Imputation of Missing Genotypes from Sparse to High Density using Long-Range Phasing

Hans D. Daetwyler¹,²,³,*, George R. Wiggans⁴, Ben J. Hayes¹, John A. Woolliams² and Mike E. Goddard¹,⁵

*Corresponding Author: Hans D. Daetwyler, Biosciences Research Division, Department of Primary Industries, 1 Park Drive, Bundoora 3083, Victoria, Australia. Phone: +61 3 9032 7037. Fax: +61 3 9032 7158. hans.daetwyler@dpi.vic.gov.au

Key words: long-range phasing, imputation, haplotyping, sparse marker map

Summary:

Knowledge of parental origin of alleles (phase) is valuable in many genomic analyses. Individuals of finite populations are expected to share segments of the genome which trace to a common ancestor. The authors propose a long-range phasing algorithm for phasing and imputing missing genotypes which exploits this property. The method is tested in simulated and real Holstein cattle data by imputing genotypes from sparse to high density and a high proportion of genotypes were imputed correctly. Imputation of missing genotypes from sparse to high density may contribute to greater uptake of genomics in species where high density genotyping costs currently outweigh the perceived benefits.

¹Biosciences Research Division, Department of Primary Industries, 1 Park Dr., Bundoora 3083, Australia; ²The Roslin Institute and R(D)SVS, The University of Edinburgh, Easter Bush, Roslin, EH25 9RG, United Kingdom; ³Animal Breeding and Genomics Centre, Wageningen University, 6700 AH, Wageningen, The Netherlands; ⁴Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA; ⁵Faculty of Land and Environment, University of Melbourne, Parkville 3010, Australia

Copyright 2011.
ABSTRACT

Related individuals share potentially long chromosome segments which trace to a common ancestor. We describe a phasing algorithm (ChromoPhase) which utilises this characteristic of finite populations to phase large sections of a chromosome. In addition to phasing, our method imputes missing genotypes in individuals genotyped at lower marker density when more densely genotyped relatives are available. ChromoPhase uses a pedigree to collect an individual’s (the proband) surrogate parents and offspring and uses genotypic similarity to identify its genomic surrogates. The algorithm then cycles through the relatives and genomic surrogates one at a time to find shared chromosome segments. Once a segment has been identified, any missing information in the proband is filled in with information from the relative. We tested ChromoPhase in a simulated population consisting of 400 individuals at a marker density of 1500 per Morgan, which is approximately equivalent to a 50K Bovine single nucleotide polymorphism chip. In simulated data, 99.9% loci were correctly phased and, when imputing from 100 to 1500 markers, more than 87% of missing genotypes were correctly imputed. Performance increased when the number of generations available in the pedigree increased, but was reduced when the sparse genotype contained fewer loci. However, in simulated data, ChromoPhase correctly imputed at least 12% more genotypes than fastPHASE, depending on sparse marker density. We also tested the algorithm in a real Holstein cattle data set to impute 50K genotypes in animals with a sparse 3K genotype. In this data 92% of genotypes were correctly imputed in animals with a genotyped sire. We evaluated the accuracy of genomic predictions with the dense, sparse and imputed simulated datasets and show that the reduction in genomic evaluation accuracy is modest even with imperfectly...
imputed genotype data. Our results demonstrate that imputation of missing

genotypes, and potentially full genome sequence, using long-range phasing is feasible.

INTRODUCTION

Single nucleotide polymorphism (SNP) arrays from sparse to high density are now
available in many species. The genotypes resulting from high throughput methods are
un-phased and, therefore, the paternal or maternal source of each allele is unknown.
Knowledge of parental origin or haplotype information can be useful in the analysis of
complex traits, such as quantitative trait loci (QTL) detection (e.g. MEUWISSEN and
GODDARD 2000), genomic selection (e.g. MEUWISSEN et al. 2001; CALUS et al. 2008;
VILLUMSEN and JANSS 2009), and detection of imprinting (e.g. REIK and WALTER
2001; WOOD and OKEY 2006).

Many methods for resolving haplotypes have been proposed and they fall into two
broad categories: those that use known relationships between individuals to perform a
linkage analysis (e.g. ELSTON and STEWART 1971; LANDER and GREEN 1987; WEEKS
et al. 1995; WINDIG and MEUWISSEN 2004) and those that rely on linkage
disequilibrium among the SNP in a population without known relationships (e.g.
CLARK 1990; SCHEET and STEPHENS 2006; TIER 2006; BROWNING and BROWNING
2009).

More recently, another feature of finite population genomics together with the
availability of denser marker maps have fostered new phasing approaches, as
demonstrated by Kong et al. (2008). 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 shared chromosome segments of average length approximately
$1/n$ Morgans. Provided that $n$ is not too large and 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 (IBS) but
do not trace to the common ancestor, without complex likelihood calculations. This
leads to a phasing approach based on the key observation of Kong et al. (2008) that if
animals have non-conflicting homozygote genotypes over a long string of consecutive
loci, they have at least one long haplotype in common. The requirement of a long
string of loci leads to a high probability that the common long haplotype has
originated in a common ancestor.

Kong et al. (2008) called their method long-range phasing, but the principle of
comparing long stretches of chromosomes between individuals to identify common
segments 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. Then, for example, genomic
predictions for selection in livestock or crop species could be made on the imputed
genotypes, at the cost of genotyping the low density markers. 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 algorithm (ChromoPhase) that can phase whole chromosomes and simultaneously impute missing genotypes if a haplotype has been observed in the densely genotyped population. We use a similar approach to that of Kong et al. (2008) and its extension (Hickey et al. 2011), but whereas their focus is on genotypes, we use haplotypes more explicitly. We also use the pedigree to identify whether a relative is likely to share a part of an individual’s paternal or maternal chromosome in addition to animals which are genomically similar, which we call genomic surrogates. In addition, while Kong et al. (2008) requires operational extensions to impute missing genotype information, our algorithm already addresses the objective of long-range phasing to both phase and impute loci simultaneously.

METHODS

ChromoPhase has the objective of inferring paternal and maternal gametes for a set of individuals based upon a subset (possibly a complete subset) of all individuals with dense genotypes. It relies upon the same principle as Kong et al. (2008) in that it makes use of the potentially long chromosome segments which related animals share. These segments are particularly long when individuals are closely related, as during meiosis the expected number of crossovers is one per Morgan of chromosome. Therefore, with dense marker genotypes, the phase can be established by comparing an individual to close relatives. An edited pseudo code of the algorithm is provided in supplementary materials.
We assumed bi-allelic loci with a reference allele coded 0 and an alternative allele coded 2. Genotypes were coded 0, 1, and 2, corresponding to 00, 02, and 22 respectively. Missing alleles and genotypes are assigned ‘5’. At the start of ChromoPhase, all alleles on the paternal and maternal gametes for all individuals are set to 5. We assume that the individuals in the data can be divided into two groups, set ‘D’ containing all the individuals with dense genotyping, and set ‘S’ containing all the individuals that are sparsely genotyped. Our dense simulated data had 15 SNP per cM but, in general, dense can be defined as being sufficient for the risk of double crossovers between adjacent loci to be negligible. The complete algorithm has different stages and an overview of the progression through the stages is shown in Figure 1. In Stage 1, potential sources of shared chromosome segments are identified using pedigree and genotypic similarity. In Stage 2 alleles are assigned using 2 different approaches: ‘2A’ employs rule-based allele assignment based upon genotypes of parents, immediate offspring and mates; and ‘2B’ in which assignments are based upon phasing using unbroken strings of matching alleles on their respective chromosomes. Stage 2 is iterated (‘A’, ‘B’, ‘A’ … ‘B’) and initially only involves individuals in set ‘D’. Once the pre-defined maximum phasing iteration for set ‘D’ is reached, Stage 3 is carried out involving a further pre-defined number of iterations of 2A and 2B for all individuals in set ‘S’. The imputed paternal and maternal gametes of the densely-genotyped individuals (i.e. set ‘D’) remain unchanged by the imputation of individuals in ‘S’, and if only phasing of densely genotyped individuals is required then Stage 3 is unnecessary.

**Stage 1, information sources:** Each individual in set ‘D’ and ‘S’ is considered as a proband. Molecular genotyping errors are checked at each locus by identifying where
the proband genotype is inconsistent with the paternal (maternal) genotype, e.g. proband genotype = 0 and father’s genotype = 2. Inconsistent genotypes are set to missing in the proband if it is the only progeny conflicting with the parent. A parent’s genotype is set to missing if it conflicts with most offspring. Three sets of densely genotyped relatives are then defined for each proband by tracing through the full pedigree. Hence, completely ungenotyped individuals are used to connect genotyped individuals. The first set of relatives consists of all descendants of the proband and these are collected starting with the youngest individual. The second set, starting with the oldest individual is called surrogate fathers and consists of individuals related to the proband through his or her father. If the father is densely genotyped, then he is the only surrogate father, because more distant relatives do not add further information on the paternal side of the proband. In a proband with an ungenotyped father, the set of surrogate fathers of the proband consists of the father’s surrogate fathers and mothers as well as their offspring. Analogous rules are applied to define the surrogate mothers. Only relatives up to 3 degrees removed from the proband are used in the sets of surrogates.

Stage 2A, 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 upon an individual’s own known genotype, the parental alleles and offspring alleles (e.g. PONG-WONG et al. 2001; BARUCH et al. 2006). In all invocations the following rules are applied starting at the top of the pedigree (i.e. the oldest individual) working through to the bottom of the pedigree. If the proband genotype is homozygous, then both its paternal and maternal alleles are equal to the homozygous allele. If the father’s genotype is homozygous then the
proband paternal allele is equal to the father’s homozygous allele. If a proband genotype is missing and only the proband’s paternal allele is known, then if an offspring only has one allele known and it is opposite to the proband’s paternal allele, the proband’s maternal allele must be the same as the offspring’s maternal allele.

Stage 2B, phasing of densely genotyped individuals: At the start of each invocation of ‘2B’, genomic surrogates are identified for each proband, by comparing its genotypes with those for individuals in set ‘D’. A genomic surrogate must have non-conflicting homozygotes (for example, no instance of 0 genotype in proband and 2 genotype in the potential genomic surrogate) with the proband for at least 100 consecutive loci if applied to an individual in set ‘D’, but 400 consecutive loci if applied to an individual in set ‘S’. This is necessary because a proband in set ‘S’ may have a lot of missing information and the stretch of matches must include a sufficient number of sparsely genotyped loci to effectively identify genomic surrogates. Missing genotypes are not included in the 100 loci in set ‘D’, but are included in the 400 loci in set ‘S’. Genomic surrogates are chosen to limit the population based search for shared segments to individuals which ‘have something to offer’ to the proband, thereby reducing the total number of comparisons. In all invocations the following procedure is then applied starting with the oldest individual working through to the youngest. Each proband is compared to each of its relatives, contained within the sets of surrogate fathers and mothers as well as offspring, to allocate alleles by identifying shared segments for three iterations. In the last two phasing iterations, comparisons to relatives continue but now the proband is also compared to its genomic surrogates to identify shared chromosome segments and fill in missing alleles in the proband if the allele can be found in a shared segment. It is possible that
even densely genotyped individuals may have some missing genotypes and these are imputed in the course of 2B.

Stage 3: phasing and imputation of sparsely genotyped individuals: In Stage 3, the iterations of Stage 2 are now applied to individuals in ‘S’, where rule-based filling is applied (Stage 2A) and the proband is compared to its relatives and genomic surrogates for the remaining 8 iterations (Stage 2B).

We will now describe the comparisons of probands with relatives and genomic surrogates as well as the criteria used to accept segments as being shared in Stage 2B. These routines are the same for phasing or imputation iterations, except for some shared segment acceptance criteria, detailed below. Probands are separated into two main groups, those which have at least one genotyped parent (non-founders) and those that do not (founders). Founders may have surrogate fathers or mothers if connected to genotyped relatives through pedigree. In non-founders phasing and imputation is simpler because rule-based methods can be used (Stage 2A). In founders, there is a need to identify erroneous crossovers within shared segments with surrogates so that proband gametes can be flipped after the crossover, if necessary, to minimise the number of crossovers and, in turn, phase the proband.

Criteria for acceptance of shared chromosome segments: The longer the stretch of matching loci between two individuals, the higher is the probability that the segment traces to a common ancestor (KONG et al. 2008). We required a minimum length of 100 consecutive matching loci for all individuals to accept a segment as shared. In the last phasing and imputation iteration we relaxed this to 50 loci to allow
the filling in of more alleles. These criteria for defining the minimum length of a
shared segment can be adapted to suit a dataset and will depend on the number of
markers per Morgan. Note that the minimum segment length is different from the
length of non-opposing homozygotes required to select genomic surrogates. When
selecting genomic surrogates the main concern is identifying a subset of ‘D’ large
enough to provide enough power to phase or impute, but small enough to reduce the
number of redundant comparisons and, in turn, reduce the computational burden. In
contrast, here the focus is on accepting specific chromosome segments as shared and
we compare gametes instead of genotypes. Also note that while there is a theoretical
basis for the length of a genome segment required to be accepted as shared, the
remainder of the algorithm’s thresholds and parameters have been chosen
heuristically.

A segment starts when the alleles match, and ends when the alleles do not match
anymore, as demonstrated in Figure 2. Missing genotypes or alleles do not end a
segment although the proportion of missing alleles is tracked. When there is a
moderate amount of missing information, we may be reasonably confident that the
segment is shared if it is long, but it is difficult to define the ends. Hence, it seems
useful to still make use of these segments but penalise their length. Therefore, the
algorithm removes a short stretch of loci at the beginning and end of an identified
segment, termed ‘offset’ in Figure 2. Testing has shown that having an offset of 20
loci results in fewer errors while not impairing phasing or imputation performance
significantly. Furthermore, in both phasing and imputation iterations the offset is
multiplied by four (chosen through empirical testing) if the proportion of missing
information exceeds half the proportion of loci to be imputed. Short segments with a
lot of missing information will be discarded by this quadrupled offset. In the last
phasing and imputation iteration the offset is reduced to zero to fill in as many loci as
possible. When there is a large amount of missing information within a segment, it
will be accepted as shared if it meets the following thresholds: (i) segments in non-
founders in set ‘D’ require less than 20% missing loci, but no threshold is imposed for
set ‘S’; (ii) segments of founders in either set (‘D’ or ‘S’) require less than 50% missing loci. In general, the sparser the genotypes in set ‘S’ are, the more relaxed the
missing allele thresholds need to be. At the same time, relaxing the thresholds will
increase errors.

Comparisons to surrogates: Non-founders are compared with their surrogates one
gamete at a time and Figure 3 shows the individuals steps in this process. A
proband’s paternal gamete is compared to both of its surrogate father’s gametes
consecutively (Figure 2). Similarly, a proband’s maternal gamete is compared to its
surrogate mother’s gametes and both proband gametes are compared to both gametes
of descendants, and then genomic surrogates, one at a time. When a segment meets
the acceptance criteria, we collect allele information within the segment as a count of
allele ‘2’ and as a count of the total non-missing alleles in all surrogates which share a
segment with the proband at a particular locus. Once comparisons to surrogates have
concluded for a particular proband, its missing alleles are filled in based on the
collective information from all surrogates. For any given locus, allele ‘2’ is assigned
to the proband gamete if the ratio of allele ‘2’ counts over the total number of counts
from all contributing surrogates exceeds 0.7, assigned the ‘0’ allele if less than 0.3,
and remains as ‘5’ otherwise. The values of 0.7/0.3 can be changed to suit a particular
dataset. Filling in based on majority information is expected to reduce errors as the
information from all surrogates is used and not just the first surrogate that matches. Nevertheless, in the last imputation iteration a slightly modified version of this routine is run where filling in is not based on majority information but is filled in as soon as a segment is accepted as being potentially shared.

Phasing of founders, which is shown in Figure 4, is difficult because rule-based methods based on genotyped parents cannot be used to distinguish paternal or maternal gametes. We define founders as individuals without genotyped parents. Our algorithm makes use of surrogate fathers and mothers to partially distinguish the parental gametes of the founder, and then minimises the number of crossovers for phasing. Phasing and imputation of founders is accomplished by simultaneously comparing both their gametes to one gamete of a surrogate. If the surrogate is a surrogate father then they are compared to the proband’s paternal gamete and to the maternal gamete if comparing to a surrogate mother. If the surrogate is a descendent, a descendent of a surrogate father or mother, or a genomic surrogate, then both proband gametes are compared to both surrogate gametes consecutively. This comparison is shown in Figure 5 and the process is the same to that for non-founders with regard to finding shared segments. Strings of consecutive and matching loci are sought, but keeping track of which proband gamete matched the surrogate. Once the current proband gamete ceases to match, either because of a switch to matching with the other proband gamete, or because neither proband gamete matches the surrogate, then the segment is subject to the acceptance criteria above and, if accepted, missing alleles in the proband gamete are filled with allele information from the surrogate gamete.
The main rationale behind comparing both gametes of a proband founder simultaneously to the surrogate gamete is that it allows the algorithm to collect information on what appear to be crossovers. As an example, in Figure 5 the surrogate gamete first matches the proband’s paternal gamete, it then switches to matching the proband’s maternal gamete. The locus where the matching switches from paternal to maternal proband gamete could be regarded as either a crossover in the surrogate’s gamete, or the proband’s paternal and maternal gametes need to be flipped from that locus onwards (i.e. erroneous crossover in proband). All such crossover points are stored during the assessment of each proband. Before moving to the next proband, in loci where the number of crossovers among surrogates outnumbers the number of non-crossovers, the paternal and maternal gametes of the proband founder are flipped to remove the crossover. This process of minimising crossovers is required to be repeated as flipping one segment in the proband can reveal new crossover points (Figure 4). In most cases, 5 ‘minimise crossover’ iterations are sufficient for the number of crossovers to converge, but we allow up to 20. While minimising crossovers for founders is repeated in each invocation of 2B, only few adjustments are required in later invocations as new information becomes available.

Simulated populations for testing: Populations in mutation drift equilibrium were simulated by randomly mating individuals for 1000 generations with crossover and mutation. Effective population size ($N_e$) was 200 and the number of male and female parents was equal across generations. Previous work established that with this $N_e$ mutation drift equilibrium was achieved with 1000 generations. One male and one female offspring were produced per mating. Pedigree and genotype information was
retained for individuals in the last five generations. In generations 997 through 999, 100 individuals were simulated. Finally generation 1000 consisted of 200 individuals. for a total of 500 individuals considered for phasing and imputation. The number of individuals in the last generation was achieved by doubling the number of offspring per mating.

The genome simulated consisted of one chromosome which measured one Morgan. In generation zero all individuals were completely homozygous for the same allele in all 40,000 potential loci per Morgan and mutations were applied at a rate of \(2.5 \times 10^{-5}\) per locus per meiosis in the following generations. Mutations switched allele zero to two and vice versa. The number of mutations and crossovers per chromosome were sampled from a Poisson distribution. The mean for the number of mutations corresponded to the product of the number of loci per chromosome (both monomorphic and polymorphic) and the mutation rate, and the mean for crossovers was one per Morgan. The sampled mutations and crossovers were then randomly placed on the chromosome. A more detailed account of the simulations can be found in Daetwyler et al. (2010).

Approximately 1500 segregating bi-allelic loci exceeded a MAF of 0.02 at generation 1000, which is equivalent to a marker density of \(7.5N_e\) per Morgan. Linkage disequilibrium (LD, \(r^2\)) statistics (Hill and Robertson 1968) were 0.169, 0.066, 0.022, 0.008, and 0.005 for neighbouring SNP, and SNP which were 1, 5, 20, or 50 cM apart, respectively (SE were low). Allele frequency was found to follow a U-shaped distribution as expected.
Testing in simulated data: The utility of ChromoPhase was evaluated in 25 replicates of the data simulated as described above (genome summary = 1 chromosome, 1 Morgan, 1500 loci, 100 QTL). Phasing utility was checked in each replicate consisting of the last four, three and two generations in the data set. The pedigree used by the program was restricted to the generations being tested. Hence, no additional information was available on ancestors beyond the animals in the genotyped dataset. Inferred alleles were compared to true alleles and this yielded the following test parameters for both paternal and maternal alleles, i) percent correct, ii) percent missing, iii) percent wrong. Statistics on phasing were compiled only for non-founders.

Imputation of missing genotypes was evaluated in 25 population replicates. Three sparse SNP densities were simulated: 14, 34, and 100 per Morgan, which correspond to 420, 1020 and 3000 SNP for a 30 Morgan genome. Markers in the sparse subset were chosen to have higher than average MAF and were evenly distributed across the genome. If an animal was to be imputed, genotypes were set to missing in loci not chosen as part of the sparse set at the beginning of the algorithm. This was done in the last four, three, and two generations to investigate how ChromoPhase copes with varying depths of pedigree and resulted in nine scenarios (i.e. three sparse densities and three pedigree depths). For example, if there are four generations in the dataset, 3 parental generations had dense genotypes (1500 SNP per Morgan) and a proportion of animals in the last generation had sparse genotypes (14 to 100 SNP per Morgan) and need imputation. Test parameters chosen for imputation of genotypes were, i) percent correct, ii) percent missing and iii) percent wrong when compared to true genotypes from simulation.
Effect of imperfect imputation of genotypes was evaluated by applying genomic evaluation to the data set. One hundred QTL per Morgan were randomly sampled from the segregating loci. This assured that the number of QTL was larger than the number of independent chromosome segments (DAETWYLER et al. 2010). Additive allele substitution effects ($\beta$) were sampled from $N(0,1)$. True breeding values were calculated for each QTL as $2(1-m_j)\beta_j$ (where $m_j$ is the major allele frequency at locus $j$), $-2m_j\beta_j$ and $((1-m_j)-m_j)\beta_j$ for the major and minor homozygote and heterozygote genotype, respectively (FALCONER and MACKAY 1996). Phenotypes were generated by adding random environmental deviations to genotypic values which were also drawn from $N(0,1)$ and scaled to achieve a heritability of 0.3. While the imputation datasets may have included more than two generations (as noted in Table 5), only the last two generations (300 individuals) were used to estimate genomic breeding values to keep sample size constant. The genomic evaluation method fitted a realised relationship matrix ($G$) based on marker information (NEJATI-JAVAREMI et al. 1997; HAYES et al. 2009b). In each sparse marker density (14, 34, 100 per Morgan), three different scenarios were considered: i) all 300 individuals were genotyped at high density (Dense, 1500 SNP per Morgan), ii) all individuals were genotyped at sparse density (Sparse), iii) individuals in the last generation were imputed to high density from the respective sparse densities (Imp). The $G$ matrices were then fitted with phenotypes in ASReml to produce genomic breeding values (GILMOUR et al. 2000). The QTL were not masked, that is, they were part of the dense genotype data. This was expected to increase the difference in genomic evaluation accuracy between the Dense and Imp scenarios. The accuracy of the genomic evaluation was calculated as the statistical correlation between genomic and true breeding values in the training set.
Comparison with other programs: Our algorithm was compared to the fastPHASE program (SHEET and STEPHENS 2006). The algorithm described by Burdick et al. (2006) was also used to impute missing genotypes. Default settings were used in all cases, as preliminary investigations showed that altering the settings changed the accuracy of imputation less than 1%.

Testing in real Holstein data: Imputation accuracy was tested in a dataset of 1183 Holstein bulls, which were densely genotyped with the Illumina Bovine 50K array. Quality control reduced the number of SNP to 39048 SNP and is described in more detail in Hayes et al. (2009a). A pedigree consisting of 3674 animals was used to gather the genotyped relatives for each proband. We tested the imputation accuracy on chromosome one, which had 2529 SNP available. Two scenarios were tested, one where missing genotypes were imputed from the SNPs on the Illumina Bovine 3K to 50K density and one where we only imputed 5% of SNP in the 50K SNP chip. The second case was designed to mimic imputing upwards from the 50K chip density, because if our algorithm can impute this scenario with a certain accuracy then we should be able to impute from 50K to say 800K or even to full sequence with at least the same accuracy. The Illumina 3K chip had 182 SNP on chromosome one. Operationally, instead of re-genotyping with the 3K chip, we blanked out and imputed the non-sparse SNP in the 50K chip. We tested the imputation of 5% of SNP by random masking. Imputed genotypes were then compared to real 50K genotypes once the algorithm had completed to assess imputation accuracy.
RESULTS

Phasing: Phasing was evaluated in all non-founders and results can be found in Table 1. The percentage of alleles phased correctly when compared to true alleles was high, ranging from 99.97% when all generations were available to 99.91% when only 2 generations were included. Errors decreased slightly as the number of generations increased.

Imputation: Similar trends to phasing were observed for the accuracy of imputation. Table 2 shows means of 25 replicates of imputing missing genotypes at three different sparse densities. The percentage of correctly imputed genotypes was greatest when four generations of data (three generations of dense genotypes and last generation sparse) were available. However, performance was only slightly reduced when three generations were in the dataset. The number of non-imputed genotypes increased slightly as the number of generations decreased. The proportion of correctly imputed loci increased and the proportion of wrongly imputed genotypes decreased as the density of the sparse genotypes increased. In Table 3, one can observe that as soon as some individuals (20%) in the last generation have dense genotypes, more information flows to the sparsely genotyped individuals and the proportion of correctly imputed genotypes increases by approximately 5%. This demonstrates that collateral relatives can be used to improve imputation of missing genotypes when the number of ancestral dense genotypes is limited. When using two or three generations of dense data to impute 80% of sparsely genotyped individuals in the last generation, very similar results to imputing all individuals were obtained. This shows that with sufficient ancestral information, dense genotyping of collateral relatives is of limited value (results not shown).
ChromoPhase correctly imputed approximately 12 to 16% more missing genotypes than fastPHASE for the scenarios tested (Table 4). Our algorithm was also tested against that of Burdick et al. (2006) and was found to impute at least 27% more loci. This may be because the Burdick et al. (2006) approach was not designed to handle very sparse marker densities. Computation time in cpu time for ChromoPhase, fastPHASE and Burdick et al. (2006) were approximately 30 cpu seconds, 3 hours 19 minutes, and 20 seconds, respectively.

The accuracy of genomic evaluation (correlation of predicted and true breeding values) using dense genotypes to calculate the genomic relationship matrix (G) was 75.5% (Table 5). While the depth of the dataset was varied between two and four generations, only the last two generations (300 animals) were used to estimate genomic breeding values to keep the sample size constant. The accuracies of the sparse scenarios were all very similar at 56.1%. At these low densities, increasing the sparse loci set slightly seemed not to increase accuracy. As expected, accuracy increased when more missing genotypes were imputed. However, it was apparent that although imputation was imperfect, the accuracy of genomic estimated breeding values achieved for the sparse subset of markers as a percentage of that achieved for the dense accuracy (i.e. accuracy with 1500 loci) was in all cases greater than the proportion of correctly imputed genotypes (Table 5). Consider the scenario with 3 generations in the dataset and a sparse genotype of 14 loci per Morgan, here 87.8% of missing genotypes were correctly imputed but 95% of the dense genomic evaluation accuracy was achieved. Thus, it seems that genomic evaluation is able to cope well with a percentage of loci missing or even wrongly imputed, which has been confirmed
independently by Weigel et al. (2010) using a Bayesian method to estimate SNP
effects for genomic predictions.

Real data presents additional challenges such as ungenotyped parents, genotyping
errors and map location errors. Thus the imputation accuracy in real Holstein cattle
data was reduced when compared to results in simulated data (Table 6). Nevertheless,
the imputation accuracy in real Holstein cattle data using the sparse 3K chip was
92.5% in non-founders and 72.8% in founders. When 5% of 50K loci were randomly
masked the imputation accuracy was 97.2% in non-founders and 90.2% in founders.
The lower performance in real data when compared to simulated data is due in part to
imperfections in the real data, such as map and genotyping errors. In addition, there
were many more genotyped founders in the simulated data set and our Holstein data
set contained no genotyped dams. It is expected that in the future more genotypes
would be available (including more dams), which would allow imputation accuracy in
real data to move towards results achieved in simulated data.

DISCUSSION

We have described a long-range phasing and imputation algorithm which seeks out
and phases long chromosomal segments which are shared between close or distant
relatives. The algorithm takes advantage of family data, in addition to population
data, to improve performance over methods which phase per locus or only consider a
few loci at a time and use population data only. The results demonstrate that
ChromoPhase is accurate for both pure phasing of genotyped loci and for imputation
of missing genotypes. The identification of shared chromosome segments is
important to phase and impute genotypes, as any missing information within a
segment can potentially be filled in the proband with information from its surrogate.  

The key aspect of identifying a shared segment is the recognition that the probability  
that a long haplotype coalesces to a common ancestor becomes high if two animals  
match at a high proportion of alleles within a segment and there are no non-matches.  

Kong et al. (2008) made use of this concept in individuals of unknown relationship by  
searching for a sufficiently long stretch of loci with no incompatible genotypes that  
can therefore be assumed to have originated in a common ancestor. All potential  
surrogates of a proband for a genome segment of predefined length were identified  
and stored at the beginning of the Kong algorithm. They then phased a proband by  
cycling through its surrogates to identify a homozygote at a particular locus. Our  
approach is similar, but operationally different, as we also make use of pedigree  
information and thus we are able to compare alleles within family relationships, and  
to compare at the level of the allele instead of genotypes. Our algorithm compares  
relatives in each iteration and genomic surrogates in later iterations to make use of  
new information as it becomes available and we do not specify a maximum length for  
shared segments. Thus, a shared segment may potentially span the full chromosome  
and allows us to use all available information. Consequently ChromoPhase uses most  
of the information used by Kong et al. (2008), but uses some additional information,  
such as pedigree and simple rule based filling. 

The use of both family and population genomic information makes our approach  
feasible in species which have incomplete pedigree information or only have few  
parents with dense genotypes. Genotyped parents allow for rule-based filling of  
alleles which is important in early iterations as little information is available to  
distinguish gametes. Our results indicate that having at least one densely genotyped
parent is crucial to achieving high accuracy imputation. This is partly because rule-
based filling is difficult in individuals without genotyped parents and therefore a
larger number of phasing or imputation errors are likely to occur in these founders.
Imputation is especially difficult in younger animals which have no genotyped parents
or genotyped offspring available (Table 6). Our algorithm may need further
development to be able to capitalise on population linkage disequilibrium more
effectively and improve imputation rates in founders. However, comparing to
relatives, if available, results in significant computer time savings, as animals do not
have to be compared to all other animals. Furthermore, restricting comparisons to
relatives in the first iterations reduced error rates.

The comparison of haplotypes in our algorithm also results in computational
efficiency because the same process is used for phasing and imputation. The main
objective of ChromoPhase is to complete as much information as possible in a
proband haplotype by using information from shared segments with relatives. It is
therefore irrelevant from the method’s point of view whether this is for phasing or
imputation, though the algorithm benefits when genotypes are available at a locus.

Haplotype libraries have been proposed as a way to phase and impute genotypes
(Hickey et al. 2011; VanRaden et al. 2011). Conceptually, these approaches are not
very different from ChromoPhase. While they explicitly build a library of haplotypes,
our algorithm contains the haplotypes in each individual with dense genotypes. A
library may provide computational efficiencies when there are only few main
haplotypes segregating in a population so that the proband only needs to be compared
to few haplotypes. This may be the case in domestic animal populations with low $N_e$. 
A draw back of the library approach is that the haplotypes are restricted to a specific length. As an example, consider a SNP which has been mapped to the wrong location on the genome or it has been incorrectly genotyped in the lab. First, this would inflate the number of haplotypes stored in the library and, second, if an individual has a genotyping error, a match cannot be identified in the library for that whole segment, resulting in reduced imputation. In contrast, our approach will compare animals up to the map or genotyping error and, if a non-match occurs, will stop the shared segment. However, the next shared segment may start right after the non-match increasing imputation when compared to a library approach.

The application of our method in real datasets may require further development to address several challenges, such as completely ungenotyped animals in the data, incomplete pedigrees, genotyping errors, rare alleles and SNP mapped to wrong genome locations. Currently, completely ungenotyped individuals are not attempted to be imputed. However, doing so should be feasible if an ungenotyped individual has both a genotyped parent and genotyped progeny. It is not expected that imputing completely ungenotyped individuals would increase imputation rates, as they offer no additional information. In fact, it was observed that more errors occurred when ungenotyped animals were imputed and used to impute sparsely genotyped animals, because proportionally more errors occur in completely ungenotyped animals which then transfer to sparsely imputed animals. It is also important that correct and as complete as possible pedigree information is available for determining surrogates. Most genotyping errors can be detected by comparing trios, though if they are not detected then they may result in erroneous haplotype assignments. Rare alleles can only be imputed if they are observed in the dense genotypes; hence the dense sample
needs to be sufficiently large. Map errors may cause a wrongly mapped locus to appear shared between relatives where it may only be a match by chance (i.e. it is only IBS) causing phasing errors. Every effort should be made to correct map errors on SNP chips.

Currently the algorithm applies to autosomes and further modification to sex chromosomes may be necessary. Crossover occurs freely between X chromosomes hence, phasing involving females is expected to be the same as autosomes. Simplification may be possible in males because X-specific markers in the non paralogous region are already phased. Pseudoautosomal crossover between X and Y is believed to be restricted to the relatively short regions at either end of the X chromosome in human representing approximately 2% of all bases in total (CHARLESWORTH 1991; ROSS et al. 2005). It has been reported that the ratio of genetic versus physical distance is inflated in the pseudoautosomal region, which may indicate high haplotype diversity (e.g. ARIAS et al. 2009; DUMONT et al. 2011).

The application highlighted here is to impute haplotypes and missing genotypes from sparse to medium density. The current study confirms that it is feasible to impute a large number of missing genotypes to a higher density, though the performance is dependent on the sparse chip density and the number of generations in the dataset. It will be feasible to use our method to impute 50K genotypes to even higher density once information from such denser SNP chips becomes available in their relatives. So there is no upper limit to how dense the genotypes can be for successful imputation, and even imputing full genome sequence data will eventually be possible once sufficient ancestors have been sequenced.
An interesting result is that when we applied genomic evaluation to the dense, sparse and imputed datasets, the reduction in genomic evaluation accuracy is only small with imperfectly imputed genotype data. It was apparent that while imputation is imperfect, the proportion of dense accuracy observed was in all cases larger than the proportion of correctly imputed genotypes. This may be due to more than one SNP being in linkage disequilibrium with each of the QTL across the genome, so that the consequence of a single incorrectly imputed SNP on predicting the effect on a QTL is reduced.

The potential for ChromoPhase to increase the number of genotyped individuals while simultaneously reducing genotyping costs is very large. Key benefits will be increased sample sizes to achieve higher accuracies in genomic selection and to increase the power of QTL studies. Reducing genotyping costs through strategic genotyping of ancestors and upgrading to denser genotyping from sparser SNP chips in the current generation with programs such as ChromoPhase will allow for the application of genomic selection in species where currently this technology is not economically feasible.

ACKNOWLEDGEMENTS

HDD was in part supported by the SABRETRAIN Project, which is funded by the Marie Curie Host Fellowships for Early Stage Research Training, as part of the 6th Framework Programme of the European Commission. The authors thank Phil Bowman for providing a Holstein pedigree and two anonymous reviewers for their
helpful comments. The software is available by request from the corresponding author.
Table 1. Phasing performance in percent in paternal alleles of non-founder individuals when the dataset consisted of 4, 3, or 2 generations (Gen.) of genotyped animals (maximum SE < 0.01).

<table>
<thead>
<tr>
<th>Gen.</th>
<th>Correct</th>
<th>Missing</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>99.97</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>99.96</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>99.91</td>
<td>0.03</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 2. Imputation performance in percent when imputing missing genotypes in all individuals in the last generation for three different sparse densities per Morgan and 4, 3 or 2 generations (Gen.) included in the dataset (maximum SE < 1.72%).

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Gen.</th>
<th>Sparse Density</th>
<th>Correct</th>
<th>Missing</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>14</td>
<td>87.79</td>
<td>0.87</td>
<td>11.34</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>34</td>
<td>94.13</td>
<td>0.29</td>
<td>5.58</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>98.67</td>
<td>0.06</td>
<td>1.27</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>87.80</td>
<td>1.00</td>
<td>11.20</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>94.07</td>
<td>0.31</td>
<td>5.62</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>98.57</td>
<td>0.08</td>
<td>1.36</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>81.36</td>
<td>1.48</td>
<td>17.2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>83.62</td>
<td>1.49</td>
<td>14.89</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>87.07</td>
<td>1.93</td>
<td>11.00</td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Imputation performance in percent when imputing missing genotypes in 80% of individuals in the last generation for three different sparse densities per Morgan and 2 generations (Gen.) included in the dataset (maximum SE < 1.5%).

<table>
<thead>
<tr>
<th>Gen.</th>
<th>Sparse Density</th>
<th>Correct</th>
<th>Missing</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>14</td>
<td>84.89</td>
<td>1.37</td>
<td>13.75</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>89.39</td>
<td>0.86</td>
<td>9.76</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>94.09</td>
<td>0.61</td>
<td>5.30</td>
</tr>
</tbody>
</table>

Table 4. Imputation performance of ChromoPhase, fastPHASE and Burdick et al. (2006) in replicate 1, shown as the percentage of correctly imputed missing genotypes in all individuals in the last generation for three different sparse densities per Morgan and 3 generations (Gen.) included in the dataset.

<table>
<thead>
<tr>
<th>Gen.</th>
<th>Sparse Density</th>
<th>ChromoPhase</th>
<th>fastPHASE</th>
<th>Burdick et al</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>14</td>
<td>87.8</td>
<td>75.3</td>
<td>60.5</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>94.1</td>
<td>78.6</td>
<td>60.8</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>98.6</td>
<td>85.0</td>
<td>61.6</td>
</tr>
</tbody>
</table>
Table 5. Accuracy of estimated breeding values in percent when using dense (Dense), sparse (Sparse), and imputed (Imp) genotypes to calculate realised relationship matrices (maximum SE < 1.0%). Gen refers to the number of generations in dataset used for imputation. Sample size 300 (i.e. last two generations) for genomic evaluation in all scenarios.

<table>
<thead>
<tr>
<th>Gen.</th>
<th>Density</th>
<th>Scenario</th>
<th>Accuracy of Genomic Evaluation</th>
<th>% of All Dense Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500</td>
<td>Dense</td>
<td></td>
<td>75.5</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
<td></td>
<td>56.1</td>
<td>74.3</td>
</tr>
<tr>
<td>4</td>
<td>Imp</td>
<td>14</td>
<td>71.7</td>
<td>95.0</td>
</tr>
<tr>
<td>4</td>
<td>Imp</td>
<td>34</td>
<td>74.1</td>
<td>98.1</td>
</tr>
<tr>
<td>4</td>
<td>Imp</td>
<td>100</td>
<td>75.5</td>
<td>100.0</td>
</tr>
<tr>
<td>3</td>
<td>Imp</td>
<td>14</td>
<td>71.7</td>
<td>95.0</td>
</tr>
<tr>
<td>3</td>
<td>Imp</td>
<td>34</td>
<td>74.0</td>
<td>98.0</td>
</tr>
<tr>
<td>3</td>
<td>Imp</td>
<td>100</td>
<td>75.5</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>Imp</td>
<td>14</td>
<td>68.8</td>
<td>91.1</td>
</tr>
<tr>
<td>2</td>
<td>Imp</td>
<td>34</td>
<td>71.2</td>
<td>94.3</td>
</tr>
<tr>
<td>2</td>
<td>Imp</td>
<td>100</td>
<td>72.7</td>
<td>96.3</td>
</tr>
</tbody>
</table>
Table 6. Imputation performance of ChromoPhase in percent on chromosome one in real Holstein cattle data tabulated as the percentage correctly imputed missing genotypes in the youngest founders (F) and non-founders (NF) for two sparse densities.

<table>
<thead>
<tr>
<th>Sparse Density</th>
<th>Type</th>
<th>Number of animals imputed</th>
<th>Correct</th>
<th>Missing</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>182</td>
<td>NF</td>
<td>112</td>
<td>92.5</td>
<td>1.1</td>
<td>6.3</td>
</tr>
<tr>
<td>182</td>
<td>F</td>
<td>212</td>
<td>72.8</td>
<td>0.6</td>
<td>26.5</td>
</tr>
<tr>
<td>2400</td>
<td>NF</td>
<td>112</td>
<td>97.2</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>2400</td>
<td>F</td>
<td>278</td>
<td>90.2</td>
<td>0.7</td>
<td>9.1</td>
</tr>
</tbody>
</table>
Figure 1. Representation of the three stages and workflow in the long-range (LR) phasing and imputation algorithm, where Dense and Sparse refer to densely and sparsely genotyped individuals.
Figure 2. Comparison of proband with a genotyped parent (non-founder) to its surrogate father.

Alleles on the proband’s paternal gamete (\(pa_s\)) are compared to alleles on surrogate’s paternal gamete (\(pa_l\)) to identify a shared chromosome segment, where ? is a possible match, + is a match and – is a non-match. The shared segment starts with the first match and ends at the last match. Offset signifies the beginning and end of a matching segment which is removed to guard against errors due to missing alleles (5) before using the segment to fill in information.
Figure 3. Representation of comparing non-founders (individuals with at least one genotyped parent) to their surrogates (densely genotyped relative or genomic surrogate).
Figure 4. Representation of comparing founders (individuals with no genotyped parents) to their surrogates (densely genotyped relative or genomic surrogate), including iterations in which the number of crossovers are minimised in the proband.
Figure 5. Comparison of proband without genotyped parents to one of its surrogates to illustrate crossover minimisation.

Alleles on both gametes of proband (\(pa\)) and (\(ma\)) are compared to alleles on surrogate paternal gamete (\(pa_s\)) to identify a shared chromosome segment, where ? is a possible match, + is a match and – is a non-match. The proband shares \(pa\) with the surrogate at the start of the chromosome and then this switches to \(ma\). The ‘switch’ loci could either be a crossover in the surrogate or, more likely, the paternal and maternal gametes need to be ‘flipped’ in the proband. Proband gametes are only flipped if, after cycling through all surrogates and storing information on all crossovers, there is more evidence for crossovers than no crossovers. Offsets are applied but not shown in the diagram for simplicity.
REFERENCES


BROWNING, B. L., and S. R. BROWNING, 2009 A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. American Journal of Human Genetics 84: 210-223.


TIER, B., 2006 Haplotyping for linkage disequilibrium mapping, pp. in 8th World Congress of Genetics Applied to Livestock Production, Belo Horizonte, Brasil.


