Recombination and sterility in inversion homo- and heterokaryotypes under a general counting model of chiasma interference

Øystein Kapperud^{1,2}

1) Centre for Ecological and Evolutionary Synthesis, University of Oslo

2) Current address: Norwegian Institute of Public Health

Sandakerveien 24C

0473 Oslo

Oystein.Kapperud@fhi.no

Abstract:

It has long been known that chiasmata are not independently generated along the chromosome, a phenomenon known as *chiasma interference*. In this paper, I suggest a model of chiasma interference that generalizes the *Poisson model*, the *counting model*, the *Poisson-skip model* and the *two-pathway counting model* into a single framework, and use it to derive infinite series expressions for the sterility and recombination pattern probabilities in inversion homo- and heterokaryotypes, and a closed-form expression for the special case of the two-pathway counting model in homokaryotypes.

1 Introduction

Meiotic recombination reshuffles the genetic material before reproduction, and may thus serve to avoid the problem of having slightly deleterious mutations accumulate through their association with favored ones (Muller 1964). Although models of multilocus evolution typically assume that the foci of recombination are independently distributed along the chromosome, in reality it has long been known that such *chiasmata* are subject to spatial *interference* (Sturtevant 1913, 1915, Muller 1916, Berchowitz and Copenhaver 2010). In 1993, Foss et al. suggested the *counting model* of chiasma interference, in which the genetic distances between neighbouring chiasmata are generated by summing a given number of exponential distributions. This model later spawned the *Poisson-skip model* of Lange et al. (1997), in which the number of intervening exponential distributions is drawn from a given probability distribution, and the *two-pathway counting model* of Copenhaver et al. (2002), which is identical to Foss et al.'s model except that it allows for a second pathway of non-interfering chiasmata. These models, which I will refer to collectively as the *counting models*, have been used to generate expressions of recombination pattern probabilities which fit well with data for inversion homokaryotypes from a wide range of organisms (Foss et al. 1993, Zhao et al. 1995a, Lande and Stahl 1993, Lange et al. 1997, Copenhaver et al. 2002).

Chromosomal inversions partly suppress recombination in heterokaryotypes (e.g. Coyne et al. 1991, 1993, Navarro and Ruiz 1997, Jaarola et al. 1998) and are often found to be linked to loci involved in differentiation of alternative mating strategies (Tuttle et al. 2016, Lamichhaney et al. 2016, Wang et al. 2013), local adaptation (Etges and Levitan 2004, Sinclair-Waters et al. 2018) and pre- and postzygotic reproductive isolation between species or subpopulations (Noor et al. 2001, Feder et al. 2003, Ayala et al. 2013, Poelstra et al. 2014). Although the effects of chiasma interference on underdominance and recombination in inversion heterokaryotypes have been recognized (Navarro et al. 1997), general mathematical expressions for an indefinite number of loci are currently lacking. Accordingly, although the evolution of chromosomal inversions is a common target for modelers (e.g. Charlesworth and

Charlesworth 1973, Tricket and Butlin 1994, Navarro and Barton 2003, Feder and Nosil 2009, Dagilis and Kirkpatrick 2016), the potential effects of chiasma interference is to my knowledge invariably ignored.

In this paper, I will present a *general counting model* of chiasma interference, which expands the *Poisson-skip model* to allow for a second pathway of non-interfering chiasmata. Since this model can be reduced to each of the other counting models in special cases, it provides a single unifying framework for investigating chiasma interference. I will here use it to derive exact infinite series expressions for gamete proportions and sterility in inversion homo- and heterokaryotypes for an indefinite number of loci at any location on the chromosome, and a closed form expression for the gamete proportions in homokaryotypes under the *two-pathway counting model*. I single out for special treatment the case of paracentric inversions in *Drosophila* females, for which meiotic peculiarities complicates the mathematics (Sturtevant and Beadle 1936, Navarro et al. 1997)

1.1 General terminology

1.1.1 Recombination

Since discussions of recombination and related issues risk being obfuscated by inconsistent and ambiguous terminology, I will begin by carefully laying out my own. A *tetrad* is a bundle of four *chromatids*, originating as a duplication of each of the two *parental homologues*, with one *pair* of chromatids denoted *sister chromatids* if they are derived from the same homologue and *non-sister chromatids* if they are not. Exchanges of genetic material – or *crossing over* – occur as a series of *chiasmata* (singular *chiasma*) or *chiasma events* distributed along the tetrad according to a model of *chiasma interference*, so that each such event *involves* two *non-sister chromatids*, or, equivalently, one out of the four possible *non-sister chromatid pairs*, chosen according to a model of *chromatid interference*.¹ In accordance with the lack of consistent evidence to the contrary (Zhao et al. 1995b), I will here assume that each non-sister chromatid pair has an equal chance of being involved in each chiasma event, i.e. I assume no *chromatid* interference (not to be confused with *chiasma* interference). The number of chiasma events in an interval $\mathbb{I}$ will be denoted $X_{\mathbb{I}}$, so that the *genetic length* of $\mathbb{I}$ in Morgans is given by $\frac{1}{2}E[X_{\mathbb{I}}]$. In the following, the terms *distance* and *length* always refer to genetic (as opposed to physical) distance, unless otherwise noted.

A chromatid is said to *be recombinant* or *show recombination* in a given interval if it is *involved* in an odd number of chiasma events within that interval (note that this is not the same as saying that an odd number of chiasma events occur within the interval), and to *be non-recombinant* or show *non-recombination* in the opposite case. A *recombination pattern* is a set of Boolean variables that represent the *recombination status* – recombination (1) or non-recombination (0) – in each of the n adjoining and non-overlapping intervals, so that if the set of all intervals – the *region of interest* – is $\{\mathbb{I}_0, \mathbb{I}_1, \mathbb{I}_2 \dots \mathbb{I}_{n-1}\}$, then we can denote a recombination pattern r as

$$r = \{r(\mathbb{I}_0), r(\mathbb{I}_1), r(\mathbb{I}_2) \dots r(\mathbb{I}_{n-1})\}$$

where

$$r(\mathbb{I}) = \begin{cases} 1, & \text{recombination in } \mathbb{I} \\ 0, & \text{non-recombination in } \mathbb{I} \end{cases}$$

¹ This is sometimes referred to as the *four-strand model of recombination*, to distinguish it from the simplified *two-strand* (sometimes *one-strand*) *model* that is implicit in most textbook accounts (see Speed 1995 for a discussion of these two models). *Strand* is here synonymous with chromatid.

Accordingly, a chromatid is said to *show* or *have* recombination pattern r if the presence or absence of recombination in all the intervals of the region of interest correspond to r . I will use the random vector $\mathbf{R} = \{R(\mathbb{I}_0), R(\mathbb{I}_1), R(\mathbb{I}_2) \dots R(\mathbb{I}_{n-1})\}$ to represent the recombination pattern of a randomly chosen chromatid, so that I can write e.g. “the probability of observing a chromatid with recombination pattern r is p ” as

$$\Pr\{\mathbf{R} = \mathbf{r}\} = p$$

or “the probability of observing recombination status $r(\mathbb{I})$ in interval $\mathbb{I}$ is p ” as

$$\Pr\{R(\mathbb{I}) = r(\mathbb{I})\} = p$$

or “the probability of observing recombination in interval $\mathbb{I}$ is p ” as

$$\Pr\{R(\mathbb{I}) = 1\} = p$$

For notational simplicity, I will sometimes use the shorthand forms $R_k = R(\mathbb{I}_k)$ and $r_k = r(\mathbb{I}_k)$ when discussing recombination for a set of indexed intervals.

I will assign *directionality* to the region of interest by denoting interval $\mathbb{I}_0$ as the *the leftmost interval* and $\mathbb{I}_{n-1}$ as the *rightmost interval*. More generally, I will say that any interval $\mathbb{I}_x$ is positioned to the *left* of interval $\mathbb{I}_y$ if $x < y$, and to the *right* of $\mathbb{I}_y$ if $x > y$. The directionality also applied to the boundaries between the intervals and the chiasmata within the intervals, so that I for instance can say that a chiasma in interval $\mathbb{I}_x$ occurs to the right of another chiasma in the same interval if it is positioned closer to the right boundary of $\mathbb{I}_x$ (which is equivalent to the left boundary of $\mathbb{I}_{x+1}$).

1.1.2 Chromosomal inversions

It is common to distinguish between *pericentric inversions*, which include the centromere, and *paracentric inversions*, which do not. In individuals heterozygous for both of these types of inversion, crossing over within the inverted region results in a proportion of *unbalanced gametes* – i.e. gametes that do not have the full set of genes. In the following, all such gametes are assumed to produce inviable zygotes, and the proportion of such gametes for any given individual or genotype will be referred to as that individual or genotype’s *sterility*. For pericentric inversion heterokaryotypes, a chromatid is, as illustrated in figure 4, unbalanced if and only if it shows recombination in the inverted region, so that the sterility is given by the recombination rate of the inverted region. For paracentric inversions, different number of chiasmata in the inverted and proximal regions produce different proportions of tetrads with five different possible *configurations* (figures 1, 4, 6, 7 and 8). In the following, a *no bridge* tetrad is a tetrad that do not form chromosome bridges at either anaphase I or II, a *single anaphase I bridge* tetrad is a tetrad in which one chromatid pair form a chromosome bridge at anaphase I, a *double anaphase I bridge* tetrad is a tetrad in which two chromatid pairs (i.e. all four chromatids) form anaphase I bridges, and a *single* and *double anaphase II bridge* tetrad is the same for anaphase II. I include in theorem 6 a to my knowledge unprecedented expression for the proportions of these five configurations given the number of chiasma events in the inverted and proximal regions. The chromatids involved in a chromosome bridge break randomly at the respective anaphase, and so become unbalanced in the sense defined above.

Figure 1: The five types of tetrad configurations. a) no bridge, b) single anaphase I bridge, c) single anaphase II bridge, d) double anaphase I bridge, e) double anaphase II bridge. Acentric fragments (see figure 6) are not shown. Chromatids involved in a bridge break randomly at the given anaphase, and are necessarily unbalanced.

When all chromatids have an equal chance of becoming gametes, the sterility and recombination pattern probabilities for paracentric inversion heterokaryotypes is the same as for pericentric inversions. I will therefore refer to these cases collectively as *standard inversions*. In the *linear meiosis* of females of *Drosophila* (Sturtevant and Beadle 1936, Roberts 1976) and *Sciara* (Carson 1946), the two unbalanced chromatids in an anaphase I tetrad are retained in the polar bodies, so the remaining two balanced chromatids are the only ones that can become gametes. Since a single chiasma event in the inverted region always generates an anaphase I tetrad (figure 6), this means that additional chiasma events in either the inverted or proximal region are required to produce unbalanced gametes (figures 7 and 8). Hence, the sterility is significantly reduced. Furthermore, the set of recombination patterns that end up in an anaphase I tetrad is not a representative sample of the full set, so the gamete proportions are also affected. In the following, I will refer to the case of paracentric inversion heterokaryotypes with linear meiosis as *paracentric linear inversions*.

1.2 Chiasma interference

A *renewal process* (see Ross 2014, chapter 7) is as a stochastic process that represents the number of events (of a given type) that occur in an interval of a given length, when the distances between neighboring events are all drawn from the same *interarrival distance distribution*. I will say that the events *occur independently* or *are independent* if the probability of observing a given number of events in any given interval is independent of the number of events in all other disjoint intervals. The renewal process that possesses this attribute is known as the *Poisson process* (Ross 2014), in which the interarrival distance distribution is exponential, and the number of events in an interval of a given (genetic) length follows a Poisson distribution, so that

$$\Pr\{y \text{ events in interval } \mathbb{I}\} = \frac{e^{-\mu} \mu^y}{y!}$$

where μ is the expected number of events in interval $\mathbb{I}$. If the chiasma events occur independently, they are distributed according to a Poisson process, and we say that there is *no chiasma interference*. This

seems to be the case in some organisms, e.g. *Schizosaccharomyces pombe* (Munz 1994), but these are exceptional (see Berchowitz and Copenhagen 2010 for a review). I will refer to the model that assume no chiasma interference as *the Poisson interference model*, as first described in Haldane (1919). If chiasma events do not occur independently – i.e. if the probability of observing a given number of chiasma events in one interval depends on the number of chiasma events in a disjoint interval – we say that there is *chiasma interference*. A common measure of the degree of interference is the *coefficient of coincidence* (Muller 1916, Foss et al. 1993), one version of which is defined as

$$C(L_0, L_1) = \frac{\Pr\{R_0 = 1, R_1 = 1\}}{\Pr\{R_0 = 1\} \Pr\{R_1 = 1\}}$$

where I_0 and I_1 are two adjoining intervals of genetic length L_0 and L_1 (This is equivalent to S_3 in Foss et al. 1993, except that they impose the additional constraint that $L_0 = L_1$). Note that if the chiasma events occur independently, then $\Pr\{R_0 = 1, R_1 = 1\} = \Pr\{R_0 = 1\} \Pr\{R_1 = 1\}$ and $C(L_0, L_1) = 1$ for all values of L_0 and L_1 . It is common to distinguish between *positive chiasma interference*, in which a chiasma in one location impede (i.e. makes less likely) the generation of a chiasma in a nearby location ($C(L_0, L_1) < 1$ for small L_0, L_1), and *negative chiasma interference*, in which a chiasma in one location facilitates (i.e. makes more likely) the generation of a chiasma in a nearby location ($C(L_0, L_1) > 1$ for small L_0, L_1). In practice, chiasma interference is almost always positive and gradually decreasing with distance, i.e. $C(L_0, L_1)$ is typically close to zero for small L_0, L_1 and approaches 1 as L_0, L_1 are increased (Berchowitz and Copenhagen 2010, Foss et al. 1993).

The existence of chiasma interference implies that the information that a chiasma has occurred at one location on a chromosome must somehow propagate to nearby locations where it interferes with (i.e. influences the probability of) the generation of other chiasmata. I will refer to this information as *the interference signal*. Exactly how the interference signal manifests itself physically is poorly understood (Hillers 2004, Berchowitz and Copenhagen 2010); suggestions include a hypothetical polymer that grows out from each chiasma (King and Mortimer 1990), and the build-up and release (at chiasma locations) of physical stress along the chromosome (Kleckner et al. 2004, Wang et al. 2015). Foss et al. (1993) suggests a model, henceforth referred to as the *pure counting model*, where *intermediate events* occur according to a Poisson process (i.e. independently) with rate $2(m+1)$ along the tetrad, but each such event is subsequently resolved as either a chiasma event – with both crossing over and gene conversion – or what I will refer to as a *dummy event* – with gene conversion but not crossing over – in a strict sequence so that there are always m dummy events between each pair of chiasma events. Letting C denote an intermediate event, C_x a chiasma event and C_0 a dummy event, the counting model hence postulates that for, say, $m = 2$ the sequence $\dots C_x C_0 C_0 C_x C_0 C_0 C_x C_0 C_0 \dots$ with exponential interarrival distances between the C 's will be repeated along the span of the region of interest. This implies a chiasma interarrival distance distribution equal to the sum of $m+1$ exponential distributions (m dummy events plus one chiasma event); in the literature on chiasma interference, this distribution is variously referred to as the chi-squared distribution (e.g. Zhao et al. 1995a) or the (scaled) Erlang distribution (Nolan 2017). Interference in the pure counting model depends on genetic, as opposed to physical, distance and is always positive (Lange et al. 1997) and stronger for higher m ; the latter two attributes follow from the central limit theorem and the law of large numbers, as illustrated in figure 2. Note that the chiasma events occur independently if $m = 0$, because then all (independent) intermediate events are resolved as chiasma events. The Poisson interference model, in which there is no chiasma interference, can therefore be thought of as a special case of the counting model.

Figure 2: The distribution of distances between chiasma events under the pure counting model with $m=0,3,7$. Note that all three curves by definition have a mean of 0.5 Morgans, and that the probability of observing two closely spaced chiasma events is smaller for higher m . Since the distances are given by the sum of $m+1$ (exponential) random variables with rate $2(m+1)$, it follows from the central limit theorem and the law of large numbers that the distribution approaches an increasingly narrow normal distribution for higher m . In the limit when m approaches infinity, all chiasma events are spaced exactly 0.5 Morgans apart.

There are two ways of interpreting the pure counting model; either as a literal description of the physical manifestation of the interference signal – a “machine that can count”, in the words of Foss and Stahl (1995) – or as a mathematical abstraction that, disregarding gene conversions, models the distribution of chiasma events without making any claims as to how interference actually works. The latter interpretation is foreshadowed in the mathematically equivalent models of Cobbs (1978) and Stam (1979), among others (see McPeck and Speed 1995 for a brief historical overview). Foss et al. (1993), however, clearly favor the former interpretation, and they support this view by presenting a remarkably good fit between the predicted and observed coefficients of coincidence for *Drosophila* and *Neurospora* when m , crucially, is independently estimated from the ratio of the number of gene conversions to the number of chiasmata in a different dataset. Additional evidence for the accuracy of the pure counting model’s prediction of the distribution of chiasma events in at least some species is given in Lande and Stahl (1993), Zhao et al. (1995), and McPeck and Speed (1995). One prediction of the literal “machine that can count” interpretation – that a region enclosed by two chiasma events will be enriched for gene conversions without crossing over – is, however, not fulfilled in *Saccharomyces cerevisiae* (Foss and Stahl 1995). Stahl et al. (2004) and Malkova et al. (2004) suggest that this discrepancy is due to the presence of two distinct and independent types of chiasmata, with and without interference, in *S. cerevisiae* and some (but not all) other organisms, and they provide evidence to that effect (see also Berchowitz and Copenhaver 2010). I will refer to the chiasma interference model that incorporates the none-interference type in addition to the interference (counting) type of chiasma events as the *two-pathway counting model*. This model has proved a significantly better fit to data from e.g. *Arabidopsis thaliana* (Copenhaver et al. 2002) and humans (Housworth and Stahl 2003), compared to the pure counting model. Another suggested generalization, anticipated in Foss et al. (1993), is the *Poisson-skip model* of Lange et al. (1997), in which the number of dummy events between each chiasma event is drawn from a probability distribution. Among the advantages of this model is that allows for more fine-tuned modelling of interference strengths, and that it allows for negative as well as positive chiasma interference. For clarity, I will henceforth refer to the Poisson-skip model as the *stochastic counting model*.

In keeping with the deficiency of evidence for any of the proposed theories (Hillers 2004, Berchowitz and Copenhaver 2010), I will in this study remain agnostic about the physical manifestation of the interference signal and, disregarding gene conversions altogether, treat the dummy events as useful mathematical abstraction that may or may not actually correspond to sites of gene conversion without

crossing over. On that basis, I will here suggest a further generalization of the four models introduced in this section – the *Poisson model*, the *pure counting model*, the *two-pathway counting model* and the *stochastic counting model* – which I will refer to as the *general counting model* or just the *general model* for short. I will then use this model to derive expressions for the coefficient of coincidence (theorem 1), recombination pattern probabilities in homokaryotypes (theorem 2), a closed-form version of theorem 1 in a special case (theorem 3), the sterility of standard inversion heterokaryotypes (theorem 4), the recombination pattern probabilities in standard inversion heterokaryotypes (theorem 5), the sterility of paracentric linear inversion heterokaryotypes (theorem 6) and the recombination pattern probabilities of paracentric linear inversion heterokaryotypes (theorem 7). Note that the expressions give recombination pattern probabilities of balanced gametes as the proportion among all gametes and not (as in Navarro et al. 1997) as a proportion of balanced gametes only; to get the latter, simply divide by $1 - \zeta$, where ζ is the sterility calculated by using the relevant theorem.

2 The model

As in the two-pathway counting model, I postulate two independent chiasma-generating pathways. The *type I pathway* is without chiasma interference, and generate only *type I chiasma events* according to a Poisson process. The *type II pathway* is (potentially) with interference, and generate *type II intermediate events* according to a Poisson process; these are subsequently resolved as either *type II chiasma events* or *type II dummy events*, where the number of the latter following each instance of the former is drawn from a user-defined probability distribution. That is, the type II chiasma events occur according to the stochastic counting model. It will be convenient to give the different events symbols for easier reference, in particular when these have to be incorporated into mathematical expression. In the following I will denote the type I chiasma events, type II chiasma events, type II dummy events, and type II intermediate events as X' , X'' , O'' , and C'' , respectively (note that the C'' events comprises the union of all X'' and O'' events.). The union of all chiasma events of either type will be denoted X , and the union of all events – the *intermediate events* (without the *type II* qualifier) – will be denoted C . The same symbols with a subscript indicating an interval will serve as random variables representing the number of the event in question within that interval. Hence, I can, for example, write “the probability of observing one or more type I chiasma events in interval $\mathbb{I}$ is p ” as $\Pr\{X'_{\mathbb{I}} > 0\} = p$, or “the expected number of type II intermediate events in interval $\mathbb{I}$ is λ ” as $E[C''_{\mathbb{I}}] = \lambda$.

In the homokaryotype case, the general model can be described by three user-defined sets of parameters. $\mu = \{\mu_0, \mu_1, \mu_2, \dots, \mu_{n-1}\}$ and $\lambda = \{\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_{n-1}\}$ are the expected number of type I chiasma events and type II intermediate events, respectively, in each of the n intervals, so that $\mu_k = E[X'_{\mathbb{I}_k}]$, $\lambda_k = E[C''_{\mathbb{I}_k}]$ for $k = 0, 1, 2 \dots n - 1$. Keep in mind that the $X'_{\mathbb{I}}$ and $C''_{\mathbb{I}}$ events are both Poisson distributed. The third set of parameters is the distribution of probabilities for observing a given number of consecutive O'' events following a X'' event. I will call this distribution the *intervening O'' events distribution*, and denote it $\gamma = \{\gamma_0, \gamma_1, \gamma_2 \dots \gamma_m\}$, so that γ_q (for $q = 0, 1, 2 \dots m$, $\sum_{q=0}^m \gamma_q = 1$) gives the probability of observing q consecutive O'' events following a X'' event, or, equivalently, q intervening O'' events between a pair of X'' events. One cycle hence consists of one X'' event and q O'' events, where q is drawn from γ , so that the expected total number of C'' events in a cycle is $\sum_{q=0}^m (q + 1)\gamma_q$. I will assume that there is a user-defined finite upper limit, m , on the possible number of intervening O'' events, so that $\gamma_q = 0$ for $q > m$. This implies that the genetic length of an interval $\mathbb{I}_k$ is given by

$$\text{Length of } \mathbb{I}_k \text{ in Morgans} = \frac{1}{2}E[X_{\mathbb{I}_k}] = \frac{1}{2}(E[X'_{\mathbb{I}_k}] + E[X''_{\mathbb{I}_k}]) = \frac{1}{2}\left(\mu_k + \frac{\lambda_k}{\sum_{q=0}^m (q + 1)\gamma_q}\right)$$

Note that the general counting model reduces to the stochastic counting model in the special case where $\mu_k = 0$ for $k = 0, 1, 2 \dots n - 1$, i.e. when there are no type I chiasma events; and to the two-pathway counting model in the special case where $\gamma_m = 1, \gamma_q = 0$ for $q \neq m$, i.e. when there are strictly m intervening O'' events between all consecutive pairs of X'' events. Figure 2 presents the general counting model, the two-pathway counting model, the stochastic counting model, the pure counting model, and the Poisson model as a model hierarchy where each downwards arrow pointing from model a to model b indicate which parameters you have to restrict (and how) in model a to get model b as a special case. I will in this text sometimes refer to all these collectively as *the counting models*, and use the term *pure counting model* when referring specifically to the model suggested by Foss et al. (1993).

Figure 3: The hierarchy of counting models.

During meiosis for chromosomes heterozygous for a chromosomal inversion, the inverted region can form a homosynaptic *inversion loop* (figure 4) inside and outside of which the formation of chiasmata is partly or fully inhibited (Novitski and Braver 1954, Coyne et al. 1991, 1993, Navarro and Ruiz 1997, Jaarola et al. 1998, Anton et al. 2005, Pegueroles et al. 2010, del Priore and Pigozzi 2015). I will henceforth define the *inhibition factor* for interval $\mathbb{I}_k$, denoted $d_k \in \mathbb{R}_{\geq 0}$, so that if the expected number of X' and C'' events in interval $\mathbb{I}_k$ is μ_k and λ_k , respectively, in a homokaryotype, then the corresponding values in a heterokaryotype are $d_k\mu_k$ and $d_k\lambda_k$. The two boundaries of the inverted region, the *breakpoint boundaries*, can for our purpose be thought of as loci that serve as interval boundaries in the same way that other loci do, so that e.g. the interval between the left breakpoint boundary and the leftmost loci within the inverted region (or the right breakpoint boundary if there are no loci within the inverted region) is an interval in the same respect that the intervals enclosed by two non-breakpoint loci are. Since inversions affect the rate of chiasma generation also in interval outside of the inverted region (e.g. Pegueroles et al. 2010), d_k also applies for such intervals. I will adopt the convention that $E[\]$ always refer to expectation in a homokaryotype, so that e.g. $E[C_{\mathbb{I}_k}'']$ and $d_k E[C_{\mathbb{I}_k}'']$ gives the expected number of type II intermediate events in interval k in a homokaryotype and heterokaryotype, respectively. This notation allows different intervals within and outside of the inverted region to be experience inhibition to different degrees, depending, for example, on their distance from the breakpoints.

Figure 4: An inversion loop for a pericentric inversion. The lines with the same line type (dotted or dashed) are sister chromatids, and the circles represent the centromere. The X symbolizes a chiasmata, which involves the chromatids touched by the upper and lower arms of the X. Gametes that show recombination in the interval comprising the full inverted region are unbalanced, as illustrated here by gamete 2) and 4), which lack locus D and A, respectively.

For pericentric inversion heterokaryotypes, I will divide the region of interest into three subregions for easier reference. *The left region* is the region comprising all intervals to the left of the inverted region, *the inverted region* is, as before, the region comprising all intervals captured by the inversion, and *the right region* is the region comprising all intervals to the right of the inverted region. These three regions will be denoted $\mathbb{d}$, $\mathbb{h}$ and $\mathbb{r}$, respectively, so that e.g. $E[C''_{\mathbb{h}}]$ and $d_{\mathbb{h}}E[C''_{\mathbb{h}}]$ denotes the expected number of type II intermediate events in the inverted region as a whole in a homokaryotype and heterokaryotype, respectively. The number of intervals in $\mathbb{d}$, $\mathbb{h}$, and $\mathbb{r}$ will be denoted $\mathfrak{d}$, $\mathfrak{h}$, and $\mathfrak{r}$, so that the total number of intervals is $\pi = \mathfrak{d} + \mathfrak{h} + \mathfrak{r}$. As before, $\mathbb{i}_k$ refer to zero-indexed interval number k from the left (one-indexed interval $k + 1$), so from the preceding we can unambiguously deduce that $\mathbb{i}_k$, $k = 0, 1, 2 \dots \mathfrak{d} - 1$ is an interval in the left region, $\mathbb{i}_k$, $k = \mathfrak{d}, \mathfrak{d} + 1, \mathfrak{d} + 2 \dots \mathfrak{d} + \mathfrak{h} - 1$ is an interval in the inverted region, and $\mathbb{i}_k$, $k = \mathfrak{d} + \mathfrak{h}, \mathfrak{d} + \mathfrak{h} + 1, \mathfrak{d} + \mathfrak{h} + 2 \dots \pi - 1$ is an interval in the right region. For easier reading, I will when convenient adopt the convention that an $\mathbb{d}$, $\mathbb{h}$, and $\mathbb{r}$ with subscript index k refer to the same interval as $\mathbb{i}_k$, the only difference being the additional explicit (and redundant) information about which of the three subregions the interval belongs to. For example, $\mathbb{i}_k$ and $\mathbb{h}_k$ refer to the same interval, zero-indexed number k from the left in the region of interest, but in the latter case you can immediately see that the interval is in the inverted region without having to scrutinize the index.

For paracentric linear inversion heterokaryotypes, I will divide the region of interest into four subregions from left to right: *the distal region* comprises all intervals to the left of the inverted region, *the inverted region* is as before, *the proximal region* comprises all intervals between the inverted region and

the centromere (which, for our purpose, serve as a loci), and the *right region* comprises all intervals to the right of the proximal region. These will in the following be denoted $\mathbb{d}$, $\mathbb{h}$, $\mathbb{p}$, and $\mathbb{r}$, respectively, with the number of intervals in each denoted $\mathfrak{d}$, $\mathfrak{h}$, $\mathfrak{p}$ and $\mathfrak{r}$. Otherwise the notation is the same as for pericentric inversions.

Since the order of the intervals within the inverted region is inverted in one homologue with respect to the other, there are at least two equivalent paths through which the interference signal can travel across this region (figure 5). This raises the empirical problem of how the interference signal travels through inversion breakpoint boundaries, which to my knowledge is not addressed in the literature – unsurprisingly, given that the physical manifestation of the interference signal is yet to be satisfactory described (Hillers 2004, Berchowitz and Copenhaver 2010). In my model, I suggest either to calculate the recombination pattern probabilities under the assumption that the interference signal travels through either of the two paths with equal probability, or to do so under the assumption that the signal is blocked by the inversion boundaries so that the number of chiasma events inside the inverted region is independent of the number outside of this region (there is evidence that this occurs in at least some cases; Gorlov and Borodin 1995). The Boolean parameter α can be set to 1 or 0 to facilitate either of these possibilities. Note that in the former case I have for notational simplicity only included one of the two possible paths in the theorems; to get the desired result, do the calculations twice with the order of the intervals inside the inverted region reversed and average the results. When the interference signal is blocked by the inversion boundaries ($\alpha = 0$), the directionality of the inverted region is irrelevant under the assumption of stationarity (see the next section), so only one calculation needs to be performed.

Figure 5: The problem of interference across breakpoint boundaries. Does the interference signal from a chiasma event (the x) in $\mathbb{I}_0$ go through $\mathbb{I}_1$ (top) or $\mathbb{I}_2$ (bottom)? Or is it blocked by the breakpoint boundaries?

2.1 Phases and stationarity in the general model

In the following, I will define the *phase* at any location on a tetrad as the number of O'' events between that location and the nearest X'' event to the *right*. I will furthermore define $Q_{\mathbb{I}^l}$ and $Q_{\mathbb{I}^r}$ as random variables representing the phase at the left and right boundary, respectively, of interval $\mathbb{I}$, and more generally Q_a as a random variable representing the phase at a location a . A *phase distribution* is now a probability distribution that gives the probabilities of observing the individual possible phases at a given location. The *stationary phase distribution* is the phase distribution satisfy the equation set

$$\begin{aligned} \pi &= \pi P \\ \sum_{q=0}^m \pi_q &= 1 \end{aligned}$$

where $\pi = (\pi_0 \ \pi_1 \ \dots \ \pi_m) = (\Pr\{Q_a = 0\} \ \Pr\{Q_a = 1\} \ \dots \ \Pr\{Q_a = m\})$ and P is the Markovian transition matrix that transforms the phase distribution at a given location a into the phase distribution at a location b , when there is one C'' event between a and b , i.e.

$$P[i, j] = \begin{cases} 1, & i \neq 0; j = i - 1 \\ \gamma_j, & i = 0; \\ 0, & \text{otherwise;} \end{cases} \quad \text{for } i, j = 0, 1, 2 \dots m$$

(See the appendix for a note on the matrix notation used in this text). Solving this equation set gives

$$\pi_k = \frac{\sum_{q=k}^m \gamma_q}{\sum_{q=0}^m (q+1)\gamma_q} \quad \text{for } k = 0, 1, 2, 3 \dots m$$

which is the stationary phase distribution for the general model. This distribution is important because if it is given that the phase distribution at any given location a corresponds to this distribution, then the phase distribution at all other locations b also corresponds to this distribution. This means that if we *assume stationarity*, then all regions that are described by the same set of λ , σ and γ values are fungible, regardless of their position on the tetrad. It also means the *left* and *right* boundaries of any interval or region have the same phase distribution, so that it does not matter which we call *left* and which we call *right*. In other words, reversing the numbering of the intervals in the region of interest does not change the results.

Cobbs' (1978) and Stam's (1979) early analyses of (the mathematical equivalent of) the pure counting model were complicated by the assumption of a non-stationary phase distribution (though they both considered the stationary phase distribution as a special case), which were motivated by Mather's (1938) hypothesis that the process of chiasma generation starts at the centromere and proceed from there in both directions, implying that the phase at the centromere is non-stationary. This hypothesis was in turn motivated by early results indicating that the interference signal is blocked by the centromere, meaning that chiasma on one side of the centromere does not interfere with chiasma on the other. However, a more recent analysis by Colombo and Jones (1997) indicate that the results cited by Mather are merely statistical artefacts, and that, in contradiction to Mather's predictions, interference does in fact work across the centromere in the same way as in other regions. This leaves it unclear exactly how and where the process of chiasma generation starts; or as Colombo and Jones (1997, p. 226) put it, "If chiasma formations does not start from the centromere, or end against the centromere, where does it start from or end? Or, to put it bluntly, does it start from anywhere?" Given that we do not know the

answers to these questions (Hillers 2004, Berchowitz and Copenhaver 2010), it seems to me that assuming a stationary phase distribution, as more recent treatments of chiasma interference tend to do, is as plausible and parsimonious as any other alternative. I will therefore do so in the rest of this text. The predictions of a stationary and non-stationary model are in any case not very different (Lande and Stahl 1993).

3 Results

3.1 Theorem 1: The coefficient of coincidence for the general model

The coefficient of coincidence for the general model is given by

$$C(L_0, L_1) = \frac{\frac{1}{2} (\Pr\{R(\mathbb{I}_0) = 1\} + \Pr\{R(\mathbb{I}_1) = 1\} - \Pr\{R(\mathbb{I}_{0 \cap 1}) = 1\})}{\Pr\{R(\mathbb{I}_0) = 1\} \Pr\{R(\mathbb{I}_1) = 1\}} \quad (1)$$

where $\mathbb{I}_{0 \cap 1}$ is the interval between the left boundary of $\mathbb{I}_0$ and the right boundary of $\mathbb{I}_1$, and

$$\Pr\{R(\mathbb{I}) = 1\} = \frac{1}{2} \left(1 - e^{-\mu_{\mathbb{I}}} \sum_{q=0}^m \pi_q \sum_{c=0}^q \frac{e^{-\lambda_{\mathbb{I}}} \lambda_{\mathbb{I}}^c}{c!} \right), \quad \text{for } \mathbb{I} = \mathbb{I}_0, \mathbb{I}_1, \mathbb{I}_{0 \cap 1} \quad (2)$$

$$\lambda_{\mathbb{I}} = E[C_{\mathbb{I}}'']$$

$$\mu_{\mathbb{I}} = E[X_{\mathbb{I}}']$$

$$E[C_{\mathbb{I}_{0 \cap 1}}''] = E[C_{\mathbb{I}_0}'] + E[C_{\mathbb{I}_1}']$$

$$E[X_{\mathbb{I}_{0 \cap 1}}'] = E[X_{\mathbb{I}_0}'] + E[X_{\mathbb{I}_1}']$$

Proof:

Equation (1) follows from the definition of $C(L_0, L_1)$ (section 1.2) and the relation

$$\begin{aligned} \Pr\{R(\mathbb{I}_{0 \cap 1}) = 1\} &= \Pr\{R(\mathbb{I}_0) = 1 \text{ XOR } R(\mathbb{I}_1) = 1\} \\ &= \Pr\{R(\mathbb{I}_0) = 1\} + \Pr\{R(\mathbb{I}_1) = 1\} - 2\Pr\{R(\mathbb{I}_0) = 1, R(\mathbb{I}_1) = 1\} \end{aligned}$$

(Foss et al. 1993). To get equation (2), first note that

$$\Pr\{X_{\mathbb{I}} = 0\} = \Pr\{X_{\mathbb{I}}' = 0\} \Pr\{X_{\mathbb{I}}'' = 0\}$$

where

$$\begin{aligned} \Pr\{X_{\mathbb{I}}' = 0\} &= e^{-\mu_{\mathbb{I}}} \\ \Pr\{X_{\mathbb{I}}'' = 0\} &= \sum_{q=0}^m \pi_q \sum_{c=0}^q \frac{e^{-\lambda_{\mathbb{I}}} \lambda_{\mathbb{I}}^c}{c!} \end{aligned}$$

The latter expression sums over all possible phases q at the left boundary of $\mathbb{I}$ weighted by their stationary probabilities (π_q), and multiplies each term with the probability of observing between 0 and q C'' events

in the interval, which by the definition of phase means all consecutive C'' events up to, but not including, the first X'' event. Equation (2) now follows from Mather's (1938) equation,

$$\Pr\{R(\mathbb{I}) = 1\} = \frac{1}{2}(1 - \Pr\{X_{\mathbb{I}} = 0\})$$

QED

3.2 Theorem 2: Recombination in homokaryotypes

Assuming stationarity and no chromatid interference, the probability of observing recombination pattern r on a homokaryotypic chromosome under the general counting model is given by

$$\Pr\{\mathbf{R} = \mathbf{r}\} = \boldsymbol{\pi} \left(\prod_{k=0}^{n-1} \mathbf{M}_{\mathbb{I}_k}(\mathbf{r}) \right) \mathbf{1}^T$$

where

$$\boldsymbol{\pi} = (\pi_0 \quad \pi_1 \quad \pi_2 \quad \dots \quad \pi_m)$$

$$\pi_l = \frac{\sum_{q=l}^m \gamma_q}{\sum_{q=0}^m (q+1)\gamma_q} \text{ for } l = 0, 1, 2 \dots m$$

$$\mathbf{M}_{\mathbb{I}_k}(\mathbf{r}) = \begin{cases} \frac{1}{2} \mathbf{G}_{\mathbb{I}_k}, & r(\mathbb{I}_k) = 1 \\ \frac{1}{2} \mathbf{G}_{\mathbb{I}_k} + \mathbf{H}_{\mathbb{I}_k}, & r(\mathbb{I}_k) = 0 \end{cases} \text{ for } k = 0, 1, 2 \dots n-1$$

$$\mathbf{G}_{\mathbb{I}_k}[i, j] = \psi_k(i, j) e^{-\mu_k} + \left(\psi_k(i, j) + \delta_{\{i \geq j\}} \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!} \right) (1 - e^{-\mu_k}) \text{ for } i, j = 0, 1, 2 \dots m$$

$$\psi_k(i, j) = \sum_{n=0}^{\infty} b_n \sum_{q=j}^m \gamma_q \frac{\lambda_k^{i+1+n+q-j} e^{-\lambda_k}}{(i+1+n+q-j)!}$$

$$b_n = \begin{cases} \sum_{q=0}^{n-1} b_q \gamma_{n-1-q}, & n = 1, 2, 3 \dots \\ 1, & n = 0 \end{cases}$$

$$\mathbf{H}_{\mathbb{I}_k}[i, j] = \begin{cases} \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!} e^{-\mu_k}, & i \geq j \\ 0, & i < j \end{cases} \text{ for } i, j = 0, 1, 2 \dots m$$

$$\delta_{\{condition\}} = \begin{cases} 1, & \text{condition is true} \\ 0, & \text{condition is false} \end{cases}$$

$$\lambda_k = E[C''_{\mathbb{I}_k}]$$

$$\mu_k = E[X'_{\mathbb{I}_k}]$$

Proof:

As in Zhao et al. (1995), Lange et al. (1997) and Copenhaver et al. (2002), the key idea in theorem 1 is to sum over all possible phases at each boundary by means of matrix multiplications. The vector $\boldsymbol{\pi}$ gives the phase distribution at the leftmost boundary², and element i, j of the $\mathbf{M}_{\mathbb{I}_k}(\mathbf{r})$ matrices gives the probability of observing the relevant recombination status (according to $\mathbf{r}$) and phase j at the right boundary of interval $\mathbb{I}_k$, given phase i at the left boundary of $\mathbb{I}_k$, i.e.

$$\mathbf{M}_{\mathbb{I}_k}(\mathbf{r})[i, j] = \Pr\{R_k = r_k, Q_{\mathbb{I}_k^r} = j | Q_{\mathbb{I}_k^l} = i\}$$

I will in this proof first show by induction that

$$\Pr\{\mathbf{R} = \mathbf{r}\} = \boldsymbol{\pi} \left(\prod_{k=0}^{n-1} \mathbf{M}_{\mathbb{I}_k}(\mathbf{r}) \right) \mathbf{1}^T$$

before I will find expressions for $\mathbf{M}_{\mathbb{I}_k}(\mathbf{r})$ in terms of the model parameters.

For the induction proof, let S_1 be the statement $\Pr\{R_0 = r_0, R_1 = r_1, Q_{\mathbb{I}_1^r} = j\} = (\boldsymbol{\pi} \mathbf{M}_{\mathbb{I}_0}(\mathbf{r}) \mathbf{M}_{\mathbb{I}_1}(\mathbf{r})) [j]$, and the induction hypothesis S_{k-1} be

$$\Pr\{R_0 = r_0, R_1 = r_1, \dots, R_{k-1} = r_{k-1}, Q_{\mathbb{I}_{k-1}^r} = j\} = (\boldsymbol{\pi} \mathbf{M}_{\mathbb{I}_0}(\mathbf{r}) \dots \mathbf{M}_{\mathbb{I}_{k-1}}(\mathbf{r})) [j]$$

To prove S_1 , note first that

$$\Pr\{Q_{\mathbb{I}_0^l} = i\} = \pi_i \quad \text{for } i = 0, 1, 2 \dots m$$

so

$$\begin{aligned} \sum_{i=0}^m \Pr\{Q_{\mathbb{I}_0^l} = i\} \Pr\{R_0 = r_0, Q_{\mathbb{I}_0^r} = j | Q_{\mathbb{I}_0^l} = i\} &= \sum_{i=0}^m \pi_i \Pr\{R_0 = r_0, Q_{\mathbb{I}_0^r} = j | Q_{\mathbb{I}_0^l} = i\} \\ &= \Pr\{R_0 = r_0, Q_{\mathbb{I}_0^r} = j\} = (\boldsymbol{\pi} \mathbf{M}_{\mathbb{I}_0}(\mathbf{r})) [j] \end{aligned}$$

Using that $\Pr\{R_1 = r_1, Q_{\mathbb{I}_1^r} = j | R_0 = r_0, Q_{\mathbb{I}_1^l} = i\} = \Pr\{R_1 = r_1, Q_{\mathbb{I}_1^r} = j | Q_{\mathbb{I}_1^l} = i\}$ (because R_0 is irrelevant if the phase at the left boundary of $\mathbb{I}_1$ is known) and that $Q_{\mathbb{I}_k^r} = Q_{\mathbb{I}_{k+1}^l}$ we have

$$\begin{aligned} \Pr\{R_0 = r_0, R_1 = r_1, Q_{\mathbb{I}_1^r} = j\} &= \sum_{i=0}^m \Pr\{R_0 = r_0, R_1 = r_1, Q_{\mathbb{I}_1^l} = i, Q_{\mathbb{I}_1^r} = j\} \\ &= \sum_{i=0}^m \Pr\{R_0 = r_0, Q_{\mathbb{I}_1^l} = i\} \Pr\{R_1 = r_1, Q_{\mathbb{I}_1^r} = j | R_0 = r_0, Q_{\mathbb{I}_1^l} = i\} \\ &= \sum_{i=0}^m \Pr\{R_0 = r_0, Q_{\mathbb{I}_1^l} = i\} \Pr\{R_1 = r_1, Q_{\mathbb{I}_1^r} = j | Q_{\mathbb{I}_1^l} = i\} \end{aligned}$$

² The theorem gives the stationary distribution, but any other distribution may be used.

$$\begin{aligned}
 &= \sum_{i=0}^m \Pr\{R_0 = r_0, Q_{i_0}^r = i\} \Pr\{R_1 = r_1, Q_{i_1}^r = j | Q_{i_1}^l = i\} \\
 &= \left(\boldsymbol{\pi} \mathbf{M}_{i_0}(\mathbf{r}) \mathbf{M}_{i_1}(\mathbf{r}) \right) [j]
 \end{aligned}$$

Hence, S_1 is true. Now assume that the induction hypothesis S_{k-1} is true and observe that

$$\begin{aligned}
 \Pr\{R_0 = r_0, R_1 = r_1, \dots, R_k = r_k, Q_{i_k}^r = j\} &= \sum_{i=0}^m \Pr\{R_0 = r_0, R_1 = r_1, \dots, R_k = r_k, Q_{i_k}^l = i, Q_{i_k}^r = j\} \\
 &= \sum_{i=0}^m \Pr\{R_0 = r_0, \dots, R_{k-1} = r_{k-1}, Q_{i_k}^l = i\} \Pr\{R_k = r_k, Q_{i_k}^r = j | R_0 = r_0, \dots, R_{k-1} = r_{k-1}, Q_{i_k}^l = i\} \\
 &= \sum_{i=0}^m \Pr\{R_0 = r_0, \dots, R_{k-1} = r_{k-1}, Q_{i_k}^l = i\} \Pr\{R_k = r_k, Q_{i_k}^r = j | Q_{i_k}^l = i\} \\
 &= \sum_{i=0}^m \Pr\{R_0 = r_0, \dots, R_{k-1} = r_{k-1}, Q_{i_{k-1}}^r = i\} \Pr\{R_k = r_k, Q_{i_k}^r = j | Q_{i_k}^l = i\} \\
 &= \left(\boldsymbol{\pi} \mathbf{M}_{i_0}(\mathbf{r}) \dots \mathbf{M}_{i_{k-1}}(\mathbf{r}) \mathbf{M}_{i_k}(\mathbf{r}) \right) [j]
 \end{aligned}$$

where the last equality follows from the induction hypothesis. Since S_1 is true and S_{k-1} implies S_k , it follows that S_{n-1} must be true. Hence,

$$\Pr\{\mathbf{R} = \mathbf{r}\} = \sum_{j=0}^m \Pr\{\mathbf{R} = \mathbf{r}, Q_{i_{n-1}}^r = j\} = \boldsymbol{\pi} \mathbf{M}_{i_0}(\mathbf{r}) \mathbf{M}_{i_1}(\mathbf{r}) \mathbf{M}_{i_2}(\mathbf{r}) \dots \mathbf{M}_{i_{n-1}}(\mathbf{r}) \mathbf{1}^T = \boldsymbol{\pi} \left(\prod_{k=0}^{n-1} \mathbf{M}_{i_k}(\mathbf{r}) \right) \mathbf{1}^T$$

which concludes the induction proof.

The next step is to express the $\mathbf{M}_{i_k}(\mathbf{r})$ matrices in terms of the model parameters. I will do this by writing $\mathbf{M}_{i_k}(\mathbf{r})$ as a linear combination of the matrices $\mathbf{G}_{i_k}$ and $\mathbf{H}_{i_k}$ with elements

$$\mathbf{G}_{i_k}[i, j] = \Pr\{X_{i_k} > 0, Q_{i_k}^r = j | Q_{i_k}^l = i\}$$

$$\mathbf{H}_{i_k}[i, j] = \Pr\{X_{i_k} = 0, Q_{i_k}^r = j | Q_{i_k}^l = i\}$$

and then find such expressions for these matrices. For this end, I use Mather's (1938) equation

$$\Pr\{R(i) = 1 | X_i = x\} = \begin{cases} 1/2, & x = 1, 2, 3 \dots \\ 0, & x = 0 \end{cases}$$

from which it follows that

$$\Pr\{R_k = 1, Q_{i_k}^r = j | Q_{i_k}^l = i\} = \frac{1}{2} \Pr\{X_{i_k} > 0, Q_{i_k}^r = j | Q_{i_k}^l = i\}$$

$$\Pr\{R_k = 0, Q_{i_k}^r = j | Q_{i_k}^l = i\} = \frac{1}{2} \Pr\{X_{i_k} > 0, Q_{i_k}^r = j | Q_{i_k}^l = i\} + \Pr\{X_{i_k} = 0, Q_{i_k}^r = j | Q_{i_k}^l = i\}$$

so

$$\mathbf{M}_{\mathbb{I}_k}(\mathbf{r}) = \begin{cases} \frac{1}{2} \mathbf{G}_{\mathbb{I}_k}, & r_k = 1 \\ \frac{1}{2} \mathbf{G}_{\mathbb{I}_k} + \mathbf{H}_{\mathbb{I}_k}, & r_k = 0 \end{cases} \quad \text{for } k = 0, 1, 2 \dots n-1$$

Element i, j of $\mathbf{H}_{\mathbb{I}_k}$ is equal to the probability of observing no chiasma events (of either type) in $\mathbb{I}_k$ and phase j at the right boundary, given phase i at the left boundary, which in the case $i \geq j$ is equivalent to the probability of observing $i - j$ C'' events multiplied by the probability of observing 0 X'' events. If $i < j$, it is impossible to arrive at phase j without at least one X'' event in the interval, so the probability is in this case 0. Hence,

$$\mathbf{H}_{\mathbb{I}_k}[i, j] = \begin{cases} \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!} e^{-\mu_k}, & i \geq j \\ 0, & i < j \end{cases} \quad \text{for } i, j = 0, 1, 2 \dots m$$

To find an expression for $\mathbf{G}_{\mathbb{I}_k}[i, j]$, I first define the functions $\psi_k(i, j)$ as

$$\psi_k(i, j) = \Pr \{X''_{\mathbb{I}_k} > 0, Q_{\mathbb{I}_k}^r = j | Q_{\mathbb{I}_k}^l = i\}$$

and b_n as

$$b_n = \begin{cases} \sum_{q=0}^{n-1} b_q \gamma_{n-1-q}, & n = 1, 2, 3 \dots \\ 1, & n = 0 \end{cases}$$

Note that b_n , equivalent to u_n in Lange et al. (1997), gives the probability that the n -th C'' event to the right of any X'' event is also a X'' event; as in Lange et al., it is derived by recursively conditioning on the index of the last X'' event before the n -th C'' event, with the base case $b_0 = 1$. To get at least one $X''_{\mathbb{I}_k}$ event and phase j at the right boundary, given phase i at the left boundary, one must observe in interval $\mathbb{I}_k$ (i O'' events to get to the first X'' event) + (the first X'' event) + (n additional C'' events of which the last is a X'' event) + ($q - j$ additional O'' events, where q is the number of O'' events in the rightmost cycle of the interval, to end up in phase j), in total $i + 1 + n + q - j$ C'' events. Hence, by summing over all possible values of n and q ,

$$\psi_k(i, j) = \sum_{n=0}^{\infty} b_n \sum_{q=j}^m \gamma_q \frac{\lambda_k^{i+1+n+q-j} e^{-\lambda_k}}{(i+1+n+q-j)!} = \Pr \{X''_{\mathbb{I}_k} > 0, Q_{\mathbb{I}_k}^r = j | Q_{\mathbb{I}_k}^l = i\}$$

which is similar to the $1_{\{j>0\}}$ term in equation 6 in Lange et al. (1997) (though note that my 'phase' is defined differently from their 'state'). By conditioning on the presence ($X'_{\mathbb{I}_k} > 0$) or absence ($X'_{\mathbb{I}_k} = 0$) of X' events in the interval,

$$\Pr \{X_{\mathbb{I}_k} > 0, Q_{\mathbb{I}_k}^r = j | Q_{\mathbb{I}_k}^l = i\} = \Pr \{X''_{\mathbb{I}_k} > 0, Q_{\mathbb{I}_k}^r = j | Q_{\mathbb{I}_k}^l = i\} \Pr \{X'_{\mathbb{I}_k} = 0\} + \Pr \{Q_{\mathbb{I}_k}^r = j | Q_{\mathbb{I}_k}^l = i\} \Pr \{X'_{\mathbb{I}_k} > 0\}$$

Since

$$\begin{aligned}\Pr\{Q_{ik}^r = j | Q_{ik}^l = i\} &= \Pr\{X_{ik}'' > 0, Q_{ik}^r = j | Q_{ik}^l = i\} + \Pr\{X_{ik}'' = 0, Q_{ik}^r = j | Q_{ik}^l = i\} \\ &= \psi_k(i, j) + \delta_{\{i \geq j\}} \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!}\end{aligned}$$

and

$$\begin{aligned}\Pr\{X_{ik}' > 0\} &= 1 - e^{-\mu_k} \\ \Pr\{X_{ik}' = 0\} &= e^{-\mu_k}\end{aligned}$$

we now have that

$$\Pr\{X_{ik} > 0, Q_{ik}^r = j | Q_{ik}^l = i\} = \psi_k(i, j) e^{-\mu_k} + \left(\psi_k(i, j) + \delta_{\{i \geq j\}} \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!} \right) (1 - e^{-\mu_k}) = \mathbf{G}_{ik}[i, j]$$

QED

3.3 Theorem 3: A closed-form version of the $\mathbf{G}$ matrices

3.3.1 Prologue

Nolan (2017) derives a closed-form expression for the recombination pattern probabilities under the counting model. Building on the results from that paper, I will in this theorem do the same for the more general two-pathway counting model.

3.3.2 Theorem

In the special case where $\gamma_m = 1, \gamma_q = 0$ for $q \neq m$ (i.e. the two-pathway counting model), the matrix $\mathbf{G}_{ik}$ can be written in closed form as

$$\mathbf{G}_{ik}[i, j] = \psi_k(i, j) e^{-\mu_k} + \left(\psi_k(i, j) + \delta_{\{i \geq j\}} \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!} \right) (1 - e^{-\mu_k})$$

where

$$\begin{aligned}\psi_k(i, j) &= e^{-\lambda_k} \left(\delta_{\{i \geq j\}} \left[f_{i-j, m+1}(\lambda_k) - \frac{\lambda_k^{i-j}}{(i-j)!} \right] + \delta_{\{i < j\}} f_{m+1+i-j, m+1}(\lambda_k) \right) \\ f_{r, q}(\lambda) &= \frac{1}{q} \sum_{j=0}^{q-1} e^{\lambda \cos(\frac{2\pi j}{q})} \cos \left[\lambda \sin \left(\frac{2\pi j}{q} \right) - \frac{2\pi r j}{q} \right] \quad \text{for } r = 0, 1, 2, \dots, q-1\end{aligned}$$

Proof:

The key to this theorem is the relation

$$\sum_{k=0}^{\infty} \frac{\lambda^{qk+r}}{(qk+r)!} = \frac{1}{q} \sum_{j=0}^{q-1} e^{\lambda \cos(\frac{2\pi j}{q})} \cos \left[\lambda \sin \left(\frac{2\pi j}{q} \right) - \frac{2\pi r j}{q} \right] \quad \text{for } r = 0, 1, 2, 3, \dots, q-1$$

which is derived in Erdelyi (1955) and Nolan (2017). If $\gamma_m = 1, \gamma_q = 0$ for $q \neq m$, then $\psi_k(i, j)$ reduces to

$$\psi_k(i, j) = \sum_{c=0}^{\infty} \frac{\lambda_k^{c(m+1)+m+1+i-j} e^{-\lambda_k}}{(c(m+1) + m + 1 + i - j)!}$$

which we can now write as

$$\begin{aligned} \psi_k(i, j) &= \sum_{c=0}^{\infty} \frac{\lambda_k^{c(m+1)+m+1+i-j} e^{-\lambda_k}}{(c(m+1) + m + 1 + i - j)!} \\ &= \delta_{\{i \geq j\}} \left(\sum_{c=0}^{\infty} \frac{\lambda_k^{c(m+1)+i-j} e^{-\lambda_k}}{(c(m+1) + i - j)!} - \frac{\lambda_k^{i-j} e^{-\lambda_k}}{(i-j)!} \right) + \delta_{\{i < j\}} \left(\sum_{c=0}^{\infty} \frac{\lambda_k^{c(m+1)+m+1+i-j} e^{-\lambda_k}}{(c(m+1) + m + 1 + i - j)!} \right) \\ &= e^{-\lambda_k} \left(\delta_{\{i \geq j\}} \left[f_{i-j, m+1}(\lambda_k) - \frac{\lambda_k^{i-j}}{(i-j)!} \right] + \delta_{\{i < j\}} f_{m+1+i-j, m+1}(\lambda_k) \right) \end{aligned}$$

where

$$f_{r,q}(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^{qk+r}}{(qk+r)!} = \frac{1}{q} \sum_{j=0}^{q-1} e^{\lambda \cos\left(\frac{2\pi j}{q}\right)} \cos\left[\lambda \sin\left(\frac{2\pi j}{q}\right) - \frac{2\pi r j}{q}\right] \quad \text{for } r = 0, 1, 2 \dots q-1$$

(note that f is not defined for $r \geq q$)

QED

3.4 Theorem 4: The sterility of standard inversion heterokaryotypes

The sterility of an individual heterozygous for a standard inversion is given by

$$\zeta_s = \Pr\{R(\mathbb{h}) = 1\} = \frac{1}{2} \left(1 - e^{-d_{\mathbb{h}} \mu_{\mathbb{h}}} \sum_{q=0}^m \pi_q \sum_{c=0}^q \frac{e^{-d_{\mathbb{h}} \lambda_{\mathbb{h}}} (d_{\mathbb{h}} \lambda_{\mathbb{h}})^c}{c!} \right)$$

Proof:

For standard inversion heterokaryotypes, a chromatid is, as illustrated in figure 2.2, unbalanced if and only if it shows recombination in the inverted region. Since by assumption all chromatids, unbalanced or not, have an equal chance of becoming gametes, the sterility, ζ , of such a chromosome is simply given by $\zeta = \Pr\{R(\mathbb{h}) = 1\}$. The rest follows from the discussion in theorem 1.

QED

3.5 Theorem 5: Recombination in standard inversion heterokaryotypes

The probability of observing recombination pattern r on a chromosome with a standard inversion is given by

$$\Pr\{R_s = r\} = \pi \left(\prod_{k=0}^{b-1} M_{\mathbb{d}_k}(r) \right) Q^{1-\alpha} \left(\prod_{k=b}^{b+h-1} M_{\mathbb{h}_k}(r) \right) Q^{1-\alpha} \left(\prod_{k=b+h}^{n-1} M_{\mathbb{r}_k}(r) \right) \mathbf{1}^T$$

where

$$Q[i, j] = \pi[j] = \pi_j \quad i, j = 0, 1, 2 \dots m$$

and the M matrices are the same as in theorem 1, except that the lambda and mu values are weighted by d_k so that e.g.

$$H_{ik}[i, j] = \begin{cases} \frac{(d_k \lambda_k)^{i-j} e^{-d_k \lambda_k}}{(i-j)!} e^{-(d_k \mu_k)}, & i \geq j \\ 0, & i < j \end{cases} \quad \text{for } i, j = 0, 1, 2 \dots m$$

Proof:

Apart from the inhibition factors (d), the only difference between theorems 1 and 5 is the matrix Q that is inserted into the latter equation at the positions corresponding to the breakpoint boundaries. The parameter α indicate the presence ($\alpha = 1$) or absence ($\alpha = 0$) of chiasma interference across the breakpoint boundaries, so in the former case $Q^{1-\alpha} = Q^0 \equiv I_{m+1}$ (i.e. the identity matrix). In the absence of interference across the breakpoint boundaries ($\alpha = 0$), Q serves to redistribute the phase probabilities according to the stationary distribution (π). We can see why this is so by decomposing Q into the two matrices Q' and Q'' so that

$$Q = Q' Q''$$

where

$$Q'[i, j] = \begin{cases} 1, & j = 0 \\ 0, & j > 0 \end{cases} \quad i, j = 0, 1, 2 \dots m$$

$$Q''[i, j] = \begin{cases} \pi_j, & i = 0 \\ 0, & i > 0 \end{cases} \quad i, j = 0, 1, 2 \dots m$$

That is, Q' maps all phase probabilities (arbitrarily) to phase 0, whereas Q'' maps them from 0 to j in proportion to the stationary probabilities. This is equivalent to making the chiasma events in the three subregions mutually independent.

QED

3.6 Theorem 6: The sterility of paracentric linear inversion heterokaryotypes

3.6.1 Prologue

Navarro et al. (1997) provide approximate expressions for the sterility and gamete proportions in paracentric linear inversion heterokaryotypes under the Poisson and pure counting interference models for a maximum of two loci in either the distal, inverted or proximal region. I will in this theorem and the next present exact infinite series expressions for the sterility and gamete proportions under the general interference model for an indefinite number of loci in each of the regions on the chromosome.

3.6.2 Theorem

The sterility of a paracentric inversion heterokaryotype with linear meiosis is given by

$$\zeta_p = \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} [\pi D_{\mathbb{h}}(x_1) Q^{1-\alpha} D_{\mathbb{p}}(x_2) \mathbf{1}^T] \mathbf{v}_{0,0} T_{\mathbb{h}}^{x_1} T_{\mathbb{p}}^{x_2} \mathbf{w}_u$$

where

$$D_{\mathbb{i}}(0)[i, j] = \begin{cases} \frac{(d_{\mathbb{i}} \lambda_{\mathbb{i}})^{i-j} e^{-d_{\mathbb{i}} \lambda_{\mathbb{i}}}}{(i-j)!} e^{-d_{\mathbb{i}} \mu_{\mathbb{i}}}, & i \geq j \\ 0, & i < j \end{cases} \quad \text{for } i, j = 0, 1, 2 \dots m; \mathbb{i} = \mathbb{h}, \mathbb{p}$$

$$D_{\mathbb{i}}(x)[i, j] = \sum_{l=0}^{x-1} \sum_{n=a}^z \sum_{q=0}^j g_n(x-l-1) \gamma_q \frac{e^{-d_{\mathbb{i}}(\lambda_{\mathbb{i}}+\mu_{\mathbb{i}})} (d_{\mathbb{i}}[\lambda_{\mathbb{i}}+\mu_{\mathbb{i}}])^h}{h!} \binom{h}{l} \left(\frac{\mu_{\mathbb{i}}}{\lambda_{\mathbb{i}}+\mu_{\mathbb{i}}}\right)^l \left(\frac{\lambda_{\mathbb{i}}}{\lambda_{\mathbb{i}}+\mu_{\mathbb{i}}}\right)^{h-l}$$

$$+ \delta_{\{i \geq j\}} \frac{e^{-d_{\mathbb{i}} \mu_{\mathbb{i}}} (d_{\mathbb{i}} \mu_{\mathbb{i}})^x}{x!} \frac{e^{-d_{\mathbb{i}} \lambda_{\mathbb{i}}} (d_{\mathbb{i}} \lambda_{\mathbb{i}})^{i-j}}{(i-j)!}, \quad \text{for } x = 1, 2, 3 \dots; i, j = 0, 1, 2 \dots m; \mathbb{i} = \mathbb{h}, \mathbb{p}$$

$$a = x - l - 1$$

$$z = (x - l - 1)(m + 1)$$

$$h = i + 1 + l + n + q - j$$

$$g_n(s) = \begin{cases} \sum_{k=s-1}^{n-1} g_k(s-1) \gamma_{n-1-k}, & n \geq s \neq 0 \\ 1, & n = s = 0 \\ 0, & n \neq 0; s = 0 \\ 0, & n < s \end{cases}$$

$$\mathbf{v}_{0,0} = (1 \quad 0 \quad 0 \quad 0)$$

$$T_{\mathbb{h}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/4 & 1/2 & 0 & 1/4 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$T_{\mathbb{p}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{w}_u = \begin{pmatrix} 0 \\ 0 \\ 1/2 \\ 1 \end{pmatrix}$$

$$d_{\mathbb{i}} \lambda_{\mathbb{i}} = d_{\mathbb{i}} E[C_{\mathbb{i}}''] \text{ for } \mathbb{i} = \mathbb{h}, \mathbb{p}$$

$$d_{\mathbb{i}} \mu_{\mathbb{i}} = d_{\mathbb{i}} E[X_{\mathbb{i}}'] \text{ for } \mathbb{i} = \mathbb{h}, \mathbb{p}$$

Proof:

The basic structure of this equation is

$$\zeta_p = \Pr\{\text{unbalanced gamete}\} = \sum_{x_1=0}^{\infty} \sum_{x_2=0}^{\infty} \Pr\{X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2\} \Pr\{\text{unbalanced gamete} | X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2\}$$

i.e. it finds the probability of an unbalanced gamete by conditioning on the number of chiasma events in each region. $\Pr\{X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2\} = \boldsymbol{\pi} \mathbf{D}_{\mathbb{h}}(x_1) \mathbf{Q}^{1-\alpha} \mathbf{D}_{\mathbb{p}}(x_2) \mathbf{1}^T$ is a special case of a more general expression that is shown in the proof for theorem 7. I therefore focus here on the conditional probability of observing an unbalanced gamete, $\Pr\{\text{unbalanced gamete} | X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2\}$, which I will derive by means of a Markov chain, as follows. Consider a stochastic process $\{Z(x_1, x_2)\}_{(x_1, x_2) \in \mathbb{Z}_{\geq 0}}$ that represents the configuration of a tetrad when $X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2$ (i.e. when there are x_1 chiasma events in the inverted region and x_2 chiasma events in the proximal region), so that

$$Z(x_1, x_2) = \begin{cases} 0, & \text{no bridge} \\ 1, & \text{single anaphase I bridge} \\ 2, & \text{single anaphase II bridge} \\ 3, & \text{double anaphase I bridge} \\ 4, & \text{double anaphase II bridge} \end{cases}$$

If I now define the matrices $\mathbf{T}_{\mathbb{h}}$ and $\mathbf{T}_{\mathbb{p}}$ so that

$$\begin{aligned} \mathbf{T}_{\mathbb{h}}[i, j] &= \Pr\{Z(x_1 + 1, 0) = j | Z(x_1, 0) = i\}, & x_1 = 0, 1, 2, \dots; i, j = 0, 1, 2, 3, 4 \\ \mathbf{T}_{\mathbb{p}}[i, j] &= \Pr\{Z(x_1, x_2 + 1) = j | Z(x_1, x_2) = i\}, & x_1, x_2 = 0, 1, 2, \dots; i, j = 0, 1, 2, 3, 4 \end{aligned}$$

and the vector $\mathbf{v}_{0,0}$ so that

$$\mathbf{v}_{0,0} = (\Pr\{Z(0,0) = 0\} \quad \Pr\{Z(0,0) = 1\} \quad \Pr\{Z(0,0) = 2\} \quad \Pr\{Z(0,0) = 3\} \quad \Pr\{Z(0,0) = 4\})$$

then the $\mathbf{T}_{\mathbb{i}}$ matrices are the Markovian transition matrices whose element i, j give the probability of observing a tetrad in configuration j after considering an additional chiasma event at the right end of region $\mathbb{i}$ ($= \mathbb{h}, \mathbb{p}$), given that the tetrad was in configuration i before considering that chiasma event and that there are no other chiasma event further to the right; and $\mathbf{v}_{0,0}$ is the tetrad configuration distribution when there are no chiasma events in either interval. Accordingly,

$$\mathbf{v}_{x_1, x_2} = \mathbf{v}_{0,0} \mathbf{T}_{\mathbb{h}}^{x_1} \mathbf{T}_{\mathbb{p}}^{x_2}, \quad \text{for } x_1, x_2 = 0, 1, 2, \dots$$

where

$$\mathbf{v}_{x_1, x_2} = (\Pr\{Z(x_1, x_2) = 0\} \quad \Pr\{Z(x_1, x_2) = 1\} \quad \Pr\{Z(x_1, x_2) = 2\} \quad \Pr\{Z(x_1, x_2) = 3\} \quad \Pr\{Z(x_1, x_2) = 4\})$$

I now find $\Pr\{\text{unbalanced gamete} | X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2\}$ by weighting each configuration by its proportion of unbalanced gametes, i.e.

$$\Pr\{\text{unbalanced gamete} | X_{\mathbb{h}} = x_1, X_{\mathbb{p}} = x_2\} = \mathbf{v}_{0,0} \mathbf{T}_{\mathbb{h}}^{x_1} \mathbf{T}_{\mathbb{p}}^{x_2} \mathbf{w}_u$$

where

$$\mathbf{w}_u = \begin{pmatrix} 0 \\ 0 \\ 1/2 \\ 1 \\ 1 \end{pmatrix}$$

$\mathbf{T}_h$ and $\mathbf{T}_p$ can be found by considering the resulting tetrads for the four possible combinations of non-sister strand involvements when an extra chiasma event is added at the right end of the appropriate interval of a tetrad that is originally in the given configuration (see figures 6 and 7). This results in

$$\mathbf{T}_h = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 1/2 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{T}_p = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

(Cobbs (1978) uses a similar line of reasoning to find the probability distribution of parental ditype, tetratype, and non-parental ditype tetrads given the number of chiasma events in a single colinear interval; note the similarity between his matrix $\mathbf{T}$ and my $\mathbf{T}_h$.) Since state 3 and 4 are identical for our purpose and form an absorbing class in $\mathbf{T}_p$, we can merge them into a single class *double bridge* to get the alternative (equivalent) expressions for $\mathbf{v}_{0,0}$, $\mathbf{T}_h$, $\mathbf{T}_p$ and $\mathbf{w}_u$ given in the theorem.

Figure 6: An inversion loop for a paracentric inversion. A single chiasma in the inverted region creates a tetrad with an anaphase I bridge and an acentric fragment (meaning that it is not connected to the centromere) which fails to segregate.

Figure 7: The leftmost X represent an anaphase I bridge, and the X 's marked a, b, c, d represent the four possible non-sister chromatid pair involvements in a single chiasma event. The resulting configurations for each pair is shown below. The figure shows that $\Pr\{Z(x_1 + 1, 0) = 0 | Z(x_1, 0) = 1\} = 1/4$, $\Pr\{Z(x_1 + 1, 0) = 1 | Z(x_1, 0) = 1\} = 1/2$, and $\Pr\{Z(x_1 + 1, 0) = 3 | Z(x_1, 0) = 1\} = 1/4$, which gives the second row (zero-indexed row 1) in matrix $\mathbf{T}_{\mathbb{N}}$. The remaining rows are found in the same manner.

Figure 8: The effect of a single additional chiasma in the proximal region when the tetrad is originally in the configuration single anaphase I bridge. The figure shows that $\Pr\{Z(x_1, x_2 + 1) = 1 | Z(x_1, x_2) = 1\} = 1/2$ and $\Pr\{Z(x_1, x_2 + 1) = 2 | Z(x_1, x_2) = 1\} = 1/2$, which give the second row (zero-indexed row 1) of matrix $\mathbf{T}_{\mathbb{P}}$. The remaining rows are found in the same manner.

3.7 Theorem 7: Recombination in paracentric heterokaryotypes with linear meiosis

3.7.1 Prologue

For this theorem we will need some additional terminology. I will say that a chromatid is in *state* s_j : $\mathbf{r}_{s_j} = \mathbf{r}$, $t_{s_j} = t$ if it shows recombination pattern $\mathbf{r}$ and is in a tetrad with configuration t , and I will refer to

$\mathbf{r}_{s_j}$ as the recombination pattern associated with state s_j , and t_{s_j} as the tetrad configuration associated with state s_j . The *statespace* is hence the set of all possible combinations of recombination patterns and tetrad configuration, except that all states with unbalanced recombination patterns that also show recombination in any interval with index higher than $\mathfrak{d} + \mathfrak{h} - 1$ (i.e. the intervals in the proximal and rightmost regions) are excluded. As an example, with two intervals in the inverted region and one in the proximal region and zero intervals in the distal and rightmost regions (i.e. $\mathfrak{d} = 0, \mathfrak{h} = 2, \mathfrak{p} = 1, \mathfrak{r} = 0$), which means that there is one loci of interest in the inverted region and none in either of the other regions (note that this implies that $\mathfrak{i}_2 = \mathfrak{p}_2 = \mathfrak{p}$), we get the (arbitrarily numbered) statespace:

$$\begin{aligned} s_0: \mathbf{r}_{s_0} &= \{r_{s_0}(\mathfrak{h}_0) = 0, r_{s_0}(\mathfrak{h}_1) = 0, r_{s_0}(\mathfrak{p}_2) = 0\}, t_{s_0} = 0 \\ s_1: \mathbf{r}_{s_1} &= \{r_{s_1}(\mathfrak{h}_0) = 0, r_{s_1}(\mathfrak{h}_1) = 0, r_{s_1}(\mathfrak{p}_2) = 0\}, t_{s_1} = 1 \\ s_2: \mathbf{r}_{s_2} &= \{r_{s_2}(\mathfrak{h}_0) = 0, r_{s_2}(\mathfrak{h}_1) = 0, r_{s_2}(\mathfrak{p}_2) = 0\}, t_{s_2} = 2 \\ s_3: \mathbf{r}_{s_3} &= \{r_{s_3}(\mathfrak{h}_0) = 0, r_{s_3}(\mathfrak{h}_1) = 0, r_{s_3}(\mathfrak{p}_2) = 1\}, t_{s_3} = 0 \\ s_4: \mathbf{r}_{s_4} &= \{r_{s_4}(\mathfrak{h}_0) = 0, r_{s_4}(\mathfrak{h}_1) = 0, r_{s_4}(\mathfrak{p}_2) = 1\}, t_{s_4} = 1 \\ s_5: \mathbf{r}_{s_5} &= \{r_{s_5}(\mathfrak{h}_0) = 0, r_{s_5}(\mathfrak{h}_1) = 0, r_{s_5}(\mathfrak{p}_2) = 1\}, t_{s_5} = 2 \\ s_6: \mathbf{r}_{s_6} &= \{r_{s_6}(\mathfrak{h}_0) = 0, r_{s_6}(\mathfrak{h}_1) = 1, r_{s_6}(\mathfrak{p}_2) = 0\}, t_{s_6} = 1 \\ s_7: \mathbf{r}_{s_7} &= \{r_{s_7}(\mathfrak{h}_0) = 0, r_{s_7}(\mathfrak{h}_1) = 1, r_{s_7}(\mathfrak{p}_2) = 0\}, t_{s_7} = 2 \\ s_8: \mathbf{r}_{s_8} &= \{r_{s_8}(\mathfrak{h}_0) = 0, r_{s_8}(\mathfrak{h}_1) = 1, r_{s_8}(\mathfrak{p}_2) = 0\}, t_{s_8} = 3 \\ s_9: \mathbf{r}_{s_9} &= \{r_{s_9}(\mathfrak{h}_0) = 1, r_{s_9}(\mathfrak{h}_1) = 0, r_{s_9}(\mathfrak{p}_2) = 0\}, t_{s_9} = 1 \\ s_{10}: \mathbf{r}_{s_{10}} &= \{r_{s_{10}}(\mathfrak{h}_0) = 1, r_{s_{10}}(\mathfrak{h}_1) = 0, r_{s_{10}}(\mathfrak{p}_2) = 0\}, t_{s_{10}} = 2 \\ s_{11}: \mathbf{r}_{s_{11}} &= \{r_{s_{11}}(\mathfrak{h}_0) = 1, r_{s_{11}}(\mathfrak{h}_1) = 0, r_{s_{11}}(\mathfrak{p}_2) = 0\}, t_{s_{11}} = 3 \\ s_{12}: \mathbf{r}_{s_{12}} &= \{r_{s_{12}}(\mathfrak{h}_0) = 1, r_{s_{12}}(\mathfrak{h}_1) = 1, r_{s_{12}}(\mathfrak{p}_2) = 0\}, t_{s_{12}} = 0 \\ s_{13}: \mathbf{r}_{s_{13}} &= \{r_{s_{13}}(\mathfrak{h}_0) = 1, r_{s_{13}}(\mathfrak{h}_1) = 1, r_{s_{13}}(\mathfrak{p}_2) = 0\}, t_{s_{13}} = 1 \\ s_{14}: \mathbf{r}_{s_{14}} &= \{r_{s_{14}}(\mathfrak{h}_0) = 1, r_{s_{14}}(\mathfrak{h}_1) = 1, r_{s_{14}}(\mathfrak{p}_2) = 0\}, t_{s_{14}} = 2 \\ s_{15}: \mathbf{r}_{s_{15}} &= \{r_{s_{15}}(\mathfrak{h}_0) = 1, r_{s_{15}}(\mathfrak{h}_1) = 1, r_{s_{15}}(\mathfrak{p}_2) = 1\}, t_{s_{15}} = 0 \\ s_{16}: \mathbf{r}_{s_{16}} &= \{r_{s_{16}}(\mathfrak{h}_0) = 1, r_{s_{16}}(\mathfrak{h}_1) = 1, r_{s_{16}}(\mathfrak{p}_2) = 1\}, t_{s_{16}} = 1 \\ s_{17}: \mathbf{r}_{s_{17}} &= \{r_{s_{17}}(\mathfrak{h}_0) = 1, r_{s_{17}}(\mathfrak{h}_1) = 1, r_{s_{17}}(\mathfrak{p}_2) = 1\}, t_{s_{17}} = 2 \end{aligned}$$

where

$$t = \begin{cases} 0, & \text{no bridge} \\ 1, & \text{single anaphase I bridge} \\ 2, & \text{single anaphase II bridge} \\ 3, & \text{double bridge} \end{cases}$$

Note that there are no states for unbalanced patterns in *no bridge* tetrads or balanced patterns in *double bridge* tetrads; this is because if the tetrad is in the *no bridge* configuration, then none of the patterns associated with the tetrad can be unbalanced, and, similarly, if the tetrad is in the *double bridge* configuration, then none of the patterns associated with the tetrad can be balanced, so these combinations are impossible and hence not included in the statespace. The number of states – or, equivalently, the *size* of the statespace – will henceforth be denoted $\mathfrak{s}$; in the example above $\mathfrak{s} = 18$.

3.7.2 Theorem

Let the statespace be as defined above. The probability of observing a gamete with recombination pattern $\mathbf{r}$ for a paracentric linear inversion heterokaryotype is now given by

$$\Pr\{\mathbf{R}_p = \mathbf{r}\} = \mathbf{v}\mathbf{w}_r^T$$

where

$$\mathbf{w}_r[i] = \begin{cases} 2, & \mathbf{r}_{s_i} = \mathbf{r}; \varphi(\mathbf{r}_{s_i}) = 1; t_{s_i} = 1 \\ 0, & \mathbf{r}_{s_i} = \mathbf{r}; \varphi(\mathbf{r}_{s_i}) = 0; t_{s_i} = 1 \\ 1, & \mathbf{r}_{s_i} = \mathbf{r}; t_{s_i} \neq 1 \\ 0, & \mathbf{r}_{s_i} \neq \mathbf{r} \end{cases} \quad \text{for } i = 0, 1, 2 \dots s-1$$

$$\varphi(\mathbf{r}) = \begin{cases} 1, & \sum_{i=\mathfrak{d}}^{\mathfrak{d}+\mathfrak{h}-1} r(\mathbb{h}_i) \text{ is even} \\ 0, & \sum_{i=\mathfrak{d}}^{\mathfrak{d}+\mathfrak{h}-1} r(\mathbb{h}_i) \text{ is odd} \end{cases}$$

$$\mathbf{v} = \sum_{x_0=0}^{\infty} \sum_{x_1=0}^{\infty} \dots \sum_{x_{n-1}=0}^{\infty} \mathbb{D}(x_0, x_1 \dots x_{n-1}) \mathbf{v}_{x_0, x_1 \dots x_{n-1}}$$

$$\mathbb{D}(x_0, x_1 \dots x_{n-1}) = \boldsymbol{\pi} \left(\prod_{k=0}^{\mathfrak{d}-1} \mathbf{D}_{\mathfrak{d}_k}(x_k) \right) \mathbf{Q}^{1-\alpha} \left(\prod_{k=\mathfrak{d}}^{\mathfrak{d}+\mathfrak{h}-1} \mathbf{D}_{\mathbb{h}_k}(x_k) \right) \mathbf{Q}^{1-\alpha} \left(\prod_{k=\mathfrak{d}+\mathfrak{h}}^{\mathfrak{d}+\mathfrak{h}+\mathfrak{p}-1} \mathbf{D}_{\mathbb{p}_k}(x_k) \right) \left(\prod_{k=\mathfrak{d}+\mathfrak{h}+\mathfrak{p}}^{n-1} \mathbf{D}_{\mathfrak{r}_k}(x_k) \right) \mathbf{1}^T$$

$$\boldsymbol{\pi} = (\pi_0 \quad \pi_1 \quad \pi_2 \quad \dots \quad \pi_m)$$

$$\pi_l = \frac{\sum_{q=l}^m \gamma_q}{\sum_{q=0}^m (q+1) \gamma_q} \quad \text{for } l = 0, 1, 2 \dots m$$

$$\mathbf{Q}[i,j] = \pi_j \quad \text{for } i, j = 0, 1, 2 \dots m$$

$$\mathbf{v}_{x_0, x_1 \dots x_{n-1}} = \mathbf{v}_{0,0,\dots,0} \left(\prod_{k=0}^{\mathfrak{d}-1} \mathbf{P}_{\mathfrak{d}_k}^{x_k} \right) \left(\prod_{k=\mathfrak{d}}^{\mathfrak{d}+\mathfrak{h}-1} \mathbf{P}_{\mathbb{h}_k}^{x_k} \right) \left(\prod_{k=\mathfrak{d}+\mathfrak{h}}^{\mathfrak{d}+\mathfrak{h}+\mathfrak{p}-1} \mathbf{P}_{\mathbb{p}_k}^{x_k} \right) \left(\prod_{k=\mathfrak{d}+\mathfrak{h}+\mathfrak{p}}^{n-1} \mathbf{P}_{\mathfrak{r}_k}^{x_k} \right)$$

$$\mathbf{D}_{\mathbb{h}_k}(0)[i,j] = \begin{cases} \frac{(d_k \lambda_k)^{i-j} e^{-d_k \lambda_k}}{(i-j)!} e^{-d_k \mu_k}, & i \geq j \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i, j = 0, 1, 2 \dots m$$

$$\begin{aligned} \mathbf{D}_{\mathbb{h}_k}(x)[i,j] &= \sum_{l=0}^{x-1} \sum_{n=a}^z \sum_{q=j}^m g_n(x-l-1) \gamma_q \frac{e^{-d_k(\lambda_k+\mu_k)} (d_k[\lambda_k+\mu_k])^h}{h!} \binom{h}{l} \left(\frac{\mu_k}{\lambda_k+\mu_k} \right)^l \left(\frac{\lambda_k}{\lambda_k+\mu_k} \right)^{h-l} \\ &\quad + \delta_{\{i \geq j\}} \frac{e^{-d_k \mu_k} (d_k \mu_k)^x}{x!} \frac{e^{-d_k \lambda_k} (d_k \lambda_k)^{i-j}}{(i-j)!}, \quad \text{for } x = 1, 2, 3 \dots; i, j = 0, 1, 2 \dots m \end{aligned}$$

$$a = x - l - 1$$

$$z = (x - l - 1)(m + 1)$$

$$h = i + 1 + l + n + q - j$$

$$g_n(s) = \begin{cases} \sum_{k=s-1}^{n-1} g_k(s-1)\gamma_{n-1-k}, & n \geq s \neq 0 \\ 1, & n = s = 0 \\ 0, & n \neq 0; s = 0 \\ 0, & n < s \end{cases}$$

$$\lambda_k = E[C''_{\mathbb{I}_k}]; \mu_k = E[X'_{\mathbb{I}_k}]$$

$$\mathbf{v}_{0,0,\dots,0}[j] = \begin{cases} 1, & r_{s_j}(\mathbb{I}_l) = 0 \text{ for all } l; t_{s_j} = 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } j = 0, 1, 2 \dots s-1$$

$$\mathbf{P}_{\mathbb{I}_k}[i,j] = \begin{cases} \frac{1}{2}, & \theta(i,j,k) = 1; t_{s_i} = t_{s_j} = 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{for } k = 0, 2, 3 \dots \mathfrak{d} - 1; i, j = 0, 1, 2 \dots s-1$$

$$\mathbf{P}_{\mathbb{I}_k}[i,j] = \begin{cases} 1/2, & \theta(i,j,k) = 1; t_{s_i} = 0; t_{s_j} = 1 \\ 1/4, & \theta(i,j,k) = 1; \varphi(\mathbf{r}_{s_j}) = 1; t_{s_i} = 1; t_{s_j} = 0 \\ 1/4, & \theta(i,j,k) = 1; t_{s_i} = 1; t_{s_j} = 1 \\ 1/4, & \theta(i,j,k) = 1; \varphi(\mathbf{r}_{s_j}) = 0; t_{s_i} = 1; t_{s_j} = 3 \\ 1/2, & \theta(i,j,k) = 1; t_{s_i} = 3; t_{s_j} = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{for } k = \mathfrak{d} + 1, \mathfrak{d} + 2 \dots \mathfrak{d} + \mathfrak{h} - 1; i, j = 0, 1, 2 \dots s-1$$

$$\mathbf{P}_{\mathbb{P}_k}[i,j] = \begin{cases} 1/2, & \theta(i,j,k) = 1; t_{s_i} = t_{s_j} = 0 \\ 1/2, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 0; r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) \text{ for all } l; t_{s_i} = t_{s_j} = 1 \\ 1/2, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 0; r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) \text{ for all } l; t_{s_i} = 1; t_{s_j} = 2 \\ 1/4, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 1; \theta(i,j,k) = 1; t_{s_i} = t_{s_j} = 1 \\ 1/4, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 1; \theta(i,j,k) = 1; t_{s_i} = 1; t_{s_j} = 2 \\ 1, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 0; r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) \text{ for all } l; t_{s_i} = 2; t_{s_j} = 1 \\ 1/2, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 1; \theta(i,j,k) = 1; t_{s_i} = 2; t_{s_j} = 1 \\ 1, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 0; r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) \text{ for all } l; t_{s_i} = t_{s_j} = 3 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{for } k = \mathfrak{d} + \mathfrak{h}, \mathfrak{d} + \mathfrak{h} + 1 \dots \mathfrak{d} + \mathfrak{h} + \mathfrak{p} - 1; i, j = 0, 1, 2 \dots s-1$$

$$\mathbf{P}_{\mathbb{R}_k}[i,j] = \begin{cases} 1, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 0; r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) \text{ for all } l; t_{s_i} = t_{s_j} \\ 1/2, & \varphi(\mathbf{r}_{s_i}) = \varphi(\mathbf{r}_{s_j}) = 1; \theta(i,j,k) = 1; t_{s_i} = t_{s_j} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{for } k = \mathfrak{d} + \mathfrak{h} + \mathfrak{p}, \mathfrak{d} + \mathfrak{h} + \mathfrak{p} + 1 \dots n-1; i, j = 0, 1, 2 \dots s-1$$

$$\theta(i, j, k) = \begin{cases} 1, & r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) \text{ for } l \neq k; r_{s_i}(\mathbb{I}_l) = r_{s_j}(\mathbb{I}_l) = 0 \text{ for } l > k \\ 0, & \text{otherwise} \end{cases}$$

Proof:

Consider first a stochastic process $\{S(x_0, x_1, x_2 \dots x_{n-1})\}_{(x_0, x_1, x_2 \dots x_{n-1}) \in \mathbb{Z}_{\geq 0}}$ where $S(x_0, x_1, x_2 \dots x_{n-1})$ represents the state of a chromatid (recombination pattern and tetrad configuration) when there are x_0 chiasma events in $\mathbb{I}_0$, x_1 chiasma events in $\mathbb{I}_1$, etc, so that

$$S(x_0, x_1, x_2 \dots x_{n-1}) = j, \quad \text{the process is in state } s_j; \text{ for } j = 0, 1, 2 \dots s-1$$

If we now define the vector $\mathbf{v}_{0,0,\dots,0}$ so that

$$\mathbf{v}_{0,0,\dots,0}[j] = \Pr\{S(0,0, \dots, 0) = j\} \quad j = 0, 1, 2 \dots s-1$$

and the matrices $\mathbf{P}_{\mathbb{I}_k}$, $k = 0, 1, 2 \dots n-1$; $\mathbb{I} = \mathbb{A}, \mathbb{B}, \mathbb{P}, \mathbb{R}$, so that

$$\begin{aligned} \mathbf{P}_{\mathbb{I}_k}[i, j] &= \Pr\{S(x_0, x_1, \dots, x_k + 1, 0, 0 \dots, 0) = j | S((x_0, x_1, \dots, x_k, 0, 0 \dots, 0)) = i\} \\ \text{for } x_0, x_1 \dots x_k &= 0, 1, 2 \dots; i, j = 0, 1, 2 \dots s-1 \end{aligned}$$

then

$$\mathbf{v}_{x_0, x_1, \dots, x_{n-1}} = \mathbf{v}_{0,0,\dots,0} \mathbf{P}_{\mathbb{I}_0}^{x_0} \mathbf{P}_{\mathbb{I}_1}^{x_1} \dots \mathbf{P}_{\mathbb{I}_{n-1}}^{x_{n-1}}$$

where

$$\mathbf{v}_{x_0, x_1, \dots, x_{n-1}}[j] = \Pr\{S(x_0, x_1, \dots, x_{n-1}) = j\} \quad \text{for } j = 0, 1, 2 \dots s-1; x_0, x_1, \dots, x_{n-1} = 0, 1, 2 \dots$$

i.e. the $\mathbf{P}_{\mathbb{I}_k}$ matrices are transition matrices in the same sense that the $\mathbf{T}_{\mathbb{I}}$ matrices in theorem 6 are, and they are found in the same manner. Hence,

$$\mathbf{v} = \sum_{x_0=0}^{\infty} \sum_{x_1=0}^{\infty} \dots \sum_{x_{n-1}=0}^{\infty} \Pr\{X_{\mathbb{I}_0} = x_0, X_{\mathbb{I}_1} = x_1, \dots, X_{\mathbb{I}_{n-1}} = x_{n-1}\} \mathbf{v}_{x_0, x_1 \dots x_{n-1}}$$

gives the unconditional state distribution. Because unbalanced patterns in anaphase I tetrads are retained in the polar bodies, balanced patterns in anaphase I tetrads are weighted by 2 and unbalanced patterns in anaphase I tetrads are weighted by 0.³ So

$$\Pr\{\mathbf{R}_p = \mathbf{r}\} = \mathbf{v} \mathbf{w}_r^T$$

where

³ If this is not clear, notice that eliminating unbalanced anaphase I patterns is probabilistically equivalent to turning the two unbalanced pattern chromatids of an anaphase I tetrad into copies (one of each) of the two balanced pattern chromatids in the same tetrad, and then drawing randomly (uniformly) from the resulting four balanced pattern chromatids.

$$w_r[i] = \begin{cases} 2, & r_{s_i} = r; \varphi(r_{s_i}) = 1; t_{s_i} = 1 \\ 0, & r_{s_i} = r; \varphi(r_{s_i}) = 0; t_{s_i} = 1 \\ 1, & r_{s_i} = r; t_{s_i} \neq 1 \\ 0, & r_{s_i} \neq r \end{cases} \quad \text{for } i = 0, 1, 2 \dots s-1$$

Note that the function

$$\varphi(r) = \begin{cases} 1, & \sum_{i=b}^{b+h-1} r(\mathbb{I}_i) \text{ is even} \\ 0, & \sum_{i=b}^{b+h-1} r(\mathbb{I}_i) \text{ is odd} \end{cases}$$

is 1 if the pattern is balanced and 0 if the pattern is unbalanced (see figure 2.5 for an induction proof).

Figure 9: An induction proof of the statement (S) that a chromatid is unbalanced if and only if it shows recombination in an odd number of intervals within the inverted region. Keep in mind that an interval shows recombination if and only if it is involved in an odd number of chiasma events within the interval, and that a chromatid (in a heterokaryotype) is unbalanced if and only if it show recombination in the inverted region. Note first that the statement is trivially true when the inverted region consists of only a single interval (S_1). Now assume for the sake of argument that the boundaries of the inverted region can be moved to include additional intervals, and that the statement is true when the inverted region consist of $n = k$ intervals (S_k). Hence, we must consider two scenarios: either the chromatid is unbalanced (involved in an odd number of chiasma events within the inverted region), which, the induction hypothesis states, must mean that it also shows recombination in an odd number of intervals within the inverted region (scenario 1), or the chromatid is balanced (involved in an even number of chiasma events within the inverted region), meaning that it show recombination in an even number of intervals within the inverted region (scenario 2). The figure shows the effect of including an additional recombinant (dashed arrow) or non-recombinant (solid arrow) interval in both of these scenarios (a: unbalanced/balanced chromatid, b: odd/even number of recombinant interval in the inverted region). Note that the statement is true for $n = k + 1$. Hence, since S_1 is true and S_k implies S_{k+1} , the statement must be true for all $n > 0$.

The next step is to find expressions for the probability of observing a given number of chiasma events in each interval. Using an induction argument similar to the one in the proof for theorem 1 gives

$$\Pr\{X_{\mathbb{I}_0} = x_0, X_{\mathbb{I}_1} = x_1, \dots, X_{\mathbb{I}_k} = x_{\mathbb{I}_k}, Q_{\mathbb{I}_k^r} = j\} = (\pi D_{\mathbb{I}_0}(x_0) D_{\mathbb{I}_1}(x_1) \dots D_{\mathbb{I}_k}(x_k)) [j]$$

where

$$D_{\mathbb{I}_k}(x)[i, j] = \Pr\{X_{\mathbb{I}_k} = x, Q_{\mathbb{I}_k^r} = j | Q_{\mathbb{I}_k^l} = i\}$$

Note that $D_{ik}(0)$ is equivalent to H_{ik} in theorem 1. To get an expression for $D_{ik}(x)$, $x > 0$, first make the following observations. If there are in total x chiasma events in the interval, then between 0 and x of them must be X' events. If there are l ($0 \leq l < x$) X' events in the interval, then we can express the total number of C events, h , as $h = (i \text{ } O'' \text{ events before the first } X'' \text{ event}) + (\text{the first } X'' \text{ event}) + (l \text{ } X' \text{ events [in no particular spatial order]}) + (n \text{ additional } C'' \text{ events up to and including the last } X'' \text{ event}) + (\text{the final } q - j \text{ } O'' \text{ events, where } q \text{ is the phase after the last } X'' \text{ event, so as to end up in phase } j \text{ at the right boundary}),$ in short

$$h = i + 1 + l + n + q - j$$

In the special case where $\gamma_m = 1$, n and q are fixed at $(m + 1)(x - l - 1)$ and m , respectively, so that we can get an expression for $\Pr\{X_{ik} = x, Q_{ik}^r = j | Q_{ik}^l = i\}$ by summing over all possible values of l , i.e.

$$\begin{aligned} \Pr\{X_{ik} = x, Q_{ik}^r = j | Q_{ik}^l = i\} \\ = \sum_{l=0}^{x-1} \Pr\{C_{ik} = h\} \Pr\{X'_{ik} = l | C_{ik} = h\} + \delta_{\{i \geq j\}} \frac{e^{-d_k \mu_k} (d_k \mu_k)^x}{x!} \frac{e^{-d_k \lambda_k} (d_k \lambda_k)^{i-j}}{(i-j)!} \end{aligned}$$

where $\Pr\{C_{ik} = h\}$ and $\Pr\{X'_{ik} = l | C_{ik} = h\}$ are given by the Poisson and Binomial distributions, respectively (Copenhaver et al 2002). The final term takes into account the case where all the x chiasma events are X' events (which means that there are no X'' events). This term is non-zero only if $i \geq j$ (otherwise there must be at least one X'' events, which is a contradiction), and is in that case equal to the probability of observing x X' events and $i - j$ O'' events, which is just the Poisson probabilities multiplied. For the general model, we have to also sum over all values of n and q , weighted by their probabilities, i.e.

$$\begin{aligned} \Pr\{X_{ik} = x, Q_{ik}^r = j | Q_{ik}^l = i\} = \sum_{l=0}^{x-1} \sum_{n=a}^z \sum_{q=j}^m g_n(x-l-1) \gamma_q \frac{e^{-d_k(\lambda_k + \mu_k)} (d_k[\lambda_k + \mu_k])^h}{h!} \binom{h}{l} \left(\frac{\mu_k}{\lambda_k + \mu_k}\right)^l \left(\frac{\lambda_k}{\lambda_k + \mu_k}\right)^{h-l} \\ + \delta_{\{i \geq j\}} \frac{e^{-d_k \mu_k} (d_k \mu_k)^x}{x!} \frac{e^{-d_k \lambda_k} (d_k \lambda_k)^{i-j}}{(i-j)!}, \quad \text{for } x = 1, 2, 3 \dots; i, j = 0, 1, 2 \dots m \end{aligned}$$

where $a = x - l - 1$ and $z = (x - l - 1)(m + 1)$. The function $g_n(s)$ gives the probability that the n -th C'' event after an X'' event is the s -th X'' event after that X'' event. This is achieved by defining the base case $g_0(0) = 1$, which means that the zeroth C'' event is the zeroth X'' event with probability 1, and then recursively conditioning on which X'' event is the last before the s -th, i.e.

$$g_n(s) = \begin{cases} \sum_{k=s-1}^{n-1} g_k(s-1) \gamma_{n-1-k}, & n \geq s \neq 0 \\ 1, & n = s = 0 \\ 0, & n \neq 0; s = 0 \\ 0, & n < s \end{cases}$$

Note the difference between $g_n(s)$ and b_n from theorem 1: the former gives the probability that the n -th C'' event after an X'' event is strictly the s -th X'' event after that X'' event, whereas the latter gives the

probability that the n -th C'' event after an X'' event is any X'' event.

QED

4 Discussion

I have in this paper suggested a generalization of the many counting models of chiasma interference, and derived novel expressions for recombination pattern probabilities in homo- and heterokaryotypes for two types of chromosomal inversions. I envision two applications for these expressions. Firstly, the model can be useful in testing simplifying assumptions in analytic studies against the results of more realistic numerical simulations. It is, for example, common in multilocus evolution models to simplify the analysis by assuming haploidy and run numerical diploid simulations to check that they give qualitatively similar results (e.g. Kirkpatrick 1982, Kirkpatrick and Servedio 1999). Similarly, such models almost invariably assume no chiasma interference for reasons of mathematical simplicity. Running numerical simulations with various degrees of chiasma interference could prove useful in identifying whether or when this assumption affects the results. This is particularly important in studies of chromosomal inversions, as both recombination rates and underdominance in inversion heterokaryotypes can depend strongly on the degree of interference (Navarro et al. 1997). The model also provides a more realistic view of the relationship between recombination and underdominance in heterokaryotypes. Models of the evolution of inversions commonly assume perfect suppression of chiasmata (e.g. Charlesworth and Charlesworth 1973, Trickett and Butlin 1994, Dagilis and Kirkpatrick 2014), which implies no recombination and no underdominance (equivalent to $d_k=0$ for all k in the inverted region in my model). Although some inversions, particularly short ones, seem to have this attribute, it appears that most do not (Coyne et al. 1993, Navarro and Ruiz 1997). Furthermore, some studies specifically designed to test the effect of recombination in heterokaryotypes assume no underdominance (e.g. Feder and Nosil 2009), or vice versa (e.g. Kirkpatrick and Barton 2006), which is unrealistic in that chiasmata that cause recombination will inevitably also cause some degree of sterility, and vice versa. The model presented in this paper thus provides an opportunity to compare the results of simulations with or without the assumption of perfect suppression of chiasmata, and to conduct more realistic simulations of the effects of recombination and underdominance in heterokaryotypes.

Secondly, the predictions of the model with various parameter settings can be compared to real data so as to elucidate problems related to interference and inversions. For example, Lange et al. (1997) perform maximum likelihood analyses using both the pure counting model and a variant of what I call the stochastic counting model to see if the latter gives a significantly better fit to the data. My model allows, for example, for a similar comparison of the two-pathway counting model and the full general model (i.e. of a two-pathway model with deterministic or stochastic counting in the interfering pathway) or of the full general model and the stochastic counting model (i.e. of a stochastic counting model with or without an additional non-interfering pathway). The model can also be used to address questions about interference in inversion heterokaryotypes. I have, for example, throughout this paper assumed that the intermediate events in interval I_k are suppressed by a factor d_k , and that the counting process works on this depleted set in the same way that it does in homokaryotypes. Whether interference actually works the same way in hetero- as in homokaryotypes is however an open question that can be addressed by comparing the predictions of the model to recombination data. A related question is that of the presence or absence of interference across breakpoint boundaries, which can be addressed by testing the predictions of the model with the parameter α set to 1 and 0, respectively.

5 Acknowledgements

Figures 1, 4, 5, 6, 7, 8 and 9 were drawn by Sofie Ensby Rostad. I thank Arcadi Navarro, Fabrice Eroukhanoff, Glenn-Peter Sætre, Thomas Hansen, Åke Brännström and Iain Johnston for discussions, exchanges and comments.

6 References

- Anton, E., Blanco, J., Egozcue, J., & Vidal, F. (2005). Sperm studies in heterozygote inversion carriers: a review. *Cytogenetic and genome research*, 111(3-4), 297-304.
- Ayala, D., Guerrero, R. F., & Kirkpatrick, M. (2013). Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution: International Journal of Organic Evolution*, 67(4), 946-958.
- Berchowitz, L. E., & Copenhaver, G. P. (2010). Genetic interference: don't stand so close to me. *Current genomics*, 11(2), 91-102.
- Carson, H. L. (1946). The selective elimination of inversion dicentric chromatids during meiosis in the eggs of *Sciara impatiens*. *Genetics*, 31(1), 95-113.
- Charlesworth, B., & Charlesworth, D. (1973). Selection of new inversions in multi-locus genetic systems. *Genetics Research*, 21(2), 167-183.
- Cobbs, G. (1978). Renewal process approach to the theory of genetic linkage: the case of no chromatid interference. *Genetics*, 89(3), 563-581.
- Colombo, P. C., & Jones, G. H. (1997). Chiasma interference is blind to centromeres. *Heredity*, 79(2), 214.
- Copenhaver, G. P., Housworth, E. A., & Stahl, F. W. (2002). Crossover interference in Arabidopsis. *Genetics*, 160(4), 1631-1639.
- Coyne, J. A., Aulard, S., & Berry, A. (1991). Lack of underdominance in a naturally occurring pericentric inversion in *Drosophila melanogaster* and its implications for chromosome evolution. *Genetics*, 129(3), 791-802.
- Coyne, J. A., Meyers, W., Crittenden, A. P., & Sniegowski, P. (1993). The fertility effects of pericentric inversions in *Drosophila melanogaster*. *Genetics*, 134(2), 487-496.
- Dagilis, A. J., & Kirkpatrick, M. (2016). Prezygotic isolation, mating preferences, and the evolution of chromosomal inversions. *Evolution*, 70(7), 1465-1472.
- Erdelyi, A. (Ed.) (1955). *Higher Transcendental Functions*, vol. 3. New York: McGraw-Hill.
- Etges, W. J., & Levitan, M. (2004). Palaeoclimatic variation, adaptation and biogeography of inversion polymorphisms in natural populations of *Drosophila robusta*. *Biological Journal of the Linnean Society*, 81(3), 395-411.
- Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, 25(11), 660-669.
- Feder, J. L., Roethele, J. B., Filchak, K., Niedbalski, J., & Romero-Severson, J. (2003). Evidence for inversion

- polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics*, 163(3), 939-953.
- Feder, J. L., & Nosil, P. (2009). Chromosomal inversions and species differences: when are genes affecting adaptive divergence and reproductive isolation expected to reside within inversions?. *Evolution: International Journal of Organic Evolution*, 63(12), 3061-3075.
- Foss, E., Lande, R., Stahl, F. W., & Steinberg, C. M. (1993). Chiasma interference as a function of genetic distance. *Genetics*, 133(3), 681-691.
- Foss, E. J., & Stahl, F. W. (1995). A test of a counting model for chiasma interference. *Genetics*, 139(3), 1201-1209.
- Gethmann, R. C. (1988). Crossing over in males of higher Diptera (Brachycera). *Journal of Heredity*, 79(5), 344-350.
- Gorlov, I. P., & Borodin, P. M. (1995). Recombination in single and double heterozygotes for two partially overlapping inversions in chromosome 1 of the house mouse. *Heredity*, 75(2), 113.
- Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet*, 8(29), 299-309.
- Hillers, K. J. (2004). Crossover interference. *Current Biology*, 14(24), R1036-R1037.
- Housworth, E. A., & Stahl, F. W. (2003). Crossover interference in humans. *The American Journal of Human Genetics*, 73(1), 188-197.
- Jaarola, M., Martin, R. H., & Ashley, T. (1998). Direct evidence for suppression of recombination within two pericentric inversions in humans: a new sperm-FISH technique. *The American Journal of Human Genetics*, 63(1), 218-224.
- Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., ... & Wilkinson, P. A. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363), 203.
- King, J. S., & Mortimer, R. K. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics*, 126(4), 1127-1138.
- Kirkpatrick, M. (1982). Sexual selection and the evolution of female choice. *Evolution*, 36(1), 1-12.
- Kirkpatrick, M., & Servedio, M. R. (1999). The reinforcement of mating preferences on an island. *Genetics*, 151(2), 865-884.
- Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1), 419-434.
- Kleckner, N., Zickler, D., Jones, G. H., Dekker, J., Padmore, R., Henle, J., & Hutchinson, J. (2004). A mechanical basis for chromosome function. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34), 12592-12597.
- Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoepfner, M. P., ... & Chen, W. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff

- (*Philomachus pugnax*). *Nature Genetics*, 48(1), 84.
- Lande, R., & Stahl, F. W. (1993). Chiasma interference and the distribution of exchanges in *Drosophila melanogaster*. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 58, pp. 543-552). Cold Spring Harbor Laboratory Press.
- Lange, K., Zhao, H., & Speed, T. P. (1997). The Poisson-skip model of crossing-over. *The Annals of Applied Probability*, 299-313.
- Malkova, A., Swanson, J., German, M., McCusker, J. H., Housworth, E. A., Stahl, F. W., & Haber, J. E. (2004). Gene conversion and crossing over along the 405-kb left arm of *Saccharomyces cerevisiae* chromosome VII. *Genetics*, 168(1), 49-63.
- Mather, K. (1938). Crossing-over. *Biological Reviews*, 13(3), 252-292.
- McPeck, M. S., & Speed, T. P. (1995). Modeling interference in genetic recombination. *Genetics*, 139(2), 1031-1044.
- Muller, H. J. (1916). The mechanism of crossing-over. *The American Naturalist*, 50(592), 193-221.
- Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(1), 2-9.
- Munz, P. (1994). An analysis of interference in the fission yeast *Schizosaccharomyces pombe*. *Genetics*, 137(3), 701-707.
- Navarro, A., Betrán, E., Barbadilla, A., & Ruiz, A. (1997). Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, 146(2), 695-709.
- Navarro, A., & Ruiz, A. (1997). On the fertility effects of pericentric inversions. *Genetics*, 147(2), 931-933.
- Navarro, A., & Barton, N. H. (2003). Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, 57(3), 447-459.
- Nolan, J. P. (2017). Erlang renewal models for genetic recombination. *Journal of Statistical Distributions and Applications*, 4(1), 10.
- Noor, M. A., Grams, K. L., Bertucci, L. A., Almendarez, Y., Reiland, J., & Smith, K. R. (2001). The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. *Evolution*, 55(3), 512-521.
- Novitski, E., & Braver, G. (1954). An analysis of crossing over within a heterozygous inversion in *Drosophila melanogaster*. *Genetics*, 39(2), 197-209.
- Pegueroles, C., Ordóñez, V., Mestres, F., & Pascual, M. (2010). Recombination and selection in the maintenance of the adaptive value of inversions. *Journal of evolutionary biology*, 23(12), 2709-2717.
- Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M. G. & Wolf, J. B. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, 344(6190), 1410-1414.
- del Priore, L., & Pigozzi, M. I. (2015). Heterologous synapsis and crossover suppression in heterozygotes

- for a pericentric inversion in the Zebra Finch. *Cytogenetic and genome research*, 147(2-3), 154-160.
- Roberts, P. A. (1976). The genetics of chromosome aberration. In M. Ashburner and E. Novitski (Eds.) *The Genetics and Biology of Drosophila* (67-184). Academic Press: London
- Ross, S. M. (2014). *Introduction to probability models*. London: Academic press.
- Schaeffer, S. W., & Anderson, W. W. (2005). Mechanisms of genetic exchange within the chromosomal inversions of *D. pseudoobscura*. *Genetics*, 171, 1729-1739.
- Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., ... & Sonesson, A. (2016). Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Molecular ecology*, 25(10), 2130-2143.
- Speed, T. P. (1996). What is a genetic map function?. In *Genetic Mapping and DNA sequencing* (pp. 65-88). New York, NY: Springer.
- Sinclair-Waters, M., Bradbury, I. R., Morris, C. J., Lien, S., Kent, M. P., & Bentzen, P. (2018). Ancient chromosomal rearrangement associated with local adaptation of a postglacially colonized population of Atlantic Cod in the northwest Atlantic. *Molecular ecology*, 27(2), 339-351.
- Stahl, F. W., Foss, H. M., Young, L. S., Borts, R. H., Abdullah, M. F., & Copenhaver, G. P. (2004). Does crossover interference count in *Saccharomyces cerevisiae*?. *Genetics*, 168(1), 35-48.
- Stam, P. (1979). Interference in genetic crossing over and chromosome mapping. *Genetics*, 92(2), 573-594.
- Sturtevant, A. H. (1913). A third group of linked genes in *Drosophila ampelophila*. *Science*, 37(965), 990-992.
- Sturtevant, A. H. (1915). The behavior of the chromosomes as studied through linkage. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, 13(1), 234-287.
- Sturtevant, A. H., & Beadle, G. W. (1936). The relations of inversions in the X chromosome of *Drosophila melanogaster* to crossing over and disjunction. *Genetics*, 21(5), 554-604.
- Trickett, A. J., & Butlin, R. K. (1994). Recombination suppressors and the evolution of new species. *Heredity*, 73(4), 339.
- Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., ... & Gonser, R. A. (2016). Divergence and functional degradation of a sex chromosome-like supergene. *Current Biology*, 26(3), 344-350.
- Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y. C., Shoemaker, D., & Keller, L. (2013). A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, 493(7434), 664.
- Wang, S., Zickler, D., Kleckner, N., & Zhang, L. (2015). Meiotic crossover patterns: obligatory crossover, interference and homeostasis in a single process. *Cell Cycle*, 14(3), 305-314.
- Zhao, H., Speed, T. P., & McPeck, M. S. (1995a). Statistical analysis of crossover interference using the chi-

square model. *Genetics*, 139(2), 1045-1056.

Zhao, H., McPeck, M. S., & Speed, T. P. (1995b). Statistical analysis of chromatid interference. *Genetics*, 139(2), 1057-1065.

7 Appendix: Vector and matrix notation

All vectors and matrices are denoted in bold. A matrix or vector followed by square brackets ($[]$) denotes the particular element of that matrix or vector, so that e.g.

$$\mathbf{M}[i,j] = i + j, \quad \text{for } i,j = 0,1,2 \dots n-1$$

indicate that $\mathbf{M}$ is a zero-indexed matrix of size n,n with element i,j (row, column) equal to $i+j$. All vectors are row-vectors unless otherwise indicated, so that e.g.

$$\mathbf{v}[j] = j, \quad \text{for } j = 0,1,2 \dots n-1$$

means that

$$\mathbf{v} = (0 \quad 1 \quad 2 \quad \dots \quad n-1)$$

When referring to a column-vector, I either do so by writing out the vector in full, like this:

$$\mathbf{w} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ \dots \\ n-1 \end{pmatrix}$$

or by first defining the vector elements and then indicate in subsequent expression that the vector is transposed, like this:

$$\begin{aligned} \mathbf{w}[j] &= j, \quad \text{for } j = 0,1,2 \dots n-1 \\ a &= \mathbf{v}\mathbf{w}^T \end{aligned}$$

Two vectors or matrices placed side by side always indicate matrix multiplication. When matrix multiplication operations are to be performed on a given number of matrices with increasing indices from left to right, I sometimes write it in the following short form

$$\prod_{i=0}^{n-1} \mathbf{M}_i = \mathbf{M}_0 \mathbf{M}_1 \mathbf{M}_2 \dots \mathbf{M}_{n-1}$$

The multiplication of x instances of the same matrix ($\mathbf{M}$) is denoted $\mathbf{M}^x$, so that

$$\mathbf{M}^x = \prod_{i=0}^{x-1} \mathbf{M} = \mathbf{M} \mathbf{M} \mathbf{M} \dots \mathbf{M}$$

By definition,

$$\mathbf{M}^0 = \mathbf{I}$$

i.e. a matrix to the zeroth power is always equal to the corresponding identity matrix. The *one-vector*, denoted $\mathbf{1}$, is a vector of only ones, i.e. $\mathbf{1} = (1 \ 1 \ 1 \ \dots \ 1)$.