bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Coalescence and linkage disequilibrium in facultatively sexual diploids

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

9th August 2018

[*] 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`

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Under neutrality, linkage disequilibrium (LD) results from physically linked sites having non-independent coalescent histories. In obligately sexual organisms, meiotic recombination is the dominant force separating linked variants from one another, and thus in determining the decay of LD with physical distance. In facultatively sexual diploid organisms that principally reproduce clonally, mechanisms of mitotic exchange are expected to become relatively more important in shaping LD. Here we outline mathematical and computational models of a facultative-sex coalescent process that includes meiotic and mitotic recombination, via both crossovers and gene conversion, to determine how LD is affected with facultative sex. We demonstrate that the degree to which LD is broken down by meiotic recombination simply scales with the probability of sex if it is sufficiently high (much greater than $1/N$ for N the population size). However, with very rare sex (occurring with frequency on the order of $1/N$), mitotic gene conversion plays a particularly important and complicated role because it both breaks down associations between sites and removes within-individual diversity. Strong population structure under rare sex leads to lower average LD values than in panmictic populations, due to the influence of low-frequency polymorphisms created by allelic sequence divergence acting in individual subpopulations. These analyses provide information on how to interpret observed LD patterns in facultative sexuals, and determine what genomic forces are likely to shape them.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.

Previous theory has investigated genetic diversity in facultatively sexual diploid organisms, which reproduce via a mixture of sexual and parthenogenetic reproduction (Brookfield 1992; Burt *et al.* 1996; Balloux *et al.* 2003; Bengtsson 2003; Ceplitis 2003). A general result arising from this work is that when an organism exhibits very rare population-level rates of sex ($\sigma \leq \mathcal{O}(1/N)$, for σ the probability of sex and N the population size), they will exhibit 'allelic sequence divergence' where both alleles within a diploid individual accumulate distinct polymorphisms from each other (Mark Welch and Meselson 2000; Butlin 2002). Hartfield *et al.* (2016) subsequently investigated a coalescent model of facultative sexuals, and quantified how the presence of gene conversion can reduce within-individual diversity to less than that expected in sexual organisms, contrary to the effects of allelic sequence divergence. Hence these results provide a potential explanation as to why allelic divergence is not widely observed in empirical studies of facultatively sexual organisms (reviewed in Hartfield (2016)).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

However, this analysis only modelled the genetic history at a single, non-recombining locus. Here, genealogies only greatly differed from those in obligately sexual organisms at very low frequencies of sex ($\sigma \leq \mathcal{O}(1/N)$). As a consequence, methods to estimate the frequency of sex can only do so based on the degree of allelic sequence divergence, and are expected to be ineffective if the frequency of sex is greater than $1/N$ and/or gene conversion is prevalent (Ceplitis 2003; Hartfield *et al.* 2016). In contrast, many facultatively sexual organisms exhibit much higher occurrences of sex. Pea aphids reproduce sexually about once every ten to twenty generations (Jaquiéry *et al.* 2012), 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 frequency 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 can be used to estimate rates of sexual reproduction in species undergoing more frequent 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 facultative 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 sites through transferring genetic material across DNA

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

strands (Wiuf and Hein 2000).

As meiotic crossing over occurs during sexual reproduction, one may expect that the extent to which linkage disequilibrium is broken down should scale with the probability of sex (see Nordborg (2000) for a related argument for the coalescent with self-fertilisation). Tsai *et al.* (2008) used this logic to calculate the frequency 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 sex if sufficiently high). Lynch *et al.* (2017) used similar arguments to conclude that even though the facultatively sexual water flea *Daphnia pulex* has a lower overall crossover recombination rate than *Drosophila melanogaster*, it has a higher crossover rate when sex does occur.

However, the logic used in these studies assumes that the frequency of sex only affects occurrences of meiotic crossing over. Low rates of sex also distort the underlying genealogies, leading to subsequent events (including allelic sequence divergence or removal of diversity via gene conversion) that also affect how polymorphisms are correlated along haplotypes. Hence these approaches may become problematic in species exhibiting low rates of sexual reproduction, or if gene conversion is an important force in shaping genetic diversity, as observed in empirical studies of facultative sexuals (Crease and Lynch 1991; Schön *et al.* 1998; Normark 1999; Schön and Martens 2003; Flot *et al.* 2013).

We describe both mathematical theory and a routine for simulating multi-site genealogies with facultative sex, allowing for both meiotic and mitotic crossover recombination and gene conversion. We use these new models to investigate how linkage disequilibrium patterns are affected in facultatively sexual organisms, and

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

how these results can be used to infer rates of sex from genome data. Specifically, we investigate when the breakdown of linkage disequilibrium scales with sex, as predicted by intuition, and when this logic does not hold.

Overview of the facultative-sex coalescent and recombination events

Our primary goal is to examine how different frequencies of sex affects linkage disequilibrium. Heuristically speaking, the expected strength of disequilibrium depends on the probability that two sampled haplotypes (hereafter 'samples') coalesce before either haplotype is disrupted by recombination. Before presenting the formal model, we begin by discussing how facultative sex affects coalescence and then recombination.

In the standard coalescent, each member of a set of (non-recombining) samples can be thought of as travelling independently backward in time through the generations. A coalescence event occurs if two samples independently "choose" the same parental allele as their ancestor. The waiting time until the next coalescence depends only the number of remaining samples but, importantly, not on "where" the samples are currently found (i.e., in which individual organisms). However, for diploids with a low frequency of sex, the "where" information is crucial (Bengtsson 2003; Ceplitis 2003; Hartfield *et al.* 2016). For example, two samples can be the two haplotypes found in a single diploid individual (which we denote as 'a paired sample') or they can each come from different individuals (which we denote as 'two unpaired samples'). The two haplotypes within a paired sample do not

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

travel back in time independently, rather they travel together for as long as reproduction is asexual. Coalescence between them is not possible for all the asexual generations they remain paired (ignoring gene conversion). A sexual event splits a paired sample into two unpaired samples that can then coalesce in a subsequent generation. For this reason, paired samples are expected to have longer average coalescence times than unpaired samples and low sex increases average coalescence time compared to high sex (Bengtsson 2003; Ceplitis 2003; Balloux *et al.* 2003). However, if the frequency of mitotic gene conversion is high relative to the frequency of sex, then these predictions are reversed (Hartfield *et al.* 2016). In this case, paired samples can coalesce faster than unpaired samples because each generation the samples are paired provides an opportunity for coalescence via mitotic gene conversion.

In sum, low sex in diploids requires a 'structured' coalescent approach because paired and unpaired samples behave differently; this structure affects the distribution of coalescence times (including the mean and variance) and is sensitive to the amount of mitotic gene conversion. Technically, this structure occurs even in diploids that are obligately sexual; however, the coalescent can be safely modelled ignoring the structure because the time spent in "paired" states is infinitesimally brief on the coalescent timescale when sex is common. To affect coalescence times, sex must be sufficiently uncommon, i.e. $\sigma \leq \mathcal{O}(1/N)$, where N is the population size and σ is the fraction of offspring sexually produced each generation via the random union of gametes (i.e., $\sigma = 1$ represents obligate sex and $\sigma = 0$ obligate asexuality; see Table 1 for a list of symbol definitions).

Different sites along a genetic segment can have different genealogical histories as long as there is some recombination. Low sex affects recombination in several

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Symbol	Usage
N	Diploid population size (with $2N$ haplotypes), denoted N_T if measured across a subdivided population
f_A, f_B	Frequency of derived allele at site A, B
f_{AB}	Frequency of haplotypes carrying derived allele at both sites
D_{AB}	Disequilibrium between two sites, $f_{AB} - f_A f_B$
r^2	Standardised linkage disequilibrium, $D_{AB}/(f_A(1-f_A)f_B(1-f_B))$
r_d^2	'Ratio of means' measure of linkage disequilibrium, $E[D_{AB}]/(E[f_A(1-f_A)f_B(1-f_B)])$
$t_A(ij)$	Coalescent time at site A (if sampled from haplotypes i, j)
σ	Fraction of offspring produced via sex
c, c_A	Probability of meiotic (mitotic) crossing over between two sites
$\tilde{c}$, $\tilde{c}_A$	Probability of meiotic (mitotic) crossing over between two adjacent sites
γ_1, γ_2	Probability of mitotic gene conversion covering one, two sites (analytical model)
γ_{1S}, γ_{2S}	Probability of meiotic gene conversion covering one, two sites (analytical model)
g, g_S	Probability of mitotic (meiotic) gene conversion initiating at a site
Ω	Population-level frequency of sex, $2N\sigma$
ρ, ρ_A	Population-level rate of meiotic (mitotic) crossing over, $4Nc$ ($4Nc_A$)
$\tilde{\rho}_A$	Population-level rate of mitotic crossing over between two adjacent sites, $4N\tilde{c}_A$
Γ_1, Γ_{1S}	Population-level rate of mitotic (meiotic) gene conversion affecting a single site, $4N\gamma_1$ ($4N\gamma_{1S}$)
G	Population-level rate of gene conversion initiation, $4Ng$
λ, λ_S	Average length of mitotic (meiotic) gene conversion event
L	Number of sampled sites; $L-1$ is number of breakpoints
Q, Q_S	Number of breakpoints in units of average gene conversion length (e.g. $Q = (L-1)/\lambda$ for mitotic gene conversion)
R	Population-level meiotic crossing rate (simulation), $4N\tilde{c}(L-1)$
Γ	Population-level mitotic gene conversion rate (simulation), $4Ng(L-1)$
ϕ	Ratio of sex to mitotic gene conversion acting at a single site, $(\Omega Q)/\Gamma$
μ	Mutation rate over L sites
θ	Population level mutation rate, $4N\mu$
m	Probability of migration (island model)
M	Population-level rate of migration, $2N_T m$
p_1, p_2	Probability of one, two gene conversion breakpoints within sample

Table 1: Glossary of Notation.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ways. We consider both crossing-over (the reciprocal exchange of genetic material between two haplotypes) and gene conversion (where genetic material is copied from one haplotype to its homolog). When sexual reproduction is rare, the frequency of meiotic recombination will necessarily be low. Mitotic crossovers and mitotic gene conversion can then become important for two reasons. First, in comparison to meiotic recombination, mitotic recombination becomes a relatively more important route of genetic exchange as meiosis becomes rare. Second, in paired samples (which are only an important consideration when sex is low), mitotic recombination can either lead to gene exchange (the splitting of a multisite sample into separate pieces) or coalescence.

Figure 1 outlines the possible outcomes for recombination under facultative sex. Going back in time, sex involving a meiotic crossover will transform an unpaired sample into a paired sample (i.e., the unpaired sample descended from the two homologs in the parent; Figure 1(a)). For a paired sample, sex segregates the two samples into separate parents, creating two unpaired samples (Figure 1(b)). However, if a crossover also occurred on one of these samples, then the affected sample becomes a paired sample in the parent; the overall outcome is a new paired sample in one parent (each containing a section of ancestral material), and one unpaired sample in the other parent (Figure 1(c)). Mitotic crossovers can also act in paired samples unaffected by sex, swapping genetic material between homologues (Figure 1(d)).

Gene conversion can affect a sample in several ways, where (i) gene conversion initiates outside a tract of ancestral material but finishes within it; (ii) gene conversion begins within a tract of ancestral material but extends beyond it; (iii) both conversion breakpoints lie within ancestral material; or (iv) gene conversion

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

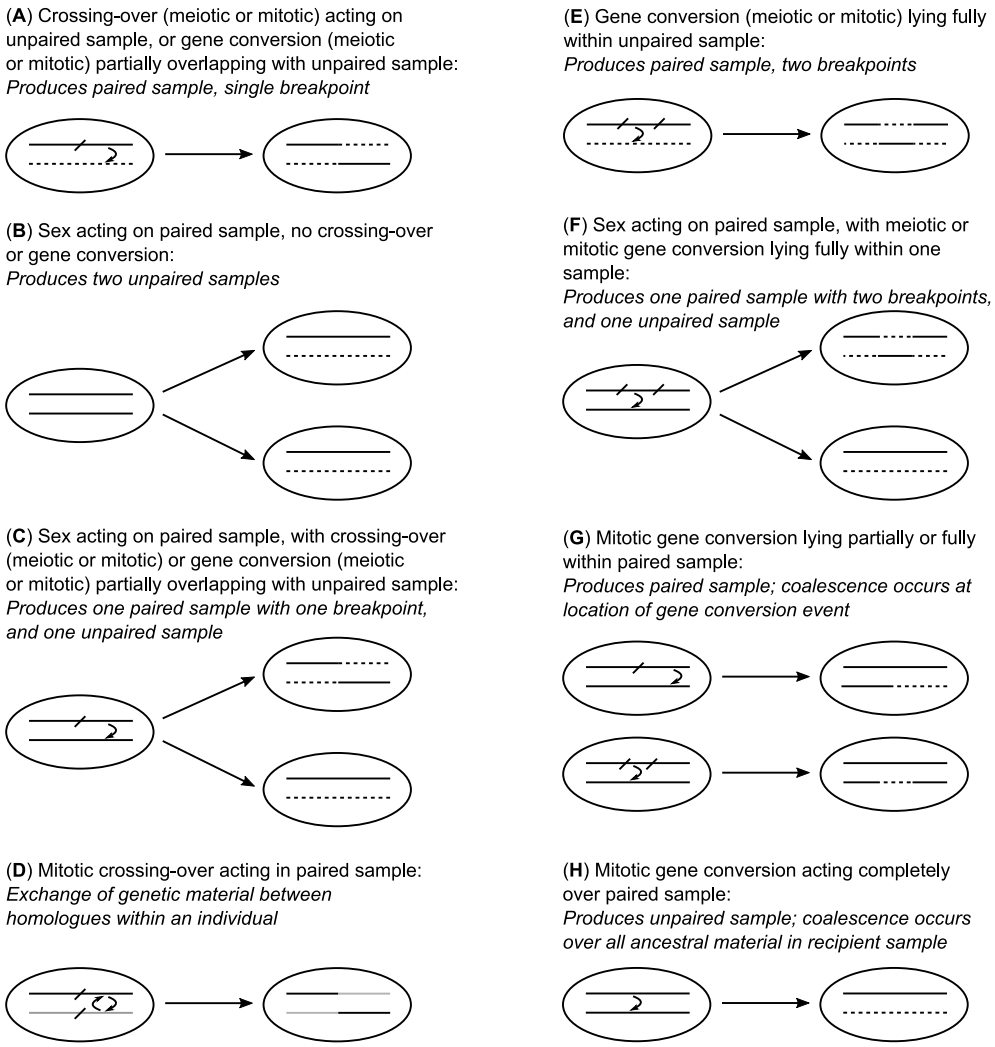

Figure 1: Schematic of different outcomes following gene exchange in the facultative-sex coalescent. Solid lines represent ancestral material; dotted lines represent non-ancestral material. Outcomes are described for (a) meiotic crossing over acting on an unpaired sample, or gene conversion acting on an unpaired sample that only partially overlaps with the haplotype; (b) sex acting on a paired sample with no crossing over or gene conversion; (c) both sex and either crossing over, or gene conversion that only partially overlaps with the haplotype, acting on a paired sample; (d) mitotic crossing over acting on a paired sample; (e) gene conversion (meiotic or mitotic) acting on an unpaired sample, which fully lies within the sampled haplotype; (f) both sex and gene conversion (lying fully within a haplotype) acting on a paired sample; (g) mitotic gene conversion acting on a segment of a paired sample; (h) mitotic gene conversion acting over the entire length of a paired sample (or over all remaining extant material).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

acts over all ancestral material in a paired sample (see Wiuf and Hein (2000) for a detailed discussion of the coalescent with gene conversion applicable to obligately sexual diploids). If gene conversion acts 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 (Figure 1(e)). There are either one or two breakpoints, depending on whether gene conversion lies partly or fully within ancestral material. If acting on paired samples, the outcome depends on whether sex has segregated the samples into different individuals. If so then one of the two parents contains a paired sample with each part carrying ancestral material (Figure 1(f)). If not, then a segment of one sample coalesces into the other (Figure 1(g)). Finally, mitotic gene conversion acting completely over a paired sample reproducing asexually causes complete coalescence of one paired sample, converting it into an unpaired sample. This outcome is equivalent to 'gene conversion' for the single-site coalescent model (Hartfield *et al.* 2016) (Figure 1(h)).

Overall, facultative sex will affect linkage disequilibrium for at least three reasons. First, the population-level rate of meiotic recombination will be proportional to the frequency of sexual reproduction. Second, when sex becomes very rare, the rate and patterns of coalescence change substantially, which is important because disequilibrium is affected by the rate of recombination relative to coalescence. Third, in the low-sex regime, mitotic gene conversion can become important as it becomes a key coalescence mechanism for a paired sample; alternatively, a single haplotype within an individual can be separated (with either one or two breakpoints) via gene conversion.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Two-site analytical model

A commonly used metric of linkage disequilibrium is (Hill and Robertson 1968):

$$r^2 = \frac{D_{AB}^2}{f_A(1-f_A)f_B(1-f_B)} \tag{1}$$

where f_i is the frequency of the derived allele at site i ($i = A$ or B), and $D_{AB} = f_{AB} - f_A f_B$ with f_{AB} being the frequency of haplotypes carrying the derived allele at both sites. A related quantity

$$r_d^2 = \frac{E[D_{AB}^2]}{E[f_A(1-f_A)f_B(1-f_B)]} \tag{2}$$

has been studied in analytical neutral models (Ohta and Kimura 1971; Weir and Hill 1986; McVean 2002) (we use r_d^2 to represent this quantity, rather than the traditional symbol σ_d^2, to avoid confusion with σ that is used to parameterize the frequency of sex in this analysis). r_d^2 overestimates the expected value of r^2 but the discrepancy is reduced if it is only applied to sites where the minor allele is not too rare (McVean 2002). In the classic analyses, which is applicable to obligately sexual diploids:

$$r_d^2 \approx \frac{10 + 2\rho}{22 + 26\rho + 4\rho^2} \tag{3}$$

where $\rho = 4Nc$ with c being the per-generation probability of meiotic crossing over between two sites. McVean (2002) showed that a coalescence approach can be used to derive this result, demonstrating that r_d^2 is a function of the covariance in coalescence times between two sites. The goal here is to quantify how r_d^2 is altered by facultative sex. We use the coalescent approach of McVean (2002) for

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

a two-site model in a diploid population of size N. Two samples at each of two sites in a diploid model can occur in 17 different states, as outlined in Figure 2. In the traditional haploid model, only 7 states are necessary, but here we must consider the full 17-state model to consider the pairing of haplotypes. The model is presented in detail in Section A of Supplementary Mathematica File S1, with an overview provided in Figure 2.

The first key step in constructing the model is to derive the transition matrix giving the probabilities (going backwards in time) of changing states. These probabilities depend on the biology of reproduction and inheritance. If meiosis occurs, there is a crossover between sites A and B with probability c. The probability of a mitotic crossover is c_A per generation (which does not require meiosis). Regardless of reproductive mode, mitotic gene conversion can occur. With probability γ_2 there is a mitotic gene conversion event whose tract length covers both sites. With probability γ_1 a mitotic gene conversion event occurs where one end of the gene conversion tract lies at the breakpoint between the two sites, and the other end lies beyond them (γ_{1S} is the analogous probability for a meiotic gene conversion event, conditional on meiosis occurring). It is worth noting that γ_1 enters the transition matrix for two separate reasons. It determines the probability that one site coalesces via mitotic gene conversion (e.g., transition from state S2 to S14; see Figure 2) and it determines the probability that samples at different sites on the same haplotype get split onto separate haplotypes by mitotic gene conversion (e.g., transition from state S3 to S9; see Figure 2). Note that gene conversion involving one site is functionally equivalent to a crossing-over event. Using these parameters, the construction of the transition matrix is tedious but straightforward. In contrast, γ_2 only enters transitions involving coalescence affecting one

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

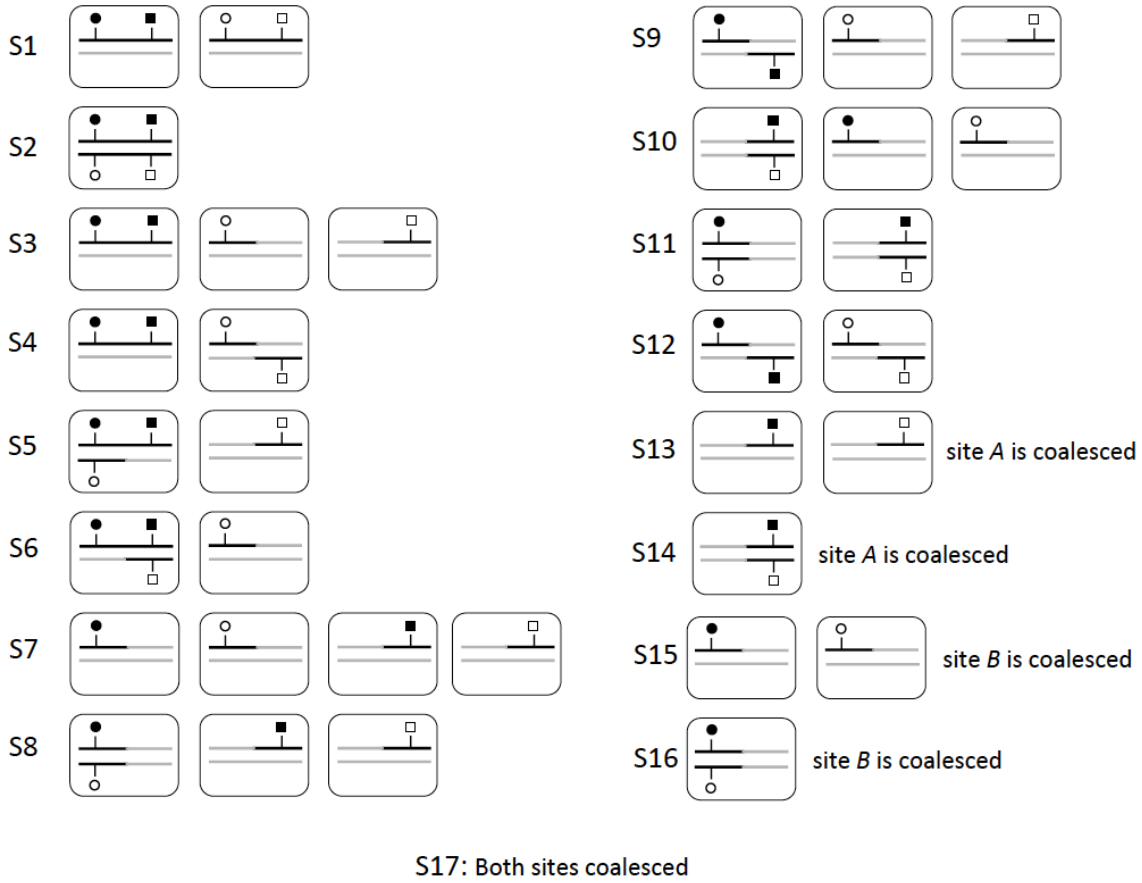

Figure 2: The 17 possible states for two copies of each of two sites, for the analytical model. Each rounded rectangle represents a separate diploid individual. The two focal copies of the A site are represented by open and closed circles. The two focal copies of the B site are represented by open and closed squares. The shading of symbols, i.e., open vs. closed, has no meaning other than to distinguish focal copies. Haplotypes or parts of haplotypes that do not carry ancestral material (i.e., not carrying focal copies) are shown in grey. Coalesced sites are not shown.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

or both sites. Using first-step analysis (Wakeley 2009, Chapter 7) and following McVean (2002), we construct a system of equations for the expected value of the product of coalescent times at the two sites, given their current state z. These equations capture the expected time for the system to move out of the state z, before calculating the expected coalescent time of either one or both sites, given the new state k. These equations have the form:

$$\begin{aligned} E[t_A t_B | Z = z] = E[\tau_z^2] &+ E[\tau_z] \sum_{k \neq z} P_{zk} E[t_A | Z = k] \\ &+ E[\tau_z] \sum_{k \neq z} P_{zk} E[t_B | Z = k] \\ &+ \sum_{k \neq z} P_{zk} E[t_A t_B | Z = k] \end{aligned} \tag{4}$$

where τ_z is the time to exit state z, P_{zk} is the probability that the system moves from state z to state k conditional on leaving state z, and $E[t_x | Z = k]$ is the expected time to coalescence of site x given it is currently in state k. As described in Section A of Supplementary Mathematica File S1, these components can be calculated from the transition matrix, either directly for discrete time or after appropriate transformation for the continuous time approximation (Möhle 1998; Wakeley 2009).

Following McVean (2002):

$$r_d^2 = \frac{E[t_{A(ij)} t_{B(ij)}] - 2E[t_{A(ij)} t_{B(ik)}] + E[t_{A(ij)} t_{B(kl)}]}{E[t_{A(ij)} t_{B(kl)}]} \tag{5}$$

where $E[t_{A(ij)} t_{B(kl)}]$ is the expected product of the coalescent times at site A, where the two copies are sampled from haplotypes i and j, and at site B, where

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the two copies are sampled from haplotypes k and l (and different indexes denote other haplotype samples). Analogous to measuring r_d^2 from haploids where each haplotype represents an independent sample, we calculate r_d^2 assuming that each haplotype comes from a different individual (Figure 2) so that the three terms in the numerator represent coalescence times from states S1, S3, and S7.

We first consider the case of partial asexuality where sex may be rare at the individual level but is not too rare at the population level (i.e., $0 < \sigma \leq 1$ but $N\sigma \gg 1$). We find

$$r_d^2 = \frac{10 + \Psi}{(2 + \Psi)(11 + \Psi)} \tag{6}$$

where $\Psi = \rho\sigma + \rho_A + (1/2)\Gamma_1 + (1/2)\sigma\Gamma_{1S}$ with scaled parameters $\rho = 4Nc$, $\rho_A = 4Nc_A$, $\Gamma_1 = 4N\gamma_1$, and $\Gamma_{1S} = 4N\gamma_{1S}$. Simplifying the model by ignoring gene conversion and mitotic crossing over ($\Gamma_1 = \Gamma_{1S} = \rho_A = 0$), the result above is the same as the obligate sex result (Equation 3) but using an effective scaled crossover rate $\rho\sigma$ in place of ρ (Figure 3(a)).

We next consider the case where sex is rare at the population level, $2N\sigma \to \Omega$ as $N \to \infty$. In the absence of mitotic gene conversion or mitotic crossing over ($\Gamma_1 = \rho_A = 0$) then:

$$r_d^2 = \frac{1728 + 3960\Omega + 3870\Omega^2 + 2091\Omega^3 + 634\Omega^4 + 95\Omega^5 + 5\Omega^6}{1728 + 4824\Omega + 5958\Omega^2 + 3927\Omega^3 + 1342\Omega^4 + 209\Omega^5 + 11\Omega^6} + \mathcal{O}(\xi) \tag{7}$$

where the rates of disruptive meiotic processes c, $\gamma_{1S} \sim \mathcal{O}(\xi)$ with ξ being a small term ($|\xi| \ll 1$). Equations 6 and 7 differ in several important ways. First their maximum values, and the conditions to achieve these maxima, differ. The max-

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

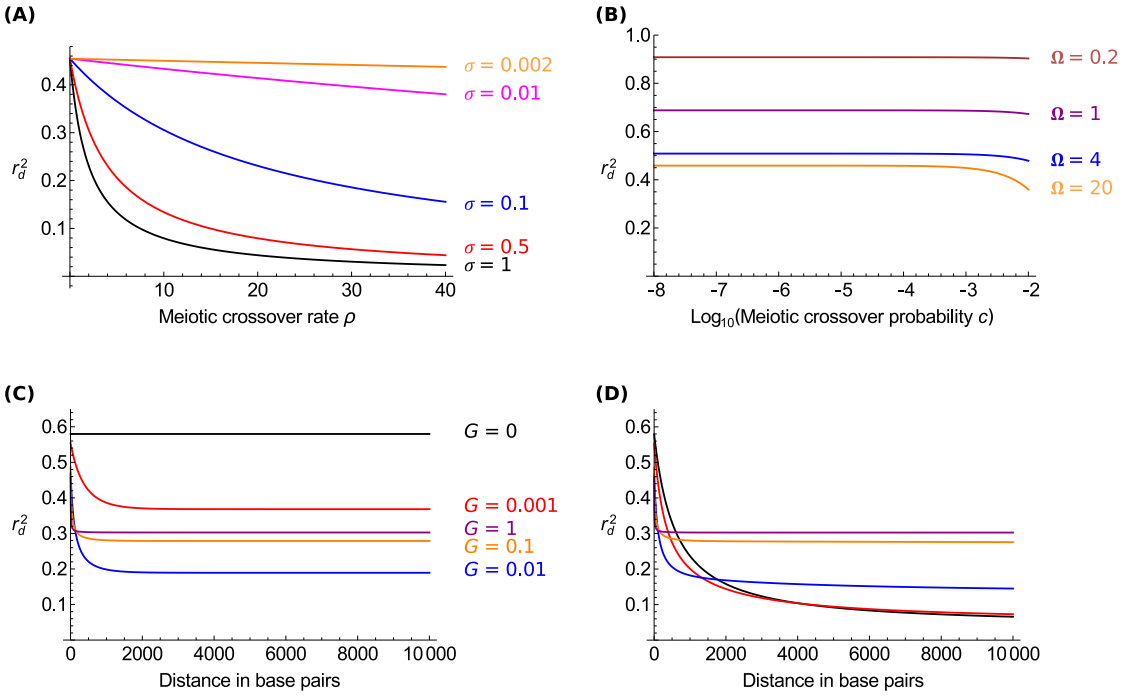

Figure 3: (a) Linkage disequilibrium, measured as r_d^2, when sex is high at the population level (i.e., $\sigma \gg \mathcal{O}(1/N)$), measured as a function of the meiotic crossover rate ρ. Different frequencies of sex (σ) are shown. Other parameters: $c_A = \gamma_1 = \gamma_2 = \gamma_{1S} = 0$. (b) r_d^2 when the frequency of sex is low ($\sigma \sim \mathcal{O}(1/N)$) and the only haplotype disrupting force is meiotic crossing over ($c > 0$). Different rates of sex ($\Omega = 2N\sigma$) are shown. Other parameters: $c_A = \gamma_1 = \gamma_2 = \gamma_{1S} = 0$. (c, d) r_d^2 when the frequency of sex is low ($\sigma \sim \mathcal{O}(1/N)$) as a function of the distance between two sites (measured in basepair distance), for different levels of mitotic gene conversion ($G = 4Ng$). Results are shown without (c) and with (d) mitotic crossing-over (i.e., $\tilde{\rho}_A = 4N\tilde{c}_A$ with $\tilde{\rho}_A = 0$ in (c) and $\tilde{\rho}_A = 0.002$ in (d)). Other parameters: $\lambda = 500$; $\Omega = 2$ (see Figure A in Supplementary File S2 for similar plots using different Ω values).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

imum value of Equation 6 occurs as the haplotype disrupting forces (crossovers and gene conversion) become small, i.e., $r_d^2 \to 10/22 \approx 0.45$ as $\Psi \to 0$. In contrast, the maximum value of Equation 7 occurs as sex becomes increasingly rare, i.e., $r_d^2 \to 1$ as $\Omega \to 0$. Second, r_d^2 in Equation 6 has a strong dependence on physical distance because disruption via crossover or gene conversion is an increasing function of distance, i.e., Ψ is implicitly an increasing function of distance. In contrast Equation 7 is very weakly dependent on physical distance through terms of $\mathcal{O}(\xi)$ (Figure 3(b)).

Equation 7 assumes no mitotic gene conversion or mitotic crossing over but important changes to r_d^2 occur with either of these processes. An analytical approximation can be obtained but the expression is unwieldy (Section A of Supplementary Mathematica File S1). Both types of mitotic gene conversion events, represented via $\Gamma_1 = 4N\gamma_1$ and $\Gamma_2 = 4N\gamma_2$, as well as mitotic crossing over ($\rho_A = 4Nc_A$) affect the leading-order term for r_d^2 and are functions of the distance between sites. Mitotic crossing over can be modelled as a linear function of distance d, $c_A(d) = \tilde{c}_A d$. Using a standard assumption of exponentially distributed gene conversion tract lengths (Wiuf and Hein 2000), the probabilities of mitotic gene conversion are given by $\gamma_1(d) = 2g\lambda(1 - \exp(-d/\lambda))$ and $\gamma_2(d) = g\lambda \exp(-d/\lambda)$ where λ is the average tract length and g is the probability of gene conversion initiation per base pair (more precisely, per breakpoint between adjacent base pairs). The derivation is provided in Section A of Supplementary Mathematica File S1.

Figure 3(c) shows that r_d^2 declines with physical distance when there is mitotic gene conversion but no mitotic crossing over. Note that r_d^2 does not decline down to 0 with distance as it does in the classic model (Equation 3) of meiotic crossing over. Because gene conversion probabilities change slowly for $d/\lambda > 2$, there is

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

little decline in r_d^2 beyond this point. Surprisingly, r_d^2 is not always a monotonically declining function of the probability of gene conversion initiation, g (or the scaled parameter $G = 4Ng$), especially when $d > \lambda$ (Figure B(a) in Supplementary File S2). Consequently, a species with a lower frequency of gene conversion events (i.e., smaller g) can have larger r_d^2 for small d but smaller r_d^2 for large d compared to an otherwise similar species with larger g (Figure 3(c)). This behaviour of r_d^2 with respect to g is likely due to the dual (and conflicting) roles of gene conversion in increasing both the probability of coalescence and disruption of haplotypes. In contrast, mitotic crossing over, which only affects haplotype disruption, affects r_d^2 monotonically as expected (Figure B(b) in Supplementary File S2). The addition of mitotic crossing over reduces the minimum value of r_d^2 (Figure 3(d)). Even with mitotic crossing over, r_d^2 does not go to zero at large distances and can be considerably greater than zero when gene conversion is high (see Section A of Supplementary Mathematica File S1). The minimum value reached by r_d^2 is independent of the rate of mitotic crossing over (provided it is not zero) though the distance at which the minimum is reached is shorter with higher rates of mitotic crossing over.

Simulation Algorithm

We have previously developed an algorithm to build genealogies of facultative sexual organisms at a single non-recombining locus (Hartfield *et al.* 2016). This algorithm simulates genealogies of n samples, of which $2x$ are paired and the remaining $y = n - 2x$ samples are unpaired. The algorithm proceeds in a similar manner as other coalescent simulations, in that it tracks the genetic histories of

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

samples into the past, sequentially enacting events that affect the genetic history (e.g. coalescence, sexual reproduction). The relative probability of each event occurring per generation is used to determine what the next event is, and at which time in the past it arises. To further investigate the effects of facultative sex on linkage disequilibrium, we extended this previous routine to consider coalescent histories of multiple sites, and how various recombination phenomena affect how genetic histories are correlated along chromosomes. In Appendix A we describe how the crossover routine of Hudson (1983) and the gene conversion routine of Wiuf and Hein (2000) are extended to consider the effects of facultative sex. As a consequence, the updated coalescent simulation now models the effects of meiotic and mitotic recombination on facultative sex, the outcomes of which are summarised in Figure 1.

Measuring linkage disequilibrium from simulations

We used the updated coalescent simulation to calculate expected linkage disequilibrium in facultatively sexual organisms. Following a single simulation of a coalescent process, a series of j genealogies are produced, one 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; this measure was then normalised to $r^2 = D^2/(f_A f_B(1 - f_A)(1 - f_B))$. Once r^2 was measured over all simulations,

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 18 so that they include the same number of pairwise comparisons as in the smallest bin that contained data, to standardise bin size per simulation. Mean values per bin were recorded for each simulation run. We then calculated the mean of means per bin over all 1,000 simulations, omitting points where data was not present in a bin for a simulation. 95% confidence intervals were calculated as $\pm\, \tilde{t}(s/\sqrt{n})$ for s the standard deviation for the bin, n the number of points in the bin (maximum of $1,000$, one for each simulation run), and $\tilde{t}$ the 97.5% quantile for a t-distribution with $n-1$ degrees of freedom.

Measuring correlation in coalescence time between sites

For some cases with low sex and mitotic gene conversion, we measured the correlation in coalescence times between sites as a function of the distance between them, to investigate how these values relate to observed linkage disequilibrium patterns. For each simulation run, we obtained the number of non-recombined regions, and the times at which the ancestral segment of that region for each individual coalesced. If more than 100 segments existed, these were subsampled down to 100. We calculated the Pearson correlation in coalescent times for all segments; values were then placed into one of 20 bins based on the distance between blocks (the location of each segment was given by its midpoint). Values were only reported

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for the first 18 bins, with further subsampling performed on bins 1 to 18 so they contained the same number of comparisons as the smallest bin that contained data for that simulation. The mean bin value for a simulation, as well as the mean of means over all simulations and 95% confidence intervals were 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 the single-locus simulation program in C to greatly increase execution speed, before adding the crossover and gene conversion routines. 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 and outputs differ from other coalescent simulations. We performed various tests of the simulation as described in Section B of Supplementary File S2.

Supplementary File S1 is a *Mathematica* notebook of analytical derivations. Supplementary File S2 contains additional results and figures. Supplementary File S3 is a copy of the simulation code and manual. Supplemental files have also been uploaded to FigShare.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Simulation Results

Linkage disequilibrium with crossing over

High frequencies of sex

We looked at how patterns of linkage disequilibrium are affected by crossovers, when sexual reproduction is frequent (that is, the scaled rate of sex $N\sigma \gg 1$). Analytical results (Equation 6) suggest that the effect of meiotic crossovers on linkage disequilibrium is equal to that observed in an obligately sexual population with a rescaled probability $c_{eff} = c\sigma$ (for $c = \tilde{c}d$ the crossover probability over distance d). To further investigate this pattern, we simulated genealogies over $L = 1,001$ sites with a fixed population-level meiotic crossover rate over the entire ancestral segment $R = 4N\tilde{c}(L-1) = 40$, which acts during sexual reproduction. Results are reported over the first 900 sites.

Figure 4(a) plots how linkage disequilibrium decays over this region with different probabilities 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 (compare Figure 4(a) to analytical expectations in Figure 3(a)). We confirm that the observed decay is equivalent to an obligately sexual population with $c_{eff} = c\sigma$ in three ways. First, we ran equivalent (but haploid and sexual) simulations in `ms` using the rescaled crossover probability, and observed that the decay in linkage disequilibrium matches results from the facultative-sex simulation (Figure 4(b)). Second, we used the 'pairwise' routine in the *LDhat* software (McVean *et al.* 2002) to estimate crossover rates from facultative-sex simulation data, and observed that they scaled linearly with

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

σ (Figure 4(c)). Finally, Figure 4(d) plots linkage disequilibrium values for all facultative-sex coalescent simulations as a function of the effective recombination rate, alongside the analytical expectation for r_d^2 (Equation 6). r^2 is calculated after removing sites with minor allele frequency $< 10\%$, as r_d^2 is known to overestimate r^2 if all allele frequencies are considered. We see that the decay in linkage disequilibrium over all simulations is close to, but slightly overshoots the theoretical expectation (Equation 6). Similar behaviour was observed by McVean (2002, Figure 3; compare solid line to square points).

Low frequencies of sex

When the probability of sex is low (i.e., $N\sigma \leq \mathcal{O}(1)$), 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 $N\sigma \sim \mathcal{O}(1)$ (specifically we investigated $2N\sigma = \Omega = 20$, 2 and 0.2), but with a fixed scaled crossover rate $4N\tilde{c}(L-1)\sigma = R\sigma = 0.1$. 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 along the simulated genotype. We ran simulations over a larger number of sites ($L = 100,001$) so there was enough distance to observe a decay in disequilibrium.

Figure 5(a) displays the linkage disequilibrium observed in low-sex cases, along with analytical expectations (Equation 7). After removing sites with minor allele frequency less than 10%, the $\Omega = 20$ simulation exhibits r^2 close to 0.45, which is as given by Equation 6 as the crossover rate goes to zero. However, lower rates of sex result in higher values of linkage disequilibrium, indicating that classic estimates of r^2 using the rescaled recombination rate $R\sigma$ do not properly quantify

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

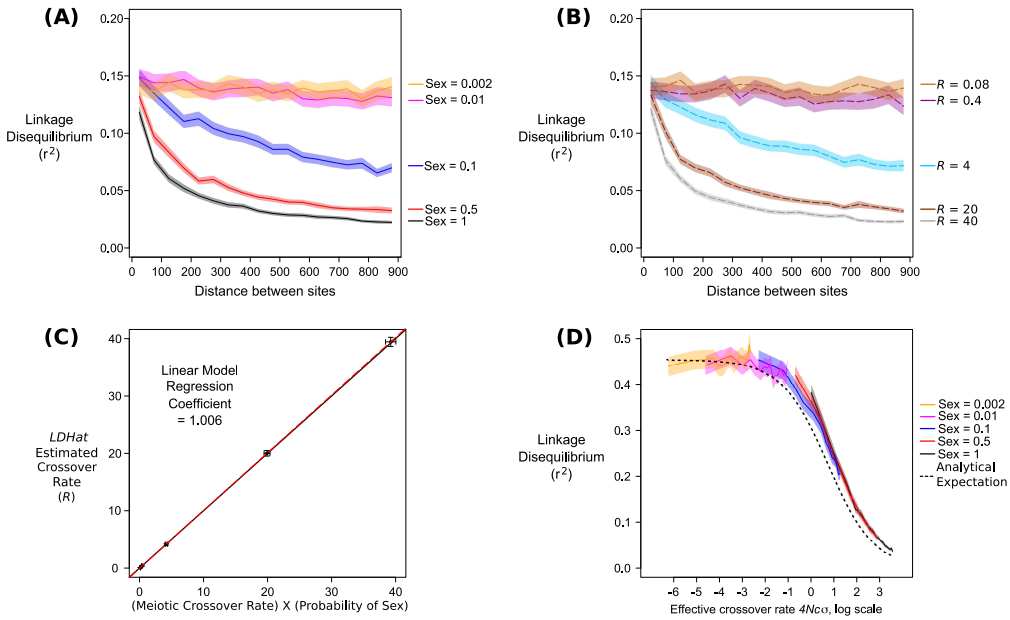

Figure 4: Effects of facultative, but not very low rates of sex (i.e. $\sigma \gg 1/N$) on estimates of meiotic crossing over. (a) Decay of linkage disequilibrium over 900 sites, as a function of distance between two sites. Different colours denote individual rates of sex, as shown in the legend. Solid 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 $R = 40$. (b) As (a) but instead shows results from obligate sex simulations ran using `ms`, using a crossover rate equal to 40σ as shown in the legend. Due to binning of samples, r^2 is shown for distances between 25 to 875 sites apart in (a) and (b). (c) Estimates of R using *LDhat*, as a function of the 'effective' crossover rate used in the facultative-sex coalescent 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. (d) Plot of all simulation results in (a) but instead as a function of the rescaled recombination rate $4Nc\sigma$ (plotted on a natural log scale), and after omitting polymorphisms with minor allele frequency less than 10%. Dotted lines show analytical expectations (Equation 6). Note the different y-axis scale compared to panel (a).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

disequilibrium when sex is low. Equation 7 captures the general behaviour of r^2 under low frequencies of sex (i.e., elevated r^2 values and a weaker dependence on the meiotic crossover rate) but there are several reasons why the results do not quantitatively match. Specifically, the analytical result is based on r_d^2 rather than r^2, and finite sample sizes also introduce additional complications ignored in the calculation of expected r_d^2. Our analytical model is intended to allow comparisons of the main patterns with the comparable sexual model, rather than providing precise predictions of the quantity as estimated by empiricists (which can instead be calculated using the simulation).

Elevated r^2 occurs under low sex due to allelic sequence divergence creating highly-differentiated haplotypes consisting of polymorphisms at intermediate frequencies ($\sim$50%). These polymorphisms arise due to a lack of genetic segregation creating highly differentiated haplotypes (Balloux *et al.* 2003; Bengtsson 2003; Ceplitis 2003; Hartfield *et al.* 2016). Figure 5(b) shows the density of minor allele frequencies over all simulation data, demonstrating that the $\Omega = 0.2$ case has many sites with minor allele frequency between 45 and 50%. Consequently, r^2 is higher over the genomic sample than expected based on obligate sex results using the effective crossover probability $c\sigma$. We also observe that linkage disequilibrium decay is only weakly affected by the meiotic crossover frequency, in line with analytical expectations (Equation 7).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

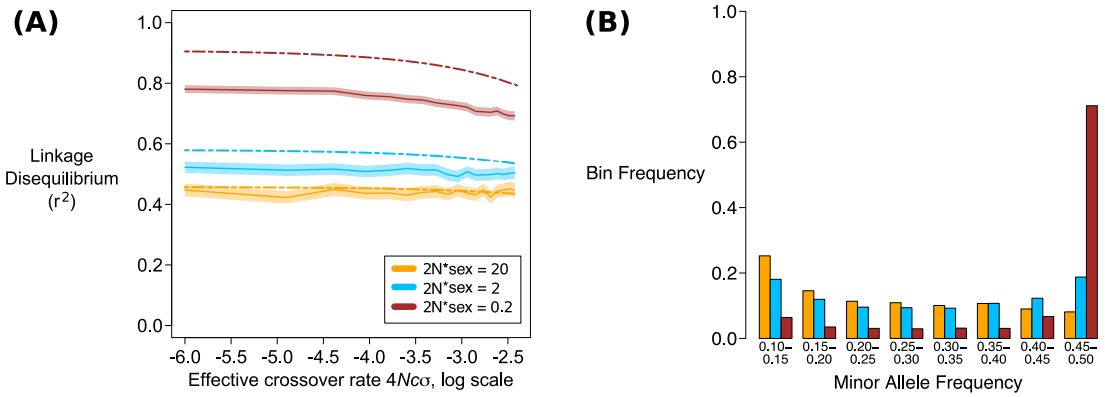

Figure 5: (a) Decay of linkage disequilibrium over 90,000 sites, as a function of the rescaled recombination rate $4Nc\sigma$ (on a natural log scale), and after omitting polymorphisms with minor allele frequency less than 10%. Different colours denote individual rates of sex, as shown in the legend. 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 over all 100,001 sites $R\sigma = 0.1$. Coloured dash-dotted lines are low-sex analytical expectations for r_d^2 (Equation 7). r^2 is shown for distances between 2,500 to 87,500 sites apart. Results for short distances (125 to 4,375 sites apart) are shown in Figure E in Supplementary File S2. (b) Histogram of minor allele frequencies for the low-sex scenarios; the bin frequency is measured over all 1,000 simulations. Bar colours correspond to the same rates of sex as used in panel (a).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Linkage disequilibrium with mitotic gene conversion

High frequencies of sex

We ran simulations with mitotic gene conversion to investigate its effect on linkage disequilibrium. We define gene conversion using the population-level rate per sample, $\Gamma = 4Ng(L-1)$. We first ran simulations with no meiotic crossovers (i.e., some degree of sexual reproduction occurs, but not any meiosis-related processes); here, the decay of linkage disequilibrium is independent of the rate of sex (provided sex is not too low; Figure 6(a)). This decay is similar to that observed in obligate sexual populations experiencing the same gene conversion rate (Figure 6(b)). When meiotic crossovers are included with rate $R = 40$, disequilibrium profiles separate out depending on the frequency of sex (Figure 6(c)), and are similar to those arising in obligate sexuals that experience the same gene conversion rate and an effective crossover probability $c_{eff} = c\sigma$ (Figure 6(d)). The pattern of linkage disequilibrium decay is more dependent on the probability of sex when the frequency of mitotic gene conversion is low, relative to the crossover probability c (contrast Figure 6 that uses $\Gamma = 20$, with Figure F in Supplementary File S2 that uses $\Gamma = 2$).

Low frequencies of sex

With low frequencies of sex ($N\sigma \leq \mathcal{O}(1)$), within-individual diversity is affected by the ratio of sex to gene conversion at a site, denoted ϕ (Hartfield *et al.* (2016); ϕ is defined mathematically below). If sex occurs more frequently than gene conversion ($\phi > 1$), elevated within-individual diversity should be observed. However, if gene conversion arises at the same frequency, or more often than sex ($\phi \leq 1$), then gene

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

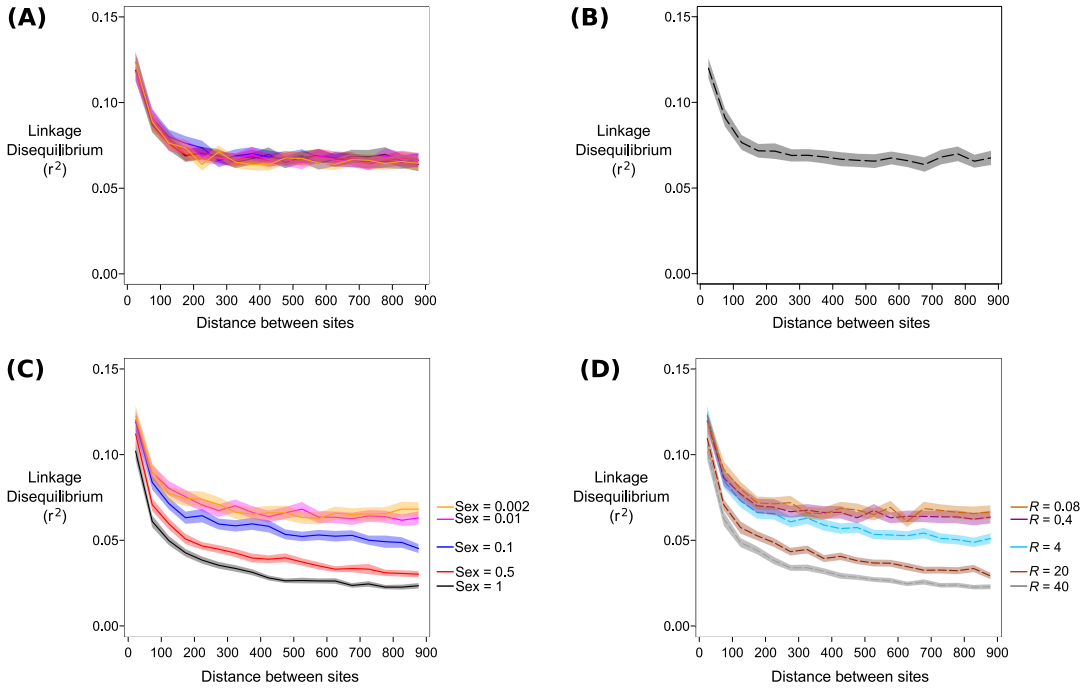

Figure 6: Decay of linkage disequilibrium over 900 sites, as a function of distance between two sites as caused by mitotic gene conversion with high rates of sex ($\sigma \gg 1/N$). 50 paired samples are taken from a population of size $N = 10,000$, scaled mutation rate $\theta = 10$, and mitotic gene conversion occurs with rate $\Gamma = 20$ (with average gene conversion tract length $\lambda = 100$ sites). Meiotic crossovers are either (a, b) absent or (c, d) present at rate $R = 40$. (a, c) Results from the facultative-sex coalescent simulation with different probabilities of sex. Colours are as shown in the legend; shaded bands are 95% confidence intervals. (b, d) Results from `ms` with 100 samples and the same gene conversion rate, with crossover probability $c_{eff} = c\sigma$ in (d). Note only one `ms` comparison is plotted in (b). Equivalent results with $\Gamma = 2$ are shown in Figure F in Supplementary File S2. r^2 is shown for distances between 25 to 875 sites apart.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

conversion will lead 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 paired samples; $\theta = 10$; 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 $\lambda = 1,000$ sites, matching estimates of non-crossover events obtained from yeast (Judd and Petes 1988; Martini *et al.* 2011). We fixed $\Omega = 2N\sigma = 2$ and varied Γ so that the ratio $\phi = (\Omega Q)/\Gamma$ (for $Q = (L-1)/\lambda$ the number of breakpoints in units of mean gene conversion length), which determines neutral diversity at a single site, equals either 10, 1 or 0.1 (requiring $\Gamma = 2$, 20 and 200 respectively). Note that we define the probability of gene conversion per haplotype rather than per diploid genotype, so the probability of gene conversion is scaled by $4N$ as opposed to the $2N$ scaling used in Hartfield *et al.* (2016) (i.e., there is an extra factor of 2 in the denominator of ϕ to account for two haplotypes per individual).

Figure 7(a) demonstrates the unusual behaviour associated with high gene conversion with low rates of sex, with r^2 being a non-monotonic function of the gene conversion frequency, in line with analytical findings (Figure 3(c)). Mitotic crossing over at rate $\rho_A = 4N\tilde{c}_A(L-1) = 10$ breaks down linkage disequilibrium over longer distance for $\Gamma = 2$ and 20, but not for $\Gamma = 200$ (Figure 7(b)). Both results are in line with analytical findings: r^2 is a non-monotonic function of gene conversion when sex is rare (Figure 3(c)), and the presence of mitotic recombination can

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

reduce long-distance r^2, unless mitotic gene conversion acts at a much higher rate (Figure 3(d)).

Elevated linkage disequilibrium is likely related to the reduced mean coalescence times that arise under frequent gene conversion. To further understand this behaviour, we can relate the observed r^2 values to Equation 11 of McVean (2002), which demonstrated how r_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:

$$r_d^2 = \frac{Corr[ij, ij] - 2Corr[ij, ik] + Corr[ij, kl]}{(E[\tau]^2/Var[\tau]) + Corr[ij, kl]} \tag{8}$$

$Corr$ in Equation 8 represent correlation in coalescent times between pairs of sites (e.g., $Corr[ij, kl]$ the correlation in coalescence times where site one is taken from haplotype i and j, and site two is taken from haplotype k and l). $E[\tau]$ and $Var[\tau]$ are the mean and variance of coalescent times. Equation 8 shows that r_d^2 is not just reduced with lower covariances between pairs of loci, but it also decreases with higher $E[\tau]^2/Var[\tau]$. This ratio equals one under the standard coalescent, but low sex alters the mean and variance of coalescent times (Hartfield *et al.* 2016) which will also affect this ratio, and subsequently alter linkage disequilibrium values. Figure 7(c) plots the covariance 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

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

by within-individual mitotic gene conversion. This ratio is shown in Figure 7(d) for $\Omega = 2$ as a function of the mitotic gene conversion rate for a single site, Γ_1. As Γ_1 increases, the ratio $(E[\tau]^2/Var[\tau])$ decreases, leading to the observed increase in r^2 (Figure 7(d)). 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.

In Supplementary File S2 we investigate how linkage disequilibrium is affected if we alter g and λ while fixing the product $g\lambda$. Linkage disequilibrium decays more rapidly for higher g values with lower λ as there are more gene conversion events that break apart coalescent histories between individual sites.

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. Facultatively 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. We assumed an island model, consisting of 4 demes with a scaled migration rate $M = 2N_T m$ 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. Population-scaled parameters are subsequently defined relative to N_T (i.e., $R = 4N_T\tilde{c}(L-1)$, $\Omega = 2N_T\sigma$, $\Gamma = 4N_T g(L-1)$).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

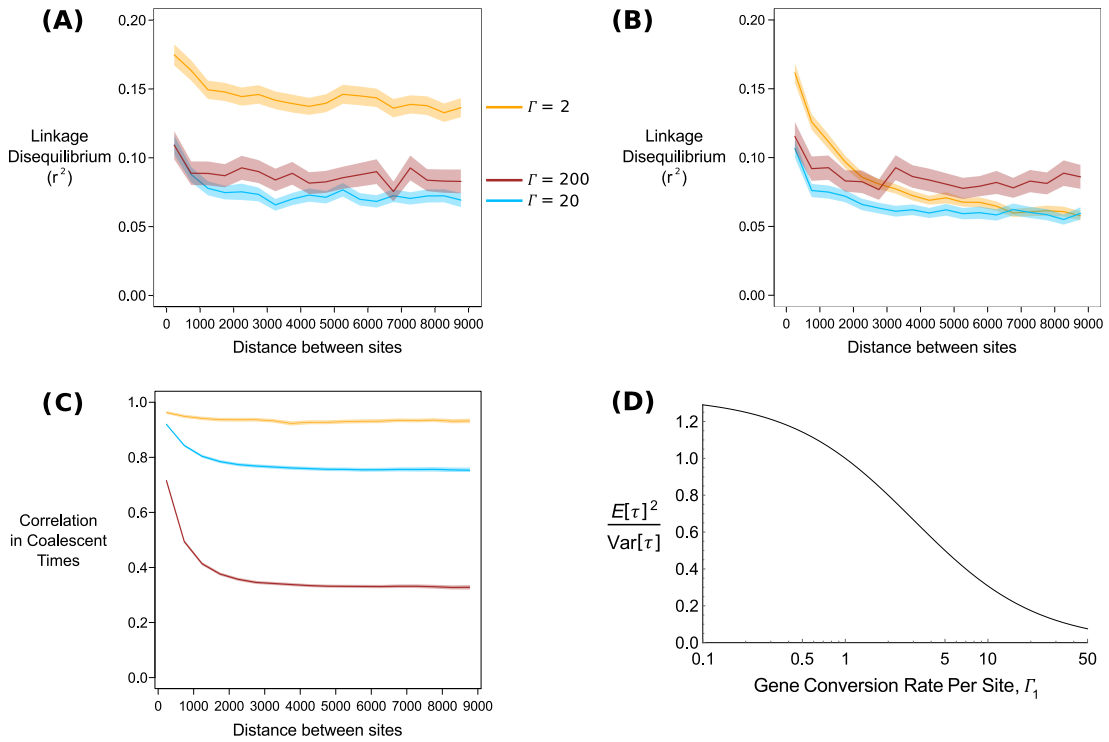

Figure 7: (a) Plot of linkage disequilibrium, measured using r^2, as a function of distance between two sites. For a fixed rate of sex $\Omega = 2$, gene conversion is set to $\Gamma = 2$ (orange line); 20 (blue line); or 200 (red line) with $\lambda = 1,000$. Shading around lines indicate 95% confidence intervals. (b) As (a) but also including mitotic recombination with rate $\rho_A = 4N(c_A) = 10$. Results over short distances for (a) and (b) (25 to 875 sites apart) are presented in Figure G in Supplementary File S2. (c) Correlation in coalescent times ($Corr[ij, ij]$ in Equation 8) between sites for the three Γ values, assuming no mitotic crossing over (i.e., $\rho_A = 0$). Note that for $\Gamma = 20$ and 200, confidence intervals are only slightly thicker than the mean line. (d) Ratio of $(E[\tau]^2/Var[\tau])$ for two samples taken from the same individual, as a function of the scaled gene conversion rate per site Γ_1, with $\Omega = 2$. (a)–(c) are shown for distances between 250 to 8,750 sites apart.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For high sex cases ($\sigma \gg 1/N_T$), and low sex cases ($N_T\sigma \leq \mathcal{O}(1)$) where mitotic gene conversion is present, results are qualitatively similar to those observed for a single population (Supplementary File S2). For the low-sex case with meiotic crossing-over, we ran simulations with $R\sigma = 0.1$ and Ω equal to 20, 2 or 0.2, and compared them to an obligate sex case with the same crossover rate with different rates of migration. With high migration ($M = 10$) the results are similar to what is observed without population structure, with disequilibrium visually decaying along the genome sample for $\Omega = 0.2$. Yet values are lower than in the panmictic case (compare the red line in Figure 8(a) with Figure 5). With lower migration ($M = 0.1$), disequilibrium values are unexpectedly reduced as the probability of sex decreases (Figure 8(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 frequencies 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 probabilities of sex (Figure 8(c, d)). These polymorphisms tend to have small values for r^2, thereby reducing the average value. After removing polymorphisms with minor allele frequency less than 15%, estimates of r^2 become similar for all rates of sex, although $\Omega = 0.2$ results are still slightly lower than other cases for $M = 0.1$ (Figure J in Supplementary File S2).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

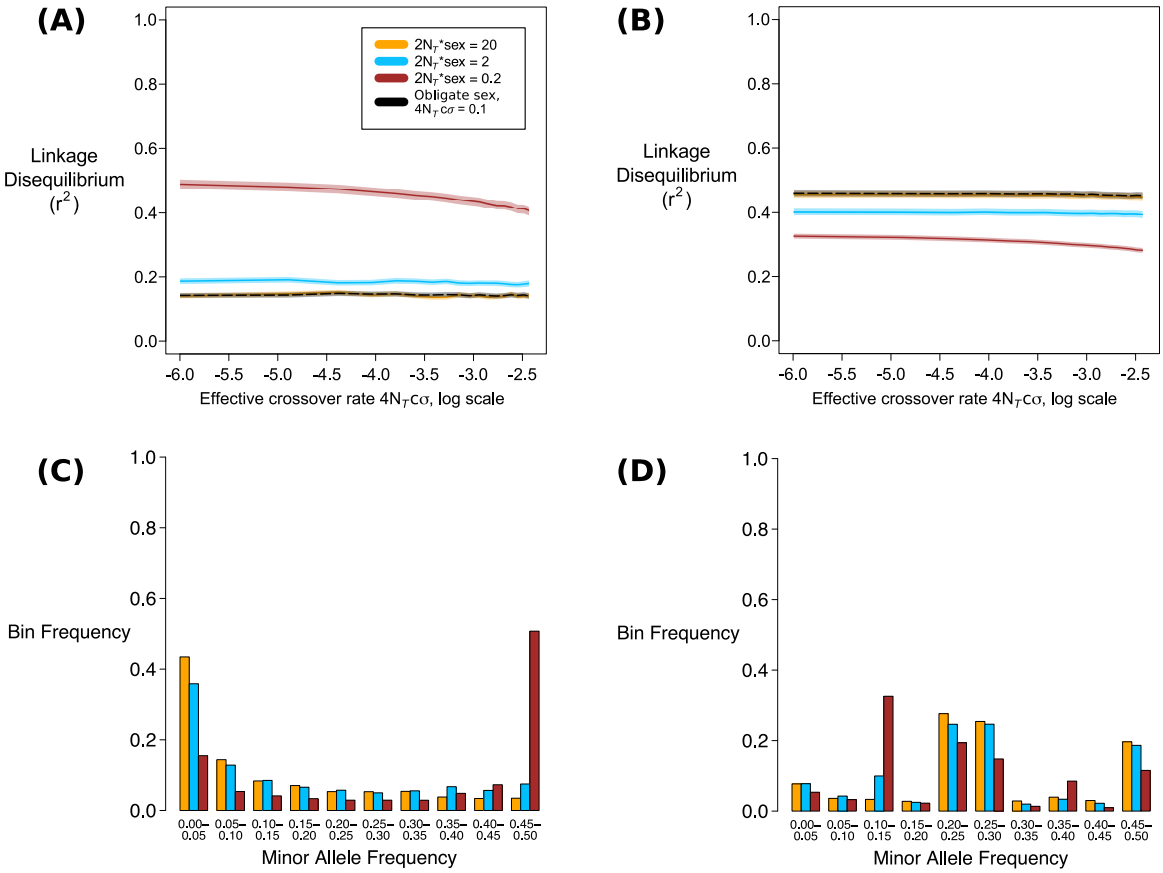

Figure 8: (a, b) Decay of linkage disequilibrium over 90,000 sites from samples taken over a subdivided population, as a function of the rescaled recombination rate $4N_T c\sigma$ (plotted on a natural log scale). Different colours denote individual rates of sex, as shown in the legend. Solid line is 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 over entire ancestral tract $R\sigma = 0.1$, scaled migration rate is either (a) $M = 10$ or (b) $M = 0.1$. 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. (c, d) Histogram of minor allele frequencies, with the bin frequency measured over all 1,000 simulations. Bar colours correspond to the same rates of sex as used in panels (a, b).

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Discussion

Summary of results

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

When the frequency of sex is high ($N\sigma \gg 1$), we observe that the breakdown in linkage disequilibrium in a genetic sample is similar to that observed in an obligate sex model using an effective crossover probability $c_{eff} = c\sigma$ (Figure 3(a), Figure 4). 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 frequency of sex. The situation becomes more complicated if mitotic recombination is present as it also breaks down linkage disequilibrium, even under low frequencies of sex. If sex is frequent but crossing over is rare, mitotic gene conversion principally affects linkage dis-

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

equilibrium (Figure 6(a)). Once crossing-over probabilities become high then these principally break down linkage disequilibrium, so the effective crossover rate scaling $c_{eff} = c\sigma$ still holds (Figure 6(b)).

When rates of sex become low ($\sigma \leq \mathcal{O}(1/N)$), the decay in linkage disequilibrium can no longer be captured by rescaling $c_{eff} = c\sigma$, as the distribution of genealogies becomes fundamentally different than when sex is common (Figure 3(b-d)). In the absence of gene conversion, r^2 becomes elevated with low rates of sex, reflecting more linked polymorphisms present at intermediate frequencies (Figure 5). If mitotic gene conversion is present, the ratio between rates of sex and gene conversion ϕ becomes a strong determinant of linkage disequilibrium, with unexpected behaviour arising if gene conversion occurs at high rates relative to sex. Increasing 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 cause elevated linkage disequilibrium values, which is a consequence of how gene conversion changes the distribution of coalescence times. Adding mitotic crossovers reduces the minimum observed linkage disequilibrium, unless mitotic gene conversion occurs at much higher rates (Figures 3(c, d) and 7). Finally, low sex combined with low migration rates in subdivided populations also reduces r^2 values, due to more low-frequency polymorphisms being present within demes (Figure 8). 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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Future Directions

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. This can be achieved by using coalescent simulations to create likelihood profiles over two loci (Hudson 2001; McVean *et al.* 2002; Wall 2004; Auton and McVean 2007). 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). 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, especially if mitotic gene conversion is pervasive (Figure 7).

We anticipate that the *FacSexCoalescent* simulation can be expanded upon in the future to account for more complex scenarios. The only population structure we considered is an island model, whereas bottlenecks or unequally-sized subpopulations are common (Pool and Aquadro 2006; Veeramah and Hammer 2014; Frantz *et al.* 2016). The gene conversion model can also be expanded upon to consider context-dependent events (for example, GC-biased gene conversion; Duret and Galtier (2009)). Given ongoing debates on how gene conversion potentially affects genetic diversity and fitness in facultatively sexual organisms (Mancera *et al.* 2008; Flot *et al.* 2013; Tucker *et al.* 2013), a deeper understanding of how gene conversion affects the distribution of genetic diversity can shed further insight into what processes influence genetic evolution in facultatively sexual organisms.

Acknowledgements. We would like to thank two anonymous reviewers and

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Associate Editor John Wakeley for providing constructive comments on the manuscript. 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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A Implementing Recombination into the Facultative Sex Simulation Algorithm

Overview of basic coalescent simulation

Here we outline the implementation of meiotic and mitotic recombination events in the facultative sex coalescent simulation routine (Hartfield *et al.* 2016). We describe the probability that set events occur per generation; that is, both the time in the past to the next event and resolution of events are based on unscaled probabilities (as opposed to rates, where a probability is multiplied by the population size to give the expected number of events per generation). We define p_{NS} as the probability that none of the x paired samples are split by sex and p_{E0} as the probability of any event (e.g., coalescence, recombination) given that none of the paired samples are affected by sex. The absolute time to the next event is drawn from a geometric distribution with parameter $p_{sum} = (1 - p_{NS}) + p_{NS}p_{E0}$; this time is rescaled by $2N$ so that it is on a coalescent timescale. It is subsequently determined whether any and, if so, how many of the x paired samples segregate into different individuals due to sexual reproduction. If k of x paired samples are produced via sex, then $2k$ new unpaired samples are created. The total probability of any other event occurring is then re-calculated, conditional on this updated sample configuration. It is determined whether any such event occurs, and which type if one does arise; the sample configuration is updated appropriately. Note that if sex is common, the first term in p_{sum} is large and all paired samples are rapidly split by sex, so the model then behaves like a haploid process. If the population is structured as an island model, the logic is similar but we instead track x_i,

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

y_i paired and unpaired haplotypes in deme i, and consider $2N_T$ total haplotypes over all subpopulations. We refer the reader to Hartfield *et al.* (2016) for further details of the basic coalescent simulation.

Implementing Meiotic and Mitotic Crossing Over

We outline the probability of either a crossover or gene conversion event occurring each generation, then implement these probabilities into the calculation of p_{sum} as described above. As with the single-locus routines, we assume that sexual reproduction occurs first, followed by subsequent gene exchange events. Let $\tilde{c}$ be the absolute meiotic crossover probability between any two adjacent sites, conditional on sex having occurred; $\tilde{c}_A$ the mitotic crossover probability (which is not conditional on the reproductive mode); and L the number of sites that the genetic samples cover. Assuming $\tilde{c}$ and $\tilde{c}_A$ are small, the total meiotic crossover probability in each individual at the start of the process is $\tilde{c}(L-1)$. We assume that the total recombination probability is low (i.e., $\tilde{c}(L-1), \tilde{c}_A(L-1) \sim \mathcal{O}(1/N)$) so we do not consider outcomes where more than one crossover event occurs per generation.

Following sexual reproduction, crossovers act on unpaired samples with probability $(\tilde{c}\,\sigma + \tilde{c}_A)L_{e,y}$. The quantity $L_{e,y}$ is the 'effective' crossover length summed over all y unpaired samples. We include this term to ensure that unnecessary crossover events are not considered, thus speeding up the routine (Hein *et al.* 2005). $L_{e,y}$ is defined as follows: let $L_{s,i}$ be the first ancestral site in unpaired sample i, and $L_{t,i}$ the last ancestral site. Then $L_{e,i} = L_{t,i} - L_{s,i}$ equals the total number of breakpoints where a crossover can create two new samples, each carrying ancestral material; note that any sites within individual haplotypes that have reached

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

their most recent common ancestor are converted into non-ancestral material (Hein *et al.* 2005). Then $L_{e,y} = \sum_{i \in y} L_{e,i}$. This crossover event creates two new samples, with each part carrying ancestral material (Figure 1(a)).

If k out of the x paired samples segregate via sex into $2k$ new unpaired samples, then the crossover probability is increased by adding on an extra $(\tilde{c}+\tilde{c}_A)L_{e,2k}$ term. Here $L_{e,2k}$ is the 'effective' crossover length over the $2k$ new unpaired samples, defined in a similar manner to $L_{e,y}$. The $2k$ samples are a transitory class of unpaired haplotypes, created through sexual reproduction segregating paired samples into distinct individuals (see also Figure 1(b, c)). 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 meiotic crossing over. Those that do not undergo crossing-over become regular unpaired samples (Figure 1(b)); those that do are transformed into regular paired samples (Figure 1(c)).

Mitotic crossing over can act on the remaining $x - k$ paired samples that do not undergo sexual reproduction. This event occurs with probability $\tilde{c}_A L_{e,x-k}$, for $L_{e,x-k}$ the effective crossover length measured over both arms within the remaining $x - k$ paired samples. $L_{e,x-k}$ is measured in a different manner than for unpaired samples. Let i be an individual where both haplotypes i_1, i_2 are sampled. Define L_{s,i_1}, L_{s,i_2} as the first ancestral site in each of these samples, and L_{t,i_1}, L_{t,i_2} the last ancestral sites. Then the first ancestral site at which mitotic crossing over is valid in individual i is $L_{s,i} = min(L_{s,i_1}, L_{s,i_2})$; similarly, $L_{t,i} = max(L_{t,i_1}, L_{t,i_2})$. Then $L_{e,i} = L_{t,i} - L_{s,i}$ and $L_{e,x-k} = \sum_{i \in (x-k)} L_{e,i}$. Mitotic crossing over exchanges genetic material between the two samples within an individual (Figure 1(d)).

These probabilities are considered alongside other events to determine whether the next event involves a meiotic crossover. If it is chosen, then one of the appro-

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

priate samples is picked at random (weighing by the length of extant breakpoints present in each sample), and the appropriate outcome is enacted. Note that if the potential for crossing over is high (i.e. if the probability of sex and crossing over is high, and there are a large number of samples) then the net recombination probability can exceed one, as the assumption that only up to one recombination event occurring per generation is violated, causing the algorithm to terminate. Hence large crossover probabilities should be avoided.

Implementing Meiotic, Mitotic Gene Conversion

To account for both meiotic and mitotic gene conversion events (Figure 1(e–g)), up to four additional parameters are specified. Two new parameters are g_S and g, the probabilities of either meiotic or mitotic gene conversion occurring with its leftmost boundary arising on the recipient homolog at a given site. We also define the average length of gene conversion events, denoted λ_S for meiotic gene conversion and λ for mitotic gene 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. Here, the length of gene conversion events (scaled by the total number of breakpoints) is drawn from an exponential distribution with parameter $Q = (L-1)/\lambda$, the number of breakpoints in units of average gene conversion length (Wiuf and Hein 2000). We define distinct $Q_S = (L-1)/\lambda_S$, $Q = (L-1)/\lambda$ for meiotic and mitotic gene conversion events. Further details of the mathematical derivations are in Section B of Supplementary Mathematica File S1.

There also exists a special class of gene conversion events, where conversion

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

initiates outside the ancestral tract and extends completely over ancestral material (Figure 1(h)). If there exist $x - k$ pairs after k of them are split by sex, then the probability of this event happening equals $2(x-k)(g(L-1)e^{-Q})/Q$ (a full derivation is presented in "Deriving probability of 'complete' gene conversion" below, and Section C of Supplementary Mathematica File S1).

Determining type of gene conversion (meiotic or mitotic)

To understand how the different events (meiotic and mitotic gene conversion) are considered, it is easiest to relate calculations to the obligate sex case with a single type of gene conversion event. Here, the product of the gene conversion probability g_0 (using g_0 to differentiate this general gene conversion probability from the mitotic gene conversion notation) and the number of breakpoints $L-1$ is such that $g_0(L-1) \sim \mathcal{O}(1/N)$. Then the probability of a disruptive gene conversion event in a sample of length L is $g_0(L-1)Q_0^*$, where g_0 is the probability that the leftmost edge of a gene conversion tract is at a given site, and:

$$Q_0^* = 1 + \frac{1}{Q_0}\left(1 - e^{-Q_0}\right) \tag{A1}$$

$Q_0 = (L-1)/\lambda_0$ is the number of breakpoints in units of average gene conversion length (here too, we use Q_0, Q_0^* and λ_0 to define this general gene conversion process). Q_0^* 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). Equation A1 assumes that the length of gene conversion events (scaled by the total number of breakpoints) is drawn from an exponential

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

distribution with parameter Q_0 (Wiuf and Hein 2000). Equation A1 also disregards possible edge effects (e.g., if the ancestral tract lies near the chromosome edge). 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:

$$\Sigma = 2Q^*g(L-1)(x-k)+y(Q^*g(L-1)+\sigma Q_S^*g_S(L-1))+2k(Q^*g(L-1)+Q_S^*g_S(L-1)) \tag{A2}$$

Here, Q^* and Q_S^* are equal to Q_0^* above, with parameters $Q = (L-1)/\lambda$ and $Q_S = (L-1)/\lambda_S$ for mitotic and meiotic gene conversion respectively. As segregation has already been resolved, the $(x-k)$ remaining paired samples reproduce asexually so only they can 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 probability 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

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

on unpaired or paired samples. The probability that gene conversion acts on a paired sample is $2Q^*g(L-1)(x-k)/\Sigma$ where Σ is given by Equation A2, 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 $\Upsilon = g(L-1)Q^*(y+2k)+g_S(L-1)Q_S^*(y\sigma+2k)$ 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 $(g(L-1)Q^*(y+2k))/\Upsilon$; a similar calculation can be made for meiotic gene conversion.

Drawing start, end breakpoints following gene conversion

The scaling terms Q^*, Q_S^* account 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 the logic 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_0 (which can be either Q or Q_S), $K(Q_0) = 1-(1-\exp(Q_0))/Q_0$ 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):

$$p_2 = \frac{K(Q_0)}{2-K(Q_0)} \tag{A3}$$

The probability that only one breakpoint falls within the sample equals $p_1 = 1-p_2$ (Wiuf and Hein 2000, Eq. 12). We first choose whether one or both

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

breakpoints lie within the sample, as determined by Equation A3. The appropriate start and end points are then chosen from the relevant probability distributions. Wiuf and Hein (2000) determined how the distribution of breakpoints depends 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. 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. Further details on the following derivations are provided in Section B of Supplementary Mathematica File S1.

If *two breakpoints are chosen*, then the joint probability distribution of start points s and end points t equals $f(s,t) = (Q_0 \exp(-Q_0(t-s)))/K(Q_0)$ (Wiuf and Hein 2000, Eq. 4). By integrating out t from s to 1, one obtains the marginal density of start points, $f(s) = (1-\exp(-Q_0(1-s)))/K(Q_0)$ (Wiuf and Hein 2000, caption of Fig. 4). The CDF of s can then be calculated as:

$$F(S) = \int_0^S f(s) = \frac{1 - e^{Q_0 S} + e^{Q_0} Q_0 S}{1 + e^{Q_0}(Q_0 - 1)} \tag{A4}$$

To choose a start point, we draw a value between 0 and 1 from a uniform

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

distribution and plug it into $F^{-1}(S)$, which equals:

$$F^{-1}(S) = \frac{e^{-Q_0}}{Q_0} \cdot (S - 1 + e^{Q_0}((Q_0 - 1)S - W[-e^{-Q_0 - e^{-Q_0}(1-S) - (1-Q_0)S}])) \quad \text{(A5)}$$

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 $f(s,t)/f(s) = (Q_0 \exp(-Q_0(t-s)))/(1 - \exp(-Q_0(1-s)))$. From this function we obtain the CDF of T given s, as well as the inverse CDF:

$$F(T|s) = \int_s^T f(t|s) = \frac{1 - \exp(-Q_0(T-s))}{1 - \exp(-Q_0(1-s))} \quad \text{(A6)}$$

$$F^{-1}(T|s) = s - \frac{\log(1 - T(1 - \exp(-Q_0(1-s))))}{Q_0} \quad \text{(A7)}$$

Equation A7 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 $f(t) = \exp(-Q_0 t)/(1 - K(Q_0))$ (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

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sample. The CDF and inverse CDF can be calculated as:

$$F(T) = \frac{\exp(Q_0) - \exp(Q_0(1-T))}{\exp(Q_0) - 1} \tag{A8}$$

$$F^{-1}(T) = 1 - \frac{\log(\exp(Q_0) + T - \exp(Q_0 T))}{Q_0} \tag{A9}$$

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 $f(s_1) = \exp(-Q_0(1-s_1))/(1-K(Q_0))$ (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:

$$F(S_1) = \frac{1 - \exp(Q_0 S_1)}{1 - \exp(Q_0)} \tag{A10}$$

$$F^{-1}(S_1) = \frac{\log(1 + (\exp(Q_0) - 1)S_1)}{Q_0} \tag{A11}$$

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) gene conversion causes a within-individual coalescent event, converting the paired sample into an unpaired sample. The recipient sample becomes non-ancestral and is no longer tracked.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Deriving probability of 'complete' gene conversion

Let Q 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. 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 $g(L-1)$ per length of focal sequence) equals $2(x-k)g(L-1)\int_{x=0}^{\infty} e^{-Q(1+x)}$. Solving the integral gives the probability $2(x-k)(g(L-1)e^{-Q})/Q$.

Note that if $L > 1$ then this probability goes to infinity as $Q \to 0$. In this case, the average gene conversion length is much larger than the genetic sample being simulated (i.e., $\lambda \gg L-1$), 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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Crease, T. J. and M. Lynch, 1991 Ribosomal DNA variation in *Daphnia pulex*. Mol. Biol. Evol. **8**: 620–640.

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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. 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.

Jaquiéry, J., S. Stoeckel, C. Rispe, L. Mieuzet, F. Legeai, *et al.*, 2012 Accelerated evolution of sex chromosomes in aphids, an X0 system. Mol. Biol. Evol. **29**: 837–847.

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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.

Möhle, M., 1998 A Convergence Theorem for Markov Chains Arising in Population Genetics and the Coalescent with Selfing. Adv. Appl. Probab. **30**: 493–512.

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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.

Normark, B. B., 1999 Evolution in a putatively ancient asexual aphid lineage: Recombination and rapid karyotype change. Evolution **53**: 1458–1469.

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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Schön, I., R. K. Butlin, H. I. Griffiths, and K. Martens, 1998 Slow molecular evolution in an ancient asexual ostracod. Proc. R. Soc. B **265**: 235–242.

Schön, I. and K. Martens, 2003 No slave to sex. Proc. R. Soc. B **270**: 827–833.

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.

bioRxiv preprint doi: https://doi.org/10.1101/232405; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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.