A Simulation Study of Genotype Phasing Software

Dent A. Earl[1], David Haussler[1,2,3]

1. Bioinformatics Graduate program, UCSC 2. Center for Biomolecular Science and Engineering, UCSC, 3. Howard Hughes Medical Institute

BIOINFORMATICS
BIOMOLECULAR ENGINEERING DEPT

UC SANTA CRUZ

Baskin Engineering
UC SANTA CRUZ

Abstract

Here we report the results of a simulation study of three genotype phasing software packages. The three were tested on ten simulated data sets built using an identical pedigree structure, a high resolution recombination map, and real genotypes taken from HapMap III data for two populations: Utah residents with ancestry from northern and western Europe (CEU, 88 individuals) and Yoruba in Ibadan, Nigerian (YRI, 100 individuals).

The most surprising result shown here is that though none of the methods tested were provided with information on the underlying pedigree that gave rise to the data, they were all affected indirectly by the structure: the algorithms performed better on individuals that were more related to everyone else in the data set (their expected number of identical genotypes in the data was large).

Introduction

Diploid organisms have, by definition, homologous pairs of chromosomes. An individual's genotype is a combination of the individual's underlying genetic haplotypes, a haplotype being the sequence of alleles that are physically connected to one another along one chromosome. Without labor, and cost intensive methods (cloning and sequencing), haplotypes are not directly observable, and instead must by inferred through a process called ***phasing***.

Phasing has experienced a large increase in attention in recent years, likely due to the boom in single nucleotide polymorphism (SNP) microarray market penetration. SNP microarrays (SNP chips) allow for the observation of an individual's genotype at predefined loci. Unfortunately SNP genotypes contain no direct information about the sequence of the observed alleles, information which is crucial for disease, population genetic, and evolutionary studies.

Two endeavors rely on haplotype information: clinical assays and evolutionary reconstruction. In a clinical assay for a cis-acting genetic disease a clinician would like to determine what the chromosome-local pattern of inherited alleles is for a patient.

For evolutionary reconstruction the goal of phasing genotype data into haplotypes is to understand and record the pattern of transmission of alleles through individuals over generations.

These endeavors are not mutually exclusive, but complementary. For example, knowing the genetic ancestry of an individual could also help to inform the clinician about the potential efficacy of certain drugs. And knowing the underlying molecular genetics and epidemiological genetics of a particular region of the genome could be illuminating in reconstructing the evolutionary history of that region.

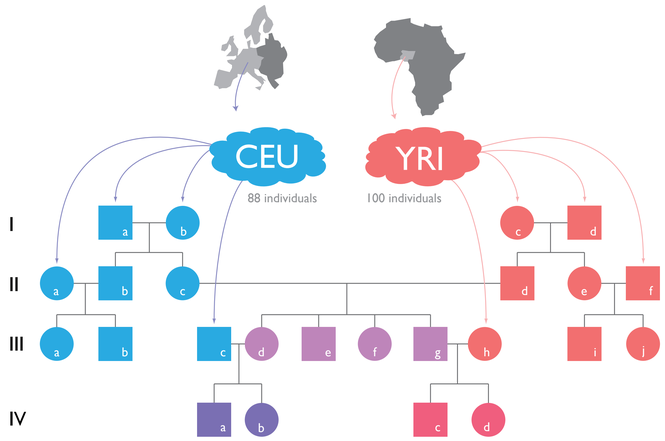

Figure 1

Pedigree showing experimental design. Colored clouds represent source populations and colored curves are indicative of founding individuals in the pedigree. Generation number is shown in roman numerals on the left. Individuals are referred to as in III-a, or IV-c.

Methods

Pedigree

The structure of the pedigree that we simulate, shown in Figure 1, was chosen in an attempt to have a gradient of difficulty of phasing problems, from the easy when an individual has one chromosome from one population and one from another, to the more difficult where the individual's chromosomes come from the same population. With this in mind I created a symmetrical pedigree with founding haplotypes taken directly from the source populations, admixture between descendants of the different populations, and with exogamous individuals from the ancestral populations breeding into the descendant family.

Founders

Haplotypes are chosen at random without replacement to make up founding members of the pedigree and subsequent exogamous individuals. Founders for the CEU side of the family are I-a and I-b and for the YRI side of the family, I-c and I-d. The subsequent exogamous individuals in the CEU side are II-a and III-c and for the YRI side are II-f and III-h.

A large pool of software with varying abilities has been released over the past decade, but there is no standard methodology for phasing and no clear gold standard algorithm at the moment.

Following the simulation of genotype data, I test the major phasing software packages as to their abilities to resolve the simulated genotype data.

Software

fastPHASE, published in 2006, was developed as a compromise between accuracy and speed. fastPHASE does this by omitting the MCMC scheme of PHASE in favor of a hidden Markov model (HMM). As the authors note, however, the price paid here is that the HMM incorporates no information about demographics, or evolutionary processes.

Beagle, published in 2007, uses a directed acyclic graph (DAG) to model localized haplotype-clusters and then an HMM to find the most likely haplotype pair, conditional on an individual's genotype. The algorithm allows the use of known haplotypes in a panel, which is used to populate the paths of the HMM in the localized haplotype-cluster. The final step of the algorithm is to use the Viterbi algorithm to find the most-likely haplotype pair for each individual.

2SNP, published in 2008, when presented with trio data (father, mother and child) uses a combination of approaches to phasing. It enforces simple patterns of Mendelian inheritance where possible, and elsewhere uses a graph structure approach and a maximum spanning tree based on the population genetic metrics of linkage disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE) to assign haplotypes. Essentially, the algorithm constrains the problem based on the known demographic structure and focuses solely on complex positions.

Switch Error

In order to quantitively assess the performance of the algorithms I implemented the metric switch error, which is one minus the switch accuracy of Lin et al. (2004). Given the correct (True) phase for a particular segment of a chromosome, and a solution given by a phasing algorithm, the switch error is calculated to be the number of times the phase pattern of adjacent alleles is not the same in the test as in the True, divided by the total number of heterozygous positions minus one. Switch error ranges from 1 to 0.

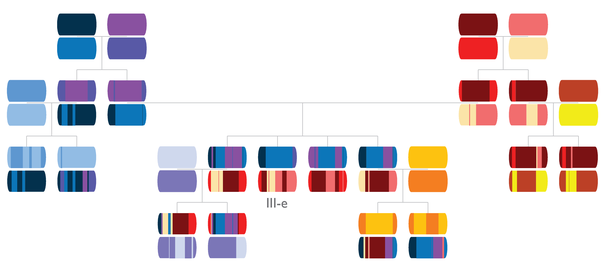

Figure 2

Ten simulated pedigrees. Each individual is comprised of a pair of stacked colored sausage shapes (each representing a copy of chromosome 1 with 120,000 SNPs). Different colors represent different haplotype blocks with warm colors (reds, oranges) indicating YRI population haplotypes and cool colors (blues, purples) indicating CEU population haplotypes. Offspring of matings recieve one recombined chromosome from each parent. The first simulation is shown at larger size for detail.

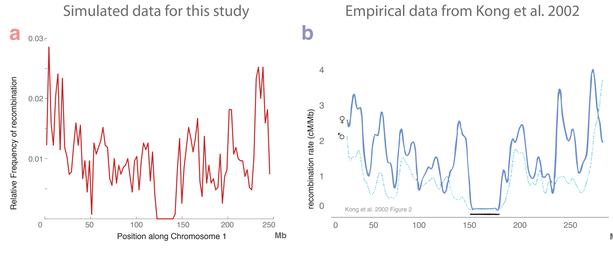

Figure 3

Comparison between simulated gametes and empirical data from the literature. Part (a) shows the relative frequency of recombination events along chromosome 1 for 1,000 simulated gametes. Part (b) shows Figure 2 from Kong et al. 2002, empirically measured recombination rates in centiMorgans per megabase (cM/Mb) for males and females.

That the simulated data appears similar to the data from the literature is indicative of the accuracy of the recombination rate data used in the simulation.

Figure 4

Haplotype line plots representing the true and inferred states of a simulated individual's genotype (individual III-e from the first sumilation shown in Figure 2). Each line shows the status of 37,000 SNPs covering a physical region of 67 megabases (Mb). The small ideogram to the right shows where on the chromosome the represented data originates.

Each line is made up of cyan lines above and blue lines below a central gray line. These colored lines indicate heterozygous SNPs and their placement on the top (cyan) or bottom (blue) area corresponds to their origin on the maternal or paternal chromosome. Homozygous positions are not drawn because they add no information to the plot. A "switch error" occurs when a mistake is made as to which chromosome the allele is placed on. Switch errors are marked by red circles.

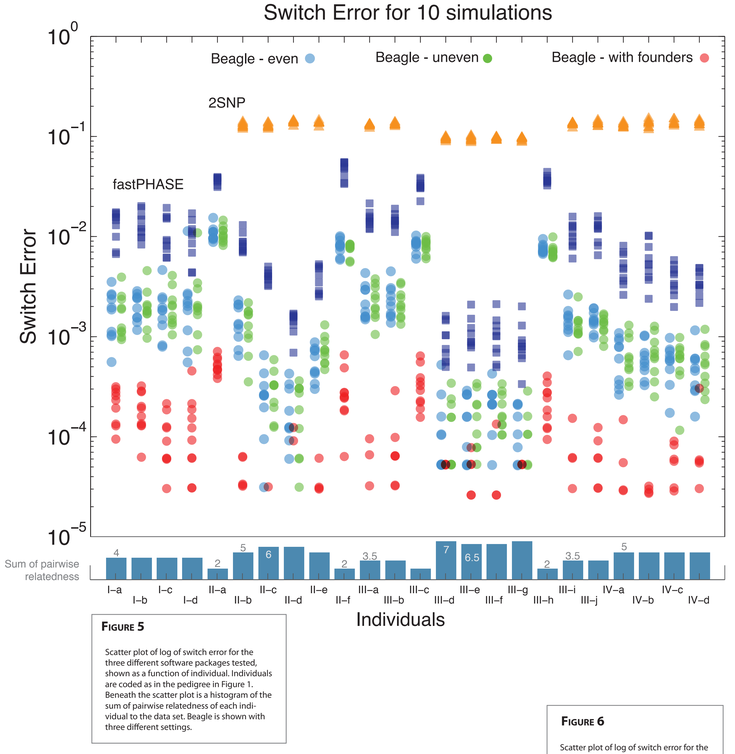

Figure 5

Scatter plot of log of switch error for the three different software packages tested, shown as a function of individual. Individuals are coded as in the pedigree in Figure 1. Beneath the scatter plot is a histogram of the sum of pairwise relatedness of each individual to the data set. Beagle is shown with three different settings.

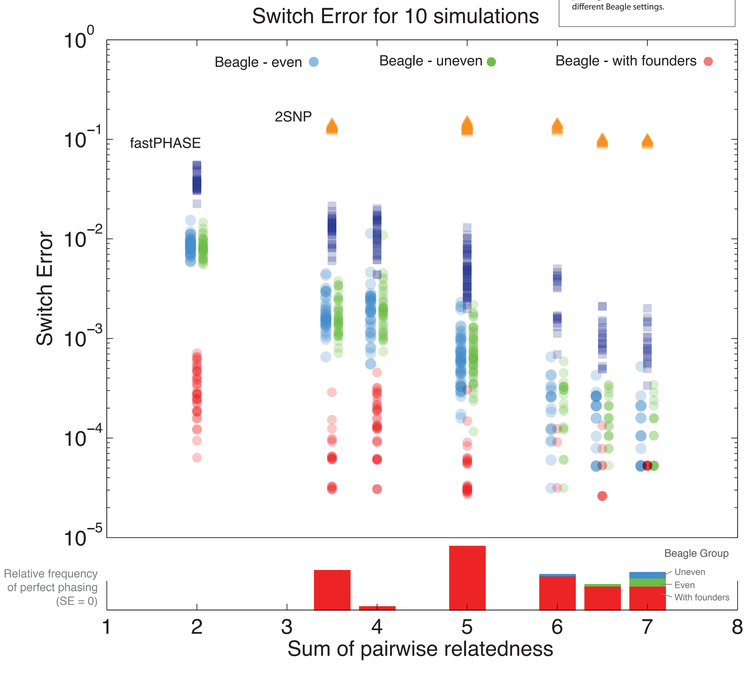

Figure 6

Scatter plot of log of switch error for the three different software packages tested, shown as a function of the individual's sum of pairwise relatedness to the data set. Beneath the scatter plot is a histogram showing the relative frequency of perfect phasings (switch error = 0) for the three different Beagle settings.

Results

Tests

All algorithms were run on the same server powered by dual, quad-core processors (eight cores total, Intel Xeon 2.83GHz) with 16 gigs of RAM.

fastPHASE

fastPHASE showed consistently good performance, with switch errors ranging between 4.927e-04 and 4.76e-02. While the authors claim the software can make use of a panel of known haplotypes to improve phasing I was unable to get the feature to work without crashing due to memory allocation issues (segmentation faults). Running without panel information the algorithm took approximately seven and a half hours for each replicate to run and performed very well, easily besting the switch error scores of the 200 times faster 2SNP.

I note here that the large number of markers in the present study (approximately 120,000) does not seem to have adversely affected fastPHASE. The small individual sample size, 24, could also work against the algorithm, as it uses the presented data set in part as a template for finding the phasing solutions. However, fastPHASE performed well even with the large number of markers and relatively small number of individuals.

Beagle

Beagle had the best performance of the algorithms, solving a few haplotypes with no errors (perfect phasing) and generally running switch errors below 1 in 10,000 to 1 in 1,000 range. Beagle does not take in any information about the pedigree and the success it enjoyed in this data set is likely due to the prevalence of similar haplotypes to the founder haplotypes being in the panel. Beagle took approximately an hour and a half to run.

Beagle Subsets

In order to determine how much of a role Beagle's library played in its performance, I tested three library types: uneven, where the library was made up of 168 CEU haplotypes and 192 YRI haplotypes; even, where the library was made up of 168 CEU haplotypes and 168 (randomly chosen) YRI haplotypes; and with founders, where the libarry was made up of 176 CEU and 200 YRI haplotypes, including the founding haplotypes (the correct solutions).

2SNP

2SNP, arguably the simplest package evaluated in this study was also the poorest performer with switch errors consistently in the range of 1 in 10. 2SNP is limited in that it does not attempt to predict the true phase of the parents in a trio but will simply predict the transmitted and un-transmitted chromosomes to the child. This is why there are missing data in the 2SNP column in Table switchError. 2SNP was remarkably efficient in terms of time taken, completing work after an average of 130 seconds.

Conclusions

We presented here the results of 10 independent simulations of a pedigree of 24 individuals over four generations, tracking 120,000 SNPs over chromosome 1.

The most exciting results from this study are the instances of near perfect haplotype reconstruction carried out by Beagle, and the apparent increase in phasing accuracy in all software when analyzing individuals more related to the data set than individuals less related to the data set. It was unexpected that any of the software tested would be able to solve the phasing correctly, but Beagle returned at leaste one perfect phasing under all subsets of its library that were tested.

The increased accuracy relative to relatedness effect appears to have some impact on 2SNP, which is unexpected because 2SNP does not consider the entire data set at once but instead runs on trios (mother father child). A more likely explanation for the slight increase in accuracy seen in 2SNP with the increase of relatedness is that the individuals with the greatest relatedness values are the ones in the center of the pedigree with parents from two different ancestral populations. Without further testing it seems compelling that the slight increase in accuracy for for 2SNP is due to the extreme difference between the mother and father haplotypes for these individuals. To test for this a second experiment would need to be run with an altered pedigree in which an individual was mated into the family from the opposite ancestral population (e.g. a new coupling between a CEU individual and III-j, or a YRI individual and III-b).

The strong downward trend of Beagle and fastPHASE with increasing relatedness seems to be best explained by the fact that both models rely on HMMs and both use the entire data set to aid in phasing individuals. Thus, more closely related individuals would have access to better templates for finding the best solution. The increased accuracy in Beagle relative to fastPHASE seems best explained by its incorporation of the haplotype library, as evidenced by the order of magnitude increase in accuracy when the original founder haplotypes are included in the haplotype library. We note that difference in switch error between the uneven and even sized library is negligible.

Another unexpected result was the sheer speed of 2SNP. The program phased the children of the trio subsets in only 134 seconds. This raw speed seems to be achieved by parsing the trios by simple Mendelian determinism (which could be implemented through boolean logic) and then to applying a maximum spanning tree to whatever is left over. This second part of the algorithm is likely the source of 2SNP's errors, and is would be an area worth exploring competing methodologies. 2SNP had a switch error ranging from 0.086 to 0.14, meaning it made mistakes on up to 14% of the data. Despite the raw speed of the algorithm, this error rate is quite high.

fastPHASE performed quite well. The program had switch errors ranging from 4.9e-04 to 4.8e-02, with the lowest number of errors occurring in the admixed offspring and the largest errors occurring in the founding individuals. The program took seven and a half hours to run on average, which was the longest amount of time.

[1] Gonc, alo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet, 30(1):97–101, Jan 2002.

[2] Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet, 73(5):1162–9, Nov 2003.

[3] G Greengrass and D Geiger. High density linkage disequilibrium mapping using models of haplotype block variation. Bioinformatics, 20 Suppl 1:i137–44, Aug 2004.

[4] Gad Kimmel and Ron Shamir. Gerbil: Genotype resolution and block identification using likelihood. Proc Natl Acad Sci USA, 102(1):158–62, Jan 2005.

[5] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet, 78(4):629–44, Apr 2006.

[6] Tom Becker, Ruta Valentonyte, Peter J P Croucher, Konstantin Strauch, Stefan Schreiber, Jochen Hampe, and Michael Knapp. Identification of probable genotyping errors by consideration of haplotypes. Eur J Hum Genet, 14(4):450–8, Apr 2006.

[7] H K Goring and R T Lie. Case-parent trios: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. Ann Hum Genet, 70(Pt 3):382–96, May 2006.

[8] Lauri Eronen, Floris Geerts, and Hannu Toivonen. Haploree: efficient and accurate large-scale reconstruction of apletypes. BMC Bioinformatics, 7:542, Jan 2006.

[9] Dumitru Brinza and Alexander Zelikovsky. 2snp: scalable phasing method for trios and unrelated individuals. EEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM, 5(2):313–8, Jan 2008.

[10] Brian L Browning and Sharon R Browning. Haplotypic analysis of wellcome trust case control consortium ata. Hum Genet, 123(3):273–80, Apr 2008.

[11] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference or whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet, 1(5):1084–97, Nov 2007.

[12] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next ereration of genome-wide association studies. In Presented at the annual meeting of The American Society of uman Genetics, Philadelphia, Pennsylvania, Nov 2008. The American Society of Human Genetics, ttp://www.ashg.org/2008meeting/abstracts/fulltext/f22281.htm.

[13] E.S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. Proceedings of the aional Academy of Sciences USA, 84:2363–2367, 1987.

[14] Jonathan Marchini, David Cutler, Nick Patterson, Matthew Stephens, Eleazar Eskin, Eran Halperin, Shin Lin, haohui S Qin, Heather M Munro, Goncalo R Abecasis, Peter Donnelly, and International HapMap consortium. A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet, 8(3):437–50, Mar 2006.

[15] A Kong, G Masson, M Frigge, A Gylfason, P Zusmanovich, G Thorleifsson, P Olason, A Ingason, S Steinberg, Rafnar, P Sulem, M Mouy, F Jonsson, U Thorsteinsdottir, D Gudbjartsson, H Stefansson, and K Stefansson. etection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet, Aug 2008.

[16] The International HapMap Consortium. HapMap Phase III. In Press, 2009.

[17] The International HapMap Consortium. Hapmap3 phasing data., Jan 2009. ttp://www.hapmap.org/downloads/index.html.en.

[18] C. Dib, S. Faur´, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach. A comprehensive genetic map of the human genome ased on 5,264 microsatellites. Nature, 380(6571):152–4, Mar 1996.

[19] K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, and J. L. Weber. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet, 63(3):861–9, Sep 1998.

[20] Augustine Kong, Daniel F Gudbjartsson, Jesus Sainz, Gudrun M Jonsdottir, Sigurjon A Gudjonsson, Bjorgvin srhardsson, Sigrun Sigurdardottir, John Barnard, Bjorn Hallbeck, Gisli Masson, Adam Shlien, Stefan T alason, Michael L Frigge, Thorgeir E Thorgeirsson, Jeffrey R Gulcher, and Kari Stefansson. A high-resolution aombination map of the human genome. Nat Genet, 31(3):241–7, Jul 2002.

[21] Gil McVean, Philip Awadalla, and Paul Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics, 160(3):1231–41, Mar 2002. 7

[22] Loukas Moutsianas and Gil McVean. Recombination rates in the human genome., Nov 2008. ttp://ftp.hapmap.org/phasing/2008-10 phasefll/phased/README.txt.

[23] J.B.S. Haldane. The combination of linkage values, and the calculation of distances between th eloci of linked acters. Journal of Genetics, 8:299–309, 1919.

[24] Shin Lin, Aravinda Chakravarti, and David J Cutler. Exhaustive allelic transmission disequilibrium tests as a ew approach to genome-wide association studies. Nat Genet, 36(11):1181–8, Nov 2004.

[25] Augustine Kong. personal communication, 2009.

http://users.soe.ucsc.edu/~dearl/posters/ dearl@soe.ucsc.edu