The Facultative Sex Coalescent with Crossing Over Recombination and Multi-Site Gene Conversion

Matthew Hartfield[1,*,†], Stephen I. Wright[*], and Aneil F. Agrawal[*]

12th December 2017

[*] Department of Ecology and Evolutionary Biology, University of Toronto, Ontario, Canada.

[†] Bioinformatics Research Centre, Aarhus University, Aarhus C 8000, Denmark.

[1] *Corresponding author:* matthew.hartfield@birc.au.dk

Running Head: Linkage Disequilibrium under partial sex

Key words: Facultative sex, coalescent theory, crossing over recombination, gene conversion, linkage disequilibrium

Website for simulation code:

`http://github.com/MattHartfield/FacSexCoalescent`

Abstract

The extent of correlations between loci, as measured by linkage disequilibrium, is commonly used to infer the effect of different types of gene exchange. It can also provide information on rates of sex in facultative sexual organisms, as meiotic crossing over occurs during sexual reproduction. In addition, recent theoretical and empirical work has provided evidence that gene conversion shapes genetic diversity in facultative sexuals. Here we outline computational models of a facultative sexual coalescent process that includes both crossover recombination and multi-site gene conversion, to determine how linkage disequilibrium is affected with facultative sex. We demonstrate that the decay in linkage disequilibrium scales with the rate of sex if it is sufficiently high, indicating that linkage disequilibrium data can be potentially used to deduce rates of sex from genome data. With very low rates of sex, mitotic gene conversion both breaks down associations between alleles and removes within-individual diversity. The two combined effects produce complex linkage disequilibrium patterns. Furthermore, strong population structure and low rates of sex lead to lower average linkage disequilibrium values than those in panmictic populations, due to the influence of low-frequency polymorphisms created by allelic sequence divergence arising in individual subpopulations. These analyses provide information on how to interpret observed linkage disequilibrium patterns in facultative sexuals, and determine what genomic forces are likely to shape them.

Introduction

Coalescent theory is a powerful mathematical framework that is used to determine how natural selection and demographic history affect genetic diversity (Kingman 1982; Rosenberg and Nordborg 2002; Hein *et al.* 2005; Wakeley 2009). Traditional coalescent models assume that the population is obligately sexual, but there has been less attention on creating models that account for different reproductive modes. While the coalescent with self-fertilisation has been extensively studied (Nordborg and Donnelly 1997; Nordborg 1997, 2000; Nordborg and Krone 2002), little theory exists on coalescent histories in organisms with other mixed reproductive systems.

Building on previous theory (Brookfield 1992; Burt *et al.* 1996; Balloux *et al.* 2003; Bengtsson 2003; Ceplitis 2003), Hartfield *et al.* (2016) investigated a coalescent model of facultative sexual diploid organisms, which reproduce via a mixture of sexual and parthenogenetic reproduction. It was determined that when an organism exhibits very rare population-level rates of sex ($\sigma \sim \mathcal{O}(1/N)$, for σ the rate of sex and N the population size), the presence of gene conversion can reduce within-individual diversity to less than that expected in sexual organisms, contrary to traditional expectations that alleles within individuals should diverge at these low rates of sex (Mark Welch and Meselson 2000; Butlin 2002). Hence these results provide a potential explanation as to why allelic divergence is not widely observed in empirical studies of facultative sexual organisms (reviewed in Hartfield (2016)).

However, this analysis only modelled the genetic history at a single, non-recombining locus. Here, genealogies only greatly differed from those in oblig-

ately sexual organisms at very low rates of sex ($\sigma \sim \mathcal{O}(1/N)$). As a consequence, diversity-based estimators for σ only work when sex is sufficiently rare (Ceplitis 2003; Hartfield *et al.* 2016). In contrast, many facultatively sexual organisms exhibit much higher rates of sex. Pea aphids reproduce sexually about once every ten generations (Simon *et al.* 2002), while Daphnia undergo one sexual generation and five to twenty asexual generations a year (Haag *et al.* 2009). The wild yeast *Saccharomyces paradoxus* has an outcrossing rate of 0.001; while low, this value is four orders of magnitude higher than $1/N_e$ (Tsai *et al.* 2008). If we wish to create a general coalescent model that has the ability to infer rates of sexual reproduction in species undergoing more frequent rates of sex, then we need to increase the power of this coalescent process to consider how patterns of genetic diversity at multiple loci are affected with infrequent sex.

This is achievable by considering how genealogies of multiple sites correlate along a chromosome. Two completely linked sites will reach a common ancestor in the past at the same time, so will share the same gene genealogy. However, if a recombination event (e.g., via meiotic crossing over) were to separate the sites, each sub-segment may have different genetic histories (Hudson 1983). Breaking apart correlations between sites is reflected with lower linkage disequilibrium, which can be measured from genomic data (Griffiths 1981; Hudson and Kaplan 1985; Hudson 1990; Simonsen and Churchill 1997; McVean 2002). Gene conversion can also break apart correlations between samples through transferring genetic material across DNA strands (Wiuf and Hein 2000).

Including meiotic recombination into the facultative sex coalescent and studying the resulting linkage disequilibrium patterns should provide more information with which one can discriminate between higher rates of sex. As meiotic crossing

over occurs during sexual reproduction, we expect that the extent to which linkage disequilibrium is broken down should scale with the rate of sex (see Nordborg (2000) for a related argument for the coalescent with self-fertilisation). Tsai *et al.* (2008) used this logic to calculate rates of sex in the yeast *Saccharomyces paradoxus*, by comparing effective population sizes inferred from linkage disequilibrium decay (which should scale with the meiotic recombination rate, and therefore the rate of sex) with those from nucleotide variation (which should be independent of the sex rate if it is sufficiently high). Lynch *et al.* (2017) used similar arguments to conclude that even though the facultative sexual water flea *Daphnia pulex* has a lower overall crossover recombination rate than *Drosophila melanogaster*, it has a higher recombination rate when sex does occur. However, we show below that this approach becomes problematic when rates of sex are very low, or if mitotic gene conversion is an important force in shaping linkage disequilibrium.

We describe a routine for simulating multi-site genealogies with facultative sex, allowing for both crossover recombination and different forms of gene conversion. We use these new models to investigate how linkage disequilibrium patterns are affected in facultative sexual organisms, and how these results can be used to infer rates of sex from genome data.

Simulation Algorithm

Coalescent simulations have proved invaluable for predicting genetic correlations under various neutral scenarios (Hudson 2002). We have previously developed an algorithm to build genealogies of facultative sexuals at a single non-recombining locus (Hartfield *et al.* 2016). The algorithm simulates genealogies of x paired

samples (i.e., where both haplotypes are sampled from a diploid individual) and y unpaired samples. We track paired and unpaired samples separately because their coalescent histories (and therefore expected diversity patterns) will differ at very low rates of sex, in contrast to obligate sex coalescent processes (Hartfield *et al.* 2016). The simulation draws times to the previous event in the past; once the time has been found it is then determined how many of the x paired samples segregate into different individuals due to sexual reproduction. Subsequent events are then resolved (e.g., coalescence, migration). We will henceforth describe how to implement meiotic crossing-over (or 'crossovers' for short) and multi-locus gene conversion into this simulation. Using this new program, we can investigate how the breakdown of linkage disequilibrium is affected by different rates of sex under a variety of scenarios, and determine when it is possible to infer rates of sex. Readers uninterested in the details of how crossing over and gene conversion were modelled can skip directly to the results.

Implementing Crossover Recombination

We implemented an extension of Hudson's (1983) algorithm to take facultative sex into account. In this process, a crossover splits a sample into two, with each part carrying ancestral material. In the facultative sexual simulation, meiotic crossovers only occur in individuals undergoing sexual reproduction. Going back in time, sex involving a crossover will transform an unpaired sample into a paired sample (i.e., the unpaired sample descended from the two homologs in the parent). For a paired sample, sex splits the two samples into separate individuals. If a crossover also occurred, then the affected sample becomes a paired sample. Figure A in

Supplementary File S1 presents a schematic of these different outcomes. This outcome differs from that in obligate sex coalescent routines (e.g., Hudson (1983)) where it is unnecessary to keep track of paired versus unpaired samples because pairings are extremely brief when sex is frequent.

We can formalise this routine in the facultative sexual framework as follows. Let the absolute crossover rate between any two adjacent sites be r, and L the number of sites that the genetic samples cover. The total crossover rate in each individual at the start of the process is $r(L-1)$. At a set time there are x paired samples, each consisting of two haploid samples, and y unpaired samples (hence $2x+y$ total haploid samples). The rate of sex is σ, where 1 represents obligate sex and 0 obligate asexuality.

Meiotic crossovers act on unpaired samples with probability $\sigma \cdot r \cdot L_{e,y}$. $L_{e,y}$ is the 'effective' crossover length summed over all y unpaired samples, defined as the total number of breakpoints where a crossover can create two new samples, each carrying ancestral material. Note that any sites that have coalesced are converted into non-ancestral material (Hein *et al.* 2005). Furthermore, if k out of the x paired samples segregate via sex into $2k$ new unpaired samples, then the crossover probability is increased by $r \cdot L_{e,2k}$. Here $L_{e,2k}$ is the 'effective' crossover length summed over the $2k$ new unpaired samples. It is different from $L_{e,y}$ since the set of valid breakpoints can differ depending on the set of samples that crossing over can act upon. The $2k$ new unpaired samples are regarded as a distinct but transitory class of samples. Because they are already determined to be have been created by sex (by an earlier step in the algorithm), there is no factor σ contributing to their probability of experiencing a crossing over. Those that undergo crossing over are transformed into regular paired samples; those that do not become regular

unpaired samples. We include these L_e terms to ensure that unnecessary crossover events are not considered, thus speeding up the routine (Hein *et al.* 2005).

The probability of crossovers is considered alongside other events to determine whether the next event involves a crossover. If it is chosen, then one of the appropriate samples is picked at random (weighing by the length of extant breakpoints present in each sample), and split at a valid breakpoint to produce a new paired sample (Figure A in Supplementary File S1).

Implementing 'Disruptive' Gene Conversion

Determining type of gene conversion (meiotic or mitotic)

Gene conversion can affect a sample in several ways. We first consider 'disruptive' gene conversion, where (i) gene conversion initiates outside the sample but finishes within it; (ii) gene conversion begins within the sample but extends beyond it; or (iii) both conversion breakpoints lie within the sample. We consider both meiotic gene conversion, which repairs double-strand breaks during meiosis and therefore only acts during sexual reproduction, and mitotic gene conversion that can occur during sexual and asexual phases. To account for these events, up to four additional parameters are specified (alongside the number of sites in the ancestral samples). Two new parameters are the probabilities of either meiotic or mitotic gene conversion occurring with its leftmost boundary at a specific site, denoted g_{me} and g_{mi} respectively. Each parameter represents the probability that the leftmost boundary of a gene conversion event occurs on the recipient homolog at a given site. We also define the average length of gene conversion events, denoted λ_{me} for meiotic gene conversion and λ_{mi} for mitotic conversion. We implement and extend

the algorithm of Wiuf and Hein (2000) to calculate the probability of either type of gene conversion event occurring each generation. To understand how the different events are considered, it is easiest to relate calculations to the obligate sex case with a single type of gene conversion event. Here, the probability of disruptive gene conversion in a sample of length L is $g(L-1)Q^*$, where g is the probability that the leftmost edge of a gene conversion tract is at a given site, and:

$$Q^* = 1 + \frac{1}{Q}\left(1 - e^{-Q}\right) \tag{1}$$

$Q = (L-1)/\lambda$ is the number of breakpoints in units of average gene conversion length. Q^* accounts for gene conversion events that only partly lie in ancestral material (i.e., only one end of the gene conversion lies in ancestral material) as well as those that lie entirely within this region (i.e., both breakpoints lie within ancestral material) (Wiuf and Hein 2000). In the facultative sex coalescent, we can partition this probability depending on the type of conversion event (meiotic or mitotic), and the number of each type of sample (paired or unpaired) present at the time. Let there be $(x-k)$ paired samples after k of them have split after sex, making $2(x-k)$ haplotypes in total; y unpaired samples; and $2k$ new unpaired samples following genetic segregation. After partitioning over all possible outcomes, the total probability of disruptive gene conversion equals:

$$\Sigma = 2Q^*_{mi}g_{mi}(x-k) + y(Q^*_{mi}g_{mi} + \sigma Q^*_{me}g_{me}) + 2k(Q^*_{mi}g_{mi} + Q^*_{me}g_{me}) \tag{2}$$

Here, Q^*_{mi} and Q^*_{me} are equal to Q^* above, with parameters $Q_{mi} = (L-1)/\lambda_{mi}$ and $Q_{me} = (L-1)/\lambda_{me}$. As segregation has already been resolved, the $(x-$

k) remaining paired samples reproduce asexually so they can only be subject to mitotic gene conversion. Unpaired samples can be subject to both meiotic and mitotic gene conversion, hence for each unpaired sample (y in total) we also have to consider the rate of sex σ. Note that there is no σ term when considering the $2k$ new unpaired samples, as they have already undergone sex by this point in the algorithm. In contrast to the crossover procedure, we do not weigh samples by the number of breakpoints within ancestral material. Gene conversion events affecting only non-ancestral material are allowed to occur.

When a disruptive gene conversion event occurs, it is first determined if it acts on unpaired or paired samples. The probability that gene conversion acts on a paired sample is $2Q^*_{mi}g_{mi}(x-k)/\Sigma$ where Σ is given by Equation 2, and one minus this probability is the chance it acts on unpaired samples. If acting on a paired sample then only mitotic gene conversion can occur. If acting on unpaired samples, a further random draw is made to determine whether the gene conversion event is meiotic or mitotic. Let $\Upsilon = g_{mi}Q^*_{mi}(y+2k) + g_{me}Q^*_{me}(y\sigma+2k)$ be the probability of a gene conversion event that occurs on an unpaired sample. If an unpaired sample undergoes conversion, the probability that the event is mitotic equals $(g_{mi}Q^*_{mi}(y+2k))/\Upsilon$; a similar calculation can be made for meiotic gene conversion.

Choosing gene conversion breakpoints

The scaling term Q^* accounts for the fact that gene conversion does not necessarily take place entirely within the gene tract, but may only partially overlap with it. We follow calculations outlined in Wiuf and Hein (2000) to accurately model the relative frequency of each of these events. Given the tract length in units of

conversion events Q (which can be either Q_{mi} or Q_{me}), $K(Q) = 1-(1-\exp(Q))/Q$ is the probability that if gene conversion starts in the sample, it will also end within it (Wiuf and Hein 2000, Eq. 2). One can then define the probability that both breakpoints occur within the sample (Wiuf and Hein 2000, Eq. 11):

$$p_2 = \frac{K(Q)}{2 - K(Q)} \tag{3}$$

The probability that only one breakpoint falls within the sample is equal to $p_1 = 1 - p_2$ (Wiuf and Hein 2000, Eq. 12). We first choose whether one or both breakpoints lie within the sample, as determined by Equation 3. The appropriate start and end points are then chosen from the relevant probability distributions. Wiuf and Hein (2000) determined how the distribution of breakpoints depend on whether one or both breakpoints lie within the genome tract. For example, if only one breakpoint lies in the tract then it is likelier to occur closer to one of its edges. In Appendix A and Section A of Supplementary Mathematica File S2, we describe how we use calculations outlined in Wiuf and Hein (2000) to determine the relevant start and end points, given the number of breakpoints.

Resolution of gene conversion

The outcome of gene conversion differs depending on whether it acts on unpaired or paired samples. If acting on an unpaired sample, then it becomes a paired sample with each haplotype carrying a segment of ancestral material, which is a similar outcome to that following a crossover. There are either one or two breakpoints, depending on whether gene conversion lies partly or fully within ancestral material. If acting on paired samples, then a segment of one sample coalesces into the other

(Figure A in the Supplementary File S1).

Before gene conversion is carried through, it is first checked whether it would result in a sample that does not carry any ancestral material. These fully non-ancestral samples can arise if either (i) conversion acts on an unpaired sample, in a region spanning entirely non-ancestral material; or (ii) conversion acts over all remaining ancestral material in a paired sample, rendering it non-ancestral. In case (i) the action stops without creating this 'ghost' sample. In case (ii) the paired sample is converted into an unpaired sample, and the recipient sample is no longer tracked.

Including 'Complete' Gene Conversion

The above routine does not consider a specific class of gene conversion events, where conversion is initiated outside the ancestral tract, and extends completely over the ancestral material. For unpaired samples with obligate sex, such events do not change the genealogy, so they are not considered in Wiuf and Hein (2000). However, if these events act over paired samples, then it would cause complete coalescence of one sample into its opposite pair. Hence these events need to be included in the facultative sex routine.

If there exists $x - k$ pairs after k of them split by sex, then the probability of this event happening equals $2(x-k)(g_{mi}(L-1)e^{-Q_{mi}})/Q_{mi}$ (a full derivation is presented in Appendix B and Section B of Supplementary Mathematica File S2; note the factor of 2 in the equation is due to two haplotypes present within a paired sample). If this event occurs, complete coalescence of one paired sample into an unpaired sample arises, as in the 'gene conversion' event for the single-

site coalescent model (Hartfield *et al.* 2016) (Figure A in Supplementary File S1). Because this event only acts on unpaired samples reproducing asexually, it can only arise via mitotic gene conversion.

Note that this probability goes to infinity as $Q_{mi} \to 0$. In this case, the average gene conversion length is much larger than the genetic sample being simulated (i.e., $\lambda_{mi} \gg L - 1$), so any gene conversion event is likely to affect the entire genetic region. The reason this discontinuity arises is because the coalescent process assumes that no more than one event occurs per generation. Wiuf and Hein (2000) ensures this logic is maintained by assuming $Q \sim \mathcal{O}(1)$. Furthermore, a small Q value would invalidate the assumption used to compute Q^*; specifically, conversion events that initiate outside the sample but end within it only do so in regions near to the genetic sample (since the probability of these events equals e^{-nQ} if initiating $n(L-1)$ breakpoints away from the sample). Hence while the simulation can be run with very small Q values, it is inadvisable to do so as erroneous genealogies may be produced.

Measuring linkage disequilibrium from simulations

We used the updated coalescent simulation to calculate expected linkage disequilibrium in facultative sexual organisms. Following a single simulation of a coalescent process, a series of j genealogies are produced for each non-recombined part of the genetic segment. Polymorphisms are added to each branch of the genealogy, drawn from a Poisson distribution with mean $(1/2)\theta_j\tau_{i,j}$, for $\theta_j = 4N_T\mu(l_j/L)$ the mutation rate of the segment covering l_j of L total sites given μ is the mutation rate for a segment of L sites, and $\tau_{i,j}$ the length of branch i in segment j.

For each simulation, we measured linkage disequilibrium $D = f_{AB} - f_A f_B$ over each pairwise combination of polymorphisms. Here, f_A, f_B are the frequencies of each derived allele, and f_{AB} the frequency of derived alleles at both sites. This measure was then normalised to $r^2 = D^2/(f_A f_B(1 - f_A)(1 - f_B))$ (Hill and Robertson 1968). Once r^2 was measured over all simulations, values were placed into 20 equally-sized bins based on the distance between the two polymorphisms. However, the number of pairwise samples were different for each of the 20 bins. Samples in the last two bins produced noisy estimates of linkage disequilibrium, so we only reported linkage disequilibrium estimates from the first 18 bins. We randomly subsampled data in bins 1 to 17 so that they contained the same number of pairwise comparisons as in the 18th bin. Mean values per bin were recorded for each simulation run. The results presented here plot the mean of means per bin over all $n = 1,000$ simulations, along with a 95% confidence intervals calculated as $\pm 1.96 s/\sqrt{n}$ for s the standard deviation for that bin.

Measuring correlation in coalescence time between sites

For some cases, we measured the correlation in coalescence times between sites, as a function of the distance between them. For each simulation run, we obtained the number of non-recombined regions, and the coalescent time for each over all haplotypes. The coalescent time for the segment upon which each polymorphism resided was subsequently obtained. We calculated the Pearson correlation coefficient in coalescent times for all polymorphisms over all samples; correlations were then placed into one of 20 bins based on the distance between polymorphisms. Values were only reported for the first 18 bins, with further subsampling performed

on bins 1 to 17 so they contained the same number of comparisons as the 18th bin. Mean values per bin were recorded for each simulation; the mean of means was then computed, along with 95% confidence intervals calculated using the same method as for linkage disequilibrium measurements.

Data Availability

This new simulation program, *FacSexCoalescent*, along with documentation is available from `http://github.com/MattHartfield/FacSexCoalescent`. We first rebuilt our single-locus simulation program in C to greatly increase execution speed, before adding the crossover and gene conversion routines. Outputs can either be printed to individual files, or to the screen in a similar manner to `ms` (Hudson 2002) to enable comparisons between programs if desired. As with the previous version of the simulation, *FacSexCoalescent* uses a timescale of $2N$ generations while `ms` uses $4N$ generations. The documentation specifies other cases where *FacSexCoalescent* inputs differ from other coalescent simulations.

Supplementary File S1 contains additional results and figures. Supplementary File S2 is a *Mathematica* notebook of analytical derivations. Supplementary File S3 is a copy of the simulation code and manual.

Results

Testing the simulation

We first tested the facultative sex simulation to ensure that it matched up with existing obligate sex routines and analytical results. First, we calculated the mean

number of non-recombining segments per simulation (measuring the minimum number of crossover events), for different crossover rates assuming obligate sex. We compared these values to those produced using `ms`. Under a wide parameter space, we confirmed that the facultative sex simulation routine produced the same mean number of non-recombining segments as `ms`. Second, we calculated the mean pairwise diversity both between and within individuals, under different rates of sex and gene conversion. We established that expected diversity over samples matches single-locus expectations $\theta E[\tau]$, where θ is the population-level neutral mutation rate and $E[\tau]$ is the expected pairwise coalescent time in units of $2N$ generations. Further information on these tests are provided in Supplementary File S1.

Linkage disequilibrium with crossing over

High rates of sex

We next looked at how patterns of linkage disequilibrium are affected by crossovers, when sexual reproduction is frequent (that is, the scaled rate of sex $N\sigma \gg 1$). Intuition suggests that the effect of meiotic crossovers on linkage disequilibrium is equal to that observed in an obligately sexual population with a rescaled rate $r_{eff} = \sigma r$. To further investigate this pattern, we simulated genealogies over $L = 1,001$ sites with a fixed crossover rate $2Nr = 20$ during sexual reproduction, with results reported over the first 900 sites.

Figure 1(a) plots how linkage disequilibrium decays over this region with different rates of sex, varying from $\sigma = 1$ (i.e., obligate sex) to $\sigma = 0.002$. As expected, the decay in linkage disequilibrium is weakened with lower sex, since there exists fewer opportunities for crossovers to act. We confirm that the observed decay is

equivalent to an obligate sexual population with $r_{eff} = \sigma r$ in two ways. First, we ran equivalent simulations in `ms` but using the rescaled crossover rate, and observed that the decay in linkage disequilibrium overlaps with the facultative sex simulation results (Figure 1(a)). Second, we used the 'pairwise' routine in the *LDhat* software (McVean *et al.* 2002) to estimate crossover rates from simulation data, and observed that they scaled linearly with σ (Figure 1(b)). These results verify that for high rates of sex, the breakdown of disequilibrium over multiple loci does indeed scale with the rate of sex.

With obligate sex, the expected value of r^2 is (Ohta and Kimura 1971; Weir and Hill 1986; McVean 2002):

$$E(r^2) \approx \frac{10 + \rho}{22 + 13\rho + \rho^2} \tag{4}$$

for $\rho = 4N_e r$. Given the above results, we expected that Equation 4 would match mean linkage disequilibrium values if using $\rho = 4N_e r_{eff} = 4N_e \sigma r$. Figure 1(c) compares the fit of simulations (both the facultative sex coalescent with different rates of sex, and `ms` under scaled crossover rates) to Equation 4 after removing sites with minor allele frequency $< 10\%$ (as the derivation of Equation 4 assumes that alleles are not at extreme frequencies; McVean (2002)). For $\sigma = 0.5$ or 1, the match between simulation and analytical results is reasonably good if using a modified version of Equation 4 to condition on the removal of rare variants (see McVean (2002) for details).

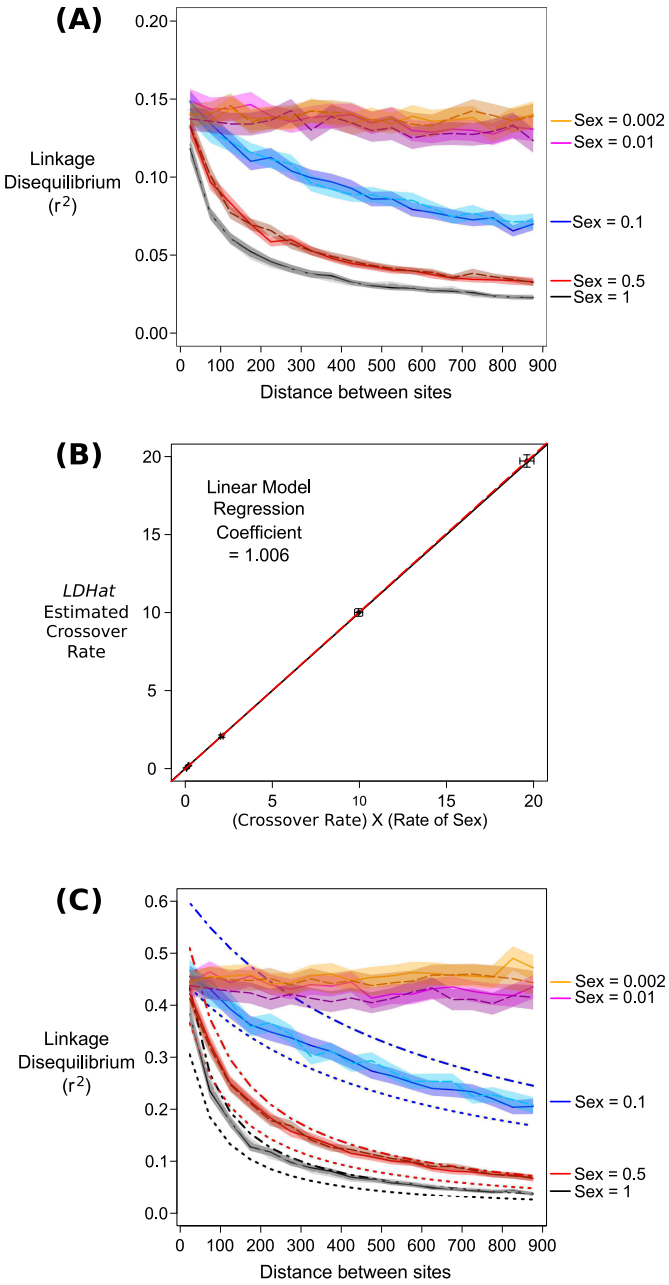

Figure 1: Effects of facultative yet high rates of sex on estimates of meiotic crossing-over. (a) Decay of linkage disequilibrium over 900 sites, as a function of distance between two sites. Different colours denote individual rates of sex, as shown in the legend. Solid line is the mean value over 1,000 simulations; fainter curves represent 95% confidence intervals. 50 paired samples were simulated (100 samples in total), $N = 10,000$, scaled mutation rate $\theta = 4N\mu = 10$, scaled crossover rate during sex $2Nr = 20$. Dashed lines are results from obligate sex simulations ran using `ms` using a crossover rate $2Nr\sigma$. Due to binning of samples, r^2 is shown for distances between 25 to 875 sites apart. (b) Estimates of $2Nr$ using *LDhat*, as a function of the 'effective' crossover rate used in the simulation. Points are mean estimates from 1,000 simulations, bars are 95% confidence intervals. Black line denotes $y = x$; dashed red line is the linear regression fit. (c) As (a) but only plotting simulation results after omitting polymorphisms with minor allele frequency less than 10%. Dotted lines show analytical expectations (Equation 4); dot-dashed lines is Equation 4 conditional on discarding minor alleles (only shown for $\sigma = 1$, 0.5, 0.1). Note the different y-axis scale compared to panel (a).

Low rates of sex

When the rate of sex is low (i.e., $N\sigma \sim \mathcal{O}(1)$), samples will diverge within individuals (Balloux *et al.* 2003; Bengtsson 2003; Ceplitis 2003; Hartfield *et al.* 2016). We examined how this allelic sequence divergence affects linkage disequilibrium by running simulations with low rates of sex ($N\sigma = 10$, 1 and 0.1), but with a fixed scaled crossover rate $2Nr\sigma = 0.05$. Although this scaled crossover rate is low, there is a high crossover rate when sex does occur, hence we expect to see some breakdown of linkage disequilibrium. We run simulations over a larger sample length ($L = 100,001$) so there was enough distance to observe a decay in disequilibrium.

Figure 2 demonstrates the linkage disequilibrium observed in low-sex cases, compared to an obligate sex case with $2Nr = 0.05$. The $N\sigma = 10$ result overlaps with the equivalent obligate sex result. After removing sites with minor allele frequency less than 10%, both simulations exhibit r^2 close to 0.45, which is as given by Equation 4 as the crossover rate goes to zero (Figure D in Supplementary File S1). However, lower rates of sex result in higher baseline values of linkage disequilibrium, indicating that Equation 4 does not properly quantify disequilibrium when sex is low. Elevated r^2 occurs due to allelic sequence divergence creating highly-differentiated haplotypes consisting of polymorphisms at intermediate frequencies ($\sim$50%). Consequently, r^2 is higher over the genomic sample than expected based on obligate sex results using the effective crossover rate σr.

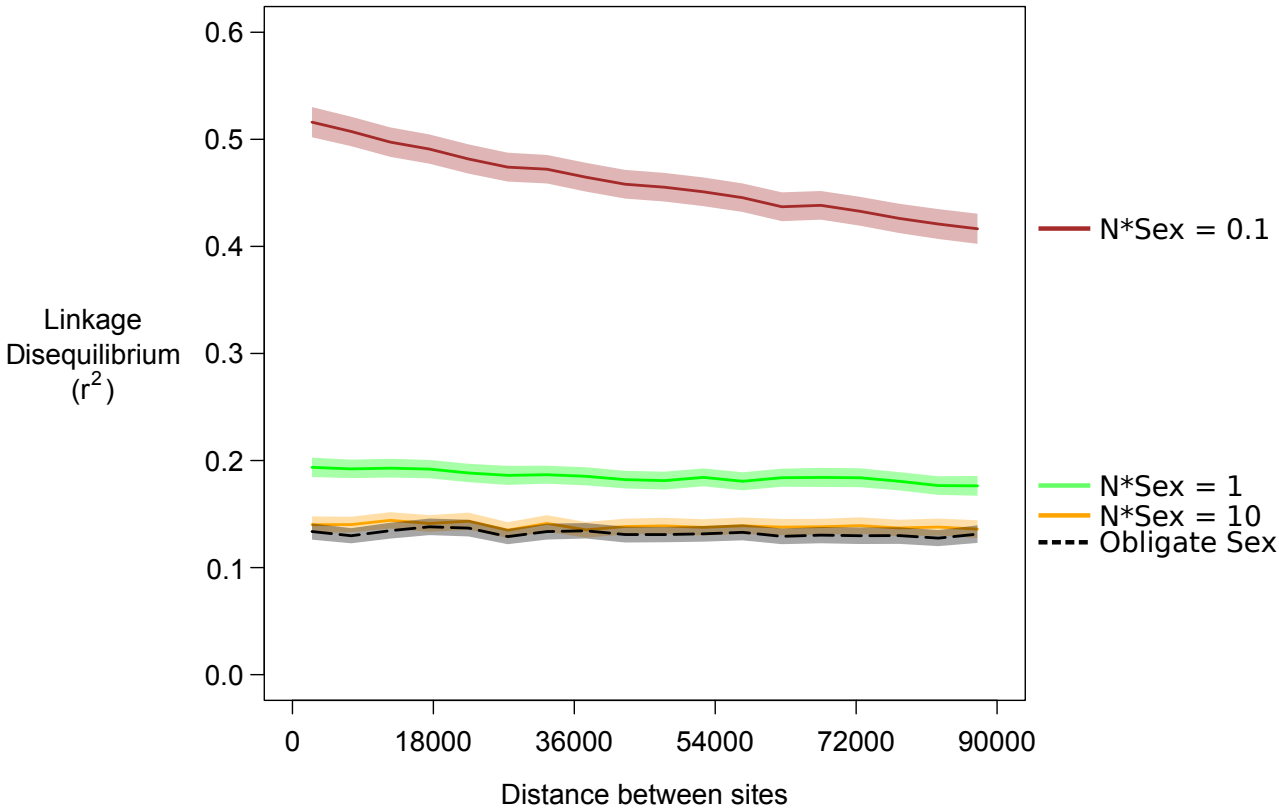

Figure 2: Decay of linkage disequilibrium over 90,000 sites, as a function of distance between two sites. Different colours denote individual rates of sex, as shown in the legend. Solid line is the mean value over 1,000 simulations; fainter curves represent 95% confidence intervals. 50 paired samples were simulated (100 samples in total), $N = 10,000$, scaled mutation rate $\theta = 4N\mu = 10$, scaled crossover rate $2Nr\sigma = 0.05$. Black dashed line is equivalent obligate sex simulation ran using `ms` with 100 haploid samples. r^2 is shown for distances between 2,500 to 87,500 sites apart. Results over short distances (125 to 4,375 sites apart) are shown in Figure E in Supplementary File S1.

Linkage disequilibrium with mitotic gene conversion

High rates of sex

We ran simulations with no meiotic crossovers but with mitotic gene conversion to investigate its effect on linkage disequilibrium. We will define gene conversion using the population-level rate per sample, $\Gamma = 2Ng_{mi}(L-1)$. With mitotic gene conversion only, the decay of linkage disequilibrium is independent of the rate of sex (provided sex is not too low; Figure 3(a)). When crossovers are included with rate $2Nr = 20$, disequilibrium profiles separate out depending on the rate of sex. The pattern of linkage disequilibrium decay is more dependent on the rate of sex when the frequency of mitotic gene conversion is low, relative to the crossover rate r (contrast Figure 3 with Figure F in Supplementary File S1, which uses a lower mitotic gene conversion rate).

Low rates of sex

With low rates of sex ($N\sigma \sim \mathcal{O}(1)$), within-individual diversity is affected by the ratio of sex to gene conversion at a site, denoted ϕ (defined mathematically below). If sex occurs at a much higher rate than gene conversion ($\phi > 1$), then elevated within-individual divergence should be observed. However, if gene conversion arises at the same rate or at a greater rate than sex ($\phi \leq 1$), then gene conversion will be the principal determinant of divergence, leading to reduced within-individual diversity compared to a sexual population (Hartfield *et al.* 2016, Eq. 11). Hence we next ran simulations with different ϕ values to explore the relative effects of both phenomena on linkage disequilibrium.

We considered a diploid population $N = 10,000$ from which we simulated 50

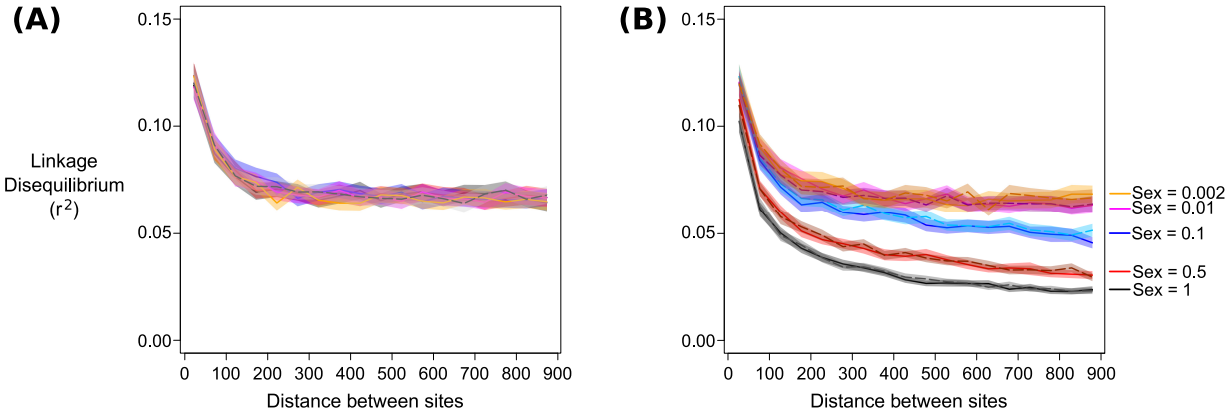

Figure 3: Decay of linkage disequilibrium over 900 sites, as a function of distance between two sites. 50 paired samples are taken from a population of size $N = 10,000$, scaled mutation rate $\theta = 10$, and mitotic gene conversion occurs with rate $\Gamma = 10$ (with average gene conversion tract length $\lambda_{mi} = 100$ sites). Meiotic crossovers are either (a) absent or (b) present at rate $2Nr = 20$. Solid lines are estimates obtained from the facultative sex coalescent simulation with different rates of sex. Colours are as shown in the legend; shaded bands are 95% confidence intervals. Dotted lines are estimates taken from `ms` with 100 samples, crossover rate $r_{eff} = \sigma r$ and same gene conversion rate. Note only one `ms` comparison is plotted in (a). Equivalent results with $\Gamma = 1$ are shown in Figure F in Supplementary File S1. r^2 is shown for distances between 25 to 875 sites apart.

paired samples (100 sampled haplotypes in total); $\theta = 10$; and a genetic segment that is $L = 10,001$ sites long. To focus on the effects of gene conversion, we assumed no meiotic crossing-over and only mitotic gene conversion was considered, with events having a mean tract length of $\lambda_{mi} = 1,000$ sites, matching estimates of non-crossover events obtained from yeast (Judd and Petes 1988; Martini *et al.* 2011). We investigate different rates of sex $\Omega = 2N\sigma$ and gene conversion rates Γ so that the ratio $\phi = (\Omega Q)/2\Gamma$, which determines neutral diversity at a single site, equals either 10, 1 or 0.1. Note that we define the probability of gene conversion per haplotype rather than per diploid genotype, so there is a factor of two in the denominator, compared to the derivation in Hartfield *et al.* (2016).

Fixed gene conversion, varying rate of sex: Here, the rate of gene conversion is fixed at $\Gamma = 10$ and the rate of sex $\Omega = 20$, 2 or 0.2 (since $Q = 10$ with $\lambda = 1,000$, these values produced the required ϕ values of 10, 1 or 0.1). As Figure 4 demonstrates, the baseline rate of linkage disequilibrium increases as the rate of sex decreases from $\Omega = 20$ to 0.2, which is a consequence of allelic sequence divergence creating different genotypes, leading to an increase in linkage disequilibrium. Curiously, disequilibrium patterns reverse at short distances (see also Figure G in Supplementary File S1 for a close-up of r^2 values at less than 1,000 sites).

Fixed rate of sex, varying gene conversion: We next investigated cases where the rate of sex was fixed at $\Omega = 2$, and gene conversion $\Gamma = 1$, 10 or 100 (with mean conversion distance $\lambda_{mi} = 1,000$). Figure 5(a) demonstrates unusual behaviour associated with high gene conversion with low rates of sex. As expected, the baseline linkage disequilibrium value decreases as Γ increases from 1 to 10 as

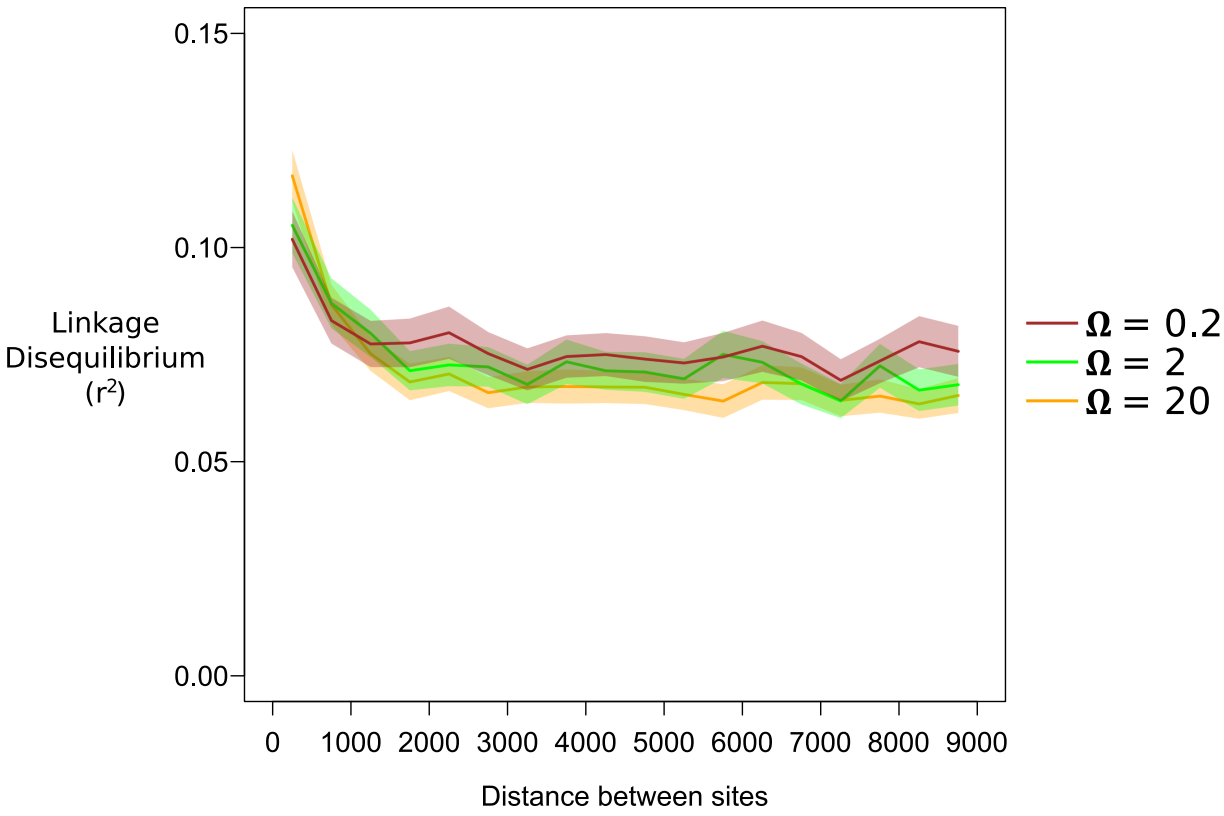

Figure 4: Plot of linkage disequilibrium, measured using r^2, as a function of distance between two sites. Only mitotic gene conversion is present (i.e., no meiotic crossovers or gene exchange), with rate $\Gamma = 10$ and $\lambda_{mi} = 1,000$. The rate of sex equals $\Omega = 20$ (orange line); 2 (green line); or 0.2 (red line). Shading around lines indicate 95% confidence intervals. r^2 is shown for distances between 250 to 8,750 sites apart.

associations between loci are broken down to a greater extent. However, baseline values then increase for $\Gamma = 100$. These results are maintained over short distances between sites (Figure G in Supplementary File S1). To understand this unusual result, we relate the observed r^2 values to the related quantity σ_d^2, which is approximately equal to $E[r^2]$ (Ohta and Kimura 1971; McVean 2002). McVean (2002) demonstrated how σ_d^2 can be written as a function of both the correlation in coalescent times between sites, and the ratio of the mean coalescent time to the variance:

$$\sigma_d^2 = \frac{E[D^2]}{E[f_A f_B (1-f_A)(1-f_B)]} \tag{5}$$

$$= \frac{\rho_{ij,ij} - 2\rho_{ij,ik} + \rho_{ij,kl}}{(E[\tau]^2/Var[\tau]) + \rho_{ij,kl}} \tag{6}$$

The ρ values represent correlation in coalescent times between pairs of sites (e.g., $\rho_{ij,kl}$ the correlation in coalescence times where site one is taken from samples i and j, and site two is taken from samples k and l). $E[\tau]$, $Var[\tau]$ are the mean and variance of coalescent times. Equation 5 shows that σ_d^2 (and thus $E[r^2]$) is reduced with lower correlations between pairs of loci from two samples ($\rho_{ij,ij}$), which can occur with higher rates of crossing over or gene conversion. σ_d^2 also decreases with higher $E[\tau]^2/Var[\tau]$.

Figure 5(b) plots the correlation in coalescent times over all simulations, for two sites sampled from a single individual. We see that they are consistently lower with higher rates of gene conversion, reflecting how genetic material is more frequently transferred between samples. We next looked at the ratio ($E[\tau]^2/Var[\tau]$), which can be calculated from Equations 11 and 12 of Hartfield *et al.* (2016). We focussed

on the within-individual coalescence times, as these are directly affected by within-individual mitotic gene conversion. This ratio is shown in Figure 5(c) for $\Omega = 2$. As Γ increases, the ratio $(E[\tau]^2/Var[\tau])$ decreases, leading to the observed increase in r^2 (Figure 5(a)). This result suggests that high rates of within-individual gene conversion distorts underlying genealogies, so that observed linkage disequilibrium is higher than that expected based on the rate of gene exchange alone. In contrast, meiotic crossing over has no direct effect on this ratio.

Effect of altering gene conversion rate, length for fixed $g_{mi}\lambda_{mi}$

The probability that mitotic gene conversion affects a single site is determined by the product $g_{mi}\lambda_{mi}$. So far we have only presented results for $\lambda_{mi} = 1,000$, hence we next investigated results with varying g_{mi} and λ_{mi} while holding $g_{mi}\lambda_{mi}$ constant. Linkage disequilibrium decays more rapidly for higher g_{mi} values with lower λ_{mi}, holding the product $g_{mi}\lambda_{mi} = 0.00005$ constant (Figure 6). If focussing on short distances (less than 100 sites), we see that r^2 values become equally high ($\sim$0.15; Figure G in Supplementary File S1). Hence although expected diversity at a site depends on the product $g_{mi}\lambda_{mi}$, increasing gene conversion rates breaks down associations faster between loci over mid-range distances. Considering mid-range distance while holding $g_{mi}\lambda_{mi}$ constant, longer tracts (larger λ_{mi}) will be more likely to affect both sites and thus less likely to separate their genealogical histories.

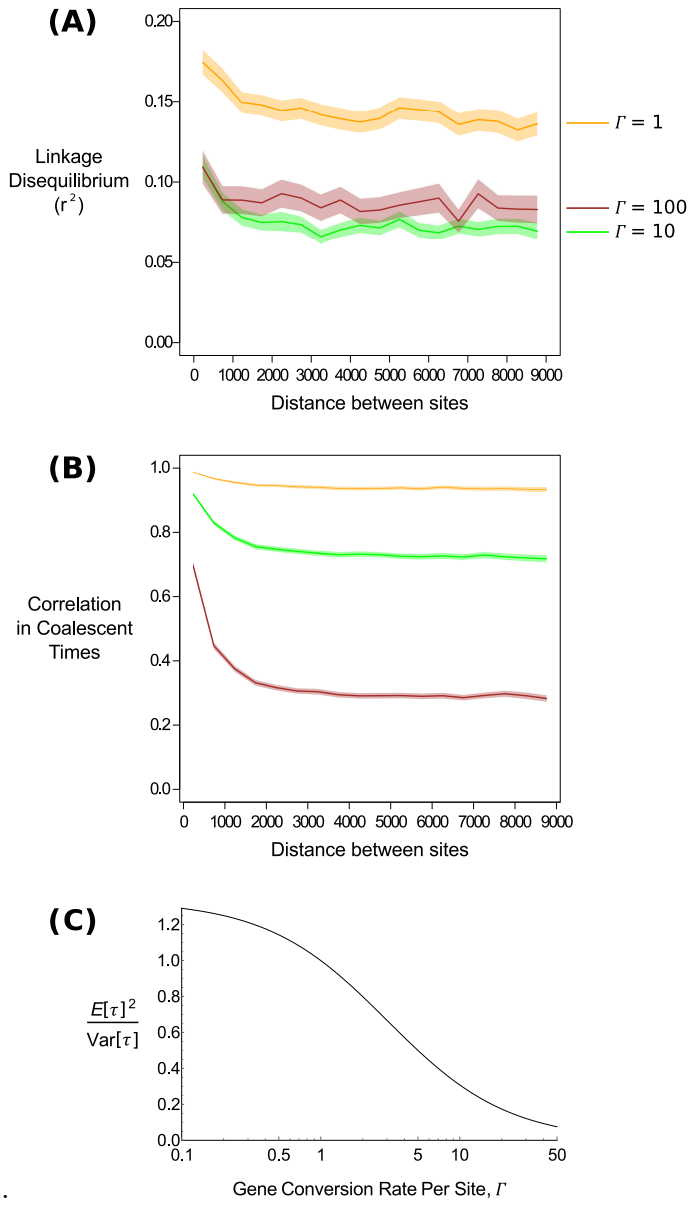

Figure 5: (a) Plot of linkage disequilibrium, measured using r^2, as a function of distance between two sites. For a fixed rate of sex $\Omega = 2$, gene conversion is set to $\Gamma = 1$ (orange line); 10 (green line); or 100 (red line) with $\lambda_{mi} = 1,000$. Shading around lines indicate 95% confidence intervals. (b) Correlation in coalescent times ($\rho_{ij,ij}$) between sites for the three cases. Note that confidence intervals are only slightly thicker than the mean line. (c) Ratio of $(E[\tau]^2/Var[\tau])$ for two samples taken from the same individual, as a function of the scaled gene conversion rate Γ, with $\Omega = 2$. r^2 (a) and the correlation in coalescent time (b) are shown for distances between 250 to 8,750 sites apart.

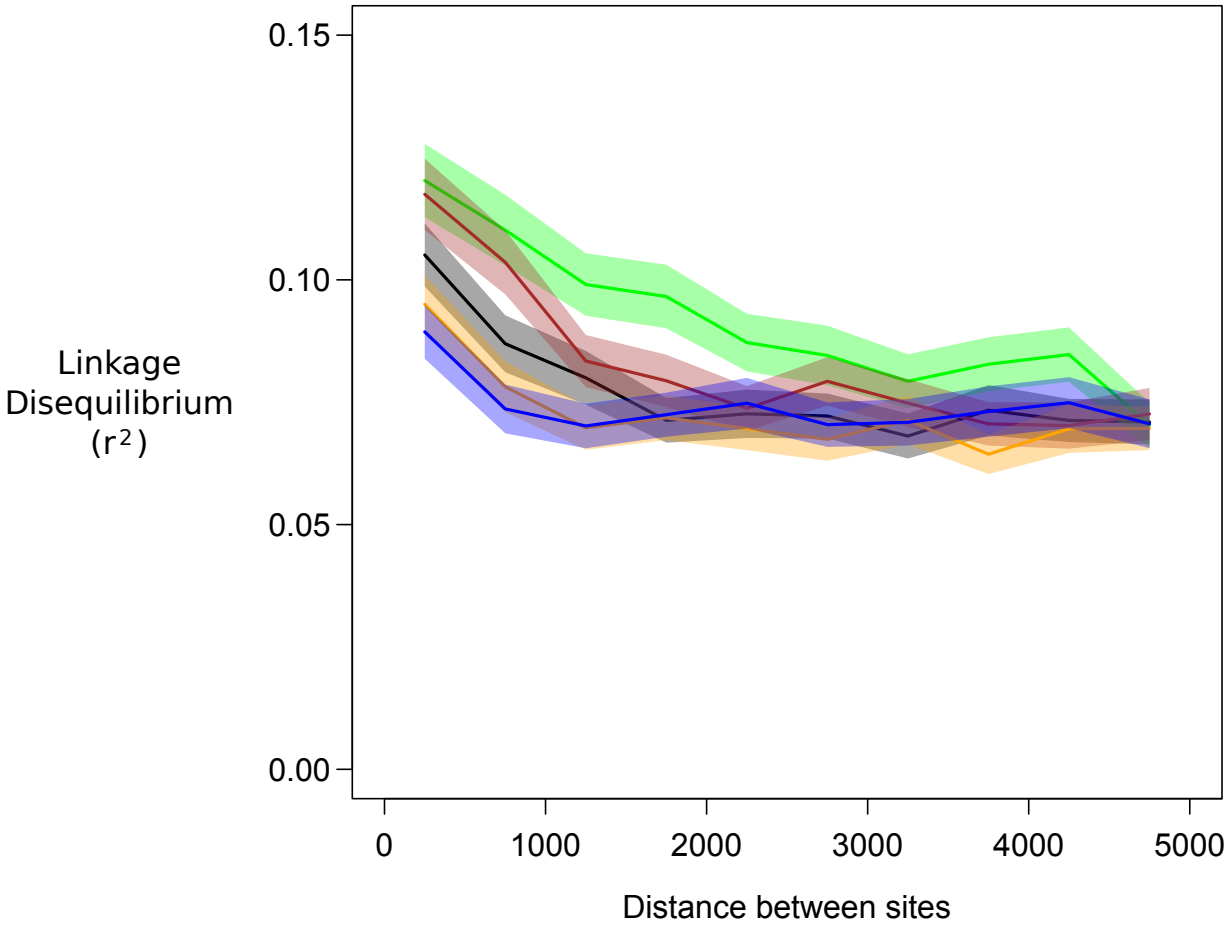

Figure 6: Plot of linkage disequilibrium as a function of distance between two sites. $\Omega = 2$ and $\Gamma = 2.5$, $\lambda_{mi} = 4,000$ (green line); $\Gamma = 5$, $\lambda_{mi} = 2,000$ (red line); $\Gamma = 10$, $\lambda_{mi} = 1,000$ (black line); $\Gamma = 20$, $\lambda_{mi} = 500$ (orange line), or $\Gamma = 40$, $\lambda_{mi} = 250$ (blue line). Colour bands represent 95% confidence intervals. Note that although simulations were performed over a tract of length 10,001 sites, results are only shown for distances between 250 and 4,750 sites apart since curves become indistinguishable at longer distances.

Effect of Population Subdivision

Measurements of linkage disequilibrium are known to increase under population structure with obligate sex (Wakeley and Lessard 2003), as polymorphisms that only appear in specific regions will naturally be in disequilibrium, increasing r^2. Facultative sexual organisms are known to show strong geographic differentiation (Arnaud-Haond *et al.* 2007). Hence we examined the effects of population structure in facultatively sexual organisms. Unless stated otherwise, we assumed an island model, consisting of 4 demes with a scaled migration rate $M = 2N_T m = 1$ between them (for $N_T = 10,000$ the total population size across all demes). 50 paired samples were simulated, with 13 samples taken from two demes, and 12 from the other two.

High rates of sex ($\sigma \gg 1/N_T$)

When sex is not too rare, baseline rates of linkage disequilibrium are higher than those observed in a single-deme population, and the decay in linkage disequilibrium is equivalent to that obtained from an obligate sex population with an effective crossover rate $r_{eff} = \sigma r$. Including gene conversion produces qualitatively similar results to the single-population case, albeit with a higher baseline rate of disequilibrium due to population structure (Figure H in Supplementary File S1).

Low rates of sex ($\sigma \sim 1/N_T$)

Meiotic crossovers only: We ran simulations with $2N_T\sigma r = 0.05$ and $N_T\sigma$ equal to 10, 1 or 0.1, and compared them to an obligate sex case with the same crossover rate with different rates of migration. With high migration ($2N_T m = 10$)

the results are similar to what is observed without population structure, with disequilibrium visually decaying along the genome sample for $N_T\sigma = 0.1$. Yet baselines values are slightly lower than in the panmictic case (compare the red line in Figure 7(a) with Figure 2). With lower migration ($2N_T m = 0.1$), disequilibrium values are unexpectedly reduced as the rate of sex decreases (Figure 7(b)). The reason for this unintuitive result is due to the partitioning of low-frequency polymorphisms under both low sex and population structure. With low migration rates, strong population structure is present so polymorphisms are localised to specific demes. Low rates of sex further partition polymorphisms within demes on diverged haplotypes (Figure 5(e) in Hartfield *et al.* (2016)). Hence the presence of rare sex, alongside high population structure, creates more polymorphisms at lower frequencies compared to populations with higher rates of sex (Figure I in Supplementary File S1). These polymorphisms will not contribute as much to linkage disequilibrium, reducing overall measurements of r^2. After removing polymorphisms with minor allele frequency less than 15%, estimates of r^2 become similar for all rates of sex, although $N_T\sigma = 0.1$ results are still slightly lower than other cases (Figure J in the Supplementary File S1).

With gene conversion included: We also ran island model simulations where either the rate of sex or gene conversion per site was fixed (at values of $N_T\sigma = 1$ and $N_T g(L-1)\lambda_{mi} = 1$ respectively), and set the sex-to-gene-conversion ratio ϕ to either 10, 1 or 0.1. Results are qualitatively similar to the panmictic population case (Figure K in Supplementary File S1). With fixed rates of gene conversion, lower rates of sex increase baseline r^2 values. With fixed rates of sex, high rates of gene conversion also raises r^2. We can understand this behaviour in terms of how the $(E[\tau]^2/Var[\tau])$ ratio affects r^2. In Supplementary File S1

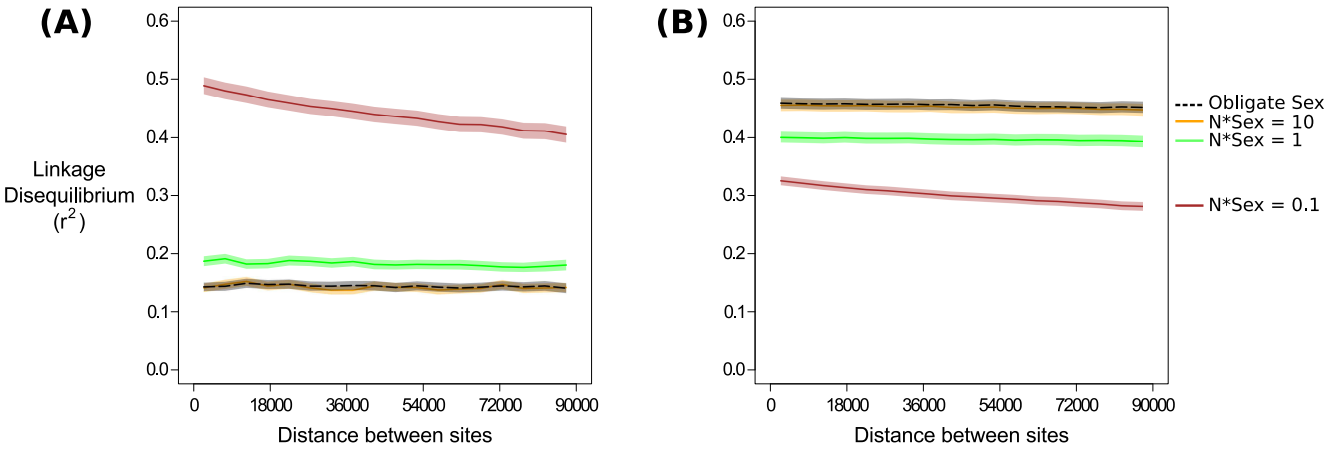

Figure 7: Decay of linkage disequilibrium over 90,000 sites from samples taken over a subdivided population, as a function of distance between two sites. Different colours denote individual rates of sex, as shown in the legend. Solid line is the mean value over 1,000 simulations; fainter curves represent 95% confidence intervals. 50 paired samples were simulated (100 samples in total) over 4 demes, $N_T = 10,000$, scaled mutation rate $\theta = 4N_T\mu = 10$, scaled crossover rate $2N_T r\sigma = 0.05$, scaled migration rate is either (a) $2N_T m = 10$ or (b) $2N_T m = 0.1$. Black dashed line is equivalent obligate sex simulation ran using `ms` with 100 samples. r^2 is shown for distances between 2,500 to 87,500 sites apart.

(see also Section C of Supplementary Mathematica File S2), we derive $Var[\tau]$ when population structure is present. While the mean within-individual coalescence time $E[\tau]$ is unaffected (Hartfield *et al.* 2016), the variance increases with lower migration rates, decreasing $(E[\tau]^2/Var[\tau])$ and causing a more pronounced increase in r^2.

Discussion

Summary of results

Existing single-locus theory for facultative sexual organisms only shows distinct behaviour from obligate sexual cases with extremely low rates of sex ($\sigma \sim \mathcal{O}(1/N)$; Bengtsson (2003); Ceplitis (2003); Hartfield *et al.* (2016)). In this paper we explain how existing crossover (Hudson 1983) and multi-site gene conversion routines (Wiuf and Hein 2000) can be included in facultative sexual coalescent processes, to investigate how they affect gene genealogies and correlations in genetic diversity across loci. We also provide an updated version of a simulation package to simulate ancestral recombination graphs of facultative sexual organisms.

When the rate of sex is high ($\sigma \gg 1/N$), we observe that the breakdown in linkage disequilibrium in a genetic sample is similar to that observed in an obligate sex model that uses an effective crossover rate $r_{eff} = \sigma r$ (Figure 1). This result reflects similar behaviour in partially self-fertilising organisms (Nordborg 2000), where the effective crossover rate is equal to $r(1 - F)$ for F the inbreeding rate (although this scaling breaks down with high self-fertilisation and crossover rates; Padhukasahasram *et al.* (2008); Roze (2009, 2016)).

Hence if there exists knowledge of meiotic crossover rates, then one can use linkage disequilibrium data to estimate the overall rate of sex. The situation becomes more complicated if mitotic gene conversion is present as it also breaks down linkage disequilibrium, even under low rates of sex. If crossover and sex rates are both high then crossovers principally break down linkage disequilibrium, so the effective crossover rate scaling $r_{eff} = \sigma r$ still holds (Figure 3(b)). Conversely, mitotic gene conversion principally affects linkage disequilibrium for low sex or crossover rates (Figure 3).

When rates of sex become low ($\sigma \sim \mathcal{O}(1/N)$), the decay in linkage disequilibrium can no longer be captured by rescaling $r_{eff} = \sigma r$, as the distribution of genealogies becomes fundamentally different than when sex is common (Figure 2). Here, the ratio between rates of sex and gene conversion, ϕ, becomes a strong determinant of linkage disequilibrium. If gene conversion acts at low rates, then decreased rates of sex will lead to a slight increase in overall linkage disequilibrium values, as caused by allelic sequence divergence (Figure 4). Unexpected behaviour arises though if gene conversion occurs at high rates relative to sex. Increasing rates of gene conversion will first reduce overall disequilibrium values due to gene exchange breaking down associations between sites. Yet very high rates of gene conversion then causes elevated linkage disequilibrium values, which is a consequence of how gene conversion changes the distribution of coalescence times (Figure 5). Finally, low sex combined with low migration rates in subdivided populations also creates lower r^2 values, due to more low-frequency polymorphisms being present within demes. These non-intuitive effects illustrate the value of explicitly modelling genetic diversity under low rates of sex when considering genomic data for facultatively sexual organisms.

Routes to inference

The creation of the new coalescent algorithm that accounts for facultative sex, crossing over and gene conversion can be used as a basis for inferring these processes from genomic data. Since linkage disequilibrium statistics are affected by all these factors in a complex manner, the most reliable inference method might be to use coalescent simulations to generate different outcomes over two loci over a wide parameter range, using these outcomes to calculate a likelihood surface. Hudson (2001) used this procedure to calculate expected crossover rates, a modified version of which was subsequently implemented into the *LDhat* software (McVean *et al.* 2002; Auton and McVean 2007). A similar method was also developed for estimating gene conversion (Wall 2004). Our simulation package can be used as a basis for similar inference procedures, especially for estimating rates of sex from genomic data. An alternative approach would be to use Approximate Bayesian Computation to recurrently simulate different outcomes, each time comparing them to the real data and keeping those that match sufficiently well, to build a pseudo-likelihood (Sunnåker *et al.* 2013). While it is guaranteed to produce parameter estimates, ABC procedures are more computationally expensive and inferred parameters may be noisier than those obtained using two-locus probability estimates. Irrespective of the technique used, given how intertwined sex and crossover recombination are then it would be necessary to have an independent estimate of crossover rates in order to infer the prevalence of sex. Simulation results also suggest that it is important to jointly consider both genome-wide diversity and linkage disequilibrium if we wish to infer the effects of sex, meiotic crossovers, and gene conversion. Very low rates of sex increase pairwise diversity due to allelic

sequence divergence, while mitotic gene conversion breaks down linkage disequilibrium and removes diversity (Figure 5). Hence considering both statistics together will prove important in disentangling the effects of meiotic crossovers from mitotic gene conversion.

Future Directions

We anticipate that the *FacSexCoalescent* simulation package can be expanded upon in the future to account for more complex scenarios. The only population structure we considered is an island model, where the population is split into several demes of equal size. More complex demography, such as population bottlenecks or subpopulations of unequal size, has been inferred in many species (Pool and Aquadro 2006; Veeramah and Hammer 2014; Frantz *et al.* 2016) and should be considered in future studies, especially since rates of sex in facultative sexuals tend to exhibit high geographic differentiation (Arnaud-Haond *et al.* 2007).

This theoretical work has elucidated the importance of gene conversion in shaping neutral diversity and disequilibrium. We implemented the model of Wiuf and Hein (2000), where a length of one DNA strand is copied over to its homologue. This model captures the general effect of gene conversion, but empirical studies in yeast have revealed more complex outcomes; for example, where several short tracts are exchanged around breakpoints (Martini *et al.* 2011). Evidence also exists that rates of gene conversion can be context specific. A well-known example is GC-biased gene conversion, where a double-strand break is likely to be repaired using GC-based nucleotides (Duret and Galtier 2009). A recent study of *Drosophila* mutation-accumulation lines has produced evidence that mitotic gene

conversion rates are reduced in lines carrying known deleterious mutations (Sharp and Agrawal 2016). In addition, we ignore the possibility of mitotic crossovers in our simulations. These events occur at a much lower frequency than mitotic gene conversions. If present, they would contribute to a decline of linkage disequilibrium over a longer distance with very low rates of sex, whereas the current model shows that linkage disequilibrium declines only over the length of a gene conversion event. Given ongoing debates on how gene conversion potentially affects genetic diversity and fitness in facultative sexual organisms (Mancera *et al.* 2008; Flot *et al.* 2013; Tucker *et al.* 2013), a deeper understanding of how more complex gene conversion processes affect the distribution of genetic diversity can shed further insight into what processes influence genetic evolution in facultative sexual organisms.

Acknowledgements. MH was supported by a Marie Curie International Outgoing Fellowship, grant number MC-IOF-622936 (project SEXSEL), and an European Research Council grant (FP7/20072013, ERC Grant 311341) awarded to Thomas Bataillon. This work was also supported by Discovery Grants (AFA & SIW) from the Natural Sciences and Engineering Research Council of Canada.

A Drawing start, end breakpoints following gene conversion

When choosing gene conversion breakpoints, they are selected by calculating the cumulative distribution function (CDF) for the event; drawing an initial start or end point from a uniform distribution; then plugging this uniform draw into the inverse CDF to obtain the true start or end point. The CDFs are obtained from the relevant probability distribution functions outlined by Wiuf and Hein (2000). Note that the resulting outputs are continuous variables lying between 0 and 1, while the simulation program assumes discrete breakpoints. Hence after the relevant breakpoint locations have been found, it is then converted into the appropriate discrete value lying between 1 and $L-1$, including these values.

If *two breakpoints are chosen*, then the joint probability distribution of start points s and end points t equals $f(s,t) = (Q\exp(-Q(t-s)))/K(Q)$ (Wiuf and Hein 2000, Eq. 4). By integrating out t from s to 1, one obtains the marginal density of start points, $f(s) = (1-\exp(-Q(1-s)))/K(Q)$ (Wiuf and Hein 2000, caption of Fig. 4). The CDF of s can then be calculated as:

$$F(S) = \int_0^S f(s) = \frac{1 - e^{QS} + e^{Q}QS}{1 + e^{Q}(Q-1)} \tag{A1}$$

To choose a start point, we draw a value between 0 and 1 from a uniform distribution and plug it into $F^{-1}(S)$, which equals:

$$F^{-1}(S) = \frac{e^{-Q}}{Q} \cdot (S - 1 + e^{Q}((Q-1)S - W[-e^{-Q-e^{-Q}(1-S)-(1-Q)S}])) \tag{A2}$$

where W is the Lambert function (Abramowitz and Stegun 1970). To draw the respective end point, we first need to determine the distribution $f(t|s)$; i.e., the density of end points *given* a starting point s. This function is also equal to $f(s,t)/f(s) = (Q\exp(-Q(t-s)))/(1-\exp(-Q(1-s)))$. From this function we obtain the CDF of T given s, as well as the inverse CDF:

$$F(T|s) = \int_s^T f(t|s) = \frac{1-\exp(-Q(T-s))}{1-\exp(-Q(1-s))} \tag{A3}$$

$$F^{-1}(T|s) = s - \frac{\log(1-T(1-\exp(-Q(1-s))))}{Q} \tag{A4}$$

Equation A4 is then used to determine the endpoint of of the gene conversion event, which automatically lies within the length of the genetic sample. If the chosen endpoint is the same as the startpoint, then another endpoint is chosen so that they are distinct.

If *one breakpoint is chosen*, it can be the start or end of gene conversion with equal probability (Wiuf and Hein 2000, Eq. 7). If it is chosen to be the end point of gene conversion initiating outside the tract, then the start point is set to zero (i.e., the far left edge of the tract). The probability density of end points t is $f(t) = \exp(-Qt)/(1-K(Q))$ (Wiuf and Hein 2000, Eq. 8). This function is left-skewed; end points are likely to be near the left-hand side of the genetic sample. The CDF and inverse CDF can be calculated as:

$$F(T) = \frac{\exp(Q) - \exp(Q(1-T))}{\exp(Q)-1} \tag{A5}$$

$$F^{-1}(T) = 1 - \frac{\log(\exp(Q) + T - \exp(QT))}{Q} \tag{A6}$$

If the single breakpoint is instead a start point, then the end point is set to the extreme right side of the sample. The probability density of start points is given by $f(s_1) = \exp(-Q(1-s_1))/(1-K(Q))$ (Wiuf and Hein 2000, Eq. 6). This function is right-skewed; start points are likely to appear towards the end of the sampled genome. The CDF and inverse CDF equal:

$$F(S_1) = \frac{1 - \exp(QS_1)}{1 - \exp(Q)} \tag{A7}$$

$$F^{-1}(S_1) = \frac{\log(1 + (\exp(Q) - 1)S_1)}{Q} \tag{A8}$$

B Deriving probability of 'complete' gene conversion

Let Q_{mi} be the mean scaled length of mitotic gene conversion events. Following Wiuf and Hein (2000), the length of gene conversion events can be drawn from an exponential distribution with parameter Q_{mi}. Let a gene conversion event start at a distance x from the focal sequence (where distances are scaled by the number of breakpoints $L-1$). The gene conversion event will therefore cover the focal sequence with probability $e^{-Q(1+x)}$. The probability of a complete gene conversion occurring over all paired haplotypes (of which there exist $2(x-k)$), and over the entire density of conversion breakpoints (of which there exists $g_{mi}(L-1)$ per length of focal sequence) equals $2(x-k)g_{mi}(L-1)\int_{x=0}^{\infty} e^{-Q(1+x)}$. Solving the integral gives

the full result as provided in the main text.

Literature Cited

Abramowitz, M. and I. Stegun, 1970 *Handbook of Mathematical Functions*. Dover Publications, Inc., New York.

Arnaud-Haond, S., C. M. Duarte, F. Alberto, and E. A. Serrão, 2007 Standardizing methods to address clonality in population studies. Mol. Ecol. **16**: 5115–5139.

Auton, A. and G. McVean, 2007 Recombination rate estimation in the presence of hotspots. Genome Res. **17**: 1219–1227.

Balloux, F., L. Lehmann, and T. de Meeûs, 2003 The population genetics of clonal and partially clonal diploids. Genetics **164**: 1635–1644.

Bengtsson, B. O., 2003 Genetic variation in organisms with sexual and asexual reproduction. J. Evol. Biol. **16**: 189–199.

Brookfield, J. F. Y., 1992 DNA fingerprinting in clonal organisms. Mol. Ecol. **1**: 21–26.

Burt, A., D. A. Carter, G. L. Koenig, T. J. White, and J. W. Taylor, 1996 Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. Proc. Natl. Acad. Sci. USA **93**: 770–773.

Butlin, R., 2002 The costs and benefits of sex: new insights from old asexual lineages. Nat. Rev. Genet. **3**: 311–317.

Ceplitis, A., 2003 Coalescence times and the Meselson effect in asexual eukaryotes. Genet. Res. **82**: 183–190.

Duret, L. and N. Galtier, 2009 Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genom. Human Genet. **10**: 285–311.

Flot, J.-F., B. Hespeels, X. Li, B. Noel, I. Arkhipova, *et al.*, 2013 Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature **500**: 453–457.

Frantz, L. A. F., V. E. Mullin, M. Pionnier-Capitan, O. Lebrasseur, M. Ollivier, *et al.*, 2016 Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science **352**: 1228–1231.

Griffiths, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19**: 169–186.

Haag, C. R., S. J. McTaggart, A. Didier, T. J. Little, and D. Charlesworth, 2009 Nucleotide polymorphism and within-gene recombination in daphnia magna and d. pulex, two cyclical parthenogens. Genetics **182**: 313–323.

Hartfield, M., 2016 Evolutionary genetic consequences of facultative sex and outcrossing. J. Evol. Biol. **29**: 5–22.

Hartfield, M., S. I. Wright, and A. F. Agrawal, 2016 Coalescent times and patterns of genetic diversity in species with facultative sex: Effects of gene conversion, population structure, and heterogeneity. Genetics **202**: 297–312.

Hein, J., M. H. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

Hill, W. G. and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**: 226–231.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23**: 183–201.

Hudson, R. R., 1990 Gene Genealogies and the Coalescent Process. In *Oxford Surveys in Evolutionary Biology*, edited by D. J. Futuyma and J. Antonovics, volume 7, pp. 1–42, Oxford Univ. Press, Oxford.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159**: 1805–1817.

Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

Hudson, R. R. and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of dna sequences. Genetics **111**: 147–164.

Judd, S. R. and T. D. Petes, 1988 Physical lengths of meiotic and mitotic gene conversion tracts in *Saccharomyces Cerevisiae*. Genetics **118**: 401–410.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. App. Prob. **19**: 27–43.

Lynch, M., R. Gutenkunst, M. Ackerman, K. Spitze, Z. Ye, *et al.*, 2017 Population genomics of daphnia pulex. Genetics **206**: 315–332.

Mancera, E., R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz, 2008 High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature **454**: 479–485.

Mark Welch, D. B. and M. Meselson, 2000 Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. Science **288**: 1211–1215.

Martini, E., V. Borde, M. Legendre, S. Audic, B. Regnault, *et al.*, 2011 Genome-Wide Analysis of Heteroduplex DNA in Mismatch Repair-Deficient Yeast Cells Reveals Novel Properties of Meiotic Recombination Pathways. PLoS Genet. **7**: e1002305.

McVean, G. A. T., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162**: 987–991.

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

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146**: 1501–1514.

Nordborg, M., 2000 Linkage disequilibrium, gene trees and selfing: An ancestral recombination graph with partial self-fertilization. Genetics **154**: 923–929.

Nordborg, M. and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146**: 1185–1195.

Nordborg, M. and S. M. Krone, 2002 Separation of time scales and convergence to the coalescent in structured populations. Modern Developments in Theoretical Population Genetics: The Legacy of Gustave Malécot pp. 194–232.

Ohta, T. and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68**: 571–580.

Padhukasahasram, B., P. Marjoram, J. D. Wall, C. D. Bustamante, and M. Nordborg, 2008 Exploring population genetic models with recombination using efficient forward-time simulations. Genetics **178**: 2417–2427.

Pool, J. E. and C. F. Aquadro, 2006 History and structure of Sub-Saharan populations of *Drosophila melanogaster*. Genetics **174**: 915–929.

Rosenberg, N. A. and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3**: 380–390.

Roze, D., 2009 Diploidy, population structure, and the evolution of recombination. Am. Nat. **174**: S79–S94.

Roze, D., 2016 Background selection in partially selfing populations. Genetics **203**: 937–957.

Sharp, N. P. and A. F. Agrawal, 2016 Low genetic quality alters key dimensions of the mutational spectrum. PLoS Biol. **14**: e1002419.

Simon, J.-C., C. Rispe, and P. Sunnucks, 2002 Ecology and evolution of sex in aphids. Trends Ecol. Evol. **17**: 34–39.

Simonsen, K. L. and G. A. Churchill, 1997 A markov chain model of coalescence with recombination. Theor. Popul. Biol. **52**: 43–59.

Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, *et al.*, 2013 Approximate Bayesian Computation. PLoS Comput. Biol. **9**: e1002803.

Tsai, I. J., D. Bensasson, A. Burt, and V. Koufopanou, 2008 Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. Proc. Natl. Acad. Sci. USA **105**: 4957–4962.

Tucker, A. E., M. S. Ackerman, B. D. Eads, S. Xu, and M. Lynch, 2013 Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. Proc. Natl. Acad. Sci. USA **110**: 15740–15745.

Veeramah, K. R. and M. F. Hammer, 2014 The impact of whole-genome sequencing on the reconstruction of human population history. Nat. Rev. Genet. **15**: 149–162.

Wakeley, J., 2009 *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers, Greenwood Village, Colorado.

Wakeley, J. and S. Lessard, 2003 Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. Genetics **164**: 1043–1053.

Wall, J. D., 2004 Estimating recombination rates using three-site likelihoods. Genetics **167**: 1461–1473.

Weir, B. S. and W. G. Hill, 1986 Nonuniform recombination within the human beta-globin gene cluster. Am. J. Hum. Gen. **38**: 776–781.

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