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

Pedigree-Based Haplotype Reconstruction, Identification of Cross-overs and Detection of Map and Genotyping Errors using PHASEBOOK

***T. Druet*[*], and M. Georges**[*]
[*]University of Liège, Liège, Belgium.

ABSTRACT: Haplotype reconstruction is important in many applications in animal genomics. In livestock species, thanks to the availability of large half-sibs families and genotyped relatives, phasing methods can rely on strong familial information and results in families with more than 10 offspring are very accurate. However, most methods are sensitive to genotyping and map errors which will be more common with next generation sequencing data. Such problems are particularly important when studying recombination rate as we plan to do in the near future. We herein describe a novel algorithm which is robust to genotyping errors and which can identify errors in marker maps. Using a large dairy cattle data set genotyped with high-density genotyping arrays, we show that the novel algorithm strongly reduces the occurrence of spurious cross-overs due to different sources of errors, and identifies map errors for most of the bovine autosomes. The implemented version is still experimental and further research will be conducted to characterize the novel method (including simulations) and to fully describe the identified map errors.

Keywords: haplotype reconstruction, recombination, map errors, genotyping errors

Introduction

In animal genomics, accurate haplotype reconstruction is required in many applications (e.g. Druet and Georges (2010)) including the imputation of missing genotypes, QTL fine-mapping, genomic selection and the analysis of recombination (Sandor et al. (2012)). In livestock species, large half-sibs families and parents are often genotyped and Mendelian and linkage information can be used to accurately reconstruct haplotypes (Druet and Georges (2010)). However, such methods are sensitive to genotyping errors, presence of structural variants (e.g., copy number variants) and to map errors. With the advent of whole genome sequence data, such errors will be more common. Most of these problems will generate spurious crossovers (for instance, a parent that is incorrectly called heterozygote generates spurious double recombinations in the offspring) which have a major impact in studies focusing on the recombination process, whereas fewer consequences are expected in applications such as imputation, QTL fine-mapping or genomic selection.

We herein describe a new haplotyping method which is robust to genotyping and map errors. The new method is implemented in LinkPHASE3 and compared to the former algorithm (LinkPHASE_2.3) described in Druet and Georges (2010). We apply our method on a large dairy cattle population genotyped with high-density SNPs arrays and illustrate that our method can identify genotyping and map errors. After removal of the identified errors, the new method resulted in fewer recombinations - closer to expectation - than with the former method.

Materials and Methods

Data. For the present study, we used individuals from the New-Zealand dairy cattle population (mainly Holstein, Jersey and crossbred individuals). We selected 58,369 individuals genotyped on either Illumina Bovine 50K (v1 and v2) or Illumina BovineHD arrays and kept markers common to the three arrays and mapping to bovine autosomes. After checking parentage errors, we removed markers with a call rate below 95%, generating more than 10 Mendelian inconsistencies, which were fixed or strongly deviating from Hardy-Weinberg proportions ($p < 1e-8$). The final data set contained 37,802 SNPs.

LinkPHASE_2.3 model. The model previously implemented in LinkPHASE (version 2.3) is fully described in Druet and Georges (2010). It works by using Mendelian rules and linkage information. Homozygous SNPs are considered phased de facto. When both parent and offspring were genotyped, heterozygous SNPs of the offspring were first assigned to paternal and maternal homologs based on Mendelian segregation rules. Next, parental phases were completed using linkage information by comparing the likelihood that an allele belongs to the paternal or maternal homolog conditionally on markers already phased and informative in offspring. This second step reconstructs haplotypes in parents that generate few recombination events in the offspring (very similar to a minimum recombination approach). It assigns marker alleles that cosegregate in offspring to the same homolog.

LinkPHASE3 model. The new model performs an additional step after using the same rules as LinkPHASE_2.3. It uses the haplotypes reconstructed by these initial steps as starting values for a hidden Markov model with two hidden states: the two parental homologs. Marker alleles are associated with a certain probability to each of the homologs (emission probabilities) and transition probabilities are equal to recombination probabilities between successive markers. The forward-backward algorithm is then used to compute for each offspring the probability, at each marker position, that it inherited either the paternal or the maternal chromosome of its parents (hereafter called the inheritance vector) based on the marker allele currently assigned to each parental homolog (emission probabilities) and the marker alleles inherited by the progeny. To accommodate genotyping errors, the offspring could have an allele different to the allele labelling the inherited parental haplotype with a probability 1-ε. The inheritance probabilities and the marker alleles inherited by the progeny were then used to estimate new emission probabilities with the Baum-Welch algorithm. In this step, the information from the parent genotype, its Mendelian phasing (based on the genotypes from the corresponding grand-parents) and the genotyping error probabilities were included as priors.

After convergence of the model, we obtain 1) inheritance patterns which are robust to genotyping errors since an offspring can still inherit a parental homolog if it carries an allele different to the one labelling the parental chromosome, 2) probabilities of alleles associated with parental haplotypes (which can be used to identify genotyping errors in the parent), 3) probabilities of recombination (localized between markers where the inheritance of one progeny changes). The identified recombinations can then be used to check the consistency of the map.

We use an EM-algorithm to estimate local recombination rates based on observed recombinations of progeny and the markers flanking the recombination. The chromosome is divided in segments according to the SNP positions. A recombination probability is associated to each segment and these are used to estimate the probability that a specific recombination occurred in a segment. These probabilities are then used to re-estimate the local recombination rates. The process is repeated until convergence.

Results and Discussion

The new algorithm was implemented in LinkPHASE included in the PHASEBOOK package (Druet and Georges (2010)). To test the performances of our new model, we compared the number of identified recombinations on BTA1 for paternal haplotypes of offspring with the new version (LinkPHASE3) and LinkPHASE_2.3 (former version). With LinkPHASE_2.3, the number of recombinations ranged from 3 to 16 recombination per gamete (mean 3.34) whereas with the LinkPHASE3 the values ranged from 2 to 10 (mean 2.10). The improvement is clear and much less spurious recombinations are identified. However, there is still an excess of recombinations per gamete (with some gametes having up to ten recombinations).

We then used recombinations identified with our new HMM and the EM algorithm described in material and methods to compute local recombination rates and check if there were indications of marker map errors on BTA1. Figure 1 represents local recombination rate estimated from the HMM (the local rate of change of the inheritance vector), the local recombination rate estimated with the EM algorithm and also estimated genotyping error rates at either the parental or the offspring level.

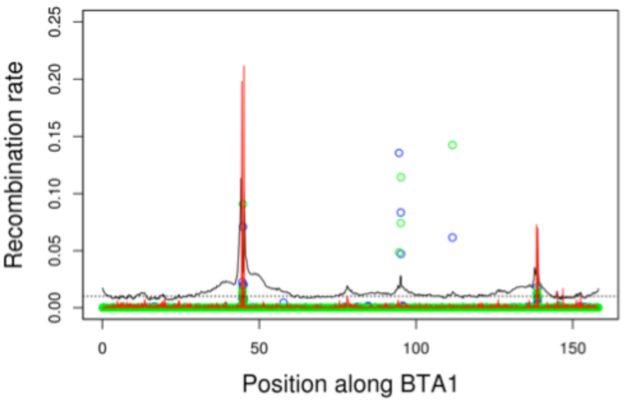

Figure 1: Estimates of local recombinations rates from changes of inheritance in the HMM (black curve), from the implemented EM algorithm (red curve) and estimates of marker genotyping error rates in parents (blue) and offsprings (green).

Based on changes of pattern of inheritance vectors (estimated with the forward-backward algorithm), we observed regions of increased recombination rate but their location remained imprecise. The results from the EM algorithm identify with more precision regions of inflated recombination rate. We clearly spotted two segments (approximately 700 kb and 400 kb) with a high excess of recombinations which are probably due to incorrect positions of these segments in the map. We also observed some clustered SNPs with inflated genotyping error rates (in both parents and offspring). These are also probably due to map errors (or frequent structural variants) but since less markers are involved, our model treats this as genotyping errors rather than (double-)recombinant gametes. Indeed, when marker alleles are incompatible between parental haplotypes and their gamete, the model can either treat this as a genotyping error in either the parent (when several offspring show the incompatibility) or in the offspring or to a double recombination (with an impact on estimated recombination rate). When there is only one incompatible marker allele, a genotyping error is more likely than a double recombination whereas with a stretch of incompatible marker alleles, a double recombination is more likely than multiple genotyping errors.

Applying the same procedure to all the autosomes, we found evidence of map errors on most chromosomes (both small and large segments). After removal of segments and markers associated with high recombination or error rates, estimated recombination rates dropped below 0.01 for all pairs of consecutive markers. On BTA1, exclusion of these positions drastically reduced the average number of crossovers per gamete from 2.07 to 1.30 (Figure 2).

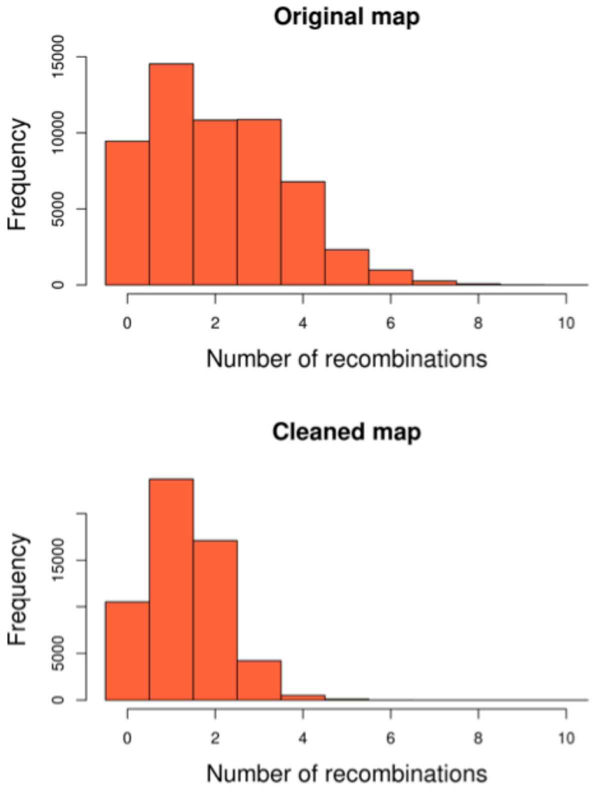

Figure 2: Distribution of number of recombinations per gamete before and after cleaning the marker map on BTA1.

Even with this cleaned-up map, the new algorithm performs better than the former one which is still sensitive to isolated genotyping errors in either parent or offspring. This is illustrated in Figure 3 which compares the inheritance vectors in one family obtained with the two methods for a small chromosomal segment on BTA1. No double recombinants are observed with the new algorithm whereas with the former one, two markers are associated with double recombinants in respectively six and five gametes. Each of those spurious double-recombinants adds two recombinations for these gametes and heavily inflates the recombination rate. As a result, the average number of crossovers per gamete on BTA1 was clearly lower with the new algorithm.

The new algorithm will be intensively tested on simulated and real data sets to fully characterize it. Application on next generation sequencing data sets (sequenced pedigree) is planned. In addition, all map errors will be fully described.

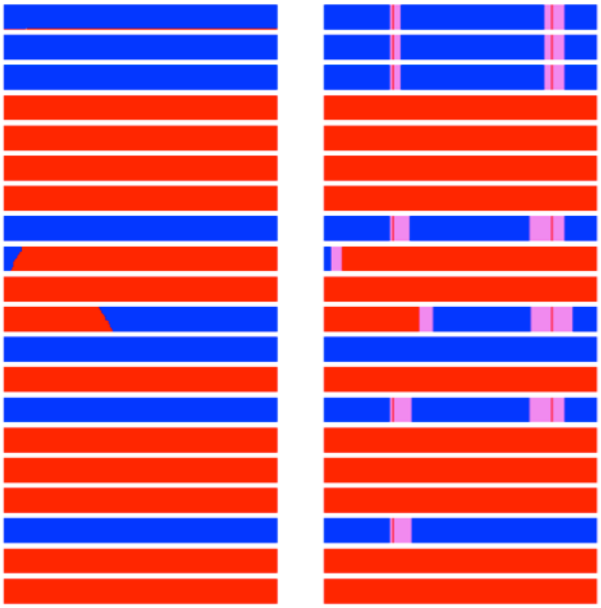

Figure 3: Paternal haplotype inherited (red for grand-paternal, blue for grand-maternal origin and violet for unknown) by 20 offspring (one offspring per line) in one half-sib family. The region is a small segment on BTA1 encompassing 300 SNPs. Origins were estimated either with LinkPHASE3 (on the left) or with LinkPHASE_2.3 (on the right) algorithm.

Conclusions

The new algorithm (LinkPHASE3) that we developed for haplotype reconstruction is more accurate than the previous version (LinkPHASE_2.3). In addition, it is robust to genotyping errors in either parent or offspring. As a result, the number of recombinations per gamete is closer to expectations. When used on large data sets, the algorithm can also identify map errors that result in inflated recombination rates. We observed such errors on many of the bovine autosomes. The new algorithm will be particularly useful in studies on recombination in livestock species.

Acknowledgments

We thank LIC (New Zealand) for providing the data for this study. Tom Druet is Research Associate from the Fonds National de la Recherche Scientifique (Belgium). This study is funded by the European Research Council (ERC-DAMONA).

Literature Cited

Druet, T., and Georges, M. (2010). Genetics, 184:789-798.

Sandor, C., Li, W., Coppieters, W. et al. (2012). PloS Genet., 8(7):e1002854.