Bioinformatics, 31(10), 2015, 1677–1679
doi: 10.1093/bioinformatics/btu859
Advance Access Publication Date: 7 January 2015
Applications Note

Genetics and population analysis

LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors

Tom Druet* **and Michel Georges**

Unit of Animal Genomics, GIGA-R, University of Liège (B34), 1 avenue de l'Hôpital, B-4000, Liège, Belgium

*To whom correspondence should be addressed.

Associate Editor: Gunnar Ratsch

Received on September 12, 2014; revised on November 20, 2014; accepted on December 23, 2014

Abstract

Summary: Many applications in genetics require haplotype reconstruction. We present a phasing program designed for large half-sibs families (as observed in plant and animals) that is robust to genotyping and map errors. We demonstrate that it is more efficient than previous versions and other programs, particularly in the presence of genotyping errors.

Availability and implementation: The software LINKPHASE3 is included in the PHASEBOOK package and can be freely downloaded from www.giga.ulg.ac.be/jcms/prod_381171/software. The package is written in FORTRAN and contains source codes. A manual is provided with the package.

Contact: tom.druet@ulg.ac.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Haplotype reconstruction is useful in many applications including quantitative trait locus (QTL) mapping, whole genome prediction, imputation of missing genotypes, identification of selective sweeps and analysis of the recombination process. In plant and animals, large half-sibs families are common and methods that exploit the corresponding within family linkage information result in accurate haplotype reconstruction. We implemented a method combining such information with Mendelian segregation rules in LinkPHASE, a program included in our PHASEBOOK package (Druet and Georges, 2010). LinkPHASE has been successfully used in many applications including studies of the recombination process in cattle (Sandor, *et al.*, 2012), imputation of missing marker genotypes (e.g. Zhang and Druet 2010), QTL mapping in half-sib families (Karim, *et al.*, 2011) or routine genomic evaluation in dairy cattle (Boichard, *et al.*, 2012). However, the methods implemented in LinkPHASE and related programs are sensitive to genotyping and map errors. These will affect accuracy and power in applications such as QTL mapping, genomic selection or genotype imputation and will inflate the estimated number of crossovers.

We herein describe an extension of LinkPHASE, referred to as LINKPHASE3, based on the implementation of a new haplotyping method which is faster, robust to genotyping errors and can identify map errors.

2 Description

2.1 Implementation

As described in Supplementary Data (part 1), LinkPHASE relies on Mendelian segregation rules (Step I) and on linkage observed in half-sibs (Step II). LINKPHASE3 performs additional steps based on a hidden Markov model (HMM) that refines haplotypes inferred in earlier steps (including correction of genotyping errors). The program is implemented in FORTRAN and a parameter file specifies which steps must be performed (options are described in the manual). When none of the options are used, the program relies solely on Mendelian segregation rules to reconstruct haplotypes. It can also run the LinkPHASE algorithm, apply the new HMM either after StepI or StepII and perform within half-sib family imputation to increase informativity (see Supplementary Data—part 1). This improves haplotype reconstruction for parents with few offspring (five or less) in case none of their own parents is genotyped.

The program generates files with inferred haplotypes with associated inheritance patterns, crossover positions (with flanking informative markers), number of crossovers per individual, putative genotyping errors and local Map Confidence Scores (MCS) combining local information on recombination rate, rate of genotyping errors in the parents, and genotype discrepancies in offspring inheriting the same homologs at each marker position. These local MCS are aimed at detecting putative map errors (see Supplementary Data—part 1).

2.2 Performance

2.2.1 Haplotype reconstruction

To test the program, we simulated datasets with and without genotyping errors and compared it with SHAPEIT2 (Delaneau, *et al.*, 2013) combined with duoHMM (O'Connell, *et al.*, 2014) which has been shown to perform well in pedigreed populations compared with other methods. Comparisons were based on number of switch errors (Stephens and Donnelly, 2003) in parents, power to detect crossovers, false-positive rate (percentage incorrect crossovers) and correlations between numbers of detected and true crossovers. Simulations and results are presented in Supplementary Data (part 2). As expected, application of the HMM performed better only in the presence of genotyping errors. LINKPHASE3 performed better than duoHMM in half-sibs families for all tested statistics. Of note, LINKPHASE3 leaves some markers unphased and does not use linkage disequilibrium information (it is unable to phase in unrelated samples).

More than 99% of the genotyping errors identified by LINKPHASE3 were true genotyping errors (in both parents and progeny).

2.2.2 Computational performance

We ran LINKPHASE3 (compiled with an Intel compiler) on a simulated dataset with 1813 individuals genotyped for 12 407 markers on a cluster with Intel E5649 processors at 2.53 GHz. The run lasted 4m33s and required 244 MB of memory. In comparison, running time and memory requirements were 9m4s and 207 MB (LinkPHASE), 616m17s and 630 MB (SHAPEIT2), 1m47s and 183 MB (duohmm for haplotype correction after SHAPEIT2) and 4m53 and 771 MB (duohmm for crossover detection).

2.3 Detection of map errors

Simulated map errors (see Supplementary Data—part 3) were efficiently detected based either on inflated recombination rates (for large segments), genotyping error rates in parents or our within haplotype entropy measures (for smaller segments). The local MCS combining these three measures clearly distinguished map errors from correct maps.

3 Example

We ran LinkPHASE and LINKPHASE3 on a real dataset from a cattle population with 58 369 genotyped individuals (Supplementary Data—part 4). Haplotype reconstruction on BTA1 (2507 SNPs) lasted 40m21s and required 852 MB of memory with LINKPHASE3 and 985m55s and 681 MB with LinkPHASE. LINPKHASE3 generated fewer spurious crossovers than LinkPHASE (examples of spurious crossovers in an halfsib family are illustrated in Figure 1). The median number of crossovers per gamete decreased from 35 (LinkPHASE) to 27 (LINKPHASE3) whereas the maximal value dropped from 108 CO to 56 CO despite the presence of map errors. Indeed, we identified 46 putative map errors on UMD 3.1 (Supplementary Data—part 4). For all the identified segments, we found evidence that these were indeed map errors (by mapping these segments on the genome based on correlation between inheritance vectors and using available information from a 5 kb mate-pair sequencing library).

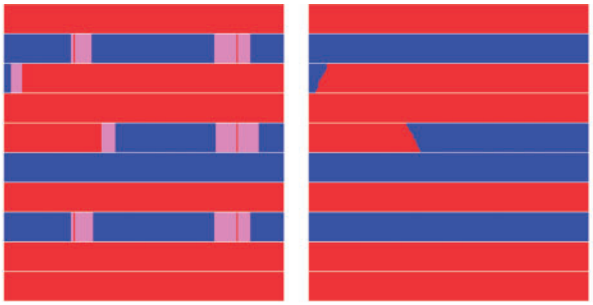

Fig. 1. Paternal haplotype inherited (distinct colours represent grand-parental and grand-maternal origins whereas an intermediate tone is used for unknown origin) by 10 offspring (one offspring per line) in one half-sib family. The region is a small segment on BTA1 encompassing 300 SNPs. Inheritance patterns were estimated either with LinkPHASE (on the left) or with LINKPHASE3 (on the right) algorithm

4 Conclusions

LINKPHASE3 is more robust to genotyping and map errors when compared with LinkPHASE. As a result, identification of crossovers, haplotype reconstruction and estimation of inheritance patterns is more accurate, which should positively affect all potential applications. These include studies of the recombination process, estimation of identity-by-descent probabilities between sibs for QTL or disease mapping and for the estimation of mutation rate, improved phasing for missing marker imputation, haplotype-based QTL mapping or genomic predictions of complex traits. In addition, the program identifies some genotyping errors and putative map errors. This is particularly important for species for which the reference genome still requires improvement. We also illustrated that we can use the inheritance vectors to find the correct position in the genome of misplaced chromosomal segments. Finally, the speed improvement makes the program more compatible with routine phasing applications on large data sets, such as those resulting from the massive application of genomic selection in livestock species.

Acknowledgements

The authors are grateful to Livestock Improvement Company—LIC (New Zealand) for providing data for this study. Tom Druet is a Research Associate from the FRS-FNRS.

Funding

The project was funded by the European Research Council via the ERC-AdG-GA323030 (DAMONA). We used the supercomputing facilities of the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the Fonds de la Recherche Scientifique de Belgique (FRS-FNRS).

Conflict of Interest: none declared.

References

Boichard,D. *et al.* (2012) Genomic selection in French dairy cattle. *Anim. Prod. Sci*, **52**, 115–120.

Delaneau,O., Zagury,J.F. and Marchini,J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5–6.

Druet,T. and Georges,M. (2010) A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, **184**, 789–798.

Karim,L. *et al.* (2011) Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nat. Genet.*, **43**, 405–413.

O'Connell,J. *et al.* (2014) A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.*, **10**, e1004234.

Sandor,C. *et al.* (2012) Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genet.*, **8**, e1002854.

Stephens,M. and Donnelly,P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Zhang,Z. and Druet,T. (2010) Marker imputation with low-density marker panels in Dutch Holstein cattle. *J. Dairy Sci.*, **93**, 5487–5494.