Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Using Half Sib Families to Evaluate the Accuracy of Haplotype Reconstruction

M. H Ferdosi*, J. HJ van der Werf*, B. Tier†, C. Gondro*

*School of Environmental and Rural Science, University of New England, Armidale, Australia,
†Animal Genetics and Breeding Unit, University of New England, Armidale, Australia

ABSTRACT: The BEAGLE program is frequently used for haplotype reconstruction using dense molecular marker genotype data in animal, plant or human populations. In this study, we evaluate the accuracy of BEAGLE for phasing with a population that consists of large half-sib families. The half-sib structure allows the accurate detection of phase thus making it easier to detect problems in more general population based algorithms such as BEAGLE. We show that the main problem in the haplotypes inferred by BEAGLE is the occurrence of switch errors where the parental origin of haplotypes is incorrectly swapped. This occurs especially in the large chromosomes and often in ~50% of individuals. Understanding the issue will allow better decision making about further analyses that relies on the haplotype origin of markers.

Key words: haplotype reconstruction; accuracy; beagle.

INTRODUCTION

Vast amounts of molecular markers genotypes are being identified in individual human, plant or animals samples, mostly as Single Nucleotide Polymorphisms (SNPs). For genomic analysis, there is often benefit in determining the phase of the genotype information, i.e. to determine the haplotypes according to the origin of parent. This can be useful for the purpose of imputation, or some models in Genome-Wide Association Studies that use parental origin in the method of inference.

Several methods have been suggested to reconstruct the haplotype from genotypic data on markers (Browning et al. 2011). BEAGLE (Browning et al. 2007) is one of the well-known algorithms developed for haplotype reconstruction of human populations, but is also frequently used by animal and plant breeders. Data from human populations mainly consists of unrelated individuals, whereas livestock populations consist of large families of half-sibs and to a lesser extent of full-sibs. This family structure can be utilized more explicitly, for example in half-sib families it is relatively easy to detect recombination sites and to recover phase (as parental origin) and block structure of the paternal haplotypes (grand parental origin) (Ferdosi et al., 2014). Such data is also suitable to test the performance of commonly used phasing algorithms. The aim of this study is to evaluate the phasing results of BEAGLE using genotypic information on sires and their progeny, and assess the ability of the algorithm to identify recombination events in the sire.

MATERIALS AND METHODS

BEAGLE algorithm. A localized haplotype cluster model with hidden Markov model (HMM) is used by the BEAGLE algorithm (Browning et al. 2007) to identify the most likely haplotypes based on the genotypes of individuals.

Real data. Nineteen half-sib groups were selected from the SheepGenomics project (White et al. 2012) and the CRC for Sheep Industry Innovation project (Van der Werf et al. 2010) in Australia. The data consisted of 1605 genotyped individuals, consisting of 19 half sib groups, each with one genotyped sire and the progeny group varying in size from 7 to 387. Genotypic data was available on those animals based on the 50k Illumina Ovine SNP chip. The details of population structure are described in (Ferdosi et al. 2014).

Inconsistency between progeny haplotype and sire genotype (IPHSG). Opposing homozygote marker genotypes between sire and offspring are useful for parentage assignment (Hayes 2011). The IPHGS is a similar criterion for identification of phasing errors in the haplotype of offspring. Each offspring inherits one haplotype from the sire; therefore, this haplotype should not have any inconsistency with the genotype of sire (other than those caused by genotyping errors). At the loci at which the genotype of offspring is heterozygous, the IPHSG criterion identifies a phasing error in the half-sib family regardless of phasing error in the sire. To evaluate the inconsistency between sire genotype and half-sibs' haplotypes the markers that were homozygous in the sire and heterozygous in the offspring were selected. Then the numbers of IPHSG between these markers for the both haplotypes of offspring were calculated and the number of IPHSG for the strand with a lowest value was recorded.

Haplotype partitioning and grouping. Half-sib families share large segments which are identical by descent (IBDs). These segments are inherited from the sire and 50% of haplotypes of a half-sib must be the same as the haplotypes of sire. To identify the IBD regions a matrix was created that contains the haplotypes of half-sibs (individual's haplotype × SNP – with two rows for each individual – Figure 2-B). Then the haplotype of the sire (sire genotype was also phased by BEAGLE) was added to the top of this matrix (Figure 2-A). This matrix (haplotypes of sire and half-sibs) was partitioned into covering segments with length of 100 SNPs. Common segments for each partition was placed in the same group. Since the first and second haplotypes belong to the sire, the segments inherited from the sire were in the same group. The results of each segment (individual's haplotype × 100 SNPs) were combined to create a new grouping matrix. In this matrix the numbers 1 and 2 show the segments that are in common between the sire and its progeny. By keeping these numbers and replacing the other numbers with zero the matrix can be

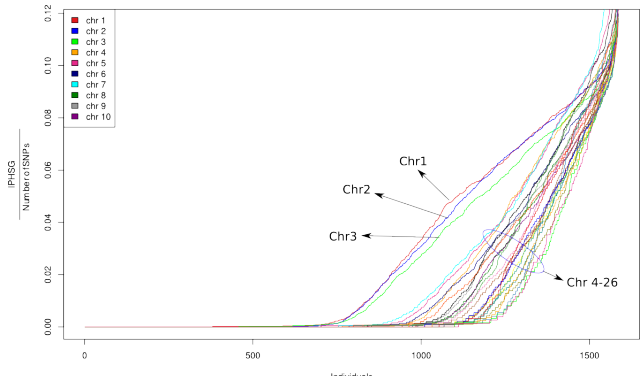

Figure 1: The sorted IPHSG for all individuals and different chromosomes

plotted. Because of genotyping errors this method ignores one SNP for grouping.

Identification of recombination events. The grouping matrix will take a code 0 for segments where the origin was unknown, a code 1 for inheriting sire strand 1 or code 2 for inheriting sire strand 2, respectively. The first and second rows that belong to the sire were removed. By replacing 2's with 1's in this matrix and adding the rows in the matrix (row sums) a vector was generated. Then this vector was partitioned into segments with length of two, with each segment referring to the two haplotypes of one individual. We expected that one of these numbers be 0 or equal to the genotyping error and the other number equal to the number of SNP on the chromosome. Observations of values more than genotyping error for both numbers indicate switch error. The switch error problem is described in detail by (Andres et al. 2007).

RESULTS AND DISCUSSION

Inconsistency between haplotype of half-sibs and genotype of sire. Figure 1 shows the result of sorted IPHSG divided by number of SNPs in each chromosome for all individuals. Around 50% of individuals have few inconsistencies with the sire genotype. In the other animals the IPHSG/SNP ratio is higher, between 0 and 12%. The amount of inconsistency is related to the length of chromosome, as the IPHSG/SNP ratio is higher for chromosomes 1-3. Since there are more recombinations in large chromosomes, we suspect that the IPHSG/SNP rate is affected by recombinations. Therefore, the BEAGLE algorithm might be sensitive to recombination events and in those cases failed to identify the phase of haplotypes correctly for some individuals.

Haplotype partitioning and grouping. The plot of common haplotypes received by the progeny from their sire can illustrate the details of haplotype reconstruction problems (Figure 2). One haplotype of each individual must be matched completely with both haplotypes of sire with respect of recombination events and the other haplotype should not have any similarity with the haplotypes of sire (except possible inbreeding that causes IBD between sire and dam). The existence of sire haplotypes in both strands of their progeny indicates a switch error in the haplotype reconstruction (Figure 2-C). Switch errors in the BEAGLE result may cause a problem for the studies related to recombination events or where it is important t correctly identify the parental origin of a haplotype, e.g. in crossbreeding studies. Since the segments of 100 SNPs were used for the partitioning of haplotypes, there is a 100 SNPs gap (white area - Figure 2-E) in the recombinant strands. These gaps can be seen between the strands of haplotype without recombination which may be caused by genotyping or phasing errors. If they are consistent across several continuous SNPs it may be caused by phasing errors in the sire. It seems that the haplotype of some individuals were completely wrong (Figure 2-F – lines 21 and 22). There are also gaps in the haplotypes originated from one strand of sire (Figure 2-F) that are possibly caused by genotyping errors or phasing errors. However, in some situations these gaps are common for some SNPs and only exist in one haplotype originated from the sire. Only genotyping error in the sire can explain this problem (Figure 2-G). At the end of the chromosome there are more often recombination and switch errors, which is probably due to the fact that it is harder to determine phase at the edges of chromosomes.

Similar to the IPHSG results, the switch error was also higher in larger chromosomes. The maximum number of switches occurred in the chromosome 2 (in ~0.51% of individuals) and the minimum was in chromosome 25 (in ~0.12% of individuals). The R^2 of linear model between number of SNPs and percent of switch was 0.89.

The accuracy in different family size was varied which suggest that family size may not affect the phasing accuracy inside a family.

CONCLUSION

BEAGLE can identify the phase of short segments accurately but for the large chromosomes that have relatively more recombinations, the switch error occurred frequently and the frequency of inconsistencies between progeny haplotype and sire genotype was increased. Therefore the phase results are useful for the analysis where the phase of short segment is important but inference based on long segment haplotypes is unreliable. The half-sib family structure allows the accurate detection of recombination events and is therefore a good reference for recognizing switch errors, map or genotyping errors, and recombination events in the sire genotype. In addition, pedigree and half-sib families are valuable information for haplotype reconstruction and utilizing them can increase the phasing accuracy and decreasing the switch errors. Extending the haplotype partitioning and grouping algorithm can also fix the switch errors in the BEAGLE result even without genotyping the sire.

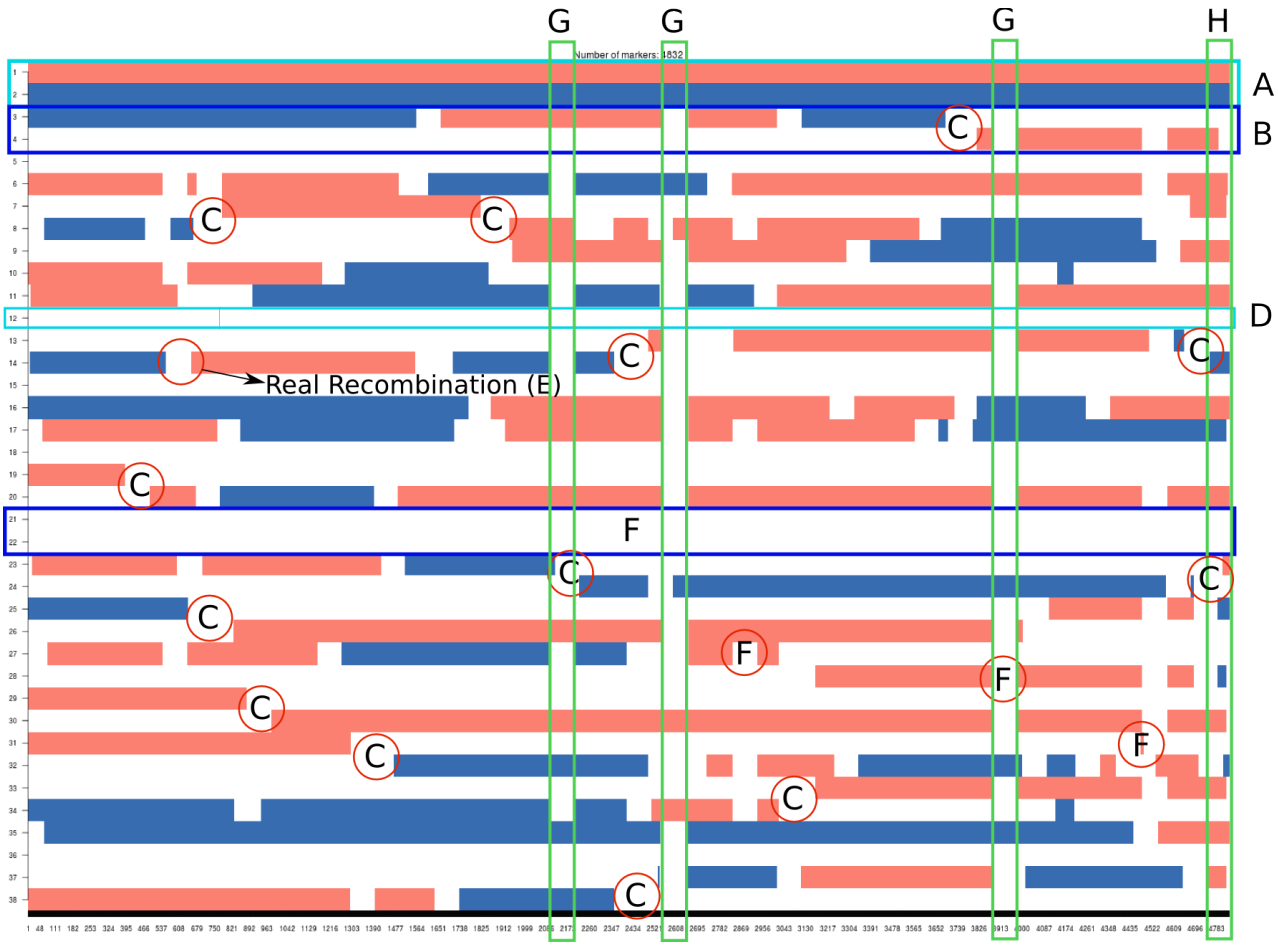

Figure 2: Image plot of group matrix for 18 half-sibs. A: Haplotypes of sire. B: Haplotypes of one half-sib. C: Switch errors. D: Maternal strand of one half-sib. E: Recombination. F: Phasing or genotyping errors. G: Genotyping error in the sire. H: Possible mapping error

Literature Cited

Andres, A. M., A. G. Clark, et al. (2007). Gen. Epid. **31**(7): 659-671.

Browning, B. L. and S. R. Browning (2007). Gen. Epid. **31**(6): 606-606.

Browning, S. R. and B. L. Browning (2011). Nat. Rev. Gen. **12**(10): 703-714.

Ferdosi, M. H., B. P. Kinghorn, et al. (2014). Gen. Sel. Evo. **46**(1): 11.

Hayes, B. J. (2011). J. Dairy Sci. **94**(4): 2114-2117.

Van der Werf, J. H. J., B. P. Kinghorn, et al. (2010). Ani. Prod. Sci. **50**: 1-6.

White, J. D., P. G. Allingham, et al. (2012). Ani. Prod. Sci. **52**: 157 – 171.