Improving haplotype phasing accuracy using many short IBD segments

Aaron Kleinman[1], Eric Y. Durand[1] and Cory Y. McLean[1]

1. 23andMe, Inc., Mountain View, CA, USA

Introduction

Phasing is the deconvolution of diploid genetic data into maternal and paternal haplotypes. Such haplotype information is required for many genetic applications, such as imputation and fine mapping. Typical approaches to phasing [1,2] leverage the idea that individuals share haplotypes, and attempt to find a minimal set of haplotypes that explain the observed data. In cases when an individual's parents' genotypes are known, one can use a different and relatively straightforward approach to phase the child almost perfectly.

Identical-By-Descent (IBD) segments are regions of identical haplotypes between two individuals that were inherited from a recent shared common ancestor. The fundamental idea of long-range phasing is that if we know someone is in IBD with an individual over a short region, they act as a "surrogate parent" over that region, and their genotype can thus be used to phase the individual. This idea was pioneered in [3], but that approach requires perfect IBD detection and has difficulty when the data are noisier.

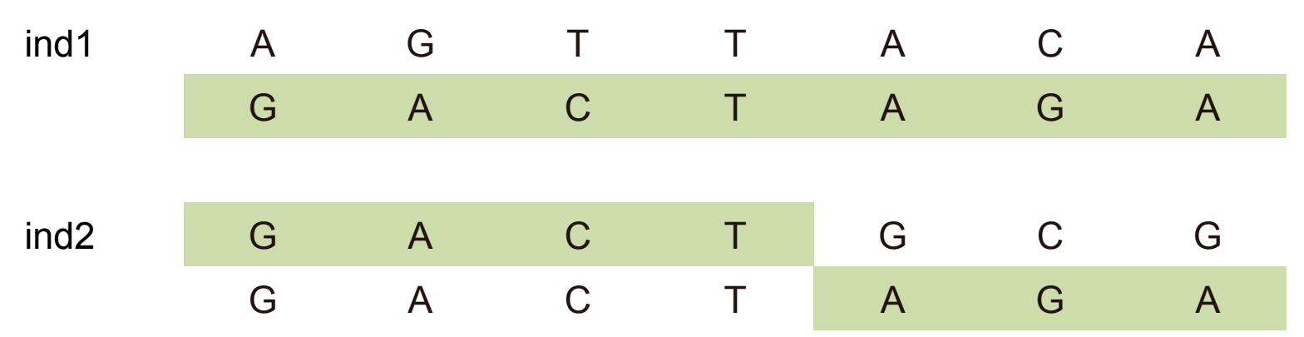
Here we present Origin, a novel method for leveraging IBD data to improve phasing accuracy. Origin calculates the most likely haplotype matches for every IBD segment spanning a genomic region and integrates switch error predictions from all IBD segments using a weighted voting method. We use trios to assess accuracy in European and non-European cohorts and find that it improves phasing accuracy in all examined populations.

Experiment Methodology

- Research participants were drawn from more than 350,000 genotyped customers participating in the 23andMe® Personal Genome Service. Participants provided informed consent under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services.
- Participants were genotyped on the Illumina HumanOmniExpress+ genotyping array.
- We phased participants using a modified version of BEAGLE [1] and computed pairwise IBD using GERMLINE [4]. IBD segments were filtered for quality using HaploScore [5].
- Local ancestry was computed using [6], and individuals were classified as European if 97% or more of their genome was classified European. A similar approach was used to determine East Asian, Latino and African American ancestry.
- Parent-child relationships were detected by finding pairs of individuals who shared over 85% half IBD and less than 10% full IBD.
- Using this, we extracted cohorts of 5,000 pairwise-disjoint father-mother-child trios of European ancestry. We also extracted cohorts of 407 Latino trios, 192 African-American trios and 464 East Asian trios.
- For each child, we removed IBD segments between that child and their parents. We then used Origin to rephase the child's genotype on chromosome 22 using all remaining IBD segments incident on that child.
- We trio-phased the children and took this to be the ground truth. We used this to compute the number of switch errors before and after Origin rephasing.

Algorithm Methodology

Origin rephases an individual (proband) in two distinct steps. First, each IBD segment incident on the individual is used to suggest switch errors (Figure 1). This is accomplished by running an HMM with four hidden states, one for each pair of (proband, surrogate) haplotypes. Certain hidden state transitions in the Viterbi path are taken as evidence of proband switch errors. Other approaches, such as using a probabilistic distribution across the states at each site, performed worse (not shown).



ind1	A	G	T	T	A	C	A
	G	A	C	T	A	G	A
ind2	G	A	C	T	G	C	G
	G	A	C	T	A	G	A

Figure 1. **IBD segments inform phasing.** We can use IBD segments to improve phasing. In this example, two individuals can be made to share a haplotype with the introduction of a switch error in individual2. The actual situation is significantly complicated by the existence of genotyping errors and the presence of false IBD segments.

Algorithm Methodology (continued)

Then, the suggested switches across all IBD segments are aggregated together in a weighted vote (Figure 2). A lack of vote in favor of a switch is taken as evidence against switching, so every segment votes at every site it covers. Because IBD endpoint determination is difficult [1], a switch error reported near the end of an IBD segment is more likely to be a false positive than a switch error near the center. Consequently, Origin weights a segment's vote by √(L*R), where L and R are the lengths of the switch error. A proband switch error is then called if the sum of the switch weights is greater than half the total weight. Other weights, such as weighting uniformly and weighting by L*R, performed worse (not shown).

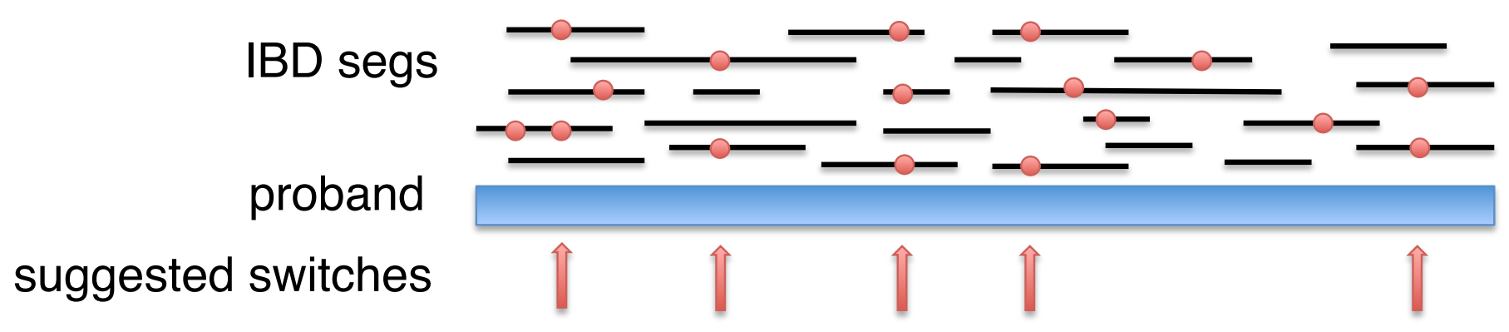


Figure 2. **Schematic illustrating the aggregation step.** This cartoon illustrates the aggregation of IBD segment votes. Each black line indicates an IBD segment, At each position, a segment either votes that the proband has a switch error there, or votes that the proband does not have a switch error there ("switch" or "stay"). In this image, red dots indicate votes for switch errors, and portions of the segments without red dots indicate votes to stay. A segment's vote is weighted by sqrt(L*R), where L is the length of the segment to the left of the vote, and R is the length of the segment to the right of the vote, and a simple weighted majority then determines whether or not to introduce a proband switch at that site. In this image, red arrows indicate places where Origin would call a switch error.

Results

Origin reduced switch errors by 27% for European probands, 16% for African American probands, 9% for Asian American probands and 19% for Latino probands (Figure 3). The proportion of errors corrected is significantly influenced by IBD coverage (Figure 4).

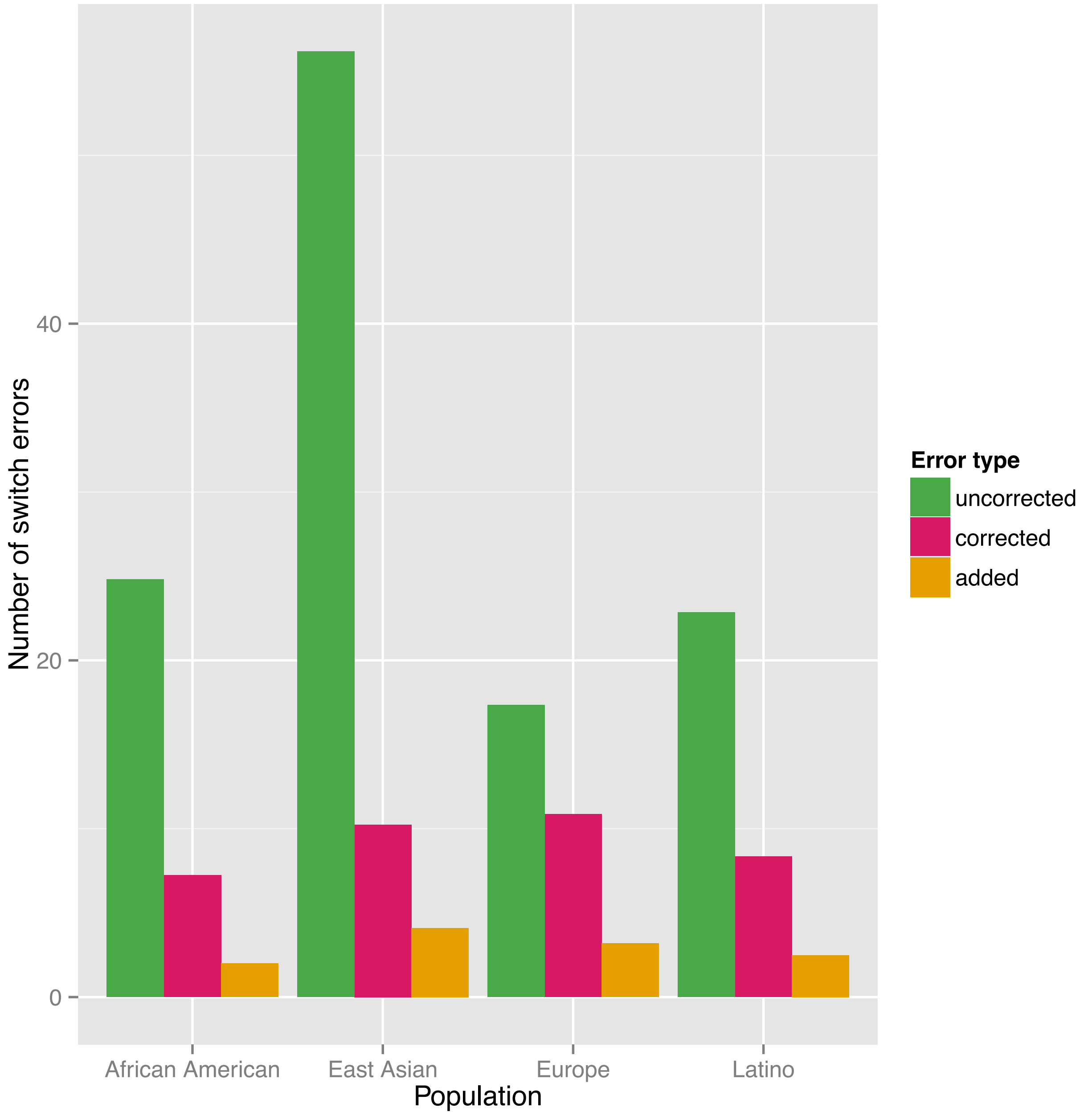


Figure 3. **Switch errors before and after rephasing.** When rephasing, there are three kinds of switch errors: those that are not corrected by rephasing, those that are corrected by rephasing, and those that are introduced by rephasing. This shows the average number of errors of each kind across chromosome 22 for probands of different ancestries. When the number corrected is greater than the number introduced, the method improves phasing accuracy. The larger number of uncorrected errors in non-European populations is partly a consequence of the fact that the non-European cohorts are smaller, so there are fewer IBD segments to help rephase.

Results (continued)

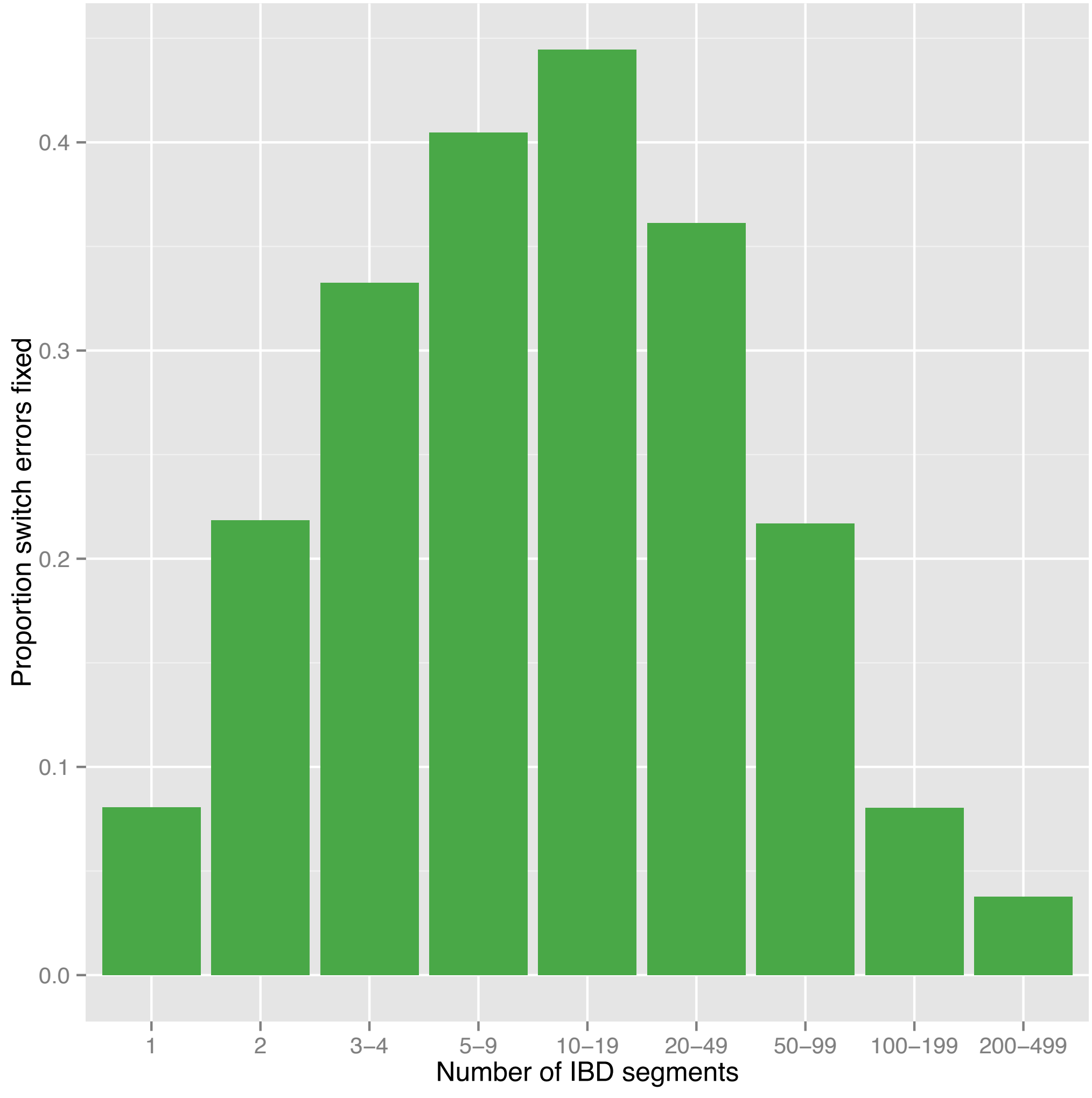


Figure 4. **Performance depends on IBD coverage.** Since Origin rephases using IBD segments, the depth of IBD coverage necessarily has a significant impact on the performance of the algorithm. This figure plots the number

1 – (#switch errors post-Origin) / (# switch errors pre-Origin)

as a function of binned IBD coverage, across the 5,000 European children on chromosome 22. Perfect rephasing would be represented by a bar of height 1. Each bin contains > 1400 switch errors.

Discussion

The algorithm's performance increases steadily with IBD coverage until about 20 segments, where it reduces switch errors by 45%, but then begins to degrade (Figure 4). The reasons for this are unclear; we speculate it is because regions with high IBD coverage are more likely to have a large number of those segments as false positives.

The individuals in this experiment were all genotyped on the same array, and missing data were imputed in the phasing step. We have adapted the algorithm to handle individuals genotyped on different arrays (which can be thought of as a special case of missing data), and the additional IBD coverage leads to further gains in accuracy. The algorithm is agnostic to the methods used for the initial phasing and IBD detection, and so it can be applied to other datasets. Many of the steps in the algorithm, including the switch-error and genotype-error rates used in the HMM, and the weights used to aggregate votes, can be easily modified to better suit the particular data at hand.

We have introduced Origin, a computationally-efficient method to improve phasing by leveraging IBD segments. Origin is online in the IBD input, making it easy to update an individual's phase as new IBD segments are calculated. The algorithm can be easily modified to run on genotypes phased by and IBD calculated by other methods, and so is widely applicable to other population-scale datasets.

Acknowledgments

We thank the customers and employees of 23andMe, who together made this research possible.

References

1. S R Browning and B L Browning. *Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering.* Am J Hum Genet 81:1084-1097.
2. O Delaneau et. al. *A linear complexity phasing method for thousands of genomes*. Nat Methods. 9(2):179-81
3. A Kong. et. all. *Detection of sharing by descent, long-range phasing and haplotype imputation.* Nature Genetics 40:1068-1075.
4. A Gusev et. al. *Whole population genomewide mapping of hidden relatedness.* Genome Research.
5. E Durand, N Eriksson and C McLean. *Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis.* Mol Biol Evol 2014 Aug; 31(8): 2212-2222.
6. E Durand et. al. *Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution.* doi: http://dx.doi.org/10.1101/010512.