0*. Hemel Hempstead: VSN International Ltd.

Hayes B.J., Chamberlain A.J., McPartlan H., Macleod I., Sethuraman L., Goddard M.E. (2007) Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet. Res.*, **89**, 215-220.

Hayes B.J., Bowman P.J., Chamberlain A.C., Goddard M.E. (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.*, **92**, 433-443.

Hickey J.M., Kinghorn B.P., Tier B., Wilson J.F., Dunstan N., van der Werf J.H.J. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.*, **43**, 12.

Hickey J.M., Gorjanc G (2012) Simulated data for genomic selection and GWAS using a combination of coalescent and gene drop methods. *G3*, **2**, 425-427.

Hickey J.M., Kinghorn B.P., Tier B., van der Werf J.H.J., Cleveland M.A. (2012) An imputation method which results in an alternative single stage genomic evaluation system. *Genet. Sel. Evol.*, **44**, 9.

Jamrozik J., Schaeffer L.R.(1991) An equivalent gametic model for animal dominance genetic linear model. *J. Anim. Breed. Genet.*, 1**08**, 3343-3348.

Kong A., Masson G., Frigge M.L., Gylfason A., Zusmanovich P., Thorleifsson G., Olason P.I., Ingason A., Steinberg S., Rafnar T., Sulem P., Mouy M., Jonsson F., Thorsteinsdottir U., Gudbjartsson D.F., Stefansson H., Stefansson K. (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.*, **40**, 1068-1075.

Lee S.H., van der Werf J.H.J. (2004) The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. *Genet. Sel. Evol.*, **36**, 145-161.

Meuwissen T.H.E., Goddard M.E. (2001) Prediction of identity by descent probabilities from marker-haplotypes. *Genet. Sel. Evol.*, **33**, 605-634.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819-1829.

Pritchard J.K., Stephens M., Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

Scheet P., Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629-644.

Van Arendonk J.A., Tier B., Kinghorn B.P. (1994) Use of multiple genetic markers in prediction of breeding values. *Genetics*, **137**, 319-29.

VanRaden P, (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci*., **91**, 4414-4423.

VanRaden P., Van Tassell C., Wiggans G., Sonstegard T., Schnabel R., Taylor J., Schenkel F. (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.*, **92**, 16-24.

Villa-Angulo R., Matukumalli L.K., Gill C.A., Choi J., Van Tassell C.P., Grefenstette J.J. (2009) High-resolution haplotype block structure in the cattle genome. *BMC Genetics*, **10**, 19.

Villumsen T.M., Janss L., Lund M.S. (2009) The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet*., **126**, 3-13.

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. *Nat. Genet*., **42**, 565-569.

APPENDIX A. Construction of H and $\mathbf{G}_{\mathbf{HAP}}$ matrices

A worked example with description of how the $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{H}_3$ haplotype similarity matrices and the resulting $\mathbf{G}_{\mathrm{HAP}}$ matrices are constructed is outlined. An algorithm for constructiong these matrices is presented as R prototype code in the supplement. The example data set comprises three individuals that have a genome comprising 10 loci (Table 2). The third individual is homozygous for haplotype 4. For simplicity the genome consists of a single genomic region.

Table 2 Example data set

			Locus harbouring SNP allele (0 or 1)									
Individual	Gamete	Haplotype	1	2	3	4	5	6	7	8	9	10
1	Paternal	1	1	1	1	1	1	1	0	1	0	0

	Maternal	2	0	0	0	0	0	0	1	1	1	0
2	Paternal	3	1	1	1	0	0	0	1	1	0	0
	Maternal	4	0	1	1	1	0	0	1	0	0	0
3	Paternal	4	0	1	1	1	0	0	1	0	0	0
	Maternal	4	0	1	1	1	0	0	1	0	0	0
Allele 1 Frequency[1]			0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0

[1] To avoid problems with multiplying by zero, allele frequencies of zero could have a small number (e.g. 1^{-10}) added to them.

$\mathbf{H}_1$ defines the relationship between a pair of haplotypes as the proportion of matching alleles.

$$\mathbf{H_1} = \begin{bmatrix} 1 & 0.2 & 0.6 & 0.5 \\ 0.2 & 1 & 0.6 & 0.5 \\ 0.6 & 0.6 & 1 & 0.7 \\ 0.5 & 0.5 & 0.7 & 1 \end{bmatrix}$$

$\mathbf{H}_2$ rewards segments of consecutive matching alleles. The reward increases with increasing length of the segment. For a pair of haplotypes spanning a genomic region an overall matching score is calculated as the sum of the sub-region matching scores. Each sub-region matching score is calculated based on the length of the segment of consecutive matching alleles squared. Single matches are treated as a segment of matching alleles of length 1. The overall matching score is standardized to a value between 0.0 and 1.0 by dividing by the maximum value in $\mathbf{H}_2$ and then taking the square root of this value. For the example:

$$\mathbf{H_2} = \begin{bmatrix} 1 & 0.1\ 4 & 0.4\ 2 & 0.3\ 6 \\ 0.1\ 4 & 1 & 0.5\ 1 & 0.3\ 3 \\ 0.4\ 2 & 0.5\ 1 & 1 & 0.4\ 1 \\ 0.3\ 6 & 0.3\ 3 & 0.4\ 1 & 1 \end{bmatrix}$$

Element $h_{1,4}$ is the relationship between haplotype 1 and haplotype 4. This pair of haplotypes has two sub-regions with segments of consecutive matching alleles. Segment 1 spans loci 2, 3, and 4, while segment 2 spans loci 9 and 10. Segment 1 is of length three and thus has a matching score of 3^2=9, while segment 2 is of length two and thus has a matching score of 2^2=4. The overall matching score for this pair of haplotypes is (9+4)=13. For this region 10^2=100 is the maximum matching score (i.e. 10 loci in the region and a haplotype matches with itself for all 10 loci). Therefore $h_{1,4}$ is standardized as a number between 0.0 and 1.0 as $\sqrt{(13/10^2)} = \sqrt{0.13} = 0.36$.

Like $\mathbf{H}_2$, $\mathbf{H}_3$ rewards consecutive segments of matching alleles but it also takes account of the allele frequency of the alleles that match, so as to penalize matches consisting of common alleles compared to those consisting of rare alleles. For the example:

$$\mathbf{H}_3 \cdot \begin{bmatrix} 1 & 0.8 & 7\,0.5 & 6\,0.2 \\ 0.8 & 7\;\;1 & 1.0 & 0\,0.8 \\ 0.5 & 6\,1.0 & 0\;\;1 & 0.0 \\ 0.2 & 0\,0.8 & 8\,0.0 & 0\;\;1 \end{bmatrix}$$

$\mathbf{H}_3$ is constructed by first making a haplotype dissimilarity matrix $\mathbf{D}$, and then obtaining similarity matrix as $\mathbf{H}_3=1-\mathbf{D}$. Element $d_{1,4}$ is the dissimilarity measure between haplotype 1 and haplotype 4. This pair of haplotypes has two sub-regions of segments of consecutive matching alleles. Segment 1 spans loci 2, 3, and 4, while segment 2 spans loci 9 and 10. The allele frequencies for allele 1 of loci 2, 3, 4, 8, and 9 are 0.8, 0.7, 0.6, 0.2, and 0.1 respectively. The dissimilarity score for each segment is calculated as the product of the allele probabilities, for segment 1 this is $0.8 \times 0.7 \times 0.6 = 0.336$, and for segment 2 this is $0.9 \times 1.0 = 0.9$. Note it is the frequency of the 1 allele that is used. Assuming linkage equilibrium these scores can be interpreted as a joint probability of observing a matching segment of alleles in another haplotype that is in fact not IBD, but IBS. The total dissimilarity score for this pair is $0.336 + 0.9 = 1.236$. Standardizing this value to a number between 0.0 and 1.0 by dividing by the maximum value, in this case 1.55 in D, gives a value of $d_{1,4} = 0.80$. Thus $h_{1,4} = 1 - 0.80 = 0.20$.

$\mathbf{G}_{\text{HAP,k}}$ built using $\mathbf{H}_2$ for the three individuals in table 1 will now be illustrated. The final matrix is:

$$\mathbf{G}_{\mathbf{HAP},\mathbf{k}} = \begin{bmatrix} 1.14 & 0.81 & 0.69 \\ 0.81 & 1.41 & 1.41 \\ 0.69 & 1.41 & 2.00 \end{bmatrix}.$$

Element $g_{1,2}$ is the relationship between individual 1 and 2 and is constructed by first constructing a 2×2 matrix $\mathbf{W}$ which is filled with the elements of $\mathbf{H}_2$ pertaining to the four haplotypes that the pair of individuals have. For these individuals, element $w_{1,1}$ is $h_{1,3}$=0.42 which is the relationship between haplotype 1 and 3; the first haplotype in individual 1 is haplotype 1 and the first haplotype in individual 2 is haplotype 3. The element $g_{1,2}$ of $\mathbf{G}_{\text{HAP,k}}$

is calculated as the sum of the 4 elements of $\mathbf{W} = \begin{bmatrix} 0.42 & 0.36 \\ 0.51 & 0.33 \end{bmatrix}$, divided by 2 thus $g_{1,2}$ = ((0.42 + 0.36 + 0.51 + 0.33) / 2) = 0.81.

APPENDIX B:

Correlations between elements of G matrices

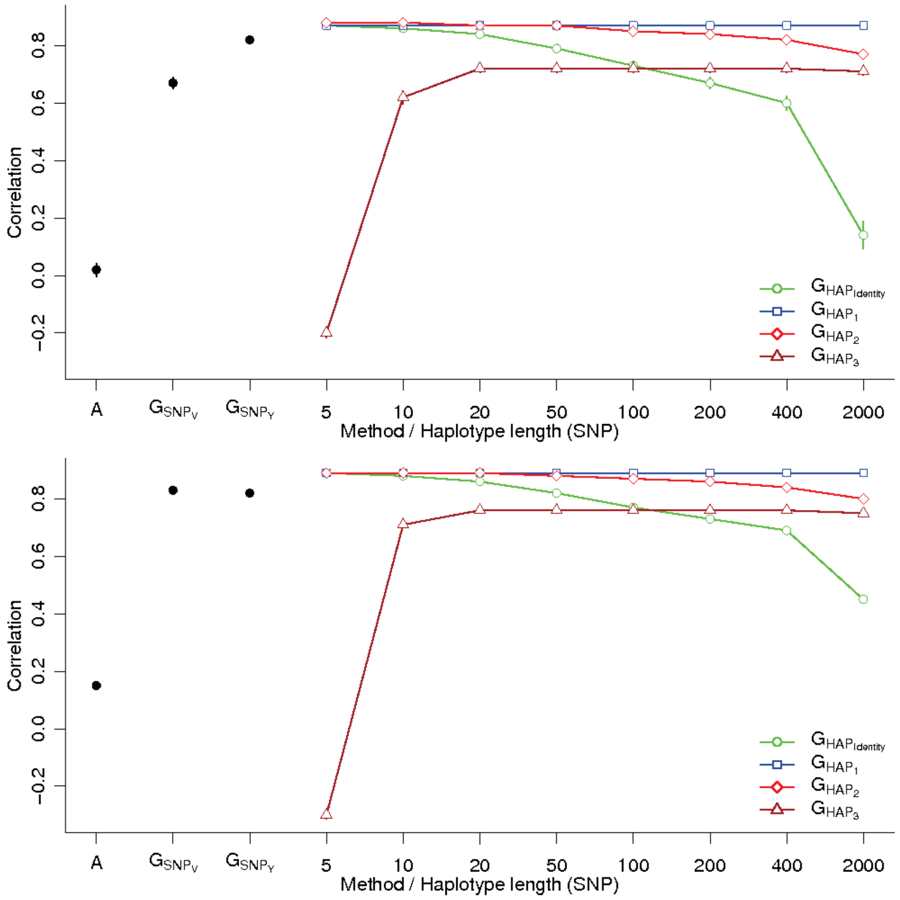

Figure 1 Average correlation (± standard deviation) over replicates between diagonal (top) and off-diagonal (bottom) elements of $\mathbf{G}_{\mathrm{QTL}}$ and other matrices for the GaussUnres trait

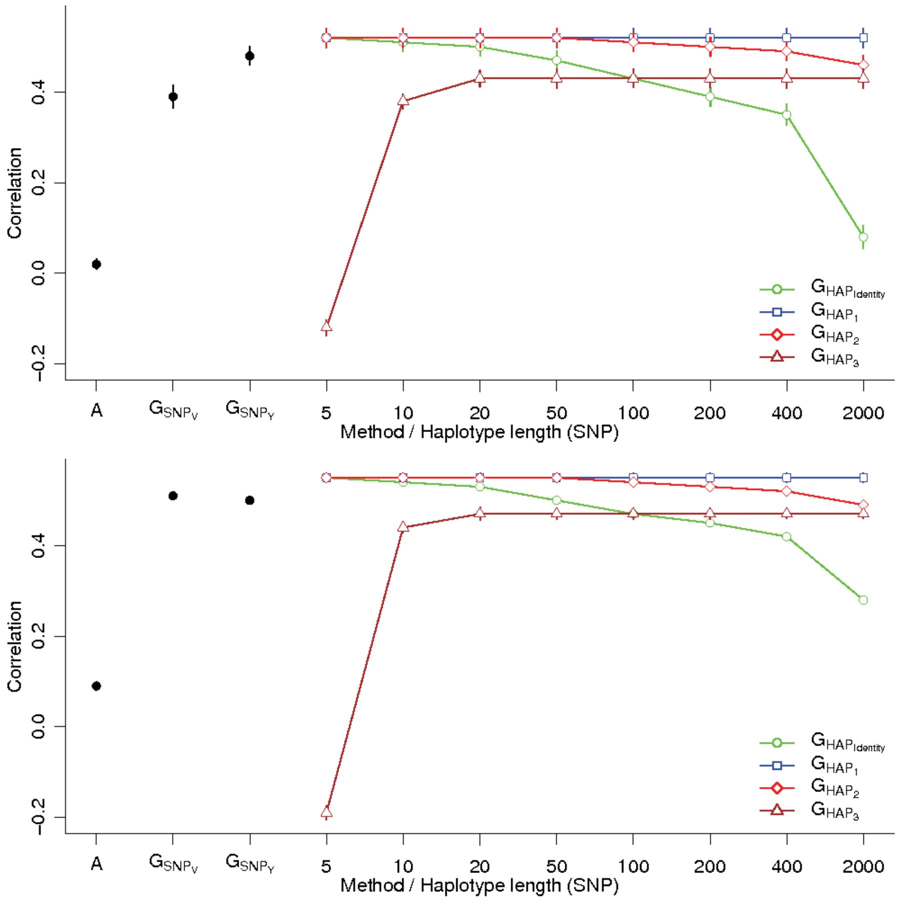

Figure 2 Average correlation (± standard deviation) over replicates between diagonal (top) and off-diagonal (bottom) elements of $\mathbf{G}_{QTL}$ and other matrices for the GammaUnres trait

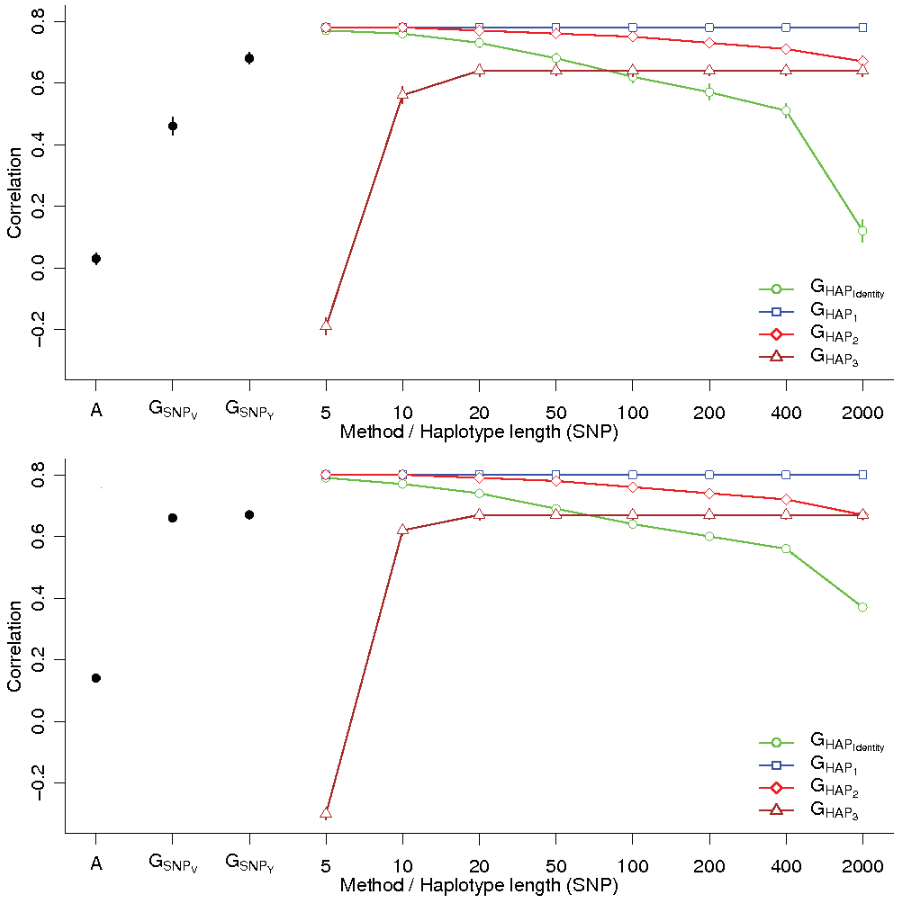

Figure 3 Average correlation (± standard deviation) over replicates between diagonal (top) and off-diagonal (bottom) elements of $\mathbf{G}_{\mathrm{QTL}}$ and other matrices for the GaussRes trait

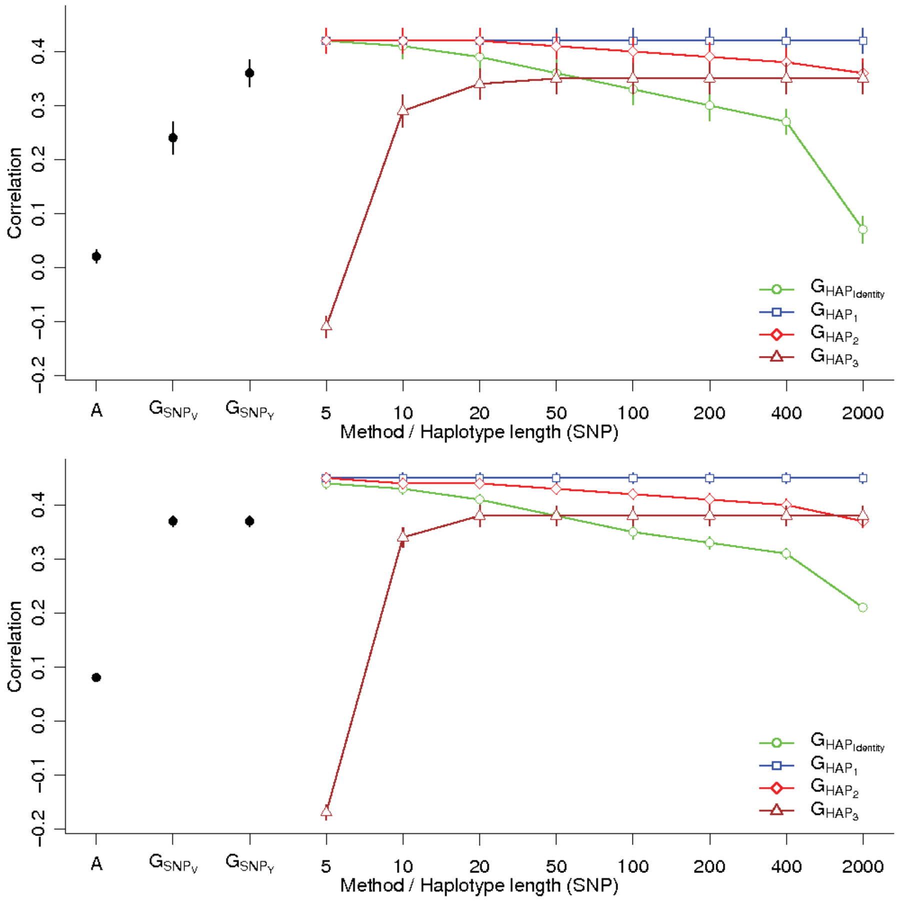

Figure 4 Average correlation (± standard deviation) over replicates between diagonal (top) and off-diagonal (bottom) elements of $\mathbf{G}_{\mathrm{QTL}}$ and other matrices for the GammaRes trait

Estimated variance components

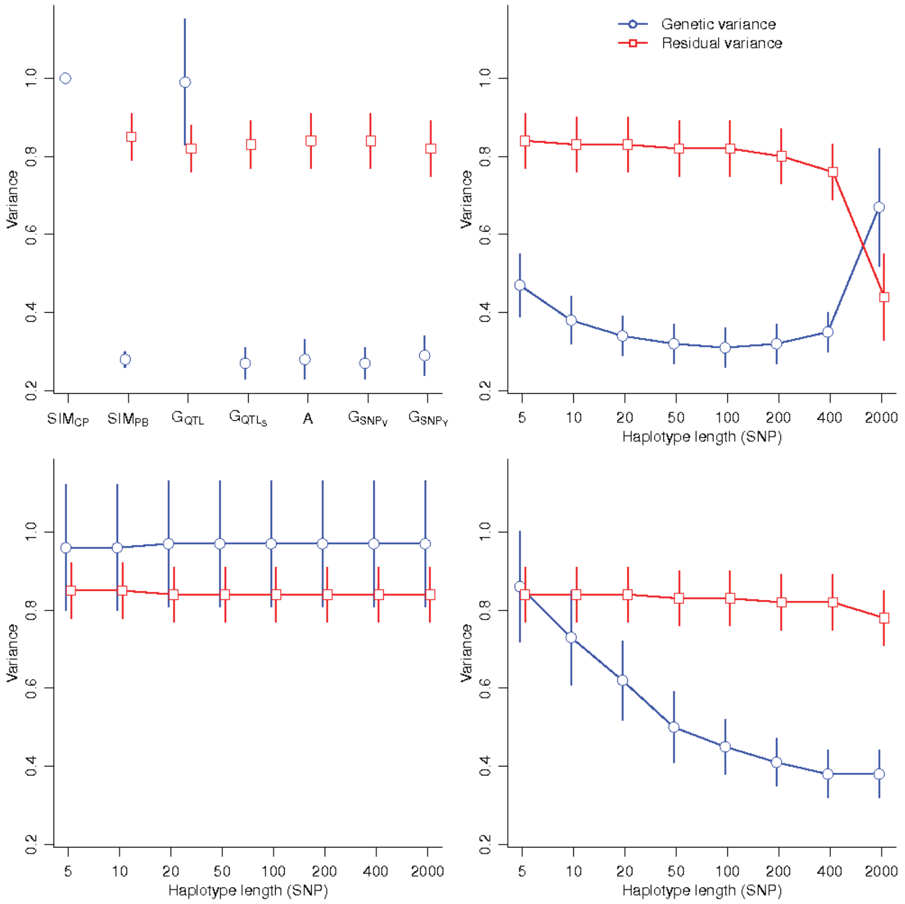

Figure 5 True and estimated genetic and residual variance components (average ± standard deviation over replicates) for the GaussUnres trait estimated using the different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathrm{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathrm{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathrm{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

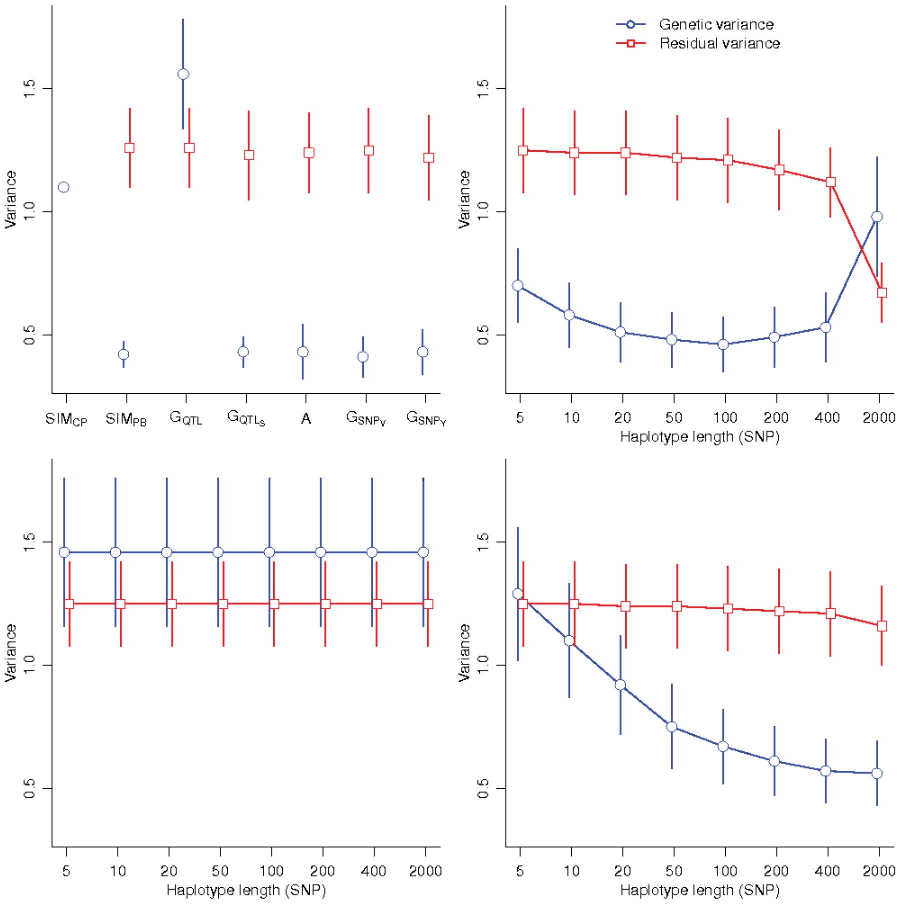

Figure 6 True and estimated genetic and residual variance components (average ± standard deviation over replicates) for the GammaUnres trait estimated using the different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathrm{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathrm{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathrm{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

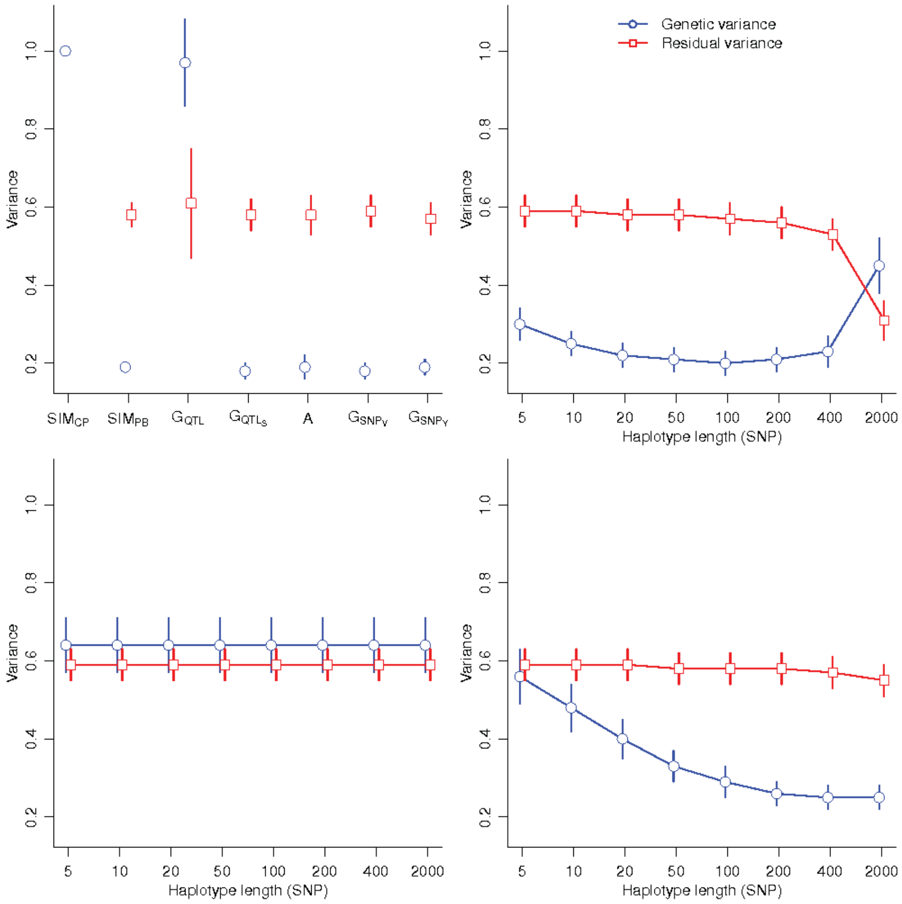

Figure 7 True and estimated genetic and residual variance components (average ± standard deviation over replicates) for the GaussRes trait estimated using the different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathrm{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathrm{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathrm{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

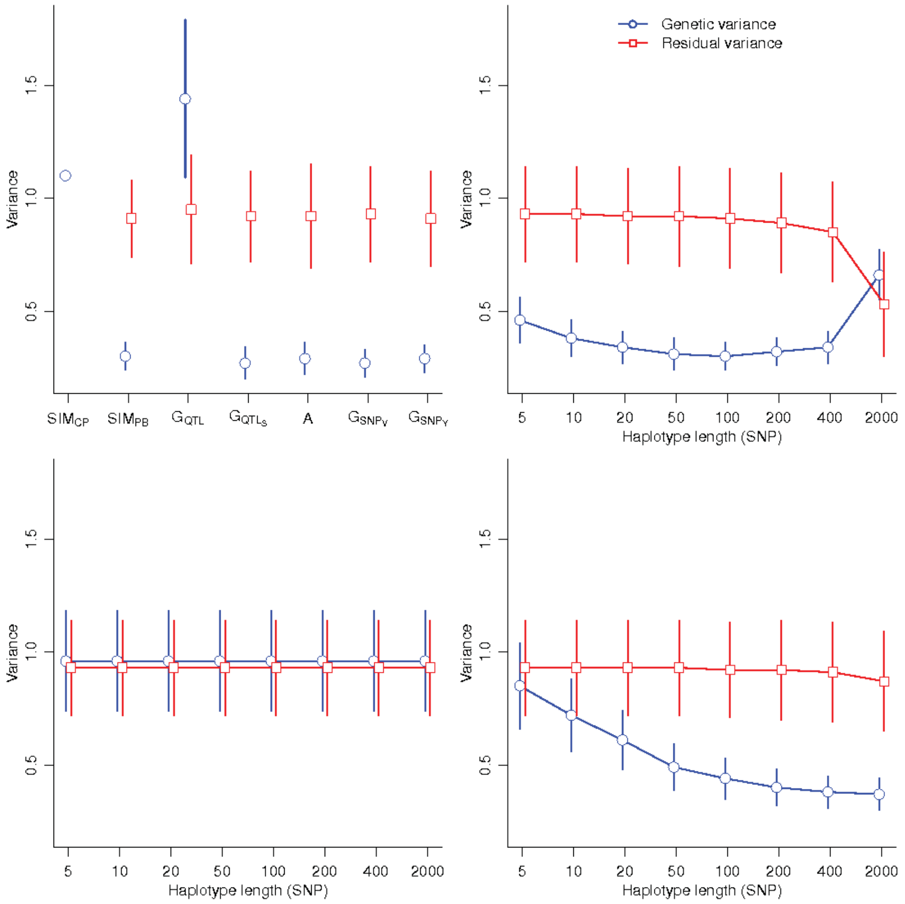

Figure 8 True and estimated genetic and residual variance components (average ± standard deviation over replicates) for the GammaRes trait estimated using the different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), G_{SNPV}, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $G_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) G_{HAP2} using 8 different haplotype lengths

Correlation between true and estimated breeding values

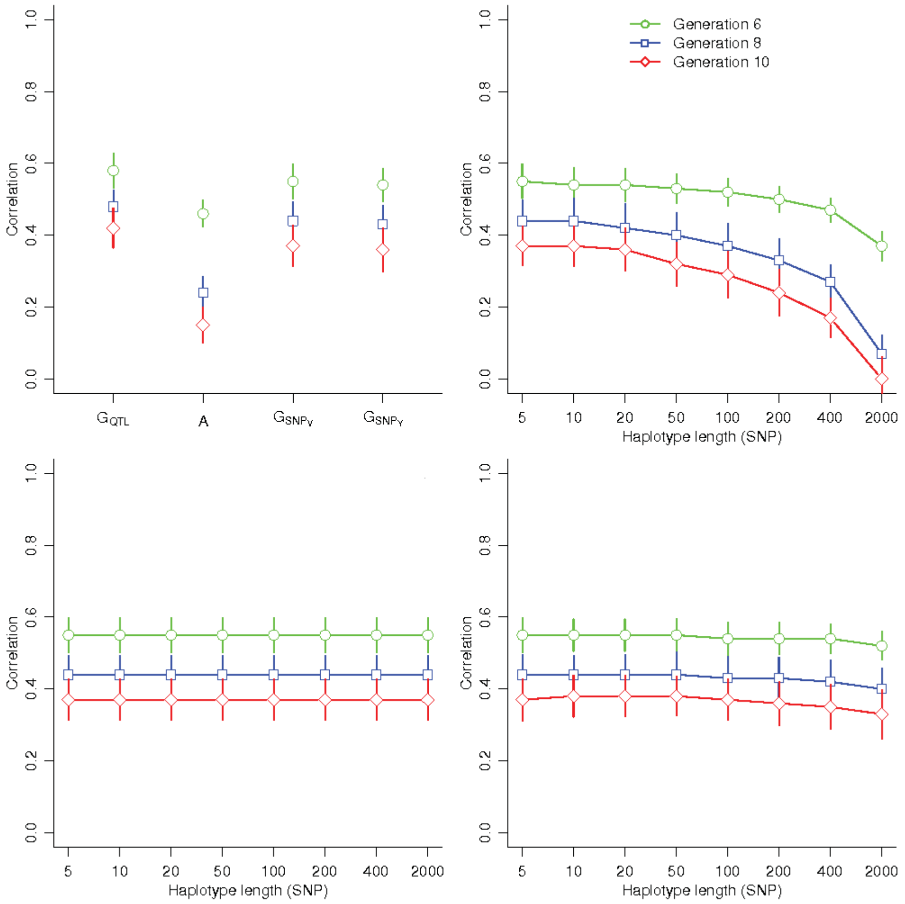

Figure 9 Average accuracy (± standard deviation) over replicates of estimated breeding values for the GaussUnres trait when estimated using different relationship matrices: (top-left) $\mathbf{G}_{\text{QTL}}$, pedigree (**A**), $\mathbf{G}_{\text{SNPV}}$, and $\mathbf{G}_{\text{SNPY}}$, (top-right) $\mathbf{G}_{\text{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\text{HAP1}}$, and (bottom-right) $\mathbf{G}_{\text{HAP2}}$ using 8 different haplotype lengths

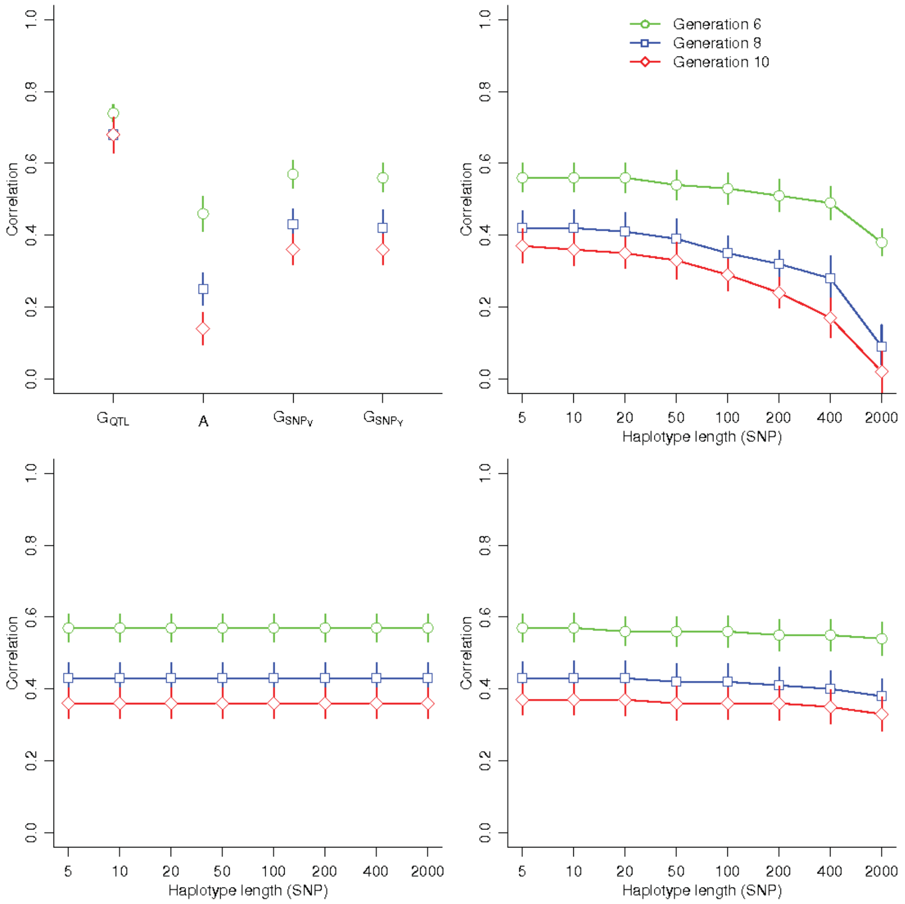

Figure 10 Average accuracy (± standard deviation) over replicates of estimated breeding values for the GammaUnres trait when estimated using different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathbf{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathbf{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathbf{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

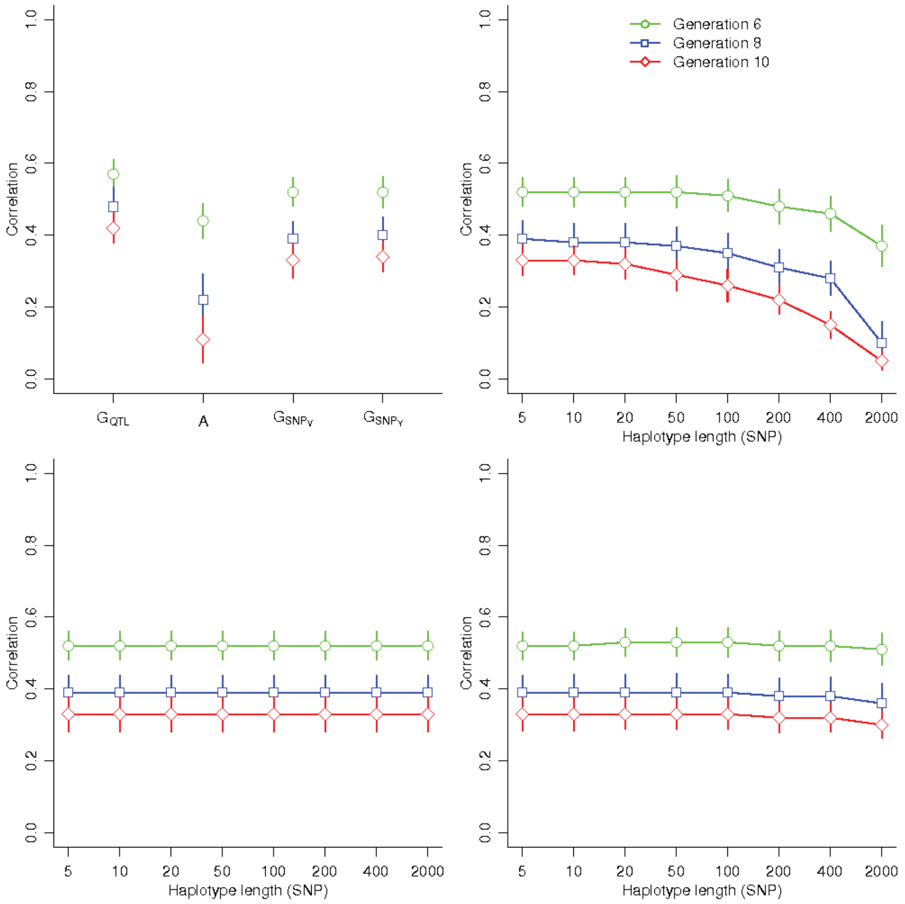

Figure 11 Average accuracy (± standard deviation) over replicates of estimated breeding values for the GaussRes trait when estimated using different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathbf{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathbf{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathbf{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

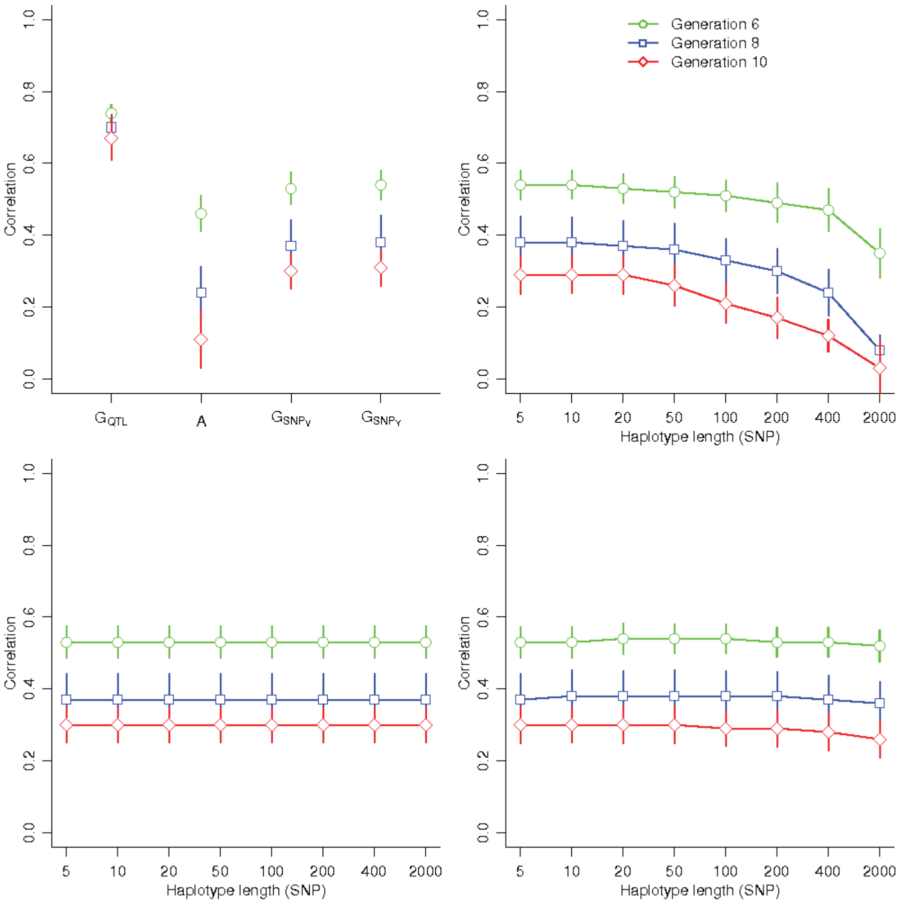

Figure 12 Average accuracy (± standard deviation) over replicates of estimated breeding values for the GammaRes trait when estimated using different relationship matrices: (top-left) $\mathbf{G}_{\mathrm{QTL}}$, pedigree (**A**), $\mathbf{G}_{\mathrm{SNPV}}$, and $\mathbf{G}_{\mathrm{SNPY}}$, (top-right) $\mathbf{G}_{\mathrm{HAPIdentity}}$, (bottom-left) $\mathbf{G}_{\mathrm{HAP1}}$, and (bottom-right) $\mathbf{G}_{\mathrm{HAP2}}$ using 8 different haplotype lengths

Prototype R code for making H and G matrices

(A fortran implementation designed for routine use that uses AlphaPhase is available from http://sites.google.com/site/hickeyjohn/home)

```
### hapHandG.R
###-----------------------------------------------------------------
### What: Prototype R code for
###         - haplotype similarity matrix (H) and
###         - genomic IBD relationship matrix (G) based on H
###-----------------------------------------------------------------

### DATA
###-----------------------------------------------------------------

## Haplotype library (nHaps x hapLength) - an example
hapLib <- matrix(data=c(1, 1, 1, 1, 1, 1, 0, 1, 0, 0,
                        0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
                        1, 1, 1, 0, 0, 0, 1, 1, 0, 0,
                        0, 1, 1, 1, 0, 0, 1, 0, 0, 0),
                 nrow=4, ncol=10, byrow=TRUE)
hapLength <- ncol(hapLib)
nHaps <- nrow(hapLib)

## Allele frequencies (2 x hapLength) - an example
## (can be extended to multiple alleles)
alleleFreq <- matrix(nrow=2, ncol=hapLength)
alleleFreq[1, ] <- seq(from=0.1, to=1, by=0.1) ## Allele 0
alleleFreq[2, ] <- 1 - alleleFreq[1, ]         ## Allele 1
print(alleleFreq)
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9   1.0
## [2,]  0.9  0.8  0.7  0.6  0.5  0.4  0.3  0.2  0.1   0.0
alleleFreq[alleleFreq == 0] <- 10^(-10) # to avoid numerical issues

## Individual haplotype diplotypes (nIndiv x 2) - an example
hapIndiv <- matrix(data=c(1, 2,
                          3, 4,
                          4, 4),
                   nrow=3, ncol=2, byrow=2)
nIndiv <- nrow(hapIndiv)

### FUNCTIONS
###-----------------------------------------------------------------

hapI <- function(x)
{
  ## --- Identity - no similarities between haplotypes ---
  ##
  ## x - matrix, haplotype library (nHaps x hapLength)
  diag(nrow(x))
}

hapH1 <- function(x)
{
  ## --- Simple match similarity ---
  ##
  ## x - matrix, haplotype library (nHaps x hapLength)

  nHaps <- nrow(x)
  hapLength <- ncol(x)
  H <- matrix(nrow=nHaps, ncol=nHaps)

  for(i in 1:nHaps) {
    for(j in i:nHaps) {
      H[i, j] <- H[j, i] <- sum(abs(x[i, ] - x[j, ]))
    }
  }
  list(Hraw=H, H=1 - (H / hapLength))
}

hapH2 <- function(x)
{
  ## --- String match similarity ---
  ##
  ## x - matrix, haplotype library (nHaps x hapLength)

  nHaps <- nrow(x)
  hapLength <- ncol(x)
  H <- matrix(nrow=nHaps, ncol=nHaps)

  ## Diagonal
  diag(H) <- hapLength*hapLength

  ## Off-Diagonal
  for(i in 2:nHaps) {
    for(j in 1:(i - 1)) {
      match <- 1 - abs(x[i, ] - x[j, ])
      sumGlobal <- 0
      sum <- 0
      for(k in 1:hapLength) {
        if(match[k] < 1) {
          sumGlobal <- sumGlobal + sum*sum
          sum <- 0
        } else {
          sum <- sum + 1
        }
      }
      H[i, j] <- H[j, i] <- sumGlobal + sum*sum
    }
  }
  list(Hraw=H, H=sqrt(H / max(H, na.rm=TRUE)))
}
```

```
hapH3 <- function(x, a)
{
  ## --- String match similarity weighted by allele frequencies ---
  ##
  ## x - matrix, haplotype library (nHaps x hapLength)
  ## a - matrix, allele frequencies (2 x hapLength)

  nHaps <- nrow(x)
  hapLength <- ncol(x)
  H <- matrix(nrow=nHaps, ncol=nHaps)

  ## Diagonal
  tmp <- diag(H)
  for(i in 1:nHaps) {
    tmp[i] <- 0
    for(k in 1:hapLength) {
      tmp[i] <- tmp[i] + log(a[x[i, k] + 1, k])
    }
    tmp[i] <- exp(tmp[i])
  }
  diag(H) <- tmp

  ## Off-Diagonal
  for(i in 2:nHaps) {
    for(j in 1:(i - 1)) {
      match <- 1 - abs(x[i, ] - x[j, ])
      sumGlobal <- 0
      sum <- 0
      for(k in 1:hapLength) {
        if(match[k] < 1) {
          if(sum < 0) {
            sumGlobal <- sumGlobal + exp(sum)
            sum <- 0
          }
        } else {
          sum <- sum + log(a[x[i, k] + 1, k])
        }
      }
      if(sum < 0) {
        sumGlobal <- sumGlobal + exp(sum)
      }
      H[i, j] <- H[j, i] <- sumGlobal
    }
  }
  list(Hraw=H, H=1 - (H / max(H, na.rm=TRUE)))
}

buildG <- function(x, H)
{
  ## --- Build genomic relationship matrix for a set of individuals
  ##     given their haplotype diplotypes and similarities ---
  ##
  ## x - matrix, individual haplotype diplotypes (nIndiv x 2)
  ## H - matrix, haplotype similarity (nHaps x nHaps)

  n <- nrow(x)
  G <- matrix(nrow=n, ncol=n)
  for(i in 1:n) {
    for(j in 1:i) {
      G[i, j] <- (H[x[i, 1], x[j, 1]] +
                  H[x[i, 1], x[j, 2]] +
                  H[x[i, 2], x[j, 1]] +
                  H[x[i, 2], x[j, 2]]) / 2
      G[j, i] <- G[i, j]
    }
  }
  G
}

### EXAMPLES
###------------------------------------------------------------------

## --- Identity - no similarities between haplotypes ---

(H <- hapI(x=hapLib))
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1

(G <- buildG(x=hapIndiv, H=H))
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    1
## [3,]    0    1    2

## --- Simple match similarity ---

(H <- hapH1(x=hapLib))
## $Hraw
## [,1] [,2] [,3] [,4]
## [1,]    0    8    4    5
## [2,]    8    0    4    5
## [3,]    4    4    0    3
## [4,]    5    5    3    0
##
## $H
## [,1] [,2] [,3] [,4]
## [1,]  1.0  0.2  0.6  0.5
## [2,]  0.2  1.0  0.6  0.5
## [3,]  0.6  0.6  1.0  0.7
## [4,]  0.5  0.5  0.7  1.0

(G <- buildG(x=hapIndiv, H=H$H))
```

```
##       [,1] [,2] [,3]
## [1,]  1.2  1.1  1.0
## [2,]  1.1  1.7  1.7
## [3,]  1.0  1.7  2.0

## --- String match similarity ---

H <- hapH2(x=hapLib)
lapply(H, round, digits=3)
##$Hraw
##     [,1] [,2] [,3] [,4]
##[1,]  100    2   18   13
##[2,]    2  100   26   11
##[3,]   18   26  100   17
##[4,]   13   11   17  100
##
## $H
##        [,1]  [,2]  [,3]  [,4]
## [1,] 1.000 0.141 0.424 0.361
## [2,] 0.141 1.000 0.510 0.332
## [3,] 0.424 0.510 1.000 0.412
## [4,] 0.361 0.332 0.412 1.000

G <- buildG(x=hapIndiv, H=H$H)
round(G, digits=3)
##        [,1]  [,2]  [,3]
## [1,] 1.141 0.813 0.692
## [2,] 0.813 1.412 1.412
## [3,] 0.692 1.412 2.000

## --- String match similarity weighted by allele frequencies ---

H <- hapH3(x=hapLib, a=alleleFreq)
lapply(H, round, digits=3)
## $Hraw
##        [,1]  [,2]  [,3]  [,4]
## [1,] 0.008 0.200 0.684 1.236
## [2,] 0.200 0.000 0.007 0.190
## [3,] 0.684 0.007 0.003 1.550
## [4,] 1.236 0.190 1.550 0.002
##
## $H
##        [,1]  [,2]  [,3]  [,4]
## [1,] 0.995 0.871 0.559 0.203
## [2,] 0.871 1.000 0.995 0.877
## [3,] 0.559 0.995 0.998 0.000
## [4,] 0.203 0.877 0.000 0.999

G <- buildG(x=hapIndiv, H=H$H)
round(G, digits=3)
##        [,1]  [,2]  [,3]
## [1,] 1.869 1.317 1.080
## [2,] 1.317 0.998 0.999
## [3,] 1.080 0.999 1.997

###-----------------------------------------------------------------------
### hapHandG.R ends here
```

3. Reliability of breeding values in selected populations, InterBull Bulletin (sprejeto v objavo)

Preučitev vpliva selekcije na točnost genetskega vrednotenja in primerjava med klasičnim in genomskim pristopom (Gorjanc in sod., 2012). Rezultati so pokazali, da je v primeru selekcioniranih populacij točnost klasičnega genetskega vrednotenja znatno zmanjšana pri mladih živalih brez fenotipskih podatkov, medtem ko selekcija nima tako izrazitega vpliva na točnost v primeru uporabe genomske informacije. To pomeni, da je genomski pristop v selekcioniranih populacijah znatno bolj točen kot je bilo znano do sedaj.

Reliability of breeding values in selected populations

***G. Gorjanc*[1]*, J.M. Hickey*[2,3] *and P. Bijma*[4]**

[1] *University of Ljubljana, Biotechnical Faculty, Animal Science Department, Domžale, Slovenia*
[2] *University of New England, School of Environmental and Rural Science, Armidale, Australia*
[3] *CIMMYT, Biometrics and Statistics Unit, Mexico D.F., Mexico*
[4] *Wageningen University, Animal Breeding and Genomics Centre, Wageningen, The Netherlands*
E-mail: gregor.gorjanc@bf.uni-lj.si

Abstract

Selection reduces genetic variance in population. However, this is not taken into account when reliabilities are computed from prediction error variance (PEV) and base population additive genetic variance. Results of simulations confirmed that when selection is present PEV based reliabilities are too high and do not reflect the true uncertainty of EBV. Drop in reliability is substantial for parent average, while EBV for progeny tested or genomically evaluated animals is reduced only slightly. This implies that genomic EBV are in comparison to pedigree EBV even more reliable than anticipated from the comparison of PEV based reliabilities.

Key words: accuracy, reliability, pedigree, genomic, selection

Introduction

Modern breeding programs base genetic improvement on estimated breeding values (EBV). In the case of linear mixed models of type:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}, \qquad (1)$$

breeding values $\boldsymbol{a}$ are inferred from the collected data $\boldsymbol{y}$ by solving the mixed model equations to obtain estimates of $\boldsymbol{a}$ (EBV). In addition variances of EBV (PEV) are also routinely reported in order to provide a measure of the potential change of EBV in the future. In most breeding programs reliabilities are reported instead of PEV as computed by:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{PEV}}{\mathrm{Var\ a}}, \qquad (2)$$

where $Var\ a$ is additive genetic variance in the base population. Reliability of EBV is an important statistic as it describes the potential change of EBV when more information becomes available and because it is one of the determining factors of a response to selection.

With the introduction of genomics comparison of reliabilities has become very common to compare different breeding programs, e.g, the reliability of EBV for progeny tested sires versus genomically tested young bulls. These comparisons often involve different types of reliabilities; based on PEV from mixed model equations or some type of validation.

Bijma (2012) showed theoretically that PEV based reliabilities are too high when selection is present in the population, especially for the EBVs that rely to a great extent on the parent average information. This work complements previous theoretical derivations of Bijma (2012) to quantify the effect of selection on PEV based reliability in genomic setting via simulation.

Theoretical basis of the effect of selection on reliability

Effect of selection on reliability of parent average a_o can be clearly demonstrated with an example of truncation selection in parents

a_s and a_d . Without selection the variance and reliability of parent average EBV is:

$$
\begin{aligned}
&Var\ \ a_s\ \ = Var\ \ a_d\ \ = Var\ \ a\ \ , \\
&Var\ \ a_o\ \ = \tfrac{1}{2} Var\ \ a\ \ , \qquad\qquad (3) \\
&R^2\ \ a_o\ \ = \tfrac{1}{2}\ \ R^2\ \ a_s\ \ + R^2\ \ a_d\ \ \ .
\end{aligned}
$$

Introduction of selection in parents reduces variability of breeding values (only a part of parents are selected) which propagates to the reliability of parent average EBV:

$$
\begin{aligned}
&Var\ \ a_s\ \ = Var\ \ a_d\ \ = Var\ \ a\ \ \ 1-k\ \ , \\
&Var\ \ a_o\ \ = \tfrac{1}{2} Var\ \ a\ \ \ 1-k\ \ , \qquad\qquad (4) \\
&R^2\ \ a_o\ \ = \tfrac{1}{2}\ \ R^2\ \ a_s\ \ + R^2\ \ a_d\ \ \ \ 1-k\ \ .
\end{aligned}
$$

With 20% parents selected $k = i\ \ i - x\ \ \approx$ 0.78, which leads to a substantial reduction in variance and reliability $\ 1 - k = 0.22\ $.

Above equalities (3 and 4) hold only for one generation of truncation selection in parents. With a continuous selection equilibrium is attained and reliability of parent average EBV when intensity of selection is equal in both sexes is (Bijma, 2012):

$$R_\infty^2\ \ a_o\ \ = \frac{R^2\ \ a}{2}\ \ \frac{1-k}{1+k\ \ 1-R^2\ \ a}\ \ , \qquad (5)$$

while reliability of EBV obtained upon progeny test is (Bijma, 2012):

$$R_\infty^2\ \ a\ \ = R^2\ \ a\ \ \ \frac{1}{1+k\ \ 1-R^2\ \ a}\ \ . \qquad (6)$$

Comparison of (5) and (6) over a range of selection intensities clearly shows that selection influences reliability, but to a greater extent for parent average than progeny test based EBV (Figure 1). When selection intensity is different in males and females the equations (5) and (6) can be modified to take this difference into account (Bijma, 2012).

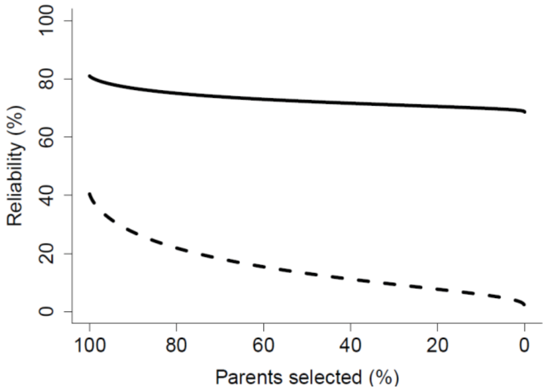

Figure 1. Effect of selection intensity on reliability of parent average (bellow) and progeny test (above) based EBV

Simulation

Simulation followed the workflow of Hickey and Gorjanc (2012) which involves a) coalescent simulation with mutation, recombination, and drop in historical effective population size (Ne) to obtain structured haplotypes for 30 chromosomes and b) haplotype dropping through pedigree. During the later phase mutations were assumed non-existent and Ne was constant. Pedigree consisted of 25 generations with each of the 50 sires mated with 10 dams having each 4 progeny per generation. Altogether, there were 2000=50×10×4 animals per generation. Phenotypes were assigned only to males. Heritability was high $\ \boldsymbol{h^2 = 0.75}\ $ in order to keep the simulated data small but still mimic progeny testing. Simulation involved random selection of parents (no selection scenario) or selection of parents on pedigree BLUP (selection scenario). In the first ten generations there was no selection in order to reach information equilibrium. At the end of simulation the available data comprised of pedigree, 60 SNP markers, true breeding values, and phenotypic values (5000 records from generations 16 to 20).

Statistical Analysis

Obtained phenotype, pedigree, and genomic data were analysed with pedigree (ABLUP) and genomic (GBLUP) based linear mixed model (1). Analyses were performed for each generation successively to obtain parent average EBV a_{A0} for each animal free of phenotypic information from descendants or collateral relatives.

Obtained EBV from ABLUP (a_{A0} and a_A) and GBLUP a_G were correlated with the true values (validation reliabilities) and compared with PEV based reliabilities (2).

Results & Discussion

Obtained reliabilities of a_{A0}, a_A, and a_G were expected – higher reliabilities of a_A in males (due to progeny testing) than females; drop in the reliabilities of parent average with successive generations (due to segregation and recombination); higher reliabilities of a_G in comparison to a_A for phenotyped individuals; and higher and more stable reliabilities of a_G in successive generations (Table 1).

In the scenario with no selection reliabilities obtained from PEV roughly matched validation based reliabilities for both ABLUP and GBLUP (Table 1). This shows that in the case with no selection PEV based reliabilities provide accurate information about the uncertainty of EBV. However, in the scenario with selection validation reliabilities were consistently lower than PEV based reliabilities (Table 2). The difference was greater for a_{A0} and a_A in females than for a_A in males, which is in agreement with developments of Bijma (2012) as shown in Figure 1. Validation reliability of genomic EBV a_G was also lower than PEV based reliability; however the difference was much smaller than for the pedigree EBV (Table 2).

Obtained validation reliabilities in the scenario with selection (Table 2) matched the expected equilibrium reliabilities with different selection intensity by sex (Bijma, 2012) - 50 sires and 500 dams selected from 2000 offspring (both sexes) each generation (Figure 2).

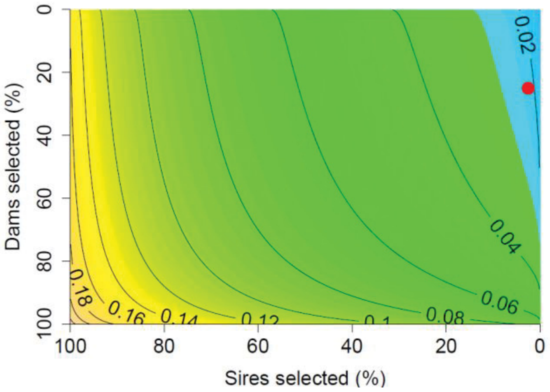

Figure 2. Expected (contours) and validation (point) reliability of parent average EBV according to selection intensity in sires and dams

Conclusions

In summary results of simulation corroborate the developments of Bijma (2012) who showed that selection reduces reliability of EBV and that PEV based reliabilities do not reflect this reduction. In addition results show that genomic EBV are in comparison to pedigree EBV even more reliable than anticipated from the comparison of PEV based reliabilities.

References

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *J. Anim. Breed. Genet.* 129, 345-358.

Hickey, J.M. & Gorjanc, G. 2012. Simulated data for genomic selection and GWAS using a combination of coalescent and gene drop methods. *Genes, Genomes, Genetics*, 2, 425-427.

Table 1. Prediction error variance and validation based reliabilities by generation and source of information in the no selection scenario

	$\mathrm{R}^2 = 1 - PEV\ \ \mathrm{Var}\ \ \mathrm{a}$			$\mathrm{R}^2 = \mathrm{Corr}\ \ EBV, TBV\ \ ^2$		
Gen.	a_{A0}	a_A	a_G	a_{A0}	a_A	a_G
20[a]	24 ± 4	50 ± 2	/	29 ± 4	56 ± 3	/
20[s]	24 ± 4	71 ± 3	83 ± 1	30 ± 5	78 ± 2	84 ± 1
20[d]	24 ± 4	30 ± 1	/	28 ± 4	35 ± 4	/
21[a]	24 ± 1		64 ± 1	27 ± 4		63 ± 3
22[a]	10 ± 1		57 ± 1	14 ± 4		55 ± 5
23[a]	4 ± 1		54 ± 1	8 ± 4		52 ± 4
24[a]	0 ± 1		52 ± 1	4 ± 3		51 ± 6
25[a]	-2 ± 1		50 ± 1	2 ± 2		47 ± 4

a – all animals; s – sires; d - dams

Table 2. Prediction error variance and validation based reliabilities by generation and source of information in the selection scenario

	$\mathrm{R}^2 = 1 - PEV\ \ \mathrm{Var}\ \ \mathrm{a}$			$\mathrm{R}^2 = \mathrm{Corr}\ \ EBV, TBV\ \ ^2$		
Gen.	a_{A0}	a_A	a_G	a_{A0}	a_A	a_G
20[a]	22 ± 3	47 ± 1	/	3 ± 1	39 ± 2	/
20[s]	22 ± 3	66 ± 2	83 ± 1	3 ± 1	68 ± 1	78 ± 1
20[d]	22 ± 3	28 ± 1	/	3 ± 2	11 ± 2	/
21[a]	22 ± 1		66 ± 1	3 ± 2		53 ± 5
22[a]	10 ± 1		61 ± 1	0 ± 1		48 ± 5
23[a]	3 ± 1		58 ± 1	0 ± 1		45 ± 5
24[a]	0 ± 1		55 ± 1	0 ± 1		41 ± 4
25[a]	-2 ± 1		54 ± 1	0 ± 1		41 ± 4

a – all animals; s – sires; d - dams

4. Accuracy of genomic prediction for milk traits with different approaches in Slovenian Brown bulls, Czech Journal of Animal Science (poslano v recenzijo)

Preučitev različnih možnosti uporabe genomskih informacij za genetsko vrednotenje živali v majhnih populacijah (Špehar in sod., 2012). Pokazali smo, da je možno genomsko selekcijo uvesti tudi v majhnih populacijah, pri čemer je nujno potrebno sodelovanje z večjimi državami ali konzorciji.

Czech Journal of Animal Science (submitted)

Accuracy of genomic prediction for milk traits with different approaches in Slovenian Brown bulls

M. ŠPEHAR[1, 2], K. POTOČNIK[2], G. GORJANC[2]

[1]Croatian Agricultural Agency, Zagreb, Croatia

[2]University of Ljubljana, Biotechnical Faculty, Animal Science Department, Domžale, Slovenia

Use of genome-wide single nucleotide polymorphism (SNP) marker information enables more accurate inference of breeding values (EBV), especially for young animals. The objective of this study was to compare the accuracy of genomic and progeny based evaluation for milk traits in a small population of Slovenian Brown bulls using different approaches to utilize genomic information. Four approaches were considered: 1) NAT - phenotypic and pedigree data from national genetic evaluation used in univariate repeatability test-day model, 2) NATss - NAT approach with the inclusion of genome-wide SNP genotypes for 183 Slovenian Brown bulls via the improved relationship matrix of single-step methodology, 3) MT1 - NAT approach with direct genomic values (DGV) as correlated trait for 183 Slovenian Brown bulls available externally from the InterGenomics consortium, and 4) MT2 - the same as MT1 but with using DGV for 399 bulls in the national pedigree. Performance of different approaches was assessed with the analysis of theoretical and validation accuracies. For validation bulls in juvenile stage (reduced dataset) increase in accuracy with the NATss approach did was negligible in comparison to the approach NAT due to the small reference population in Slovenia. With the MT1 and MT2 approach the average theoretical accuracies were 0.90 (MT1) and 0.79 (MT2) for milk, 0.86 (MT1) and 0.77 (MT2) for fat, and 0.85 (MT1) and 0.74 (MT2) for protein yield. These results confirm the expected increase in accuracy due to the inclusion of genomic information (via DGV) in the national evaluation system. However, accuracies with the MT1 approach were unrealistically high. Validation accuracies were lower in comparison to the average theoretical accuracies, especially for the NAT and NATss approaches. With the MT1 and MT2 validation accuracies were 0.92 (MT1) and 0.74 (MT2) for milk, 0.91 (MT1) and 0.81 (MT2) for fat, and 0.87 (MT1) and 0.72 (MT2) for protein yield. Use of larger number of animals with DGV information (national and foreign bulls used in Slovenia) as in the MT2 approach resulted in realistic accuracy of genetic evaluation. These results show that the integration of genomic information into national evaluation was successful. Further research is needed to quantify the effect of potential double counting of available information.

Keywords: genomic selection; breeding value, accuracy; single-step, bivariate blending

INTRODUCTION

Genetic improvement of quantitative traits in dairy cattle is commonly based on phenotypic and pedigree information that are used to infer (estimate) breeding values (EBV). The process of providing reliable breeding values based on polygenic model is time consuming. Phenotypic data in dairy cattle are collected through various recording schemes (milk and fertility recording, type classification, etc.) on daughters of progeny tested bulls. The later are four to six years old when the accuracy of their EBV is 0.90 or more. Before progeny testing, the accuracy of EBV (parent average) is around 0.60 (e.g., Schefers and Weigel, 2012). In recent years, the availability of affordable high-density panels of single-nucleotide polymorphisms (SNP) has led to abundant use of this information in selection decisions commonly called genome-wide or genomic selection (Meuwissen et al., 2001). Genomic selection is based on the inference of breeding value based on the sum of SNP or haplotypes effects across whole genome (Meuwissen et al., 2001; Solberg et al., 2008). Due to the direct use of marker data obtained marker based breeding value is often called direct genomic value (DGV). For the implementation of genomic selection, phenotyped and genotyped reference population is used to derive the prediction equation of DGV. This equation is then used to estimate DGV for non-phenotyped individuals (Meuwissen et al., 2001; Goddard and Hayes, 2007). Both types of breeding values (EBV and DGV) can be blended in one value - genomically enhanced breeding value (GEBV) using various approaches (VanRaden 2008; Kachman, 2008; Aguilar et al., 2010). The advantage of using genome-wide data is to increase the accuracy of EBV from about 0.60 to 0.80 for young or non-phenotyped animals (Hayes et al., 2009a). VanRaden et al. (2009) reported the accuracies of GBV equal to 0.70 averaged across traits in Holstein-Friesian dairy cattle in USA. If the accuracy of GEBV is high enough, early use of young bulls in artificial insemination will shorten generation interval and increase genetic gain per unit of time (Schaeffer, 2006).

The accuracy of DGV depends on the heritability of trait of interest, but for a particular trait it mainly depends on the number of genotyped and phenotyped individuals in the reference population (Schefers and Weigel, 2012). Therefore, the highest gains with genomic selection can be expected for traits with longer recording history in large dairy populations with substantial number of genotyped animals, especially progeny tested bulls. An example of such a reference population is the European consortium in Holstein breed which includes more than 17,000 progeny tested bulls via the EuroGenomics project (Liu et al., 2011). In Brown Swiss breed there is a global initiative with 3,400 reference bulls via the Intergenomics project operated at the Interbull (Zumbach et al., 2010) and this numbers are constantly increasing. In these consortia phenotype information is used in the form of EBV from the Interbull multiple across country evaluation (MACE) based on national EBV and international pedigree. MACE EBVs are used to train the prediction equation of DGVs that are later blended with pedigree based EBV. Countries not involved in these consortia try to develop prediction equations using similar approaches. In some countries there is lately an increase in the use of the single-step approach (Legarra et al., 2009; Aguilar et al., 2010) of inferring GEBV directly from the available phenotype, pedigree, and marker data (e.g. Su et al., 2012; Přibyl et al., 2012).

In beef cattle reference populations are usually much smaller than in dairy cattle due to more disconnected local populations of the same breed over several countries. This leads to smaller reference populations. Kachman (2008) presented an alternative method of using genomic information in such setting. In these populations prediction equation for DGV could be developed on some experimental population of reasonable size and later used to compute the DGV of other animals in national population that is later blended with national EBV via bivariate analysis. Similar approach was also suggested by Mäntysaari and Strandén (2010). Examples of applying these methods for beef populations are MacNeil et al. (2010a) and Johnston et al. (2010; 2012).

The objective of this study was to compare the accuracy of genomic and progeny based evaluation of bulls for milk traits in a small population of Brown breed in Slovenia based on different approaches to utilize genomic information.

MATERIAL AND METHODS

Phenotypic data in the analysis consisted of 1,342,134 test-day records for daily milk, fat, and protein yields from 57,670 cows recorded between years 1997 and 2011 and used in the routine national genetic evaluation. The total number of animals in the pedigree was 79,573. In addition SNP genotypes (BovineSNP50 BeadChip, Illumina, San Diego, CA) were available for 191 Brown bulls born between 1990 and 2007. These bulls were of predominantly Slovenian origin with nine bulls originating from other countries (Austria and Germany), but were imported to Slovenia as live animals and used only in Slovenia. Other foreign bulls that were used in Slovenia via semen were not genotyped due to the

InterGenomics project in which Slovenia is a partner. The third data source consisted of DGV and GEBV values for 6,153 bulls from the InterGenomics project as evaluated in April 2011.

Preliminary edits of national SNP data were carried as follows. SNPs were considered if a call rate was higher than 90% by SNP and 80% by bull. All together, 6,889 SNPs and 8 bulls did not meet these criteria and were therefore excluded. SNPs with minor allele frequencies lower than 0.05 were also excluded (13,340 cases). The departure from the Hardy-Weinberg equilibrium at a threshold of $P < 0.0001$ was the next edit with 7,490 SNPs failing this criterion. Finally, SNPs that could not be mapped or that were on the X chromosome were excluded, leaving a final set of 34,450 SNP for 183 bulls. Missing genotypes were imputed using the gpig program (Strandén, 2010), which implements the imputation method based on linear best unbiased prediction (Gengler et al., 2007).

Based on the available data four different approaches of utilizing genomic information were tested. For each of the following approaches, analyses involved inference of required (co)variance components with residual maximum likelihood method and conditional on those values breeding values were inferred. The first approach (NAT) was the standard method of utilizing phenotypic and pedigree information as currently used in the routine national system. The statistical model is a simple univariate repeatability test-day model per trait:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_a\mathbf{a} + \mathbf{Z}_c\mathbf{c} + \mathbf{Z}_p\mathbf{p} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypic observations for either daily milk, protein, or fat yield, $\mathbf{b}$ is a vector of unknown parameters for fixed effects, $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$ is a vector of unknown parameters for additive genetic effect (breeding values) with covariance matrix equal to a pedigree based numerator relationship matrix ($\mathbf{A}$), $\mathbf{c} \sim N(\mathbf{0}, \mathbf{I}\sigma_c^2)$ is a vector of unknown parameters for herd effect, $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2)$ is a vector of unknown parameters for permanent environment effect, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ is a vector of residuals, while $\mathbf{X}$, $\mathbf{Z}_a$, $\mathbf{Z}_c$, and $\mathbf{Z}_p$, are incidence matrices linking $\mathbf{y}$ and $\mathbf{b}$, $\mathbf{a}$, $\mathbf{c}$, and $\mathbf{p}$.

The second approach (NATss) was based on the single-step methodology (Legarra et al., 2009; Aguilar et al., 2010) utilizing national phenotypic, pedigree, and SNP genotype information. The difference between this and the NAT approach was that prior distribution for breeding values was equal to $\mathbf{a} \sim N(\mathbf{0}, \mathbf{H}\sigma_a^2)$, where covariance matrix $\mathbf{H}$ is a function of $\mathbf{A}$ and $\mathbf{G}$ (Legarra et al., 2009; Aguilar et al., 2010), where $\mathbf{G}$ is a genome-wide marker based numerator relationship matrix (VanRaden, 2008). The later matrix was available only for 183 bulls in the national system.

The third approach (MT1) was based on the bivariate analysis (Kachman, 2008; Mäntysaari and Strandén, 2010) incorporating genomic information into the national genetic evaluation through DGV information as a correlated trait. DGV were obtained from the InterGenomics project for 191 bulls genotyped in Slovenia. Finally, the fourth approach (MT2) was the same as MT1 with the difference that DGV information from the InterGenomics project was considered for all bulls that appear in the national pedigree. Altogether 399 bulls with DGV information were considered in the MT2 approach.

Performance of different approaches to utilize genomic information was assessed with the analysis of breeding value accuracies for a set of different groups of animals: genotyped bulls born between 2004 and 2006 (validation bulls), all genotyped bulls, and all animals. Two types of accuracies were obtained: theoretical accuracies based on the prediction error variance (averaged over a group of animals) and validation accuracies based on the correlation between breeding values for validation bulls in forward validation using a reduced and a full dataset. In the reduced dataset, phenotypic data from years 2008 to 2011 were removed in order to exclude progeny data for a group of validation bulls.

All computations were performed with a combination of different programs for different tasks: BLUPF90 (Misztal et al., 2002), SAS (2008) and VCE (Kovač et al., 2002).

RESULTS AND DISCUSSION

Descriptive statistics for the studied milk traits for full and reduced dataset is shown in Table 1. The statistics were similar because reduced dataset contained more than 80% of all records. In full dataset phenotype records (1,342,134) were available from 56,670 cows that were progeny 736 bulls among which 183 were genotyped. There were 9 genotyped bulls that were not yet progeny tested at the time of analysis. In reduced data phenotype records (1,082,701) were available from 47,779 cows that were progeny of 615 bulls. Altogether there were 9 plus 35 genotyped bulls that were not yet progeny tested according to the reduced dataset. Genotyped bulls with progeny were born between 1990 and 2007 (Figure 1), with 4 to 17 bulls per year. The average number of phenotyped daughters per bull ranged between 200 and 400 for bulls born before the year 2000. In the later years, the average number of phenotyped daughters decreased. Bulls born in years 2004 to 2007 (35 bulls) were used as a validation set in forward validation.

Table 1. Descriptive statistics for milk traits by dataset

Dataset	Full		Reduced	
No. records	1,342,134		1,082,701	
No. cows	56,670		47,779	
No. bulls	736		615	
No. gen. bulls	183		183	
- tested	174		139	
- non-tested	9		35+9	
Trait / kg	Mean	Std.	Mean	Std.
Milk yield	17.51	5.93	17.26	5.84
Protein yield	0.73	0.27	0.72	0.26
Fat yield	0.59	0.20	0.58	0.19

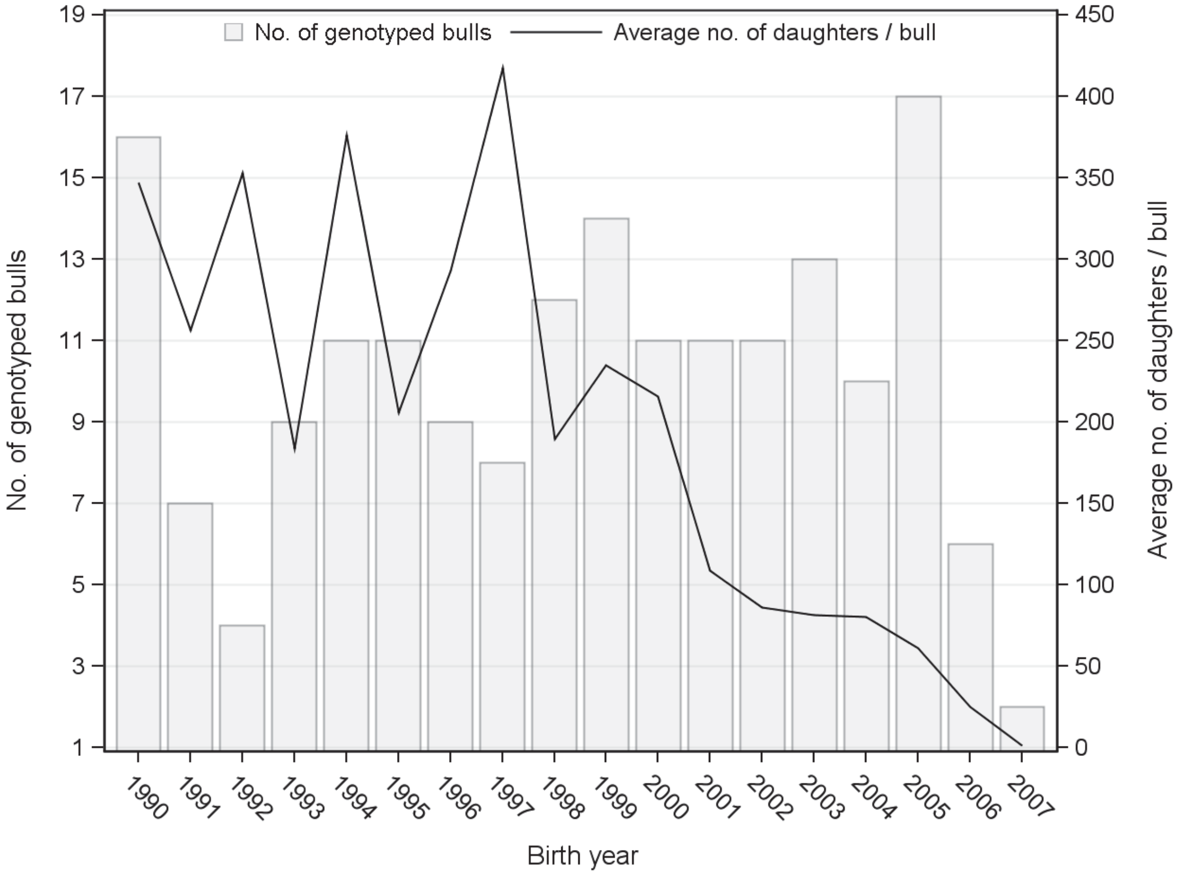

Figure 1. Number of genotyped bulls born per year (bars) and average number of daughters per genotyped bull

The estimates of (co)variance component ratios and correlations for all four approaches are shown in Appendix in Table A1. Heritability estimates with the NAT approach were moderate for daily milk (0.27 and 0.28), fat (0.20 and 0.21), and protein yield (0.23 and 0.25) in reduced and full datasets. The results are consistent with the other studies reported for milk traits (Interbull, 2000; 2008). NATss approach gave virtually equal estimates so we did not report them separately. In the MT1 and MT2 approach, DGVs were considered available from the external evaluation system and were included in the genetic evaluation as a correlated trait (Kachman, 2008). Heritabilities for polygenic breeding value as well as other components virtually did not change with the MT1 and MT2 approach. As expected, heritabilities for marker based breeding value (i.e., DGV) was equal to one since DGV is computed for all animals based on the formula employing the same marker effects for all animals and therefore behaves as a fully heritable trait with complete penetrance. In parallel the environmental variation for DGV was practically non-existent and the corresponding variance component is not shown in the Table A1. Genetic correlation between polygenic and marked based breeding value for all traits ranged from 0.89 to 0.82 in the reduced datasets and from 0.94 to 0.89 in the full datasets with the MT1 approach. Lower genetic correlations were obtained with the MT2 approach than with MT1 approach (from 0.70 to 0.60 in reduced datasets and from 0.86 to 0.79 in full datasets). These results suggest that DGV is as expected a useful early predictor of EBV. Applications of bivariate approach for dairy traits are not present in the literature. For beef populations, MacNeil et al. (2010b) reported lower genetic correlation of 0.38 between EBV and DGV for marbling trait in American Angus. For Australian Angus, genetic correlations between EBV and Pfizer Animal Genetics DGV for marbling trait were in range from 0.02 to 0.19 (Beef CRC, 2012). These correlations are much lower than in our case because of several factors. Training set for the development of DGV equations was much larger in the InterGenomics consortium (5320 progeny tested bulls; Jorjani et al., 2012) giving high accuracy of DGV. In addition, dairy populations have smaller effective population size and therefore higher linkage disequilibrium which is a prerequisite that SNP markers can capture QTLs (Goddard, 2009). Finally, part of the data used for the development of DGV equation comes from Slovenian population (191 bulls) which leads to double counting of information in our analysis. However, given that the number of Slovenian bulls in the InterGenomics consortium is small the amount of double counting should not be severe. Mäntysaari and Strandén (2010) proposed a method for correction of double counting with bivariate models for EBV and DGV and found that required corrections tend to be small. Therefore, we neglected the issue of double counting in our analysis.

The inclusion of SNP information should in theory improve the accuracy of EBV especially for young animals due to the ability to account for Mendelian variation. This has been evaluated first with the theoretical accuracy of EBV that can be calculated based on prediction error variance from diagonal elements of the coefficient matrix inverse. Since the accuracy is a reflection of the amount of information (data), it is reasonable to analyse accuracies for bulls over time; first as young bulls with only parent average information (reduced dataset) and later as progeny tested bulls (full dataset).

Average theoretical accuracies of EBV of 35 validation bulls with different approaches in the reduced dataset were 0.58 (NAT), 0.61 (NATss), 0.90 (MT1), and 0.79 (MT2) for milk yield (Table 2). Inclusion of SNP information in the national evaluation system via the modified relationship matrix **H** (see methods) in the single step approach (NATss) improved the accuracy of EBV only marginally in comparison to the NAT approach (from 0.58 to 0.61). This is due to the fact that the number of genotyped bulls was simply too small (only 182) and there is not enough additional information. Increase in the theoretical accuracy of EBV was observed for milk yield using bivariate analysis (Table 2). Increase in accuracy with the bivariate on top of univariate analysis depends on the genetic correlation between traits. Since genetic correlations between EBV and DGV were high (Table A1) inclusion of DGV as a correlated trait leads to the increase in the theoretical accuracy of EBV for young (validation) bulls.

With the MT1 and MT2 approach, the average theoretical accuracy was 0.90 and 0.79 which again confirms the expected increase in theoretical accuracy due to the inclusion of SNP information (via DGV) in the national evaluation system. Higher accuracies with the MT1 approach than with MT2 approach are a direct consequence of higher estimated genetic correlation between EBV and DGV. Results from the international InterGenomics consortium (Jorjani et al., 2012) suggest that obtained accuracies with the MT1 approach are unrealistically high, while obtained accuracies with the MT2 approach are in line with international results. More realistic accuracies with the MT2 approach could be explained by a larger number of bulls with DGV information in the system (399 bull for the MT2 approach versus 182 bulls for the MT1 approach), which should lead to more precise estimate of genetic correlation between EBV and DGV. In addition some bulls of foreign origin whose DGV was included in the MT2 approach had sizeable numbers of daughters providing information for estimation of genetic correlation.

With the full dataset, the average theoretical accuracies for validation bulls for milk yield were high already with the NAT approach (0.92) based on phenotype and pedigree information and the inclusion of SNP information did not lead to the substantial increase of accuracies. The same is true for a group of all genotyped bulls or all animals, where the average accuracies were dominated by the older progeny tested bulls or phenotyped cows, respectively, and inclusion of SNP information either in the reduced or the full dataset did not lead to the substantial increase of accuracies.

Table 2. Average theoretical accuracies for milk traits among different datasets

Approach[1]	NAT		NATss		MT1		MT2	
Dataset	Reduced	Full	Reduced	Full	Reduced	Full	Reduced	Full
				Milk yield				
Valid. bulls	0.58	0.92	0.61	0.92	0.90	0.97	0.79	0.95
All gen. bulls	0.87	0.96	0.88	0.96	0.96	0.98	0.93	0.97
All animals	0.72	0.77	0.72	0.77	0.73	0.76	0.73	0.76
				Fat yield				
Valid. bulls	0.51	0.92	0.60	0.91	0.86	0.96	0.77	0.95
All gen. bulls	0.85	0.96	0.88	0.96	0.95	0.98	0.92	0.97
All animals	0.68	0.76	0.71	0.76	0.72	0.75	0.71	0.75
				Protein yield				
Valid. bulls	0.58	0.92	0.61	0.92	0.85	0.96	0.74	0.95
All gen. bulls	0.87	0.96	0.88	0.96	0.95	0.98	0.92	0.97
All animals	0.72	0.76	0.72	0.76	0.73	0.76	0.73	0.76

[1] NAT - national genetic evaluation, NATss - single-step methodology (national genetic evaluation with modified relationship matrix), MT1 - bivariate analysis (national genetic evaluation and DGV for Slovenian genotyped bulls only, MT2 - bivariate analysis (national genetic evaluation and DGV for all genotyped bulls in the national pedigree)

Very similar trends in average theoretical accuracies were observed for fat and protein yield as with milk yield (Table 2). For fat yield, average theoretical accuracies for validation bulls were considerably higher with the NATss (0.60) than with the NAT approach (0.51), which might be explained that the NATss approach captured the large effect of DGAT gene (Grisart et al., 2002) even though the number of genotyped bulls was small. In milk and protein yield, such increase was not observed likely due to the genetic architecture of the traits (e.g., Hayes et al., 2010). With the MT1 approach, the increase in average theoretical accuracy for fat and protein yield traits in validation bulls was consistently higher (0.86 and 0.85) than with the

NAT or NATss approach showing the value of using available DGV information as a correlated trait. With the MT2 approach, average theoretical accuracy were more realistic than with MT1 (0.77 for fat yield and 0.74 protein yield) as it was the case for milk yield. In the full dataset, the average theoretical accuracies of fat and protein yield for validation bulls were high with the NAT approach (0.92 for both traits) and the inclusion of SNP information did not lead to significant increase of accuracies. For a group of all genotyped bulls or all animals, inclusion of SNP information and DGV in both datasets did not lead to substantial increase of accuracies due to reasons explained before.

The validation accuracy, i.e., correlation between EBVs from the reduced and the full datasets are summarized by group of animals and approach in the Table 3. For validation bulls, validation accuracies with the NAT and NATss approach were lower in comparison to the average theoretical accuracies (Table 2). The difference between theoretical and validation accuracies was generally lower for the NAT approach (-0.09 for milk yield, -0.01 for fat yield, and -0.02 for protein yield) than for the NATss approach (-0.14 for milk yield, -0.07 for fat yield, and -0.07 for protein yield). Difference was bigger for milk yield, which could be explained by the fact that theoretical accuracies are inflated in the case of selection (Bijma, 2012) - selection in Slovenian Brown Swiss is often more intense for milk yield than for fat and protein yields. Comparison of theoretical and validation accuracies also shows that perceived small increase in accuracies with the NATss approach vanishes when validation is being performed. This is due to the fact that there have been simply too few genotyped animals for the NATss approach. With the MT1 approach, the differences between theoretical and validation accuracies were small: +0.02 for milk yield, +0.05 for protein yield, and +0.02 for fat yield with values around 0.90. The values of validation accuracies are unrealistically high for reasons explained before. With the MT2 approach, validation accuracies were more realistic and of the same magnitude as theoretical accuracies.

Our validation accuracies (from the MT2 approach) are comparable with the results of Harris et al. (2008) in New Zealand, Hayes et al. (2009b) in Australia, and VanRaden et al. (2009) in USA. These authors used so called multi-step methods of inferring EBV on the national scale by first developing DGV equation and later blending it with conventional EBV. Alternative method is single-step that combines all the available national information in one analysis (Legarra et al., 2009; Aguilar et al., 2010). Přibyl et al. (2012) used such approach in Czech Holstein population and obtained an improvement in the accuracy of the resulting

EBV. In our case the improvement with this method (NATss) was negligible due to too small national reference population. Our results and reports from the literature suggest that the MT2 approach is the most suitable among tested approaches to integrate external DGV information into Slovenian evaluation system with small reference population. Due to potential double counting of information in the MT2 approach further research should be performed with the approach suggested by Vandenplas and Gengler (2012).

Table 3. Validation accuracies for milk traits between full dataset and different reduced datasets

Approach[1]	NAT	NATss	MT1	MT2
		Milk yield		
Valid. bulls	0.49	0.47	0.92	0.74
All gen. bulls	0.75	0.75	0.95	0.87
All animals	0.91	0.91	0.92	0.92
		Fat yield		
Valid. bulls	0.50	0.53	0.91	0.81
All gen. bulls	0.75	0.77	0.94	0.89
All animals	0.92	0.92	0.92	0.93
		Protein yield		
Valid. bulls	0.56	0.54	0.87	0.72
All gen. bulls	0.78	0.78	0.92	0.87
All animals	0.91	0.91	0.92	0.92

[1] NAT - national genetic evaluation, NATss - single-step methodology (national genetic evaluation with modified relationship matrix), MT1 - bivariate analysis (national genetic evaluation and DGV for Slovenian genotyped bulls only, MT2 - bivariate analysis (national genetic evaluation and DGV for all genotyped bulls in the national pedigree)

CONCLUSION

Different approaches of integrating genomic information into Slovenian national evaluation system for Brown breed were tested and compared to the conventional evaluation based only on phenotypic and pedigree data. Due to small reference population in Slovenia increase in accuracy of evaluation with single-step methodology was negligible. Therefore, integration of external marker based breeding values (DGV) was performed with the use of bivariate model. With this approach genomic information was successfully integrated and lead to the increase in accuracy of evaluation. Use of larger number of animals with DGV information (national and foreign bulls used in Slovenia) showed more realistic levels of theoretical and validation accuracies. Further research is needed to quantify the effect of potential double counting of available information.

REFERENCES

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., Lawlor T.J. (2010): Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93, 743–752.

Beef CRC (2009): Australian beef DNA results. Available from http://www.beefcrc.com.au/Aus-Beef-DNA-results (accessed Mar. 28, 2012).

Bijma P. (2012): Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. Journal of Animal Breeding and Genetics, 1, 1-14.

Gengler H., Mayeres P., Szydlowski M. (2007): A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal, 1, 21–28.

Goddard M.E., Hayes B.J. (2007): Genomic selection. Journal of Animal Breeding and Genetics, 124, 323–330.

Goddard M.E. (2009): Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136, 245–257.

Grisart B., Coppieters W., Farnir F., Karim L., Ford C., Berzi P., Cambisano N., Mni M., Reid S., Simon P., Spelman R., Georges M., Snell R. (2002): Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research, 12, 222–231.

Harris B.L., Johnson D.L., Spelman R.J. (2008): Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 325-330 in Proc. 36th ICAR Biennial Seassion, Niagara Falls, USA.

Hayes B.J., Bowman P.J., Chamberlain A.C., Goddard M.E. (2009a): Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science, 92, 433–443.

Hayes B.J., Bowman F.J., Chamberlain A.C., Verbyla K., Goddard M.E. (2009b): Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution, 41, 51–59.

Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J., Goddard M.E. (2010): Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. Available from PLoS One 6: e1001139. doi:10.1371/journal.pgen.1001139 (accessed Mar. 10, 2012).

Interbull (2000): National genetic evaluation programmes for dairy production traits practiced in Interbull member countries 1999–2000. Interbull Bulletin No. 24. Available from http://wwwinterbull.slu.se/ojs/index.php/ib/issue/view/29 (accessed Mar. 10, 2012).

Interbull (2008): Genetic Evaluations. Information of National and International Evaluations. Description of GES as applied in member countries. Available from http://www-interbull.slu.se/national_ges_info2/framesida-ges.htm (accessed Mar. 10, 2012).

Johnston D.J., Jeyaruban G.J., Graser H.U. (2010). Evaluation of Pfizer Animal Genetics HD 50K MVP calibration. Available from http://agbu.une.edu.au/pdf/Pfizer_50K_September%202010.pdf (accessed Mar. 28, 2012).

Johnston D.J., Tier B., Graser H.U. (2012): A Beef cattle breeding in Australia with genomics: opportunities and needs. Animal Production Science, 52, 100–106.

Jorjani H., Jakobsen J., Hjerpe E., Palucci V., Dürr J. (2012): Status of genomic evaluation in the Brown Swiss populations. Interbull Bulletin, No. 46, 46–54.

Kachman S.D. (2008): Incorporation of marker scores into national genetic evaluations. Pages 92-98 in Proc. 9th genetic prediction workshop. Prediction of genetic merit of animals for selection. Kanas City, Missouri.

Kovač M., Groeneveld E., Garcia Cortes L. (2002): VCE-5, a package for the estimation of dispersion parameters. Pages 741–742 in Proc. 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

Legarra A., Aguilar I., Misztal I. (2009): A relationship matrix including full pedigree and genomic information. Journal of Dairy Science, 92, 4656-4663.

Liu Z., Seefried F.R., Reinhardt F., Rensing S., Thaller G., Reents R. (2011): Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genetics Selection Evolution, 43, 19–28.

MacNeil, M.D., Nkrumah J.D., Northcutt S.L. (2010a): Genetic evaluation of Angus cattle for carcass marbling using ultrasound and genomic indicators. Journal of Animal Science, 88, 517–522.

MacNeil M.D., Northcutt S.L., Schnabel R.D., Garrick D.J., Woodward B.W., Taylor J.F. (2010b): Genetic correlations between carcass traits and molecular breeding values in Angus cattle. Pages 482-485 in Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

Mäntysaari E.A., Strandén I. (2010): Use of bivariate EBV-DGV model to combine genomic and conventional breeding values evaluations. Pages 353-356 in Proc. 9th World Congress on Genetics Applied to Livestock Production (WCGALP), Leipzig, Germany.

Meuwissen T.H.E., Hayes B.H., Goddard M.E. (2001): Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T., Lee D.H. (2002): BLUPF90 and related programs (BGF90). Pages 1-2 in Proc. 7th World Congress on Genetics Applied to Livestock Production (WCGALP), Montpellier, France.

Přibyl J., Haman J., Kott T., Přibylová J., Šimečková M., Vostrý L., Zavadilová L., Čermák V., Růžička Z., Šplíchal J., Verner M., Motyčka J., Vondrášek L. (2012): Single-step prediction of genomic breeding value in a small dairy cattle population with strong import of foreign genes. Czech Journal of Animal Science, 57, 151–159.

SAS (2008): The SAS System for Windows, Version 9.2. SAS Institute, Inc., Cary, USA.

Schaeffer L.R. (2006): Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics, 123, 218–223.

Schefers J.M., Weigel K.A. (2012): Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Animal Frontiers, 1, 4–9.

Solberg T.R., Sonesson A.K., Woolliams J.A., Meuwissen T.H.E. (2008): Genomic selection using different marker types and densities. Journal of Animal Science, 86, 2447–2454.

Strandén I. (2010): Manual for gpig program – pedigree based imputation of genotypes. Biotechnology and Food Research, Biometrical Genetics, MTT Agrifood Research Finland, 31600 Jokioinen, Finland.

Su G., Madsen P., Nielsen U.S., Mäntysaari E.A., Aamand G.P., Christensen O.F., Lund M.S. (2012): Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy Science, 95, 909–917.

Vandenplas J., Gengler N. (2012): Comparison and improvements of different Bayesian procedures to integrate external information into genetic evaluations. Journal of Dairy Science, 95, 1513-1526.

VanRaden P.M. (2008): Efficient methods to compute genomic predictions. Journal of Dairy Science, 91, 4414–4423.

VanRaden P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F., Schenkel F. (2009): Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science, 92, 16–24.

Zumbach B., Jorjani H., Dürr J. (2010): Brown Swiss Genomic Evaluation. Interbull Bulletin No. 42, 45–51.

APPENDIX

Table A1. Estimates of variance component ratios and correlations (± standard errors) for breeding value (a), direct genomic value (dgv), herd (c), permanent environment (p) effects and residual (e) using different datasets

Approach	NAT and NATss		MT1		MT2	
Dataset	Reduced	Full	Reduced	Full	Reduced	Full
			Milk yield			
h_a^2	0.27 ± 0.003	0.28 ± 0.003	0.27 ± 0.003	0.28 ± 0.003	0.27 ± 0.003	0.28 ± 0.003
h_{dgv}^2	/	/	1.00 ± 0.000	1.00 ± 0.000	1.00 ± 0.000	1.00 ± 0.000
c^2	0.23 ± 0.005	0.22 ± 0.005	0.23 ± 0.005	0.23 ± 0.005	0.23 ± 0.005	0.23 ± 0.005
p^2	0.18 ± 0.002	0.18 ± 0.002	0.18 ± 0.002	0.18 ± 0.002	0.18 ± 0.002	0.18 ± 0.002
e^2	0.32 ± 0.002	0.32 ± 0.002	0.32 ± 0.002	0.32 ± 0.002	0.32 ± 0.002	0.32 ± 0.002
Cor(a, dgv)	/	/	0.89 ± 0.030	0.94 ± 0.010	0.70 ± 0.060	0.86 ± 0.030
Cor(e, e_{dgv})	/	/	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
			Fat yield			
h_a^2	0.20 ± 0.003	0.21 ± 0.001	0.20 ± 0.002	0.21 ± 0.002	0.20 ± 0.003	0.21 ± 0.002
h_{dgv}^2	/	/	1.00 ± 0.000	1.00 ± 0.000	1.00 ± 0.000	1.00 ± 0.000
c^2	0.22 ± 0.005	0.21 ± 0.001	0.22 ± 0.005	0.21 ± 0.005	0.22 ± 0.005	0.21 ± 0.005
p^2	0.15 ± 0.001	0.15 ± 0.001	0.15 ± 0.001	0.15 ± 0.002	0.15 ± 0.001	0.15 ± 0.001
e^2	0.43 ± 0.003	0.43 ± 0.001	0.43 ± 0.003	0.43 ± 0.003	0.43 ± 0.003	0.43 ± 0.003
Cor(a, dgv)	/	/	0.83 ± 0.040	0.91 ± 0.020	0.65 ± 0.060	0.80 ± 0.030
Cor(e, e_{dgv})	/	/	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000
			Protein yield			
h_a^2	0.23 ± 0.003	0.25 ± 0.003	0.23 ± 0.003	0.25 ± 0.003	0.23 ± 0.003	0.25 ± 0.003
h_{dgv}^2	/	/	1.00 ± 0.000	1.00 ± 0.000	1.00 ± 0.000	1.00 ± 0.000
c^2	0.29 ± 0.006	0.28 ± 0.005	0.29 ± 0.006	0.28 ± 0.005	0.29 ± 0.006	0.28 ± 0.006
p^2	0.15 ± 0.002	0.15 ± 0.002	0.15 ± 0.002	0.15 ± 0.002	0.15 ± 0.002	0.15 ± 0.001
e^2	0.32 ± 0.003	0.32 ± 0.003	0.32 ± 0.003	0.32 ± 0.003	0.32 ± 0.003	0.32 ± 0.003
Cor(a, dgv)	/	/	0.82 ± 0.004	0.89 ± 0.022	0.60 ± 0.070	0.79 ± 0.003
Cor(e, e_{dgv})	/	/	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000	0.00 ± 0.000

NAT - national genetic evaluation, NATss - single-step methodology (national genetic evaluation with modified relationship matrix), MT1 - bivariate analysis (national genetic evaluation and DGV for Slovenian genotyped bulls only, MT2 - bivariate analysis (national genetic evaluation and DGV for all genotyped bulls in the national pedigree)

5. Genomska selekcija, Kmečki glas (sprejeto v objavo)

Opis genomske selekcije za rejce.

Kmečki glas (sprejeto)

Genomska selekcija

Selekcija je ena od ključnih rejskih metod za izboljšanje prireje domačih živali. Že od nekdaj rejci odbirajo živali, katere imajo boljšo konformacijo, prirejo, odpornost na bolezni, ipd. Odbira živali je v preteklosti temeljila na vsakodnevnih opažanjih in posledično preprostih sklepanjih rejcev, katera žival je boljša in s tem primernejša za nadaljnjo rejo. Rejci so v preteklih stoletjih tako na razmeroma enostaven način razvili populacije različnih pasem domačih živali, katere so nam danes dobro poznane. Z uvedbo modernejših metod selekcije (preko zbiranja podatkov o rodovnikih in prireji, kontrolami ter naprednimi statističnimi metodami) se je genetski napredek pri selekciji domačih živali v zadnjih desetletjih še dodatno povečal. Začela se je tako imenovana selekcija plemenskih živali znotraj pasem.

Genetski napredek v določeni populaciji je odvisen od: 1) **genetske variabilnost v populaciji**, 2) **točnosti** (kako točno je ocenjena plemenska vrednost živali), 3) **intenzivnosti selekcije** (delež odbranih živali) in 4) **generacijskega intervala** (starosti staršev ob odbiri). V določeni populaciji je genetska variabilnost več ali manj konstantna, kar pomeni, da lahko rejci na genetski napredek vplivajo predvsem z vplivom na točnost, intenzivnost selekcije in generacijski interval. Točnost plemenskih vrednosti in dolžina generacijskega intervala sta obratno sorazmerni - v primeru daljšega časovnega zbiranja podatkov pripomoremo k večji točnosti, a hkrati podaljšamo tudi generacijski interval in obratno. Želimo si, da bi bil generacijski interval čim krajši, saj lahko tako v nadaljnjo rejo zelo hitro vključimo nove in boljše živali. Zaradi bioloških omejitev, ki so povezane z nastopom spolne zrelosti živali je skrajševanje generacijskega intervala omejeno. V zadnjih desetletjih so rejci povečevali genetski napredek predvsem z izdatnejšim in čim bolj natančnim testiranjem živali (s kontrolo rodovnikov in prireje), kar je ugodno vplivalo na povečevanje točnosti in intenzivnosti selekcije.

Z uporabo osemenjevanja pri mlečnem govedu, se je razširil preizkus na potomcih. Seme izbranih mladih bikov uporabimo za osemenitev omejenega števila telic ali krav. Na podlagi rezultatov prireje pri njihovih potomkah odberemo le najboljše bike. Seme odbranih bikov je kasneje na voljo vsem rejcem. Takšen preizkus lahko traja tudi do šest let, a nam nudi zelo natančne ocene plemenskih vrednosti bikov (~90% in več). Pri prašičih in drobnici se je zaradi krajšega generacijskega intervala in večjega števila potomcev ter manj podprte selekcijske infrastrukture uveljavila predvsem lastna preizkušnja, kjer spremljamo prirejo direktno na kandidatih za selekcijo. Slednje pri mlečnem govedu ni mogoče, saj lahko lastnosti mlečnosti spremljamo le pri populaciji ženskih osebkov. Pri lastnem preizkusu so točnosti ocen plemenskih vrednosti praviloma manjše, kot pri preizkusu na potomcih, a je hkrati krajši tudi generacijski interval. Z omenjenim načinom selekcije je ravno tako možno doseči znaten genetski napredek. Pri mesnem govedu sta v uporabi oba sistema selekcije - preizkus na potomcih ali lastni preizkus. Sistem lastnega preizkusa je bolj pogost v primeru, kadar bike uporabljamo za naravne pripuste. Poleg naštetih so poznane še druge vrste preizkusov in sistemov selekcije. Vsem sistemom je skupno, da skušamo poiskati ravnotežje med točnostjo ter biološkimi omejitvami (dolžino generacijskega intervala, intenzivnost selekcije, itd.).

Genetski napredek je bil v zadnjih desetletjih znaten, a vendarle omejen zaradi obratno sorazmerne povezave med točnostjo plemenskih vrednosti živali in dolžino generacijskega intervala. Pri tem je potrebno izpostaviti da pri selekcijskem delu praviloma ne razpolagamo z informacijami o genotipih živali, ki so pravzaprav osnova variabilnosti. Do sedaj je praktično ves genetski napredek izhajal zgolj iz podatkov o fenotipu. Z uporabo ustreznih statističnih metod lahko namreč iz podatkov o fenotipu izluščimo oceno genetske (plemenske) vrednosti posamezne živali.

Razvoj na področju molekularne biologije in genetike v preteklih desetletjih nam je omogočil podrobnejše vpogled v genom in iskanje genov. Kljub napredkom količina zbranih informacij ni nudila bistveno večjih premikov pri točnosti selekcije. V zadnjem desetletju je tehnološki razvoj na omenjenem področju vendarle prispeval zelo učinkovite in tudi cenovno dostopnejše metode za določitev znatnega dela genotipa/genoma domačih živali. Takšna (genomska) informacija sama po sebi ni uporabna za selekcijo, jo pa lahko povežemo z obstoječimi informacijami o rodovnikih in prireji (fenotipu) in le te koristno uporabimo pri oceni plemenskih vrednosti genotipiziranih živali, tudi novorojenih. Postopek uporabe genomske informacije pri selekciji živali imenujemo **genomska selekcija**. Postopek pridobivanja potrebnih informacij je sledeč: 1) najprej živali odvzamemo tkivo (dlačni mešički, nosna sluznica, del ušesa tekom rovašenja, seme, kri, …), 2) nato iz tkiva izoliramo DNK molekulo, ki nosi genetski zapis živali, 3) iz izolirane DNK molekule zberemo informacije o določenih delih genoma (genotipizacija) in 4) pridobljeno genomsko informacijo poleg informacij o rodovniku in prireji (kontroli) uporabimo za oceno plemenske vrednosti živali.

Uvedba genomske selekcije omogoča številne prednosti v primerjavi s klasičnimi pristopi k selekciji. Pri govedu lahko z uporabo genomskih informacij novorojenem teletu ocenimo plemensko vrednost s točnostjo okoli ~60%. Točnost takšne ocene je manjša kot pri preizkusu na potomcih (~90%), a nam hkrati omogoča znatno skrajšanje generacijskega intervala. Posledično je lahko zaradi skrajšanja generacijskega intervala (starše nove generacije lahko odberemo prej) genetski napredek pri mlečnem govedu tudi dvakrat večji kot v primerjavi s preizkusom na potomkah. Zaradi tega se je genomska selekcija v zadnjih letih zelo razširila ravno pri mlečnem govedu, saj je pri do sedaj uveljavljenem preizkusu na potomcih obratno sorazmerni odnos med točnostjo in generacijskim intervalom najbolj izrazit. Nekateri rejski programi so recimo povsem opustili klasičen preizkus na potomkah (npr. v Franciji in Avstriji), medtem ko v nekaterih drugih državah uporabljajo genomsko selekcijo za pred odbiro bikov, ki so kasneje še vedno vključeni v preizkus na potomcih. V tujini je uporaba genomske selekcije prisotna tudi pri prašičih, perutnini kakor tudi ovc in drugih vrst domačih živali (psi, konji, …).

V Sloveniji na področju genomske selekcije, zaradi majhnosti populacij in potrebnih finančnih vložkov, nekoliko zaostajamo za razvitimi evropskimi državami. Ministrstvo za kmetijstvo in okolje (MKO) je v sodelovanju z Javno agencijo za raziskovalno dejavnost Republike Slovenije (ARRS) financiralo ciljni raziskovalni projekt preučitve možnosti uvedbe genomske selekcije v Sloveniji. Delo na projektu je potekalo na Oddelku za zootehniko Biotehniške fakultete Univerze v Ljubljani. Rezultati projekta so jasno pokazali, da zaradi majhnosti populacije sami ne moremo uvesti genomske selekcije. Kljub temu je možno z relativno majhno lastno udeležbo in s sodelovanjem z rejskimi programi v tujini uvesti genomsko selekcijo tudi v majhnih populacijah. Povezovanje je še posebej možno pri govedu, saj so populacije različnih držav genetsko dobro povezane preko pogoste uporabe semena elitnih svetovnih bikov. Takšen primer pri nas je rjava pasma govedi, kjer smo lahko zahvaljujoč sodelovanju Zveze rejcev govedi rjave pasme Slovenije v mednarodnemu projektu InterGenomics uvedli genomsko selekcijo v sodelovanju z rejskimi programi v tujini tudi pri nas. V Sloveniji lahko tako danes za živali rjave pasme, že ob rojstvu ocenimo plemenske vrednosti na osnovi genomske informacije in ostalih informacij (rodovnik in prireja sorodnikov). Zaradi finančnih omejitev smo trenutno v projekt InterGenomics vključeni zgolj z lastnostmi mlečnosti (količina prirejenega mleka, mlečnih maščob in mlečnih beljakovin v laktaciji). Upamo, da se bo v prihodnosti nabor lastnosti razširil tudi na lastnosti dolgoživosti, zdravja vimena, konformacije, ipd.

Rutinski postopek zbiranja tkiv, genotipizacije in izračuna plemenskih vrednosti za živali rjave pasme je v zadnjih korakih razvoja. Rutinsko implementacijo za živali rjave pasme lahko pričakujemo konec letošnjega leta, medtem ko bomo morali pri drugih pasmah mlečnega goveda (lisasti in črno-beli) počakati na vzpostavitev sodelovanja s populacijami (rejskimi programi) v drugih državah. Pri prašičih in drobnici ter drugih vrstah domačih živali zaostajamo za bolj razvitimi državami (ZDA, Avstralija, Nemčija, Danska). Zaradi specifičnosti nekaterih naših populacij, ki se razlikujejo od populacij v tujini, bomo morali pri teh populacijah genomsko selekcijo uvesti sami.

Za zaključek velja poudariti, da je genetski napredek v populaciji odvisen od genetske variabilnosti, točnosti ocen plemenskih vrednosti, intenzivnosti selekcije in dolžine generacijskega intervala. V praksi lahko najbolj vplivamo na točnost in intenzivnost selekcije in le nekoliko na generacijski interval. Med temi tremi faktorji je vedno potrebno poiskati ravnotežje. Z uporabo genomskih informacij je možno s sorazmerno veliko točnostjo oceniti plemensko vrednost že pri mladi živali. Kljub temu, da je točnost manjša kot pri klasičnih preizkusih (npr. na potomcih) se generacijski interval tako močno skrajša, da je genetski napredek posledično večji.

Za rejce govedi uvedba genomske selekcije v praksi pomeni, da bodo osemenjevalni centri v obtok pošiljali čedalje večji delež semena genomsko testiranih mladih bikov. Točnost pri genomsko testiranih bikih je manjša, zato rejcem priporočamo uporabo večjega (dva do trikrat več) števila različnih bikov kot so jih bili vajeni uporabljati do sedaj. Rejci bodo lahko v svoji čredi na osnovi genomske informacije, odbirali telice za obnovo lastne črede kmalu po rojstvu z enako točnostjo kot pri mladih genomsko testiranih bikih. Hkrati bodo rejci lahko po istem postopku odbirali najboljša moška teleta za prodajo vzrejališčem plemenskih bikov ali direktno osemenjevalnim centrom. Cena teličke ali teleta bi se morala oblikovati glede na genomsko oceno plemenske vrednosti. Slednje omogoča rejcem dodaten vir zaslužka – še posebej tistim rejcem, ki se temeljito posvečajo selekciji v svojih čredah.

Doc. dr. Gregor Gorjanc
Univerza v Ljubljani
Biotehniška fakulteta
Oddelek za zootehniko