DOI: 10.1534/genetics.109.113555

Patterns of Neutral Genetic Variation on Recombining Sex Chromosomes

Mark Kirkpatrick,[1] Rafael F. Guerrero and Samuel V. Scarpino

Section of Integrative Biology, University of Texas, Austin Texas 78712

Manuscript received December 23, 2009
Accepted for publication January 26, 2010

ABSTRACT

Many animals and plants have sex chromosomes that recombine over much of their length. Here we develop coalescent models for neutral sites on these chromosomes. The emphasis is on expected coalescence times (proportional to the expected amount of neutral genetic polymorphism), but we also derive some results for linkage disequilibria between neutral sites. We analyze the standard neutral model, a model with polymorphic Y chromosomes under balancing selection, and the invasion of a neo-Y chromosome. The results may be useful for testing hypotheses regarding how new sex chromosomes originate and how selection acts upon them.

STUDIES of sex chromosomes have revealed a myriad of interesting evolutionary patterns. The chromosome region responsible for sex determination can shift frequently between different pairs of homologous chromosomes (Bull 1983; Mank *et al.* 2006). Genes involved in isolation between species seem to be unusually common on sex chromosomes (Coyne and Orr 2004; Saetre *et al.* 2007; Presgraves 2008). Sex chromosomes also appear to be enriched with genes for species-specific and secondary sexual traits [*e.g.*, in lepidoptera (Prowell 1998), poeciliid fish (Lindholm and Breden 2002), and plants (Scotti and Delph 2006)].

We want to understand patterns like these, both because they hold clues about how sex chromosomes evolve and because they provide a window into evolutionary forces that may be important throughout the genome. One approach is to use patterns of neutral DNA polymorphism to make inferences. That strategy, however, has limitations in groups like mammals and Drosophila that have highly heteromorphic sex chromosomes where the lack of recombination between X and Y chromosomes makes it difficult to disentangle the interactions of factors such as selection, drift, and demography.

In other groups of animals and plants, however, sex chromosomes recombine over much of their length (Ohno 1967; Bull 1983). We can exploit that situation to gain new tools to study the evolutionary forces acting on sex chromosomes using DNA polymorphism. But even in the simplest null model with no selection or demographic effects, recombining sex chromosomes challenge our intuition about what patterns of polymorphism to expect. Although the sex-determining region can be regarded as a single locus at which X and Y alleles form a balanced polymorphism, the classic coalescent theory for the effects of balanced polymorphisms at linked sites (Hudson and Kaplan 1988) does not apply. That is because nonrandom mating between the X and Y changes the effects of recombination (for example, Y chromosomes cannot exchange material directly), and recombination rates typically differ between the sexes. Thus we lack basic predictions for what patterns of DNA polymorphism to expect under a neutral model, not to mention in more evolutionarily complicated (and potentially interesting) situations.

When sex chromosomes are heteromorphic, recombination between them occurs in segments known as pseudoautosomal regions. The models presented in this article apply to those regions as well as to the recombining portions of homomorphic sex chromosomes. We avoid the term "pseudoautosomal", however, because we are largely interested in regions of the sex chromosome that recombine but are tightly linked to the sex-determining region. These segments follow hereditary rules that are neither strictly autosomal nor strictly sex linked.

Here we develop some basic results for patterns of neutral genetic variation expected on recombining sex chromosomes. We begin with the standard neutral model (SNM) in which there is no selection and the population is at demographic equilibrium. We then go on to explore two biological scenarios suggested by empirical studies. In the first, multiple Y chromosome types are maintained by balancing selection. This model is inspired by species of poecilliid fish (guppies, platyfish, and swordtails) that have striking polymorphisms in male size that are coded by genes in the sex-determining region of the Y chromosome (Lindholm and Breden 2002; Tripathi *et al.* 2009). We then study

[1]*Corresponding author:* Section of Integrative Biology, C-0930, University of Texas, Austin, TX 78712. E-mail: kirkp@mail.utexas.edu

the consequences of the invasion of a neo-Y chromosome that converts an ancestral pair of autosomes into sex chromosomes. This model is motivated by the numerous species of plants (see Ming and Moore 2007; Bernasconi *et al.* 2009) and animals (see Peichel *et al.* 2004; Mank *et al.* 2006; Cnaani *et al.* 2008; McAllister *et al.* 2008; Takehana *et al.* 2008; Kitano *et al.* 2009; Ser *et al.* 2010) in which sex chromosomes have been recently derived from autosomes (see also van Doorn and Kirkpatrick 2007).

The emphasis of this article is on finding expected coalescence times. These are proportional to the amount of neutral genetic variation that is expected to accumulate under mutation and drift. We also present results for expected patterns of linkage disequilibria. Together our findings provide a foundation for statistical tests of alternative hypotheses for the evolution of sex chromosomes.

MODELS

The key feature of sex chromosomes is their sex-determining region (SDR). For our purposes, the SDR is regarded as a single locus, although biologically it may be composed of multiple coding regions that do not recombine. In this article we assume XY sex determination (male heterogamety), but all results also apply exactly to ZW sex determination (female heterogamety) if the names of the sexes are reversed, the Y chromosome is replaced by the W, and the X chromosome is replaced by the Z.

Modeling sex-linked inheritance involves complications not present with autosomal inheritance, and to accommodate those we introduce some vocabulary. A *site* is a selectively neutral nucleotide position or larger nonrecombining region. Sites are denoted by lowercase italic letters. The sex-determining region is denoted *s*. A site can appear in different *contexts*, which are determined by the genotype that the site is linked to (for example, an X or a Y allele at the SDR). A site in a particular context is called a *position*. These are denoted by subscripting the site with a list of the information relevant to the context. Thus i_X is the position corresponding to site i on X chromosomes. Positions are also written more compactly using single lowercase double-struck letters, for example, $\mathbb{i} = i_X$. A *gene* refers to the DNA at a site that is ancestral to the original sample. Genes remain at the same site but change contexts as a result of recombination. A *carrier* is a chromosome that carries ancestral material, which is to say one or more genes. A carrier is fully described by a list of the positions it carries; for example, $\{i_X, j_X\}$ represents a carrier that is an X chromosome with genes (ancestral material) at sites i and j. At any point in the past, the *state* of all the DNA ancestral to a sample taken at the present time can described by a list of carriers. For example, we write $\{\{i_X, j_X\}, \{j_Y\}\}$ for a state in which one carrier is an X chromosome with ancestral material at sites i and j and a second carrier is a Y chromosome with ancestral material only at site j. States are abbreviated by italic uppercase letters; *e.g.*, $S = \{\{i_X, j_X\}, \{j_Y\}\}$. Frequencies of genotypes are denoted as p subscripted by the relevant alleles; *e.g.*, p_X denotes the frequency of all sex chromosomes that are X at the SDR.

Our analyses use a genealogical, or retrospective, approach based on the classic Kingman model of coalescence (reviewed by Hein *et al.* 2005 and Wakeley 2009). The point of departure is the "standard neutral model", modified to allow for the transmission rules of recombining sex chromosomes. There is no selection, and the population is at a demographic equilibrium with a population size N. In classical forward-sense population genetics, the model is an approximation to the standard Wright–Fisher model of random genetic drift. A key assumption imposed by our coalescent model is that the number of chromosomes in the smallest class of chromosomes that can share a common parent (*e.g.*, Y chromosomes) is much larger than one.

Time is measured in units of $2N$ generations relative to the present. We use the term *forward recombination rate*, denoted by r, to refer to per-generation recombination rates measured in the standard genetic sense. We use *backward recombination rate*, denoted by ρ, for a recombination rate that has been rescaled to coalescent time units using $\rho = 4Nr$.

Our main emphasis in this article is on expected coalescence times. For these, it suffices to study the coalescent process starting from a sample of just two carriers. appendix a lays out the general results from the basic coalescence theory that we need to calculate expected coalescence times. The sections below adapt those results to specific settings. We also present results regarding linkage disequilibria, and the basic calculations for them are presented in appendix b.

The calculations typically involve a large number of terms. We wrote code in *Mathematica* (Wolfram Research 2008) to automate the analytic calculations, and that code is available from the authors on request.

We used stochastic simulations of the coalescent process to check the analytic results and find results for quantities that we could not calculate analytically. The simulation generates the ancestral recombination graph (Griffiths and Marjoram 1996) for one or more chromosome regions linked to the SDR. That graph is then used to determine the coalescence time at the site(s) of interest. Simulation results shown below for expected coalescence times are based on 10^5 independent runs for each point shown. Those for linkage disequilibria are based on 10^7 runs. The simulation code, written in C++, is available from the authors on request.

The standard neutral model for recombining sex chromosomes: Consider the evolution of neutral site i on a recombining sex chromosome evolving under the standard neutral model. There are only two positions in

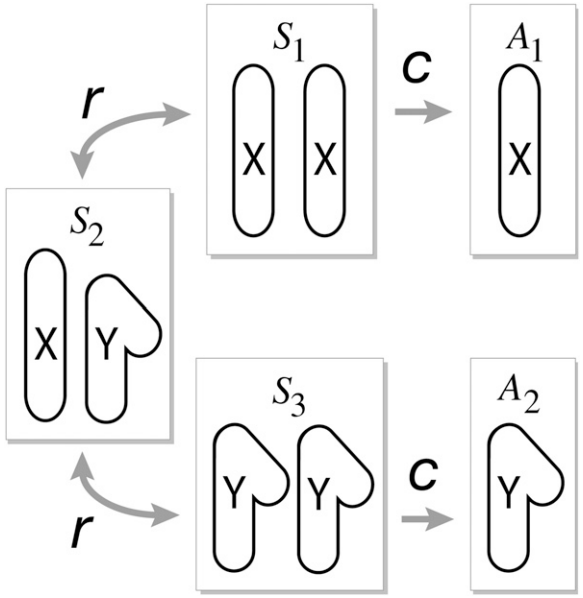

FIGURE 1.—Schematic of the states for the standard neutral model. Transitions marked "r" result from recombination, and those marked "c" result from coalescence.

this model, i_X and i_Y. The system has three transient states and two absorbing states:

$$S_1 = \{\{i_X\}, \{i_X\}\}, \quad S_2 = \{\{i_X\}, \{i_Y\}\},$$
$$S_3 = \{\{i_Y\}, \{i_Y\}\}, \quad A_1 = \{\{i_X\}\}, \quad A_2 = \{\{i_Y\}\}.$$

The states and the transitions between them are shown schematically in Figure 1.

Transitions between the transient states result from recombination. The forward recombination rate between i and the SDR is r^{m}_{si} in males. Recombination in females is not relevant in this model because it simply moves genes between different X chromosomes. Using $\rho^{\mathrm{m}}_{si} = 4Nr^{\mathrm{m}}_{si}$ for the backward recombination rate, the nonzero transition rates between the transient states are

$$P_{S_1 \to S_2} = \frac{2}{3}\rho^{\mathrm{m}}_{si}, \quad P_{S_2 \to S_1} = \rho^{\mathrm{m}}_{si}, \quad P_{S_2 \to S_3} = \frac{1}{3}\rho^{\mathrm{m}}_{si},$$
$$P_{S_3 \to S_2} = 2\rho^{\mathrm{m}}_{si}.$$

These values can be understood as follows. The rate at which a gene carried on an X recombines onto a Y is one-third of the rate that a gene recombines from a Y onto an X. That is because a gene carried on an X occurs together with a Y and so has the opportunity to recombine onto that chromosome in one generation of three. The other factor involved in the transition rates is the number of genes available to recombine: the value of $P_{S_1 \to S_2}$ is twice that of $P_{S_2 \to S_3}$ because S_1 has twice as many X-linked genes available to recombine as does S_2.

Transition rates from transient to absorbing states result from coalescence. The nonzero rates are

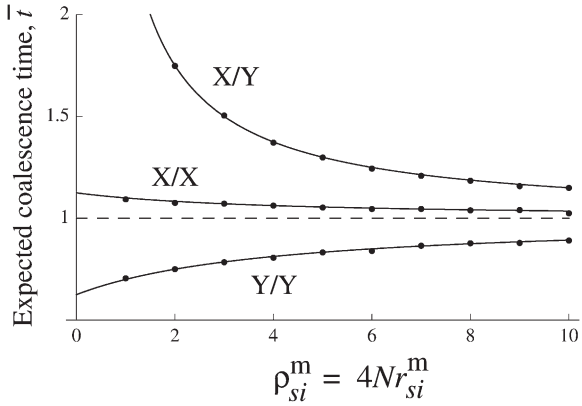

FIGURE 2.—Expected coalescence times under the standard neutral model. The sex-determining region is at the left at $\rho^{\mathrm{m}}_{si} = 0$. The dashed line shows the expectation for autosomal sites. The curves are from Equations 1, and the points show simulation results based on 10^5 runs each.

$$P_{S_1 \to A_1} = \frac{4}{3}, \quad P_{S_3 \to A_2} = 4.$$

These two values reflect the effective population sizes of X and Y chromosomes relative to autosomes.

The expected coalescence times can be calculated by evaluating Equation A5 from APPENDIX A, using these transition rates. There are three times, corresponding to the cases in which both genes are sampled from X chromosomes, where one is sampled from an X and the other from a Y, and where both genes are sampled from Y chromosomes:

$$\bar{t}_{XX} = \frac{9 + 2\rho^{\mathrm{m}}_{si}}{8 + 2\rho^{\mathrm{m}}_{si}}, \quad \bar{t}_{XY} = 1 + \frac{3}{2\rho^{\mathrm{m}}_{si}}, \quad \bar{t}_{YY} = \frac{5 + 2\rho^{\mathrm{m}}_{si}}{8 + 2\rho^{\mathrm{m}}_{si}}. \quad (1)$$

Figure 2 shows these results. In this and subsequent figures, the curves show the analytic results and the points are simulation results. Despite the fact that X chromosomes have a smaller effective population size than autosomes, the expected time to coalescence for a pair of genes sampled from the X is actually greater than that of autosomes (when $\rho^{\mathrm{m}}_{si} \neq 0$). That is because if one of these genes recombines onto a Y, coalescence cannot occur until a second recombination event either brings that gene back onto an X or brings the other gene from its X onto a Y. This additional time more than compensates for the reduction in the effective population size of X chromosomes and in fact becomes stronger with smaller values of ρ^{m}_{si}. Likewise, for pairs of genes sampled from Y chromosomes, the expected coalescence times for small recombination rates are substantially $>\frac{1}{4}$, the value for Y chromosomes with no recombination.

These results are validated by considering the weighted average of the coalescence times for pairs of genes on the X and pairs on the Y, where the weights are the frequencies of X and Y among all sex chromosomes:

$(3\bar{t}_{XX}/4) + (\bar{t}_{YY}/4)$. From Equation 1, we find that this weighted average is unity. That result follows from a general invariance principle of coalescence in structured populations (Strobeck 1987; Charlesworth *et al.* 2003). Our models are mathematically equivalent to coalescent models of migration in which there are two demes represented by X and Y chromosomes, and movement between the demes is caused by recombination.

When one gene is sampled from an X and the other from a Y, expected coalescence times become very large near the SDR. In this case, the only pathways to coalescence require at least one recombination event. Since those events are rare when ρ^m_{si} is small, expected times to coalescence are large.

As we move far from the SDR ($\rho^m_{si} > 5$, say), times for all three kinds of samples become close to the expected coalescence time of 1 for autosomes under the SNM. This might at first seem surprising, since (for example) a gene on a Y cannot coalesce with a gene on an X regardless of recombination rates. The explanation is that a pair of genes carried on X chromosomes coalesce at $\frac{4}{3}$ the rate of autosomal genes because their effective population size is smaller. Likewise, a pair of genes both on Y chromosomes coalesce at a rate four times the autosomal. These accelerated rates exactly compensate for the increased times that result because X and Y genes cannot coalesce.

Next we turn to patterns of linkage disequilibrium. One statistic that is useful to describe linkage disequilibrium is R^2, the square of the correlation coefficient between the allelic states. McVean (2002), building on results by Hudson (1985), showed how an approximation for the expected value of that quantity can be derived using coalescent methods. In our case, we are interested in the correlation between pairs of positions rather than sites since (for example) the correlation between alleles at sites i and j sampled from X chromosomes will generally be different from when they are sampled from Y chromosomes.

Consider two sites i and j at sites that lie on the same side of the SDR; the linkage map is *s–i–j*. Recombination in females now enters the model because it can dissociate a gene at site i from a gene at site j when they are both carried on an X chromosome. The recombination rate between the SDR and site i in males is written ρ^m_{si}, and the rate between i and j in females is written ρ^f_{ij}, etc. We denote the expected value of the squared correlation between them as R^2_{ij}. Appendix b shows how the approximation for R^2_{ij} is calculated using the method of McVean (2002, 2007) adapted to sex linkage. The model now includes four positions: i_X, i_Y, j_X, and j_Y. Proceeding backward from a sample of two chromosomes, the system can assume 20 states in which neither site has coalesced, 20 states in which one but not the other site has coalesced, and 6 absorbing states in which both sites have coalesced. Because of the large number of states, it is not possible to get simple expressions for R^2_{ij} even under the standard neutral model. Our calculations were done analytically (with *Mathematica*), but the results are so large that we are able to present them only when evaluated numerically.

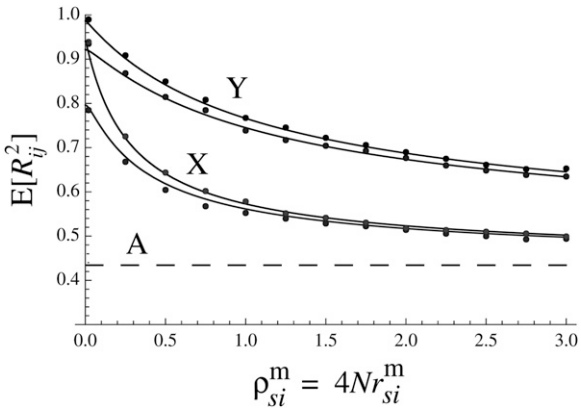

Figure 3.—The expected value for the squared correlation of allelic states between pairs of neutral sites under the standard neutral model. The two sites are separated by $\rho^f_{ij} = 0.1$. The horizontal axis is the distance ρ^m_{si} between the SDR and the nearer of the two sites. For each pair of curves, the bottom curve is for equal recombination rates in males and females ($\rho^m_{ij} = 0.1$), and the top curve is for reduced recombination in males ($\rho^m_{ij} = 0.01$). The curves are the analytic results and the points are simulation results based on 10^7 runs each. The dashed line shows the expectation for autosomes from the approximation of McVean (2002).

Figure 3 shows the expected value of R^2_{ij} at different distances from the SDR. The recombination rate in females between the two neutral sites is fixed at $\rho^f_{ij} = 0.1$. The horizontal axis is the recombination rate in males between the SDR and the nearer of the two sites. Two cases are shown: when recombination rates are equal in males and females ($\rho^m_{ij} = 0.1$) and when the recombination rate in males is 10 times smaller ($\rho^m_{ij} = 0.01$). In both cases, linkage disequilibrium between sites is highest near the SDR and converges toward the autosomal value for pairs of sites that are far from the SDR. The correlation between sites sampled from pairs of Y chromosomes is larger than that from pairs of X chromosomes. This agrees with intuition, since Y chromosomes have a smaller effective population size and therefore are expected to build up more linkage disequilibrium by drift. Decreasing the recombination rate in males elevates the disequilibria expected between sites that are very close ($\rho^m_{ij} < 1$) to the SDR.

A balanced Y polymorphism: Our next model is inspired by species of fish in which there several male morphs whose phenotypes are determined by the Y chromosome. Here we assume some form of balancing selection maintains two types of Y chromosomes, Y_1 and Y_2, at stable frequencies in the population. (The model is easily extended to any number of Y chromosome

types.) We assume this polymorphism is much older than $2N$ generations.

We now have six classes of expected coalescence times, corresponding to samples from different pairs of sex chromosomes, *e.g.*, (X, X), (X, Y_1), (X, Y_2), etc. To calculate coalescence times using the method described in APPENDIX A, we need the backward transition rates between six transient states and three absorbing states. These rates are found by modifying the formulas of KAPLAN *et al.* (1988) and HUDSON and KAPLAN (1988) for autosomal sites linked to selected loci. Transitions from transient states to absorbing states result from coalescence. A coalescent event can occur only if two genes share the same position. If they do, they coalesce at a rate $1/p_k$, where $k = \mathrm{X}, \mathrm{Y}_1, \mathrm{Y}_2$ is the state of the SDR. Transitions between transient states result from recombination. The nonzero forward recombination rates relevant to this model are

$$\Pr[i_{\mathrm{X}} \to i_{\mathrm{Y}_k}] = \frac{1}{3} r^{\mathrm{m}}_{si}(4p_{\mathrm{Y}_k}), \quad \Pr[i_{\mathrm{Y}_k} \to i_{\mathrm{X}}] = r^{\mathrm{m}}_{si}.$$

The factor of $(4p_{\mathrm{Y}_k})$ in the first transition rate is the probability that a gene that recombined from an X onto a Y will become linked to a Y_k chromosome ($i = 1, 2$). [Recall that p_{Y_k} is the frequency of all sex chromosomes (both X and Y) that are Y_k.] The backward transition rates are calculated using these forward transition probabilities. Let $\mathbb{i}_1$ be the position occupied by the gene that changes state before the transition and $\mathbb{i}_2$ its position afterward (in the backward sense). Then

$$P_{S \to U} = n_S \tilde{P}_{\mathbb{i}_2 \to \mathbb{i}_1} \left(\frac{p_{\mathbb{i}_2}}{p_{\mathbb{i}_1}} \right). \tag{2}$$

Here n_S (= 1, 2) is the number of genes in state S that could change positions in the transition from state S to state U. $\tilde{P}_{\mathbb{i}_2 \to \mathbb{i}_1}$ is the forward rate of transition from $\mathbb{i}_2$ to $\mathbb{i}_1$, and $p_{\mathbb{i}_k}$ is the frequency of chromosomes with the context of $\mathbb{i}_k$. The notation becomes clearer with an example:

$$P_{\{\{i_{\mathrm{X}}\},\{i_{\mathrm{X}}\}\} \to \{\{i_{\mathrm{X}}\},\{i_{\mathrm{Y}_1}\}\}} = 2\left[\frac{1}{3}\left(\frac{\rho^{\mathrm{m}}_{si}}{4}\right)(4p_{\mathrm{Y}_1})\right]\left(\frac{p_{\mathrm{X}}}{p_{\mathrm{Y}_1}}\right).$$

On the right, the first term is $n_S = 2$, whose value follows because there are two genes eligible to make the transition from i_{X} to i_{Y_1}. The term in square brackets is the forward transition rate from i_{X} to i_{Y_1} (that is, the forward recombination rate r^{m}_{si}) rescaled to the backward recombination rate ρ^{m}_{si}. The final term is the ratio of frequencies of the contexts that the gene changes between. We calculated the expected coalescence times for the six types of samples (two genes sampled from X chromosomes, one gene from an X and the other from a Y_1, etc.) using these transition probabilities and Equation A5 from APPENDIX A. The results are presented in APPENDIX C.

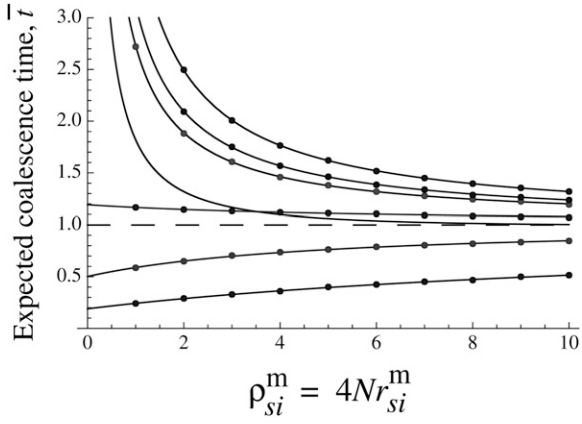

FIGURE 4.—Expected coalescence times for neutral sites when two types of Y chromosomes are maintained by balancing selection. The frequencies of Y_1 and Y_2 among all Y chromosomes are 0.75 and 0.25. The curves are, starting at the top and descending at the left, for samples from $\mathrm{Y}_1/\mathrm{Y}_2$, X/$\mathrm{Y}_2$, X/$\mathrm{Y}_1$, Y./Y., X/X, $\mathrm{Y}_1/\mathrm{Y}_1$, and $\mathrm{Y}_2/\mathrm{Y}_2$ (where Y./Y. denotes the average of randomly sampled Y chromosomes). The dashed line shows the expectation for autosomal sites. The curves are the analytic results and the points are from simulations.

Figure 4 shows an example in which Y_1 makes up three-fourths of all Y chromosomes. The largest expected coalescence times are for the case in which one gene is sampled from a Y_1 chromosome and the other from a Y_2 chromosome. Here, at least two recombination events must occur before the two genes can be carried on the same kind of sex chromosome (that is, in the same context) and so can coalesce. Because two recombination events are needed, longer coalescence times result. The smallest coalescence times occur when two genes are sampled from Y_2 chromosomes. In this event, no recombination is needed, and coalescence happens quickly because Y_2 chromosomes have an effective population size of only $N/8$ in this example.

Now we ask about $\bar{t}_{Y.Y.}$, the expected coalescence time when genes are sampled from random Y chromosomes. This quantity is of interest because in practice we may not know if multiple types of Y chromosomes are present in a population and want to test for that situation using patterns of neutral variation. This average time is calculated in APPENDIX C (Equation C4) and shown in Figure 4. This time becomes large for sites substantially closer to the SDR than $\rho^{\mathrm{m}}_{si} = 1$. For example, when $\rho^{\mathrm{m}}_{si} = 0.1$ and the frequency of the rare Y_2 allele makes up 25% of all Y chromosomes, the expected time for genes sampled from random Y chromosomes is 9.2 times greater than under the standard neutral model. This suggests that levels of neutral polymorphism among Y chromosomes near to the SDR could be used to test for balancing selection acting on variants within the SDR.

Invasion of a neo-Y: In this section we consider the recent invasion of a new Y chromosome at a pair of ancestral autosomes. The neo-Y invaded at time t^* in the

past. To simplify the model, we assume the invasion was so rapid that it appears instantaneous on the timescales of coalescence and recombination that we are considering. While our model makes no assumption about the form of selection responsible for the invasion, we can think of its speed in terms of an "effective selection coefficient", s, representing the selective advantage that the new Y had if it spread by simple genic selection. Our assumption of a strong selective sweep then assumes that $s \gg 1/2N$, $\rho^{m}_{si}/4N$. The second condition ensures that there is a negligible chance of a recombination event happening during the invasion.

There are three situations in which coalescence occurs in this model. First, proceeding from the present to the past, coalescence can happen before the invasion (that is, between the present and time t^*). In that case, the process follows the standard neutral model developed above. Second, if coalescence has not happened at before t^* and if both genes are carried by Y chromosomes, then they coalesce at t^*. Third, if neither of those eventualities hold, then both genes become carried by autosomes at t^*, and they then coalesce following the SNM for autosomes. To account for those possibilities, we write $f_{U|S}(t)$ for the probability that the system is in state U at time t, given that the system was in state S at time 0. Then the expected coalescence time for a system initially in state S is

$$\bar{t}_S = \int_0^{t^*} t c_S'(t)\,dt + t^* f_{\{i_Y,i_Y\}|S}(t^*) + \left[1 - c_S'(t^*) - f_{\{i_Y,i_Y\}|S}(t^*)\right]\left(t^* + \bar{t}_{\{i_A,i_A\}}\right). \quad (3)$$

The three terms on the right side of (3) correspond to the three situations. The first term accounts for coalescent events occurring between the present and t^*. Here $c_S(t)$ is the probability that coalescence has occurred before time t,

$$c_S(t) = f_{\{i_X\}|S}(t) + f_{\{i_Y\}|S}(t), \quad (4)$$

where $\{i_X\}$ and $\{i_Y\}$ represent, respectively, the two absorbing states in which the genes have coalesced on X chromosomes and on Y chromosomes. The derivative $c_S'(t)$ appearing inside the integral gives the rate of coalescence at time t. The second term on the right of (3) corresponds to coalescence that is forced to happen if both genes are carried by Y chromosomes when the invasion occurs at t^*. The last term in (3) corresponds to genes that coalesced farther back in the past than the invasion. Appearing in that term is $\bar{t}_{\{i_A,j_A\}}$, the expected coalescence time for a pair of autosomal genes under the SNM; its value is unity.

We now need to calculate $f_{U|S}(t)$. Let $\mathbf{f}_S(t)$ be the vector of probabilities that the system is in the transient states $\{\{i_X\}, \{i_X\}\}$, $\{\{i_X\}, \{i_Y\}\}$, and $\{\{i_Y\}, \{i_Y\}\}$ at time t given an initial state S. The theory of continuous-time Markov chains (Ross 1989, Chap. 6) and ordinary differential equations (Boyce and Diprima 2003, Chap. 7) gives us

$$\mathbf{f}_S(t) = \sum_i k_i \mathbf{e}_i \exp\{\lambda_i t\}, \quad (5)$$

where $\mathbf{e}_i$ and λ_i are the ith eigenvector and eigenvalue (respectively) of the matrix $\mathbf{M}$ of transition rates between the transient states, and the k_i are constants determined by the initial state of the system at $t = 0$. The elements of the transition rate matrix are

$$M_{ij} = \begin{cases} -\left(\sum_k P_{S_i \to S_k} + \sum_k P_{S_i \to A_k}\right) & \text{for } i = j \\ P_{S_j \to S_i} & \text{for } i \neq j, \end{cases} \quad (6)$$

where S_i is a transient state and A_i is an absorbing state. Equations 5 and 6 give us the probabilities that the system is in the transient states at any point in the past. Given those, we can calculate the probabilities that the system is an absorbing state A at time t:

$$f_{A|S}(t) = \sum_U \int_0^t f_{U|S}(t') P_{U \to A}\,dt'. \quad (7)$$

The summation in (7) is over all the transient states of the system. The quantity inside the integral represents the flow of probability from transient state U to absorbing state A, which is integrated over the relevant time period.

Substituting results from Equations 4–7 into (3) then gives the expected coalescence times. We again calculated these analytically using *Mathematica*. Again, because of their size we present results here only for particular cases evaluated numerically.

One way to view the results is to ask how expected coalescence times at a particular point on the chromosome depend on the age of the invasion. Figure 5 shows the situation for a neutral site that is closely linked to the SDR at $\rho^{m}_{si} = 1$. As expected, for recent invasions the coalescence times for pairs of genes sampled from Y chromosomes are very small. For invasions that are $\geq 2N$ generations old, these times are close to those under the SNM. The sweep has the least effect on the expected coalescence times for pairs of genes sampled from X chromosomes. The most enduring imprint of the sweep is seen on coalescence times between a gene sampled from the X and one sampled from the Y. In this case, $\bar{t}_{XY}$ goes to 1 as t^* goes to 0. That follows because, going backward in time, both genes are converted to autosomal carriers (whose expected coalescence time is unity) at the time of the invasion. The curve for coalescence times increases from $t^* = 0$ with a slope of 1. That follows because, in the absence of recombination, the genes cannot coalesce so long as they are on different sex chromosomes. That constraint is lifted at t^* when the carriers are converted to autosomes. The expected coalescent time does not approach its equi-

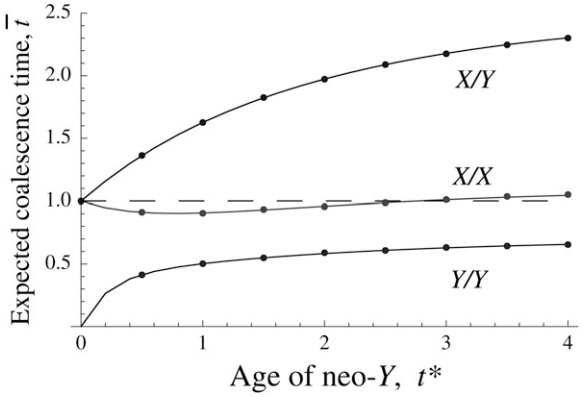

FIGURE 5.—Expected coalescence times as a function of the age of the invasion of a neo-Y at a site $\rho_{ij}^m = 1$ from the SDR. The curves are the analytic results and the points are from simulations.

librium until t^* is substantially >1. This fact may be useful for dating the origin of older neo-Y's.

For sites that are less tightly linked than $\rho_{si}^m = 1$, the picture differs in two ways from Figure 5. The key features are that equilibrium values are approached more rapidly following the sweep and that the equilibria lie closer to the SNM for autosomes.

A second way to visualize the effects of the invasion of a neo-Y is to fix the age of sweep, t^*, and look at how expected coalescence times change as we move along the chromosome. Figure 6 (top) shows the situation for a recent sweep at $t^* = 0.1$. Comparing this to Figure 2 shows that the biggest effects of the sweep are the reduced values of $\bar{t}_{XY}$ for chromosome regions near the SDR (ρ_{si}^m is not much bigger than 1). As the age of the sweep grows, the region of the chromosome that departs substantially from the SNM shrinks to a region close to the SDR. We see that effect in Figure 6 (bottom), which shows the situation for an older sweep at $t^* = 5$. By this time, coalescence times are very close to those for the standard neutral model at sites farther than $\rho_{si}^m = 1$ from the SDR (compare with Figure 2).

Invasion of a neo-Y will also alter the pattern of linkage disequilibrium between neutral sites. We studied R^2 in this situation by simulation. Figure 7 shows the results for pairs of sites at different distances from the SDR. Immediately following the invasion and near the SDR, R^2 is decreased relative to its equilibrium value. For pairs of sites sampled from X chromosomes, the expected value of R^2 is equal to its autosomal value immediately following invasion of a neo-Y. For pairs of sites sampled from Y chromosomes, the expected value of R^2 goes to 0 at $t^* = 0$. For both X and Y chromosomes, R^2 approaches its equilibrium value by $t^* = 10$ for sites at a distance $\geq \rho_{si}^m = 1$ from the SDR.

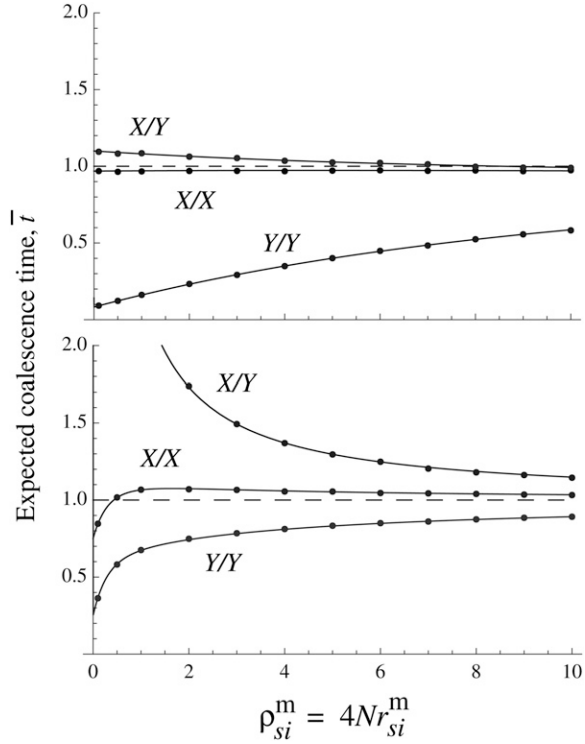

FIGURE 6.—Expected coalescence times as a function of the distance between the SDR and the neutral site following invasion of a neo-Y. (Top) A recent invasion at $t^* = 0.1$. (Bottom) An older invasion at $t^* = 5$. The curves are the analytic results and the points are from simulations.

DISCUSSION

Sex chromosomes that recombine along much of their lengths are widely distributed among animals, plants, and fungi. Some of these systems are young and have arisen in two different ways. Sex chromosomes in several families of plants (including campion, papaya, and poplar) evolved recently from hermaphroditic ancestors (CHARLESWORTH and GUTTMAN 1999; MING and MOORE 2007). A second way that new sex chromosomes originate is found in several groups of animals, where the sex-determining region has moved between linkage groups, recruiting autosomes into the role of sex chromosomes [*e.g.*, in fishes such as sticklebacks (PEICHEL *et al.* 2004), cichlids (CNAANI *et al.* 2008), and medakas (TAKEHANA *et al.* 2008)]. Ancient recombining sex chromosome are also known, for example, in boid snakes (OHNO 1967) and ratite birds (JANES *et al.* 2009). Thus the evolution of heteromorphic X and Y chromosomes, and the concomitant loss of recombination between them, does not occur in some taxa. The reasons why recombining sex chromosomes are retained in some groups but not others are not well understood.

Whatever their origins and ages, recombining sex chromosomes offer the possibility of making inferences using data on DNA polymorphism that are not possible when in systems with suppressed recombination. The models in this article suggest some of the ways by which this polymorphism is influenced by selection and historical processes. Analyzing patterns in these polymorphisms could provide a new tool for studying a wide range of

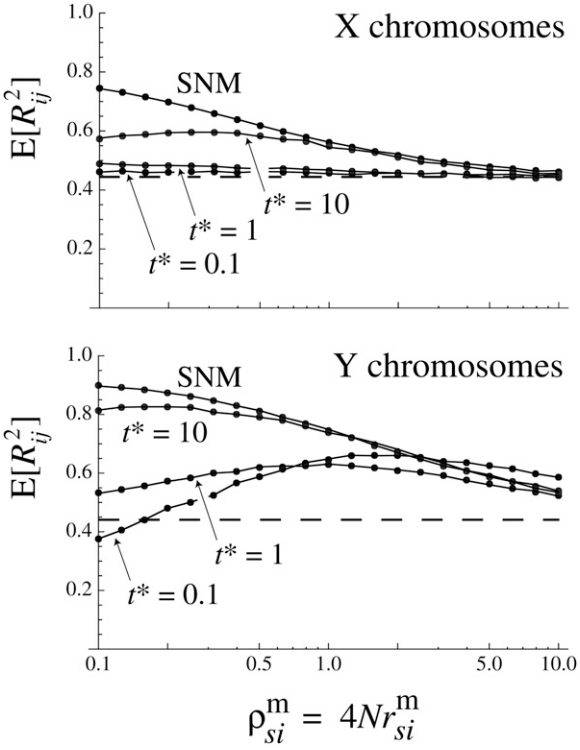

Figure 7.—The squared correlation of allelic states at two neutral sites following the invasion of a neo-Y at different times in the past. Results marked SNM are for the standard neutral model (no invasion). The horizontal axis (with a logarithmic scale) is the recombination distance between the SDR and the nearer of the two sites. Other parameters are as in Figure 3, with recombination rates equal in males and females. Simulation results (points) are connected by lines; no analytic results are shown.

important evolutionary phenomena associated with sex chromosomes. These include speciation (Coyne and Orr 2004; Presgraves 2008), sex-antagonistic selection (Rice 1987; van Doorn and Kirkpatrick 2007), the evolution of recombination (Bergero and Charlesworth 2008), and meiotic drive (Presgraves 2008).

The goal of this article is to generate expectations for patterns of neutral DNA polymorphism under some simple evolutionary scenarios. The first case we considered is X and Y chromosomes evolving under the standard neutral model, that is, in the absence of selection and demographic disturbances. Close to the SDR, expected coalescence times are shorter on Y chromosomes and longer on X chromosomes than for autosomal sites. These differences are not large, however. As the recombination rate between a neutral site and the SDR approaches 0, the expected coalescence time for a pair of genes sampled from the X is nine-eighths that of autosomes, and for a pair sampled from the Y it is five-eighths. Coalescence times are greater when one gene is sampled from an X and the other from a Y at sites very close to the SDR. For example, with $\rho^{m}_{si} = 1$, the expected time is 3.6 times longer than for a pair of genes sampled from Y chromosomes. As we might expect intuitively, expected coalescence times converge to their autosomal values at genetic distances $\gg \rho^{m}_{si} = 1$.

The second case considered here is when balancing selection maintains two types of Y chromosomes in the population. This type of selection decreases the expected coalescence time between genes sampled from the same type of Y chromosome. On the other hand, it inflates the expected coalescence times between genes sampled from different types of Y chromosome and between genes randomly sampled from Y chromosomes. These qualitative patterns follow what is seen for sites linked to loci under balancing selection on autosomes (Hudson and Kaplan 1988). This suggests that neutral genetic variation could be used to detect balancing selection that maintains multiple types of Y alleles at the sex-determining region.

The third case we modeled is that of a neo-Y chromosome that recently invaded a pair of autosomes, converting that linkage group into the sex chromosomes. The qualitative patterns seen in our results are similar to those seen following a selective sweep at an autosomal locus (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Kim and Stephan 2002; Przeworski 2002; McVean 2007; Pfaffelhuber *et al.* 2008). Immediately following the invasion, coalescence times for pairs of genes sampled from Y chromosomes are substantially reduced at sites close to the SDR ($\rho^{m}_{si} < 5$, say). By $2N$ generations after the invasion, times approach those of the standard neutral model for all but sites very closely linked to the SDR ($\rho^{m}_{si} < 1$, say). Coalescence times return to their equilibrium values more slowly, however, when one gene is sampled from an X and one from a Y. That fact could be used to date the invasion of older neo-Y chromosomes. The correlations in coalescence times between pairs of neutral sites are also depressed in the neighborhood of the SDR by invasion of a neo-Y. The effect is stronger on Y than on X chromosomes and again is most evident for moderately recent invasions ($t^* < 2$) at sites closely linked to the SDR ($\rho^{m}_{si} < 1$). Thus patterns of linkage disequilibria contain information about the age of a neo-Y that extends what can be gleaned from patterns of diversity at single sites.

These observations suggest that patterns of divergence between X and Y chromosomes, and disequilibria within X and within Y chromosomes, can be used to date when a neo-Y was established. A second possible application of these models is to test hypotheses about polymorphic sex determination systems. For example, cichlid fish in the genus Oreochromis (which includes the tilapia) have two linkage groups that contribute to sex determination, one that functions as an X/Y system and the other as Z/W (Cnaani *et al.* 2008). Other species that have sex-determining loci on more than one linkage group include the house fly (Tomita and Wada 1989), the platyfish (Kallman 1965), and the frog *Rana rugosa* (Ogata *et al.* 2003). These systems might be at stable

equilibria, or alternatively one sex determination system may be replacing another. Those possibilities might be distinguished by testing patterns of variation on the sex chromosomes against models such as ours.

Our results suggest that much of the information about the history and ongoing selection of the sex-determining region to be found in neutral genetic variation is in sites that are closely linked to the SDR (ρ^{m}_{si} not much greater than 1). For most of the genome of many organisms, that would imply a relatively small amount of DNA with which to work. That constraint could be offset, however, if regions of sex chromosomes close to the SDR show reduced recombination per DNA base. Even limited recombination may protect these regions flanking the SDR from the rapid accumulation of rearrangements and repetitive sequences often seen within the nonrecombining regions of Y chromosomes (Charlesworth *et al.* 2005). If so, these flanking regions may be a fruitful focus for empirical research.

Throughout we have presented results for the X and Y chromosomes separately. Many sequencing approaches can get sequence alleles on X chromosomes from females, but are not able to distinguish alleles on X and Y chromosomes sampled from males. Expectations for those admixtures can be found by averaging our results for the X and Y. Advancing technologies, however, may soon be able to deliver many phased sequences from the X and Y, which would make admixed data obsolete.

Exactly how will coalescent models be used to test these and other questions using molecular data? We have developed results only for expected coalescence times at single sites and the expected correlation between pairs of sites. These are useful for suggesting the chromosomal regions and types of comparisons that may provide useful data. Our results fall far short, however, of providing the full sampling distribution that could be used as the foundation of hypothesis testing using a likelihood framework. It seems unlikely that it will be possible to develop those results even under simple biological hypotheses because of algebraic complexities. A more promising approach may be analyses based on summary statistics from the data. Approaches such as approximate Bayesian computation (Beaumont *et al.* 2002; Marjoram *et al.* 2003) using summary statistics suggested by these analytic models may be a powerful way to test hypotheses about the evolution of recombining sex chromosomes. The models developed here cover but a fraction of the diverse sex determination mechanisms known in animals and plants (Bull 1983). The basic approach used here can be extended to those systems to gain understanding about evolutionary transitions between them.

We are grateful to Jon Wilkins and an anonymous reviewer for very helpful comments on the manuscript and to Katie Peichel and Monty Slatkin for discussions. This work was supported by National Science Foundation grant DEB-0819901 and the Miller Institute for Basic Research in Science.

LITERATURE CITED

Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162:** 2025–2035.

Bergero, R., and D. Charlesworth, 2008 The evolution of restricted recombination in sex chromosomes. Trends Ecol. Evol. **24:** 94–102.

Bernasconi, G., J. Antonovics, A. Biere, D. Charlesworth, L. F. Delph *et al.*, 2009 *Silene* as a model system in ecology and evolution. Heredity **103:** 5–14.

Boyce, W. E., and R. C. DiPrima, 2003 *Elementary Differential Equations and Boundary Value Problems*, Ed. 7. John Wiley & Sons, New York.

Bull, J. J., 1983 *The Evolution of Sex Determining Mechanisms*. Benjamin/Cummings, Reading, MA.

Charlesworth, B., D. Charlesworth and N. Barton, 2003 The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Evol. Syst. **34:** 99–125.

Charlesworth, B., D. Charlesworth and G. Marais, 2005 Steps in the evolution of heteromorphic sex chromosomes. Heredity **95:** 118–128.

Charlesworth, D., and D. Guttman, 1999 The evolution of dioecy and plant sex chromosome systems, pp. 25–49 in *Sex Determination in Plants*, edited by C. C. Ainsworth. BIOS Scientific Publishers, Oxford.

Cnaani, A., B. Y. Lee, N. Zilberman, C. Ozouf-Costaz, G. Hulata *et al.*, 2008 Genetics of sex determination in tilapiine species. Sex. Dev. **2:** 43–54.

Coyne, J. A., and H. A. Orr, 2004 *Speciation*. Sinauer, Sunderland, MA.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3:** 479–502.

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

Hudson, R. R., 1985 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Janes, D. E., T. Ezaz, J. A. Marshall Graves and S. V. Edwards, 2009 Recombination and nucleotide diversity in the sex chromosomal pseudoautosomal region of the emu, *Dromaius novaehollandiae*. J. Hered. **100:** 125–136.

Kallman, K. D., 1965 Genetics and geography of sex determination in the poeciliid fish, *Xiphophorus maculatus*. Zoologica **50:** 151–190.

Kaplan, N. L., T. Darden and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kitano, J., J. A. Ross, S. Mori, M. Kume, F. C. Jones *et al.*, 2009 A role for a neo-sex chromosome in stickleback speciation. Nature **461:** 1079–1083.

Lindholm, A., and F. Breden, 2002 Sex chromosomes and sexual selection in poeciliid fishes. Am. Nat. **160:** S214–S224.

Mank, J. E., D. E. L. Promislow and J. C. Avise, 2006 Evolution of alternative sex-determining mechanisms in teleost fishes. Biol. J. Linn. Soc. **87:** 83–93.

Marjoram, P., J. Molitor, V. Plagnol and S. Tavare, 2003 Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci USA **100:** 15324–15328.

Maynard Smith, J., and J. Haigh, 1974 The hitchhiking effect of a favorable gene. Genet. Res. **23:** 23–35.

McAllister, B. F., S. L. Sheeley, P. A. Mena, A. L. Evans and C. Schlötterer, 2008 Clinal distribution of a chromosomal rearrangement: A precursor to chromosomal speciation? Evolution **62:** 1852–1865.

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

McVean, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175:** 1395–1406.
Ming, R., and P. H. Moore, 2007 Genomics of sex chromosomes. Curr. Opin. Plant Biol. **10:** 123–130.
Ogata, M., H. Ohtani, T. Igarashi and Y. Hasegawa, 2003 Change of the heterogametic sex from male to female in the frog. Genetics **164:** 613–620.
Ohno, S., 1967 *Sex Chromosomes and Sex-Linked Genes*. Springer-Verlag, Berlin.
Peichel, C. L., J. A. Ross, C. K. Matson, M. Dickson, J. Grimwood *et al.*, 2004 The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Curr. Biol. **14:** 1416–1424.
Pfaffelhuber, P., A. Lehnert and W. Stephan, 2008 Linkage disequilibrium under genetic hitchhiking in finite populations. Genetics **179:** 527–537.
Presgraves, D. C., 2008 Sex chromosomes and speciation in Drosophila. Trends Genet. **24:** 336–343.
Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179–1189.
Prowell, D. P., 1998 Sex linkage and speciation in Lepidoptera, pp. 309–319 in *Endless Forms: Species and Speciation*, edited by D. J. Howard and S. H. Berlocher. Oxford University Press, New York.
Rice, W. R., 1987 The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution **41:** 911–914.
Ross, S. R., 1989 *Introduction to Probability Models*, Ed. 4. Academic Press, Boston.
Saetre, S. A., G. P. Saetre, T. Borge, C. Wiley, N. Svedin *et al.*, 2007 Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. Science **318:** 95–97.
Scotti, I., and L. F. Delph, 2006 Selective trade-offs and sex chromosome evolution in *Silene latifolia*. Evolution **60:** 1793–1800.
Ser, J. R., R. B. Roberts and T. D. Kocher, 2010 Multipleinteracting loci control sex determination in Lake Malawi cichlid fish. Evolution **64:** 486–501.
Strobeck, C., 1987 Average number of nucleotide differences in a sample from a single subdivided population: a test for population subdivision. Genetics **117:** 149–153.
Takehana, Y., S. Hamaguchi and M. Sakaizumi, 2008 Different origins of ZZ/ZW sex chromosomes in closely related medaka fishes, *Oryzias javanicus* and *O. hubbsi*. Chromosome Res. **16:** 801–811.
Tomita, T., and Y. Wada, 1989 Multifactorial sex determination in natural populations of the house fly (*Musca domestica*) in Japan. Jpn. J. Genet. **64:** 373–382.
Tripathi, N., M. Hoffman, D. Weigel and C. Dreyer, 2009 Linkage analysis reveals independent origin of Poeciliid sex chromosomes and a case of atypical sex inheritance in the guppy (*Poecilia reticulata*). Genetics **182:** 365–374.
van Doorn, G. S., and M. Kirkpatrick, 2007 Turnover of sexual chromosomes induced by sexual conflict. Nature **449:** 909–912.
Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts & Co., Greenwood Village, CO.
Wolfram Research, 2008 *Mathematica*, version 7.0. Wolfram Research, Champaign, IL.

Communicating editor: M. K. Uyenoyama

APPENDIX A: EXPECTED COALESCENCE TIMES FOR SINGLE POSITIONS

This appendix summarizes the calculations used to find expected coalescence times. All these results were developed previously by Kaplan *et al.* (1988) and Hudson and Kaplan (1988). We include them here to make clear how they apply using the notation of this article.

For a system that starts in transient state S, we write the time back to the most recent common ancestor of the genes at site i as $T_{i|S}$. Its value is

$$T_{i|S} = H_S + \sum_U Q_{S\to U} T_{i|U}. \tag{A1}$$

Here H_S is the holding time in transient state S, that is, the waiting time until the system exits that state. $Q_{S\to U}$ is the jump probability from state S to state U, that is, the probability that when the process leaves S (in the backward sense) it goes to U. The summation is over all states U, both transient and absorbing.

Taking expectations of (A1) gives the expected value of the coalescence time,

$$\overline{T}_{i|S} = \overline{H}_S + \sum_U Q_{S\to U} \overline{T}_{i|U}, \tag{A2}$$

where overbars denote expectations. The expected holding times and jump probabilities can be written in terms of the transition rates between states of the system,

$$\overline{H}_S = 1 \Big/ \sum_U P_{S\to U}, \tag{A3}$$

$$Q_{S\to U} = \frac{P_{S\to U}}{\sum_V P_{S\to V}} = \overline{H}_S P_{S\to U}, \tag{A4}$$

where $P_{S\to U}$ is the (backward) transition rate from state S to state U. We define $P_{S\to S} = 0$.

To solve for the expected coalescence times, we first order the transient states in some arbitrary way. Using this ordering, denote the vector of expected coalescent times as $\bar{\mathbf{t}}$, the matrix of conditional jump probabilities as $\mathbf{Q}$, and the vector of expected holding times as $\overline{\mathbf{h}}$. Then writing the system of Equation A2 in matrix notation, simple algebra gives the solution for the expected coalescent times as

$$\bar{\mathbf{t}} = (\mathbf{I} - \mathbf{Q})^{-1}\bar{\mathbf{h}}. \tag{A5}$$

We evaluated this equation for the different models described in the text by specifying the appropriate transition rates $P_{S\to U}$.

APPENDIX B: LINKAGE DISEQUILIBRIA BETWEEN PAIRS OF POSITIONS

This appendix gives results for linkage disequilibrium at pairs of neutral sites that are sex linked. Our approach follows McVean (2002; see also Wakeley 2009, pp. 236–241). We need to distinguish between the different positions at the sites, since, for example, the association between sites i and j sampled from X chromosomes will generally be different from the association between those sites sampled from Y chromosomes. This requires that we make minor extensions to previous results derived for autosomal inheritance. The notation used here is defined in the body of the article and appendix a.

Here we calculate the expected value of R^2_{ij}, the square of the correlation of coalescent times at positions i and j. That quantity gives a good approximation to the square of the correlation of the allelic states at those sites if their allele frequencies lie between 0.1 and 0.9 (Hudson 1985). In that case, we have

$$E[R^2_{\mathrm{ij}}] = \frac{\mathrm{Cov}\left[T_{\mathrm{i}|S_1}, T_{\mathrm{j}|S_1}\right] - 2\mathrm{Cov}\left[T_{\mathrm{i}|S_2}, T_{\mathrm{j}|S_2}\right] + \mathrm{Cov}\left[T_{\mathrm{i}|S_3}, T_{\mathrm{j}|S_3}\right]}{E\left[T_{\mathrm{i}|S_1}\right]\mathrm{E}\left[T_{\mathrm{j}|S_1}\right] + \mathrm{Cov}\left[T_{\mathrm{i}|S_3}, T_{\mathrm{j}|S_3}\right]}, \tag{B1}$$

where

$$S_1 = \{\{\mathrm{i,j}\}, \{\mathrm{i,j}\}\}, \quad S_2 = \{\{\mathrm{i,j}\}, \{\mathrm{i}\}, \{\mathrm{j}\}\}, \quad S_3 = \{\{\mathrm{i}\}, \{\mathrm{i}\}, \{\mathrm{j}\}, \{\mathrm{j}\}\}.$$

(McVean 2002, Equation 9). The first covariance in the numerator pertains to coalescent times between genes at two positions i and j that are carried on just two chromosomes. The second covariance in the numerator involves coalescence times when three chromosomes are involved. One of them carries a gene at both positions i and j, another carries a gene only at i, and the third carrier only at j. The final covariance in the numerator pertains to the situation in which each of the four genes is carried on a different chromosome.

The next step is to find expressions for these covariances. From the definition of a covariance we have

$$\mathrm{Cov}[T_{\mathrm{i}|S}, T_{\mathrm{j}|S}] = \overline{TT}_{\mathrm{ij}|S} - \overline{T}_{\mathrm{i}|S}\overline{T}_{\mathrm{j}|S}, \tag{B2}$$

where $\overline{TT}_{\mathrm{ij}|S} = E[\overline{T}_{\mathrm{i}|S}\overline{T}_{\mathrm{j}|S}]$. The last two terms on the right side of (B2) are expected coalescent times for pairs of genes at a single position, which are derived in appendix a (Equation A5). The first quantity on the right is the expectation for a product of coalescent times. Results for these have been derived previously (reviewed by McVean 2002). Here we follow those derivations with our notation to show how results for sex chromosomes are obtained.

We use Equation A1 to write

$$T_{\mathrm{i}|S}T_{\mathrm{j}|S} = \sum_U Q_{S\to U}(H_S + T_{\mathrm{i}|U})(H_S + T_{\mathrm{j}|U}). \tag{B3}$$

Taking the expectation of (B3) gives

$$\begin{aligned}\overline{TT}_{\mathrm{ij}|S} &= \sum_U Q_{S\to U}(E[H_S^2] + E[H_S T_{\mathrm{i}|U}] + E[H_S T_{\mathrm{j}|U}] + E[T_{\mathrm{i}|U}T_{\mathrm{j}|U}]) \\ &= 2\overline{H}_S^2 + \overline{H}_S\sum_U Q_{S\to U}(\overline{T}_{\mathrm{i}|U} + \overline{T}_{\mathrm{j}|U}) + \sum_U Q_{S\to U}\overline{TT}_{\mathrm{ij}|U}\end{aligned} \tag{B4}$$

(see Wakeley 2009, Equation 7.27). Here we have used two facts. First, $E[H_S T_{\mathrm{i}|U}] = E[H_S]E[T_{\mathrm{i}|U}]$, which follows from the Markovian nature of the system. Second, $E[H_S^2] = 2\overline{H}_S^2$, which follows from the properties of the exponential distribution. We have already solved for three kinds of terms appearing in the last line of (B4): $\overline{H}_S$ via (A3), $Q_{S\to U}$ via (A4), and $\overline{T}_{\mathrm{i}|U}$ via (A5).

Equation B4 represents a linear system of equations in the unknown $\overline{TT}_{\mathrm{ij}|S}$. To solve for them, we form the vector $\overline{\mathbf{tt}}$ whose elements are $\overline{TT}_{\mathrm{ij}|U}$ for the states U in which neither of the two sites has yet coalesced. The ordering of the states in this vector is arbitrary, but must be consistent with the other vectors and matrices that we will define shortly. The dimension of $\overline{\mathbf{tt}}$ (that is, the number of states in which neither site has coalesced) is denoted n_0. Equation B4 can then be written in matrix form as

$$\overline{\mathbf{tt}} = 2\overline{\mathbf{H}}\,\overline{\mathbf{h}} + \overline{\mathbf{H}}(\mathbf{Q}_0 \,|\, \mathbf{Q}_1)(\overline{\mathbf{t}}_i + \overline{\mathbf{t}}_j) + \mathbf{Q}_0\overline{\mathbf{tt}}. \tag{B5}$$

Here $\overline{\mathbf{h}}$ is the vector of expected holding times (also of dimension n_0) and $\overline{\mathbf{H}}$ is the diagonal matrix populated by these holding times. The $\overline{\mathbf{t}}_i$ and $\overline{\mathbf{t}}_j$ are vectors of expected coalescence times for sites i and j. They are of dimension $(n_0 + n_1)$, where n_1 is the number of states in which one but not both sites have coalesced. The first n_0 elements of these two vectors follow the ordering of states defined by $\overline{\mathbf{h}}$, while the remaining n_1 elements (corresponding to states in which one of the two sites has coalesced) follow some arbitrary but consistent ordering. The left part of the partitioned matrix $(\mathbf{Q}_0|\mathbf{Q}_1)$ is the matrix $\mathbf{Q}_0$, with dimensions $n_0 \times n_0$. Its elements are the jump probabilities between the states in which no coalescence has occurred; the elements are again defined by (A4). The right side of $(\mathbf{Q}_0|\mathbf{Q}_1)$ is the matrix $\mathbf{Q}_1$, whose elements are the jump probabilities from states in which neither site has coalesced to states in which one of the two sites has coalesced. This matrix has rows corresponding to the ordering of states defined by $\overline{\mathbf{h}}$ and columns corresponding to the ordering of states in $\overline{\mathbf{t}}_i$ and $\overline{\mathbf{t}}_j$, and it has dimensions $n_0 \times n_1$.

A simple bit of matrix algebra then gives a closed-form result for the expectations we need:

$$\overline{\mathbf{tt}} = (\mathbf{I} - \mathbf{Q}_0)^{-1}\overline{\mathbf{H}}[2\overline{\mathbf{h}} + (\mathbf{Q}_0 \,|\, \mathbf{Q}_1)(\overline{\mathbf{t}}_i + \overline{\mathbf{t}}_j)]. \tag{B6}$$

This gives us expressions for the $\overline{TT}_{\mathrm{ij}|S}$. By substituting these and those from Equation A5 into (B2) and then those results into (B1), we arrive at the approximation for $E[R^2_{\mathrm{ij}}]$, the expected value for the squared correlation between allelic states at positions $\mathbb{i}$ and $\mathbb{j}$. We evaluated this equation for the different models studied in the body of the article by specifying the appropriate transition rates $P_{S\to U}$.

We complemented these analytic calculations with coalescent simulations. The simulation results were used to calculate the approximation for $E[R^2_{\mathrm{ij}}]$ given by Equation B1.

APPENDIX C: EXPECTED COALESCENCE TIMES FOR A BALANCED Y POLYMORPHISM

This appendix gives the expected coalescence times when two types of Y chromosomes are maintained at a stable equilibrium by balancing selection. We use p to denote the frequency of Y_1 and q for the frequency of Y_2 among all Y chromosomes. Four of the six expected coalescence times for this case are

$$\overline{t}_{\mathrm{XX}} = \frac{28(9 + 2\rho_{\mathrm{m}}) + pq(80 + 23\rho_{\mathrm{m}} + 14\rho_{\mathrm{m}}^2)}{224 + 56\rho_{\mathrm{m}} + pq(14\rho_{\mathrm{m}} + 2)\rho_{\mathrm{m}}}, \tag{C1}$$

$$\overline{t}_{\mathrm{XY_1}} = \frac{496 - 160p + (372 - 26p - 38p^2)\rho_{\mathrm{m}} + (56 + 37pq)\rho_{\mathrm{m}}^2 + 14pq\rho_{\mathrm{m}}^3}{224\rho_{\mathrm{m}} + 2(28 + pq)\rho_{\mathrm{m}}^2 + 14pq\rho_{\mathrm{m}}^3}, \tag{C2}$$

$$\overline{t}_{\mathrm{Y_1Y_1}} = \frac{p(5 + 2\rho_{\mathrm{m}})(36 + 7\rho_{\mathrm{m}} - p(8 + 7\rho_{\mathrm{m}}))}{224 + 56\rho_{\mathrm{m}} + pq(14\rho_{\mathrm{m}} + 2)\rho_{\mathrm{m}}}, \tag{C3}$$

$$\overline{t}_{\mathrm{Y_1Y_2}} = \frac{q(5 + 2\rho_{\mathrm{m}})(28 + p(8 + 7\rho_{\mathrm{m}}))}{224 + 2(28 + pq)\rho_{\mathrm{m}} + 14pq\rho_{\mathrm{m}}^2}. \tag{C4}$$

There are two remaining times. The value for $\overline{t}_{\mathrm{XY_2}}$ is found by interchanging p and q in the expression above for $\overline{t}_{\mathrm{XY_1}}$, and the value of $\overline{t}_{\mathrm{X_2Y_2}}$ is found by doing the same with the expression for $\overline{t}_{\mathrm{X_1Y_1}}$.

These results can be checked using the weighted average of coalescence times for pairs of genes sampled from the same type of chromosome, $(3\overline{t}_{\mathrm{XX}}/4) + (\overline{t}_{\mathrm{Y_1Y_1}}p/4) + (\overline{t}_{\mathrm{Y_2Y_2}}q/4)$, which equals 1. As explained in the text following Equation 1, that follows as a special case of a general invariance principle for coalescence times in structured populations (Strobeck 1987; Charlesworth *et al.* 2003).

The expected coalescence time for two genes sampled from randomly chosen Y chromosomes is found by averaging $\overline{t}_{\mathrm{Y_1Y_1}}$, $\overline{t}_{\mathrm{Y_1Y_2}}$, and $\overline{t}_{\mathrm{Y_2Y_2}}$, weighting them, respectively, by p^2, $2pq$, and q^2. Doing that gives

$$\overline{t}_{\overline{\mathrm{Y}}\overline{\mathrm{Y}}} = \frac{(5 + 2\rho_{\mathrm{m}})(28\rho_{\mathrm{m}} + pq(128 - \rho_{\mathrm{m}}(48 - 7\rho_{\mathrm{m}} + pq(8 + 7\rho_{\mathrm{m}}))))}{2\rho_{\mathrm{m}}(112 + \rho_{\mathrm{m}}(28 + pq(1 + 7\rho_{\mathrm{m}})))}. \tag{C5}$$