Mean gene conversion tract length in humans estimated to be 459 bp from UK Biobank sequence data

Nobuaki Masaki[1], Sharon R. Browning[1]

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

Address for correspondence: masakin@uw.edu (NM), sguy@uw.edu (SRB)

Abstract

Non-crossover gene conversion is a type of meiotic recombination characterized by the non-reciprocal transfer of genetic material between homologous chromosomes. Gene conversions are thought to occur within relatively short tracts of DNA, estimated to be in the order of 100-1,000 bp in humans. However, the number of observable gene conversion tracts per study has so far been limited by the use of pedigree or sperm-typing data to detect gene conversion events. In this study, we propose a statistical method to estimate the mean length of gene conversion tracts in humans. Our method can handle a large number of gene conversion tracts, leading to more precise estimates of the mean tract length. We apply our method to gene conversion tracts detected in whole autosome sequence data from the UK Biobank using clusters of identity-by-descent segments. From this dataset, we estimate the mean gene conversion tract length in humans to be 459 bp (95% CI: [457, 461]). Stratifying detected gene conversion tracts by whether they overlapped with a recombination hotspot, we estimate the mean gene conversion tract length to be 418 bp (95% CI: [416, 420]) and 492 bp (95% CI: [489, 494]) respectively, for tracts that overlap and do not overlap with a recombination hotspot.

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 100-1,000 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. Gene conversion events may also occur via the resolution of two Holliday junctions.[2]

Gene conversions can be detected in humans by analyzing sequence data from pedigrees or sperm samples and identifying positions in which the allele of one homologous chromosome has been replaced by the other.[1,3–5] The distance between these positions, where alleles are thought to have been converted by 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] It can be difficult to distinguish between allele conversions and genotype errors when using two-generation pedigrees or sperm samples.

Williams et al. further identified apparent clusters of gene conversion tracts spanning 20-30 kb, which may have resulted from discontinuous gene conversion events occurring in close proximity during the same meiosis.[1] This phenomenon has previously been referred to as complex gene conversions. Complex gene conversions as long as 100 kb were also found in the deCODE study.[5] 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 does not exceed 1.5 kb.

Large numbers of gene conversion tracts can be detected from biobank-scale sequence data using inferred identity-by-descent (IBD) clusters.[6] A gene conversion event occurring after the most recent common ancestor of an IBD cluster will transfer new alleles onto the haplotype if the individual undergoing meiosis 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. To detect an allele conversion, this method requires at least two haplotypes within an IBD cluster to have the same alternate allele. This means that genotype errors will not be falsely identified as allele conversions, unless the same genotype error occurs twice in the same IBD cluster.[6]

Efforts have been made to model the length distribution of gene conversion tracts using detected gene conversion tracts in humans and other species.[7,8] However, these studies use pedigree datasets, which only contain information about a small number of meioses. This limits the number of detectable gene conversion tracts, leading to high uncertainty in estimates of the mean gene conversion tract length. A statistical model assuming a mixture of two negative binomial distributions for tract lengths was applied to 257 paternal and 247 maternal gene conversion tracts detected from the deCODE study. Confidence intervals for the mean gene conversion tract length are wide, spanning more than two orders of magnitude for maternal gene conversion tracts.[8]

In our study, we propose a parametric model to infer the mean length of gene conversion tracts detected from the UK Biobank whole autosome data. Our model is inspired by the model by Betran et al., which was fit to tract lengths detected in *Drosophila subobscura*.[9] Like 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. Within a gene conversion tract, allele conversions only occur at positions where the individual is heterozygous. Thus, the observed length of a gene conversion tract will likely be shorter than the actual gene conversion tract length. 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. Betran et al. use the same allele conversion probability for nearby gene conversion tracts, but we allow this probability to differ for each detected gene conversion tract. Betran et al. use a geometric distribution

to model the length distribution of gene conversion tracts.[9] We allow the length distribution to be either a single geometric random variable or a sum of two geometric random variables.

We validate our model by fitting it to detected gene conversion tracts from a coalescent simulation incorporating gene conversions, originally described in Browning and Browning (2024).[6] Next, we use our model to estimate the mean length of gene conversion tracts detected from the UK Biobank whole autosome data. Finally, we stratify these detected gene conversion tracts by whether they overlap with a recombination hotspot, and use our model to estimate the mean length of gene conversion tracts in each stratum.

Subjects 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

We used gene conversion tracts previously detected in the UK Biobank whole autosome data using IBD clusters.[6] IBD clusters are sets of haplotypes at a locus that have a recent common ancestor. If a recent gene conversion event transfers new alleles onto a haplotype in the IBD cluster, there will be discordant alleles within the IBD cluster, which can then be used to detect this gene conversion event. The detection method splits the genome into short, interleaved regions where IBD clusters are inferred or where gene conversion tracts are detected based on the inferred IBD clusters. These regions were each 9 kb long, for a total of 18 kb per IBD inference and gene conversion detection region pair, and this 18 kb pattern was

repeated throughout each chromosome. Furthermore, this 18 kb pattern was offset by 0, 6, and 12 kb, and the analysis repeated for each offset to ensure that allele conversions at all positions could be detected.[6]

For each marker within the gene conversion detection region, allele conversions were detected based on the IBD clustering of the marker (within the IBD inference region) that was closest in terms of genetic distance. To detect an allele conversion at a position, two haplotypes were required to share one allele, and two other haplotypes were required to share another allele in the corresponding IBD cluster. This requirement prevents the method from falsely detecting genotype errors as allele conversions. Furthermore, only markers that had a MAF of at least 5% were considered when detecting allele conversions to prevent mutations from being detected as allele conversions.[6]

After allele conversions were detected, they were clustered to form detected gene conversion tracts. Allele conversions were considered to belong to the same gene conversion tract if they were located within 1.5 kb 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.

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

Across all the autosomes, 9,313,066 allele conversions were detected. These allele conversions were inferred to belong to 5,961,128 detected gene conversion tracts. Furthermore, 82.9% of the detected gene conversion tracts were comprised of a single allele conversion.[6] We refer to the length spanning the furthest allele converted markers in a detected gene conversion tract as the observed tract length of the

gene conversion tract. If a detected gene conversion tract is comprised of a single allele conversion, the observed tract length is one bp.

We label the observed tract lengths of all detected gene conversion tracts as $\{\ell_j | j = 1, \ldots, m\}$. The procedure used to detect gene conversion tracts in each offset assumes that gene conversion tract lengths do not exceed 1.5 kb. To take this into account, we exclude any observed tract lengths exceeding 1.5 kb when estimating the mean gene conversion tract length. We also exclude observed tract lengths of one bp prior to estimation, because the proportion of observed tract lengths of one bp is overestimated by our model (see Supplementary Materials). This is likely because we do not account for linkage disequilibrium in our model. The effect of linkage disequilibrium on the distribution of 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. (1997).[9] We model N, the length of a gene conversion tract, as a geometric random variable, or (extending the model by Betran et al.) a sum of two independent and identically distributed geometric random variables. We further let L 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 the gene conversion tract. The event $L = 0$ represents no allele conversions occurring within the tract, and $L = 1$ represents one allele conversion occurring within the tract. In the following sections, we derive the conditional distribution of L given N and the marginal distribution of L. We further describe the procedure we use to obtain a maximum likelihood estimate of ϕ, $\hat{\phi}$, using the observed tract lengths $\{\ell_j | j = 1, \ldots, m\}$ detected from the UK Biobank whole autosome data.

The distribution of the observed tract length conditional on the gene conversion tract length

The observed tract length of a gene conversion tract, represented by the random variable L, depends on where allele conversions occur on the gene conversion tract. We will first assume that allele conversions happen with probability ψ at every position within some gene conversion tract that is exactly n bp long. Under this scenario, the following conditional distribution is derived by Betran et al.[9]

$$P(L = \ell | N = n) = \begin{cases} (1-\psi)^n & \text{if } \ell = 0 \\ n\psi(1-\psi)^{n-1} & \text{if } \ell = 1 \\ (n-l+1)\psi^2(1-\psi)^{n-\ell} & \text{if } 2 \leq \ell \leq n \end{cases}.$$

In the probability above, we conditioned on the gene conversion tract length, represented by the random variable N, being n bp long. Obtaining an observed tract length of zero bp is equivalent to allele conversions not occurring within the gene conversion tract, which happens with a probability of $(1-\psi)^n$. Next, obtaining an observed tract length of one bp is equivalent to an allele conversion occurring at exactly one position within the gene conversion tract. There are n possible positions in which the allele conversion can occur, and each configuration happens with a probability of $\psi(1-\psi)^{n-1}$. Finally, to obtain an observed tract length of ℓ bp, where $2 \leq \ell \leq n$, we need to observe two allele conversions that span exactly ℓ positions, and allele conversions cannot occur at the $n-\ell$ positions flanking the two allele conversions. There are $n-\ell+1$ ways to overlay these two allele conversions on the gene conversion tract, and each configuration occurs with a probability of $\psi^2(1-\psi)^{n-\ell}$.

Deriving the marginal distribution of the observed tract length

If the gene conversion tract length N is drawn from geometric distribution with mean ϕ, we have,

$$P(N = n) = \left(1 - \frac{1}{\phi}\right)^{n-1} \frac{1}{\phi}.$$

158 Letting $\lambda = 1/\phi$,

159 $$P(L = \ell) = \sum_{n=\ell}^{\infty} P(L = \ell | N = n) P(N = n)$$

160 $$= \begin{cases} \dfrac{\lambda(1-\psi)}{\lambda + \psi - \lambda\psi} & \text{if } \ell = 0 \\ \dfrac{\lambda\psi}{(\lambda + \psi - \lambda\psi)^2} & \text{if } \ell = 1 \\ \dfrac{\lambda(1-\lambda)^{\ell-1}\psi^2}{(\lambda + \psi - \lambda\psi)^2} & \text{if } \ell \geq 2 \end{cases}.$$

161 This is the marginal distribution of the observed tract length L. However, we do not observe tracts with
162 an observed tract length of zero bp in our dataset. Furthermore, recall that we only retain observed tract
163 lengths between 2 and 1,500 bp during estimation (as described above), so we account for this by
164 truncating the distribution of L between 2 and 1,500 bp.

165 We have,

166 $$P(2 \leq L \leq 1500) = \sum_{l=2}^{1500} \frac{\lambda(1-\lambda)^{\ell-1}\psi^2}{(\lambda + \psi - \lambda\psi)^2} = \frac{\psi^2[(1-\lambda) - (1-\lambda)^{1500}]}{(\lambda + \psi - \lambda\psi)^2}.$$

167 Then,

168 $$P(L = \ell | 2 \leq L \leq 1500) = \frac{P(L = \ell)}{P(2 \leq L \leq 1500)} = \frac{\lambda(1-\lambda)^{\ell-1}}{[(1-\lambda) - (1-\lambda)^{1500}]}.$$

169 Notice that conditioning on $2 \leq L \leq 1500$ removed the parameter ψ from our model.

170 As mentioned earlier, $\{\ell_j | j = 1, \dots, m\}$ represents the observed tract lengths in our dataset. When fitting
171 the model, we use the filtered set of observed tract lengths, $\{\ell_j | j = 1, \dots, m,\ 2 \leq \ell_j \leq 1500\}$.
172 Henceforth, we will also index our random variable L using j. L_j represents the random variable
173 corresponding to the observed tract length of detected gene conversion tract j in our dataset. We have,

$$P\left(L_j = \ell_j \middle| 2 \le L_j \le 1500, \lambda\right) = \frac{\lambda(1-\lambda)^{\ell_j - 1}}{[(1-\lambda) - (1-\lambda)^{1500}]}.$$

Finally, we consider the case when N follows a sum of two independent and identically distributed geometric random variables. The derivation of $P\left(L_j = \ell_j \middle| 2 \le L_j \le 1500\right)$ under this setting is included in the Appendix. Under this setting, $P\left(L_j = \ell_j \middle| 2 \le L_j \le 1500\right)$ depends on ψ_j, so we estimate ψ_j for each tract j before estimating ϕ. The procedure to estimate ψ_j for each tract j is described in the following section.

Estimating the allele conversion probability for each detected tract

Recall that ψ_j represents the probability that an allele conversion will occur at each position within detected gene conversion tract j. When N is a sum of two geometric random variables, the likelihood of the observed tract length for detected gene conversion tract j, $P\left(L_j = \ell_j \middle| 2 \le L_j \le 1500\right)$, depends on ψ_j (see Appendix), so we need to estimate ψ_j for $j = 1, \ldots, m$ to obtain a maximum likelihood estimate for the mean gene conversion tract length ϕ.

Allele conversions occur at positions within each gene conversion tract where the individual is heterozygous. Therefore, the probability that a randomly selected individual from the population is heterozygous at a given 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 L when we only allow allele conversions to occur at SNV positions, and with differing rates at each SNV position. Thus, we let allele conversions occur with the same probability ψ_j at all positions within detected gene conversion tract j. We use the average heterozygosity rate of positions near detected tract j to estimate ψ_j.

Letting a_j and b_j ($a_j \leq b_j$) represent the positions on the chromosome corresponding to the furthest allele converted markers within detected gene conversion tract j, we average the heterozygosity rate across the set of positions $[a_j - 5000, b_j + 5000]$ to estimate ψ_j:

$$\hat{\psi}_j = \frac{1}{b_j - a_j + 10001} \sum_{i=a_j-5000}^{b_j+5000} 2p_i(1-p_i).$$

Here, p_i denotes the MAF of position i on the chromosome in which the gene conversion event occurred. p_i is calculated using the sample of 125,361 White British individuals from the UK Biobank. 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 i, we set $p_i = 0$. The formula $2p(1-p)$ for heterozygosity at a marker assumes that Hardy-Weinberg equilibrium holds, which is a reasonable approximation for common variants in a relatively homogeneous population.

If either $a_j - 5000$ or $b_j + 5000$ exceeds the end of the chromosome, the averaging only takes place within the bounds of the chromosome (e.g. if $a_j = 100$ and $b_j = 200$, we only average the heterozygosity rate from positions 1 to 5,200).

Maximum likelihood estimation of the mean gene conversion tract length

Given observed tract lengths $\{\ell_j | j = 1, \ldots, m\}$, we propose the following maximum likelihood estimator for ϕ, the mean gene conversion tract length, when the gene conversion tract length N is drawn from a geometric distribution. Recall that the version of the model in which N is geometric was parameterized by $\lambda = 1/\phi$, but we can simply maximize with respect to ϕ. In other words,

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}} \sum_{j \in I_2^{1500}} \log P(L_j = \ell_j | 2 \leq L_j \leq 1500, \phi),$$

where $I_2^{1500} = \{j = 1, \ldots, m | 2 \leq \ell_j \leq 1500\}$. When N is a sum of two geometric random variables, we parameterize the distribution of L using $\gamma = 2/\phi$ (see Appendix). Unlike the geometric case, our marginal distribution of L_j truncated between 2 and 1,500 still depends on ψ_j, so for each j, we plug in our estimated $\hat{\psi}_j$ in place of ψ_j. Then, we can again maximize with respect to ϕ:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}} \sum_{j \in I_2^{1500}} \log P(L_j = \ell_j | 2 \leq L_j \leq 1500, \phi, \psi_j = \hat{\psi}_j).$$

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

To choose between the two distributions of N, we propose calculating the Akaike Information Criterion (AIC) under each version of the model.[14] Lower AIC indicates that the distribution of N that is used is a better fit to the data.

Bootstrap confidence intervals

We calculate 95% bootstrap confidence intervals for ϕ. We denote the number of detected gene conversion tracts with observed tract length between 2 and 1,500 bp as $\left|I_2^{1500}\right|$. To obtain each bootstrap sample, we sample with replacement $\left|I_2^{1500}\right|$ observed tract lengths from the set $\{\ell_j | j = 1, \ldots, m,\ 2 \leq \ell_j \leq 1500\}$. Each bootstrap sample consists of the set of observed tract lengths $\{\ell_j\}$ and allele conversion probabilities $\{\psi_j\}$ corresponding to the resampled indices.

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

Simulation study

We use simulated data described in Browning and Browning (2024).[6] 20 regions of length 10 Mb were generated for 125,000 individuals using the coalescent simulator msprime v1.2.[15] 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 recombination rate of 1 cM/Mb and a mutation rate of 1.5×10^{-8} 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. The processes used to add uncalled deletions and genotype errors are described in Browning and Browning (2024).[6] Variants with MAF $\leq$ 0.01 were excluded, the phase information was removed, and Beagle 5.4 was used to statistically phase the genotypes.[11] The multi-individual IBD analysis detected 284,838 allele conversions belonging to 226,007 detected gene conversion tracts across the 20 regions. We fit our model to the detected gene conversion tracts in each of the 20 regions to estimate the mean gene conversion tract length in each region. For the purposes of this simulation study, we refer to the detected gene conversion tracts in each region as a separate replicate dataset. We refer to fitting our model to the detected gene conversion tracts in each of the 20 regions as a separate replicate of this simulation study.

We fit our model under two settings, one assuming a geometric distribution and the other assuming a sum of two geometric random variables for the gene conversion tract lengths N. 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 N is geometric will be favored using AIC.

UK Biobank analysis

We apply our methods to the UK Biobank whole autosome data to estimate the autosome-wide mean gene conversion tract length. In addition, we run a stratified analysis, stratifying observed tract lengths by whether the corresponding gene conversion tract overlapped with a recombination hotspot.

We use the deCODE genetic map to define recombination hotspots on each autosome.[16] For each autosome, we first calculate 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 calculate 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 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 into one hotspot. We stratify the observed tract lengths $\{\ell_j | j = 1, \ldots, m\}$ based on whether the detected gene conversion tract overlapped with a recombination hotspot. We then obtain a maximum likelihood estimate and a 95% bootstrap confidence interval for the mean gene conversion tract length, separately for tracts that overlap and do not overlap with a recombination hotspot.

Results

Simulation study

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

Based on the AIC, the geometric setting was a better fit in all 20 replicates. The difference in AIC (the AIC for the geometric setting subtracted from the AIC assuming a sum of two geometric random variables) ranged from 11 to 41 across the 20 replicates.

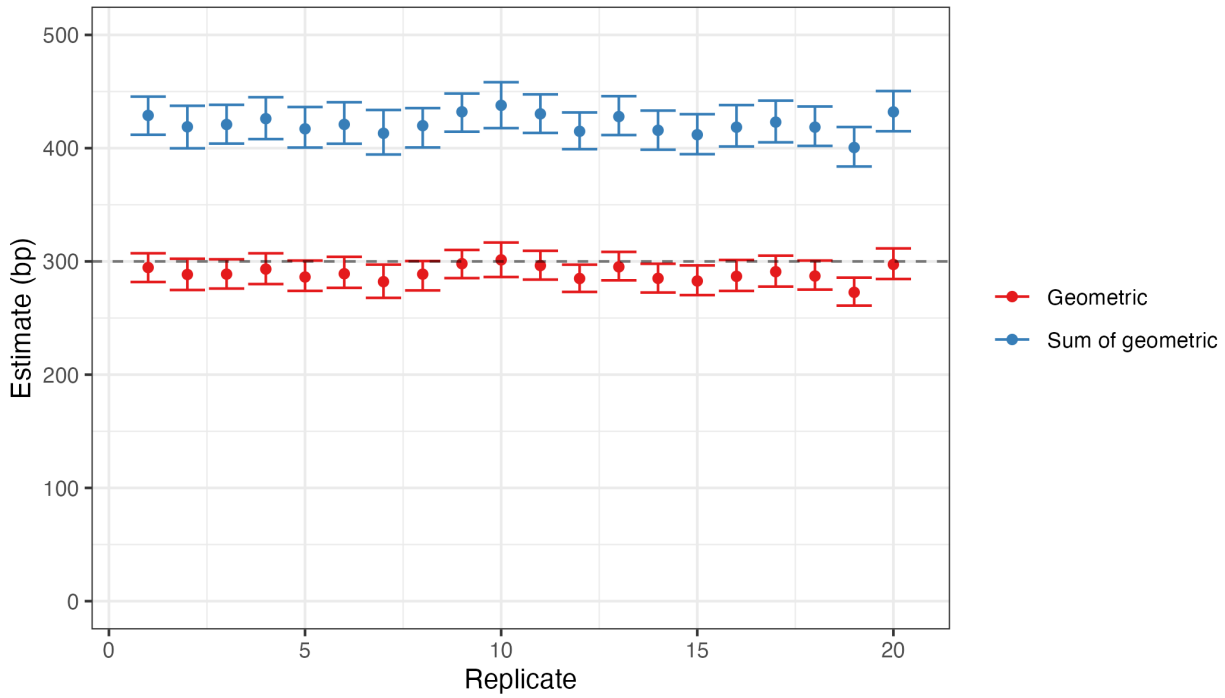

Figure 1. Estimated mean gene conversion tract lengths across replicate simulations. The dotted horizontal line represents the true mean gene conversion tract length. Gene conversion tract lengths were simulated under a geometric distribution, and analyses were conducted assuming that tract lengths are 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 simulation.

UK Biobank analysis

We applied our estimation method to the observed tract lengths detected from the UK Biobank whole autosome data. When assuming that gene conversion tract lengths are geometric, our model estimates the mean gene conversion tract length to be 459 bp (95% CI: [457, 461]). When assuming that tract lengths are drawn from a sum of two geometric random variables, our model estimates the mean gene conversion tract length to be 649 bp (95% CI: [647, 651]). The geometric setting had lower AIC, indicating a better fit to the data, and the difference in AIC between the two settings was 66,237.

We next detected recombination hotspots on all 22 autosomes. We found 32,279 recombination hotspots on all autosomes, with the longest hotspot being 51,470 bp on Chromosome 13. To illustrate how we detect recombination hotspots, we plot the recombination hotspots that we found on a region on Chromosome 21 in Figure S1.

Taking the subset of observed tract lengths in which the corresponding detected tracts overlapped with a recombination hotspot, we reran the analysis. For these observed tract lengths, we estimate the mean gene conversion tract length to be 418 bp (95% CI: [416, 420]) assuming a geometric gene conversion tract length distribution.

For the subset of observed tract lengths in which the corresponding detected tracts did not overlap with a recombination hotspot, we estimate the mean gene conversion tract length to be 492 bp (95% CI: [489, 494]) assuming a geometric gene conversion tract length distribution. In both subsets, the AIC was smaller under the geometric setting relative to the setting in which we assume that gene conversion tract lengths are drawn from a sum of two geometric random variables.

Discussion

Previous studies have tried to measure gene conversion tract lengths in humans by detecting allele conversions 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 also been limited by the availability of appropriate data. In sperm-typing studies, distinguishing genotype errors from allele conversions is also difficult. A statistical method has been proposed to infer the length distribution of gene conversion tracts in humans,[8] but the

relatively 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 were able to 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. In the largest pedigree study conducted to detect gene conversions, only around 2,000 gene conversion events were detected from a combination of 7,219 three-generation pedigrees genotyped with a SNP chip and 101 whole-genome sequenced three-generation pedigrees.[5]

We proposed a likelihood-based estimation method, inspired by a previous method by Betran et al.,[9] to infer the mean gene conversion tract length from a large number of detected gene conversion tracts. In our method, 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 distribution based on AIC.

We used a coalescent simulation incorporating gene conversion events to validate our estimation method. We found that our model accurately estimated the mean gene conversion tract 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 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 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, 461]) from this dataset. The

width of our confidence interval is much narrower than confidence intervals from previous studies, while our estimate is higher than previous estimates for humans. 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, 2,925]) respectively, based on 257 paternal and 247 maternal gene conversion tracts detected from sequenced three-generation pedigrees.[8] 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.[3] Our estimate of 459 bp is not inside this range.

It is important to acknowledge that our method omits observed tract lengths exceeding 1.5 kb, because we cannot accurately detect observed tract lengths corresponding to longer gene conversion tracts. Complex gene conversion events, which result in both allele converted and non-allele converted markers, often span more than 1.5 kb.[5]

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. Applying our model on just the detected 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]). Thus, we found a significant difference in mean tract lengths between hotspots and non-hotspots, with smaller tract lengths in hotspots. This is a preliminary finding and we caution that the difference could be attributable to unknown technical factors. We recommend further analysis to confirm this result. Recombination hotspots correlate with other genomic features such as GC rate,[17] so the difference, if real, may be caused by factors other than recombination rate.

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. 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).

4. 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).

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. 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).

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

9. 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).

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).

17. Fullerton, S. M., Bernardo Carvalho, A. & Clark, A. G. Local Rates of Recombination Are Positively Correlated with GC Content in the Human Genome. *Molecular Biology and Evolution* **18**, 1139–1142 (2001).

Appendix

Deriving the marginal distribution of $\boldsymbol{L}$ when $\boldsymbol{N}$ is a sum of two geometric random variables

We consider the case in which N is distributed as a sum of two independent and identically distributed geometric random variables each with mean $\phi/2$. We have,

$$P(N=n) = (n-1)\left(1-\frac{2}{\phi}\right)^{n-2}\left(\frac{2}{\phi}\right)^{2}.$$

Letting $\gamma = \frac{2}{\phi}$,

$$P(L=l) = \sum_{n=l}^{\infty} P(L=l|N=n)P(N=n)$$

$$= \begin{cases} \frac{\gamma^2(1-\psi)^2}{(\gamma+\psi-\gamma\psi)^2} & \text{if } l=0 \\ \frac{2\gamma^2\psi(1-\psi)}{(\gamma+\psi-\gamma\psi)^3} & \text{if } l=1 \\ \frac{\gamma^2(1-\gamma)^{l-2}\psi^2[(l-3)(\gamma+\psi-\gamma\psi)+2]}{(\gamma+\psi-\gamma\psi)^3} & \text{if } l\geq 2 \end{cases}.$$

Then,

$$P(2\leq L\leq M) = \sum_{l=2}^{M} \frac{\gamma^2(1-\gamma)^{l-2}\psi^2[(l-3)(\gamma+\psi-\gamma\psi)+2]}{(\gamma+\psi-\gamma\psi)^3}$$

$$= \frac{(\gamma+\psi-\gamma\psi)\psi^2[(3-M)\gamma(1-\gamma)^{M-1}-(1-\gamma)^{M-1}-2\gamma+1]+2\gamma\psi^2[1-(1-\gamma)^{M-1}]}{(\gamma+\psi-\gamma\psi)^3}.$$

Finally,

$$P(L=l|2\leq L\leq 1500) = \frac{P(L=l)}{P(2\leq L\leq 1500)}$$

$$= \frac{(\gamma+\psi-\gamma\psi)(l-3)\gamma^2(1-\gamma)^{l-2}+2\gamma^2(1-\gamma)^{l-2}}{(\gamma+\psi-\gamma\psi)[(3-M)\gamma(1-\gamma)^{M-1}-(1-\gamma)^{M-1}-2\gamma+1]+2\gamma[1-(1-\gamma)^{M-1}]}.$$

Notice that unlike the case where N is geometric, $P(L = l|2 \leq L \leq 1500)$ depends on ψ.

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

$$P\left(L_j = l_j \middle| 2 \leq L_j \leq 1500, \gamma, \psi_j\right)$$

$$= \frac{\left(\gamma + \psi_j - \gamma\psi_j\right)\left(l_j - 3\right)\gamma^2(1-\gamma)^{l_j-2} + 2\gamma^2(1-\gamma)^{l_j-2}}{\left(\gamma + \psi_j - \gamma\psi_j\right)[(3-M)\gamma(1-\gamma)^{M-1} - (1-\gamma)^{M-1} - 2\gamma + 1] + 2\gamma[1-(1-\gamma)^{M-1}]}.$$

Simulation study to assess the robustness of the model

We run a simulation study to assess how well our model can estimate the mean tract length ϕ 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 $\{\ell_j | j = 1, \ldots, m\}$ using four distributions for the length distribution of gene conversion tracts (Figure S2):

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

All four distributions have mean 300 bp. Recall that in the previous coalescent simulation, we generated 20 regions of length 10 Mb for 125,000 individuals using the coalescent simulator msprime v1.2.[15] In this simulation study, we generate observed tract lengths by simulating gene conversion tracts on the first

region (out of the 20 regions) from the previous coalescent simulation. To simulate one set of observed tract lengths, we first sample 100,000 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.

Markers with MAF less than 5% were not considered in step 3, similarly to how we do not detect allele conversions at these markers using the multi-individual IBD method.[6] This procedure results in 100,000 observed tract lengths, some of which may be zero bp due to the absence of heterozygous markers within the corresponding gene conversion tracts. For each of the four distributions listed earlier, we repeat this procedure 100 times to obtain 100 sets of 100,000 observed tract lengths. Then, we fit our model under both settings for N (geometric and sum of two geometric random variables), to each set of observed tract lengths (after retaining tract lengths between 2 and 1,500 bp). For each set of observed tract lengths, we obtain both a point estimate and a 95% bootstrap confidence interval for ϕ. The empirical bias and standard error of our estimates under each setting of N is shown in Table 1. Under the AIC selected setting, we use the estimate from the setting of N with the smaller AIC value in each of the 100 sets.

	Bias (SE)		
	Geometric N	Sum of geometric N	AIC selected
Geometric	-16 bp (7)	114 bp (9)	-14 bp (16)
Sum of two geometric	-102 bp (4)	-8 bp (6)	-17 bp (29)

Sum of three geometric	-133 bp (4)	-53 bp (6)	-53 bp (6)459
Uniform	-143 bp (3)	-70 bp (4)	-70 bp (4)460

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

465 We also calculated the coverage of our 95% bootstrap confidence intervals. When the gene conversion
466 tracts were simulated from a geometric distribution, and we specified N to be geometric in our model,
467 our 95% confidence intervals covered the true mean of 300 in 34 out of the 100 replicates. When the gene
468 conversion tracts were simulated from a sum of two geometric random variables, and we specified N to
469 be this distribution in our model, our 95% confidence intervals covered the true mean of 300 in 79 out of
470 the 100 replicates. When we simulated the gene conversion tract lengths from the remaining two
471 distributions, the coverage was 0% under both settings of the model.