Distribution of meiotic crossovers : heterogeneity, modeling interference, inter-pathway crosstalk

Sayantani Basu Roy

Logo de l'établissement de co-tutelle (le cas échéant)

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE : *Gènes, Génomes, Cellules*
Laboratoire : UMR de Génétique Végétale du Moulon
(Équipe Génétique Quantitative Fondamentale)

DISCIPLINE (Biologie)

THÈSE DE DOCTORAT

soutenue le 31/03/2014

par

Sayantani BASU ROY

Répartition des crossovers méiotiques : hétérogénéité, modélisation de l'interférence, interaction entre voies de formation

Directeur de thèse :	Olivier MARTIN	DR, Université Paris-Sud/ INRA, Ferme du Moulon
Co-directeur de thèse :	Matthieu FALQUE	IR, INRA (Ferme du Moulon)

Composition du jury :

Président du jury :	Denise ZICKLER	DR, Université Paris-Sud (Orsay)
Rapporteurs :	Eric JENCZEWSKI	CR, INRA (Centre de Versailles-Grignon)
	Paul-Henry COURNÈDE	PR, Ecole Centrale de Paris (Châtenay-Malabry)
Examinateurs :	Christine MÉZARD	DR, CNRS (Centre de Versailles-Grignon)

Acknowledgement

I truly believe that my experience here has made me more perceptive and curious about research in particular and life in general.

My advisor, *Olivier Martin* has made me believe that learning and improving should be the only constants in a researcher's life. This thesis would never have begun, nor concluded without his guidance and optimism.

The biological knowledge I have gathered over the last three years was largely inspired by the insightful discussions with and explanations from *Matthieu Falque*, my co-advisor.

Franck Gauthier initiated me into object-oriented programming, which stood me in good stead through my thesis and will, even beyond.

Adrienne Ressayre has always been very kind and generous, making me feel at home in the GQF team. It would have been difficult without her around.

Dedicated to ...

my dearest Ma,Baba,Tanu. my beloved Shanto.

my pal Vasu.

'Because life ceases to be as I know it without you'

Thesis Contents

GENERAL INTRODUCTION

The purpose…

STATEMENT OF OBJECTIVE

This chapter is to explain the primary premise of my thesis and what it is that we set out to explore. The following text will also put the specialized introduction (Chapter 1) in appropriate perspective.

Life, sex and meiosis

Life comes in many forms, from bacteria to multi-cellular organisms. The essence of life is the ability to replicate, forming more or less faithful copies of the original. While eubacteria and archae (all unicellular) divide, a single mother cell producing two daughter cells that are nearly identical copies of it, eukaryotes have the unique characteristic that they indulge in sex. Specifically, *two* parents rather than just one are used to go from one generation to the next, and at the heart of this process is "meiosis". During meiosis of unicellular eukaryotes, a single "diploid" cell having *n* pairs of chromosomes (homologues) will produce "haploid" daughter cells having only one chromosome for each homologous pair; when two of these haploid cells fuse, one obtains a diploid cell that can start the process over again. In the case of multi-cellular organisms, specialized cells in the germ line of the organism undergo meiosis, producing the haploid gametes that then can produce the next generation by fusing to form the diploid zygote that is the single cell starting point for developing a new multicellular organism of the next generation. Note that for both the unicellular and multicellular cases, the "offspring" contains a *mixture* of the chromosomal content of its two "parents". Thus "reproduction" in eukaryotes is fundamentally different from replication in bacteria or archae. Making sure that each offspring has exactly the same karyotype as the parents (the same number of pairs of homologues and nothing else) is a priori a challenge, so it will come as no surprise that meiosis involves a remarkably intricate system to position and properly separate chromosomes. If segregation errors arise as such in meiosis (trisomy being a common case in humans), generally the effects are deleterious if not lethal.

Is it a coincidence that all multicellular organisms reproduce via meiosis? It is tempting to postulate that the ability to produce offspring that are mixtures of their parents is evolutionarily advantageous, and there is a large body of research that has tried to demonstrate that this is so. Clearly the shuffling of chromosomes will allow for rapid enrichment in a population of favorable alleles for instance when the environment undergoes changes, so sex (in fact meiosis) should provide a selective advantage in adapting to new environments. It should be thus no surprise that meiosis is a focal point for many studies ranging from cell biology (because of the incredibly intricate molecular machinery required for faithful segregation of chromosomes) or in evolutionary biology (where much research focuses on modes and strategies of sexual reproduction). But meiosis is also of applied interest: by recurrent selection of offspring in breeding programs, humanity has domesticated numerous species (plants and animals) without which our lifestyles would be very different. In the last century the issue of fast effective breeding has become a major challenge, for instance to provide plant varieties that have high yield. Presently the stakes are in producing varieties that are more efficient in water and fertilizer use or are more resistant to disease. Although such programs can work by selecting one generation after another, recently it has transpired that manipulation of meiosis could dramatically speed up such programs. With this in mind, the present thesis is focused on the general question of understanding the mixing of alleles generated during meiosis and more specifically characterizing the pathways for crossover production. These questions are to a large extent questions of fundamental cell biology, but they also have potential impact for practical applications and for the long term sustainability of our way of life.

Meiosis in a nutshell

When a eukaryotic cell is not undergoing any kind of division – sexual or otherwise – it typically is in a diploid form with a complete set of chromosomes. By "complete" we mean that chromosomes come in pairs, one being maternal and one being paternal.

Meiosis is preceded by replication of chromosomal DNA. After this replication (in essence each chromosome is duplicated or cloned), each chromosome appears as two sister chromatids (thus there are four copies of each chromosome) held together by cohesins. Cohesins are protein complexes that tie together these sisters. The centromere is a chromosomal region that is characterized by repeated sequences and is associated with different histones than the rest of the chromosome, but its main characteristic is that it recruits the proteins that assemble the kinetochore that attaches to microtubules. Microtubules thus can move the chromosomes and do so actively in meiosis. The centromere may be positioned towards the middle of the chromosome for metacentric chromosomes, rather to one side for acrocentric chromosomes or even at one end in telocentric chromosomes. In the first phase of meiosis, homologous chromosomes pair up and exchange chromosomal segments reciprocally (COs) or non-reciprocally (non-COs) between non-sister chromatids. Thereafter, first the pairs of homologues separate and in a second phase of

meiosis the remaining sister chromatids also separate in an orderly manner (in detail in Chapter 1) to finally give gametes with only one copy per chromosome.

Recombination or intra-chromosomal shuffling

This exchange mixes the alleles, known as recombination. In the absence of COs, each gamete will inherit just one of the 2 homologues of its parent for each chromosome, this choice being random. If there are *n* pairs of homologues, there are 2^n possible gametes that can be produced so we see the shuffling introduced by meiosis is important. However in the absence of COs, chromosomes would stay intact. By producing COs, meiosis introduces an additional level of shuffling that is intra-chromosomal, arising between homologues. When a chromosome of the gamete has for gene 1 the allele of homologue 1 and for gene 2 the allele of homologue 2, one says that it is recombinant for these 2 genes. Recombination thus enhances the possibilities of genetic shuffling. Most certainly such shuffling is under selection constraints: too little shuffling reduces adaptability to changing environments while too much may break favorable associations of alleles. Thus one expects recombination rate to be regulated. This clearly must be the case as when one compares diverse organisms, the sizes of their genomes can differ by factors of 100 or more while their recombination rates vary typically only by factors of 2 to 3. How is this regulation implemented and why is it present? A better characterization of CO formation in many species should shed light on these questions.

Variation in Recombination Rates

Recombination rates are known to vary locally along chromosomes displaying several hotspots. However, these hotspots do not affect the global recombination landscape. Specific regions of a chromosome tend to recombine much more than others (telomeric regions for example as opposed to pericentromeric regions). But this tendency does not reflect on the relationship between genetic and physical distance, which remains monotonic. Genetic distance is measured in Morgan (1 unit = distance over which there is 1 CO on average) while physical distance units are base pairs (bp). So, with each Morgan, several hundreds of bp could be traversed on a chromosome. Thus it is befuddling how an almost constant slope is maintained despite the proved presence of recombination hotspots.

In addition difference in recombination rates is also seen between sexes. For instance in *Arabidopsis*, female meiosis shows more recombination than male meiosis in general with the differences becoming more pronounced in certain chromosomal regions.

Interference amongst Crossovers

COs do not form "randomly" nor independently along chromosomes. Two COs tend not be close to one another. This phenomenon is known as CO interference – formation of a CO *interferes* with the formation of COs in its neighborhood in either direction along the chromosome. Thus interference reduces variability in inter-CO distances by restricting nearby COs. But it is not known if it exercises an upper limit on inter-CO distances. It is however prudent to take note that higher inter-CO distance would indicate stronger interference.

It has been seen that during meiosis each chromosome receives at least 1 CO. This is known as the obligate CO rule. But too many COs are also known to be deleterious. Hence interference could indeed be understood as a mechanism that helps strike a balance between these two scenarios.

Diversity of crossover dynamics

COs may also form without interference. Presently, at least two distinct CO formation pathways are known – interfering and non-interfering. So an organism might have only interfering COs or strictly non-interfering COs or a contribution from both kinds. When both pathways are found, the non-interfering COs constitute a minority of the total number of COs rather than majority. Further, inter-specific variations have been observed in interference strength as well as the proportion of the non-interfering pathway.

These pathways are mediated by different gene families. But if we agree that interference is to restrict the number of COs, then what is the purpose of COs without any interference? Why is it that all organisms do not have both pathways? And how does sometimes only COs which do not interfere suffice to keep deleterious mutations at bay?

What this thesis is about

The work presented in this thesis focusses on understanding the two CO formation pathways in greater detail. For instance, how do they tend to vary at several levels – between chromosomes or between different regions of the same chromosome. Are these trends common to all species?

We do our studies here mainly based on modeling CO interference. Modeling can provide insights which looking only at inter-CO distributions cannot. Till now, various models have been postulated – some statistical and others mechanistic. Each kind has its own ups and caveats. But due to the main disadvantage of no true likelihood for mechanistic models, we have dealt in the gamma model here based on stationary renewal processes (SRP). As the name suggests, the CO formation process is governed by a gamma SRP with the shape parameter quantifying

interference. This model is used as a powerful tool to make comparisons about interference at all levels and discovering distinct characteristics of the two pathways.

If we can find an explanation to the arrangement of COs along chromosomes it would have far-reaching applications in breeding programs.

Data-inspired Models

The kind of models formulated to study interference is also dependent on the kind of data. And the quality of data is indicated by several parameters. Firstly more the number of individuals in a dataset, the subsequent statistical analyses would be that much decisive and coherent. Second, whether information is available about every CO along each chromosome – one requires sufficient number of genetic markers or efficient microscopy (electron or immunofluorescence) techniques to detect all COs. Thirdly, the stage at which CO positions have been annotated - at the bivalent (four-stranded, during the first half of meiotic division or Meiosis I) level or at the gamete stage (double-stranded, at the end of sexual division or Meiosis II). On average, half the number of COs are observed on gametes as compared to bivalents.

But models need to evolve as datasets become better and more detailed. Since though one can develop models including several experimentally proved details, this might entail increase in the number of parameters. And it is possible that the dataset then is not large enough to estimate that many parameters efficiently. Thus even if more intricate models can be developed and validated by simulations, for the purpose of application, larger datasets would need to be generated. Till sometime back, it was not possible to differentiate if COs were from one pathway or another on the same chromosome at the same time. However it was recently performed for the first time on tomato by the Anderson group (discussed later in the thesis). Thus now possibilities become manifold on the modeling front. Now we can actually validate the predicted landscapes for the two pathways.

Guide to Chapters

Chapter 1 gives a detailed introduction to the sexual division process of meiosis, CO formation and the various approaches followed to modeling CO interference over the years.

Chapter 2 provides description of the different aspects of the gamma model of CO interference.

Part I:

We published a paper on the study of CO interference in *Arabidopsis thaliana* in Genetics. This data was from huge male and female backcross populations (1500 individuals each), prepared by

the Mézard group at INRA Versailles. So Chapter 3 details the methods proposed and implemented in the paper to analyze interference. Thereafter Chapter 4 summarizes the results presented in this work followed by its journal version.

Part II:

This portion describes the statistical analysis performed on the tomato (*Solanum lycopersicum*) data by the Anderson group (Colorado State University, USA). Thus Chapter 5 lays out the analysis details and the resulting insights from this first-of-its-kind data.

Part III:

The last part is about a model that we postulate by drawing from our observations in the Genetics publication. Chapter 6 explains the model intricacies. Finally Chapter 7 gives results from simulations and application to a dataset.

Chapter 8 concludes the thesis.

Chapter 1: INTRODUCTION

What it takes to begin…

BACKGROUND: ON HINDSIGHT

This introductory section intends to give an idea about the knowledge base we had at our disposal to begin building upon. In what follows, I explain to the best of my abilities, the three most important pillars of my thesis.

MEIOSIS

Eukaryotes reproducing sexually undergo meiosis where the first segregation Meiosis I (MI) is reductional and the second division or Meiosis II (MII) is equational. MI reduces the number of cytologically visible chromosomes to half while MII splits chromosomes into two chromatids without changing their number.

The cell begins as diploid (say, 2N) with two sets of homologous chromosomes (maternal and paternal) followed by one round of replication. This leads to the appearance of sister chromatids held all along their length by cohesins and at the centromere for each homologue thus maintaining diploid status. And each chromosome then has four copies instead of the usual two.

During the first stage of division (Meiosis I or MI), there is exchange of chromosomal segments (crossing over) between the homologues which finally separate into two cells (Figure 1.1). Crossing over allows recombination between parental chromosomes and creation of novel allelic combinations or genotypes. The cells during MI are diploid but possess only half of the chromosomes they began with, during the passage from MI to MII (say, n) having only one of the homologues for each homologue pair but with the sister chromatids still together. Further, there is stochasticity as to which cell receives which homologue in addition to crossing over as neither is purely maternal or paternal now. The second division phase (Meiosis II or MII) entails random segregation of the sister chromatids in each of the two nuclei dividing into two nuclei each, which are now haploid with a single chromatid per homologous chromosome pair. Hence meiosis derives four haploid cells (or gametes) from a diploid cell.

Meiosis is an important process in the sexual life cycle of eukaryotes for largely two consequences (Figure 1.1). Firstly, the diploid status of organisms is restored when fertilization

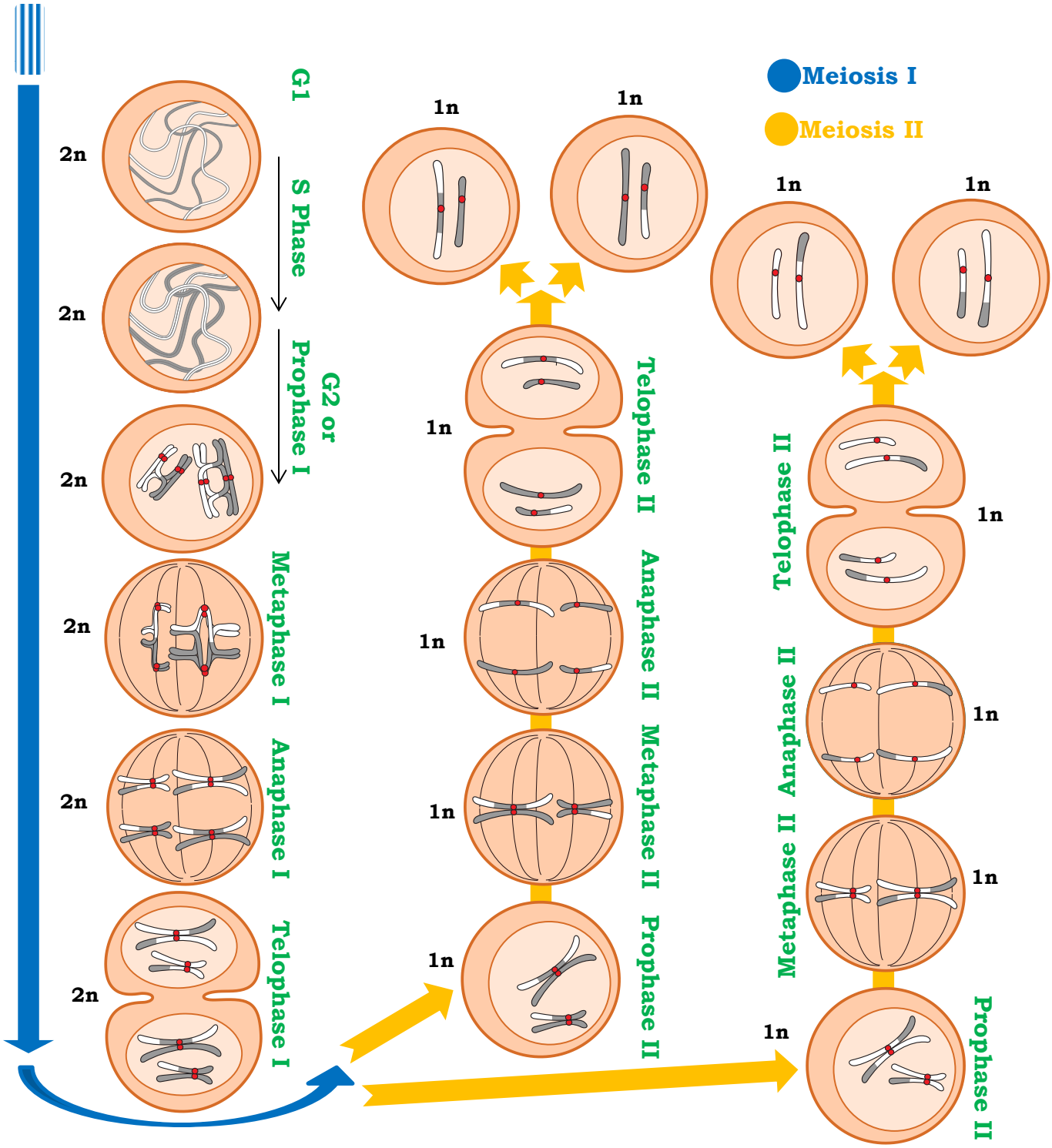


Figure 1.1 Stages of Meiosis including Meiosis I and II.

The diagram depicts details of the different stages of the process of meiosis using a cell having 4 chromosomes in its diploid form and 2 for its haploid. The numbers indicated beside each 'cell' gives information about its contents. 2n denoted diploid while 1n denotes haploid.

Cell diagrams taken and modified from the Thesis of Nicolas Macaisne (Defended on 26 November 2010)

(fusion of the haploid gametes) takes place. Secondly, meiosis introduces genotypic variation in two specific ways: crossing over and random segregation at MI.

1.1 The Process Overview

1.1 (a) *Interphase*

The cell prepares to carry out sexual division by rigorous manufacture of proteins such as enzymes and structural proteins essential to maintain cellular well-being. This is also known as the Growth I (G_1) phase (Figure 1.1). At this stage nuclear De-oxy-ribonucleic acid (DNA) appears as chromatin, composed of euchromatin and heterochromatin. Subsequently the entire genetic material of the cell undergoes replication during the rather long Synthesis (S) phase (Cha *et al.* 2000). Thus each chromosome transforms into a complex of two sister chromatids held together by cohesins and the centromere (typically embedded in heterochromatin (Talbert & Henikoff 2010)), retaining diploidy.

1.1 (b) *Meiosis I*

The first *prophase* is the longest meiotic phase sub-divided into *leptonema*, *zygonema*, *pachynema* and *diplonema*, each with its characteristic changes in the emergence of double-strand breaks (DSBs), positioning of the homologues with respect to each other, developmental stages of the synaptonemal complex (SC) and formation of late recombination nodules (SC-associated cytological manifestation of DNA crossovers) (Figure 1.2).

Leptonema sees initiation of an axial element (AE) along each sister chromatid pair followed by the formation of DSBs. These elements eventually become parts of the SC in the form of lateral elements. Chromosomes also begin to thicken, undergo various kinds of movements (e.g., attachment of telomeres to the nuclear membrane, Zickler &Kleckner 1998) which aid at least partly the homology search to complete *homologous juxtaposition*.

Subsequently, the transverse elements (TEs) are gradually assembled between homologous AEs and in an attempt to repair the DSBs, intermediate joint molecules (JMs) such as single end invasion develop during *zygonema*. As the TEs appear in the region between the two pairs of sister chromatids in a homologue, each chromatid pair loops must align itself to opposite sides, away from the AEs such that the AEs face each other allowing space for the TEs to form in between (Bishop & Zickler 2004). SC formation is completed by *pachynema* and DSBs are repaired to either COs or non-crossovers (NCO) mediating recombination between mainly non-sister chromatids. These two DSB repair classes differ in that COs result when the two chromosomal segments on either sides of the recombination site are inter-changed. Recombination plays an important role in ensuring meiosis is carried out efficiently as chiasmata

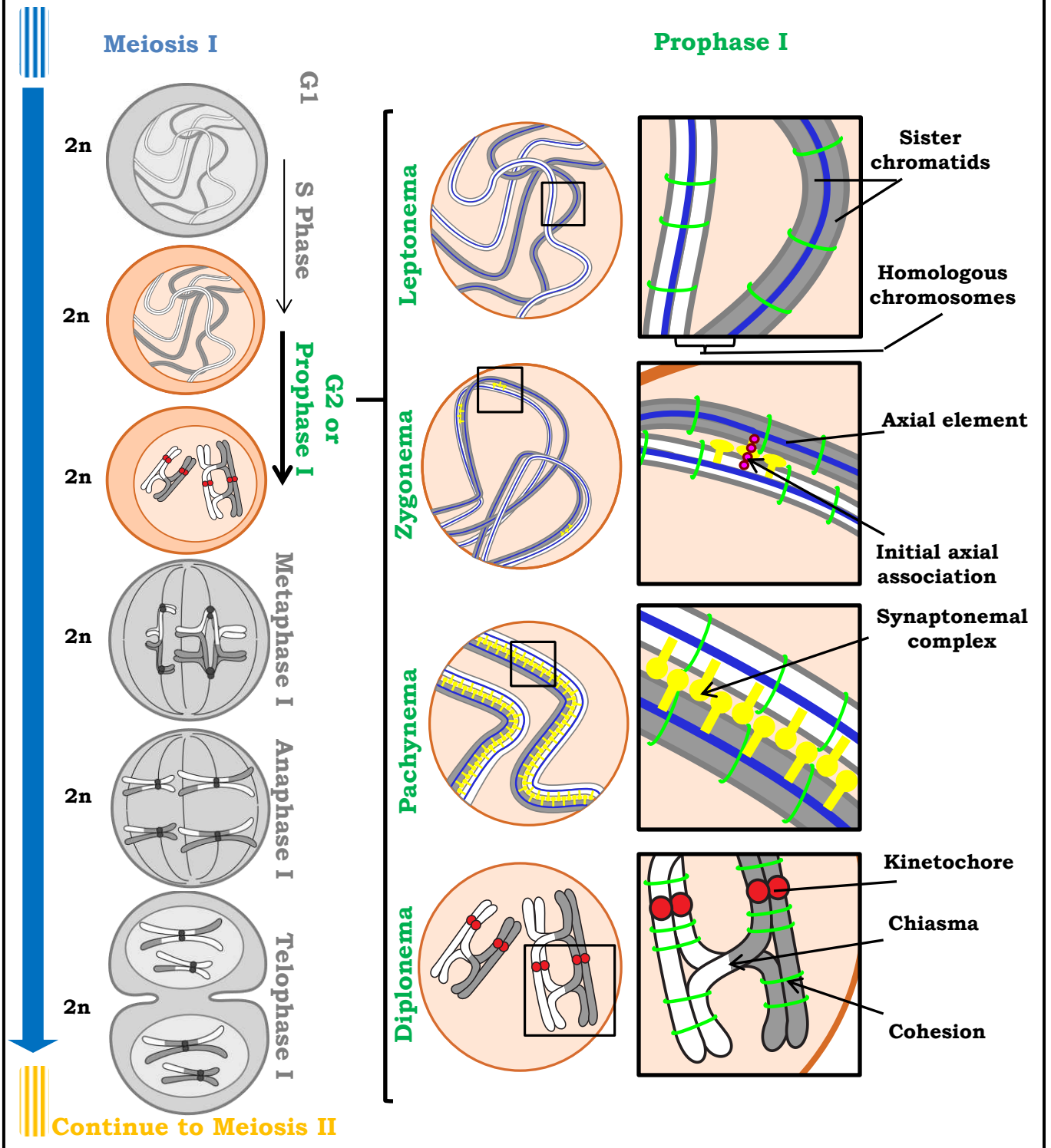


Figure 1.2 The sub-phases of the Prophase during Meiosis I.

Leptonema - Interactions begin between the homologues (with 2 sister chromatids each) to move towards juxtaposition.

Zygonema - The homologous axial elements begin to interact while the lateral element of the synaptonemal complex (SC) begins to take shape.

Pachynema - Formation of SC is complete and homologues have paired successfully.

Diplonema – The homologues have completed exchange of segments at the chiasma where they remain attached.

The cells have 4 chromosomes in their diploid state.

Cell diagrams taken and modified from the Thesis of Nicolas Macaisne (Defended on 26 November 2010)

establish connections between homologues. A fresh collage of alleles, drawing from the two parental homologues is more of a fortunate coincidence. This also explains the obligate-crossover rule, adhered to by most eukaryotes.

During the last phase, namely *diplonema*, there is an abrupt de-condensing and re-condensing of the homologues as the SC disintegrates leaving the homologues attached at the COs and the cohesins, responsible for the structural integrity of COs (Bishop 2006). Cohesins bind the sister chromatids along their entire lengths except where they are attached to a non-sister chromatid at a chiasma. Connected homologs hence become a single unit known as bivalent which facilitate bipolar spindle attachment of the centromeres to the MI spindle. Crucial to *metaphase* I are spindle-centromere interactions which stabilize only under tension, provided by "wrangling" of the microtubules between the two spindle poles (Bishop 2006). The inter-sister cohesin connections along arms prevent the chiasmata from separating by succumbing to the bi-polar forces before *anaphase* I.

The homologues thus align towards opposite poles along the equatorial plate aided by their respective centromeres and conjoined via chiasmata while sister chromatids are kept together by the remnant cohesins near the centromeres as chiasmata are pulled apart to separate the homologues. The arm cohesins along sister chromatids are degraded by Separin as the cell enters *anaphase* I from *metaphase* I. And as the spindle now contracts to pull apart the homologous pairs, we observe opposite polar migration of one chromosome (sister chromatid pair) from each pair. *Telophase* I shows as the "dyad" stage with two polar chromosome groups in the cell which form the two daughter nuclei with half the number of chromosome sets such that each still has its sister chromatids together. These chromosomes de-condense to a certain degree but do not reach the interphase condition before entering the second meiotic division (Ross *et al.* 1996).

1.1 (c) *Meiosis II*

Prophase II leads to disappearance of the two nuclei and thickening of the chromosomes and the microtubule organizing centers (MTOC) begin arranging spindle formation again (Figure 1.1). One spindle each is assembled for each of the two chromosome sets to separate their sister chromatids this time. Each chromatid has its centromere consisting of a kinetochore. The kinetochores of sister chromatids have opposite orientation thus enabling each sister chromatid in a particular pair to attach itself to opposing poles, randomizing which sister chromatid faces which pole. The spindle axis of MII is perpendicular to that of MI. At the end of *metaphase* II, the chromosomes lie at the middle, attached to the respective spindle, awaiting the impending split of the sister chromatids, which will then be known as sister chromosomes. Through *telophase* II, the sister chromatids finally separate to give four clusters each with a single chromatid per homologous pair. This is followed by formation of four nuclei surrounding the

clusters. Finally cytoplasmic division or cytokinesis is observed resulting in four haploid gametes.

1.2 Details of the Important Events

1.2 (a) *DNA Replication*

This is an immensely vital and lengthy process during the pre-meiotic S-phase (meiS). For instance, for budding yeast, this window lasts 65-80 minutes (Padmore *et al.* 1991, Williamson *et al.* 1983). Further, an essential aspect is fidelity; more so, because an anomalous gamete may result in an organism-wide defect. In order to keep a tab on replication, there exist several check-points at crucial turn of events which are activated in case the molecular machinery errs. This might lead to temporary stalling of meiosis until rectification is ensured (reviewed in Liu *et al.* 2003). In addition there also exists communication between DNA replication and DSB formation which initiates recombination in turn. The *pachynema check-point* (reviewed in Roeder & Bailis 2000) monitors completion of recombination and resolution of its various intermediates. This pathway thwarts passage through *prophase* I unless recombination is successfully concluded.

In recent years, it has been advocated that there is direct association between meiS and genetic recombination. Previous studies on budding yeast (*Saccharomyces cerevisiae*) have observed that recombination can be prohibited if meiS is hampered genetically or chemically (Budd *et al.* 1989, Schild & Byers 1978, Simchen *et al.* 1976, Stuart & Wittenberg 1998). Closer temporal examination of these major events revealed that formation of DSBs necessarily *followed* bulk DNA synthesis (Borde *et al.* 2000, Smith *et al.* 2001). Though this suggests that replication and recombination are closely linked, there can be chiefly two possibilities. Either there is forthright ex-change of information between the development of the replication fork and the appearance of chromosomes having recombination capability. Or, an unsatisfactory meiS might hinder the subsequent events by activating the *replication check-point* system which may present itself further as an indirect hesitation for recombination to initiate. Finally the scenario might vary between organisms though there a hierarchical molecular vigil is inevitable to ensure replication is carried out efficiently (Strich 2004).

1.2 (b) *Inter-Homologue Interactions and Juxtaposition*

Homologous chromosomes appear to conclude the action of juxtaposition right at the outset of *prophase* I. They seem to perform this in two stages which are notionally distinct – the pairs of homologues *co-localizing* to a shared spatial region and thereafter *co-aligning*. These specific movements may not always be completely temporally separated (Kleckner & Weiner 1993, Scherthan *et al.* 1994). It has been suggested that this occurs via an onslaught of interstitial interactions which begin as unstable but become increasingly sure. Beginning slow aids just so

unfavorable interactions may be minimized early before turning enduring (Weiner & Kleckner 1994, Kleckner & Weiner 1993). Several such interactions are observed using fluorescence *in-situ* hybridization (FISH) analysis of yeast chromosome spreads (Weiner & Kleckner 1994). This communication is perhaps mediated by direct DNA-DNA contacts among undisturbed duplexes along with homology searching, made possible by appropriate accompanying proteins (Kleckner 1996).

These chromosomal movements chiefly involve homology recognition and as previously mentioned, a spatial meeting of the homologues (Kleckner & Weiner 1993, von Wettstein *et al.* 1984, Wang *et al.* 2009). In a majority of organisms, pairing of homologues depends on and is carried out through the DNA recombination process. This is mediated by the physical localization of DNA recombination complexes (or "recombinosome") to the underlying homologous axes (Henderson and Keeney 2005, Anderson and Stack 2005, Börner *et al.* 2004, Franklin *et al.* 2006). Recombinosome-axis tête-à-têtes occur immediately previous to or following the beginning of recombination expressed duly by the appearance of DSBs. Hereafter, both, recombination and attempt to juxtapose the homologous axes using recombination as a medium, progress together, synchronizing in time and function.

Though DSBs usually play an indispensable role in the occurrence of homologous pairing, the extent to which DSB-mediated pathways may or may not be the favored tend to vary between organisms (Burgess 2002, Page & Hawley 2003). For *Saccharomyces cerivisiae*, even though the primary mode to establish homologue pairing is via DSBs, at least some pairing interactions do occur without their assistance (Peoples *et al.* 2002, Burgess 2002). In contrast, homologue pairing happens successfully without any DSBs whatsoever in *C. elegans* as well as both sexes of *Drosophila melanogaster* (Dernburg *et al.* 1998, McKim *et al.* 1998, MacQueen *et al.* 2002, Vazquez *et al.* 2002). The chromosomes of *C. elegans* begin meiosis unpaired and then submit to brisk alignment. Here aligning chromosomes requires no DSBs, thus recombination is not initiated and not even the function of proteins which will finally lead to synapsis (MacQueen *et al.* 2002). Rather in tandem for *Drosophila* females, the presence of previous somatic pairing seems to help evade DSB-mediated homology searches. Since forming recombination intermediates to reach synapsis is dispensable in these organisms, this probably reflects the ability of flies and worms to use homology recognition methods independent of DSBs.

1.2 (c) *Formation of the Synaptonemal Complex*

The principle purpose of this complex structure, which begins to form almost as soon as *prophase* begins, is not understood completely. Its assembling is concluded by *pachynema* and then it is taken apart through *diplonema* (Figure 1.3).

When chromosomes begin *leptonema*, they appear as thin long chromatin strips which are starting to condense but without any hint towards homolog pairing. As prophase continues, the

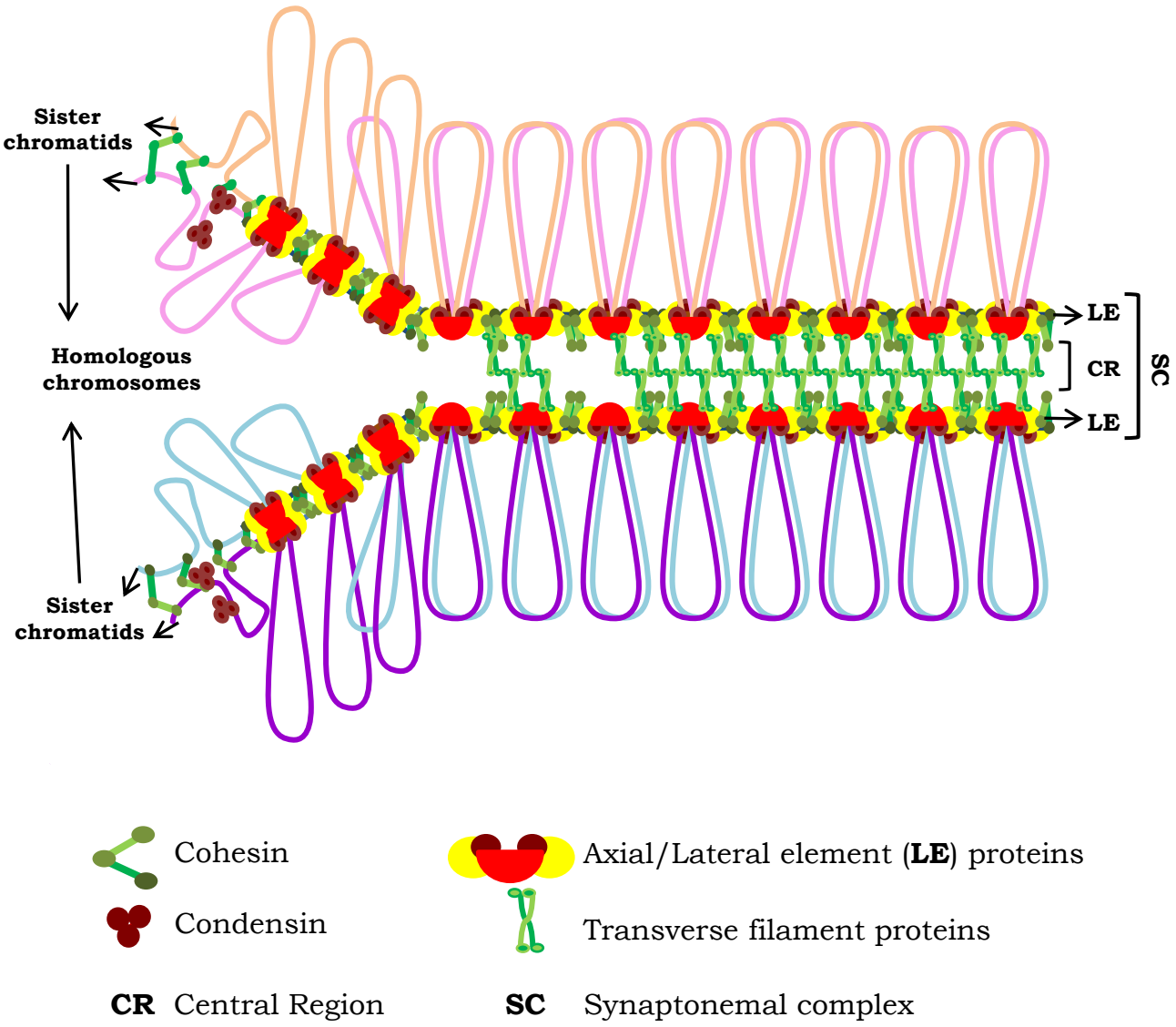


Figure 1.3 The Synaptonemal Complex.
This is a schematic representation of this involved scaffolding formed between the homologous chromosomes during prophase I but degrading by metaphase I. The presence of this structure under the microscope indicates the completion of pairing between homologues. It has two prominent regions. The first beginning as the axial element holding sister chromatids together and then maturing into the lateral element. And the second portion known as the central region is formed between the two homologous lateral elements consisting of transverse filament proteins. Further the two proteins which are indispensable to the complex are cohesins and condensins.

Taken and Modified from the Thesis of Nicolas Macaisne (Defended on 26 November 2010)

chromatin condenses and the sister chromatids localize along protein scaffolds called axial elements (AEs).

For most organisms, the induced DSBs, with complementary sequences on homologues, facilitates their alignment. The resulting interactions are seen as ~400nm inter-axis bridges (Albini & Jones 1987, Tesse *et al.* 2003). Each bridge probably consists of a DSB inducing a growing interaction with a homologous sequence. Gradually through *leptonema*, only a few of these bridges transform into axial associations (AAs), which then link the paired lateral elements (LEs) (Rockmill *et al.*1995). These AAs will finally serve as synapsis initiation sites leading to the formation of the SC intimately bound by the AEs on either side. Thus the AEs are integrated as part of the LEs in the SC architecture. The regions of synapsis lengthen between the homologous pair of chromosomes till the entire stretch is covered to conclude the event of synapsis, which happens by *pachynema*. Synapsis indicates that the chromosomes form a compact uninterrupted structure (the SC), wherein the four chromatids are arranged in an orderly fashion. The SC is now composed of the transverse filaments and the paired LEs. The filaments hold the LEs together by *stitching* in between them.

Following *pachynema*, many organisms undergo de-condensation during which the SC disintegrates and the homologues are attached via chiasmata.

1.2 (d) *Double-strand Breaks and the Synaptonemal Complex*

Recombination is begun by the appearance of DSBs which are formed by Spo11 and its homologues in most organisms (Keeney *et al.* 1997, Grelon *et al.* 2001, Dernburg *et al.* 1998). And whether a DSB will become a reciprocal CO or a gene conversion event (or NCO) seems to be known quite early, almost at the time of DSB formation (Allers & Lichten 2001, Hunter & Kleckner 2001) (Figure 1.4). Though the succession of events which result in gene conversions is still unclear, it is known how DSBs mature into COs. A DSB usually undergoes resection leading to the formation of a gap (White and Haber 1990) which gives rise to two intermediates, single-end invasions (SEIs) and double Holliday junctions (dHJs, Holliday 1964) emerging from consecutive single-strand events taking place at each end of the double-strand gap (Hunter & Kleckner 2001, Schwacha & Kleckner 1995). Both intermediates are quite stable and in yeast, their sequential emergence and disappearance indicates advance in the meiotic *prophase* I. So in yeast, the DSBs show up before any SC is detected at *leptonema*; SEIs begin to show themselves as the first sign of SC formation is seen and become stable when SC is formed in its entirety; dHJs appear during *pachynema* and are resolved into COs (or NCOs) by the end of *pachynema*.

As observed by Blat *et al.* (2002), the transformation of DSBs into COs is concurrent to structural changes in the AEs of meiotic chromosomes such that “interruption of DNA via a DSB may be accompanied by interruption of the underlying axis at the corresponding position”.

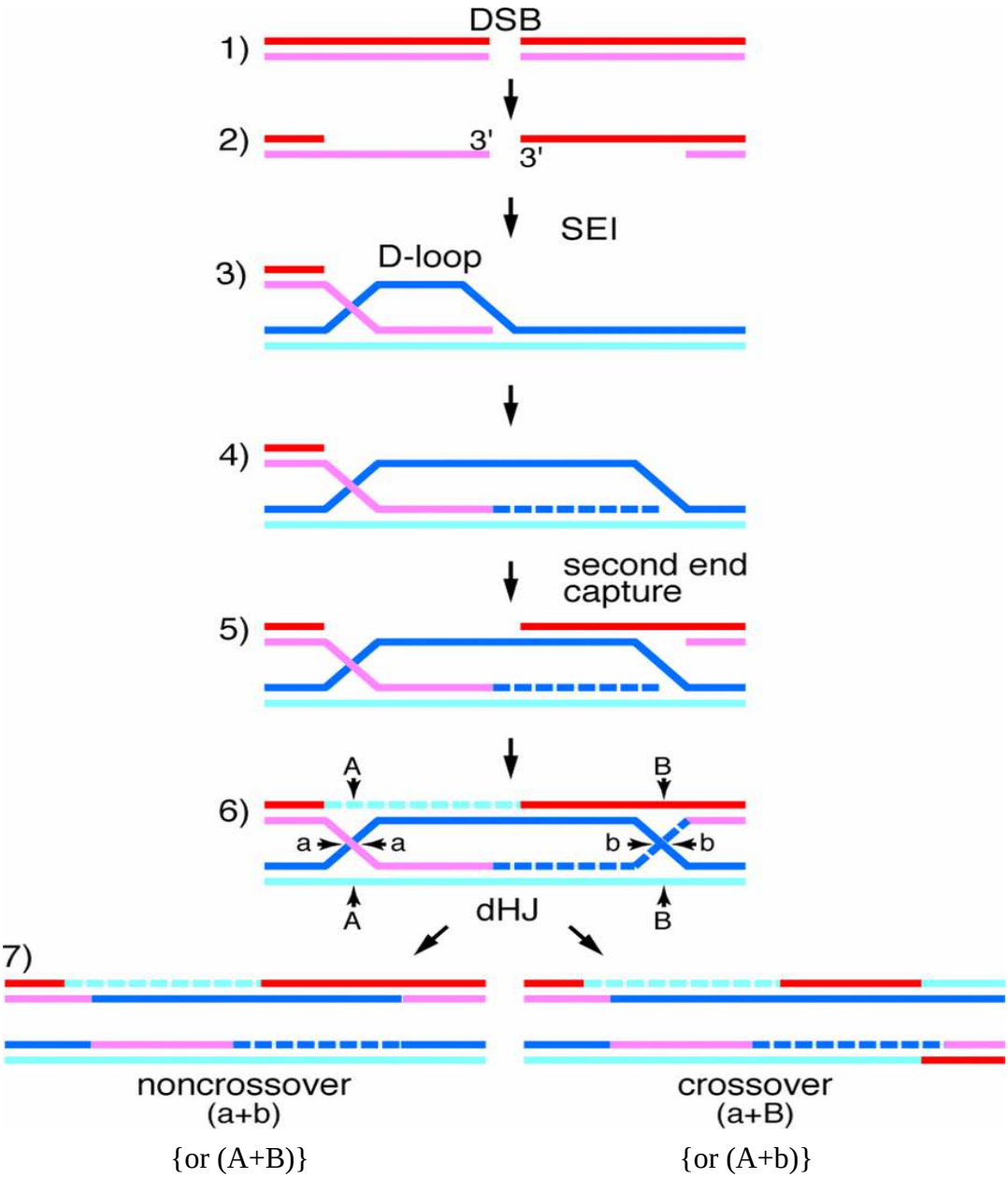


Figure 1.4 The DSBR Model of Recombination.

The classical explanation to the formation of crossovers and non-crossovers as proposed by Szostak *et al.* (1983). Colours pink and red are for the first homologue and the other is light and dark blue. Dashed lines indicate freshly formed DNA.

Step 1 - Double strand break (DSB) develops in one homologue.

Step 2 - DNA ends are modified by degrading their 5' strand ends giving single-stranded 3'-OH termini.

Step 3 - One of the 3' termini invades a complimentary sequence in the homologue forming a displacement loop (D-loop). Also called single end invasion (SEI).

Steps 4 & 5 - The invading strand extends the D-loop by initiating DNA synthesis. This aids it in annealing to the single-strand on the other side of the break (second end capture).

Step 6 - Continuing synthesis of DNA coupled with ligation of the breaks leads to the homologues being connected by two four-way DNA junctions, namely the double Holliday junction (dHJ).

Steps 6 & 7 - Cleavage of different strand pairs at each junction leads to chiasma formation and a crossover. When the same pair of strand is cleaved in the two junctions, it gives a non-crossover and no chiasma.

Figure from Whitby M.C. (2005) Making crossovers during meiosis. Biochem. Soc. Trans.

Intuitively, it would be expected that the events which lead DSBs to the SEI stage and thereafter to dHJ are also associated with events which develop an axial break in the exchange partner and then repair the break to bring about an axial exchange. Finally it is indeed probable that these axis changes dictate which DSB will move towards becoming a CO and which will become a gene conversion instead.

1.2 (e) *Homologous Recombination Models and 'Synthesis-dependent Strand Annealing'*

In order to delve more into the story of gene conversions, it is important to go back to the decision-making involved once DSBs are formed. Also, gene conversions are more aptly described as chromosomal sites where recombination brings about non-Mendelian segregation of one or several genetic markers or alleles. Classically, it has been thought that there are alternative resolution pathways followed from a common precursor dHJs (Holliday 1964). The Double-Strand Break Repair (DSBR) model (Figure 1.4) found support from several studies (Bell and Byers 1983, Szotak *et al.* 1983) which suggested that a recombination event can have either of two orientations of the dHJ intermediate resolution leading to a CO or a NCO.

Important transitional species and characteristics of the DSBR model have found acceptance through budding yeast (*S. cerevisiae*) studies (reviewed in Allers & Lichten 2001). But there have also been observations which challenge this model. Firstly several mutants have been recognized which exhibit high levels of DSBs and NCOs but reduced frequency of COs (Engebrecht *et al.* 1990, Fung *et al.* 2004). Proteins involved specifically in carrying out meiotic CO recombination include the following (Börner *et al.* 2004): (1) DNA helicase Mer3; (2) Msh4 and Msh5, relatives of *E. coli* MutS not contributing to mismatch repair; (3) Zip1, a structural protein of the central region of the SC; and (4) Zip2 and Zip3, two proteins aiding Zip1 to begin polymerization along homologues. Mutation in any one of these four reduces the number of COs (but not that of NCOs). This appears to advocate that COs are formed via a more intricate pathway than NCOs. Secondly, the heteroduplex DNA configuration was found to be different on NCO recombinants from what is expected from COs by the DSBR model (Gibertson & Stahl 1996, Porter *et al.* 1993). Again thirdly, physical detection methods have been utilized to detect that almost all of the dHJ intermediates are steps to form a CO in particular instead of being transitions on the way to form either COs or NCOs. Finally, more recent budding yeast studies (Allers & Lichten 2001, Hunter & Kleckner 2001) say that CO positions are determined much before in prophase than the time point at which dHJ resolution occurs. These studies profess the Early CO decision (ECD) model of recombination (Bishop & Zickler 2004) (Figure 1.5).

The Lichten group postulated that NCO recombinants are formed via a pathway not involving dHJ intermediates. Perhaps NCOs occur by ephemeral invasion of one or both of the two DNA ends resulting from a DSB. Such an invasion might allow the invading 3' end to be lengthened using the invaded duplex as template. It is possible that NCOs are formed via an alternative

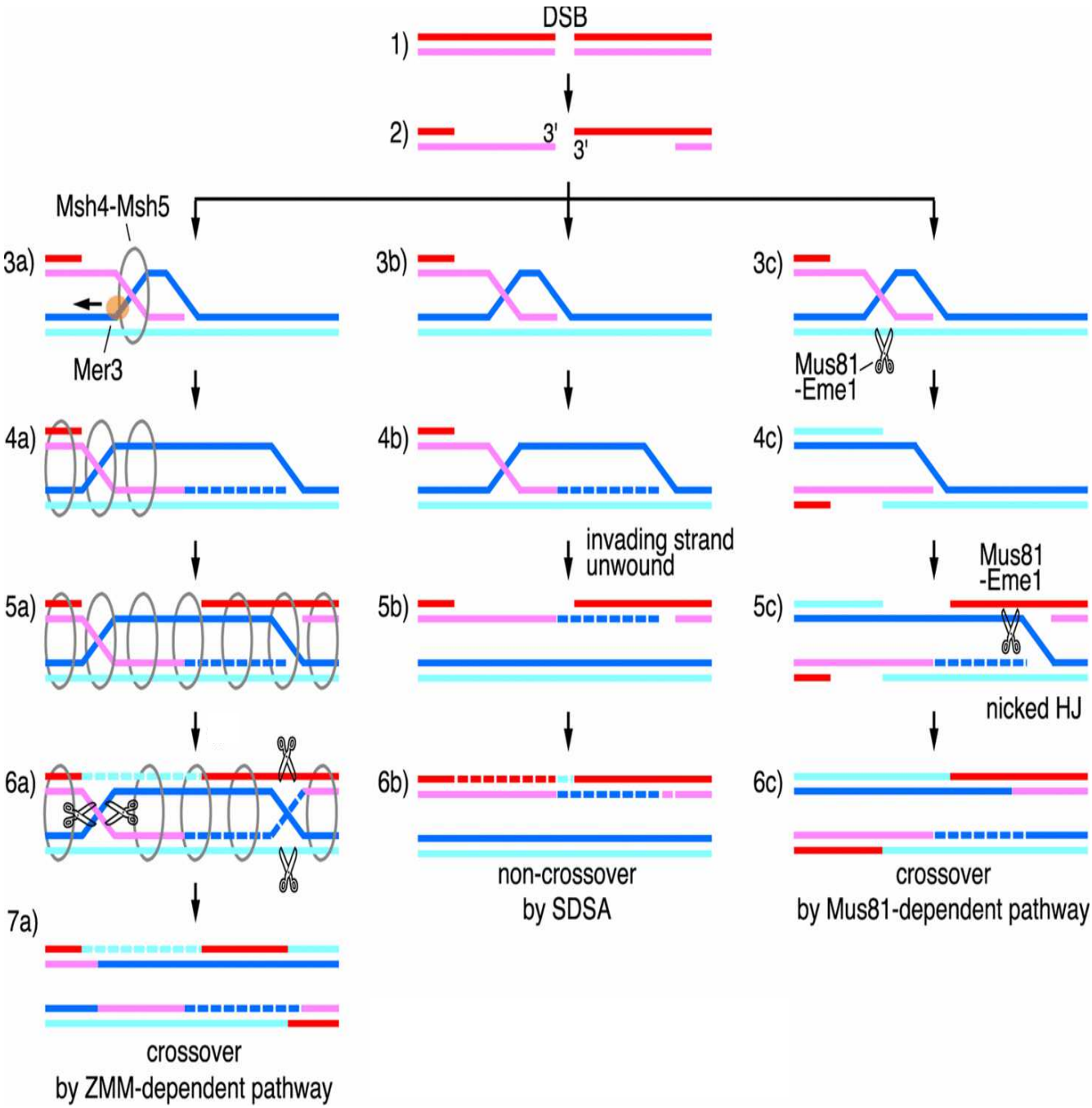


Figure 1.5 Present Meiotic DSB repair pathways (incl. ECD Model).

A DSB formed in one of the homologues is resected to give single-stranded 3' ends. One such end invades the other homologue (single end invasion or SEI; **Steps 3a, 3b, 3c**).

SEI is not very stable and may or may not recruit molecules to stabilize itself. In case of non-recruit of additional molecule for stability, the invasion is withdrawn after initial DNA synthesis giving the synthesis-dependent strand annealing (SDSA) pathway and thus, non-crossovers (**Steps 4b, 5b, 6b**).

The two other pathways lead to either interfering or non-interfering crossovers. If an SEI recruits the ZMM proteins (Msh4, Msh5, Mer3) to stabilize strand invasion intermediates, it results in the formation of the stable structure of a double Holliday junction cleaving it in a way so that only (interfering) crossovers are formed (**Steps 4a, 5a, 6a, 7a**). Another molecule that may be recruited by SEIs is the complex formed by Mus-81 with Eme1/Mms4. This preferentially avoids Holliday junction formation and cleaves D-loops instead to give (non-interfering) crossovers (**Steps 4c, 5c, 6c**).

Figure from Whitby M.C. (2005) Making crossovers during meiosis. Biochem. Soc. Trans.

mechanism, called “synthesis-dependent strand annealing” (SDSA), which involves dissociation of the SEI after DNA synthesis is followed by annealing of the two broken ends (Figure 1.5). This hypothesis sounds impressive and successfully explains most of present data. Subsequently it has been reported that once SEIs are formed, most of them remain unstable and result in NCOs following SDSA while an appropriate fraction is stabilized via various meiosis-specific proteins such as the ZMM proteins leading to dHJs and finally COs (McMahill *et al.* 2007). In addition there exists another but minority CO formation pathway as well dictated by Mus81 and Mms4 (details under Crossover Formation) to which branch some COs are pushed towards directly from the SEI stage or the nicked dHJ state (Whitby 2005) (Figure 1.5).

1.2 (f) *Assembling the Microtubule apparatus*

The microtubule organizing centers (MTOCs) for yeast and animals are respectively the spindle pole body (SPB) and the centrosome. MTOCs nucleate microtubules and mediate spindle orientation for cell division (Pereira & Schiebel 2001). Within the centrosome and the SPB lies a pair of centrioles that consist of γ-tubulin, the nucleation site for microtubules (Job *et al.* 2003). Even though γ-tubulin is conserved across eukaryotes, plants do not possess a discernible MTOC in the form of a centrosome. Microtubules nucleate in the cortical cytoplasm and on the nuclear membrane from sites along existing microtubules which are marked by γ-tubulin (Paradez *et al.* 2006). Thus microtubule nucleation for plants appears to be spatially distributed within the cell. Several studies have even hypothesized that perhaps γ-tubulin functions as MTOC for plants but the process is still not completely understood (Ehrhardt & Shaw 2006).

Spindle assembly is very integral to cell division, both mitosis and meiosis (Hyams 1996, Vernos & Karsenti 1995). Presently there exist two models of assembling the spindle, the ‘search and capture’ model and the ‘self-assembly’ model. In the ‘search and capture’ model, the centrosomes nucleate and organize the microtubules of the spindle (Kirschner & Mitchison 1986). The microtubules have their minus ends at the centrosomes as the microtubules are captured and stabilized when the plus ends come in contact with kinetochores, which are specialized protein complexes which assemble onto centromeric DNA. In this model, bipolarity and spindle orientation is established by the freshly divided centrosome, before nuclear envelope is lost (Vernos & Karsenti 1995). This model is based on observations on somatic and early embryonic cells of animals. And though centrosomes are absent in somatic cells of higher plants, bipolar spindle arrays are still formed before the nuclear membrane degrades, indicating that in the larger scheme of things the same process is followed in plant cells as well (Baskin & Cande 1990, Palevitz 1993).

The second model, ‘self-assembly’ is observed during meiosis of some animal species and higher plants. This model advocates that once the nuclear envelope ceases to exist, microtubules develop from multiple sites around the condensed chromatin. Thereafter the microtubules self-

organize into a spindle even though there is no centrosome or any well-defined MTOCs (Theurkauf & Hawley 1992, Vernos & Karsenti 1995, Chan & Cande 1998). During female meiosis in *Drosophila* (Theurkauf & Hawley 1992) and *Xenopus* (Vernos & Karsenti 1995), forming the meiotic spindle involves randomly oriented growth of microtubules which then self-organize into bipolar arrays. The same is observed during Maize meiosis (Chan & Cande 1998). The conundrum as to how microtubules developed bipolarity in the absence of centrosomes in this model subdued when it was suggested that bipolarity could be an intrinsic property of newly forming microtubule arrays (Heald *et al.* 1996).

Another aspect of spindle assembly which has been studied extensively is its elongation which follows after being assembled. There are at least two known processes among different organisms. For yeast and diatoms, spindle elongation depends on sliding of the inter-zonal microtubules which extend from both poles and interlink near the equator (Baskin & Cande 1990, Shelden & Wadsworth 1990). As the microtubules slide, the spindle elongates and the overlap reduces temporarily, which implies that continuous microtubule growth is required to maintain such overlap. In contrast in higher animals it has been proposed that elongation of the spindle is achieved by the pulling force exerted on the two opposite poles with no significant sliding of the inter-zonal microtubules (Mastronarde *et al.* 1993).

Further it is also known that microtubules are dynamic as the tubulins polymerize and depolymerize in 'bursts', enabling the tubules to grow in discrete movements. In fact it is possible that spindle elongation may occur in discrete steps and hence spindle lengths do not contribute to a continuous distribution (Yang & Ma 2001). And rather interestingly this characteristic has been observed among many diverse organisms, including slime mould (Moens 1976), the diatom *Fragilaria capucina* (Tippit *et al.* 1978), rat kangaroo (Armstrong & Snyder 1989), budding yeast *S. cerevisiae* (Winey *et al.* 1995) and *Arabidopsis* pollen (Yang & Ma 2001). And the striking commonality was seen to be that all spindles elongated by multiples of 0.7 µm!

1.3 *Comparison with Mitosis*

The challenge is the same during both mitosis and meiosis – to communicate an intact genome to the next generation. While mitosis keeps the chromosome composition and retains diploidy, meiotic products are a result of recombining parental features and result in haploid gametes.
Ploidy is maintained during mitosis via one round of DNA replication followed by one equal division (Figure 1.6). In contrast, meiosis leads to haploidy owing to one round of replication and two subsequent equal division rounds without any intervening replication (Figure 1.6). The first of these divisions (MI) leads to two products consisting of chromosomes with their sister chromatids intact and the next equal division of each of these finally gives four haploid gametes with one copy of each chromosome.

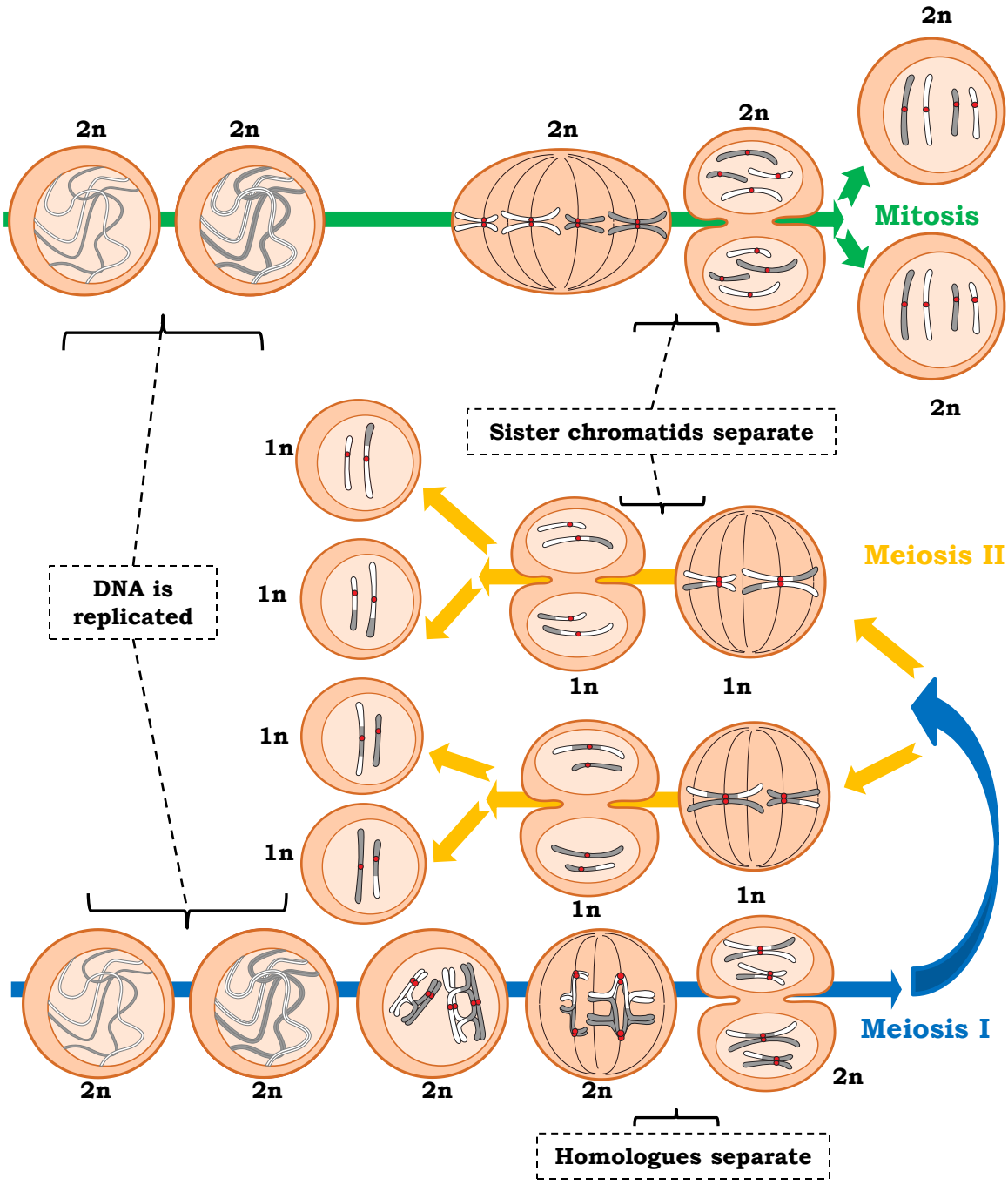


Figure 1.6 Comparing the Meiotic division against the Mitotic division.
Since mitosis gives begins with a diploid cell and gives 2 diploid cells, it is a cycle. But meiosis is not so as it begins with one diploid cell and gives four haploid gametes.
Nevertheless there exist common events such as DNA replication and the separation of the resulting sister chromatids albeit during different stages. Also, the homologous chromosomes never separate during mitosis whereas they do at the end of the first meiotic division. We illustrate these differences using the example of a cell again with $2n = 4$. The number beside each cell gives the ploidy.

Cell diagrams taken and modifed from the Thesis of Nicolas Macaisne (Defended on 26 November 2010)

So mitosis and MII are conceptually similar except the former works on a diploid cell and the latter applies to a half the number of chromosomes present in a somatic cell, but in both cases each chromosome has two copies (two chromatids) and separating sister centromeres successfully is the purpose of the division (Figure 1.6).

Faithful chromosome segregation is vital to the continued viability of a cell, an organism as well as its offspring. Segregation has thus evolved into a highly precise process. Mitotic chromosomes separate abnormally in yeast once every 10^5 divisions while during meiosis, segregation aberrations happen at the rate of one in 10^4 meioses in both yeast and Drosophila (Murray & Szostak 1985). In addition, though mitosis is a 'cycle', meiosis is not. This is to indicate that each mitotic division leads to two identical copies of the initial cell itself, each of which can in turn undergo mitosis and the process continues. Whereas meiosis leads to four haploid products each of which is unlike the initial cell – neither diploid nor identical genetically, hence, this process is not a cycle. Another distinct difference between the two divisions is in the role played by the cohesin complex. During MI, the sister chromatids are held together along their arms and centromere during prophase I. The cohesins are first released along their arms during metaphase I to ensure that sister chromatids remain together and then released from the centromere during metaphase II to separate sisters. In contrast during mitosis though cohesins hold chromatids in the same way to begin with, cohesins give way along the whole chromosome in one go during metaphase and not in stages. Thus cohesins mediate an intrinsic mechanistic hand.

Pre-mitotic *interphase* involves replication of the genetic material in each chromosome and leads to the appearance of sister chromatids and their dual kinetochores. Through *prophase* there is condensation of chromosomes till they transform to their shortest at *metaphase*. Onset of *metaphase* is ensured as the bipolar microtubules, which develop either from centrosomes in animal cells or induced directly from the cytoplasm as seen among plants, apply forces on the paired sister kinetochores of each chromosome. This phase concludes when the opposing stresses balance each other as a kinetochore per chromosome attaches itself according to the pole it orients itself towards. The chromosomes at this stage present themselves aligned at the equatorial plane of the cell with microtubular aid. The next chromosomal movement ushers in *anaphase* when the sister chromatids are pulled towards opposite poles as the spindle contracts separating the two identical chromosome copies from each other to the two poles. Hereafter, *telophase* leads to the formation of two nuclei having one diploid set of chromosomes each, usually followed by cytokinesis.

CROSSOVER FORMATION

COs are extremely important for faithful homologue segregation during MI as they ensure developing proper orientation and appropriate spindle force. Once the bipolar forces are

balanced, segregation is carried out successfully when the homologues separate at the COs. The role played by COs in implementing proper segregation cannot be stated enough as most human aneuploidy cases result from defective number and/or distribution of COs (Lamb *et al.* 2005). So when a gamete is formed, accurate transmission of an intact genome entails that enough number of COs are placed precisely throughout the genome so that each chromosome receives at least one (the obligate CO). However, the number of COs is not determined simply by the genome size as intra-specific differences in recombination rates are seen between sexes (Lenormand & Dutheil 2005, Giraut *et al.* 2011) and between genotypes within a species (Bauer *et al.* 2013). Also, COs do not occur randomly across genomes or chromosomes. Usually there are regions which show lower probability of CO occurrence than others (the centromere and its neighborhood for instance). And in addition CO positions are not independent either which constitutes the phenomenon of *interference*.

Coming back to recombination, the source of COs, we recall that meiotic recombination follows the Early Crossover Decision (ECD) model as opposed to the Double-Strand Break Repair (DSBR) model, also known as the Szostak Model (Szostak *et al.* 1983). What follows is a short tour through the many facets of CO formation and interference!

The DSBR Model of Recombination

It involves a single pathway of DNA intermediates which leads to both COs and NCOs wherein both are initiated by the formation of DNA DSBs (Figure 1.4). Each DSB results in two duplex ends each of which are processed to give single-stranded extensions by degrading one of the duplex ends to make them unequal in a complementary manner and generate scope for homologous interactions. Then single-end invasion (SEI) occurs where one of the two processed DSB ends invades the homologue, forming a displacement-loop (D-loop) structure (Figure 1.4). Subsequently the second processed DSB end adheres to the D-loop. This joint molecule (JM) then receives DNA repair to complete the single strand gaps. This molecule is held together by two structures in which strands are exchanged between the interacting duplexes. Individually these structures are called Holliday junctions (HJs) and together they are known as the double Holliday junction (dHJ). Severing pairs of strands in one of two orientations which is picked at random decides a CO or NCO outcome. Gene conversions can be associated with a CO or a NCO since they appear via mismatch repair in heteroduplex DNA regions in the JM. Though expected intermediates such as DSBs have been observed in budding yeast, fission yeast and concluded in mice (Schwacha & Kleckner 1995, Mahadevaiah *et al.* 2001, Qin *et al.* 2004), it is increasingly becoming evident that this model is not generic but specific to only CO formation.

CO and NCO formation pathways

DSBs are precursors to both COs (reciprocal exchange of chromosomal segments between non-sister chromatids) as well as NCOs (a non-reciprocal exchange of the same kind) but the other

transitory molecules of the DSBR model (D-loops and dHJs) seem to be a part of only the CO pathway. For instance in budding yeast, comparing the timing of dHJ appearance with the appearance of COs and NCOs suggests that they are precursors only to COs (Allers & Lichten 2001). Several mutant analyses in budding yeast, mice and *Drosophila* also lend support to this (Allers & Lichten 2001, Guillon *et al.* 2005, Bhagat *et al.* 2004). Thus it appears that the *decision* to form a CO or a NCO is taken *early* which is after DSB formation but before HJ or even SEI formation.

For NCOs, as previously mentioned, the SDSA or 'synthesis-dependent strand annealing' diversion is suggested. DSB formation if followed by development of the usual single-stranded overhangs, strand invasion, transient DNA repair and a subsequent strand 'pull-out' before reaching the HJ stage (Figure 1.5). A budding yeast study (Terasawa *et al.* 2007) has been reported in support for this conjecture.

Usually all DNA repair relies on using a non-sister chromatid from the homologous counterpart as a template. This is ensured as there is a barrier in place to avoid sister chromatid-assisted repair during meiosis (Carballo *et al.* 2008). In situations when this is not possible, DSBs might use sister chromatids to carry out repair, which does not lead to CO formation and thus may result in error-prone homologue separation.

Two CO formation pathways

Studying various organisms has led to the conclusion that indeed there exist at least two CO formation pathways, each of which is monitored by a group of specific proteins and thus imparting different spatial properties to the resultant COs.

One pathway depends on the Msh4-Msh5 complex (from the 'ZMM' protein group) and it exhibits a form of CO position control, known as *interference* (details later). The second pathway, exhibiting no *interference*, is mediated by the Mus81-Eme1 (Eme1 is called Mms4 in budding yeast) complex. Nevertheless both these pathways and the NCO formation pathway begin with DSBs (Figure 1.5).

But the situation varies between organisms as for fission yeast and some eukaryotes such as fruit flies (Sekelsky *et al.* 2000) lack the *interfering* CO pathway, which is active in mammals, budding yeast and plants. And the pattern of interference seen in each organism relies on whether one or both pathways is/are found. The worm *C. elegans* prefers only the interfering pathway thus suffering extreme interference (de los Santos *et al.* 2003) and the other extreme behavior is seen in fission yeast with *no* interference (Munz 1994). Further, budding yeast, plants and mammals seem to be in possession of proteins for both pathways to form COs (de los Santos *et al.* 2003, Guillon *et al.* 2005, Copenhaver *et al.* 2002, Housworth & Stahl 2003, Mercier *et al.* 2005, Lhuissier *et al.* 2007, Falque *et al.* 2009). Hence, both interfering as well as non-interfering COs are present.

Though both CO formation pathways initiate at DSBs, they might proceed via differing DNA intermediates. In fission yeast with only the Mus81 pathway, single HJs (sHJs) were found to be prevalent via electron microscopic (EM) and gel electrophoresis studies (Cromie *et al.* 2006). In addition, an EM analysis indicates a sizeable minority of sHJs in budding yeast (Osman et al. 2003). These observations suggest that sHJs were targeted by the Mus81 pathway while dHJs mediate the ZMM pathway. An sHJ might arise by cleaving a D-loop before instead of after the second end is also involved in invading the homologue.

Crossover Position, Number and Interference Regulation

There are many successive levels of control exerted on the different aspects of crossovers. To begin with, the DSBs already have a non-random distribution, appearing more frequently in specific regions known as *hotpots* along chromosomes in many organisms (de Massy 2003).

DSB distribution and number

Early studies revealed local chromatin structure to be an important indication for DSB formation. DSBs were more abundant (the *hotspots*) in regions that correlated with nuclease-hyper-sensitive regions (Petes 2001). In fact, histone H3 trimethylation of lysine 4 (H3K4me3), an epigenetic mark for active chromatin appears to mark DSBs in *S. cerevisiae* (Borde *et al.* 2009). Higher order chromosome features also yield control on the DSB number and positions, as seen in *C.elegans* mutants (Tsai *et al.* 2008).

Recently several groups have undertaken genome-wide studies in budding yeast to explore meiotic DSB positions using Spo11-based chromatin immune-precipitation (ChIP) and microarray analysis or the technique commonly known as ChIP-chip. Spo11 is a topoisomerase-like active-site like protein that is responsible for DSB formation. Hence ChIP-enriched loci are concluded to be the positions where DSBs have appeared. Such mapping of DSBs in *S. cerevisiae* shows that majority of DSBs occur in intergenic regions consisting of promoters and in regions 20-120 kb away from telomeres but are absent from the 20 kb region adjacent to telomeres (Blitzbau *et al.* 2007). Rather unexpectedly, DSBs were also found in pericentromeric regions which are bereft of COs and which were previously thought to be DSB *coldspots* (Blitzbau *et al.* 2007, Smith *et al.* 2003). These results clearly demonstrate that DSB positions are monitored both locally (via chromatin) and globally (position relative to telomeres).

To be or not to be a CO

In most organisms studied till now, the number of DSBs outnumber resulting CO number manifold. For example, in maize for a particular meiosis, it was observed that of 560 post-DSB intermediates, only about 20 matured into COs (Franklin *et al.* 1999). Interestingly it has been concluded in budding yeast, human and mice that, COs and NCOs arise from the same recombination-rich regions (Allers & Lichten 2001, Jeffreys & May 2004, Guillon & de Massy 2002, Hunter & Kleckner 2001). This is consistent with the hypothesis that the CO v/s NCO fate

of a DSB is not pre-determined merely from its genomic position but follows from a more intricate set of molecular modifications. On the other hand, in maize, the distribution of DSBs (from RAD51 ChIP-Seq) is almost uniform even near centromeres. But this is very different from the distributions of the COs which subsequently form showing scarcity (of COs) in large peri-centromeric regions (W.P. Pawlowski *et al.*, personal communication). Thus determining which DSB will transform into a CO is the second vital level of control exercised in the journey of CO formation.

Interference

In addition, the distribution of COs is extremely non-random in most organisms studies till now (Figure 1.7). And though the overall CO number tends to be low, even the smallest chromosome receives at least one (obligate CO rule). Thus there seem to be two tenets to follow here. First is the presence of the mandatory CO on each chromosome. Second, there is a suppression gradient in the vicinity of each CO which diminishes the probability of nearby COs without affecting the occurrence of DSBs or NCOs. This suppressive behavior is known as CO *interference* (Figure 1.7).

Initially it seemed that it is the SC which communicates interference along chromosomes. Especially since the only organisms known without interference, the two fungi *S. pombe* and *Aspergillus nidulans* also lack SCs and a mutation in budding yeast surrounding Zip1, an SC component, obliterates SC as well as interference. But contemporary work suggests that the interference plan is set out before the appearance of the SC. The Zip2 and Zip3 are proteins from the ZMM group, used to devise interfering COs in particular (Börner *et al.* 2004). Msh4-Msh5, also a ZMM group member, is found to co-localize with Zip2-Zip3 (Novak *et al.* 2001). While Zip2-Zip3 foci mark the interference-sensitive CO loci in budding yeast, Msh4-Msh5 do the same for mice (Rockmill *et al.* 2003, Fung *et al.* 2004, Henderson & Keeney 2004, de Boer *et al.* 2006). Since SC nucleation is also initiated at these foci, it implies that SC begins to appear from CO positions exhibiting interference.

Further an interesting study in the worm *C. elegans* reveals that the chromosome axis structure might propagate interference. In *C. elegans* only the Msh4-Msh5 pathway is present with severe interference which always gives exactly one CO per chromosome (Meneely *et al.* 2002). Interestingly, when two or even three chromosomes are fused end-to-end, instead of forming two or three COs respectively, it again gave only one CO in a chromosome CO (Hillers & Villeneuve 2003). So for this meiotic system it seems that CO interference is controlled by considering a chromosome as a unit, no matter what its length, it is ensured that each chromosome receives one. This exemplifies a chromosome-wide control on CO distribution that can extend beyond the usual chromosome length specific to an organism.

A *stress relief* model (discussed in detail later, Figure 1.9) professes a link between CO interference and changes in the physical properties of a chromosome – mechanical stress along

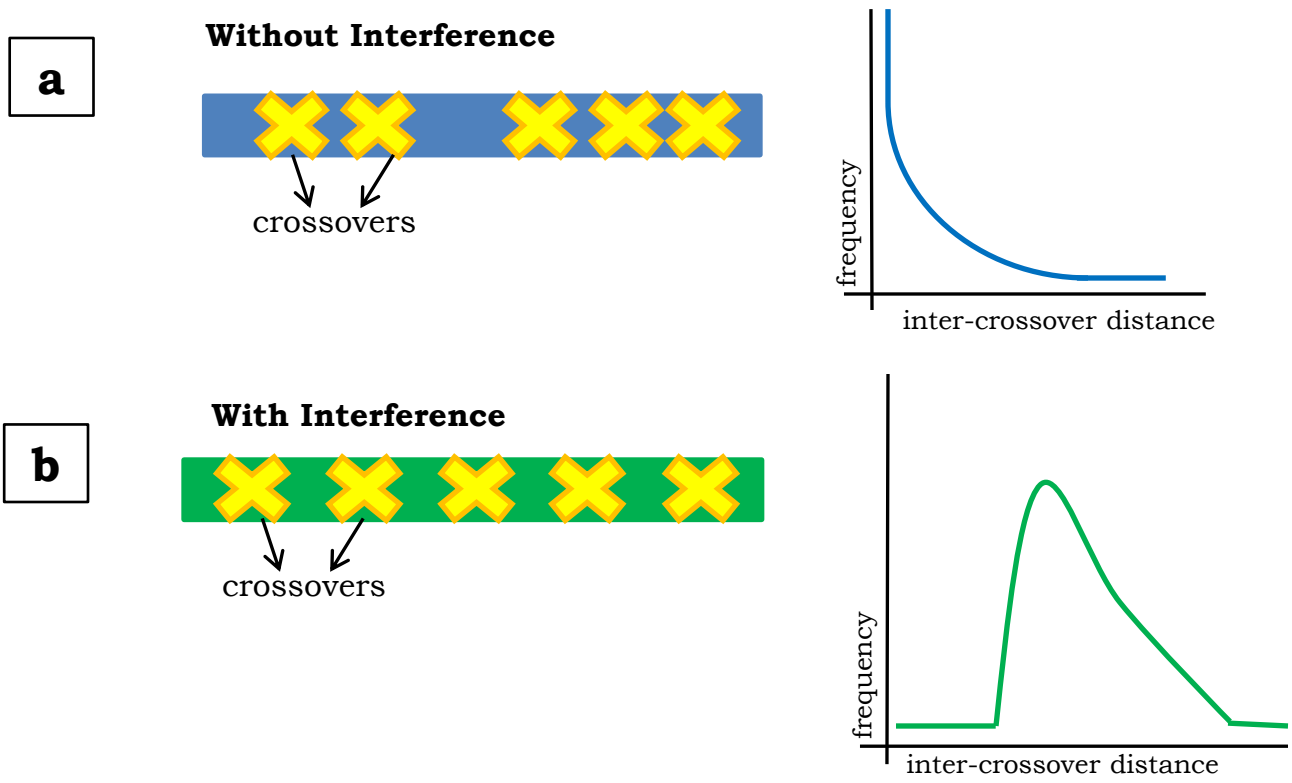


Figure 1.7 Inter-crossover distance distributions.
Distances between non-interfering crossovers give an exponential distribution (a). And interfering crossovers tend to be separated by distances which are much more regular and have a lower variance.

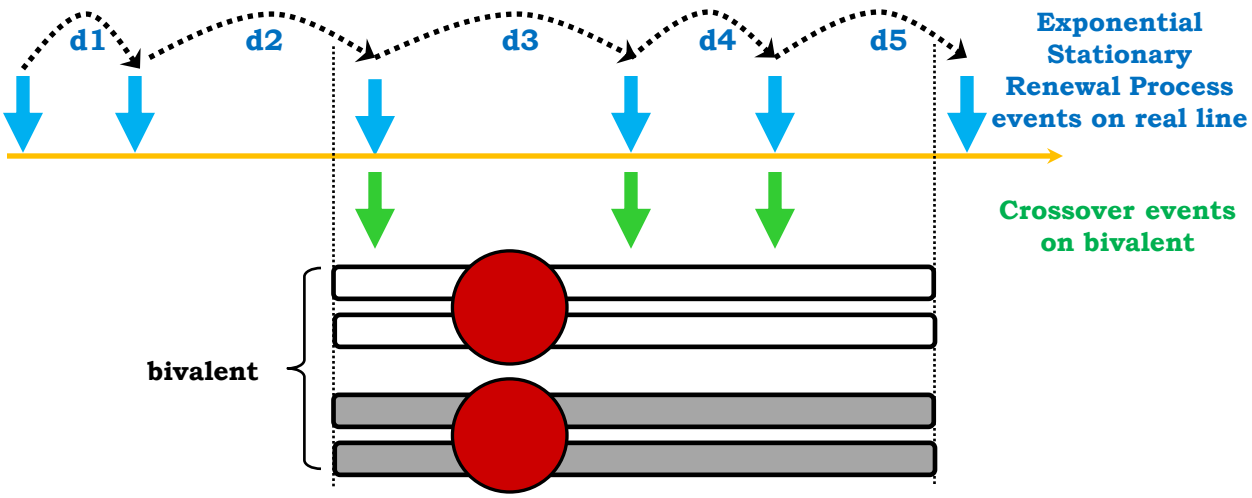


Figure 1.8 The Poisson Model (without interference).
If the number of crossovers is drawn from a Poisson distribution and then they are distributed randomly along the chromosome to mimic the without interference case. Or, the inter-event distances are considered from an exponential distribution and placed along the real line (orange). Those event positions included in the bivalent length alone are taken as crossovers, again with no interference.

chromosomes initiates CO formation and COs relieve this stress locally and thus prevent occurrence of additional COs closeby (Börner *et al.* 2004, Kleckner *et al.* 2004). An added level of complexity here is that since all COs do not show interference, this model has difficulty in generalizing the CO formation process to all COs. This also leads us to the subtlety that interference is not a characteristic that is intrinsic to every CO event *per se*. Moreover, genome-wide mapping of recombination events has also revealed that interference is present between COs and NCOs as well (Mancera *et al.* 2008). This very short range interference may be the consequence of interference between DSBs (de Boer *et al.* 2006). Note that this kind of interference is unlikely to contribute significantly to CO-CO interference because the distance scale of interference between DSBs is much smaller than the average distance between COs. It remains a non-trivial problem to understand interference and its molecular mechanism.

Crossover Homeostasis

A recent work discovered a hitherto unknown and non-linear relationship between DSBs and COs (Martini *et al.* 2006). Mutants of *S. cerevisiae* with DSB levels as ~80%, ~30%, ~20% of those present in the wild-type were utilized to examine the effect on the outcome of COs. CO frequencies were observed for eight intervals along the length of three chromosomes. Surprisingly, it was noted that despite reduced DSBs, the meiotic CO levels were maintained. This occurrence was termed as crossover homeostasis.

But some genome regions were less capable of exhibiting this phenomenon than others. For example, when the *ARG4* locus (a meiotic recombination hotspot in the genome) was looked at more closely, it seemed that as DSBs reduced, CO frequency was maintained by compromising on the number of NCOs. More interestingly, when DSB levels decreased, there was slight or negligible impact on interference intensity or the distance till which it could be diagnosed. Thus interference in a given chromosomal interval might be quite independent of the DSBs in the region. In an overall glance, CO homeostasis seems to lend a bias towards CO formation, probably to ensure proper chromosome segregation during the first half of meiosis. This also indicates that this homeostasis might have a role to play in making the obligate CO and also might be involved in the molecular apparatus bringing about CO interference (Martinez-Perez & Colaiacovo 2009).

Heterochiasmy: opposing Recombination

Recombination rates may or may not vary between sexes. There are several comparative data on male and female meiosis for plants as well as animals showing that recombination levels remain the same in both sexes for some organisms, while in others, extreme differences exist. The disagreeability in the frequency of recombination may be restricted to a specific portion of the genome. Alternatively the difference could be negligible, constrained or unconfined. These possibilities manifest in various combinations with respect to recombination.

A typical observation has been that reduced recombination (higher linkage) occurs in the heterogametic sex than the homogametic (1918 Haldane rule). Several studies have reported achiasmate meiosis, usually in the heterogametic sex in plants as well as animals. Among human populations too, mapping linkages using segregation and pedigree analysis reflects lower *intra*-chromosomal recombination (for instance in chromosome 1) in males than in females (Karlin & Liberman 1994, Lenormand & Dutheil 2005).

Finally, theoretical understanding of differential sex-related recombination reveals that such dimorphism promotes polymorphic expression. This conclusion is in adherence to the finding that recombination rate might be affected by agents such as temperature, age, internal physiology, overall energy, enzymes, radiation, enzymes and so on. These factors influence genders disparately to maximize overall fitness. This characteristic aids populations to efficiently harness inherent environmental properties via fitness plasticity and this flexibility is truly expanded by sexuality.

A HISTORICAL PERSPECTIVE ON MODELING CROSSOVER INTERFERENCE

CO distribution along chromosomes has puzzled researchers since interference was observed as a phenomenon. The observation was made by Alfred H Sturtevant in 1913 that crossovers tend to be further away from each other in Drosophila than would be expected if their positions had been random (Sturtevant 1913). Thereafter, he addressed this behavior as 'interference' and duly acknowledged Muller's contribution to the discovery and the term (Sturtevant 1915). It is to be noted that here interference refers strictly to *positive* interference where the COs are further away from each than would be expected if they were distributed randomly across chromosomes. It follows that if COs are placed closer than if they were positioned randomly, then interference is *negative*.

Since then it has been a pertinent question in the meiosis community to delineate the molecular machinery behind interference. But traditional genetic screens are quite labor-intensive. Also, each protein playing a role in the interference story will more often have an integral part in meiosis as well, which adds to the complexity of the problem. Thirdly, till now mutants have been obtained with disrupted interference though the few of these which are not accompanied by changes in CO frequency are all in *S. cerevisiae* (Berchowitz & Copenhaver 2010). Lastly, interference is not an absolute phenotype in most species, it depicts a reduced probability of nearby COs in the most common scenario when interfering and non-interfering COs co-exist.

On the other hand, modelling is a non-invasive technique which can ably complement experimental insights. It is possible to test statistical or mechanistic hypotheses about CO formation and subsequently accept or reject them. A complicated system such as the one at hand needs to be approached with caution. As a model becomes more complicated, several challenges arise - either computational or data-related limitations. Hence it is the earnest endeavor of

modelers to always begin with the simplest structure and introduce modifications (read complexities) gradually and slowly.

There are two perspectives to understand with respect to modelling about CO formation. One class of models is based on statistics of various associated measurements, the inter-CO distances in particular. The second category pertains to the mechanistic models.

Determining Crossover Positions

For any kind of modeling, it is important to take into consideration the nature of the data at disposal. Earlier, mainly recombination data was available where genetic markers would be placed along chromosomes and experimental algorithms were utilized to deduce recombination between adjacent markers (usually at *gamete* level) following meiosis. Often the number of markers were far and few such as *nine* marker loci along the X-chromosome in Morgan's (Morgan *et al.* 1935) *Drosophila* dataset as opposed to present-day data with as many as *hundred* markers along each chromosome (Maize in Bauer *et al.* 2013). As the number of markers (apart from sample size) increase, the uncertainty associated with 'real' CO positions reduces. Further, more than the number of markers, the number of individuals often limit the scope of statistical analysis deemed possible later. The Giraut Arabidopsis dataset (Giraut *et al.* 2011) had an unusually large backcross population (about 1500) for both male and female meiosis but such sample sizes are rather rare often due to infeasible experimental costs and also the sheer labour involved in preparing the samples.

In addition, more understanding about the meiotic process has resulted in a different kind of data as well. Using immunofluorescence labeling and EM techniques, late recombination nodules (all COs) and MLH1 loci (interfering COs only) can be demarcated which provide us with rather precise CO positions along the SCs (*bivalent* level). So apart from the indirect way of observing CO formation via recombination, direct methods have also become possible.

From the modeling perspective, it is important to keep in mind whether we are dealing with data at the gamete or the bivalent level. However it is also important to first understand the system at the bivalent level and then come down to the gamete level. This is the process which is termed as *thinning*. The number of COs is halved on a gamete as compared to a bivalent. We go from the bivalent to the gamete using precisely this understanding. Each CO is equally probable to be present or absent on the resulting gamete. So from the CO pattern on a bivalent, COs are kept or removed on the derived gamete with probability ½ to perform thinning.

Recombination: When recombination is observed across two adjacent genetic markers, it is implied that an odd number of COs (1 or 3 and so on) have been formed in the intervening region. In a likewise manner, absence of recombination denotes the presence of an even number of COs (0 or 2 and so on).

Statistical Methods

Coincidence: Let there be two disjoint genomic regions *A* and *B*. The effect of CO recombination was assessed by Muller (1916) via the *coefficient of coincidence* (C)*:*

$$C(A,B) = r(A\ and\ B)/(r(A)r(B)) = r(A|B)/r(A) \quad \ldots (1.1)$$

where, $r(A\ and\ B)$ stands for the event that there is a recombination in region A as well as region B. So C is expressed as the ratio between the probability of recombination occurring in A given that recombination is known to exist in B, and the unconditional probability of recombination in A. An equivalent expression is:

$$C(A,B) = (r(A) + r(B) - r(A+B))/(r(A)r(B)) \quad \ldots (1.2)$$

where, $(A+B)$ is a combination of regions A and B resulting in $r(A+B)$ denoting the $Prob$((there is recombination in region A accompanied by no recombination in B) or (no recombination in A while recombination is seen in region B)). As recombination rates are really probabilities, this coefficient can be equivalently presented in such terms as well. Let $p_{i_1 i_2}$ denote the probability of i_1 recombination events in the first interval A and i_2 occurrences in second region B, $i_1, i_2 = 0,1$, then we define C as:

$$C = p_{11}/((p_{01} + p_{11})(p_{10} + p_{11})) \quad \ldots (1.3)$$

Finally, interference is classified based on the value of coincidence as follows:

$C(A,B) > 1 \Leftrightarrow$ *negative* interference

$C(A,B) < 1 \Leftrightarrow$ *positive* interference $\quad \ldots (1.4)$

$C(A,B) = 1 \Leftrightarrow$ *no* interference

Cytological data more often exhibit positive interference. But negative interference has also been observed between chromosomal arms for instance such that occurrence of a CO on one arm increases the probability of there being one on the other arm (Karlin & Liberman 1994). More recently (Auger & Sheridan 2001), apparent negative interference has been concluded in Maize while mapping reciprocal translocation breakpoints using classical genetic markers.

Another related interference measure, S_4 may be defined as below (Foss *et al.* 1993):

$$S_4(d) = lim_{h\to 0}\ lim_{k\to 0}\ p_{11}/((p_{01} + p_{11})(p_{10} + p_{11})) \quad \ldots (1.5)$$

Here, $p_{i_1 i_2}$ have the same definition as for C where the lengths of the intervals are h and k respectively which are at a genetic distance d from each other. S_4 provides a better perspective on interference than C (McPeek & Speed 1995).

Approaches galore

The Mather Case: Consider the simplest case – a homologous pair is represented by a four-stranded structure of which crossing of two of the four chromatids leads to a CO. Mather's formula (1936) predicts that for all COs, under no chromatid interference assumption, the probability of recombination in a well-defined region A on a specific chromatid is given by:

$$r(A) = 1/2[1 - P_0(A)] \quad \ldots (1.6)$$

where, $P_n(A)$ denotes the probability of a total of n COs in region A on the four strands. This expression shows that the highest value for the recombination rate of a genomic region is ½. Mather also gave an extension which is useful to deal with the recombination distributions from multiple markers. Let there be m distinct genomic regions, namely, $A_1, A_2, \ldots, A_m$. If E_i represents a recombination event (*odd* number of COs) on A_i and F_i indicates the presence of *at least one* CO in A_i considering the four-strand ensemble, then we know as follows (Karlin & Liberman 1983, Risch & Lange 1983):

$$Prob\{E_1, and E_2 and \ldots E_m\} = \left(\frac{1}{2^m}\right) Prob\{F_1, and F_2 and \ldots F_m\} \quad \ldots (1.7)$$

Formula (1.5) gives us bounds such as, Prob(double recombinant) $\leq 1/4$, Prob(triple recombinant) $\leq 1/8$ and so on. In contrast when chromatid interference is allowed, the bounds increase to 1. But the hunch on this interference has thereafter been unsubstantiated (Zhao *et al.* 1995).

Map Functions: These functions (reviewed in Zhao & Speed 1996) scale an observed recombination ratio between genetic markers into the distance between COs. Such ability would allow us to convert the observed recombination fraction r to the genetic distance d between COs, written as, $r = M(d)$. Also a consistent process to do so for all organisms might provide a novel outlook on meiosis.

Haldane proposed the first map function (Haldane 1919) and a differential equation method which served as a scaffold for the subsequent map functions. So the Haldane map function assuming no interference between COs is given as:

$$r = \frac{1}{2}(1 - e^{-2d}) \; with \; corresponding \; inverse \; d \; = -\frac{1}{2} log(1 - 2r) \quad \ldots (1.8)$$

When using map functions, it is assumed that the map function or the functional relationship between genetic distances and recombination fractions remains the same along the entire genome of a particular organism. So if we consider three genetic markers, then this in turn means that the

associated coincidence C would depend only on the inter-marker distances. Let these markers be A, B, C while the distances between them be d and h. Then we have using the $p_{i_1 i_2}$convention introduced before:

$p_{01} + p_{11} = M(h),\ p_{10} + p_{11} = M(d)\ and\ \ p_{11} = (M(h) + M(d) - M(d+h))/2,$

Thus, $C(d,h)\ = (M(d) + M(h) - M(d+h))/\ 2M(d)M(h),$

Finally if $h\ \to 0$ and we assume that $lim_{h\to 0}\ (M(h)/h) = 1$, we have the differential equation:

$M'(d) = 1 - 2C(d)M(d)$ …. (1.9)

Varied values substituted for $C(d)$ will give different map functions. In practice, when markers are dense, d is proportional to r easing things out for modeling.

We next put forward the two most well-known map functions while associating them with Stationary Renewal Processes (SRPs).

Relating Map functions & SRPs: Usually there are two levels to consider recombination, either at the four-strand level or at the single strand level. Crossover modeling has been approached as a renewal process since a long time. For example, Fisher (1947) modeled crossovers via renewal processes at the single-strand level where in crossovers were considered to for along the strand in an order, beginning from the centromere and the distances between them followed a specific distribution. Now of course it is known that crossover formation does not begin at the centromeres (Whitehouse 1982) and a single-strand system was considered without connecting it to the more realistic chiasma process. Subsequent work attempted ensured that the association to a chiasma process was included (Cobbs 1978, Stam 1979).

Previously it was thought that chromatid interference also exists but it was later found to be inconsistent with experimental observations (Whitehouse 1982, Zhao *et al.* 1995a). Thus crossover interference exists of the two interferences. The assumption of no chromatid interference (NCI) imposes some conditions on map functions (Speed 1995) just as it would affect the mathematical notion of recombination:

$$0 \le M(d) \le \frac{1}{2}$$

$$M'(d) \ge 0$$

$$M''(d) \le 0$$

Also, if the point process under consideration is simple and stationary, then, $M(0) = 0$ and $M'(0) = 1$ (Daley and Vere-Jones 1988). Finally, a map function M defined on $[0, \infty)$ satisfies condition ‘A’ if:

$$M(0) = 0, \qquad (A1)$$

$$M'(d) \geq 0, for\ all\ d\ , \qquad (A2)$$

$$M'(0) = 1, \qquad (A3)$$

$$M''(d) \leq 0, for\ all\ d, \qquad (A4)$$

$$lim_{d\to\infty} M(d) = \frac{1}{2}, \qquad (A5)$$

Under the NCI assumption and that condition that the crossover process is a simple stationary point process, except $(A5)$. This left-out condition professes that two genetic markers placed very far apart on a chromosome usually segregate independently as they tend to behave as markers on different chromosomes altogether. Further here is a theorem to define this class of map functions (Zhao & Speed 1996):

'*Let map function, M, adhere to the NCI condition on a chromosome arm of indefinite length and denote an SRP. This indicates that M satisfies 'A'. Conversely, if there exists a function which maps $[0,\infty)$ into $[0,1/2]$ such that it meets condition 'A'. Then there is an associated SRP with map function, M and renewal density, – M''.*' … (T1)

The two most well-known map functions are Haldane (1919) and Kosambi (1944).

1. *modified* Haldane map function

This function gives us that:

$$d = M_H^{-1}(r) = 0.7r + 0.3(-\frac{1}{2}\log(1-2r))$$

This implies that, then $M^{-1}(0) = 0$ ... $(A1)$

Again we have, $(M^{-1})'(r) = 0.7 + 0.3/(1-2r) \geq 0$ ... $(A2)$

and thus, $(M^{-1})'(0) = 1$ ... $(A3)$

Now, $(M^{-1})''(r) = \frac{0.6}{(1-2r)^2} > 0,$ ... $(A4)$

Finally, $lim_{r\to\frac{1}{2}} M^{-1}(r) = \frac{0.7}{2} + 0.3\left(-\frac{1}{2}\log 0\right) = \infty,$ ... $(A5)$

Clearly, condition 'A' is satisfied by M_H which implies that due to Theorem (T1), there is a stationary renewal process resulting in M_H much like the traditional Haldane which is only the second part $(-\frac{1}{2}\log(1-2r))$.

2. Kosambi map function

Here we have that, $M_K(d) = \frac{1}{2}tanh(2d) = \frac{1}{2}(1 - e^{-4d})/(1 + e^{-4d})$.

Then, $M_K(0) = 0$ ... $(A1)$

and, $M'_K(d) = \left(\frac{1}{2}\right)[4e^{-4d}\{(1 + e^{-4d}) + (1 - e^{-4d})\}](1 + e^{-4d})^{-2}$

$= 4(e^{2d} + e^{-2d})^{-2} > 0$ ... $(A2)$

Next, $M'_K(0) = \frac{4}{4} = 1$ ... $(A3)$

Also, $M''_K(d) = -16(e^{2d} - e^{-2d})(e^{2d} + e^{-2d})^{-3} < 0$ ... $(A4)$

Lastly, $lim_{d\to\infty}M_K(d) = (1/2)lim_{d\to\infty}\tanh(2d) = \frac{1}{2}$ ... $(A5)$

Thus we can safely say that condition 'A' is satisfied and the inter-event distribution is given by $-M''_K(d)$.

Each genetic map function represents varying degrees of interference. Hence, the same map function does not fit data from different organisms as each organism has its own interference magnitude range. Since map function do not give joint recombination probabilities in case of genetic markers number more than three, it is rather inconvenient to use them. Map functions have been extended in several ways to make them more generic in their usage.

One such approach involves assuming that the recombination pattern across two marker intervals is independent of the distance between them, but this does not support observed data (Liberman & Karlin 1984). Since the idea was found to be flawed, this criterion had expectedly out-ruled some map functions which fitted well to recombination data, such as the Kosambi function.

Nevertheless, we saw above that there is another way to modify map functions towards crossover interference modeling. Theorem (T1) gives a way to extend these functions to analyze multilocus data (Zhao & Speed 1996). If it becomes possible to find a point process such that it generates the map function under consideration, then multilocus joint recombination probabilities which are readily compatible with the map function follow. Various map functions can be derived from SRPs, the most common among which have been explained above – Haldane and Kosambi.

Towards Crossover Formation Processes: There are several ways in which this process has been approached. Here we present the most persistent of them in limited detail – renewal and count-location processes and, a less studied but interesting possibility, namely Markov process.

(a) *Serial Recombination Processes:* This kind of a (renewal) CO formation process is described as follows. Let us assume that each chromosome arm is defined in the interval

$[0, l]$ where 0 represents a natural starting point such as the centromere. Subsequently the COs begin to arise the centromere onwards, in a sequential manner. The inter-crossover distances are regarded to be independent and identically distributed according to a probability distribution function, F. In fact the renewal process is completely defined by this function. When $F(x)$ is an exponential distribution, the CO positions are derived from a Poisson process (Haldane case, discussed later). CO data for Drosophila and *Neurospora crassa* have been modeled as a specific renewal process with $F(x)$ being a Gamma distribution allowing for only integer-valued shape (Foss *et al.* 1993, McPeek & Speed 1995).
Nonetheless, CO events occur simultaneously at several positions along eukaryotic chromosomes rendering serial models less generic. In contrast, viruses and fungi might be appropriate for applying renewal processes to (Fisher *et al.* 1947, Owen 1950).

Further, a renewal process is termed as stationary when the occurrence of successive events is independent of the position of the first event (in this case it denotes CO formation). A special kind of stationary renewal process (SRP) called chi-square models were examined recently. The models were named after the distribution followed by the distances between COs but it is a special case of gamma, being restricted to have only even degrees of freedom. This model was found to show efficient goodness of fit to data of several organisms (Zhao *et al.* 1995b).

(b) *The Count-Location (C-L) Process:* Consider a chromosome of length l and markers numbered $1, 2, \dots, n+1$ arranged along it randomly. The process is outlined by a discrete probability set, $\{c_0, c_1, c_2, \dots, : \ c_i \geq 0, \sum c_i = 1\}$ and a family of distribution functions $\{F_k(x)\}_{k=1}^{\infty}$ defined continuously on the interval $[0, l]$. $\{c_i\}$ denote the probabilities that k COs are formed. Given that k COs have been formed, these COs are distributed along the chromosomes as k independent samples from the distribution $F_k(x)$.
A special case is when $F_k = F$, without any dependence on k, then $F(x)$ or its probability density $f(x)$ determines the regions where COs would form. So for regions for which $f = 0$ such as the centromeric portion, no COs will be formed there. The frequency series $\{c_i\}$ is called the crossover count *C-distribution* and $\{F_k(x)\}$ is the conditional crossover location *L-distribution* (Karlin & Liberman 1978, Risch & Lange 1979). If $c_k = e^{-\mu}\mu^k/k!$, that is Poisson and $F_k(x)$ is a uniform distribution then the C-L process becomes the classical Haldane model (explained later). Here while the C-distribution introduces overall variation in recombination rates possibly mediated by environmental factors such as temperature, the L-distribution ensures more intricate control over meiotic CO locations.

(c) *Markov Growth CO Formation Process:* Under this framework, it is assumed that the CO events are formed according to a time-dependent sequence where ‘time’ simply relates to

the distance of a CO along the chromosome. Continuing with this thought, we can consider CO formation as a continuous Markov growth process, beginning from one end or from the centromere where time, $t = 0$. So if $N(t)$ gives the number of CO events in the duration $[0, t]$, then $\{N(t)\}$ for $t \geq 0$ becomes a Markov process with the following transition probabilities, which are also stationary:

$$P_{ij}(s) = Prob(N(t+s) = j | N(t) = i), \qquad i, j = 0,1,2, \ldots \qquad \ldots (1.10)$$

Also, being an increasing process, $P_{ij}(s) = 0$ if $j < i$. Hence the two factors impacting the CO count in an interval are the length of the interval and the cumulative number of COs which have already formed. The actual positions of previously formed COs do not influence the position of the CO being formed presently.

Map functions are usually not efficient at imitating crossover interference. Further, an SRP or a count-location process could result in a map function, though interference is expressed in contrasting ways in the two cases. Take for example, the simple map function $M(d) = d/(1 + 2d)$. This function can be given by a count-location model with $c_k = \frac{1}{2^k}$ for $k \geq 0$, length of the chromosome is defined to be ½ which means $M(d)$ exists only in the interval $[0,1/2]$. On the other hand, the corresponding SRP has the following inter-CO distribution, $4(1 + 2d)^{-3}$ and d can take values all along $[0, \infty)$.

An interesting observation is that the distribution of the distances between crossovers for an SRP can more often be estimated by a gamma distribution. Thus, map functions such as Haldane and Kosambi can be neared via the gamma model or the chi-square model. Finally the use of map functions and SRPs requires the underlying process to be uniform, that is, interference magnitude must not alter along the chromosome. Though this may not reflect reality, it would require a dataset with a large sample size and numerous markers to understand the true nature (predictably non-stationary) of the process under study. But various SRPs have been modified to accommodate the concept of interference which fit data across organisms as well.

Renewal Process Models: These models represent recombination as a point process and each point represents a recombination event across a bivalent – the structure formed when paired homologues are composed of two sister chromatids each. It is assumed that only non-sister chromatids interact leading to the formation of a crossover (NCI). And crossing over leads to the four chromatids being a mosaic of the parental genotypes. The point process occurring on a single chromatid can be derived from the four-strand bundle by independently deleting (or *thinning*) each point with probability ½ . A given chromatid has an equal opportunity to be or not to be involved in a crossover independent of the role played by any other chromatid. Also due to NCI, recombination probability increases as the marker interval size becomes larger with an upper limit of ½.

We consider these point processes to be stationary in constant multiples of genetic distance. This distance metric for a particular chromosomal stretch is defined as the average number of crossovers formed on a single chromatid in that interval.

(a) Poisson model (Haldane 1919): This is a simple point process which does not allow any interference (Figure 1.8). That is, crossovers occur in nearby positions independently and randomly leading to the value of coincidence C being 1 throughout the chromosome. Under this model, we let μ_x be the strength of the process at position x along the chromosome such that:

$$\mu_x = lim_{\delta x \to 0} \frac{P(at\ least\ 1\ point\ in\ (x,\ x+\delta x])}{\delta x} \quad \ldots (1.11)$$

We only consider cases where the above limit exists and μ_x can be integrated. So the probability of there being no crossovers at all in an interval $[a, b)$ is given by $\exp(-\int_a^b \mu_x\, dx)$, or $\exp(-\mu(b-a))$ when $\mu_x = \mu$ for all x. If there are m markers, then let their physical positions be $z_1, z_2, \ldots, z_m$ and the inter-marker genetic distances be $d_1, d_2, \ldots, d_{m-1}$. Using recombination data, we can estimate these distances as follows:

$$d_1 = \frac{1}{2}\int_{z_1}^{z_2} \mu_x dx,\ d_2 = \frac{1}{2}\int_{z_2}^{z_3} \mu_x dx, \ldots,\ d_{m-1} = \frac{1}{2}\int_{z_{m-1}}^{z_m} \mu_x dx \quad \ldots (1.12)$$

The additional factor ½ comes from the assumption that a given chromatid shows probability ½ of being involved in crossover formation. From here, it is difficult to conclude about heterogeneity in the Poisson process strength and also the actual physical positions of markers. To arrive at a more amicable scenario, let us scale the interval [0,1] to the distances between the first and last markers such that, $0 = z_1 < z_2 < \ldots < z_m = 1$, let $\mu = \int_0^1 \mu_x dx$ and $M(y) = \mu^{-1}\int_0^y \mu_x dx$, for all $y \in [0,1]$. This will give us a function $M: [0,1] \to [0,1]$ which is an increasing monotone and for each $w \in [0,1]$, there is a $y \in [0,1]$ satisfying, $w = M(y)$. Hence M is continuous on $[0,1]$ and can also be considered akin to a cumulative distribution function (cdf) on this interval on the real line.

Finally transforming all Poisson process points by M, will give a homogenous process with overall intensity μ. Now there are two processes defined on $[0,1]$, the homogenous $M(z_1),\ M(z_2), \ldots, M(z_m)$ and the heterogeneous $z_1,\ z_2, \ldots, z_m$ of which we proceed with the former as both give the same recombination data distribution. Thus the homogenous Poisson process $\{M(z)\}$ considers $\mu_x = \mu\ \forall\ x$.

To obtain the probability of an event g which encompasses two sub-events, either no crossovers in an interval i, i.e., $g_i = 0$ or at least one crossover in the interval which is

represented by $g_i = 1$. The joint probability for the recombination data for a whole chromosome is given by multiplying the probabilities for individual intervals (as *independent* events). Thus the probability of the event g is:

$$\prod_{i=1}^{m-1} \exp(-2d_i g_i)(1 - \exp(-2d_i))^{(1-g_i)} \quad \ldots (1.13)$$

where $d_i$is as defined above in (1.12) and it can also be observed that $\mu = 2\sum_{i=1}^{m-1} d_i$.

(b) Gamma model: The inter-crossover distance distribution of the homogeneous Poisson model is an exponential random variable which is also a gamma variable with its shape parameter as 1 (Figure 1.9). This process can be generalized by having a renewal process with its inter-arrivals governed by the gamma distribution. The SRP with gamma inter-arrivals can be formulated as follows. Let there be a position x, then the density of the distance of the first point to the right will be:

$$f_I(w) = \mu^{\gamma} w^{\gamma-1} e^{-\mu w}/\Gamma(w) \quad \ldots (1.14)$$

where I is for 'inter-arrival' and $\Gamma(w) = \int_0^{\infty} \mu^{\gamma} t^{\gamma-1} e^{-\mu t} dt$. Also the event whose density is given by (1.14) occurs independently of all events to the left of x as an expression of the *memory-less* property. Distribution of the next point to the left of x is also governed by (1.12), thus there is no *directionality*. If it is assumed that no event has taken place at x, then the density to the next event (left or right) is given by:

$$f_{1st}(w) = \frac{\mu}{\Gamma(\gamma+1)} \int_w^{\infty} \mu^{\gamma} t^{\gamma-1} e^{-\mu t} dt \quad \ldots (1.15)$$

The strength of the process now becomes μ/γ. When $\gamma = 1$, we obtain the Poisson model.

As before, let us scale the four chromatid bundle to the interval [0,1] and restrict the range of the inter-arrival distribution to $[0, 1]$ as well. Then the probability of having at least one crossover along the bundle of chromatids is, $\alpha = \int_0^1 f_{1st}(w)dw$. Again, if it is known that at least one crossover forms, the probability density of the distance from the end equated to 0 to the first crossover position on the bundle will be $\alpha^{-1} f_{1st}(w)$. Now if it is known that there is a crossover at position $x \in [0, 1]$, then the probability of the presence of another crossover between x and 1 is given by $\beta(x) = \int_0^{1-w} f_I(w)dw$. Similarly, in addition to a crossover at x, if it is also known that there is at least one crossover between x and 1, then the density of the distance from x to the next crossover (between x and 1) will be $\beta(x)^{-1} f_I(w)$. It must be noted here that though the

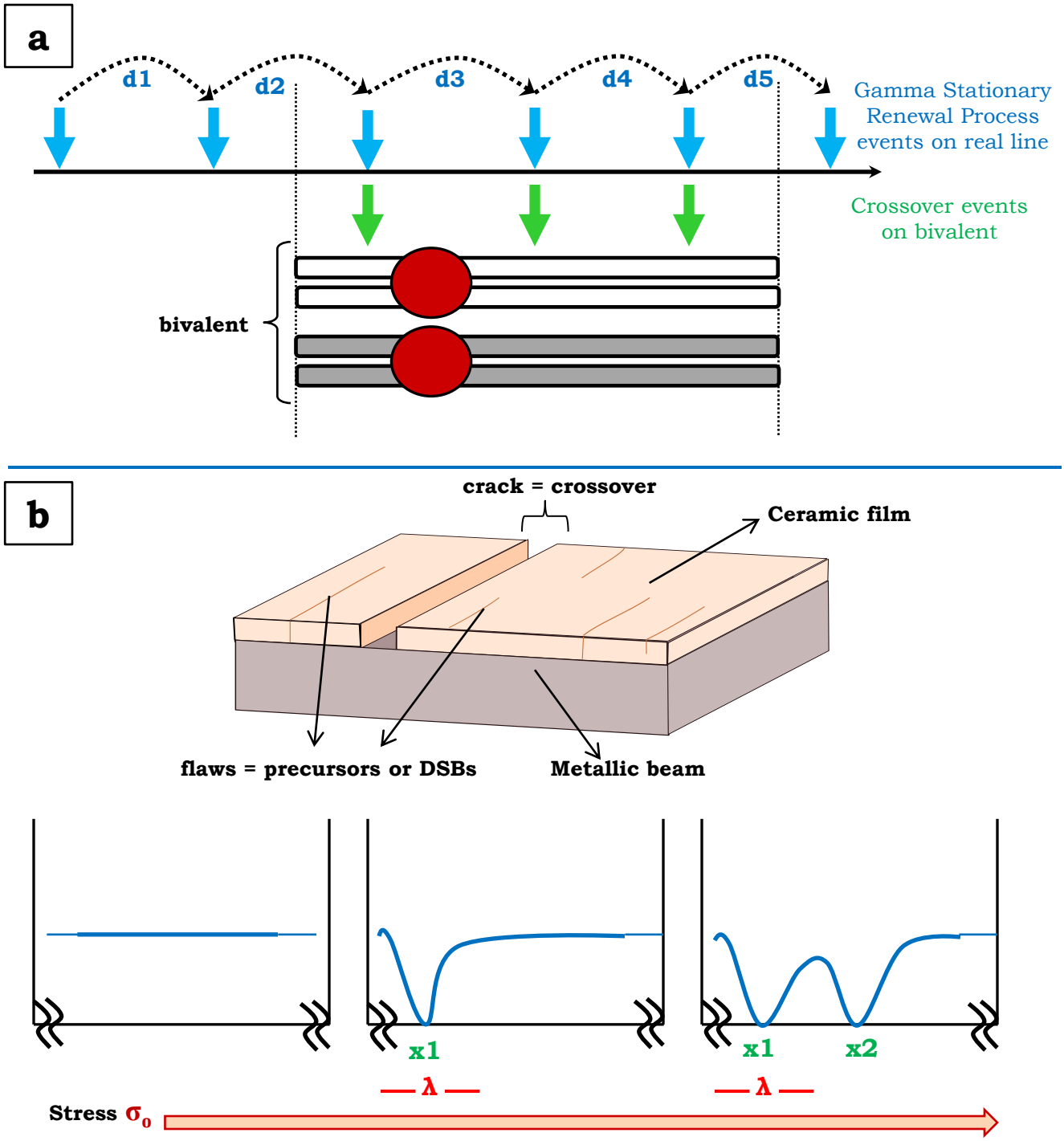


Figure 1.9 Interference Models.

a | Gamma Model

The crossover positions are obtained from a Stationary Renewal Process (SRP) with the inter-crossover distances from a gamma distribution. The rate (shape) of this distribution denotes the interference strength among the resulting crossovers. The gamma shape gives much lower variance than exponential giving interfering crossovers.

b | Mechanical Stress Model (Beam-film)

Formation of crossovers from DSBs is compared to flaws along a thin ceramic film coating a sizeable metallic beam developing into cracks as stress along the beam increases. With increase in stress when a flaw turns into a crack it releases stress on nearby flaws preventing their conversion into cracks, thus mediating interference.

relationship of the metric of these described distances with physical distance remain uncertain, this metric is indeed a constant multiple of the genetic distance.
Just as in the Poisson model, thinning is implemented where each crossover event on the four-strand bundle is equally probable to be kept or to be removed to arrive at the crossover events for a single chromatid.

Though here we do not restrict ourselves to the integral shape parameter case of gamma, the gamma SRP with its shape parameter value constrained only to integers has been implemented and examined since long. Fisher had suggested (Fisher *et al.* 1947) that the point process of crossovers observed along a chromatid can be equated to an inter-arrival renewal process with density $f(w) = \left(\frac{\pi}{2}\right) \operatorname{sech}\left(\frac{\pi w}{2}\right) \tanh\left(\frac{\pi w}{2}\right)$. Thereafter, it was found (Owen 1950) that the gamma density with 2 as the shape and rate parameters for the inter-arrivals closely resembled Fisher *et al*'s renewal process. The gamma SRP has been applied at the four-strand level with various models of chromatid interference (Carter and Robertson 1952), without chromatid interference (Cobbs 1978) and also at the two-strand level (Payne 1956). The gamma model at the four-strand level with no chromatid interference has also been fitted to data from *Drosophila* and *Neurospora* by way of comparison between the observed and expected crossover distribution (disregarding those that could not be observed due to experimental limitations) in a particular chromosomal segment (Cobbs 1978). Various mathematical aspects of the same model were also explored and fits were presented for data from the same organisms by contrasting the coincidence curves obtained from the experiment and from the model (Stam 1979, Foss *et al.* 1993).

Thus, this set-up with the inter-arrivals from a gamma distribution and only integer shapes has been utilized generously mainly since it is mathematically manageable. But it has also been projected to have the capability to elucidate specific experimental observations about the formation of gene conversions (non-reciprocal exchange between homologues) and recombination (Foss *et al.* 1993). To begin with, gene conversions have been seen to be accompanied by high recombination frequency in surrounding markers (Mortimer & Fogel 1974). In addition, gene conversions tend to appear independently in marker intervals whereas gene conversions with recombination do not. So a new perspective was presented by Foss *et al.* (1993). This entails that there be a Poisson process for the initial pre-crossover events which result in gene conversions. These conversions may or may not become crossovers. In this model, also known as a *counting* model, every $(m+1)th$ event matures into a crossover. If the first event is given the probability $1/(m+1)$ to lead to a crossover, then this model is nothing but a gamma SRP for the inter-arrivals with shape parameter as $m+1$.

Fitting a stationary gamma renewal process also results in permitting the associated real physical map to be non-stationary. As in the Poisson model, for convenience, we assume that a non-decreasing and *onto* (continuous) transformation function $M: [0, 1] \rightarrow [0, 1]$ (cdf on $[0, 1]$) exists which maps the model to a stationary process.

Mechanistic Methods

Under this we mainly have two approaches which have been proposed: the King-Mortimer model and the mechanical stress model. Here is a conceptual overview of these models:

(a) *King-Mortimer model:* This model is also known as the polymerization model (King & Mortimer 1990) as it suggests recombination travels as a polymerizing signal along the chromosome. What is proposed is as follows. Early recombination structures (such as DSBs or early recombination nodules) are distributed independently along the chromosomes and are granted constant probability per unit of time to initiate bi-directional polymerization. The polymer extends from its initiation site and is capable of preventing further addition of early structures into the bivalent. These sites where initiation begins are suggested to finally mature into crossovers. This model is remembered as it explains interference as well as the obligatory crossover rule. Also the foreseen crossover pattern is a result of interference which is the strongest when nearest to initiation sites and reduces in strength in a distance-dependent manner. Simulations of the associated parameters have been fit efficiently against data from Drosophila and *S. cerevisiae*.

One of the motivations behind this model was that an optimal interference model will be useful in systems with numerous magnitudes of differences in their genome sizes in base-pairs or *bp*. Also physical distances measured in terms of SC length rather than *bp* are closer among organisms, hence the signal is considered to be moving along the SC axis. This idea which entails that interference, though propagating via physical distance but is measured along SC length, and the interference signal disseminates along the SC axis is engaging as species with varied sizes of genome can be normalized by changing the axis-associated DNA loop sizes. But model-supporting experimental observations have not been made yet.

(b) *Beam-film (stress-based) model:* Among physical systems, it is commonly seen that a disturbance such as an increasing or decreasing stress originates at a point and dissipates outward from that point (Figure 1.9). Stress is usually generated along chromosomes as they contract and expand. And this model (Kleckner *et al.* 2004) advocates that crossovers form to release this stress locally. Stress relief in the neighborhood of a crossover travels down the chromosome in both directions down the SC axis.

Conceptual Background – When there is a thin film bound to a thick substrate (metallic beam here), it is able to support a tensile stress σ_o parallel to the film (Figure XX). This stress can build up when the film is deposited or it can occur later when temperature increases are imposed on the system. Subsequently thermal expansion differences as explained above mediate initiation and formation of crack from the flaws. In case the film is fragile and well-bound, the cracks which penetrate to the beam will travel across the film perpendicular to the stress.

This model has a very complex implementation as certain heterogeneities must be introduced in the mechanical properties of the film to lead to precursors (flaws) of varying fragility (explained later). As the beam is subjected to stress and expands as a result, each precursor can resist a critical amount of stress beyond which it "cracks" (forming a CO). In the limit that the film is thin, how the stress propagates along the film is given by a simple linear partial differential equation. Once a crack develops, the stress disappears at the crack while on its either side the stress in the film relaxes to its crack-independent value exponentially with distance. Let us denote the distance over which this relaxation happens as λ. Then the stress in an interval along the film without any cracks is given by $\sigma_o + ae^{x/\lambda} + be^{-x/\lambda}$. For additional cracks to form, they will more often be initiated beyond the stress-relieved regions near already-formed cracks. As stress increases (e.g. continued temperature rise), this physical process is led towards relatively uniformly spaced crack sites in the film. Thus the crack positions can be considered analogous to interfering CO positions on a bivalent.

In addition, it is necessary to specify the stress at the chromosome ends (Kleckner *et al.* 2004). The boundary conditions given by Neuman (Landau & Lifshitz 1986) could be utilized for this purpose which ensures that the stress at the film ends (bivalent) has zero derivative. The partial differential equation associated with the system can be solved analytically with the general solution of the form:

$$\sigma(x) = \sigma_o \left(1 - \cosh\left(\frac{x - x_o}{\lambda}\right)\right) + B \sinh\left(\frac{x - x_o}{\lambda}\right), \quad \ldots (1.16)$$

for the stress $\sigma(x)$ at the point x on one side of a crack at position x_o. A subtle point in the value of B specific for each interval, restricted between cracks or an edge of a chromosome. Each value of B must be adjusted for the boundary conditions to be satisfied and $\sigma(x) = 0$ at the crack. An explicit formula at any given point provided the crack positions are known, is derived from this piecewise solution. Lastly, it is important to adjust the model by setting the maximum possible stress in the beam to yield the correct mean number of COs (twice the genetic length of a chromosome in Morgans) while λ is chosen based on how interfering the CO positions must be. Very thin films correspond to a small value for λ and a very limited interference range.

In addition to the restrictions exercised on the final outcome, there is a need to set the *ratio* of the number of precursors (DSBs) to the final number of crossovers as well, which remains organism-dependent (varying from 4 to 25 in maize experiments {Franklin *et al.* 1999, Stack & Anderson 2002}) and 20 was used by our group for modeling maize data {Falque *et al.* 2009}; also, 17 in tomato studies {Stack & Anderson 1986}). These precursors are distributed randomly on the bivalent and a random threshold is also assigned to each based on the formula:

$$\theta_i = 1/\sqrt{x_i} \quad \ldots (1.17)$$

where, x_i represents a random variable distributed uniformly in the interval [0, 1]. When the stress at the precursor exceeds this threshold, it turns into a crack. There is an inherent stochasticity in the model which lends strong randomness in the determination of which precursor will lead to a crack and which won't. If the stress at a precursor does not reach the threshold they are considered to be conversions.

This model is appealing as it ensures crossover interference and homeostasis. Since each chromosome is under stress, the formation of the first crossover becomes indispensable (preserving the obligate CO rule). The signal of a CO being formed travelling along a chromosome acts as an inhibitory signal. Numerous events can happen on a chromosome but the tendency to be evenly spaced would be retained.

Though mathematical simulations of this model have been fit to experimental data, its predictions remain difficult to verify. Further, DSBs may relieve tensile stress but the role of crossovers in stress relief is elusive to gauge. Lastly it is also arduous to conjecture about a mechanism which allows some COs to release stress while some escape as in the case of organisms with both interfering and non-interfering crossover formation pathways.

References

Albini S.M., Jones G.H. (1987) Synaptonemal complex spreading in *Allium cepa* and *A. fistulosum*. I. The initiation and sequence of pairing. *Chromosoma* 95: 324-38.

Allers T., Lichten M. (2001) Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* 106: 47-57.

Anderson L.K., Stack S.M. (2005) Recombination nodules in plants. *Cytogenet. Genome Res.* 109: 198-204.

Armstrong L., Snyder J.A. (1989) Selective reduction of anaphase B in quinacrine-treated PtK1 cells. *Cell Motil. Cytoskeleton* 14: 220–29.

Auger D.L., Sheridan W.F. (2001) Negative crossover interference in maize translocation heterozygotes. *Genetics* 159: 1717-26.

Baskin T.I., Cande W.Z. (1990) The structure and function of the mitotic spindle in flowering plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 41: 277-315.

Bauer E., Falque M., Walter H. *et al.* (2013) Intra-specific variation of recombination rate in maize. *Genome Biol.* 14: R103.

Bell L.R., Byers B. (1983) Homologous association of chromosomal DNA during yeast meiosis. *Cold Spring Harb. Symp. Quant Biol.* 47: 829-40.

Berchowitz L.E., Copenhaver G.P. (2010) Genetic interference: don't stand so close to me. *Curr. Genomics* 11: 91-102.

Bhagat R., Manheim E.A., Sherizen D.E. *et al.* (2004) Studies on crossover-specific mutants and the distribution of crossing over in Drosophila females. *Cytogenet. Genome Res.* 107: 160-71.

Bishop D.K. (2006) Meiotic Recombination Pathways. *eLS.*

Bishop D.K., Zickler D. (2004) Early decision: meiotic crossover interference prior to stable strand exchange and synapsis. *Cell* 117: 9-15.

Blat Y., Protacio R.U., Hunter N. *et al.* (2002) Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell* 111: 791-802.

Borde V., Goldman A. S., Lichten M. (2000) Direct coupling between meiotic DNA replication and recombination initiation. *Science* 290: 806-09.

Borde V., Robine N., Lin W. *et al.* (2009) Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J.* 28: 99-111.

Börner G.V., Kleckner N., Hunter N. (2004) Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. *Cell* 117: 29-45.

Budd M. E., Wittrup K. D., Bailey J. E. *et al.* (1989) DNA polymerase I is required for premeiotic DNA replication and sporulation but not for X-ray repair in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 9: 365-76.

Burgess S.M. (2002) Homologous chromosome associations and nuclear order in meiotic and mitotically dividing cells of budding yeast. *Adv. Genet.* 46: 49-90.

Carballo J.A., Johnson A.L., Sedgwick S.G. *et al.* (2008) Phosphorylation of the axial element protein Hop1 by Mec1/Tel1 ensures meiotic interhomolog recombination. *Cell* 132: 758-70.

Carter T. C., Robertson A. (1952) A mathematical treatment of genetical recombination using a four-strand model. *Proc. R. Soc. Lond. Ser. B* 139: 410-26.

Cha R. S., Weiner B. M., Keeney S. *et al.* (2000). Progression of meiotic DNA replication is modulated by interchromosomal interaction proteins, negatively by Spo11p and positively by Rec8p. *Genes Dev.* 14: 493-503.

Chan A., Cande W.Z. (1998) Maize meiotic spindles assemble around chromatin and do not require paired chromosomes. *J. Cell Sci.* 111: 3507-15.

Cobbs G. (1978) Renewal process approach to the theory of genetic linkage: case of no chromatid interference. *Genetics* 89: 563-81.

Copenhaver G.P., Housworth E.A., Stahl F.W. (2002) Crossover interference in Arabidopsis. *Genetics* 160: 1631-39.

Cromie G.A., Hyppa R.W., Taylor A.F. *et al.* (2006) Single Holliday junctions are intermediates of meiotic recombination. *Cell* 127: 1167-78.

Daley D.J., Vere-Jones D. (1988) *An Introduction to the Theory of Point Processes.* Springer-Verlag, New York.

de Boer E., Stam P., Dietrich A.J. *et al*. (2006) Two levels of interference in mouse meiotic recombination. *Proc. Natl. Acad. Sci. USA* 103: 9607-12.

de los Santos T., Hunter N., Lee C. *et al.* (2003) The Mus81/Mms4 endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics* 164: 81-94.

de Massy B. (2003) Distribution of meiotic recombination sites. *Trends Genet.* 19: 514-22.

Dernburg A.F., McDonald K., Moulder G. *et al.* (1998) Meiotic recombination in *C. elegans* initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. *Cell* 94: 387-98.

Ehrhardt D.W., Shaw S.L. (2006) Microtubule dynamics and organization in the plant cortical array. *Annu. Rev. Plant Biol.* 57: 859-75.

Engebrecht J., Hirsch J., Roeder G.S. (1990) Meiotic gene conversion and crossing over: their relationship to each other and to chromosome synapsis and segregation. *Cell* 62: 927-37.

Falque M., Anderson L.K., Stack S.M. *et al.* (2009) Two types of meiotic crossovers coexist in maize. *Plant Cell* 21: 3915-25.

Fisher R.A. (1947) The theory of linkage in polysomic inheritance. *Phil. Trans. Roy. Soc. B.* 233: 55-87.

Fisher R.A., Lyon M.F., Owen A.R.G. (1947) The sex chromosomes in the house mouse. *Heredity* 1: 355-65.

Foss E., Lande R., Stahl F.W. *et al.* (1993) Chiasma interference as a function of genetic distance. *Genetics* 133: 681-91.

Franklin F.C., Higgins J.D., Sanchez-Moran E. *et al.* (2006). Control of meiotic recombination in Arabidopsis: role of the MutL and MutS homologues. *Biochem. Soc. Trans.* 34: 542-44.

Franklin A.E., McElver J., Sunjevaric I. *et al*. (1999) Three-dimensional microscopy of the Rad51 recombination protein during meiotic prophase. *Plant Cell* 11: 809-24.

Fung J.C., Rockmill B., Odell M. *et al.* (2004) Imposition of crossover interference through the nonrandom distribution of synapsis initiation complexes. *Cell* 116: 795-802.

Gilbertson L.A., Stahl F.W. (1996) A test of the double-strand break repair model for meiotic recombination in *Saccharomyces cerevisiae*. *Genetics* 144: 27-41.

Giraut L., Falque M., Drouaud J. *et al.* (2011) Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet.* 7: e1002354.

Grelon M., Vezon D., Gendrot G. *et al.* (2001) AtSPO11-1 is necessary for efficient meiotic recombination in plants. *EMBO J.* 20: 589-600.

Guillon H., Baudat F., Grey C. *et al.* (2005) Crossover and noncrossover pathways in mouse meiosis. *Mol. Cell* 20: 563-73.

Guillon H., de Massy B. (2002) An initiation site for meiotic crossing over and gene conversion in the mouse. *Nat. Genet.* 32: 296-99.

Haldane J.B.S. (1918) Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* 7: 106-09.

Haldane J.B.S. (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8: 299-309.

Heald R., Tournebize R., Blank T. *et al.* (1996) Self-organization of microtubules into bipolar spindles around artificial chromosomes in *Xenopus* egg extracts. *Nature* 382: 420-25.

Henderson K.A., Keeney S. (2004) Tying synaptonemal complex initiation to the formation and programmed repair of DNA doublestrand breaks. *Proc. Natl. Acad. Sci. USA* 101: 4519-24.

Henderson K.A., Keeney S. (2005) Synaptonemal complex formation: where does it start? *Bioessays* 27: 995-98.

Hillers K.J., Villeneuve A.M. (2003) Chromosome-wide control of meiotic crossing over in *C. elegans*. *Curr. Biol.* 13: 1641-47.

Higgins J.D., Armstrong S.J., Franklin F.C. *et al.* (2004) The Arabidopsis MutS homolog AtMSH4 functions at an early step in recombination: evidence for two classes of recombination in Arabidopsis. *Genes Dev.* 18: 2557-70.

Holliday R. (1964) A mechanism for gene conversion in fungi. *Genet. Res.* 5: 282-304.

Housworth E.A., Stahl F.W. (2003) Crossover interference in humans. *Am. J. Hum. Genet.* 73: 188-97.

Hunter N., Kleckner N. (2001) The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. *Cell* 106: 59-70.

Hyams J. (1996) Look Ma, no chromosomes! *Nature* 382: 397-98.

Jeffreys A.J., May C.A. (2004) Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* 36: 151-56.

Job D., Valiron O., Oakley B. (2003) Microtubule nucleation. *Curr. Opin. Cell Biol.* 15: 111-7.

Karlin S., Liberman U. (1994) Theoretical recombination processes incorporating interference effects. *Theor. Popul. Biol.* 46: 198-231.

Keeney S., Giroux C.N., Kleckner N. (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* 88: 375-84.

King J.S., Mortimer R.K. (1990) A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* 126: 1127-38.

Kirschner M., Mitchison T. (1986) Beyond self-assembly: from microtubules to morphogenesis. *Cell* 45: 329-42.

Kleckner N. (1996) Meiosis: how could it work? *Proc. Natl. Acad. Sci. USA* 93: 8167-74.

Kleckner N., Weiner B. (1993) Potential advantages of unstable interactions for pairing of chromosomes in meiotic, somatic, and premeiotic cells. *Cold Spring Harb. Symp. Quant Biol.* 58: 553-65.

Kleckner N., Zickler D., Jones G.H. *et al.*(2004) A mechanical basis for chromosome function. *Proc. Natl. Acad. Sci. USA* 101:12592-97.

Kosambi D.D. (1944) The estimation of the map distance from recombination values. *Ann. Eugen.* 12: 172-75.

Lamb N.E., Sherman S.L., Hassold T.J. (2005) Effect of meiotic recombination on the production of aneuploid gametes in humans. *Cytogenet. Genome Res.* 111: 250-55.

Landau L.D., Lifshitz E.M. (1986) *Course of Theoretical Physics Volume 7: Theory of Elasticity*. Pergamon Press, New York.

Lenormand T., Dutheil J. (2005) Recombination difference between sexes: a role for haploid selection. *PLoS Biol.* 3: e63.

Lhuissier F.G.P., Offenberg H.H., Wittich P.E. *et al.* (2007) The mismatch repair protein MLH1 marks a subset of strongly interfering crossovers in tomato. *Plant Cell* 19: 862-876.

Liberman U., Karlin S. (1984) Theoretical models of genetic map functions. *Theor. Popul. Biol.* 25: 331-46.

Liu L., Parekh-Olmedo H., Kmiec E.B. (2003) The development and regulation of gene repair. *Nat. Rev. Genet.* 4: 679–89.

MacQueen A.J., Colaiacovo M.P., McDonald K. *et al.* (2002) Synapsis-dependent and -independent mechanisms stabilize homolog pairing during meiotic prophase in *C. elegans. Genes Dev.* 16: 2428-42.

Mahadevaiah S.K., Turner J.M.A., Baudat F. *et al.* (2001) Recombinational DNA double-strand breaks in mice precede synapsis. *Nat. Genet.* 27: 271-76.

Mancera E., Bourgon R., Brozzi A. *et al.* (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454: 479-85.

Martinez-Perez E., Colaiacovo M.P. (2009) Distribution of meiostic recombination events: talking to your neighbors. *Curr. Opin. Genet. Dev.* 19: 105-12.

Martini E., Diaz R.L., Hunter N. *et al.* (2006) Crossover homeostasis in yeast meiosis. *Cell* 126: 285-95.

Mastronarde D.N., McDonald K. L., Ding R. *et al.* (1993) Interpolar spindle microtubules in PTK cells. *J. Cell Biol.* 123: 1475-89.

Mather K. (1936) The determination of position in crossing-over. *J. Genet.* 33: 207-35.

McKim K.S., Green-Marroquin B.L., Sekelsky J.J. *et al.* (1998) Meiotic synapsis in the absence of recombination. *Science* 279:876-78.

McMahill M.S., Sham C.W., Bishop D.K. (2007) Synthesis-dependent strand annealing in meiosis. *PLoS Biol.* 5: e299.

McPeek M.S., Speed T.P. (1995) Modeling interference in genetic recombination. *Genetics* 139: 1031-44.

McPeek M.S., Speed T.P. (1995) Modeling interference in genetic recombination. *Genetics* 139: 1031-44.

Meneely P.M., Farago A.F., Kauffman T.M. (2002) Crossover distribution and high interference for both the X chromosome and an autosome during oogenesis and spermatogenesis in *Caenorhabditis elegans*. *Genetics* 162: 1169-77.

Mercier R., Jolivet S., Vezon D. *et al.* (2005) Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on MER3, whereas the other one is not. *Curr. Biol.* 15: 692-701.

Moens P.B. (1976) Spindle and kinetochore morphology of *Dictyostelium discoideum*. *J. Cell Biol.* 68: 113-22.

Morgan T.H., Kidges C.B., Schultz J. (1935) Constitution of the germinal material in relation to heredity. *Carnegie Inst. Wash. Yearbook* 34: 284-91.

Mortimer R.K., Fogel S. (1974) *Mechanisms in Recombination.* Plenum, New York.

Munz P. (1994) An analysis of interference in the fission yeast *Schizosaccharomyces pombe*. *Genetics* 137: 701-07.

Murray A.W., Szostak J.W. (1985) Chromosome segregation in mitosis and meiosis. *Annu. Rev. Cell Biol.* 1: 289-315.

Novak J.E., Ross-Macdonald P.B., Roeder G.S. (2001) The budding yeast Msh4 protein functions in chromosome synapsis and the regulation of crossover distribution. *Genetics* 158: 1013-25.

Osman F., Dixon J., Doe C.L. *et al.* (2003) Generating crossovers by resolution of nicked Holliday junctions: a role for Mus81-Eme1 in meiosis. *Mol. Cell* 12: 761-74.

Owen A.R.G. (1950) The theory of genetical recombination. *Adv. Genet.* 3: 117-57.

Padmore R., Cao L., Kleckner N. (1991) Temporal comparison of recombination and synaptonemal complex formation during meiosis in *S. cerevisiae*. *Cell* 66: 1239-56.

Page S.L., Hawley R.S. (2003) Chromosome choreography: the meiotic ballet. *Science* 301: 785-89.

Palevitz B.A. (1993) Morphological plasticity of the mitotic apparatus in plants and its developmental consequences. *Plant Cell* 5: 1001-09.

Paradez A., Wright A., Ehrhardt D.W. (2006) Microtubule cortical array organization and plant cell morphogenesis. *Curr. Opin. Plant Biol.* 9: 571-78.

Payne L. C. (1956) The theory of genetical recombination: a general formulation for a certain class of intercept length distributions appropriate to the discussion of multiple linkage. *Proc. R. Soc. Lond. Ser. B* 144: 528-44.

Peoples T.L., Dean E., Gonzalez O. *et al.* (2002) Close, stable homolog juxtaposition during meiosis in budding yeast is dependent on meiotic recombination, occurs independently of synapsis, and is distinct from DSB-independent pairing contacts. *Genes Dev.* 16:1682-95.

Pereira G., Schiebel E. (2001) The role of the yeast spindle pole body and the mammalian centrosome in regulating late mitotic events. *Curr. Opin. Cell Biol.* 13: 762-69.

Petes T.D. (2001) Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* 2: 360-69.

Porter S.E., White M.A., Petes T.D. (1993) Genetic evidence that the meiotic recombination hotspot at the *HIS4* locus of *Saccharomyces cerevisiae* does not represent a site for a symmetrically processed double-strand break. *Genetics* 134: 5-19.

Qin J., Richardson L.L., Jasin M. *et al.* (2004) Mouse strains with an active H2-Ea meiotic recombination hot spot exhibit increased levels of H2-Ea-specific DNA breaks in testicular germ cells. *Mol. Cell. Biol.* 24: 1655-66.

Risch N., Lange K. (1979) An alternative model of recombination and interference. *Ann. Hum. Genet.* 43: 61-70.

Risch N., Lange K. (1983) Statistical analysis of multiple recombination. *Biometrics* 39: 949-63.

Rockmill B., Fung J.C., Branda S.S. *et al.* (2003) The Sgs1 helicase regulates chromosome synapsis and meiotic crossing over. *Curr. Biol.* 13: 1954-62.

Rockmill B., Sym M., Scherthan H. *et al.* (1995) Roles for two RecA homologs in promoting chromosome synapsis. *Genes Dev.* 9: 2684-95.

Roeder G. S., Bailis J. M. (2000) The pachytene checkpoint. *Trends Genet.* 16: 395-403.

Ross K. J., Fransz P., Jones G. H. (1996) A light microscopic atlas of meiosis. *Chrom. Res.* 4: 507-16.

Scherthan H., Bahler J., Kohli J. (1994) Dynamics of chromosome organization and pairing during meiotic prophase in fission yeast. *J. Cell Biol.* 127: 273-85.

Schild D., Byers B. (1978) Meiotic eVects of DNA-defective cell division cycle mutations of *Saccharomyces cerevisiae*. *Chromosoma* 70: 109-30.

Schwacha A., Kleckner N. (1995) Identification of double Holliday junctions as intermediates in meiotic recombination. *Cell* 83:783-91.

Sekelsky J.J., Brodsky M.H., Burtis K.C. (2000) DNA repair in Drosophila: insights from the Drosophila genome sequence. *J. Cell Biol.* 150: F31-F36.

Shelden E., Wadsworth P. (1990) Interzonal microtubules are dynamic during spindle elongation. *J. Cell Sci.* 97: 273-81.

Simchen G., Idar D., Kassir Y. (1976) Recombination and hydroxyurea inhibition of DNA synthesis in yeast meiosis. *Mol. Gen. Genet.* 144: 21-27.

Smith G.R., Boddy M.N., Shanahan P. *et al.* (2003) Fission yeast Mus81–Eme1 holliday junction resolvase is required for meiotic crossing over but not for gene conversion. *Genetics* 165: 2289-93.

Smith K. N., Penkner A., Ohta K. *et al.* (2001) B-type cyclins CLB5 and CLB6 control the initiation of recombination and synaptonemal complex formation in yeast meiosis. *Curr. Biol.* 11: 88-97.

Speed T.P. (1995) What is a genetic map function? in *Genetic Mapping and DNA Sequencing.* Springer-Verlag, New York.

Stack S.M., Anderson L.K. (1986) Two-dimensional spreads of synaptonemal complexes from solanaceous plants. *Chromosoma* 94: 253-58.

Stack S.M., Anderson L.K. (2002) Crossing over as assessed by late recombination nodules is related to the pattern of synapsis and the distribution of early recombination nodules in maize. *Chromosome Res.* 10: 329-45.

Stam P. (I979) Interference in genetic crossing over and chromosome mapping. *Genetics* 92: 573-94.

Strich R. (2004) Meiotic DNA replication. *Curr. Top. Dev. Biol.* 61: 29-60.

Stuart D., Wittenberg C. (1998) CLB5 and CLB6 are required for premeiotic DNA replication and activation of the meiotic S/M checkpoint. *Genes Dev.* 12: 2698-710.

Sturtevant A.H. (1913) A third group of linked genes in Drosophila Ampelophila. *Science* 37: 990-92.

Sturtevant A.H. (1915) The behavior of chromosomes as studied through linkage. *Z. Induct. Abstammungs-Vererbungsl.* 13: 234-87.

Szostak J.W., Orr-Weaver T.L., Rothstein R.J. *et al.* (1983) The double-strand-break repair model for recombination. *Cell* 33: 25-35.

Talbert P.B., Henikoff S. (2010) Centromeres Convert but Don't Cross. *PLoS Biol.* 8: e1000326.

Terasawa M., Ogawa H., Tsukamoto Y. *et al.* (2007) Meiotic recombination-related DNA synthesis and its implications for cross-over and non-crossover recombinant formation. *Proc. Natl. Acad. Sci. USA* 104: 5965-70.

Tesse S., Storlazzi A., Kleckner N. *et al.* (2003) Localization and roles of Ski8p protein in *Sordaria* meiosis and delineation of three mechanistically distinct steps of meiotic homolog juxtaposition. *Proc. Natl. Acad. Sci. USA* 100: 12865-70.

Theurkauf W.E., Hawley R.S. (1992) Meiotic spindle assembly in *Drosophila* females: behavior of nonexchange chromosomes and the effects of mutations in the *nod* kinesin-like protein. *J. Cell Biol.* 116: 1167-80.

Tippit D.H., Schultz D., Pickett-Heaps J.D. (1978) Analysis of the distribution of spindle microtubules in the diatom *Fragilaria*. *J. Cell Biol.* 79: 737-63.

Tsai C.J., Mets D.G., Albrecht M.R. *et al.* (2008) Meiotic crossover number and distribution are regulated by a dosage compensation protein that resembles a condensin subunit. *Genes Dev.* 22: 194-211.

Vazquez J., Belmont A.S., Sedat J.W. (2002) The dynamics of homologous chromosome pairing during male *Drosophila* meiosis. *Curr. Biol.* 12: 1473-83.

Verno I., Karsenti E. (1995) Chromosomes take the lead in spindle assembly. *Trends Cell Biol.* 5: 297-301.

von Wettstein D., Rasmussen S.W., Holm P.B. (1984) The synaptonemal complex in genetic segregation. *Annu. Rev. Genet.* 18: 331-413.

Wang C.R., Carlton P.M., Golubovskaya I.N. *et al.* (2009) Interlock formation and coiling of meiotic chromosome axes during synapsis. *Genetics* 183: 905-15.

Weiner B., Kleckner N. (1994) Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. *Cell* 77: 977-91.

Whitby M.C. (2005) Making crossovers during meiosis. *Biochem. Soc. Trans.* 33(6): 1451-55.

Whitehouse H.L.K. (1982) *Genetic Recombination: Understanding the Mechanism.* John Wiley, New York.

Williamson D. H., Johnston L. H., Fennell, D. J. *et al.* (1983). The timing of the S phase and other nuclear events in yeast meiosis. *Exp. Cell Res.* 145: 209-17.

Winey M., Mamay C.L., O'Toole E.T. *et al.* (1995) Three-dimensional ultrastructural analysis of the *Saccharomyces cerevisiae* mitotic spindle. *J. Cell Biol.* 129: 1601-15.

Yang M., Ma H. (2001) Male Meiotic Spindle Lengths in Normal and Mutant Arabidopsis Cells. *Plant Physiol.* 126: 622-630.

Zhao H., McPeek M.S., Speed T.P. (1995a) Statistical analysis of chromatid interference. *Genetics* 139: 1057-65.

Zhao H., Speed T.P. (1996) On genetic map functions. *Genetics* 142: 1369-77.

Zhao H., Speed T.P., McPeek M.S. (1995b) Statistical analysis of crossover interference using the chi-square model. *Genetics* 139: 1045-56.

Zickler D., Kleckner N. (1998) The leptotene-zygotene transition in meiosis. *Annu. Rev. Genet.* 32: 619-97.

Chapter 2

Readily available…

MODELS OF CROSSOVER INTERFERENCE

Here I discuss the theory which was used frequently throughout my thesis. Apart from the interference models, this chapter covers statistical techniques related to model or model parameter optimization.

Single Pathway Interference models

Gamma model (McPeek & Speed 1995)

As presented before (in the last section of the previous chapter) the so-called Gamma model is a framework which considers the inter-crossover distances as being generated from a stationary renewal process (SRP) (Figure 9). We begin by launching the renewal process from an initial position a long distance to the left of the chromosome to ensure that once we reach the chromosome domain, the CO events there become independent of initial position. This is to take advantage of the memorylessness of the SRP.

Continuing with previous notations, the gamma probability distribution function has shape parameter γ and rate as . This defines in the Gamma model the distribution of inter-crossover distances $\{d_i\}$:

$$f(\{d_i\},\gamma,\mu) = \mu^{\gamma} d^{\gamma-1} e^{-\mu d}/\Gamma(d) \quad \ldots (2.1a)$$

Statistically it is always in the interest of modelers to make a model more parsimonious. In the same spirit, the Gamma model rate parameter is set so the density of COs is 2 per Morgan so it is twice the shape which is usually denoted as ν or *nu*. Historically, it has been found that the Gamma model provides reasonably good fits to experimental data, though such early data had rather limited statistics. Thus the modified distribution form from (2.1a) we will be using in further analysis is given as:

$$f(\{d_i\},\nu,2\nu) = \frac{(2\nu)^{\nu} d^{\nu-1} e^{-2\nu d}}{\Gamma(d)} = f(\{d_i\},\nu) \quad \ldots (2.1b)$$

This model quantifies crossover interference by the parameter, ν (hereafter referred to as *nu*), the interference strength or intensity.

The model is mathematically very elegant, and in particular allows for an easy likelihood for any realization of the CO positions. This allows one to use the maximum likelihood framework for inferring the "optimum" parameter values. Further, the Gamma model is formulated in the space of genetic positions and in practice one may scale all distances to lie in the interval $[0,1]$ since the distribution of crossovers in this space is uniform. Note that by definition, a distance of one *Morgan* (the genetic distance unit) is defined as an interval for which the mean number of COs per meiosis is 1 at the level of gametes (and thus 2 at the level of bivalents). Thus each additional Morgan indicates that the average number of crossovers increased by one.

Beam-film model (Kleckner *et al.* 2004)

This model hypothesizes that the formation of crossovers can be explained via a framework to quantify mechanistic effects which go as described (refer also detailed description in the last section of the previous chapter) (Figure 9). It considers an (e.g., metallic) elastic beam or plate which is laminated tightly on one side by a thin fragile film (for instance, porcelain) with a considerable number of flaws lining its edges. When this ensemble is heated, the beam will expand more than the film owing to its higher thermal expansion coefficient. As the metal expands, it forces the film to stretch along with it, leading to a *tensile stress* on the film. The film resists to a certain extent, creating an opposing compressional force in the metallic beam. As the film is fragile compared to the beam, the tensile stress will lead to the initiation of crack formation at the flaws in the film. Once begun, a crack extends down to the film-beam interface and then travels across the whole, reaching the other edge and completing the crack. When a crack forms, it leads to local stress release on both sides of the crack site. The elasticity of the beam (in practice the value of the Young modulus) introduces a characteristic scale for stress relief; beyond that distance, the influence of crack on the stress on the film decays exponentially. This characteristic distance is put in correspondence with an interference parameter λ. This "relief" of stress spreads outwards from a crack in both directions and decreases steadily with distance from the crack. Newer cracks may subsequently develop but outside the stress relief region. Development of a crack hence *interferes* with the formation of nearby cracks.

The likelihood for this model has not been ever determined and it is unlikely a closed form can be obtained. In such situations, it is possible to use an approach based on summary statistics: a score is introduced which is used to quantify a goodness of fit and then one has to maximize this score as if it were a likelihood (details under Parameter Estimation).

Incorporating Two Pathways: interfering & non-interfering

Since it is known that the interfering and non-interfering pathways co-exist in most organisms with their respective contributions being in general unknown, it is important for models to have a way to achieve incorporating these two pathways. Thus in the following we cover two kinds of such models.

A two-pathway model includes an interfering (the first pathway or $P1$) as well as non-interfering pathway (denoted as the second pathway or $P2$) of crossover formation. An additional parameter p gives the proportion of the non-interfering pathway (thus also called $pnip$). If the first pathway is taken to be described by the Gamma model, the rate parameter of the gamma distribution for COs from the first pathway is then $2 * nu * (1 - p)$ while the Poisson density rate for the non-interfering fraction of crossovers is $2 * nu * p$. The pathway 2 COs can be thought of as superposed (independently) onto those of the first pathway, which is why the non-interfering COs are said to be "sprinkled" (Copenhaver *et al.* 2002) on top of the interfering crossovers. Such a sprinkling of non-interfering COs on top of interfering COs does not care what the first pathway is, so it can be implemented also when the interfering COs are produced by the beam-film model.

Parameter Estimation

Likelihood Maximization (Rothenburg & Leenders 1964, Theil 1971): Under this approach it is assumed that the data has been produced via a procedure involving unknown parameters and that these are to be inferred from the data. In our case, the parameters are for instance those of the family of gamma distributions. Under this assumption, we can compute the likelihood function of the parameters by applying Bayes' theorem and using the probability of the measured data for given values of the parameter(s). The maximization of the likelihood (we seek the global maximum) provides the estimates of the parameter values. Usually we work with the log likelihood as it is more convenient to handle numerically. In the case of a single pathway Gamma model, the inter-crossover distances come from a gamma SRP but one has to deal with the fact that the chromosome is finite whereas the SRP is defined on the infinite line. The densities with which the SRP enters (first crossover position) and then leaves (the last crossover) the chromosome have to be treated with some care (details in the Supplementary Materials of Basu-Roy *et al.* 2013). Hence a systematic algorithm to navigate through the parametric space is required. This need is heightened when we consider two-pathway models where two parameters must be estimated.

There are two more issues that need to be kept in mind: the matter of *thinning* (going from the four chromatids or bivalent to a single chromatid or gamete) and considering all possible ways of allocating experimentally observed crossovers to the two pathways when the non-interfering pathway is also incorporated. *Thinning* has been discussed before. But for the latter, what is done in practice is if a gamete (or chromosome) has k crossovers, then each of them can come from either the interfering or the non-interfering pathway, giving 2^k possibilities. And to compute the likelihood of such a collection of COs, the probability for each possibility needs to be calculated and then summed.

Score maximization (Falque *et al.* 2009): In models such as the Gamma model constructed from SRP, the likelihood can generally be calculated, though as is shown in Supplementary Materials of Basu-Roy *et al.* 2013, the calculation can become quite complex if there is missing data. However in the case of the beam-film model, there is no known procedure for computing the likelihood of data. In its absence, one can define a score which is based on summary statistics deduced from comparing simulated (at least 10^6 simulations) and observed histograms of inter-crossover distances. Let $P(0), P(1), P(2), ...$ be the model probabilities of the formation of $0, 1, 2, ...$ crossovers on a bivalent and ρ_k be the inter-crossover distance density on a bivalent with exactly k crossovers. Then the (score proposed in Falque et al. 2009) is given by:

$$PLS = P(k) \sum_{i=1}^{k-1} \rho_k(\Delta_i) \quad \ldots (2.2)$$

where, Δ_i is the i^{th} inter-crossover length between the i^{th} and $(i+1)^{th}$ crossover.

Optimization algorithms

The maximization of the likelihood or of a score requires searching through the parametric space. Some methods are more efficient than others, and we have had to adapt when necessary our implementations. Nevertheless, if details are omitted, here are the two major classes of searching in the parameter space.

Two-dimensional Scan: When using this algorithm, we replace all possible values for a parameter if there is only one parameter by the values on a mesh, defined using a parameter "step size" which is defined by the user. The search on the whole space is replaced by the scan throughout the whole mesh In case there are two parameters as in a two-pathway model, one moves along a two-dimensional mesh based on the two "step sizes" given as input.

The precision reached by such an approach depends on the step size input, but decreasing this step size increases the amount of computation. Clearly it would be appropriate to reduce the step size but force the search to stay localized to the best point found. This is in effect close in spirit to the hill climbing approach.

Hill-climbing: Here, beginning with an initial value with likelihood L_0, several neighbors are considered at a time, to the left, to the right and a trial point via polynomial interpolation, which are decided based on the initial step size. This initial step size depends on the range provided for each parameter as well as on the upper and lower limits for the respective range. The initial step size is initially relatively large because one may be far away from the optimum. So the likelihood for all the neighbors are compared against the present likelihood L_0 and the move is decided. When a better neighbor is not found, the step size is revised to a smaller value as the absence of a better neighbor usually indicates close proximity to the optimum hence the need to make the search more thorough and the moves more cautious. As the procedure is iterated, the trajectory

converges to a local maximum. Experience shows whether in general there is a single such optimum parameter point; interestingly, as long as there are not too many parameters, this seems to be the case for the models we have studied in this thesis.

When there is one parameter, it is very straight forward to check all neighbors and keep moving till the optimum is reached. In case of two parameters, changes are made to parameters in an alternate manner, checking for neighbors with higher likelihoods. This can be done one parameter at a time or not, with generally no consequence on the final result.

Confidence intervals

The one parameter case: In our work, we will generally use confidence intervals derived from the Fisher information matrix (Rao 1973, Theil 1971). Fisher information, denoted as $I_n(\theta)$ represents the amount of information that an observed random variable X_i contains on the unknown parameter θ in case of the single parameter model. The partial derivative of the natural logarithm of the likelihood $w.r.t.\theta$ is known as the expected score. The first moment of the score is 0 under some regularity conditions (Theil 1971):

$$E[\frac{\delta}{\delta\theta}\{ln\prod_i f(X_i,\theta)|\theta\}] = \sum_i \left[E\left\{\frac{\delta}{\delta\theta}\ln f(X_i,\theta)\,\middle|\,\theta\right\}\right] = \int \frac{\frac{\delta}{\delta\theta}f(X_i,\theta)}{f(X_i,\theta)}\, f(X_i,\theta)\, dX_i$$

$$= \int \frac{\delta}{\delta\theta} f(X_i,\theta)\, dX_i = \frac{\delta}{\delta\theta}\int f(X_i,\theta)\, dX_i = \frac{\delta}{\delta\theta}\, 1 = 0 \qquad \ldots (2.3)$$

If we consider the notation, $\lambda_n(\boldsymbol{X}|\theta) = ln\prod_i f(X_i,\theta)|\theta$

the Fisher information is given by the second moment of the score, and this will be as follows when $\lambda_n(\boldsymbol{X}|\theta)$ is doubly differentiable,

$$I_n(\theta) = E\{\frac{\delta}{\delta\theta}\lambda_n(\boldsymbol{X}|\theta)\}^2 = -E(\lambda_n''(\boldsymbol{X}|\theta)) \qquad \ldots (2.4)$$

$I_n(\theta)$ is in fact the *expected* Fisher information for the whole sample $\{X_i\}$ and $I(\theta)$ denotes the information corresponding to one random variable, X_i. Further, the *observed* Fisher information is:

$$J_n(\theta) = -\lambda_n''(\boldsymbol{X}|\theta) \qquad \ldots (2.5)$$

By using finite differences and limits, we have:

$\lambda_n'(\boldsymbol{X}|\theta) \approx \epsilon^{-1}\{\lambda_n(\boldsymbol{X}|\theta+\epsilon) - \lambda_n(\boldsymbol{X}|\theta+\epsilon)\}$, for small ϵ … (2.6a)

$J_n(\theta) \approx -\epsilon^{-1}\{\lambda_n'(\boldsymbol{X}|\theta+\epsilon) - \lambda_n'(\boldsymbol{X}|\theta)\} \approx -\epsilon^{-2}\{\lambda_n(\boldsymbol{X}|\theta+\epsilon) - 2\lambda_n(\boldsymbol{X}|\theta) + \lambda_n(\boldsymbol{X}|\theta-\epsilon)\}$,
for small ϵ … (2.6b)

The only condition for (2.5) to be computable is the existence of a smooth likelihood. The observed Fisher information can also be represented as:

$$J_n(\theta) = -\sum_{i=1}^{n} \frac{\delta^2}{\delta\theta^2} \ln f(X_i|\theta) \quad \ldots (2.6c)$$

Since $\{X_i\}$ are considered to be independent and identically distributed, the summation terms also have the same property. So by the law of large numbers, we have that their average converges to the expectation of one term, which is, $I(\theta)$. In addition since the observed and expected Fisher information are related by the theory of consistent estimation, we have:

$$\frac{1}{n} J_n(\theta) \to I(\theta) \text{ in probability,} \quad \ldots (2.7)$$

We may also write: $J_n(\theta) \approx n\, I(\theta) = I_n(\theta)$. Hence for large samples, $J_n(\theta)$ and $I_n(\theta)$ approximate each other. Since in our case we utilize the maximum likelihood estimate of the parameter, denoted as $\hat{\theta}$, we can also say that, $J_n(\hat{\theta}) \approx I_n(\hat{\theta})$. Thus in the case of large samples, we obtain that, $\hat{\theta} \sim Normal(\theta, I_n(\hat{\theta})^{-1})$ using the expected Fisher information. Or, in terms of the observed Fisher information, we can also write that, $\hat{\theta} \sim Normal(\theta, J_n(\hat{\theta})^{-1})$. Hence the Fisher confidence intervals can be expressed as:

$$\hat{\theta} \pm c I_n(\hat{\theta})^{-1/2} \quad \ldots (2.8)$$

where, c denotes the appropriate z-value corresponding to the confidence limits required (e.g., 1.96 for 95%). Finally the confidence intervals may also be: $\hat{\theta} \pm c J_n(\hat{\theta})^{-1/2}$. This expression is more realistic since we know the actual value of the observed Fisher information, not the expected one.

The several parameter case: Here the parameter is a vector $\boldsymbol{\theta_{k,1}}$, a column vector with the k-parameters. The logic of the computation of the confidence intervals remains similar apart from the fact that now all concepts are extended to the multi-variable case. So the first partial derivative is a vector of partial derivatives instead of one first derivative written as:

$$\nabla\lambda_n(\boldsymbol{X}|\boldsymbol{\theta}) = \{\frac{\boldsymbol{\delta}}{\boldsymbol{\delta\theta_i}} \boldsymbol{\lambda_n}(\boldsymbol{X}|\boldsymbol{\theta})\}_{\boldsymbol{k,1}},\ i = 1(1)k \quad \ldots (2.9)$$

Again, the double derivative will be a matrix now instead, denoted as $\nabla^2\lambda_n(\boldsymbol{X}|\boldsymbol{\theta})$. While the diagonal terms will be partial derivatives with respect to the same parameter twice, the off-diagonal terms will be double partial derivatives corresponding to all paired combinations of parameters in a symmetric manner.

As before we have: $E_\theta\left\{\frac{\boldsymbol{\delta}}{\boldsymbol{\delta\theta_i}} \boldsymbol{\lambda_n}(\boldsymbol{X}|\boldsymbol{\theta})\right\} = 0, i = 1(1)k$ … (2.10)

Hence, from (2.9), we also have that:

$$Var_{\theta}\{\nabla\lambda_n(\boldsymbol{X}|\boldsymbol{\theta})\} = -E_{\theta}\{\nabla^2\lambda_n(\boldsymbol{X}|\boldsymbol{\theta})\} \quad \ldots (2.11a)$$

We can also write the above k^2 equation from the double partial derivative matrix as:

$$E_{\theta}\left\{\frac{\delta\boldsymbol{\lambda_n(X|\theta)}}{\delta\theta_i}\frac{\delta\boldsymbol{\lambda_n(X|\theta)}}{\delta\theta_j}\right\} = -E_{\theta}\left\{\frac{\delta^2\boldsymbol{\lambda_n(X|\theta)}}{\delta\theta_i\delta\theta_j}\right\}, i,j = 1(1)k \quad \ldots (2.11b)$$

Thus, finally from (2.11a), we obtain that: $I_n(\boldsymbol{\theta}) = -E_{\theta}\{\nabla^2\lambda_n(\boldsymbol{X}|\boldsymbol{\theta})\}$, and as seen previously, $J_n(\boldsymbol{\theta}) = -\nabla^2\lambda_n(\boldsymbol{X}|\boldsymbol{\theta})$.

Proceeding as before and applying multivariate convergence theory, we have the following result that, $\widehat{\boldsymbol{\theta}} \sim Normal(\boldsymbol{\theta}, I_n(\widehat{\boldsymbol{\theta}})^{-1})$, and also $\widehat{\boldsymbol{\theta}} \sim Normal(\boldsymbol{\theta}, J_n(\widehat{\boldsymbol{\theta}})^{-1})$. This gives us the expression to arrive at the confidence interval of the $it\hbar$ parameter as follows:

$$\hat{\theta}_i \pm c\sqrt{\{I_n(\hat{\theta}_i)^{-1}\}_{ii}} \quad \ldots (2.12)$$

Also the same expression holds by replacing I_n by J_n. What the term $\{I_n(\hat{\theta}_i)^{-1}\}_{ii}$ represents in (2.13) must be emphasized. First the Fisher matrix is to be inverted and then the diagonal element corresponding to the $it\hbar$ parameter must be considered to obtain the confidence interval.

Model selection

The *Bayesian Information Criterion* (BIC) was introduced by Schwarz (1978) as a competitor to the *Akaike* (1974) *Information Criterion* (AIC). The computation of BIC is based on the log likelihood without any need to define priors and it is given by the following formula:

$$BIC = -2\lambda_n\left(\boldsymbol{X}\middle|\widehat{\boldsymbol{\theta}}\right) + k\, ln(n) \quad \ldots (2.13)$$

where, k is the number of parameters (vector, $\boldsymbol{\theta}$ with its estimate $\widehat{\boldsymbol{\theta}}$) in the model and n is the sample size while $\lambda_n\left(\boldsymbol{X}\middle|\widehat{\boldsymbol{\theta}}\right) = ln\prod_i f(X_i, \widehat{\boldsymbol{\theta}})|\widehat{\boldsymbol{\theta}}$ or the log likelihood as before.

The other criterion, AIC has a similar expression for its computation:

$$AIC = -2\lambda_n\left(\boldsymbol{X}\middle|\widehat{\boldsymbol{\theta}}\right) + 2k \quad \ldots (2.14)$$

with the same notation convention as before.

As can be clearly observed, the term for goodness of fit is the same for both criteria. But the penalty term is more stringent for BIC than for AIC since for large samples, that is when $n \geq 8$, we see that, $k\, ln(n) > 2k$. This leads to BIC preferring more parsimonious models than AIC.

Hence the differences in models selected by the two criteria would be especially glaring in large sample settings.

Further, let us consider two definitions: consistency and asymptotic efficiency. Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collections under consideration. A *consistent* criterion will asymptotically select the fitted candidate model having the correct structure with probability one in the limit of infinite amount of data. On the other hand, if we assume that the generating model is of infinite dimension, and thus lies outside of the candidate collection under consideration. Then, an *asymptotically efficient* criterion will asymptotically select the fitted candidate model which minimizes the mean square error of prediction. AIC is asymptotically efficient yet not consistent whereas BIC is consistent and not asymptotically efficient.

A pragmatic point of view would advocate the following:

a. AIC should be preferred when the ultimate goal of the model is predictive, that is to profess a model which predicts new outcomes efficiently.
b. BIC is useful when more importance is granted to the description of a model, that is, to build a model which will include the most important contributing factors to the outcome.

Lastly, with increasing sample size, predictive accuracy is enhanced as subtle effects are included in the model. While AIC will support the inclusion of such influences, BIC will not.

References

Akaike H. (1974) A new look at the statistical model identification. IEEE Trans. Autom. Control 19(6): 716-23.

Copenhaver G.P., Housworth E.A., Stahl F.W. (2002) Crossover interference in Arabidopsis. Genetics 160: 1631-39.

Falque M., Anderson L. K., Stack S. M. *et al.* (2009) Two types of meiotic crossovers coexist in maize. Plant Cell 21: 3915-25.

Kleckner N., Zickler D., Jones, G.H. *et al.* (2004) A mechanical basis for chromosome function. Proc. Natl. Acad. Sci. USA 101: 12592-97.

McPeek M. S., Speed T.P. (1995) Modeling interference in genetic recombination. Genetics 139: 1031-44.

Rao C. R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.

Rothenburg T. J., Leenders C. T. (1964) Efficient estimation of simultaneous equation systems. Econometrica 32: 57-76.

Schwarz G. (1978) Estimating the dimension of a model. Ann. Statist. 6(2): 461-64.

Theil, H. (1971) *Principles of Econometrics*. New York: Wiley.

PART I

STUDY OF INTERFERENCE HETEROGENEITY IN ARABIDOPSIS

Chapter 3

Making things happen…

PEDAGOGICAL EXPLANATION OF OUR METHODS

This chapter describes the innovative developments reported in the first publication of my thesis (Basu-Roy *et al.* 2013). Since the entire published text will follow after Chapter 4, I remain relatively succinct here.

Ad-hoc interference heterogeneity detection

There exists a very useful statistical measure of the degree of variation of observables, known as the coefficient of variation (CV) which is the ratio of the standard deviation, σ and the mean, $\bar{X}$ of a dataset. For the present discussion, the dataset consists of inter-crossover distances. In the case of data generated using the Gamma model, CV has been found to be inversely proportional to interference strength, so that higher CV indicates less interference and vice-versa.

We wanted to utilize CV ($\sigma/\bar{X}$) to probe whether interference strength varies with position along the chromosome. When using CV to provide an overall intuition about interference variation along a chromosome, it becomes important to modify the CV formula so that it focuses on a local region. What is done is position-related weights are assigned to each distance between adjacent crossovers and then the CV is computed with these weights. These weights are taken to be exponential with the distance of the current position and the mid-point of the COs under consideration, making the procedure akin to a sliding window along the chromosome.

Let the present position be denoted as G which is usually the mid-point between two genetic markers. And suppose that the mid-point between the ith adjacent pair of crossovers is Y_i, then the associated weight is considered to be $w_{G_i} = \exp(-10(G - Y_i)^2)$. Thus using these weights, a CV is computed corresponding to each position X on the chromosome. So we have a weighted mean and weighted standard deviation as follows:

$\bar{X}_G = \sum_i w_{G_i} d_i \ / \sum_i w_{G_i}$ and,

$$\sigma^2(X_G) = \sum_i (w_{G_i} - \bar{X}_G)^2 / \sum_i w_{G_i} \quad \text{... (3.1)}$$

Likelihood for discontinuous chromosome regions

This likelihood computation was done as a part of our endeavour to compare interference between different regions for the same chromosome. When it was required to see if there is was significant difference between the central portion and distal regions of chromosomes, it was imperative to deduce likelihood for the extremities. These regions were the first and last quarter of the chromosome in genetic length with an intervening *hidden* region. This implied that this central region had to be treated as missing data. The only information available at hand was whether that missing region was recombinant or not. For our work which was implemented in our computer code, we had to consider both the recombinant and the non-recombinant cases.

When the hidden region is recombinant, the hidden region must have an odd number of crossovers (one or three or five and so on..). While if the central portion is non-recombinant, one has to consider all even numbers of COs (zero or two or four and so on…). Here what stood us in good stead was the additive property of the shape parameter of the gamma distribution. The second layer of complexity is added to the situation when we first need to understand the problem at the bivalent or four-chromatid level (before *thinning*) and then assess the situation at gamete or single-chromatid level (after *thinning*). It is necessary to have a clear picture in both scenarios as the data available could be at either level. In this particular case, we worked with the *Arabidopsis* data (Giraut *et al.* 2011, Data I in Supporting Datasets).

We began with the likelihood provided by Broman & Weber (2000) for the full chromosome and proceeded with the 'without thinning' case and moved on to 'with thinning'. This quite involved computation is laid out comprehensively with explanatory figures and text in the Supplementary Information section of my Genetics paper (pages 10 SI – 19 SI).

Discerning non-interfering pathway (P2) heterogeneity

We undertook this analysis by comparing simulated and observed histograms using Pearson's chi-square test within the R statistical software.

Each interval between markers is considered the *reference* interval in turn. For each interval, considering only those gametes with a crossover in this interval, we count the number of crossovers in each of the *other* intervals. Keeping the reference fixed, the number of crossovers in the other intervals in gametes with a *total* of two and three crossovers are enumerated separately. The frequency corresponding to the reference interval of course is ignored. Once we have these separate two and three-crossover frequency series or histograms corresponding to each inter-marker interval from the experimental data, we follow the same steps to derive their

counterparts using data simulated with the Gamma model where the parameters are obtained from the global fit for each chromosome. Then each pair of frequency series are compared using the R chi-square test function, $chisq.test(.)$.

The combined p-values from two and three-crossover gametes were procured using the R function, $pchisq(.)$. This function first retrieves the Pearson's chi-square statistic based on the two *p*-values (two and three-crossover cases), adds these statistic values and recalculates the *p*-value. Further this function in perspective to our case can only be used if there exist both two as well as three-crossover gametes for a particular reference interval. If both cases do not exist, then the previous *p*-values are retained. Finally, we apply the Bonferroni correction at a family-wise error rate of 5% for male and female meioses and each chromosome (more details on page 20 SI of Supplementary Information of Genetics paper).

References

Basu-Roy S., Gauthier F., Falque M. *et al.* (2013) Hot regions of non-interfering crossovers coexist with a nonuniformly interfering pathway in *Arabidopsis thaliana*. Genetics 195: 769-79.

Giraut L., Falque M., Drouaud J. *et al.* (2011) Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. PLoS Genet. 7: e1002354.

Chapter 4

Mission published…

INTERFERENCE IN ARABIDOPSIS IS HETEROGENOUS

I will give a concise tour of the work we published in the paper (Basu-Roy *et al.* 2013; called 13BR hereafter), the journal version of which follows this chapter. Without going into elaborate details, I describe the various analyses undertaken on this data to further the understanding of interference in *Arabidopsis* in particular and of all organisms in general.

Note: No Beam-film model results were provided in the paper. Nevertheless I present them here under one particular head as their comparison with the Gamma model results draws interesting parallels between the two approaches.

Qualitative Results

The position-dependent coefficient of variation explained in the previous chapter provides some intuition about the variation in interference along chromosomes (Figure 4.1). This approach indicates that interference shows variation, it does not remain constant. Also in a previous publication by Drouaud *et al.* (2007), the same conclusion had been reached for chromosome number 4. What are the scales of this variation and how does it differ between sexes or between chromosomes? Are there any trends for interference variation in different regions of chromosomes regardless of the sex? This paper has tried to explore the different levels of interference variation in Arabidopsis.

Parameter estimation given the interference models: Global view on comparisons

Recall that the Gamma model is a statistical model while the beam-film model is mechanistic. Within the Gamma model, we typically maximize the likelihood to obtain the best-fitted parameter(s) whereas the projected likelihood score (PLS) is maximized to infer the beam-film parameter(s).

To make a justified comparison between these two models, the importance of using the same score cannot possibly be overstated. To compare these two models the simplest approach is to fit both using the same PLS score; as shown previously, the PLS provides nearly the same inference of parameters in the Gamma model as the maximum likelihood approach (Falque *et al.* 2009).

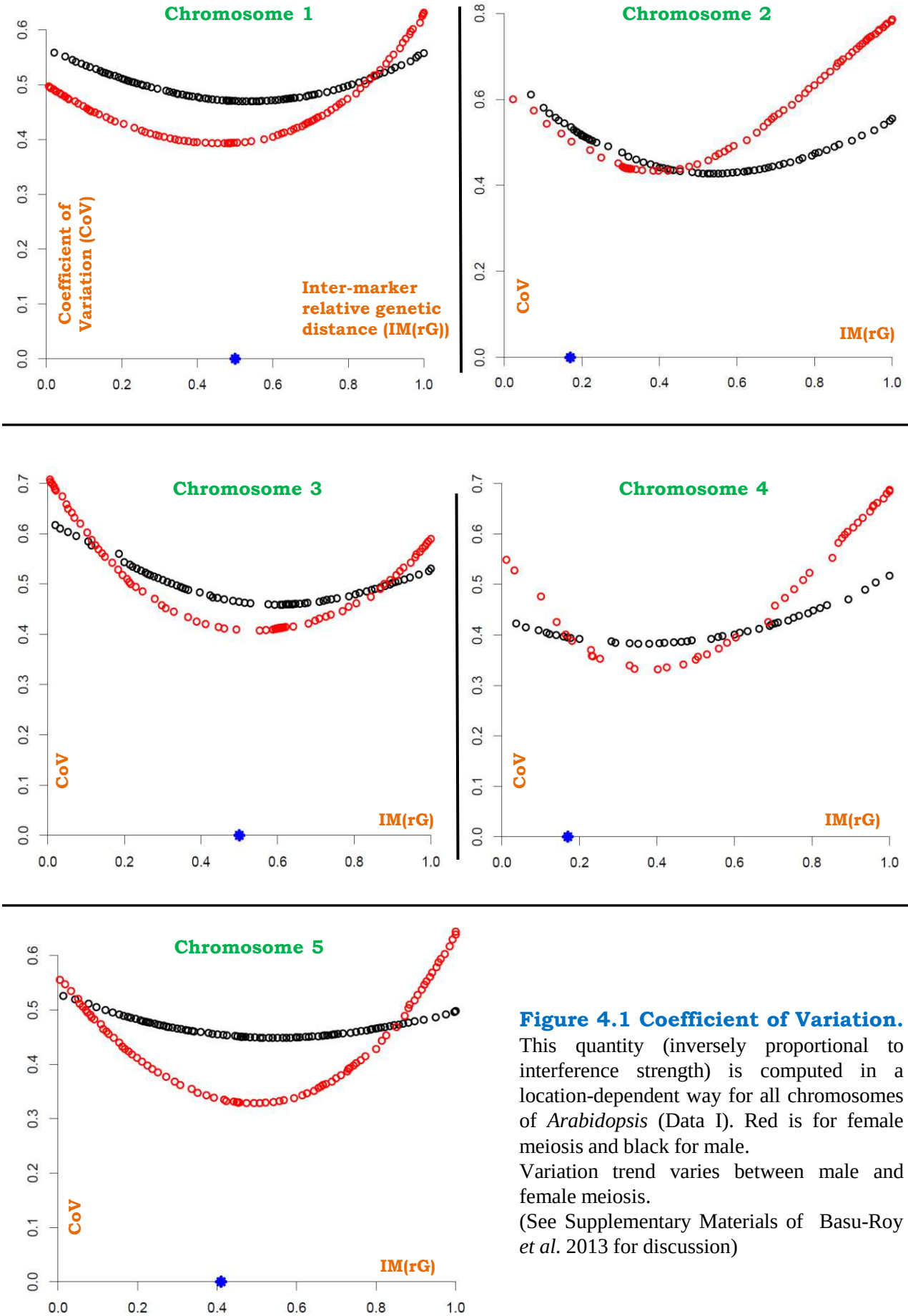


Figure 4.1 Coefficient of Variation. This quantity (inversely proportional to interference strength) is computed in a location-dependent way for all chromosomes of *Arabidopsis* (Data I). Red is for female meiosis and black for male.

Variation trend varies between male and female meiosis.

(See Supplementary Materials of Basu-Roy *et al.* 2013 for discussion)

In the published work in absence of the beam-film, all Gamma model results are obtained from maximizing the likelihood.

Single pathway (Gamma v/s Beam-film)

Using PLS score maximization, we first obtain the best fits of the gamma parameter, nu and the beam-film parameter, λ assuming a single (interfering) pathway only. Then in order to compare the two models, we inquire if they show similar trends when it comes to the different sexes (Figure 11). Such a direct comparison is possible as both parameters, nu and λ, are directly monotone in interference strength.

From both models, the trend is seen to be similar (Figure 4.2). Female meiosis shows higher interference than male meiosis. The interference estimates are more clustered for the Gamma model than for beam-film.

Note that since it is known that *Arabidopsis* shows non-interfering crossovers as well as interfering ones, the single pathway results give only a qualitative perspective.

Two-pathway (Gamma v/s Beam-film)

Here we estimate the parameter values using the PLS score for the Gamma model and the beam-film model. In addition to the parameter *nu* associated with interference strength (Figure 4.2), there is the parameter p which gives the proportion of the non-interfering pathway (Figure 4.3). The 95% confidence interval limits are also determined for each of these parameters. Subsequently we compare separately the two parameter classes across sexes and chromosomes for each model.

Our study leads to similar conclusions in the two models. While the interfering parameter is found to be higher in female meiosis, the parameter p is lower in female meiosis (Figure 4.3 (a), (b)). Further, since the parameter p is common to both models and has the same meaning in both, it is worthwhile to delve into whether the value of this parameter is consistent in these models. Interestingly, the value of p is found to be strikingly close and strongly correlated when considering the different sexes and chromosomes (Figure 4.3 (c)). This shows that even though the Gamma and beam-film models use completely different frameworks, there is good agreement in the final conclusions they lead to.

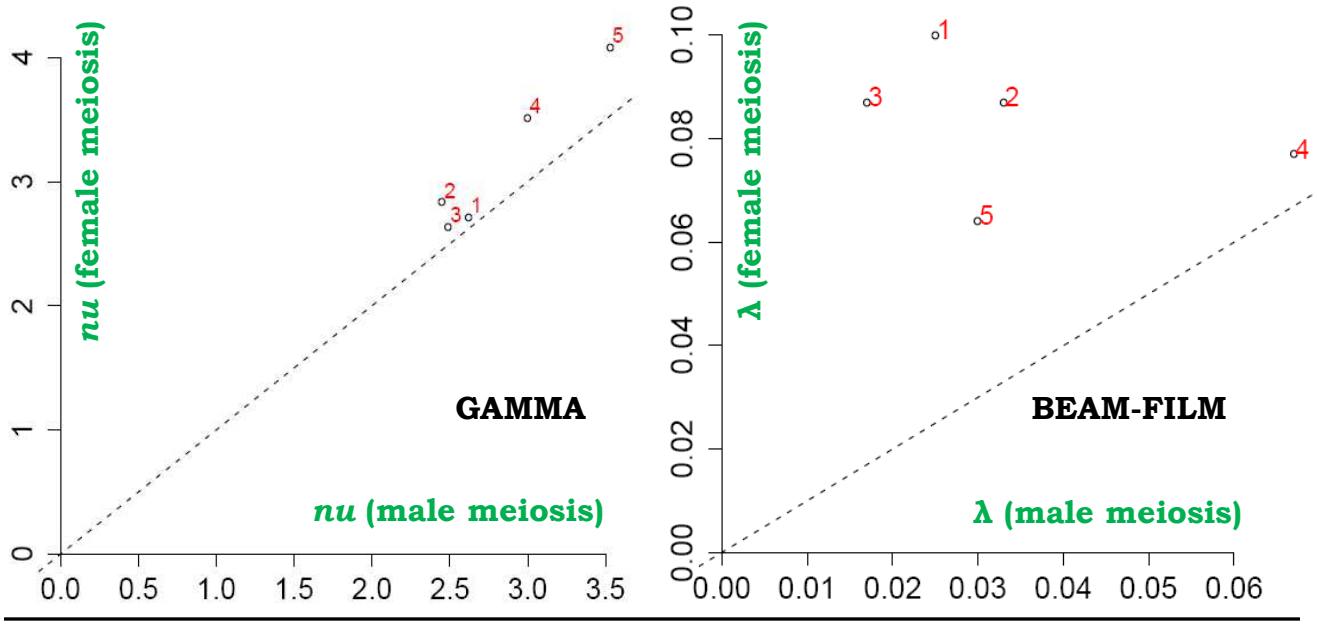


Figure 4.2 a| Single Pathway Interference Model Estimates.
The interference strength parameters *nu* (gamma model) and λ (beam-film) have been estimated for male and female meiosis for the five chromosomes (numbered 1 through 5) of Arabidopsis (Data I) . The estimated values are compared between male and female meiosis. Female meiosis shows higher interference.

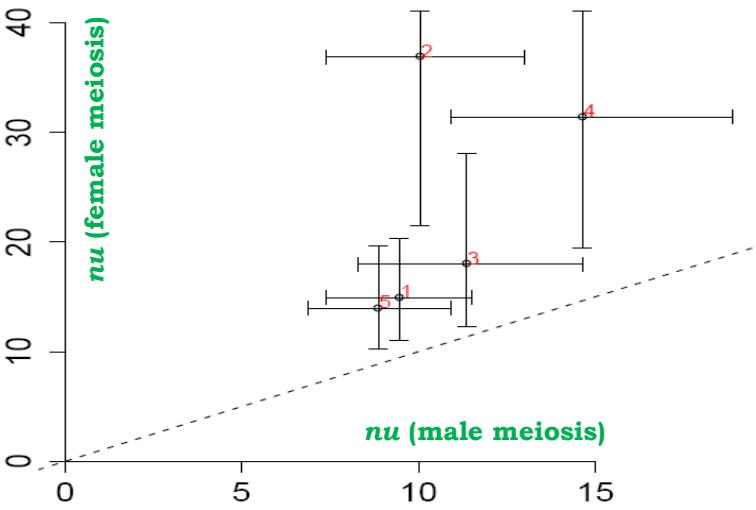
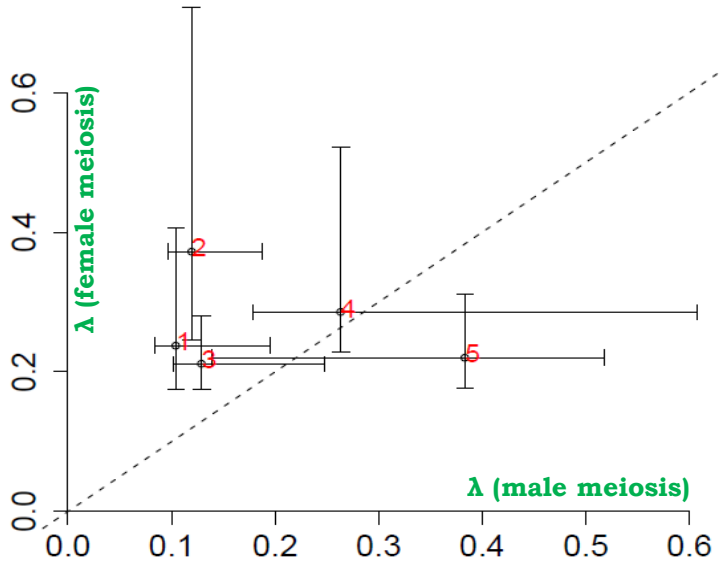


Figure 4.2 b| Gamma Two-pathway Model Estimates. The gamma interference strength parameter ***nu*** has been estimated for male and female meiosis for the five chromosomes (numbered 1 through 5) of Arabidopsis (Data I) . The estimated values are compared between male and female meiosis. Female meiosis again shows higher interference.

Figure 4.2 c| Beam-film Two-pathway Model Estimates. The beam-film interference strength parameter **λ** has been estimated for male and female meiosis for the five chromosomes (numbered 1 through 5) of Arabidopsis (Data I) . The estimated values are compared between male and female meiosis. Female meiosis shows higher interference yet again.



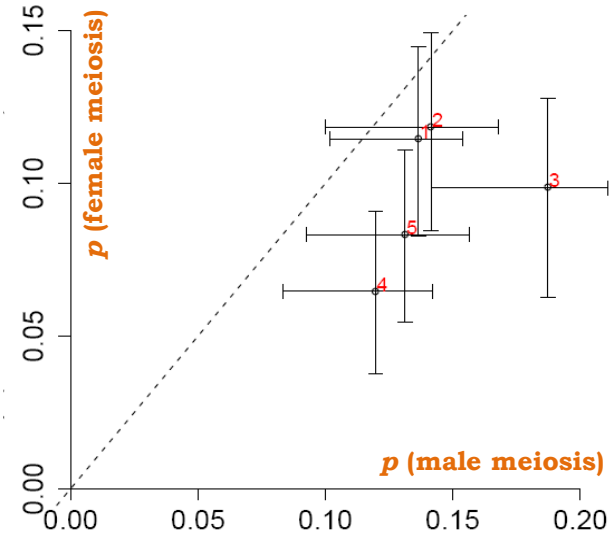
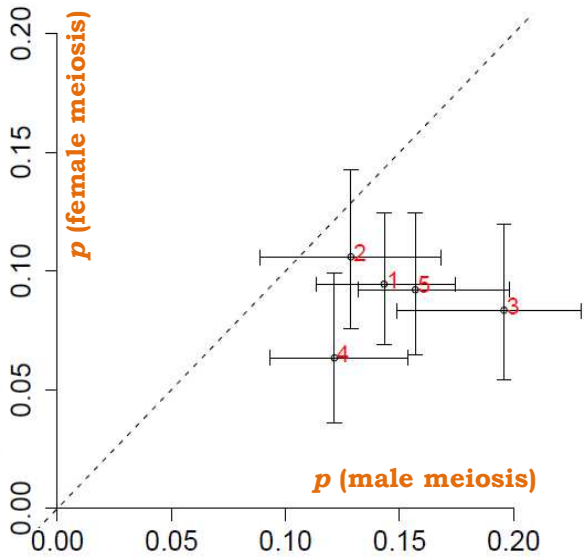


Figure 4.3 a| Gamma Two-pathway Model Estimates. The gamma parameter ***p*** denoting the proportion of non-interfering pathway has been estimated for male and female meiosis for the five chromosomes (numbered 1 through 5) of Arabidopsis (Data I) . The estimated values are compared between male and female meiosis. The parameter values are higher for male meiosis than for female.

Figure 4.3 b| Beam-film Two-pathway Model Estimates. The beam-film parameter ***p*** denoting the proportion of non-interfering pathway has been estimated for male and female meiosis for the five chromosomes (numbered 1 through 5) of Arabidopsis (Data I) . The estimated values are compared between male and female meiosis. Male meiosis shows higher values than female.



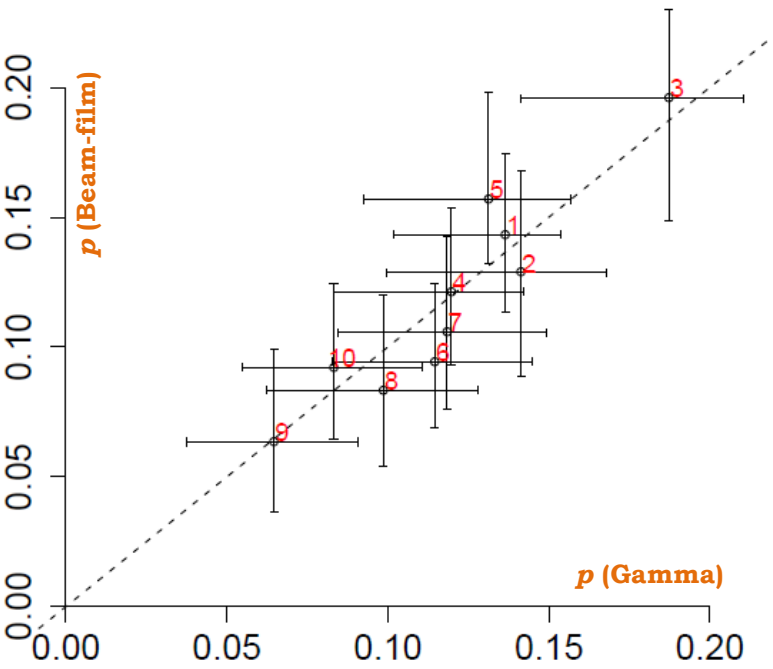


Figure 4.3 c| Comparison between Models. The second parameter *p* in two-pathway models common to gamma and beam-film are compared in all the five chromosomes (numbers 1 through 5 for male; 6 through 10 for female meiosis) of Arabidopsis (Data I).

Two different models give highly correlated results for the parameter common among them.

Whole chromosome interference comparison (Gamma model only)

All these comparisons were done using the Welch t-test, since unlike the traditional t-test, the Welch test does not assume that the two datasets of interest have equal variances.

Male v/s Female

This comparison has become possible as male and female meiosis data sets have become available. Although such differences have long been speculative, at present there are numerous organisms where differences between sexes have been exhibited. Posteriori, such differences are not unexpected because male and female meiosis usually occur in different parts of an organism and within very different cells. Meiosis regulation is likely to be different in these cells, leading to potentially both different genetic lengths and interference strengths.

We have discussed the male-female comparison to a certain extent in the previous section using a single pathway. The change here is that we focus on the Gamma model only, and then the parameter estimation can be performed using the likelihood. This approach results in the same observational trends for both single and two-pathway models. The highest female to male interference ratio (1.3) is seen in the fifth chromosome under the single pathway model (Figure S1, 13BR). The two-pathway framework in contrast finds the highest ratio (3.9) to concern the second chromosome (Figure 1, 13BR). For the additional parameter, p which is higher for male than female meiosis, the highest male/female ratio is observed in the third chromosome (Figure 2, 13BR).

Between chromosomes for male and for female meiosis

Variation in interference amongst chromosomes is not expected if interference is a simple mechanical consequence of crossover formation. Nevertheless, previous reports have found differences in interference strength between chromosomes in various species, so it would not be surprising to find the same trend in Arabidopsis. Furthermore, it is possible that interference strength as we measure it is affected by chromosome size even if the underlying mechanism is not.

The fourth chromosome exhibits the highest interference in male and female meiosis via the single pathway model. Interestingly, this higher interference among female chromosomes may justify why they have fewer crossovers if interference is an evolutionary mechanism for minimizing the number of COs. Within male or female meiosis, the chromosome with the highest or lowest values of interference (or the parameter p) show significant differences against intermediate-valued chromosomes (Tables 1 and S1 in 13BR).

Intra-chromosomal interference comparisons (Gamma model only)

When comparing sub-parts of chromosomes, the amount of data available (in particular the number of COs one can work with) reduces a lot leading to a reduction in statistical power of any test that is done on the two sub-datasets. Thus, when we compared the two chromosome arms or central and distal chromosomal regions, many comparisons yielded non-significant *p*-values.

In order to see if some overall trends existed even if individual chromosomes did not statistically suffice to detect them, we grouped the *Arabidopsis* chromosomes according to the position of the centromere - metacentric and acrocentric. While the second and fourth chromosomes are acrocentric, the first, third and fifth are metacentric. We thus grouped the chromosomes accordingly and re-did the comparisons to unravel any trends which might exist. Grouping chromosomes in this manner led to some interesting trends.

Between the two chromosome arms (Table S2, 13BR)

Under the single pathway model, the fourth chromosome shows significant difference in the parameter nu between the two unequal arms – the longer arm (on the right) is more interfering. And two-pathway analyses showed a notable difference between the arms of the second chromosome during male meiosis for both nu as well as p. When acrocentric chromosomes 2 and 4 were grouped together, we saw significantly lower value for p in the shorter left arm for male as well as female meiosis.

Between the central and distal chromosomal regions (Table S2, 13BR)

Considering the single pathway model (with the interfering pathway only), five of the ten chromosomes led to higher interference in the central region and some of these differences were significant. Further, the two-pathway model obtained higher values for both nu and p in the distal regions of chromosomes, a few of which were significant as well. And merging metacentric chromosomes gave significantly higher nu (male meiosis) as well as p (male and female meioses) in the extremities of chromosomes.

Heterogeneity in the non-interfering pathway (P2)

A special kind of scatter plot for all gametes having more than one crossover, where co-ordinates of adjacent crossover positions were displayed, led us to decipher the reason behind some unusually close crossovers. These plots clearly showed enrichment of nearby crossovers which we interpreted as COs contributing to the non-interfering pathway. Indeed, if there is a local

enrichment of COs from the non-interfering pathway, this is exactly the pattern that should emerge.

To delve deeper into this hypothesis, the Pearson's chi-square test was applied to compare the observed and model-predicted inter-crossover distance distributions, as explained in the previous chapter. This test results in highly significant *p*-values for several intervals (Figure 5, 13BR), for most of the chromosomes. That conclusion shows explicitly that the current two-pathway modeling is unable to explain all statistical aspects of observed crossover arrangement along chromosomes. To err on the cautious side, note that our test would yield significant results if gene conversions affected our data. To take a conservative approach, we thus re-did our analysis after removing all those cases that were compatible with gene conversions, *i.e.*, where crossovers occurred in adjacent marker intervals (Figure S4, 13BR). This gave 29 intervals with significant *p*-values. Thus the heterogeneity we see is not only driven by gene conversions. Clearly the present two-pathway models assuming uniform interference and fraction *p* of non-interfering COs along chromosomes has to be amended.

The endeavor of this published work was to explore interference variations and test the limits of the current modeling on state of the art data sets. Both crossover formation pathways consider the corresponding parameters to be constant along chromosomes. As we dug more into the variation scales, it became increasingly clear that in addition to varying between chromosomes and between male and female meiosis, interference shows variations along the chromosome length as well. Interference strength and the non-interfering crossover fraction exhibit local fluctuations from the global value considered by the two-pathway model framework. Given what we found, it is now time to suggest models that incorporate such heterogeneities. I have done so towards the end of my thesis.

References

Basu-Roy S., Gauthier F., Falque M. *et al.* (2013) Hot regions of non-interfering crossovers coexist with a nonuniformly interfering pathway in *Arabidopsis thaliana*. Genetics 195: 769-79.

Falque M., Anderson L. K., Stack S. M. *et al.* (2009) Two types of meiotic crossovers coexist in maize. Plant Cell 21: 3915-25.

GENETICS PAPER

Basu-Roy *et al.* 2013

Hot Regions of Noninterfering Crossovers Coexist with a Nonuniformly Interfering Pathway in *Arabidopsis thaliana*

Thesis Page Numbers (TPN) **70** through **104** (35 pages)

(In order to avoid confusion with the journal page numbers already present at the bottom of the page, TPN have not been inserted on every page but only on the *first* and *last* pages on the top left corner. Inserting page numbers into a pdf also results in reduction of letter and image quality.)

Hot Regions of Noninterfering Crossovers Coexist with a Nonuniformly Interfering Pathway in Arabidopsis thaliana

Sayantani Basu-Roy,* Franck Gauthier,* Laurène Giraut,† Christine Mézard,† Matthieu Falque,*,1,2 and Olivier C. Martin*,1

*Institut National de la Recherche Agronomique, Unité Mixte de Recherche de Génétique Végétale, Université Paris-Sud, 91190 Gif-sur-Yvette, France, and †Station de Génétique et Amélioration des Plantes, Institut Jean Pierre Bourgin, Institut National de la Recherche Agronomique, 78026 Versailles, France

ABSTRACT In most organisms that have been studied, crossovers formed during meiosis exhibit interference: nearby crossovers are rare. Here we provide an in-depth study of crossover interference in Arabidopsis thaliana, examining crossovers genome-wide in . 1500 backcrosses for both male and female meiosis. This unique data set allows us to take a two-pathway modeling approach based on superposing a fraction p of noninterfering crossovers and a fraction (1 2 p) of interfering crossovers generated using the gamma model characterized by its interference strength nu. Within this framework, we fit the two-pathway model to the data and compare crossover interference strength between chromosomes and then along chromosomes. We find that the interfering pathway has markedly higher interference strength nu in female than in male meiosis and also that male meiosis has a higher proportion p of noninterfering crossovers. Furthermore, we test for possible intrachromosomal variations of nu and p. Our conclusion is that there are clear differences between left and right arms as well as between central and peripheral regions. Finally, statistical tests unveil a genome-wide picture of small-scale heterogeneities, pointing to the existence of hot regions in the genome where crossovers form preferentially without interference.

SEXUALLY reproducing organisms undergo meiosis, thereby producing gametes having a level of ploidy equal to half that of the parental cells. This reduction in ploidy emerges from a complex and tightly controlled sequence of events. In most organisms, prophase I of meiosis begins by the active formation of double-strand breaks (DSBs) mediated by Spo11, a topoisomerase-like transesterase (Keeney et al. 1997). Then homologous chromosomes align and pair as the DSBs are repaired, typically using a homolog as template. Such DSB repairs can lead to either a crossover (CO), a reciprocal exchange of large chromosomal fragments between homologs), or to a noncrossover (NCO), a nonreciprocal exchange of small chromosomal segments between homologs, detected through associated gene conversions (Bishop and Zickler 2004)). COs mediate intrachromosomal rearrangement of parental alleles, giving rise to novel haplotypes in the gametes and thus driving genetic diversity. They also provide a physical connection between the homologs, holding them in a stable pair (bivalent) and allowing their correct segregation during anaphase I (Page and Hawley 2004; Jones and Franklin 2006).

Studies in several plants and animals have shown that the average number of COs in meiosis may vary between male and female meiosis (see review by Lenormand and Dutheil 2005), but there is no general rule that governs the direction or degree of CO number variation. Similarly, the distribution of COs along chromosomes can also differ when comparing male and female meiosis (Drouaud et al. 2007). Such distributions are generally nonuniform: some portions of the physical chromosome seem more likely to recombine while others hardly ever do (e.g., close to the centromere: Jones 1984; Anderson and Stack 2002). Furthermore, COs do not arise as independent events—there is some "interference" between them. Generally, CO interference reduces the probability of occurrence of nearby COs

doi: 10.1534/genetics.113.155549
Manuscript received July 21, 2013; accepted for publication September 2, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.155549/-/DC1.
[1]These authors contributed equally to the work.
[2]Corresponding author: UMR de Génétique Végétale, INRA/Université Paris-Sud, Ferme du Moulon, 91190 Gif-sur-Yvette, France. E-mail: falque@moulon.inra.fr

(Sturtevant 1915; Muller 1916), leading to a lower variability in inter-CO distances than in the absence of interference. It has been suggested that CO interference plays a role in controlling the number of COs formed for each pair of homologs. Indeed, most organisms obey the obligate CO rule whereby each bivalent must have at least one CO to ensure proper segregation of the homologs; one way to ensure such an obligate CO is to have many DSBs and let them develop into COs. But at the same time having many COs may be deleterious, inducing genomic errors or breaking advantageous allelic associations. Interference might then be the signature of a mechanism allowing for the appearance of an obligate CO followed by suppression of CO formation in favor of NCOs. Such an interpretation is supported by the fact that most organisms produce far more DSBs than COs (Baudat and De Massy 2007), so that high interference strengths probably reflect selection pressures to control the number of COs. Interference is thus integral to a mechanistic understanding of meiosis, and it comes as no surprise that most organisms that have been studied do exhibit CO interference.

In the past decade, much progress has been made in identifying key genes and the associated pathways for CO formation. A first pathway (P1) is interfering and depends on proteins of the ZMM family such as MLH1 and MLH3. A second pathway (P2) seems to be noninterfering and depends on MUS81 with other associated proteins. The two pathways have been found to coexist in *Saccharomyces cerevisiae* (Hollingsworth and Brill 2004; Stahl *et al.* 2004), *Arabidopsis thaliana* (Higgins *et al.* 2004; Mercier *et al.* 2005), tomato (*Solanum lycopersicum*, Lhuissier *et al.* 2007), and mouse (*Mus musculus*, Guillon *et al.* 2005). These studies indicate that the proportion of P2 COs varies from species to species. For example, in tomato it hovers at ∼30% (Lhuissier *et al.* 2007) while in mouse it is ∼10% (estimated by putting together results from Broman *et al.* 2002; Froenicke *et al.* 2002; Falque *et al.* 2007). Outliers are *Caenorhabditis elegans* with only interfering COs and *Schizosaccharomyces pombe*, which shows only noninterfering COs. Variations in the proportion of P2 COs also seem to arise within a given species when comparing different chromosomes (Copenhaver *et al.* 2002; Falque *et al.* 2009).

Early ways to detect CO interference were based on the coefficient of coincidence (Ott 1999). More recently, various models of CO formation have been introduced; the fitting of these mathematical models to the experimental data allows one to (1) extract a quantitative estimate of interference strength and (2) dissect interference effects that are entangled in the mixture of interfering and noninterfering pathways, following present biological knowledge summarized in the previous paragraph. Multiple models have been proposed to incorporate interference in CO formation modeling. The most widely used ones take the COs to be generated by a stationary renewal process, which assumes that the genetic distances between successive COs are independently and identically distributed. This is the case of the gamma model (McPeek and Speed 1995), called so because the inter-CO distances follow a gamma distribution. When the shape parameter of this gamma distribution is restricted to be an integer, the model reduces to the chi square (Bailey 1961; Foss *et al.* 1993) or the counting model. All such models of interference have been designed to describe the formation of COs within the P1 pathway. In the case of modeling COs from two pathways (P1 being interfering and P2 not), one first simulates COs from P1 and then one uniformly "sprinkles" additional P2 COs without interference (Copenhaver *et al.* 2002). Using the gamma model for the P1 pathway accompanied by P2 sprinkling has led to estimates of p (the proportion of P2 COs) varying from 19 to 20% in *Arabidopsis* chromosomes 1, 3, and 5 (Copenhaver *et al.* 2002), 3 and 5% for *Arabidopsis* chromosomes 2 and 4 (Lam *et al.* 2005), ∼12% for the 10 maize chromosomes (Falque *et al.* 2009), 0–21% for humans depending on the chromosome (Housworth and Stahl 2003), and ∼10% for baker's yeast chromosome 7 (Malkova *et al.* 2004). Such modeling studies have also provided estimates of interference strengths via the fitted value of the shape parameter of the gamma model distribution. Note that the availability of confidence intervals on these parameters is often an issue, and systematic testing of differences between chromosomes has not yet been attempted. Are the currently estimated differences in interference strength or the P2 proportion p significant? Would it be possible instead to assign an overall (chromosome and sex-independent) interference and/or P2 proportion value to each species? Finally, is there any evidence of variation in interference at the intrachromosomal level?

As a first answer to these questions, previous work on *Arabidopsis* male chromosome 4 (Drouaud *et al.* 2007) found that both the local recombination rates and the synaptonemal complex lengths were significantly different when comparing male and female meiosis. Furthermore, those authors concluded that interference strength varied along the length of chromosome 4 through tests using the classical coefficient of coincidence (Ott 1999). But these tests had two limitations. First, large intervals were used to avoid statistical noise, erasing any small-scale variations in interference. Second, different interval sizes were pooled together to gain statistical power, introducing biases. It is thus worthwhile to see whether the modeling approach (based on fitting parameters of CO formation models rather than on measuring coefficients of coincidence) gives results similar to those of Drouaud *et al.* (2007) while adding more insight. A first step in this direction was provided by Giraut *et al.* (2011) using the single pathway gamma model. Although these researchers found higher interference strengths in female than male meiosis for all five chromosomes of *Arabidopsis*, they provided no tests and thus made no claims of statistically significant differences. As we shall see, the single-pathway approach has serious shortcomings, making it essential to use two-pathway modeling, which allows one to resolve interference into properties of the two pathways P1 and P2.

In this work, we exploit the two large reciprocal backcross populations produced by Giraut *et al.* (2011) to study CO interference in *A. thaliana*. The gamma single and two-pathway models are used to fit the data of male and female meiosis for all five chromosomes. The variability in interference at several levels is then analyzed: between different chromosomes (separately for male and female meiosis), between male and female meiosis for each chromosome pair, and finally, between different regions of the same chromosome. Significant differences are found at all levels. We also obtain a genome-wide picture of candidate intervals that are anomalously hot for the proportion of the noninterfering pathway. Finally, the discrepancies unveiled in this work between the *Arabidopsis* data and the fits demonstrate the need for more sophisticated models than the ones available today.

Materials and Methods

Experimental data

Plant material: The two *A. thaliana* accessions, Columbia-0 (Col) (186AV) and Landsberg *erecta* (Ler) (213AV), were obtained from the Centre de Ressources Biologiques at the Institut Jean Pierre Bourgin, Versailles, France (http://dbsgap.versailles.inra.fr/vnat/). The Col accession was crossed with Ler to obtain F_1 hybrids. These hybrids were backcrossed with Col: their pollen was used for male meiosis, while Col pollen was used for female meiosis. Further details of the crossings are given in Giraut *et al.* (2011).

DNA extraction: The plant material from the Colx(ColxLer) and (ColxLer)xCol populations was lyophilized (specifics in Giraut *et al.* 2011). Then DNA was extracted as explained in Giraut *et al.* (2011).

Selection of single-nucleotide polymorphism markers and genotyping: For the two populations associated with male and female meiosis, a set of 384 SNP markers (Supporting Table S1 of Giraut *et al.* 2011) were chosen from the Monsanto and Salk Institute databases based on even physical spacing along the chromosomes (details in Giraut *et al.* 2011). Markers and plants with too many undetermined genotypes were removed from the final data set. The resulting populations consisted of 1505 and 1507 plants having genotype data from 380 and 386 markers for the male and the female populations, respectively (380 markers in common). Totals of 222 and 163 singletons were verified in the male and female populations, respectively, using PCR and DNA sequencing.

Single-pathway interference modeling

Model: We have worked within the standard hypothesis that the two CO formation pathways produce COs independently (Copenhaver *et al.* 2002; Argueso *et al.* 2004) and that P2 has no interference at all (Copenhaver *et al.* 2002). P2 participates only in the two-pathway modeling, which will be discussed below; here we need to consider only the first pathway that is interfering. To specify the P1 pathway framework, we used the gamma model (McPeek and Speed 1995) at the level of the bivalent (two homologs, four chromatids). To completely define the model, one has to provide the genetic length L_G of the chromosome considered and the interference parameter *nu*, which can take any value in $[1, \infty]$; these two quantities are independent. L_G is simply set to the experimentally observed genetic length. The parameter *nu* quantifying the pathway's interference strength corresponds to the "shape parameter" of the gamma distribution used in the process of generating the genetic positions of successive COs. In addition, $2*(nu)$ is the "rate" of that gamma distribution on the bivalent, ensuring that the density of COs is two per Morgan as it should be by definition of genetic distances. Note that the backcross data lead to information on only one of the four gametes produced during each meiosis. Properties of CO patterns at the gamete level were deduced using the assumption of no "chromatid interference" (Zhao *et al.* 1995; Copenhaver 1998) (details in Supporting Information).

Estimation: Given a value *nu* of interference strength, the likelihood for each backcross genotype was computed. Since each backcross is associated with a different meiosis, the likelihood L for the whole data set is the product of the likelihoods of each meiosis. Then we obtained the "best" value of the interference strength *nu* by maximizing L (this is the classical maximum-likelihood method), adjusting the model parameter *nu* using a "hill-climbing" procedure (Gauthier *et al.* 2011) (details in Supporting Information). Although the computation of likelihoods has been provided previously for whole chromosomes, in the present work we were also able to compute the likelihoods when chromosome portions under consideration did not form a continuous stretch (details given in Supporting Information). Such calculations allowed us to perform comparisons of interference strength between the central regions and the extremities of chromosomes. Not surprisingly, our whole-chromosome single-pathway estimates agree with those reported by Giraut *et al.* (2011) as these were based on the same data and same maximum-likelihood approach. Confidence intervals were computed using the Fisher information matrix.

Two-pathway modeling via sprinkling

The P2 (noninterfering) COs were put down randomly with uniform density in genetic position (that is, along the genetic map) and then superimposed or "sprinkled" (Copenhaver *et al.* 2002) onto the P1 COs. On the bivalent, the density of P2 COs (defined as their mean number per Morgan) was 2 times the proportion (chromosome-wide) of the noninterfering pathway COs, p, where p lies in [0,1]. Similarly, the density of P1 COs was 2 times $(1 - p)$, leading to a value of the rate parameter $2*(nu)*(1 - p)$. This two-pathway gamma model is then specified by the genetic length L_G, *nu*, and p; these three quantities are independent.

The first is set to its experimentally observed value while the other two are adjustable. The adjustment was obtained by maximizing the likelihood L for a given chromosome as described above except that here L had two parameters (nu, p) spanning a two-dimensional parameter space. Again, the hill-climbing algorithm (Gauthier *et al.* 2011) was used for maximization. After the adjustment, confidence intervals were obtained from the Fisher information matrix.

Statistical procedures and comparison tests

Comparing two data sets (separate chromosomes or different regions of one chromosome): We examined differences of interference strength at three levels. We compared the effective interference (using the single-pathway model) as well as the P1 interference and the proportion of non-interfering COs (using the two-pathway model) between (1) male and female meiosis for the same chromosomes, (2) between the different chromosomes but for a given sex, and (3) between different regions or the two arms of the same chromosomes. We tested the null hypothesis (H_0) that the two data sets being compared have equal means, using the Welch t-test rather than a t-test. Indeed, the classical two-sample t-test assumes the sample variances to be equal, which is not valid for the comparisons here. The Welch t-test generalizes the standard t-test to allow for unequal variances for the two samples. When the null hypothesis is rejected, it indicates that there is a statistically significant difference between the means of the two samples (details given in Supporting Information).

Detecting intervals hot for P2 COs: We compared simulated and experimental data sets to detect hot intervals specific for P2. For each interval between adjacent markers, we first selected the plants having a CO in that interval. The frequency of COs in each of the other intervals was then computed, separating the cases of gametes with a total of two and three COs. The same was done for simulated data, generated with the *simdata* option in CODA (Gauthier *et al.* 2011) using the nu and p values obtained by fitting the experimental data. Expected (simulated or "theoretical") and observed (experimental) distributions of COs for each intermarker interval were contrasted by Pearson's chi-square test (Lindsey 1995) within the R statistical software, *chisq.test()*. This allowed us to test for each interval the null hypothesis that the two distributions (experimental and simulated) are similar. Furthermore, to better exploit the data, we merged the values from the two-CO and three-CO cases by taking the sum of the corresponding chi-square values (for intervals having data for two COs and three COs). The corresponding P-value was then computed by the R function, *pchisq()*. For intervals with data for one case only, the previous chi square and P-values were retained (details in Supporting Information). Since Pearson's test is performed for all the intervals, we applied the Bonferroni correction at a family-wise error rate (FWER) of 5% for male and female meioses and each chromosome.

Results

Whole-chromosome analyses

Single-pathway analyses: For each of the five chromosomes in *A. thaliana*, we estimated the values of the effective interference strength (given by the parameter nu) in male and female meiosis with the corresponding 95% confidence intervals, using the gamma model (see *Materials and Methods*). The fitted values of nu fall in a rather small range—from 2.4 to 4.1. Interestingly, female meiosis consistently exhibits higher values of nu than male meiosis (Figure S1). The highest female/male (F/M) interference ratios are seen for chromosomes 5 (1.3) and 4 (1.2) although these differences are not significant statistically (diagonal entries of Table S1). Comparing now different chromosomes for male meiosis, chromosome 4 has the largest nu, which is statistically different from the nu of chromosomes 1, 2, and 3 (top triangular part of Table S1). Furthermore, chromosome 5 has the second highest effective interference, and when compared to the two chromosomes with the lowest values (2 and 3), the differences in nu are statistically significant. Finally, for female meiosis, chromosome 4 has the highest effective interference while chromosome 3 has the lowest and the difference between them is significant (bottom triangular part of Table S1).

Two-pathway analyses: For the gamma-sprinkling two-pathway model, we found values of nu between 8 and 37, with the range for male meiosis being 8–15, and for female meiosis, 13–37 (Figure 1). These values are systematically higher than the ones obtained in the single-pathway modeling. Such a trend is expected since the single-pathway modeling provides only an effective interference strength that mixes contributions from the two pathways; whenever p is appreciable, effective interference strength will necessarily be low. Considering now the estimates of p, the values found lay within the range 0.06–0.19, with 0.12–0.19 for male and 0.06–0.12 for female meiosis (Figure 2).

Comparing the male and female meioses, just as in the single-pathway analyses, we find that nu is consistently higher in female meiosis than in male meiosis for all chromosomes; in particular, the female-to-male ratios for nu are highest for chromosomes 2 and 4 (3.9 and 2.2, respectively). These differences are statistically significant for three chromosome pairs (diagonal entries of Table 1 associated with nu). Furthermore, the F/M ratios for nu are much higher than those obtained within the single-pathway analyses that do not dissect the interference signal into two pathways.

We obtain confidence intervals on p that do not contain the point $p = 0$. We therefore exclude at the 5% significance level the possibility of having only P1 COs: it indeed is necessary to use the two-pathway framework for all of the chromosomes for a sensible modeling. Furthermore, just as nu is larger for female meiosis than for male meiosis, we find that female meiosis has lower values of p than male meiosis; the highest male-to-female ratio (1.9) occurs for chromosome

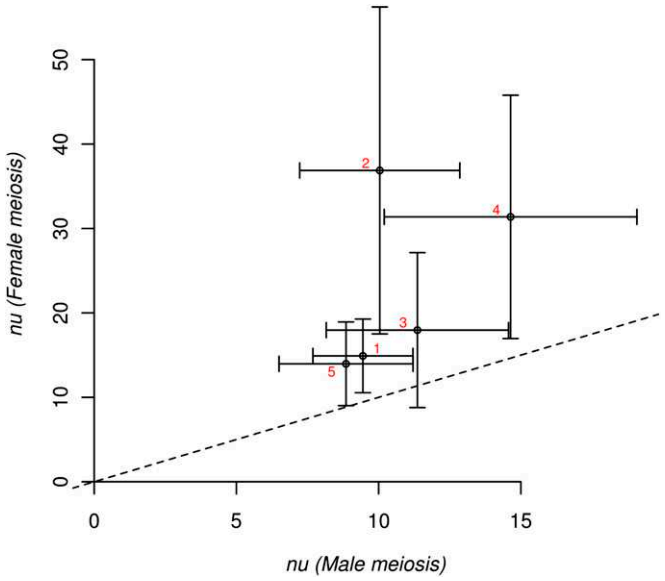


Figure 1 Estimated values of *nu* for the two-pathway gamma model. The *nu* estimates (black circles) for the five chromosome pairs (1–5 in red) for male (*x*-axis) and female (*y*-axis) meiosis with their 95% confidence intervals (black solid lines). The diagonal (black dashed line) is for $y = x$.

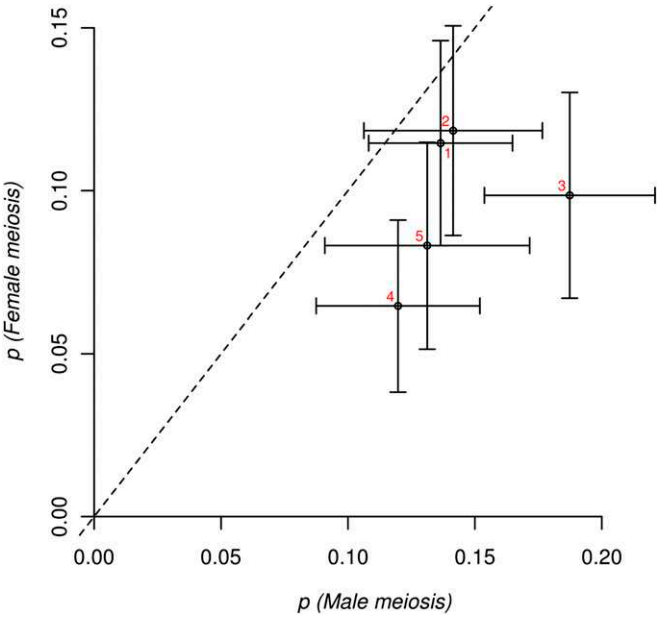


Figure 2 Estimated values of *p* for the two-pathway gamma model. The estimates for the model parameter *p* (black circles), which is the proportion of COs from the noninterfering pathway for the five chromosome pairs (1–5 in red) for male (*x*-axis) and female (*y*-axis) meiosis with their 95% confidence intervals (black solid lines). The diagonal (black dashed line) is for $y = x$.

3. This difference is significant for two among the five chromosome pairs (diagonal entries of Table 1 associated with *p*).

Compare now the different chromosomes for their level of P1 interference strength *nu* and proportion *p* of P2 COs for male and female meiosis separately. Beginning with male meiosis, chromosome 4 has the highest *nu* value (14.6) that is statistically different from that for chromosomes 1 and 5 (male-male comparisons, top triangular part of Table 1, entries associated with *nu*). Considering the values of *p* in male meiosis, chromosome 3 has a significantly larger proportion of P2 COs than chromosomes 1, 4, and 5 (top triangular part of Table 1, entries associated with *p*). For female meiosis, chromosomes 2 and 4 have higher values of *nu* as compared to chromosomes 1, 3, and 5, and many of the associated comparisons are statistically significant (female-female comparisons, bottom triangular part of Table 1, entries associated with *nu*). We also find that chromosomes 1 and 2 have greater values of *p* than the others, while chromosome 4 has the lowest; most of the statistically significant comparisons arise when including chromosome 4 (bottom triangular part of Table 1, entries associated with *p*).

Intrachromosomal variation of interference

Uniformity of interference along chromosomes was tested via the difference in interference strength *nu* or parameter *p* (1) between the two arms of the chromosome (denoted "left" and "right" and separated by the centromere) or (2) between the central region (corresponding to half of the genetic length, taken between the fractions 0.25 and 0.75 of the whole chromosome) and the rest of the chromosome (extremities). These analyses were performed on individual chromosomes and when pooling the acrocentric chromosomes 2 and 4 on the one hand and the metacentric chromosomes 1, 3, and 5 on the other.

Single-pathway analyses: A few of the comparisons suggest interference strength heterogeneities. For example, chromosome 4F shows a significant difference between the *nu* values of the left and right arms (the suffix M or F denotes male or female meiosis, respectively); the right arm that is longer shows a higher interference strength (first column of Table S2). But when merging data sets into two groups — metacentric chromosomes 1, 3, and 5 and acrocentric chromosomes 2 and 4 — no significant differences are found between left and right arms in either groups, be it for male or female meiosis.

When comparing the central region to the extremities, there is no overall trend for *nu*: five chromosomes show higher interference in the central region while the remaining exhibit the opposite behavior. In spite of that, the difference is significant for certain chromosomes (see the second column of Table S2). Here again, merged data sets for chromosomes 1, 3, 5 and for chromosomes 2 and 4 do not yield significant differences.

Being based on a single pathway, all these results should be considered in a qualitative spirit only since the two-pathway analyses exclude the possibility that $p = 0$.

Two-pathway analyses: First consider the P1 interference strength parameter, *nu*. Among the comparisons between left and right arms, only chromosome 2M shows a significant difference, with a higher value for the right arm (third

Table 1 Results (*P* values) for two-pathway gamma-sprinkling model comparisons

	1M		2M		3M		4M		5M		Male
1F	*nu** 1F	—	—	—	—	*p** 3M	*nu** 4M	—	—	—	1M
2F	*nu** 2F	—	*nu** 2F	—	—	—	—	—	—	—	2M
3F	—	—	—	—	—	*p** 3M	—	*p** 3M	—	*p** 3M	3M
4F	*nu** 4F	*p** 1F	—	*p** 2F	—	—	*nu** 4F	*p** 4M	*nu** 4M	—	4M
5F	—	—	*nu** 2F	—	—	—	*nu** 4F	—	—	—	5M
Female	**1F**		**2F**		**3F**		**4F**		**5F**		

Comparisons (1) between male and female meioses for the same chromosome (diagonal values boxed) and (2) between different chromosomes in male meiosis (upper left values) and female meiosis (lower left values) separately. There are two boxes corresponding to each comparison, the left one for *nu* and the right one for *p*. The *P* values were computed for the null hypothesis that the two meioses under consideration are associated with the same value of *nu* or *p* (separately). Note that an asterisk indicates that the *P*-value is significant at a 95% level of significance. For every significant comparison, the chromosome indicated is the one having the higher nu or *p*. Empty cells refer to non-significant statistical tests.

column of Table S2). Merged data sets for chromosomes 1, 3, 5 and for chromosomes 2, 4 give no significant differences. However, for comparisons between the central region and extremities, *nu* tends to be higher in the extremities for the majority of the chromosomes (fifth column of Table S2; see also Figure 3). For the metacentric merged data consisting of chromosomes 1M, 3M, and 5M, we find a significant difference between these regions, with higher *nu* for the extremities. The other merged data show no trend.

Second, for the parameter *p*, the difference between left and right arms is significant only for chromosomes 2M and 3M (fourth column of Table S2). Considering merged data sets for acrocentric chromosomes 2 and 4, the right long arm shows significantly higher *p* than the left short arm in male as well as female meiosis. For comparisons between the central region and extremities, *p* is observed to be higher in the extremities for most chromosomes with several significant differences (sixth column of Table S2). For the merged data sets, metacentric chromosomes (1, 3, and 5) exhibit (significantly) higher values for *p* in the extremities for both male and female meiosis.

The passage from single to two-pathway modeling leads to fewer statistically significant differences because there is an additional parameter to fit and thus loss of power. Nevertheless, the merged data analysis provides an unambiguous trend of intrachromosomal interference heterogeneity, namely higher interference as well as a higher proportion of noninterfering COs in the extremities.

Hot intervals for the noninterfering (P2) pathway

Scatter plots of positions of CO pairs (see Figure S2 and Figure S3) reveal the presence of pairs close to the diagonal, indicating an anomalously low effective interference and thus presumably a high contribution of P2 in the corresponding regions. Furthermore, if an interval I* is hot for P2, that is, if the fraction of P2 COs arising in that interval is much higher than the value *p* inferred from the standard two-pathway modeling, then there will be an enrichment phenomenon whereby gametes with two or three (or more) COs will have a particularly high probability of having I* be recombinant. Such an enrichment leads to an excess of points on the horizontal or vertical line associated with that interval in the scatter plot of pairs of COs; this is indeed what is observed for a number of intervals (*cf*. Figure 4 and Figure S2 and Figure S3). Note that some of the events displayed correspond to CO position pairs that arise in several gametes; that is, there are points with multiplicities going up to 4 [Figure 4, made visible by introducing random noise in both axes using the R function *jitter()*].

To have an objective criterion for considering an interval to be hot for P2, we apply Pearson's chi-square test, comparing the theoretical and observed distribution of genetic distances between successive COs. The Bonferroni correction is then applied to take into account that there are as many *P*-values calculated for each chromosome as there are intervals (*cf*. *Materials and Methods* and Supporting Information). These tests reveal highly significant *P*-values for several intervals along most of the chromosomes, showing that the current two-pathway modeling does not adequately describe all of the statistical features in the experimental CO patterns. From the *P*-values derived for each interval, we obtain a putative genome-wide profile of hot P2 intervals (Figure 5). We see that the intervals for which *P*-values are highly significant suggest the presence of hot regions for the noninterfering (P2) pathway. In addition to the heterogeneity within these P2 hot regions, the pattern varies between chromosomes and between male and female meiosis. Some chromosomes show several average and major peaks while others show only one high peak. The positions of the peaks also vary, sometimes occurring next to the centromere in particular for male meiosis, sometimes farther down each arm as seems to be typical in female meiosis (Figure 5).

The profiles of Figure 5 suggest hot regions for P2 COs, but our test would also generate small *P*-values if there were many gene conversion events (due to NCOs) affecting our data. Noting that these events would give rise to double recombinants in adjacent intervals, we have reanalyzed the data after removing all such cases. For this modified data set, 29 intervals lead to significant *P*-values (Figure S4). Thus we reject the hypothesis that current two-pathway modeling (where P2 COs are uniformly sprinkled along chromosomes) describes the statistical features of the experimental CO patterns. This result was also reached before removing double recombinants in adjacent intervals, so gene

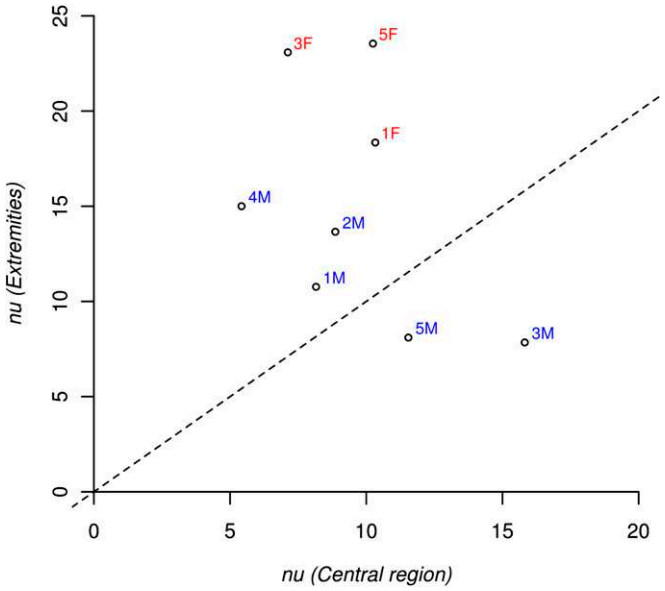


Figure 3 Comparison of the interference parameter (*nu*) between central region and extremities of chromosomes (two-pathway gamma model). The *nu* estimates (black circles) for the eight chromosomes (1–5 for male meiosis and 1, 3, and 5 for female meiosis) for the central region (*x*-axis) and the extremities (*y*-axis). The central region is defined as the middle half of the chromosome in genetic length, the rest of the chromosome forming the extremities. The diagonal (black dashed line) is for $y = x$.

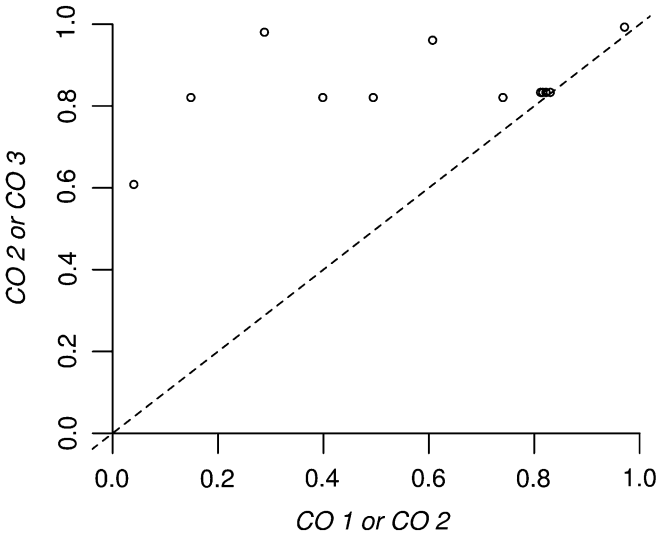


Figure 4 Scatter plot of successive CO positions for gametes having three COs on chromosome 2 in female meiosis. Each point gives the positions of the pair (*CO 1*, *CO 2*), *i.e.*, (first CO, second CO) or of the pair (*CO 2*, *CO 3*), *i.e.*, (second CO, third CO) where the positions are genetic and rescaled to lie in [0,1]. CO positions in adjacent or nearby intermarker intervals are close to the diagonal (black dashed line, $y = x$).

conversions on their own do not explain the heterogeneities we find in either Figure 5 or Figure S4.

Discussion

Female meiosis exhibits higher "effective" interference than male meiosis

Fitting data to the single-pathway gamma model provides a value for the associated interference parameter. The estimated values here are in agreement with those reported earlier (Giraut *et al.* 2011). We further find that the five chromosomes show higher effective interference in female meiosis than in male meiosis (Figure S1 and Table S1). However, we also find that the single-pathway model gives rise to poor adjustments to the experimental data. Such behavior is not surprising since, when using the two-pathway model, fitting leads to estimates of p (the P2 parameter) that are always incompatible with zero. To be on the safe side, single-pathway approaches should be considered to provide qualitative information only.

Female meiosis exhibits higher P1 interference and lower P2 proportion than male meiosis

We find that two-pathway modeling points to higher P1 interference strength in female than in male meiosis (Figure 1). Furthermore, the values for p are significantly higher in male than in female meiosis (Figure 2). These systematic differences, arising in all chromosomes, suggest that the action of interference is affected by the cellular environment; *i.e.*, the male and female meiocytes provide environments where the interference strength and presumably the proportions for each pathway are modulated at a systemic level. Clearly, the cellular environment effects recombination rates, and thus its effecting interference strength does not come as a surprise.

In the light of these results for P1 and P2, we can look back at the results of the single-pathway analysis. Because P1 is more interfering in females and p is higher in males, one expects the effective interference inferred by the single-pathway modeling to be stronger in females. As shown above, this is indeed what the single pathway finds; in fact, it does so for all chromosomes.

In previous studies (Vizir and Korol 1990; Giraut *et al.* 2011), it was observed that the M/F overall recombination ratio in *A. thaliana* is ~1.93. Could the extra genetic length of the male genetic maps be due to just an increase in COs from P2, keeping P1 unchanged both for the number of COs and their level of interference? The answer is "no": we know that P1 COs see both their numbers increased and their interference level reduced when going from female to male meiosis because the male–female differences in *nu* are often statistically significant.

Comparison to previous two-pathway studies

Genome-wide CO interference in *Arabidopsis* has been studied previously by other authors (Copenhaver *et al.* 2002; Lam *et al.* 2005) but only for male meiosis. Using the single-pathway gamma model, it was concluded that the effective interference parameter *nu* lies in the range 4–10 (Copenhaver *et al.* 2002) while in our analyses we have *nu* going from 2.4 to 3.5 (Figure S1); it is not possible to make

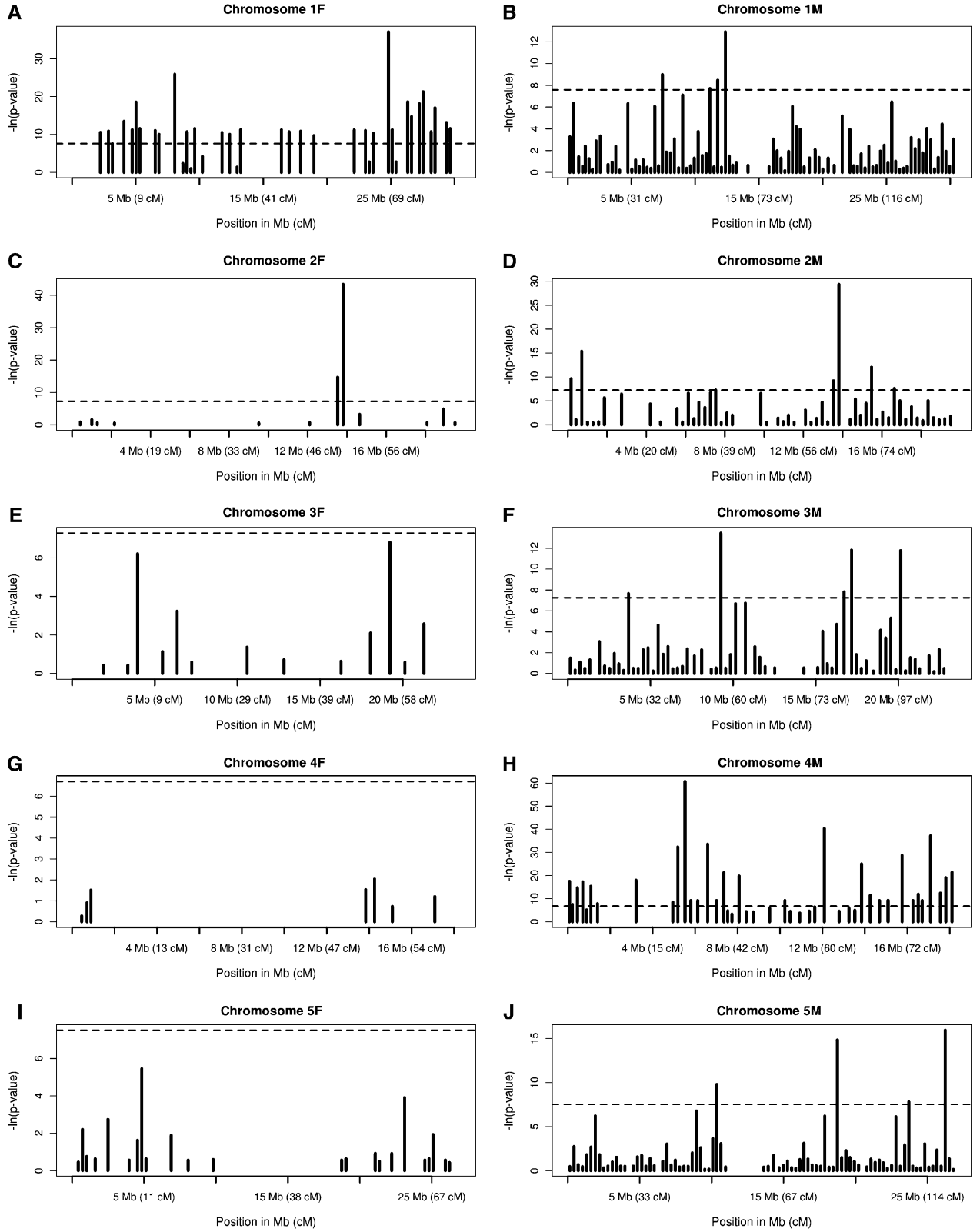


Figure 5 Genome-wide view of intervals hot for the (noninterfering) P2 pathway. *x*-axis: marker interval positions in mega basepairs along the chromosome considered. *y*-axis: minus the natural logarithm of the *P*-value of Pearson's chi-square comparison test for that interval. This *P*-value corresponds to the null hypothesis that P2 crossovers are uniformly distributed along the chromosomes. The dashed horizontal line corresponds to the FWER of 5% when using the Bonferroni correction for the multiple tests on the chromosome considered. Panels A-J refer to the different chromosomes and to male/female meiosis, as indicated in each plot title.

a more quantitative comparison because those authors do not provide confidence intervals. They also performed two-pathway analyses, estimating *nu* to be between 10 and 21 and *p* between 0 and 0.2. Our values range from 8.8 to 14.6 for *nu* (Figure 1) and from 0.12 to 0.19 for *p* (Figure 2), so the results are qualitatively similar, but again a more detailed comparison cannot be given in the absence of confidence intervals. Also, our confidence intervals for *p* do not include zero, precluding the presence of only the interfering CO-formation pathway in agreement with Copenhaver *et al.* (2002). Another two-pathway analysis was performed (Lam *et al.* 2005) with a larger data set for chromosomes 2M and 4M, each of which bears a nucleolus-organizing region (NORs). Those authors find *p* to be 0.029 [confidence interval (0.003, 0.059)] for chromosome 2M and 0.054 [confidence interval (0.023, 0.097)] for chromosome 4M. These estimates are lower than what we find here, namely 0.14 [confidence interval (0.106, 0.176)] for 2M and 0.12 [confidence interval (0.087, 0.151)] for 4M (Figure 2). The difference might be attributed to several factors: (i) we have 71 SNP markers (they had 17) on chromosome 2 and 44 markers (they had 21) on chromosome 4; (ii) our data set contains >1500 gametes while theirs contains 143 tetrads (tetrad data bring roughly four times more power to the analysis as compared to the same number of gametes, so 1500 gametes here should be compared to the equivalent of ~572 tetrads); and (iii) the plants were not subject to exactly the same growth conditions.

Chromosome-specific effects

Analysis using the two-pathway framework leads to markedly higher P1 interference parameter values in female meiosis for chromosomes 2 and 4. It may be a coincidence, but these are the two short, acrocentric, NOR-bearing chromosomes. Within the NOR regions, one does not have exploitable markers, so no COs are detected there. Such missing data can very well lead to fitting biases, especially if the remaining COs are few as in the female meioses. This effect may thus explain why chromosomes 2 and 4 are *nu* outliers for female meiosis (Figure 1 and bottom triangular part of Table 1); note furthermore that chromosome 4 is a *nu* outlier for male meiosis also (top triangular part of Table 1), giving further credence to the hypothesis that chromosomal architecture and in particular NOR regions are responsible for high P1 interference parameters.

A similar analysis for the parameter *p* reveals that chromosome 4 has a markedly lower proportion of P2 COs than the other chromosomes in male meiosis (Figure 2 and top triangular part of Table 1). The same reasons as above are plausible causes.

Heterogeneity of interference within chromosomes

In a previous study (Drouaud *et al.* 2007) of chromosome 4M of *A. thaliana*, it had been observed by analyzing coefficients of coincidence that the left side of that chromosome had higher effective interference than the right side. Even though our data set comes from a plant panel different from the one of Drouaud *et al.*, the two have similar interference characteristics (see Supporting Information and Figure S5 and Figure S6). In particular, using the coefficient of variation of inter-CO distances, which provides a qualitative measure of effective interference, we find very good agreement between the two data sets (see Figure S6) and that effective interference is stronger on the left side than on the right side of chromosome 4M.

The present work extends intrachromosomal interference comparison to all five chromosomes while also basing such comparisons on model fitting. We compare left *vs.* right arms and central region *vs.* extremities (Table S2). When doing so, we consider only a portion of each chromosome, and so the number of CO events is reduced, thus diminishing statistical power to a large extent. This difficulty explains why our comparison tests are often inconclusive. One striking result of the two-pathway modeling is the generally higher value of the interference strength parameter *nu* as well as *p* in the extremities, compared to the central region, for both male and female meiosis (see Figure 3). In addition to several significant comparisons yielded by analyzing one chromosome at a time, the merged data sets also provided conclusive results. The metacentric chromosomes 1, 3, and 5 together give significant differences between the central region and the extremities for *nu* and *p*. Both parameters are higher in the extremities. This may indicate that, while interference strength (or *nu*) increases toward the extremities, reducing the number of interfering COs, the fraction of noninterfering COs (or *p*) rises in compensation. Whereas toward the central region, the opposite behavior presents itself, with the proportion of noninterfering COs (or *p*) decreasing and interfering COs populating the region, which is more in keeping with the decreased interference (or *nu*). Perhaps this effect is governed by some architectural properties of the chromosomes. These could involve, for example, the level of compaction of the chromatin or mechanical stiffness that certainly plays a role in a number of other phenomena (Kleckner *et al.* 2004). Or the centromere itself could play a role, given that the merged data for the three metacentric chromosomes (1, 3, and 5) gives lower *nu* in the central part than in their extremities in both male and female meiosis.

The heterogeneities hereby demonstrated have never been considered in any interference model. Models to date consider interference to be constant along the chromosome with a single representative parameter (*nu* in the gamma model). Our results suggest that modifications of the gamma model should be considered, for example, by replacing the single value of *nu* for the entire chromosome by a vector of local *nu* values. Unfortunately, this increases greatly the number of parameters to be estimated so that a much larger data set would be required to perform reliable parameter estimation.

There were some instances when it was difficult to obtain estimates for the parameters *nu* and/or *p* when comparing

subparts of the chromosomes, especially for chromosomes 2 and 4. In addition to being the smaller chromosomes, for female meiosis in particular, where interference is higher, the number of COs plummet rapidly in general and further when we look at smaller regions rather than the whole chromosome. Also, the left arm in general is very small, which leads the maximum-likelihood algorithm to allow for large values for interference.

P2-associated hot regions within chromosomes

In light of the evidence for intrachromosomal variations of (i) interference strength and (ii) proportion p of noninterfering COs, we tested within our modeling framework for intervals that may be anomalously hot for P2. This possibility is not considered by any of the currently available interference models, but may be of strong biological relevance. Indeed, one already knows that double-strand breaks mature into crossovers or into noncrossovers in proportions depending on the locus (Mancera *et al.* 2008); such a propensity may extend to the choice of using one CO pathway rather than the other. Differences in the treatment of double-strand breaks may in fact tie in with the different mechanisms that are used for mis-match repair in the two pathways of CO formation (Getz *et al.* 2008).

Performing our tests for male and female meiosis in each interval, we found a number of very strong candidate intervals where P2 COs likely arise at significantly higher frequencies than expected (Figure 5 and Figure S4). This result suggests that not only does interference strength vary along chromosomes, but so does p, the relative contribution of P2 COs to recombination rates. Our genome-wide exploration revealed a heterogeneous pattern; on average some large-scale regions are likely to be hotter than others, but otherwise there do not seem to be any global trends. Clearly, current models, in particular the two-pathway gamma model in which P2 COs are sprinkled uniformly, are simply too crude. A new class of models has to be formulated to incorporate this knowledge. The molecular mechanisms specifying the relative proportions of P1 and P2 COs are only beginning to be unveiled (Crismani *et al.* 2012); one may also speculate that the chromosomal architectural properties can play an important role in determining these proportions. With high-quality data, some of these speculations could provide useful guidance for the modeling.

In summary, our use of crossover formation models to analyze meiotic recombination in *A. thaliana* has led to a genome-wide view of interference. Although some trends are obtainable from the coefficients of coincidence as developed for a single chromosome in (Drouaud *et al.* 2007), the use of the two-pathway modeling provides numerous new insights. For example, there are marked differences in the inferred model parameters when comparing male and female meiosis as well as when comparing different chromosomes. A number of trends emerge such as higher P1 interference strength in female meiosis and higher proportions of P2 COs in male meiosis. Furthermore, we find that the model parameters have clear intrachromosomal variations. For example, the interference strength as well as proportion of noninterfering COs is higher in the extremities compared to the central region for most chromosomes. And, when merging data sets, we find this trend to be significant for male and female meiosis for the metacentric chromosomes 1, 3, and 5. In fact, we reveal genome-wide intrachromosomal heterogeneities arising at scales going from centimorgan distances to the size of a whole chromosome. In particular, the large data set used in this study (taken from Giraut *et al.* 2011) allowed us to present the first genome-wide picture of candidate hot regions for the (noninterfering) P2 pathway. It remains to be seen whether this phenomenon is specific to *Arabidopsis* or more general. Finally, given these strong heterogeneities, it will be necessary to introduce more sophisticated models of crossover formation that allow for such behavior. Just as when going from single- to two-pathway modeling, these improvements will bring deeper biological insights, but because of the increase in their number of parameters, the use of such models will require still larger data sets.

Acknowledgments

We thank Denise Zickler, Raphaël Mercier, and Mathilde Grelon for fruitful discussions of this work and Eric Jenczewski for a critical reading of the manuscript.

Literature Cited

Anderson, L. K., and S. M. Stack, 2002 Meiotic recombination in plants. Curr. Genomics 3: 19.

Argueso, J. L., J. Wanat, Z. Gemici, and E. Alani, 2004 Competing crossover pathways act during meiosis in *Saccharomyces cerevisiae*. Genetics 168: 1805–1816.

Bailey, T. J., 1961 *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

Baudat, F., and B. de Massy, 2007 Regulating double-stranded DNA break repair towards crossover or non-crossover during mammalian meiosis. Chromosome Res. 15: 565–577.

Bishop, D. K., and D. Zickler, 2004 Early decision. Cell 117: 9–15.

Broman, K. W., L. B. Rowe, G. A. Churchill, and K. Paigen, 2002 Crossover interference in the mouse. Genetics 160: 1123–1131.

Copenhaver, G. P., 1998 Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc. Natl. Acad. Sci. USA 95: 247–252.

Copenhaver, G. P., E. A. Housworth, and F. W. Stahl, 2002 Crossover interference in Arabidopsis. Genetics 160: 1631–1639.

Crismani, W., C. Girard, N. Froger, M. Pradillo, J. L. Santos *et al.*, 2012 FANCM limits meiotic crossovers. Science 336: 1588–1590.

Drouaud, J., R. Mercier, L. Chelysheva, A. Bérard, M. Falque *et al.*, 2007 Sex-specific crossover distributions and variations in interference level along *Arabidopsis thaliana* chromosome 4. PLoS Genet. 3: e106.

Falque, M., R. Mercier, C. Mézard, D. de Vienne, and O. C. Martin, 2007 Patterns of recombination and MLH1 foci density along mouse chromosomes: modeling effects of interference and obligate chiasma. Genetics 176: 1453–1467.

Falque, M., L. K. Anderson, S. M. Stack, F. Gauthier, and O. C. Martin, 2009 Two types of meiotic crossovers coexist in maize. Plant Cell 21: 3915–3925.

Foss, E., R. Lande, F. W. Stahl, and C. M. Steinberg, 1993 Chiasma interference as a function of genetic distance. Genetics 133: 681–691.

Froenicke, L., L. K. Anderson, J. Wienberg, and T. Ashley, 2002 Male mouse recombination maps for each autosome identified by chromosome painting. Am. J. Hum. Genet. 71: 1353–1368.

Gauthier, F., O. C. Martin, and M. Falque, 2011 CODA (crossover distribution analyzer): quantitative characterization of crossover position patterns along chromosomes. BMC Bioinformatics 12: 27.

Getz, T. J., S. A. Banse, L. S. Young, A. V. Banse, J. Swanson *et al.*, 2008 Reduced mismatch repair of heteroduplexes reveals "non"-interfering crossing over in wild-type *Saccharomyces cerevisiae*. Genetics 178: 1251–1269.

Giraut, L., M. Falque, J. Drouaud, L. Pereira, O. C. Martin *et al.*, 2011 Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. PLoS Genet. 7: e1002354.

Guillon, H., F. Baudat, C. Grey, R. M. Liskay, and B. de Massy, 2005 Crossover and noncrossover pathways in mouse meiosis. Mol. Cell 20: 563–573.

Higgins, J. D., S. J. Armstrong, F. C. H. Franklin, and G. H. Jones, 2004 The Arabidopsis MutS homolog AtMSH4 functions at an early step in recombination: evidence for two classes of recombination in Arabidopsis. Genes Dev. 18: 2557–2570.

Hollingsworth, N. M., and S. J. Brill, 2004 The Mus81 solution to resolution: generating meiotic crossovers without Holliday junctions. Genes Dev. 18: 117–125.

Housworth, E. A., and F. W. Stahl, 2003 Crossover interference in humans. Am. J. Hum. Genet. 73: 188–197.

Jones, G. H., 1984 The control of chiasma distribution. Symp. Soc. Exp. Biol. 38: 293–320.

Jones, G. H., and F. C. H. Franklin, 2006 Meiotic crossing-over: obligation and interference. Cell 126: 246–248.

Keeney, S., C. N. Giroux, and N. Kleckner, 1997 Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell 88: 375–384.

Kleckner, N., D. Zickler, G. H. Jones, J. Dekker, R. Padmore *et al.*, 2004 A mechanical basis for chromosome function. Proc. Natl. Acad. Sci. USA 101: 12592–12597.

Lam, S. Y., S. R. Horn, S. J. Radford, E. A. Housworth, F. W. Stahl *et al.*, 2005 Crossover interference on nucleolus organizing region-bearing chromosomes in Arabidopsis. Genetics 170: 807–812.

Lenormand, T., and J. Dutheil, 2005 Recombination difference between sexes: a role for haploid selection. PLoS Biol. 3: e63.

Lhuissier, F. G. P., H. H. Offenberg, P. E. Wittich, N. O. E. Vischer, and C. Heyting, 2007 The mismatch repair protein MLH1 marks a subset of strongly interfering crossovers in tomato. Plant Cell 19: 862–876.

Lindsey, J. K., 1995 *Introduction to Applied Statistics: A Modelling Approach*, Oxford University Press, New York.

Malkova, A., J. Swanson, M. German, J. H. McCusker, E. A. Housworth *et al.*, 2004 Gene conversion and crossing over along the 405-kb left arm of *Saccharomyces cerevisiae* chromosome VII. Genetics 168: 49–63.

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

McPeek, M. S., and T. P. Speed, 1995 Modeling interference in genetic recombination. Genetics 139: 1031–1044.

Mercier, R., S. Jolivet, D. Vezon, E. Huppe, L. Chelysheva *et al.*, 2005 Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on MER3, whereas the other one is not. Curr. Biol. 15: 692–701.

Muller, H. J., 1916 The mechanism of crossing-over. Am. Nat. 50: 193–221.

Ott, J., 1999 *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

Page, S. L., and R. S. Hawley, 2004 The genetics and molecular biology of the synaptonemal complex. Annu. Rev. Cell Dev. Biol. 20: 555–558.

Stahl, F. W., H. M. Foss, L. S. Young, R. H. Borts, M. F. F. Abdullah *et al.*, 2004 Does crossover interference count in *Saccharomyces cerevisiae*? Genetics 168: 35–48.

Sturtevant, A. H., 1915 The behaviour of the chromosomes as studied through linkage. Mol. Gen. Genet. 13: 234–287.

Vizir, I. Y., and A. B. Korol, 1990 Sex difference in recombination frequency in Arabidopsis. Heredity 65: 379–383.

Zhao, H., M. S. McPeek, and T. P. Speed, 1995 Statistical analysis of chromatid interference. Genetics 139: 1057–1065.

Communicating editor: A. Houben

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.155549/-/DC1

Hot Regions of Noninterfering Crossovers Coexist with a Nonuniformly Interfering Pathway in *Arabidopsis thaliana*

Sayantani Basu-Roy, Franck Gauthier, Laurène Giraut, Christine Mézard, Matthieu Falque, and Olivier C. Martin

DOI: 10.1534/genetics.113.155549

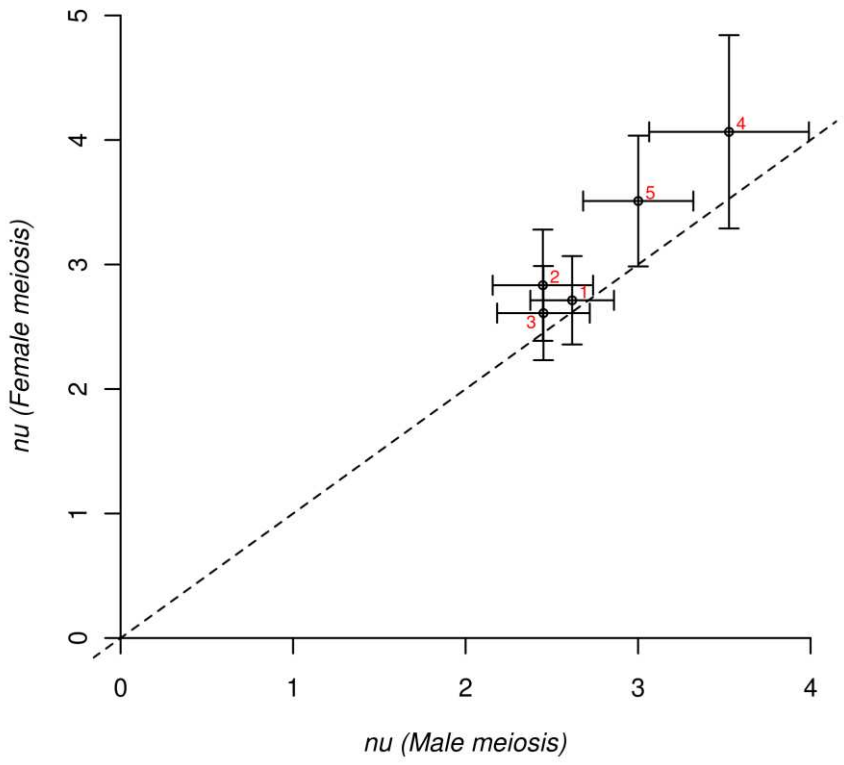


Figure S1 Estimated values of *nu* for the Single-Pathway Gamma model. The *nu* estimates (black circles) for the 5 chromosome pairs (1 to 5, in red) for male (*x*-axis) and female (*y*-axis) meiosis with their 95% confidence intervals (black solid lines). The diagonal (black dashed line) is for *y=x.*

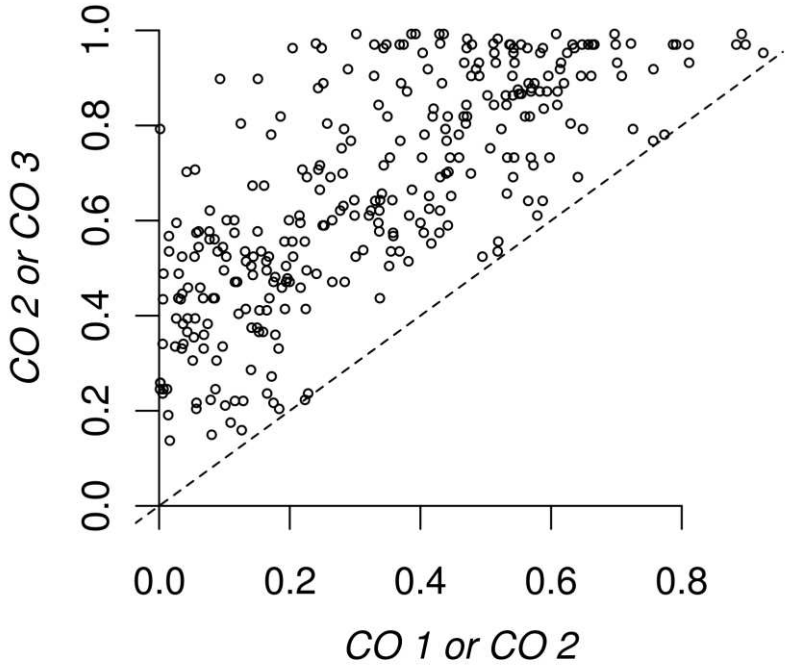


Figure S2 Scatter plot of successive CO positions for all gametes having 3 COs on their chromosome 1 (male backcross population). Each point gives the positions of the pair (*CO 1*, *CO 2*), *i.e.*, (first CO, second CO) or of the pair (*CO 2*, *CO 3*), *i.e.*, (second CO, third CO) where the positions are *genetic* and rescaled to lie in [0,1]. CO positions in adjacent or nearby inter-marker intervals are close to the diagonal (black dashed line, $y=x$).

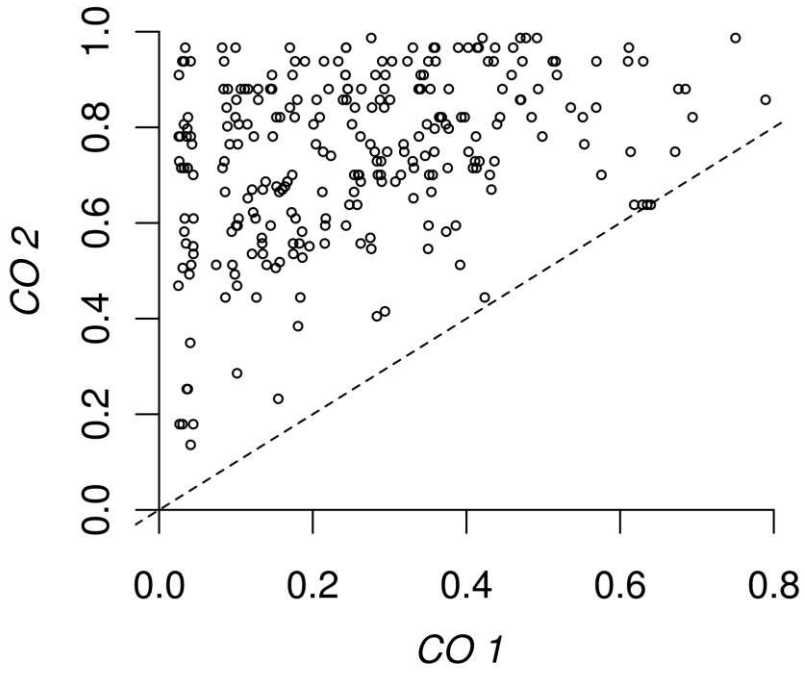


Figure S3 Scatter plot of successive CO positions for all gametes having 2 COs on their chromosome 2 (male backcross population). Each point gives the positions of the pair (*CO 1*, *CO 2*), *i.e.*, (first CO, second CO) where the positions are *genetic* and rescaled to lie in [0,1]. CO positions in adjacent or nearby inter-marker intervals are close to the diagonal (black dashed line, $y=x$).

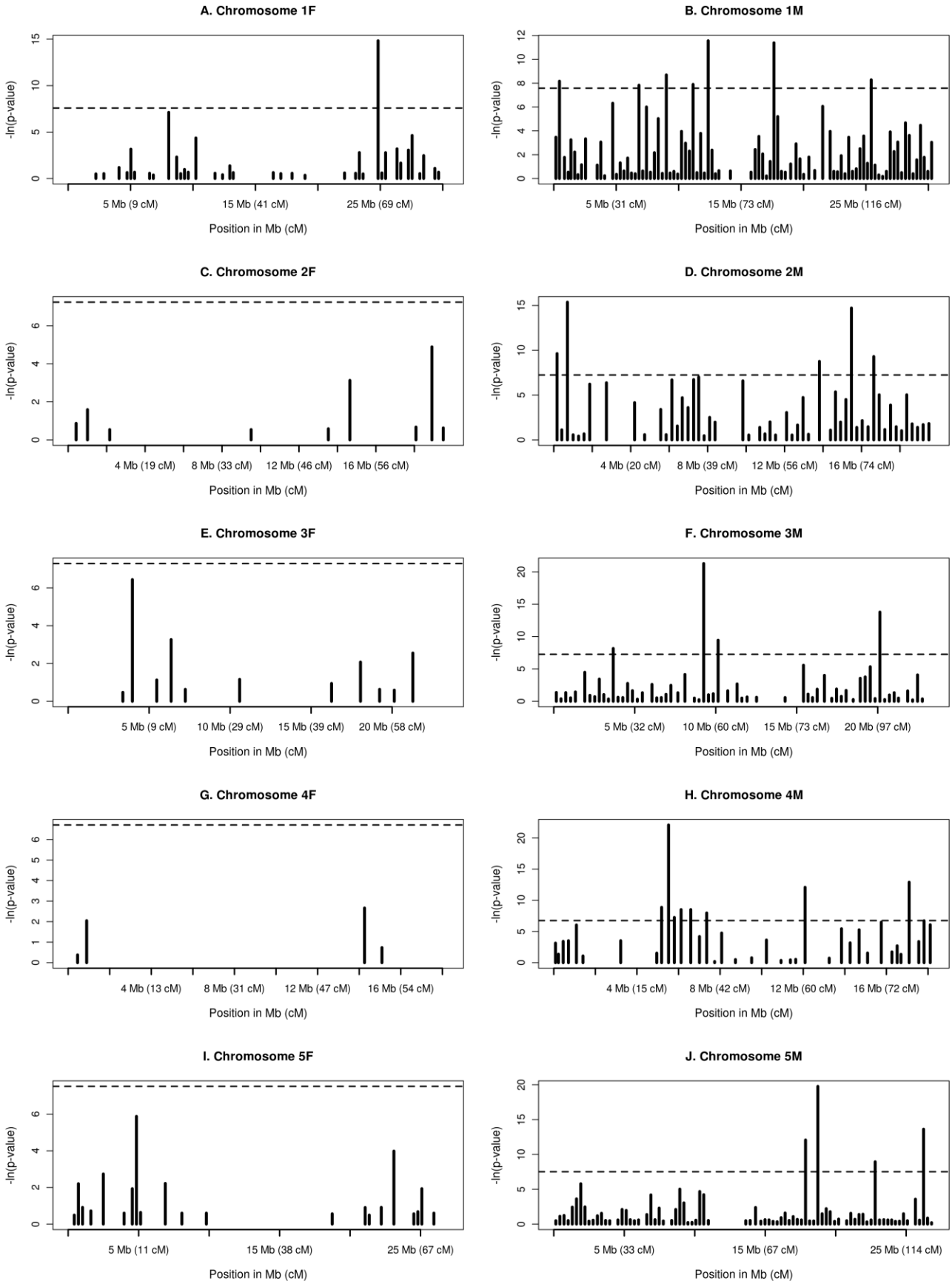

A. Chromosome 1F
-ln(p-value)
0
5
10
15
5 Mb (9 cM)
15 Mb (41 cM)
25 Mb (69 cM)
Position in Mb (cM)
B. Chromosome 1M
-ln(p-value)
0
2
4
6
8
10
12
5 Mb (31 cM)
15 Mb (73 cM)
25 Mb (116 cM)
Position in Mb (cM)
C. Chromosome 2F
-ln(p-value)
0
2
4
6
4 Mb (19 cM)
8 Mb (33 cM)
12 Mb (46 cM)
16 Mb (56 cM)
Position in Mb (cM)
D. Chromosome 2M
-ln(p-value)
0
5
10
15
4 Mb (20 cM)
8 Mb (39 cM)
12 Mb (56 cM)
16 Mb (74 cM)
Position in Mb (cM)
E. Chromosome 3F
-ln(p-value)
0
2
4
6
5 Mb (9 cM)
10 Mb (29 cM)
15 Mb (39 cM)
20 Mb (58 cM)
Position in Mb (cM)
F. Chromosome 3M
-ln(p-value)
0
5
10
15
20
5 Mb (32 cM)
10 Mb (60 cM)
15 Mb (73 cM)
20 Mb (97 cM)
Position in Mb (cM)
G. Chromosome 4F
-ln(p-value)
0
1
2
3
4
5
6
4 Mb (13 cM)
8 Mb (31 cM)
12 Mb (47 cM)
16 Mb (54 cM)
Position in Mb (cM)
H. Chromosome 4M
-ln(p-value)
0
5
10
15
20
4 Mb (15 cM)
8 Mb (42 cM)
12 Mb (60 cM)
16 Mb (72 cM)
Position in Mb (cM)
I. Chromosome 5F
-ln(p-value)
0
2
4
6
5 Mb (11 cM)
15 Mb (38 cM)
25 Mb (67 cM)
Position in Mb (cM)
J. Chromosome 5M
-ln(p-value)
0
5
10
15
20
5 Mb (33 cM)
15 Mb (67 cM)
25 Mb (114 cM)
Position in Mb (cM)

Figure S4 Genome-wide map of interval hotness for the non-interfering P2 pathway. *x*-axis: marker intervals in Mbp (Mb) along the chromosome considered (corresponding genetic positions also provided in centiMorgan or cM). *y*-axis: minus the natural logarithm of the *p*-value of Pearson's chi-square comparison test for that interval. This *p*-value corresponds to the null hypothesis that the two-pathway Gamma model fits the data and in particular that the P2 COs are uniformly distributed in genetic positions along the chromosomes. The dashed horizontal line shows the FWER (family-wise error rate) of 5% when using the Bonferroni correction for the multiple tests on the chromosome considered. Compared to Figure 5 of the main text, the data set has been filtered: all cases where a gamete is doubly recombinant in two adjacent intervals have been removed.

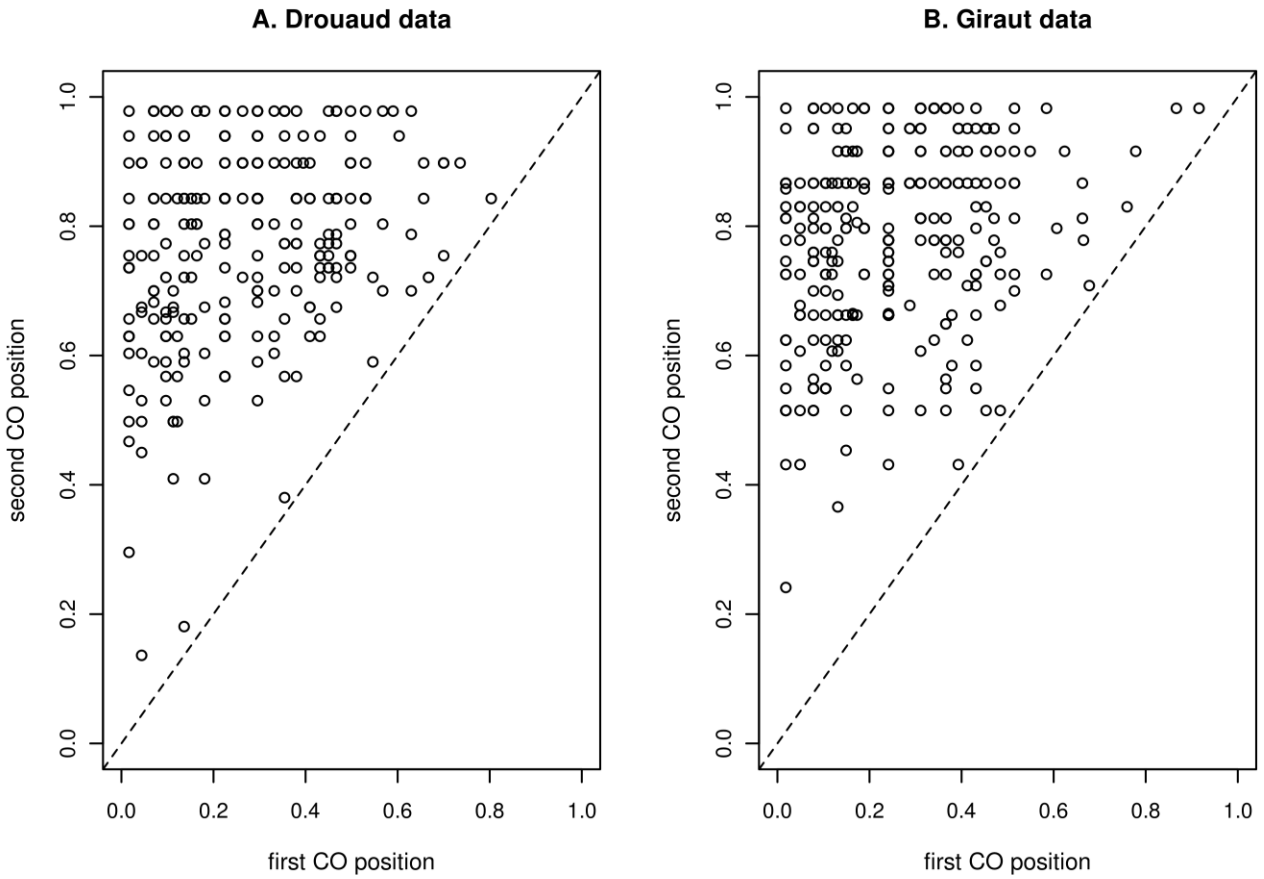


Figure S5 Comparison of data for chromosome 4 during male meiosis between Drouaud *et al.* (2007) and Giraut *et al.* (2011). **A.**Scatter plot of successive CO positions for gametes having 2 COs on their chromosome 4 (male backcross population) in Drouaud *et al.* (2007) data set. **B.**Scatter plot of successive CO positions for gametes having 2 COs on their chromosome 4 (male backcross population) in Giraut *et al.* (2011) data set. Each point gives the positions of the pair (*CO 1*, *CO 2*), *i.e.*, (first CO, second CO) where the positions are *genetic* and rescaled to lie in [0,1]. CO positions in adjacent or nearby inter-marker intervals are close to the diagonal (black dashed line, $y=x$).

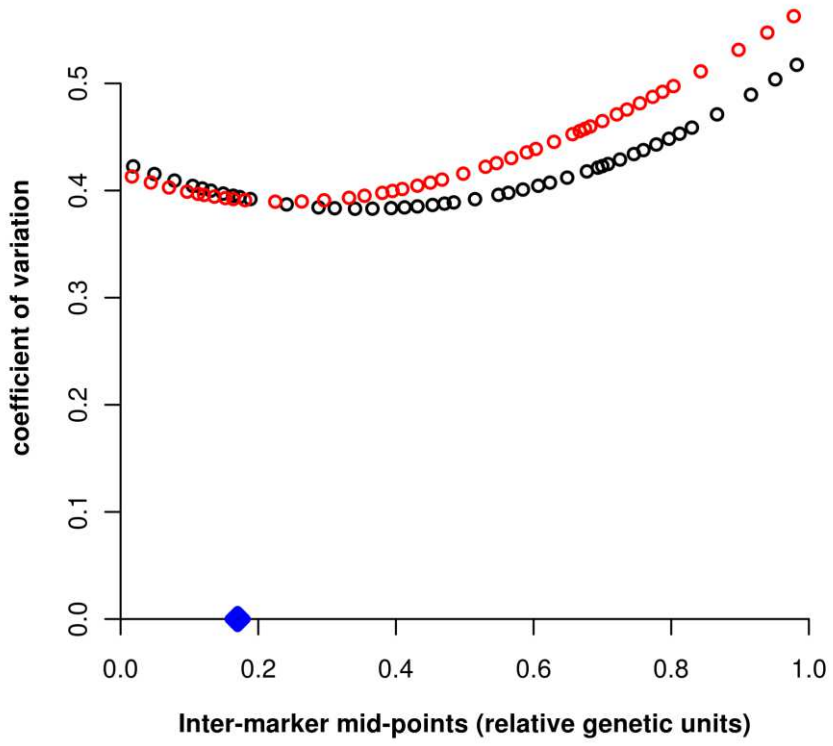


Figure S6 Comparison of data for chromosome 4 during male meiosis between Drouaud *et al.* (2007) and Giraut *et al.* (2011). *x*-axis: mid-points of inter-marker intervals in relative *genetic* position along the length of chromosome 4M. *y*-axis: coefficient of variation all inter-crossover distances, weighted (see detailed method in Supporting Information) using the distance between the mid-point of the inter-marker interval under consideration and the mid-point of each of the inter-CO distances (Drouaud in ***red***; Giraut in ***black***). The position of the centromere is indicated in a ***blue*** dot along the *x*-axis.

File S1

SUPPORTING INFORMATION

In the following text we use the symbol 'ν' instead of '*nu*' to denote the interference strength (Pathway 1, P1) parameter of the Gamma distribution for ease of handling mathematical expressions, the figures maintain the '*nu*' as it is. ν must be strictly positive. $\nu > 1$ corresponds to positive interference effects, $\nu < 1$ corresponds to negative interference effects, while no interference is produced by setting $\nu = 1$.

METHODS

Models

COs are formed via two pathways: the first (P1) is interfering and depends on the ZMM family of genes, while the second (P2) has little or no interference and depends on the Mus81 pathway. Following previous two-pathway modeling studies (Copenhaver *et al.* 2002), we worked with the hypothesis that the two pathways produce COs independently and that P2 has no interference at all. Within such a framework, to simulate meiosis, COs can be produced from each pathway separately and the two corresponding lists merged to get the complete set of CO positions. Simulating P2 is particularly simple: one puts down COs at random with a uniform density in *genetic* space that depends only on the proportion of COs coming from P2. The main challenge is to produce COs in pathway P1; the procedures are nearly identical in the single and two-pathway approaches.

Single-pathway modeling

Framework: To incorporate interference in the P1 pathway, we used the Gamma model (McPeek and Speed 1995). It is formulated at the level of the bivalent, involving the two homologues (each having 2 sister chromatids) for a total of 4 chromatids during the meiosis. This is a statistical model based on a *stationary renewal process* in which successive crossovers are separated by *genetic distances* that are independent, identically distributed random variables following the so-called Gamma *distribution*. The Gamma model has an interference (strength) parameter such that the larger this parameter, the greater the intensity of the "suppressive" effect, *i.e.*, the less likely it is to find two COs close to one another. When working in genetic space, by definition, CO density is 2 per Morgan at the bivalent level, and as a consequence the average distance between adjacent COs is 0.5 Morgan. Because this average distance between adjacent COs is fixed, when interference suppresses short distance events, this effect is accompanied by a reduction in the frequency of long distance events. As a result, increasing interference strength will reduce the coefficient of variation of distances between adjacent COs.

In the Gamma *model*, the parameter quantifying interference strength is called ν: it is the shape parameter of the (Gamma) distribution of distances between adjacent COs, while $2 * (\nu)$, that is, 2ν, is the "rate" of that Gamma distribution on the bivalent when the density of COs is 2 per Morgan. A convenient feature is that the COs generated along the bivalent are directly specified via their *genetic* positions (in Morgans or centiMorgans) for this model.

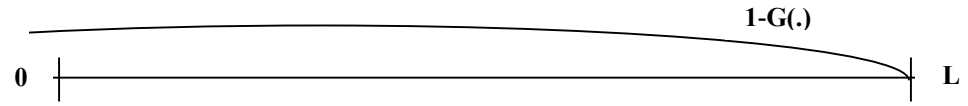
In the present work, the experimental data are not given at the bivalent level: we only have one of the four gametes produced during each meiosis. In this situation one says that the CO data are "thinned". There is evidence that COs arise between non-sister chromatids without any particular bias in favor of any of these non-sisters (Zhao *et al.* 1995; Copenhaver 1998). Because of this evidence, which also simplifies modeling, most work is performed enforcing no bias at all. This assumption is referred to as "no chromatid interference". It is then possible to derive the statistics of the CO patterns at the level of the gametes using the properties at the level of the bivalents even in the absence of knowledge of which chromatids were involved in the genetic exchange.

Likelihood Computation: Since the Gamma model produces COs using a stationary renewal process, it is possible to construct the exact likelihood function for any list of COs assuming a given value of the interference strength ν. This likelihood takes into account the effects of *thinning* (Broman and Weber 2000). Since each backcross is associated with a *different* (independent) meiosis, the likelihood of the whole data set is the product of likelihoods of each backcross plant. Thus one may obtain the "best" value of the interference strength by maximizing this product; and this is what we do in the maximum likelihood method.

The likelihood computation becomes involved mathematically when the chromosome portions under consideration do not form a continuous stretch; this is the situation when estimating the interference parameters for the distal regions of the chromosomes because the central region has been removed and has to be treated as missing data. Now we explicitly address this situation, using the notations in Broman and Weber (2000).

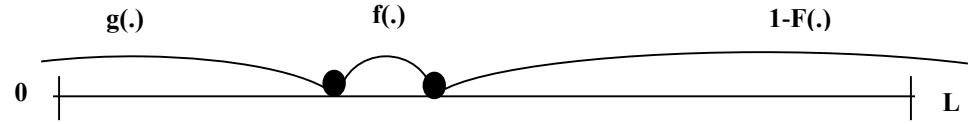
Likelihood of data in distal regions of a chromosome (no thinning): We begin with the COs on the bivalents assumed to be formed by the P1 pathway. If there is at least one CO, let $X_1,\ X_2, \ldots, X_n$ be the n corresponding *genetic* positions. Then the inter-CO distances, $d_i = X_{i+1} - X_i$ are independent random variables that follow a Gamma distribution with shape ν and rate parameter 2ν, so have probability density, $f(d;\ \nu) = e^{-2\nu\, d}\ (2\nu)^{\nu} d^{\nu-1}/\Gamma(\nu)$. We furthermore define $d_0 = X_0$ which has its own distribution corresponding to having an interval size at least as large as d_0. Following Broman and Weber (2000), the density of d_0 is, $g(d;\nu) = 2(1 - F(d;\ \nu))$, where $F(d;\ \nu)$ is the cumulative density function (cdf) of $f(d;\ \nu)$. The likelihood of these bivalents with at least one CO contains a last factor for the last interval of length $d_n = L - X_n$ where L is the *genetic* length of the chromosome; this factor is $(1 - F(d_n;\ \nu))$. If the bivalent has no COs, its likelihood is the probability that the whole chromosome of length L can be placed in an interval of density, $f(d;\ \nu)$. Again following Broman and Weber (2000), this probability can be obtained in closed form. To summarize, the likelihoods for different values of n is given in (Eqn. 1(a), (b), (c)) along with the corresponding diagrams.

$\mathrm{L}(\nu;\ \{d_i\}) = 1 - G(L;\ \nu)$, when $n = 0$ (G is the cdf of g) … 1(a)



$\mathrm{L}(\nu;\ \{d_i\}) = g(d_0;\ \nu)(1 - F(d_1;\ \nu))$, when $n = 1$ … 1(b)

$\mathrm{L}(\nu;\ \{d_i\}) = g(d_0;\ \nu)\{\prod_{i=1}^{n-1} f(\,(d_i;\ \nu))\}\,(1 - F(d_n;\ \nu))$, when $n > 1$ … 1(c)

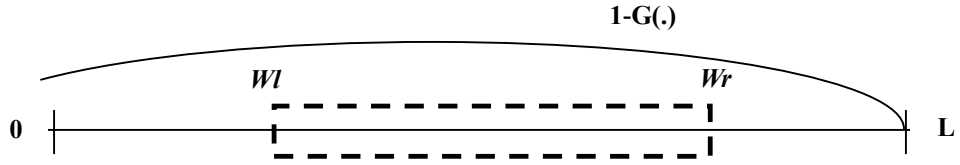


We provide a schematic representation following each case to provide intuition towards the more involved likelihood computations that follow. The straight line from 0 to L denotes the bivalent here and the shaded circles show the CO genetic positions while an arc between adjacent COs (or across the entire structure) indicates the length to be used as argument to the function given there.

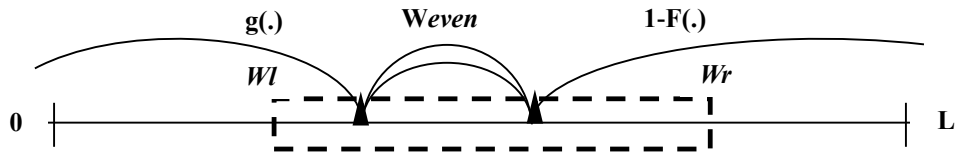
How does this framework for the likelihood generalize when data is available only for the distal parts of the chromosome (e.g., the first and last quarters of the *genetic* length of the bivalent)? In effect, one has "missing data" because the central region is *hidden*, and so we only know whether that region is recombinant or not; in particular, all information on CO positions is lost. CO positions are only known in the two *visible* regions that are disjoint: left (say the fraction 0 – 0.25) and right (say the fraction 0.75 – 1.0) each of which may or may not have COs. Since the recombination data (Giraut *et al.* 2011) has high marker density, we assign COs to the mid-point of the recombinant marker intervals in these *visible* regions. This enables us to retain a continuous picture. Considering the *visible* regions to begin with, there are four possible situations: (i) no COs *visible* to the left or to the right of the *hidden* region, (ii) at least one CO in the left *visible* region but no CO in the right *visible* region, (iii) no CO in the left *visible* region but at least one right *visible* CO and finally, (iv) at least one CO in the left as well as right *visible* regions.

In each of these four cases, the *hidden* region could be recombinant or non-recombinant. Consider first the case when it is non-recombinant. This implies that there is an *even* number of *hidden* COs, that is, 0 or 2 or 4 and so on. When the *hidden* region is recombinant, we know that the number of COs there is *odd* (1 or 3 and so on). The computation is cumbersome due to the uncertain number of these *hidden* COs and we must consider all possibilities compatible with the recombination state of the *hidden* region. When there are no *hidden* COs, the situation is the simplest. With 1 *hidden* CO, one has to calculate a one-dimensional integral. With at least 2 *hidden* COs (whether *even* or *odd*), one has to calculate a two-dimensional integral.

To proceed, we follow the previously described logic (Eqn. (1)) and treat successively the '$n = 0$', '$n = 1$' and '$n > 1$' cases. Now, *n* denotes the total (combining the left and the right *visible* regions) number of *visible* COs. Some new notations need to be introduced. First, let Wl and Wr be the left and right ends of the *hidden* region in *genetic* units (the W stands for "window").Second, if there is one *hidden* CO (respectively at least two, as the case maybe), let the integration variable associated with the first *hidden* CO genetic position be $xCO1$ (respectively the variable for the last *hidden* CO genetic position be $xCO2$).We begin by considering the likelihood for the case $n = 0$ and where the *hidden* region is non-recombinant. The likelihood is the sum of two likelihoods: P_1, when there are no *hidden* COs and P_2, when there are an *even* number of *hidden* COs. In the figures to follow, the dashed outlined box represents the *hidden* region and the shaded triangles are the first or last *hidden* COs while the shaded circles are the *visible* COs. The conventions are the same as before unless otherwise mentioned. When $n = 0$ and the non-recombinant *hidden* region has no *hidden* COs as in Eqn. 1(a) above, we have:



$$\mathrm{P}_1(\nu;\ \{d_i\}) = (1 - G(L;\ \nu)) \qquad \text{… 2(a)}$$



When $n = 0$ and instead there are an *even* number (not 0) of COs in the *hidden* region, we have to integrate over all possible genetic positions of $xCO1$ and $xCO2$. $\mathrm{P}_2(\nu;\ \{d_i\})$ is thus given as:

$$\int_{Wl}^{Wr}\int_{xCO1}^{Wr}[g(xCO1;\ \nu)Weven(xCO2 - xCO1)\{1 - F(L - xCO2;\ \nu)\}\,dxCO1]\,dxCO2$$

… 2(b)

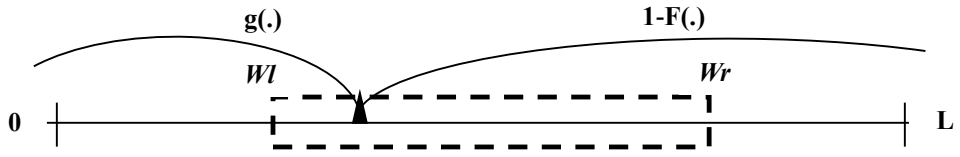
In this expression,

$$Weven(xCO2 - xCO1) = f(xCO2 - xCO1;\ \nu) + f(xCO2 - xCO1;\ 3\nu) + \cdots$$

corresponding to having 0, 2, 4,… COs in addition to $xCO1$ and $xCO2$. In practice, this infinite sum is truncated as and when the required precision is obtained. Note that the rate $2\ \nu$ remains the same for all the densities in Eqn. 2(b) but the shape parameter changes, a convenient feature of the Gamma distribution. Note also that for the *hidden* region, we have introduced a double arc between $xCO1$ and $xCO2$ to stress that there may be additional COs in that interval.

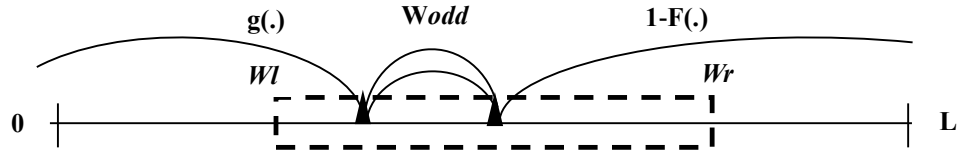
Thus, the likelihood for $n = 0$ and for the window to be non-recombinant is: $\mathrm{L}(\nu;\ \{d_i\}) = \mathrm{P}_1(\nu;\ \{d_i\}) + \mathrm{P}_2(\nu;\ \{d_i\})$.

Continuing with the case, $n = 0$, but with the *hidden* region being recombinant, again there will be two classes of events to consider with an odd number of COs in the *hidden* region: having only one or at least three. The total likelihood will be the sum of the corresponding two probabilities.



For the first class, one has to integrate over all genetic positions of $xCO1$. This gives $\mathrm{P}_1(\nu;\ \{d_i\})$ as:

$$\int_{Wl}^{Wr} g(xCO1;\ \nu)\{1 - F(L - xCO1;\ \nu)dxCO1\} \qquad \text{… 3(a)}$$



Similarly for the second class one has to integrate over all genetic positions of $xCO1$ and $xCO2$. $\mathrm{P}_2(\nu;\ \{d_i\})$ is thus computed as:

$$\int_{Wl}^{Wr}\int_{xCO1}^{Wr}[g(xCO1;\ \nu)Wodd(xCO2 - xCO1)\{1 - F(L - xCO2;\ \nu)\}\, dxCO1]\, dxCO2$$

… 3(b)

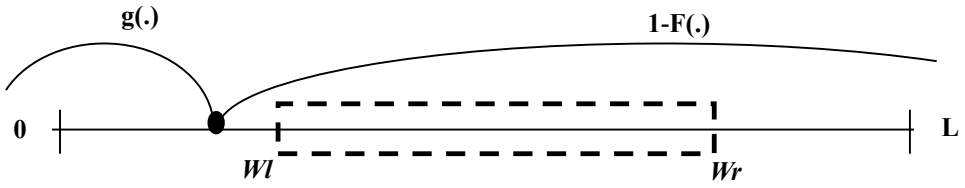
In this expression, in direct analogy with what we saw before, one has

$$Wodd(xCO2 - xCO1) = f(xCO2 - xCO1;\ 2\nu) + f(xCO2 - xCO1;\ 4\nu) + \cdots$$

corresponding to having 1, 3, 5, … COs in addition to $xCO1$ and $xCO2$. Just as for $Weven$ the series is truncated and the rate $2\,\nu$, remains the same for all the densities in 3(b) while the shape parameter varies.

This concludes the $n = 0$ case.

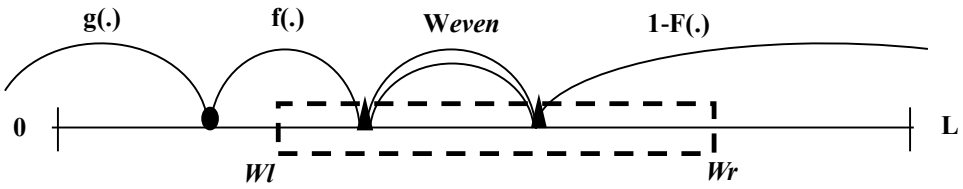
Similar calculations apply to the cases $n > 0$. Let us illustrate the calculations when there is only one *visible* CO and it is to the left of the *hidden* region. We will have the same sub-cases as before with regard to the *hidden* region which can be recombinant or non-recombinant.



Consider first the non-recombinant case. When the *hidden* region has no COs, we have a first likelihood as illustrated in the figure above:

$$\mathrm{P}_1(\nu;\,\{d_i\}) \;=\; g(d_0;\,\nu)(1 \;-\; F(d_1;\,\nu)) \qquad \text{… 4(a)}$$

Note that $d_1 = L - d_0$.

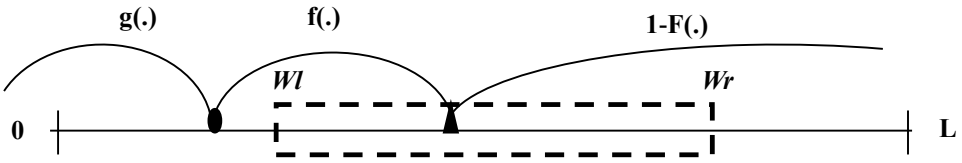


If instead the *hidden* region has COs, they must be an *even* number; the corresponding likelihood requires performing a double integral in analogy to the $n = 0$ case. This leads to the formula for $\mathrm{P}_2(\nu;\,\{d_i\})$:

$$g(d_0;\,\nu)\left[\int_{Wl}^{Wr}\int_{xCO1}^{Wr}[f(xCO1 - d_0)Weven(xCO2 - xCO1)\{1 - F(L - xCO2;\,\nu)\}dxCO1]\,dxCO2\right]$$

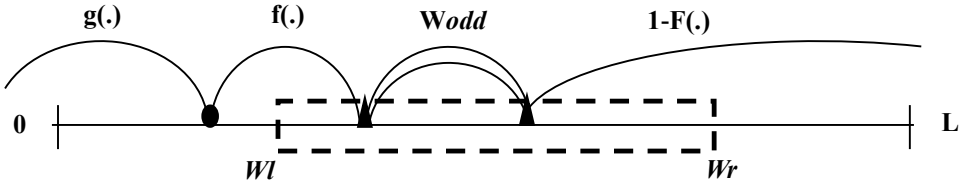
… 4(b)

where $Weven(xCO2 - xCO1)$ is as defined in Eqn. 2(b).



The same logic applies to the situation when the *hidden* region is recombinant. The case of exactly one *hidden* CO leads to $\mathrm{P}_1(\nu;\ \{d_i\})$, given by:

$$g(d_0;\ \nu)\int_{Wl}^{Wr} f(xCO1 - d_0;\ \nu)\{1 - F(L - xCO1;\ \nu)dxCO1\} \qquad \text{… 5(a)}$$



When there at least 3 COs, we see that the term, $\mathrm{P}_2(\nu;\ \{d_i\})$ is the product of $g(d_0;\ \nu)$ and the double integral:

$$\left[\int_{Wl}^{Wr}\int_{xCO1}^{Wr}[f(xCO1 - d_0;\ \nu)Wodd(xCO2 - xCO1)\{1 - F(L - xCO2;\ \nu)\}dxCO1]\,dxCO2\right]$$

… 5(b)

where $Wodd(xCO2 - xCO1)$ is as defined previously in Eqn. 3(b)

This case of there being exactly one *visible* left CO can be generalized very simply to any number of *visible* left COs (there is simply an additional factor $f(.)$ for every arc between successive *visible* left COs).

Similarly if the bivalent has only right *visible* COs (none on the left), then the relevant diagrams are the mirror images of the ones presented above for *visible* left COs. The likelihood formulae then follow straightforwardly.

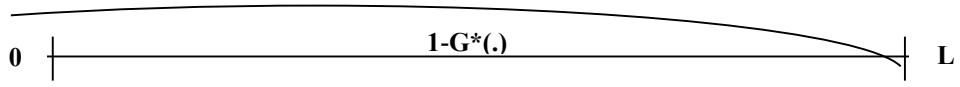
Finally when we have *visible* COs to the left as well as to the right, one has that (i) each of the end intervals contributes a factor $(1 - F(.))$ and to one of these we associate an additional factor 2 (analogous to using $g(.)$ on one side and $(1 - F(.))$ on the other), (ii) all simple arcs connecting COs carry a factor $f(.)$, (iii) all double arcs carry a factor $Weven$ (respectively $Wodd$) when the interval $[Wl, Wr]$ is non-recombinant (respectively recombinant), (iv) when there is one hidden CO, the integration range is $Wl \leq xCO1 \leq Wr$ and if

there are at least two COs, the integration range is $Wl \leq xCO1 \leq xCO2 \leq Wr$.

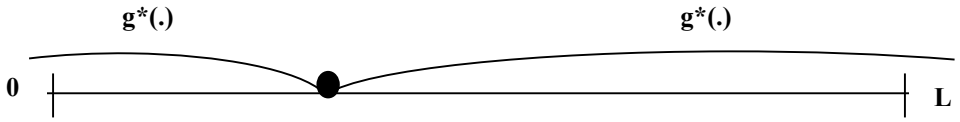
Likelihood of data in the distal regions of a chromosome (with thinning): Having treated what happens on the bivalent, we have to consider now the likelihood of a gamete, one of the four products of the bivalent (*cf* . sub-head, 'Fit of the Gamma Model' in the Materials & Methods section of Broman and Weber 2000). In the absence of chromatid interference, each CO on the bivalent has a 50% probability of being passed to the considered gamete. This process of removing COs randomly is called "thinning" and leads to modified formulae for the likelihoods. Broman and Weber derived these formulae when the whole chromosome is visible. The purpose of this section is to generalize these formulae to the case where the chromosome contains a *hidden* region. For completeness, we begin by recalling the results of Broman and Weber.

Consider a random meiotic product, that is, a gamete. The COs of the bivalent have been *thinned* independently with probability 1/2. If the inter-crossover *genetic* distances along the gamete are $l_1, l_2, \ldots$, then they have "similar" statistical properties as the d_i introduced above. Specifically, the l_i are independent and $l_1, l_2, \ldots$ have density, $f^*(l;\, \nu) = \sum_{k=1}^{\infty}(1/2)^k f_k(l;\, \nu)$, where $f_k(l;\, \nu)$ is a Gamma density with rate 2ν and shape $k\nu$. Furthermore, the density of l_0 (the distance between the left end of the chromosome and the first CO) is $g^*(l;\, \nu) = 1 - F^*(l;\, \nu)$, where $F^*(l;\, \nu)$ is the cdf of $f^*(l;\, \nu)$ (details are given in Broman and Weber 2000). In effect, the calculation of the likelihood of a gamete after *thinning* is as without *thinning* but where f^* replaces f in the derivation. Let us go over the different cases. Suppose the gamete has q COs after *thinning* and let the inter-crossover distances be $l_1, l_2, \ldots, l_{q-1}$.We furthermore set l_0 to be the genetic position of the first CO and l_q to be the length of the interval between the end of the chromosome and CO number q. As before, L is the *genetic* length of the gamete. Then depending on q, the likelihood of the gamete will be:

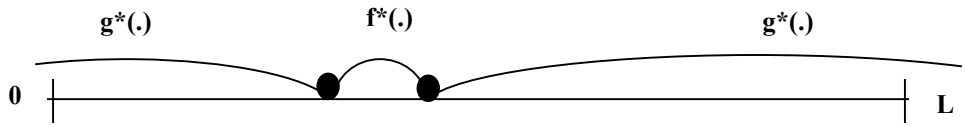
$L(\nu; \{l_i\}) = (1 - G^*(l_0;\, \nu))$, when $q = 0$ (G^* is the cdf of g^*) … (6a)



$L(\nu;\, \{l_i\}) = g^*(\, l_0;\, \nu)\, g^*(\, l_1;\, \nu)$, when $q = 1$ … (6b)



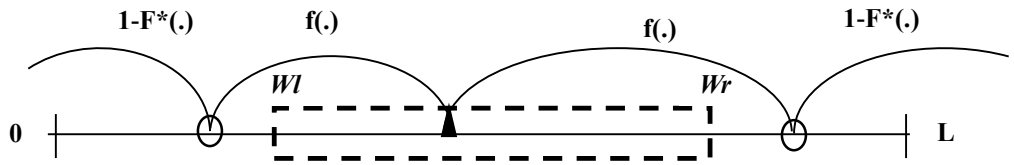
$L(\nu;\, \{l_i\}) = g^*(l_0;\, \nu) \prod_{i=1}^{q-1} f^*(l_i;\, \nu))\, g^*(l_q;\, \nu)$, when $q > 1$ … (6c)



Given these formulae for the whole chromosome, we now generalize to the case of discontinuous regions as before. The notations are

the same as before. Thus, we have *q visible* COs 'after' *thinning*. Apart from the *visible* CO positions 'after' thinning, we must consider the possibility of there being a CO 'before' *thinning* in each *visible* inter-marker interval. So let there be m markers (namely, $M(1), M(2), M(3), \ldots, M(m)$) along the gamete in the *visible* region of which ml are on the left of the hidden region and mr are on the right. Finally, as far as the *hidden* region is concerned, no *thinning* is explicitly carried out there. In fact all COs we refer to in the *hidden* region are 'before' *thinning*.

Just as in the description without *thinning*, let us begin our explanations when there are no COs ('after' *thinning*) in the *visible* regions. There are 2 associated probabilities which are important to understand: (a) the probability P_a that the *hidden* region is recombinant 'after' *thinning*; note that this event can occur only if there is at least one *hidden* CO (*not* zero) 'before' *thinning* and (b) the probability P_b that the *hidden* region is non-recombinant 'after' *thinning*: this event is possible no matter what the number of *hidden* COs (*including* zero) 'before' *thinning*. These probabilities depend on 3 sub-cases in which the *hidden* region has zero, one or at least two COs ('before' *thinning*). If these 3 sub-cases have probabilities P_0, P_1 and P_2, then: $P_a = (P_1 + P_2)/2$ and $P_a = P_0 + (P_1 + P_2)/2$. For the calculation of P_0, P_1 and P_2, there are 4 sub-sub-cases, each based on whether there are or not any (left or right) *visible* CO(s) 'before' *thinning*. In the representation below, the empty circles represent COs in the *visible* region (they could be either absent altogether or present only on one side) present 'before' *thinning* but which are *thinned* out, that is disappear 'after' *thinning*. The shaded triangles indicate *hidden* COs 'before' *thinning* (that may or may not be *thinned* out).



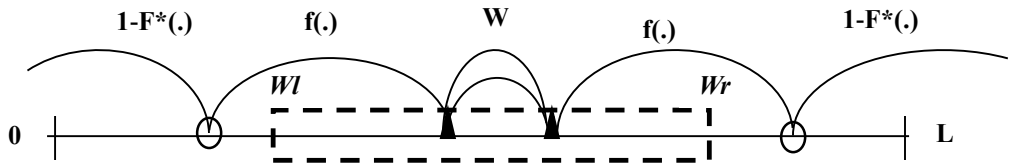
To be specific, consider the sub-case of the probability P_1 and its sub-sub-case when there are *visible* COs to be *thinned* out on both sides of the *hidden* region. Let x be the genetic position of the last left *visible* CO to be *thinned* and y that of the first right *visible* CO to be *thinned*. The probability of this sub-sub-case can be written as follows:

$$\int_0^{Wl} dx[\, \int_{Wl}^{Wr} dxCO1\{\, \int_{Wr}^{L} dy(\, 1 - F^*(x;\, \nu))\, f(xCO1 - x;\ \nu)\, f(y - xCO1;\, \nu)(1 - F^*(L - y;\, \nu))\, \}]$$

… (7)

In practice, the integrations involving x and y are replaced by summations (using the inter-marker interval sizes and their mid-points) which give good approximations in this case as the markers are densely placed along the length of the gamete. The other sub-sub-cases are treated analogously. If there is no *visible* CO to be *thinned* on one side (or both), then there is one (or two) less variable(s) to integrate over and the product $(1 - F^*(.))\, f(.)$ is replaced by $(1 - F(.))$.

Moving to the sub-case where at least two COs arise in the *hidden* region, P_2 breaks down into the



same sub-sub-cases. The difference compared to the P_1 case lies in W. Here the term W is the average of $Weven$ $(xCO2 - xCO1)$ and $Wodd$ $(xCO2 - xCO1)$ since we allow for *even* or *odd* number of *hidden* COs in this case. Each possibility (even or odd) will lead to a recombinant (or non-recombinant) *hidden* region with probability ½. So here it is necessary to integrate over $xCO1$ and $xCO2$, which gives us a quadruple integral:

$$\int_0^{Wl} dx[\int_{Wl}^{Wr} dxCO1 \int_{xCO1}^{Wr} dxCO2 \{\int_{Wr}^{L} dy\ (1 - F^*(x;\ \nu))\ f(xCO1 - x;\ \nu)\ W(xCO2 - xCO1;\ \nu)\ f(y - xCO2)\ (1 - F^*(L - y;\ \nu))\}] \quad \ldots (8)$$

Again, as before, here we actually do a double integral, the integrals over *x* and *y* are evaluated by summing over the inter-marker interval mid-points. Also, in case there is no *visible* CO to *thinned* on either side (or not at all), then, the product $(1 - F^*(.))\ f(.)$ changes to only $1 - F(.)$ and there is one (or two) less variable(s) to integrate for. This completes the cases when there are no *visible* COs 'after' *thinning* ($q = 0$).

Lastly, there may be *visible* COs 'after' *thinning* ($q > 0$). If so, consider the last left *visible* CO when there is one and the first right *visible* CO when it exists. These COs contribute to the q *visible* COs ('after' *thinning*). The computation of the corresponding likelihoods generalizes what we have seen before when there were left or right COs (without the *hidden* region) to include the sub-sub-cases based on the *thinned* that were just discussed when $q = 0$. Clearly, these *thinned* COs have a modified integration range. For instance, when integrating over x (the genetic position of the last *thinned* CO on the left), the lower bound is the genetic position of the last left *visible* CO. The same comment applies to *y* which must not go beyond the position of the first right *visible* CO. The expressions for the *hidden* region as well as the number of variables to be integrated over follow the same rules as for $q = 0$.

Fitting Procedures, Confidence intervals: Adjusting the model parameter to fit the experimental data requires searching for the maximum likelihood score L. We do this search via a "hill-climbing" procedure (Gauthier *et al.* 2011) as follows. Given a point in the search space (specified by the model's parameter value or values), we calculate L_0, the likelihood there, and also the score at neighboring points obtained by increasing and decreasing the parameter by a characteristic step. We also consider a trial point using polynomial interpolation and measure the score there. If at least one of these neighbors has likelihood larger than L_0, we move to the

point with the highest score. If none of the neighboring points has a score larger than L_0, we reduce the size of the characteristic step while also trying an intermediate point based on polynomial interpolation. The procedure is iterated and the position converges to a (local) maximum of L. We perform checks to verify that this maximum is in fact a global maximum, attainable from different initial positions. This fitting leads to the value (ν) of the interference strength which gives the maximum likelihood in the parametric space. The confidence intervals are computed using Fisher's Information matrix.

Two-pathway modeling via sprinkling

Framework: To include COs from the P2 pathway, the sprinkling procedure (Copenhaver *et al.* 2002) is used, which is simply the superposition of the non-interfering COs, generated using the Poisson distribution, onto the ones from the interfering P1 pathway. The fraction of P2 COs is thus an additional parameter to the one in the single-pathway modeling; on the bivalent, it is the proportion of non-interfering COs, that is p. When $p > 0$, the density of P1 COs is no longer 2 per Morgan, but $2 * (1 - p)$. Comparing to the procedure for producing P1 COs in the single pathway model, we see that the shape parameter of the (Gamma) distribution of distances between adjacent COs is still ν, but the rate parameter is changed from $2 * (\nu)$ to $2 * (\nu) * (1 - p)$, that is from 2ν to $2\nu(1 - p)$.

Likelihood Computation: As described above for the single pathway case, the process and logic remain the same except that in all likelihoods, the shape parameter changes (from 2ν to $2\nu(1 - p)$).

Fitting Procedures, Confidence Intervals: The principles used here are the same as those mentioned for single pathway modeling; the main difference is that the likelihood (Falque *et al.* 2009) L now is a function of two variables so the parameter space to search is two-dimensional. Again the hill-climbing algorithm was used (Gauthier *et al.* 2011). And the Fisher Information matrix was computed to obtain the confidence intervals.

Statistical analyses and comparison tests

Comparing two datasets (separate chromosomes or different regions of one chromosome): We performed three levels of comparisons to examine the variation in interference between and within chromosomes. Using mainly the two-pathway model, we compared the interference strength: (1) between male and female meiosis, (2) between the different chromosomes but for a given sex, and (3) between segments of the same chromosome, looking at variations in interference values between the two arms of a chromosome and also between the central and distal regions of a chromosome.

To make these comparisons, we tested the null hypothesis (H_0) that the means (ν or p here) of the populations, from which the two

samples under consideration have been drawn, are equal. Here the population variances are unequal which requires changing the formulae for the test statistic as well as the accompanying degrees of freedom for this modified two-sample t-test; for this we follow Welch's t-test. Let the sample means be $\overline{X1}$ and $\overline{X2}$, the respective standard deviations be $S1, S2$ and lastly the sample sizes be $n1$ and $n2$. Then under the H_0: $\mu 1 = \mu 2$ (μ are population means: interpreted here as ν or p), the statistic $t(observed)$ follows a t-distribution with degrees of freedom df. Their corresponding formulae are given as follows:

$$t(observed) = (\overline{X1} - \overline{X2}) \Big/ \sqrt{\frac{S1^2}{n1} + \frac{S2^2}{n2}}$$

$$df = \left(\frac{S1^2}{n1} + \frac{S2^2}{n2}\right)^2 \Big/ \left(\frac{S1^4}{n1^2(n1-1)} + \frac{S2^4}{n2^2(n2-1)}\right)$$

Finally, the function *pt(.)*, of the R statistical package, was used to compute the two-sided *P*-value at the 5% level of significance. When the *P*-value is significant, we reject the null hypothesis. This indicates that the difference between the samples under consideration may not be considered negligible.

Comparing two discrete distributions (simulated and experimental): We performed another kind of comparison between simulated datasets and experimental data. This was necessary when we tested for "hot" regions specific for the non-interfering (P2) pathway. The starting point is the distribution of COs given there is a CO in a "reference" interval under consideration. For each interval spanning adjacent markers (assuming at least 1 gamete has a CO in this interval), the frequency of COs in each of the *other* intervals is computed, using gametes which have multiple COs (treating separately those with 2 and 3 COs). The analogous frequencies are obtained in the context of the model's predictions. Specifically, the model's behavior is obtained from simulated data, generated using the *simdata* option of CODA (Gauthier *et al.* 2011) with ν and p set to the values obtained from fitting the experimental data. Then, we tested for a significant difference between the expected (simulated or theoretical) and observed (experimental) frequency distributions of CO occurrences for each inter-marker interval at a time. We used the Pearson's chi-square test function (Lindsey 2004) within the R statistical software, *chisq.test(.)* to test the null hypothesis that the observed distribution is not statistically different from the observed one, separately for all intervals. Furthermore, to obtain a global view of the results, we merged the values from the 2 COs and 3 COs cases by taking the sum of the corresponding chi-square values (for intervals having data for the 2COs and 3 COs cases). The new *P*-values were computed by the R function, *pchisq(.)*. And if an interval had data only for 2 COs or only for 3 COs, we retained the previous chi-square and *P*-values.

Coefficient of variation of inter-CO distances along chromosomes: Given a list of distances between adjacent COs, let μ and σ be the associated mean and standard deviation. Then the coefficient of variation CV is defined as the ratio σ/μ. Low values of CV correspond to high levels of effective interference. To allow CV to be position dependent so as to get a picture of effective interference strength *along a*

chromosome, we apply a weight to each element of the list (a distance between adjacent COs) as follows. Let X be the current position where the local CV is to be measured and let Y be the midpoint of the 2 COs under consideration. Then the weight assigned to the associated CO-CO distance is taken as $\exp(-10(X-Y)^2)$. Then the X-dependent mean μ and standard deviation σ defining CV(X) are simply computed using these weights. Explicitly, we have:

$$\mu(X) = \frac{\sum_i w_i d_i}{\sum_i w_i} \text{ and } \sigma^2(X) = \frac{\sum_i w_i (d_i - \mu(X))^2}{\sum_i w_i}$$

where i is the index of the element in the list of distances d_i between adjacent COs. The corresponding CV(X) curves are displayed in Fig S6. Note that the weights mimic a sliding window without having the drawback of a discontinuous behavior in X.

Comparison to Drouaud *et al*. results

In a previous study (Drouaud *et al.* 2007) of chromosome 4M of *A. thaliana*, variability in interference strength was revealed by using three-point coefficients of coincidence (or "C3"s that use 2 adjacent windows). Those authors found that the effective interference was higher on the left side than on the right side. We reach the same conclusion from analysing our 4M data that was generated independently of that of Drouaud *et al*. The overall similarity of these data sets is demonstrated in Figure S5 which shows the distribution of COs in 4M gametes having exactly two COs, for the two data sets. (Figure S5 (A)uses the data of Drouaud *et al*. and thus provides quantitatively the same information as given in Figure 4 (A) of their paper). At a more mathematical level, we have extracted from each data set the distribution of distances between adjacent COs from which we determine the associated coefficient of variation (CV) as a function of position along the chromosome (see previous paragraph). In Figure S6 we display the associated curve for each data set: the two curves are qualitatively similar. Furthermore, CV is clearly lower on the left than on the right, pointing to a higher effective interference on the left side than on the right side, in agreement with the conclusion of Drouaud *et al*. Nevertheless, the "C3" values in Drouaud *et al.*'s work also suggested that interference strength was weakest in the middle, not all the way on the right end (see Figure 6 (A) in Drouaud *et al.* 2007 and Figure S6 here). Such differences can arise because "C3"s and CV are sensitive to different aspects of interference. Indeed, CV incorporates *all* adjacent inter-CO distances for each inter-marker interval to compute the CV value for that interval, whereas the "C3" values in Drouaud *et al*. depend on recombination within windows surrounding the point of interest. Thus the "C3" approach is less sensitive to nearby COs because only rarely do these lead to recombination in both of the associated windows. To illustrate this difference, consider again the COs displayed in Figure S5. On the far left side, the data of Drouaud *et al*. has a few close-by COs but nevertheless the "C3" values are tiny there, showing that indeed such close-by COs do not lead to double recombination events contributing to "C3".

REFERENCES

Broman K. W., J. L. Weber, 2000 Characterization of human crossover interference. Am J Hum Genet 66: 1911–1926.

Copenhaver, G. P., 1998 Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc. Natl. Acad. Sci. 95: 247–252.

Copenhaver, G. P., E. A. Housworth, F. W. Stahl, 2002 Crossover interference in Arabidopsis. Genetics 160: 1631–1639.

Falque, M., L. K. Anderson, S. M. Stack, F. Gauthier, O. C. Martin, 2009 Two types of meiotic crossovers coexist in maize. Plant Cell 21: 3915–3925.

Gauthier, F., O. C. Martin, M. Falque, 2011 CODA (crossover distribution analyzer): quantitative characterization of crossover position patterns along chromosomes. BMC Bioinformatics 12: 27.

Lindsey, J. K., 1995 *Introduction to Applied Statistics: a Modelling Approach*. Oxford University Press, New York.

McPeek, M. S., and T. P. Speed, 1995 Modeling interference in genetic recombination. Genetics 139: 1031–1044.

Zhao H., M. S. McPeek, T. P. Speed, 1995 Statistical analysis of chromatid inference. Genetics 139: 1057-1065.

Table S1 Results (*P*-values) for Single-Pathway Gamma Comparisons. Comparisons between (a) male and female meioses for the same chromosome: diagonal values (enclosed in ***bold***) and (b) different chromosomes' male meiosis (upper triangle values) and female meiosis (lower triangle values) separately. The *P*-values were computed for the null hypothesis that the 2 meioses under consideration are associated with the same value of *nu (or ν)*.

	1M	**2M**	**3M**	**4M**	**5M**	**Male**
1F	-	-	-	*nu** 4M	-	**1M**
2F	-	-	-	*nu** 4M	*nu** 5M	**2M**
3F	-	-	-	*nu** 4M	*nu** 5M	**3M**
4F	*nu** 4F	*nu** 4F	*nu** 4F	-	-	**4M**
5F	*nu** 5F	-	*nu** 5F	-	-	**5M**
Female	**1F**	**2F**	**3F**	**4F**	**5F**	

Note: '' indicates P-value is significant at the 5% threshold. For every significant comparison, the chromosome indicated has the higher nu.*

Table S2 Results (*P*-values) for Intra-chromosomal Comparisons. Comparisons are (a) between the left (L) and right (R) arms (L/R) of each chromosome and (b) between the central (C) region (fraction 0.25 to 0.75 of the *genetic* length) and the distal regions or "extremities" (E, union of 0 to 0.25 and 0.75 to 1 of the *genetic* length) of the chromosome (C/E), for Gamma models: Single as well as Two-pathway. The *P*-values were computed for the null hypothesis that the 2 chromosome regions under consideration are associated with the same value of *nu (or ν)* or p (separately).

	Single pathway		**Two-pathway**			
	L/R	**C/E**	**L/R**		**C/E**	
	Nu	***nu***	***nu***	***p***	***nu***	***P***
1M	-	* C	-	-	-	* E
2M	-	-	* R	* R	-	* E
3M	-	-	-	* L	-	-
4M	-	-	-	-	* E	-
5M	-	* E	-	-	-	* C
1F	-	-	-	-	-	-
2F	-	* C	-	-	-	-
3F	-	-	-	-	* E	-
4F	* R	-	-	-	-	-
5F	-	* C	-	-	* E	* E

Note: '' indicates P-value is significant at the 5% threshold. For every significant comparison, the portion of the chromosome indicated (L/R for left/right or C/E for centre/extremities) is the one having the higher nu or p.*

PART II

CHARACTERIZATION OF THE TWO PATHWAYS IN TOMATO

Chapter 5

Between interfering and not…

METHODS & RESULTS OBTAINED ON THE TOMATO DATA

This chapter describes the tests used to investigate the (lack-of) interactions between the two crossover formation pathways. These tests have been applied to the tomato data produced by the Anderson group. The main results have subsequently been highlighted. To our knowledge, this unique dataset allowed testing of some very interesting hypotheses for the first time. The latest manuscript on the analysis follows this chapter.

Due to similar arm length ratios, the synaptonemal complexes (SCs) of certain chromosomes could not be differentiated from each other. Specifically, it was not possible to distinguish chromosome 5 from chromosome 12 and chromosome 7 from chromosome 9. Thus, we pooled the corresponding SCs together.

Statistical Analyses

Non independence between interfering and non-interfering crossovers

A fundamental assumption in the Gamma two-pathway model is that the two crossover classes form independently. Hence, the non-interfering crossovers are 'sprinkled' (Copenhaver *et al.* 2002) over the interfering ones to obtain the combined distribution of crossovers along a chromosome. Here we propose three tests to examine whether this assumption indeed holds. The null hypothesis H_0 is that the pathways P1 and P2 are independent.

First test of independence

This test probes whether the *number* of MLH1-positive (pos) and MLH1-negative (neg) foci on SCs are independent of each other. Let us assume that an SChas m 'pos' foci and n 'neg' foci. Let the associated experimental probabilities be $p^+(m)$ and $p^-(n)$ while the joint experimental probability of these two events is $P(m,n)$. Under the hypothesis H_0 of independence of the two random variables, we will have,

$P(m,n) = p^{+}(m) * p^{-}(n)$, accompanied with quantifiable statistical noise.

To inquire if the observed joint probability distribution deviates significantly from the one expected under H_0 , we rely on the Fisher exact test. What we have at our disposal is the experimental contingency table enlisting the number of SCs having a specific pair of (m,n). Define $M(m)$ as the number of SCs with m 'pos' LNs and $N(n)$ as the number of SCs exhibiting n 'neg' LNs. From the contingency table, $M(m)$ (respectively $N(n)$) will be the sum of SC numbers in row 'm' (respectively column 'n'). In the Fisher exact test, all possible tables giving the same $M(m)$ and $N(n)$ as the experimental table are taken into consideration to test the hypothesis. If $\{T(m,n)\}$ represent a table entries, the likelihood of the table is given by,

$$L(T) = \prod_m \{M(m)!\} \prod_n \{N(n)!\} \,/\, [N! \prod_{m,n} \{T(m,n)!\}]$$

where, $k!$ is the factorial of k and N is the total SC number. Using this expression, the likelihoods of all contingency tables under the null hypothesis are computed. And if the equivalent experimental likelihood is $L(T_E)$, then the *p*-value for H_0 is obtained as $\sum_{L(T)\leq L(T_E)} L(T)$. This procedure is attained by applying the R function $fisher.test(.)$ and it is performed for each chromosome separately from the data.

A global version of this test was also done using all chromosomes. However, because the combinatorics explodes, the R function $fisher.test(.)$ is not adapted to carry out the Fisher exact test in case of multiple experimental contingency tables (or T_Es). So the test at this level was conducted by sampling instead of enumerating the rather numerous *possible* contingency tables. For each SC, we shuffle around the values of n so that they are no longer correlated with the value of m leading to one table per chromosome per shuffle. The likelihood for each table is derived using the above-mentioned formula and a product is taken across chromosomes for the total probability. Further such shuffling is repeated 10^5 times. The *p*-value is given by the fraction of the times a shuffle leads to a likelihood at least as small as the experimental likelihood. This test expectedly gives a much smaller *p*-value than when the test is done for individual chromosomes.

Second test of independence

In this test, we compare the observed and shuffled distance distributions in micrometers between each 'pos' foci and all other 'neg' foci along an SC using the chi-square test. The shuffling is done as follows. For every studied SC, we shuffle the 'pos' LN positions keeping the 'neg' positions fixed. The shuffling done here varies from the previous in that here the foci positions are also considered, not just their number. The null *p*-value is found via a chi-square test to compare the observed and shuffled distributions.

Third test of Independence

The third test is the same as the second one except here the foci positions are considered in genetic units (cM) instead of cytogenetic units (micrometers).

Finer Points

The procedures derived for the above tests are valid even if P1 and P2 are heterogeneous along the chromosome. Secondly, there is a subtle issue to be considered about the distance distributions between the 'pos' and 'neg' foci. In a dataset, if Y denotes the genetic position of a 'pos' LN along an SC of length L, then the possible values for the distance between this position and a 'neg' focus lie not in $[0, L]$ but in $[0, Y]$ for LNs to the left and $[0, L - Y]$ for LNs to the right. This constraint must be taken into account when computing the density of CO distances.

Inter-crossover distance distributions are represented as histograms. We begin by defining the range $[0, L]$ divided into 10 equal bins. For each 'pos' LN, introduce bins for the left range $[0, Y]$ and bins for the right range and $[0, L - Y]$. (Generally the last bin will not be complete, but we allow for such bins.) Now for each bin and each SC, our first counter i_1 is incremented by one if for the SC under consideration, it is possible to have an inter-crossover distance in that bin. And if there are 1, 2, … 'neg' LNs in a bin, then we add that number to our counter i_2. Finally, the ratio of i_2 to i_1 gives the frequency of 'neg' LNs in every bin.

Role of an intervening P2 crossover on interactions among flanking P1 crossovers

If the above series of tests rejects the null hypothesis of independence, then we can conclude that there is P1-P2 crossover interference. There could then be variouspostulates on the nature of the interaction between pathwaysTo scrutinize this possibility, we could consider that P2 crossovers appear only after the formation of those from P1, resulting in P1-P1 crossover interference not being influenced by P2 crossovers. But the late-forming P2 crossovers would then locate themselves away from the P1 ones to ensure P1-P2 interference while the P1 crossovers would remain unaware of the P2 crossovers. Another possibility could be that the two crossover classes arise simultaneously and each kind is subjected to some interference from the other. Further P2 crossovers might neutralize the interference from adjacent P1 crossovers, acting as a screen and they in turn could emit their own (weakly) interfering signal; in such a scenario, one could enrich the frequency of two P1 crossovers flanking one P2 crossover.

These possibilities lead us to ask if the presence of a P2 crossover in between two P1 COs reduces the mean P1-P1 crossover distance. Let Δ represent the difference between the average P1-P1 crossover distances when there is and when there is not an interposed P2 crossover. The experimental value is determined from the data while the "theoretical" value is obtained by shuffling as before (One value per shuffle for 10^5 shuffles). This shuffling gives the test statistic

to examine the null hypothesis that the P1-P1 crossover interference is immune to the presence of P2 crossovers. The hypothesis is rejected if a two-sided *p*-value, given by the fraction of shuffles which result in a Δ higher than the experimental Δ, is significant.

Highlights of our Results

Applying the first test of independence, the null hypothesis that the two pathways are independent was rejected for chromosomes 3, for the pooled set of 5 with 12, as well as 10 with 11. In addition, when all chromosomes were pooled together, the *p*-value was highly significant, suggesting that the numbers of the two kinds of foci, MLH1-positive and MLH1-negative, are *not* independent. Using linear regression it was seen that this relationship was explained with a negative slope for most chromosomes. This implies that an SC with more MLH1-positive LNs will tend to have less MLH1-negative foci.

The second as well as third test for examining independence, where the distances between the two crossover classes are considered, yield *p*-values which are significant for almost all chromosomes whether the cytogenetic distance (second test) in considered or the genetic distance (third test). And when all chromosomes are pooled together, both tests give significant results again, indicating dependence among the crossover realizations in the two pathways.

When examining if an interposed P2 crossover between two P1 crossovers leaves P1-P1 interference unchanged, no significant *p*-value was found. Thus one cannot provide evidence that P2 crossovers intrude on the interference among P1 crossovers.

Reference

Copenhaver G. P., Housworth E.A., Stahl F.W. (2002) Crossover interference in *Arabidopsis*. Genetics 160: 1631-39.

THE MANUSCRIPT ON TOMATO

Anderson *et al.*

Combined Fluorescence and Electron Microscopic Imaging Unveils the Specific Properties of Two Classes of Meiotic Crossovers

Thesis Page Numbers **110** through **130** (21 pages)

Kindly consider the page numbers in the top left corner.

Combined fluorescence and electron microscopic imaging unveils the specific properties of two classes of meiotic crossovers

Lorinda K. Anderson, Leslie D. Lohmiller, Xiaomin Tang, D. Boyd Hammond, Lauren Javernick, Lindsay Shearer, Sayantani Basu-Roy, Olivier C. Martin, and Matthieu Falque

While light microscopy (LM) is efficient for evaluating cellular function using fluorescent probes, it lacks the high resolution capability of electron microscopy (EM). Here, we maximize the advantages of both methods by imaging the same sample of pachytene meiotic chromosomes first by immunofluorescence LM and then by EM to localize crossovers (COs). COs are formed *via* either of two molecularly distinct pathways, requiring different protein complexes. Our imaging technique allows, for the first time, the identification of the pathway origin for all COs in meiotic cells from wild-type tomato, unveiling specific characteristics of each pathway. In particular, we observe an enrichment in COs from one pathway in pericentric heterochromatin, a feature potentially exploitable in plant breeding. Furthermore, we demonstrate cross-talk between the two pathways, showing that they are not independent. Our technique is general and can provide functional insights of proteins and cellular structures at a very fine scale.

Eukaryotic sexual reproduction involves meiosis in which the genetic material is halved after two specialized cell divisions to produce genetically balanced gametes. Crossing over

between homologous chromosomes is critical in meiosis because the mechanical link generated allows for the accurate separation of homologs during the first division. Furthermore, crossovers drive genetic diversification by shuffling alleles within homologs and thereby play a key role in adaptation and evolution. To detect meiotic COs we image two cytological structures, synaptonemal complexes (SCs) that link each pair of homologous chromosomes during pachytene and recombination nodules (RNs) that are ellipsoidal structures along SCs [1]. Each RN marks a CO site [2-4]. RNs are too small (50-100 nm) to be observed by LM, but they can be readily visualized by EM, particularly in two-dimensional spreads of SCs [3].

Meiotic recombination is initiated by the formation of DNA double-strand breaks which may then follow one of several repair pathways, two of which result in COs [5-8]. Pathway 1 (P1) produces class I COs that show interference (defined as a lowered probability of nearby COs compared to random placement [9, 10]). Antibodies to MLH1 protein have been used as immunofluorescent probes to map class I COs on SCs (*e.g.*[4, 11]). Pathway 2 (P2) produces class II COs that are thought to be non-interfering. No satisfactory immunolabeling is currently available for these COs. The P1 pathway produces the majority of COs, and the P2 pathway accounts for 5-30% of COs [4, 12, 13]. Because individual COs cannot be assigned to a particular pathway using marker segregation, little is known in wild-type organisms about the properties of each pathway or whether they interact.

Here we developed a new approach utilizing SC spreads from wild-type tomato (*Solanum lycopersicum*, n = 12) to identify the pathway of origin for each CO in a meiotic cell: we superimposed the immunofluorescent LM image of an SC spread showing MLH1 foci locations onto the EM image of the same SC spread showing RN locations. Unlike other correlative LM and EM techniques that have been applied to whole cells [14], here the same preparation is

visualized without further complicated treatments (such as immunogold labeling and/or cell sectioning), allowing us to unveil directly the separate properties of the P1 and P2 pathways.

We analyzed nuclei in which each of the twelve chromosomes had at least one MLH1 focus to make sure that any absence of MLH1 immunolabeling was not due to defects in spreading [4]. RNs were first identified by EM based on their association with SCs, size, shape, and staining intensity [3], and then RNs were classified as MLH1-positive or MLH1–negative using the corresponding LM image (**Fig. 1**; **Supplementary Fig. 1**). From 162 nuclei, we were able to map RN positions on 1882 individual chromosomes that were identified based on relative length and kinetochore position (= arm ratio**; Supplementary Tables 1 and 2**). We observed a mean of 18.8 RNs per nucleus, with 82% MLH1-positive RNs and 18% MLH1-negative RNs (see also **Supplementary Fig. 2**). A large fraction of these SCs (1419/1882 = 75%) had no MLH1-negative RN (**Supplementary Table 3**).

MLH1-positive RNs were more dense and significantly larger in both length (parallel to the SC) and width (transverse to the SC) than MLH1-negative RNs (**Supplementary Fig. 3; Supplementary Table 4**). These differences in RN size and density may indicate that protein components other than MLH1 differ between the two RN types, as would be expected from molecular and genetic studies regarding proteins involved in the two different pathways of crossing over [6-8].

One of the most obvious CO trends observed in many organisms is a high level of crossing over in distal euchromatin and a low level in pericentric heterochromatin, whether assayed using linkage maps, RNs, or MLH1 foci (*e.g.*, [4, 11, 15, 16]). Similarly, we find for tomato that most RNs arise in euchromatin with only 5% of all RNs (=162/2953) located in the roughly 30% of SC length in pericentric heterochromatin (**Fig. 2**; **Supplementary Fig. 4**;

Supplementary Table 5). We find, however, that the presence of pericentric heterochromatin does not affect both RN types equally since only 3% of all MLH1-positive RNs were located in the pericentric heterochromatin compared to nearly 17% of all MLH1-negative RNs. It is notable that COs from both pathways occur in the pericentric heterochromatin, albeit at much different rates. Modifying local recombination frequencies in pericentric regions is an important challenge for plant breeding (*e.g.* for positional cloning). Knowing that heterochromatic regions are more prone to class II COs may lead to new strategies using P2 to promote COs in those regions, *e.g.* by acting on the *FANCM* gene [17].

When considered separately, the distributions of MLH1-positive and MLH1-negative RNs were significantly different for all chromosomes (Kolmogorov-Smirnov p-value < 0.04) except *11*. In addition to the heterochromatin differences, MLH1-negative RNs were disproportionately observed on the short arms of most chromosomes while MLH1-positive RNs were slightly overrepresented on the long arms of most chromosomes (**Fig. 2**; **Supplementary Fig. 4).** This difference is particularly striking for chromosome *6* (short arm = 23% of chromosome length) in which 54% of all MLH1-negative RNs were observed in the short arm compared to only 7% of all MLH1-positive RNs. The relationship between RN frequency and SC length was also different between MLH1-positive RNs (significant positive correlation) and MLH1-negative RNs (no significant correlation) (**Supplementary Fig. 5**; see also [4]). While no one has been able to definitively map class II COs in a wild-type organism before, our results are broadly consistent with results from Arabidopsis *msh4* (P1) mutants showing that when class I crossovers were blocked, chromosome length and the frequency of residual chiasmata (from class II COs) were not related [18]. The tomato results also fit with the model of Housworth and Stahl[19] in which the mean number of class I COs varies according to chromosome length while

the mean number of class II COs remains the same for all chromosomes, regardless of chromosome length.

We also investigated whether synaptic initiation patterns correlated with the patterns of the two RN types. In tomato, synapsis usually begins in distal euchromatin, and pericentric heterochromatin synapses last [20]. Using chromosome arm-specific DNA probes for markers in distal euchromatin, we observed that synapsis was initiated in long arms as much as two times more often than expected based on arm length (**Supplementary Table 6**; **Supplementary Fig. 6**). Combining the observed distributions of RN types with these synaptic patterns indicates that MLH1-positive RNs are more likely to be in earlier synapsing parts of chromosomes, and MLH1-negative RNs are more likely to be in the later synapsing parts (short arms and pericentric heterochromatin).

For all chromosomes, MLH1-positive RNs showed significant interference ($\nu \sim 7$; **Supplementary Table 7**; **Supplementary Fig. 7**; see **Online Methods**). This is expected since MLH1 marks interfering class I COs in several animals (*e.g.*, [21, 22]) and in tomato [4]. In contrast, we found no significant interference between MLH1-negative RNs for any chromosome ($\nu \sim 1$; **Supplementary Table 7**). This was expected since MLH1-negative RNs mark class II COs that were previously suggested to be non-interfering in mutants defective for the P1 pathway (*e.g.*, [12, 13, 23]. What is novel here is the demonstration that class II COs are not interfering in a wild-type background in which both pathways are intact, and that this lack of interference extends over the whole genome.

We next asked if there is any cross-talk between the two pathways. First, we tested the hypothesis that the distribution of the *numbers* of MLH1-negative RNs per SC is the same whether the SCs have one, two, or three MLH1-positive RNs. This hypothesis was rejected at the

5% level for chromosomes *3*, *5/12*, *10*, and *11* (Fisher's exact test, all *p*-values provided in **Supplementary Table 8**). Pooling all chromosomes together (see **Online Methods**), the *p*-value was less than 10^{-5}, indicating that the number of MLH1-negative RNs is *not* independent from the number of MLH1-positive RNs. Furthermore, these two numbers were negatively correlated (*p*-value<0.05 for four chromosomes; details in **Supplementary Table 9**).

Second, we tested whether SC *distances* (measured in µm of SC length) between MLH1-positive and MLH1-negative RNs present on the same SC (**Fig. 1**) were compatible with no interference. For most chromosomes, we found a lower-than-expected frequency of MLH1-negative RNs close to MLH1-positive RNs, indicating interference between the two RN types (**Supplementary Table 10**; **Supplementary Fig. 8**). This result was confirmed when all chromosomes were pooled together (**Fig. 3a**) and also when using genetic distance (**Supplementary Figs. 9-10**). We performed a similar analysis for all pairs of MLH1-positive RNs. As expected, highly significant interference between MLH1-positive RNs was found for all individual chromosomes and for the pooled set of chromosomes (**Fig. 3b**; **Supplementary Figs. 11-13**) using both SC distance (all *p*-values below 10^{-63}) and genetic distance (all *p*-values below 10^{-45}). Comparison of **Figure 3a with Figure 3b** (and **Supplementary Figs. 10** and **13** for genetic distances) shows that interference is much stronger between two MLH1-positive RNs (extending out to about 10 µm) than between MLH1-positive RNs and MLH1-negative RNs (extending out to about 6 µm). Note that the much smaller *p*-values obtained with SC distances are consistent with the hypothesis that the "repulsion" between COs depends on physical distances rather than genetic distances, as is assumed in structural models for interference [24, 25].

The question of interference between class I and class II COs had not been amenable to investigation before our study because it was not possible to reliably specify individual class I

and class II COs on the same chromosome. Our combined LM and EM method presented here overcomes this hurdle and explicitly demonstrates cross-talk between the two CO pathways. Previously, it had been speculated that class I and class II COs might interfere because of a signal produced by class I COs [26]. While we cannot confirm that there is such a signal, our results demonstrate that inter-pathway interference does exist.

Given that cross-talk between the two CO pathways is inhibitory (the numbers of COs in each pathway are negatively correlated and there is interference between nearby COs), it seems likely that some kind of balance exists between the pathways. Consequently, one would expect suppression or absence of P1 to allow more double-strand breaks to be repaired as class II COs (at the expense of non-COs) and compensate to some degree for the reduction in class I COs. Evidence of such balance comes from increased numbers of MLH1 foci when class II crossovers are blocked in *mus81*-/- male mice[27]. Future experiments using our combined LM and EM method with mutants in P1 and/or P2 CO pathways should make it possible to further test these hypotheses.

Overall, our simultaneous mapping of MLH1-positive RNs (class I COs) and MLH1-negative RNs (class II Cos) has provided access for the first time to the characteristics of CO formation pathways in a wild-type organism. This breakthrough reveals that (1) the patterns of distribution of class I and class II COs differ along the chromosomes, (2) class II COs do not interfere with each other, and (3) there is interference between class I and class II COs. Application of our technique to other challenges requiring fine scale co-localizations should provide analogous advances in our understanding of cellular processes.

1. Page,S.L. & Hawley,R.S. The genetics and molecular biology of the synaptonemal complex. *Annu. Rev. Cell Dev. Biol.* **20**, 525-558 (2004).

2. Marcon,E. & Moens,P. Mlh1p and Mlh3p localize to precociously induced chiasmata of okadaic acid treated mouse spermatocytes. *Genetics* **165**, 2283-2287 (2003).

3. Anderson,L.K. & Stack,S.M. Recombination nodules in plants. *Cytogenet. Genome Res.* **109**, 198-204 (2005).

4. Lhuissier,F.G.P., Offenberg,H.H., Wittich,P.E., Vischer,N.O.E., & Heyting,C. The mismatch repair protein MLH1 marks a subset of strongly interfering crossovers in tomato. *Plant Cell* **19**, 862-876 (2007).

5. Goldfarb,T. & Lichten,M. Frequent and efficient use of the sister chromatid for DNA double-strand break repair during budding yeast meiosis. *PLoS Biology* **8**, e1000520 (2010).

6. Schwartz,E.K. & Heyer,W.D. Processing of joint molecule intermediates by structure-selective endonucleases during homologous recombination in eukaryotes. *Chromosoma* **120**, 109-127 (2011).

7. Zakharyevich,K., Tang,S., Ma,Y., & Hunter,N. Delineation of joint molecule resolution pathways in meiosis identifies a crossover-specific resolvase. *Cell* **149**, 334-347 (2012).

8. De Muyt,A. *et al.* BLM helicase ortholog Sgs1 is a central regulator of meiotic recombination intermediate metabolism. *Mol. Cell* **46**, 43-53 (2012).

9. Sturtevant,A.H. The behavior of the chromosomes as studied through linkage. *Zeitschrift fur Induktive Abstammungs-Vererbungslehre* **13**, 234-287 (1915).

10. Berchowitz,L.E. & Copenhaver,G.P. Genetic interference: Don't stand so close to me. *Curr. Genom.* **11**, 91-102 (2010).

11. Froenicke,L., Anderson,L.K., Weinberg,J., & Ashley,T. Male mouse recombination maps for each autosome identified by chromosome painting. *Am. J. Hum. Genet.* **71**, 1353-1368 (2002).

12. de los Santos,T. *et al.* The Mus81/Mms4 endonuclease acts independently of double-Holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics* **164**, 81-94 (2003).

13. Berchowitz,L.E., Francis,K.E., Bey,A.L., & Copenhaver,G.P. The role of *AtMUS81* in interference-insensitive crossovers in A. thaliana. *PLoS Genetics* **3**, 1355-1364 (2007).

14. Mironov,A.A. & Beznoussenko,G.V. Correlative microscopy. Methods in Cell Biology. [113], 209-255. 2013. Elsevier.

15. The Tomato Genome Consortium The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641 (2012).

16. Sherman,J.D. & Stack,S.M. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). *Genetics* **141**, 683-708 (1995).

17. Crismani,W. *et al.* FANCM limits meiotic crossovers. *Science* **336**, 1588-1590 (2012).

18. Higgins,J.D., Armstrong,S.J., Franklin,F.C.H., & Jones,G.H. The *Arabidopsis MutS* homolog *AtMSH4* functions at an early step in recombination: evidence for two classes of recombination in *Arabidopsis*. *Genes Devel.* **18**, 2557-2570 (2004).

19. Housworth,E.A. & Stahl,F.W. Is there variation in crossover interference levels among chromosomes from human males? *Genetics* **183**, 403-405 (2009).

20. Stack,S.M. & Anderson,L.K. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. II. Synapsis in *Lycopersicon esculentum* (tomato). *Am. J. Bot.* **73**, 264-281 (1986).

21. Falque,M., Mercier,R., Mezard,C., de Vienne,D., & Martin,O.C. Patterns of recombination and MLH1 foci density along mouse chromosomes: Modeling effects of interference and obligate chiasma. *Genetics* **176**, 1453-1467 (2007).

22. Borodin,P.M. *et al.* Recombination map of the common shrew, *Sorex araneus* (Eulipotyphla, Mammalia). Genetics 178, 621-632. 2008.

23. Osman,K., Higgins,J.D., Sanchez-Moran,E., Armstrong,S.J., & Franklin,F.C.H. Pathways to meiotic recombination in *Arabidopsis thaliana*. *New Phytol.* **190,** 523-544 (2011).

24. Petkov,P.M., Broman,K.W., Szatkiewicz,J.P., & Paigen,K. Crossover interference underlies sex differences in recombination rates. *Trends Genet.* **23**, 539-542 (2007).

25. Kleckner,N. *et al.* A mechanical basis for chromosome function. *PNAS USA* **101**, 12592-12597 (2004).

26. Fung,J.C., Rockmill,B., Odell,M., & Roeder,G.S. Imposition of crossover interference through the nonrandom distribution of synapsis initiation complexes. *Cell* **116**, 795-802 (2004).

27. Holloway,J.K., Booth,J., Edelmann,W., McGowan,C.H., & Cohen,P.E. MUS81 generates a subset of MLH1-MLH3-independent crossovers in mammalian meiosis. *PLoS Genetics* **4**, e1000186 (2008).

……

ONLINE METHODS

SC spreading, immunolabeling, microscopy and measuring: Primary microsporocytes in the pachytene of Prophase I from the highly inbred cherry tomato line (LA4444) were used to prepare SC spreads as described [27, 28]. In brief, protoplasts were prepared by enzymatic digestion, then the protoplasts were mixed with a dilute detergent solution on microscope slides that had been coated with 0.6% Falcon plastic. After nuclei swelled and burst to produce SC spreads, the slides were air-dried for two hours, then stored at -80C for up to two months before immunolabeling. Slides were thawed at room temperature, then SC spreads were treated for 10 min with DNase I (1 µg/ml in 10 mM Tris-HCl, 2.5 mM $MgCl_2$, 5.0 mM $CaCl_2$, pH 7.5), washed, then immunolabeled with affinity-purified chicken anti-SlSMC1 diluted 1:50 and affinity-purified rabbit anti-SlMLH1 diluted 1:50 [27, 29, 30] followed by goat anti-chicken Dylight 649 and goat anti-rabbit AlexaFluor 488 (both from Jackson Labs and diluted 1:500). Primary antibody incubations were performed overnight at 4˚C, and secondary antibody incubations were performed at 37˚C for one hour. Coverglasses were mounted with Vectashield (Vector Labs) and visualized by fluorescence light microscopy (LM) with a 100X Plan-Apo objective with an adjustable iris and a Leica DM5000 microscope equipped for both phase contrast and fluorescence microscopy with FITC and TRITC filter cubes with zero pixel shift. Using an automated stage (Prior), coordinates for spreads in which all twelve SCs were well-separated and each SC had at least one MLH1 focus [28] were recorded. Red and green signals for each spread

were captured individually using a cooled Hamamatsu monochrome 1344X1044 pixel camera and IP Lab software (ver 4). Because some MLH1 foci were quite small and dim, we used long exposure times (1.5 – 2.5 sec. depending on the slide) to be sure that even dim foci would be imaged. Images for each spread were artificially colored using IPLab and merged using Photoshop CS2. After images were captured, the coverglass of each slide was removed carefully by placing the slides in a horizontal Coplin jar containing an aqueous solution of 0.05% Triton X-100, 150 mM NaCl, and 50 mM Tris (pH 7.5; TBST). After 10-15 min, the coverglasses were loose enough that each could be carefully removed from the slide. Slides were then placed into TBST in an upright Coplin jar and washed 3 x 5 min with TBST. Slides were given a final wash of deionized water, placed into a slide rack, air-dried, and covered with a 25X50 mm coverglass. Then, using previously recorded coordinates, each spread was re-imaged by phase microscopy using a 40X objective and 2X magnifier inserted before the camera. Coverglasses were gently removed to avoid damaging the plastic surface, and the slides were stained with phosphotungstic acid [29]. After drying, phase microscopy was used to visualize SC spreads, and 50-mesh copper finder grids (GC50 – Ted Pella) were placed over as many of the previously imaged SC spreads as possible. Previously captured phase images aided identification of SC spreads for grid placement. The plastic around the grids was scored with a sharp probe, and then the plastic with grids was lifted from the surface of the slide using 1% hydrofluoric acid followed by 5% acetic acid. Grids with attached plastic were floated onto a distilled water surface where the grids were picked up using small strips of Parafilm [29]. After air drying, the grids were scanned by phase contrast microscopy, and the grid coordinates of each SC spread were recorded for examination by EM. SC spreads were photographed at a magnification of 3000 (generally requiring 3 – 6 images per SC spread) using a JEOL 2000 electron microscope, and the negatives were scanned

at 800 dpi using an Epson Perfection V700 Photo scanner. A montage of each SC spread was assembled either by hand using separate layers in Adobe Photoshop CS2 or automatically using Microsoft ICE with Planar Motion 1 setting that ensures a rigid scale with no warping when assembling montages. For each EM montage, RNs were identified based on their association with SCs, size (around 100 nm), staining (more dense than lateral elements), and indefinite margins (in contrast to stain precipitate that has sharp margins)[30], and the positions of RNs and kinetochores were marked. The corresponding LM fluorescence image was pasted as a separate layer over the 800 dpi EM montage, and the size and rotation of the LM image was adjusted as necessary to fit over the EM image. Enhancement of EM or LM images to increase contrast was done uniformly to the entire LM or EM image using the Photoshop levels command. While matching the sizes of the LM and EM images was straightforward, matching the rotation was more difficult because the position of each SC spread on each grid as well as the rotation of the grid in the EM holder differed. Therefore, the LM and EM images often were not perfectly matched for every SC. Consequently, each SC was analyzed individually by precisely aligning the red (SC) and green (MLH1) combined fluorescence image over the EM image of the same SC. Then, the MLH1 fluorescent signal at each previously identified RN position was assessed. Each "unlabeled" RN was then more carefully evaluated using only the green channel (instead of the red and green combined image), and any RNs that corresponded to a dim green MLH1 signal were also marked as MLH1-positive RNs. When no green signal could be observed under these conditions (including additional temporary enhancement of the green signal), the RN was marked as an MLH1-negative RN. On a separate layer, dots (using the pencil tool) of different colors were used to mark the positions of kinetochores, MLH1-positive RNs, MLH1-negative RNs, and MLH1 foci on SCs that did not correspond to RNs. Layers corresponding to the EM

montage, the RN and kinetochore dots, and SC numbers (used only to identify the order in which the SCs were to be measured) were merged into one layer and saved as a separate bitmap file for measuring using MicroMeasure 3.0 (http://www.colostate.edu/Depts/Biology/MicroMeasure/). The image resolution of the bitmap file was adjusted to 200 dpi so the files were not too large to be processed by the MicroMeasure program. After measurement, the chromosome identification of each SC was determined using relative length and arm ratio (**Supplementary Table 1**). In some cases, one or more SCs in a set had to be excluded due to a lack of distinct kinetochores or to the presence of stain precipitate at the EM level that obscured large portions of SCs. SCs from these groups were used for counting the number of RNs per SC set but were not used for mapping RN positions. Positions of RNs on SCs were measured as a percentage of the arm length from the kinetochore [31]. Each RN position was converted to a µm position on the appropriate chromosome using average arm lengths [31] (**Supplementary Table 1**). Data on RN positions for each SC were combined into an Excel spreadsheet for analysis.

RN Density: Density measurements of lateral elements, MLH1-negative RNs, and MLH1-positive RNs were taken from 800 dpi EM images using Image J after correcting for background.

Fluorescence *in situ* hybridization (FISH) on zygotene SC spreads: SC spreads from tomato primary microsporocytes in the zygotene stage of meiosis were prepared on uncoated glass slides as described above for plastic-coated slides. FISH was performed as described [29, 33] using biotin and digoxigenin labeled DNA probes from bacterial artificial chromosomes (BACs) LE_HBa0176H22 to identify the short arm of chromosome *10*, LE_HBa0013B20 to identify the

long arm of chromosome *10*, LE_HBa0045N22 to identify the short arm of chromosome *12* and LE_HBa0030J22 to identify the long arm of chromosome *12* (http://solgenomics.net/cview/map.pl?map_id=13). Hybridized biotin-labeled probes were detected using consecutive antibody labeling of mouse anti-biotin (1:50), biotinylated donkey anti-mouse (1:100) and streptavidin-DTAF (1:100) and hybridized digoxigenin-labeled probes were detected using consecutive antibody labeling of sheep anti-digoxigenin (1:125) and donkey anti-sheep TRITC. The microscope system and imaging parameters were the same as those used for the SC and MLH1 focus imaging (above). All antibodies were from Jackson ImmunoResearch Labs except sheep anti-digoxigenin that was from Roche.

Genetic coordinates and local recombination rates along chromosomes: SC coordinates of COs (MLH1 foci and/or RN) and the kinetochore were constructed from the curvilinear distance as described above. Given the SC positions of all COs, the *genetic* coordinate of an arbitrary point is defined as half the mean number of COs detected between the left end of the SC and the point of interest. Furthermore, *local* recombination rates are measured in cM (centiMorgan) per µm along the SC by introducing intervals and taking 50 times the mean number of COs per µm in that interval.

Interference strength inferred using the Gamma model: The Gamma model [32] provides a standard framework for modeling CO formation and CO interference in a pathway (see more details in **Supplementary Methods**). The model's parameter ν quantifies the strength of interference: absence of interference corresponds to the value $\nu=1$ and increasing interference corresponds to increasing $\nu>1$. For any realization of COs on an SC, it is possible to calculate its

likelihood within the Gamma model assuming a value of v [33]. This allows us to use the maximum likelihood method to determine for each chromosome its "optimal" v *, i.e.,* that which best fits the experimental data [34]. Confidence intervals for the fitted v's are computed based on the Fisher information matrix using the CODA software [34].

Tests of no cross-talk between the two CO formation pathways: Our first test is based on the numbers of MLH1-positive and MLH1-negative RNs on each SC, and applying a Fisher's exact test on the frequency of these numbers. To be able to perform this test on data pooled over all chromosomes, we wrote an implementation of Fisher's exact test in which we compute the log-likelihood of the data by summing log-likelihoods over all chromosomes. We then obtain the *p*-value of the test by comparing this log-likelihood with 10^5 log-likelihood values obtained after shuffling the list of MLH1-negative RNs amongst all SCs to remove any correlation between MLH1-negative and MLH1-positive RNs while keeping the exact same distribution of numbers per SC of MLH1-positive RNs and of MLH1-negative RNs (see **Supplementary Methods**). We checked this procedure by ensuring that for individual chromosomes, we obtained the same *p*-values as when using the *fisher.test()* function in R.

Our second test is based on distances between MLH1-positive and MLH1-negative RNs. We test for independence of the two pathways by comparing the distributions of those distances when there is and when there is not shuffling just as in the previous test. The chi-square (function *chisq.test()* in R) test applied to the histograms representing these distributions was used to produce *p*-values.

Test of no P1-P1 interference based on inter-CO distances: In this case, the observed distribution of distances between each class I CO and all other class I COs on the same SC was compared to the distribution expected in the absence of interactions, which was obtained again by a shuffling procedure. Specifically, while keeping the same number of MLH1-positive RNs for each SC, we shuffled their positions across the whole data set.

References

28. Lhuissier,F.G.P., Offenberg,H.H., Wittich,P.E., Vischer,N.O.E., & Heyting,C. The mismatch repair protein MLH1 marks a subset of strongly interfering crossovers in tomato. *Plant Cell* **19**, 862-876 (2007).

29. Anderson,L.K. & Stack,S.M. Preparing SC spreads with RNs for EM analysis in *Plant Meiosis: Methods and Protocols* (eds. Pawlowski,W.P., Grelon,M. & Armstrong,S.) 147-158 (Spring Science + Business Media, New York, 2013).

30. Stack,S.M., Sherman,J.D., Anderson,L.K., & Herickhoff,L.S. Meiotic nodules in vascular plants in *Chromosomes Today* (eds. Sumner,A.T. & Chandley,A.C.) 301-311 (Chapman & Hall, London, 1993).

31. Anderson,L.K. *et al.* High resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics* **165**, 849-865 (2003).

32. McPeek,M.S. & Speed,T.P. Modeling interference in genetic recombination. *Genetics* **139**, 1031-1044 (1995).

33. Broman,K.W. & Weber,J.L. Characterization of human crossover interference. *Am. J Hum. Genet.* **66**, 1911-1926 (2000).

34. Gauthier,F., Martin,O.C., & Falque,M. CODA (crossover distribution analyzer): quantitative characterization of crossover position patterns along chromosomes. *BMC Bioinformatics* **12**, 27 (2011).

.....

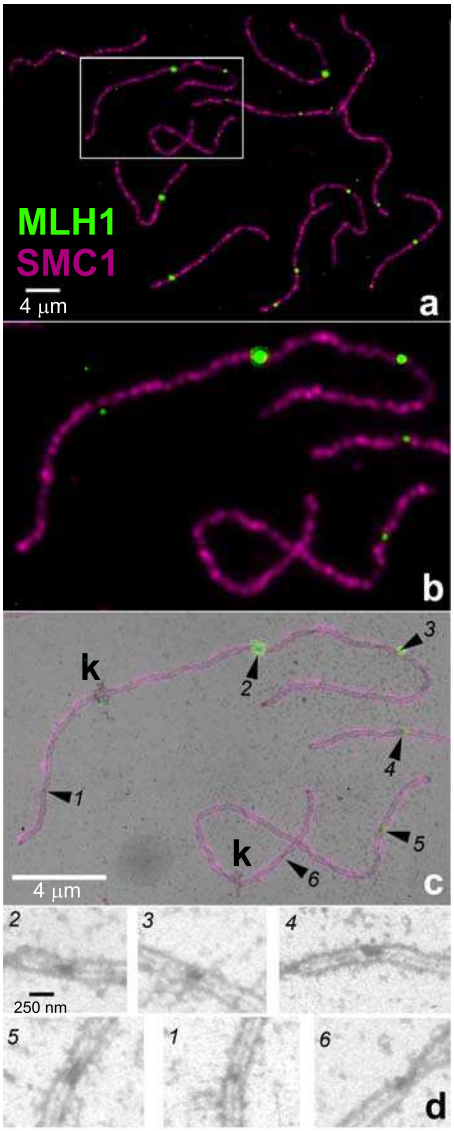
Figure legends.

Figure 1. Identifying MLH1-positive and MLH1-negative RNs using consecutive light and electron microscopy of the same SC spread. (**a**) LM view of complete set of 12 tomato SCs labeled with antibodies to MLH1 (green) and SMC1, an SC component (magenta). (**b**) Area boxed in **a**, enlarged to match EM image. (**c**) Fluorescent LM image from **b** made partially transparent (25% opacity) and superimposed over the corresponding EM image. Two complete SCs, each with a kinetochore (**k**) are shown along with a part of another SC. Numbered arrowheads point to RNs that are visible in the EM image. Four RNs (numbers 2-5) correspond to sites of MLH1 foci (= MLH1-positive RNs) while two RNs (numbers 1 and 6) do not correspond with MLH1 foci (= MLH1-negative RNs). (**d**) Enlarged view of RNs without the LM overlay. MLH1-positive RNs (numbers 2-5) are larger than MLH1-negative RNs (numbers 1 and 6). The complete SC spread with LM, superimposed and EM images is illustrated in **Supplementary Figure 1**. Scale bars are 4 µm for (**a-c**) and 250 nm for (**d**).

Figure 2. Distributions of MLH1-positive RNs (blue bars) and MLH1-negative RNs (red bars) for tomato chromosomes 1, 6 and 10. (See **Supplementary Fig. 2** for the distributions for all tomato SCs.) Histogram bars are in 2% SC length intervals (X-axis) and show the number of RNs observed per interval (left Y-axis). The red histogram has been superimposed on the blue histogram for each chromosome. To highlight differences between the individual histograms, lines showing the cumulative frequency (right Y-axis) of all RNs (black), of MLH1-positive RNs (blue), and of MLH1-negative RNs (red) have been added. The black numbers in each histogram indicate the length of the short arm as a percentage of the entire SC. The red and blue numbers in each histogram indicate the percentage of MLH1-negative RNs and MLH1-positive RNs, respectively, that were observed on the short arm. The length of each chromosome has been scaled to represent its appropriate length relative to the other chromosomes.

Figure 3. Distribution of distances (in µm SC) between MLH1-positive RNs (class I COs) and MLH1-negative RNs (class II COs) or between pairs of MLH1-positive RNs (class I COs) pooled over all chromosomes. Vertical bars represent 95% confidence intervals for each bin of the histogram. The expected distribution in the absence of any interaction between class I and class II COs (see **Online Methods**) is plotted as a broken line with the associated 95% confidence intervals (dashed lines). (**a**) Interference between class I and class II COs and (**b**) between class I COs is indicated by the lower than expected frequency of events at short distances. Similar figures for all individual chromosomes are shown in **Supplementary Figure 8** for class I – class II CO distances and in **Supplementary Figure 11** for class I – class I CO distances.

Figure 1.



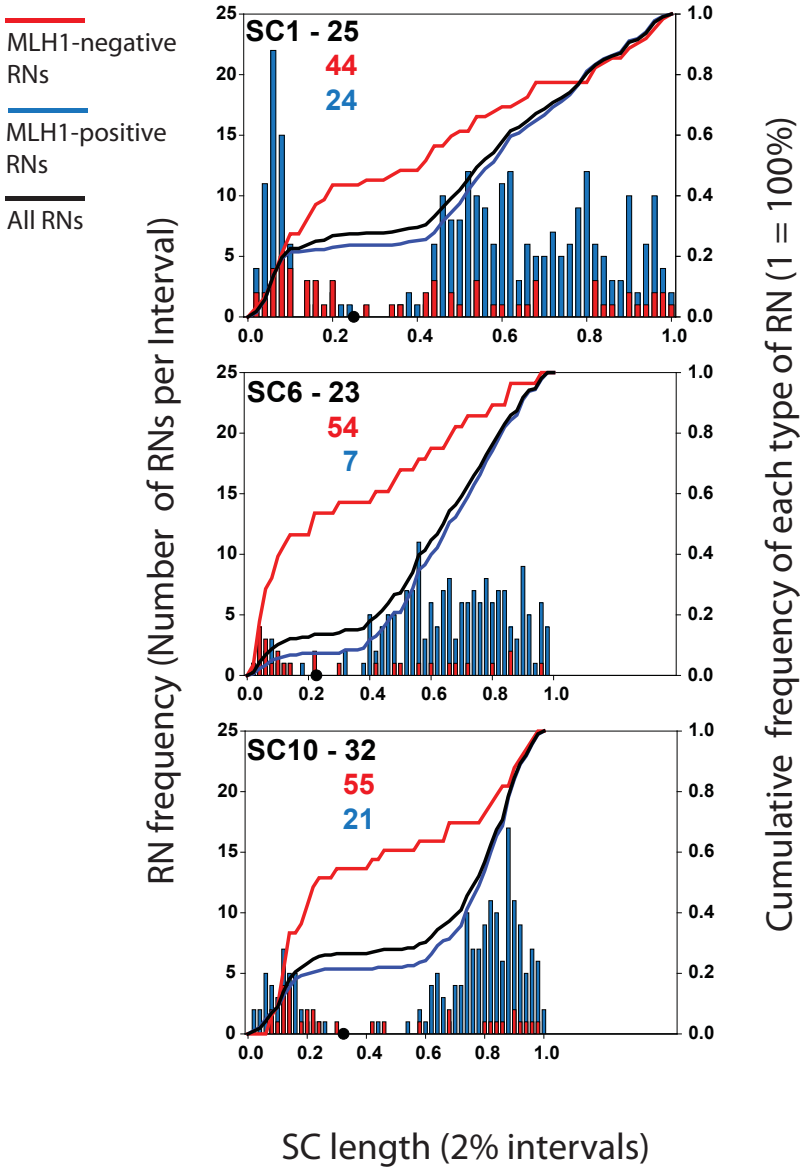
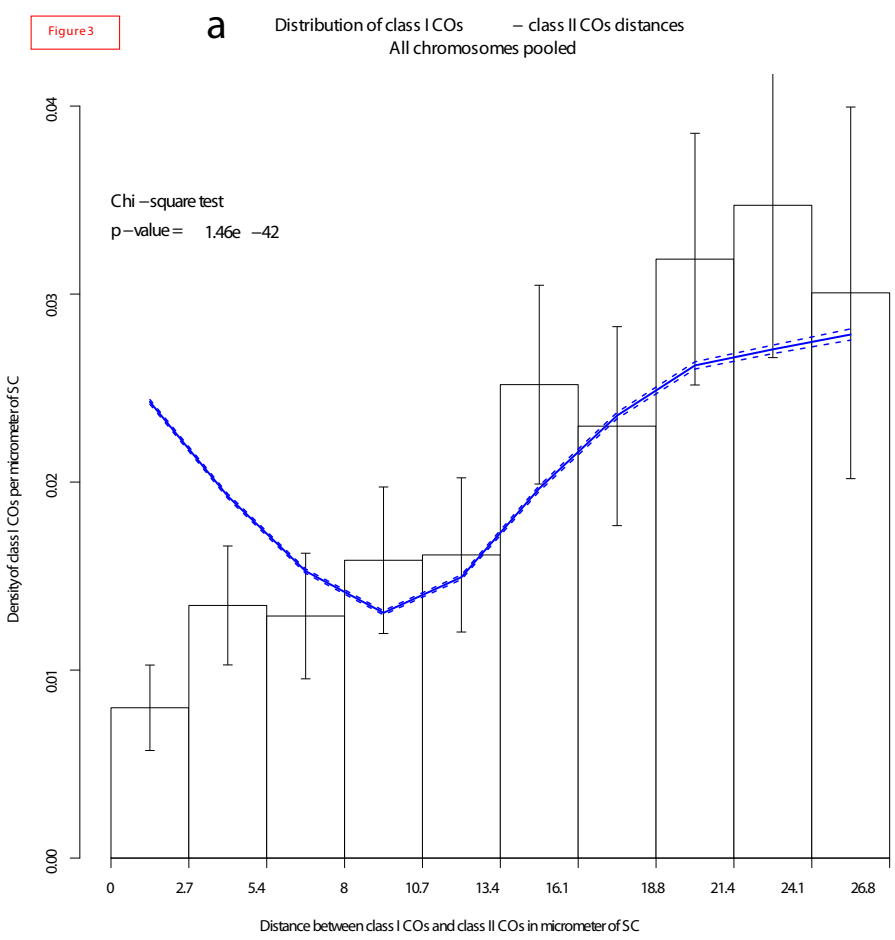
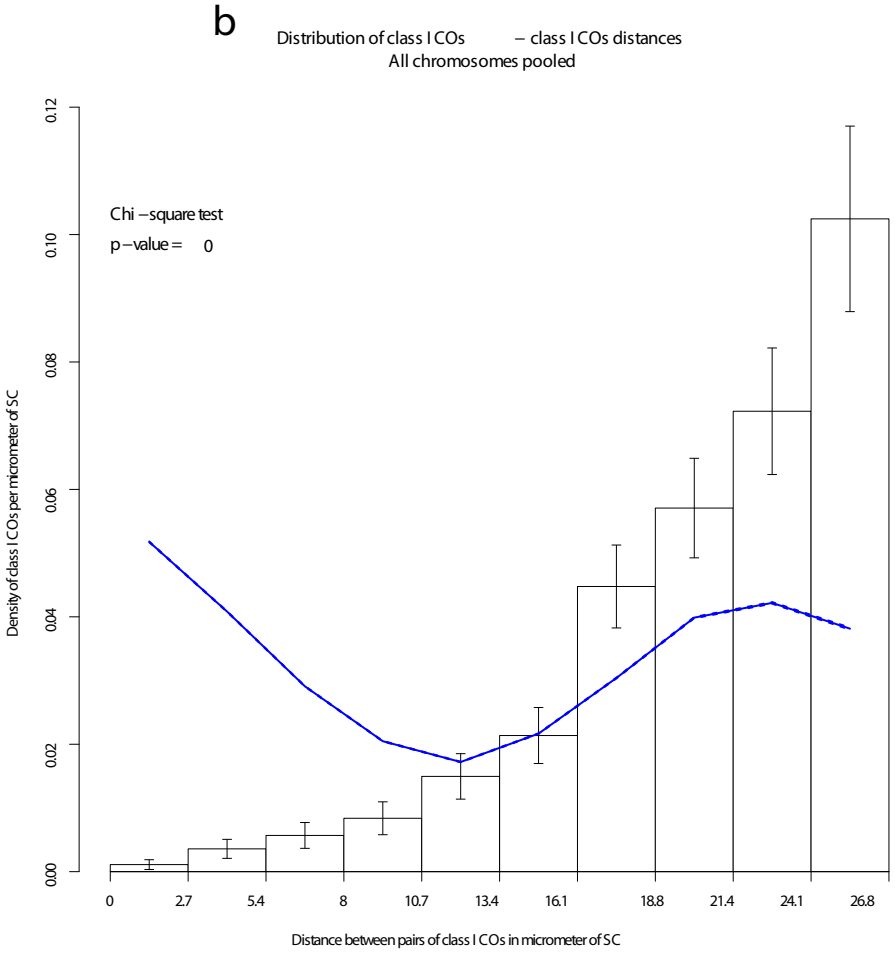

MLH1-negative RNs
MLH1-positive RNs
All RNs
SC1 - 25
44
24
SC6 - 23
54
7
SC10 - 32
55
21
RN frequency (Number of RNs per Interval)
Cumulative frequency of each type of RN (1 = 100%)
SC length (2% intervals)

Figure 2.

Figure3


a
Distribution of class I COs – class II COs distances
All chromosomes pooled
Chi –square test
p –value = 1.46e –42
Density of class I COs per micrometer of SC
0.00
0.01
0.02
0.03
0.04
0
2.7
5.4
8
10.7
13.4
16.1
18.8
21.4
24.1
26.8
Distance between class I COs and class II COs in micrometer of SC


b
Distribution of class I COs – class I COs distances
All chromosomes pooled
Chi –square test
p –value = 0
Density of class I COs per micrometer of SC
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2.7
5.4
8
10.7
13.4
16.1
18.8
21.4
24.1
26.8
Distance between pairs of class I COs in micrometer of SC

PART III

A MODEL HAVING DIFFERENT RECOMBINATION LANDSCAPES FOR INTERFERING AND NON-INTERFERING PATHWAYS

Chapter 6

Setting a Model up…

THE PATHWAY-SPECIFIC LANDSCAPES GAMMA SPRINKLING MODEL (PSL-GS)

A Gamma Recall

As we have discussed before, the single pathway Gamma model is based on an inter-crossover distance distribution and a stationary gamma renewal process with its shape parameter twice its scale, denoted as *nu*. Since this model is accompanied by a computable likelihood, the maximum likelihood approach can be used to infer the best value of *nu* for the data at hand. Furthermore, one can extend it to the two-pathway Gamma model by incorporating the non-interfering pathway. This is done by the random sprinkling of a Poisson (with rate p) number of non-interfering crossovers, superposing these on top of the first (interfering) pathway. Again the likelihood of any realization of COs can be computed. We use the hill climbing algorithm to navigate the parametric space and maximize the likelihood, leading to the optimal parameter values. In addition we are able to compute Fisher information matrix, leading to the 95% confidence intervals for each parameter.

The PSL-GS Model

We introduce a model here which incorporates the knowledge that the recombination landscapes of the two pathways are not necessarily proportional. Indeed, the two-pathway models to date take the framework that the proportion of non-interfering COs is *p*, assumed to be constant along the whole chromosome. Here we want the parameter p which determines the fraction of crossovers contributing to the non-interfering pathway to be an arbitrary function along the chromosome (Figure 13). Note that since the non-interfering crossover fraction is allowed to vary along the chromosome, standard modeling will lead to the conclusion that the interference strength *nu* is also not constant along chromosomes.

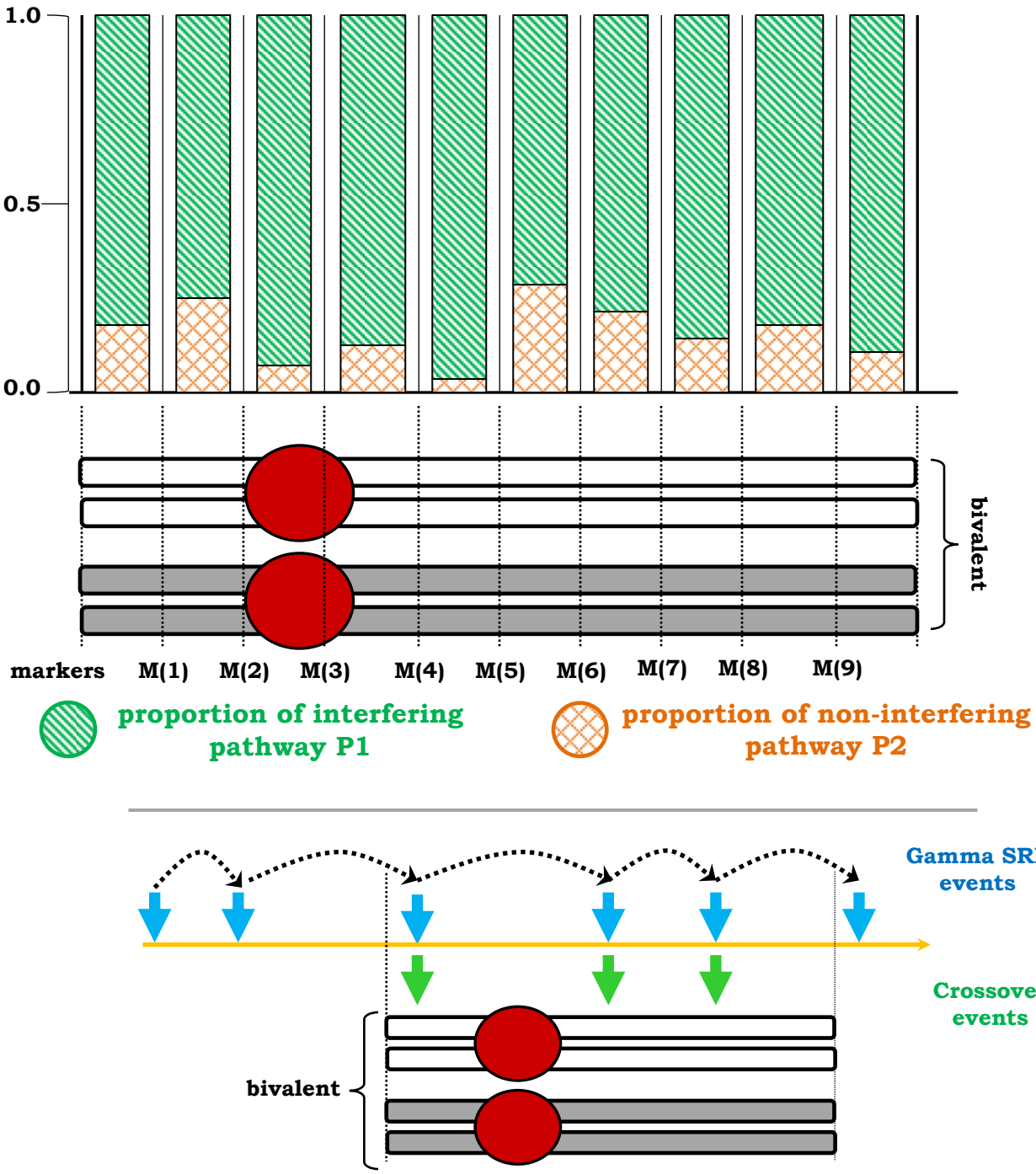


Figure 6.1 The Pathway-specific Landscapes Gamma Sprinkling Model.

The **PSL-GS** model is a generalization of the Gamma two-pathway model to allow for interference heterogeneity along the chromosome. Under this model, we assign different proportions of the non-interfering pathway P2 (denoted as p) in each inter-marker interval (the varying green and orange histograms). This is done in practice by considering each p as a polynomial written as a linear combination of Legendre polynomials of degree $\leq K$, the additional specific in this model.

The crossovers assigned to the interfering and the non-interfering pathway in each marker interval is done in the same way as the gamma model. Interfering crossovers are from a gamma SRP and the non-interfering crossovers come from an exponential SRP.

Additional parameters

The parameters specific to this model are related only to p. We retain the same interference parameter *nu* as in a gamma model. As seen in our previous work (see e.g. our paper in Genetics), different marker intervals are likely to have different values for the fractions of the interfering and non-interfering COs. Taking this into consideration, let each marker interval along a chromosome have a different value for p_i such. Note that the genetic distance between two markers is receives a contribution from the interfering and non-interfering pathways (Figure 13). Thus we have:

$$X_{G_{M(i),M(i+1)}} = X_{G(P1)_{M(i),M(i+1)}} + X_{G(P2)_{M(i),M(i+1)}} \quad \ldots (6.1)$$

where, $M(i)$ indicates the ith marker, G denotes the genetic distance between the markers $M(i)$ and $M(i+1)$. Also, $P1$ stands for the interfering crossover formation pathway while $P2$ represents the non-interfering pathway.

Unfortunately, such a framework would result in a huge number of parameters to estimate as the number of marker intervals along a chromosome may be as many as 100 or even more. This severely reduces the statistical usefulness of our model, limiting in particular the possibility of estimating the model's parameters. To counter this hindrance, we consider that the parameter p along the chromosome is given by a polynomial in the genetic position $G_{Mi,Mi+1}$ of degree ≤ K. Without loss of generality, the polynomial which represents each interval's non-interfering fraction p can be considered as a function $p(x)$, written as a linear combination of $(K+1)$ Legendre polynomials of degrees, 0, 1, 2, …, K. Here x denotes the position along the chromosome in genetic units. A nice feature of this framework is that, when K = 0, this model becomes equivalent to the Gamma two-pathway model.

Thus the parameters to be estimated for this model are nu along with the $(K+1)$ coefficients corresponding to the contributing Legendre polynomials $(P_n(x))$. In practice we use *shifted* Legendre polynomials (written as $\tilde{P}_n(x)$ and explained in the next sub-head) which are orthogonal in the domain [0, 1]. This is convenient for us as we want our model to be defined in the (relative) genetic space. Thus, the new model's function $p(x)$ is defined as follows:

$$p(x) = \sum_{n=0}^{K} \alpha_n \tilde{P}_n(x) \quad \ldots (6.2)$$

where, α_i is the coefficient corresponding to the *ith* shifted Legendre polynomial. The vector $\boldsymbol{\alpha}$ (with as many elements as (K+1)) gives the additional parameters to be estimated in this model, apart from *nu*.

Legendre (L) polynomials

This polynomial class are solutions to the Legendre differential equation that is often used in quantum mechanics. The equation is given as follows:

$$(1-x^2)y'' - 2xy' + l(l+1)y = 0 \quad \ldots (6.3)$$

where, $y = f(x)$ and $l = 0, 1, 2, ...$

Since the differential equation is of second order (the highest derivative is of that order), when considering the Legendre polynomials, two constants of integration, a_0, a_1 must be considered. These polynomials are denoted as $P_l(x)$ wherein l gives the L polynomial degree. Without loss of generality, a_1 is taken to be zero when l is even and a_0 is put to zero if l is odd. Then, by convention, one chooses the value of the non-zero a_0 or a_1 such that we have, $P_l(1) = 1$. And the recursive formula which these polynomials satisfy is:

$$a_{k+2} = \frac{k(k+1)-l(l+1)}{(k+1)(k+2)} a_k \quad \ldots (6.4)$$

For example, consider the $0th$ order L polynomial, $l = 0$ and $a_1 = 0$, but $a_0 \neq 0$, then we have

$a_{k+2} = \frac{k(k+1)-0(0+1)}{(k+1)(k+2)} a_k$. Clearly all odd terms are zero and a_2 as well when this relation is used, thus we conclude that, $a_{0+2} = \frac{0(0+1)-0(0+1)}{(0+1)(0+2)} a_0 = 0$. This implies that all the even terms also reduce to zero. Thus we can say that $P_0(x) = a_0$. Now the condition $P_0(1) = 1$ gives $P_0(x) = 1$.

Next for the first order L polynomial, $l = 1$ and $a_0 = 0$ with $a_1 \neq 0$. This leads to:

$a_{k+2} = \frac{k(k+1)-1(1+1)}{(k+1)(k+2)} a_k$. All the even terms disappear here and for the odd terms we have $a_{1+2} = \frac{1(1+1)-1(1+1)}{(1+1)(1+2)} a_1 = 0$ which means all odd terms apart from a_1 must be zero. Hence, $P_1(x) = a_1 x$. Again with $P_1(1) = 1$, we can see that $a_1 = 1$. This gives us that $P_1(x) = x$.

Going on to the second order polynomial, we have:

$l = 2,\ a_1 = 0,\ a_0 \neq 0$.

All odd terms disappear and for the even terms we can say that:

$a_{0+2} = \frac{0(0+1)-2(2+1)}{(0+1)(0+2)} a_0 = -3a_0$ but the remaining terms will again be zero, as $a_{2+2} = \frac{2(2+1)-2(2+1)}{(2+1)(2+2)} a_2 = 0$. So we get, $P_2(x) = a_0 - 3a_0 x^2$. Then, under the convention, we have, $P_2(1) = 1 = a_0 - 3a_0$, which gives, $a_0 = -1/2$. Finally we have $P_2(x) = \frac{1}{2}(3x^2 - 1)$.

One may derive the following third and fourth order L polynomials in exactly the same way. One obtains the following results:

$P_3(x) = \frac{1}{2}(5x^3 - 3x)$ and, $P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$.

It is convenient to use L polynomials which are orthogonal on the domain [0,1]. This property is provided by the *shifted L* polynomials where the x is replaced by $(2x - 1)$, that is, $\tilde{P}_l(x) = P_l(2x - 1)$. This transformation maps between the intervals [0,1] and $[-1,1]$. With this convention, the recursive expression for these shifted polynomials is given as:

$$\tilde{P}_l(x) = (-1)^n \sum_{i=0}^{n} C(n,i)C(n+i,i)(-x)^i \quad \ldots (6.5)$$

where, $C(p,q) = \frac{factorial(p)}{factorial(q)factorial(p-q)}$.

For completeness, the first 5 shifted L polynomials are the following:

$$\tilde{P}_0(x) = 1 \quad \ldots (6.6a)$$

$$\tilde{P}_1(x) = 2x - 1 \quad \ldots (6.6b)$$

$$\tilde{P}_2(x) = 6x^2 - 6x + 1 \quad \ldots (6.6c)$$

$$\tilde{P}_3(x) = 20x^3 - 30x^3 + 12x - 1 \quad \ldots (6.6d)$$

$$\tilde{P}_4(x) = 70x^4 - 140x^3 + 90x^2 - 20x + 1 \quad \ldots (6.6e)$$

Non-interfering pathway & L polynomials

L polynomials are mutually orthogonal but the important point is that they form a basis of the space of polynomials. Hence decomposing the non-interfering crossover fraction function $p(x)$ along a chromosome as a linear combination of a finite number of such polynomials provides a very convenient representation. In practice, we work with genetic data providing genotypes at a set of markers. If an interval is recombinant, we infer a CO and assign its position to the interval's mid-point and then the relevant value of p is the value of the polynomial at that midpoint. The polynomial's degree K to be considered for each $p(x)$ is data-dependent and will be inferred based on systematic examination of the fitting of the data under consideration.

Likelihood & Hill-climbing for more than 2 parameters

In the PSL-GS model, the likelihood for any given list of n crossover positions is obtained by considering all possible assignments of each crossover to either P1 or P2. Then one must take into account the fact that a different $P2$ fraction for each interval also affects the $P1$ landscape. And as the number of parameters increases to more than two, moving through the parametric

space becomes more challenging. For our work, implemented in a C++ computer program, we perform the hill climbing search by changing one parameter at a time, beginning with nu and proceeding with the successive coefficients multiplying the shifted Legendre polynomials, whatever their number may be. nu has its own range while for simplicity each of the coefficients is taken to have the same range.

The initial likelihood L_0 is defined at the point with coordinates as the initial value for all parameters. All parameters' initial values are defined at 10% of the range of the corresponding parameter plus the lower range limit. Beginning with the first parameter, left and right neighbors are considered which are at a distance given by a step size (this step size is initialized but then can be reduced). Once we have the values of these two neighbors, a curvature is computed based on the central parameter with the left and right neighbors. If this curvature is positive, we simply move to the neighbor which gives the highest likelihood. If the curvature is negative, we compute the likelihood at a trial position based using quadratic interpolation; then we move to the best point among the central one, the neighbors and the trial one. Once we have scanned through the neighbors of all parameters the first time, it is important to revise the step sizes if appropriate. The criterion for reducing the step size for each parameter is based on not having found any better neighbors. Following this iterative procedure, as the likelihood increases monotonically, the step size decreases. The process halts when the step sizes for all parameters are all below a predefined value.

In the algorithm it is important to update the genetic positions of each marker, for both the interfering and the non-interfering pathways, every time one of the coefficients multiplying a L polynomial is modified.

Fishers' confidence intervals for more than 2 parameters

This has already been described in Chapter 2 (*The several parameter case* under Confidence Intervals). For each parameter, we consider its optimum value obtained using the likelihood method and left and right neighbors defined using a step-size. The likelihood corresponding to these three parameter values are then used to compute the corresponding diagonal term in the Fisher matrix. The procedure to get the off-diagonal terms of the Fisher matrix requires considering all possible parameter *pairs*. We do that too, defining step sizes to estimate the mixed second order derivatives. As a last operation, as seen in formula (2.12) in Chapter 2, the Fisher matrix is inverted. The 95% confidence interval of a parameter is then given by 1.96 times the square root of the corresponding diagonal term in the inverse matrix.

$$\hat{\theta}_i \pm c\sqrt{\{I_n(\hat{\theta}_i)^{-1}\}_{ii}} \quad \ldots (2.12)$$

where, $\hat{\theta}_i$ is the ith parameter estimate, c is 1.96 for 95% confidence intervals and $\{I_n(\hat{\theta}_i)^{-1}\}_{ii}$

is the ith diagonal term in the inverse of the Fisher matrix.

Simulating PSL-GS model data

We have implemented a simulation program in C++ that produces a PSL-GS dataset, that is which generated bivalents or gametes where COs are produced, allowing the fraction of non-interfering crossovers to vary from one marker interval to the next. The ith interval of a chromosome (bivalent) will have a $P2$ crossover with probability $2 * p_i * G_{M(i),M(i+1)}$. The remaining genetic length is contributed by $P1$. It is important to understand that having a different p_i for each interval affects the $P1$ landscape as well. The two maps need to be recalculated every time one of the parameters defining the position dependence of p are changed. Only then can the likelihood for each set of COs (from each pathway) be determined. Note that one can also define overall value for p given by its mean in genetic space along the chromosome; this can be useful for comparing to the standard Gamma sprinkling model results.

Chapter 7

Exploring PSL-GS…

BENCHMARKING & EXPLOITATION OF THE PSL-GS MODEL

In this chapter we first perform a detailed study of the efficiency of our inference approach (maximum likelihood method) to estimate the parameters of the PSL-GS model using simulated data. This kind of benchmarking is imperative when a new model is proposed as it reveals its potential as well as drawbacks. Simulated datasets provide an indispensable tool to study the practical usefulness of inferring parameters in a model, regardless of whether the model is biologically relevant. In the second part of this chapter we use this PSL-GS model on *Arabidopsis* data.

Accuracy & Precision of the Inference

Accuracy of inference refers to the reliability of parameter estimation, assuming that the model at hand correctly describes the data. Explicitly, consider that data is produced by simulating the model with parameter τ^*. For each realization k of a simulated data set, the inference will provide an estimate τ_k for τ^* and so τ_k can be thought of as a random variable. Accuracy is defined as the difference between τ^* and the mean of τ_k and so is a measure of the bias of the inference. Similarly, one defines the precision of the inference as the spread (standard deviation) of the random variable τ_k. For maximum likelihood methods, the inference will become both more accurate and more precise as the sample size increases (in our case the sample size is the number of gametes or bivalents). Also since data at the bivalent level (without thinning) gives information about all four chromatids, it should give better estimates (higher accuracy and precision) than a dataset with an equal number of gametes (with thinning). Given that the *Arabidopsis* data set we have is based on gametes, we provide here the benchmarking only for the case with thinning.

Procedures

When only one Legendre polynomial of zeroth order is allowed ($K = 0$), then the PSL-GS model is in principle the same as the two-pathway Gamma model with parameters nu and p. We checked that our coding of the PSL-GS model satisfied this constraint; specifically, maximum likelihood inference using the Gamma-Sprinkling code or the PSL-GS code gave the same optimal parameters as shown in Figure 7.1.

In our first benchmarking of the PSL-GS model, the sample sizes (number of gametes produced by simulating the PSL-GS CO formation procedure) are: 100, 200, 400, 800 and 1600. Datasets

of each size are simulated 20 times using the GS model with the value of nu as 9.5 and p as 0.14. These 20 simulations are different from each other only in terms of the random number seeds used for generating them, leading to 20 independent realizations of simulated data within the GS model. The choice of the parameter values used in these runs is directly dictated by the inferred values on the *Arabidopsis* data set, chromosome 1 (male meiosis).

Given these simulated data, how well does our inference work? And in particular, what happens to accuracy and precision when we increase the number of parameters in this new model? We performed the analyses for Legendre polynomials of degree 0, 1, 2 and 3 in the PSL-GS model on the simulated data described above. The maximum likelihood method was used to determine the best fit for the α_i. The 20 repetitions provided estimates of the mean and spread of these inferred parameters. Since the simulated datasets have been produced with α_i=0, for $i > 0$, the fits should lead to values for these parameters that are compatible with zero values when the data sets are sufficiently large. Furthermore the trend with data set size provides information about the accuracy and precision of the inference procedure.

For a second way to benchmark the inference procedure in the PSL-GS model, the previous analysis was repeated but on simulated data in which $\alpha_0 = \alpha_1 = 0.14$ and $\alpha_i = 0$ for $i > 1$. Again we fit these data to the PSL-GS model using Legendre polynomials of degree 0, 1, 2 and 3. If the inference works well, the fitting will lead to parameter estimates that are close to the values used for the simulations.

Results

We obtained good agreement between the parameters of simulated datasets and the corresponding parameter values estimated by the PSL-GS (Figure 7.1). For both parameters, nu as well as p (or α_0 in case of PSL-GS), the values used for simulation are reproduced very reliably as these values always remain in the estimated confidence intervals. The overall trend followed by the confidence interval lengths is also similar in the two models. With increase in sample size, the estimated value creeps closer to the actual parameter value and the confidence interval becomes smaller as well.

Secondly, we look at the performance of the PSL-GS model for increasing number of parameters (Figure 7.2). The uncertainty in estimation increases when there are additional parameters to be estimated; this is to be expected and is always a challenge when models become more complicated. When there are 2 parameters (Figure 7.2 (a)) considered in addition to *nu* (K = 1), then the first Legendre coefficient, α_0 is estimated reliably as the sample sizes increase from 100 to 1600 whereas the value of the second coefficient α_1 is compatible with zero as expected from

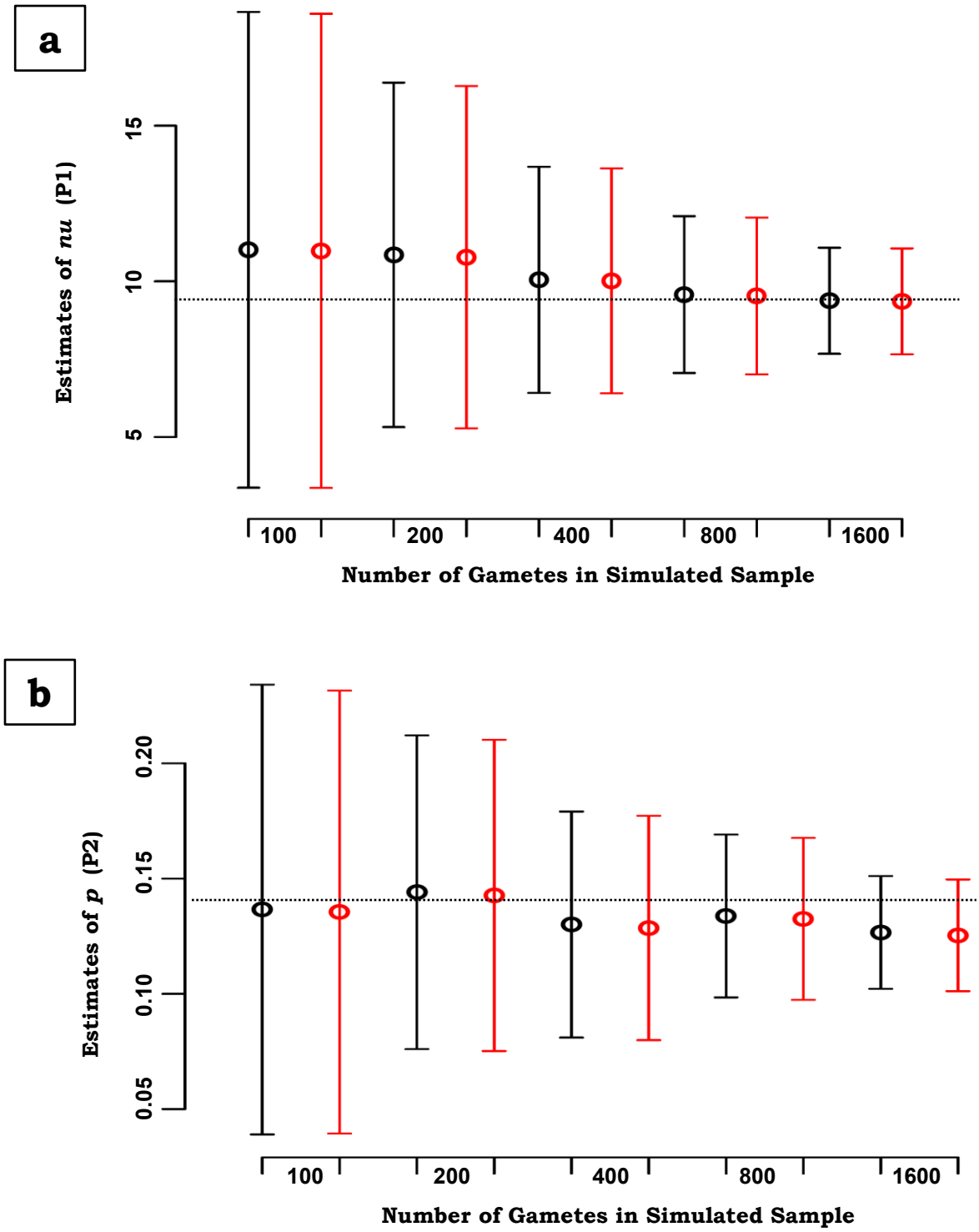


Figure 7.1 PSL-GS estimates against Gamma two pathway estimates.
These figures establish that indeed the new model Pathway-Specific Landscapes Gamma Sprinkling (PSL-GS) is a generalization of the Gamma Sprinkling (GS) or Gamma two pathway model. All datasets were simulated using the Gamma model. The same data was then used to estimate best-fit parameters using the two models. Black gives PSL-GS estimates while Red gives those from GS.

a | Interfering pathway (P1) parameter *nu*
This P1 parameter gives very consistent estimates. The error bars reduce with increase in sample size as expected. They increasingly close in on the value used for *nu* during simulation (9.506, dotted line).

b | Non-interfering pathway (P2) parameter *p* (proportion of P2)
This parameter also gives very consistent estimates. Again the error bars reduce with increase in sample size, growing closer to the real value (0.137, dotted line).

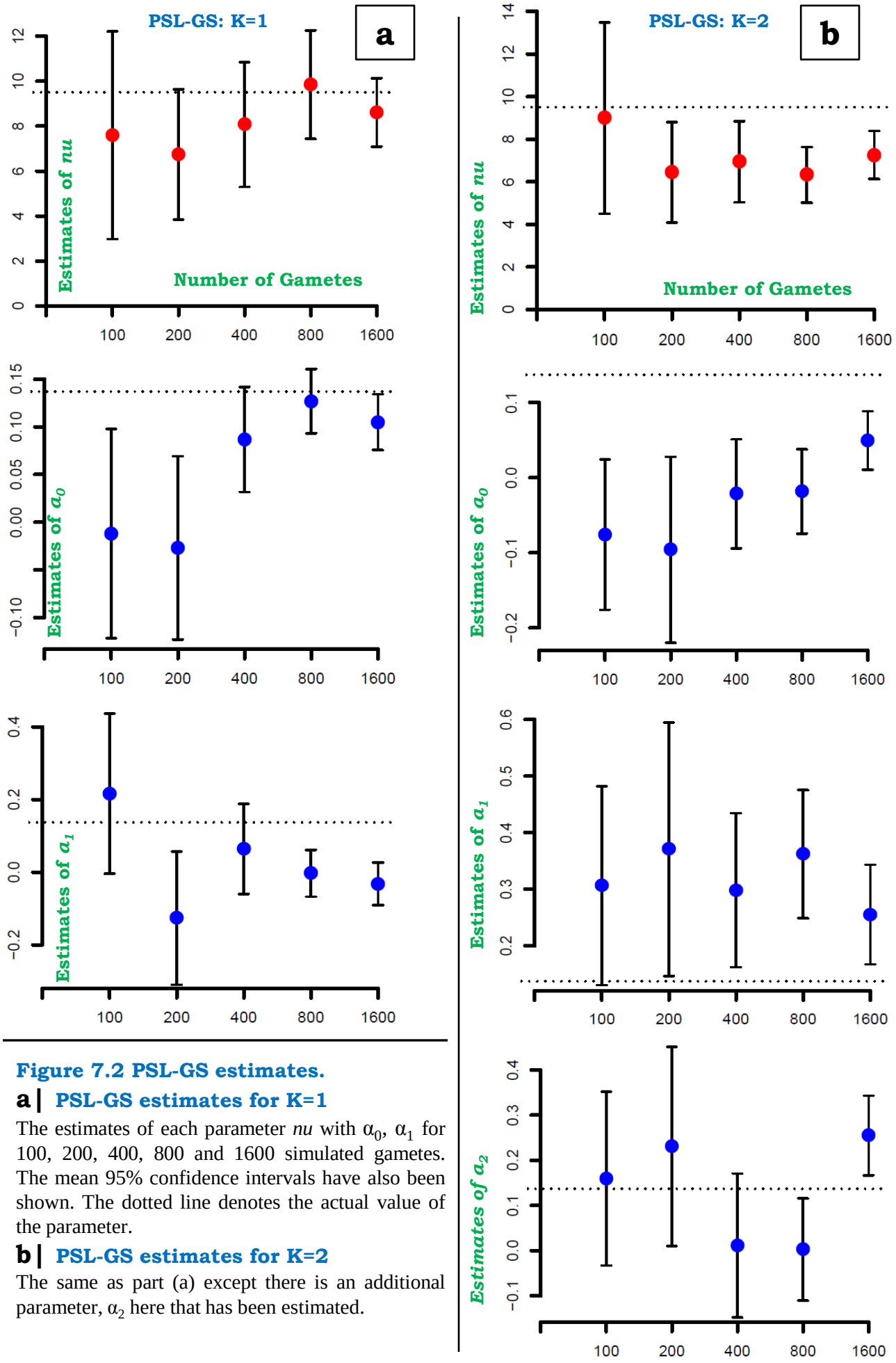


Figure 7.2 PSL-GS estimates.

a | PSL-GS estimates for K=1

The estimates of each parameter *nu* with α_0, α_1 for 100, 200, 400, 800 and 1600 simulated gametes. The mean 95% confidence intervals have also been shown. The dotted line denotes the actual value of the parameter.

b | PSL-GS estimates for K=2

The same as part (a) except there is an additional parameter, α_2 here that has been estimated.

data construction. Values estimated for *nu* are well-estimated also. Further, if we introduce another parameter in the model (K = 2), then *nu* is underestimated even for 1600 gametes. And the first coefficient (α_0) also shows lower values than 0.137 (Figure 7.2 (b)). As a result the subsequent 2 coefficients (α_1 and α_2) are also not estimated reliably. This shows that if we have data at the gamete level, then estimates are reliable for the case with 2 coefficients (K=1) with ~1600 gametes. If more parameters are to be estimated, then tests need to be done with greater sample sizes.

Model selection

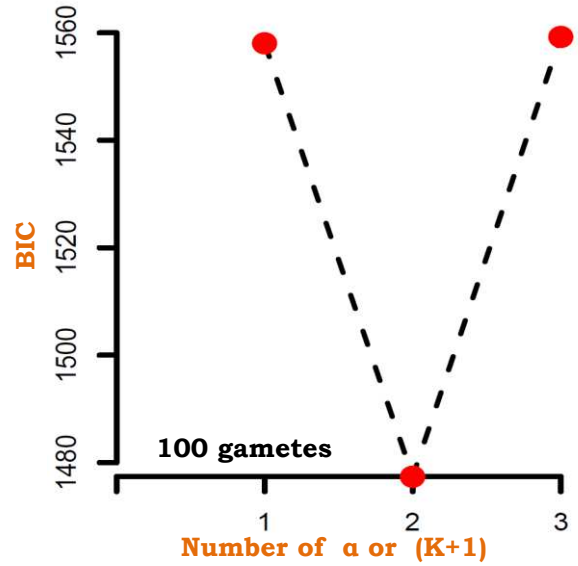
When confronted with real data, one does not know *a priori* the number of parameters that should be included. Furthermore, adding more parameters is sure to improve the fit. It is thus necessary to take into account this advantage of models with more parameters and avoid over fitting data. This challenge of selecting which model (or how many parameters should be used) is "best" is precisely what the Bayesian Information criterion (BIC) does. We are interested in the optimal value of K to represent a dataset when PSL-GS model parameters are estimated. We use the same simulated datasets as above (20 simulations for sample sizes 100, 400 and 1600 only) restricting ourselves to the case of K=1, before looking towards real data. Subsequently, each dataset is analyzed using different parameter numbers (K = 0, 1, 2) from which we take the average of likelihoods for each sample size (for a particular value of K) to compute the associated BIC.

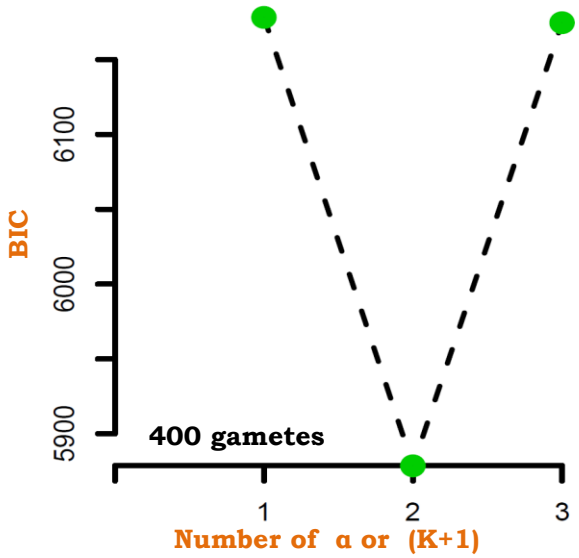
Results

The model with the lowest BIC is considered to be the best: having fewer parameters introduces systematic biases while having more parameters leads to over fitting. Since the simulated data comes from the PSL-GS case with K = 1 (K+1 is indicated on the *x*-axis of the plots), it is expected that all datasets show the lowest BIC for this case. In fact it is indeed concluded to be so (Figure 7.3 (a)) for all three sample sizes.

Application to *Arabidopsis*

Here we have fitted our model to the data from the first chromosome of Arabidopsis produced during male meiosis with 1499 gametes. We chose this chromosome as it is the longest among the five chromosomes. Further, we considered male meiosis as it shows more recombination, the larger number of COs allowing more precise inference. We study the trend in the BIC as the number of parameters are increased (K = 0, 1, 2, 3) in the PSL-GS model.





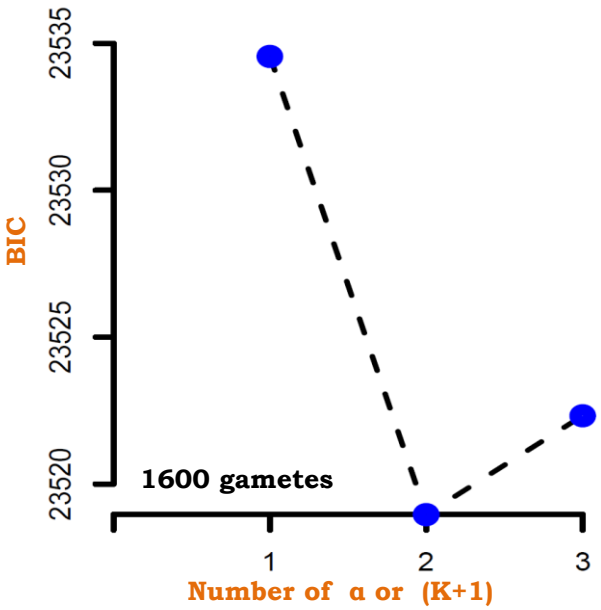


Figure 7.3 a| BIC Values for PSL-GS Simulations.

Data was simulated for PSL-GS (K = 1) with parameter *nu* with α_0, α_1 for 100, 400 and 1600 simulated gametes. Then these data sets were analyzed with PSL-GS with K = 0, 1 and 2.

The corresponding BIC were plotted for each simulated number of gametes. BIC is seen to be the minimum (thus chosen) for K = 1 for all data sets.

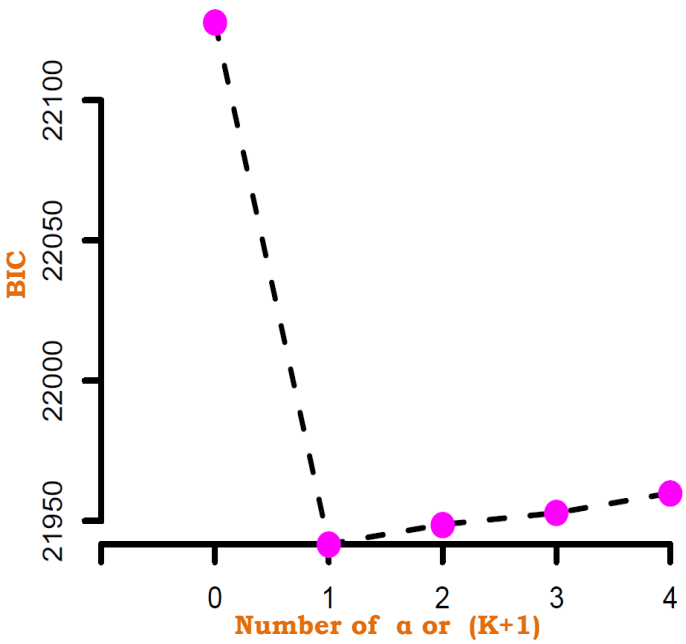


Figure 7.3 b| BIC values for Arabidopsis data.

The data for chromosome 1 (male meiosis) of Arabidopsis was analyzed with PSL-GS using K = -1, 0, 1, 2, 3. When K = -1, it is the Gamma Single pathway model and for K = 0 is the Gamma Sprinkling (GS) model.

The corresponding BIC were computed and plotted. It is seen that the GS model (best model in this case) gives the lowest BIC value.

Results

We observe that there is a big dip in the value of BIC (Figure 7.3 (b)) as we go from no non-interfering pathway associated parameter (K = -1) to at least one (K = 0). Thereafter, with addition of each α_i (K= 1, 2, 3), the BIC keeps increasing in much smaller amounts than the first dip, giving lower goodness of fit than before. Thus when K = 0 and the model is equivalent to a Gamma Sprinkling model, this dataset is able to facilitate parameter estimation efficiently.

Our PSL-GS model is actually a family of models that allows for a transition from the Gamma Sprinkling model to models incorporating heterogeneity in the interfering and non-interfering crossover landscapes. The model tested on simulated dataset the size of 1600 gives parameter estimates consistent with their actual values when K = 1. However, for the same maximum sample size (1600), when another parameter has to be estimated (K = 2), the estimate of *nu* suffers and so does that of α_0 , α_1 and α_2 . Clearly with 1600 gametes, we are able to go till K = 1 but not much beyond. And when considering the BIC for simulated datasets, it perhaps overestimates the efficiency of the PSL-GS as when we move on to the experimental dataset at hand, it provides the best fit with K = 0. Thus the preliminary testing on real data still favors the PSL-GS version equivalent to the Gamma-sprinkling model. It is possible that the heterogeneity (or differences in landscapes of the two pathways) contains mainly small scale differences that are not accessible to low order polynomial representations.

The Ending.

CONCLUSION

Recombination is an integral player in creating newer allelic combinations in each successive generation of sexually reproducing organisms. And it is mediated by crossing over between homologous chromosomes during meiosis. As this process is able to bring together previously apart genes, it is a strategic tool in artificial breeding. However it remains difficult to foresee exactly which alleles will come together or under what conditions a particular portion of a chromosome will recombine or not. The phenomenon of interference further adds to the complexity of the scenario. It has been known since early 20th century that interference monitors crossover number and position along chromosomes but several aspects still remain mysterious.

It can also be argued that interference is not a cause but only a coincidence. But then the till date intriguing fact remains that there are two crossover formation pathways, one that interferes and one that does not. And thus it must be an effect with certain consequence. Though modeling of crossover interference has been performed quite scarcely, related biological data have improved steadily. As genotyping sequencing became the norm, increasingly higher throughput experiments along with techniques such as immuno-precipitation and electron microscopy resulted in various modern data types. Thus the main objective of this thesis is to explore the scope of current modeling techniques.

We began by studying the interference in *Arabidopsis*. Analysis of mutants in the mid-last decade showed that there exist two active pathways of crossover formation (Higgins *et al.* 2004, Mercier *et al.* 2005). Further there had been one modeling study on male meiosis providing estimates of the properties of each pathway (Copenhaver *et al.* 2002). It was concluded that the proportion of the non-interfering pathway (*p*) in this plant usually measures upto 20% or less. Subsequently better data has become available (Giraut *et al.* 2011: male and female backcross population genotyping data) encouraging more stringent testing of models. The Gamma single pathway model fits gave values of *nu* in the range between 2.4 and 4.1. Further Gamma two-pathway modeling yielded values of *nu* in the interval (8, 37) accompanied by *p* varying from 6% to 19%. The non-zero values of *p* for all chromosomes supported the presence of the non-interfering pathway in addition to the interfering one.

Subsequently we attempted to explore if two-pathway modeling indeed describes the Giraut data completely. Were heterogeneities present which were not detected when the model is applied to the whole chromosome? How much does interference vary between chromosomes or even between different regions of the same chromosome? Variations in crossover interference were

investigated at all levels possible – between meiosis in the male and female reproductive organs, between chromosomes in general and also among regions of the same chromosome. In addition to interesting variation patterns in the Gamma single and two-pathway interference parameters at all these levels, we also deduced heterogeneity in the fraction of the non-interfering pathway along each chromosome. An overall major difficulty to overcome here was computation of the likelihood for discontinuous chromosome regions. This happens when a part of the chromosome is *hidden* and this region in turn is flanked by *visible* regions. The situation is complicated by the fact that all possible crossover positions and number are to be considered for the hidden regions (details in Supplementary Material of Basu-Roy *et al.* 2013). And this analysis was published in Genetics (2013).

Then we moved towards the Tomato data from the Anderson group at our disposal. This is the only group to have developed and mastered the technique via which we have the first ever data to allocate crossovers to interfering and non-interfering pathways on the same synaptonemal complex of a wild-type organism. This data clearly shows that the two pathways have different recombination landscapes. We developed novel statistical tools to examine subtle interaction aspects between the interfering and non-interfering crossover formation pathways referred to as "cross-talk" between the two pathways. Since data of this nature where crossovers have been known to contribute to one pathway or the other on the same synaptonemal complex in a wild-type species became available for the first time, there was no previous scope to explore such queries. This rendered the developed tests rather unique and specific in capturing crucial information from the data. It was concluded that the two pathways do interact as opposed to being independent. The manuscript summarizing this work will soon be submitted for publication.

Since analyzing the Tomato data reveals the need for at least two different landscapes for the two crossover formation pathways, we venture to postulate a model in the same spirit. This model is also follow-up to the observed interference heterogeneity in *Arabidopsis* at sub-chromosome level. This new model, known as the Position-specific Landscapes Gamma-Sprinkling (PSL-GS) is a generalization of the standard two-pathway Gamma Sprinkling (GS) model. It encompasses the variation by considering a different non-interfering pathway fraction for each inter-marker interval. There was a need to parametrize these variations as otherwise the model has too many parameters and would lead to over fitting. We have chosen to parametrize by polynomials of a given degree, K. A "natural" choice is to take the basis provided by the shifted Legendre polynomials. The model has the advantage that it is identical to the usual GS model when K= 0. Once the model was set up it was important to benchmark it to ask how many parameters can indeed be efficiently estimated with present datasets. Finally applying the model to Arabidopsis shows higher (maximized) likelihood for this model as the number of parameters to be fitted increase, indicating improvement in the model to describe the data. But model selection criterion does not permit its selection over the GS model. More work is underway to

explore the model and drive it towards publishing as this is the very first endeavor to incorporate heterogeneity in an interference model.

As increasingly detailed and high-throughput biological data becomes the norm, we head towards more involved modeling approaches to crossover interference. Our PSLGS model is a tool that should become more useful as data sets increase in size. In addition, as more data sets with the two pathways annotated separately become available, it would help delve deeper into the various interactions between pathways. Such modeling, in tandem with biologists would further and deepen our present understanding of interference in particular and sexual reproduction in general.

This thesis has been a wonderful opportunity for me, providing the perfect amalgam of two of my cherished interests: modeling and biology. I earnestly hope to continue this journey of learning, contemplation and discovery.

References

Basu-Roy S., Gauthier F., Falque M. *et al.* (2013) Hot regions of non-interfering crossovers coexist with a nonuniformly interfering pathway in *Arabidopsis thaliana*. Genetics 195: 769-79.

Giraut L., Falque M., Drouaud J. *et al.* (2011) Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet.* 7: e1002354.

Higgins J.D., Armstrong S.J., Franklin F. C. H. *et al.* (2004) The Arabidopsis MutS homolog AtMSH4 functions at an early step in recombination: evidence for two classes of recombination in Arabidopsis. *Genes Dev.* 18: 2557-70.

Mercier R., Jolivet S., Vezon D. et al. (2005) Two meiotic crossover classes cohabit in Arabidopsis: one is dependent on MER3, whereas the other one is not. *Curr. Biol.* 15: 692-701.

Résumé

Dans la plupart des organismes, les crossovers se formant au cours de la méiose sont interférérents : deux crossovers sont rarement à proximité. Nous avons étudié ce phénomène en détail dans la plante modèle *Arabidopsis thaliana* en utilisant une grande population de rétro-croisements, en méiose mâle et en méiose femelle. Nous avons utilisé le modèle Gamma à deux voies, superposant à la voie interférente une proportion *p* de crossovers d'une deuxième voie non interférente. La méiose femelle montre une interférence plus élevée et une proportion *p* plus faible que la méiose mâle. Par ailleurs nos comparaisons intra-chromosomiques concluent qu'il existe des variations d'interférence entre le bras gauche et le bras droit d'une part et entre la partie centrale et les régions distales d'autre part. Nous avons ensuite développé des tests statistiques qui ont révélé des hétérogénéités dans la proportion *p* le long des chromosomes. De telles variations ont ensuite été trouvées dans un jeu de données de tomate où notre analyse statistique nous a permis de montrer pour la première fois sur un organisme «sauvage» que la deuxième voie était très peu interférente et qu'en fait les deux voies avaient un peu de « cross-talk ». Vu que la limitation la plus sévère des modèles utilisés jusqu'à présent est l'hypothèse de constance du paramètre *p*, nous avons développé un modèle généralisé où les deux voies peuvent avoir des paysages de recombinaison différents. Ce modèle nommé PSL-GS a été implémenté en logiciel et testé sur des données simulées et sur trois jeux de données de plantes. L'utilisation du critère BIC suggère que tout comme chez la tomate, les paysages de recombinaison d'*Arabidopsis* et du maïs sont différents entre les 2 voies.

Mots clés: méiose, crossovers, interférence entre crossovers, deux voies de formation des crossovers

Abstract

For most organisms, crossovers forming during meiosis exhibit crossover interference – nearby crossovers are rare. This phenomenon was studied in great detail in *Arabidopsis thaliana* using a large backcross population for male and female meiosis. We used the gamma two-pathway model by superposing proportion (1- *p*) of interfering crossovers with *p* of non-interfering crossovers. It was observed that female meiosis shows higher interference but lower proportion of non-interfering crossovers than male meiosis. Further intra-chromosomal interference comparisons conclude that there are variations between (a) left and right arms, and (b) the central versus distal regions of a chromosome. Then statistical tests revealed heterogeneities in the non-interfering crossover proportion along chromosomes. Thereafter various statistical tools developed to examine a very novel wild-type tomato data annotating crossover positions from both pathways provided evidence for ‘cross-talk’ between the two pathways of crossover formation as opposed to being independent. Finally a model named Pathway-Specific Landscapes Gamma-Sprinkling (PSL-GS) incorporating chromosomal non-uniformity in the individual pathways has been proposed to extend present state-of-the-art interference modeling which consider both pathways to be uniform along chromosomes.

Keywords: meiosis, crossovers, crossover interference, two crossover formation pathways