Optimal design of low density marker panels for genotype imputation

H. Aliloo[1], R. Mrode[2,3], A.M. Okeyo[2], J. Ojango[2], T. Dessie[4], J.E.O. Rege[5], M.E. Goddard[6] & J.P. Gibson[1]

[1] University of New England, Armidale, NSW 2350, Australia
haliloo@une.edu.au (Corresponding Author)
[2] International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya
[3] Scotland's Rural College, Easter Bush, Midlothian EH25 9RG, Scotland, UK
[4]International Livestock Research Institute (ILRI), PO Box 5689, Addis Ababa, Ethiopia
[5]PICOTeam East Africa, Nairobi, Kenya
[6]The University of Melbourne, Parkville, VIC 3010, Australia

Summary

Cost-effective genotyping of livestock species can be done through a process which involves genotyping part of the population using a high density (HD) panel and the remainder with a lower density panel and then use imputation to infer the missing genotypes that are not included on the low density panel. Therefore, it is desirable to have a method of selecting markers for an assay that maximises imputation accuracy. Here we present a marker selection method that relies on the pairwise (co)variances between single nucleotide polymorphisms (SNPs) and the minor allele frequency (MAF) of SNPs. The performance of the developed method was tested in a 5 fold cross-validation process using genotypes of crossbred dairy cattle in East Africa, a population in which it is unclear whether existing low density SNP assays designed for purebred populations will maintain high imputation accuracies. Various densities of SNPs were selected using the (co)variance method and alternative SNP selection methods and then imputed up to the HD panel. The (co)variance method provided the highest imputation accuracies at all marker densities, with accuracies being up to 19% higher than the random selection of SNPs. The presented method is straightforward in its application and can ensure high accuracies in genotype imputation of crossbred dairy population in East Africa.

Keywords: imputation, low density marker panel design, accuracy of genotype imputation

Introduction

Genetic evaluation using DNA-based information, i.e. genomic selection, is now a standard tool in genetic improvement of many livestock species. Genomic selection usually uses evenly spaced SNPs spread across the genome to estimate breeding values (GEBV) for the target individuals (Meuwissen *et al.,* 2016). The accuracy of GEBVs can increase with more genotyped animals and higher density SNP panels. Although the cost of genotyping has decreased substantially since the technology emerged, HD SNP panels are still very costly for genetic improvement of livestock species. A cost-effective alternative is to genotype animals with cheaper low density panels and then to infer the missing genotypes that have not been directly assayed, based on information from a reference population genotyped by an HD panel; a method called genotype imputation.

The design of low density SNP panels to date has been mostly based on the use of evenly spaced markers and maximization of minor allele frequencies (MAF) with some

enrichments at chromosomal ends (*e.g.* Boichard *et al.,* 2012; Bolormaa *et al.,* 2015). Recently, interest has increased in developing methods for the optimal design of low density SNP panels that can be used for accurate genotype imputations (*e.g.* Wu *et al.,* 2016). For example Wu *et al.* (2016) described a multiple objective optimization algorithm to design lower density SNP chips that achieved substantially higher imputation accuracies than selecting SNPs solely based on uniform distribution of map information.

Here, we developed a method based on (co)variances between SNPs and weighted by MAFs to select subsets of SNPs that can be used for accurate imputation of genotypes to higher densities. We tested the method in different scenarios of imputation in crossbred dairy cattle populations of East Africa. Given that almost all existing SNP assays have been designed specifically for use in purebred populations, it was of additional interest to investigate what accuracy of imputation can be achieved in East African crossbred dairy cattle populations that are complex admixtures of dairy *Bos taurus* and indigenous African breeds.

Materials and methods

Genotypes

Genotypes were obtained from 3,083 crossbred animals sampled in four East African countries, Kenya, Uganda, Ethiopia and Tanzania, by the Dairy Genetics East Africa (DGEA) project for 777,962 SNPs using Illumina BovineHD BeadChip (Illumina, San Diego, CA, USA). Quality controls applied on raw data were: Only SNPs with GC score > 0.6 and call rate > 95 % were kept; mitochondrial, unmapped, duplicate map position and SNP located on sex chromosomes (X and Y) were removed. Further, SNPs with a MAF less than 0.01 were excluded. These resulted to 691,230 SNP genotypes over 29 autosomal chromosomes which were coded as 0, 1, and 2 respectively for AA, AB and BB allele combinations.

SNP selection

In order to design lower density SNP panels that can be efficiently used in genotype imputation to higher densities, a method of selecting SNPs based on the pairwise SNP (co)variance and weighted by MAF was developed. Consider *n* SNPs from which we want to select *k* SNPs such that the selected *k* SNPs together explain a higher proportion of the variance of the *n* SNPs than any other set of *k* SNPs. To start the SNP selection process, SNP genotypes are scaled so that the mean and variance of genotype at each SNP are 0 and 1, respectively. Then the covariance between all pairs of scaled SNP genotypes are calculated and stored in a matrix (**V**), which is an $n \times n$ (co)variance matrix and Vij is the covariance between SNP i and SNP j. The diagonal elements of matrix **V** are the variances of SNPs, and initially are all equal to 1. The sum of the diagonal elements or the trace of **V** matrix is defined as the total variance of n SNPs which is equal to the total number of SNPs. The SNP selection method is a sequential process where: 1) For each SNP a parameter is calculated which measures the strength of its correlation with all other SNP, and this is then summed across all SNPs and weighted by the MAF of the SNP:

$$i = 1,\ldots,n;\ \ j = 1,\ldots,n \text{ and } i \neq j, \quad (1)$$

$$E_{ij} = V_{ii} - \frac{V_{ij}^2}{V_{jj}}, \quad (2)$$

$$D_j = \sum_{i=1}^{n} E_{ij},$$ and (3)

$$D_{j_{adj}} = D_j \times (1 - MAF_j)$$ (4)

where E_{ij} is the unexplained variance (FUV) for SNP *i* after accounting for SNP *j* and D_j is the sum of FUVs across all SNPs for SNP *j* which is then weighted by MAF of SNP *j* in $D_{j_{adj}}$.

The SNP with the lowest D_{adj}, say SNP *k*, is selected because it has highest average covariance with all the other SNPs, so it explains more variance than any other SNP and it is also highly informative because of being highly polymorphic. 2) Then the pairwise (co)variances between the remaining SNPs are corrected by removing the amount of (co)variance explained by covariance of each SNP with the selected SNP, *k*:

$i = 1,\ldots n-1$ and $j = 1,\ldots n-1$, (5)

$$V_{ij_{adj}} = V_{ij} - \frac{V_{ik} \times V_{jk}}{V_{kk}}$$ (6)

3) At this stage, it is determined whether the selected SNPs have explained enough variance and the SNP selection process should be stopped or if more SNP are required. The proportion of variance explained by the selected SNPs at time *t* ($\omega^2_{exp_t}$) is calculated as:

$$\omega^2_{exp_t} = \frac{(\sigma_0^2 - \sigma_t^2)}{\sigma_0^2}$$ (7)

where σ_0^2 is the total variance with no SNP selected and $\sigma_t^2 = tr(\mathbf{V}_{adj_t})$ which is the trace of $\mathbf{V}_{adj}$ after selecting t SNPs.

We used a sliding window approach in which SNPs were selected within overlapping intervals of 1 Mbp. The interval moved forward by 500 Kbp until the end of the chromosome was reached. The number of SNPs selected from each window was determined based on the proportion of variance that was required to be explained by the selected SNPs. Different thresholds for the proportion of explained variance ($\omega^2_{exp_t}$) were set to achieve different densities of selected SNP panels. To account for the edge effect, twice the number of SNPs required for explaining variance were selected from the first and last 1 Mbp interval in each chromosome. We also selected equal number of SNPs to that selected by the (co)variance method (COV) within each interval either based on highest MAF (MAFI) or randomly (RANI). Further, SNPs were also selected randomly (RANC) or based on highest MAF (MAFC) across the whole chromosome without accounting for their map position on the chromosome, according to the total number of SNPs selected by the (co)variance method at each density.

Imputation

To assess the efficacy of the above methods for selecting SNPs, the selected SNP panels were used in turn for imputation to HD genotypes. To implement the SNP selection and validation procedures in independent populations, a cross-validation approach was implemented for the (co)variance and MAF methods. Animals were randomly divided into 5 groups such that the number of animals in each group was as similar as possible (~ 617 animals in each fold).

Then at each rotation, 4 folds were used to select SNPs and 1 fold was used in imputation. The random selections of SNPs within interval or across chromosome were also repeated 5 times to minimize the sampling error. In each fold of the cross-validation, only selected SNP genotypes were retained for the validation animals and their remaining genotypes in the HD panel were masked. The masked genotypes were then imputed using Minimac V3 (Das *et al.,* 2016). The accuracy of imputation was measured by correlation between real and imputed genotypes. The imputation accuracies obtained from SNPs selected by the five selection methods were averaged across the 5 folds and reported.

Results and Discussion

The number of SNPs selected at each threshold of the explained variance and correlations between the real and imputed genotypes obtained from different SNP selection methods are shown in Table 1. Selection of SNPs based on the (co)variance method always achieved the highest imputation accuracy at all thresholds such that it provided up to 3.2, 18.6, 16.8 and 15.5 percentage points higher correlations compared to SNP selections based on MAFI, RANI, RANC and MAFC, respectively. The difference between the accuracy of imputation from the (co)variance method and those of other SNP selection methods was highest at lower marker densities and there was little difference in accuracy of imputation between methods at high marker densities. Selection of SNPs based on highest MAF provided the second highest correlations at lower densities after the (co)variance method. MAFI was inferior to the COV method because it doesn't account for the linkage disequilibrium (LD) between SNPs and hence can select subsets of SNPs that have high LD with each other and have lower information content. This problem is worse when SNPs are selected based on highest MAF across the chromosome (MAFC) because MAFC is not optimized for uniformity across the chromosome and hence can leave gaps across genome with little information for imputation. Random selection of SNPs within intervals (RANI) or across chromosomes (RANC) provided very similar accuracies to each other at all densities. This suggests that even at the lowest densities used here, uniformity of marker spacing is not a particularly important factor for accuracy of imputation if SNPs are selected at random.

The results of the current study confirm that the genotype imputation in crossbred dairy cattle from East Africa can be done with a relatively high accuracy. The imputation accuracies reported in Table 1, however, are somewhat lower than accuracies reported in the literature for purebred dairy populations but are still within the same range of those from populations with similar genetic diversities (*e.g.* Hoze *et al.,* 2013). High effective population size (N_e) and low levels of long-distance LD across the genome lead to lower imputation accuracy. Crossbred populations resulting from many generations of admixture are expected to have larger N_e and weaker long-distance LD compared to purebred populations (*e.g.* Lu *et al.,* 2012). Hoze *et al.* (2013) reported lower imputation accuracies in beef breeds compared to dairy breeds where the former group in general showed higher rate of decay of LD across their genome. Bolormaa *et el.,* (2015) also reported lower imputation accuracies for a crossbred sheep population than those obtained for purebred sheep breeds. The presented method is straightforward in its application and can ensure higher accuracies in genotype imputation of crossbred dairy population in East Africa compared to other SNP selection methods.

Table 1. Average number of selected SNPs and correlations between real and imputed genotypes obtained from different SNP selection methods at different thresholds for the explained variance.

		SNP selection method[1]				
Threshold (%)	No. SNPs	COV	MAFI	RANI	RANC	MAFC
1	3,757	0.6379	0.6077	0.4611	0.4680	0.4886
5	4,013	0.6778	0.6458	0.4920	0.5094	0.5225
10	6,166	0.7834	0.7553	0.7166	0.7033	0.7005
15	8,738	0.8292	0.8016	0.7871	0.7803	0.7701
20	11,773	0.8599	0.8325	0.8281	0.8245	0.8101
25	15,373	0.8832	0.8561	0.8578	0.8552	0.8385
30	19,812	0.9017	0.8756	0.8816	0.8802	0.8626
35	25,410	0.9170	0.8928	0.9021	0.9015	0.8830
40	32,573	0.9299	0.9088	0.9203	0.9199	0.9010
45	41,383	0.9405	0.9226	0.9355	0.9351	0.9161
50	52,134	0.9495	0.9344	0.9481	0.9478	0.9291

[1] Selection of SNPs based on COV: (co)variance method; MAFI: minor allele frequency within interval; RANI: random within interval; RANC: random across chromosome and MAFC: minor allele frequency across chromosome.

List of References

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K.J. Gietzen, B.J. Hayes, C.T. Lawley, T.S. Sonstegard, C.P. Van Tassell, P.M. VanRaden, K.A. Viaud-Martinez, G.R. Wiggans, and L.D.C. for the Bovine. 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLOS ONE 7(3):e34130.

Bolormaa, S., K. Gore, J.H.J. van der Werf, B.J. Hayes, and H.D. Daetwyler. 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Animal Genetics 46(5):544-556.

Das, S., L. Forer, S. Schonherr, C. Sidore, A.E. Locke, A. Kwong, S.I. Vrieze, E.Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W.G. Iacono, A. Swaroop, L.J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G.R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat Genet 48(10):1284-1287.

Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution 45(1):33.

Lu, D., M. Sargolzaei, M. Kelly, C. Li, G. Vander Voort, Z. Wang, G. Plastow, S. Moore, and S. P. Miller. 2012. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Frontiers in Genetics 3:152.

Meuwissen, T., B.J. Hayes & M.E. Goddard, 2016. Genomic selection: A paradigm shift in animal breeding. Animal Frontiers 6(1):6-14.

Wu, X.-L., J. Xu, G. Feng, G.R. Wiggans, J.F. Taylor, J. He, C. Qian, J. Qiu, B. Simpson, J. Walker, and S. Bauck. 2016. Optimal Design of Low-Density SNP Arrays for Genomic Prediction: Algorithm and Applications. PLOS ONE 11(9):e0161719.