Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw484
Advance Access Publication Date: 6 August 2016
Applications Note

OXFORD

Genetics and population analysis

scphaser: haplotype inference using single-cell RNA-seq data

Daniel Edsgärd[1], Björn Reinius[1] and Rickard Sandberg[1,2,*]

[1]Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden and [2]Ludwig Institute for Cancer Research, Box 240, 171 77 Stockholm, Sweden

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on May 17, 2016; revised on July 2, 2016; accepted on July 9, 2016

Abstract

Summary: Determination of haplotypes is important for modelling the phenotypic consequences of genetic variation in diploid organisms, including *cis*-regulatory control and compound heterozygosity. We realized that single-cell RNA-seq (scRNA-seq) data are well suited for phasing genetic variants, since both transcriptional bursts and technical bottlenecks cause pronounced allelic fluctuations in individual single cells. Here we present scphaser, an R package that phases alleles at heterozygous variants to reconstruct haplotypes within transcribed regions of the genome using scRNA-seq data. The devised method efficiently and accurately reconstructed the known haplotype for ≥93% of phasable genes in both human and mouse. It also enables phasing of rare and *de novo* variants and variants far apart within genes, which is hard to attain with population-based computational inference.
Availability and Implementation: scphaser is implemented as an R package. Tutorial and code are available at https://github.com/edsgard/scphaser
Contact: rickard.sandberg@ki.se
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The haplotype phase, the sequence of alleles present on the same nucleic acid molecule, such as the maternal or paternal copy of a chromosome, is of importance to reveal relationships between DNA sequence and phenotype. Major efforts have been made using expression-quantitative-trait-loci studies to identify *cis*-regulatory variants that affect gene expression. Making use of allele-specific expression (ASE) increases the power of such studies; however, to reach the full potential of ASE-based approaches require or depend on phased alleles within genes (Kumasaka *et al.*, 2016; van de Geijn *et al.*, 2015). Phase information is also important for associating clinical outcomes to genetic variation, e.g. to identify cases of compound heterozygosity where risk alleles at different loci do not co-occur on the same DNA molecule but affect both homologous copies of a gene. This information can help to elucidate the impact of mutations in cancer, Mendelian disease and in personalized medicine.

Several approaches exist to determine haplotypes, including direct experimental phasing of a single individual, such as physical separation of the chromosomes, dilution to single-haplotype concentration equivalents, barcoding schemes and long-read sequencing, as well as computational approaches including population phasing using genome reference panels, transmission between related individuals, or utilizing the presence of multiple variants in overlapping reads (Browning and Browning, 2011; Snyder *et al.*, 2015). However, the direct experimental phasing techniques are relatively laborious and the computational methods depend on either DNA data or sequencing read length.

RNA-sequencing (RNA-seq) allows quantification of the number of transcribed copies from each allele; however, short read lengths preclude direct observation of haplotype sequences. Studies to date have evaluated ASE in tissues or cell populations where ASE in individual cells is averaged out. By contrast, single-cell RNA-seq provides frequent monoallelic or skewed allelic expression (Fig. 1A), due to stochastic bursting of gene expression and technical losses (Reinius and Sandberg, 2015). Here, we demonstrate that the pronounced allelic fluctuations in scRNA-seq data can be used to accurately infer haplotypes of the transcribed parts of a genome (Fig. 1B).

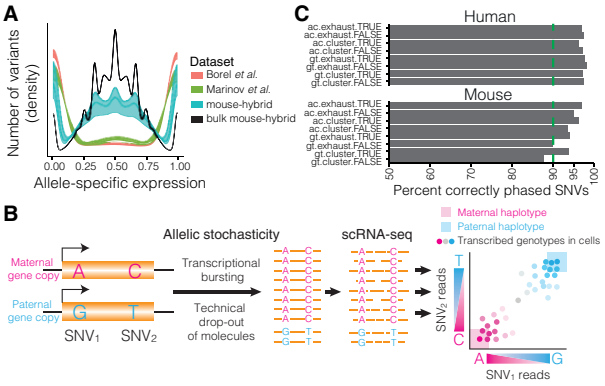

Fig. 1. Concept and performance of scphaser. (**A**) Number of genes against observed ASE in scRNA-seq (two human and a mouse dataset) and bulk RNA-seq data. Line indicates mean and band the inter-quartile range across cells. (**B**) Transcriptional bursts and technical drop-out cause frequent monoallelic or allele-biased observations in scRNA-seq data, which can reveal the phase of transcribed sequences. (**C**) Percent correctly phased SNVs in the human and mouse dataset, *X*-axis labels denote the input, method and weighing settings for the phasing (Methods)

2 Methods

scphaser assumes a diploid genome, for which there are two possible states of the DNA haplotype sequence. Input to the program is either the expression of each allele, or the expressed genotype, for every variant and cell. If a gene is monoallelically expressed in a cell, the genotype vector of such a cell is identical to the haplotype sequence. Cells with an imbalanced allelic expression will be closer to the haplotype towards which it is imbalanced. Determining which of the two underlying states a cell is closest to can then be viewed as a two-class clustering problem.

To solve this, we implemented an exhaustive search where every possible haplotype of a gene is evaluated, where the haplotype is chosen that minimize the variation of the resulting cell distribution. Additionally, we include PAM-clustering (R package 'cluster'), which is used for genes with ten or more variants to reduce runtime. We also include an option to minimize the variation using discrete transcribed genotypes, instead of the continuous ASE, and a simple transcribed-genotype caller if allele read counts are input. The package also includes the option to weigh the calculations based on the read counts, to account for sampling error. To enable short run-times, parallelization was implemented where the input is split into subsets of genes. Thus, scphaser provides in total eight alternative phasing strategies: clustering: {exhaustive, PAM}, input: {genotype, read allele counts} and weigh: {true, false}.

Further details on the algorithms, including gene and variant filters, are described in the Supplementary Data. Usage instructions are detailed in the vignette, as part of the R package.

3 Results

We assessed the performance of scphaser on two full-length scRNA-seq datasets, where the phase was known. The first dataset contained 336 cells from a mouse F1 cross of two inbred strains for which the genomes are known (CAST/EiJ × C57BL/6J, reciprocal cross) and the second dataset contained 28 single cells from the human individual NA12878, where the phase was inferred via transmission between the sequenced genomes of the family-trio (Marinov *et al.*, 2014). All eight phasing approaches implemented in scphaser (Methods) were assessed with respect to varying pre-filtering settings and genotype calling cutoffs for the two datasets (Supplementary Fig. S1). A suitable trade-off between phasing accuracy and number of phasable genes (at least two heterozygous variants left in a gene) was obtained by pre-filtering for at least five cells with imbalanced ASE for a variant, where imbalance should be at least 3-fold ($fc \leq 1/3$ or $fc \geq 3$, where fc = alternative allele count/reference allele count) and there are at least three reads for any of the alleles in such a cell. Furthermore, we set the default phasing arguments to cluster = exhaustive, input = allele counts and weigh = false, since that combination performed the best on average across the two datasets, with 95.1% and 97.5% correctly phased variants in the mouse and human dataset, respectively (Fig. 1C). At gene-level, 93.6% and 94.9% of phasable genes had all variants correctly phased.

Originally, there were 20 268 and 15 597 RefSeq genes with at least two heterozygous variants at the DNA level and using the default pre-filtering values 8563 and 534 phasable genes remained in the mouse and human dataset, respectively (336 versus 28 sequenced cells). In a second human dataset, containing 163 single cells sequenced from a single individual (Borel *et al.*, 2015), we found that 3155 RefSeq genes were phasable out of 15 556 RefSeq genes with at least two heterozygous variants. The dependency of number of phasable genes with respect to number of sequenced cells, degree of genome heterozygosity and sequencing depth is shown in Supplementary Figure S2. The run-time on the human 163-cell dataset using default settings was 61 seconds using 80 cores and the computational complexity with respect to the number of variants, genes and cells are shown in Supplementary Figure S3.

4 Discussion

We conclude that phasing by leveraging the imbalanced ASE frequently observed in full-length scRNA-seq data is both accurate and fast. Using RNA instead of DNA enables phasing of variants located far apart from each other within a gene due to introns. As data from only a single individual is needed scphaser can also phase rare and *de novo* variants. The number of genes available for phasing depends on (i) the number of sequenced single cells, (ii) cell type, as the number of expressed genes vary considerably between cell types, (iii) the number of heterozygous variants in the individual, where genetically distant parents yield higher heterozygosity, (iv) sequencing depth and (v) the efficiency of the full-length scRNA-seq protocol. The retrieved gene phase information has important applications in functional and clinical genomics, such as empowering *cis*-regulatory variation studies and in elucidating the impact of haplotype structures on phenotypic outcome and response.

Author contributions

DE conceived algorithms, programmed the scphaser R package, evaluated its performance and wrote the manuscript. BR conceived the allele-phasing concept and programmed an initiatory phasing script. RS supervised the project and contributed to the manuscript.

Funding

This work has been supported by the Swedish Foundation for Strategic Research, European Research Council (648842) and the Swedish Research Council.

Conflict of Interest: The authors have filed a patent application on phasing using single-cell RNA-seq data.

References

Borel,C. *et al*. (2015) Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet*., **96**, 70–80.

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

Kumasaka,N. *et al*. (2016) Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet*, **48**, 206–213.

Marinov,G.K. *et al*. (2014) From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res*., **24**, 496–510.

Reinius,B. and Sandberg,R. (2015) Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet*., **16**, 653–664.

Snyder,M.W. *et al*. (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet*., **16**, 344–358.

van de Geijn,B. *et al*. (2015) WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, **12**, 1061–1063.