Luan *et al. Genetics Selection Evolution* 2014, **46**:64
http://www.gsejournal.org/content/46/1/64

GSE Genetics Selection Evolution

RESEARCH

Open Access

Genomic prediction based on runs of homozygosity

Tu Luan[1*], Xijiang Yu[1], Marlies Dolezal[2], Alessandro Bagnato[2] and Theo HE Meuwissen[1]

Abstract

Background: Genomic prediction is based on the accurate estimation of the genomic relationships among and between training animals and selection candidates in order to obtain accurate estimates of the genomic estimated breeding values (GEBV). Various methods have been used to predict GEBV based on population-wide linkage disequilibrium relationships ($\mathbf{G}_{\mathbf{IBS}}$) or sometimes on linkage analysis relationships ($\mathbf{G}_{\mathbf{LA}}$). Here, we propose a novel method to predict GEBV based on a genomic relationship matrix using runs of homozygosity ($\mathbf{G}_{\mathbf{ROH}}$). Runs of homozygosity were used to derive probabilities of multi-locus identity by descent chromosome segments. The accuracy and bias of the prediction of GEBV using $\mathbf{G}_{\mathbf{ROH}}$ were compared to those using $\mathbf{G}_{\mathbf{IBS}}$ and $\mathbf{G}_{\mathbf{LA}}$. Comparisons were performed using simulated datasets derived from a random pedigree and a real pedigree of Italian Brown Swiss bulls. The comparison of accuracies of GEBV was also performed on data from 1086 Italian Brown Swiss dairy cattle.

Results: Simulations with various thresholds of minor allele frequency for markers and quantitative trait loci showed that $\mathbf{G}_{\mathbf{ROH}}$ achieved consistently more accurate GEBV (0 to 4% points higher) than $\mathbf{G}_{\mathbf{IBS}}$ and $\mathbf{G}_{\mathbf{LA}}$. The bias of GEBV prediction for simulated data was higher based on the real pedigree than based on a random pedigree. In the analyses with real data, $\mathbf{G}_{\mathbf{ROH}}$ and $\mathbf{G}_{\mathbf{LA}}$ had similar accuracies. However, $\mathbf{G}_{\mathbf{LA}}$ achieved a higher accuracy when the prediction was done on the youngest animals. The $\mathbf{G}_{\mathbf{IBS}}$ matrices calculated with and without standardized marker genotypes resulted in similar accuracies.

Conclusions: The present study proposes $\mathbf{G}_{\mathbf{ROH}}$ as a novel method to estimate genomic relationship matrices and predict GEBV based on runs of homozygosity and shows that it can result in higher or similar accuracies of GEBV prediction than $\mathbf{G}_{\mathbf{LA}}$, except for the real data analysis with validation of young animals. Compared to $\mathbf{G}_{\mathbf{IBS}}$, $\mathbf{G}_{\mathbf{ROH}}$ resulted in more accurate GEBV predictions.

Background

With the development of high-throughput genotyping technologies and the reduction of genotyping costs, genomic selection (GS) has become a practical and effective tool for animal and plant breeding [1,2]. In genomic selection [3], markers that densely cover the genome are expected to be in complete or partial population-wide linkage disequilibrium (LD) with the QTL (quantitative trait loci), which allows a high fraction of the genetic variance to be explained by the markers [4]. The population-wide LD information can be approximated by a relationship matrix based on identity by-state (IBS) ($\mathbf{G}_{\mathbf{IBS}}$ matrix), where the relationships are reflected by the actual proportion of shared marker alleles that are IBS, as a deviation from expected IBS allele sharing in the population. With an animal model similar to the classical mixed model, best linear unbiased prediction (BLUP) of the GEBV can be achieved by replacing the pedigree-based numerator relationship matrix with the $\mathbf{G}_{\mathbf{IBS}}$ matrix (G-BLUP) [5,6].

Habier et al. [7] and Luan et al. [8] found that, although genomic prediction based on IBS information does not in principle require pedigree data, it does use the family structure of the population, since the markers capture the LD that arises from the family structure. This LD allows close genetic relationships between animals within the pedigree, which are explained by linkage analysis (LA). Fernando and Grossman [9] reported a genomic identity-by-descent (IBD) matrix ($\mathbf{G}_{\mathbf{LA}}$ matrix) that contains IBD probabilities within a known pedigree

* Correspondence: tu.luan@nmbu.no
[1]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås N-1432, Norway
Full list of author information is available at the end of the article

and that depicts this LA information. Thus, based on a limited number of generations within the known pedigree, GEBV can be predicted using the $\mathbf{G_{LA}}$ matrix [6].

For genomic prediction based on the $\mathbf{G_{LA}}$ matrix, marker alleles are IBD if they can be traced back to common ancestors in a clearly defined base generation. The probability of IBD is based only on pedigree information and the inheritance of marker alleles is traced within the pedigree. For genomic prediction based on the $\mathbf{G_{IBS}}$ matrix, it is not possible to identify whether IBS marker alleles are IBD or not, since there is no defined base generation, which means that the $\mathbf{G_{IBS}}$ matrix potentially depicts non-recorded relationships that occurred before the base generation of the known pedigree. Thus, a key difference between the $\mathbf{G_{IBS}}$ and $\mathbf{G_{LA}}$ matrices is the number of generations they take into consideration. For the $\mathbf{G_{IBS}}$ matrix, this is limited by the age of the SNPs used, whereas estimates of IBD coefficients with the $\mathbf{G_{LA}}$ matrix are based only on the known pedigree, and founders are considered to be unrelated. Because of selection, mutations that cause genetic variation in the trait of interest may be considerably younger than the mutations that underlie SNPs on the SNP chip that have a high minor allele frequency (MAF), which may be very old since they drifted to high MAF. Moreover, the $\mathbf{G_{LA}}$ matrix may focus on too few generations [8]. Hence, in this work, we developed a relationship matrix, $\mathbf{G_{ROH}}$, based on runs of homozygosity (ROH), which considers a range of relationship ages that is between that considered by $\mathbf{G_{IBS}}$ and $\mathbf{G_{LA}}$.

For $\mathbf{G_{ROH}}$, IBD probabilities are calculated using a multi-locus measure of LD called ROH or haplotype homozygosity [10]. ROH is defined as the probability that all consecutive markers on a pair of homologous chromosome segments, in the same or different individual(s), have identical alleles, which indicates IBD [11]. The probability of IBD can be calculated from the distribution of the length of homozygous chromosome segments that surround an IBD locus, since the mean of this length is approximately $(\log N_e - 1)/2\ N_e$, where N_e is the effective population size [12]. Hayes et al. [11] found that ROH over long distances reflects recent N_e, whereas ROH over small distances reflects the N_e in the more distant past. ROH can be used to measure multi-locus LD between a marker and a QTL. Compared to two-locus measures of LD such as r^2, a major advantage of ROH is that it is generally a less variable indicator of IBD than r^2, since the latter is known to be very variable [13]. The variability of LD measured by ROH decreases as the marker density increases, whereas variability of LD based on r^2 is unaffected by the number of markers [11].

In this paper, we propose a novel method to predict GEBV based on ROH, hereafter referred to as $\mathbf{G_{ROH}}$. Using simulated datasets with various thresholds of MAF for markers and QTL, we compared the accuracy of GEBV prediction using $\mathbf{G_{ROH}}$, $\mathbf{G_{IBS}}$ based on (population-wide) LD, and $\mathbf{G_{LA}}$ based on linkage analysis. In addition, we evaluated the accuracy of prediction of these methods using real data of deregressed EBV of 1086 Italian Brown Swiss bulls.

Methods

Simulation

A forward simulator (http://ihaiwtheoserv.umb.no/tools/xform/xform.tar.gz) was used to simulate populations according to Wright's ideal population model, i.e. with random mating, uniform mutation rate and base pair position, drift/mutation balance, manageable effective size, SNP mutations that are accumulated through generations of spontaneous mutations and recombinations under random mating. The ideal populations had an effective size of 500, a 1:1 sex ratio and a mutation rate of 10^{-8} per base pair per meiosis. To maintain a reasonable computation time, only one chromosome of length 1 Morgan was simulated.

After 10 000 generations of random mating, the genotypes of the newly produced individuals, referred to as generation 0, were recorded. Genotypes for two kinds of pedigrees were created with generation 0, and were used to produce two simulated datasets i.e. Data I and Data II. Data I was based on a sampled pedigree that was based on 25 sires randomly sampled from the previous generation that were randomly mated to 250 dams randomly sampled from the same generation. Each dam had two offspring. This procedure was repeated for eight generations. Genotypes of the last five generations were recorded to form the simulated dataset. The simulation was performed 10 times to obtain 10 replicates of Data I.

For Data II, the genotypes after 10 000 generations of random mating were gene-dropped through a real pedigree of the Italian Brown Swiss population. The population consisted of 11 599 animals, including 3626 founders and their offspring. There were 27 generations in the pedigree. Genotypes simulated for generation 0 were diffused into this pedigree through its founders. The simulated genotypes of the individuals that were genotyped in the real data were recorded to obtain Data II. This simulation was also performed 10 times to obtain 10 replicates.

One thousand SNPs per chromosome and 30 QTL were sampled disjointedly (i.e. QTL loci could not be sampled as marker loci) from the genotypes created above. Five sampling strategies were used according to the SNP allele frequencies to obtain the following five populations in each dataset (Data I and Data II): Population 1 consisted of randomly sampled markers and QTL ($MAF_{SNP} > 0$, $MAF_{QTL} > 0$); Population 2 consisted of markers all sampled with a minimum MAF of 0.1 and QTL with a maximum MAF of 0.1 ($MAF_{SNP} > 0.1$, $MAF_{QTL} < 0.1$);

Population 3 consisted of markers with a minimum MAF of 0.1 and QTL sampled at random ($\mathrm{MAF}_{\mathrm{SNP}} > 0.1$, $\mathrm{MAF}_{\mathrm{QTL}} > 0$); Population 4 consisted of markers sampled at random and QTL sampled with a maximum MAF of 0.1 ($\mathrm{MAF}_{\mathrm{SNP}} > 0$, $\mathrm{MAF}_{\mathrm{QTL}} < 0.1$); Population 5 consisted of markers with a minimum MAF of 0.15 and QTL sampled with a maximum MAF of 0.05 ($\mathrm{MAF}_{\mathrm{SNP}} > 0.15$, $\mathrm{MAF}_{\mathrm{QTL}} < 0.05$). These five populations reflect the varying degrees to which SNPs can be selected for inclusion on the SNP chip based on high MAF and the variable low frequency of QTL due to selection.

The simulated QTL effects were additive and followed a Laplacian distribution with mean 0 and shape parameter 1. The phenotypes were finally simulated by adding random environmental effects that were independently, identically and normally distributed, in order to achieve a heritability of 0.10.

Real data

The real data on 1086 Italian Brown Swiss bulls consisted of genotyping (i.e. 35 706 SNPs) and phenotyping data i.e. de-regressed proofs (DP) for three traits: milk yield (kg), milk fat yield (kg) and milk protein yield (kg). A detailed description of these data is reported in Luan et al. [8].

Cross-validation data

To obtain the cross-validation datasets, the phenotypes of a defined number of individuals were masked. For the simulated data, in Data I we randomly selected 500 of 4500 individuals at a time for each replicate, without replacement, to produce nine non-overlapping cross-validation datasets, i.e., every phenotype was masked once. Therefore, a total of 90 cross-validation datasets were produced for 10 replicates. Similarly, in Data II we randomly selected 181 of 1086 individuals at a time to produce six non-overlapping cross-validation datasets for each replicate, resulting in 60 datasets. The GEBV of the masked individuals were predicted by the genomic prediction methods described in the next section. The correlation coefficient between the GEBV and true breeding values (TBV) was calculated and used as a measure of the GEBV prediction accuracy, and the deviation of the coefficient of regression of TBV on GEBV from 1 was used as a measure of bias. The mean and standard error of the prediction accuracies and biases in the 90 and 60 datasets for Data I and Data II, respectively, were calculated for each population and each prediction method.

For the analysis with real data on 1086 Italian Brown Swiss bulls, two strategies were used to produce cross-validation datasets. The first strategy was the same as that applied to Data II, except that the GEBV were correlated to the masked DP, to obtain a measure of the accuracy of the GEBV (this measure does not have a maximum of 1 since the reliability of the DP is less than 100%). To obtain standard errors, the division into sets and GEBV predictions were replicated 10 times. The second strategy consisted of selecting the youngest bulls as the validation dataset. In practice, to obtain a validation dataset of reasonable size, we selected bulls born in the three most recent years in the pedigree (2003, 2004 and 2005). One hundred and sixteen young bulls were selected, and their GEBV were predicted using data on 970 older animals.

ROH-based relationships

A run of homozygosity is defined as two haplotypes carrying IBS marker alleles from some position i through to some position j. Let ROH(i,j) denote the probability of this occurring (without making any assumptions about marker identity at the border positions (i-1) and (j + 1)). The method to calculate ROH(i,j) was described in detail by Macleod et al. [10]. Briefly, it calculates the probability that no mutation at the marker positions has occurred since the two homologous chromosome segments coalesced into a common ancestor, integrated over all possible coalescence times. The calculations also account for the fact that the segment between markers i and j may consist of a combination of several shorter IBD segments that each coalesced into different common ancestors. The calculations require knowledge on the genetic distances between the markers, their mutation rates, and the effective population size (N_e), which was assumed to be 100. An approximate estimate for the mutation rate at the markers, $\mathrm{m_m}$, is obtained by equating the average homozygosity of all markers to its expected value $1/(1 + 4N_e\mathrm{m_m})$. Here, we will also consider run of homozygosity probability ROH(i,k,j), where a putative focal position k which is in the middle between two consecutive markers forming a marker bracket, is also assumed to be IBS (with $\mathrm{i} < \mathrm{k} < \mathrm{j}$). Since the focal position k is in the middle between two markers, there is no actual marker data at this position. Thus, ROH(i,j) is the sum of the probability ROH(i,k,j) and the probability that all positions except position k carry IBS alleles. Let ROH(−i,−j) denote the probability that all marker alleles between positions i and j are IBS between two haplotypes, but the haplotypes are not IBS at positions (i−1) and (j + 1) (for ROH, we usually observe that all markers between positions i and j are IBS but that IBS does not extend beyond the boundaries i and j). Bounded ROH probabilities can be calculated from unbounded ROH(i,j) probabilities as [14]:

$$\mathrm{ROH}(-\mathrm{i},-\mathrm{j}) = \mathrm{ROH}(\mathrm{i},\mathrm{j}) - \mathrm{ROH}(\mathrm{i}-1,\mathrm{j}) - \mathrm{ROH}(\mathrm{i},\mathrm{j}+1) + \mathrm{ROH}(\mathrm{i}-1,\mathrm{j}+1).$$

Inclusion of an extra position k among the IBS markers is straightforward:

$$\mathrm{ROH}(-\mathrm{i},\mathrm{k},-\mathrm{j}) = \mathrm{ROH}(\mathrm{i},\mathrm{k},\mathrm{j}) - \mathrm{ROH}(\mathrm{i}-1,\mathrm{k},\mathrm{j}) - \mathrm{ROH}(\mathrm{i},\mathrm{k},\mathrm{j}+1) + \mathrm{ROH}(\mathrm{i}-1,\mathrm{k},\mathrm{j}+1).$$

Now, given that we know that all actual markers are IBS between positions i and j and not IBS at positions (i−1) and (j + 1), the IBD probability at position k is defined as the probability that there has been no mutation at this position since its coalescence:

$$\mathrm{PIBD}(\mathrm{k}|{-\mathrm{i}},{-\mathrm{j}}) = \mathrm{ROH}({-\mathrm{i}},\mathrm{k},{-\mathrm{j}})/\mathrm{ROH}({-\mathrm{i}},{-\mathrm{j}}).$$

Here, we need to make an (arbitrary) assumption about the mutation rate at position k ($\mathrm{m_k}$), which is chosen such that the a priori IBD probability at position k is close to 0.5, i.e. $1/(1 + 4N_e\mathrm{m_k}) \approx 0.5$, in order to give the marker data ample opportunity to change the a priori probability either towards 0 (few or no IBS markers in the vicinity of k) or towards 1 (k is in the middle of a long stretch of IBS markers).

IBD probability PIBD(k|−i,−j) is calculated and averaged over all marker brackets in the genome, with the focal position k in the middle of each bracket. The averaged PIBD(k|−i,−j) of all combinations of genotyped animals are stored in a ROH-based relationship matrix, called $\mathbf{G_{ROH}}$. $\mathbf{G_{ROH}}$ is not always positive definite, because its elements are calculated on a one-by-one basis. Therefore, the eigenvalues of $\mathbf{G_{ROH}}$ are checked, negative eigenvalues are set to 0, and the matrix is reconstructed using only the positive eigenvalues. Finally, a small value (0.0001) is added to the diagonals to make $\mathbf{G_{ROH}}$ positive definite.

The calculation of PIBD(k|−i,−j) and $\mathbf{G_{ROH}}$ is implemented in the LDMIP software (http://ihaiwtheoserv.umb.no/tools/ldmip) [15]. Program LDMIP can also use the PIBD(k|−i,−j) probabilities for imputation of missing marker data, i.e. it finds the $\mathrm{N_{hap}}$ haplotypes that resemble the haplotype with a missing marker based on the highest PIBD probability at every position k. Next, it uses the Viterbi algorithm [16] to find, for the current haplotype, a path through these $\mathrm{N_{hap}}$ haplotypes without mismatches between the current and the proposed haplotype and with the fewest number of switches between the $\mathrm{N_{hap}}$ haplotypes. I.e., the algorithm finds a mosaic of the $\mathrm{N_{hap}}$ haplotypes that most closely resembles the current haplotype, and uses this mosaic to impute the missing markers. Because marker phase is often unknown (i.e. a heterozygous genotype is not known to be '1 2' or '2 1'), the Viterbi algorithm is actually applied to resolve both haplotypes of an individual simultaneously, resulting in a mosaic (as explained above) for each of the two haplotypes and resolving the phase of heterozygous genotypes ('1 2' or '2 1'). For this, the Viterbi algorithm considers $\mathrm{N_{hap}}^2$ combinations of the $\mathrm{N_{hap}}$ haplotypes that were selected based on PIBD. Based on some preliminary testing, we found $\mathrm{N_{hap}} = 40$ as a suitable tradeoff between accuracy and computing time. The LDMIP algorithm also yields probabilities of paternal and maternal inheritance at the marker alleles for all animals in the pedigree [15], which can be used to set up a linkage analysis based on the genomic relationship matrix $\mathbf{G_{LA}}$ by setting up such a relationship matrix at all marker positions and averaging across positions [8,9].

GEBV prediction based on IBS, LA and ROH relationships

The model used to predict GEBV with IBS, LA and ROH information can be expressed as:

$$\mathbf{y}=\mathbf{1}\mu\ +\ \mathbf{Za}\ +\ \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes (DP) for a trait; μ is the overall mean; $\mathbf{Z}$ is a design matrix linking the animals to the phenotypes; $\mathbf{a}$ is a vector of additive genetic effects of the animals and $\mathbf{e}$ is the vector of random residuals. It is assumed that $\mathbf{a}_{(.)} \sim N\ (\mathbf{0},\ \mathbf{G}_{(.)}\sigma_{(.)}{}^2)$ where (.) refers to ROH, LA or IBS and $\sigma_{(.)}{}^2$ is the additive genetic variance associated with $\mathbf{G}_{(.)}$.

For GEBV prediction based on IBS relationships, $\mathbf{G}_{(.)}$ is $\mathbf{G_{IBS}}$. Two ways to set up the $\mathbf{G_{IBS}}$ matrix were used here. The first was to construct the $\mathbf{G_{IBS}}$ matrix with standardized marker genotypes ($\mathbf{G_{IBS\text{-}STD}}$) as $\mathbf{G_{IBS\text{-}STD}} = \mathbf{XX'}/N_m$, where N_m is the number of markers and $\mathbf{X}$ is a matrix of the standardized marker genotypes, $X_{ij} = \left(g_{ij}-2p_j\right)/\sqrt{2p_j\left(1-p_j\right)}$, where g_{ij} is the genotype of animal i for SNP j, with $g_{ij} = 0$, 1 or 2 for genotypes "0 0", "1 0" or "1 1", respectively, and p_j is the frequency of allele 1 of SNP j. Standardization is such that the mean and the variance of X_{ij} are 0 and 1, respectively [6]. The second method used to construct the $\mathbf{G_{IBS}}$ matrix was as in VanRaden [5], where markers are not standardized and the IBS matrix is calculated as $\mathbf{G_{IBS-NSTD}} = \mathbf{YY}'/\Sigma(2p_j(1-p_j))$, where $\mathbf{Y}$ is a matrix of non-standardized marker genotypes, i.e. $Y_{ij} = g_{ij}-2p_j$.

For GEBV prediction based on LA relationships, $\mathbf{G}_{(.)}$ is $\mathbf{G_{LA}}$, the LA-based genomic IBD relationship matrix. For a detailed description about models for GEBV prediction based on LA relationships, see Luan et al. [8]. For GEBV prediction based on ROH relationships, $\mathbf{G}_{(.)}$ is $\mathbf{G_{ROH}}$, the ROH-based genomic IBD relationship matrix. To implement the models, $\mathbf{G_{IBS\text{-}STD}}$, $\mathbf{G_{IBS\text{-}NSTD}}$, $\mathbf{G_{LA}}$ and $\mathbf{G_{ROH}}$ were inverted and were then used in ASReml [17] to predict GEBV of both phenotyped and non-phenotyped individuals.

Results

Accuracy and bias of GEBV prediction using simulated data

We evaluated the accuracy and bias of the GEBV obtained by $\mathbf{G_{ROH}}$ in 90 simulated datasets of Data I and in 60 datasets of Data II and compared them with GEBV based on $\mathbf{G_{IBS\text{-}STD}}$, $\mathbf{G_{IBS\text{-}NSTD}}$ and $\mathbf{G_{LA}}$ relationship matrices. Means and standard errors of the accuracies and biases of the GEBV are in Tables 1 and 2 for Data I and II, respectively.

For all datasets, accuracy of the GEBV was higher with $\mathbf{G_{ROH}}$ than with $\mathbf{G_{LA}}$, $\mathbf{G_{IBS\text{-}STD}}$ and $\mathbf{G_{IBS\text{-}NSTD}}$, although the differences were not always statistically significant. Accuracies of GEBV based on $\mathbf{G_{IBS\text{-}STD}}$ and $\mathbf{G_{IBS\text{-}NSTD}}$ were similar. Accuracies of GEBV were higher and biases lower with Data I (Table 1) than with Data II (Table 2), which is expected since the size of the reference population was larger in Data I (4000 individuals) than in Data II (905 individuals). It is also notable that the presence of rare QTL alleles reduced the accuracy much more with real pedigree structures (Data II) than with random pedigrees (Data I). For the simulated dataset with a random pedigree (Data I), accuracies were higher with $\mathbf{G_{LA}}$ than with $\mathbf{G_{IBS}}$ for all QTL allele frequency scenarios (Table 1). For the simulated dataset with real pedigree (Data II), accuracies were also higher with $\mathbf{G_{LA}}$ than with $\mathbf{G_{IBS}}$ when a maximum MAF was applied to the QTL but were lower when QTL were randomly sampled (Table 2).

Results from simulations (Tables 1 and 2) demonstrated that the accuracy of GEBV was affected by the MAF of markers and QTL. The highest accuracy was obtained in the population with a minimum MAF of 0.1 for markers and no MAF threshold applied to QTL. Application of a maximum MAF threshold to QTL appeared to reduce the accuracy of GEBV. For example, for GEBV using $\mathbf{G_{ROH}}$, the accuracies decreased by ~38% and ~11% when QTL had a MAF below 0.1 in Data I and Data II, respectively, compared to when both QTL and SNPs were randomly sampled. When a minimum MAF was applied to markers and a maximum MAF to QTL, accuracies of GEBV were reduced.

Table 1 Means and standard errors of accuracies and biases of GEBV obtained from four methods for populations with different thresholds of MAF for markers and QTL in Data I

MAF		Method			
SNP	QTL	$\mathbf{G_{ROH}}$	$\mathbf{G_{LA}}$	$\mathbf{G_{IBS\text{-}STD}}$	$\mathbf{G_{IBS\text{-}NSTD}}$
Accuracy					
> 0	> 0	0.750 ± 0.003	0.676 ± 0.004	0.725 ± 0.005	0.720 ± 0.005
> 0.10	< 0.10	0.704 ± 0.004	0.623 ± 0.006	0.669 ± 0.005	0.665 ± 0.005
> 0.10	> 0	0.760 ± 0.004	0.669 ± 0.005	0.751 ± 0.005	0.751 ± 0.005
> 0	< 0.10	0.721 ± 0.003	0.649 ± 0.004	0.671 ± 0.005	0.661 ± 0.005
> 0.15	< 0.05	0.699 ± 0.005	0.643 ± 0.007	0.653 ± 0.006	0.652 ± 0.006
Bias[1]					
> 0	> 0	1.017 ± 0.007	1.005 ± 0.009	1.009 ± 0.008	1.006 ± 0.007
> 0.10	< 0.10	1.002 ± 0.010	0.984 ± 0.011	1.014 ± 0.011	1.013 ± 0.012
> 0.10	> 0	1.020 ± 0.010	1.030 ± 0.011	1.023 ± 0.010	1.023 ± 0.010
> 0	< 0.10	1.029 ± 0.010	1.025 ± 0.011	1.043 ± 0.012	1.039 ± 0.013
> 0.15	< 0.05	1.026 ± 0.013	1.056 ± 0.017	1.021 ± 0.014	1.021 ± 0.014

[1] Bias is calculated as the regression of TBV on GEBV; bias is equal to 1 if the GEBV prediction is unbiased.

Table 2 Mean and standard error of the accuracies and biases of GEBV prediction for populations with different thresholds of MAF for markers and QTL in Data II

MAF		Method			
SNP	QTL	$\mathbf{G_{ROH}}$	$\mathbf{G_{LA}}$	$\mathbf{G_{IBS\text{-}STD}}$	$\mathbf{G_{IBS\text{-}NSTD}}$
Accuracy					
> 0	> 0	0.600 ± 0.009	0.570 ± 0.010	0.582 ± 0.011	0.579 ± 0.011
> 0.10	< 0.10	0.466 ± 0.013	0.457 ± 0.012	0.425 ± 0.012	0.424 ± 0.012
> 0.10	> 0	0.603 ± 0.008	0.571 ± 0.009	0.597 ± 0.009	0.594 ± 0.009
> 0	< 0.10	0.533 ± 0.016	0.528 ± 0.016	0.496 ± 0.017	0.490 ± 0.017
> 0.15	< 0.05	0.406 ± 0.015	0.395 ± 0.016	0.369 ± 0.015	0.368 ± 0.015
Bias[1]					
> 0	> 0	1.101 ± 0.041	1.055 ± 0.037	1.114 ± 0.045	1.137 ± 0.051
> 0.10	< 0.10	1.327 ± 0.149	1.357 ± 0.131	1.364 ± 0.153	1.393 ± 0.170
> 0.10	> 0	0.961 ± 0.026	0.983 ± 0.033	0.955 ± 0.030	0.948 ± 0.030
> 0	< 0.10	1.140 ± 0.052	1.110 ± 0.049	1.157 ± 0.056	1.225 ± 0.064
> 0.15	< 0.05	1.479 ± 0.215	1.795 ± 0.356	1.776 ± 0.286	1.728 ± 0.271

[1] Bias is calculated as the regression of TBV on GEBV.

Correlation between GEBV and DP and bias of GEBV using real data

To investigate the performance of the $\mathbf{G_{ROH}}$-based method in practice, we applied $\mathbf{G_{ROH}}$ to real DP datasets of 1086 Italian Brown Swiss bulls for fat yield, milk yield and protein yield. Table 3 presents the correlations between GEBV and DP and biases of GEBV with $\mathbf{G_{ROH}}$, $\mathbf{G_{IBS\text{-}STD}}$, $\mathbf{G_{IBS\text{-}NSTD}}$ and $\mathbf{G_{LA}}$, since the TBV is unknown in the real dataset. The bias was calculated as the regression of DP on GEBV. However, it should be noted here that any under- or over-scaling of the DP by the deregression process will appear as regression coefficients that deviate from 1, i.e. as bias [8]. Table 3 shows that the correlations obtained with $\mathbf{G_{ROH}}$ were higher than those with $\mathbf{G_{IBS}}$ and very similar to those with $\mathbf{G_{LA}}$ but with a substantially

Table 3 Mean and standard error of the correlation between GEBV and DP, and biases of the GEBV evaluated with cross-validation in real data

Trait	$\mathbf{G_{ROH}}$	$\mathbf{G_{LA}}$	$\mathbf{G_{IBS\text{-}STD}}$	$\mathbf{G_{IBS\text{-}NSTD}}$
Correlation				
Fat yield	0.768 ± 0.007	0.768 ± 0.010	0.751 ± 0.010	0.752 ± 0.009
Milk yield	0.763 ± 0.007	0.762 ± 0.010	0.748 ± 0.009	0.748 ± 0.009
Protein yield	0.784 ± 0.007	0.784 ± 0.010	0.768 ± 0.009	0.767 ± 0.009
Bias[1]				
Fat yield	1.002 ± 0.031	1.132 ± 0.026	1.022 ± 0.026	1.015 ± 0.027
Milk yield	1.004 ± 0.024	1.124 ± 0.020	1.026 ± 0.021	1.018 ± 0.021
Protein yield	1.009 ± 0.023	1.130 ± 0.020	1.036 ± 0.020	1.027 ± 0.020

[1] Bias is calculated as the regression of DP on GEBV.

lower regression coefficient. In agreement with results from the simulated data, the $\mathbf{G_{IBS\text{-}STD}}$- and $\mathbf{G_{IBS\text{-}NSTD}}$-based methods resulted in similar correlations and biases. The correlation obtained for protein yield was higher than that for fat and milk yields. The standard error of the correlations obtained with the $\mathbf{G_{ROH}}$-based method was smaller than that of the other methods, which indicates that the results were less variable. This is in line with the expectation of Hayes et al. [11] that multi-locus measures of LD are less variable.

Correlations between GEBV and DP and biases of GEBV obtained when $\mathbf{G_{ROH}}$, $\mathbf{G_{LA}}$ and $\mathbf{G_{IBS}}$ were used to predict the group of 116 young bulls are in Table 4. In contrast to the analyses with randomly selected cross-validation bulls, GEBV prediction for the group of young bulls was performed only once and thus no standard errors were available. The correlation between GEBV and DP was higher with $\mathbf{G_{LA}}$ than with $\mathbf{G_{ROH}}$ and $\mathbf{G_{IBS}}$ (Table 4). Also, the $\mathbf{G_{IBS\text{-}NSTD}}$-based method resulted in slightly higher correlations than the $\mathbf{G_{IBS\text{-}STD}}$-based method. Figures 1 and 2 show the scatter-plots of 1086 diagonal and 589 156 off-diagonal entries of $\mathbf{G_{ROH}}$, $\mathbf{G_{IBS\text{-}STD}}$ and $\mathbf{G_{IBS\text{-}NSTD}}$ versus the $\mathbf{G_{LA}}$ matrix for the real data. Regression lines of the entries of $\mathbf{G_{ROH}}$, $\mathbf{G_{IBS\text{-}STD}}$ and $\mathbf{G_{IBS\text{-}NSTD}}$ matrix on those of the $\mathbf{G_{LA}}$ matrix are also shown in these figures.

Discussion

In this study, we proposed $\mathbf{G_{ROH}}$-based genomic prediction, a novel method to compute GEBV based on runs of homozygosity. Runs of homozygosity yield a multi-locus measure of LD, from which a measure of IBD is derived, which we expected to be less variable than the IBD derived by single-locus measures of LD and thus to result in an increased accuracy of GEBV. Using simulated and real data, the accuracy and bias of GEBV based on $\mathbf{G_{ROH}}$ were compared to those based on $\mathbf{G_{LA}}$ and $\mathbf{G_{IBS}}$. Results from simulation analyses showed that, in general, $\mathbf{G_{ROH}}$ resulted in more accurate GEBV,

Table 4 The correlation between GEBV and DP, and biases of GEBV evaluated for the group of young animals in the real data

Trait	G_{ROH}	G_{LA}	$G_{IBS\text{-}STD}$	$G_{IBS\text{-}NSTD}$
Correlation				
Fat yield	0.447	0.500	0.400	0.415
Milk yield	0.410	0.415	0.376	0.389
Protein yield	0.385	0.410	0.363	0.368
Bias[1]				
Fat yield	0.946	1.277	0.835	0.855
Milk yield	0.829	1.013	0.778	0.789
Protein yield	0.763	1.007	0.736	0.733

[1] Bias is calculated as the regression of DP on GEBV.

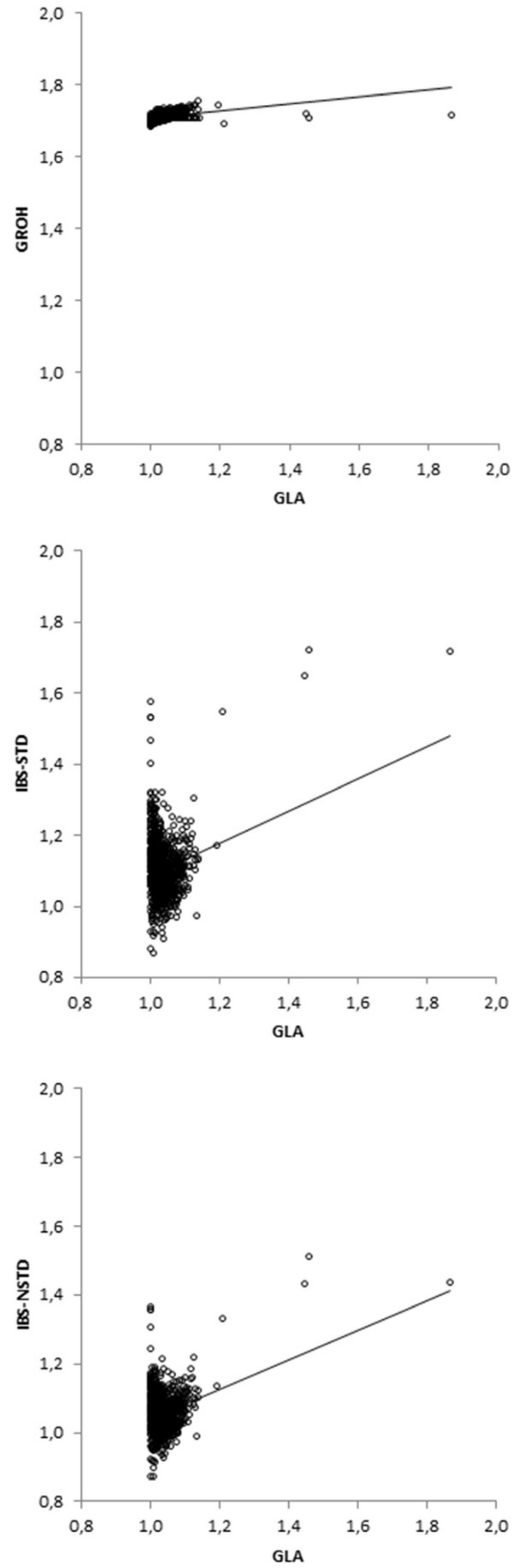

Figure 1 Scatter plot of diagonal entries of G_{ROH} matrix and G_{IBS} matrices versus G_{LA} matrix.

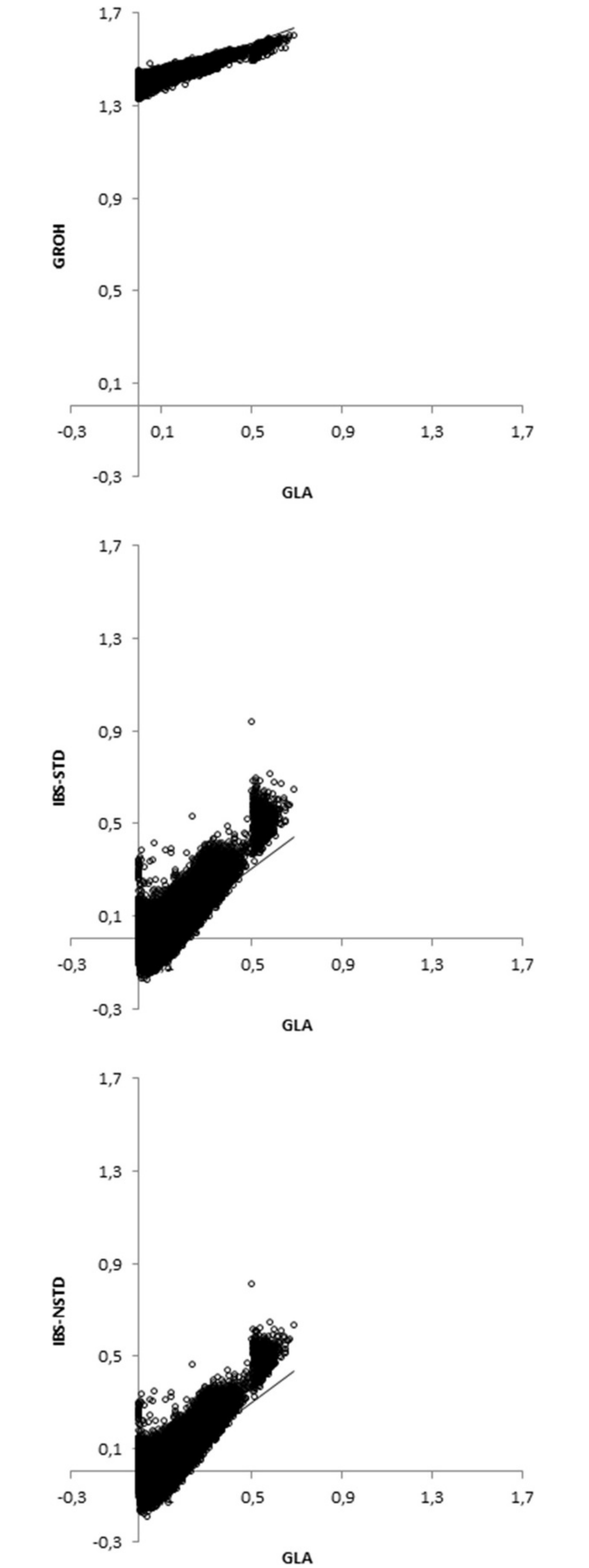

Figure 2 Scatter plots of off-diagonal entries of G_{ROH} matrix and G_{IBS} matrices versus G_{LA} matrix.

up to 4% points higher. With the real data, the accuracy of GEBV was higher with $\mathbf{G_{ROH}}$ than with $\mathbf{G_{IBS}}$ but only slightly higher than with $\mathbf{G_{LA}}$. Predictions using $\mathbf{G_{ROH}}$ and, especially, $\mathbf{G_{LA}}$ were less affected by the difference in allele frequencies between QTL and markers, probably because they do not rely on pair-wise LD between QTL and markers. A possible explanation of the difference in results between real and simulated data may lie in the difference in the population structures in the datasets. In the real dataset, we found that recent family relationships are strong in the population of bulls used [8]. Thus, in the real data, older and more distant relationships may contribute little to the accuracy of GEBV. This favors $\mathbf{G_{LA}}$, which relies on recent family relationships to predict GEBV, whereas $\mathbf{G_{IBS}}$ relies on more distant relationships and hence yielded lower accuracies than $\mathbf{G_{LA}}$ for the real data. Matrix $\mathbf{G_{ROH}}$ captures relationships that span a range of ages of relationships that is intermediate between those captured by $\mathbf{G_{LA}}$ and $\mathbf{G_{IBS}}$. Matrix $\mathbf{G_{ROH}}$ also takes recent family relationships much more in account and resulted in accuracies of GEBV that were as high as with $\mathbf{G_{LA}}$ for the real data. In simulated data based on a sampled pedigree (Data I), the older and more distant relationships contributed more to the accuracy of GEBV, which favors $\mathbf{G_{IBS}}$ and thus the accuracy obtained with $\mathbf{G_{IBS}}$ was higher than with $\mathbf{G_{LA}}$ for Data I.

For the simulated data based on the real pedigree from the population with strong recent family relationships (Data II), the results in Table 2 show that the performance of $\mathbf{G_{LA}}$ and $\mathbf{G_{IBS}}$ depended on the simulation scenario. If QTL were randomly sampled, the older and more distant relationships were important because this scenario allows old QTL mutations to still contribute to the genetic variance. Matrix $\mathbf{G_{IBS}}$ can capture such relationships and hence achieved higher accuracy than $\mathbf{G_{LA}}$ for that scenario. If QTL were sampled with a maximum threshold of MAF, then mainly recent QTL mutations contribute to the genetic variance. This implies that recent family relationships are important for the accuracy of GEBV and that more distant relationships contribute little, as indicated by the fact that $\mathbf{G_{LA}}$ achieved higher accuracy than $\mathbf{G_{IBS}}$ for this scenario (Table 2).

In the simulation results, $\mathbf{G_{ROH}}$ achieved higher accuracies than $\mathbf{G_{LA}}$ and $\mathbf{G_{IBS}}$ in all cases. In cases where older and more distant relationships are important (Table 1 and $MAF_{QTL} > 0$ in Table 2), $\mathbf{G_{IBS}}$ can capture such relationships and, therefore, achieved higher accuracies than $\mathbf{G_{LA}}$. However, the capacity of $\mathbf{G_{IBS}}$ to capture information on old relationships depends on the use of uncertain relationships between ancestors, which may undermine the performance of $\mathbf{G_{IBS}}$. In contrast, $\mathbf{G_{ROH}}$ does not look back as many generations as $\mathbf{G_{IBS}}$ and thus yields higher accuracies than $\mathbf{G_{IBS}}$. In cases where recent family relationships are important (QTL sampled with a MAF threshold in

Table 2), $\mathbf{G_{LA}}$ achieved higher accuracies than $\mathbf{G_{IBS}}$. The capacity of $\mathbf{G_{ROH}}$ to capture non-recorded relationships before the base generation of the known pedigree means that $\mathbf{G_{ROH}}$ can use such information and thus achieve higher accuracies than $\mathbf{G_{LA}}$. The latter would be especially useful for across-breed prediction, as in the case when the training population contains several breeds. Thus, it appears that our novel method $\mathbf{G_{ROH}}$ can benefit from the favorable properties of both $\mathbf{G_{LA}}$ and $\mathbf{G_{IBS}}$.

The relationship matrices used in this study differ in the chromosomal distance they consider. Matrix $\mathbf{G_{LA}}$ can consider a distance up to one complete chromosome, while $\mathbf{G_{IBS}}$ relies on pair-wise LD between markers and QTL, which stretches only over small chromosomal distances. Matrix $\mathbf{G_{ROH}}$ uses multi-locus LD and thus can account for larger chromosomal distances to capture LD than $\mathbf{G_{IBS}}$. Matrix $\mathbf{G_{ROH}}$ seems to strike a balance between short and long range LD and resulted in the highest prediction accuracies for a range of situations, except for prediction of the youngest animals in the data set, which is however a typical scenario for genomic selection in practice.

Similar to the way both marker and pedigree information are used in matrix $\mathbf{G_{LA}}$, Goddard et al. [4] proposed a method to obtain an unbiased estimate of relationships for genomic prediction by regressing the IBS matrix onto the pedigree-based relationship matrix. Their method uses the relationship matrix $\hat{\mathbf{G}} = \mathbf{A}+b(\mathbf{G_{IBS\text{-}STD}}-\mathbf{A})$ for genomic prediction, where $\mathbf{A}$ is the relationship matrix based on pedigree, and the regression coefficient b reflects the proportion of genetic variance explained by the markers. This method attempts to take the whole range of population structures into consideration: matrix $\mathbf{A}$ accounts for recent relationships and matrix $\mathbf{G_{IBS\text{-}STD}}$ for distant relationships. Therefore, when recent family relationships are more important, use of matrix $\hat{\mathbf{G}}$ may yield higher accuracies than use of $\mathbf{G_{IBS}}$. The regression coefficient b depends on marker density and putative differences in the properties of markers vs. QTL. If QTL and markers do not systematically differ, for example markers and QTL are randomly sampled, as explored in this study, b can be predicted from the marker data. If QTL and markers differ systematically (e.g. $\mathrm{MAF_{SNP}} > 0.15$ and $\mathrm{MAF_{QTL}} < 0.05$ in this study), the regression coefficient b should be estimated from the data, which can be achieved by fitting the $\mathbf{A}$ and $\mathbf{G_{IBS\text{-}STD}}$ matrices jointly in a variance component estimation model.

Accuracies were lower for prediction of young animals than for random cross-validation datasets. This may be due to the fact that the Mendelian sampling component of their TBV, i.e. about half of the total variance of the TBV, is uncorrelated to any of the training records. The results for the group of young animals in the real data show that the highest accuracy was obtained with $\mathbf{G_{LA}}$, followed by $\mathbf{G_{ROH}}$ and $\mathbf{G_{IBS}}$. A possible explanation may be that the three methods capture different relationship ages. Matrix $\mathbf{G_{LA}}$ only focuses on the known pedigree, for which the relationships of young animals can be well-defined. With $\mathbf{G_{IBS}}$, uncertain relationships between ancestors prior to the known pedigree may deteriorate the ability to capture relationships between young animals. Matrix $\mathbf{G_{ROH}}$ captures information on relationships for ages that are intermediate between those of $\mathbf{G_{LA}}$ and $\mathbf{G_{ROH}}$, and thus achieves an intermediate accuracy.

The performance of the methods also depends on the effective size of the population (N_e). If N_e is small, common ancestors tend to be in the recent past and recent family relationships tend to dominate the population structure. It is expected that $\mathbf{G_{LA}}$ performs better than $\mathbf{G_{ROH}}$. If N_e is large, distant family relationships occur frequently and the performance of $\mathbf{G_{LA}}$ deteriorates. Thus, it is expected that $\mathbf{G_{LA}}$ will perform worse than $\mathbf{G_{ROH}}$ in a population with a large N_e and when the training population consists of a mixture of different breeds. In the analysis with real data, $\mathbf{G_{LA}}$ was found to give higher accuracies than $\mathbf{G_{ROH}}$, while in the simulation study with $N_e = 500$, $\mathbf{G_{ROH}}$ performed better. This suggests that the Italian Brown Swiss bull population has a smaller N_e than 500, which is also suggested by its small number of sires (21). Goddard et al. [4] pointed out that variation in relationships between animals in a population increases with N_e. Thus, $\mathbf{G_{ROH}}$ is expected to result in higher accuracies of prediction than $\mathbf{G_{LA}}$ when variation in relationships is small, such as between breeds.

In the simulation study, we used two methods to calculate the matrix $\mathbf{G_{IBS}}$, which differed in whether markers were standardized or not prior to its calculation. It is known that standardizing markers increases the weight placed on low MAF markers [18]. The effects of markers with low MAF are estimated with much lower accuracy. This suggests that the standardization of markers may result in different accuracies. However, our results show that the two methods of computing $\mathbf{G_{IBS}}$ resulted in similar accuracies. This agrees with the expectation of Sonesson et al. [18].

The simulated and real data results were quite different even when the real pedigree was used in the simulations. The changes in allele frequencies between markers and QTL introduced in Table 2 did not result in simulated results being closer to the real data results of Table 3. Possibly, in the real data, the QTL do not have a MAF as low as that simulated in Table 2 because they have been recently selected and the population has high rates of inbreeding, which causes low allele frequencies to drift towards intermediate values. It seems that LA information was much more important for the analysis of real data than that of simulated data. A possible explanation is the much higher reliability of the deregressed proofs compared to that of the simulated trait ($h^2 = 0.1$), which also resulted in the higher cross-validation accuracies in the real data. This high reliability of the de-regressed proofs resulted in accurate estimation of chromosomal segments in the linkage analysis, while the low heritability of the simulated trait implies that the long-term, LD-based genetic effects also need to be estimated to achieve high cross-validation accuracy.

Conclusions

The present study proposes a novel method, $\mathbf{G}_{\mathbf{ROH}}$, to predict GEBV based on runs of homozygosity. Through computer simulations, we showed that the accuracy of GEBV was higher with $\mathbf{G}_{\mathbf{ROH}}$ than with $\mathbf{G}_{\mathbf{IBS}}$ and $\mathbf{G}_{\mathbf{LA}}$. In the analyses of real data, accuracies obtained with $\mathbf{G}_{\mathbf{ROH}}$ and $\mathbf{G}_{\mathbf{LA}}$ were similar and for the youngest animals, they were highest with $\mathbf{G}_{\mathbf{LA}}$. The accuracies obtained with the LD matrix calculated with or without standardizing marker genotypes were similar.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TL planned the study and drafted the manuscript. XY performed the simulation and contributed to the manuscript writing. MD, AB prepared the genotyping and phenotyping data. THEM planned and coordinated the whole study, and contributed to writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 222664 ("Quantomics"). This article reflects only the author's views and the European Community is not liable for any use that may be made of the information contained herein. We gratefully thank Atillio Rossoni from the Italian Brown Swiss Association for kindly providing the data.

Author details

[1]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås N-1432, Norway. [2]Dipartimento di Scienze e Tecnologie Veterinarie per la Sicurezza Alimentare, Università degli Studi di Milano, Via Celoria 10, 20133 Milano, Italy.

Received: 6 August 2013 Accepted: 11 September 2014
Published online: 04 October 2014

References

1. Hayes B, Goddard M: **Genome-wide association and genomic selection in animal breeding.** *Genome* 2010, **53**:876–883.
2. Jannink JL, Lorenz AJ, Iwata H: **Genomic selection in plant breeding: from theory to practice.** *Brief Funct Genomics* 2010, **9**:166–177.
3. Meuwissen THE, Hayes BJ, Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps.** *Genetics* 2001, **157**:1819–1829.
4. Goddard ME, Hayes BJ, Meuwissen THE: **Using the genomic relationship matrix to predict the accuracy of genomic selection.** *J Anim Breed Genet* 2011, **128**:409–421.
5. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91**:4414–4423.
6. Yang JA, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42**:565–569.
7. Habier D, Fernando RL, Dekkers JCM: **The impact of genetic relationship information on genome-assisted breeding values.** *Genetics* 2007, **177**:2389–2397.
8. Luan T, Woolliams JA, Odegard J, Dolezal M, Roman-Ponce SI, Bagnato A, Meuwissen THE: **The importance of identity-by-state information for the accuracy of genomic selection.** *Genet Sel Evol* 2012, **44**:28.
9. Fernando RL, Grossman M: **Marker assisted selection using best linear unbiased prediction.** *Genet Sel Evol* 1989, **21**:467–477.
10. MacLeod IM, Meuwissen THE, Hayes BJ, Goddard ME: **A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors.** *Genet Res (Camb)* 2009, **91**:413–426.
11. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME: **Novel multilocus measure of linkage disequilibrium to estimate past effective population size.** *Genome Res* 2003, **13**:635–643.
12. Sved JA: **Linkage disequilibrium and homozygosity of chromosome segments in finite population.** *Theor Popul Biol* 1971, **2**:125–141.
13. Hill WG, Weir BS: **Maximum-likelihood estimation of gene location by linkage disequilibrium.** *Am J Hum Genet* 1994, **54**:705–714.
14. Meuwissen THE, Goddard ME: **Prediction of identity by descent probabilities from marker-haplotypes.** *Genet Sel Evol* 2001, **33**:605–634.
15. Meuwissen THE, Goddard ME: **The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data.** *Genetics* 2010, **185**:1441–1449.
16. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition.** *Proc IEEE* 1989, **77**:257–286.
17. Gilmour AR, Gogel BJ, Cullis BR, Thompson R: *ASReml User Guide Release 2.0.* 2006.
18. Sonesson AK, Woolliams JA, Meuwissen THE: **Genomic selection requires genomic control of inbreeding.** *Genet Sel Evol* 2012, **44**:27.

doi:10.1186/s12711-014-0064-6
Cite this article as: Luan *et al.*: Genomic prediction based on runs of homozygosity. *Genetics Selection Evolution* 2014 **46**:64.