J. Dairy Sci. 100:2837–2849
https://doi.org/10.3168/jds.2016-11590

A comparison of different algorithms for phasing haplotypes using Holstein cattle genotypes and pedigree data

Younes Miar,*†[1] Mehdi Sargolzaei,†‡ and Flavio S. Schenkel†
*Department of Animal Science and Aquaculture, Dalhousie University, Truro, Nova Scotia, Canada B2N 5E3
†Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, Ontario, Canada N1G 2W1
‡The Semex Alliance, Guelph, Ontario, Canada N1H 6J2

ABSTRACT

Phasing genotypes to haplotypes is becoming increasingly important due to its applications in the study of diseases, population and evolutionary genetics, imputation, and so on. Several studies have focused on the development of computational methods that infer haplotype phase from population genotype data. The aim of this study was to compare phasing algorithms implemented in Beagle, Findhap, FImpute, Impute2, and ShapeIt2 software using 50k and 777k (HD) genotyping data. Six scenarios were considered: no-parents, sire-progeny pairs, sire-dam-progeny trios, each with and without pedigree information in Holstein cattle. Algorithms were compared with respect to their phasing accuracy and computational efficiency. In the studied population, Beagle and FImpute were more accurate than other phasing algorithms. Across scenarios, phasing accuracies for Beagle and FImpute were 99.49–99.90% and 99.44–99.99% for 50k, respectively, and 99.90–99.99% and 99.87–99.99% for HD, respectively. Generally, FImpute resulted in higher accuracy when genotypic information of at least one parent was available. In the absence of parental genotypes and pedigree information, Beagle and Impute2 (with double the default number of states) were slightly more accurate than FImpute. Findhap gave high phasing accuracy when parents' genotypes and pedigree information were available. In terms of computing time, Findhap was the fastest algorithm followed by FImpute. FImpute was 30 to 131, 87 to 786, and 353 to 1,400 times faster across scenarios than Beagle, ShapeIt2, and Impute2, respectively. In summary, FImpute and Beagle were the most accurate phasing algorithms. Moreover, the low computational requirement of FImpute makes it an attractive algorithm for phasing genotypes of large livestock populations.

Received June 11, 2016.
Accepted December 9, 2016.
[1]Corresponding author: miar@dal.ca

Key words: haplotype inference, imputation, livestock, phasing accuracy

INTRODUCTION

Haplotypes are combinations of alleles that are present on each of 2 homologous chromosomes in a diploid individual. Some statistical methods exist for inferring haplotypes from observed genotypes. The inference of a haplotype from genotype data is called "phasing." The importance of haplotype phasing is increasing with the availability of enormous amounts of genotype data generated by high-throughput technologies. Several applications of phasing include imputation of untyped genetic variation (Marchini et al., 2007; Browning and Browning, 2009; Li et al., 2010), interplay of genetic variation and phenotype (Tewhey et al., 2011), population evolutionary history (Tishkoff et al., 1996), linkage disequilibrium mapping, calling genotypes in microarray and sequence data (Kang et al., 2004; Li et al., 2011), detecting genotyping errors (Scheet and Stephens, 2008), inferring points of recombination (Kong et al., 2008), detecting recurrent mutation (Kong et al., 2008), signatures of selection (Sabeti et al., 2002), and modeling cis-regulation of gene expression (Tao et al., 2006). Recently, advances in genotyping technologies and computational approaches have improved the accuracy of haplotyping but experimental methods are expensive and time consuming. On the other hand, computational (in silico) phasing methods are inexpensive but may be time consuming. Computational methods are generally divided into family-based and population-based methods that use linkage information from close relatives and linkage disequilibrium information from population, respectively (Li et al., 2009).

Family-based methods are mostly rule-based methods such as those proposed by Burdick et al. (2006) and Kong et al. (2008). Population-based methods are often based on the stochastic model, and their accuracy depends on sample size, marker density, genotype accuracy, allele frequency, ethnicity, and relatedness (Browning and Browning, 2011). Population-based ap-

proaches can be highly accurate if high-density markers and large sample sizes are used but they are computationally intensive (Sargolzaei et al., 2014).

The task of phasing in livestock populations is something of a special case, because individuals exhibit much higher levels of relatedness and tend to share much longer stretches of chromosomes compared with individuals in the human population. Lander et al. (1987) used hidden Markov models (**HMM**) to construct primary genetic linkage maps of experimental and natural populations, which is implemented in Mapmaker. Currently, the most accurate methods use HMM to infer the haplotypes using linkage disequilibrium information (Browning and Browning, 2009; Delaneau et al., 2013; O'Connell et al., 2014). A method proposed by Kong et al. (2008) uses the surrogate parents to infer long haplotypes with high accuracy using Mendelian inheritance rules. Palin et al. (2011) proposed a model based on this approach called "systematic long-range phasing." Meuwissen and Goddard (2010) proposed a combined family and population phasing approach in which family information is used by iterative peeling algorithm following by an approximation of identical-by-descent probabilities. Sargolzaei et al. (2014) developed a rule-based method for phasing that exploited the relationships between individuals based on the fact that close relatives share longer haplotypes and distant relatives share shorter haplotypes.

Because of the relatively small effective population size and planned breeding in most livestock populations, a wide range of structured relationships between individuals is usually observed. For example, large half-sib families are common in livestock. Assessing the performance (i.e., accuracy and computational requirements) of alternative phasing algorithms for livestock populations is important before performing haplotype phasing in research or applied settings. There exist a few investigations on the performance of algorithms for phasing genotypes in livestock, especially using large genotypes and pedigreed data sets. Therefore, the objective of this study was to investigate the phasing accuracy and computing requirements of 5 previously published statistical algorithms for inferring haplotypes from genotype data in a large Holstein cattle population, which were implemented in Beagle (Browning and Browning, 2009), Findhap (VanRaden et al., 2013), FImpute (Sargolzaei et al., 2014), Impute2 with both default and high-accuracy settings (Howie et al., 2009), and ShapeIt2 with both default and high-accuracy settings and the new duoHMM algorithm for scenarios with pedigree information (Delaneau et al., 2013; O'Connell et al., 2014).

MATERIALS AND METHODS

Data Sets

To provide a comprehensive assessment of the accuracy of algorithms, 6 different scenarios (3 data subsets with or without pedigree) that vary in the extent of the relatedness between individuals from the North American Holstein genotype database were analyzed. The scenarios are summarized in Table 1. The data set was provided by the Canadian Dairy Network (CDN, Guelph, ON, Canada), contained 2,495 and 118,946 animals genotyped with Illumina BovineHD (**HD**) and BovineSNP50 (**50k**) BeadChips (Illumina Inc., San Diego, CA), respectively. The North American Holstein database contains individuals registered in Canada and the United States. Quality control was performed on 50k genotypes by the Council on Dairy Cattle Breeding (CDCB, Bowie, MD). Details of quality control measures are given in Wiggans et al. (2009). A total of 45,187 SNP were retained for analysis after filtering. Genotyped animals that were born from 2012 to 2015 with both parents also genotyped comprised the validation set. Genotypes of parents were used to determine phase of heterozygous loci in validation animals, for which parents carried opposing homozygous genotypes. These inferred phases were considered highly accurate and were used to assess haplotype accuracy from different algorithms. The validation set for the 50k panel included 9,266 dairy cattle. Influential genotyped animals, which had more than 40 offspring and were not parents of animals in validation sets, were included in all 6 scenarios for the 50k panel as influential animals (Table 1). Also, 1,916 animals with HD genotypes that were not parents of validation animals were considered influential animals for the HD panel (Table 1). The percentage of influential animals that were grandparents of validation animals was 0.00 and 1.59% for the 50k and HD data sets, respectively. The average pedigree-based relationship between validation and influential animals that had pedigree information were 0.13 (ranged from 0.02 to 0.62) and 0.11 (ranged from 0.01 to 0.46) for the 50k and HD data sets, respectively. No-parents scenarios included both the validation animals and the influential animals; the sire-progeny pair scenarios included validation animals, influential animals, and sires; and sire-dam-progeny trio scenarios included validation animals, influential animals, sires, and dams (Table 1).

The validation set for the HD panel included 284 dairy cattle that were genotyped with the HD panel. Parents of validation animals had HD or imputed HD genotypes from 50k available. Imputation from 50k