Estimating the rates of crossover and gene conversion from individual genomes

Derek Setter^{*,1,2}, Sam Ebdon^{*,1}, Ben Jackson^{*}, and Konrad Lohse^{*}

^{*}Institute of Evolutionary Biology, University of Edinburgh,
Edinburgh, EH9 3FL, UK

November 9, 2021

¹ These authors contributed equally to this work

² correspondce to derek.setter@ed.ac.uk

Abstract

Recombination can occur either as a result of crossover or gene conversion events. Population genetic methods for inferring the rate of recombination from patterns of linkage disequilibrium generally assume a simple model of recombination that only involves crossover events and ignore gene conversion. However, distinguishing the two processes is not only necessary for a complete description of recombination, but also essential for understanding the evolutionary consequences of inversions and other genomic partitions in which crossover (but not gene conversion) is reduced. We present **heRho**, a simple composite likelihood scheme for co-estimating the rate of crossover and gene conversion from

individual diploid genomes. The method is based on analytic results for the distance-dependent probability of heterozygous and homozygous states at two loci. We apply **heRho** to simulations and data from the house mouse *Mus musculus castaneus*, a well studied model. Our analyses show i) that the rates of crossover and gene conversion can be accurately co-estimated at the level of individual chromosomes and ii) that previous estimates of the population scaled rate of recombination $\rho = 4N_e r$ under a pure crossover model are likely biased.

Introduction

Genetic recombination, the exchange of genetic material between homologous chromosomes during meiosis, is one of the most fundamental evolutionary processes. By creating novel combinations of alleles, recombination increases the efficacy of positive selection (Hill and Robertson, 1966) and reduces the fitness burden of deleterious variants (Charlesworth and Charlesworth, 1997). Recombination breaks down linkage disequilibrium (LD) in the genome and so determines the physical scale over which selective events interfere with each other and affect linked neutral sites (Charlesworth et al., 1993; Simonsen et al., 1996). Since recombination modulates virtually all evolutionary processes, understanding how and why it varies between organisms and between different regions of the genome remains a topic of intense research (see Stapley et al., 2017; Peñalba and Wolf, 2020, for recent reviews). Beyond interest in recombination rate variation *per se*, estimates of recombination are also relevant for other inferences from genomic data. In particular, the power of quantitative or population genetic analyses depends crucially on recombination. Thus, while association studies or inference about past selection (e.g. DeGiorgio et al., 2016, 2014; Setter et al., 2020; Campos and Charlesworth, 2019) and demography (Gutenkunst et al., 2009) often treat single nucleotide polymorphisms (SNP) as

independent for the purpose of obtaining point estimates, they rely on parametric bootstrapping or resampling procedures that are conditioned on a model of recombination to quantify uncertainty.

Recombination occurs via double-strand breaks which are either Holliday-junction mediated, resulting in crossovers (CO) and crossover gene conversion (GC) events, or synthesis-dependent strand-annealing mediated, resulting in non-crossover GC events (Resnick, 1976; Szostak et al., 1983; Nassif et al., 1994). In a CO event, two non-sister chromatids break during pairing and reciprocally exchange sequence regions on either side of the break point (Griffiths et al., 2002). In contrast, GC, which typically occurs due to mismatch errors during replication (Carpenter, 1982), involves the non-reciprocal copying of a short stretch of sequence, the GC tract (typically tens to hundreds of bases), from one non-sister chromatid to the other (Szostak et al., 1983; McMahon et al., 2007). The ratio of GC to CO rates varies widely across the tree of life: estimates range from 4-15x in humans (Jeffreys and May, 2004) and mice (Li et al., 2019a) to 1/2 - 1/10x in yeast, algae, and plants (Liu et al., 2018). Similarly, estimates of GC tract lengths range from ten to several thousand base pairs between taxa (Mansai et al., 2011; Casola et al., 2010). Furthermore, the ratio of CO and GC may also vary drastically along the genome. In particular, CO may be severely reduced in centromeric and telomeric regions and within chromosomal inversion, while rates of GC may be unchanged (Talbert and Henikoff, 2010; Korunes and Noor, 2017) or even increased (Crown et al., 2018). However, given that joint estimates for the rates of GC and CO within genomes and across taxa are sparse, the evolutionary causes and consequences of variation in these two components of recombination remain poorly understood.

Much effort has been devoted to estimating CO and GC directly from lab crosses (Hilliker et al., 1994), pedigrees (Kong et al., 2002; Johnston et al., 2016),

or sperm-typing (Jeffreys and May, 2004) data. However, such direct estimates are time consuming and expensive given that data from many meiotic events are required. While some pedigree-based (Smeds et al., 2016) and sperm typing methods distinguish CO and GC events, most direct estimates of recombination are necessarily limited to CO events (Kong et al., 2002, 2014; Ma et al., 2015; Johnston et al., 2016). Since individual GC tracts are undetectable unless they span variants, the resolution to detect GC events is inherently limited by the scale of SNP variation.

Given the limitations of direct approaches for estimating recombination, methods that infer recombination indirectly from patterns of LD in whole genome resequence data from natural populations are attractive. LD-based estimators of recombination implemented in popular tools such as *LDhat* (McVean et al., 2002; Auton and McVean, 2007) and *LDhelmet* (Chan et al., 2012) are based on analytic expectations for pairs of loci which, given a large number of pairwise observations, can be used to compute the composite likelihood of the population scaled rate of recombination $\rho = 4N_e c$. However, current LD based approaches for inferring recombination are limited in at least two ways:

Firstly, both *LDhat* (McVean et al., 2002; Auton and McVean, 2007) and *LDhelmet* (Chan et al., 2012) assume a simple model of recombination that only considers CO and ignores GC. A notable exception is the work of Gay et al. (2007), which extends the copying model of Li and Stephens (2003) to co-estimate CO and GC rates. Secondly, since two-locus approaches are generally conditioned on variant sites, they require phased data from many samples. Such data are still only available for a small minority of taxa.

Here, we address both these limitations by developing a simple composite likelihood method that allows co-estimation of CO and GC rates from individual diploid genomes. The calculation is based on analytic expectations for

99 observing heterozygosity at two loci under the simplest model of recombination
100 (crossover only) and genetic drift (Strobeck and Morgan, 1978; Lohse et al.,
101 2011) and has previously been implemented by Haubold et al. (2010). We first
102 use coalescent simulations to demonstrate that GC biases estimates of the CO
103 rate and show that this bias depends on the physical distance between loci.
104 We then exploit this non-linear dependence of GC on distance by incorporating
105 GC into the two locus expectations which gives a framework for co-estimating
106 the rates of CO and GC. Finally, we apply our method to genome-wide data
107 from wild-caught individuals of the house mouse *Mus musculus castaneus* and
108 compare our estimates to previous estimates of recombination based on a CO
109 only model (Booker et al., 2017). We also quantify the power of our approach
110 using simulations and investigate the extent to which the rates of CO and GC
111 are correlated with each other and with chromosome length.

112 **Materials and Methods**

113 **Analytic expectation of two locus heterozygosity**

114 We extended the models of Strobeck and Morgan (1978) and Haubold et al.
115 (2010) to account for GC and use a composite-likelihood approach to co-estimate
116 the rates of CO and GC and the mean GC tract length from individual genomes.
117 We consider a neutral Wright-Fisher model for the evolution of two linked loci
118 separated by d nucleotides in a population of N diploid individuals. Mutations
119 occur at per-base rate μ . CO occurs at per-base rate c and results in an exchange
120 of genetic material between sister chromatids. GC initiates at per-base rate g
121 and GC tracts are replaced by the sequence from the sister chromatid. For the
122 analysis, we re-scale time by $1/2N$ generations and use the population-scaled
123 parameters $\theta = 4N\mu$, $\kappa = 4Nc$, and $\gamma = 4Ng$ for mutation, CO, and GC rates,

124 respectively. We follow Wiuf (2000) in assuming that the GC tract length is
125 an exponentially distributed random variable with mean L base pairs (Hilliker
126 et al., 1994; Wiuf and Hein, 2000).

127 The heterozygosity at a single site H with $\mathbb{E}[H] = \frac{\theta}{1+\theta} \approx \theta$ is informative
128 only about the depth of a local genealogy: a site is more likely to be heterozy-
129 gous when the time to the most recent common ancestor, T_{mrca} , is large and
130 homozygous (i.e. identical in state) when T_{mrca} is small. Consider a second site
131 at a fixed distance d and define H_0 , H_1 , and H_2 as the proportion of all such
132 pairs where neither site, one site, or both sites are heterozygous (respectively).
133 These two-locus measures of heterozygosity are informative about the joint dis-
134 tribution of the two underlying genealogies and allow estimation of the rate of
135 recombination (Haubold et al., 2010; Lohse et al., 2011).

Using Eq. 4 of Strobeck and Morgan (1978), Haubold et al. (2010) derive
analytic expressions for the expected frequency of H_0 , H_1 , and H_2 as a function
of ρ , the total rate of events which lead to recombination between two sites
separated by d base pairs, agnostic to the underlying contributions of CO and
GC:

$$\mathbb{E}_{\theta,\rho}[H_0] = \frac{1}{(1+\theta)^2} + \Delta \frac{\theta}{(1+\theta)^2}, \quad (1)$$

$$\mathbb{E}_{\theta,\rho}[H_2] = \frac{\theta^2}{(1+\theta)^2} + \Delta \frac{\theta}{(1+\theta)^2} \quad (2)$$

$$\mathbb{E}_{\theta,\rho}[H_1] = 1 - \mathbb{E}_{\theta,\rho}[H_0] - \mathbb{E}_{\theta,\rho}[H_2], \quad (3)$$

where

$$\Delta = \frac{\theta(18 + \rho + 18\theta + \rho\theta + 4\theta^2)}{18 + 13\rho + \rho^2 + 54\theta + 40\theta^2 + 8\theta^3 + \rho(\rho\theta + 19\theta + 6\theta^2)}$$

136 represents the *zygosity* correlation (Lynch, 2008) or the deviation from inde-

pendence due to linkage. For large d , the genealogies at the two sites become independent, so $\mathbb{E}_{\theta,\rho}[H_0] = \mathbb{E}[H]^2$ and $\mathbb{E}_{\theta,\rho}[H_2] = (1 - \mathbb{E}[H])^2$, the first term in equations 1 and 2. In contrast, if the second site is tightly linked, the two sites likely share the same genealogy and we expect to see an excess in H_0 and H_2 . As d increases, so too does the probability that recombination occurs between the sites, resulting in differing genealogies and an increase in H_1 .

Co-estimating crossover and gene conversion rates

The above expectations for H_0 , H_1 , and H_2 make no assumption about the nature of recombination between pairs of sites, and ρ represents the rate at which the alleles transfers to different genetic backgrounds. For sites separated by a given distance d , we can obtain a maximum likelihood estimate for the total rate of recombination observed over these distances. Let $n_{d,0}$, $n_{d,1}$, and $n_{d,2}$ be the counts of pairs in the genome corresponding to H_0 , H_1 , and H_2 for a given distance d . The log-likelihood is then

$$\ln L_d(\rho, d) = n_{d,0} \ln(\mathbb{E}_d[H_0]) + n_{d,1} \ln(\mathbb{E}_d[H_1]) + n_{d,2} \ln(\mathbb{E}_d[H_2]). \quad (4)$$

If we assume that recombination between the two loci occurs only through CO, recombination always transfers alleles onto different genetic backgrounds, and the per-base recombination rate ρ/bp is constant. This is not true for GC, because the two sites will still share a genealogy if the GC tract both initiates and terminates between them. In other words, recombination through GC occurs only if the GC tract spans only one of the two focal sites, in which case GC has the same effect as a CO event. Accounting for the probability of recombination during GC (Wiuf and Hein, 2000; Wiuf, 2000), we can re-write the total rate of recombination ρ as a function of distinct rates of CO (κ) and GC (γ) and the

161 expected GC tract length L (Frisse et al., 2001; Langley et al., 2000).

$$\rho = \kappa d + 2\gamma L(1 - e^{-\frac{d}{L}}). \quad (5)$$

162
163 Given the dependence on distance d , observations $n_{d,0}$, $n_{d,1}$, and $n_{d,2}$ for a
164 single d are insufficient to estimate a three-parameter model of recombination.
165 However, by compositing the likelihood over many distances and substituting
166 Eq. 5 into Eq.s 1 - 3, we can co-estimate the rate of CO κ , the rate of GC γ ,
167 and the mean tract length L .

168 The composite likelihood is thus given by

$$\ln CL(\kappa, \gamma, L) = \sum_{d_{min}}^{d_{max}} \ln CL_d(\kappa, \gamma, L) \quad (6)$$

169
170 We have implemented the composite likelihood estimation described above in
171 python as a simple open source tool, **heRho** which is available at <https://github.com/samebdon/heRho>.

172 **Estimating recombination rates in the eastern house mouse**

173 As a proof of principle, we tested our composite likelihood estimation of recom-
174 bination on whole genome data from a well studied model species, the eastern
175 house mouse *M. m. castaneus*. Both direct and indirect estimates for the to-
176 tal rate of recombination exist for this species (Booker et al., 2017) and several
177 studies provide estimates for GC tract lengths (Paigen et al., 2008; Mansai et al.,
178 2011; Li et al., 2019b; Cole et al., 2014).

179 The data – originally described in Halligan et al. (2010) (ENA accession
180 number PRJEB2176) – consists of Illumina (PE) resequence data for ten in-
181 dividuals sampled from a wild *M. m. castaneus* population in India. Variant

182 calling is described in (Booker et al., 2017).

183 To minimize potential biases arising from background selection and the effect
184 of selection on nearby linked sites, all analyses were restricted to intronic regions,
185 which are putatively neutral. Specifically, we considered all introns ≥ 1 kb. Given
186 that the X chromosome has a different recombination rate than the autosomes
187 due to its lack of recombination in males, it was excluded from this analysis.

188 The final dataset included a total of 123,488 introns on autosomes 1-19,
189 spanning a total of 9×10^8 bases. For each intron, the positions of heterozygous
190 sites in each individual were converted into two-locus counts $n_{d,0}$, $n_{d,1}$, and $n_{d,2}$
191 for each distance d using a custom *Python* script (available on github). **heRho**
192 obtains maximum composite likelihood (MCL) estimates for ρ were obtained
193 using the *Python* library *NLopt*.

194 Power analysis

195 To quantify how accurately CO and GC rates can be estimated, we performed
196 a power analyses and parametric bootstrap on data simulated under the full
197 model in **msprime** (Kelleher et al., 2018): we simulated 100 replicates for each
198 chromosome under the MCL estimates obtained from the house mouse data (see
199 Results). Each replicate consisted of 10 diploid samples assuming $\theta = 0.071$ (the
200 observed heterozygosity), $\mu = 5.410^{-9}$ (Uchimura et al., 2015), $L = 108.4$ for all
201 chromosomes. The rates of CO and GC were set to those inferred for each *M.*
202 *castaneus* chromosome and the length of simulated sequence corresponded to
203 the total length of intronic sequence analysed for each chromosome (simulation
204 code is available in the github repository).

Results

Gene conversion explains the non-linear relationship between estimates of ρ and physical distance

As a first step, we used Eqs 1 - 4 to investigate how the per-base rate of recombination between pairs of sites depends on the distance between them (Fig. 1). Given a model that only includes CO events (red, dashed), we expect estimates of ρ/bp to be constant with respect to the distance between sites (red, dashed). However, when GC is included, nearby sites experience a higher per-base rate of recombination than pairs that are distant or unlinked (blue, dashed).

Figure 1: Maximum composite likelihood estimates of ρ/bp at fixed distances d between pairs of sites; simulations with CO only (red), simulations with CO and GC (blue), and empirical data for *M. m. castaneus* chromosome 19 (black). In each case, data was combined across a sample of 10 individuals. For simulations, $\theta = 0.0071$, $\kappa = 0.0014$, $\gamma = 0.0036$, and $L = 200$. The dashed lines show the expectation under the corresponding model (Eq.'s 1 - 4).

We find that estimates of ρ/bp based on a single replicate simulation either

under a model with GC (solid blue) or without (solid red) follow the expected relationship with the distance between sites d . When inferring ρ/bp between pairs of loci at different distances d in the mouse data (Fig. 1, black line), the relationship between ρ/bp and d is similar to that seen for data simulated under a model of recombination that includes both CO and GC. We note that the pattern of distance dependent ρ/bp is not exclusive to *M. m. castaneus*, but has been inferred previously by Haubold et al. (2010) for the ascidian *Cionia intestinalis* (although the authors do not mention GC).

Comparing the distance profiles of ρ/bp estimates between simulated and real data to each other and to analytic expectations (Fig. 1), we find two striking patterns:

Firstly, in the real data, estimates of ρ/bp are close to zero for nearby pairs of sites and increase sharply over the first ≈ 50 bases. In contrast, while the accuracy and precision of ρ/bp estimates in simulated data is strongly dependent on d (i.e. there is high variability for estimates over short distances $d < 50$), we find no similar monotonic increase in ρ/bp estimates over the first *approx* 50 bases. This discrepancy in estimates of ρ/bp in real vs simulated data suggests that over short distances ρ/bp estimates in the real data are biased downwards due to data quality/filtering effects: tightly-linked polymorphisms are difficult to distinguish from complex mutations (e.g. indels) and/or are removed by so called "best practices" variant calling/filtering approaches, skewing the observed values of H_0 , H_1 , and H_2 . This is compatible with the findings of Haubold et al. (2010) who have layered a sequencing error profile on coalescent simulations with CO only and shown that the noise generated by low-coverage data leads to a downward bias in estimates of ρ/bp over short distances.

Secondly, we find that estimates of ρ/bp are generally upwardly biased compared to expectations (compare solid and dashed lines in Fig. 1) due to sim-

plifying assumptions about the mutational process: unlike real genomes which consist of a discrete number of bases, equations 1-3 assume a continuous genome that evolves under the infinite sites mutation model. In that case, the occurrence of two mutations at a pair of sites always results in an H_2 state. In contrast, under a finite-sites mutation model (which `msprime` assumes by default) a back-mutation could generate an H_0 state. Indeed, the resulting upward bias is observed only in simulations that assume finite sites (Fig. S1). The bias is strongest at short distances, where recombination is rare and the expected values of the H_i are primarily governed by the mutational process. However, at greater distances, recombination primarily drives H_i counts and estimates converge to the model predictions.

Co-estimating crossover and gene conversion rates

By decomposing recombination to distinguish CO and the distance-dependent effects of GC (Eq. 5) and compositing the likelihood over the single-distance counts of H_i (Eq. 6), we may co-estimate both the rates of CO (κ) and GC (γ) and the mean tract length L . However, there are two challenges in implementing this inference: i) the noisiness of the data and the inaccuracy of the analytic results at short distances and ii) the inherent difficulty of co-estimating strongly correlated parameters.

Given the biases over short-distance in the real data, an obvious strategy is to introduce a minimum distance d_{min} in the composite likelihood (Eq. 6). However, since most of the information to co-estimate the rate of GC and the mean GC tract length is contained in short-distances, there is a trade-off between minimizing bias and retaining information. Our exploration of this trade-off both in real and simulated data shows that parameter estimates are stable across a broad range of d_{min} values (Fig. S2). To minimize the loss of information,

we chose $d_{min} = 100bp$ for all further analyses. Since genomes are finite and analysis is often restricted to a particular genomic partition, an upper distance threshold d_{max} is also unavoidable. To avoid biasing inference towards very long introns (which are selectively constrained), we limited the analysis to the first $d_{max} = 1000bp$ of each intron.

For our chosen range of distance thresholds $(d_{min}, d_{max}) = (100, 1000)$, as the next step in our preliminary analysis we examined the of the mouse data, we asked whether sufficient information is retained to confidently co-estimate the three recombination parameters. To do this, we focused on chromosome 19 for which **heRho** gives the following maximum composite likelihood estimates: $\kappa = 0.00267$, $\gamma = 0.0044$, and $L = 113.24$. Examining the support, as measured by the logarithm of the composite likelihood ($lnCL$), surface around this maximum illustrates the challenge of co-estimating L and κ . Although estimates are negatively correlated (Fig. 2 B), we were positively surprised that it is not only possible to co-estimate both parameters (the $lnCL$ surface is smooth and contains a distinguishable optimum, Fig. 2), but that the estimates are indeed plausible, i.e. are compatible with direct, experimental estimates.

Less surprisingly, we observe that estimates for the GC rate γ and the mean tract length L are negatively correlated both with each other (Figure 2 A) and with estimates of the CO rate κ (Figure 2 B and C). Given the degree to which parameter estimates are confounded and the fact that we have no biological reason to expect the length of GC tracts to vary between chromosomes, we chose to co-estimate a global L and chromosome-specific GC and CO parameters in the subsequent analysis of the mouse data described below.

Figure 3: Recombination parameters co-estimated for the 19 autosomes of *M. m. castaneus* using data pooled across ten individuals (black dots) and corresponding parametric bootstrap results from 100 replicate simulations. The per chromosome estimates of the GC rate (γ) and mean tract length are shown in yellow, estimates for the rate of CO (κ) in blue.

The recombination profile of *M. m. castaneus*

Our per chromosome co-estimates of the rates of CO (κ) and GC (γ) in *M. m. castaneus* (based on data from all 10 individuals, Fig. 3) range from 0.00145 to 0.00269 and 0.00211 to 0.00461 and respectively. Assuming that the mean GC tract length is the same for all autosomes, our global MCL estimate for this parameter is 108 bases, which is within the range of previous direct estimates (≈ 10 -300 for NCO and 200-1200 for CO gene conversion tracts (Paigen et al., 2008; Mansai et al., 2011; Li et al., 2019b; Cole et al., 2014)). When restricting our analysis to a single individual, we obtain broadly concordant estimates of the CO rate and mean tract length, but with less data available, the GC rate estimates vary much more across chromosomes (Fig. S3). Indeed, for simulations, increasing the sample size from one (Fig. S4) to ten individuals (Fig. 3) reduces the variance but not the bias of the results.

Our per chromosome estimates recover several well known, broad-scale patterns: Firstly, as some GC events occur during CO, we expect the rates of CO and GC to be mechanistically and positively correlated, and this is indeed the case (Figure S5). Note that this signal contrasts with the negative correlation in the estimation error of both parameters (Fig. 2 C) and therefore must reflect

the underlying dynamics of meiotic recombination rather than any statistical artefact.

Secondly, as chromosomes have a minimum bound of map length at 50cM due to obligate CO, we expect the CO rate per-base to be negatively correlated with chromosome length. We recover this pattern (Figure S5) that is widely documented not only in mammals (Johnston et al., 2017), including humans (The International Genome Sequencing Consortium, 2001), but also in flycatchers (Kawakami et al., 2014), yeast (Kaback et al., 1999) and butterflies (Martin et al., 2016). In contrast, we find that per chromosome estimates of the rate of GC are not significantly correlated with chromosome length ($p = 0.148$) (Figure S5 C). Since a high proportion of GC products are the result of non-CO recombination events, we do not expect GC rates to correlate significantly with patterns of chiasma formation.

Discussion

A significant challenge in population genetics is to develop inference methods that are both efficient in extracting signals about population processes from sequence variation and simple, i.e. rely on a minimum number of assumptions. Given that high-coverage whole-genome data have become the norm, we now have the ability to study the fundamental forces of evolution, such as recombination, both at fine genomic scales and across a broad taxonomic range. We have developed a method for quantifying CO and GC from the distribution of heterozygous sites in small samples – even from individual diploid genomes.

heRho’s strengths and weaknesses

Not only does our framework allow for more complete/realistic estimates of recombination than previous CO-restricted methods (Auton and McVean, 2007;

Haubold et al., 2010), it is also simpler and less error prone since it does not rely on phase information which can bias results. For example, Booker et al. (2017) find that in the presence of switch errors, LDhelmet consistently overestimates the CO rate. Furthermore, by including homozygous states in the analysis, we garner sufficient information to co-estimate CO and GC when data are restricted to short distances. As such, **heRho** can potentially generate a whole-genome annotation-specific recombination profile, even for small genomic features that partition the genome (e.g. 4D sites in exons).

Our method does suffer from many of the same potential biases as other population genetic estimators of recombination: given that we are assuming a neutrally evolving Wright-Fisher population of constant size, any demographic and selective processes that generate LD (e.g. positive selection, population bottlenecks or admixture), will bias estimates of recombination obtained with **heRho** downwards. Perhaps a more important question is whether there are processes that bias the inference of GC rates specifically or, in the extreme case, may mimic signatures of GC, where none exist. While there are processes that may increase short-range measures of LD (e.g. heterogeneity in mutation rate that leads to an excess of H_2 and H_0), we cannot think of forces or mechanisms that, like GC, have the opposite effect of decreasing LD only over short genomic distances on the order of GC tract lengths.

As demonstrated, our method **heRho** relies on large amounts of sequence data and is fundamentally limited by the frequency of the rarest two-locus observation H_2 , which for any distance d , is of order H^2 . We therefore expect that it will not be possible to obtain estimates of GC and CO at small genomic scales (say in windows of 100kb). While pooling observations across individuals increases the number of H_i observations and reduces variance in the estimates, we expect many heterozygous sites to be shared among individuals, and thus the returns

diminish quickly with sample size.

Reconciling heRho's recombination estimates for *M. m. castaneus* with LDHelmet

How do our estimates in *M. m. castaneus* compared to those obtained previously by Booker et al. (2017) using LDHelmet and a CO-only recombination model? Co-estimated under a model of GC, our genome-wide average of the CO rate per-base (0.00186) is lower than the ρ estimate of Booker et al. (2017) (0.00924, averaged across autosomes) by a factor of 5. If we compare the total recombination rate estimated between any two adjacent bases, which corresponds to the upper bound of the recombination rate in our model (for $d = 1$ eq. 5 reduces to $\rho = \kappa + 2\gamma$), the results are quite similar (0.00841 vs. 0.00924). While this suggests that GC may contribute substantially to the ρ as estimated by Booker et al. (2017), this is unrealistic. LDHelmet uses long-range SNP-only data, making it attune to the broader signal of CO and less sensitive to the short-range effects of GC. Rather, the difference between the estimates likely reflects biology. The Booker et al. (2017) estimates are obtained using data from large contiguous windows of the genome aggregated over all sites and genomic partitions which vary in proportion along the genome but will be dominated by intergenic sequence. Our estimates instead reflect the (per-chromosome) recombination profile specifically for the beginning of introns. Direct recombination estimates in humans suggest that the recombination rate in introns is lower than the genome-wide average (Myers et al., 2005). Furthermore, intron length is negatively correlated with recombination rates in some taxa (Comeron and Kreitman, 2000), and our filtering strategy enriched for long introns.

386 Further applications and outlook

387 There are several potential avenues for further work. Firstly, it should be pos-
 388 sible to relax the assumption of an infinite sites mutation model. While our
 389 analysis of the *M. m. castaneus* data reveals very small/tolerable biases (Figure
 390 3); basing estimates of GC and CO on more realistic mutation models might
 391 be important when analyzing more heterozygous genomes. Secondly, as a nat-
 392 ural choice, we have assumed that loci are individual nucleotides. One could in
 393 principle extend the two locus inference to longer blocks of sequence and use
 394 the framework developed by (Lohse et al., 2011) to base inference on the joint
 395 distribution of pairwise differences. However, this comes at the cost of intro-
 396 ducing additional assumptions and biases. Finally, it would be interesting to
 397 explore whether the machinery could be extended to three loci. If analogous
 398 analytic results for three loci are tractable, this would allow extracting substan-
 399 tially more signal and better estimate the rate and tract length of GC events
 400 from genomic data.

401 While we have limited our analyses to long introns, any genomic data parti-
 402 tion for which pairwise heterozygosity can be accurately measured over a suffi-
 403 cient range of physical distances is suitable. It remains to be seen whether our
 404 method is informative about smaller genomic partitions such as centromeres
 405 and chromosomal inversions which differ from the genome-wide rates of recom-
 406 bination in systematic and where GC may occur but CO is restricted (Korunes
 407 and Noor, 2017).

408 Acknowledgements

409 This work was supported by an ERC starting grant (ModelGenomLand 757648).
 410 SE is supported by an EastBio studentship from the British Biological Sciences

Research Council (BBSRC). KL is supported by a fellowship from the Natural Environment Research Council (NERC, NE/L011522/1).

We thank Stuart Baird for helpful discussions and Susan Johnston for insightful comments on the manuscript.

References

- Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. *Genome Research*, 17(8):1219–1227.
- Booker, T., Ness, R., and Keightley, P. (2017). The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data. *Genetics*, 207(1):297–309.
- Campos, J. L. and Charlesworth, B. (2019). The effects on neutral variability of recurrent selective sweeps and background selection. *Genetics*, 212(1):287–303.
- Carpenter, A. (1982). Mismatch repair, gene conversion, and crossing-over in two recombination-defective mutants of drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 79(19):5961–5965.
- Casola, C., Ganote, C. L., and Hahn, M. W. (2010). Nonallelic gene conversion in the genus drosophila. *Genetics*, 185(1):95–103.
- Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in drosophila melanogaster. *PLoS genetics*, 8(12):e1003090.
- Charlesworth, B. and Charlesworth, D. (1997). Rapid fixation of deleterious alleles can be caused by Muller’s ratchet. *Genetical Research*, 70(1):63–73.
- Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303.
- Cole, F., Baudat, F., Grey, C., Keeney, S., De Massy, B., and Jasin, M. (2014). Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nature genetics*, 46(10):1072–1080.
- Comeron, J. M. and Kreitman, M. (2000). The correlation between intron length and recombination in drosophila: dynamic equilibrium between mutational and selective forces. *Genetics*, 156(3):1175–1190.

443 Crown, K. N., Miller, D. E., Sekelsky, J., and Hawley, R. S. (2018). Local in-
444 version heterozygosity alters recombination throughout the genome. *Current*
445 *Biology*, 28(18):2984–2990.e3.

446 DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., and Nielsen, R.
447 (2016). SweepFinder2: Increased sensitivity, robustness, and flexibility. *Bioin-*
448 *formatics*, 32(12):1895–1897.

449 DeGiorgio, M., Lohmueller, K. E., and Nielsen, R. (2014). A model-based
450 approach for identifying signatures of ancient balancing selection in genetic
451 data. *PLoS Genet*, 10(8):e1004561.

452 Frisse, L., Hudson, R., Bartoszewicz, A., Wall, J., Donfack, J., and Di Rienzo,
453 A. (2001). Gene conversion and different population histories may explain
454 the contrast between polymorphism and linkage disequilibrium levels. *The*
455 *American Journal of Human Genetics*, 69(4):831–843.

456 Gay, J., Myers, S., and McVean, G. (2007). Estimating Meiotic Gene Conversion
457 Rates From Population Genetic Data. *Genetics*, 177(2):881–894.

458 Griffiths, A. J., Gelbart, W. M., Lewontin, R. C., and Miller, J. H. (2002). *Mod-*
459 *ern genetic analysis: integrating genes and genomes*, volume 2. Macmillan.

460 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D.
461 (2009). Inferring the joint demographic history of multiple populations from
462 multidimensional SNP frequency data. *PLoS Genetics*, 5(10).

463 Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., and Keightley, P. D.
464 (2010). Evidence for pervasive adaptive protein evolution in wild mice. *PLoS*
465 *Genetics*, 6(1):e1000825.

466 Haubold, B., Pfaffelhuber, P., and Lynch, M. (2010). MRho - A program for
467 estimating the population mutation and recombination rates from shotgun-
468 sequenced diploid genomes. *Molecular Ecology*, 19(SUPPL. 1):277–284.

469 Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial
470 selection. *Genetical Research*, 8(3):269–294.

471 Hilliker, J., Haraw, G., Reaume, A. G., and Clark, S. H. (1994). Meiotic Gene
472 Conversion Tract Length Distribution Within the rosy Locus of *Drosophila*
473 *melanogaster*. page 8.

474 Jeffreys, A. and May, C. (2004). Intense and highly localized gene conversion
475 activity in human meiotic crossover hot spots. *Nature Genetics*, 36(2):151–
476 156.

477 Johnston, S. E., Bérénos, C., Slate, J., and Pemberton, J. M. (2016). Conserved
478 Genetic Architecture Underlying Individual Recombination Rate Variation in
479 a Wild Population of Soay Sheep (*Ovis aries*). *Genetics*, 203(1):583–598.

Johnston, S. E., Huisman, J., Ellis, P. A., and Pemberton, J. M. (2017). A High-Density Linkage Map Reveals Sexual Dimorphism in Recombination Landscapes in Red Deer (*Cervus elaphus*). *G3 Genes—Genomes—Genetics*, 7(8):2859–2870.

Kaback, D. B., Barber, D., Mahon, J., Lamb, J., and You, J. (1999). Chromosome size-dependent control of meiotic reciprocal recombination in *Saccharomyces cerevisiae*: The role of crossover interference. *Genetics*, 152(4):1475–1486.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., Olason, P., and Ellegren, H. (2014). A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*, 23(16):4035–4058.

Kelleher, J., Thornton, K. R., Ashander, J., and Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLoS Computational Biology*, 14(11):e1006581.

Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Richardson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S., Frigge, M., Thorgeirsson, T., Gulcher, J., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3):241–247.

Kong, A., Thorleifsson, G., Frigge, M. L., Masson, G., Gudbjartsson, D. F., Villemoes, R., Magnusdottir, E., Olafsdottir, S. B., Thorsteinsdottir, U., and Stefansson, K. (2014). Common and low-frequency variants associated with genome-wide recombination rate. *Nature Genetics*, 46(1):11–16.

Korunes, K. and Noor, M. (2017). Gene conversion and linkage: effects on genome evolution and speciation. *Molecular ecology*, 26(1):351–364.

Langley, C. H., Lazzaro, B. P., Phillips, W., Heikkinen, E., and Braverman, J. M. (2000). Linkage Disequilibria and the Site Frequency Spectra in the *su(s)* and *su(w a)* Regions of the *Drosophila melanogaster* X Chromosome.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233.

Li, R., Bitoun, E., Altemose, N., Davies, R. W., Davies, B., and Myers, S. R. (2019a). A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nature Communications* 2019 10:1, 10(1):1–15.

Li, R., Bitoun, E., Altemose, N., Davies, R. W., Davies, B., and Myers, S. R. (2019b). A high-resolution map of non-crossover events reveals impacts of

genetic diversity on mammalian meiotic recombination. *Nature communications*, 10(1):1–15.

Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., Yu, L., Liti, G., Tian, D., Hurst, L. D., and Yang, S. (2018). Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature Ecology and Evolution*, 2(1):164–173.

Lohse, K., Harrison, R., and Barton, N. (2011). A General Method for Calculating Likelihoods Under the Coalescent Process. *Genetics*, 189(3):977–987.

Lynch, M. (2008). Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Molecular Biology and Evolution*, 25(11):2409–2419.

Ma, L., O’Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., Bickhart, D. M., Cole, J. B., Null, D. J., Liu, G. E., Da, Y., and Wiggans, G. R. (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLOS Genetics*, 11(11):1–24.

Mansai, S. P., Kado, T., and Innan, H. (2011). The Rate and Tract Length of Gene Conversion between Duplicated Genes. *Genes*, 2(2):313–331.

Martin, S. H., Möst, M., Palmer, W. J., Salazar, C., McMillan, W. O., Jiggins, F. M., and Jiggins, C. D. (2016). Natural selection and genetic diversity in the butterfly *Heliconius melpomene*. *Genetics*, 203(1):525–541.

McMahill, M. S., Sham, C. W., and Bishop, D. K. (2007). Synthesis-dependent strand annealing in meiosis. *PLoS Biology*, 5(11):2589–2601.

McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160(3):1231–1241.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). Genetics: A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324.

Nassif, N., Penney, J., Pal, S., Engels, W., and Gloor, G. (1994). Efficient copying of nonhomologous sequences from ectopic sites via p-element-induced gap repair. *Molecular and cellular biology*, 14(3):1613–1625.

Paigen, K., Szatkiewicz, J. P., Sawyer, K., Leahy, N., Parvanov, E. D., Ng, S. H., Graber, J. H., Broman, K. W., and Petkov, P. M. (2008). The recombinational anatomy of a mouse chromosome. *PLoS genetics*, 4(7):e1000119.

Peñalba, J. V. and Wolf, J. B. W. (2020). From molecules to populations: appreciating and estimating recombination rate variation. *Nature Reviews Genetics*, 21(8):476–492.

556 Resnick, M. A. (1976). The repair of double-strand breaks in dna: a model
557 involving recombination. *Journal of Theoretical Biology*, 59(1):97–106.

558 Setter, D., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., Hermisson, J.,
559 Buerkle, A., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., and Her-
560 misson, J. (2020). VolcanoFinder: Genomic scans for adaptive introgression.
561 *PLoS Genetics*, 16(6):697987.

562 Simonsen, K. L., Churchill, G. A., et al. (1996). A markov chain model of
563 coalescence with recombination.

564 Smeds, L., Mugal, C. F., Qvarnström, A., and Ellegren, H. (2016). High-
565 resolution mapping of crossover and non-crossover recombination events
566 by whole-genome re-sequencing of an avian pedigree. *PLoS genetics*,
567 12(5):e1006044.

568 Stapley, J., Feulner, P. G., Johnston, S. E., Santure, A. W., and Smadja, C. M.
569 (2017). Variation in recombination frequency and distribution across eukary-
570 otes: patterns and processes. *Philosophical Transactions of the Royal Society*
571 *B: Biological Sciences*, 372(1736):20160455.

572 Strobeck, C. and Morgan, K. (1978). The effect of intragenic recombination on
573 the number of alleles in a finite population. *Genetics*, 88(4):829–844.

574 Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983).
575 The double-strand-break repair model for recombination.

576 Talbert, P. B. and Henikoff, S. (2010). Centromeres convert but don’t cross.
577 *PLOS Biology*, 8(3):1–5.

578 The International Genome Sequencing Consortium (2001). Initial sequencing
579 and analysis of the human genome. *Nature*, 409(6822):860–921.

580 Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A.,
581 Miura, I., Wakana, S., Nishino, J., and Yagi, T. (2015). Germline mutation
582 rates and the long-term phenotypic effects of mutation accumulation in wild-
583 type laboratory mice and mutator mice. *Genome research*, 25(8):1125–1134.

584 Wiuf, C. (2000). A coalescence approach to gene conversion. *Theoretical Pop-*
585 *ulation Biology*, 57(4):357–367.

586 Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics*,
587 155(1):451–462.

588 Supporting Information

Figure S1: The bias of the estimator for discrete genomes. Here, we compare the analytic predictions for single-distance estimates of ρ/bp to those observed in simulations. The model predictions with and without GC are shown in dark blue and dark red, respectively. Light blue and light red show estimates from a coalescent model with a continuous genome, and the turquoise and orange lines to a discrete genome, correspondingly. Here, estimates were obtained from the combined data of 100 replicates under each simulated scenario. Parameters as in fig 1

Figure S2: The effect of minimum distance on the per-autosome composite likelihood estimate of recombination. Panel A shows, from top to bottom, the estimated values of γ , κ , and L as a function of the minimum distance included in the likelihood calculation. Each indexed color corresponds to one chromosome, with chromosome 1 the darkest and chromosome 19 the lightest. Panel B shows the corresponding results obtained using the a single replicate of the simulations used for parametric bootstrapping. The relative error in the estimated value, that is, the (observed - expected)/expected value, and the chromosomes are colored by index from the lowest (dark) to highest (light) recombination rate.

Figure S3: Comparison of the co-estimated CO rate (green) and GC rate (purple) *M. m. castaneus* for single-individual data (dashed) vs data pooled from ten individuals (solid). Each marker corresponds to a unique chromosome and estimate. The single- and multiple-individual estimates of the mean tract length were 107.8 and 108.4, respectively.

Figure S4: Bootstrapping results when data is limited to a single individual. The black dots correspond to the recombination parameters co-estimated for the autosomes using data pooled across ten *M. m. castaneus* individuals (Fig. 3). Here, we randomly subsampled one individual from each of the simulation replicates. The per-chromosome GC rate and mean tract length estimates are shown in yellow, and the corresponding CO rate estimates are shown in blue.

Figure S5: Left) Per chromosome estimates for the rates of CO and GC in *M. m. castaneus* are positively correlated; Center) Given that chromosomes have a roughly fixed map length, we expect ρ per base to correlate negatively with the physical length of chromosomes; Right) we find no analogous correlation between the rate of GC (γ) and chromosome length.