Published in final edited form as:
Nat Methods. 2013 January ; 10(1): . doi:10.1038/nmeth.2308.

High-resolution whole-genome haplotyping using limited seed data

Weinian Rao[1], **Yamin Ma**[1], **Li Ma**[1], **Jian Zhao**[2], **Qiling Li**[1], **Weikuan Gu**[3], **Kui Zhang**[4], **Vincent C Bond**[5], and **Qing Song**[1]

Qing Song: qsong@msm.edu

[1]Cardiovascular Research Institute, Morehouse School of Medicine, Atlanta, Georgia, USA.

[2]Independent investigator, Atlanta, Georgia, USA.

[3]Department of Orthopaedic Surgery, University of Tennessee Health Science Center, Memphis, Tennessee, USA.

[4]Section of Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama, USA.

[5]Department of Microbiology, Biochemistry & Immunology, Morehouse School of Medicine, Atlanta, Georgia, USA.

To the Editor: The term ‘haplotype’ refers to a group of alleles inherited on the same chromosome. Although most sequencing technologies cannot resolve which chromosomal copy a given sequence comes from, the value of haplotype information for genetic studies is increasingly being appreciated by researchers. Haplotypes are important for inferring disease status, determining which allele combinations tend to segregate together and helping to ‘fill in’ or impute missing values in regions that lack genotype information to power association studies. Statistical methods exist that can resolve or ‘phase’ haplotypes, but these have been limited to short sequence stretches and can be computationally demanding. Technical difficulties still exist for obtaining accurate whole-genome and long-range haplotypes experimentally[1,2].

Two high-throughput experimental haplotyping approaches have been developed recently. A single-chromosome isolation approach can yield entire chromosomal haplotypes, but resolution is low because of locus dropout during single-molecule whole-genome amplification[3–5]. A fosmid-based approach yields high-resolution haplotypes but not of chromosome length[6]. To improve the resolution of our previously described single-chromosome approach[4], we have developed the haplotype imputation from incomplete data (HiFi) software. Using limited experimental seed data, HiFi can yield two integral, high-resolution personal chromosomal haplotypes in a cost- and time-efficient manner.

Our aim was to integrate the chromosomal-range accuracy of experimental haplotyping with the efficiency of computational approaches. HiFi exhaustively seeks unambiguous matches

Note: Supplementary information is available at http://www.nature.com/doifinder/10.1038/nmeth.2308.

AUTHOR CONTRIBUTIONS

W.R. and Q.S. developed the algorithm. W.R., Y.M. and J.Z. coded the program.

W.R., L.M., W.G., Q.L., Y.M., V.C.B., K.Z. and Q.S. performed the research.

W.R., L.M. and Q.S. analyzed the data. Q.S. designed the research and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

to an individual's seed haplotypes and genotypes among a panel of reference haplotypes along a sliding window (Fig. 1 and Supplementary Methods). Once HiFi identifies a single match in a window, it uses the identified reference haplotypes to impute the phases at all loci within this window. If HiFi does not find a unique match, it adjusts the window size and repeats the search automatically.

We examined HiFi performance on three data sets (Supplementary Table 1), measuring accuracy as the concordance of HiFi output with high-confidence phase results from trio families (Supplementary Methods). We simulated the first data set from HapMap trio haplotypes; then we blinded the phases randomly at 70% of the entire (homozygous and heterozygous) single-nucleotide polymorphism (SNP) set. The second data set contained haplotype-resolved SNPs at ~40.7% of heterozygous loci with the single-chromosome isolation approach[3]. We observed 99.5% (Caucasian) and 99.2% (African) concordances among all SNPs (including homozygous and heterozygous loci) and 98.1% (Caucasian) and 97.9% (African) concordances among imputed heterozygous SNPs in the first data set (Supplementary Table 2). In the second data set, we observed 99.7% concordance among all SNPs and 98.2% concordance among imputed heterozygous SNPs (Supplementary Table 3). Single-chromosome haplotyping data[4] from the third data set further validated HiFi accuracy, which achieved 98.23% concordance among imputed heterozygous loci starting from 30% of phased heterozygous SNPs. The existence of missing data in the input seed haplotype and genotype data sets showed only a modest influence on the accuracy of HiFi (Supplementary Tables 4 and 5). Thus, relatively limited seed information can be used in combination with HiFi to gain accurate long-range haplotypes.

Because it is unrealistic to recruit a well-matched reference panel for every population in the world, mixed reference panels from diverse populations have been suggested for when well-matched references are unavailable. The capacity to handle a large reference panel would thus be an essential feature for imputation software. We created a series of simulated reference haplotype panels of various sizes and observed high speed and a linear relationship between panel size and computing time using HiFi (Supplementary Fig. 1).

Errors cannot be avoided in input data. To examine the error tolerance of HiFi, we introduced allele-calling errors into seed genotype and seed haplotype data sets, and flip errors into the reference panel. When both seed genotype and seed haplotype input contained a 0.5% error rate, HiFi accuracy was 97.7% among heterozygous sites (Supplementary Table 6). Flip errors in the reference panel did not significantly affect the accuracy of HiFi (Supplementary Table 7).

We summarize the features of HiFi in Supplementary Table 8. HiFi phased 116,415 SNPs in 10.6 s on a desktop computer (CPU 3.0 GHz, 8 GB of RAM), and it could be used to phase each human genome in ~2.5 min with >99% accuracy (Supplementary Tables 3 and 9). Owing to its simplicity, HiFi reached a very high computing speed (Supplementary Fig. 2), and the computing time correlated linearly to the number of SNPs (Supplementary Fig. 1 and Supplementary Table 9). HiFi performed well on rare SNPs, with an accuracy of 94.1% among imputed rare heterozygous loci (Supplementary Table 10). In parallel to HiFi, we have developed a quality-score system that is outputted together with the phasing results (Supplementary Figs. 3 and 4). As a reference-based method, HiFi cannot phase *de novo* mutations and structural variation until they are included in future reference panels (see Supplementary Discussion for advantages of long-range haplotyping information and limitations of the software). The entire procedure and projected labor and costs of this integrated haplotyping pipeline are described in Supplementary Table 11 and Supplementary Note 1.

The HiFi program and test datasets are available as Supplementary Software.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by US National Institutes of Health grants (R21HG006173, RC4MD005964, HL003676, RR014758, RR003034, GM74913, G12MD007602, U54MD007588), an American Heart Association grant (09GRNT2300003), and in part by the Baltimore Research Enhancement Award Program in Stroke of the Baltimore Department of Veterans Affairs. We also want to thank G.H. Gibbons for his critical discussions.

References

1. Bansal V, Tewhey R, Topol EJ, Schork NJ. Nat. Biotechnol. 2011; 29:38–39. [PubMed: 21221098]
2. Rusk N. Nat. Methods. 2011; 8:107. [PubMed: 21355116]
3. Fan HC, Wang J, Potanina A, Quake SR. Nat. Biotechnol. 2011; 29:51–57. [PubMed: 21170043]
4. Ma L, et al. Nat. Methods. 2010; 7:299–301. [PubMed: 20305652]
5. Yang H, Chen X, Wong WH. Proc. Natl. Acad. Sci. USA. 2011; 108:12–17. [PubMed: 21169219]
6. Kitzman JO, et al. Nat. Biotechnol. 2011; 29:59–63. [PubMed: 21170042]

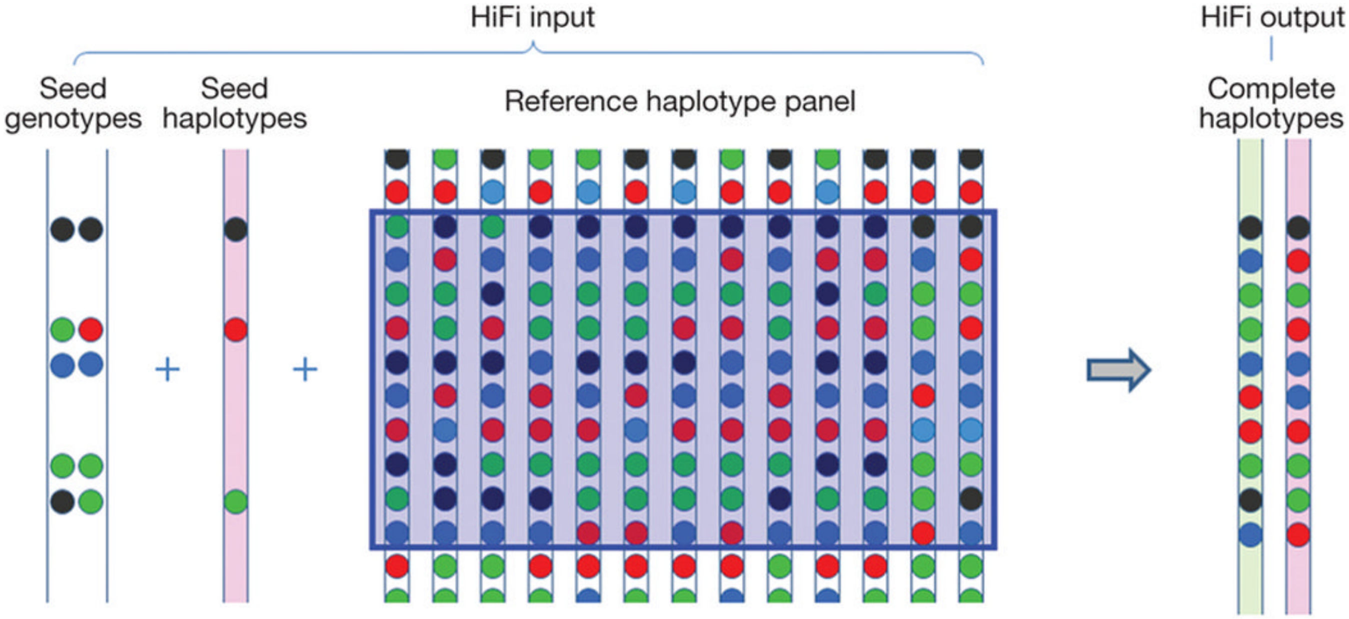

Figure 1.

The principle behind HiFi. Unrelated individuals may share short stretches of DNA sequences derived from their common ancestors. HiFi requires three inputs: low-resolution seed genotypes, low-resolution seed haplotypes and a reference haplotype panel. HiFi uses the seed haplotypes to predict allele phases at unresolved loci. When no match or multiple matches are found in a window, HiFi expands the window until a unique match is found.