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 efficiently and accurately infer haplotype phase in large genomic samples, and their performance is critical for downstream analyses. However, rigorously evaluating phasing accuracy remains challenging. Here we demonstrate the use of synthetic diploids generated from male X chromosome sequences to benchmark and compare three commonly-used phasing methods (Beagle 5.4, SHAPEIT 5.1.0, and Eagle v2.4.1). While errors across methods are highly correlated, we observe overall higher rate of errors in Eagle. When contrasting switch error and flip (double-switch) errors, switch rates are higher in Eagle, whereas flip rates are higher in SHAPEIT. These patterns ae consistent across populations. For all methods we observe an enrichment of errors at both CpG sites and rare variant sites, with a stronger enrichment for flip errors. We compare these error estimates with estimates from Mendelian-resolved trio probands: single switch error rates observed in X chromosome synthetic diploids are consistent with those observed in all autosomes, with higher rates of double switch errors in the autosomes than in X.

1 Introduction

Modern genomic studies have contributed a wealth of genetic information on humans across a wide range of ancestral backgrounds, offering 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 [5–7]. 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 [8]. In addition, phased haplotypes are required for the identification of compound heterozygous events, in which the two copies of a gene contain nonidentical mutations [9, 10]. Modern statistical phasing methods such as Eagle [5], Beagle [6], and SHAPEIT [7] are able to efficiently determine the haplotype phase in thousands to millions of samples with high precision, and the increasing availability of larger and more diverse 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 [11]. 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 genotyped trios, or genomes phased using a combination of multiple sequencing platforms and experimental techniques [8, 11]. While 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 [12]. In addition, the cost of obtaining parent-child samples often precludes their widespread use in genetic studies, whereas large samples of unrelated individuals are readily available [1, 2, 13]. Leveraging such large samples for the evaluation of phasing algorithms enables a more detailed characterization of phasing errors across chromosomes by providing a greater number of examples for analysis. 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, 14]. 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 evaluation of the genomic context in which phasing errors occur as the number and location of phasing errors are unknown.

To directly evaluate statistical phasing methods on unrelated individuals and identify the quantity and location of phasing errors, true haplotypes are required 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). By sampling pairs of male X chromosomes, synthetic diploids with

known phase can be constructed [15]. These synthetic diploids can then be phased and compared to the known true haplotypes, which allows for both the estimation of phasing error rates and a characterization of the locations where these 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.

We used this approach to benchmark three modern statistical phasing software implementations (Beagle 5.4, Eagle 2.4.1, and SHAPEIT 5.1.0), constructing synthetic diploids from male X chromosomes sampled from 2,504 unrelated individuals in the 1000 Genomes Project (1kGP) deep sequence release [13]. Each synthetic diploid was phased individually with all three algorithms, using the remaining 2,502 samples as a reference panel. The inferred haplotypes were then compared to the original haplotypes to identify the location of switch and flip errors. We also re-phased the autosomes of 602 trio probands from the expanded 1kGP cohort to compare the statistically inferred haplotypes with the Mendelian inheritance-adjusted haplotypes.

In our benchmark of the three methods, we found that Beagle and SHAPEIT introduced fewer errors per sample on average (197.7 and 201.2, respectively) than Eagle (236.5). When considering switches and consecutive switches (flips) separately, we observed that Eagle and Beagle on average generated more switches than SHAPEIT, while for flips we observed that SHAPEIT introduced more flip errors than either Beagle or Eagle. Comparisons across populations were found to be 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 were found to be highly correlated among the three methods, and we observed 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 were found to be similar to switch error rates observed in the probands in the autosomes whereas flip rates are elevated in the autosomes relative to the X chromosome. Overall, we showcase an approach that can be used to assess the performance of phasing algorithms in other data sets and provide an up-to-date benchmark of current statistical phasing algorithms used to phase modern large sequencing data sets.

2 Results

We simulated synthetic diploids with known phase by sampling male X chromosomes from the 1kGP dataset and phased each synthetic diploid using Eagle, Beagle and SHAPEIT. We compared each method's inferred chromosomes to the known chromosomes to identify switch and double-switch (flip) errors. To assess consistency across chromosomes, we re-phased the autosomes of probands from each of the 602 trios of the extended 1kGP sample while ignoring the available pedigree information, comparing the results to pedigree-adjusted phased chromosomes.

2.1 Distributions of Errors in Synthetic Diploids

We evaluated the accuracy of three phasing algorithms—Beagle, Eagle, and SHAPEIT—on synthetic diploid chromosome X samples. Across a mean of 68,930 heterozygous sites per sample, all algorithms showed a high degree of accuracy. The mean total error counts per chromosome were lowest for Beagle (197.7), and highest for Eagle (236.5), with SHAPEIT having an average of 201.2 errors per sample. These counts correspond to low mean error rates of 0.0031 for Beagle and SHAPEIT, and 0.0037 for Eagle. Despite a relatively small difference in the mean error rates, one-way repeated measures ANOVA indicated that the error rates differed significantly across the methods ($p < 2.2 \times 10^{-16}$), and all pairs of methods were found to be statistically significantly different.

Across individual samples, we observe substantial variation of the error rate within each method, with samples clustering by 1kGP population (Fig. A1). Comparing across populations, both the mean number of heterozygous sites per sample and error counts differed greatly, with AFR samples having on average the largest number of heterozygous positions at which a phasing error could occur (101673.52 heterozygous sites per sample), whereas EAS and SAS samples had an overall higher mean error count and mean error rate (Table 1). Comparing methods within populations we observed that in all populations Beagle had the lowest error rate, followed by SHAPEIT while Eagle had the largest error rate.

	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.

We next considered switch errors and flip errors separately. Overall, switch errors constituted 54.3% of all observed errors across all samples and methods. This composition varied across methods, with 53% of Beagle's errors, 61% of Eagle's errors, and 47% of SHAPEIT's errors being switch errors. Switch rates differed significantly across the three methods ($p < 2.2 \times 10^{-16}$). SHAPEIT introduced the fewest errors (0.0016 switches per heterozygous site), followed by Beagle (0.0017) and Eagle (0.0023). These

	Switch Error Rates			Flip Error Rates		
	Beagle	Eagle	SHAPEIT	Beagle	Eagle	SHAPEIT
AFR	0.0009 (0.0003)	0.0014 (0.0004)	0.0008 (0.0003)	0.0010 (0.0003)	0.0010 (0.0003)	0.0012 (0.0003)
AMR	0.0012 (0.0008)	0.0017 (0.0009)	0.0011 (0.0008)	0.0011 (0.0005)	0.0011 (0.0004)	0.0013 (0.0005)
EAS	0.0028 (0.0005)	0.0037 (0.0006)	0.0027 (0.0005)	0.0020 (0.0003)	0.0019 (0.0003)	0.0021 (0.0004)
EUR	0.0015 (0.0004)	0.0021 (0.0005)	0.0014 (0.0004)	0.0012 (0.0003)	0.0012 (0.0003)	0.0014 (0.0004)
SAS	0.0020 (0.0006)	0.0027 (0.0007)	0.0019 (0.0006)	0.0016 (0.0004)	0.0016 (0.0004)	0.0018 (0.0004)

Table 2 Mean (sd) switch and flip rates for each 1kGP population.

rates correspond to an average of 98.3, 107.4, and 146.3 switches per sample for SHAPEIT, Beagle, and Eagle, respectively. Similarly, flip rates showed a significant difference across the methods ($p < 2.2 \times 10^{-16}$). In contrast to switch errors, SHAPEIT yielded the highest flip rate (0.0016 flips per heterozygous site), while Beagle (0.0014) and Eagle (0.0014) demonstrated comparable, lower rates. These rates translate to an average of 102.9, 90.4, and 90.2 flips per sample for SHAPEIT, Beagle, and Eagle, respectively.

While all pairs of methods were found to have a significantly different switch rate ($p < 0.0001$), the difference between flip rates for Beagle and Eagle was not found to be significantly different ($p = 0.72$). This ranking of methods by both switch rate and flip rate was observed to be consistent across the five 1kGP populations. Switch rates and flip rates differed significantly among populations ($p < 2.2 \times 10^{-16}$ for both). For both error types, the EAS population exhibited the highest average rate, while the AFR population showed the lowest average rate. (Table 2).

Within individuals, the number of total errors was highly correlated between methods, with correlation coefficients ranging from 0.979 to 0.982. The number of switches and the number of flips were also highly correlated. Specific phasing errors were often shared between methods (i.e. they were the same type of errors occurring at the same location in at least two methods for the same sample). Among the 1000 synthetic diploids, 59. 5% of the Beagle switches were shared with at least one other method, while 63. 9% of the SHAPEIT switches were shared with at least one other method, and 39. 6% of Eagle switches were shared with at least one other method. Overlap in flip errors were generally higher, where we observe that 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 rate of errors for all three methods were heterogeneous across the X chromosome (Figure 2). Calculated in 1mb bins, average errors rates for switches range from $8 \cdot 10^{-5}$ to $7 \cdot 10^{-3}$ switches per heterozygous position while error rates for flips range from $6 \cdot 10^{-4}$ to $7 \cdot 10^{-3}$ flips per heterozygous position. Between methods, the distribution of both switch and flip rates across bins was highly correlated ($\rho \approx 1$). Nonetheless, Eagle's average rate was higher than SHAPEIT's in all bins, and 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 A4, A5). 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 A6, A7).

2.2 Genomic predictors of phasing error

To assess the impact of mutation rate on phasing error, we examined the proportion of errors that occurred at CpG sites, as these sites have on average 10-fold higher mutation rates [16–20]. Across the synthetic diploids, an average of 14.2% of heterozygous sites are CpGs while 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 occurring at CpGs across synthetic diploids, we observed a significant enrichment of switch errors at CpG sites for Beagle (mean difference in proportions = 0.024, $p < 2.2 \times 10^{-16}$), Eagle (0.027, $p < 2.2 \times 10^{-16}$), and SHAPEIT (0.023, $p < 2.2 \times 10^{-16}$). For flip errors, we observed a larger enrichment, with an average difference of proportions from the background rate of CpGs at heterozygous sites of 0.20 for Beagle ($p < 2.2 \times 10^{-16}$), 0.16 for Eagle ($p < 2.2 \times 10^{-16}$), and 0.20 for SHAPEIT ($p < 2.2 \times 10^{-16}$). Comparing across populations, we observed a similar enrichment for flip errors in each method (Supplementary Figure A10).

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 switches, we consider the marker at which the discrepancy between the original haplotype and the inferred haplotype is observed, while for flips, we considered the marker that was assigned to the wrong haplotype. 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 B2 and B3). Both the proportions of switches and flips at rare variants exceeded the background proportion of heterozygous sites at rare variants. On average, the mean difference between the error proportions and the background proportion was notably higher for flip errors (ranging from 0.50 to 0.51 across methods) than for switch errors (ranging from 0.16 to 0.26). Specifically, the average difference (Error Rate − Heterozygous Proportion) for switch errors was 0.16 (Beagle), 0.26 (Eagle), and 0.16 (SHAPEIT), whereas flip errors exhibited differences of 0.51 (Beagle), 0.50 (Eagle), and 0.51 (SHAPEIT). (Supplementary Figure A9).

To examine the relationship between recombination rate and phasing accuracy, we divided the X chromosome into non-overlapping 1 MB bins. Across the 755 resulting bins, we observed a positive correlation between average recombination rate and switch error rates ($r = 0.78 (p < 0.001)$, $r = 0.45 (p < 0.001)$, and $r = 0.81 (p < 0.001)$ for Beagle, Eagle, and SHAPEIT, respectively). In contrast, the correlation between recombination and flip error rates was negative, but not statistically significantly different from 0 ($r = -0.11 (p = 0.19)$, $r = -0.06 (p = 0.44)$, and $r = -0.136 (p = 0.17)$) (Supplementary Figures **??**). We then examined whether switches and flips were

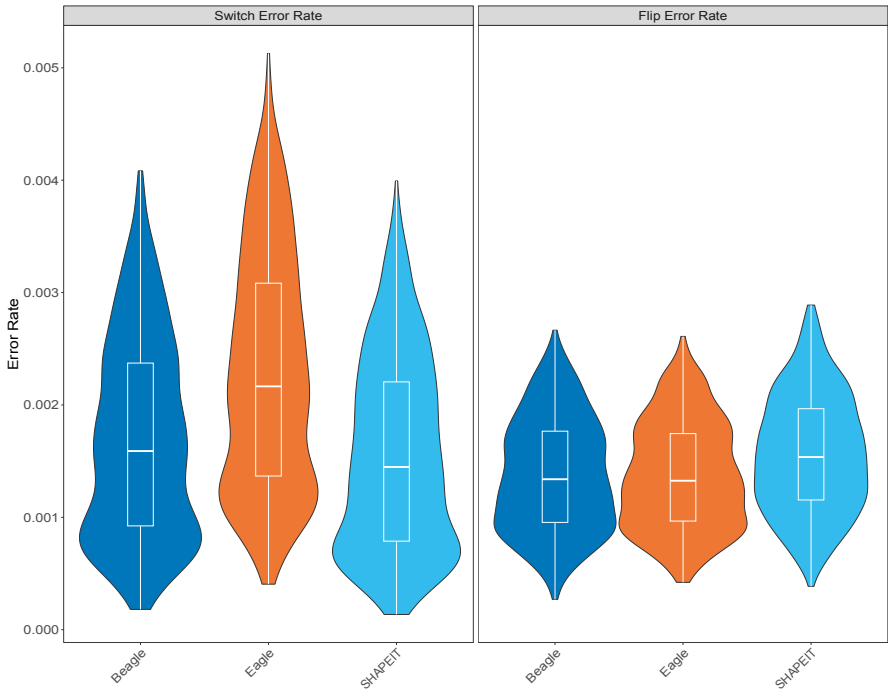

Fig. 1 Switch and flip error rate distributions observed across 1000 synthetic X chromosome diploids.

Fig. 2 The rate of switches and flips observed across the X chromosome in 1Mb sized bins. Rates observed across the chromosome are highly correlated between each pair of methods ($\rho \approx 1$)

enriched in regions with the highest recombination rates. To do so, we assigned each error to a recombination rate decile. For Beagle and SHAPEIT, we observed that on average over 40% of switch errors occurred at heterozygous positions with recombination rates in the top decile, and over 30% of Eagle switch errors on average were also in this top decile. For flips we observed lower enrichment of errors in the top recombination rate decile, but in all methods on average over 15% of flip errors occurred at positions with recombination rates in the top decile (Supplementary Figures **??**).

2.3 Comparison to Autosomes

To evaluate the robustness of our observations and comparisons of methods in chromosome X to other chromosomes, we re-phased the autosomes of 602 trio probands from the 1kGP deep-sequence release while excluding their parents' haplotypes. Compared to the X chromosome, the mean number of heterozygous sites per 10,000 base pairs ranged from 4.9 to 7.1 across the autosomes, whereas in our X synthetic diploids we observed on average of 4.5 heterozygous sites per 10,000 bases (Supplementary Table B4). The average proportion of heterozygous sites at positions with a minor allele frequency less than 0.05 ranged from 0.066 to 0.075 in the autosomes, whereas this average was slightly lower in X (0.060). With regards to the occurrence of heterozygous sites at CpGs, on average the mean proportion of CpG heterozygous sites ranged from 0.142 to 0.212 in the autosomes, while on average the proportion of CpG heterozygous positions in X was 0.142 (Supplementary Figure A19).

Error rates were observed to vary across the autosomes, with Beagle's error rates varying from 0.0065 to 0.0105 errors per heterozygous site, Eagle's from 0.0063 to 0.0105, and SHAPEIT's from 0.0069 to 0.0109. These error rates were all higher than what we had observed in the X chromosome for the synthetic diploids (Beagle: 0.0031 errors per heterozygous site; Eagle: 0.0037; SHAPEIT: 0.0031). This difference was largely due to an increased flip rate in the autosomes relative to X, with switch rates in the autosomes aligning with what was observed in the X chromosome (Figure 5). Average flip rates in the autosomes were observed to vary between 0.0050 and 0.0067 flips per heterozygous site in Beagle, 0.0044 and 0.0060 in Eagle, and 0.0054 and 0.0072 in SHAPEIT, whereas the average flip rates in X were 0.0014, 0.0014, and 0.0016 for Beagle, Eagle, and SHAPEIT. In contrast, switch rates in the autosomes ranged from 0.0015 to 0.0038 switches per heterozygous site in Beagle, 0.0020 to 0.0045 in Eagle, and from 0.0015 to 0.0037 in SHAPEIT, with the corresponding average switch rates in X being 0.0017, 0.0023, and 0.0016 in Beagle, Eagle, and SHAPEIT.

Despite differences in error rates observed across chromosomes, the rankings of methods was found to be consistent both across all autosomes and between proband and synthetic diploid results. In particular, in all chromosomes Eagle has a higher average switch error rate, while SHAPEIT and Beagle have similar average switch error rates. In contrast, SHAPEIT has a higher average flip rate in all chromosomes and Eagle has the lowest flip error rate. Switch errors showed a similar proportional enrichment at CpG sites across all chromosomes, and the Eagle method consistently yielded the highest proportion of switch errors at CpGs. The pattern for flip errors differed sharply: although flip errors were enriched at CpG sites in all chromosomes, the enrichment observed in the autosomes (mean proportional enrichment: 0.031) was

markedly smaller than the substantial enrichment observed on Chromosome X (mean proportional enrichment: 0.185).

3 Discussion

Statistical phasing allows for the inference of haploid phase using data from genotypes enabeling downstream analyses such as imputation and the investigation of compound heterozygosity. However, it is challenging to evaluate the performance of statistical phasing for a specific dataset. While sequencing trios is typically used to generate gold standard haplotypes, such trios are not always available and this practice has been shown to create biased estimates[cite]. However, large sequencing studies are rich in haplotypes of known phase from male X chromosomes. Here we showed how such data cen be used to evaluate the performance of phasing algorithms by leveraged male X chromosomes from the 1kGP deep-sequence release to construct synthetic diploids of known phase. Across these synthetic diploids, we compared the performance of three modern statistical phasing methods, ShapeIt, Eagle and Beagle.

Comparing the methods we find that the majority of phasing errors is shared between at least two of the three methods. However, each method still has a unique profile of errors if we distinguish switch errors (switch of haplotypes) and flip errors (one allele assigned to the wrong haplotype). SHAPEIT intoduces fewer switch errors than the other two methods while Eagle introduces the most switch errors. On the other hand, SHAPEIT introduces more flip errors than Eagle and Beagle with Beagle introducing the fewest. As flip errors are overall less common on the X chromosome than flip errors, this pattern results in SHAPEIT having the overall smallest number of errors while Eagle has the largest number of errors. We observe the same error profiles for all populations. However, the rate of phasing errors differs, with XXX having the highest error rate and YYY having the lowest.

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 [21], while the enrichment at CpG sites is likely 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.

Many of the findings on the X can readily be extended to the autosomes. The error profiles of the three methods with eagle having the largest number of switch errors and Shapeit having the largest number of flip errors are the same across all autosomes. While the rates of switch errors differ between the autosomes by up to 2-fold, the rates of all three methods on the X are very similar to the rate on chromosomes 1-8 and 9-13 (70% of the genome). The most notable difference is the number of flip errors, where we infer ¿3x as many errors on autosomes compared to the X-chromosome. This may be driven part by the bias on error calling [cite browning] as genotype errors both in parrents and in offspring can lead to a percieved flip error. Moreover *Something about allele frequency.* Overall, this increase in flip errors on the autosomes changes the overall ranking by error rate.

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.

4 Methods

We used male X chromosomes to construct synthetic diploids of known phase in order to evaluate three model statistical phasing algorithms [5–7]. 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 characterized 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.

4.1 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 [13]. 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 [22]. 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 applied the same criteria, filtering the phased VCF for all autosomes to retain only biallelic, non-singleon SNPs that are not in chromosomal regions covered by the 1kGP pilot accessibility mask. The unphased VCF is filtered to the same set of variants.

4.2 Synthetic Diploid Construction and Phasing

to create synthetic diploid individuals from a give super-population, we first sampled a population from this super-population. From this population, we then sample two individual male X chromosomes and combine them to form a synthetic diploid. We ensured that no pair of samples were chosen more than once. Based on the sampled haplotypes, we created 2 files: 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 of 2,502 samples not used in construction of 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 using default options and genetic maps provided by the authors of the software [5–7].

4.3 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. To avoid potential errors, we excluded triple heterozygous positions at which patterns of Mendelian inheritance cannot be used to resolve haplotypes using only a trio. We also exclude these positions from the original, trio-resolved true haplotypes against which we benchmarked the phasing algorithms.

4.4 Evaluation of Population-based Inferred Phase

For our synthetic X diploids and the re-phased trios, we identified 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 identified 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.

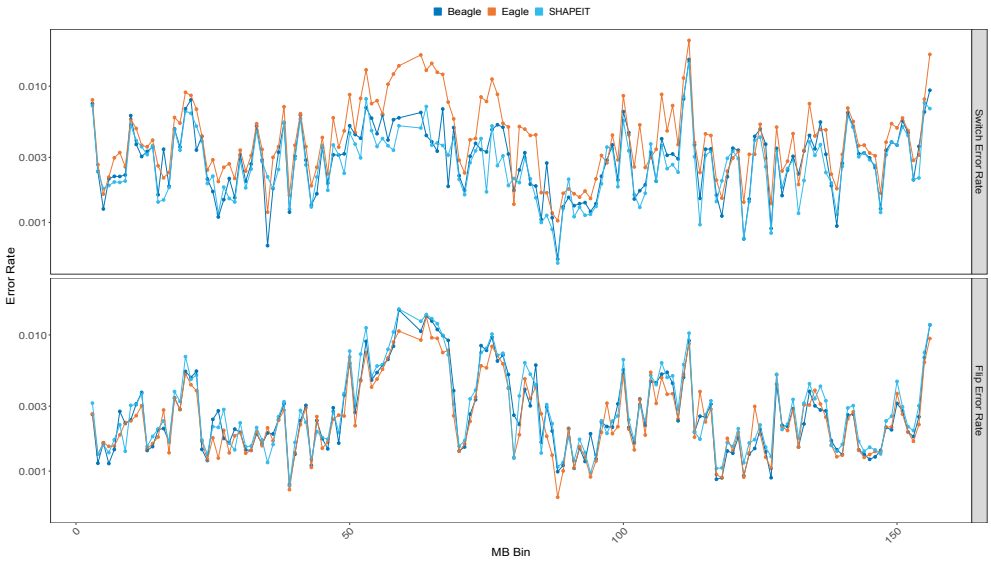

Fig. 3 The rate of switches and flips observed across the X chromosome in 1Mb sized bins. Rates observed across the chromosome are highly correlated between each pair of methods ($\rho \approx 1$)

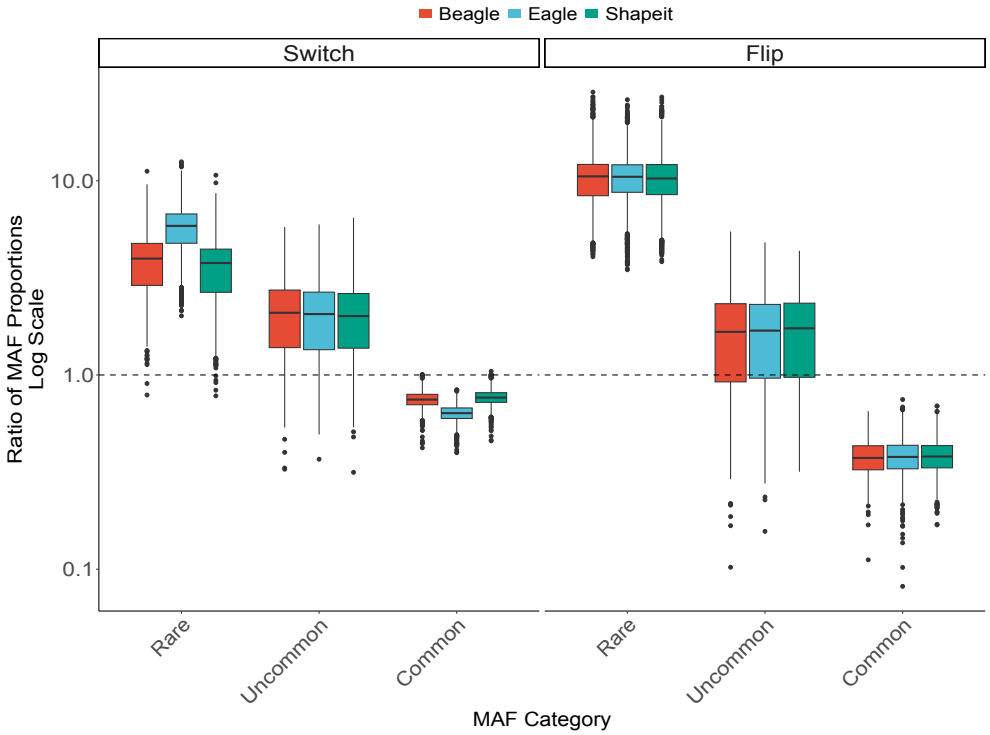

Fig. 4 Distributions of proportions of heterozyogous positions and error locations across minor allele frequency (maf) bins. In the left panel we compare the switch maf distributions to the background heterozygous position maf distributions, while on the right we do the same for flip error maf distributions. In both types of error we observe higher proportions of errors across all three methods in the rare category compared to the background distribution.

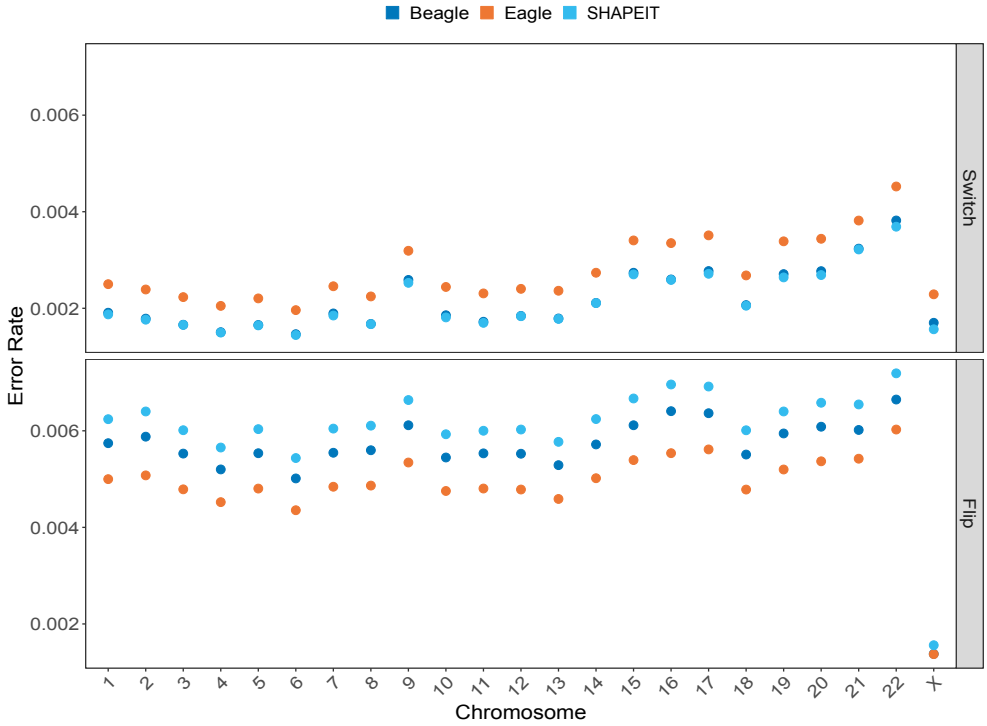

Fig. 5 Mean switch and flip error rates observed in the re-phasing of probands (autosomes) and synthetic diploids (X).

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nature Reviews Genetics* **12**, 215–223 (2011). URL http://dx.doi.org/10.1038/nrg2950.

[10] Guo, M. H. *et al.* Inferring compound heterozygosity from large-scale exome sequencing data. *Nature Genetics* **56**, 152–161 (2023). URL http://dx.doi.org/10.1038/s41588-023-01608-3.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2011). URL http://dx.doi.org/10.1038/nmeth.1785.

[16] 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.

[17] 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.

[18] Tomatsu, S. *et al.* Methylation patterns of the human β-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.

[19] 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.

[20] Si, Y., Kang, H. M. & Zöllner, S. Inferring cpg methylation signatures accumulated along human history from genetic variation catalogs (2023). URL http://dx.doi.org/10.1101/2023.03.24.534151.

[21] 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.

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

Appendix A Supplementary Figures

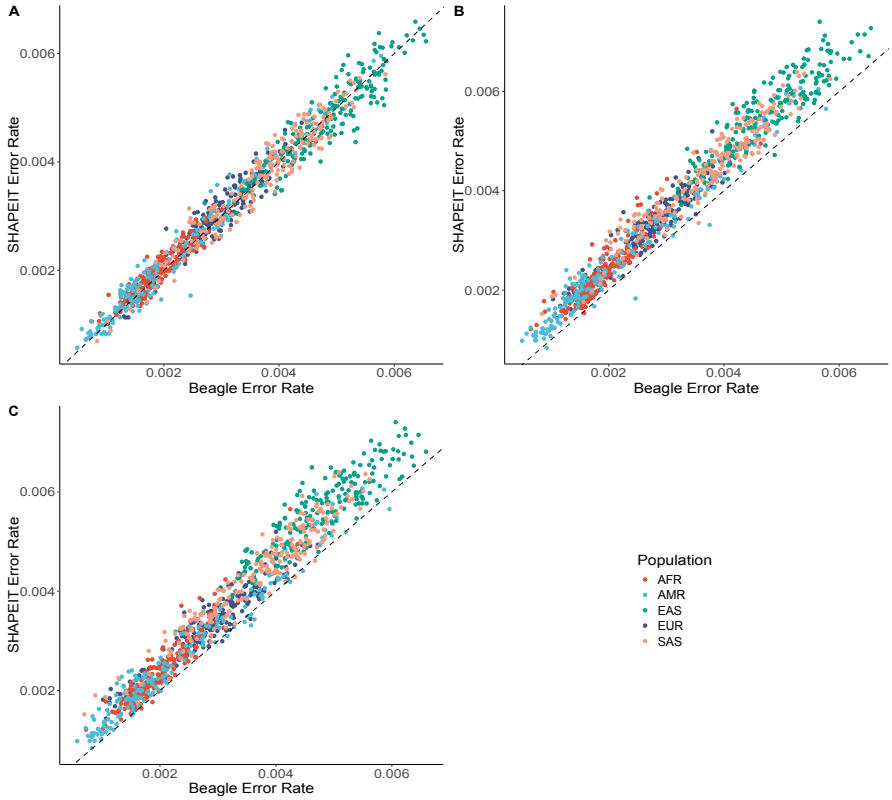

Fig. A1 Error rates observed in each synthetic X chromosome diploid across each pair of methods. Rates are computed as the total number of errors divided by the number of heterozygous positions in each synthetic diploid. Each point represents one synthetic diploid with the x- and y-position determined 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.

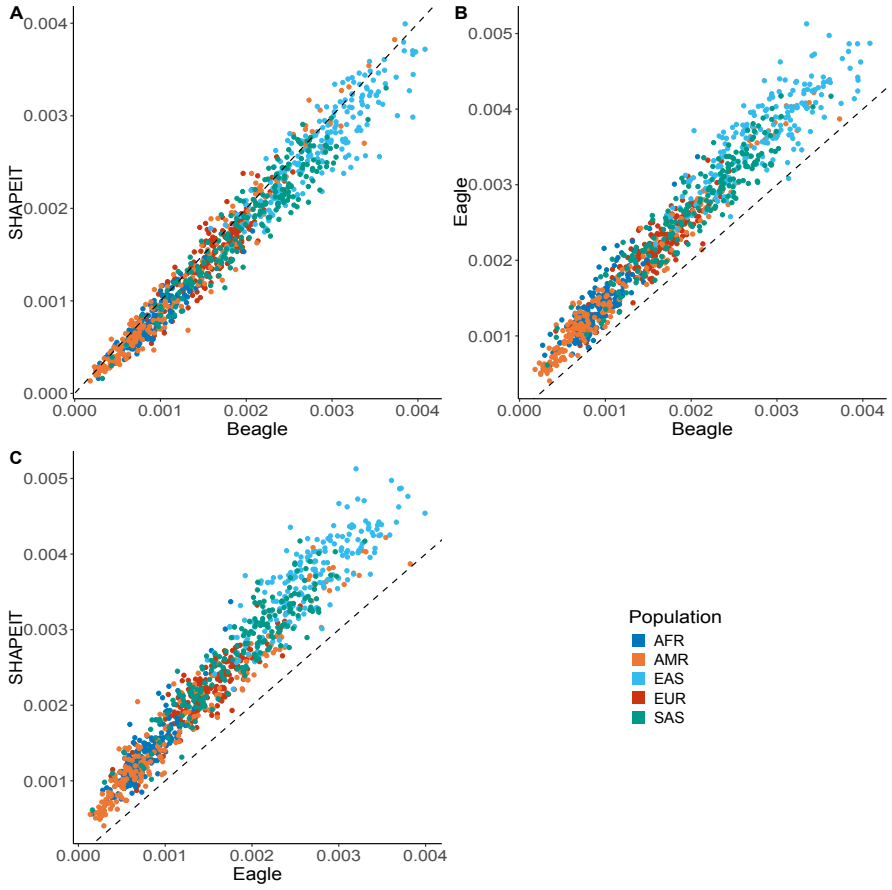

Fig. A2 Switch error rates observed in each synthetic X chromosome diploid across each pair of methods.

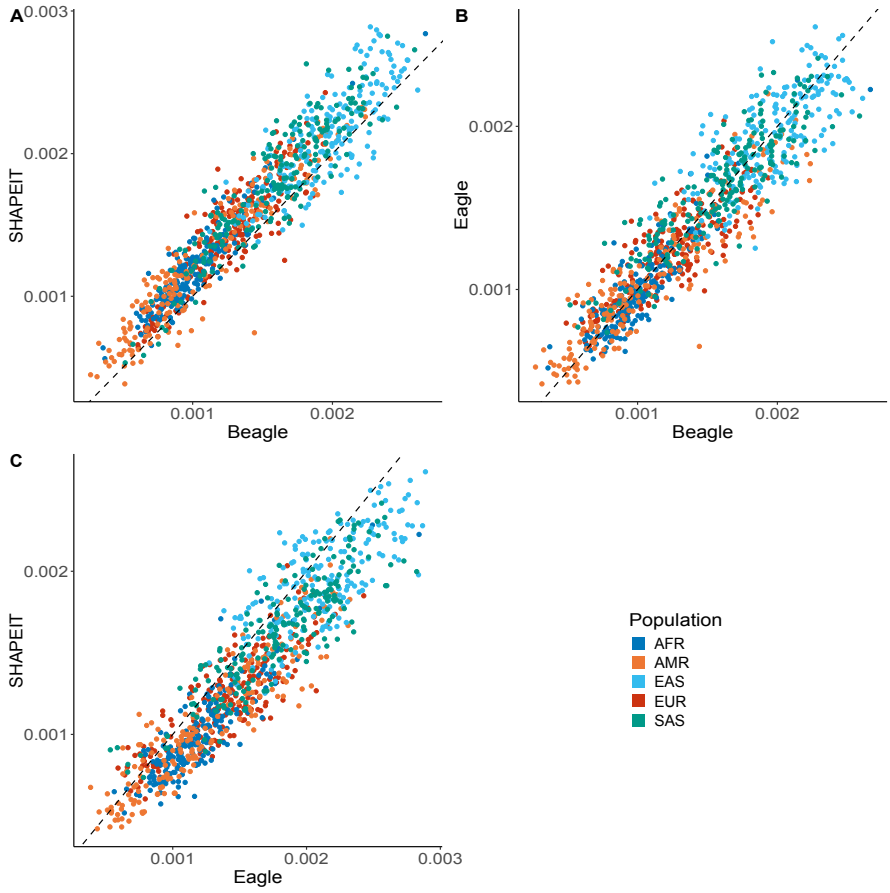

Fig. A3 Flip error rates observed in each synthetic X chromosome diploid across each pair of methods.

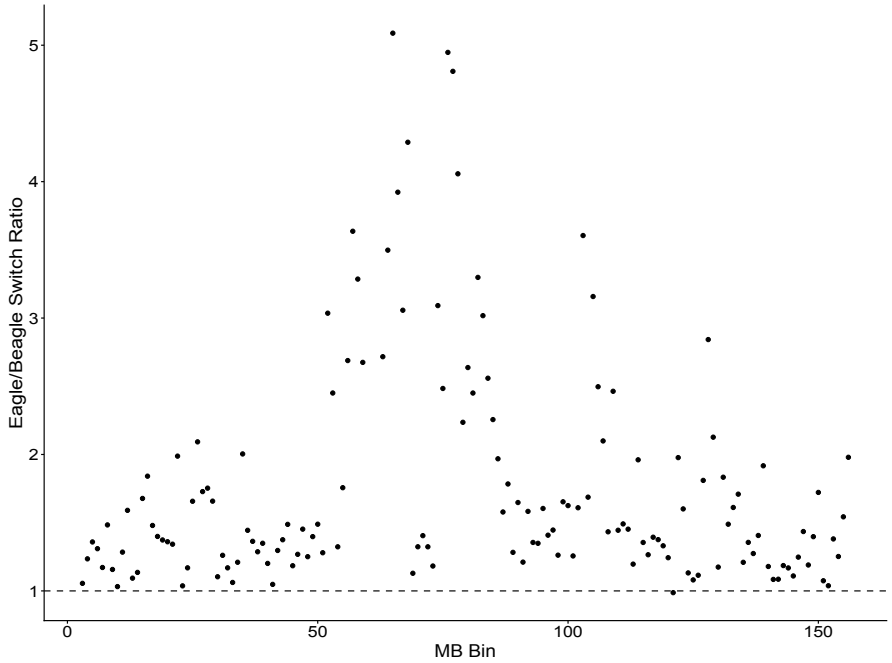

Fig. A4 Ratio of average Eagle switch rates to Beagle switch rates in megabase bins along the X chromosome. On average each bin's average switch rate for Eagle was 1.75 times that of Beagle

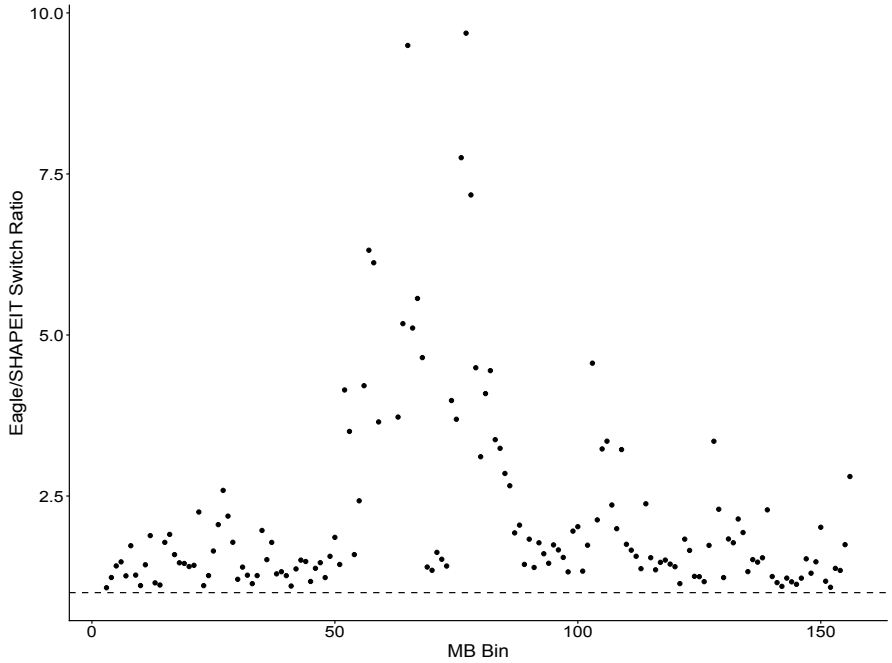

Fig. A5 Ratio of average Eagle switch rates to SHAPEIT switch rates in megabase bins along the X chromosome. On average each bin's average switch rate for Eagle was 2.18 times that of SHAPEIT

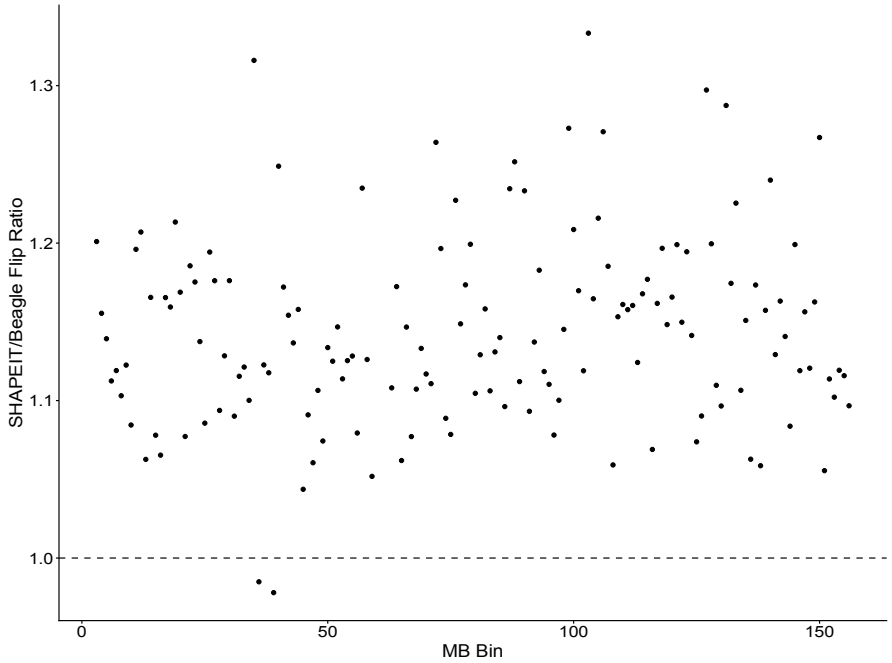

Fig. A6 Ratio of average SHAPEIT flip rates to Beagle flip rates in megabase bins along the X chromosome. On average each bin's average flip rate for SHAPEIT was 1.14 times that of Beagle

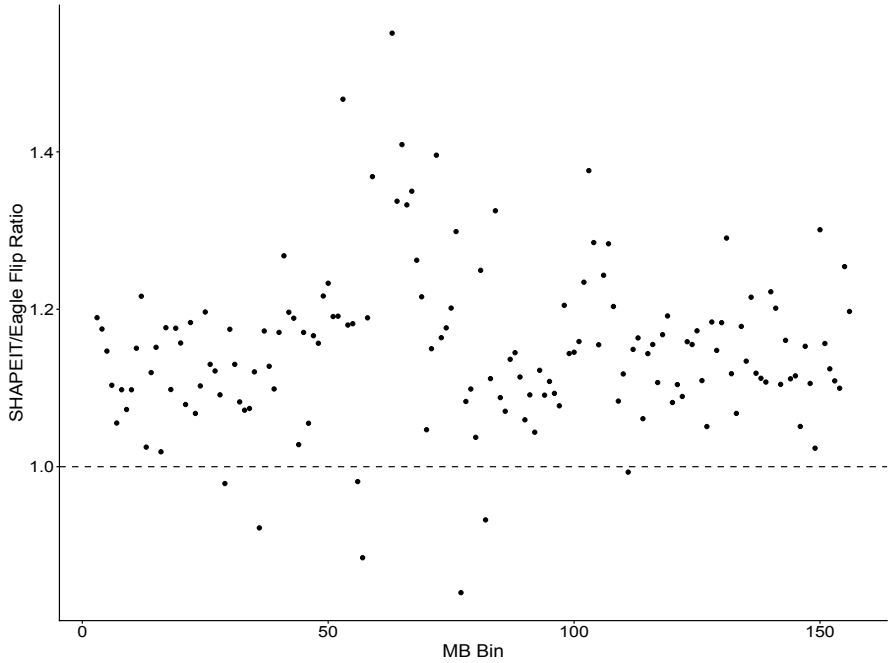

Fig. A7 Ratio of average SHAPEIT flip rates to Eagle flip rates in megabase bins along the X chromosome. On average each bin's average switch rate for SHAPEIT was 1.15 times that of Eagle

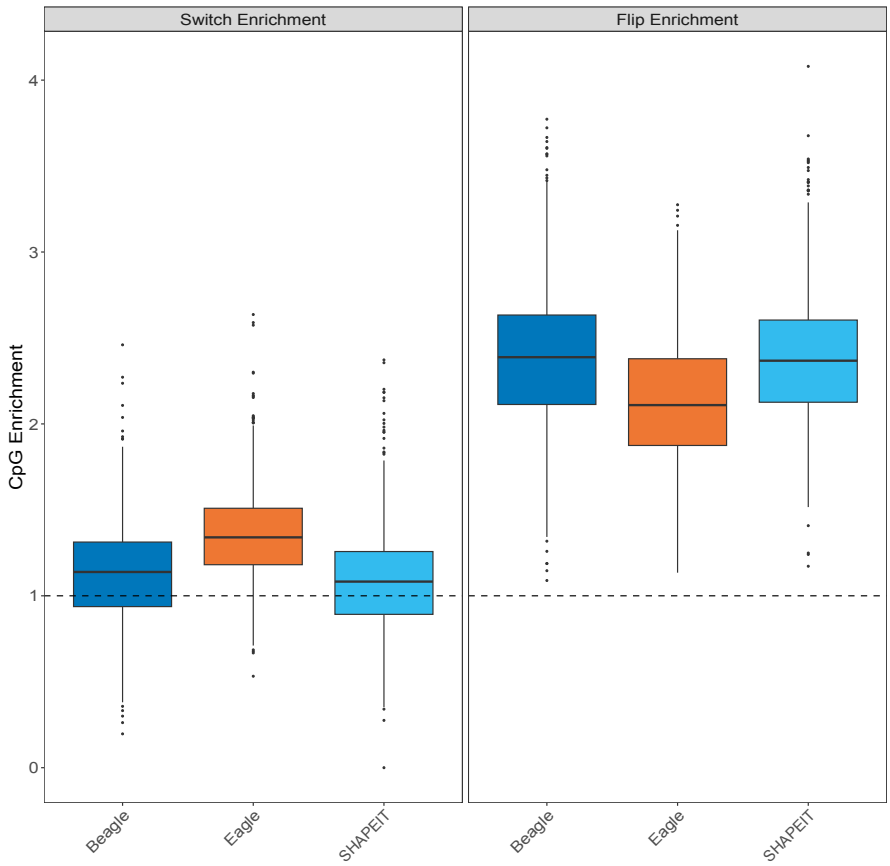

Fig. A8 Average enrichment of switch and flip errors, comparing the proportion of errors at CpG sites to the proportion of heterozygous positions at CpG sites.

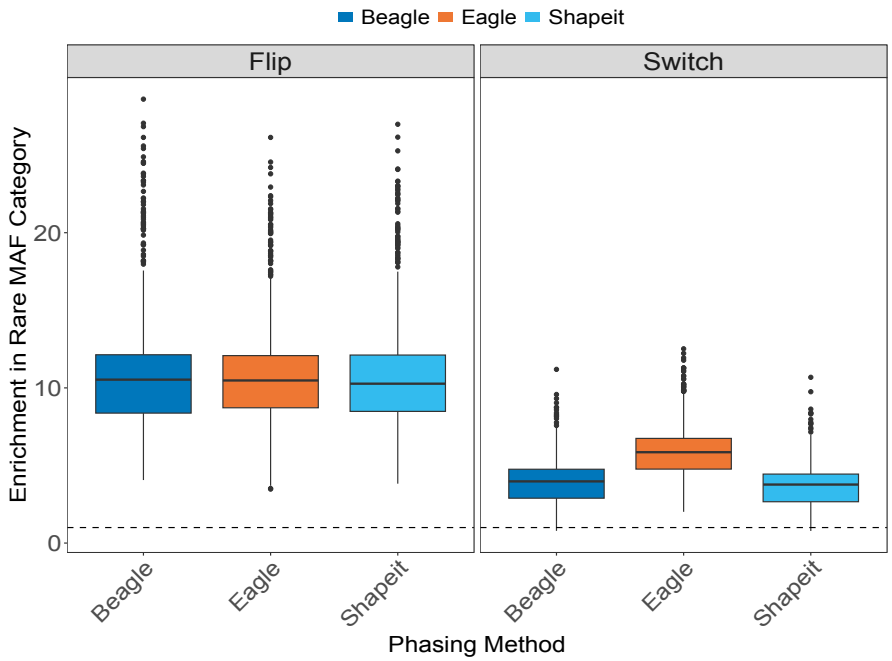

Fig. A9 Enrichment of errors at rare variants relative to the proportion of heterozygous sites at rare variants.

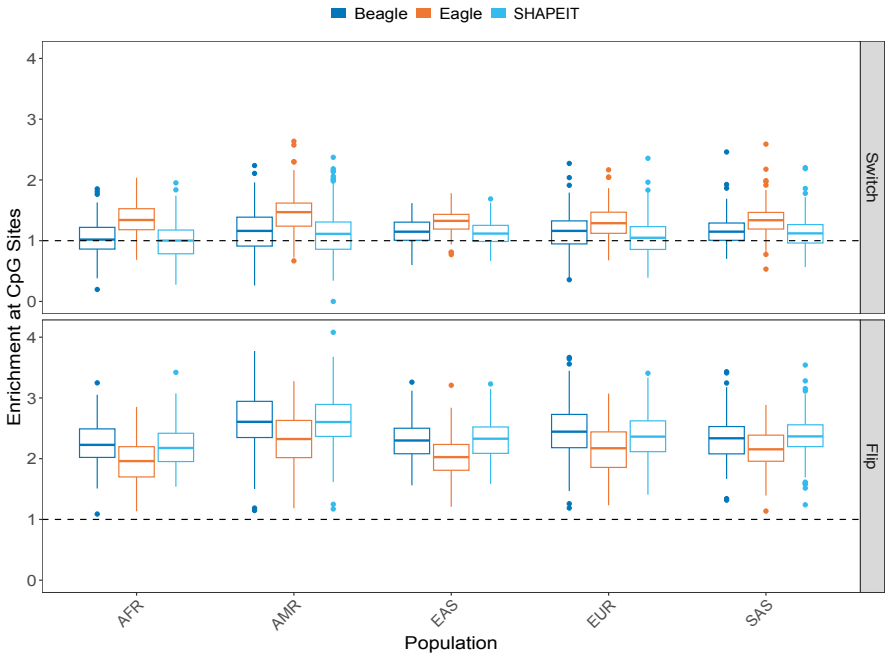

Fig. A10 Enrichment of switch and flip errors at CpG sites in synthetic X diploids by population. Enrichment is computed as the ratio of the proportion of errors occurring at a CpG site to the proportion of heterozygous sites occurring at CpG sites.

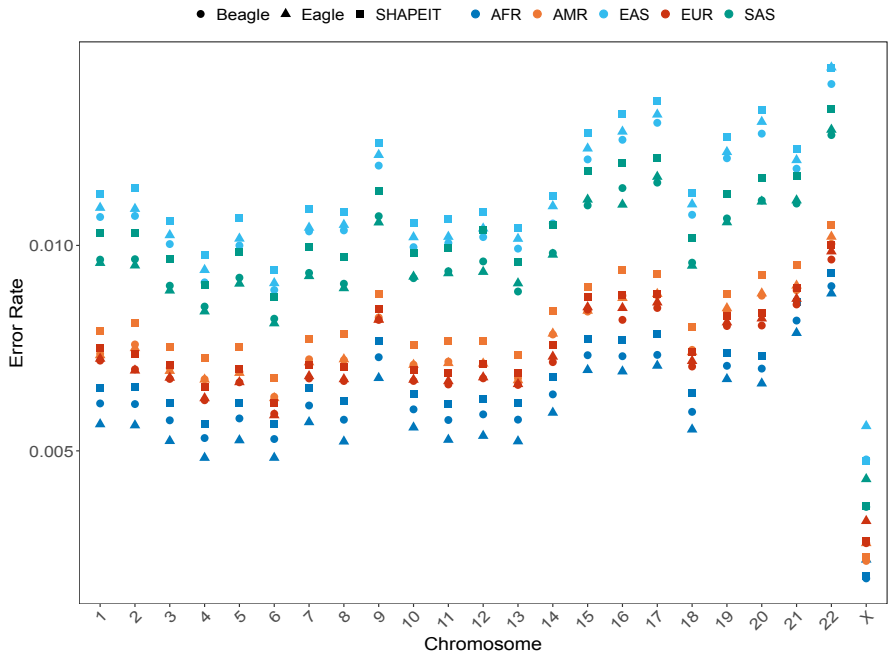

Fig. A11 Average error rate across all chromosomes and X by population and method.

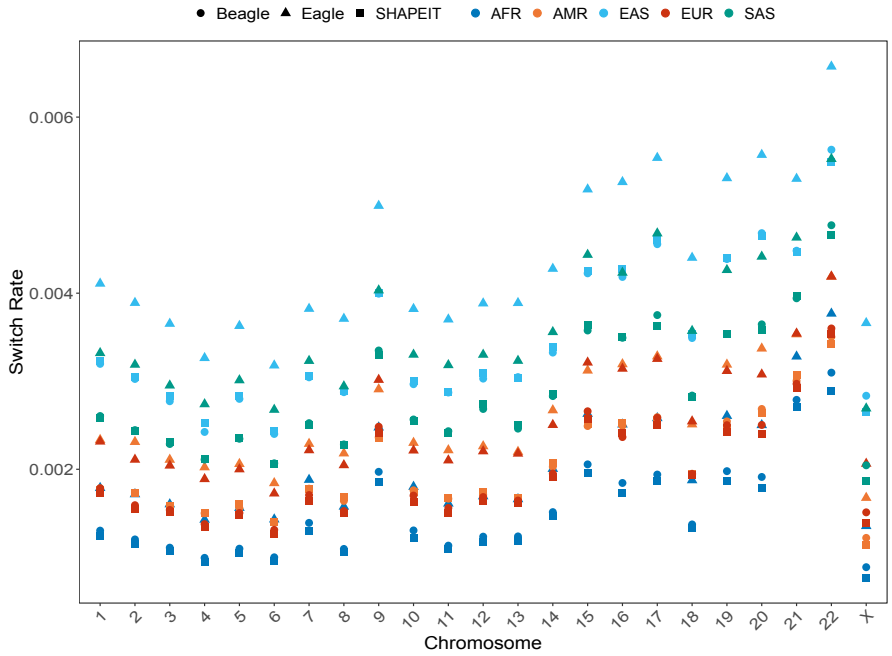

Fig. A12 Average switch rate across all chromosomes and X by population and method.

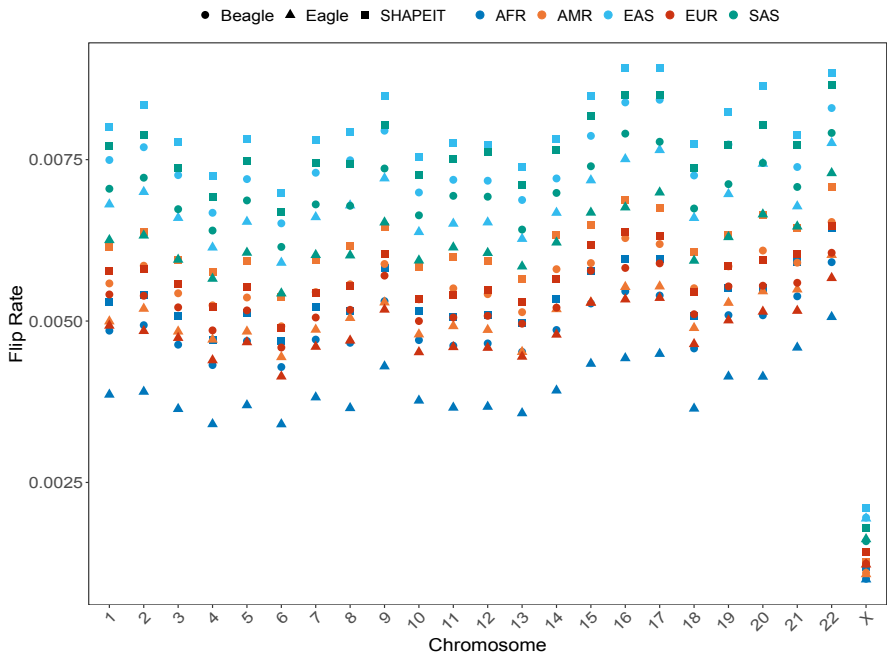

Fig. A13 Average flip rate across all chromosomes and X (23) by population and method.

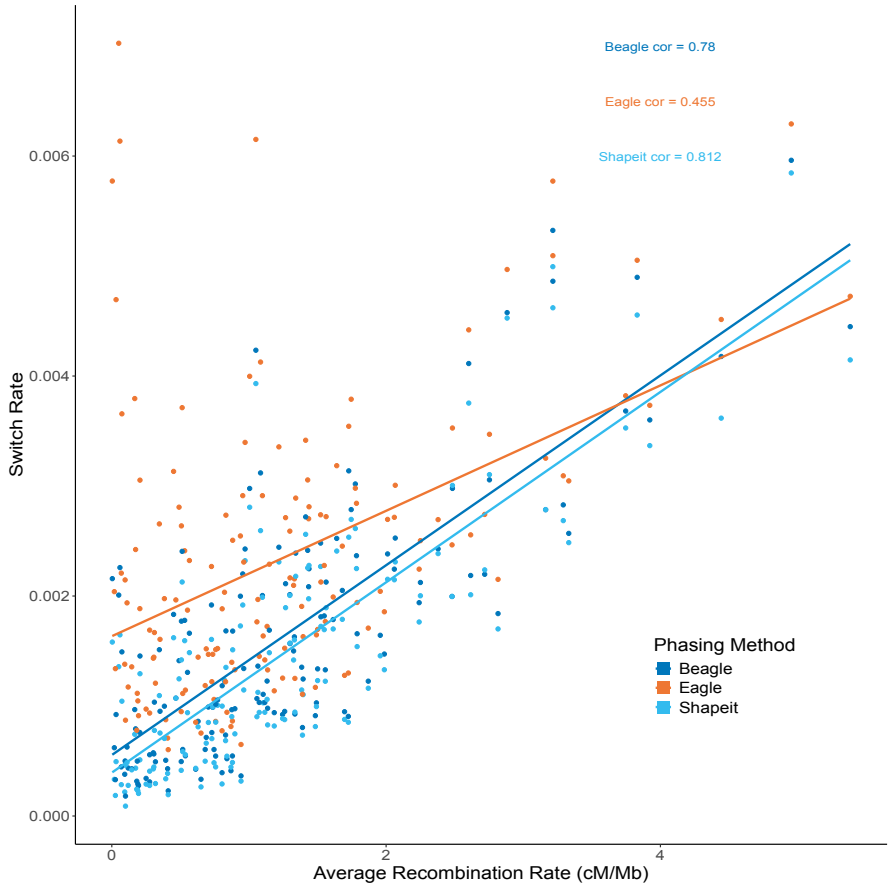

Fig. A14 Average flip error rates and recombination rate values in non-overlapping MB bins in chromosome X. Switch error rates are observed to be correlated with average recombination rates in non-overlapping MB bins in chromosome X (Beagle: $r = 0.78, p < 0.001$; Eagle: $r = 0.45(p < 0.001)$; SHAPEIT: $r = 0.81(p < 0.001)$).

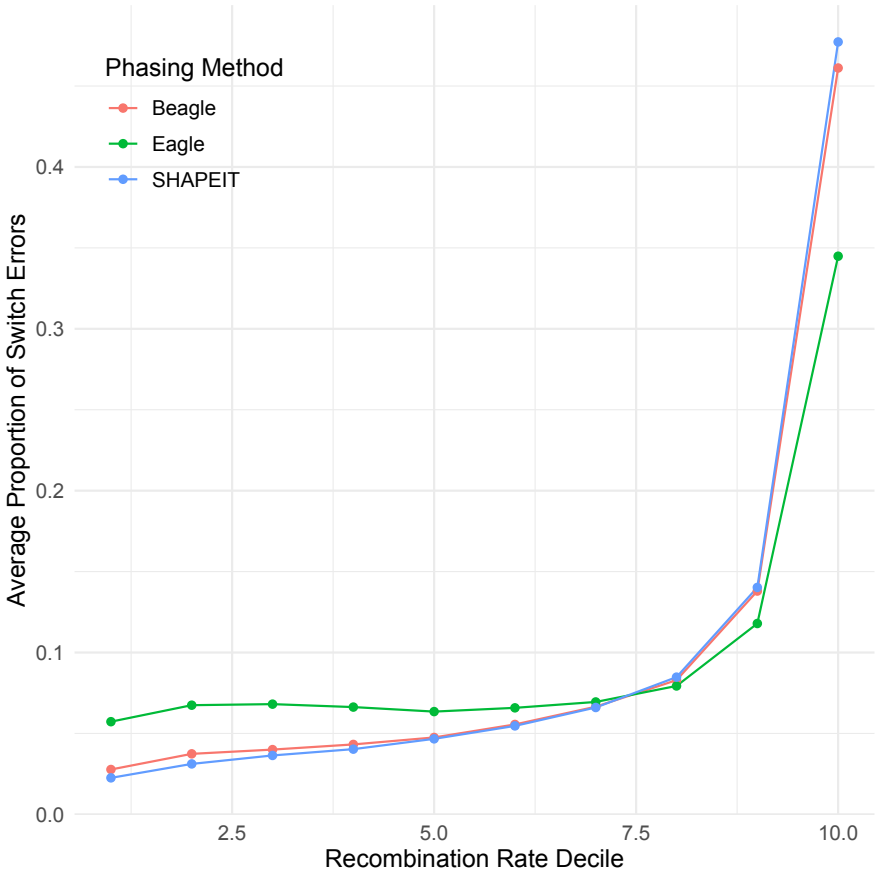

Fig. A15 Average proportion of switch errors by recombination rate decile. For all three methods we observe an enrichment in the top decile.

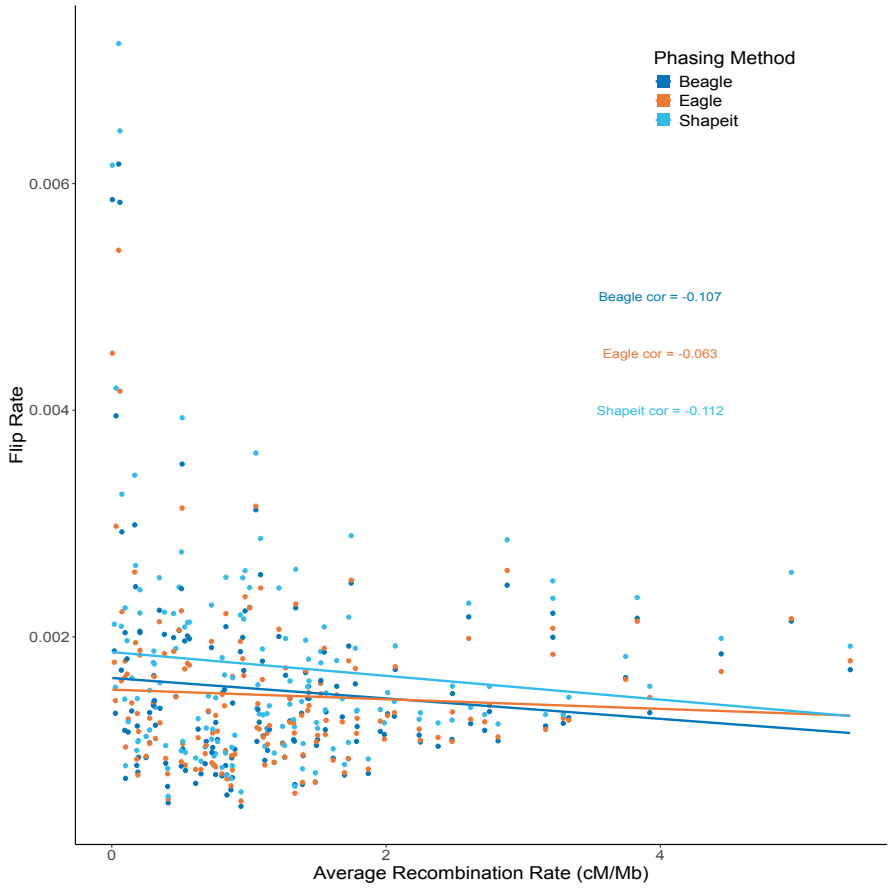

Fig. A16 Average flip error rates and recombination rate values in non-overlapping MB bins in chromosome X. Flip error rates were not found to be correlated with average recombination rates in non-overlapping MB-sized bins.

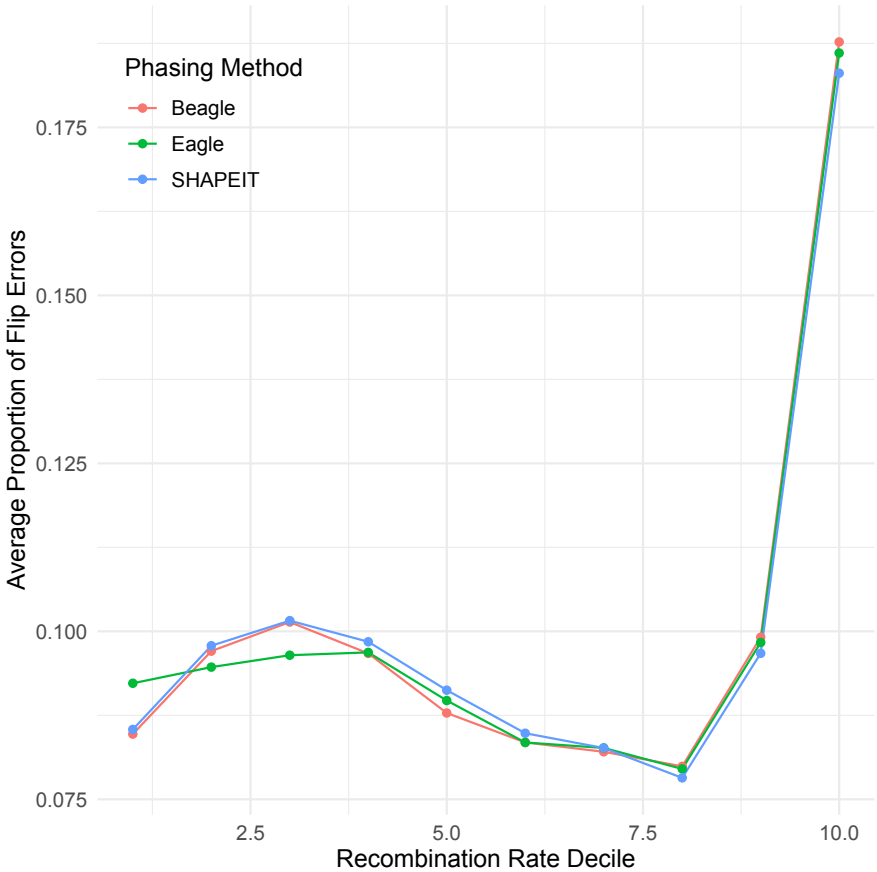

Fig. A17 Flip errors were found to be enriched for at sites with recombination rate values in the highest decile.

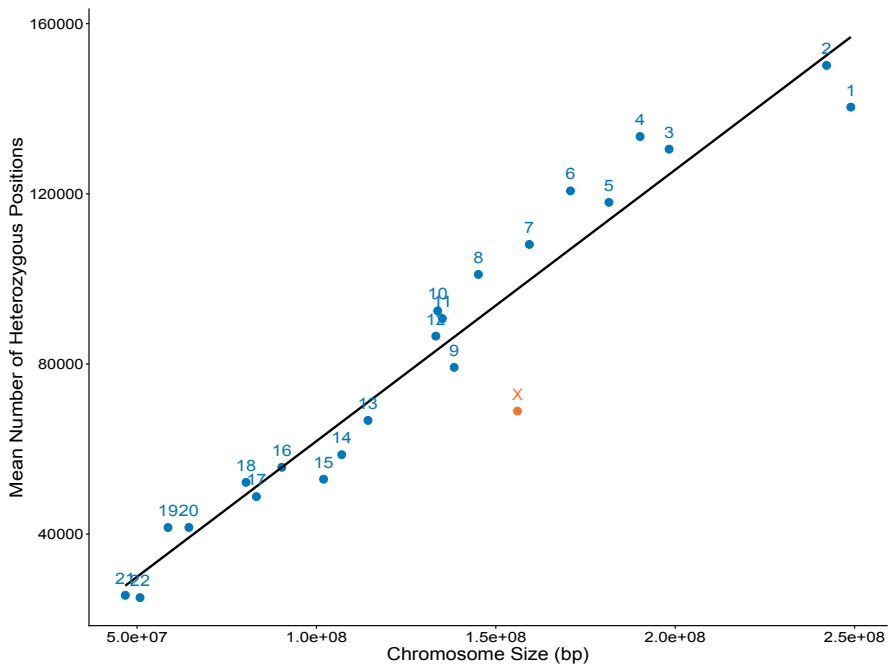

Fig. A18 The mean number of heterozygous sites in probands and synthetic diploids by size of the autosome.

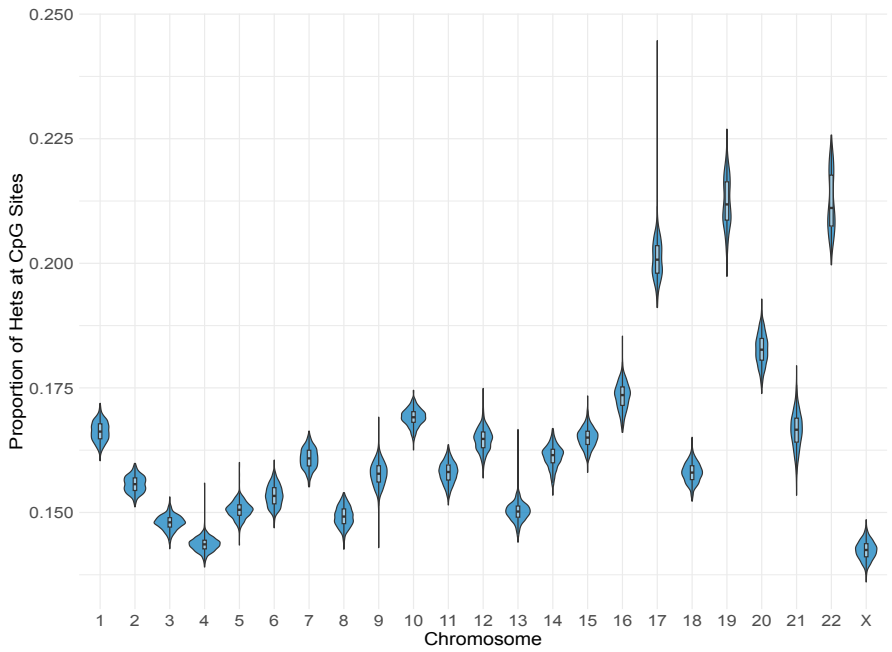

Fig. A19 Distribution of proportion of heterozygous sites at CpG by chromosome.

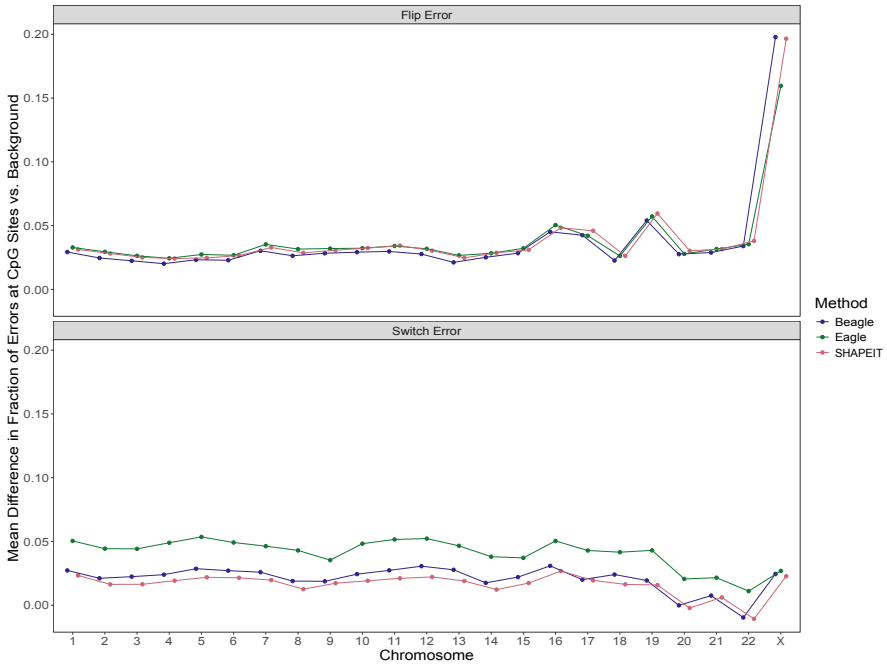

Fig. A20 Enrichment of flips and switch errors at CpG sites across the autosomes (trio re-phase) and X (synthetic diploids). Enrichment is computed as the mean difference between the proportion of errors at CpG sites and the proportion of heterozygous positions at CpG sites.

Appendix B Supplementary Tables

Error Type	Method	Mean Difference	T	P value	CI Lower	CI Upper
Flip Error	Beagle	0.1977867	110.07309	0	0.1948284	Inf
Flip Error	Eagle	0.1595476	93.97320	0	0.1567524	Inf
Flip Error	SHAPEIT	0.1965362	115.14546	0	0.1937261	Inf
Switch Error	Beagle	0.0244927	32.94071	0	0.0232685	Inf
Switch Error	Eagle	0.0268967	36.95026	0	0.0256983	Inf
Switch Error	SHAPEIT	0.0227155	31.24358	0	0.0215185	Inf

Table B1 T test results comparing the proportion of errors at CpG sites to the background proportion of heterozygous sites at CpG sites.

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 B2 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 B3 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.

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

Chromosome	Average Heterozygous	Size	Mean Density
1	140381.27	248,956,422	0.0005638788864
2	150199.66	242,193,529	0.000620163803
3	130503.74	198,295,559	0.0006581273966
4	133478.04	190,214,555	0.0007017235879
5	117995.98	181,538,259	0.0006499785811
6	120689.8	170,805,979	0.0007065900193
7	108102.46	159,345,973	0.0006784135047
8	101055.01	145,138,636	0.0006962653969
9	79201.45	138,394,717	0.0005722866574
10	92489.26	133,797,422	0.0006912633937
11	90662.32	135,086,622	0.0006711421061
12	86556.05	133,275,309	0.0006494530056
13	66732.95	114,364,328	0.0005835119321
14	58679.32	107,043,718	0.0005481808844
15	52914.63	101,991,189	0.0005188156989
16	55717.3	90,338,345	0.0006167624612
17	48818.39	83,257,441	0.0005863546779
18	52180.24	80,373,285	0.0006492236817
19	41554.18	58,617,616	0.0007089025934
20	41571.99	64,444,167	0.0006450853807
21	25626.59	46,709,983	0.0005486319702
22	25074.31	50,818,468	0.0004934094038
X	68930	152,939,905	0.0004506999007

Table B4 The mean number of heterozygous sites per proband (autosomes) and synthetic diploids (X) and chromosome sizes from GRCh38.

	Mean Error Rate		
Chromosome	Beagle	Eagle	SHAPEIT
1	0.0077	0.0075	0.0081
2	0.0077	0.0075	0.0082
3	0.0072	0.0070	0.0077
4	0.0067	0.0066	0.0072
5	0.0072	0.0070	0.0077
6	0.0065	0.0063	0.0069
7	0.0074	0.0073	0.0079
8	0.0073	0.0071	0.0078
9	0.0087	0.0085	0.0092
10	0.0073	0.0072	0.0077
11	0.0073	0.0071	0.0077
12	0.0074	0.0072	0.0079
13	0.0071	0.0070	0.0076
14	0.0078	0.0078	0.0084
15	0.0089	0.0088	0.0094
16	0.0090	0.0089	0.0095
17	0.0091	0.0091	0.0096
18	0.0076	0.0075	0.0081
19	0.0087	0.0086	0.0090
20	0.0089	0.0088	0.0093
21	0.0093	0.0092	0.0098
22	0.0105	0.0105	0.0109
X	0.0031	0.0037	0.0031

Table B5 Mean error rates observed in probands (autosomes) and synthetic diploids (X).

	Mean Switch Rate			Mean Flip Rate		
Chromosome	Beagle	Eagle	SHAPEIT	Beagle	Eagle	SHAPEIT
1	0.0019	0.0025	0.0019	0.0057	0.0050	0.0062
2	0.0018	0.0024	0.0018	0.0059	0.0051	0.0064
3	0.0017	0.0022	0.0017	0.0055	0.0048	0.0060
4	0.0015	0.0021	0.0015	0.0052	0.0045	0.0057
5	0.0017	0.0022	0.0016	0.0055	0.0048	0.0060
6	0.0015	0.0020	0.0014	0.0050	0.0044	0.0054
7	0.0019	0.0025	0.0019	0.0055	0.0048	0.0060
8	0.0017	0.0022	0.0017	0.0056	0.0049	0.0061
9	0.0026	0.0032	0.0025	0.0061	0.0053	0.0066
10	0.0019	0.0024	0.0018	0.0054	0.0048	0.0059
11	0.0017	0.0023	0.0017	0.0055	0.0048	0.0060
12	0.0018	0.0024	0.0018	0.0055	0.0048	0.0060
13	0.0018	0.0024	0.0018	0.0053	0.0046	0.0058
14	0.0021	0.0027	0.0021	0.0057	0.0050	0.0062
15	0.0027	0.0034	0.0027	0.0061	0.0054	0.0067
16	0.0026	0.0034	0.0026	0.0064	0.0055	0.0070
17	0.0028	0.0035	0.0027	0.0064	0.0056	0.0069
18	0.0021	0.0027	0.0021	0.0055	0.0048	0.0060
19	0.0027	0.0034	0.0026	0.0059	0.0052	0.0064
20	0.0028	0.0034	0.0027	0.0061	0.0054	0.0066
21	0.0032	0.0038	0.0032	0.0060	0.0054	0.0065
22	0.0038	0.0045	0.0037	0.0066	0.0060	0.0072
X	0.0017	0.0023	0.0016	0.0014	0.0014	0.0016

Table B6 Mean switch and flip error rates observed in probands (autosomes) and synthetic diploids (X).