SPEARS: Standard Performance Evaluation of Ancestral Reconstruction through Simulation

H. Manching¹, and R. J. Wisser¹,²

¹Department of Plant & Soil Sciences, University of Delaware, Newark, DE, 19716 USA

Motivation: Ancestral haplotype maps provide useful information about genomic variation and biological processes. Reconstructing the descendent haplotype structure of homologous chromosomes, particularly for large numbers of individuals, can help with characterizing the recombination landscape, elucidating genotype-to-phenotype relationships, improving genomic predictions and more. Inferring haplotype maps from sparse genotype data is an efficient approach to whole-genome haplotyping, but this is a non-trivial problem. A standardized approach is needed to validate whether haplotype reconstruction software, conceived population designs and existing data for a given population provides accurate haplotype information for further inference.

Results: We introduce SPEARS, a pipeline for whole simulation-based appraisal of genome-wide ancestral haplotype inference. The pipeline generates virtual genotypes (truth data) with real-world missing data structure. It then proceeds to mimic analysis in practice, capturing sources of error due to imputation and reconstruction of ancestral haplotypes. Standard metrics allow researchers to assess which features of haplotype structure or regions of the genome are sufficiently accurate for analysis and reporting. Haplotype maps for 1,000 outcross progeny from a multi-parent population of maize is used to demonstrate SPEARS.

Availability: https://github.com/maizeatlas/spears

Introduction

The genome is a mosaic of ancestral haplotypes that capture the evolutionary or breeding history of an individual. Reconstructing ancestral blocks is important for imputing untyped regions of the genome (1), mapping quantitative trait loci (2), investigating the recombination landscape (3), and inferring the evolutionary history and structure of haplotypes (4, 5). Accurately inferring ancestral haplotypes is non-trivial, but several approaches for this have been developed: MERLIN (6); HAPPY (2); GAIN (7); DOQTL (8); R/qtl2 (9). These tools have been shown to perform well, but are mostly limited to specific population types or breeding schemes and can be computationally intensive with complex pedigrees or large numbers of markers.

RABBIT (Reconstructing Ancestry Blocks BIT by bit) is a flexible tool that uses a Markovian model to reconstruct haplotypes for complex pedigrees involving various mating scenarios (10). Compared to current tools, it has shown the highest accuracy for inbred lines from multiparent pedigrees (including the mouse CC design [11] and Arabidopsis thaliana MAGIC design [12]) (10). Here, tailored for RABBIT but extensible for other tools, we present a standard performance evaluation of ancestral reconstruction through simulation (SPEARS) pipeline. SPEARS is designed to determine expectations for the accuracy of haplotype reconstruction software and different population designs. As proof-of-concept and a new demonstration of RABBIT, we develop a detailed picture on the variation in accuracy for genome-wide haplotype maps for progeny from a multi-parent population of maize.

Materials and methods

A. Simulation Data. This study introduces SPEARS (Figure S1) which incorporates SAEGUS (https://github.com/maizeatlas/saegus) as the genome simulator. Using SAEGUS, we first generated a virtual multiparent outcross population of 1,000 progeny (Figure S2). Real genotype data on inbred line parents of the population was used to initiate the simulation (n = 47,078 markers obtained from both genotyping-by-sequencing (GBS) [13] and the MaizeSNP50 BeadChip [14]; filtered to remove residual heterozygous sites in the parents for compatibility with MaCH imputation [15]). We also examined the impact of using fewer markers (n = 23,584 GBS markers).

In order to represent analysis in practice, the level of missing data per marker observed in a real maize population (not shown) was reproduced for the virtual progeny. That is, at each marker, virtual genotype scores were randomly converted to missing according to the missing data rate observed at the corresponding marker. These missing data were then imputed using MaCH (15) and filtered to exclude markers with imputation accuracy < 0.5, resulting in 47,074 markers (or 23,584 GBS markers) being retained for RABBIT.

B. Reconstruction of Ancestral Haplotypes. To build haplotype maps for the simulated population (Figure S3) the joint model of RABBIT was used to assign an optimal Viterbi path using the origViterbiDecoding algorithm (10). Custom scripts in R were used to reformat output from RABBIT into phased genotypes for SPEARS.

C. Evaluation of Ancestral Reconstruction. Based on comparing simulated and inferred data per individual,
SPEARS uses four metrics to assess distinct features of haplotype maps (Figure S4): (i) Ancestral Assignment Accuracy (AAA); (ii) Genotype Assignment Accuracy (GAA); (iii) Switch-Error Rate (SER); and (iv) Correlation between Crossover Counts (CCC). AAA is calculated as the proportion of markers that have the correct inferred parent assigned on each homologue (given as a percentage). To calculate GAA, genotypes are assigned based on the inferred ancestry and corresponding parental genotype data used as input, from which the proportion of genotype matches are calculated (given as a percentage). To assess phasing accuracy, SER is calculated only among markers with correctly inferred genotype scores (phasing cannot be assessed for markers with genotyping errors) and is defined as: \[ \frac{1-\left(n^{-1}\cdot sw\right)}{(n-1)} \] where \( n \) is the number of heterozygous sites and \( sw \) is the number of switches required to obtain the true phase based on known data. The CCC is calculated as the correlation coefficient for the total number of crossovers across both homologues of all chromosomes per individual. For additional analysis, we also calculated “parent certainty” as the difference between the posterior probabilities of the two most likely parents at each marker (obtained from the original DeCoding algorithm within RABBIT).

Results

SPEARS metrics are reported in Table S1. The average per sample AAA was 97.0%. Regions with lower accuracy showed decreases in parent certainty alone or in combination with a lower density of markers (Figure S5), indicating that both identity-by-state among parents and marker density, but neither RABBIT nor the Viterbi algorithm per se were main sources of error in reconstructing haplotype maps. Additionally, the average GAA (99.3%) was higher than AAA (Table S1), suggesting shared parental haplotypes were indeed a contributing factor to error. RABBIT performed very well in phasing among markers that had correctly assigned genotypes with an average SER of 0.0090 across all samples. The CCC was positive and showed high correlation between known and RABBIT-inferred crossover counts (\( r = 0.89, p < 2.2e^{-16} \)) (Figure S6). However, the number of crossover counts per sample was downward biased with an average of \( 260 \pm 16 \) versus \( 228 \pm 14 \) for known and RABBIT data, respectively.

Conclusions

Reconstruction of ancestral haplotypes from genomic data is useful for a number of applications. SPEARS showed that RABBIT produced highly accurate genome-wide haplotype maps for an admixed non-inbred population. Genomic segments with minimal differentiation between the parents resulted in localized reductions in the accuracy of ancestral assignment (Figure S5). This may be overcome with a higher marker density unless parental haplotypes are truly identical-by-descent. For this study, using approximately half the number of markers had minimal influence on the average genome-wide AAA, GAA, CCC, and SER (Table S1); some local regions showed slight reductions in accuracy (Figure S7). This exemplifies how SPEARS estimates of the expectation for different types of error can help guide investigations on haplotype structure. It allows analysis of certain features of haplotype data to be included/excluded in a study based on a corresponding metric, and the expectations can be reported. Moreover, one can assess whether specific chromosomes or regions of the genome, but not others, are sufficiently accurate for downstream analysis. SPEARS, the protocol and suite of scripts, has been made publicly available at https://github.com/maizeatlas/spears.

ACKNOWLEDGEMENTS

We thank Dr. Chaozhi Zheng for support in operating RABBIT.

Funding

This work was supported by the Agriculture and Food Research Initiative Competitive Grant (grant no. 2011-67003-30342); and the Agriculture and Food Research Initiative Fellowships Grant Program (grant no. 2018-67011-28052) from the United States Department of Agriculture National Institute of Food and Agriculture.

Bibliography

1. Robert W Davies, Jonathan Flint, Simon Myers, and Richard Mott. Rapid genotype imputation from sequence without reference panels. Nature genetics, 48(8):965, 2016.
2. Richard Mott, Christopher J Talbot, Maria G Turri, Allan C Collins, and Jonathan Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. Proceedings of the National Academy of Sciences, 97(23):12649–12654, 2000.
3. Andrew P Morgan, Daniel M Gatti, Maya L Najarian, Thomas M Keane, Raymond J Galante, Allan I Pack, Richard Mott, Gary A Churchill, and Fernando Pardo-Manuel de Villena. Structural variation shapes the landscape of recombination in mouse. Genetics, 206(2):603–619, 2017.
4. Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. The structure of haplotype blocks in the human genome. Science, 296(5576):2225–2229, 2002.
5. David L Aytor, William Valdar, Wendy Foulds-Mathes, Ryan J Buus, Ricardo A Verdugo, Ralph S Baric, Martin T Ferris, Jeff A Frelinger, Mark Heise, Matt B Friedman, et al. Genetic analysis of complex traits in the emerging collaborative cross. Genome research, 21(8):1213–1222, 2011.
6. Gonçalo R Abecasis, Stacey S Chen, William O Cookson, and Lon R Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics, 30(1):97–101, 2002.
7. Eric Yi Liu, Q stature, Leonard McMillin, Fernando Pardo-Manuel De Villena, and Wei Wang. Efficient genome ancestry inference in complex pedigrees with inbreeding. Bioinformatics, 26(12):1199–1207, 2010.
8. Daniel M Gatti, Karen L Svenson, Andrey Shabalin, Long-Yang Wu, William Valdar, Petr Simecek, Neal Goodwin, Riyan Cheng, Daniel Pomp, Abraham Palmer, et al. Quantitative trait locus mapping methods for diversity outbred mice. G3: Genes, Genomes, Genetics, 4(9):1623–1633, 2014.
9. Karl W Broman, Daniel M Gatti, Petr Simecek, Nicholas A Furlotte, Pjotr Prins, Saunak Sen, Brian S Yandell, and Gary A Churchill. R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics, 211(2):495–502, 2019.

10. Chaozhi Zheng, Martin P Boer, and Fred A van Eeuwijk. Reconstruction of genome ancestry blocks in multiparental populations. Genetics, 200(4):1073–1087, 2015.

11. Gary A Churchill, David C Airey, Hooman Alayee, Joe M Angel, Alan D Attie, Jackson Beatty, William D Beavis, John K Belknap, Wade Berrettini, et al. The collaborative cross, a community resource for the genetic analysis of complex traits. Nature genetics, 36(11):1133, 2004.

12. Paula X Kover, William Valdar, Joseph Trakalo, Nora Scarcelli, Ian M Ehrenreich, Michael D Purugganan, Caroline Durrant, and Richard Mott. A multiparent advanced generation intercross to fine-map quantitative traits in arabidopsis thaliana. PLoS genetics, 5(7), 2009.

13. Heather Manching, Subhajit Sengupta, Keith R Hopper, Shawn W Polson, Yuan Ji, and Randall J Wisser. Phased genotyping by sequencing enhances analysis of genetic diversity and reveals divergent copy number variants in maize. G3: Genes, Genomes, Genetics, 7(7):2161–2170, 2017.

14. Martin W Ganal, Gregor Durstewitz, Andreas Polley, Aurélie Bérard, Edward S Buckler, Alain Charcosset, Joseph D Clarke, Eve-Maria Graner, Mark Hansen, Johann Joets, et al. A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PloS one, 6(12), 2011.

15. Yun Li, Cristen J Willer, Jun Ding, Paul Scheet, and Gonçalo R Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology, 34(8):816–834, 2010.

16. Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. The American Journal of Human Genetics, 73(5):1162–1169, 2003.

17. Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shih Lin, Zhaohua Qin, Heather M Munroe, Gonçalo R Abecasis, et al. A comparison of phasing algorithms for trios and unrelated individuals. The American Journal of Human Genetics, 78(3):437–450, 2006.