Supplementary Materials for “Mean gene conversion tract length in humans estimated to be 459 bp from UK Biobank sequence data”

Nobuaki Masaki1\*, Sharon R. Browning1

1 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America

# Supplementary methods

## Estimating the proportion of singleton observed tracts

When we assume a geometric distribution for the gene conversion tract length, recall that,

where and is the probability of allele conversion at each position inside the th observed tract. We condition the observed tract lengths to be no more than 1500 bp, because we remove any observed tracts larger than 1500 bp prior to estimation of (see the section, Detecting gene conversion tracts).

We also described a method for obtaining , our estimate of , for all observed tracts . Using , we can estimate the probability that observed tract will be 1 bp, conditioned on and :

We can estimate the proportion of observed tracts that are 1 bp long (among the observed tracts no longer than 1500 bp) by taking the mean of across all observed tracts less than or equal to 1500 bp. We will refer to observed tracts that are 1 bp long as singleton observed tracts. Denoting our estimate of the proportion of singleton observed tracts as ,

where and represents the number of observed tracts no longer than 1500 bp.

## Assessing model fit on the simulated data

We only use observed tract lengths smaller than or equal to 1500 bp when estimating , the mean tract length. This is because observed tract lengths are likely to be truncated for longer gene conversion tracts during the detection process (see the section, Detecting gene conversion tracts).

We also remove all observed tract lengths that are 1 bp before estimating because our model overestimates the frequency of observed tract lengths that are 1 bp, which leads to biased estimates for the mean tract length . We refer to observed tract lengths that are 1 bp as singleton observed tracts. In this section, we take a closer look at this discrepancy between the proportion of singleton observed tracts detected in the coalescent simulation and the estimated proportion of singleton observed tracts according to our model. Our method for estimating this proportion is described in the previous section, Estimating the proportion of observed tracts that are 1 bp.

Recall that in the coalescent simulation, a geometric distribution with a mean tract length of 300 was used to simulate gene conversion tracts. Thus, we compare the proportion of singleton observed tracts detected in the coalescent simulation to , where we set , or equivalently, .

We are also interested in comparing the remaining distribution of observed tract lengths (between 2 and 1500 bp) detected in the coalescent simulation to what we expect according to the model. This can be done by comparing the empirical CDF of observed tract lengths between 2 and 1500 bp to the corresponding model CDF. Denoting the model CDF as ,

Our model may not fit well to the observed tract lengths detected in the coalescent simulation due to a multitude of reasons. One potential cause for this is that our model does not adjust for linkage disequilibrium (LD), even though this is present in the coalescent simulation. Recall that the distribution of observed tract lengths in our model only depends on the population level heterozygosity rate of markers. However, LD will change the distribution of heterozygous markers within individuals, which will likely affect the distribution of observed tract lengths in the coalescent simulation.

Another potential reason for a poor model fit is due to detection bias in the observed tract lengths. Observed tracts may be detected at different rates using the multi-individual IBD method,1 depending on their lengths (e.g. singleton observed tracts may be detected at a different rate compared to longer observed tract lengths). This will affect the empirical distribution of observed tract lengths in the coalescent simulation.

To explore these potential causes for the model fitting poorly to the observed tract lengths detected in the coalescent simulation, we run two other simulation studies, described in the sections, “Simulating observed tract lengths from chromosome 1 without accounting for linkage disequilibrium” and“Simulating observed tract lengths from individuals in the coalescent simulation.” We analyze the observed tract lengths generated from these two simulation studies to assess potential reasons for why the model may not fit well to the observed tract lengths detected from the coalescent simulation.

## Simulating observed tract lengths from chromosome 1 without accounting for linkage disequilibrium

In this simulation study, we simulate observed tract lengths, only using the population heterozygosity rate of markers on chromosome 1 from the UK Biobank whole autosome data. We simulate these observed tract lengths using the following steps:

1. We generate gene conversion tracts by uniformly sampling the starting position on chromosome 1 and drawing the length of the gene conversion tract from a geometric distribution with mean 300. The start and end positions of each tract are saved.
2. We let an allele conversion occur at each position within each gene conversion tract with probability , where is the minor allele frequency at position .
3. For each gene conversion tract, we obtain the observed tract length of the gene conversion tract by taking the length spanning the furthest allele converted positions.

We excluded variants with MAF less than 5% when detecting allele conversions in the UK Biobank whole autosome data. For this simulation, we also set if the MAF is less than 5% at position to prevent detecting allele conversions at these markers.

## Simulating observed tract lengths from individuals in the coalescent simulation

From the first replicate of the coalescent simulation, we have the genomes of all 125,000 individuals simulated using msprime.2 Instead of detecting observed tract lengths from past gene conversion events, we can simulate observed tract lengths directly from these individuals using the following steps:

1. We sample individuals with replacement from the 125,000 individuals.
2. We generate gene conversion tracts by uniformly sampling the starting position on the genome and drawing the length of the gene conversion tract from a geometric distribution with mean 300. The start and end positions of each tract are saved. Now we have individual, gene conversion tract pairs.
3. For each individual, gene conversion tract pair, we simulate the gene conversion tract on the individual. We store the length spanning the furthest heterozygous positions as the observed tract length.

We excluded variants with MAF less than 5% when detecting allele conversions in the UK Biobank whole autosome data. For this simulation, we did the same to prevent detecting allele conversions at these markers.

# Supplementary results

## Assessing model fit on the simulated data

We first assess model fit on the observed tract lengths detected from the coalescent simulation. We remove all observed tract lengths greater than 1500 bp and compare the proportion of remaining observed tracts that are 1 bp to the estimated proportion according to the model, denoted The proportion of observed tract lengths that were 1 bp was 0.807, whereas. This indicates that the model overestimates the proportion of observed tract lengths that are 1 bp long.

Next, we plot the empirical CDF of observed tract lengths between 2 and 1500 bp detected from the coalescent simulation, along with the corresponding model CDF, denoted , in Figure 2.

A graph with a red line

Description automatically generated

**Figure 2. Expected vs. empirical CDF from the coalescent simulation.** We plot the model CDF (in grey) and the empirical CDF of observed tract lengths between 2 and 1500 bp detected in the coalescent simulation (in red).

From Figure 2, we see that the model is a good fit to the empirical proportions of observed tract lengths between 2 and 1500 bp. In practice, when estimating the mean gene conversion tract length , we use the observed tract lengths between 2 and 1500 bp and truncate the distribution of between 2 and 1500 bp.

We want to know why the model overestimates the proportion of singleton observed tracts detected in the coalescent simulation. We previously listed two possible explanations for why the model may not fit the observed tract lengths well. The first explanation is that we do not consider LD when deriving the marginal distribution of the observed tract lengths in our model. However, LD will change the distribution of heterozygous markers within individuals, which will likely affect the distribution of observed tract lengths in the coalescent simulation. Another potential reason for a poor fit is because the multi-individual IBD method may detect observed tracts at different rates, depending on their lengths.1 This will also affect the empirical distribution of observed tract lengths in the coalescent simulation.

We can see whether LD is a plausible explanation for the model overestimating the proportion of singleton observed tracts by seeing whether the model will accurately estimate this proportion for a set of observed tracts that is generated without accounting for linkage disequilibrium. In the section, “Simulating observed tract lengths from chromosome 1 without accounting for linkage disequilibrium,” we described a way to generate such a set of observed tract lengths. For this set of observed tract lengths, the empirical proportion of singleton observed tracts was 0.812 whereas . Our model better predicts this proportion for the observed tract lengths generated in this simulation, compared to the observed tract lengths detected in the coalescent simulation. Next, we compare the empirical CDF of observed tract lengths between 2 and 1500 bp from this simulation to the model CDF in Figure 3.

A graph with a red line

Description automatically generated

**Figure 3. Expected vs. empirical CDF from the set of observed tract lengths generated without accounting for LD.** We plot the model CDF (in grey) and the empirical CDF of observed tract lengths between 2 and 1500 bp generated in the simulation (in red).

From Figure 3, we see that the model CDF closely matches the empirical CDF of observed tract lengths between 2 and 1500 bp generated from this simulation. Our model closely fits the empirical distribution of observed tract lengths generated from this simulation, including the singleton observed tracts.

From the above, LD may explain why the model is overestimating the proportion of singleton observed tracts in the coalescent simulation. However, an alternative explanation for this is that the gene conversion detection method described in Browning and Browning (2024) is detecting singleton observed tracts less frequently relative to larger observed tract lengths.1 We can explore whether this is the case by simulating gene conversion tracts directly on individual genomes from the coalescent simulation, using the same length distribution used to simulate gene conversion tracts in the coalescent simulation. If there is no detection bias, we expect the empirical distribution of observed tract lengths to be similar to what we detect using the multi-individual IBD method.1 We describe our method for generating observed tract lengths directly on individual genomes from the coalescent simulation in the section, “Simulating observed tract lengths from individuals in the coalescent simulation.”

In Figure 4, we plot the empirical CDF of observed tract lengths from the above simulation, and the empirical CDF of observed tract lengths detected from the coalescent simulation.

A graph with a line

Description automatically generated**Figure 4. Empirical from the coalescent simulation vs. simulating observed tracts directly on genomes from the coalescent simulation.** We plot the empirical CDF of observed tract lengths detected in the coalescent simulation (in red) and the empirical CDF of observed tract lengths simulated directly on genomes from the coalescent simulation (in blue).

We see from Figure 4 that generating the observed tract lengths directly on individual genomes from the coalescent simulation leads to an empirical CDF that is very similar to what we obtain by detecting observed tract lengths using the multi-individual IBD method.1 We also see that the proportion of singleton observed tracts is almost identical. This likely means that detection bias cannot explain why our model overestimates the proportion of singleton observed tracts from the coalescent simulation.

# References

1. Browning, S. R. & Browning, B. L. Biobank-scale inference of multi-individual identity by descent and gene conversion. *The American Journal of Human Genetics* **111**, 691–700 (2024).

2. Baumdicker, F. *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**, iyab229 (2022).