Imputation of Missing Genotypes From Sparse to High Density Using Long-Range Phasing

H.D. Daetwyler[*†‡], G.R. Wiggans[§], B.J. Hayes[*], J.A. Woolliams[†] and M.E. Goddard[*,**]

Introduction

Population characteristics such as geographical proximity can result in a high probability that individuals within a given population share a common ancestor not many generations ago. Similarly, in commercial animal populations selective breeding has reduced effective population sizes by limiting the number of parents, again causing individuals to share one or more common ancestors in the last few generations. If individuals share a common ancestor *n* generations ago, they are likely to have a shared chromosome segments of average length 1/*n* Morgans. With dense genotyping of markers, these segments will contain many markers and so it should be possible to recognise them and distinguish them from short segments that are identical-by-state but do not trace to the common ancestor, without complex likelihood calculations. These observations lead to new approaches to phasing haplotypes which are based on the premise that if a large section of two gametes is identical-by-state then there is a high probability that this section originated in a common ancestor (Kong et al. 2008).

Kong et al. (2008) called their method long-range phasing but the principle can also be used to impute and phase missing genotypes or even to impute genotypes on individuals that have not been genotyped at all. One particularly useful application is to impute dense genotypes on individuals with sparse genotypes using dense genotype information on their relatives. In the extreme, full genome sequences could be imputed for individuals which have been genotyped at moderate density, provided they had enough relatives that had been fully sequenced (Goddard 2008).

Here we describe a computationally efficient long-range phasing algorithm (ChromPhase) that can phase whole chromosomes and simultaneously impute a large number of markers. We test our method by imputing markers in sparsely genotyped individuals with many missing genotypes. Futhermore, we investigate the reduction in accuracy of genomic evaluation with imperfectly imputed and sparse data.

Material and methods

Our approach is similar to that of Kong et al. (2008), but we also use the pedigree to identify whether a relative is likely to share a part of an individual's paternal or maternal chromosome. We will describe all processes for the paternal side of the pedigree but the maternal side is treated in the same manner. The algorithm consists of three stages.

[*] Biosciences Research Division, Department of Primary Industries, 1 Park Dr., Bundoora 3083, Australia;
[†] The Roslin Institute and R(D)SVS, The University of Edinburgh, Roslin, EH25 9PS, United Kingdom;
[‡] Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, The Netherlands;
[§] Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA;
[**] Faculty of Land and Food Resources, University of Melbourne, Parkville 3010, Australia

Stage 1, Information Sources: Pedigree and genotype data is read and molecular genotyping errors are checked at each locus. For each individual, considered in turn as the proband, three sets of relatives are defined. The first set consists of all offspring of the proband. The second set, called surrogate fathers, consists of individuals related to the proband through his or her father.

Stage 2, Single locus, rule-based allele assignment: ChromoPhase applies rule-based allele assignment to the paternal or maternal gamete if they can be unambiguously resolved based on an individual's own known genotype, parental alleles (e.g. Pong-Wong et al. 2001, Baruch et al. 2006).

Stage 3, Comparison of relatives: An iterative process follows in which each individual is considered as the proband once per iteration starting at the top of the pedigree. Single locus, rule-based filling of alleles (stage 2) continues at the start of each iteration as more information becomes available. ChromoPhase then compares each proband to each of its relatives in the three sets (i.e. offspring, surrogate fathes and mothers) one locus at a time to identify shared chromosome segments consisting of a consecutive string of matching loci. When a matching segment is identified any missing alleles in proband are filled in with allele information from the relative if available. We will describe the different types of matches at a locus by using an example where the proband is compared to a surrogate father. Consider a proband whose paternal allele is compared to both alleles of a surrogate father at one locus. This comparison yields one of four outcomes: a conclusive match occurs when the paternal allele is not missing and it is equal to one or both of the surrogate fathers alleles at that locus. A distinguishing match is defined as a match between the paternal allele and one surrogate father allele but not both. Thus, a distinguishing match is also a conclusive match, but in a distinguishing match the source of the paternal allele can be clearly determined and it is used to define the start and end of a shared segment to reduce errors. Missing information counts as an inconclusive match which is not allowed to end a shared segment. A definite non-match occurs when the paternal allele of the proband is not equal to the surrogate father's allele found on the chromosome which matched at the last distinguishing locus. A minimum length of 50 consecutive matching (i.e. conclusive, inconclusive, and distinguishing) loci between two distinguishing matches was required to accept a shared segment. Within that run, the number of loci with conclusive matches needed to exceed 40. The minimum number of conclusive matches guards against too many missing loci being counted as matches within a considered segment. Requiring longer segments will reduce errors but it will also result in fewer phased or imputed loci.

Populations and Genome: Populations in mutation drift equilibrium were simulated by randomly mating individuals for 1000 generations with recombination and mutation. Effective population size (N_e) was 200 and the number of male and female parents was equal across generations. One male and one female offspring were produced per mating. Pedigree and genotype information was retained for individuals in the last four generations. In generation 997 through 999, 100 individuals were simulated and generation 1000 consisted of 200 individuals for a total of 500 individuals.

One chromosome was simulated measuring one Morgan. In generation zero all individuals were completely homozygous for the same allele at all 40,000 potential loci. Mutations were then applied at a rate of 2.5×10^{-5} per locus per meiosis in the following generations. The number of mutations and recombinations per chromosome were sampled from a Poisson distribution. Approximately 1500 segregating bi-allelic loci were present at generation 1000, which is equivalent to a density of $7.5N_e$ per Morgan. Loci were selected to exceed 0.02 minor allele frequency.

Testing Imputation: The utility of ChromoPhase for imputation of missing genotypes was evaluated in 25 replicates of simulated data described above. Three different depths of data were tested (2, 3 and 4 generations) and, within each, three different sparse marker densities were investigated, 13, 34 and 100 markers per chromosome. These three densities correspond to whole genome densities of 400, 1000 and 3000 markers in a 30 Morgan genome. The markers chosen to be in the sparse set were selected based on higher than average minor allele frequency and were evenly distributed across the chromosome. In each imputation scenario, all animals in the last generation were set to missing for genotypes not chosen as part of the sparse set. This resulted in nine scenarios (i.e. three sparse densities and three pedigree depths). In each scenario, impuation was evaluated by checking imputed genotypes against true genotypes from simulation.

Testing Genomic Evaluation Accuracy: One hundred quantitative trait loci per Morgan were randomly sampled from the segregating loci. Additive substitution effects were sampled from N(0,1). Phenotypes were generated by adding a random environmental deviation to genotypic values, where this environmental deviate was scaled to achieve a heritability of 0.3. In the imputed dataset resulting from three generations, realised relationship matrices were calculated following the same procedure as Nejati-Javaremi et al. (1997). Only the last two generations were used (300 individuals). This was done for three scenarios at each sparsity: i) all individuals were genotyped at high density (All Dense), ii) all individuials were genotyped at sparse density (All Sparse), and iii) individuals in last generation had imputed genotypes (Imputed). These realised relationship matrices were fitted to phenotypes in ASReml. Accuracy was computed as the correlation of true and genomic breeding values.

Results and discussion

In general, a high proportion of missing genotypes were imputed using ChromPhase (table 1). However, the proportion of correctly imputed genotypes decreased as markers became sparser. A decreasing trend was also seen when fewer generations of data were available.
While the efficiency of imputation shows a clear decreasing trend as sparse density decreased, this trend is much less pronounced when genomic evaluation accuracy is considered (table 2). The reduction in accuracy from dense to imputed genotypes is small when the number of sparse genotypes were 34 and 100, but is more pronounced when only 14 sparse markers were available. The accuracy in the All Sparse scenarios is likely due to tracking of relationship information by the markers and increasing the sparse density does not seem to improve the accuracy of All Sparse scenarios at this low density.

Table 1: Imputation performance of missing genotypes with sparse marker densities of 14, 34 and 100 markers per Morgan, when imputed genotypes were compared to true genotypes from simulation (SE < 0.004 in all scenarios).

Gen.	Sparse Density	Prop. Correct	Prop. Wrong	Prop. Missing
2	14	0.615	0	0.385
3	14	0.863	0.037	0.102
4	14	0.899	0.040	0.062
2	34	0.618	0	0.382
3	34	0.908	0.020	0.071
4	34	0.944	0.022	0.034
2	100	0.628	0	0.372
3	100	0.941	0.008	0.052
4	100	0.976	0.009	0.016

Table 2: Accuracy using different marker densities per Morgan and three generations of data for imputation (SE < 0.013 in all scenarios).

Density	Scenario	Accuracy	% of All Dense
1500	All Dense	0.755	100.0
14	All Sparse	0.561	74.3
34	All Sparse	0.561	74.3
100	All Sparse	0.564	74.7
14	Imputed	0.684	90.6
34	Imputed	0.740	98.0
100	Imputed	0.743	98.4

Conclusion

The results show that imputation of missing genotypes from sparse to high density is feasible using our algorithm. In addition, while imputation is imperfect this does not cause a great reduction in genomic evaluation accuracy.

References

Baruch, E., Weller, J.I., Cohen-Zinder, M. et al. (2006) *Genetics* 172: 1757-1765.

Goddard, M.E. (2008) The use of high density genotyping in animal health, in Animal Genomics for Animal Health, Basel.

Kong, A., Masson, G., Frigge, M.L. et al. (2008) *Nature Genetics* 40: 1068-1075.

Nejati-Javaremi, A., Smith, C. and Gibson, J.P. (1997) *J. Anim. Sci.* 75:1738-1745.

Pong-Wong, R., George A.W. Woolliams, J.A. et al. (2001) *Genet. Sel. Evol*. 33: 453-471.