Bioinformatics, 37(24), 2021, 4860–4861
doi: 10.1093/bioinformatics/btab454
Advance Access Publication Date: 19 June 2021
Applications Note

Genome analysis

Unfazed: parent-of-origin detection for large and small *de novo* variants

Jonathan R. Belyeu [1,2,†], Thomas A. Sasani[1,2,3,†], Brent S. Pedersen [1,2] and Aaron R. Quinlan [1,2,*]

[1]Department of Human Genetics, Salt Lake City, UT 84112, USA, [2]Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA and [3]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received on February 5, 2021; revised on May 25, 2021; editorial decision on June 13, 2021; accepted on June 15, 2021

Abstract

Summary: *Unfazed* is a command-line tool to determine the parental gamete of origin for *de novo* mutations from paired-end Illumina DNA sequencing reads. *Unfazed* uses variant information for a sequenced trio to identify the parental gamete of origin by linking phase-informative inherited variants to *de novo* mutations using read-based phasing. It achieves a high success rate by chaining reads into haplotype groups, thus increasing the search space for informative sites. Unfazed provides a simple command-line interface and scales well to large inputs, determining parent-of-origin for nearly 30 000 *de novo* variants in under 60 h.

Availability and implementation: Unfazed is available at https://github.com/jbelyeu/unfazed.

Contact: aaronquinlan@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identifying the origin parent for DNA variants, known as 'phasing', is an important task for understanding molecular mechanisms that generate mutations. Phasing *de novo* mutations can also reveal the effects of parental sex and age on germline mutation rates (Jónsson *et al.*, 2017; Sasani *et al.*, 2019), and elucidate parental effects on allele-specific expression (Castel *et al.*, 2016). Tools also exist to define haplotypes of variants within samples by connecting sequencing reads that overlap the variants (Edge *et al.*, 2017; Hager *et al.*, 2020; Martin *et al.*, 2016). Although these tools can assign variants to one of two possible haplotypes, in each case they either do not support *de novo* variation or do not directly report the origin parent for those alleles. These haplotype creation tools also are generally applicable only to single-nucleotide variants (SNVs) and small insertion/deletion (INDEL) variants, not structural variants (SVs), which are rearrangements of at least 50 base pairs. Unfazed applies a novel extended read-based phasing method to identify and report the parent of origin for *de novo* SNV, INDEL and SV mutations identified in family 'trios' (mother, father and child) and uses additional non-read-based phasing information from SNVs internal to deletion or duplication SVs to improve phasing rates. This allows direct prediction of the origin parent for *de novo* variants of all sizes.

2 Application

Unfazed identifies the parental gamete of origin for *de novo* mutations via read-based phasing (Fig. 1A), using individual reads that contain the *de novo* allele and an allele from a phase-informative variant where the origins of the child's alleles are identified by inheritance. The gamete of origin for the *de novo* allele is inferred by linkage to a phase-informative allele. Recovery of phase information is thus limited by species heterozygosity and the existence of a phase-informative variant near enough to the *de novo* allele to be overlapped by a sequencing read.

Unfazed extends the read-based phasing method by including heterozygous loci near the *de novo* allele that are not phase-informative on their own (Fig. 1B). These loci are 'phase-chainable', meaning they can connect reads overlapping the *de novo* allele with other reads. This increases the potential search distance for informative sites up to kilobases from the *de novo* site and improves the recovery of phase information by enabling the use of distant phase-informative sites.

Phasing SVs requires specialized logic, as these variants cannot usually be represented by a single Illumina sequencing read. Instead, Unfazed uses logic inspired by SV detection tools (Belyeu *et al.*, 2020b; Layer *et al.*, 2014) to identify SV evidence in the form of split-reads and discordant read pairs. These reads can then be used

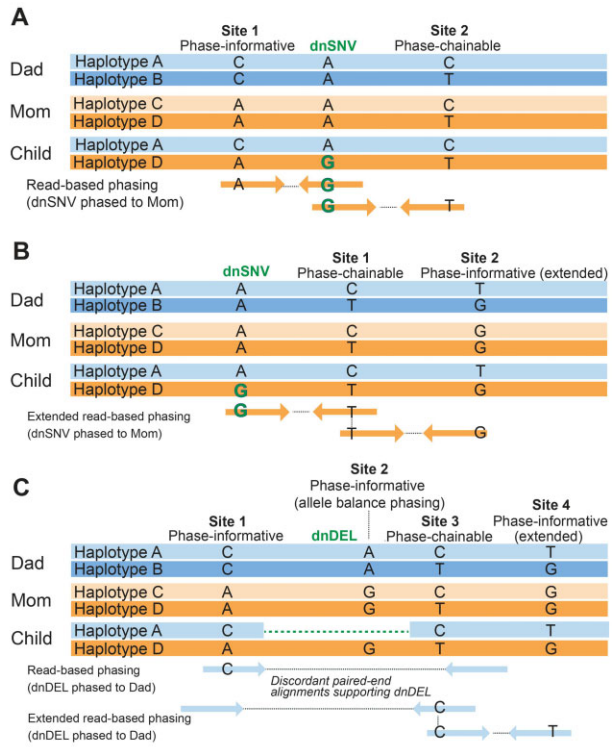

Fig. 1. Unfazed identifies the origin parent for variants by extended read-based phasing. (**A**) Read-based phasing uses reads overlapping a site of interest and a phase informative site to identify the origin parent. (**B**) Extended read-based phasing chains reads to include information from non-overlapped phase-informative sites. (**C**) Extended read-based phasing can be applied to SVs by using discordant pairs or split reads

to connect the *de novo* SV with phase-informative alleles (Fig. 1C), allowing the parental gamete of origin to be identified for deletions, duplications and inversions. Insertions, however, are not supported due to greater complexity in the read alignment patterns they produce (Chen *et al.* 2016).

Deletions and duplications, often referred to as copy-number variants (CNVs), change the number of copies of genomic material. This enables another technique for CNV phasing via the allele balance of inherited variants within the CNV. Allele-balance phasing works by finding phase-informative sites inside the CNV (Fig. 1C) and identifying deviations from the expected allele balance in the offspring. For example, given different homozygous alleles in each parent, the offspring should inherit distinct alleles from each parent. The deletion of one parental copy of a region results in offspring hemizygosity for the other parent's single-nucleotide alleles in the region. The origin parent for the deletion is therefore the one whose single-nucleotide alleles are lost. Duplications can be phased similarly, using the observation that the duplicated copy includes an extra allele from one parent and increasing the allele balance in favor of the *de novo* mutation's origin parent (Supplementary Fig. S1).

3 Results

Measuring accuracy is challenging for *de novo* variant phasing due to a lack of high-confidence truth sets. However, *de novo* variants from the second generation of a large three-generation pedigree, (Dausset *et al.*, 1990) sequenced to 30× coverage and phased using haplotype sharing through all three generations (Sasani *et al.*, 2019, Supplementary Fig. S2), contributed a powerful validation set, with large numbers of third-generation offspring ensuring variant transmission. Unfazed reported a parent-of-origin determination for 1210 out of 4370 second-generation *de novo* SNVs/INDELs whose origin parent was known from haplotype sharing, while confident phase-informative sites were not found for the remaining 2260 variants. The Unfazed prediction was correct in 1207 cases (99.75%). Unfazed phased 7902 of 28 583 unique SNVs/INDELs in both the second and third generations, yielding a phase rate of 27.6% (compared to 21% with un-extended read-based phasing). Unfazed achieved a phase rate of 40% when applied to a large set of *de novo* SVs (Belyeu *et al.*, 2020a).

Command-line example: `unfazed -d denovos.vcf -s snvs.vcf -p pedigree -b bam_directory`.

4 Discussion

Unfazed is a simple tool for variant phasing with Illumina sequencing reads, with a unique focus on determining the origin parent of *de novo* variants of any size. Unfazed combines ease-of-use and fast runtime with high phase rates for both large and small *de novo* variation. This is accomplished by extending read-based phasing to use distant phase-informative sites and leveraging distinct SV read signatures. We anticipate that this tool will prove highly useful to researchers who investigate the rates, patterns, mechanisms and origins of *de novo* variation.

Funding

This work was supported by awards to A.R.Q. from the US National Human Genomic Research Institute (NIH HG006693, NIH HG009141) and the US National Institute of General Medical Sciences (NIH GM124355).

Conflict of Interest: none declared.

References

Belyeu,J.R. *et al.* (2020a) *De novo* structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.*, **108**, 597–607.

Belyeu,J.R. *et al.* (2020b) Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol.*, **22**, 1–13.

Castel,S.E. *et al.* (2016) Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.*, **7**, 1–6.

Chen,X. *et al.* (2016) Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, **32**, 1220–1222.

Dausset,J. *et al.* (1990) Center d'Etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, **6**, 575–577.

Edge,P. *et al.* (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.*, **27**, 801–812.

Hager,P. *et al.* (2020) SmartPhase: accurate and fast phasing of heterozygous variant pairs for genetic diagnosis of rare diseases. *PLoS Comput. Biol.*, **16**, 1–12.

Jónsson,H. *et al.* (2017) Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature*, **549**, 519–522.

Layer,R.M. *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Martin,M. *et al.* (2016) WhatsHap: fast and accurate read-based phasing. *bioRxiv*, **085050**.

Sasani,T.A. *et al.* (2019) Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife*, 1–24.