A Benchmark of Modern Statistical Phasing Methods

Andrew T. Beck[1*], Hyun Min Kang[1,2] and Sebastian Zöllner[1,3]

[1*]Department of Biostatistics, University of Michigan, Ann Arbor, 48108, MI, United States of America.
[2]Center for Statistical Genetics, University of Michigan, Ann Arbor, 48108, MI, United States of America.
[3]Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, 48108, MI, United States of America.

*Corresponding author(s). E-mail(s): beckandy@umich.edu;
Contributing authors: hmkang@umich.edu; szoellne@umich.edu;

Abstract

Modern statistical phasing methods have evolved to efficiently and more accurately infer the phase of haplotypes in large samples. However, these methods introduce errors that can impact downstream analyses. Here we employ both synthetic diploids generated from male X chromosome sequences and Mendelian-resolved trio probands to directly benchmark three commonly-used phasing methods (Beagle 5, SHAPEIT5, and Eagle 2). Switch error rates are found to be lower in SHAPEIT and Beagle, whereas when we distinguish between single switch and double switch errors, we observe a higher rate of double switch errors in SHAPEIT. For all methods we observe an enrichment of errors at both CpG sites and rare variant sites, with a stronger enrichment for double switch errors. Single switch error rates observed in X chromosome synthetic diploids are consistent with those observed in all autosomes other than 8, with higher rates of double switch errors in the autosomes than in X.

1 Introduction

Modern sequencing studies have contributed a wealth of genetic information on humans across a wide range of ancestral backgrounds, enabling discovery of new insights into our understanding of genetic diversity and disease heritability. Among

these efforts are large biobanking initiatives whose sample sizes number in the hundreds of thousands [1, 2]. Such large sample sizes have allowed for the discovery of rare variants and their contributions to disease risk and trait heritability [3, 4]. In tandem with the growth of sample sizes, modern population-based statistical phasing methods have evolved to handle the complex task of efficiently constructing phased haplotypes for many samples. Knowledge of chromosome phase enhances, or in some cases enables additional analyses. For example, imputation can increase the statistical power to detect variant associations by improving SNP coverage, and most imputation methods require phased haplotypes (De Marino, et al). In addition, phased haplotypes are required for the identification of compound heterozygous events, in which the two copies of a gene contain nonidentical mutations. Modern statistical phasing methods such as Eagle, Beagle, and SHAPEIT5 are able to efficiently determine the haplotype phase in thousands to millions of samples to a high precision by sharing information between samples, and the increasing availability of larger samples and diverse, high-quality reference panels has also led to improvements in phasing accuracy. Despite these improvements in both the phasing methods and the sample sizes, statistical phasing remains a challenging task, and modern phasing algorithms still introduce thousands of errors in each phased genome [5]. These errors impact downstream analyses, and therefore a comprehensive analysis of their frequency and the genomic contexts in which they occur more frequently is of considerable interest.

Previous efforts to benchmark phasing methods have relied on the availability of either trio genotyping or sequencing data, or genomes phased using a combination of multiple sequencing platforms and experimental techniques [5]. While the pedigree information allows for the construction of phased haplotypes under the principles of Mendelian inheritance, switch and flip error rates estimated from trio data have been shown to be inflated due to genotyping error [6]. In addition, the cost of obtaining parent-child trios often precludes their widespread use in genetic studies, whereas large samples of unrelated individuals are readily available [7]. A phasing evaluation which leveraged such a large sample would allow for a more detailed characterization of phasing errors across chromosomes. However, the absence of gold-standard haplotypes in this setting does not allow for a direct evaluation of phasing methods. Thus previous efforts to benchmark phasing methods using these large samples of unrelated individuals have relied on imputation quality as a downstream metric of phasing quality [8, 9]. Imputation accuracy is a reflection of the underlying phasing quality and thus allows for a comparison across methods, but it does not allow for a detailed characterization of the genomic context in which phasing errors occur as the number and location of phasing errors are unknown.

A direct evaluation of statistical phasing methods on samples of unrelated individuals that identifies the quantity and location of phasing errors requires true haplotypes to serve as a benchmark. While haplotype phase is initially unknown for most sequenced chromosomes, the phase of male X chromosomes is known outside of the pseudo-autosomal region (PAR). This suggests a strategy of constructing "synthetic diploids" with known phase by sampling two male X chromosomes. These synthetic diploids can then be phased and compared to the known phase to both estimate rates of phasing errors and characterize the locations where errors occur. While limited to

the X chromosome, such an approach allows for the evaluation of phasing in studies without requiring the presence of trios. Large studies of unrelated individuals across diverse populations are readily available, and synthetic diploids offer an avenue by which the rate of statistical phasing errors can be evaluated and characterized with regards to the genomic contexts in which they occur.

In our benchmark of the three methods, we find that Beagle and SHAPEIT introduce fewer errors on average than Eagle. When considering switches and flips separately, we observe that Eagle and Beagle tend to generate more non-flip switches than SHAPEIT, while for flips we observe that SHAPEIT introduces more flip errors on average than either Beagle or Eagle. Comparisons across populations are consistent between the methods, with EAS and SAS having higher switch and flip error rates than the other populations. The distributions of errors across the chromosome are found to be highly correlated amongst the three methods, and we observe a consistent enrichment of errors at both CpG sites and at rare variants, especially for flip errors. Switch error rates for the synthetic diploids in the X chromosome are found to be similar to switch error rates observed in the probands in all autosomes other than chromosome 8, whereas flip rates are elevated in the autosomes relative to the X chromosome. Overall, we provide an up-to-date benchmark of current statistical phasing algorithms used to phase modern large sequencing data sets utilizing an approach that can be used to assess the performance of phasing algorithms in other data sets.

2 Methods

We utilized male X chromosomes to construct synthetic diploids of known phase in order to evaluate three model statistical phasing algorithms [10–12]. The chromosomes were sampled from the 2,504 phase 3 samples of the 1000 Genomes Project (1kGP), allowing for comparisons to be made across distinct ancestral groups. In addition to calculating error rates, we also characterize the genomic context in which errors occur by comparing error densities across the chromosome to other genomic features. Additionally we re-phased the autosomes of the probands from 602 trio pedigrees and evaluated the discordance between the population-based phasing inferred haplotypes and the pedigree-adjusted phase.

2.1 1000 Genomes 30x High-Coverage Sequencing Data

The high-coverage WGS expanded 1kGP cohort consists of the 2,504 original phase 3 1kGP samples, along with 602 trios with probands having had their phase refined using Mendelian inheritance logic. All of these samples are members of one of five super-populations (AFR, AMR, EAS, EUR, SAS), with each super-population consisting of non-overlapping subsets of 26 distinct populations [7]. For the construction of the synthetic chromosome X diploids we used only the 2,504 phase 3 unrelated individuals. From the phased VCF for chromosome X we removed sites that are in either pseudo-autosomal region (PAR1 and PAR2), and we also remove sites that are difficult to sequence according to the 1kGP pilot accessibility mask [13]. Non-SNP variants were also removed, as were SNPs who were either non-biallelic or singletons.

For our analysis of the 602 1kGP trios, we filtered the phased VCF for all autosomes to retain only SNPs that are biallelic, are not singletons, and are not in masked regions of the chromosome in the 1kGP pilot accessibility mask. The unphased VCF is filtered to the same set of variants, and we additionally removed SNPs from the phased VCF which were found to be multiallelic in the unphased VCF.

2.2 Synthetic Diploid Construction and Phasing

Each synthetic diploid consists of two male X haploids sampled from two individual samples from the same population. For each super-population, we first sampled a population from which we then sample two individual male X chromosomes. With the sampled haplotypes, we created a phased VCF file with the known phase and an unphased VCF where at heterozygous sites the order of the two alleles is selected at random. Each synthetic diploid has its own reference panel constructed from the 2,502 samples not used in the synthetic diploid. We constructed 200 synthetic diploids for each of the five 1kGP super-populations in this manner. Each synthetic diploid was phased with Eagle v2.4.1, Beagle 5.4, and SHAPEIT5. For genetic maps, we used the maps provided by the authors of the software [10–12].

2.3 Trio Proband Phasing

For each of the 602 1kGP trios, we re-phased the autosomes of each proband without pedigree-based correction. Each proband was phased individually using a reference panel constructed from the 2,504 unrelated 1kGP phase III samples, excluding any member of the trio.

2.4 Evaluation of Population-based Inferred Phase

For our synthetic X diploids, we identify the location of switch and flip errors by comparing the inferred diploids from each of the three phasing methods to the original haploids used in construction of the synthetic diploid. Using Vcftools (version 0.1.17), we identify the location of any discordance between the inferred phase and the known true phase. Each discordance was classified as either a switch or a flip, where flips are defined as sets of two consecutive errors at consecutive heterozygous positions, while any remaining errors were classified as switches.

3 Results

We simulated synthetic diploids with known phase by sampling male X chromosomes from the 1kGP dataset. Each synthetic diploid was composed of two male X chromosomes sampled from individuals within the same 1kGP population. 200 such synthetic diploids were constructed for each of the five 1kGP super populations (AFR, AMR, EAS, EUR, and SAS). Each synthetic diploid was then individually phased using a reference panel consisting of the 1kGP X chromosomes which were not used in the synthetic diploid (i.e., the 2504 unrelated 1kGP samples minus the two males used to construct the synthetic diploid). The output of each of the three phasing methods was compared to the original male haploids to identify the locations of switch and flip

	AFR	AMR	EAS	EUR	SAS
Beagle	0.0019 (0.0005)	0.0023 (0.0013)	0.0048 (0.0008)	0.0028 (0.0007)	0.0036 (0.0010)
Eagle	0.0024 (0.0006)	0.0028 (0.0013)	0.0056 (0.0008)	0.0033 (0.0008)	0.0043 (0.0010)
SHAPEIT	0.0020 (0.0005)	0.0024 (0.0013)	0.0048 (0.0008)	0.0028 (0.0008)	0.0037 (0.0010)

Table 1 Mean (sd) error rates per heterozygous position by 1kGP population.

errors. Additionally, probands of each of the 602 trios from the extended 1kGP sample were re-phased using a reference panel consisting of the 2,504 unrelated individuals (excluding any member of the trio). The results of this population-based re-phasing to that of the pedigree-corrected phase.

3.1 Distributions of Errors in Synthetic Diploids

The three algorithms reconstructed X haplotypes to a high degree of accuracy, with error rates of 0.0031, 0.0037, and 0.0031 errors per heterozygous position for Beagle, Eagle, and SHAPEIT, respectively. While similar, these error rates were found to be significantly different ($F(2, 1998) = 2542$, $p < 0.001$), with Eagle having a higher error rate than both Beagle (mean difference 5.87×10^{-4}, $t(1998) = 64.10$, $p < 0.001$) and SHAPEIT (mean difference 5.4×10^{-4}, $t(1998) = 59.09$, $p < 0.001$), and SHAPEIT having a higher error rate than Beagle (mean difference 4.6×10^{-5}, $t(1998) = 5.00$, $p < 0.001$). Error rates were also found to differ across populations ($F(4, 995) = 351.58$, $p < 0.001$), with EAS and SAS having the two highest error rates within each of the three methods (Table 1). The differences between methods were also found to be significantly different between populations ($F(8, 1990) = 71.08$, $p < 0.001$), and while many pairwise comparisons of methods remain significant within each population with no change in ranking between methods, the difference between Beagle and SHAPEIT is not significant in the EAS ($t(1990) = 1.42$, $p = 0.331$) and SAS ($t(1990) = 1.51$, $p = 0.286$) populations. Within each method, error rates were found to be significantly different between all pairs of populations.

We next evaluated switch errors and flip errors (consecutive switches) separately. Switch rates were found to be significantly different between the three methods ($F(2, 1998) = 4651.30$, $p < 0.001$), with SHAPEIT (1.57×10^{-3} switches/heterozygous position) having a lower average switch rate than both Beagle (1.70×10^{-3}, $t(1998) = 16.84$, $p < 0.01$) and Eagle (2.29×10^{-3}, $t(1998) = 90.67$, $p < 0.001$), while Beagle was found to have a significantly lower switch rate than Eagle ($t(1998) = 73.82$, $p < 0.001$). Flip rates were also found to vary significantly between the three methods ($F(2, 1998) = 638.99$, $p < 0.001$), but unlike switches, SHAPEIT (1.56×10^{-3} flips/het) was found to have a significantly higher flip rate than Beagle (1.38×10^{-3} flip/het; $t(1998) = 30.57$, $p < 0.001$) and Eagle (1.37×10^{-3} flips/het; $t(1998) = 31.34$, $p < 0.001$), whereas the difference between Beagle and Eagle's flip rates was not significant ($t(1998) = 0.77$, $p = 0.72$). Both switch rates ($F(4, 995) = 382.04$, $p < 0.001$) and flip rates ($F(4, 995) = 256.57$, $p < 0.001$) were found to have significant differences between populations, but only switch rates were found to have a significant difference in method contrasts between populations ($F(8, 1990) = 89.15$, $p < 0.001$). The ordering of the methods by switch rates and flip rates was found to be consistent across

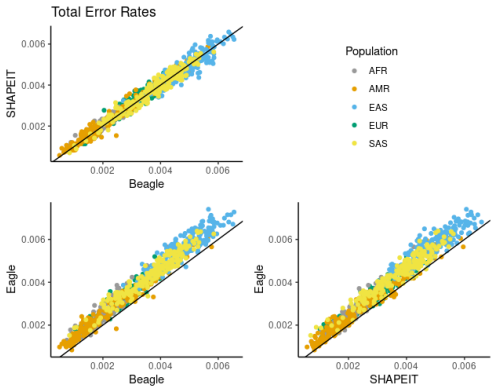

Fig. 1 Error rates observed in each synthetic X chromosome diploid across each pair of methods. Rates are computed as the total number of errors divided by the number of heterozygous positions in each synthetic diploid. Each point represents one synthetic diploid with the x- and y-position determined by the total number of errors observed in each method when comparing the inferred haplotypes to the original male X chromosome haplotypes sampled to generate the synthetic diploid.

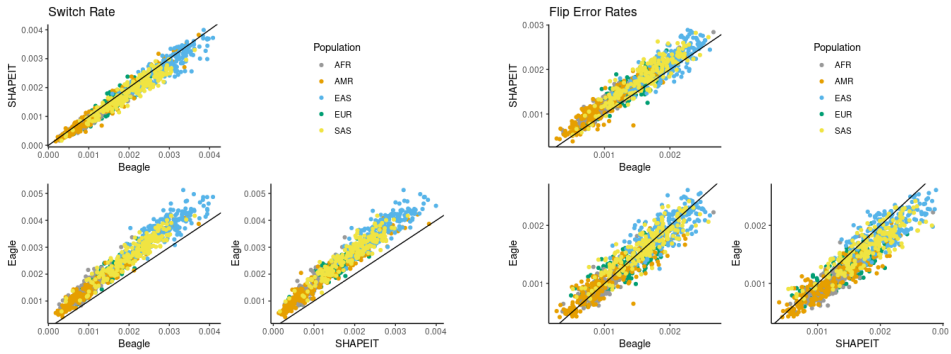

Fig. 2 Switch error rates and flip error rates observed in each synthetic X chromosome diploid across each pair of methods. While for switches a higher error rates is observed in Eagle within most synthetic diploids, a higher flip rate is generally observed in SHAPEIT compared to the other two methods.

the five populations. For both switch and flip errors, the population with the highest average rate is EAS, while the population with the lowest average rate is AFR.

Within individuals, errors were mixed between those shared between methods and those unique to a particular method. Among the 1000 synthetic diploids, on average we observe that 59. 5% of the Beagle switches were also observed in at least one other method, while 63. 9% of the SHAPEIT switches were observed in at least one other method, and 39. 6% of Eagle switches were shared with at least one other method. For flips, on average 65.8% of Beagle's flips are shared with at least one other method, 57.1% of Eagle's flips are shared with at least one other method, and 61% of SHAPEIT's flips are shared with at least one other method.

The distributions of errors for all three methods were heterogeneous across the chromosome, as seen in Figure 3 which displays the mean number of switches and

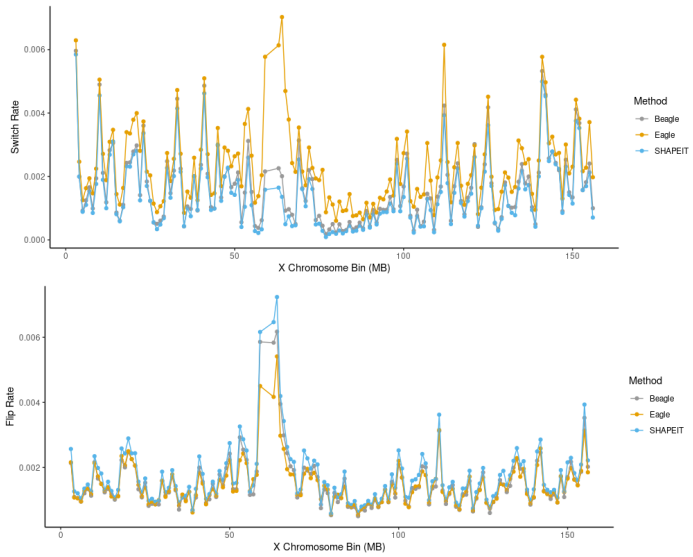

Fig. 3 The rate of switches and flips observed across the X chromosome in megabase sized bins. The rates observed across the chromosome are highly correlated between each pair of methods ($\rho \approx 1$, $p < 0.001$)

flips observed in megabase bins across the X chromosome. Between the methods, the distribution of both switch and flip rates across bins was highly correlated ($\rho \approx 1$, $p < 0.001$). While the switch rates are highly correlated across bins, we observed that Eagle's average rate in each bin was higher than SHAPEIT's in all bins, while also being higher than Beagle's in all but one bin. On average, the switch error rate observed for Eagle within a bin was 1.75 times that of Beagle's and 2.18 times that observed for SHAPEIT (Supplementary Figures A1, A2). In contrast, we observed a higher flip rate in most bins for SHAPEIT compared to Beagle and Eagle, with SHAPEIT's flip rate in each bin being on average 1.14 times that of Beagle and 1.15 times that of Eagle (Supplementary Figures A3, A4).

3.2 CpGs and Rare Variants Are Enriched for In Errors

As a proxy for high mutation rate, we examined the proportion of errors that occured at CpG sites, as these positions are associated with higher mutation rates [14–17]. On average 17% of all errors observed in any of the three methods occurred at CpG sites. For switches, on average 16% of Beagle switches, 19% of Eagle switches, and 15% of SHAPEIT switches occurred at CpG sites, whereas for flips 34% of Beagle flips, 30% of Eagle flips, and 34% of SHAPEIT flips occurred at CpG sites. Comparing the proportion of switch errors observed at CpG sites to the background proportion of heterozygous sites occuring at CpGs, we observed significant enrichment of switch errors at CpG sites for Beagle (mean odds ratio (OR) = 1.13, $t(909) = 13.99$, $p < 0.001$), Eagle (mean OR = 1.36, $t(999) = 42.76$, $p < 0.001$), and SHAPEIT (mean

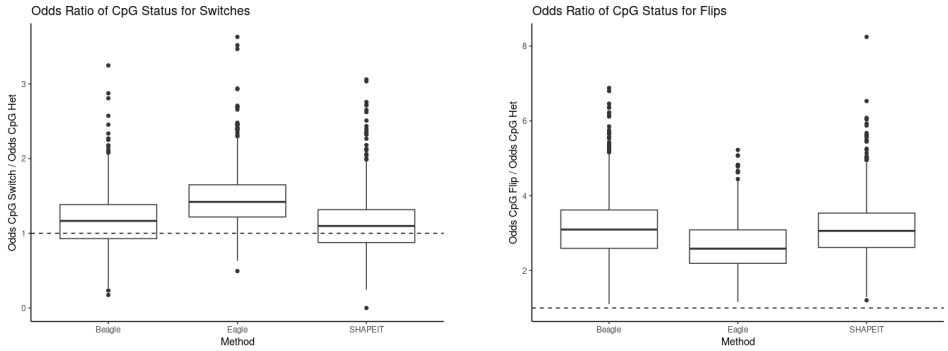

Fig. 4 Average odds ratios of switches and flips occurring at CpG compared to the background distribution odds of a heterozygous position being a CpG. Although significant for all three methods between both types of errors, we observe higher enrichments for Eagle within the switch error category, and an overall higher enrichment for flip errors than for switch errors.

OR = 1.09, $t(999) = 9.49$, $p < 0.001$). For flip errors, we observed an even larger enrichment compared to the same background, with an average OR of 2.39 for Beagle ($t(999) = 109.25$, $p < 0.001$), 2.12 for Eagle ($t(999) = 93.36$, $p < 0.001$), and 2.38 for SHAPEIT ($t(999) = 114.11$, $p < 0.001$). We observed a similar enrichment for flip errors in each method when we evaluated each population separately, but for switch errors we observed a reduced enrichment for switches at CpG sites for the AFR population in Beagle and SHAPEIT (Supplementary Figure A5).

We then examined the contribution of rare variants to phasing errors by comparing the distributions of minor allele frequencies (maf) at phasing errors to the maf distributions at heterozygous positions within each synthetic diploid. We categorized variants as either rare (maf ≤ 0.05), uncommmon ($0.05 <$ maf ≤ 0.1), or common (maf > 0.1) For both switches and flips, we observed a higher proportion of errors at rare variants than the proportion of heterozygous positions observed at rare variants (Supplementary Tables B1 and B2). The average odds ratio for switch errors to be at a rare variant relative to the odds of a heterozygous position within a synthetic diploid being rare were significantly greater than one for all three methods, with Beagle having an average odds ratio of 4.85 ($t = 57.55$, $p < 0.001$), Eagle 8.11 ($t = 79.20$, $p < 0.001$), and SHAPEIT 4.44 ($t = 57.73$, $p < 0.001$). We observed an even larger enrichment of flip errors in rare variants across all three methods, with the average odds ratio for Beagle flips being 24.45 ($t = 63.33$, $p < 0.001$), Eagle 23.79 ($t = 63.30$, $p < 0.001$), and SHAPEIT 23.76 ($t = 63.89$, $p < 0.001$)

3.3 Rephasing of Trio Probands

In addition to the synthetic diploids constructed from sampled X chromosomes, we re-phased the probands from 602 trios included as part of the expanded 1kGP deep sequence release and compared the inferred phase to the Mendelian inheritance adjusted phase. The average error rate across the autosomes was found to be higher than what was observed in X for Beagle ($t(1661.5) = 85.202$, $p < 0.001$), Eagle

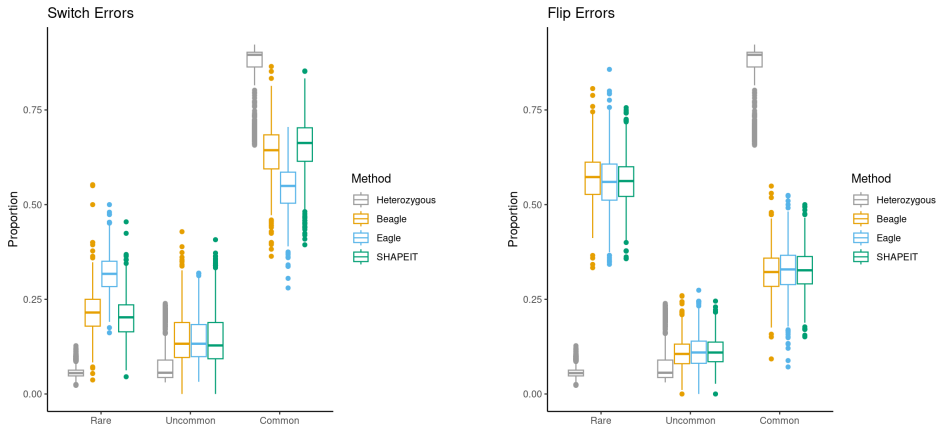

Fig. 5 Distributions of proportions of heterozyogous positions and error locations across minor allele frequency (maf) bins. In the left panel we compare the switch maf distributions to the background heterozygous position maf distributions, while on the right we do the same for flip error maf distributions. In both types of error we observe higher proportions of errors across all three methods in the rare category compared to the background distribution.

($t(1526.6) = 62.034$, $p < 0.001$), and SHAPEIT ($t(1661.5) = 85.202$, $p < 0.001$). Autosomal error rates were generally consistent, except for chromosome 8, which exhibited the lowest average error rate. Excluding chromosome 8, autosomal error rates varied from 0.00103 to 0.0404 errors per heterozygous position. The rankings of methods by error rates were also consistent over all autosomes excluding 8, with SHAPEIT consistently demonstrating the highest error rate. In contrast, we observe a higher error rate from Eagle in both chromosome 8 and X. Stratified by 1kGP population, consistent patterns of rankings by error rates are observed, with SAS and EAS having higher overall error rates than the other populations, with AFR having the lowest error rates (Supplementary Figure A7).

Compared to the overall error rates, the switch error rates observed among the autosomes, excluding chromosome 8, are more similar to what we observed for the synthetic diploids in chromosome X. In contrast, the flip rates across the autosomes are predominantly larger than what we observed in chromosome X, with the exception of chromosome 8. While there is a significant difference in the switch rates among all the chromosome ($F(22, 9328) = 1569$, $p < 0.001$)

4 Discussion

Statistical phasing allows for the inference of haploid phase using data where only genotypes are known in advance, and allows for downstream analyses such as imputation and the investigation of compound heterozygosity. While methods exist to generate haploids with known phase experimental, readily available genotype and sequencing data for hundreds of thousands of individuals facilitate higher power analyses using statistically phased haplotypes. However, statistical phasing is not perfect, and errors in phasing impacts our ability to use these haplotypes in downstream analyses. With

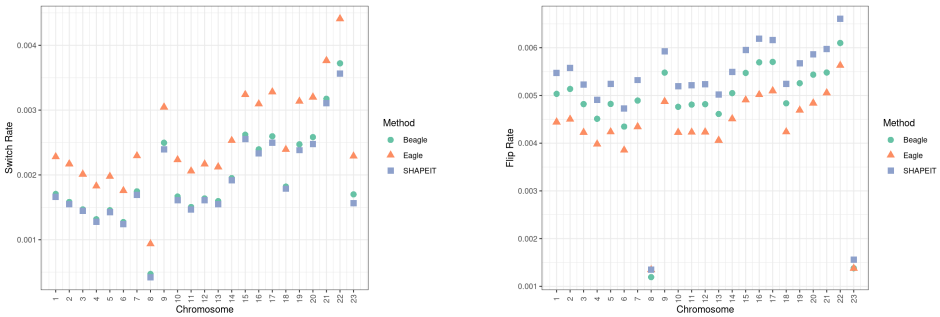

Fig. 6 Average switch and flip rates observed in the proband rephased autosomes and the X chromosome synthetic diploids.

the availability of multiple methods which can efficiently phase large samples, it is of interest to compare how accurately these methods are able to phase haplotypes. While family sequencing trios offer an avenue by which gold standard haplotypes can be generated, large sequencing studies are rich in haplotypes of known phase from male X chromosomes. Here, we leveraged male X chromosomes from the 1kGP deep-sequence release to construct synthetic diploids of known phase across with which we compared the accuracy of three modern statistical phasing methods. In addition, we evaluated performance not only across methods, but also between broad population groups.

In our evaluations of phasing errors in our synthetic diploids, we find that SHAPEIT tends to outperform both Eagle and Beagle in regards to the total number of errors. This is especially the case when we consider switch errors, those which result in the phase at the switch site and at least one subsequent heterozygous position are incorrectly phased with respect to the heterozygous position preceding the error. On the other hand, SHAPEIT tends to introduce more flip errors, i.e. those in which errors occur at consecutive heterozygous positions, leaving the segments on either side of the flip point correctly phased. Both switches and flips are likely to impact downstream analyses of haplotypes, but by disrupting linkage disequilibrium (LD) patterns, switch errors are more likely to have a negative impact on procedures such as imputation that rely on matching haplotype segments between a target sample and a reference panel.

The synthetic diploids also allow for use to evaluate the genomic context among regions where errors occur. The strong enrichment of flip errors at rare variant sites is consistent with the observation that rare variants are more difficult to phase [18], while the enrichment at CpG sites is likely a result of the impact of recurrent mutation on phasing accuracy. CpG sites and rare variants had less of an impact on switch error rates, but across methods we observed a higher enrichment for both of these factors for Eagle, suggesting that these features could identify regions of the genome in which we'd expect to see differential accuracy between Eagle and Beagle/SHAPEIT. Further characterization of the genomic context in which errors occur could yield important insights towards improving the results of statistical phasing methods.

While the ready availability of male X chromosomes from large samples of unrelated individuals offers a wealth of information with which we can evaluate the performance of statistical phasing methods, it only allows for comparisons of the frequency of error

generation in phasing two X chromosome haploids. Here we also evaluate modern statistical phasing methods on the autosomes using probands from trios, allowing for the comparison of strictly population-based phasing to Mendelian-inheritance logic adjusted phase. While we observe an inflated total error rate in most of the autosomes relative to what is observed in the synthetic diploid X chromosomes, this is largely driven by an increased rate of flip errors in the autosomes, whereas switch errors remain consistent with what we observed in the X chromosome. Additionally, the rankings of methods by switch error rate is consistent across the autosomes and X, with SHAPEIT and Beagle having lower switch error rates. These results suggest that the switch error rates in other datasets could be estimated by utilizing a similar X chromosome synthetic diploid approach, although it remains to be determined if the inflated flip error rate in the autosomes is a product of the procedure in which probands were re-phased and then compared to the pedigree-adjusted phase, or if the generation of synthetic X chromosome diploids lends to a decreased rate of flip errors that does not reflect errors generated in phasing large samples of unrelated individuals.

Here we provide an evaluation of modern statistical phasing methods using synthetic diploids of known phase. This evaluation is timely, as data sets of sequenced individuals of unknown phase grow and the need for phased haplotypes grows with increasing interest in downstream analyses requiring haplotypes of known phase. Our results offer practitioners a view towards the relative performance of three modern methods and a sense of the rate of switch and flip errors across individuals of different ancestries. Additionally, our characterization of the genomic context in which errors occur offers insight into putative error generation processes and ways by which statistically inferred haplotypes may be fine-tuned to reduce errors and increase the size of haplotypes of known phase.

References

[1] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). URL http://dx.doi.org/10.1038/s41586-018-0579-z.

[2] Taliun, D. *et al.* Sequencing of 53, 831 diverse genomes from the nhlbi topmed program. *Nature* **590**, 290–299 (2021). URL http://dx.doi.org/10.1038/s41586-021-03205-y.

[3] Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics* **54**, 263–273 (2022). URL http://dx.doi.org/10.1038/s41588-021-00997-7.

[4] Wang, Q. *et al.* Rare variant contribution to human disease in 281, 104 uk biobank exomes. *Nature* **597**, 527–532 (2021). URL http://dx.doi.org/10.1038/s41586-021-03855-y.

[5] Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLOS Genetics* **14**, e1007308 (2018). URL http://dx.doi.org/10.1371/journal.pgen.1007308.

[6] Browning, B. L. & Browning, S. R. Genotype error biases trio-based estimates of haplotype phase accuracy. *The American Journal of Human Genetics* **109**, 1016–1025 (2022). URL http://dx.doi.org/10.1016/j.ajhg.2022.04.019.

[7] Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022). URL http://dx.doi.org/10.1016/j.cell.2022.08.004.

[8] Stahl, K., Gola, D. & König, I. R. Assessment of imputation quality: Comparison of phasing and imputation algorithms in real data. *Frontiers in Genetics* **12** (2021). URL http://dx.doi.org/10.3389/fgene.2021.724037.

[9] De Marino, A. *et al.* A comparative analysis of current phasing and imputation software. *PLOS ONE* **17**, e0260177 (2022). URL http://dx.doi.org/10.1371/journal.pone.0260177.

[10] Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics* **108**, 1880–1890 (2021). URL http://dx.doi.org/10.1016/j.ajhg.2021.08.005.

[11] Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the uk biobank. *Nature Genetics* **55**, 1243–1249 (2023). URL http://dx.doi.org/10.1038/s41588-023-01415-w.

[12] Loh, P.-R. *et al.* Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics* **48**, 1443–1448 (2016). URL http://dx.doi.org/10.1038/ng.3679.

[13] Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL http://dx.doi.org/10.1038/nature15393.

[14] Huttley, G. A. Modeling the impact of dna methylation on the evolution of brca1 in mammals. *Molecular Biology and Evolution* **21**, 1760–1768 (2004). URL http://dx.doi.org/10.1093/molbev/msh187.

[15] Ying, H. & Huttley, G. Exploiting cpg hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome Biology and Evolution* **3**, 938–949 (2011). URL http://dx.doi.org/10.1093/gbe/evr021.

[16] Tomatsu, S. *et al.* Methylation patterns of the human -glucuronidase gene locus: Boundaries of methylation and general implications for frequent point mutations at cpg dinucleotides. *Genomics* **79**, 363–375 (2002). URL http://dx.doi.org/10.1006/geno.2002.6706.

[17] Tsunoyama, K., Bellgard, M. I. & Gojobori, T. Intragenic variation of synonymous substitution rates is caused by nonrandom mutations at methylated cpg. *Journal of Molecular Evolution* **53**, 456–464 (2001). URL http://dx.doi.org/10.1007/s002390010235.

[18] Wertenbroek, R., Hofmeister, R. J., Xenarios, I., Thoma, Y. & Delaneau, O. Improving population scale statistical phasing with whole-genome sequencing data. *PLOS Genetics* **20**, e1011092 (2024). URL http://dx.doi.org/10.1371/journal.pgen.1011092.

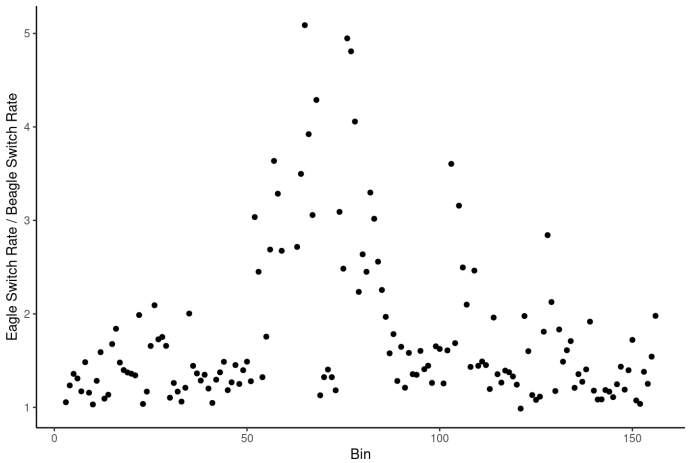

Fig. A1 Ratio of average Eagle switch rates to Beagle switch rates in megabase bins along the X chromosome. On average each bin's average switch rate for Eagle was 1.75 times that of Beagle

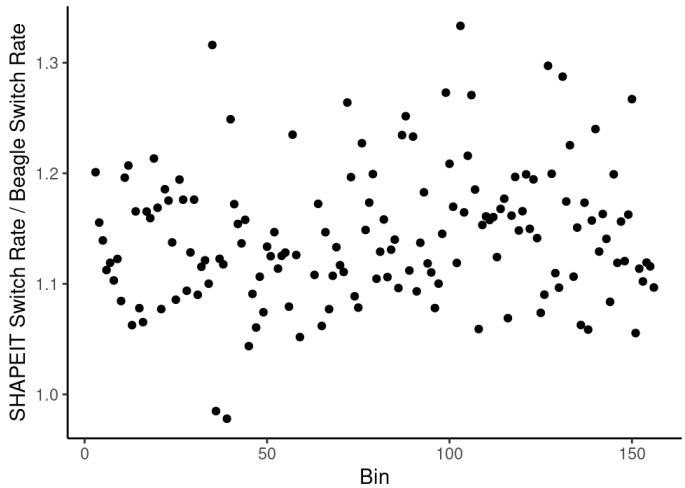

Fig. A2 Ratio of average Eagle switch rates to SHAPEIT switch rates in megabase bins along the X chromosome. On average each bin's average switch rate for Eagle was 2.18 times that of SHAPEIT

Appendix A Supplementary Figures

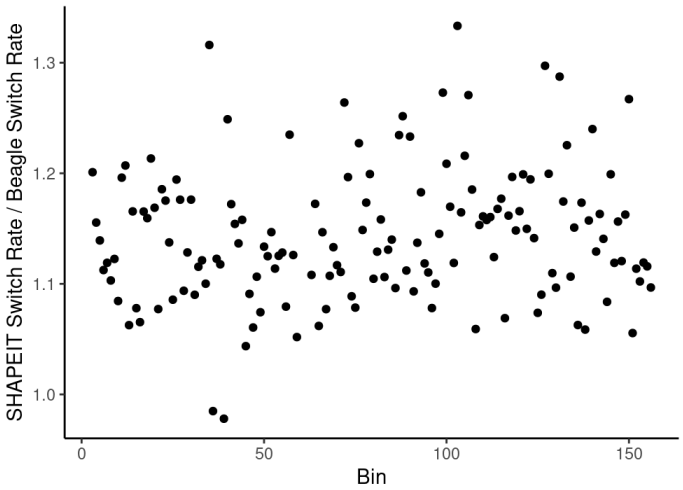

Fig. A3 Ratio of average SHAPEIT flip rates to Beagle flip rates in megabase bins along the X chromosome. On average each bin's average flip rate for SHAPEIT was 1.14 times that of Beagle

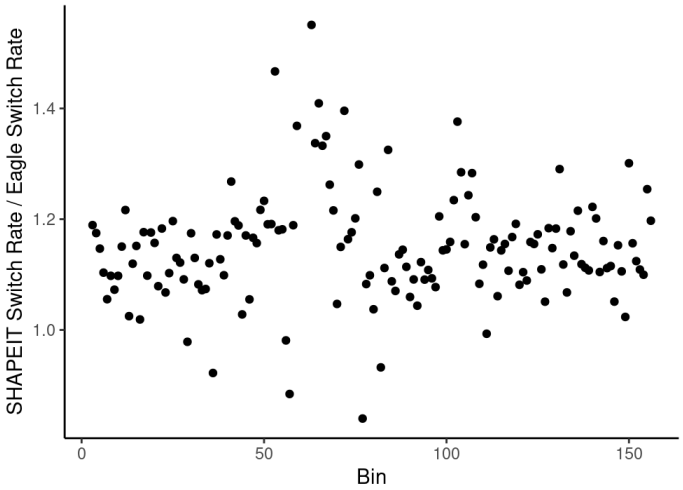

Fig. A4 Ratio of average SHAPEIT flip rates to Eagle flip rates in megabase bins along the X chromosome. On average each bin's average switch rate for SHAPEIT was 1.15 times that of Eagle

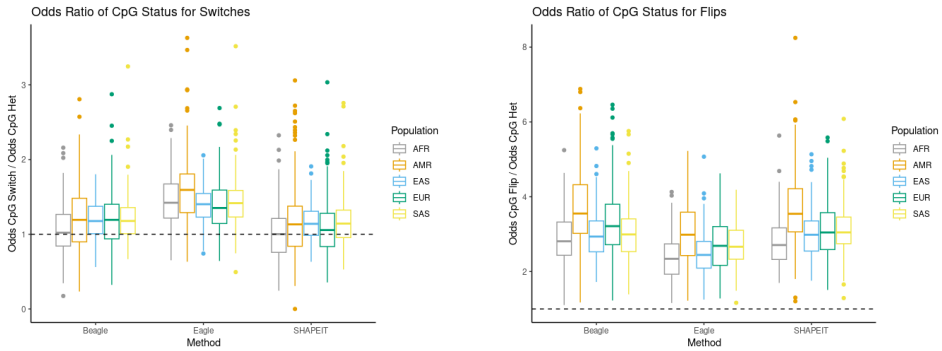

Fig. A5 Average odds ratios of switches and flips occurring at CpG compared to the background distribution odds of a heterozygous position being a CpG, stratified by 1kGP super population.

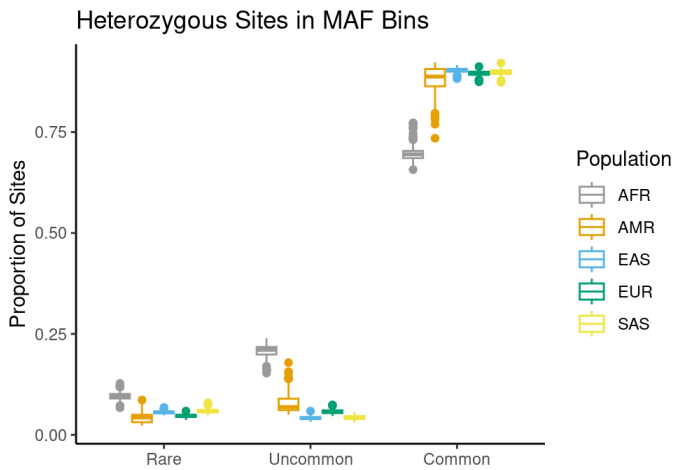

Fig. A6 Distribution of minor allele frequencies at heterozygous sites in X chromosome synthetic diploids by 1kGP population.

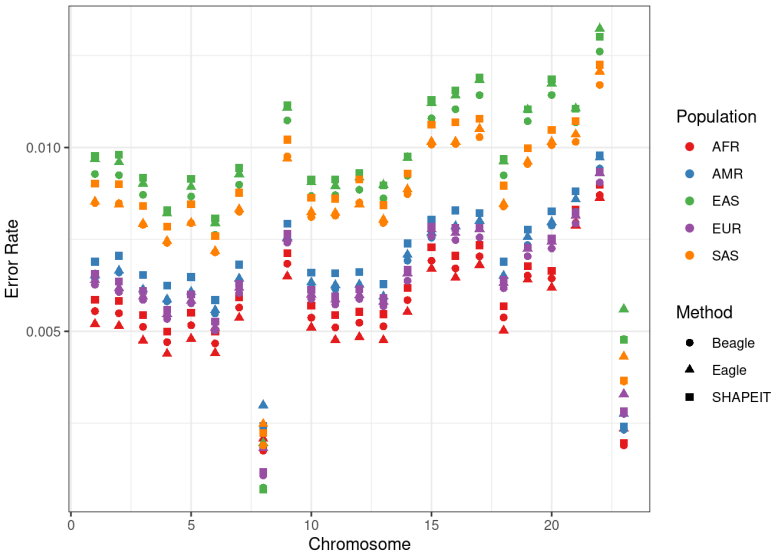

Fig. A7 Average error rate across all chromosomes and X (23) by population and method.

Category	Rare	Uncommon	Common
Background	0.0602	0.0853	0.855
Beagle	0.216	0.148	0.637
Eagle	0.317	0.142	0.541
SHAPEIT	0.202	0.144	0.654

Table B1 Average proportions of heterozygous positions and switch error locations by minor allele frequency bins in X chromosome synthetic diploids. Each synthetic diploid has a proportion of its heterozygous positions and its errors for each method within each of these bins, and here we present the average of these values over all 1000 synthetic diploids.

Category	Rare	Uncommon	Common
Background	0.0602	0.0853	0.855
Beagle	0.569	0.108	0.323
Eagle	0.56	0.112	0.328
SHAPEIT	0.562	0.112	0.326

Table B2 Average proportions of heterozygous positions and flip error locations by minor allele frequency bins in X chromosome synthetic diploids. Each synthetic diploid has a proportion of its heterozygous positions and its errors for each method within each of these bins, and here we present the average of these values over all 1000 synthetic diploids.

Appendix B Supplementary Tables

Chromosome	Method	estimate	std.error	df	statistic	adj.p.value
1	Beagle	6.05E-06	4.39E-05	2067.936	0.1379321	1
2	Beagle	-1.18E-04	4.39E-05	2067.936	-2.682988	0.509086
3	Beagle	-2.33E-04	4.39E-05	2067.936	-5.3036684	0.00003075944
4	Beagle	-3.84E-04	4.39E-05	2067.936	-8.7445067	0
5	Beagle	-2.45E-04	4.39E-05	2067.936	-5.5883773	0.000006438516
6	Beagle	-4.27E-04	4.39E-05	2067.936	-9.7370398	0
7	Beagle	4.63E-05	4.39E-05	2067.936	1.0558307	0.9999935
8	Beagle	-1.23E-03	4.39E-05	2067.936	-27.9974873	0
9	Beagle	7.96E-04	4.39E-05	2067.936	18.1477958	0
10	Beagle	-3.30E-05	4.39E-05	2067.936	-0.7519545	1
11	Beagle	-1.95E-04	4.39E-05	2067.936	-4.4416657	0.002080135
12	Beagle	-6.27E-05	4.39E-05	2067.936	-1.4289566	0.999052
13	Beagle	-1.05E-04	4.39E-05	2067.936	-2.389463	0.737284
14	Beagle	2.50E-04	4.39E-05	2067.936	5.6992767	0.000003425662
15	Beagle	9.20E-04	4.39E-05	2067.936	20.9594343	0
16	Beagle	6.95E-04	4.39E-05	2067.936	15.8318406	0
17	Beagle	8.97E-04	4.39E-05	2067.936	20.4323871	0
18	Beagle	1.19E-04	4.39E-05	2067.936	2.7216113	0.4784671
19	Beagle	7.72E-04	4.39E-05	2067.936	17.6014712	0
20	Beagle	8.83E-04	4.39E-05	2067.936	20.1332684	0
21	Beagle	1.47E-03	4.39E-05	2067.936	33.599325	0
22	Beagle	2.02E-03	4.39E-05	2067.936	46.0793141	0

Chromosome	Method	estimate	std.error	df	statistic	adj.p.value
1	Eagle	-1.03E-05	5.06E-05	1941.936	-0.20368774	1
2	Eagle	-1.19E-04	5.06E-05	1941.936	-2.35934156	0.7582929
3	Eagle	-2.78E-04	5.06E-05	1941.936	-5.49261683	0.00001107514
4	Eagle	-4.62E-04	5.06E-05	1941.936	-9.12511328	0
5	Eagle	-3.13E-04	5.06E-05	1941.936	-6.18691326	0.0000001876917
6	Eagle	-5.33E-04	5.06E-05	1941.936	-10.52451182	0
7	Eagle	4.69E-06	5.06E-05	1941.936	0.09270737	1
8	Eagle	-1.35E-03	5.06E-05	1941.936	-26.7441368	0
9	Eagle	7.57E-04	5.06E-05	1941.936	14.94926699	0
10	Eagle	-5.64E-05	5.06E-05	1941.936	-1.11330034	0.9999834
11	Eagle	-2.29E-04	5.06E-05	1941.936	-4.53055468	0.001404393
12	Eagle	-1.23E-04	5.06E-05	1941.936	-2.42878149	0.7087799
13	Eagle	-1.66E-04	5.06E-05	1941.936	-3.2759475	0.1414117
14	Eagle	2.40E-04	5.06E-05	1941.936	4.73921221	0.0005349949
15	Eagle	9.50E-04	5.06E-05	1941.936	18.75251336	0
16	Eagle	8.06E-04	5.06E-05	1941.936	15.90954436	0
17	Eagle	9.90E-04	5.06E-05	1941.936	19.55697216	0
18	Eagle	1.02E-04	5.06E-05	1941.936	2.01310391	0.9338632
19	Eagle	8.46E-04	5.06E-05	1941.936	16.71490914	0
20	Eagle	9.11E-04	5.06E-05	1941.936	17.98706464	0
21	Eagle	1.47E-03	5.06E-05	1941.936	29.11985953	0
22	Eagle	2.12E-03	5.06E-05	1941.936	41.85677564	0

Chromosome	Method	estimate	std.error	df	statistic	adj.p.value
1	SHAPEIT	9.80E-05	4.40E-05	2127.154	2.2283598	0.8400881
2	SHAPEIT	-1.56E-05	4.40E-05	2127.154	-0.353992	1
3	SHAPEIT	-1.21E-04	4.40E-05	2127.154	-2.7552344	0.4521556
4	SHAPEIT	-2.90E-04	4.40E-05	2127.154	-6.5876412	0.00000001423474
5	SHAPEIT	-1.42E-04	4.40E-05	2127.154	-3.2257272	0.1616549
6	SHAPEIT	-3.24E-04	4.40E-05	2127.154	-7.3572285	0.0000000001324301
7	SHAPEIT	1.28E-04	4.40E-05	2127.154	2.9095039	0.338613
8	SHAPEIT	-1.14E-03	4.40E-05	2127.154	-25.9787331	0
9	SHAPEIT	8.28E-04	4.40E-05	2127.154	18.8333616	0
10	SHAPEIT	4.65E-05	4.40E-05	2127.154	1.056835	0.9999934
11	SHAPEIT	-9.84E-05	4.40E-05	2127.154	-2.2384943	0.8343866
12	SHAPEIT	4.32E-05	4.40E-05	2127.154	0.9830908	0.9999982
13	SHAPEIT	-1.68E-05	4.40E-05	2127.154	-0.3821787	1
14	SHAPEIT	3.55E-04	4.40E-05	2127.154	8.0643672	0
15	SHAPEIT	9.87E-04	4.40E-05	2127.154	22.448602	0
16	SHAPEIT	7.69E-04	4.40E-05	2127.154	17.4896767	0
17	SHAPEIT	9.30E-04	4.40E-05	2127.154	21.1528589	0
18	SHAPEIT	2.22E-04	4.40E-05	2127.154	5.0373435	0.0001229692
19	SHAPEIT	8.16E-04	4.40E-05	2127.154	18.5592496	0
20	SHAPEIT	9.10E-04	4.40E-05	2127.154	20.6904978	0
21	SHAPEIT	1.54E-03	4.40E-05	2127.154	35.0336366	0
22	SHAPEIT	2.00E-03	4.40E-05	2127.154	45.4824576	0