HHS Public Access

Author manuscript

Res Comput Mol Biol. Author manuscript; available in PMC 2017 August 16.

Published in final edited form as:

Res Comput Mol Biol. 2015 ; 9029: 28–29. doi:10.1007/978-3-319-16706-0_4.

HapTree-X: An Integrative Bayesian Framework for Haplotype Reconstruction from Transcriptome and Genome Sequencing Data

Emily Berger[1,2,3], **Deniz Yorukoglu**[2], and **Bonnie Berger**[1,2]

[1]Department of Mathematics, MIT, Cambridge, MA, USA

[2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

[3]Department of Mathematics, UC Berkeley, Berkeley, CA, USA

Background

By running standard genotype calling tools, it is possible to accurately identify the number of “wild type” and “mutant” alleles for each single-nucleotide polymorphism (SNP) site. However, in the case of two heterozygous SNP sites, genotype calling tools cannot determine whether “mutant” alleles from different SNP loci are on the same chromosome or on different homologous chromosomes (i.e. compound heterozygote). In many cases, the latter can cause loss of function while the former is healthy; therefore, it is necessary to identify the phase (or diplotype) – the copies of a chromosome on which the mutant alleles occur – in addition to the genotype. Identifying phase information for an individual is important in biomedical studies due to disease association of complex haplotype effects such as compound heterozygosity, as well as matching donor and host in organ transplantation.

As more sequencing data becomes available, we seek to design efficient algorithms to obtain accurate and comprehensive phase information directly from transcriptomic, as well as the commonly-used genomic, NGS read data. Transcriptome sequencing data differs from genomic read data in that genes often have differential haplotypic expression [3] (*expression bias* between the maternal and paternal chromosomes of a particular gene). We are able to leverage this asymmetry to increase the number of SNPs of an individual that can be phased.

Method

We develop the first method for solving the haplotype reconstruction problem using differential allele-specific expression (DASE) information within RNA-seq data. We follow the intuition that DASE in the transcriptome can be exploited to improve phasing power because SNP alleles within maternal and paternal haplotypes of a gene are present in the read data at (different) frequencies corresponding to the differential haplotypic expression (DHE). To solve this haplotype reconstruction problem, we introduce a new maximum-

Correspondence to: Bonnie Berger.

E. Berger and D. Yorukoglu—contributed equally.

likelihood formulation which takes into account DASE (generalizing that from HapTree [2]) and is thus able to newly exploit reads covering only one SNP. This formulation results in a novel integrative algorithm, HapTree-X, which determines a haplotype of maximal likelihood based on both RNA-seq and DNA-seq read data.

Results

To measure phasing accuracy and assess theoretical accuracy bounds, we define *concordant expression* to be when the DASE of a SNP agrees with the DHE of the gene to which the SNP belongs; that is when the majority allele (allele present in the majority of the reads overlapping the SNP locus) is in agreement with the expected majority allele as determined by the DHE. We show that under realistic biological assumptions, the solution of maximal likelihood is, intuitively, that which has concordant expression at each SNP locus. Furthermore, we show that the theoretical probability of concordant expression increases exponentially with the coverage level.

We compare the accuracy of phasing (along with the total number of SNPs phased and phased block sizes) DNA-seq and RNA-seq datasets from NA12878 using HapTreeX to that of HapCut [1]. Our results indicate that incorporating DASE information into haplotype phasing increases the total number of SNPs phased, without increasing the switch error rate (with respect to the trio-phased gold-standard annotation). Furthermore, HapTree-X reduces the total number of phased blocks while increasing their overall sizes. Our work shows for the first time that RNA-seq data can be used as a complement to DNA-seq data to improve phasing.

References

1. Bansal V, Bafna V. Hapcut: an efficient and accurate algorithm for the haplotype assembly problem. Bioinformatics. 2008; 24(16):i153–i159. [PubMed: 18689818]
2. Berger E, Yorukoglu D, Peng J, Berger B. Haptree: A novel Bayesian framework for single individual polyplotyping using NGS data. PLoS Computational Biology. 2014; 10(3):e1003502. [PubMed: 24675685]
3. Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, et al. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. PLoS Genetics. 2008; 4(2):e1000006. [PubMed: 18454203]