Data in Brief 8 (2016) 1412–1415

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data supporting the high-accuracy haplotype imputation using unphased genotype data as the references

Wenzhi Li [a,b,1], Wei Xu [b,1], Shaohua He [c], Li Ma [b,c,*], Qing Song [a,b,c,*]

[a] Center of Big Data and Bioinformatics, First Affiliated Hospital of Medicine School, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[b] Cardiovascular Research Institute, Morehouse School of Medicine, Atlanta, GA, USA
[c] 4DGenome Inc., Atlanta, GA, USA

ARTICLE INFO

Article history:
Received 3 November 2015
Received in revised form
10 May 2016
Accepted 21 June 2016
Available online 29 June 2016

ABSTRACT

The data presented in this article is related to the research article entitled "High-accuracy haplotype imputation using unphased genotype data as the references" which reports the unphased genotype data can be used as reference for haplotyping imputation [1]. This article reports different implementation generation pipeline, the results of performance comparison between different implementations (A, B, and C) and between HiFi and three major imputation software tools. Our data showed that the performances of these three implementations are similar on accuracy, in which the accuracy of implementation-B is slightly but consistently higher than A and C. HiFi performed better on haplotype imputation accuracy and three other software performed slightly better on genotype imputation accuracy. These data may provide a strategy for choosing optimal phasing pipeline and software for different studies.

DOI of original article: http://dx.doi.org/10.1016/j.gene.2015.07.082

[*] Corresponding authors at: Cardiovascular Research Institute, Morehouse School of Medicine, Atlanta, GA, USA.
E-mail addresses: lma@msm.edu (L. Ma), qsong@msm.edu (Q. Song).

[1] Equal contribution.

http://dx.doi.org/10.1016/j.dib.2016.06.029
2352-3409/

Specifications Table

Subject area	Biology
More specific subject area	Bioinformatics
Type of data	Tables
How data was acquired	Genotype and haplotype data were obtained from the International HapMap Project database
Data format	Analyzed
Experimental factors	The original data were reformatted to fit the requirement of different software
Experimental features	We generated different implementations from HapMap data set. Then: [1] We compared the performance of different implementations [2]. We compared the phasing performances among HiFi, MACH 1.0, IMPUTE2, BEAGLE.
Data source location	Atlanta, Georgia, USA
Data accessibility	The data are with this article

Value of the data

- This data is beneficial to researchers who are interested in haplotyping The data may provide guidance on how to choose the optimal phasing pipeline.
- This data is beneficial to researchers who are interested in imputations and comparison between HiFi and three major phasing software tools (MACH, Impute2 and Beagle) on the accuracy and speed. The data may provide guidance on how to choose the suitable software for different study.
- This data is helpful to compare between HiFi and three major phasing software tools (MACH, Impute2 and Beagle) on their tolerance on statistical reference panels.

1. Data

Data presented are summaries of comparison of HiFi performances with three different implementations A, B and C; comparison of HiFi and three standard imputation software performances with molecular reference and statistical reference. The data showed that implementation-B is slightly but consistently higher than A and C; and the data also showed that HiFi performed better on haplotype imputation accuracy and speed,three other tools performed slightly better on genotype imputation.

2. Experimental design, materials and methods

2.1. Acquisition and processing of HapMap data for different implementations

We downloaded CEU (CEPH, U.S. Utah residents with ancestry from northern and western Europe) chromosome 1 genotype data and haplotype data from HapMap in text format [5,6]. We use the original haplotype data as molecular reference. To generate the statistical haplotype reference panel, we erased the phase information from those trio haplotypes downloaded from HapMap, and then used the software Beagle version 3.3.2 to resolve the haplotypes from the unphased genotypes. Then we generated following three different implementations by Beagle version 3.3.2: (A) Beagle statistical phasing of unrelated persons and Mendelian-inheritance-based phasing of trios, and then pools the results together; (B) Beagle statistical phasing of pooled unrelated persons and trios, but presumes all as unrelated; and (C) Beagle statistical phasing of pooled unrelated persons and trios, and specifying the family structure in the input. And we chose same 6 samples [2] for further analysis.

2.2. *Comparison of HiFi performances with three different implementations A, B and C*

We compared the HiFi performances with three different implementations. Our data showed that the performances of these three implementations are similar on accuracy, in which the accuracy of implementation-B is slightly but consistently higher than A and C (Table S1).

2.3. *Comparison of HiFi and three standard imputation software performances with molecular reference and statistical reference*

We compared the performance between HiFi and three standard imputation software tools (MACH, IMPUTE2 and BEAGLE) [7–9]. As the result, HiFi performed better on haplotype imputation accuracy (Table S2) and speed (Table S4), whereas MACH, IMPUTE2 and BEAGLE performed slightly better on genotype imputation accuracy (Table S3), in which MACH and IMPUTE2 performed the best on genotype imputation.

Acknowledgments

This work was supported by National Institutes of Health (R21HG006173, R43HG007621, HL117929, MD007602, MD005964, RR003034, U54MD07588); and the American Heart Association Grant (09GRNT2300003).

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.06.029.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.06.029.

References

[1] W. Li, W. Xu, G. Fu, L. Ma, J. Richards, W. Rao, T. Bythwood, S. Guo, Q. Song, High-accuracy haplotype imputation using unphased genotype data as the references, Gene 572 (2015) 279–284.

[2] W. Li, W. Xu, S. He, L. Ma, Q. Song, References for haplotype imputation in the big data era, Mol. Biol. (2015), http://dx.doi.org/10.4172/2168-9547.1000143 (in press).

[5] W. Li, G. Fu, W. Rao, W. Xu, L. Ma, S. Guo, Q. Song, GenomeLaser: fast and accurate haplotyping from pedigree genotypes, Bioinformatics (2015), pii: btv452. [Epub ahead of print].

[6] W. Xu, L. Ma, W. Li, T.A. Brunson, X. Tian, J. Richards, Q. Li, T. Bythwood, Z. Yuan, Q. Song, Functional pseudogenes inhibit the superoxide production, Precis. Med. 1 (2015).

[7] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLoS Genet. 5 (2009) e1000529.

[8] S.R. Browning, B.L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, Am. J. Hum. Genet. 81 (2007) 1084–1097.

[9] Y. Li, C.J. Willer, J. Ding, P. Scheet, G.R. Abecasis, MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes, Genet. Epidemiol. 34 (2010) 816–834.

Further reading

[3] Y. Ma, J. Zhao, J.S. Wong, L. Ma, W. Li, G. Fu, W. Xu, K. Zhang, R.A. Kittles, Y. Li, Q. Song, Accurate inference of local phased ancestry of modern admixed populations, Sci. Rep. 4 (2014) 5800.

[4] W. Rao, Y. Ma, L. Ma, J. Zhao, Q. Li, W. Gu, K. Zhang, V.C. Bond, Q. Song, High-resolution whole-genome haplotyping using limited seed data, Nat. Methods 10 (2013) 6–7.