An overview of phasing methods for polyploids

Soumya Ranganathan[1], Fabian Grandke[1,2], Dirk Metzler[1]

[1]Department of Biology, Ludwigs-Maximilians-University

[2]Genetwister Technologies, Wageningen, The Netherlands

soumya@bio.lmu.de

1. Introduction

Polyploidy is common in most of the commercially significant plant species like Potato(4n) and strawberries(8n). There are very limited number of tools to analyse the polyploid data, especially those which perform the phasing. Haplotype phasing plays an important role in finer localisation of causative mutations[1].

Genotype	Haplotype	Diagnosis
1. (G,C),(G,A)	G-A & C-G	NO disease
2. (G,C),(G,A)	**G-G** & C-A	Disease
3. (G,C),(G,G)	**G-G** & C-G	Disease

As experimental methods to infer haplotypes are expensive, we rely on analytical methods. We compare two polyploid phasing tools with different approaches, statistical method called 'Polyhap'[2] and a combinatorial method, 'SATlotyper' [4]. We evaluate their performance on simulated data sets which were generated using PedigreeSim [5].

2. SATlotyper

This tool is implemented on the principle of pure Parsimony and extends the ideas of SAT model [7] which is described for diploid phasing. The goal is to find a set of common haplotypes, explaining each unphased genotype present in the given population.

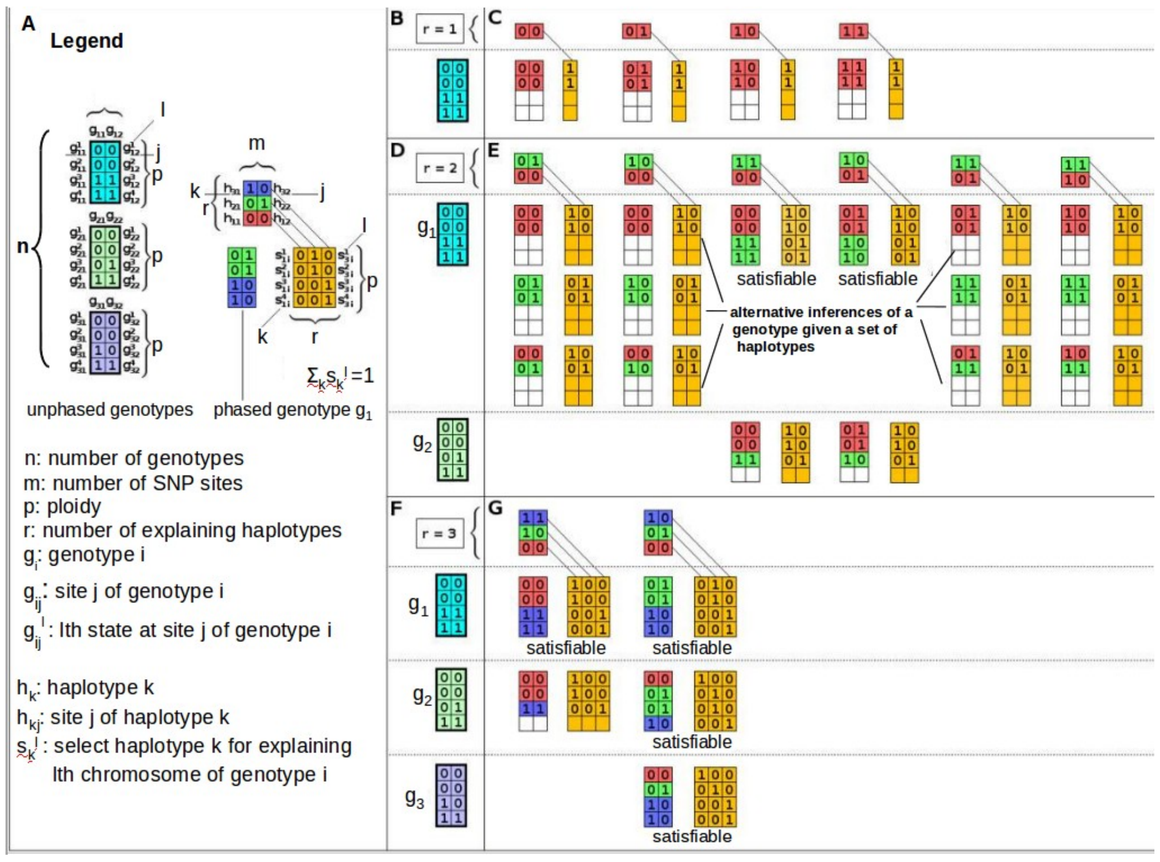
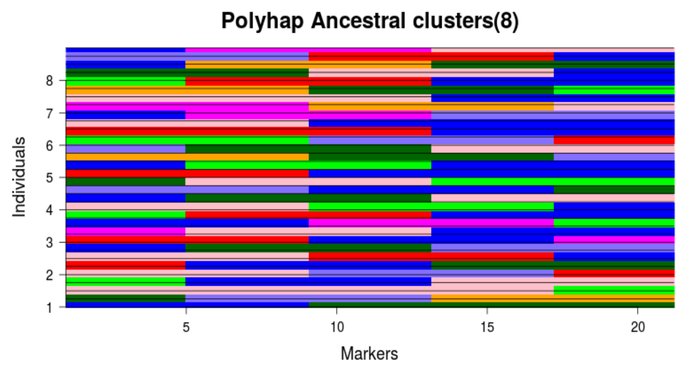


Figure depicting phasing mechanism of SATlotyper

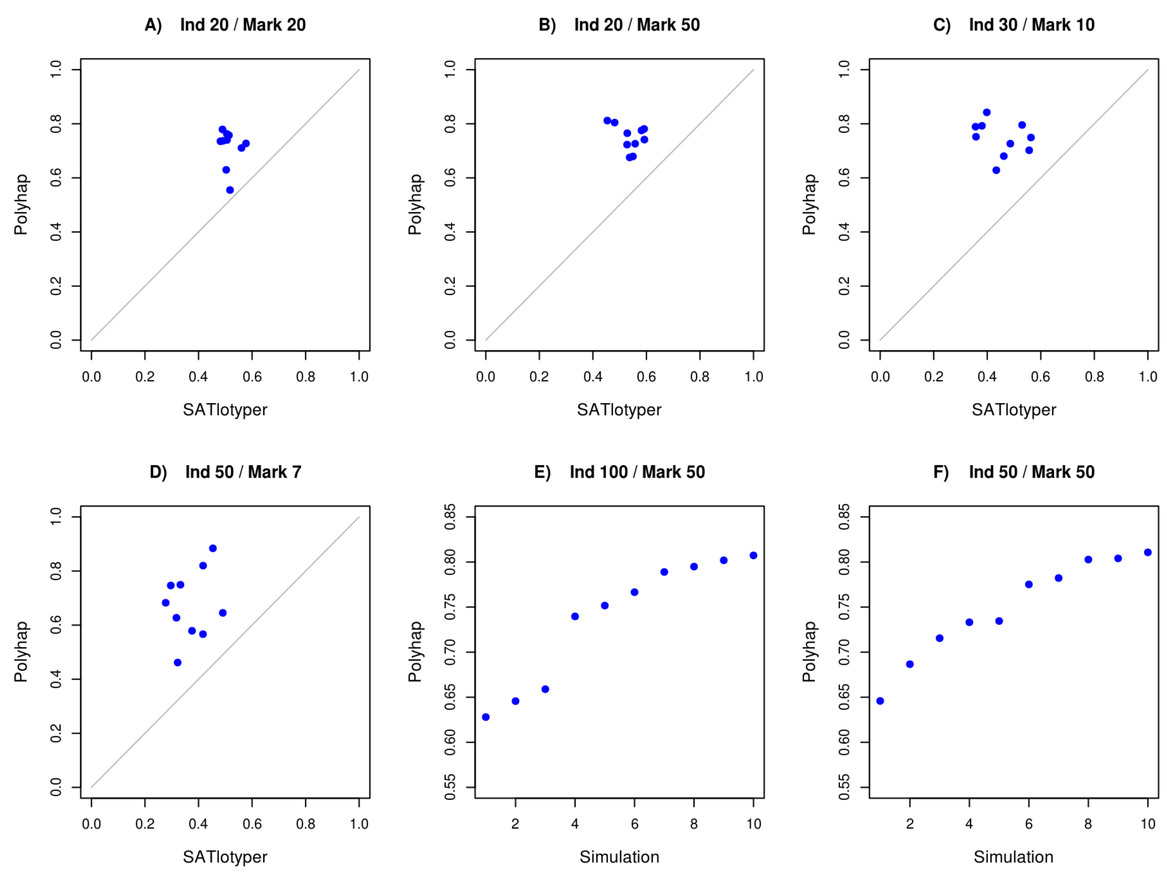
3. Polyhap

This method is motivated by the model behind fastPHASE[6], a tool to phase diploids. It is based on the observation that over short regions in genomes, haplotypes tend to cluster into groups. The cluster membership varies along the chromosome according to a recombination rate. This model perceives each cluster like one of the common haplotypes in the population and uses Hidden Markov Model(HMM) assumption for the cluster memberships.



5. Results

We simulated unphased genotype data with various numbers of individuals and SNPs (10 repetitions). We evaluated the performance of both the methods, using switch error rate as shown in the plot below. Switch error per individual is the minimum number of switches made to reconstruct its true haplotypes. On smaller data sets, SATlotyper outperformed Polyhap, but it failed to converge even after three days for larger ones.



Switch error rates : A to D compare Polyhap against SATlotyper. For E & F, SATlotyper did not converge

6. Conclusion

SATlotyper does not always converge in a reasonable time and so is not applicable to large data sets with hundreds of individuals and markers. Polyhap may be used only for ploidies up to four and we observe high rates of switch error. There is a room for improvement.

Literature

[1]. Waldron ERB, Whittaker JC, Balding DJ: Fine mapping of disease genes via haplotype clustering. Genet Epidemiol 2006, 30:170-179.

[2]. Neigenfind J, Gyetvai G, Basekow R, Diehl S, Achenbach U, Gebhardt C, Selbig J, Kersten B: Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. BMC Genomics 2008, 9:.

[4]. Shu-Yi Su , Jonathan White , David J Balding, Lachlan JM Coin: Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions.BMC Bioinformatics 2008, 9:513

[5]. Voorrips, R.E. , Maliepaard, C.A. (2012) BMC Bioinformatics 13 (2012). - ISSN 1471-2105 - p. 12.

[6]. Scheet P, Stephens M: A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am J Hum Genet 2006, 78:629-644.

[7] . Lynce I, Marques-Silva JP: Efficient haplotype inference with Boolean Satisfiability. National Conference on Artificial Intelligence (AAAI) 2006.