European Journal of Human Genetics (2020) 28:790–803
https://doi.org/10.1038/s41431-020-0574-3

ARTICLE

Genotype phasing in pedigrees using whole-genome sequence data

August N. Blackburn[1,2] · Lucy Blondell[1] · Mark Z. Kos[1] · Nicholas B. Blackburn[1] · Juan M. Peralta[1] · Peter T. Stevens[1] · Donna M. Lehman[3] · John Blangero[1] · Harald H. H. Göring[1]

Received: 12 July 2019 / Revised: 4 December 2019 / Accepted: 24 December 2019 / Published online: 29 January 2020

Abstract

Phasing is the process of inferring haplotypes from genotype data. Efficient algorithms and associated software for accurate phasing in pedigrees are needed, especially for populations lacking reference panels of sequenced individuals. We present a novel method for phasing genotypes from whole-genome sequence data in pedigrees, called PULSAR (Phasing Using Lineage Specific Alleles/Rare variants). The method is based on the property that alleles specific to a single founding chromosome within a pedigree are highly informative for identifying haplotypes that are shared identical by descent. Simulation studies are used to assess the performance of PULSAR with various pedigree sizes and structures, and the effect of genotyping errors and the presence of nonsequenced individuals is investigated. In pedigrees with complete sequencing and realistic genotyping error rates, PULSAR correctly phases >99.9% of heterozygous genotypes, excluding sites at which all individuals are heterozygous, and does so with a switch error rate frequently below 10^{-4}. PULSAR is highly accurate, capable of genotype error correction and imputation, and computationally competitive with alternative phasing software applicable to pedigrees. Our method has the significant advantage of not requiring reference panels that are essential for other population-based phasing algorithms. A software implementation of PULSAR is freely available.

Introduction

Haplotypes are combinations of alleles at different polymorphic sites occurring on the same DNA molecule and have been an important tool in genetics research [1]. Haplotypes are useful for imputation of alleles at ungenotyped loci, identification of genomic regions shared identical by descent (IBD), genotype error detection and correction, identification of compound heterozygosity, and analysis of parent-of-origin effects, among many other applications. Haplotypes can also be used instead of alleles at individual variable sites in association testing. At present, the most popular sequencing platforms and associated software packages report genotypes for individual polymorphisms, from which haplotypes must be inferred algorithmically. *Phasing*, or the process of analyzing known genotypes to infer haplotypes, is an important concern in genetics and an area of active research [2].

In principle, pedigrees should facilitate highly accurate estimation of haplotypes. In contrast to unrelated (or distantly related) individuals, pedigrees provide considerable additional information for phasing. The transmission of alleles from one generation to the next is directly observed in pedigrees and inheritance patterns can be used to establish phase empirically. Although pedigree-based studies have not been popular for association-based gene mapping of complex traits, this is changing with the emerging interest in rare variants, which are arguably more easily and more powerfully studied in family data. Indeed, as the trend toward sequencing greater numbers of subjects in a study population leads—if simply by chance—to the inclusion of ever more closely related individuals [3], knowledge of the pedigree relationships becomes increasingly essential. Furthermore, phasing of pedigree data need not require the allocation and processing of samples to form a population

Supplementary information The online version of this article (https://doi.org/10.1038/s41431-020-0574-3) contains supplementary material, which is available to authorized users.

✉ Lucy Blondell
lucy.blondell@utrgv.edu

[1] Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Brownsville, TX, USA

[2] Department of Biological Sciences, St. Mary's University, San Antonio, TX, USA

[3] Department of Medicine, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

reference panel and can be performed with a smaller number of study participants.

Pedigree data can introduce unique computational challenges, however. Currently, there are limited options for phasing pedigrees, especially in the absence of reference panels (which is typical for most species other than humans). Consequently, there is a need for computational methods and easy-to-use software optimized for phasing related individuals in pedigrees using whole-genome sequence (WGS) data. Here we describe a novel and fast algorithm, PULSAR (Phasing Using Lineage Specific Alleles/Rare variants), that phases WGS data in families. We demonstrate the utility of the algorithm and software implementation using simulated data as well as real WGS data, and compare its performance with alternative approaches to phasing, including long-range phasing (LRP).

Background

The key step in pedigree-based phasing is identification of large haplotype segments that are IBD within the pedigree. The fundamental premise of PULSAR is the supposition that alleles specific to a single founding chromosome within a pedigree, here referred to as lineage-specific alleles (LSAs), are highly informative for inferring phase and are sufficiently abundant in the genomes of humans (and in many other species) to do so with high confidence. In other words, our method focuses on variants for which a single founder is heterozygous and all other founders are homozygous wild-type. In noninbred individuals each chromosome carries many rare variants, many of which will be LSAs within a pedigree and can be interrogated with whole-genome sequencing. The abundance of LSAs allows one to observe large haplotype segments shared IBD within pedigrees empirically, simplifying the phasing effort compared with inferring the inheritance of common alleles. Non-LSAs provide additional information for elucidating the inheritance of large haplotype segments, but it is not necessary to use non-LSAs to identify these large segments with confidence. As non-LSAs tend to be more common variants they can be mapped to the haplotype segments after the haplotype inheritance has been determined using LSAs.

Our phasing approach is similar in one respect to the technique of LRP [4], in that both PULSAR and LRP use IBS information to identify genuine IBD sharing. The methods differ, however, in the strategy used to identify IBS sharing, and therefore tend to perform optimally in different contexts. LRP identifies IBD sharing by searching for opposing homozygotes (i.e., loci at which two individuals are homozygous for different alleles) and excluding regions that are *not* IBS. This approach works well with common variants, and is highly accurate when at least one individual sharing a segment IBD is homozygous for a given variant. PULSAR identifies putative IBD sharing by searching for shared LSAs (which are necessarily IBS), and uses IBS information from opposing homozygous genotypes only to extend the haplotype boundaries initially established by the observed inheritance of LSAs. This strategy is most efficient for rare variants, but more common variants may be informative in specific pedigrees.

Phasing strategy

The algorithm used by PULSAR comprises four main stages: (1) identify LSAs, i.e., alleles that are likely to be lineage specific; (2) use these LSAs to identify haplotypes, set their initial boundaries, and trace their pattern of inheritance; (3) extend the estimated boundaries of these haplotypes using the fact that individuals must share at least one allele at loci at which they share a haplotype IBD (this is the idea behind LRP); and (4) comprehensively assign alleles to the established haplotypes. Our method assumes the physical location of variants is known, and optionally makes use of external information regarding allele frequencies. The approach is robust to some degree of genotyping error, provided the genotypes have been called from WGS data of reasonably high accuracy, such as from high- or moderate-coverage rather than low-coverage WGS data. Below we describe each of these steps in more detail.

Identify putative LSAs

When an allele is present in a single chromosome among the founders of a given pedigree, then any direct descendent of that founder who also carries the allele can be assumed to share IBD the chromosomal segment harboring that locus. This inference enables empirical estimation of the inheritance of haplotypes within the pedigree. Exceptions to this rule, such as de novo variation, are expected but assumed infrequent enough not to vitiate the overall strategy.

In many pedigrees not all founders will be sequenced, and thus the initial challenge is identifying those alleles that are specific to a single founding chromosome. (Note that at this point we seek only to identify *putative* LSAs.) In PULSAR, potential LSAs are identified by analyzing the pattern of individuals carrying a given allele, i.e., within a pedigree we search for alleles for which all individuals carrying the allele also have at least one pedigree founder in common. If such is not the case, then clearly not all copies of the allele within the pedigree can be IBD and the allele cannot be an LSA. (Distinct lineages within a pedigree, however, could carry alleles IBD in consequence of having common ancestors external to the observed pedigree structure). Putative LSAs are then checked for Mendelian

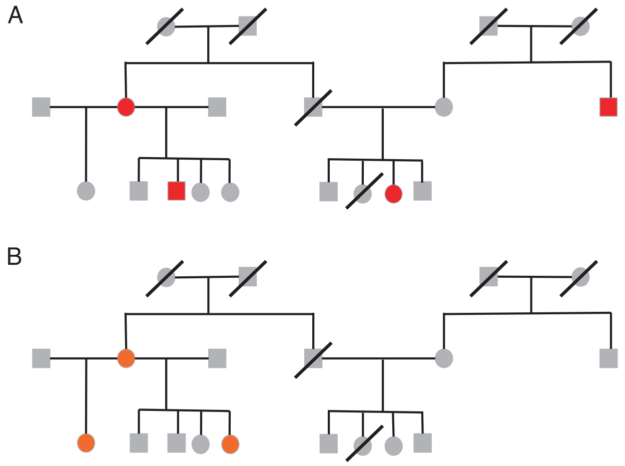

Fig. 1 **Inheritance patterns of lineage-specific alleles. a** A pattern of individuals who share an allele (red) but do not share a common founder (within the limits of the available pedigree information). The shared allele cannot therefore be lineage specific. **b** A pedigree in which the individuals sharing an allele (orange) have two founders in common. The shared allele is potentially lineage specific. [Legend: ■ male; ● female; slash indicates an unsequenced individual].

consistency within their inheritance patterns, i.e., between each founder carrying the allele and the carriers in subsequent generations (at least for those individuals that have sequence data). We also check that the allele is not homozygous in any individual. The algorithm does not presently accommodate the presence of inbreeding loops (wherein LSAs could be homozygous in inbred individuals). These ideas are illustrated in Fig. 1. Figure 1a illustrates a hypothetical pedigree in which individuals carry an allele that *cannot* be lineage specific because the carriers do not share a common founder. Figure 1b illustrates a hypothetical pedigree in which individuals share an allele that is *potentially* an LSA.

Establish haplotype boundaries and haplotype inheritance

Using the set of putative LSAs identified in the previous step, we establish boundaries within which haplotypes are shared IBD in a set of related individuals. At this point the putative LSAs may include false positives, but for now we assume we have only true LSAs and relax this assumption later. Assuming the absence of meiotic recombination between the loci, de novo variation, and genotyping error, the set of individuals carrying a true LSA must share alleles IBD at nearby loci. Conversely, a change in the set of individuals sharing an LSA from one locus to the next indicates one or more recombination events within the meioses that gave rise to the haplotype lineage. By tracking those individuals who share neighboring LSAs along the chromosome, the boundaries of a given haplotype are determined approximately. The procedure is straightforward, but complications arise because at any given region in the diploid genome all individuals carry two sets of haplotypes, one derived maternally and the other paternally. It is therefore necessary to track two separate haplotypes simultaneously. This approach is implemented in PULSAR using a rule-based algorithm to identify changes in the pattern of LSA sharing along a chromosome.

It will often be the case that not all individuals in a given pedigree are sequenced, and consequently one may expect false positives among the putative LSAs identified in the previous step. In other words, some of the putative LSAs are in fact not IBD but merely identical by state (IBS). To reduce the risk of mischaracterizing IBS loci as IBD and inferring incorrect haplotypes, we require a predetermined number of neighboring putative LSAs to be shared before demarcating a new haplotype. A change in the observed inheritance pattern of a small number of neighboring putative LSAs is required to infer, with high confidence, the presence of a true recombination event. This heuristic is reasonable since it is highly unlikely that the same set of individuals will share putative LSAs over a given region of a chromosome if the genotypes are simply a product of chance (such as being IBS, or perhaps due to genotyping error) and not truly lineage specific. For PULSAR, we have required five observations of neighboring putative LSAs showing the same changed pattern (among individuals sharing the putative LSA) before accepting a recombination event and declaring a haplotype change.

Once the haplotypes carried by a proband have been identified based on the pattern of LSA inheritance, we then establish chromosome-wide haplotypes from these smaller haplotype segments. This task comprises two steps. First, if during the course of identifying the haplotypes, a change is observed in a subset of carriers along a particular haplotype segment, we then assume that the resulting two haplotype segments are on the same chromosome in the individual in which the change was observed. Second, for nonfounders having a sufficient number of informative relatives with sequencing data, we determine if the haplotype is shared with the maternal or paternal side of the proband's relatives. Since the proband inherits one chromosome from each parent, haplotypes that are shared with either maternal or paternal relatives are declared to reside on the maternally or paternally inherited chromosome, respectively.

Extend haplotype boundaries

Although the human genome contains a large number of rare variants, many of which will be lineage specific in a given pedigree, the density of LSAs limits the precision with which haplotype boundaries can be determined. We extend the haplotype boundaries observed in a proband with a simple heuristic: individuals sharing a haplotype must also share an allele IBD (and therefore IBS as well) at each locus in the haplotype. This heuristic is central to the strategy behind LRP [4], and our approach is similar in that we also search for opposing homozygous genotypes. The primary difference, however, is that we do not use this rationale to discover shared haplotypes, but only to extend predetermined haplotypes, already determined from the observed LSA patterns, for relatively small (typically <1% of the genome) unresolved segments of the genome. (We examine the effect of the density of LSA coverage in the results). Haplotype extension proceeds in both directions into the unassigned gap between neighboring haplotypes identified in the previous step, and haplotypes are extended only as far as is possible without overlap. In the case of overlapping haplotypes, the region of overlap is not assigned to either haplotype.

Map alleles to haplotypes

After determining the haplotype boundaries and pattern of inheritance throughout the pedigree, the alleles for all variants are mapped onto the haplotypes. This task could be integrated into the prior steps, but in our implementation we assign alleles to haplotypes (including LSAs) as a separate and final step of the procedure. Initially, homozygous genotypes are straightforwardly mapped onto haplotypes, with each of the two haplotypes carrying the same allele. Next, considering each variant independently, we phase heterozygous genotypes in individuals for which at least one allele has been mapped onto one of the two haplotypes carried at that genomic location. This mapping process is iterated, at each step incorporating the mapped alleles from heterozygous genotypes from the previous iteration, until no additional genotypes can be phased. When all individuals within a pedigree are sequenced, and all of the haplotypes carried by each individual are known, then—barring genotyping errors—this procedure can resolve the phase of all combinations of genotypes (except in the case where every individual is heterozygous, a situation that becomes less likely in larger pedigrees).

When multiple individuals carry a given haplotype the information from these individuals is considered jointly when assigning alleles to the haplotype. In the presence of genotyping error, however, it is possible for the genotypes of multiple individuals to provide a conflicting or ambiguous indication as to which allele is carried on a haplotype. (For example, two individuals may share a haplotype yet have opposing homozygous genotypes.) In such cases we apply a majority rule: the allele supported by the majority of carriers is assigned to the haplotype. If resolution of ambiguity requires assignment of new alleles (i.e., alleles not originally observed in an individual), then the revision of genotypes within the considered haplotypes effectively amounts to a correction of genotyping error.

Note that in certain cases there can be no majority support for resolving allelic ambiguity. In trios, for example, no haplotype is shared by more than two people, and no majority can be formed. The same is true for haplotypes shared between two or fewer individuals within larger pedigrees, such as between married-in founders and only one offspring when the lineage (and its LSAs) are not passed to subsequent generations.

Methods

Pedigrees used for simulation studies

We used a variety of simulated sibships and extended pedigrees to investigate the phasing and imputation accuracy of PULSAR. Sibships comprised two parents and 1–7 children. As a variance reduction technique, smaller pedigrees were generated as a subset of larger pedigrees, e.g., a pedigree with one child is formed as a subset of the pedigree with two children, and so forth. Seven large, multigenerational pedigree structures were chosen from the San Antonio Mexican American Family Studies (SAMAFS) [5, 6] to serve as simulation templates. These pedigrees have 11, 14, 20, 28, 55, 78, 94 individuals, comprising 3, 4, 4, 4, 6, 6, 5 generations, and 4, 3, 3, 16, 29, 31, 61 sequenced individuals, respectively. Diagrams of these pedigrees, generated using the application Cranefoot [7], are included in the Supplementary information. To investigate the computational scalability of the software, two additional pedigree structures were simulated and studied: pedigrees having two parents and 1–1000 children, and pedigrees with 2–50 generations in which each generation comprised one offspring and one married-in founder.

Simulation of whole-genome sequence data

For the individuals in the test pedigrees we simulated genotypes and chromosomal haplotypes having many of the characteristics of real data. Whole-genome sequence data for 84 male X chromosomes (excluding the pseudoautosomal regions) from British (GBR) and Finnish (FIN) populations from the 1000 Genomes Project [8] were used as the set of potential founder chromosomes. Use of real male X chromosomes has the advantage that chromosomal haplotypes are known, while preserving the complexities of real whole-genome sequencing data such as the minor allele frequency distribution of variants, linkage disequilibrium, and the existence of small segments shared IBD between distantly related individuals [9], even if the X chromosome differs slightly from the autosomes in these characteristics.

After filtering out pseudoautosomal regions and loci with more than two alleles, there remained 318,912 polymorphic variants in the seed dataset. To accommodate limitations in the software package AlphaPhase1.1 [10], the chromosomes were shortened in simulation to 200,000 variants by considering a smaller section (from 152.23 to 94.37 Mb) of the original chromosome. Although not designed for phasing WGS data, AlphaPhase1.1 provides an important benchmark because it implements the 'long-range phasing' approach to phasing in pedigrees [4].

Within each pedigree, founder chromosomes were sampled randomly without replacement from the 84 X chromosomes, assuming the pedigree founders represent a random sample from the population. Inheritance of haplotypes within pedigrees was simulated by gene-dropping, with a recombination probability between variants corresponding to one recombination per 100 Mb, or roughly the observed recombination rate in humans. Chromosomes not used as founder chromosomes were used to calculate minor allele frequencies (MAF), or were used as a reference panel of additional samples for SHAPEIT2 and Beagle 4.0. Genotyping errors were introduced by choosing genotypes at random and replacing them with one of the two alternative genotypes (for diallelic loci) with equal probability.

Whole-genome sequence data

Whole-genome sequence data have been generated for many participants in the SAMAFS. Single-nucleotide variants and diallelic indels with at least five observations of the minor allele were homogenized and merged from vendor-provided (Complete Genomics, Illumina) sequencing genotype calls from 2330 directly sequenced genomes, resulting in 27,160,796 genetic variants. Some additional, minimal, quality control had been performed on the vendor-provided genotypes beyond simple filtering based on allele count and merging. Illumina sequencing was performed at an average read depth of 30×, with 98% of the not-N Refseq [11] coverage being ≥20×. (These data are available through dbGaP Study Accession: phs000462.v1.p1.) Chromosome 21 (356,545 variants) was selected for analysis. Individuals and variants with genotyping rates less than 99% were excluded, resulting in 354,466 variants. To maximize the comparability of the simulated data with the real data we focused on the same seven pedigrees of SAMAFS for both real and simulated datasets. These seven pedigrees comprise three hundred individuals, of whom 147 are sequenced.

Benchmarking

The performance of PULSAR was compared with AlphaPhase1.1 [10], Beagle 4.0 [9], and SHAPEIT2 [12]. AlphaPhase1.1 is an implementation of the LRP [4] algorithm which phases haplotypes under the assumption that individuals sharing a haplotype IBD must also share at least one allele IBS at each locus in the shared region. Functionally, the method searches for opposing homozygous alleles, the presence of which excludes two individuals from sharing a segment IBD at that genomic location (assuming no de novo variants or genotyping error). LRP is highly accurate when at least one individual sharing the segment IBD is homozygous for a given variant. Beagle 4.0 is a phasing algorithm for population sequencing that is based on a hidden Markov model and can accommodate pedigree information. SHAPEIT2 is a phasing method for singleton individuals and is a variant of the approximate coalescent model. Default settings were used for all three software packages. Pedigree information was supplied for Beagle 4.0 and AlphaPhase1.1, but not for SHAPEIT2.

Three metrics were used to compare the performance of these methods: the switch error rate (SER) [13] to assess phasing accuracy, the proportion of heterozygous genotypes phased, and the GNU *time* command to measure execution times. The SER is a measure of the discrepancy between reconstructed and original haplotypes that is due strictly to misphasing of neighboring heterozygous regions; an SER of zero indicates no phasing error, and an SER of one indicates that no neighboring heterozygous genotypes were correctly phased.

Benchmarking results

Coverage by LSAs

The PULSAR algorithm as described above is based on the idea that variants introduced into a pedigree via a single founding chromosome, i.e., LSAs, can be used to trace the inheritance of haplotypes within the pedigree. For this approach to be practical, it is crucial that LSAs exist in sufficient density in realistic pedigree structures. We have estimated the degree of LSA coverage using real-sequencing data with known phase, specifically the male X-chromosome data from the British (GBR) and Finnish (FIN) cohorts from the 1000 Genomes Project [8]. If we take the pedigree founders to be a random sample of the population, then various key aspects of LSA coverage can easily be determined by permutation for different pedigree structures.

Supplementary Fig. 1 shows the distribution of the number of LSAs per Mb along the X chromosome as a function of the number of pedigree founders for the GBR and FIN cohorts. For the GBR X chromosomes, the median density ranges from 135.4 LSA/Mb in 2-founder pedigrees (median inter-LSA distance of 874 bp, with a maximum of 5.14 Mb which includes a gap in coverage caused by the centromere) to 14.8 LSA/Mb with 15 founders (median distance of 19.0 Kb, with a maximum of 5.39 Mb which includes a gap in coverage caused by the centromere). To estimate the coverage with larger pedigrees, we fit a power function ($R^2 > 0.99$) to the relationship between the number of founders and the median number of LSA/Mb from our simulations. By extrapolating this power function, we estimate that a pedigree with 170 founders would have ~1.0 LSA/Mb. In Finns, a population with decreased genetic diversity due to a series of founder effects [14], there are comparatively fewer LSAs (median 133.1 LSA/Mb in 2-founder pedigrees and 13.4 LSA/Mb with 15 founders). The trend to fewer LSAs in larger pedigrees is reasonable given that the definition of LSA is based on the observation or inference of an allele in a single pedigree founder—thus, as a pedigree contains more founders, fewer polymorphic sites will qualify as LSAs.

Allele frequency distribution of LSAs

The probability of a single founding event in a pedigree depends on the population prevalence of an allele, with rare alleles more likely to be specific to a single founding chromosome. The enrichment for rare alleles among all LSAs will also be greater in pedigrees having more founders, and this is indeed what we observe. For the GBR pedigrees with 2 founders, ~13% of LSAs have MAF less than 5% (median 21.7%), whereas with 15 founders ~91% of LSAs have MAF < 5% (median 2.2%). These observations suggest that filtering on MAF, with allele frequencies estimated from sample data or an independent reference panel, could be an effective means of decreasing the false positive rate when LSAs cannot be identified unambiguously (perhaps due to the presence of unsequenced individuals). In any case, these imputations indicate that for various human pedigree types there will generally be sufficient numbers of LSAs to make them useful as a starting point for phasing WGS data with our algorithm.

Pedigrees with complete sequencing data and no genotyping error

Using the 1000 Genomes Project data, we evaluated the performance of PULSAR, AlphaPhase1.1, Beagle 4.0, and SHAPEIT2.0 in pedigrees of varying size and structure under the ideal scenario in which all individuals are sequenced without error. We chose to examine simulated nuclear families having 1–7 children and seven larger and more complex pedigree structures comprising 11–94 individuals and 3–6 generations. The larger pedigrees are based on actual pedigrees from the SAMAFS [5, 6]. (Diagrams of a seven-child nuclear family, and the seven SAMAFS pedigrees, are provided in Supplementary Fig. 2–9.) Beagle 4.0 and SHAPEIT2.0 were provided with additional reference samples created using male X chromosomes that were not used for simulating the pedigree genotypes. For comparison, Beagle 4.0 was run with and without the reference samples.

Table 1 presents a comparison of the SER [13] and the proportion of heterozygous markers phased for these simulations. Across most simulations PULSAR produced the lowest SER by a considerable margin. Note that PULSAR phased fewer heterozygous markers than either SHAPEIT2.0 or Beagle 4.0 (each of which phases all genotypes), but a higher percentage than AlphaPhase1.1. In the case of nuclear families some markers were heterozygous in all individuals, a situation that is unresolvable without additional data. Excluding such cases, the observed rate of heterozygous genotypes phased by PULSAR was >99.9% of the achievable upper limit. The greatly reduced accuracy of Beagle 4.0 in the absence of reference samples shows the clear need for reference panels in population-based phasing algorithms.

Effect of genotyping error

Table 2 summarizes the SER and the proportion of heterozygous markers phased for those simulations in which the genotyping accuracy is 99.0% (i.e., 1% error and assuming, for simplicity, an equal error rate for all variants). PULSAR, SHAPEIT2.0, and Beagle 4.0 yielded similar SERs. PULSAR yielded a lower SER, and phased a higher proportion of heterozygous markers, than did AlphaPhase1.1.

To further examine the influence of genotyping error on the accuracy of imputation, nuclear pedigree datasets were simulated with genotyping accuracies of 99–100%. For each level of accuracy, 20 datasets were simulated for each pedigree type and analyzed with the results shown in Fig. 2. Significantly, SER increases linearly with the genotyping error rate. Genotyping error has somewhat less impact on SER in pedigrees with more children; in larger pedigrees there is an increased potential for sharing genomic sections IBD, enabling PULSAR to identify haplotypes accurately and more reliably despite errors in genotyping.

Using the haplotypes reported by PULSAR we reconstructed genotypes and compared these with the true (error-free) genotypes. As shown in Fig. 3, error in the reconstructed genotypes increases approximately linearly with the simulated error rate. Note also that the error in genotype reconstruction for the sib trio is comparable with the error in the simulated genotypes; this is expected because (in the trio) no haplotype can be observed more than twice, i.e., the pedigree lacks sufficient individuals to establish a majority haplotype. In fact, the reconstructed

Table 1 Phasing accuracy and heterozygous genotypes phased (simulated data, all individuals sequenced without error).

Pedigree characteristics			Switch error rate (per 1000)					Proportion of heterozygous genotypes phased				
Size	Founders	Generations	PULSAR	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference	PULSAR	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference
Nuclear families												
3	2	2	0.0	41.7	113.0	3.39	13.7	0.8164	0.7425	1	1	1
4	2	2	0.0	84.6	98.3	2.86	3.34	0.8499	0.8193	1	1	1
5	2	2	0.0104	92.0	70.2	1.53	2.22	0.9420	0.8994	1	1	1
6	2	2	0.0128	84.4	65.3	0.834	1.06	0.9887	0.9312	1	1	1
7	2	2	0.0179	89.3	68.9	0.356	0.725	0.9998	0.9398	1	1	1
8	2	2	0.0156	92.8	75.0	0.278	0.517	0.9999	0.9514	1	1	1
9	2	2	0.0168	78.4	77.3	0.296	0.0	0.9999	0.9464	1	1	1
SAMAFS-based pedigrees												
11	5	3	0.0190	55.2	12.2	0.555	4.53	0.9996	0.9794	1	1	1
14	6	4	0.0485	54.6	6.44	0.867	5.52	0.9979	0.9815	1	1	1
20	6	4	0.107	41.0	4.79	0.534	3.29	0.9966	0.9846	1	1	1
28	8	4	0.0343	47.3	1.61	0.321	2.34	0.9989	0.9849	1	1	1
55	17	6	0.115	55.0	0.845	0.677	3.30	0.9989	0.9747	1	1	1
78	21	6	0.715	55.3	0.586	0.346	2.64	0.9973	0.9744	1	1	1
94	28	5	0.264	62.6	0.693	0.596	2.55	0.9966	0.9696	1	1	1

Table 2 Phasing accuracy and heterozygous genotypes phased (simulated data, all individuals sequenced, and genotyping error 1%).

Pedigree characteristics			Switch error rate (per 1000)					Proportion of heterozygous genotypes phased				
Size	Founders	Generations	PULSAR	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference	PULSAR	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/ reference
Nuclear families												
3	2	2	20.5	37.0	370	21.5	18.1	0.8193	0.6140	0.9818	1	1
4	2	2	16.2	77.2	350	16.0	8.37	0.8549	0.7218	0.9938	1	1
5	2	2	10.4	77.4	338	5.38	7.45	0.9429	0.8217	0.9940	1	1
6	2	2	7.75	74.8	338	5.68	5.75	0.9886	0.8355	0.9940	1	1
7	2	2	5.23	76.6	337	3.34	5.15	0.9985	0.8670	0.9936	1	1
8	2	2	2.76	76.2	347	3.73	4.72	0.9992	0.8880	0.9932	1	1
9	2	2	1.74	72.2	352	2.40	3.58	0.9997	0.8960	0.9929	1	1
SAMAFS-based pedigrees												
11	5	3	8.33	69.5	218	10.8	9.95	0.9961	0.8800	0.9999	1	1
14	6	4	8.44	62.2	85.7	9.98	10.4	0.9837	0.9093	0.9999	1	1
20	6	4	6.99	53.3	15.0	5.07	7.20	0.9830	0.9316	1	1	1
28	8	4	5.95	63.6	8.85	4.98	6.08	0.9912	0.9242	1	1	1
55	17	6	7.51	67.1	7.75	7.13	6.67	0.9629	0.9149	1	1	1
78	21	6	9.28	67.3	5.46	4.86	5.69	0.9437	0.9228	1	1	1
94	28	5	8.41	76.0	6.70	6.47	6.05	0.9026	0.9091	1	1	1

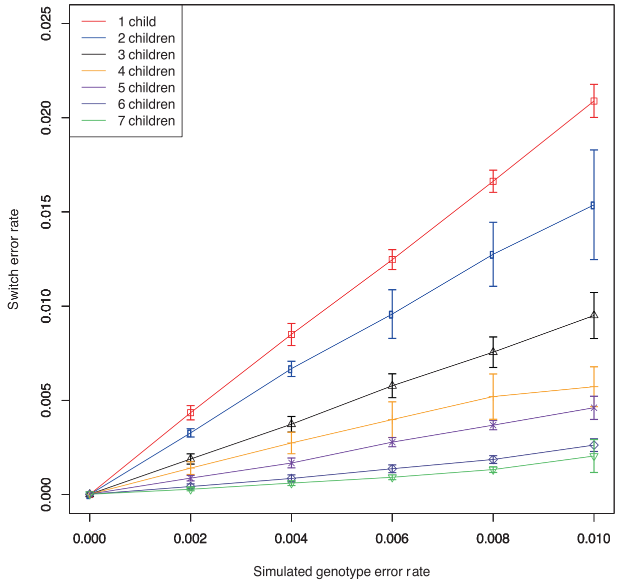

Fig. 2 Effect of genotyping accuracy and IBD sharing (number of sibs) on the switch error rate (the proportion of adjacent heterozygous genotypes correctly phased). SER of zero indicates perfect phasing, SER of one indicates no adjacent heterozygous genotypes were correctly phased. Results are based on 20 simulations in nuclear families having 1–7 children.

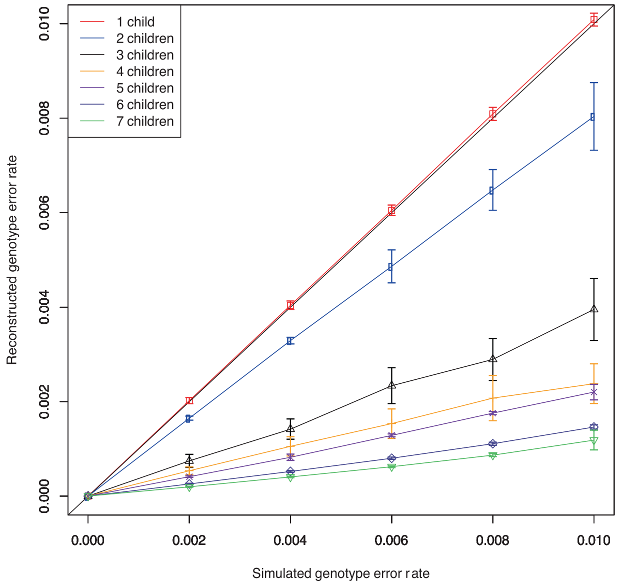

Fig. 3 Effect of IBD sharing on the accuracy of genotype reconstruction according to the haplotypes generated by PULSAR. Results are based on 20 simulations in nuclear families having 1–7 children.

genotype error rate can be slightly higher than the true genotype error rate simply because the algorithm must impute haplotype boundaries, which can introduce a small number of errors. PULSAR is able to correct genotypes in pedigrees with two or more children, performing better in pedigrees having more individuals.

Effect of ungenotyped individuals

Ungenotyped individuals are a common complication in real datasets. For phasing, nonsequenced individuals can affect the false positive rate among putative LSAs identified by PULSAR, which will ultimately decrease the accuracy of reconstructing haplotypes, inferring haplotype boundaries, and estimating haplotype sharing. Moreover, PULSAR will not phase heterozygous genotypes for individuals in genomic regions that are not shared IBD with another individual (unless this happens in consequence of falsely inferring IBD sharing).

Table 3 presents a comparison of the SER and the proportion of heterozygous markers phased for simulations in which genotypes are known without error but with some individuals in each pedigree having no available sequence data. For the seven pedigrees based on the SAMAFS data (11–94 individuals), we recorded which pedigree members had actual sequencing data and used this information to model the presence of unsequenced individuals in our WGS simulations. For the hypothetical nuclear families, we performed two simulations in which one or both parents were sequenced, with the results shown in Table 3. (For comparison, refer to Table 1 for the results for nuclear families with complete sequence data.)

From Table 3 it is evident that the number of ungenotyped individuals and, more significantly, the *location* of the ungenotyped individuals within the pedigree, have pronounced effects on the ability to phase haplotypes. In the case of nuclear families, children missing sequence data have only a small negative effect on the accuracy of PULSAR, as this situation does not lead to an increase in the number of falsely inferred LSAs (although the number of observations of each haplotype is reduced on average). The presence in a nuclear family of children who are missing sequence data is effectively equivalent to a reduction in the size of the nuclear family. As a consequence, a greater number of variants will likely be heterozygous in all the sequenced individuals, although this effect is of practical importance only if the number of sequenced individuals is small. Since PULSAR cannot resolve cases in which all individuals are heterozygous, the main impact of missing children in nuclear families is a decreased percentage of loci that are successfully phased. The effect of missing parents is more significant because this situation creates ambiguity in identifying LSAs. For example, in the simulated nine-member nuclear family, the false positive rate for putative LSAs was 15.9% with one parent missing and 0% with both parents sequenced (data not shown); from Tables 1 and 3 the SER correspondingly increased from 1.68×10^{-5} to 4.04×10^{-2} and the percentage of heterozygous markers phased dropped slightly from 99.99% to 99.45%.

Table 3 Phasing accuracy and heterozygous genotypes phased (simulated data with ungenotyped individuals).

Pedigree characteristics			Switch error rate (per 1000)						Proportion of heterozygous genotypes phased					
Size[a]	Founders[a]	Generations	PULSAR	PULSAR w/MAF filtering[b]	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference	PULSAR	PULSAR w/MAF filtering[b]	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference
Nuclear families														
3 (2)	2 (1)	2	0	0	0	269	7.51	15.9	0.5377	0.5377	0.5377	1	1	1
4 (3)	2 (1)	2	44.7	2.29	114	158	4.01	8.57	0.8324	0.8487	0.7378	1	1	1
5 (4)	2 (1)	2	26.5	2.75	133	118	2.06	4.70	0.9285	0.9336	0.9109	1	1	1
6 (5)	2 (1)	2	11.2	0.0970	139	84.0	1.26	2.80	0.9525	0.9818	0.9089	1	1	1
7 (6)	2 (1)	2	19.1	1.57	143	55.2	0.693	1.91	0.9971	0.9923	0.9598	1	1	1
8 (7)	2 (1)	2	9.52	0.00767	144	43.8	0.617	1.39	0.9983	0.9933	0.9721	1	1	1
9 (8)	2 (1)	2	40.4	1.46	14.1	35.3	0.360	1.14	0.9945	0.9858	0.9443	1	1	1
SAMAFS-based pedigrees														
11 (4)	5 (1)	3	35.8	1.52	82.2	124	9.44	18.8	0.6915	0.5473	0.6822	1	1	1
14 (3)	6 (0)	4	53.0	53.0	53.4	229	22.3	28.1	0.3266	0.3256	0.4128	1	1	1
20 (3)	6 (0)	4	59.5	54.8	45.4	353	44.5	26.8	0.6601	0.6587	0.0040	1	1	1
28 (16)	8 (2)	4	11.1	3.08	63.5	8.58	3.03	4.48	0.9944	0.9988	0.9730	1	1	1
55 (29)	17 (3)	6	20.0	14.4	75.0	8.85	4.99	8.68	0.9788	0.9609	0.8901	1	1	1
78 (31)	21 (5)	6	8.17	5.76	72.1	7.70	4.96	7.01	0.9945	0.9873	0.8989	1	1	1
94 (61)	28 (9)	5	2.54	2.98	65.7	3.11	2.30	6.44	0.9738	0.9726	0.9280	1	1	1

[a]Number in parentheses is number of individuals sequenced.

[b]Allele frequencies estimated from reference samples.

Mitigating the effect of ungenotyped individuals

One approach to mitigating the effect of ungenotyped individuals on the performance of PULSAR is to filter putative LSAs based on their minor allele frequency. Table 3 includes a comparison of the effect of missing individuals in nuclear families when putative LSAs are filtered on the threshold MAF < 5%. In the nine-member nuclear family with one sequenced parent, the rate of falsely inferred LSAs was 15.9% prior to filtering and only 1.50% with filtering (data not shown); from Tables 1 and 3 the SER correspondingly decreased from 4.04×10^{-2} to 1.46×10^{-3}. Filtering of LSAs is not without some compromise, however; in this example, filtering reduced the total number of putative LSAs from 61,160 to 7,850 (data not shown), and decreased the percentage of heterozygous markers phased from 99.45% to 98.58%. A simple metric for quantifying the tradeoff between phasing accuracy (as measured by SER), and the proportion of heterozygous markers phased (correctly or not), is the number of heterozygous markers *phased correctly*, given by the product of phasing accuracy (1-SER) and number of heterozygous markers phased. By this measure, we found that filtering of LSAs clearly improved overall performance, at least in the case of nuclear families (data not shown).

Filtering LSAs according to MAF is most advantageous with smaller pedigrees having fewer founders; in larger pedigrees with many founders, the putative LSAs identified by PULSAR tend to have lower allele frequencies anyway, and filtering on MAF < 5% has little effect. In the largest pedigree considered in this study, for example, in which 61 of 94 (64.9%) individuals have sequence data, prefiltering with MAF < 5% had only a minor effect on the results. The false positive rate for putative LSAs decreased from 0.0099 to 0.0091, the total number of putative LSAs decreased from 44889 to 40060 (data not shown); from Table 3 the SER increased from 2.54×10^{-3} to 2.98×10^{-3}, and the percentage of heterozygous variants phased declined slightly from 97.38% to 97.26%. We have not investigated the effect of different thresholds for MAF, but with larger pedigrees more stringent filtering (i.e., smaller MAF) could well be advantageous.

Performance with missing data and genotyping error

In typical situations, pedigree-based studies must cope both with missing data (e.g., individuals unavailable for sequencing) and genotyping errors. We investigated this situation using the simulated nuclear families and the pedigrees based on SAMAFS data. For the nuclear families the genotypes for one parent were blanked to simulate missing data. Genotypes were simulated with an accuracy of 99%;

as most sequencing platforms can outperform this benchmark easily with appropriate read depth, an accuracy of 99% is somewhat conservative. Results for these simulations are presented in Table 4, from which it is seen that PULSAR yields an SER competitive with SHAPEIT2.0 and Beagle 4.0, and all three methods outperformed AlphaPhase1.1.

Application to real WGS data

Last, we investigated the phasing performance of PULSAR with actual whole-genome sequence data. We used whole-genome sequence genotypes for chromosome 21 from the SAMAFS data, and the same multigenerational pedigrees that were used as template pedigree structures in the other simulations. This test dataset embodied the usual problems and errors inherent with real-world data such as ungenotyped individuals, missing (uncalled) genotypes within sequenced individuals, and genotyping errors. SAMAFS pedigree relationships had been examined previously and corrected as necessary, but a small degree of kinship between some presumably unrelated founders is possible. (Of course, the disadvantage of real data for characterizing the performance of statistical methods is that the true state —mainly with respect to phasing, and to a lesser extent the genotypes—is unknown.) To better characterize the performance of PULSAR with these data, we introduced missing data by blanking genotypes at random with a probability of 1/1000, and then applied PULSAR to the remaining data. The phased haplotypes imputed by PULSAR were used to reconstruct the blanked genotypes, and the concordance of the reconstructed and originally observed genotypes was computed.

The results for this experiment are summarized in Table 5. PULSAR accurately phased nearly all of the observed genotypes (>97% except in one sparsely genotyped pedigree), and the genotypes imputed from the phased haplotypes were in excellent agreement with the originally observed genotypes (>98% concordance). Although fewer than half of the blanked genotypes were phased or imputed, the concordance of the imputed and blanked genotypes ranged from 95.45% to 98.58% across pedigrees, indicating high-phasing accuracy and reliable genotype imputation.

The original SAMAFS sequence data also contained some genuinely missing genotypes, and we found a notable difference between the proportions of artificially missing (blanked) genotypes and truly missing genotypes that PULSAR was able to impute (data not shown). This difference is interpreted as a consequence of the different distributions for the two kinds of missing data. The blanked genotypes were (by construction) distributed randomly, whereas the truly missing genotypes were more likely to be

Table 4 Phasing accuracy and heterozygous genotypes phased (simulated data with missing individuals and genotyping error 1%).

Pedigree characteristics			Switch error rate (per 1000)					Proportion of heterozygous genotypes phased.				
Size[a]	Founders[a]	Generations	PULSAR w/MAF filtering[b]	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference	PULSAR w/MAF filtering[b]	AlphaPhase1.1	Beagle 4.0 w/o reference	Beagle 4.0 w/reference	SHAPEIT2.0 w/reference
Nuclear families												
3 (2)	2 (1)	2	16.4	17.1	401	20.8	19.2	0.5425	0.4950	0.9928	1	1
4 (3)	2 (1)	2	13.7	114	369	18.5	14.6	0.8402	0.6534	0.9966	1	1
5 (4)	2 (1)	2	18.6	125	348	11.4	11.0	0.9318	0.8206	0.9970	1	1
6 (5)	2 (1)	2	11.6	133	331	11.5	8.59	0.9752	0.8620	0.9973	1	1
7 (6)	2 (1)	2	9.84	137	325	10.1	7.13	0.9704	0.8999	0.9971	1	1
8 (7)	2 (1)	2	9.35	136	314	10.5	6.11	0.9826	0.9081	0.9969	1	1
9 (8)	2 (1)	2	18.5	132	320	9.16	5.43	0.9370	0.8879	0.9968	1	1
SAMAFS-based pedigrees												
11 (4)	5 (1)	3	18.5	62.4	248	24.8	23.7	0.5207	0.5331	0.9999	1	1
14 (3)	6 (0)	4	19.6	57.1	247	38.8	29.6	0.3327	0.3166	1	1	1
20 (3)	6 (0)	4	72.5	*	384	62.3	30.3	0.6569	*	1	1	1
28 (16)	8 (2)	4	77.1	75.0	21.0	11.0	8.82	0.9617	0.9026	1	1	1
55 (29)	17 (3)	6	38.3	83.8	20.9	16.9	13.0	0.8734	0.8155	1	1	1
78 (31)	21 (5)	6	19.3	82.2	18.4	14.4	12.7	0.9274	0.8295	1	1	1
94 (59)	28 (9)	5	14.5	78.3	11.0	10.4	9.91	0.8908	0.8393	1	1	1

Asterisk indicates test failed to produce a result.

[a]Number in parentheses is number of individuals sequenced.

[b]Allele frequencies estimated from reference samples.

Table 5 Performance of PULSAR using SAMAFS WGS Data.

Pedigree characteristics			Proportion of observed genotypes phased	Concordance of reconstructed and observed genotypes	Proportion of blanked genotypes imputed	Proportion of blanked genotypes phased/ imputed	Concordance of imputed and blanked genotypes
Size[a]	Founders[a]	Generations					
11 (4)	5 (1)	3	0.9769	0.9948	0.0714	0.1853	0.9813
14 (3)	6 (0)	4	0.9297	1	0	0	*
20 (3)	6 (0)	4	0.9950	0.9859	0.2025	0.4221	0.9545
28 (16)	8 (2)	4	0.9937	0.9899	0.3396	0.4385	0.9798
55 (29)	17 (3)	6	0.9885	0.9873	0.2074	0.2450	0.9689
78 (31)	21 (5)	6	0.9909	0.9943	0.2603	0.3116	0.9858
94 (61)	28 (9)	5	0.9858	0.9941	0.1437	0.2866	0.9823

[a] Number in parentheses is number of individuals sequenced.

Asterisk indicates no genotypes were imputed.

found at the same marker in multiple individuals. This difference can be explained in various ways, but a reasonable hypothesis is simply that the presence of indels or copy number variants acts to disrupt the diploid state of these markers in a subset of related individuals.

Execution times

All software packages were benchmarked on a server with 2.40 GHz Intel Xeon Core i7 CPUs running CentOS Linux 7. Although computation times can become critical with increasing volumes of data, we found all software packages tested to be sufficiently fast for practical use on whole-genome sequence data in the pedigrees we studied. For simulated data on the largest pedigree (94 individuals), AlphaPhase was fastest (93 s), followed by eagle 4.0 (826 s), PULSAR (938 s), and SHAPEIT2.0 (8833 s). SHAPEIT2.0 and Beagle 4.0 were each provided with reference samples as would be expected in typical usage.

Scalability

Computational scalability is an important consideration with any statistical genetic procedure applicable to pedigree data. An algorithm that performs well with nuclear families and small sibships may rapidly become impractical for use with large sibships or deep, extended pedigrees. To estimate the scalability of the PULSAR algorithm we simulated and analyzed sibships comprising two parents and 1–100 children. Such extreme sibships are unrealistic for human populations, although they may exist in other species. This design was chosen simply to enable us to investigate the performance of PULSAR as IBD sharing within the pedigree increases. The results (data not shown) indicate that overall computation time can be well-modeled by a polynomial of quadratic order in the number of individuals, with coefficient 0.48 for the quadratic term. Based on this polynomial fit, we can estimate that a pedigree with 2 parents and 1000 children will require ~41,000× the execution time for a trio comprising two parents and one offspring.

We also studied artificial pedigrees having 2–50 generations, in which each generation comprised one offspring and one married-in founder. This structure was chosen because the IBD sharing within such a pedigree does not, on average, increase between generations. In this case the simulation results disclosed a strongly linear relationship ($R^2 > 0.99$) between the number of individuals and execution time. Based on this result, analysis of a hypothetical pedigree having 500 generations and 999 individuals is expected to require 450× the execution time for a trio comprising two parents and one offspring.

In general, our simulation results indicate that the overall execution time of the PULSAR algorithm scales approximately linearly with pedigree size and approximately quadratically with factors that increase IBD sharing (i.e., number of meioses). These tendencies also affect the relationship between execution time and presence of unsequenced individuals. Unsequenced individuals cannot be used by PULSAR and act generally to reduce execution time, but the precise effect depends to some extent on pedigree size, structure, and the location of the unsequenced individuals within the pedigree.

Discussion

Accurate phasing of genotype data is an essential step in many genetic studies, yet there are currently few tools designed specifically to phase genotypes in pedigrees of arbitrary size and structure. Alternative methods for phasing are typically restricted to use with unrelated individuals or nuclear families, and many require supplemental data in the form of reference panels in order to produce accurate results. While large reference panels are now available for major human ethnic groups, this is not the case for many smaller human populations or for virtually all other species of biomedical relevance.

We have presented a novel algorithm for phasing whole-genome sequence data in a broad range of pedigree sizes and structures, and have described the associated software, PULSAR. Our algorithm yields low SER, and phases a high percentage of heterozygous variants, without the use of additional data in the form of reference panels. The high accuracy of the haplotypes produced by PULSAR, often spanning entire chromosomes, is encouraging. Based on the results of our benchmarking tests, PULSAR is a practical and effective approach to phasing that works well for a range of pedigree sizes and structures provided that a reasonable proportion of the pedigree members is sequenced. The algorithm used in PULSAR can also form the basis of a tool for genotype error checking and correction.

While the method clearly has merit, PULSAR shares some of the limitations typical of other phasing approaches. Unsequenced individuals and/or genotype errors are unavoidable in real datasets and will weaken the performance not only of PULSAR but also of alternative approaches for phasing. However, in our simulations based on real data, in which we examined ‘difficult’ pedigrees with sparse sequencing coverage (e.g., only 3 out of 20 individuals sequenced), PULSAR performed well, erring conservatively on the side of phasing fewer heterozygous markers rather than phasing them incorrectly.

The critical factor in the performance of PULSAR is the proportion of alleles an individual shares IBD with at least one other sequenced individual. Since it is not always straightforward to know what patterns of missing individuals will be detrimental to the performance of PULSAR, we also developed a tool that uses Monte Carlo gene-dropping to estimate the expected proportion of the genome for which each individual will share at least one haplotype IBD with another sequenced individual. This tool can be used for a priori estimation of the applicability of PULSAR to a given pedigree-based dataset. If the sequenced individuals in the pedigree are not predicted to share a high percentage of alleles IBD across the genome, one may wish to consider alternative phasing approaches, such as those designed for unrelated individuals.

Our simulations, using realistic genotypes and chromosomal haplotypes, show that PULSAR is quite robust to genotyping errors. Indeed, PULSAR can be used to correct some kinds of genotyping errors. PULSAR currently ‘fixes’ incorrect genotypes using a simple majority rule within the individuals sharing a haplotype, but the ability to fix genotype errors could potentially be improved by implementing a weighting scheme based on the confidence of genotype calls, or on the number of reads supporting a given allele call.

We have not investigated the impact of errors in the pedigree structures, including the existence of unknown relationships between pedigree founders, on the performance of PULSAR. As a rule, we advocate that pedigree relationships be confirmed or estimated analytically before undertaking any efforts to establish phase, correct genotyping errors, or impute missing genotypes [15, 16].

PULSAR was designed specifically to work on genotype calls generated from high- or medium-coverage WGS data, based on the rationale that the vast number of rare sequence variants in the human genome will favor a high density of LSAs in pedigrees. Low-coverage WGS data leads to less precise genotype calls, particularly for rare variants observed only once or a few times in a given dataset, and PULSAR’s performance with such low-coverage data would be impacted correspondingly. At the present time, and due mainly to cost factors, many studies only involve exome-sequencing data or dense SNP genotyping rather than WGS data, but we have not investigated the performance of PULSAR with data generated on these other genotyping platforms. SNP genotyping panels are biased towards common SNPs, and for this reason we expect that PULSAR would be of limited utility for such data in extended pedigrees, where very few common SNPs would be introduced only once into a given pedigree and thus serve as LSAs. In smaller pedigrees, however, having relatively few founders, we expect the performance of PULSAR to be much less impacted.

One can envisage a number of potential improvements and extensions to PULSAR, but these are quite beyond the scope of the present discussion. However, we did explore the utility of prescreening variants based on allele frequency when identifying putative LSAs. When pedigree founders are not available for sequencing, as is often the case, then such prefiltering based on minor allele frequency was found to be useful for reducing the number of false positives among putative LSAs and thereby improving the performance of the algorithm. Of course, the allele frequency estimates *must* be reliable to be effective as a filtering criterion. With a sufficient number of founders, population allele frequencies can be estimated from the analysis sample; alternatively they can be estimated from independent reference panels. In either case, one must remain cognizant of the effects of admixture, variations in sequencing technology platform, and quality of sequencing, on the resulting estimates.

The algorithm used in PULSAR for inferring phase in pedigrees is a rule-based procedure, but alternative methods based on a maximum-likelihood criterion could be investigated. For example, one might infer phase using the Elston–Stewart algorithm [17] as implemented in LINKAGE [18] or FASTLINK [19]. As a practical consideration, however, the Elston–Stewart algorithm is limited in the number of variants that can be jointly analyzed, which would make it necessary to subdivide the genome into a number of overlapping segments, phase these segments separately, and then assemble the resulting haplotype data. Alternatively, the Lander–Green algorithm [20], as implemented in the software package MERLIN [21], can analyze many variants simultaneously, but is limited in the number of individuals (in a single pedigree) that can be handled. For large pedigrees, one would need to break the pedigree into several overlapping subpedigrees for phasing, then recombine the phased genotypes. Pedigree breaking and combining are not trivial tasks, since joining marker segments of pedigree fragments can lead to Mendelian inconsistencies and other problems. To overcome some of the limitations of the Elston–Stewart and Lander–Green algorithms, a number of Monte Carlo Markov Chain (MCMC) methods have been developed, including LOKI [22] and SimWalk2 [23]. These methods are computationally intensive, however, and not easily applied to whole-genome sequence data. All of these methods are also impacted to varying degrees by genotyping errors and missing genotype data.

There are a number of methods for phasing unrelated individuals [2], some of which, such as Beagle 4.0 [9], are based on hidden Markov models. Currently, the most prominent phasing methods for singleton individuals use variants of the approximate coalescent models, such as in SHAPEIT2.0 [12] and MaCH [24]. The accuracy and computational speed of these packages has greatly improved recently, aided by the availability of ever-larger reference panels for some of the major ethnic groups. One difficulty with applying these methods to pedigree data is that suitable reference panels are not presently available for many smaller populations which are often well-characterized and particularly well-suited for pedigree studies. Without large and accurate reference samples, however, statistical-phasing methods requiring reference panels perform much worse, as seen in the phasing results for Beagle 4.0 in the absence of reference samples. Unfortunately, large and accurate reference panels do not exist for most species.

Yet another simplification that is often made is that of treating family members as unrelated individuals during phasing. This strategy will often generate haplotypes that are inconsistent with Mendelian rules of inheritance. Some packages, such as Beagle 4.0, take family relationships into consideration and can correct for Mendelian inconsistencies. The phasing algorithm duoHMM [25], as implemented in SHAPEIT2 [12], attempts to use the restrictions on inheritance observed in duos to correct the haplotypes produced by statistical phasing, but when applied to the test data described in this study SHAPEIT2 frequently failed to run on most of the simulated pedigree structures.

In view of the various strengths and limitations of the available phasing methods for pedigree data, we see the opportunity for development of a meta-analytic algorithm for combining the phasing results from multiple methods and issuing results that are statistically favored in some well-defined sense. Such an algorithm would weigh the conclusions of different algorithms according not only to the method and assumptions embodied by specific algorithms, but also to features of the sequence data, e.g., sequencing technology and platform, read depth, and so on. Results from a combination of methods may ultimately prove to be more accurate and complete while remaining computationally practical. Alternatively, the phasing results from one algorithm could be refined by additional methods. For example, the phasing output from PULSAR could be used as input for MCMC methods, providing an initial, plausible, phased genotyping state to be refined by an MCMC-based method. Alternatively, the output from statistical phasing of singletons could serve as input for pedigree-based phasing methods, or conversely. Such tandem analyses could harness the high-phasing accuracy achievable by direct observation of inheritance patterns within pedigrees, with statistical-phasing methods that infer phase in parts of the genome where inheritance may not be directly observable given the particular sequencing dataset.

Data availability

PULSAR is available at https://github.com/AugustBlackburn/PULSAR_1.0.

Acknowledgements This work was funded in part by NIH grant R01 DK099051 (to HHHG) and conducted in part in facilities constructed with the support of NIH grant C06 RR020547. The SAMAFS whole-genome sequence data were obtained as part of the T2D-GENES Consortium, supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. Pedigree data were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, supported by NIH grants R01 HL0113323, P01 HL045222, R01 DK047482, and R01 DK053889. We are grateful to the anonymous reviewers for their constructive reviews.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. Nat Rev Genet. 2011;12:215–23.

2. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12:703–14.
3. Ramstetter MD, Shenoy SA, Dyer TD, Lehman DM, Curran JE, Duggirala R, et al. Inferring identical-by-descent sharing of sample ancestors promotes high-resolution relative detection. Am J Hum Genet. 2018;103:30–44.
4. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet. 2008;40:1068–75.
5. Mitchell BD, Kammerer CM, Blangero J, Mahaney MC, Rainwater DL, Dyke B, et al. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. Circulation. 1996;94:2159–70.
6. Hunt KJ, Lehman DM, Arya R, Fowler S, Leach RJ, Göring HHH, et al. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. Diabetes. 2005;54:2655–62.
7. Mäkinen V-P, Parkkonen M, Wessman M, Groop P-H, Kanninen T, Kaski K. High-throughput pedigree drawing. Eur J Hum Genet. 2005;13:987–9.
8. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
9. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
10. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JHJ. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol. 2011;43:12.
11. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Proc Natl Acad Sci USA. 1986;83:4–8.
12. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9:179–81.
13. Lin S, Cutler DJ, Zwick ME, Chakravarti A. Haplotype inference in random population samples. Am J Hum Genet. 2002;71:1129–37.
14. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat Rev Genet. 2000;1:182–90.
15. Sun L, Wilder K, McPeek MS. Enhanced pedigree error detection. Hum Heredity. 2002;54:99–110.
16. Sun L, Dimitromanolakis A. PREST-plus identifies pedigree errors and cryptic relatedness in the GAW18 sample using genome-wide SNP data. BMC Proc. 2014;8:S23.
17. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Heredity. 1971;21:523–42.
18. Lathrop GM, Lalouel JM, Julier C, Ott J. Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA. 1984;81:3443–6.
19. Cottingham RW, Idury RM, Schäffer AA. Faster sequential genetic linkage computations. Am J Hum Genet. 1993;53:252–63.
20. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA. 1987;84:2363–7.
21. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30:97–101.
22. Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet. 1997;61:748–60.
23. Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet. 1996;58:1323–37.
24. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816–34.
25. O’Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10:e1004234.