Estimating gene conversion tract length and rate from PacBio HiFi data

Authors: Anders Poulsen Charmouh[1], Peter Porsborg Sørud[1], Thomas Bataillon[1], Asger Hobolth[2], Lasse Thorup Hansen[2], Søren Besenbacher[3], Sofia Boeg Winge[4], Kristian Almstrup[4], Mikkel Heide Schierup[1].

[1]Bioinformatics Research Centre Aarhus University, University City 81, building 1872, 3rd floor. DK-8000 Aarhus C, Denmark. [2]Department of Mathematics, Aarhus University, Aarhus, Denmark. [3]Department of Molecular Medicine (MOMA), Brendstrupgårdsvej 21A, 8200 Aarhus N, Denmark. [4]Copenhagen University Hospital, Inge Lehmanns Vej 7, 2100 Copenhagen, Denmark.

Keywords: Gene conversion, genome evolution, recombination, genomics methods, non-crossover.

Abstract

Gene conversions are broadly defined as the transfer of genetic material from a 'donor' to an 'acceptor' sequence and can happen both in meiosis and mitosis . They are a subset of non-crossover events and like crossover events, gene conversion can generate new combinations of alleles, erode linkage disequilibrium, and even counteract the mutation load by reverting germline mutations through GC-biased gene conversion. Estimating the rate of gene conversion and the distribution of gene conversion tract lengths remains challenging. Here, we present a new method for estimating tract length, rate and detection probability of non-crossover events directly in HiFi PacBio long read data. The method can be applied with data from a single individual, is unbiased even under low single nucleotide variant densities and does not necessitate any demographic or evolutionary assumptions. We apply the method to gene conversion events observed directly in Pacbio HiFI read data from a human sperm sample and find that human gene conversion tracts are shorter (mean of 50 base pairs) than estimates from yeast or *Drosophila*. We also estimate that typical human male gametes undergo on average 280 non-crossover events where approximately 7 are expected to become visible as gene conversions moving variants from one donor haplotype to an acceptor haplotype.

Introduction

Recombination between chromosomes during meiosis is initiated by a double-strand break (DSB) of the chromosome, after which a single strand from the homologous chromosome can invade the broken double strand (strand invasion) to form a heteroduplex of DNA from different chromosomes (Jasin and Rothstein 2013). Depending on how the heteroduplex is resolved, strand invasion can result in either a crossover (CO) event or a non-crossover (NCO) event (Cole et al. 2010). A crossover event results in large segments of homologous chromosomes being swapped, whereas an NCO occurs when the DSB is repaired, using the homologous chromosome as a template, through the synthesis-dependent strand annealing pathway or the double Holliday junction pathway (Chen et al. 2007; McMahill et al. 2007). A subset of NCOs are observable as gene conversions which occur when one or more SNVs are transferred non-reciprocally from one haplotype/chromosomal segment to another (Mansai et al. 2011; Lorenz and Mpaulo 2022). Gene conversions have evolutionary consequences on patterns of linkage disequilibrium (Padhukasahasram and Rannala 2011), directly by causing GC-biased gene conversion which change the GC content of the genome which serves to revert new germline mutations by opposing the AT mutation bias (Arbeithuber et al. 2015) and biasing the distribution of fitness effects towards more beneficial mutations (Joseph 2024).

NCO events are hard to detect. Firstly, an NCO event is only visible in genomic data as a new haplotype when it overlaps and converts at least one SNV from one donor haplotype to an acceptor haplotype. Because humans harbour typically an SNV density of approximately 1 in 1500 bp, NCOs are hard to detect unless the number of bases affected is on the same order which it does not seem to be (Haldorsson et al. 2016; Arndt et al. 2023). More specifically, human NCO tract length, which is the length of double-stranded DNA converted in an NCO event, has been found to be in the range 3-2086 bp with a frequently

cited estimate being 55-290 bp, making it less likely that any one NCO event will overlap and convert at least one SNV, hence becoming a gene conversion (Bosch et al. 2004; Jeffreys and May 2004; Hallast et al. 2013; Harpak et al. 2017; Hardarson et al. 2023). A related problem is that finding gene conversion events requires extensive sequencing if the gene conversion rate is low. Estimates of NCO rates in human span $5.9 \cdot 10^{-6}$ to $8.75 \cdot 10^{-6}$ NCO per base pair per generation, where this rate can be interpreted as the probability per generation that a base pair lies within an NCO tract (Williams et al. 2015; Halldorsson et al. 2016; Narasimhan et al. 2017). Furthermore, the sequence of both haplotypes needs to be known to determine which SNVs have been converted. Flanking unconverted SNVs on either side of the gene conversion event are typically necessary to distinguish a gene conversion event from a crossover event (Halldorsson et al. 2016; Hardarson et al. 2023).

The sequencing of highly accurate long reads from sperm samples offers a new way to identify gene conversion events. Specifically, if enough gametes from a single individual are sequenced, both parental haplotypes are known such that gene conversions between two haplotypes at heterozygous sites can be directly observed (Porsborg et al. 2024, BIORXIV/2024/601967). Furthermore, flanking SNVs on either side of the gene conversion event will typically be present such that the gene conversion event can be called, because HiFI PacBio long read sequencing results in very long reads, usually with a mean read length above 10kb (Hon et al. 2020; Marx 2023) that we expect greatly exceeds the expected size of conversion tracts. Here, we build a model suitable for estimating NCO rate, tract length and detection probability directly from HiFi PacBio long read data (Figure 1).

The Model

We count the number of gene conversion events where a single SNV has been converted and compare it to the number of gene conversion events where multiple SNVs have been converted. The ratio of these two counts provides information about the distributions of gene conversion tract lengths. This can be understood intuitively by realizing that if single SNV conversions are far more common than multi-SNV conversion events, this indicates that tract length are typically quite short since it is unlikely for a short tract to convert multiple SNVs. This verbal expectation can be formalized as the expected probability of observing a single SNV conversion relative to the probability of observing a multi-SNV conversion.

We first derive this expected ratio single to multi SNV conversions in the idealized case and then subsequently show that using the relative occurrence of single to multi-SNV conversions to estimate gene conversion tract length, rate and detection probability can be generalized beyond a uniform SNV distribution to any scenario wherein SNVs are clustered, as they are in real genomes. We do this by using simulations to contain the conversion probabilities and then use the relative numbers of single and multi-SNV conversions in a maximum likelihood framework to obtain maximum likelihood estimates for the mean tract length, rate and detection probability of gene conversion.

The model – idealized case of uniform SNV distribution

Consider a gene conversion event with a tract length L base pairs (bp). For each bp, we assume that the probability of the gene tract terminating is s such that the probability of the gene conversion tract extending is $1 - s$. Given this, the distribution of gene conversion tract lengths L follows a geometric distribution with parameter s. For any base pair, the probability of observing a SNV is p and the probability of not observing a SNV is $1 - p$. We initially

assume that all SNVs overlapping the conversions event will be converted (this assumption can be relaxed, see supporting Information S1).

We ask, given a gene conversion tract of length *m*, what is the probability, *P(#SNV=1 | L*=m*)* , of converting a single SNV? In the idealized case where the SNV distribution is uniform (all positions in the tract have a probability *p* of containing an SNV), this probability is given by the binomial distribution,

$$P(\#SNP = 1 \mid L = \mathrm{m}) = \binom{m}{m-1} p(1-p)^{m-1} = mp(1-p)^{m-1}.$$

The joint probability of $L = m$ and $\#SNP = 1$ is then

$$P(\#SNP = 1, L = \mathrm{m}) = m(1-p)^{m-1} p(1-s)^{m-1} s \qquad (1)$$

because tracts follow a geometric distribution, so we scale the probability of converting a single tract by the probability of sampling a tract of length *m* from a geometric distribution with mean $1/s$.

Similarly, the joint probability of converting more than one SNVs given a tract of length *m*, *P(#SNV > 1, L*=m*)*, is

$$P(\#SNP > 1, L = \mathrm{m}) = (1-s)^{m-1} s(1-(1-p)^{m} - m(1-p)^{m-1} p). \qquad (2)$$

We now calculate P(#SNV=1) and P(#SNV > 2) by summing the probability over all possible values of *m*. This yields

$$P(\#SNP = 1) = \sum_{m=1}^{\infty} P(\#SNP = 1, L = m) = \sum_{m=1}^{\infty} mp(1-p)^{m-1}(1-s)^{m-1} s \qquad (3)$$

and

$$P(\#SNP > 1\,) = \sum_{m=2}^{\infty} P(\#SNP > 1, L = m)$$

$$= \sum_{m=2}^{\infty} (1-s)^{m-1} s(1-(1-p)^m - m(1-p)^{m-1}p). \qquad (4)$$

Notice, that the sum in eq. (4) begins at *m=2*, because a tract length of at least 2 bp is necessary before the conversion of more than one SNV is possible.

The ratio of single to multi-SNV gene conversion events which we would expect to observe in the data, is now given as

$$R(p,s) = \frac{P(\#SNP = 1\,)}{P(\#SNP > 1\,)} = \frac{\sum_{m=1}^{\infty} mp(1-p)^{m-1}(1-s)^{m-1}s}{\sum_{m=2}^{\infty}(1-s)^{m-1}s(1-(1-p)^m - m(1-p)^{m-1}p)}\,. \qquad (5)$$

Notice, that since the numerator and denominator of *R(p,s)* is the probability of converting one or more SNVs respectively, we can also calculate the proportion of "silent" NCO events, $S(p,s)$, which we define as the proportion of NCOs converting no SNVs (i.e. the probability that an NCO does not become a gene conversions),

$$S(p,s) = 1 - \sum_{m=1}^{\infty} mp(1-p)^{m-1}(1-s)^{m-1}s$$

$$- \sum_{m=2}^{\infty} (1-s)^{m-1}s(1-(1-p)^m - m(1-p)^{m-1}p). \qquad (6)$$

Since we know the proportion of NCOs which fail to convert a SNV relative to the proportion of NCOs resulting in observable gene conversions, we can now write an expression for the total number of expected NCOs in the whole genome D_E, based on the observed number of gene conversions, D_o

$$D_E = \frac{S(p,s)}{1-S(p,s)} D_o + D_o. \qquad (7)$$

The total NCO rate is then simply D_E divided by the number of base pairs in the genome of interest, multiplied by the mean number of base pairs affects by a NCO event.

Estimating SNV conversion probabilities using the empirical SNV distribution

The distribution of inter-SNV distances in a recombining genome is not uniform. This is because different regions, have very different coalescent histories. Times to the most recent common ancestor vary and accordingly modulate the SNV density, ranging from regions harbouring runs of homozygosity (recent coalescent event) to regions with high density of SNVs (deep coalescent events) (Wiuf 2000b; Arndt et al. 2023). Additionally, variation in the mutation rate across the genome further contributes to a non-uniform SNV distribution (Oman et al. 2022). To estimate the probability of the different conversion events, we stimulated a genome with the exact same inter-SNV distances as observed in the sample genome, and calculated the proportion of different conversion events as a function of different values of *s*.

Initialization and parameters

To obtain conversions probability for our estimates of mean tract length, NCO rate and NCO detection probability, we simulated genomes with the same length and the same position of SNVs as the sampled genomes. In each simulated genome, NCOs occurred at 10^5 positions chosen from a uniform distribution $U(1, M)$ where M is the genome size. The tract length of the gene conversion event was sampled from a geometric distribution with mean *s*. The proportion of NCOs converting one, two, *n* SNV under each value of *s* (hereafter denoted $P_1(s), P_2(s), \ldots, P_n(s)$) was calculated. All classes up to *n=11* were calculated for each simulation. Simulation sets with *s* in the range of [5, 200] were simulated. For each value of *s*, we simulated the entire genome 10^3 times (i.e 10^3 simulation replicates) and mean values of $P_1(s), P_2(s), \ldots, P_n(s)$ were calculated for each *s*.

Simulations were conducted on a slurm cluster (slurm 23.02.5) and the full source code is available at [https://github.com/r02ap19/GeneConv] along with a ReadMe file detailing how to change model parameters as well as how to compile and run the code.

Generalizing to an arbitrary SNV distribution

Using a maximum likelihood framework, the idea that the relative numbers of single and multi-SNV conversion contains information about the tract length, rate and detection probability of gene conversion can be generalized to the case where SNVs are not uniformly distributed. From the HiFi PacBio sequence data, we called a set of gene conversion events (table 1). Out of these k gene conversion events, a certain number of events $d_1, d_2, \dots, d_n$ resulted in the conversion of 1, 2,…, n SNVs, respectively (Fig 1A). Since the distribution of SNVs in the genome of the sampled individual is known, the probabilities of converting 1, 2,…,n SNVs contingent on the SNV distribution and given a mean tract length of 1/s $(P_1(s), P_2(s), \dots, P_n(s))$ can be found (Figure 1B). By definition, the likelihood $L(s)$ of each s given the data $d_1, d_2, \dots, d_k$ can now be calculated as

$$L(s|d_1,\ d_2,\ \dots,d_n) = \binom{k}{d_1, d_2, \dots, d_n} \prod_{i=1}^{n} p_i(s)^{d_i}. \qquad (8)$$

Finding the value s^* which maximizes $L(s)$ yields a maximum likelihood estimate (MLE) of s given the data $d_1, d_2, \dots, d_n$ (Figure 1C).

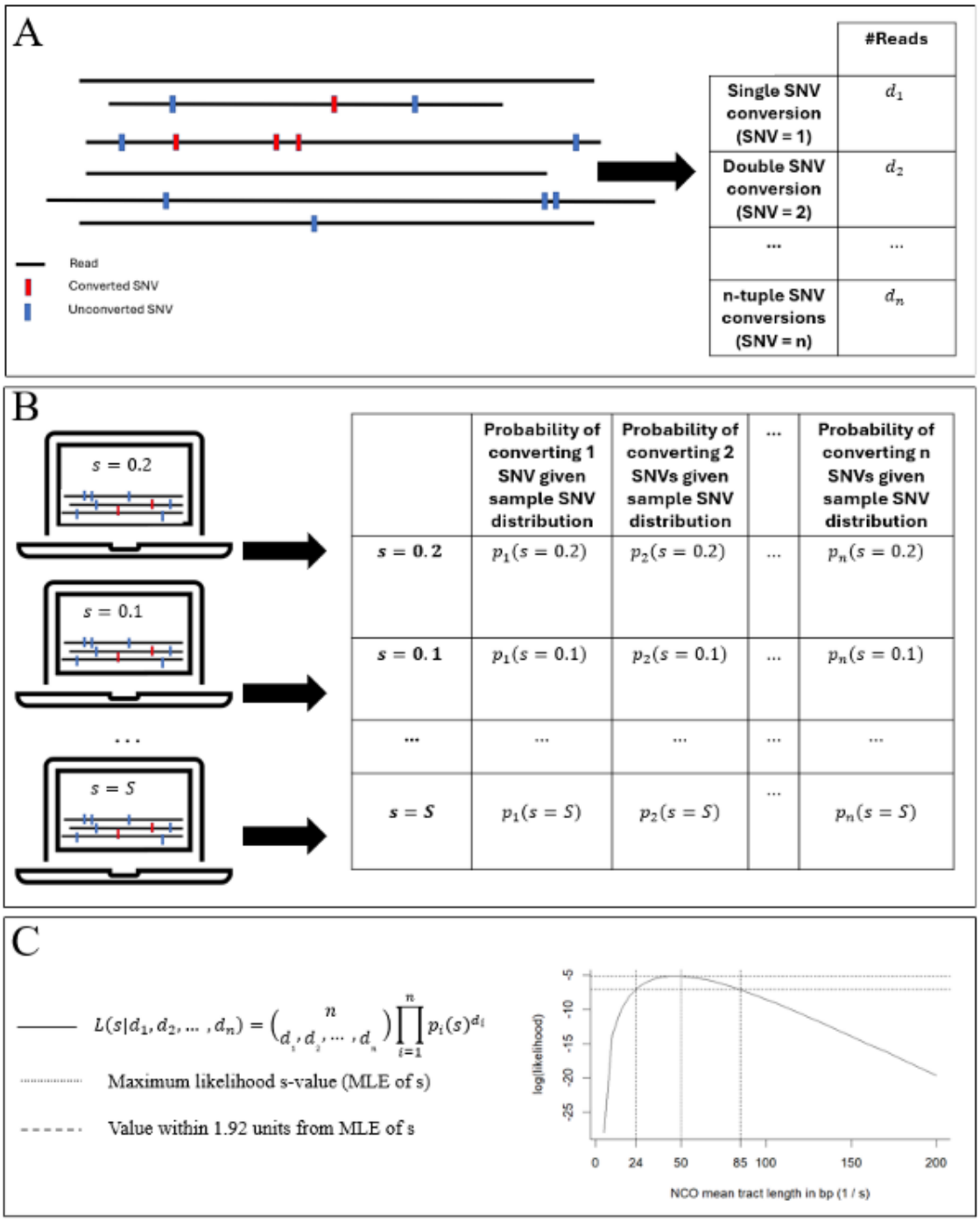

Figure 1: A method for estimating NCO tract length, rate and detection probability. A) HiFi long reads are screened for gene conversion and read data is summarized via the counts of the number of reads where 1, 2,…, n SNVs are converted (conversion counts) are made. B) Using the SNV distribution along the sequenced sample genome, simulations are conducted with varying mean tract length and the probability of converting 1, 2,…,n SNVs (terms in eq. (5)) are estimated contingent on the SNV density and tract length distribution (conversion probabilities). C) Using the conversion probabilities which take the non-uniform SNV

distribution into account, the mean tract length which maximizes the likelihood of the data (the conversion counts) is estimated using eq.(8).

Calling gene conversions in HiFi PacBio data

We use a pipeline for filtering of reads, calling of gene conversion candidate reads, quality checks on reads and manual curation of candidate gene conversion reads which has been tested and described in details in another study (Porsborg et al. 2024, BIORXIV/2024/601967).

Results

We find an analytical model which can be used to infer the mean gene conversion tract length, the variance of gene conversion tract lengths, the NCO rate and the NCO detection probability based on the SNV density and the observed ratio of single to multi-SNV conversions (figure 2). When the SNV distribution is uniform, we find that this method allows for unbiased inference of gene conversion tract length, rate, and detection probability even when using data from populations with very low SNV density. This makes it suitable for inferring the mean NCO tract length for species with low levels of genetic diversity, such as humans (figure 2).

We tested the accuracy of the ML method across different mean tract lengths as well as the accuracy under varying amounts of data and found this approach to result in unbiased inference too, also in cases where less data is available than we have used here (Supporting Information S2).

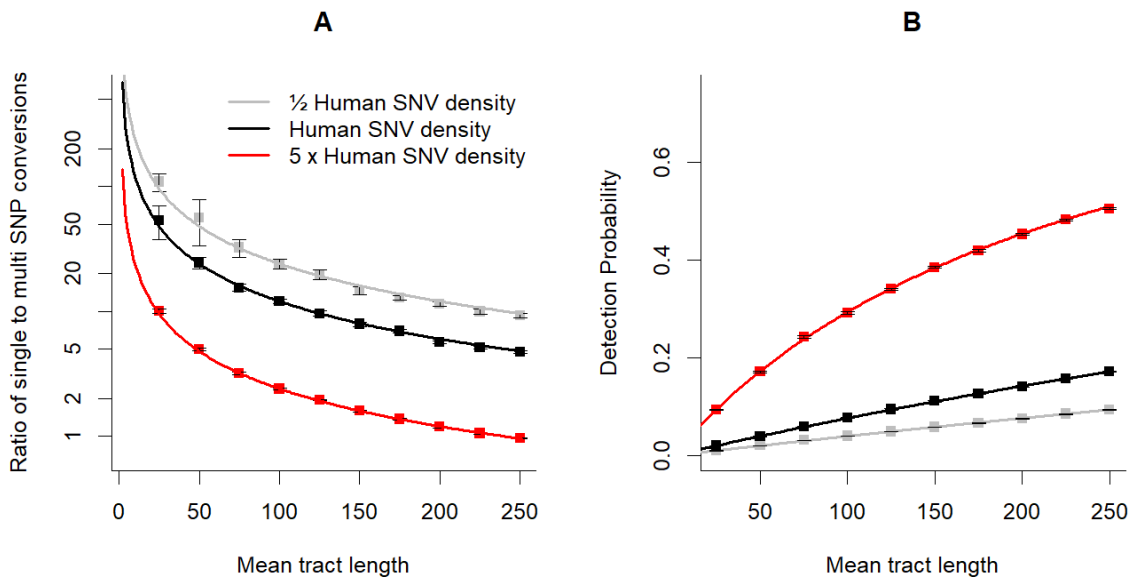

Figure 2: Analytical predictions of the ratio model (full lines) compared to simulation results (points) with 95% confidence intervals. A genomic fragment of 30Mbp was simulated assuming different SNV densities, using approximately half the human SNV density (SNVs/bp), human SNV density (0.00083 SNVs/bp) and five times the human SNV density (SNVs/bp). The dashed line represents the true simulated mean tract length of 200 bp (i.e.). The results show that in the idealized case where all positions in the genome have some probability of being an SNV, using the ratio of single to multi-SNV conversions can yield unbiased estimates despite very low SNV densities, such as those observed in populations of humans.

Under the likelihood model presented, we estimate the mean tract length in humans based on gene conversion events called directly from HiFi PacBio long read data from a human sperm sample. In the HiFi long reads obtained from a human sperm sample [reference our biology/non-method study], we observed a total of 182 gene conversions with between one a four SNVs converted per event (Table 1). We estimate the mean tract length to be 50 bp ([CI 95%: 24, 85], figure 3A).

Table 1: Conversion counts obtained from a human sperm sample by calling gene conversions directly in HiFI long read data [ref to main manuscript]. We observed 167 single SNV conversions, 12 double SNV conversions, 2 triple SNV conversion and 1 quadruple SNV conversion.

	1-SNV	**2-SNV**	**3-SNV**	**4-SNV**

	conversion	conversion	conversion	conversion
#reads	167	12	2	1

We further find that given the MLE, the conversion counts (i.e. 167, 12, 2, 1) do not differ significantly (*p=0.73*) from the expected conversion counts when tracts are sampled from a single geometric distribution with the MLE s (see supporting information S3).

The model also yields MLEs of the proportion of detected (and undetected) NCO events, that is, the proportion of NCO events which can be observed in data because they convert at least one SNV. We find that most NCOs are 'invisible' because they happen to only convert identical sequences (figure 3B). We also find MLEs for the total NCO rate, which implies that (using eq.(7) and scaling by genome size and coverage), each male human gamete is expected to harbour on average 280 NCO events, of which approximately 7 are expected to become visible as gene conversions.

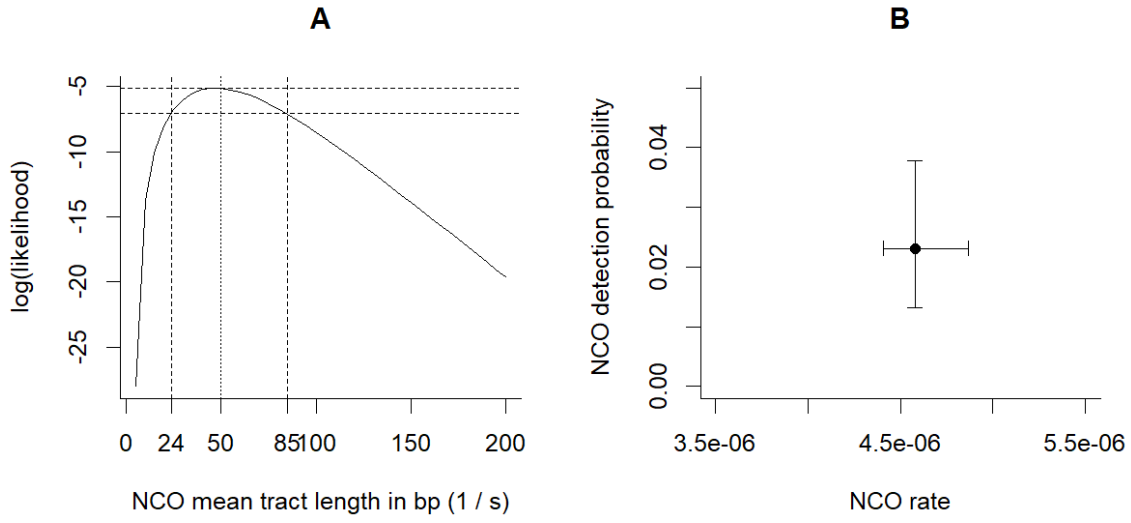

Figure 3: MLEs of mean gene conversion tract lengths, rate and detection probability based on gene conversion events called directly from HiFi PacBio data of sperm sample from an approximately 25-year-old human. A) Likelihood profile for mean tract length. Each point

shows the log likelihood of the data (counts of single, double, triple,…,n-tuple gene conversion events for the sampled individual) conditional on the SNV distribution and density of the individual, i.e. eq. (8). Dotted vertical line represents the MLE and the vertical dashed lines show the 95% confidence interval. The results suggest that human gene conversion tracts are typically quite short (mean tract length of 50 bp). B). MLE of NCO detections probability (probability that an NCO becomes a gene conversion) and total NCO rate (including gene conversions) Bars denote the 95% confidence interval. The results indicate that most NCO events fail to convert at least one SNV meaning that these are not observable as gene conversions. Each MLE is based on 10^3 simulations using the SNV distribution along the genome in the sample data.

Ascertainment bias and the effect of finite read lengths

We note that our pipeline for calling gene conversion events [reference the other paper] requires that unconverted SNVs are present at the end of each read, such that it is possible to ascertain whether a read represents a potential crossover or gene conversion event. This leaves open the possibility of a slight ascertainment bias because some gene conversion events (e.g. those converting more than one SNV) might be indistinguishable from crossovers when occurring at the edge of a read. It turns out this ascertainment bias is very weak, since the average read length of HiFi PacBio 16.36 kb greatly exceeds a typical NCO tract length. We checked that simulating under the same read distribution as the sample (rather than assuming much longer reads) yield an almost identical likelihood profile and hence almost the same MLEs (see Supporting Information S4).

Discussion

We developed a method for estimating the rate and length distribution of NCO events and show that the average gene conversion tract in humans appears to be much shorter than in yeast and Drosophila (Mansai et al. 2011). We showed that this method is unbiased even for organisms with low SNV densities, such as humans and that sufficient input data for the method can be obtained from sequencing a single individual at approximately 7x coverage and then calling gene conversion events directly from HiFi PacBio sequencing data.

While we use HiFi PacBio data here, inference of NCO rate, tract length and detection probability could easily be done with a different type of data and different organisms – all that is required is the number of gene conversion events counts (i.e. counts of NCO events resulting in single, double, triple, etc. SNVs conversions) and the background SNV distribution of the sample where NCOs were called. Because of this, our method could also be used to obtain estimates by using gene conversions called in trio data (Halldorsson et al. 2016), in hybrid cross experiments (Li et al. 2019) and in pollen or sperm-typing (Lien et al. 2000; Jeffreys and May 2004; Tiemann-Boege et al. 2006). However, the comparison of estimates obtained from very different types of data is not necessarily straightforward since different ways of obtaining data can lead to different ascertainment biases. For example, trio data makes it in principle possible to observe NCOs inducing gene conversion tracts of an arbitrary length whereas HiFi PacBio long read data is constrained by the finite read length. On the other hand, when all individuals within a trio are heterozygous at a specific position, phasing relies on algorithms which cannot phase all positions (Miller and Piccolo 2021). A more detailed investigation of the extent to which different types of datasets might bias the estimates of NCO rate and tract lengths upwards or downwards would be a worthwhile aim for future work.

We exploit the fact that multi SNV conversion becomes more common as the tract length increases, which is a finding that has also been made in early simulation studies (Gay et al. 2007). More recently, Li et al. (2019) proposed a method which assumes an exponential distribution of gene conversion tracts, and uses information about the co-conversion of SNVs. Rather than using the ratio of single to multi-marker conversion events, the method of Li et al. (2019) creates a composite likelihood function by taking all consecutive pairs of SNVs adjacent to or within gene conversion tracts and considering whether these were part of the same conversion tract or not, as a function of the distance between the SNVs. While the

method also takes the distance between potentially co-converted SNVs into account, it results in a composite likelihood (pseudolikelihood) because SNVs within the same conversion tract are assumed to be independent, which is not the case. Considering the ratio of single to all multi SNV events (as in our model) has the further advantage that it grants analytical solutions for the probability of an unobserved NCO (i.e. 'invisible' gene conversion) and the total NCO rate in the idealized case. Combining these methods, such that the distances between converted markers are also considered, could result in more precise predictions. Exploring this could be fertile grounds for future work.

Gene conversion tract length and the geometric distribution

Our method assumes that tract length follows a geometric distribution. However, we consider this assumption to be reasonable since several studies have found the geometric distribution (or the continuous version of the distribution, the exponential distribution) to fit well with data (Hilliker et al. 1994; Taghian and Nickoloff 1997; Li et al. 2019) and this was also the case for our study when considering the conversion counts (see supporting results S3). This makes sense, because the transfer of SNVs from one sequence to another via gene conversion is, at the core, the result of a polymerase moving along a sequence, which is being repaired using some donor sequence as template (Jasin and Rothstein 2013). If the polymerase has a certain probability of stopping at each base, and the remaining probability of extending the tract at each base, this will result in a geometric distribution of tract lengths (Wiuf 2000a; Frisse et al. 2001; Padhukasahasram and Rannala 2011; Setter et al. 2022). A recent study found that a mixture of two negative binomial distributions fitted datasets on gene conversion tract lengths better than a geometric distribution (Hardarson et al. 2023).

In some studies, a small subset of the observed gene conversion tracts seems to be incompatible with viewing all tracts as coming from a single geometric distribution with a

low mean. Several recent studies have pointed out that while the majority of gene conversion tracts appear to fit a geometric distribution, the distribution of gene conversion tracts as a whole does not, since a very small fraction of tracts are much larger than expected under a single geometric distribution (Halldorsson et al. 2016b; Wall et al. 2022; Versoza et al. 2023). In a study of olive baboons (*Papio anubis*), Wall et al. (2022) found >99% of all gene conversion tracts to be short whereas a few tracts were very long (mean length 47.58.6kb). Similar results have been found in a study of rhesus macaques (*Macaca mulatta*) where two very long tracts were observed, the mean length of which were 5.166 kb (Versoza et al. 2023). In human data, Halldorsson et al. (2016) also reported a minority of very long NCO tracts, the mode of which was 30 kb.

Since the distribution of gene conversion tract lengths appears to be bimodal, this suggests (as proposed by Hardorsson et al. 2023) that two distinct processes might give rise to gene conversions – one which is responsible for nearly all tract and results in short tracts, and one which is very rate, but results in long tracts. When using HiFi PacBio long read data, these rare long NCO events can in principle be detected if the distribution of read lengths overlaps with the distribution of long NCO event lengths (see Supporting Information S5).

Acknowledgements

We would like to thank Regev Schweiger and Richard Durbin for their insightful comments.

Author contributions:

MHS, TB and APC conceived the study. APC wrote the manuscript with substantial input from MHS, TB, AH, SB and LTH. APC developed the model with substantial input from AH, TB and MHS. APC wrote the simulation software, conducted simulations and analysis of simulation output. PSP called SNVs and GC events and curated the HiFi long-read data. KA and SBW were responsible for all lab-related handling and processing of the sperm sample.

Competing interests statement

None

Corresponding author: Anders Poulsen Charmouh: apc@birc.au.dk

Funding

This work was supported by the Novo Nordisk Foundation (NNF18OC0031004 to MHS) and the Independent Research Fund Denmark, Natural Sciences (6108-00385 to MHS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement

All simulations software is available at [https://github.com/r02ap19/GeneConv] along with a readme file describing how to compile and run the code. Gene conversion counts were obtained from Porsborg et al. 2024 (BIORXIV/2024/601967). The repository also contains the input data file.

References

Arbeithuber, B., A. J. Betancourt, T. Ebner, and I. Tiemann-Boege. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc Natl Acad Sci U S A 112:2109–2114. National Academy of Sciences.

Arendt, J. D. 2015. Effects of dispersal plasticity on population divergence and speciation. Heredity (Edinb) 115:306–311. Nature Publishing Group.

Arndt, P. F., F. Massip, and M. Sheinman. 2023a. An analytical derivation of the distribution of distances between heterozygous sites in diploid species to efficiently infer demographic history. , doi: 10.1101/2023.09.20.558510.

Arndt, P. F., F. Massip, and M. Sheinman. 2023b. An analytical derivation of the distribution of distances between heterozygous sites in diploid species to efficiently infer demographic history. , doi: 10.1101/2023.09.20.558510.

Bosch, E., M. E. Hurles, A. Navarro, and M. A. Jobling. 2004. Dynamics of a Human Interparalog Gene Conversion Hotspot. Genome Res 14:835–844.

Chen, J. M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos. 2007. Gene conversion: Mechanisms, evolution and human disease.

Cole, F., S. Keeney, and M. Jasin. 2010. Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis. Mol Cell 39:700–710.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack, and A. Di Rienzo. 2001. Gene Conversion and Different Population Histories May Explain the Contrast between Polymorphism and Linkage Disequilibrium Levels.

Gay, J., S. Myers, and G. McVean. 2007. Estimating meiotic gene conversion rates from population genetic data. Genetics 177:881–894.

Hallast, P., P. Balaresque, G. R. Bowden, S. Ballereau, and M. A. Jobling. 2013. Recombination Dynamics of a Human Y-Chromosomal Palindrome: Rapid GC-Biased Gene Conversion, Multi-kilobase Conversion Tracts, and Rare Inversions. PLoS Genet 9. Public Library of Science.

Halldorsson, B. V., M. T. Hardarson, B. Kehr, U. Styrkarsdottir, A. Gylfason, G. Thorleifsson, F. Zink, A. Jonasdottir, A. Jonasdottir, P. Sulem, G. Masson, U. Thorsteinsdottir, A. Helgason, A. Kong, D. F. Gudbjartsson, and K. Stefansson. 2016a. The rate of meiotic gene conversion varies by sex and age. Nat Genet 48:1377–1384. Nature Publishing Group.

Halldorsson, B. V, M. T. Hardarson, B. Kehr, U. Styrkarsdottir, A. Gylfason, G. Thorleifsson, F. Zink, A. Jonasdottir, A. Jonasdottir, P. Sulem, G. Masson, U. Thorsteinsdottir, A. Helgason, A. Kong, D. F. Gudbjartsson, and K. Stefansson. 2016b. The rate of meiotic gene conversion varies by sex and age. Nat Genet 48:1377–1384.

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

Harpak, A., X. Lan, Z. Gao, and J. K. Pritchard. 2017. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. Proc Natl Acad Sci U S A 114:12779–12784. National Academy of Sciences.

Hilliker, A. J., G. Harauz, A. G. Reaume, M. Gray, S. H. Clark, A. Chovnick, and W. R. Engels. 1994. Meiotic gene conversion tract length distribution within the rosy locus of Drosophila melanogaster. Genetics 137:1019–1026.

Hon, T., K. Mars, G. Young, Y. C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, D. Kudrna, M. A. Hardigan, C. C. Steiner, S. J. Knapp, D. Ware, B. Shapiro, P. Peluso, and D. R. Rank. 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data 7. Nature Research.

Innan, H. 2003. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. Proceedings of the National Academy of Sciences 100:8793–8798.

Jasin, M., and R. Rothstein. 2013. Repair of strand breaks by homologous recombination. Cold Spring Harb Perspect Biol 5. Cold Spring Harbor Laboratory Press.

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

Joseph, J. 2024. Increased Positive Selection in Highly Recombining Genes Does not Necessarily Reflect an Evolutionary Advantage of Recombination, *Molecular Biology and Evolution*, Volume 41, Issue 6, June 2024, msae107, https://doi.org/10.1093/molbev/msae107

Li, R., E. Bitoun, N. Altemose, R. W. Davies, B. Davies, and S. R. Myers. 2019. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nat Commun 10. Nature Publishing Group.

Lien, S., J. Szyda, B. Schechinger, G. Rappold, and N. Arnheim. 2000. Evidence for Heterogeneity in Recombination in the Human Pseudoautosomal Region: High Resolution Analysis by Sperm Typing and Radiation-Hybrid Mapping. The American Journal of Human Genetics 66:557–566.

Lorenz, A., and S. J. Mpaulo. 2022. Gene conversion: a non-Mendelian process integral to meiotic recombination. Heredity (Edinb) 129:56–63. Springer Nature.

Mansai, S. P., T. Kado, and H. Innan. 2011. The rate and tract length of gene conversion between duplicated genes. Genes (Basel) 2:313–331.

Marx, V. 2023. Method of the year: long-read sequencing. Nature Research.

McMahill, M. S., C. W. Sham, and D. K. Bishop. 2007. Synthesis-Dependent Strand Annealing in Meiosis. PLoS Biol 5:e299.

Miller, D. B., and S. R. Piccolo. 2021. trioPhaser: using Mendelian inheritance logic to improve genomic phasing of trios. BMC Bioinformatics 22. BioMed Central Ltd.

Narasimhan, V. M., R. Rahbari, A. Scally, A. Wuster, D. Mason, Y. Xue, J. Wright, R. C. Trembath, E. R. Maher, D. A. Van Heel, A. Auton, M. E. Hurles, C. Tyler-Smith, and R. Durbin. 2017. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. Nat Commun 8. Nature Publishing Group.

Padhukasahasram, B., and B. Rannala. 2011. Bayesian population genomic inference of crossing over and gene conversion. Genetics 189:607–619.

Oman, M. Alam, A. Ness, R.W. 2022. How Sequence Context-Dependent Mutability Drives Mutation Rate Variation in the Genome. Genome Biology and Evolution 14,3.

Setter, D., S. Ebdon, B. Jackson, and K. Lohse. 2022. Estimating the rates of crossover and gene conversion from individual genomes. Genetics 222. NLM (Medline).

Taghian, D. G., and J. A. Nickoloff. 1997. Chromosomal Double-Strand Breaks Induce Gene Conversion at High Frequency in Mammalian Cells. Mol Cell Biol 17:6386–6393. Informa UK Limited.

Takahashi, K. K., and H. Innan. 2022. Frequent somatic gene conversion as a mechanism for loss of heterozygosity in tumor suppressor genes. Genome Res 32:1017–1025.

Tiemann-Boege, I., P. Calabrese, D. M. Cochran, R. Sokol, and N. Arnheim. 2006. High-Resolution Recombination Patterns in a Region of Human Chromosome 21 Measured by Sperm Typing. PLoS Genet 2:e70.

Versoza, C. J., S. Weiss, R. Johal, B. La Rosa, J. D. Jensen, and S. P. Pfeifer. 2023. Novel insights into the landscape of crossover and non-crossover events in rhesus macaques Macaca mulatta. Genome Biol Evol, doi: 10.1093/gbe/evad223.

Wall, J. D., J. A. Robinson, and L. A. Cox. 2022. High-Resolution Estimates of Crossover and Noncrossover Recombination from a Captive Baboon Colony. Genome Biol Evol 14.

Williams, A. L., G. Genovese, T. Dyer, N. Altemose, K. Truax, G. Jun, N. Patterson, S. R. Myers, J. E. Curran, R. Duggirala, J. Blangero, D. Reich, and M. Przeworski. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. Elife 4.

Wiuf, C. 2000a. A coalescence approach to gene conversion. Theor Popul Biol 57:357–367.

Wiuf, C. 2000b. A Coalescence Approach to Gene Conversion. Theor Popul Biol 57:3577367.

Wiuf, C., and J. Hein. 2000. The Coalescent With Gene Conversion.

Zhang, J., A. Lindroos, S. Ollila, A. Russell, G. Marra, H. Mueller, P. Peltomaki, M. Plasilova, and K. Heinimann. 2006. Gene conversion is a frequent mechanism of inactivation of the wild-type allele in cancers from MLH1/MSH2 deletion carriers. Cancer Res 66:659–664.

Zhao, Z., Y.-X. Fu, D. Hewett-Emmett, and E. Boerwinkle. 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene 312:207–213.