The identification of key contributors increases the imputation accuracy of target populations

M. Neuditschko[1,2], M.S. Khatkar[2] & H.W. Raadsma[2]

[1] Agroscope – Swiss National Stud Farm, Les-Longs Prés, 1580 Avenches, Switzerland
markus.neuditschko@agroscope.admin.ch (Corresponding Author)
[2] University of Sydney, Sydney School of Veterinary Science, 425 Werombi Road, Camden, NSW 2570, Australia

Summary

Currently different methods are used to select informative individuals for re-sequencing and genotype imputation. In this study we compared the utility of the recently described identification of key contributors (KCO) method with two commonly applied strategies, namely the identification of pedigree-based marginal gene contributions (PED) and the optimization of genetic relatedness (REL) and against animals selected at random (RAN). Based upon a simulated population structure (5,100 individuals and 10,000 SNPs) we show that, KCO provided the highest phasing (lowest switch error rates) and imputation accuracies (0.5% and 91.5%), followed by PED (2.6% and 88.1%), RAN (1.6% and 87.5%) and REL (5.4% and 87.0%) when including a maximum number of 100 individuals in the reference population. Furthermore, it was demonstrated that with the selection of key contributors especially the imputation accuracy (correlation between true and imputed genotype) of rare variants (minor allelic frequency <0.1) can be significantly increased by more than 10%. Therefore, we suggest to include the individual genetic contribution score in the decision criteria when selecting individuals for re-sequencing and genotype imputation.

Keywords: re-sequencing, genotype imputation, phasing accuracy, key contributors

Introduction

Presently, global efforts are focused on re-sequencing individuals within species and breed groups to improve our knowledge on the genetic architecture of populations (Deatwyler *et al.,* 2014). A typical approach in such scenarios is to re-sequence informative individuals within populations, and to impute genotypes at whole-genome sequence level of additional animals genotyped with high density SNP panels (Frischknecht *et al.,* 2014). The prevailing methods for the selection of reference individuals for genotype imputation solely focus on the identification of key ancestors through pedigree or genomic relationship information to maximize genetic diversity. Typically such strategies do not involve genotype information of the most influential and connected progeny, which may lead to a loss of phasing accuracy of the reference population and has posed problems in genotype imputation (van Binsbergen *et al.,* 2014).

We have recently shown that informative individuals for re-sequencing and genotype imputation can be selected based on the Eigenvalue Decomposition (EVD) of a genomic relationship matrix (Neuditschko *et al.,* 2017). EVD like Principal Component Analysis

(PCA) is a multivariate technique that provides an optimal subspace to investigate population structures by maximizing variation on the highest components. Based upon this mathematical principle we identified so called key contributors that capture most of the variation in the relevant genetic relationship structures. We have already demonstrated that the selection of key contributors increases phasing accuracy of the reference population compared to methods currently applied. Here, we further investigate the selection of key contributors on imputation accuracy including different sample sizes and five minor allelic frequency (MAF) classes in the analyses.

Material and methods

Simulated data

The simulated data consisted of a total of 5,100 individuals and 10,000 SNPs as described by Usai *et al.* 2014 at http://qtl-mas-2012.kassiopeagroup.com/en/dataset.php. The simulation starts with a base population (F0) of 1,020 unrelated individuals (20 males and 1,000 females). The first generation (F1) was generated by randomly mating each of the 20 founder males with 50 females. Each of the next three generations (F2-F4) also consisted of 20 males and 1,000 females and was generated following the same principle. The simulated genome consisted of five chromosomes each spanning 100 Mb with 2,000 equally distributed SNPs. To select subsets of informative individuals under REL and KCO (see description below) we computed an identity by descent (IBD) genomic relationship matrix (G) using Germline Gusev *et al.* 2009, whilst PED was applied on the simulated pedigree structure.

Selecting informative individuals for genotype imputation

Subsets of informative individuals were selected according to four different strategies including the identification of key contributors (KCO), two commonly applied methods previously outlined by Boichard (2002) (PED) and Goddard and Hayes (2008) (REL) and individuals selected at random (RAN). In order to evaluate the performance of the different subsets of informative individuals for genotype imputation 20% of the SNPs (2,000 SNPs) were randomly set to missing across the genome in the target populations, whilst the missing SNPs are equally distributed over five MAF classes (MAF 0>0.1, MAF 0.1>0.2, MAF 0.2>0.3, MAF 0.3>0.4 and MAF 0.4>0.5). After selecting subsets of informative individuals, phasing of the selected reference populations was performed using the program FImpute (Sargolzaei *et al.,* 2014) including the pedigree information of the selected individuals. The phasing accuracy between the inferred haplotype phase and true haplotype phase was examined using the switch-error-rate metric (Browning & Browning, 2002). Finally the phased reference populations were used to impute the missing SNPs in the remaining population. The accuracy of imputation was assessed by calculating the genotype concordance between true and imputed genotypes ($\mathbf{gc}_{TI}$). To assess imputation accuracy across the five selected MAF classes we additionally calculated the correlation between true and imputed genotype ($\mathbf{r}_{TI}$). Phasing and imputation accuracy was evaluated for five different scenarios by subsequently increasing the number of informative individuals from 20 to 100 in increments of 20. Imputation of the target populations was also performed with FImpute (Sargolzaei *et al.,* 2014), as described above.

Results and Discussion

Phasing accuracy increased as the number of informative individuals increased within all selected reference populations, including when subsets of individuals were selected at random (Figure 1A). Selected reference populations under KCO consistently resulted in the highest phasing accuracy (lowest switch error rates) of all selected reference populations, whilst PED and REL performed worse than RAN individuals.

Similarly increasing the number of individuals included in the reference populations increased imputation accuracy under all four methods. However, the rate of improvement in imputation accuracy was slightly different across the applied methods (Figure 1B). Imputation accuracy was highest when reference populations were selected under KCO strategy, followed by PED, whilst selecting individuals under REL performed no better than selecting individuals at RAN, except including a subset of 20 individuals in the reference population. Imputation accuracy reached a plateau at 87% under PED when the highest number of individuals were selected resulting in only small differences between the imputation accuracies of PED, RAN and REL, whilst KCO performed significantly better than the other applied strategies (91.5%).

Table 1 illustrates the imputation accuracy according to each of the five selected MAF classes including a maximum number of 100 individuals in the reference population. Once again the REL strategy was the least efficient strategy to impute the missing genotypes across all MAF bands, except to the MAF 0.4<0.5 class. Again, KCO strategy resulted in the highest imputation accuracy across all MAF classes.

Figure 1. Phasing and imputation accuracy of selected subsets according to the four different applied strategies (KCO, PED, RAN and REL).

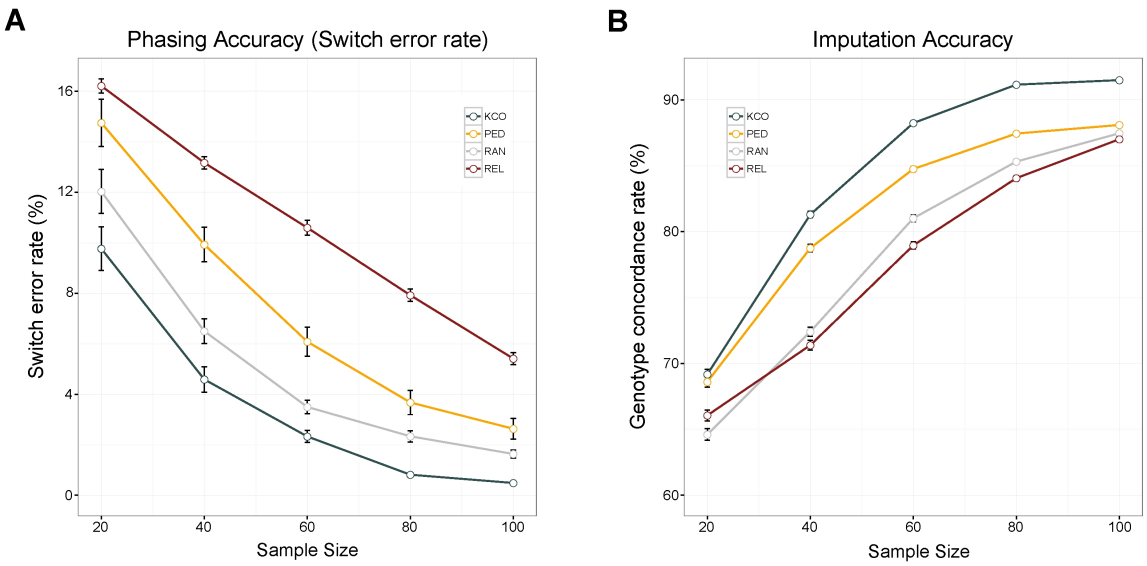

Table 1. Imputation accuracy ($\mathbf{gc}_{TI}$ and $\mathbf{r}_{TI}$) of the five selected minor allelic frequency classes (MAF) including a maximum number of 100 individuals in the reference population.

MAF classes	**KCO**		**PED**		**REL**		**RAN**	
	$\mathbf{gc}_{TI}$	$\mathbf{r}_{TI}$	$\mathbf{gc}_{TI}$	$\mathbf{r}_{TI}$	$\mathbf{gc}_{TI}$	$\mathbf{r}_{TI}$	$\mathbf{gc}_{TI}$	$\mathbf{r}_{TI}$
MAF 0>0.1	96.97	84.19	95.66	73.54	95.10	66.04	95.45	72.16

MAF 0.1>0.2	92.92	85.40	89.97	77.56	88.62	73.77	89.78	77.44
MAF 0.2>0.3	90.38	86.84	86.21	79.74	85.11	77.71	85.61	78.81
MAF 0.3>0.4	89.17	87.87	84.79	81.93	83.67	80.28	83.97	80.81
MAF 0.4>0.5	88.02	87.70	83.83	82.46	82.57	80.85	82.50	80.68

Conclusion

In this study we demonstrated that the identification of key contributors (KCO) also increases imputation accuracy of target populations, besides phasing accuracy of selected reference populations (Neuditschko *et al.,* 2017), compared to other commonly used methods (PED and REL) and that it becomes feasible to significantly increase the imputation accuracy of rare variants with MAF >0.1. For general use and application we have assembled an online open access platform for the identification of key contributors within complex populations (https://github.com/esteinig/netview).

List of References

Deatwyler H.D., Capitan A., Pausch H., Stothard P., van Binsbergen R., Brondum R.F., *et al.* (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics. 46(8): 858-865

Frischknecht M., Neuditschko M., Jagannathan V., Drogemuller C., Tetens J., Thaller G., *et al.* (2014). Imputation of sequence level genotypes in the Franches-Montagnes horse breed. Genetics Selection Evolution. 46(1): 63

van Binsbergen R., Bink M., Calus M., van Eeuwijk F., Hayes B., Hulsegge I., *et al.* (2014). Accuracy of imputation to whole-genome sequence data in Holstein Frisian cattle. Genetics Selection Evolution. 46(1): 41

Neuditschko M., Raadsma H.W., Khatkar M.S., Jonas E., Steinig E.J., Flury C., *et al.* (2017). Identification of key contributors in complex population structures. PLOS ONE. 12(5): e0177638

Usai M.G., Gaspa G., Macciotta N.P., Carta A. & Casu S. (2014) XIVIth QTLMAS: simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation. BMC Proceedings. 8(5): 1-9.

Gusev A., Lowe J.K., Stoffel M., Daly M.J., Altshuler D., Breslow J.L., *et al*. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Research. 19(2): 318-326

Boichard D. Pedig: a fortran package for pedigree analysis suited for large population. (200). Proceedings of the 7th World Congress on Genetics Applied to Livestock Production.

Goddard M. & Hayes B. Genomic selection based on dense genotypes inferred from sparse genotypes. (2009). Association for the Advancement of Animal Breeding and Genetics. 18: 26-29

Browning S.R. & Browning B.L. (2011). Haplotype phasing: existing methods and new developments. Nature Review Genetics. 12

Sragolzaei M., Chesnais J., & Schenkel F. (2014). A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15(1): 478