Bioinformatics, 31(24), 2015, 3984–3987
doi: 10.1093/bioinformatics/btv452
Advance Access Publication Date: 18 August 2015
Applications Note

OXFORD

Genome analysis

GenomeLaser: fast and accurate haplotyping from pedigree genotypes

Wenzhi Li[1,2,†], Guoxing Fu[3,†], Weinian Rao[3], Wei Xu[2], Li Ma[2,3], Shiwen Guo[1,*] and Qing Song[2,3,4,*]

[1]Department of Neurosurgery, First Affiliated Hospital of Medical School, Xi'an Jiaotong University, Xi'an, Shaanxi, 710061 China, [2]Cardiovascular Research Institute and Department of Medicine, Morehouse School of Medicine, Atlanta, GA, 30310 USA, [3]4DGENOME Inc, Atlanta, GA, 30033 USA and [4]Center of Big Data and Bioinformatics, First Affiliated Hospital of Medical School, Xi'an Jiaotong University, Xi'an, Shaanxi, 710061 China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on April 24, 2015; revised on July 21, 2015; accepted on July 28, 2015

Abstract

Summary: We present a software tool called GenomeLaser that determines the haplotypes of each person from unphased high-throughput genotypes in family pedigrees. This method features high accuracy, chromosome-range phasing distance, linear computing, flexible pedigree types and flexible genetic marker types.
Availability and implementation: http://www.4dgenome.com/software/genomelaser.html.
Contact: qsong@msm.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Haplotype refers to a group of alleles inherited together on a single chromosome (Browning and Browning, 2011). Haplotype is essential for mapping disease signals, finding functional cis-acting loops among non-coding elements, imputing the missing values in big data. Genotypes can be obtained directly from high-throughput genotyping or next-generation sequencing, but haplotypes cannot be obtained directly. Given that current high-throughput haplotyping technologies are still expensive (Fan *et al.*, 2011; Kirkness *et al.*, 2013; Kitzman *et al.*, 2011; Kuleshov *et al.*, 2014; Ma *et al.*, 2010; Rao *et al.*, 2013; Selvaraj *et al.*, 2013; Suk *et al.*, 2011; Yang *et al.*, 2011) and statistical haplotyping approach suffers from ambiguities (Browning and Browning, 2011; Li *et al.*, 2010; Liu *et al.*, 2008; Niu *et al.*, 2002; Qin *et al.*, 2002), phasing by pedigrees is the most cost-effective and accurate approach for haplotype determination whenever pedigree genotypes are available (Chen *et al.*, 2013). Most of the phasing programs by pedigrees assume zero or minimal recombinations and difficult for large complicated and looping pedigrees (Abecasis *et al.*, 2002; Druet and Georges, 2010, 2015; Gao *et al.*, 2009; Kong *et al.*, 2008; Kruglyak *et al.*, 1996; Li and Li, 2009; O'Connell, 2000; O'Connell *et al.*, 2014; Sobel and Lange, 1996; Zhang *et al.*, 2005). The lack of efficient and accurate methods for long-range haplotype determination has hampered the functional interpretation of non-coding cis-acting elements in human genome. Here, we present a computational tool, GenomeLaser, to reconstruct personal haplotypes from unphased genotypes, it can also handle the looped pedigrees and missing genotype data.

2 Features

Pedigree dissection is the central switchboard throughout the entire procedure (Supplementary Fig. S1). Large pedigrees will be first dissected to nuclear families. If a person is shared by two nuclear families, he/she will be included in both nuclear family units. In the algorithm of Laser I, each nuclear family will be further dissected into trios (Supplementary Fig. S2). A nuclear family with n children will be dissected into n trios. If only one parent is recruited in a study, the parent–child pair will be used. Then the child haplotypes

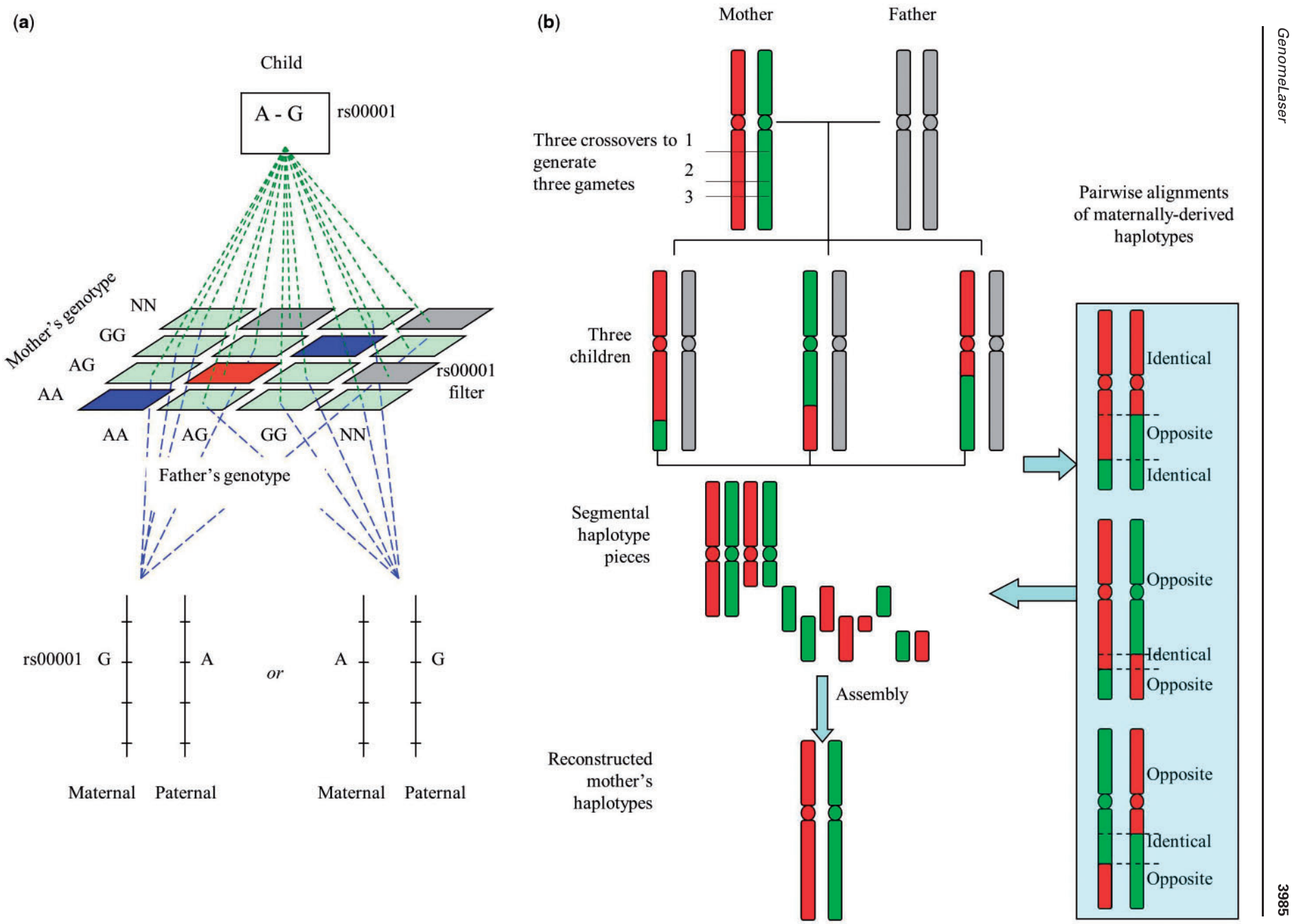

Fig. 1. The GenomeLaser principle. (**a**) Resolving children's haplotypes including red, all-three heterozygotes; blue, misinheritance and gray, missing data. (**b**) Resolving parental haplotypes. To simplify this figure, we focus only on the q arm of maternally derived chromosomes

resolved in Laser I will be sent back to the pedigree dissection algorithm, in which children's haplotypes inherited from the same parent will be grouped (Supplementary Fig. S2) and sent to the Laser II algorithm. In case that a person has children in more than one nuclear family, all of his/her children in different nuclear family units will be grouped together.

The algorithm of GenomeLaser is based on the Mendelian Law of Inheritance (Hodge *et al.*, 1999). It is composed of four components, pedigree dissection, Laser I, Laser II and Laser III (Supplementary Fig. S1).

Laser I resolves child haplotypes. It first determines the allele origins of the child using parental genotypes (Supplementary Tables S1 and S2) (Hodge *et al.*, 1999). For example, when a child is AG, his/her mother is AA and father is AG, under the Mendelian Law of Inheritance, the G allele must come from father and thus the A allele must be inherited from mother. After Laser I determines the parental origin of each allele, all maternally originated alleles will be grouped together and constitute the haplotype of the chromosome inherited from his/her mother and all paternally originated alleles will be grouped and constitute the haplotype of the chromosome inherited from his/her father. If all three members (mother, father, child) of a trio are heterozygous at a single-nucleotide polymorphism (SNP) locus, Laser I will output XX at the triple-heterozygous SNP site on the child haplotypes of this trio.

Laser II resolves parental haplotypes. It first groups the child haplotypes inherited from the same person and then compares the nucleotide sequences among these child haplotypes. A zebra pattern will appear in which identical parts (all alleles are the same) and non-identical segments (all alleles are different) will separate each other along a chromosome. The boundary between an identical segment and an opposite segment will indicate a crossover breakpoint occurred in either gamete of these two children. Laser II then cut the children's haplotypes into pieces at these boundaries and reassemble them into two chromosomal haplotypes (Fig. 1).

After Laser II, there are some loci ($8.7 \pm 5.9\%$ of all SNPs) that are still labeled as NN or XX due to missing data or triple-heterozygotes. Laser III will impute all of those NN and XX sites with an additional input, the reference panel, using the phase-resolved loci by Laser I and II as seeds into Laser III. The algorithm for Laser III was the same as HiFi, whose algorithm was described in a previous publication (Rao *et al.*, 2013).

The Laser program is coded in Python. The input files include the pedigree relationship, genotypes and an imputation reference panel. The output will be two haplotypes of each person.

To examine the performance of the Laser program, we created a simulated dataset containing 30 nuclear families (150 individuals, 116 415 SNPs) in a pedigree structure of two parents and three children (Supplementary Methods) and resolved the haplotypes of all members in this pedigree dataset. Missing data (NN) was introduced into the simulated genotypes at randomly selected loci. The results showed that accuracy was >99.99%, and the error rate was 0.003% on parental haplotypes and 0.0009% on child haplotypes (Supplementary Table S3). We then created two simulated complicated pedigrees and resolved the haplotypes of all members (Supplementary Fig. S3). The results showed that accuracy was 99.99%, and the error rate was 0.002% (Supplementary Table S4). Its accuracy is substantially higher compared with other software for phasing pedigrees. We further examined the computing speed on a regular desktop computer (Intel Core i7-2600K CPU at 3.40 GHzx8, 31.3 GB RAM). The computing time of Laser is linear to the number of Trios (Supplementary Fig. S4).

3 Conclusion

GenomeLaser provides an efficient computational tool for determination of molecular haplotypes from unphased genotype data in pedigrees. Accurate chromosome-range personal haplotypes will be extremely useful to explore those long-distance cis-regulatory networks and epigenetic controls of chromosome function. It is not suitable for those scenarios (such as very small genomic regions) in which pedigree information are not available. Briefly, Laser III may be affected by admixture. Laser I and Laser II are based on the Mendelian Law of Inheritance and they will not be affected by genetic admixtures, but Laser III is a reference-based imputation, it will be affected by the admixture. The performance may be improved by the following operations. First, locus-specific ancestry may be determined in parallel with GenomeLaser to monitor the admixtures (Ma *et al.*, 2014). Second, the reference panels should be either chosen according to the admixture background or based on the pooled reference panels (Huang *et al.*, 2009). The statistical haplotyping approaches also use similar strategies to improve imputation performance for admixed samples by either improving the algorithm or careful selecting reference panels (Huang and Tseng, 2014; Krithika *et al.*, 2012; Liu *et al.*, 2013; Zhang *et al.*, 2011).

Funding

This work was supported by National Institutes of Health (R21HG006173, R43HG007621, HL117929, MD007602, RR003034, U54MD07588). The research was conducted in a facility constructed with support from Research Facilities Improvement Grant 1 C06 RR07571 from the National Center for Research Resources, National Institutes of Health.

Conflict of Interest: none declared.

References

Abecasis,G.R. *et al.* (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.

Browning,S.R. and Browning,B.L. (2011) Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, **12**, 703–714.

Chen,W. *et al.* (2013) Genotype calling and haplotyping in parent-offspring trios. *Genome Res.*, **23**, 142–151.

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.

Druet,T. and Georges,M. (2015) LINKPHASE3: an improved pedigree-based phasing algorithm robust to genotyping and map errors. *Bioinformatics*, **31**, 1677–1679.

Fan,H.C. *et al.* (2011) Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.*, **29**, 51–57.

Gao,G. *et al.* (2009) Haplotyping methods for pedigrees. *Hum. Hered.*, **67**, 248–266.

Hodge,S.E. *et al.* (1999) Loss of information due to ambiguous haplotyping of SNPs. *Nat. Genet.*, **21**, 360–361.

Huang,G.H. and Tseng,Y.C. (2014) Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proc.*, **8**, S64.

Huang,L. *et al.* (2009) Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.*, **84**, 235–250.

Kirkness,E.F. *et al.* (2013) Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.*, **23**, 826–832.

Kitzman,J.O. *et al.* (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.*, **29**, 59–63.

Kong,A. *et al.* (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet*, **40**, 1068–1075.

Krithika,S. *et al.* (2012) Evaluation of the imputation performance of the program IMPUTE in an admixed sample from Mexico City using several model designs. *BMC Med. Genomics*, **5**, 12.

Kruglyak,L. *et al.* (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.

Kuleshov,V. *et al.* (2014) Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.*, **32**, 261–266.

Li,X. and Li,J. (2009) An almost linear time algorithm for a general haplotype solution on tree pedigrees with no recombination and its extensions. *J. Bioinform. Comput. Biol.*, **7**, 521–545.

Li,Y. *et al.* (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

Liu,E.Y. *et al.* (2013) MaCH-admix: genotype imputation for admixed populations. *Genet. Epidemiol.*, **37**, 25–37.

Liu,N. *et al.* (2008) Haplotype-association analysis. *Adv. Genet.*, **60**, 335–405.

Ma,L. *et al.* (2010) Direct determination of molecular haplotypes by chromosome microdissection. *Nat. Methods*, **7**, 299–301.

Ma,Y. *et al.* (2014) Accurate inference of local phased ancestry of modern admixed populations. *Sci. Rep.*, **4**, 5800.

Niu,T. *et al.* (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–169.

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

O'Connell,J.R. (2000) Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genet. Epidemiol.*, **19**(Suppl 1), S64–S70.

Qin,Z.S. *et al.* (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **71**, 1242–1247.

Rao,W. *et al.* (2013) High-resolution whole-genome haplotyping using limited seed data. *Nat. Methods*, **10**, 6–7.

Selvaraj,S. *et al.* (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, **31**, 1111–1118.

Sobel,E. and Lange,K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.*, **58**, 1323–1337.

Suk,E.K. *et al.* (2011) A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.*, **21**, 1672–1685.

Yang,H. *et al.* (2011) Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci. USA*, **108**, 12–17.

Zhang,B. *et al.* (2011) Practical consideration of genotype imputation: sample size, window size, reference choice, and untyped rate. *Stat. Interface*, **4**, 339–352.

Zhang,K. *et al.* (2005) HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics*, **21**, 90–103.