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

Abstract

Determination of haplotypes is important for correctly modelling the phenotypic consequences of genetic variation in diploid organisms, including *cis*-regulatory control and compound heterozygosity. Single cell RNA-seq (scRNA-seq) data is exceptionally 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. scphaser is implemented as an R package. Tutorial and code are available at https://github.com/edsgard/scphaser* (*Private repository, access available upon request)

Background

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 elucidate 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, state-of-the-art ASE-based methods to identify cis-regulatory variants require or depend on phased alleles within genes to reach their full potential (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. Such analysis may be especially important to elucidate the impact of mutations in cancer, Mendelian disease and personalized medicine.

Several approaches exist to determine the 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 (S. R. Browning and B. L. Browning, 2011). However, the direct experimental phasing techniques are relatively laborious and the computational methods depend on either DNA data or sequencing read length.

RNA-sequencing allows quantification of the number of transcribed copies from each of the two alleles of a diploid genome; however, short read lengths preclude direct observation of the haplotype sequence. Studies to date evaluated ASE in tissues or cell populations, where the ASE from individual cells is averaged out and it is difficult to obtain gene-based estimates from data at independent heterozygous loci. Instead, scRNA-seq has several unexplored advantages, such as frequent monoallelic or skewed allelic expression (Figure 1A), due to stochastic bursting of gene expression and technical losses of RNA and cDNA molecules (Reinius and Sandberg, 2015). Here, we leverage the pronounced allelic fluctuations in scRNA-seq data to infer the haplotypes of the transcribed parts of a genome (Figure 1B).

Results

We assessed the performance of scphaser on two datasets where the phase was known. This included full-length single-cell RNA-seq data of 336 fibroblast cells from a mouse F1 cross of two inbred strains for which the genomes are known (CAST/EiJ × C57BL/6J, reciprocal cross) and 28 single cells from the human individual NA12878 where phase was inferred via transmission between the sequenced genomes of the family-trio (Marinov, et al., 2014). Using default settings of scphaser 95.1% and 97.5% of variants were correctly phased in the mouse and human dataset, respectively (Figure 1C). At a gene-level 93.6% and 94.9% of genes had all variants correctly phased. Originally, there were 12,247 and 6,065 RefSeq genes with at least two exonic heterozygous variants in the mouse and human dataset, respectively, and 11,512 and 534 genes were phasable (336 vs 28 sequenced cells). In a human dataset with

163 single cells sequenced from an individual (Borel, et al., 2015) we found that 3,155 RefSeq genes were phasable (Figure 1D).

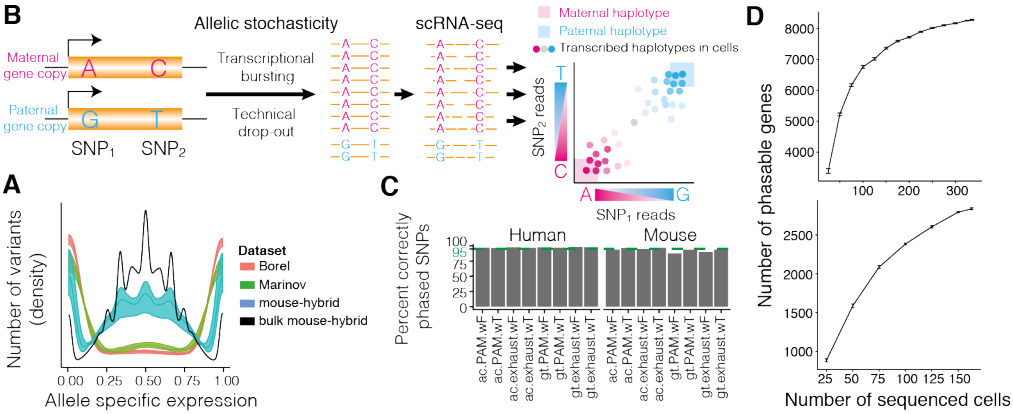

Figure 1. Concept and performance of scphaser. (A) Number of genes against ASE in two human and a mouse scRNA-seq dataset. Line indicates mean and band the inter-quartile range across cells. (B) Stochastic allelic expression bursting and technical drop-out events often cause monoallelic or allele-biased expression in scRNA-seq data. ASE observations in several individual cells can reveal phase of a transcribed sequence, since alleles originating from the same parental copy are co-expressed. (C) Fraction correctly phased SNPs for eight implemented phasing approaches with respect to a human and mouse dataset. X-axis labels denote the input, method and weighing settings for the phasing (Methods). (D) Number of phasable RefSeq genes against number of sequenced cells in a mouse-hybrid (upper) and human (Borel et al.) dataset (lower).

Conclusions

We conclude that phasing by leveraging the imbalanced ASE frequently observed in single-cell RNA-seq data is both accurate and fast. Using RNA instead of DNA enables phasing of variants that are far apart from each other within a gene due to introns. As data from only a single individual is needed we can also phase rare and *de novo* variants. Phasing capacity is facilitated by data from full-length scRNA-seq methods. The more cells that are sequenced the likelihood increase that there are a number of cells where an imbalance is present in the ASE for a particular gene in that individual. 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.

Methods

scphaser assumes a diploid genome, for which there are two possible states of the DNA haplotype sequence. If a gene is mono-allelically expressed the genotype vector of such a cell is identical to the haplotype sequence. Cells in the variant-space, where each variant is a variable with the ASE as domain, with an imbalance in its allelic expression will be closer to the haplotype prototype vector 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 combination of the two possible states for each variant in a gene is evaluated, where the combination is chosen that minimize the variation of the resulting cell distribution. We also include PAM-clustering as an alternative option (R-package "cluster"). 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 a weighing option, based on the read counts, as to account for sampling error. Thus, scphaser provides eight ways to conduct phasing as there are three binary options: clustering: {exhaustive, PAM}, input: {genotype, read allele counts} and weigh: {true, false}. Usage instructions are detailed in the vignette, as part of the R package.

References

Borel, C., *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* 2015;96(1):70-80.

Browning, S.R. and Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 2011;12(10):703-714.

Kumasaka, N., Knights, A.J. and Gaffney, D.J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* 2016;48(2):206-213.

Marinov, G.K., *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* 2014;24(3):496-510.

Reinius, B. and Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* 2015;16(11):653-664.

van de Geijn, B., *et al.* WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 2015;12(11):1061-1063.