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

\* Corresponding author:

E-mail: masakin@uw.edu (NM)

# Abstract

Non-crossover gene conversion is a type of meiotic recombination characterized by the non-reciprocal transfer of genetic material between homologous chromosomes, thought to occur within a relatively short tract of DNA. Previous studies have estimated these short tracts to be in the order of 100-1000 bp in humans. However, the number of observable gene conversion tracts per study has been limited by the use of pedigree or sperm-typing data to detect past gene conversion events. In this study, we propose a statistical model to estimate the mean length of gene conversion tracts in humans, designed to be fit on a very large number of detected gene conversion tracts. We fit this model to gene conversion tracts detected using clusters of identity-by-descent (IBD) segments in whole autosome sequence data from the UK Biobank. From this dataset, we estimated the mean gene conversion tract length in humans to be 459 bp (95% CI: [457, 460]). Fitting our model on the subset of gene conversion tracts that overlapped and did not overlap with a recombination hotspot, we estimated the mean gene conversion tract length to be 418 bp (95% CI: [416, 420]) and 492 bp (95% CI: [489, 494]) respectively.

# Introduction

During meiosis, homologous chromosomes undergo genetic recombination resulting in the transfer of genetic material. Double strand breaks that occur during recombination are resolved in two distinct ways. Crossovers result in a long tract of DNA (typically spanning millions of base pairs) being exchanged between homologous chromosomes. On the other hand, non-crossover gene conversions typically result in a non-reciprocal transfer of alleles within a short tract of around 100-1000 bp.1 These gene conversion events are thought to most commonly occur via the synthesis-dependent strand annealing mechanism, where a double stranded break is repaired by the invasion of a protruding 3’ end into the donor chromatid, but may also occur through the resolution of two Holliday junctions.2

Gene conversions can be detected in humans by amplifying sperm DNA and identifying positions in which the allele of one homologous chromosome has been replaced by the other.3,4 The distance between these positions, where alleles are thought to have been converted from a gene conversion event, can be used to estimate the length of the gene conversion tract. Using SNP array and whole genome sequence data from 34 three-generation pedigrees, Williams et al. determined that tract lengths are in the order of 100-1,000 bp based on detected allele conversions. Using three-generation pedigrees helps to distinguish between allele conversions and genotype errors.1

Williams et al. also identified clusters of gene conversion tracts spanning 20-30 kb, which may have resulted from clustered but discontinuous gene conversion events during the same meiosis.1 This phenomenon has previously been referred to as complex gene conversions. Complex gene conversions as long as 100 kb, where allele conversions are detected at some markers but not others, were also found in the deCODE study.5 These complex gene conversions could arise from mechanisms such as GC biased repair across long stretches of DNA.1 In this study, we will focus on individual gene conversion tracts where the length spanning the furthest allele converted markers within the gene conversion tract is no more than 1500 bp.

Large numbers of gene conversion tracts can be detected from biobank-scale sequence data using inferred IBD clusters, a set of haplotypes at a locus that have a recent common ancestor.6 A gene conversion event occurring after the most recent common ancestor of an IBD cluster will transfer new alleles onto the haplotype, assuming that the individual in which meiosis occurs has at least one heterozygous marker within the gene conversion tract. Allele conversions cause discordant alleles within the IBD cluster in the current population, which can be used to detect past gene conversion events. Because discordant alleles can prevent the detection of the IBD cluster, Browning and Browning devised a method to use non-overlapping regions of each chromosome for detecting IBD clusters and gene conversions that have occurred on each IBD cluster. Applying their method to whole autosome sequence data from 125,361 individuals from the UK Biobank, they found 9,313,066 allele conversions inferred to belong to 5,961,128 gene conversion tracts.6

Efforts have also been made to model the length distribution of gene conversion tracts using detected gene conversion tracts in humans and other species.7–9 However, these studies use pedigree datasets (or in the case of Betran et al., a small number of *Drosophila* sequences), which only contain information about a few meioses, limiting the number of detectable gene conversion tracts. This can lead to more uncertain estimates of the mean gene conversion tract length. For example, a statistical model was used to infer the mean length of gene conversion tracts using 257 paternal and 247 maternal gene conversion tracts detected from the deCODE study, but confidence intervals span more than an order of magnitude in some cases.9

Various distributions have been suggested for gene conversion tract lengths. A geometric distribution was used to model gene conversion lengths in *Drosophila*.7 A sum of two geometric random variables and a mixture distribution with negative binomial components have also been used to model gene conversion tract lengths.8,9

In this study, we propose a parametric model to infer the mean length of gene conversion tracts using tract lengths detected from the UK Biobank whole autosome data.6 Our model is inspired by a model proposed by Betran et al., which was fit to tract lengths detected in *Drosophila subobscura*. Like in Betran et al., we refer to the length spanning the furthest allele converted markers within a gene conversion tract as the observed length of the gene conversion tract, and incorporate this as a random variable in our model.7 Within a gene conversion tract, allele conversions only occur at heterozygous positions. Thus, the observed length of a gene conversion tract will likely be shorter than the actual gene conversion tract length. Like in Betran et al., we account for this difference in length by allowing allele conversions to occur with the same probability at each position within the same gene conversion tract.7 In our study, we allow this probability to differ for each detected gene conversion tract. A geometric distribution is exclusively used to model the length distribution of gene conversion tracts in Betran et al., but in our study, we also allow the length distribution to be a sum of two geometric random variables.

For model validation, we fit our model to detected gene conversion tract lengths from a coalescent simulation incorporating gene conversions, originally described in Browning and Browning (2024).6 Our model fits the observed tract lengths well after excluding observed tract lengths of one bp and truncating the distribution of observed tract lengths in our model to be two bp or longer. Our model overestimates the frequency of observed tract lengths that are one bp, likely because we do not account for linkage disequilibrium (see Supplementary materials). Truncating our model to only consider observed tract lengths of two bp or longer, we get an accurate estimate of the mean tract length when the length distribution of gene conversion tracts is correctly specified.

By maximizing the likelihood of the set of observed gene conversion tract lengths longer than 1 bp detected in the UK Biobank whole autosome data, we estimate the mean gene conversion tract length to be 459 bp long (95% CI: [457, 460]). Furthermore, we stratify these observed tract lengths based on whether the corresponding tracts overlapped with a recombination hotspot. We classified the region spanning two adjacent markers as a recombination hotspot if the markers were more than 2 kb apart and the local recombination rate between the markers exceeded five times the background recombination rate of the autosome. Fitting our model on the subset of observed tract lengths in which the corresponding tracts overlapped and did not overlap with a recombination hotspot, we estimated the mean gene conversion tract length to be 418 bp (95% CI: [416, 420]) and 492 bp (95% CI: [489, 494]) respectively.

# Materials and methods

## UK Biobank whole autosome data

We ran our analysis on whole autosome sequence data from 125,361 individuals from the UK Biobank, who identified themselves as ‘white British’ in the initial release of 150,119 sequenced genomes.10 The data were obtained under UK Biobank application number 19934, and the 150,119 genomes were phased using Beagle 5.4.11,12

## Detecting gene conversion tracts

Browning and Browning devised a multi-individual IBD method to detect gene conversion tracts in the UK Biobank whole autosome data.6 Their method utilizes IBD clusters, a set of haplotypes at a locus that have a recent common ancestor, to detect past allele conversions (i.e., a change in alleles on a haplotype due to a gene conversion event). If a recent gene conversion event transfers new alleles onto a haplotype in the IBD cluster, we will observe discordant alleles within the IBD cluster, which can be used to detect past gene conversion events. In their method, the genome is split into short, interleaved regions in which IBD clusters are inferred or in which gene conversion tracts are detected based on the inferred IBD clusters. These regions are 9 kb long, for a total of 18 kb, and this 18 kb pattern is repeated throughout each chromosome. Furthermore, this 18 kb pattern is offset by 0, 6, and 12 kb, and the analysis repeated across each offset to ensure that allele conversions at all positions can be detected.6

For each marker within the gene conversion detection region, we detect allele conversions based on the IBD clustering of the marker (within the IBD clustering region) that is closest in terms of genetic distance. Only markers with MAF greater than or equal to 5% are considered when detecting allele conversions to prevent mutations from being detected as allele conversions. To detect an allele conversion at a position, the corresponding IBD cluster must contain at least two copies of two alleles in order to protect against sporadic genotype errors.6

After allele conversions are detected, they are clustered to form observed gene conversion tracts. Allele conversions are considered to belong to the same gene conversion tract if they are located within 1500 bp of each other, and if the membership of the two sub-clusters (representing the two alleles present in the IBD cluster) overlaps for the two allele conversions.6

After clustering allele conversions to form gene conversion tracts within each offset, the results are combined across offsets. Only tracts that start within the central 6 kb of the 9 kb gene conversion detection region for the corresponding offset are retained. This is because tracts starting at the ends of the detection region are likely to protrude into the neighboring region in which allele conversions are not detected. This also prevents double counting any tracts.

The detection of gene conversion tracts in the UK Biobank whole autosome data was previously performed in the multi-individual IBD paper, where additional settings are described.6 Across all the autosomes, 9,313,066 allele conversions were detected and these allele conversions were inferred to belong to 5,961,128 gene conversion tracts. 82.9% of the detected tracts had only one allele conversion.6 We refer to the length spanning the furthest allele converted markers in each detected gene conversion tract as the observed tract length of the gene conversion tract. If only one allele conversion is in the detected tract, the observed tract length is one.

We label the detected observed tract lengths as . Our gene conversion detection method is likely to truncate larger tracts, because of how detected gene conversion tracts are aggregated across the three offsets. To account for this, we exclude any observed tract lengths larger than 1500 bp when estimating the mean gene conversion tract length. We also exclude observed tract lengths of 1 bp prior to estimation, because the proportion of these tracts is overestimated by our model. This is likely because we do not account for linkage disequilibrium in our model. The effect of linkage disequilibrium on the distribution of the observed tract lengths is further discussed in the Supplementary Materials.

## Definitions and overview of model

Our model follows the general framework described in Betran et al.7 We let be a geometric random variable, or (extending Betran et al.) a sum of two identically distributed geometric random variables, representing the length of a single gene conversion tract. We parameterize the distribution of by its mean . We further let be a random variable representing the observed tract length of a gene conversion tract, which is the length spanning the furthest allele converted markers within a gene conversion tract. The event represents no allele conversions occurring within the tract, and represents one allele conversion occurring within the tract. In the following sections, we derive the conditional distribution of and the marginal distribution of . We further describe the procedure we use to obtain a maximum likelihood estimate of ,  , using the observed tract lengths detected from the UK Biobank whole autosome data.

## Deriving the distribution of

The observed tract length of a gene conversion tract, represented by the random variable , depends on where allele conversions occur on the gene conversion tract. We will first assume that allele conversions happen with probability  at every position within some gene conversion tract that is exactly bp long. Under this scenario, the following conditional distribution is derived in Betran et al.7

In the probability above, we conditioned on the gene conversion tract length, represented by the random variable , being bp long. Obtaining an observed tract length of zero is equivalent to allele conversions not occurring within the gene conversion tract, which happens with a probability of . Next, obtaining an observed tract length of one is equivalent to an allele conversion occurring at exactly one position within the gene conversion tract. There are possible positions in which the allele conversion can occur, and each configuration happens with a probability of . Finally, to obtain an observed tract length of , where , we need to observe two allele conversions that span exactly positions, and allele conversions cannot occur at the positions flanking the two allele conversions. There are ways to overlay these two allele conversions on the gene conversion tract, and each configuration occurs with a probability of .

## Deriving the marginal distribution of

If is geometric with mean , we have,

Letting ,

We have derived the marginal distribution of . However, we do not observe tracts with length zero in our dataset. Furthermore, recall that we only retain observed tract lengths between 2 and 1500 bp during estimation (as mentioned at the end of the section, Detecting gene conversion tracts), so we account for this by truncating the distribution of between 2 and 1500.

We have,

Then,

Notice that conditioning on removed the parameter from our model.

As mentioned earlier, represents the observed tract lengths in our dataset. When fitting the model, we use the filtered set of observed tracts, . Henceforth, we will also index our random variable using . represents the random variable corresponding to the observed tract length for tract in our dataset. We have,

Finally, we consider the case when follows a sum of two identically distributed geometric random variables. The derivation of under this setting is included in the Appendix. When is a sum of two identically distributed geometric random variables, depends on , so we estimate for each tract before estimating . The procedure to estimate for each tract is described in the following section.

## Estimating for each observed tract

Recall that is the probability that an allele conversion will occur at each marker within a gene conversion tract. We will allow this probability to differ by tract because marker density varies across the genome. represents the probability that an allele conversion will occur at each position within gene conversion tract . When is a sum of two geometric random variables, the likelihood of each observed tract,, depends on (see Appendix), so we need to estimate for to fit the model.

Allele conversions occur at positions within each gene conversion tract where the individual is heterozygous. Therefore, the probability that an individual is heterozygous at a given SNV marker can be used to estimate the probability that an allele conversion will happen at this marker, once it is included in a gene conversion tract. However, it is difficult to derive a closed form expression for the marginal distribution of when we only allow allele conversions to occur at SNV markers, and with differing rates at each SNV marker. Thus, we let allele conversions occur with the same probability at all positions within gene conversion tract . We use the heterozygosity rate of positions near tract to estimate .

Denoting the th observed tract as , where and represent the positions corresponding to the ends of the observed tract, we average the heterozygosity rate across the set of positions to estimate :

Here, denotes the minor allele frequency of position on the chromosome in which the gene conversion event occurred. is calculated using the sample of 125,361 White British individuals from the UK Biobank. Furthermore, variants with MAF less than 5% were excluded when detecting allele conversions, so we cannot observe allele conversions at these positions (see the section, Detecting gene conversion tracts). Therefore, if the MAF is less than 5% at position , we set .

If either or exceeds the end of the chromosome, the averaging only takes place within the bounds of the chromosome (e.g. if and , we only average the heterozygosity rate from positions 1 to 5200).

## Maximum likelihood estimation of

Given observed tract lengths detected from the UK Biobank whole autosome data, we propose the following maximum likelihood estimator for when is geometric. Recall that the version of the model in which is geometric was parameterized by , but we can simply maximize with respect to . In other words,

where . When is a sum of two geometric random variables, we parameterize the distribution of using (see Appendix). Unlike the geometric case, our distribution of truncated between 2 and 1500 still depends on , so for each , we plug in our estimated in place of . Then, we can again maximize with respect to :

To find the argmax, we use Brent’s method, implemented in the optim function in R.13

To select the distribution of , we propose calculating the Akaike Information Criterion (AIC) under each version of the model.14 Lower AIC indicates that the selected distribution of is a better fit to the data.

## Bootstrap confidence intervals

We calculate 95% bootstrap confidence intervals for . We denote the number of observed tracts with length between 2 and 1500 bp as . To obtain each bootstrap sample, we sample with replacement observed tracts from the set of observed tracts . Each bootstrap sample consists of the set of tract lengths and probabilities corresponding to the resampled tracts.

We refit our model to 500 bootstrap samples and obtain a new maximum likelihood estimate of for each bootstrap sample. We take the 0.025 and 0.975 quantiles of the resulting bootstrap distribution of  and use this as the bounds of our 95% bootstrap confidence interval.

## Simulation study using a coalescent model with gene conversions

We use simulated data described in Browning and Browning (2024). 6 20 replicates of length 10 Mb were simulated for 125,000 individuals. The demographic model for the simulation was an exponentially growing population with an initial size of 10,000 and a growth rate of 3% per generation for the past 200 generations. To simulate recombination and mutation, a constant recombination rate of 1 cM/Mb and a mutation rate of per bp per meiosis were used. Gene conversions were simulated with an initiation rate of 0.02 per Mb and gene conversion lengths were simulated from a geometric distribution with a mean tract length of 300 bp. msprime v1.2 was used to perform the simulation.15 The processes used to add uncalled deletions and genotype errors are described in Browning and Browning.6 Variants with MAF 0.01 were excluded, the phase information was removed, and Beagle 5.4 was used to statistically phase the genotypes.11 Variants with MAF smaller than 5% were removed when detecting allele conversions to prevent mutations being detected as allele conversions.6 The multi-individual IBD analysis detected 284,838 allele conversions comprising 226,007 gene conversion tracts across the 20 replicates.6

We fit our model under two settings, one assuming a geometric distribution and the other assuming a sum of two geometric random variables for , to the data from each replicate, obtaining  as described in the section, Maximum likelihood estimation of . Because the true tract lengths in this simulation study are drawn from a geometric distribution, we are interested in whether the version of the model in which is geometric will be favored using AIC. 95% bootstrap confidence intervals for under both versions of the model are calculated as described in the section, Bootstrap confidence intervals.

## UK Biobank analysis

Recall that we detect the observed tract lengths from the UK Biobank whole autosome data using the multi-individual IBD method. In the section, Maximum likelihood estimation of , we describe our method for estimating , the mean gene conversion tract length, using the observed tract lengths. We further obtain a 95% bootstrap confidence interval for using the method described in the section, Bootstrap confidence intervals.

We ran an additional stratified analysis, stratifying observed tract lengths by whether they overlapped with a recombination hotspot. We used the deCODE genetic map to define recombination hotspots on each autosome.16 For each autosome, we first calculated a background recombination rate by dividing the genetic distance between the two most distant markers on the genetic map (in cM) by their physical distance (in Mb). Next, we similarly calculated local recombination rates between nearby markers on this autosome by dividing the genetic distance between the two markers by their physical distance. Initially, we calculate the local recombination rate between the first marker on the autosome in the genetic map, and the marker closest to it that is distant by at least 2 kb. We next calculate the local recombination rate between this newly identified marker and the marker closest to it that is distant by at least 2 kb. We repeat this process until the last marker on this autosome is included in a local recombination rate calculation, or until we cannot identify further markers that are at least 2 kb away.

If the local recombination rate between two markers is more than five times the background recombination rate of the autosome, we classify the region spanning these markers as a recombination hotspot. We cluster adjacent recombination hotspots together. We stratify the observed tract lengths  based on whether each tract overlapped with a recombination hotspot. For each subset, we then obtain a maximum likelihood estimate and a 95% bootstrap confidence interval for .

# Results

## Simulation study using a coalescent model with gene conversions

We fit our model to the observed gene conversion tracts from each replicate in the simulation study. Recall that a geometric distribution was used to simulate the gene conversion tract lengths in this simulation study. We plot our estimates of (assuming both a geometric distribution and a sum of two geometric random variables for ) from each replicate in Figure 1. The mean estimate of across the 20 replicates was 289 under the geometric setting, which is slightly lower than the true value of 300 used to simulate the gene conversion tracts. Additionally, the true value of 300 was contained in our 95% bootstrap confidence intervals in 15 out of the 20 replicates. However, when we assume a sum of two geometric random variables for , the mean estimate of across the 20 replicates was 421, which is much higher than the true value of 300. Furthermore, none of our 95% bootstrap confidence intervals captured the true value of 300.

A graph of a gene conversion

Description automatically generated

**Figure 1. Estimated mean tract lengths across replicates in simulation study.** The dotted horizontal line represents the true mean gene conversion tract length that was used to generate the observed tract lengths in the simulation. Gene conversion tract lengths were simulated under a geometric distribution, and analyses were conducted assuming that the tract lengths are geometric (red) or a sum of two geometric random variables (blue). We plot our estimate and 95% bootstrap confidence interval under both settings of the model for each replicate of the simulation.

Based on the AIC, the version of the model in which was set to be geometric was a better fit in all 20 replicates. The difference in AIC (the AIC when is geometric subtracted from the AIC when is a sum of two geometric random variables) ranged from 11 to 41 across the 20 replicates.

## UK Biobank analysis

We applied our model to the observed tract lengths detected from the UK Biobank whole autosome data. When assuming that is geometric, our model estimated the mean gene conversion tract length to be 459 bp (95% CI: [457, 460]). When assuming that is a sum of two geometric random variables, our model estimated the mean gene conversion tract length to be 649 bp (95% CI: [648, 651]). The setting in which was set to be geometric had lower AIC, and the difference in AIC between the two settings was 66,237.

We next detected recombination hotspots on all 22 autosomes using the procedure described earlier (see the UK Biobank analysis section in Materials and methods). We found 32,279 recombination hotspots on all autosomes, with the longest hotspot being 51,470 bp on chromosome 13. In Supplementary figure 1, we plot the recombination hotspots that we found on chromosome 21.

A graph of a number of dots

Description automatically generated

**Supplementary figure 1. Recombination hotspots found on chromosome 21.** Hotspots are highlighted in red. Local recombination rates, represented by the dots, were calculated between nearby markers on the genetic map that were at least 2 kb apart. The x-axis positions of the dots represent the midpoint between each pair of markers in which a local recombination rate was calculated. The black horizontal line indicates the threshold of five times the background recombination rate for the autosome (9.82 cM/Mb). If the local recombination rate between two markers exceeds this threshold, we classify the region spanning these markers as a recombination hotspot.

Taking the subset of observed tract lengths in which the corresponding tracts overlapped with a recombination hotspot (on any of the autosomes), we reran the analysis. For these observed tract lengths, our model, when assuming a geometric , estimated the mean gene conversion tract length to be 418 bp (95% CI: [416, 420]). When assuming that is a sum of two geometric random variables, our model estimated the mean gene conversion tract length to be 598 bp (95% CI: [596, 601]).

For the subset of observed tract lengths in which the corresponding tracts did not overlap with a recombination hotspot, our model, when assuming a geometric , estimated the mean gene conversion tract length to be 492 bp (95% CI: [489, 494]). When assuming that is a sum of two geometric random variables, our model estimated the mean gene conversion tract length to be 689 bp (95% CI: [687, 692]). In both subsets, the AIC was smaller when was set to be geometric.

# Discussion

Previous studies have tried to measure gene conversion tract lengths in humans by detecting allele converted markers from pedigree and sperm-typing data.1,3–5 However, in these studies, it is only possible to detect gene conversion events occurring in a relatively small number of meioses. Efforts to detect gene conversions from pedigree data have been limited by the number of multi-generational pedigrees that have been genotyped. Sperm-typing studies have been limited by the availability of appropriate data and by the difficulty in distinguishing genotype errors from allele conversions in this setting. A statistical model has been proposed to infer the length distribution of gene conversion tracts in humans,9 but the small number of detected gene conversion tracts has made it difficult to estimate the mean gene conversion tract length with precision.

By applying the multi-individual IBD method to the UK Biobank whole autosome data, we can detect gene conversion events across multiple meioses in the ancestral history of this population.6 Using this method, 5,961,128 gene conversion tracts were detected, which is at least several orders of magnitude larger than what had been detected in humans in the past. Only around 2,000 gene conversion events were detected from a combination of 7,219 proband-family sets genotyped with a SNP chip and 101 whole-genome sequenced proband-family sets.5

We next proposed a parametric model, inspired by a previous model by Betran et al.,7 to infer the mean gene conversion tract length from a large number of detected gene conversion tracts. In our model, the length distribution of gene conversion tracts can be specified to either be geometric or a sum of two geometric random variables, and it is possible to select the better fitting setting based on AIC.

When using our model to estimate the mean gene conversion tract length, we first removed observed tract lengths greater than 1500 bp because the multi-individual IBD method is likely to truncate longer observed tracts when detecting these from sequence data (see the section, Detecting gene conversion tracts). Furthermore, we removed any observed tract lengths of 1 bp because our model overestimates the proportion of these tracts, likely because it does not account for linkage disequilibrium (see Supplementary materials). To account for omitting these tract lengths, we truncate the distribution of observed tract lengths between 2 and 1500 bp in our model.

We used a coalescent simulation incorporating gene conversion events to validate our parametric model. We found that our model accurately estimated the mean length when the length distribution of gene conversion tracts was correctly specified to be geometric. Our model resulted in biased estimates of the mean gene conversion tract length when the length distribution was incorrectly specified. To assess the robustness of our model to misspecification of the tract length distribution, we ran a separate simulation study (see the section, Simulation study to assess the robustness of the model, in the Appendix). We see from this study that the AIC selected model results in relatively unbiased estimates across a range of true tract length distributions.

We then fit our model to detected gene conversion tracts from the UK Biobank whole autosome data. We estimated the mean gene conversion tract length to be 459 bp (95% CI: [457, 460]) using the setting in which gene conversion tract lengths were assumed to be geometric, which resulted in a smaller AIC compared to when we assumed this to be a sum of two geometric random variables.

This result is consistent with some previous findings. Hardarson et al. estimate the mean paternal and maternal gene conversion tract length to be 177 bp (95% CI: [61.0, 389]) and 41.9 bp (95% CI: [16.4, 2925]) respectively, based on 504 gene conversion tracts detected in sequenced families (257 paternal and 247 maternal gene conversion tracts).9 Because of the wide confidence intervals for the mean tract length, our estimate is consistent with their findings. In contrast, Jeffreys and May estimate the mean length to be in the range of 55-290 bp based on minimum and maximum possible lengths of detected gene conversion tracts determined from allele converted markers.4 Our estimate of 459 bp is not inside this range.

We further ran a stratified analysis based on whether the detected gene conversion tracts from the UK Biobank whole autosome data overlapped with a recombination hotspot, defined to be the region spanning two nearby markers in which the local recombination rate is five times that of the background recombination rate of the autosome. Applying our model on just the tracts that overlapped with a recombination hotspot, we estimated the mean gene conversion tract length to be 418 bp (95% CI: [416, 420]). On the other hand, when applying our model to just the tracts that did not overlap with a recombination hotspot, we estimated the mean gene conversion tract length to be 492 bp (95% CI: [489, 494]). For both subsets of observed tracts, we used the setting in which was assumed to be geometric, which resulted in a smaller AIC value. Although we found the mean gene conversion tract lengths from these two subsets to differ significantly, this difference may be attributable to technical factors, such as reduced phasing accuracy at recombination hotspots. We recommend further analysis to confirm this preliminary result.

# References

1. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4**, e04637 (2015).

2. McMahill, M. S., Sham, C. W. & Bishop, D. K. Synthesis-Dependent Strand Annealing in Meiosis. *PLoS Biol* **5**, e299 (2007).

3. Odenthal-Hesse, L., Berg, I. L., Veselis, A., Jeffreys, A. J. & May, C. A. Transmission Distortion Affecting Human Noncrossover but Not Crossover Recombination: A Hidden Source of Meiotic Drive. *PLOS Genetics* **10**, e1004106 (2014).

4. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* **36**, 151–156 (2004).

5. Halldorsson, B. V. *et al.* The rate of meiotic gene conversion varies by sex and age. *Nat Genet* **48**, 1377–1384 (2016).

6. 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).

7. Betran, E., Rozas, J., Navarro, A. & Barbadilla, A. The Estimation of the Number and the Length Distribution of Gene Conversion Tracts from Population DNA Sequence Data. *Genetics* **146**, 89–99 (1997).

8. Wall, J. D., Robinson, J. A. & Cox, L. A. High-Resolution Estimates of Crossover and Noncrossover Recombination from a Captive Baboon Colony. *Genome Biology and Evolution* **14**, evac040 (2022).

9. Hardarson, M. T., Palsson, G. & Halldorsson, B. V. NCOurd: modelling length distributions of NCO events and gene conversion tracts. *Bioinformatics* **39**, btad485 (2023).

10. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).

11. 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).

12. Browning, B. L. & Browning, S. R. Statistical phasing of 150,119 sequenced genomes in the UK Biobank. *The American Journal of Human Genetics* **110**, 161–165 (2023).

13. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2024).

14. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974).

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

16. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).

# Appendix

## Deriving the marginal distribution of when is a sum of two geometric random variables

We consider the case in which is distributed as a sum of two independent and identically distributed geometric random variables each with mean . The derivation of is similar to when is geometric. We have,

.

Letting ,

.

Then,

Finally,

Notice that unlike the case where is geometric, depends on .

Similarly to the case where is geometric, we index our random variable using so that represents the random variable corresponding to the observed tract length for tract in our dataset. This time, we also index using so that an allele conversion happens with probability  at every position within the th gene conversion tract (the estimation of is described in the section, Estimating for each observed tract ). We have,

## Simulation study to assess the robustness of the model

We ran a simulation study to assess how well our model can estimate the mean tract length when we misspecify the length distribution of gene conversion tracts. Recall that in our model, we allow this distribution to be geometric or a sum of two geometric random variables.

In this simulation study, we simulate observed tract lengths using four distributions for the length distribution of gene conversion tracts:

1. Geometric distribution with mean 300
2. Sum of two geometric random variables, each with mean 150
3. Sum of three geometric random variables, each with mean 100
4. Uniform distribution with support from 1 to 599

A graph with lines and lines on a black background

Description automatically generated

**Figure 2. Density of the four distributions used to simulate gene conversion tract lengths.** We plot the density of the geometric distribution, the sum of two geometric random variables, the sum of three geometric random variables, and the uniform distribution that we draw the gene conversion tract lengths from in the simulation study.

All four distributions have mean 300. In Figure 2, we plot the four distributions that we draw the gene conversion tract lengths from. We simulate the observed tract lengths by simulating gene conversion tracts on individuals from the coalescent simulation (see the section, Simulation study using a coalescent model with gene conversions). Recall that in each replicate of the coalescent simulation, we simulated genotype data for 125,000 individuals within a 10 Mb region. To simulate the observed tract lengths, we only use the genotype data from the first replicate of the coalescent simulation. To simulate one set of observed tract lengths, we first sample individuals with replacement from the 125,000 individuals. For each resampled individual, we follow these steps:

1. We randomly select a starting position for the gene conversion tract, chosen uniformly across the 10 Mb region.
2. We draw the length of the gene conversion tract from one of the four specified distributions.
3. We determine the observed tract length as the length spanning the furthest heterozygous markers within the simulated gene conversion tract.

This procedure results in observed tract lengths (some of which may be length zero due to the absence of heterozygous markers within the simulated gene conversion tracts). For each of the four distributions listed earlier, we repeated this procedure 100 times to obtain 100 sets of observed tract lengths. Then, we fit our model under both settings for (geometric and sum of two geometric random variables), to each set of observed tract lengths (after retaining tract lengths between 2 and 1500 bp). For each set of observed tract lengths, we obtained both a point estimate and a 95% bootstrap confidence interval for . The empirical bias of our estimates under each setting of is shown in Table 1. Under the AIC selected setting, we use the estimate from the setting of with the smaller AIC value in each of the 100 sets.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Bias | | |
| Geometric | Sum of geometric | AIC selected |
| Geometric | -14.1 | 116.3 | -7.5 |
| Sum of geometric | -103.3 | -9.6 | -28.5 |
| Negative binomial | -132.5 | -52.9 | -52.9 |
| Uniform | -143.9 | -71.0 | -71.0 |

**Table 1. Bias from simulation study to assess robustness.** We report the empirical bias of our estimates across 100 replicates for each distribution used to simulate the gene conversion tract lengths and for each setting of . Under the AIC selected setting, we use the estimate from the model with the smaller AIC value in each of the 100 replicates.

We also calculated the coverage of our 95% bootstrap confidence intervals. When the gene conversion tracts were simulated from a geometric distribution, and we specified to be geometric in our model, our 95% confidence intervals covered the true mean of 300 in 11 out of the 20 replicates (55% of the time). Similarly, when the gene conversion tracts were simulated from a sum of two geometric random variables, and we specified to be this distribution in our model, our 95% confidence intervals covered the true mean of 300 in 14 out of the 20 replicates (70% of the time). When we simulated the gene conversion tract lengths from the remaining two distributions, the coverage was 0% under both settings of the model.