Investigating the spatial regulation of meiotic recombination in *S. cerevisiae*

A thesis submitted for the degree of Doctor of Philosophy in Biochemistry

Timothy J. Cooper

Genome Damage & Stability Centre

University of Sussex

September 2017

Declaration

I hereby declare that this thesis has not been and will not, be submitted in whole or in part to another University for the award of any other degree.

Signature: ……

Acknowledgements

First and foremost, I must express my utmost gratitude to Dr Matthew Neale. This has not been an easy journey for me, and I certainly would not have made it under any other supervisor. Matt has been extremely supportive of me both personally and professionally, going above and beyond his role as a supervisor. His expert advice, patience and encouragement throughout kept me on track and has been an invaluable factor in the completion of this thesis.

I would also like to thank all members of the Neale lab, past and present. I have never been the most engaged or agreeable member of the team but they have always remained professional, welcoming and friendly toward me. In particular I would like to thank Valerie Garcia, who has always been there for helpful advice and encouragement. I am grateful to the European Research Council (ERC) for funding my PhD studentship. Without this, I would not have been able to complete my research projects.

Dedicated to my loved ones.

Abstract

Investigating the spatial regulation of meiotic recombination in S. cerevisiae

In order for a species to engage in and reap the evolutionary benefits of sexual reproduction, a subset of cells in each individual must undergo a complex ordeal known as meiosis—a specialised cell division. By halving the genome content and "shuffling the deck", meiosis generates genetically diverse haploid gametes (eggs, sperm) or spores from diploid cells. Such a monumental task is by no means easy or risk free: during the meiotic programme, cells intentionally damage their own genomes through widespread induction of DNA double-strand breaks (DSBs) in order to initiate homologous recombination—a DNA-repair process—and subsequent crossover (CO) formation. The success of meiosis is, however, not left up to chance. Rather, a complicated web of regulation acts at multiple stages to ensure this dangerous tradeoff pays dividends. Notably, the spatial pattern of meiotic recombination across the genome is complex and non-random. Whilst ultimately stochastic in nature, recombination events within any given meiotic cell display relatively even distributions along each chromosome—a phenomenon mediate by processes of "interference" acting at two key stages in meiosis: DSB and CO formation. Despite wide ranging historical observation, relatively little is known about how either form of interference is accomplished. Genome-wide mapping of recombination within *S. cerevisiae* has, however, provided a unique opportunity to investigate the underlying mechanisms.

By computationally and mathematically analysing genome-wide data, work presented throughout this thesis seeks to: (i) investigate CO distribution and CO interference within various DNA damage response and DNA repair mutants (Tel1^{ATM}, Mec1^{ATR}, Rad24, Msh2) **(Chapter 2)** (ii) develop novel approaches to DSB mapping **(Chapter 3)** (iii) characterise the hyperlocal regulation of DSB formation **(Chapter 3)** and (iv) examine the mechanics of DSB interference **(Chapter 4)**. Moreover, widely applicable simulation platforms for investigating DSB and CO formation have been developed **(Chapter 2, 4)**. Collectively, this thesis further elucidates the mechanisms that underpin the spatial regulation of meiotic recombination in *S. cerevisiae*.

Publications

T.J. Cooper, K. Wardell, V. Garcia & M.J. Neale, 2014. Homeostatic regulation of meiotic DSB formation by ATM/ATR. Experimental cell research, 329(1), pp.124–131.

V. Garcia, S. Gray, R.M. Allison, **T.J. Cooper** & M.J. Neale, 2015. Tel1(ATM)-mediated interference suppresses clustered meiotic double-strand-break formation. Nature, 520(7545), pp.114-8.

T.J. Cooper, V. Garcia & M.J. Neale, 2016. Meiotic DSB patterning: A multifaceted process. Cell Cycle, (1), pp. 13-21.

Table of Contents

Figure & Table List

Abbreviations

(ϒ)	Gamma (Distribution)
CDF	Cumulative Distribution Function
ChIP	Chromatin Immunoprecipitation
CO	Crossover
CoC	Coefficient of Coincidence
DDR	DNA Damage Response
dHJ	Double Holliday Junction
DSB	Double-strand Break
DSBR	Double-strand Break Repair
dsDNA	Double-stranded DNA
GoF	Goodness of Fit
h(x)	Hazard Function
HJ	Holliday Junction
HR	Homologous Recombination
IED	Inter Event Distance
IHB	Inter-homologue Bias
INDEL	Insertion/Deletion
KS	Kolmogorov-Smirnov
MI	Meiosis I
MII	Meiosis II
MMR	Mismatch Repair
NCO	Non-Crossover
NGS	Next-generation Sequencing
NHEJ	Non-homologous End Joining
ORF	Open Reading Frame
PDF	Probability Distribution Function
rDNA	Ribosomal DNA
RMM	Rec114-Mer2-Mei4
SC	Synaptonemal Complex
SDS	Synapsis-dependent Shutdown
SDSA	Synthesis-dependent Strand Annealing
SNP	Single Nucleotide Polymorphism
ssDNA	Single-stranded DNA
VCF	Variant Call Format
WT	Wild Type

CHAPTER 1

Introduction

1.1—Mitosis & Meiosis

Cell division underscores much of biology—facilitating vegetative growth, development and maintenance or repair of tissues. Division is predominately accomplished through the process of mitosis; in short, following a protracted interphase consisting of G1/S/G2 phases during which cellular growth, organelle production and DNA replication occurs, mitosis, or M phase, segregates chromosomes and forms two identical daughter cells (Figure 1.1A). M phase proceeds through several stages including chromosomal condensation (prophase), alignment of chromosomes on the metaphase plate (metaphase), separation of sister chromatids (anaphase), movement of chromatids toward opposing cellular poles (telophase) and physical division of the cell (cytokinesis). However, an alternative pathway of cell division also exists within sexually reproducing organisms. Meiosis is a unique, specialised cell cycle programme responsible for the production of genetically diverse, haploid gametes from single diploid progenitors. Meiosis involves two sequential nuclear divisions preceded by a single round of DNA replication (Figure 1.1B). Meiosis II (MII) resembles a standard mitotic division, involving the equational separation of sister chromatids. Meiosis I (MI), however, must power an unusual, reductional segregation of homologous chromosomes, necessitating significantly complex and multifaceted procedures to find, pair and ultimately segregate the homologues. Meiosis I is primarily defined by a lengthy prophase I, which is cytologically separated into several distinct phases, namely leptotene, zygotene, pachytene, diplotene and diakinesis. (i) Leptotene ("*thin ribbons*")—subsequent to interphase replication of DNA, chromosomes begin to condense. During this stage, intentional formation of double-strand breaks (DSBs) occurs throughout the genome as part of a stringently controlled process to initiate homologous recombination (HR)—a DNA repair mechanism critical to meiosis I; (ii) Zygotene ("*paired ribbons*")—HR-dependent engagement of homologues triggers synapsis—a physical pairing of homologous chromosomes through a tripartite, zipper-like structure known as the synaptonemal complex (SC). The leptotene-zygotene transition is marked by a "bouquet" structure, whereby telomeric regions cluster along the nuclear periphery, aiding the "search and find" process; (iii) Pachytene ("*thick ribbons*")—synapsis completes yielding bivalent chromosomes.

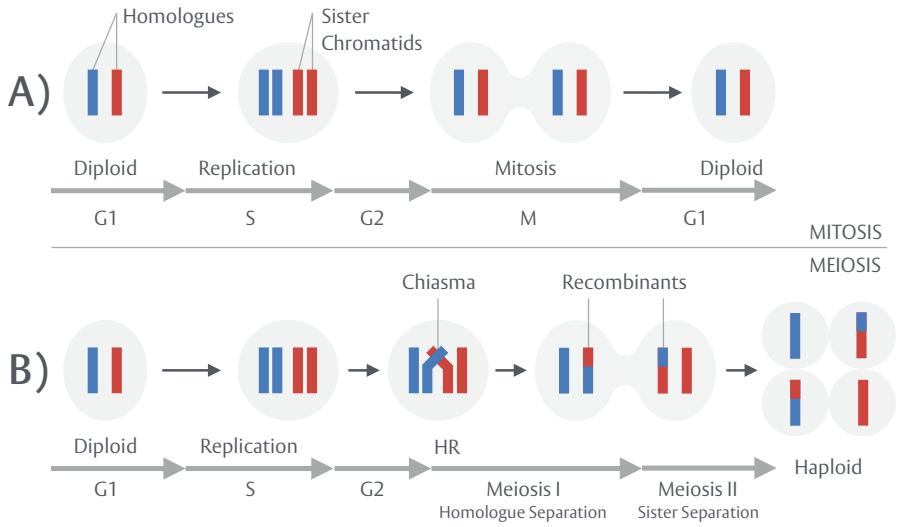


Figure 1.1. Mitotic and meiotic cell cycles

A) Mitotic cell cycle. During mitotic S phase, diploid cells undergo DNA replication to produce sister chromatids. Sister chromatids subsequently condense, align on the metaphase plate and separate to opposing poles during M phase—generating two near identical, diploid daughter cells. **B)** Meiotic cell cycle. In meiosis, a single round of DNA replication during pre-meiotic S phase is proceeded by two rounds of chromosomal segregation (meiosis I and meiosis II). Prior to the first division and during prophase I, homologous recombination (HR) generates recombinant chromosomes and physical connections between homologues, known as chiasmata, before they separate to opposite poles. Meiosis II is characterised by a mitotic-like division and separation of sister chromatids—generating four, non-identical haploid gametes. Lengths of each cell cycle stage are not drawn to scale.

Joint molecules created during HR are resolved to yield non-crossovers (NCOs) or crossovers (COs), exchanging genetic information between maternal and paternal genomes—creating unique genetic complements; (iv) Diplotene ("*double ribbons*")—homologous chromosomes migrate apart as SC deconstruction proceeds but remain attached at discrete points through chiasmata—physical bridges marking the site of crossovers; (v) Diakinesis ("*to move*")—the final stage of meiotic prophase I, marked by full bivalent condensation, disassembly of the nuclear envelope and spindle migration in preparation for metaphase I. Prophase I is thus marked by a complex set of processes unique to meiosis; in the following sections these are further explored.

1.2—DNA double-strand break (DSB) repair

DNA double-strand breaks (DSBs), comprising cleavage of both DNA strands, are highly toxic lesions that can drive genomic instability. Several exogenous and endogenous factors can induce spontaneous DSB formation including ionising radiation (IR), reactive oxygen species (ROS) and replication fork collapse (Mehta & Haber 2014). Persistent, unrepaired DSBs are essentially oncogenic—driving translocations, mutation, loss of heterozygosity or gross chromosomal loss (Jeggo et al. 2015; Aplan 2006). Thus, in order to preserve genomic integrity, distinct DNA repair mechanisms have evolved to detect and subsequently repair DSBs—primarily non-homologous end joining (NHEJ) and homologous recombination (HR) (Chapman et al. 2012). Classical NHEJ (c-NHEJ) mediates the direct ligation of cleaved DNA with clean, adduct free ends or microhomology-dependent ligation with blocked ends—the latter of which is often error-prone (Chang et al. 2017). Within *S. cerevisiae* NHEJ initiates with the binding of Ku (a heterodimer of YKu70-YKu80) and the highly conserved Mre11-Rad50-Xrs2 (MRX-complex), which tether the DSB ends (Lisby et al. 2004; Mari et al. 2006). Subsequently, Dnl4 (DNA Ligase IV) and Lif1 (XRCC4) are recruited by Ku and MRX, and ligation is attempted (Chen et al. 2001). Further end processing may occur if ligation remains blocked including 5' flap cleavage by Rad27 (FEN1) (Wu et al. 1999) or gap filling by Pol4 (Wilson & Lieber 1999). In contrast, homologous recombination constitutes a largely error-free method that relies upon homology mediated repair from a template substrate (see: Section 1.2.1). The choice of

pathway is heavily influenced by cell cycle phase, cell type and species. Within *S. cerevisiae*, NHEJ predominates during interphase G0 and G1, while HR is preferred during and post-S phase when an intact sister chromatid template is present (Veuger et al. 2003). Transcriptional repression of Ku and Lif1 during meiosis reduces the capacity of meiotic cells to perform NHEJ, thereby heavily promoting usage of HR (Heyting et al. 1999; Valencia et al. 2001). In addition to repair pathways, surveillance mechanisms exist to ensure repair occurs before other processes, such as cell cycle progression, and coordinate cellular responses to damage. Eukaryotic life has, however, also co-opted these systems and repair pathways to drive specific processes. Intentional, programmed formation of DSBs occurs during V(D)J recombination (Franco et al. 2006), mating type switching within *S. cerevisiae* (Haber 2012) and meiosis in sexually reproducing organisms (Keeney & Neale 2006). DSB formation during meiosis is a particularly crucial process—inducing HR and in turn, the accurate segregation of chromosomes and exchange of genetic information, resulting in genetically diverse haploid gametes (see: Section 1.2.2).

1.2.1—Homologous Recombination (Overview)

HR utilises homologous sequences, present on sister chromatids, homologous chromosomes or at *cis*|*trans* repeat sequences, as templates for repair (Stahl 1996; Szostak et al. 1983). Error-free repair is predicated on the existence of a template containing perfect homology. Non-perfect homology, caused by single nucleotide polymorphisms (SNPs) or insertions/deletions (INDELs), can result in gene conversion, and thus loss of heterozygosity, through non-reciprocal exchange. Moreover, ectopic recombination between non-allelic regions can generate translocations and genomic rearrangements (Hastings 2010). In brief, the double-strand break repair (DSBR) model of HR depicts initiation by DSBs and subsequent lesion recognition by the Mre11-Rad50-Xrs2/NBS1 (MRX/N) complex—also utilised in NHEJ. Nucleolytic end processing of the 5' strand, mediated by MRX/N and Sae2 (CtIP in mammals), generates short 3' ssDNA tails/overhangs (Figure 1.2A) (Mimitou & Symington 2009). Under a mitotic context, further resection of the 5' strand is catalysed by exonuclease 1 (Exo1) and/or the Sgs1-Top3-Rmi1 (STR)-Dna2 complex, generating longer 3' ssDNA tails/overhangs of ~2-4kb (Krogh & Symington 2004). ssDNA serves as the primary substrate for HR.

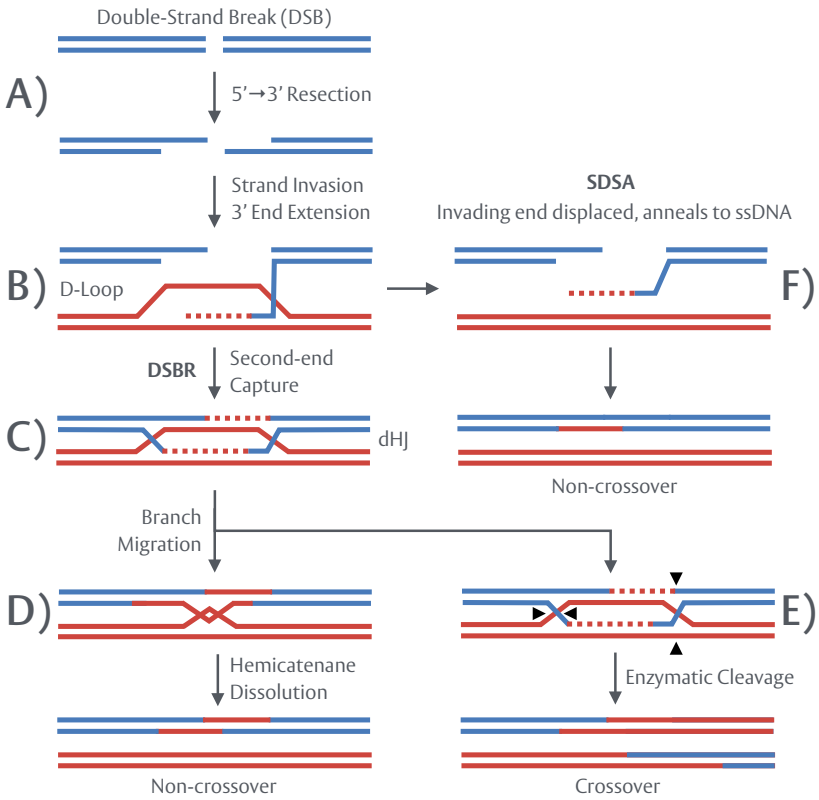


Figure 1.2. Models of homologous recombination

A) During repair, DNA double-strand breaks (DSBs) undergo 5'→3' resection, generating 3' ssDNA tails **B)** A single ssDNA tail invades a homologous sequence, priming leading strand synthesis and D-loop formation. In the DSBR model of HR, the non invasive 3' end anneals to the D-loop—a process known as second-end capture—priming a second round of leading strand synthesis. **C)** Ligation generates a double Holliday junction (dHJ) intermediate that is resolved in one of two ways. **D)** Branch migration can form a DNA hemicatenane which is subsequently dissolved to form a NCO. **E)** Alternatively, dHJs may be enzymatically cleaved (black arrows) in an asymmetric manner to form a crossover (CO) **F)** Synthesis-dependent strand annealing (SDSA) is an alternative pathway to DSBR and involves the displacement of the invading end prior to second-end capture, re-annealing, ligation and gap filling to produce a NCO. Separate chromatids are shown as red (parent A) and blue (parent B).

ssDNA tails are immediately bound by the trimeric replication protein A (RPA) to minimise formation of secondary structures and facilitate the loading of Rad51, a RecA-family recombinase, by Rad52 (Kowalczykowski et al. 1998). Rad51 subsequently polymerises, forming a nucleoprotein filament capable of homology searching, strand invasion and engagement with the repair template. Strand invasion creates a displacement loop (D-loop) (Figure 1.2B). Invading strands directly prime DNA synthesis, while the 3'-end from the opposing side of the DSB anneals to the D-loop (second-end capture), priming a second round of synthesis (Figure 1.2C) (San Filippo et al. 2008).

Subsequent to ligation, a branched double Holliday junction (dHJ) intermediate is formed which may be resolved in a multitude of ways, governing the outcome of HR: (i) Dissolution—a STR/BLM-dependent process involving the migration of dHJ structures toward one another and cleavage-dependent decatenation resulting in non-crossovers (NCOs) (Figure 1.2D) (Cejka et al. 2010; Wu & Hickson 2003) or (ii) enzymatic cleavage by structure-specific resolvases in specific orientations at each HJ, generating crossovers (COs) (Figure 1.2E) (Symington et al. 2014). Under the meiotic context, HR primarily differs in the source of the initiating DSB—a programmed rather than spontaneous event—and components of the machinery used (see: Sections 1.2.2+). Moreover, while COs rarely form in mitosis, around ~50% of DSBs repair as COs in meiosis (Chen et al. 2008; Mancera et al. 2008).

1.2.2—Meiotic DSB Formation—Spo11 & Accessory Proteins

Meiotic DSB formation requires the concerted effort of the evolutionarily conserved type II topoisomerase-like enzyme, Spo11, and nine additional Spo11-accessory factors (in *S. cerevisiae*): Mre11, Rad50, Xrs2/Nbs1, Rec102, Rec104, Ski8, Rec114, Mer2 and Mei4 (Arora et al. 2004; Lam & Keeney 2015; Maleki et al. 2007). Spo11 catalysis is thought to mimic that of topoisomerase—specifically occurring via attack of the phosphodiester bond through an activated tyrosine residue (Tyr135 in Spo11), resulting in a transient, covalently linked protein:DNA complex with Spo11 attached to the 5' end either side of the break (Bergerat et al. 1997; Keeney et al. 1997; Keeney & Kleckner 1995; Liu et al. 1995). Akin to its homolog, archaeal TopVIA—the catalytic subunit of the

type-II DNA topoisomerase most closely related to Spo11—dimerisation of Spo11 is essential for meiotic DSB formation and is dependent upon a Rec102-Rec104 complex (Robert et al. 2016; Vrielynck et al. 2016). Each monomer within the Spo11 homodimer cleaves one strand, collectively forming the DSB with a 2bp 5' overhang (Liu et al. 1995; Sasanuma et al. 2007). The Rec102-Rec104 sub-complex is further required for efficient nuclear retention of Spo11 and DNA binding (Kee et al. 2004). Repair of Spo11-dependent DSBs necessitates removal of the Spo11 moiety. Removal of Spo11 is accomplished through the nucleolytic activity of Mre11 which, in cooperation with Sae2 (CtIP), creates a nick distal to the DSB —initiating loading of Exo1 and bidirectional exonucleolytic resection by Mre11 in the 3'->5' direction toward the break and by Exo1 in the 5'->3' direction away from the break forming ssDNA tails of ~800bp (as opposed to ~2-4kb in mitosis) (Garcia et al. 2011; Symington et al. 2014). As a result of these processes, Spo11 is released covalently attached to a short ~24-40bp or ~10-15bp "Spo11 oligonucleotide" (Neale et al. 2005), cleaning the end for HR. Rec114-Mei4-Mer2 form the conserved RMM subcomplex and co-localise to the chromosomal axis (Maleki et al. 2007; Panizza et al. 2011). While Mer2 serves as a major regulation target, coupling DSB formation with DNA replication (see Section 1.4.5) and mediating interactions required for DSB formation (see Section 1.2.10), the way in which Rec114 and Mei4 are essential for DSB formation remains unclear.

1.2.3—Genome-Wide Mapping of Spo11 DSBs

Genome-wide mapping of Spo11-dependent DSBs has yielded vast insights into the process of meiotic DSB formation. Initial attempts to map the spatial positions of DSB formation at the genome scale focused on the use of microarrays to detect Spo11-associated fragments or ssDNA produced within *dmc1*Δ backgrounds—within which HR stalls due to a lack of strand invasion (Buhler et al. 2007; Blitzblau et al. 2007; Gerton et al. 2000). Microarray usage however suffers from limited resolution, dynamic range and quantification. To alleviate these problems, modern immunoprecipitation techniques have been developed to exploit two key features of meiotic DSB formation and HR—the generation of Spo11-oligos and ssDNA. Affinity of Rad51 and Dmc1 for ssDNA has allowed ChIP-based analysis of meiotic DSBs at a high resolution within *M. musculus* (Smagulova et al. 2011; Brick et al. 2012) and *H. sapiens* (Khil et al. 2012; Pratto et al. 2014)—a

technique termed ssDNA sequencing (SSDS). Another widely employed and sensitive method is Spo11-oligo sequencing—which relies upon immunoprecipitation of Spo11 and capture of the bound oligonucleotides for next-generation sequencing (NGS) (Neale et al. 2005; Pan et al. 2011). This technique has been successfully applied in *S. cerevisiae* (Pan et al. 2011), *S. pombe* (Fowler et al. 2014) and *M. musculus* (Lange et al. 2011; Lange et al. 2016; Kauppi et al. 2013). However, while Spo11-oligo sequencing has greatly improved the resolution of mapping, poly(G)-tailing of Spo11-oligonucleotides—a required step of the technique—produces base pair discrepancies and coordinate ambiguity at DSB sites where a genomic 5' cytosine is present (Pan et al. 2011; Fowler et al. 2014). Moreover, the shorter ~10-15bp Spo11-oligonucleotides are lost due to poor recovery, alignment and multi-mapping ambiguities. The existence of two different Spo11-oligonucleotide species is thought to be due to asymmetric nucleolytic cleavage by Mre11 on either side of the DSB (Neale et al. 2005; Garcia et al. 2011). Loss of shorter Spo11-oligonucleotides may thus result in an incomplete picture of DSB formation, mapping only one side of the break. Indeed, a multitude of mapped hotspots retain information on only one strand (Pan et al. 2011). Further refinement of Spo11 mapping technologies is thus required to address these issues (see: Chapter 3).

1.2.4—Outcome of Homologous Recombination—NCO vs. CO

Within mitosis, subsequent to resection, Rad51 slowly replaces the ssDNA-bound RPA to catalyse strand invasion and homology search. While Rad51 still functions within meiosis, an additional meiosis-specific but related RecA-family recombinase, Dmc1, is also required (Bishop et al. 1992). Following D-loop formation—the stage at which CO-NCO differentiation is thought to occur (Bishop & Zickler 2004; Börner et al. 2004)—a multitude of resolution mechanics can come into play to govern the outcome of HR: (i) If the invading, single stranded molecule reanneals with the originally broken chromatid following extension, further DNA synthesis and ligation can occur—a process known as synthesis-dependent strand annealing (SDSA) (Figure 1.2F). This pathway is responsible for the vast majority of non-crossovers (NCOs) within *S. cerevisiae* (Pâques & Haber 1999; McMahill et al. 2007; Martini et al. 2011) (ii) if SDSA does not occur, second-end capture of the D-loop generates a joint molecule between the two participating chromatids. Structure-specific

endonucleases may cleave these singular junctions, generating a crossover (CO) (Schwartz & Heyer 2011) (iii) if the initial joint molecules are not resolved, double Holliday junctions (dHJs) form (see: Figure 1.2C) (Bzymek et al. 2010). dHJ resolution is heavily biased toward CO formation (accomplished via enzymatic cleavage) (Allers & Lichten 2001), however, STR/BLM-dependent dissolution can still produce NCOs from these structures , albeit at low rates (see: Figure 1.2D, E) (Wu & Hickson 2003; Cejka et al. 2010).

1.2.5—Inter-homologue Bias

Following meiotic S phase, three allelic templates are available (one sister, two homologs) and yet, despite the spatial proximity of the sister, ~70-90% of meiotic DSBs repair off the homologue within *S. cerevisiae* (Goldfarb & Lichten 2010)—a phenomenon known as inter-homologue bias (IHB) (Schwacha & Kleckner 1994; Schwacha & Kleckner 1997). This is in contrast to the "default" mitotic bias, where non-deleterious inter-sister recombination predominates (Kadyk & Hartwell 1992). Such a meiotic bias is crucial to ensuring connections are formed between homologous chromosomes and exchange of genetic material occurs. IHB relies upon the DNA damage response (DDR) kinases Tel1 and Mec1 (see Section 1.3.1) (Carballo et al. 2008), the structural transducers Hop1 (Niu et al. 2005) and Red1 (Schwacha & Kleckner 1997), and the meiosis-specific effector kinase, Mek1 (Niu et al. 2007; Callender & Hollingsworth 2010) as well as HR components including Rad51 (Schwacha & Kleckner 1997; Zierhut et al. 2004). Notably, when Rad51 is mutated, Dmc1-dependent inter-sister recombination occurs. The exact mechanics of IHB, however, remain unclear. A "kinetic impediment" model has been proposed whereby Tel1/Mec1-dependent phosphorylation of Hop1 occurs in response to DSB formation and subsequent activation of Mek1 locally slows the otherwise high rate of inter-sister recombination (Lao & Hunter 2010).

1.2.6—Crossover Formation

Crossovers (COs) constitute a central tenet to meiotic progression and subsequent generation of genetic diversity—forming physical links between homologous chromosomes, driving accurate segregation during anaphase I, and shuffling maternal/paternal markers on a potentially large scale.

COs are subject to strict regulatory control, including the obligate crossover rule (crossover assurance) which ensures formation of at least one CO per chromosome (Jones & Franklin 2006). Canonical CO formation not only requires a particular resolution pathway (see Section 1.2.5), but also a distinct group of genes, termed the ZMM family. This family includes the meiosis-specific MutS homologs of the DNA mismatch repair (MMR) family, Msh4-Msh5, which stabilise dHJs, the Mer3 helicase, the structural/axis protein Zip1, Zip2, Zip4 and the SUMO or ubiquitin E3 ligase, Zip3 (Hollingsworth et al. 1995; Tsubouchi et al. 2006; Lynn et al. 2007; Sym et al. 1993). Collectively these proteins orchestrate the formation of class I crossovers. The Msh4-Msh5 heterodimer (MutSγ) is directly recruited to ~30-50% of DSB repair sites and is thought to form a sliding clamp encircling dHJ substrates, nucleating ZMM foci (Manhart & Alani 2016). Crucially, MutSγ subsequently recruits the non-ZMM and putative endonuclease Mlh1-Mlh3 complex (MutLγ)—the central class I HJ resolvase (Manhart & Alani 2016). The precise activity of Mlh1-Mlh3 remains unclear however it appears to directly cleave dHJs in conjunction with Exo1 and Sgs1, leading to repair completion (Zakharyevich et al. 2012; Rogacheva et al. 2014).

Although the class I pathway accounts for the majority of COs in most organisms, a secondary minor class II pathway also exists, dependent upon the structure-specific endonuclease Mus81-Mms4 (MUS81-EME1 in mammals) (Boddy et al. 2001)—generating ~5%, ~5-10% and 15-35% of COs within *A. thaliana* (Higgins et al. 2008), *M. musculus* (Holloway et al. 2008) and *S. cerevisiae* (de los Santos et al. 2003) respectively. Mus81 appears to act with Sgs1/BLM to resolve complex multichromatid joint molecule intermediates that have arisen through atypical HR, such as secondary strand invasions (Jessop & Lichten 2008), and generate class II COs.

1.2.7—Crossover Designation

Crossover designation refers to the process by which a DSB site is selected for canonical class I CO formation, as opposed to NCO or class II CO formation. NCOs, class I and class II COs all arise from the same recombination rich regions and thus HR outcome is not predetermined according to genomic position alone—but rather an active process. A critical designation step appears to be a

transition from a high number of early MSH4-MSH5 (MutSγ) complexes during zygotene, to a mature lower number of recruited and late MLH1-MLH3 (MutLγ) complexes (e.g. ~150-MSH4-MSH5 vs. ~22-24 MLH1-MLH3 per cell in *M. musculus*) (Gray & Cohen 2016). The mechanics underscoring selective, subset binding of MLH1-MLH3 to MSH4-MSH5 remain unclear. Interestingly, late wave reductions in MSH4-MSH5 are still observed within *MLH3*$^{-/-}$ null *M. musculus* mutants, suggesting that, firstly, MutLγ loading is not essential for the disassembly of unused MutSγ sites or the stabilisation of selected sites, and secondly, CO designation relies on an alternative process—such as the proposed post-translational modification of target proteins (Gray & Cohen 2016). For example, within *S. cerevisiae*, the E3 SUMO/Ubiquitin ligase, Zip3, co-localises with MutSγ-complexes and directly interacts with MutSγ-components (Agarwal & Roeder 2000)—as do the Zip3 equivalent proteins ZHP-3 and RNF212 in *C. elegans* (Bhalla et al. 2008) and *M. musculus* (Reynolds et al. 2013) respectively. Moreover, mutation of Zip3, ZHP-3 and RNF212 significantly perturbs CO formation, resulting in chromosomal non-disjunction, infertility and spore inviability (Agarwal & Roeder 2000; Jantsch et al. 2004; Reynolds et al. 2013). Targeted and coordinated SUMOylation or ubiquitination of pro-class I CO factors may thus coordinate CO designation but the precise mechanisms remain a mystery.

1.2.8—Genome-Wide Mapping of COs/NCOs

Designation processes permit the generation of NCOs, class I COs and class II COs from the underlying precursor array of DSBs—it is therefore necessary to explicitly assess recombination outcome as this cannot be directly implied from DSB mapping alone (see: Section 1.2.4). Moreover, Spo11-DSB mapping technologies generate population averaged data, excluding the possibility of assessing recombination on a per cell basis. Classically, post-meiotic linkage analysis was employed to detect crossovers through co-inheritance of linked heterozygous markers—where a loss of linkage indicates the occurrence of an intervening CO residing between the test loci (Hunt Morgan 1916). However, modern techniques based around ChIP and next-generation sequencing (NGS) have largely replaced classical analyses. Tetrad analysis—the mapping of recombination within all

four post-meiotic daughter cells, successfully applied within *S. cerevisiae* (Mancera et al. 2008; Martini et al. 2011; Oke et al. 2014) and *Z. mays* (Li et al. 2015)—is a powerful, high resolution and genome-wide technique capable of (i) distinguishing COs/NCOs (ii) identifying atypical events such as 3/4-strand COs and complex gene conversion tracts and (iii) generating accurate positional information of where recombination took place in a single cell (Lichten 2014). Tetrad analysis typically relies upon the use of hybrid strains—derived from two parental strains of divergent origin—which contain substantial polymorphisms (SNP/INDEL) between them and NGS or microarray based detection of the variants to ascribe parental origin to any given loci, on any given chromatid. Single cell analysis of recombination has also been accomplished within *H. sapiens* via use of multiple annealing and looping-based amplification cycles (MALBEC) of template DNA prior to genotyping (Hou et al. 2013; Lu et al. 2012).

1.2.9—Meiotic Chromatin Architecture

Meiotic chromosomes display a unique architecture—organising into linear arrays of protruding chromatin loops, each basally attached to a proteinaceous axis, known as the axial or lateral element, via AT rich axial association sites (Figure 1.3A) (Blat et al. 2002; Kleckner 2006; Borde & de Massy 2013). Within *S. cerevisiae*, these loops are ~10-15kb in size (12.1kb average) (Ito et al. 2014). Loops appear to increase or decrease in size to accommodate genomes of variable length as opposed to changes in total loop count (Blat et al. 2002; Kleckner 2006; Novak et al. 2008; Kauppi et al. 2011). Synapsis, the HR-dependent pairing of homologous chromosomes, requires additional structural components. As meiosis progresses, homologous chromosomes become progressively connected through the central region (CR)—a large, proteinaceous zipper-like structure whose assembly completes the tripartite scaffold known as the synaptonemal complex (SC) (Page & Hawley 2004). Upon completion, the SC runs the entire length of the paired chromosomes. Key meiotic processes, such as CO formation, therefore do not occur in bare isolation but rather within the context of these complex structures.

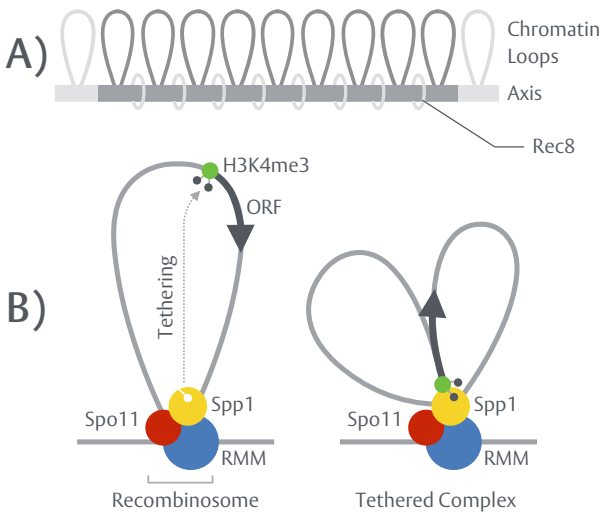


Figure 1.3. Meiotic chromatin architecture and tethered DSB formation

A) Meiotic chromosomes are organised into linear arrays of protruding chromatin loops, ~12-15kb in size within *S. cerevisiae*, attached to a proteinaceous axis comprised of, amongst other factors, Rec114-Mei4-Mer2 (RMM), Hop1, Red1 and the meiosis-specific cohesin component, Rec8. **B)** Within this structural arrangement, hotspots predominately reside within loop regions while, rather counterintuitively, the machinery essential for the regulation and enzymatic induction of DSBs is bound to the axis. To explain this discrepancy, the tethered loop axis model proposes that Spp1, bound to Mer2 on the axis, acts as a "molecular bridge"—docking H3K4me3 via a PHD-finger domain and tethering the loop for DSB formation.

Moreover, these structures are functionally active and intimately linked to meiosis—interplaying with recombination in a multitude of ways. While the structural organisation of the SC remains conserved, the constituent proteins widely vary between species. Within *S. cerevisiae*, the meiosis-specific HORMA (HOp1-Rev7-MAd2)-domain containing protein, Hop1, is a key axial element (AE) protein localised along meiotic chromosomes in complex with a second AE factor, Red1, during early prophase I (prior to DSB formation) (Page & Hawley 2004). Hop1 facilitates DSB formation through recruitment of the Spo11-accessory proteins Rec114, Mei4 and Mer2 (RMM complex) (Panizza et al. 2011). Indeed, DSB frequency is reduced to ~10% of WT levels within *hop1*Δ strains (Mao-Draayer et al. 1996; Woltering et al. 2000). RMM is therefore also enriched at axial-association sites (Panizza et al. 2011). Rec8, a meiosis-specific component of the tetrameric Smc1-Smc3-Rec8-Scc3 cohesin complex (collectively referred to as Rec8), is thought to demarcate loop boundaries, modulate axis construction and regulate DSB formation (Novak et al. 2008; Glynn et al. 2004) (see: Figure 1.3A). Specifically, induction of DSBs proximal to Rec8 binding sites is notably inefficient (Ito et al. 2014), and removal of Rec8 profoundly alters both Spo11 binding patterns and DSB distribution (Kugou et al. 2009; Blat et al. 2002; Sun et al. 2015; Kim et al. 2010).

Unexpectedly, DSB formation predominately occurs at sites residing within loop domains while, rather counterintuitively and as discussed above, the machinery essential for the regulation and enzymatic induction of DSBs (e.g. RMM) is bound to the axis. To explain this paradoxical discrepancy, the tethered-loop axis model proposes that, in *S. cerevisiae*, Spp1—a PHD finger domain protein and COMPASS (chromatin-modifying complex) family member, interacts with (i) H3K4me3—a histone modification enriched at sites of *S. cerevisiae* DSB formation and (ii) axial factors (Mer2), bridging the two entities together and effectively "tethering" the loop to the axis for DSB formation (Figure 1.3B) (Sommermeyer et al. 2013; Acquaviva et al. 2013; Borde et al. 2009; Tischfield & Keeney 2012). Consistent with this model, Spp1 is required for wild type levels of DSB formation (Sommermeyer et al. 2013).

1.2.10—Controlling Meiosis

The inherently dangerous but essential act of DSB formation is subject to multiple forms of stringent and self corrective regulation that collectively ensures fruitful and appropriate levels of genetic exchange without risk to cellular survival. The complex and multistep nature of meiotic processes affords many potential points of regulation. Perhaps unsurprisingly, meiotic regulation centres on the key processes of DSB and CO formation. However, this regulation is in of itself complex, multifaceted and can take on multiple forms including homeostatic, spatial and temporal. In the following sections, these key regulatory pathways are discussed.

1.3—Quantitative Control

1.3.1—DNA-Damage Response (DDR) & Meiotic Checkpoints

In response to critical lesions (e.g. DSBs), the DNA damage response (DDR)—a checkpoint pathway—couples deactivation of cell cycle progress with the activation of repair pathways in order to allow DNA repair to occur (Ciccia & Elledge 2010). As DNA damage resides at the heart of meiosis (see Section 1.2.2), it is unsurprising to find that the central DNA damage response (DDR) kinases, ataxia-telangiectasia mutated (ATM)/RAD3-related (ATR) and respective orthologues (Tel1/Mec1 in *S. cerevisiae*), feature prominently in the meiotic landscape (Cooper et al. 2014; MacQueen & Hochwagen 2011). ATM/ATR are highly conserved members of the phosphoinositide-3-kinase-related protein kinase (PIKK) family which invoke DDR responses through phosphorylation of target proteins at hydrophobic-X-hydrophobic-[S/T]-Q consensus sequences (S/T-Q or SQ/TQ sites) (Kim et al. 1999). Indeed, a structural hallmark of DNA damage response proteins are large ~100aa S/T-Q cluster domains containing multiple modification sites (Traven & Heierhorst 2005). Within mitotic cells, ATR/ATM primarily signal through CHK1 or CHK2 to suppress the activity of CDKs—master regulators of cell cycle progression, arresting the cell at G1/S or G2/M—as well as target a wide array of other repair proteins (Bartek & Lukas 2007; Chen et al. 2010; Matsuoka et al. 2007; Smolka et al. 2007).

Tel1^{ATM}

ATM (Tel1 in *S. cerevisiae*) primarily signals via CHK2 and is recruited to DSB ends via the Mre11–Rad50–Xrs1/Nbs1 complex (MRX/N), whose Mre11 subunit exhibits direct DSB binding activity (Figure 1.4A) (Maréchal & Zou 2013). ATM/Tel1 interacts with MRX/N via the Xrs2/Nbs1 subunit (Nakada et al. 2003; You et al. 2005) and upon recruitment, ATM is activated through monomerisation and auto-phosphorylation (Bakkenist & Kastan 2003). Loss of Tel1 signalling activity is concomitant with 5'→3' resection as its DSB substrate is eroded (Mantiero et al. 2007).

Mec1^{ATR}

Akin to mitotic cycles, checkpoint mechanisms also exist within meiosis. The pachytene checkpoint, operating during prophase I, surveys the status of DSB repair and homolog synapsis in order to arrest cells until such processes are completed (MacQueen & Hochwagen 2011; Roeder & Bailis 2000). Given that premature anaphase I entry is lethal, this checkpoint is of critical importance (Lydall et al. 1996). Transmission of pachytene checkpoint signals primarily depends upon the ssDNA-sensing ATR system comprising ATR, the RAD9–RAD1–HUS1 (9–1–1) clamp complex, ATRIP and the RAD17 clamp loader; respectively designated Mec1, Rad17, Mec3, Ddc1, Ddc2 and Rad24 in *S. cerevisiae* (Figure 1.4B) (Lydall et al. 1996; Majka et al. 2006; Majka & Burgers 2003). ssDNA is exposed proceeding 5'→3' resection during HR, leading to the recruitment of ATRIP and the loading of the 9-1-1-complex at ssDNA:dsDNA junctions by RAD17 (Maréchal & Zou 2013). RAD17 in turn stimulates ATR activity. A central target of the checkpoint in *S. cerevisiae* is Ndt80, a meiosis-specific transcription factor responsible for exit from pachytene into anaphase I via the induction of key genes involved in cell cycle progression and Holliday junction resolution (Xu et al. 1995; Winter 2012; Allers & Lichten 2001). Checkpoint signals inhibit Ndt80 via suppression of its hyper-phosphorylation—a modification required for its transcription factor activity—ultimately arresting cells within prophase I (Tung et al. 2000). Ndt80 shutdown critically allocates the cell an extended period of time to generate and subsequently repair DSBs without interruption by unscheduled anaphase I entry.

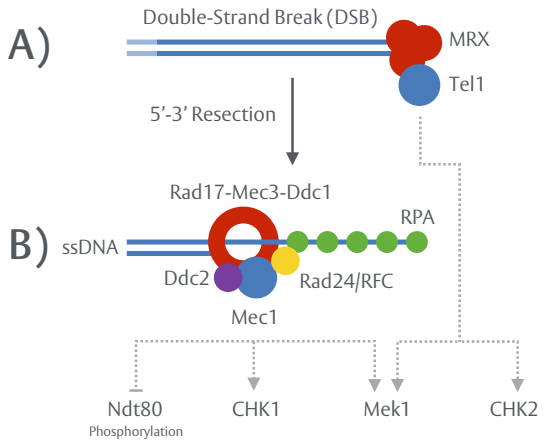


Figure 1.4. ATR/ATM activation during meiotic HR

A) DSB ends, produced by Spo11, are detected by the Mre11-Rad50-Xrs2/NBS1 (MRX/N) complex which in turn recruits and activates the DNA damage response (DDR) kinase, Tel1 (ATM). Tel1^{ATM} primarily signals through CHK2. Loss of Tel1 signalling is concomitant with the formation of ssDNA, produced by enzymatic 5'→3' resection. **B)** ssDNA is detected by additional components of the DDR, namely Mec1 (ATR), the Rad17-Mec3-Ddc1 (Rad9-Hus1-Rad1 9-1-1) clamp complex and the Rad24 (Rad17) clamp loader. Activation of the DDR promotes further DSB formation by inhibiting the hyper-phosphorylation of Ndt80, resulting in transient cell cycle arrest. Mec1^{ATR} primarily signals through CHK1 but also integrates into the meiosis specific kinase, Mek1, along with Tel1^{ATM}. Only one side of the DSB is shown for clarity.

In addition to this, a mitotic-like replication checkpoint also appears to function within pre-meiotic S phase in *S. cerevisiae* and requires Mec1^{ATR}-signalling (Blitzblau & Hochwagen 2013). The pachytene checkpoint also relies upon ATM signalling, albeit to a lesser extent (MacQueen & Hochwagen 2011). A meiosis-specific paralogue of CHK2, known as Mek1, has been identified within *S. cerevisiae* (Rockmill & Roeder 1991; Usui et al. 2001). Both Mec1 and Tel1 signals feed into Mek1 activation during meiosis, integrating multiple stimuli into a common target (Usui et al. 2001; Refolio et al. 2011; Carballo et al. 2008). Ablation of Mek1 activity reduces the viability of *S. cerevisiae* spores (the haploid products of yeast meiosis) suggesting that Mek1 is a major effector of Mec1/Tel1 meiotic activity (Rockmill & Roeder 1991; Wan et al. 2004).

In a number of *S. cerevisiae* lab strains, prophase arrest is additionally regulated by the evolutionarily conserved hexameric ATPase, Pch2 (TRIP13 in mammals). Pch2/TRIP13 appears to promote the remodelling of the HORMA domain-containing meiotic chromosome component, Hop1 (HORMAD1/2 in mammals), found on the axis, which are targets of the ATM/ATR response—thereby aiding prophase arrest in response to defects in chromosome pairing and synapsis (Carballo et al. 2008; Wojtasz et al. 2009).

1.3.2—DSB Homeostasis

While the number of DSBs typically formed per meiotic cycle differs between species, such differences do not significantly scale with genome size (Buhler et al. 2007; Pan et al. 2011; Joyce et al. 2011; Kauppi et al. 2013; Lange et al. 2011). Moreover, DSB frequency is maintained at a moderate level despite an apparent excess of Spo11 protein (Neale et al. 2005), hinting at strict quantitative control. This phenomenon—termed DSB homeostasis (Carballo et al. 2013; Robine et al. 2007)—is proposed to maintain levels of DSBs within genetically encoded ranges in order to prevent the deleterious effects associated with excessive or insufficient DSB formation (Lange et al. 2011; Gray et al. 2013; Rockmill et al. 2013). DSB homeostasis is intimately linked with the meiotic pachytene checkpoint and ATR/ATM activation, as detailed in the following sections.

1.3.3—Positive Regulation of DSB Formation

A Mec1, Rad24 and Rad17-dependent positive feedback loop that promotes DSB formation under conditions of suboptimal DSB catalysis appears to exist within *S. cerevisiae*, contributing to DSB homeostasis (Gray et al. 2013; Argunhan et al. 2013). Specifically, strains carrying hypomorphic forms of Spo11 (*spo11–HA* and *spo11–D290A*) or mutation of the *PCH2* gene display significantly reduced levels of DSB formation when Spo11 removal and ssDNA resection are blocked. By contrast, only minor reductions in DSB frequencies and spore viability are observed in cells capable of removing Spo11 from DSB ends to expose ssDNA. This apparent ability of resection proficient cells to compensate for reduced Spo11 activity is abolished when components of the ATR branch of the checkpoint pathway (Mec1, Rad24 and Rad17) are mutated, resulting in severely reduced DSB frequencies and synergistic reductions in spore viability despite the fact that *spo11-HA* or *spo11-D290A* alone display no appreciable reduction (Gray et al. 2013; Argunhan et al. 2013). Collectively, these results suggest that the transient formation of ssDNA at meiotic DSBs creates a signal—transduced by $Mec1^{ATR}$—that enables further Spo11-DSB catalysis within meiotic cells. Indeed, deletion of Ndt80 within *rad24Δspo11-HA/D290A* or *rad17Δspo11-HA/D290A* backgrounds was found to significantly rescue these defects in DSB formation (Gray et al. 2013; Argunhan et al. 2013), and even moderate extension to meiotic prophase, mediated by transient depletion of Ndt80 activity, is sufficient to restore spore viability (Gray et al. 2013). While these observations provide an attractive model for the positive regulation of DSB formation, contradictory data suggesting Rad17 is a negative regulator have also been reported (Argunhan et al. 2013). Furthermore, the synergistic effect of a Spo11 hypomorph within a *rad17*Δ background is less pronounced than within *mec1*Δ or *rad24*Δ, resulting in a smaller reduction in DSB formation (Gray et al. 2013). Taken together, these observations raise the possibility that Rad17 mediates both positive and negative regulation in Mec1/Rad24- dependent and Mec1/Rad24-independent manners respectively. Deletion of Rad17 could thus eradicate a positive effect, producing a synergistic reduction in DSB frequency that is stunted due to the removal of a negative effect.

1.3.4—Negative Regulation of DSB Formation

Tel1^{ATM} is primarily implicated in the negative regulation of DSB formation and key examples of this have been recently observed within *M. musculus*, *S. cerevisiae* and *D. melanogaster* model systems. DSB formation during prophase I, as ascertained by assessing Spo11-oligonucleotide levels, significantly increases within *ATM*$^{-/-}$ null mice compared to wild type (Lange et al. 2011). It is known that *ATM*$^{-/-}$ mice display severe meiotic defects that ultimately result in infertility (Barlow et al. 1996; Xu et al. 1996; Barlow et al. 1998). In a striking contrast to the loss of Mec1 activity in *S. cerevisiae*, which displays a synergistic defect with Spo11 hypomorphs (see Section 1.3.3), reduction in DSB formation via means of Spo11 heterozygosity largely rescues the defects normally observed in *ATM*$^{-/-}$ mice—attributing the *ATM*$^{-/-}$ meiotic phenotype to the excessive formation of DSBs (Lange et al. 2011). Consistent with the idea of increased DSB formation, CO frequencies are also increased within *ATM*$^{-/-}$ mice (Barchi et al. 2008). Interestingly, ataxia telangiectasia (AT) patients, who contain a mutated form of *ATM*, also display infertility—hinting that ATM may play a similar meiotic role within humans (Boder 1975). Comparable mechanisms also appear to function within *D. melanogaster* and *S. cerevisiae*. Inactivation of the *D. melanogaster* ATM homologue, tefu, results in substantial increases in γH2AV foci formation within both nurse cells and oocytes (Joyce et al. 2011). As γH2AV is a functional homolog of mammalian γH2AX, which forms in response to DSBs, these observations indirectly implicate tefu within the negative regulation of DSB formation (Joyce et al. 2011; Madigan et al. 2002). Tel1/ATM is thought to negatively regulate DSB formation through the phenomenon of DSB interference, discussed later (see Section 1.4.2).

1.3.5—Crossover Homeostasis

A non-linear relationship between DSB frequency and CO frequency exists—termed CO homeostasis. CO homeostasis seemingly buffers against decreases or increases in DSB number to maintain a pre-established and genetically encoded CO count, which varies between species (e.g. ~80 in *S. cerevisiae* and ~26 in male *M. musculus*) (Martini et al. 2006; Mancera et al. 2008; Holloway et al. 2008). Notably, CO homeostasis has been observed within multiple species

including *S. cerevisiae* (Martini et al. 2006), *S. pombe* (Kan et al. 2011), *C. elegans* (Yokoo et al. 2012), *M. musculus* (Cole et al. 2012) and *Z. mays* (Sidhu et al. 2015) suggesting it is a highly conserved process. For example, *M. musculus* strains containing multiple copies of Spo11 exhibit increased DSB formation but not significantly increased MLH1 foci—indicative of no increase in class I CO formation (Cole et al. 2012). Similarly, mutations within the *S. pombe* Spo11 ortholog, Rec12, reduce DSB levels but CO levels are maintained above that governed by the obligate rule alone (Kan et al. 2011) and equivalent findings exist within *S. cerevisiae* strains containing hypomorphic forms of Spo11 (Martini et al. 2006). However, in these studies, class II counts were not assessed in parallel. Moreover, deactivation of class II CO formation (via *mus81*Δ) results in a compensatory increase in class I formation (Holloway et al. 2008)—suggesting cross-talk exists between the class I and class II pathways. Furthermore, mutations in *TEL1*^{ATM} increase DSB frequency and specifically, class II CO formation (Anderson et al. 2015). Therefore, should class II COs prove insensitive to homeostatic mechanisms, excess DSB precursors may simply be shunted into the Mus81-Mms4-dependent pathway—resulting in excess CO formation. It thus remains unclear how or if an upper homeostatic limit is established or if CO homeostasis may be more appropriately termed class I CO homeostasis. Much like the negative impact Tel1^{ATM} has upon DSB formation is primarily ascribed to DSB interference (see Section 1.4.2), an upper homeostatic limit on CO formation may be imposed by CO interference (see: Section 1.5.1)

1.4—Spatiotemporal Control of DSBs

The genome-wide distribution of Spo11-dependent DSBs is non-random and subject to multifaceted, layered control. At fine scale resolutions, DSBs concentrate within discrete, scattered and non-randomly distributed regions of permissiveness terms “DSB hotspots” (see: Figure 1.5—Bottom) (~3600 hotspots exist within *S. cerevisiae* haploid equating to ~14,400 unique locations within *S. cerevisiae* replicated diploids, and ~10,000-40,000 within mammals) (Khil et al. 2012; Pan et al. 2011; Pratto et al. 2014; Smagulova et al. 2011; Fowler et al. 2014). An ever growing collection of factors have been shown to influence the designation of a hotspot via a multitude of mechanisms

including Spo11 recruitment and the promotion of cleavage susceptibility (de Massy 2013). Only a small subset of hotspots are utilised per meiosis (~150-200 DSBs/cell in *S. cerevisiae*) and several extra layers of regulation exist to ensure even spacing of meiotic events across all chromatids—a process referred to in this thesis as spatial regulation (Pan et al. 2011; Cooper et al. 2016).

Mechanisms of spatial regulation, at the level of DSB formation, do not operate in isolation but rather coalesce into a multifaceted system, progressively layering to guide the DSB distribution both proactively and reactively in DSB-independent and -dependent manners respectively (Figure 1.5). Understanding how cells utilise this hierarchy of processes to spatially guide DSB formation is of critical importance: not only can this "DSB patterning" system potentially protect the genomic integrity of the germ line by suppressing aberrant or excessive DSB formation, but it also constructs a foundation—the genome-wide DSB distribution—upon which all downstream processes build, thereby influencing not just the identity of recombinant chromosomes arising from a given individual, but also the rates and distribution of genetic change arising long term within a population. In the following sections, each process shown in (Figure 1.5) is discussed further.

1.4.1—Hotspot Designation

Spo11 itself possess only moderate ability to discriminate between DNA sequences (Pan et al. 2011; Murakami & Nicolas 2009; Prieler et al. 2005) and yet, preferential formation of DSBs within discrete windows of opportunity (DSB hotspots) distributed non-randomly, while not universally conserved, is a distinctive feature of meiosis in many organisms (Keeney et al. 2014; de Massy 2013). The historical analysis of recombination and advent of high resolution mapping technology has revealed a wealth of information in answer to this apparent contradiction (Choi & Henderson 2015) (see: Section 1.2.3). Remarkably, a molecular system to explicitly govern and direct hotspot designation does not appear to be essential, but rather meiotic recombination is able to "piggyback" upon factors embedded within the organisational code of chromosomes—whose primary functions are to orchestrate unrelated cellular processes including gene regulation, transcription and DNA replication (Pan et al. 2011; de Massy 2013).

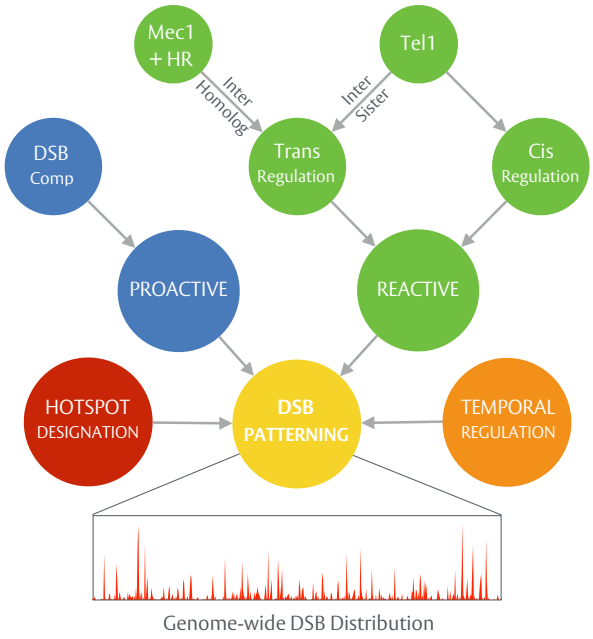


Figure 1.5. Hierarchical DSB patterning

Regulation of DSB position during prophase I is achieved by means of a hierarchical collection of processes via three major nodes: hotspot designation, DSB-independent proactive regulation and DSB-dependent reactive regulation. Rather than acting in isolation, these processes interconnect—sculpting the final DSB distribution with a high degree of complexity. Moreover, temporal regulation—such as synapsis-dependent shutdown of DSB formation or the coupling of DSB formation to DNA replication—may also contribute to spatial regulation in a generalised manner (see text for further details).

Within many species, no single factor is key and instead it is the co-occurrence of certain "gatekeeper" factors in a specific fashion that seemingly unlocks the potential for a region to accommodate DSB formation (Figure 1.6A). The dependency of DSB formation upon factors not specifically designed to guide recombination may also go some way toward explaining a certain peculiarity of meiosis: while a handful of common principles exist, no universal, cross-species mechanism underpins hotspot designation and distinctions in strategy are observed between species as well as across evolutionary classes (Lam & Keeney 2015; de Massy 2013; Nishant & Rao 2006).

Many of the concepts surrounding hotspot designation (outlined above) are particularly well illustrated within *S. cerevisiae* which relies upon a hierarchical collection of low impact factors (Figure 1.6B) (Pan et al. 2011; Lam & Keeney 2015; de Massy 2013)—that is to say, no one factor is sufficient. Of particular prominence is the striking correlation of yeast hotspots with regions of nucleosomal depletion (NDRs), a genomic feature primarily associated with promoter regions (Pan et al. 2011; Kaplan et al. 2009; Fan & Petes 1996). However, a significant proportion of detectable NDRs are not associated with robust DSB activity, revealing an insufficiency of chromatin accessibility as an isolated gatekeeper (Pan et al. 2011). Furthermore, incomplete correlations are observed between Spo11 binding sites and subsequent DSB positions within *S. cerevisiae* (50-55% overlap) (Kugou et al. 2009), Spo11 fusion constructs are incapable of inducing DSB formation at all targeted loci (Fukuda et al. 2008; Robine et al. 2007) and the localisation of Spo11 to meiotic chromosomes appears to be a distinct process from that of Spo11 activation (Prieler et al. 2005), collectively suggesting that gatekeeper factors not only facilitate simplistic substrate-enzyme interaction but also create an environment favourable for catalysis. The influence of gatekeeper factors may also extend beyond that of local effects.

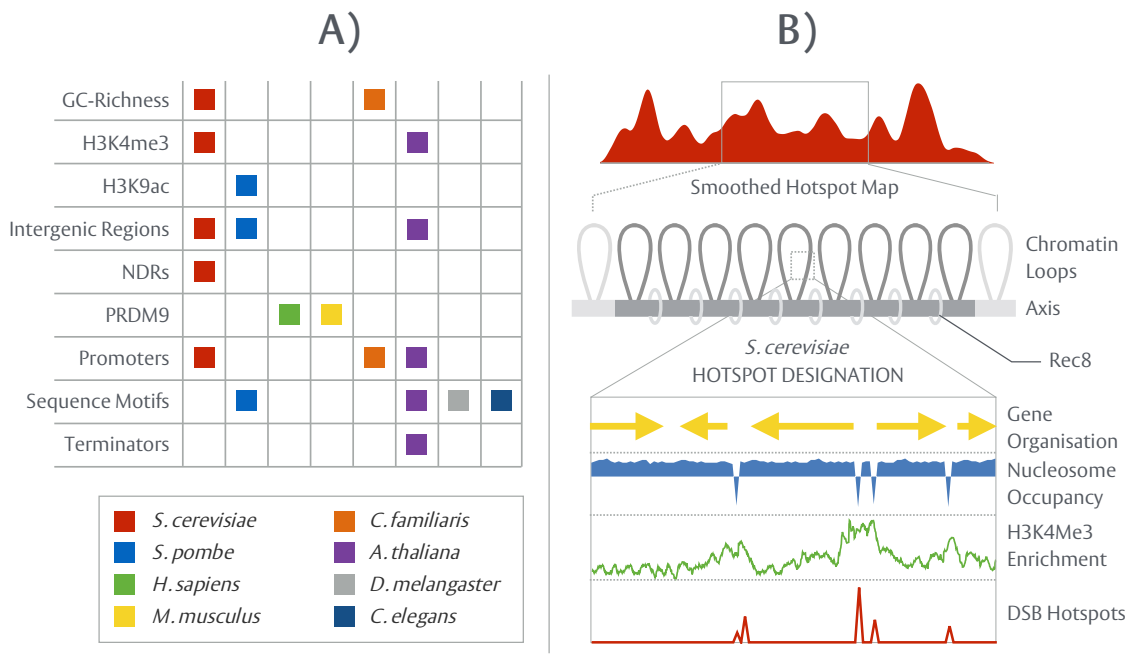


Figure 1.6. Meiotic hotspot designation

A) *Gatekeeper Factors—predictors of recombination*. Hotspot designation differs significantly between species. While eukaryotes (*S. pombe* and *S. cerevisiae*) rely upon a set of passive, low impact factors, higher eukaryotes (*H. sapiens* and *M. musculus*) utilise the multi-functional histone-trimethyltransferase, PRDM9, to guide recombination through the binding of PRDM9 consensus sequences (see text for further details). Outside of these well characterised systems, several further organisms display a number of unique properties. Within the canine lineage (*C. familiaris*), PRDM9 is unexpectedly non-functional—having inactivated between ~7-9Mya—with GC-richness instead serving as a robust predictor of Spo11 activity (Auton et al. 2013; Muñoz-Fuentes et al. 2011; Axelsson et al. 2012). In contrast to the majority of model organisms, insects (*D. melanogaster*) and worms (*C. elegans*), appear devoid of traditional hotspots—consistent with the co-localisation of short, repeat sequences with sites of recombination (Kaur & Rockman 2014; Barnes et al. 1995; Comeron et al. 2012). A role for non-PRDM9 sequence motifs within recombination, however, does not preclude the existence of hotspots, as noted within *A. thaliana* and *S. pombe* (Steiner et al. 2009; Yamada et al. 2013; Choi et al. 2013; Drouaud et al. 2013). **B)** Layers of hotspot designation within *S. cerevisiae*. Canonical hotspot designation, as seen within *S. cerevisiae*, requires the co-occurrence of several factors in a specific fashion in order to unlock the potential for a region to initiate recombination (see text for further details).

At low resolution, *S. cerevisiae* hotspots themselves cluster, organising each chromatid into periodic trough and peak subdomains of recombination potential (see: Figure 1.6B—Top) (Pan et al. 2011; Gerton et al. 2000; Baudat & Nicolas 1997)—an observation that may reflect a non-uniformity in gene density and the impact gene organisation seemingly exerts over both hotspot position and chromatin structure (see Section 1.2.10). For example, the observation that axis proteins are enriched at the 3' end of *S. cerevisiae* genes, while strong hotspots preferentially populate transcriptionally divergent intergenic regions at the 5' end of genes, suggests that the anti-correlation between axis site and hotspot is, in part, driven by the underlying organisation of genes and the associated distribution of markers (Pan et al. 2011; Sun et al. 2015; Champeimont & Carbone 2014). In this manner, the placement of genes may not only constitute a gross organiser of meiotic hotspot position, but also a regulator of hotspot usage.

In striking contrast to *S. cerevisiae*, hotspot designation within mammals (*H. sapiens* and *M. musculus*) relies heavily upon a single protein: the rapidly evolving histone trimethyl-transferase and C2H2 zinc finger domain factor, PRDM9 (de Massy 2013; Pratto et al. 2014; Neale 2010; Baudat et al. 2010; Parvanov et al. 2010). PRDM9 has emerged as a "swiss army knife" of mammalian hotspot designation, and may be more appropriately thought of as a gatekeeper organiser. PRDM9 directs hotspot designation by depositing H3K4me3 markers (Smagulova et al. 2011; Grey et al. 2011; Buard et al. 2009) and potentially recruiting Spo11 machinery (de Massy 2013), both of which promote the required co-occurrence of factors around a consensus DNA sequence specified by the PRDM9 zinc finger motif. The identities of these PRDM9 consensus sequences are predominantly dictated by the allelic variant of its repetitive zinc finger array, of which ~30 have been identified within *H. sapiens* (Berg et al. 2010), allowing differing allelic combinations to produce unique DSB distributions (Pratto et al. 2014; Parvanov et al. 2010; Grey et al. 2011; Brick et al. 2012; Buard et al. 2014). Interestingly, analysis of hotspot locations within *M. musculus PRDM9*$^{-/-}$ mutants uncovered a reversion toward *S. cerevisiae*-like mechanics, with events instead concentrating within H3K4me3-laden promoter regions (Brick et al. 2012). Such an observation reveals the yeast system to be an

ancestral means of determining recombination position, overwritten by the development of PRDM9, as well as a passive system that has persisted despite the evolution of an alternative, dominant method.

Despite the ability of hotspot designation to guide meiotic recombination toward certain sites in the genome at the level of the population, only a subset of hotspots within any given cell are utilised and wild type DSB distributions within individual *S. cerevisiae* cells do not follow models describing their random, independent placement (Zhang et al. 2011; Garcia et al. 2015; Cooper et al. 2016). Further layers of spatial regulation thus exist to control DSB formation on a per cell basis. Any such additional regulation can conceivably function in one of two distinct ways: (i) reactively—directly activated by or in response to DSBs forming or (ii) proactively—activated independently of DSB formation. A potentially distinguishing feature of reactive regulation over that of proactive is an inability to grossly impact population average data due to the low frequencies at which even the strongest hotspots are cleaved (~10-15%). The following sections explore evidence that suggests both forms of regulation function in parallel during meiosis.

1.4.2—DSB interference (Cis/Trans)

Throughout prophase I, several spatial surveillance mechanisms appear to sense the position of DSBs in each cell, relaying this information along and between chromatids (in-*cis*|*trans* respectively) to reactively sculpt the DSB distribution in a DSB-dependent manner. Central to the *cis* branch of spatial regulation, within *S. cerevisiae*, is the recently discovered phenomenon of DSB interference: a localised, suppressive effect dependent upon the DNA damage response (DDR) kinase, Tel1^{ATM}, which operates over ~70-100kb, reducing the frequency of coincident DSB formation below that expected by chance (Figure 1.7A) (Garcia et al. 2015). In effect, DSB interference serves to space DSBs evenly along each chromatid. While not explicitly investigated, the inability of Tel1^{ATM}-dependent DSB interference to strongly manifest within the population average suggests it is a reactive, DSB-dependent process—a hypothesis in line with known models of Tel1^{ATM} activation (Paull 2015).

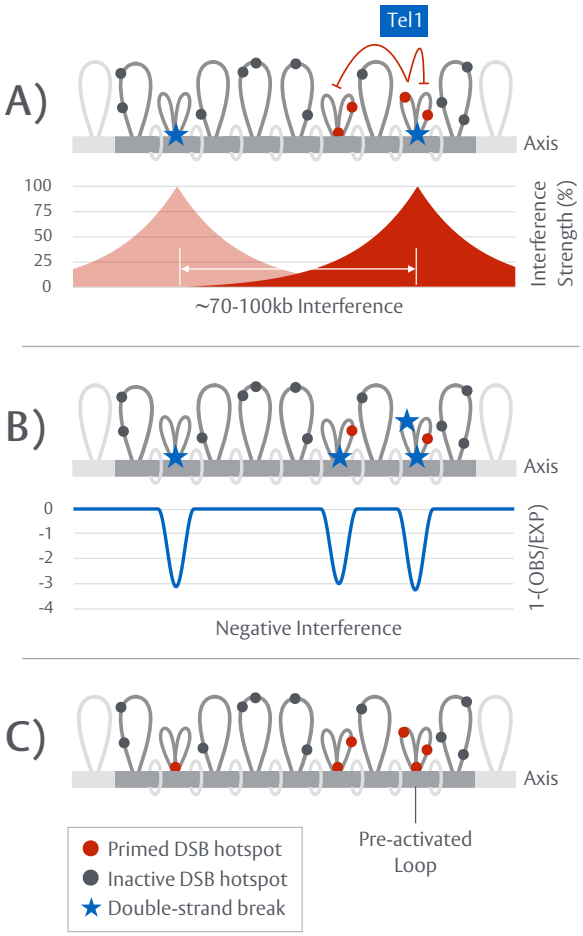


Figure 1.7. Tel1^{ATM}-dependent DSB interference

A) Within wild type cells, a DSB at any given hotspot triggers a Tel1^{ATM} and distance-dependent suppressive effect (DSB interference), repressing DSB formation at adjacent intra-loop hotspots and within neighbouring regions across ~70-100kb in a reactive, DSB-dependent manner. **B)** In the absence of Tel1, DSB interference is abrogated, enabling adjacent DSBs to arise independently over mid-long range distances (>20-100kb). Over short distances (<20kb), loss of Tel1 activity results in patches of negative interference, confined to singular loop domains. **C)** Prior to DSB formation, it is proposed that a sub-population of chromatin loops in any given cell exist in a pre-activated state, "priming" hotspots for usage. An attractive candidate for pre-activation is the tethering of loop sequences to the chromosome axis as proposed by the tethered-loop axis model. Inactivation of Tel1 subsequently unmasks this upstream process, resulting in concerted formation of intra-loop DSBs at frequencies greater than expected from the population average.

Removal of DSB interference by the inactivation of Tel1^{ATM} unexpectedly results in two distinct outcomes: (i) over most distances (±20-100kb) DSBs are no longer subject to interference—forming independently of one another, with coincident DSB formation arising at frequencies similar to those expected by chance (ii) by contrast, at short range (±~7.5kb) DSBs exhibit concerted activity—arising coincidentally at frequencies significantly greater than expected from independent behaviour. Remarkably, this latter phenomenon—which results in calculated DSB interference values that are negative—is only witnessed between DSB hotspots residing within the same chromosomal loop domain (Figure 1.7B) (Garcia et al. 2015). Confinement of concerted activity to within the boundaries of a loop could conceivably arise if a process upstream of DSB formation "activates" the contained hotspots and where the activation of any given loop region occurs in only a subset of the population. The identity and mechanics of this hypothesised activation process remain largely unknown, however, it is pertinent, given the nature of DSB formation, to consider a state of "pre-tethering"—that is, the intimate and stable association of a loop with axial elements, prior to the induction of DSBs, which may serve to "prime" hotspots for use (Figure 1.7C). The generation of localised zones of negative DSB interference may thus be a simple manifestation of a previously unconsidered and proactive consequence of the tethered-loop axis model—one otherwise masked by Tel1, whose repressive activity ensures that only one of the primed hotspots (in any given loop) undergoes DSB formation.

Interestingly, Tel1^{ATM} and its partner DDR kinase, Mec1^{ATR}, have also been implicated in the parallel *trans* branch of spatial regulation. *Trans* interference describes the ability of a DSB on one chromatid to suppress formation at the corresponding locus on its sister, homolog or—in many instances—both (one-per-pair or one-per-quartet respectively) (Zhang et al. 2011). Inactivation of Mec1 or Tel1 abolishes the occurrence of one-per-quartet constraints, indicating a loss of one specific form of *trans* interference; however, whether or not the same form (inter-homolog or inter-sister) is abrogated in each mutant is not yet clear.

Much less is known about the mechanics of how *trans* interference is accomplished, although, recent data raise the possibility that the process is HR-dependent. In vegetatively growing *S. cerevisiae* cells, induction of a DSB by means of a SceI cleavage site, not only results in loading of Rad51 and H2A phosphorylation (a Mec1/Tel1 target) in the vicinity of the break, but also at discrete locations across all chromosomes (Renkawitz et al. 2013; Cheng et al. 2013). This observation infers that ssDNA molecules, engaging in homology searching, can "deliver" biological activity to other chromosomes—an ability that sits well with the requirements of *trans* interference. Such a mechanism has the potential to mediate the *trans* inhibition of meiotic DSB formation specifically between corresponding allelic loci on homologous chromosomes. By contrast, given the spatial proximity of sister chromatids, local pools of activated Mec1/Tel1 may be sufficient to mediate inter-sister *trans* interference without a dependency upon downstream HR steps (Cooper et al. 2014). *Trans* interference likely ensures the availability of an intact repair substrate while simultaneously suppressing the potential for complex double recombination events to arise from DSB formation at the same genetic locus on both homologues.

1.4.3—Evolution & Cellular Role of DSB Interference

Substantial alterations to the DSB distribution, in the manner observed within *tel1*Δ backgrounds, might be expected to significantly perturb recombination and thereby reinforce a presumed importance for DSB interference within the meiotic program; yet, *tel1*Δ mutants display no gross, meiotic defects and exhibit only small reductions in spore viability (~5%) (Garcia et al. 2015; Carballo et al. 2008). Thus whether it is strictly necessary for DSB interference to operate during meiosis is unclear, opening the door to an intriguing possibility: DSB interference may have emerged as an unintended byproduct of another process—persisting in meiosis by means of indirect selection for an indispensable cellular role or target of Tel1^{ATM}. In line with the hijacking of transcriptional markers by meiosis for hotspot designation (see Section 1.4.1), any process or factor altering the accessibility, presence or identity of these markers has the potential to disrupt DSB formation. Interestingly, in mitotic and vegetative states, Tel1^{ATM} has extensive links to

transcriptional regulation via the underlying epigenetic code, notably mediating the in *cis* silencing of transcription in proximity to non-programmed DSBs within humans (Borde et al. 2000), modulation of nucleosomal dynamics and the extensive deposition of DSB induced histone modifications potentially spanning hundreds of kilobases (Shiloh & Ziv 2013; Price & D'Andrea 2013; Lee et al. 2014; Shroff et al. 2004)—a distance in *S. cerevisiae* similar to that of meiotic DSB interference. The availability of shared, universal substrates (e.g. histones) may thus provide a platform for the unavoidable, inadvertent acquisition—or intentional adaptation—of such Tel1^{ATM}-dependent mitotic processes whether they are explicitly required in meiosis or not, leading to the generation of novel but potentially non-essential mechanisms (i.e. DSB interference).

While the above presents an attractive model to explain the origins of DSB interference, an ability for interference to fulfil a beneficial role is not precluded. Indeed, several important considerations remain: (i) While the impact of losing DSB interference on *S. cerevisiae* spore viability is relatively subtle, it may prove cumulative across generations—manifesting after successive interference deficient meioses as a significant alteration in genetic diversity and elimination of affected lineages from the gene pool—highlighting a putative role for DSB interference in the long term stability of the population (ii) Any potential for meiotic failure to arise from the clustering of DSB events may be suppressed or compensated for, masking an otherwise greater impact upon viability. A notable candidate for this role is crossover interference (see Section 1.5.1). Specifically, CO interference may have the ability to partially "correct" the faults resulting from loss of DSB interference via a second round of spatial regulation, selecting only a single DSB per cluster to enter the CO pathway. Consistent with such a role, crossover interference is notably absent within *S. pombe*, an organism which, despite possessing a similar genome size, exhibits a significantly lower DSB frequency and hotspot density than *S. cerevisiae* (~58 DSBs/cell and 1 hotspot/23kb vs. ~150-200 DSBs/cell and 1 hotspot/3.4kb respectively) (Pan et al. 2011; Fowler et al. 2014; Wood et al. 2002; Goffeau et al. 1996). A previously unconsidered consequence of this difference may be a lower reliance upon downstream spatial regulation to spread events along each chromatid. Instead, *S. pombe* may exert

more stringent control at the level of DSB formation simply by placing the process in the hands of more rarely co-occurring factors.

DSB interference may thus collectively guard against the risks associated with otherwise stochastic DSB deposition—preventing deleterious circumstances from arising by chance no matter how infrequently. As previously stated, and in contrast to *S. cerevisiae*, *ATM*$^{-/-}$ null mice develop severe meiotic complications, rendering individuals infertile (Barlow et al. 1996; Barlow et al. 1998), a phenotype predominantly ascribed to excessive DSB formation and thus compatible with loss of a repressive, interfering effect (see Section 1.3.3). Such safeguards may therefore become increasingly important for the selective fitness of mammalian organisms, which have larger chromatin loop sizes (Kleckner 2006; Novak et al. 2008; Kauppi et al. 2011) (potentially permitting unmanageable numbers of clustered DSBs per loop), smaller populations, and greater time between sequential cycles of sexual reproduction relative to *S. cerevisiae*.

1.4.4—DSB Competition

The introduction of a novel hotspot, either by insertion of a strong, high frequency site (e.g. *HIS4*::*LEU2*) or the tethering of Spo11/Spp1 fusion constructs to cold regions within the *S. cerevisiae* genome, not only induces DSB formation but also a repressive, distance-dependent effect that profoundly alters the DSB distribution over a considerable margin (Fukuda et al. 2008; Robine et al. 2007; Acquaviva et al. 2013; Wu & Lichten 1995; Fan et al. 1997). Despite the prominent similarities to DSB interference, recent data suggests this effect exhibits substantial Tel1-independency, revealing a third, distinct layer of spatial regulation (Cooper et al. 2016; Mohibullah & Keeney 2017). Furthermore, this repressive effect appears to strongly manifest itself within the population average (Robine et al. 2007; Acquaviva et al. 2013), suggesting it is a perpetually present and proactive process that does not rely upon DSB formation for activation. This phenomenon—referred to here as "DSB competition"—may arise upstream of DSB formation out of a need for

hotspots to compete over restricted and limited pools of pro-recombination factors (Keeney et al. 2014; Robine et al. 2007; Wu & Lichten 1995).

Rec114, Mer2, and Mei4, which coalesce into the RMM complex, are factors essential to Spo11-dependent DSB formation enriched on the chromatin axis (see Sections 1.2.2 and 1.2.10) (Panizza et al. 2011; Maleki et al. 2007). While population averaged binding profiles (ChIP-chip) reveal RMM to occupy ~900 genomic loci (Panizza et al. 2011), an observation in line with the estimated ~700 meiotic loops present within *S. cerevisiae* (per haploid genome copy) (Ito et al. 2014), RMM foci peak as low as ~40-60/nucleus within individual cells (Carballo et al. 2013; Li et al. 2006), identifying RMM as potential players in DSB competition. Despite this apparent limitation, *S. cerevisiae* is observed to form ~150-200 DSBs/cell (Pan et al. 2011; Martini et al. 2011). One way to reconcile these conflicting observations with the existence of DSB competition is to consider that, in each cell, loops aggregate around RMM foci into clustered super domains which, perhaps, nucleate at or generate those chromosomal regions that are first to assemble short, incomplete elements of the chromosomal axis during early prophase I (Figure 1.8A) (Cooper et al. 2016). Within this model, it is proposed that DSB competition arises through successive rounds of intra-cluster but inter-loop competition for limited RMM and/or tether points, confining the repressive effect of DSB competition to the average size of a cluster (Figure 1.8B)—drawing considerable parallels to models previously proposed for crossover interference (Stahl et al. 2004). Furthermore, differences in the density of gatekeeper factors (e.g. H3K4me3) (see Section 1.4.1) may govern the extent to which any given loop can compete, introducing significant overlap between hotspot designation and downstream spatial regulation.

Interestingly, a comparable disparity exists within mice: an estimated 10,000 loops span the genome while MEI4 foci are present at significantly lower levels (~300/nucleus), suggesting a similar regulatory layer could operate within other species (Novak et al. 2008; Ito et al. 2014; Kauppi et al. 2011; Kumar et al. 2010).

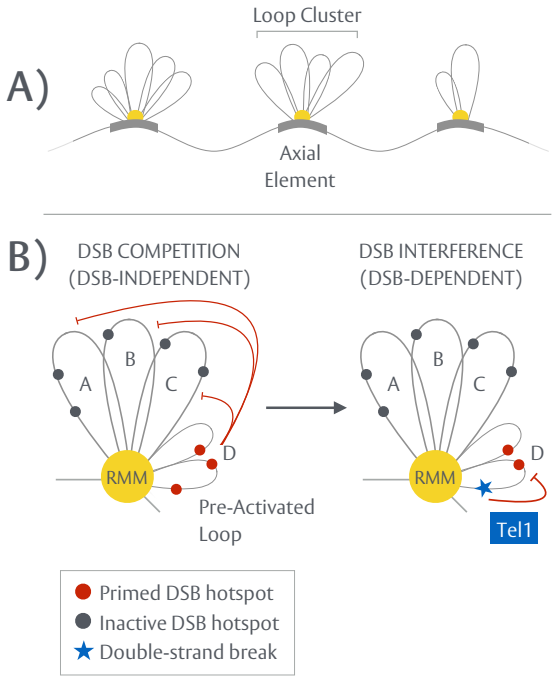


Figure 1.8. Proposed "loop cluster" model of DSB competition

A) During early prophase I, short stretches of axial element nucleate at scattered regions across each chromosome. Upon this platform, the first meiotic loop may begin to assemble, associating together into individually acting, isolated units. **B)** Building upon the tethered-loop axis model, it has been proposed that within any such clustered unit, limited availability of, or access to, essential factors such as RMM (Rec114-Mei4-Mer2 complex), coupled to a differential ability of each loop to establish a tether, could generate DSB competition by means of competitive tethering—lowering the frequency of DSB formation within the remainder of the associated loops in a proactive, DSB-independent manner. Under wild type conditions, DSB formation subsequently induces Tel1^{ATM}-dependent DSB interference—a process that may inhibit or dismantle cluster units thereby suppressing further DSB formation in the immediate region. As illustrated here, the apparent Tel1-independency of DSB competition suggests the strong, repressive effect observed around strong hotspots is in fact, a composition of two distinct processes.

A mechanism of pre-tethering may thus underpin both the negative interference values observed within individual loops when Tel1^{ATM}-dependent DSB interference is lost and, in part, the DSB-independent competition that arises between DSB hotspots residing in adjacent loop domains within *S. cerevisiae*.

1.4.5—DNA Replication Coupling

DSB formation is tightly coupled to the state of pre-meiotic S phase, whereby delayed replication of a chromosome arm delays DSB formation within the same region (Borde et al. 2000; Murakami et al. 2003; Murakami & Keeney 2014). Such regulation is of the upmost importance: replication fork progression is fully stalled by DSBs and subsequent topological constraints can result in full fork collapse (Mirkin & Mirkin 2007). Mistimed and widespread induction of meiotic DSBs would thus prove highly toxic if DNA replication remained uncompleted. Moreover, differences in the usage and timing of replication origin activity may in turn permit such processes to contribute, in a generalised manner, to spatial regulation.

Within vegetatively growing cells the replication checkpoint, which depends upon ATR and the downstream CHK1, inhibits cell cycle progression in response to aberrant replication forks (Smith et al. 2010). Mec1^{ATR} and Rad53 appear to mimic their mitotic roles, limiting the activity of Dbf4-dependent Cdc7-kinase (DDK) within pre-meiotic cells treated with the replication inhibitor hydroxyurea (HU) (Blitzblau & Hochwagen 2013). Mer2 requires DDK-dependent phosphorylation in order to recruit Spo11 to axial sites and thus this downregulation of DDK by Mec1/Rad53 inhibits a crucial step in DSB formation in response to replication stress (Blitzblau & Hochwagen 2013; Marston 2009). Inhibition of DNA replication also causes a ~10-fold reduction in Spo11 transcript levels—a reduction partially ablated within *mec1*Δ mutants, further implicating Mec1 within Spo11 downregulation at the level of transcription via unknown mechanisms (Blitzblau & Hochwagen 2013). Mec1 activity also appears to inhibit loading of Rec114 and Mre11 in response to replication stress (Blitzblau & Hochwagen 2013).

1.4.6—Centromeres, Telomeres and Repeat Sequences

Several constitutive elements of the genome serve as gross repressors to recombination—including centromeres, telomeres and various forms of repeat element—contributing to the overall distribution DSBs and COs. While Spo11 transiently associates with a ±10-15kb region flanking the 120bp-long centromeres in *S. cerevisiae* during meiotic S phase, it rapidly relocates to chromosomal arms (Kugou et al. 2009) and centromeric COs are virtually undetectable and/or non-existent owing to Zip1-dependent suppression (Chen et al. 2008; Vincenten et al. 2015). Moreover, DSBs in proximity to centromeres are detected at levels much lower than genome-wide averages (Pan et al. 2011). Akin to centromeres, telomeres also suppress DSB formation across a ~20kb region ~3.5-6.5-fold within *S. cerevisiae* with corresponding reductions in CO formation through unknown mechanisms (Blitzblau et al. 2007; Buhler et al. 2007; Pan et al. 2011). Several other regions, containing repeats, also regulate recombination. Ribosomal DNA (rDNA) genes are organised as tandemly repeating gene clusters (~1Mb ChrXII in *S. cerevisiae*) and serve as major barriers to recombination. DSB formation is essentially absent within the rDNA region (Blitzblau et al. 2007; Pan et al. 2011)—a repression dependent upon the histone deacetylase Sir2 and the AAA+ ATPase Pch2 (Mieczkowski et al. 2007; Vader et al. 2011). However, not all low complexity, repeat regions are suppressed for DSB formation. While a subset of TY-family retro-transposable elements display extremely low levels of DSBs (Pan et al. 2011), this is not a feature universal to all TY-elements and some TY-elements may actually promote DSB formation at a number of loci (Sasaki et al. 2013). Overall, repression of recombination within functional elements of each chromosome (telomeres, centromeres) or regions with low sequence divergence (repeats) is likely crucial to the maintenance of genomic integrity and repression of ectopic recombination.

1.5—Spatiotemporal Control of COs/NCOs

1.5.1—Crossover Interference

Akin to DSBs, CO distribution is similarly subject to stringent spatial control, predominately through the phenomenon of CO interference—a near-universal process of spatial regulation observed within *S. cerevisiae* (Sym & Roeder 1994; Berchowitz & Copenhaver 2010), *C. elegans* (Meneely et al. 2002), *D. melanogaster* (Page & Hawley 2001), *A. thaliana* (Copenhaver et al. 2002), *H. sapiens* (Rasmussen & Holm 1984; Housworth & Stahl 2003) and *M. musculus* (Broman et al. 2002; de Boer et al. 2006). CO interference manifests as a non-random spatial distribution of COs, whereby an interfering "signal" precludes formation of COs in proximity to pre-existing events in a distance-dependent manner—dispersing COs evenly across each chromatid. In contrast, NCOs do not exhibit detectable interference (Berchowitz & Copenhaver 2010).

Beyond distinct genetic requirements (see: Section 1.2.7), a defining characteristic of Msh4-Msh5 (MutSγ), Mlh1-Mlh3 (MutLγ) and ZMM-dependent class I COs is their imposition of and sensitivity to CO interference. In contrast, MUS81-MMS4-dependent class II COs form independently of CO interference. For example, within *S. cerevisiae*, abolition of the class I pathway (via *msh4*Δ, *msh5*Δ) reduces CO formation by ~60% and all remaining COs are randomly distributed (i.e. loss of CO interference) (Argueso et al. 2004; Chen et al. 2008; de los Santos et al. 2001). Likewise, removal of the class II pathway (via *mus81*Δ, *mms4*Δ) reduces CO formation by ~25%, but CO interference is retained (de los Santos et al. 2003; Argueso et al. 2004). Analogous mechanics are observed within *A. thaliana* (~5% Class II) (Higgins et al. 2008) and *M. musculus* (~5-10% Class II) (Holloway et al. 2008). However, despite wide ranging conservation, not all organisms display CO interference or class II formation (Table 1.1). For example, *S. pombe* relies heavily on the class II pathway (~80-95% of COs exhibit Mus81-Mms4 dependency) and is devoid of detectable CO interference (Smith et al. 2003; Munz 1994) while *C. elegans* displays "perfect" interference mediated by CO assurance (1 CO per bivalent) (see: Section 1.2.7) and is fully dependent upon Msh4-Msh5 (Meneely et al. 2002).

Organism	Class I COs	Class II COs	COs/Meiosis (WT)	COs/Chr
S. cerevisiae	✓	✓	70-90	4.3-5.6
S. pombe	X	✓	38	12.6
M. musculus	✓	✓	22-28	1.1-1.4
H. sapiens	✓	~	50-70	2.1-3.0
A. thaliana	✓	✓	10	2
C. elegans	✓	x	6	1
D. melanogaster	X	X	6	1.5

Table 1.1. Features of CO formation across common model organisms

Meiotic crossovers (COs) may form via two distinct pathways: (i) MSH4-MSH5 (MutSγ), MLH1-MLH3 (MutLγ) and ZMM-dependent interfering class I COs or (ii) Mus81-Mms4-dependent non-interfering class II COs. CO frequency and subclass usage varies widely amongst species. Several organisms, including *S. cerevisiae*, *A. thaliana* and *M. musculus*, produce a mixture of class I and class II COs during any given meiosis. Indirect evidence suggests class II COs also form within *H. sapiens.* A number of organisms, however, display more atypical mechanics. Class I COs, and in turn CO interference, are notably absent within *S. pombe*. In direct contrast, *C. elegans* exhibits a state of total interference, marked by the formation of a single class I CO per chromosome. Interestingly, COs within *D. melanogaster* appear to form independently of either pathway yet still display detectable interference.

Data References:

S. cerevisiae (Mancera et al. 2008; Martini et al. 2011; de los Santos et al. 2003; Argueso et al. 2004)

S. pombe (Cromie et al. 2006; Smith et al. 2003)

M. musculus (de Boer et al. 2006; Holloway et al. 2008; Koehler et al. 2002; Broman et al. 2002)

H. sapiens (Holloway et al. 2008; Vallente et al. 2006; Tease et al. 2006; Barlow & Hultén 1998)

A. thaliana (Higgins et al. 2008; Berchowitz et al. 2007; Copenhaver et al. 2002)

C. elegans (Meneely et al. 2002; Tsai et al. 2008)

D. melanogaster (Foss et al. 1993; Carpenter 1975)

The mechanistic differences underlying sensitivity or insensitivity to CO interference are unclear. In *S. cerevisiae*, Mus81-Mms4 acts late in recombination and resolves aberrant or complex joint molecules inaccessible to the primary Msh4-Msh5 pathway (Jessop & Lichten 2008; Oh et al. 2008). If CO interference is established early in prophase I within a strict time window, the lack of interference for class II events may simply reflect a temporal distinction between Msh4-Msh4 and Mus81-Mms4 activity and/or the time required to resolve simplistic vs. complex substrates. Consistent with this idea, *spo16*Δ*ndt80*Δ mutants, which are defective in synaptonemal complex (SC) extension, display wild type CO interference (Shinohara et al. 2008) and the distributions of SC initiation complexes are themselves non-random (Fung et al. 2004)—collectively suggesting interference is fully implemented and "read" prior to the leptotene-zygotene transition (early prophase I).

1.5.2—Crossover Interference Machinery

Exactly how the interfering "signal" is generated and propagated remains similarly unclear. Interestingly, while *spo16*Δ*ndt80*Δ mutants accumulate high levels of joint molecules—indicating that HR has taken place—*msh5*Δ*ndt80*Δ mutants do not nor do they exhibit interference (Oh et al. 2007). Furthermore, deletion of TID1, a factor which facilitates strand invasion, significantly weakens interference without reducing CO frequency below wild type levels (Shinohara et al. 2003). These observations collectively suggest that imposition or initiation of CO interference occurs at the level of strand invasion. Nevertheless, regulation of a hyper-localised event such as strand invasion alone is unlikely to account for the vast distances (~±200kb in *S. cerevisiae*) over which CO interference appears to act. Indeed, within *S. cerevisiae*, inactivation of several factors abolish CO interference and thus result in random, independent CO distributions, including topoisomerase-II (Zhang, Wang, et al. 2014), the hexameric AAA+ ATPase Pch2 (Joshi et al. 2009; Zanders & Alani 2009) and the ZMM family of proteins (see Section 1.5.1). SUMOylation of topoisomerase II and the axial factor, Red1, by the SUMO-E2 Ubc9 and ubiquitin-mediated removal of these factors by STUbL (Slx5/8) also appears essential (Zhang, Wang, et al. 2014). The seeming reliance of CO interference upon

structural and topological factors suggests the interfering "signal" may be mechanical in nature and transmitted along the axis. In support of this idea, and as previously stated, *C. elegans* exhibits "absolute" interference (1 CO/chromosome) (Meneely et al. 2002), yet, when several *C. elegans* chromosomes are fused together, CO frequency does not increase but rather only a single CO forms as before (Hillers & Villeneuve 2003). CO interference in this organism therefore considers a fused chromosome, bound by a single axis, to be a singular unit independent of base pair length—an observation that fits well with axial transmission of CO interference.

Despite the identification of factors required and wide ranging historical observation, many fundamental questions regarding the underlying mechanisms of CO interference remain. However, traditional genetic screens are often labour intensive and experimental data can be difficult to interpret. In contrast, statistical or mathematical modelling is a non-invasive technique that can provide new insights and bolster pre-existing conclusions. Several models (see: Section 1.5.3) have been devised to describe or investigate CO interference but as of yet no predominant paradigm has been established.

1.5.3—Modelling Crossover Interference

The modelling of CO interference is predicated on (i) an ability to positionally map CO formation (see Section 1.2.8) at either the bivalent or base pair level on a per cell basis, (ii) a statistical, mathematical or mechanical description of the system—of which several exist, and (iii) a mechanistic hypothesis.

Descriptors

Classically, CO interference is described via the coefficient of coincidence (CoC) (Muller 1916; Stahl & Foss 2009), which is defined as follows: if the recombination rates of two disjoint genomic regions (A and B) are denoted r(A) and r(B), then CoC = r(A,B) / (r(A)r(B))—in other words, CoC(A,B) is defined as the ratio between the observed rate of double recombination, that occurred between loci A and B, and the expected rate. The inversion of CoC (int = 1-CoC) is often used to describe

interference strength: (i) int > 0 indicates positive interference—recombination is concurrently occurring at A and B at rates lower than expected by chance (ii) int = 0 indicates no CO interference (iii) int < 0 indicates negative interference—recombination is occurring at rates higher than expected by chance. Alternatively, CO interference can be described using stochastic models, underpinned by continuous probability distribution functions: (i) originally, a homogenous Poisson model was put forward to define a state of complete independence (i.e. a CoC value of 1 across the chromosome) (Haldane 1919). The probability that an event occurs within any given interval can be calculated for varying levels of pre-existing events, thus allowing assessment of experimental data against expectations of independency (ii) inter-event distances (IEDs)—the distance between successive CO events—can be calculated and described by the gamma (γ)-distribution (McPeek & Speed 1995; Zhao et al. 1995). γ-distributions are characterised by independent γ(α) (shape) and γ(β) (scale) parameters. Akin to CoC analysis, the IED distribution of the homogenous poisson model (no CO interference) is an exponential random variable with a γ(α) value of 1.0 while γ(α) > 1 and γ(α) < 1 indicates positive and negative interference respectively.

Mechanistic Models

A prominent model of CO interference—the stress relief model—postulates that generation and subsequent relief of macroscopic, biomechanical stress along the prophase I chromosomal axes serves as the primary communicator of CO interference (Kleckner et al. 2004). In this model, initial CO formation occurs within a zone of high mechanical stress and serves as a nucleation point for localised stress relief which propagates outward into the immediate vicinity surrounding the event —defining a distance-dependent zone within which further CO formation is probabilistically repressed (i.e. interference) (Figure 1.9A). Localised stress relief subsequently remodels the chromosome wide stress distribution. All subsequent events will tend to form in any remaining zones of relatively higher stress—which are, by definition, distal to pre-established events thus generating an evenly spaced array. Such a model is largely in agreement with factors known to be involved within CO interference (see Section 1.5.2).

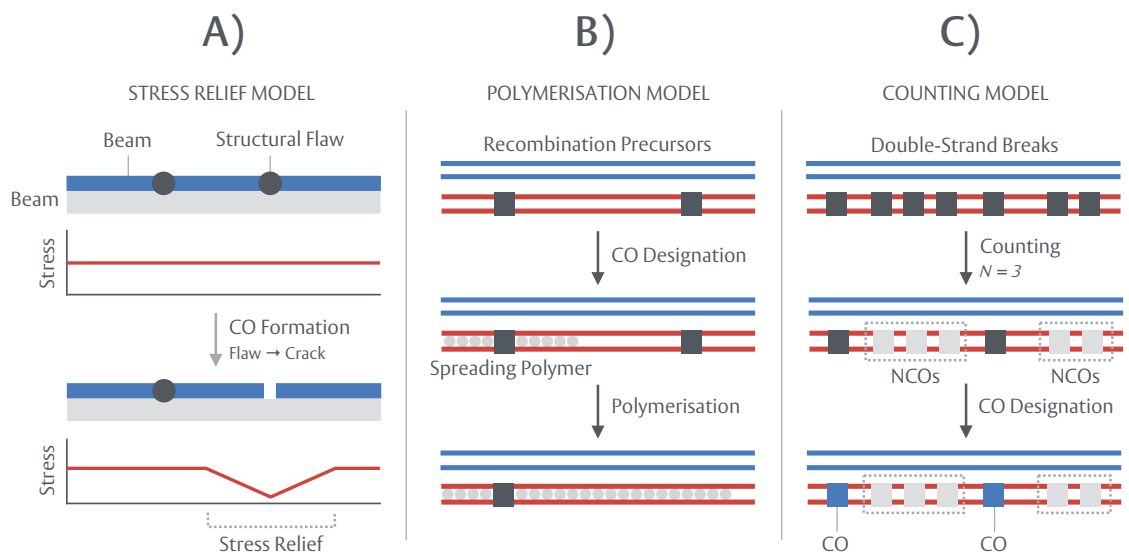


Figure 1.9. CO interference models

A) "*Beam–film*" *stress relief model*. In this model, the "beam" (chromosomal axis) imposes mechanical stress on the "film" (chromatin), generating structural flaws (recombination precursors). Localised fluctuations in stress eventually induce a "crack" (CO) and the subsequent bidirectional spreading of stress relief inhibits further CO formation in-proximity to the event. **B)** *Polymerisation model*. Chromatids initially contain an array of recombination precursors (black), all of which possess the potential to form a CO. Upon maturation of a precursor (CO formation), a polymerising signal spreads bidirectionally to prevent maturation of adjacent intermediates. **C)** *Counting model*. Chromatids initially contain a randomly distributed and dense array of recombination precursors, which may mature into COs or NCOs. This model proposes that CO interference effectively "counts" the number of intervening NCO events (N) (shown for N = 3) to space COs along each chromatid.

Notably, topoisomerase-II alleviates topological stress within chromosomes and its ATP-dependent activity is explicitly required for CO interference (Zhang, Wang, et al. 2014). A mathematical, simulated description of the stress relief model, termed the "beam-film model", robustly recaptures CO distributions within several organisms including *S. cerevisiae*, *D. melanogaster* and *S. lycoperiscum* (Zhang, Liang, et al. 2014). However, the specific identity of the interfering signal is ultimately interchangeable within this model—describing any distance-dependent inhibition—which could conceivably constitute an exponentially decaying kinase signal, spreading histone or other protein modifications and/or stress relief.

A model of similar principle—the polymerisation or King-Mortimer model—describes a situation whereby early recombination precursors are initially distributed at random but can undergo bi-directional polymerisation that spreads an interfering signal and prevents nearby events from engaging the bivalent (Figure 1.9B) (King & Mortimer 1990). Simulations based on this model efficiently explain CO distributions within *D. melanogaster* and *S. cerevisiae*. Consistent with this, synaptonemal complex (SC) polymerisation has been proposed to mediate CO interference (Egel 1978; Maguire 1988). However, as previously outlined (see Section 1.5.1), CO interference is seemingly established prior to SC assembly and thus the biological identity of this putative "polymer" remains unknown.

A third model—the counting model—postulates that COs are separated by a fixed number (n) of intervening, non-interfering NCO events (Figure 1.9C) (Foss et al. 1993; Foss & Stahl 1995). While the inclusion of non-interfering COs has allowed this model to recapture CO distributions within *S. cerevisiae* (Stahl et al. 2004), *A. thaliana* (Lam et al. 2005; Copenhaver et al. 2002) and *H. sapiens* (Housworth & Stahl 2003) well, it is not clear how a biological system could mediate such a count and no further mechanistic details are known. Moreover, the model cannot account for CO homeostasis (Martini et al. 2006).

1.6—Thesis Aims

The risk-reward tradeoff inherent in meiotic recombination evidently places a strong demand on the cell for stringent and adaptive control at multiple stages of the process. Spatial regulation of DSB and CO formation is rapidly emerging as a key part of this control. Nevertheless, many fundamental questions regarding the mechanisms of DSB and CO interference remain. Components of the DNA damage response (DDR), including Tel1^{ATM}, Mec1^{ATR}, Rad24 and Rad17 reside at the heart of many meiotic processes. While Tel1^{ATM} and Mec1^{ATR} spatially guide the formation of DSBs within *S. cerevisiae*, how DDR factors may influence CO formation is largely unknown. Furthermore, the machinery and mechanisms underlying spatial regulation form branches of a larger, interconnecting hierarchy within which extensive cross-talk and overlap is a possibility. Analysis of spatial regulation through traditional means is therefore difficult. Despite publication of several mathematical CO interference models, many fail to consider the existence of class II COs, estimate class II frequency or sufficiently reduce the system to a fundamental set of parameters. Moreover, no comprehensive model for Tel1^{ATM}-dependent DSB interference has been constructed and current Spo11-DSB mapping technologies contain ambiguities that preclude full analysis of hotspot designation. Thus, in order to further our understanding of spatial regulation, novel simulation platforms and mapping technologies are required and several questions must be addressed.

- To develop a novel simulation platform for the analysis of CO and NCO distributions using genome-wide mapping data **(Chapter 2)**
- To analyse how components of the DDR (Tel1^{ATM}, Mec1^{ATR}, Rad24) may influence CO and NCO distribution **(Chapter 2)**
- To develop an analytical software package for a novel *sae2*Δ-dependent Spo11 DSB mapping technology **(Chapter 3)**
- To characterise the hyperlocal features guiding Spo11 DSB formation and hotspot designation **(Chapter 3)**
- To develop a novel simulation platform for the analysis of DSB distributions using genome-wide mapping data **(Chapter 4)**
- To analyse proposed models for negative interference and characterise features of Tel1^{ATM}-dependent DSB interference **(Chapter 4)**

CHAPTER 2

Investigating the role of DDR proteins within the spatial regulation of COs

2.1—Introduction

Spatial patterning is a ubiquitous feature of many biological systems, including meiosis—which employs complex, layered processes to govern the distribution of recombination events at multiple levels (see: Section 1.4, 1.5). Notably, at the level of crossover (CO) formation, Msh4-Msh5 (MutS homolog, Mlh1-Mlh3 (MutL homolog) and Zip-dependent (ZMM) class I COs exhibit the phenomenon of CO interference—a near universally observed process of spatial regulation characterised within a wide range of organisms including *S. cerevisiae*, *M. musculus* and *H. sapiens* (Sym & Roeder 1994; Berchowitz & Copenhaver 2010; Rasmussen & Holm 1984; Housworth & Stahl 2003; Broman et al. 2002; de Boer et al. 2006). CO interference manifests as a non-random distribution of COs, whereby an interfering process precludes formation of COs in proximity to pre-existing events—dispersing COs evenly across each bivalent within any given individual cell (see: Section 1.5.1). A number of factors, in addition to those required for class I CO formation, appear essential for the generation or maintenance of WT-like CO interference in *S. cerevisiae*, including topoisomerase II (Zhang, Wang, et al. 2014) and the hexameric ATPase Pch2 (Joshi et al. 2009; Zanders & Alani 2009). Inactivation of such factors results in a less evenly spaced or random distribution of COs. Efforts to understand the distribution of recombination within any given mutant often revolves around the analysis of genome-wide data specifying the position of COs (see: Section 1.2.9). However, a minority of CO events (~15-35% in *S. cerevisiae*) form through an alternative Mus81-Mms4-dependent mechanism, generating class II CO events insensitive to CO interference (de los Santos et al. 2003)—which may complicate straightforward interpretation of the data.

Despite the identification of factors involved in CO interference and wide ranging historical observation of the phenomenon, many fundamental questions remain as to how the process mechanistically functions. Two branches of the DNA damage response (DDR) are intimately linked to and involved in meiosis: (i) the ssDNA sensing system, comprising Mec1 (ATR), the Rad17-Mec3-Ddc1 (RAD9–RAD1–HUS1 (9–1–1) complex) clamp complex and the clamp loader, Rad24 (RAD17) and (ii) the DSB sensing system, comprising Mre11-Rad50-Xrs1/NBS1 (MRX/N) and Tel1 (ATM) (See: Section 1.3). DDR components from both systems conspire to form the pachytene I

checkpoint—homeostatically modulating DSB formation—and regulate the distribution of precursor DSB events (Cooper et al. 2014; Cooper et al. 2016; Garcia et al. 2015; MacQueen & Hochwagen 2011). However, knowledge of how these integral factors may impact the downstream distribution of COs or influence the overall landscape of CO interference is incomplete or non-existent. Work presented throughout this chapter thus seeks to utilise computational and mathematical methods to investigate how key DDR factors may spatially regulate the formation of recombination events during meiosis.

2.2—High Resolution Mapping of Recombination

Analysis of CO and NCO distribution relies upon access to or generation of genome-wide data detailing the accurate positions of meiotic events. Such high resolution mapping of recombination may be achieved through the use of hybrid strains (see: Section 1.2.9); whereby each homologous chromosome is of a divergent parental origin—permitting single nucleotide polymorphism (SNP) and insertion/deletion (INDEL) based detection of inter homologue events (Mancera et al. 2008; Martini et al. 2011; Oke et al. 2014). This assay encompasses several key steps, namely: (i) the induction of meiosis within hybrid *S. cerevisiae* cells (e.g. S288c x SK1) (ii) recovery of all four spore progeny via tetrad dissection (iii) isolation of genomic material from these individual, haploid cells and (iv) whole genome next-generation sequencing (NGS) (Figure 2.1A). Following the alignment of read data to reference genomes, a pre-existing pipeline calls event types (CO and NCO) based on polymorphism patterns and determines event position using the midpoint of each called recombination event (Figure 2.1B) (Marsolier-Kergoat et al. 2017; M. Crawford, M.J. Neale unpublished). NCOs typically generate narrow and asymmetric gene conversion tracts and may thus be under represented due to the density of known SNPs/INDELs or if mismatch correction occurs. In order to bolster NCO detection, mismatch repair (MMR) is thus inactivated (via *msh2*Δ—MutS homolog 2) within several strains. Given the reliance of this method upon the sequence divergence that exists between homologues, inter sister events—which account for an estimated ~12.5-25% of all events within *S. cerevisiae* (Goldfarb & Lichten 2010)—remain invisible.

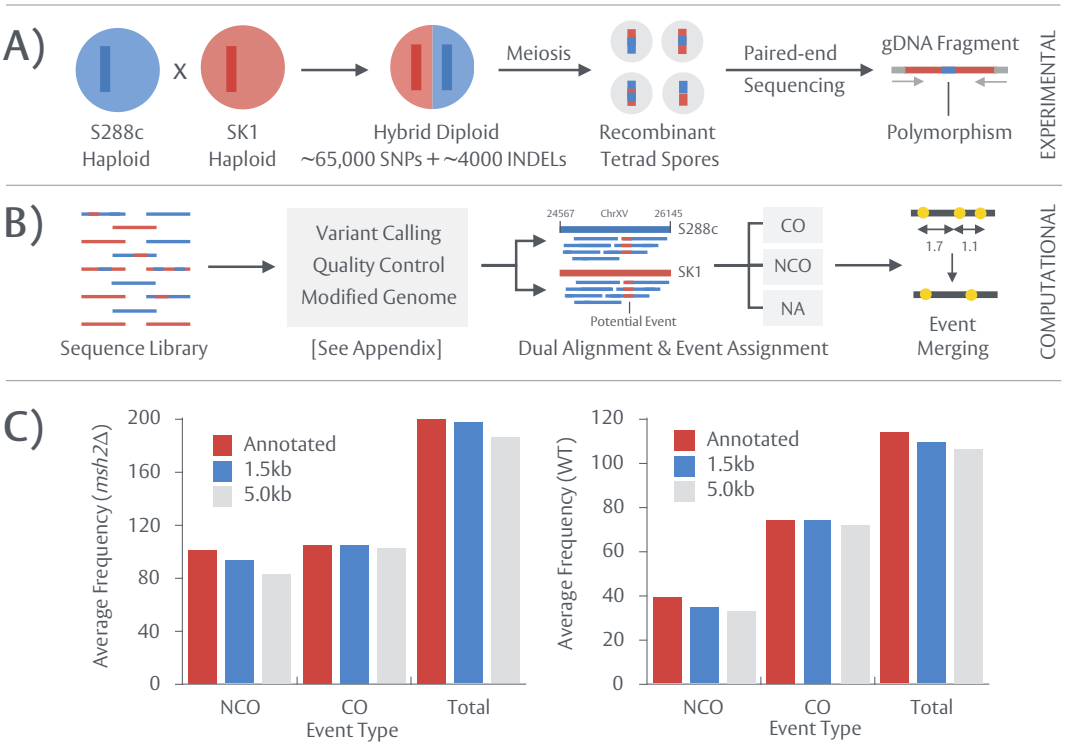


Figure 2.1. Genome-wide mapping of meiotic recombination

A) Hybrid S288c x SK1 *S. cerevisiae* strains (diploid) are constructed and meiosis is induced. Haploid cell samples are harvested following completion of a meiotic time-course and prepped for Illumina paired end, next-generation sequencing (NGS) **B)** Resulting reads are aligned against a S288c reference genome (S288c—SGD Jan 2015, R64-2-1) to construct a (i) SNP/INDEL table via GATK and *HybridVar* and (ii) a pseudo SK1 genome, containing all detected polymorphisms (see: Section 2.3). Reads from each haploid spore are subsequently re-aligned against both genomes (S288c, pseudo SK1) and undergo event assignment, designating COs or NCOs based on SNP/INDEL patterns. Event positions are called as the midpoint of any detected recombination event. Successive events within a given distance (e.g. 1.5kb) are merged and positions are recalculated as the midpoint of the combined events. **C)** Average event counts, per event type were calculated via *RecombineSim* for *msh2*Δ and WT data using a manually annotated dataset and datasets at two merging thresholds (1.5kb, 5kb).

Over analysis or misinterpretation of the data may occur, particularly at regions exhibiting complex SNP/INDEL conversion patterns. In order to obtain a conservative estimate of event count and position, any events within a given distance may be sequentially merged into a single event, with a subsequent readjustment to midpoint, position values. To determine an appropriate merging threshold, average event frequencies were analysed for unmerged, annotated datasets—whereby ambiguous events have been manually annotated to reflect their most probable identity—and unannotated, merged datasets (1.5kb, 5kb) (Figure 2.1C). As expected, given the large scale genetic exchanges that occur when a CO forms and the dispersing activity of CO interference, detection of individual COs appears robust and largely unaffected by merging in both WT and *msh2*Δ, and reductions in CO frequency relative to annotated are negligible (1.5kb—0.21%, 5kb—2.03% in *msh2*Δ) (see: Figure 2.1C). In contrast, the frequency of NCOs, which are not subject to any appreciable spatial regulation, are more heavily impacted by merging within WT and *msh2*Δ—with greater reductions in average frequency, relative to annotated, observed (1.5kb—9.00%, 5kb—19.66% in *msh2*Δ) (see: Figure 2.1C). Despite yielding higher levels of information, manual annotation is a subjective process. Thus, as a point of compromise between over and under analysis, a merging threshold of 1.5kb is employed throughout this chapter, unless stated otherwise. In other words, any given event is considered to be a distinct entity if separated from others by at least 1.5kb:

Raw Calls

Genotype	Repeat	Chr	Position	
*msh2*Δ	1	2	173594	
				519bp
*msh2*Δ	1	2	174113	
				2778bp
*msh2*Δ	1	2	176891	

1.5kb Threshold

Genotype	Repeat	Chr	Position	
*msh2*Δ	1	2	173853	Merged
*msh2*Δ	1	2	176891	

All experimental work, NGS library prep and event calling was performed by (M. Crawford, M.J. Neale unpublished). The following sections detail the downstream computational and mathematical work done to analyse the resulting data.

2.3—HybridVar: Calling S288c x SK1 SNPs/INDELs

Detection and assignment of event type (CO or NCO) is accomplished by determining the parental origin of any given loci or genomic stretch—requiring a high quality list of S288c x SK1 SNPs. SNPs may be detected through the alignment of reads—derived from pure, non-hybrid SK1 strains—to the S288c reference. Such an approach, however, necessitates additional sequencing. In contrast, genome-wide mapping of recombination within a multitude of hybrid S288c x SK1 spores inherently yields SK1 reads at an incredibly high depth. Moreover, small INDELs—often discarded in tetrad analyses—may improve the accuracy of event calling by providing additional information. Thus, in order to utilise the information already present in obtained datasets, a novel SNP/INDEL screening approach was developed (*HybridVar*, see: Section B2.1).

Variants were initially called via GATK *HaplotypeCaller* (v3.4-46) for 72 individual spores derived from complete *msh2*Δ or *tel1*Δ*msh2*Δ octads (post replicative tetrads), resulting in 72 separate variant call format (VCF) files which were subsequently parsed *en masse* by *HybridVar*, calculating: (i) call frequency (% of spores any given allele is present within) (ii) cumulative total read depth of each loci (tRD), in order to assess coverage and (iii) cumulative allelic read depth (vRD) (% of reads that contain a specific allele at a specific loci). In effect, *HybridVar* internally pools individual VCF samples to calculate library wide stats for each variant. *HybridVar* performs user specified filtering based on these parameters. Crucially, any legitimate S288c x SK1 variant is expected to exhibit a call frequency of ~50%—that is, present in half of all individual, haploid spores. ~90,000 unique variants, called by *HaplotypeCaller*, were therefore screened using a call frequency threshold of 48-52% and, to assure confidence in calls, a minimum tRD of 250 and a minimum vRD of 95%, discarding ~21,000 calls. To further reduce mis-genotyping of parental origin, variants within repetitive regions were removed using *RepeatMasker* (v.3.2.9) data, resulting in 64,591 SNPs (average density of 1/187bp) and 3973 INDELs (average density of 1/3043bp) present at similar densities across all 16 chromosomes. Such stringent filtering allows for the inclusion of INDELs, which are traditionally difficult to handle, with high confidence. The average length of filtered INDELs is 4.13bp. As a

measure of accuracy, the finalised list was compared to a previously established S288c x SK1 SNP table (Marsolier-Kergoat et al. 2017), showing a 91.4% commonality between calls.

In previous studies, hybrid spore NGS or microarray data was aligned against a singular reference (e.g. S288c) (Anderson et al. 2015; Mancera et al. 2008; Martini et al. 2011), however, the accuracy of event calling may be significantly improved using a dual alignment method (M. Crawford, M.J. Neale unpublished)—specifically improving the mappability of SK1 variant dense reads. In lieu of a high quality SK1 reference, *HybridVar* calculates SK1 equivalent coordinates for all called variants, progressively taking into account INDEL-dependent shifts in relative position, and constructs a dual coordinate variant list. *HybridVar* subsequently modifies a user provided genomic reference (e.g. S288c—SGD Jan 2015, R64-2-1) to include all filtered SNPs/INDELs, creating a "pseudo" SK1 reference for secondary alignment (see: Section B2.1).

2.4—Modelling Recombination: Descriptors

Analysis of complex biological phenomena, such as CO/NCO distribution, often necessitates computational or mathematical modelling in order to provide a quantitative and comparative description of the system. Several descriptors have been employed to evaluate CO/NCO distribution including the coefficient of coincidence (CoC) (Muller 1916; Stahl & Foss 2009) and the fitting of gamma (γ) distributions to inter event distances (IEDs)—the distance between successive events along each chromosome (see: Section 1.5.3) (McPeek & Speed 1995; Zhao et al. 1995). However, several caveats exist for both approaches. As previously noted (Zhang et al. 2014), (γ) distributions may be sensitive to changes in unrelated processes, such as alterations to the class I:class II CO ratio, skewing interpretation of global CO interference strength, while calculation of CoC involves extensive binning of data—risking removal of crucial short range information arising through processes such as event clustering. Deviations from theoretical expectations, as defined by fitted distributions can, however, be a valuable source of information in identifying the biological causes of such deviation. Methods applied throughout this chapter therefore utilise mathematical distributions in an attempt to model and describe the pattern of meiotic recombination.

In order to independently verify the suitability of (γ) distributions in the description of meiotic data, commonly applied mathematical distributions were fitted to IED datasets (*msh2*Δ) and scored using three distinct but related model selection criteria—negative log likelihood (NLogL), bayesian information criteria (BIC) and akaike information criterion (AIC)—which are commonly used to assess model quality (Figure 2.2A,B) (Mills & Prasad 1992). Lower NLogL, BIC or AIC scores indicate a more preferred model. Of the fitted models, (γ) distributions produce the lowest selection scores (NLogL—9905, BIC—19824, AIC—19815) and are therefore the most suitable description of IED data (see: Figure 2.2A). (γ) distributions constitute a multivariate, continuous probability distribution characterised by two independent parameters: (i) (γ)α (shape parameter) determines the shape or skewness of the distribution—γ(α) = 1 is a special case reflective of randomness, exhibiting a purely exponential form. When γ(α) > 1, the distribution adopts a rightward skew (Figure 2.2C). The extent to which a system has deviated from randomness is typically proportional to the value of γ(α) and thus γ(α) can constitute a measure of interference strength. (ii) γ(β) (scale parameter) determines the (x) range over which the distribution is stretched.

2.5—Visualising CO Interference

The multivariate nature of γ(α,β) distributions allows for different parameter combinations to describe similar distributions. In other words, reductions in γ(α) may be partially compensated for by increases in γ(β)—or vice versa. Therefore, it is useful to describe interference in a standardised form, taking into accounting this γ(α,β) relationship. In addition to the fitting of (γ) distributions, survival analysis may also be employed in the modelling of recombination data (Chen et al. 2008). Survival analysis is a specialised branch of statistics seeking to determine or describe the expected duration (x) until a specific event occurs and provides several useful descriptors, including the hazard function (h(x)) (Bewick et al. 2004). Derived from fitted mathematical distributions (see: Section B2.2.5), a hazard function describes the *conditional* probability that a specific event (e.g. CO formation) will occur in the interval [x], predicated on the event having not yet occurred.

A)

Distribution	NLogL	BIC	AIC
Gamma	9905	19824	19815
Weibull	9908	19828	19819
Nakagami	9908	19836	19822
GEV	9914	19841	19832
Rayleigh	9922	19851	19846
Log-Logistic	9948	19909	19900
Logistic	9965	19943	19933
Normal	9981	19976	19967
Log-Normal	9992	19997	19987
Exponential	10056	20119	20115
Gaussian	10099	20212	20203

*NLogL = Negative Log-Likelihood
*BIC = Bayesian Information Criteria
*AIC = Akakike Information Criteria

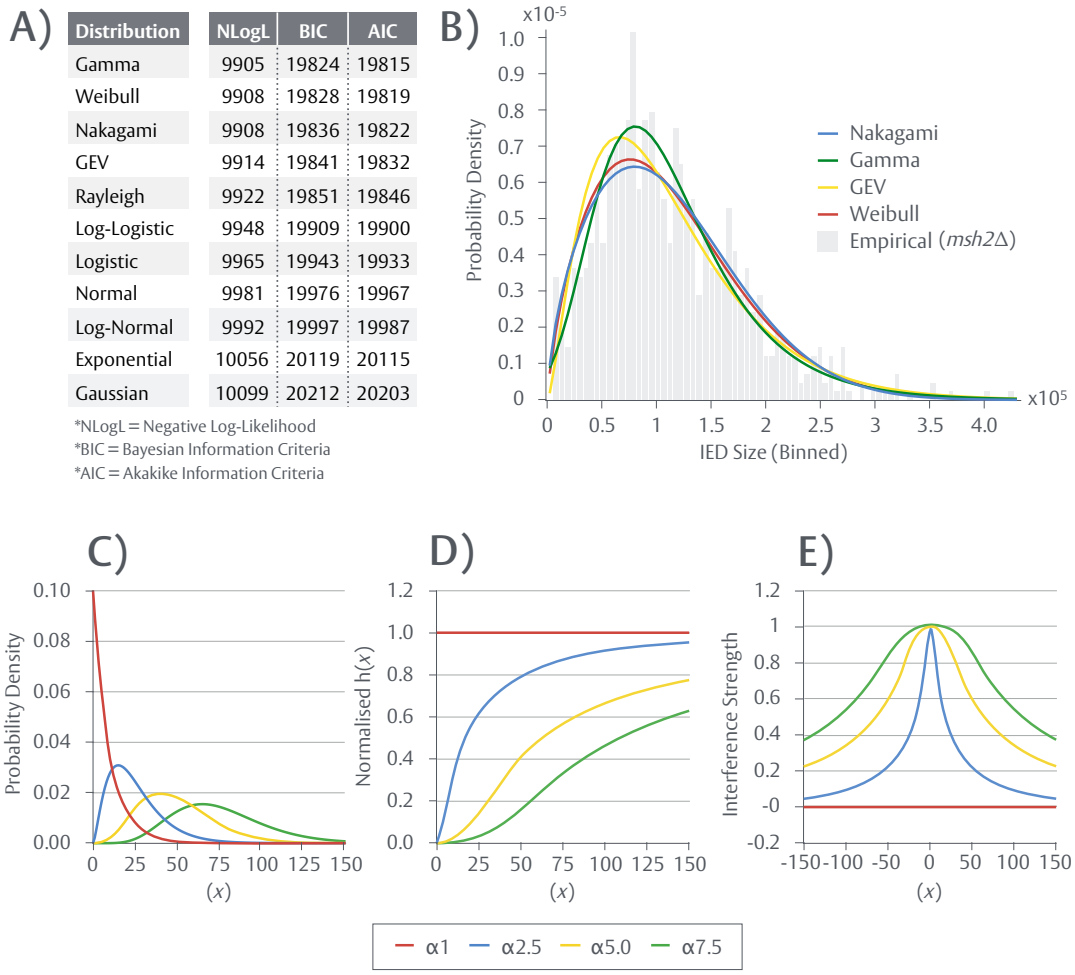


Figure 2.2. A gamma (γ) distribution is the most applicable model for IED data

A) A finite set of parametric probability distributions were fitted to *msh2*Δ CO inter event distances (IEDs) via maximum likelihood estimation (MLE) and scored using three distinct but related model selection criteria (NLogL, BIC, AIC) (MATLAB 2017a Package: *allfitdist*). Scores are sorted in ascending order. The model with the lowest score is preferred and can be considered the most applicable. Model selection criteria provides no information on the goodness of fit. **B)** *msh2*Δ CO IED data was binned at 1kb intervals and plotted as a histogram. Best fit distributions from several putative models, including Gamma (γ), are overlaid. **C)** Probability distribution functions (PDFs) for several (γ) distributions were calculated (MATLAB 2017a Package: *gampdf*) for γ(α) values of 1.0, 2.5, 5.0 and 7.5. γ(β) and (x) values were kept constant at 10 and [0-150] respectively. **D)** Corresponding hazard functions (h(x)) for γ(α) = 1.0, 2.5, 5.0 and 7.5 were calculated as the ratio between the probability distribution function and the inverse cumulative distribution function (PDF/1-CDF). Normalisation of h(x), to a scale of [0.0-1.0] (as shown) is performed by estimating the asymptotic limit of the function and designating this as the maximal value (i.e. 1.0). **E)** Corresponding bidirectional interference functions were constructed from two inverse hazard functions (1-h(x)), oppositely oriented.

Simply, given a pre-existing CO at position x(0), the hazard function describes the probability that another CO will form at any given distance (x) away. For example, a normalised hazard function derived from a (γ) distribution of $\gamma(\alpha) = 1$, representative of randomness, is a flat line at 1.0 i.e. there is an equal and unhindered probability that a second event will occur at any distance (x) away —a state of no interference (Figure 2.2D). In contrast, at $\gamma(\alpha) = 2.5$, the corresponding hazard function adopts an asymptotically increasing form, indicative of a lower probability for events to form in close proximity to one another—an interfering state (Figure 2.2D). An inverted (1-h(x)) function may thus serve as an intuitive characterisation of CO interference—describing the conditional and distance-dependent probability that CO formation will *not* occur adjacent to a pre-existing event. As CO interference acts bidirectionally, interference functions may be constructed by conjugating two mirrored 1-h(x) functions (Figure 2.2E).

2.6—RecombineSim: A novel simulation platform

While mathematical descriptors, such as (γ) distributions or h(x) functions, are an integral part of modelling, they cannot fully reveal or investigate what processes may have given rise to the observed distribution(s). In contrast, simulation platforms—utilising these descriptors as components—permit vigorous testing of hypotheses and possess significant flexibility in their approach, allowing for a closer approximation and understanding of *in vivo* systems.

In order to closely dissect CO/NCO distribution, a novel simulation platform (*RecombineSim*)—specifically adapted to the employed experimental and event processing pipeline—was established (see: Section B2.2, Figure 2.3). *RecombineSim* constitutes a reductionist platform, designed to break processes of spatial regulation down into their fundamental components. A typical simulation run, as depicted in (Figure 2.3), is split into several key processes: (i) Virtual chromosomes are constructed as binned, numerical arrays at a 100bp resolution based on *S. cerevisiae* (S288c) chromosomal lengths, which are adjusted to reflect the limit of experimental detection governed by the leftmost and rightmost genetic markers (SNPs/INDELs), effectively creating short, subtelomeric "dead zones".

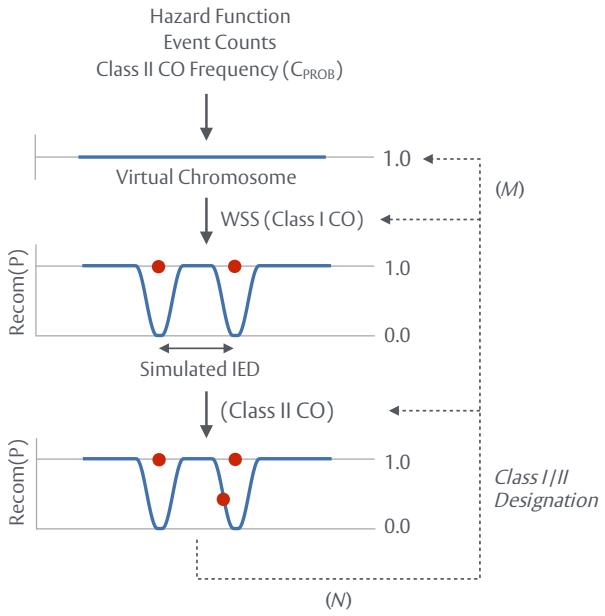


Figure 2.3. RecombineSim—An overview

Hazard functions (h(x)), interference functions, inter event distances (IEDs) and event counts are initially calculated for each experimental sample. Virtual chromosomes are constructed at a 100bp resolution as binned, numerical arrays proportional in size to *in vivo* chromosome length x 0.01 (S. cerevisiae—S288c). Array lengths are adjusted to reflect the limit of experimental detection governed by the leftmost and rightmost genetic markers (SNPs/INDELs). Any given 100bp bin contains a value in the range of [0.0-1.0], designating its recombination potential (recom(P)). Prior to event formation, bins are initially populated with [1.0]—denoting an equal and full recombination potential. *RecombineSim* supports the formation of interfering class I and non-interfering class II COs. Class II CO frequency is set as a decimal fraction in the range of [0.0-1.0] (0-100%) by the parameter $C_{PROB.}$ The position of a class I event is determined by the recom(P) values held in each bin using a weighted site selection algorithm. No such check is performed during NCO simulations or for class II COs. Subsequent to the generation of a class I event, CO interference is applied by multiplying the recom(P) values held within adjacent bins by a bidirectional interference function (see: Figure 2.2E). Successive events falling within a set threshold of one another (e.g. 1.5kb) are merged into a single event residing at the midpoint position. Event formation continues for additional events (*N*) until the experimentally observed number of IEDs is obtained. Finalised datasets are generated by repeating this procedure for additional cells (*M*) (e.g. 10,000) and averaging results.

Any given 100bp bin contains a value that corresponds to its recombination potential (recom(P)) (ii) Event tables, containing data on event assignment, are subsequently imported for a specific genotype and merging threshold. CO/NCO positions are converted into IEDs and event counts per chromosome. Event counts are utilised to exactly match the condition observed within each biological sample (iii) CO or NCO formation is simulated on a per chromosome basis, under interfering, non-interfering or mixed conditions (see: below). The ability to form an event at any given position is governed by the recom(P) value held within the corresponding bin. Event merging is included, as necessary, and event formation continues until the experimentally observed number of IEDs has been obtained (*N*) as opposed to a fixed number of events (iv) Processes outlined in step (iii) are repeated for a specific number of independently simulated cells (*M*), before results are averaged to reduce stochastic noise. All simulations conducted throughout this chapter were performed for M = 10,000 cells.

At its core, *RecombineSim* exploits the mathematical properties of (γ) distributions and associated h(x) functions to guide the spatial formation of virtual COs and NCOs. Several simulation modes are available: (a) *Random* (*No Interference*): Events exhibit total independency, forming without altering the probability (recom(P)) that any adjacent 100bp bin will incur an event (b) *Hazard* (Interference): On a per sample basis, maximum likelihood estimation (MLE) is used to obtain best fit $\gamma(\alpha,\beta)$ parameters for experimental IED data. MLE, a method for estimating statistical parameters, converges on $\gamma(\alpha,\beta)$ values that maximise the likelihood that any given set of IED data would be observed within a model described by those particular parameters. By deriving the corresponding normalised, inverted h(x) function, bidirectional interference functions are constructed (see: Figure 2.2D). Under this mode, event formation alters the probability (recom(P)) that secondary events will occur in flanking 100bp bins in accordance to the interference function, which is superimposed centred on the event (see: Figure 2.3). *RecombineSim* thus builds upon the principles outlined by the beam-film model (see: Section 1.5.3) (Kleckner et al. 2004)—that is, the imposition of interference through a bidirectionally spreading and distance-dependent signal (c) *UniHazard*

(Interference): All cells are simulated under identical interfering conditions based on user specified γ(α,β) parameters rather than individual MLE (γ) fitting of experimental data. *RecombineSim* also supports the formation of non-interfering class II COs, at a given frequency (0-100%)—as set by the parameter C_{PROB}—which form independently of recom(P), thus at random positions, during otherwise interfering CO simulations. Regardless of mode, *RecombineSim* provides a multitude of outputs including event count tables, experimental IEDs (per cell and aggregated), MLE best fit γ(α,β) parameters and simulated IEDs (per cell and aggregated) (see: Section B2.2).

2.7—Single cell variability: Assessing the quantitative reproducibility of repeats

To investigate the impact DDR mutations may have upon CO and NCO distribution, recombination was mapped using the previously described assay (see: Section 2.2) within wild type (WT), 8 hour *ndt80* arrested (*ndt80AR*) and *msh2*Δ cells plus several DDR mutants: *tel1*Δ*msh2*Δ, *rad24*Δ*msh2*Δ, *msh2*Δ*mec1MN*, *rad24*Δ*ndt80AR* and *ndt80ARmec1MN*, totalling 49 biological samples (N) (Table 2.1) (M. Crawford, M.J. Neale unpublished). Within *ndt80*AR strains, expression of Ndt80—a master regulator of prophase I exit—is placed under the control of a β-estradiol inducible *GAL4.ER GAL–NDT80* promoter construct, allowing for controlled arrest. Prophase I arrest is necessary to rescue the inviability of several pachytene checkpoint deficient strains (see: Section 1.3.3). Due to the inviability of *mec1*Δ strains, expression of Mec1 is placed under control of the *CLB2* promoter (*mec1MN*)—a meiotic null (MN) allele where expression is specifically shutdown during meiosis. A full strain table is available (see: Table 2.2—Appendix).

In order to assess the quantitative reproducibility of biological repeats, coefficients of variation (CV)—the relative deviation from the mean—for event counts per event type (CO/NCO) were calculated (see: Table 2.1). Consistent with the imposition of homeostatic control over CO formation (see: Section 1.3.5), all *msh2*Δ strains, bar *rad24*Δ*msh2*Δ, exhibit tight clustering of CO frequency (<±8% of the mean). Variation of ±22% within *rad24*Δ*msh2*Δ may reflect a *rad24*Δ-dependent deregulation of homeostatic control.

	Genotype (Type)	N	Avg. Events (±CV)	Avg. CO (±CV)	Avg. NCO (±CV)
	WT (T)	4	109.25 ±8%	74.50 ±9%	34.50 ±7%
MMR-deficient	*msh2*Δ (O)	9	197.78 ±9%	104.78 ±7%	93.00 ±21%
MMR-deficient	*tel1*Δ*msh2*Δ (O)	10	236.40 ±14%	113.90 ±8%	121.40 ±22%
MMR-deficient	*rad24*Δ*msh2*Δ (T)	6	178.33 ±24%	83.00 ±22%	94.67 ±30%
MMR-deficient	*msh2*Δ*mec1MN* (T)	5	328.00 ±12%	146.80 ±6%	180.00 ±21%
MMR-deficient	*ndt80AR* (T)	4	142.00 ±25%	95.75 ±22%	45.75 ±33%
MMR-deficient	*rad24*Δ*ndt80AR* (T)	6	186.00 ±24%	121.67 ±22%	62.83 ±35%
MMR-deficient	*ndt80ARmec1MN* (T)	5	205.80 ±30%	135.00 ±29%	69.80 ±32%

Table 2.1. Experimental samples and event counts

Recombination was experimentally mapped, via tetrad analysis, within eight DNA damage response or DNA repair strains—WT, *msh2*Δ, *tel1*Δ*msh2*Δ, *rad24*Δ*msh2*Δ, *msh2*Δ*mec1MN*, *ndt80AR* (8 hour arrest), *rad24*Δ*ndt80AR* and *ndt80ARmec1MN*, totalling 49 samples. Four strains, as marked, are deficient in mismatch repair in order to bolster NCO detection. Event counts were calculated, via *RecombineSim*, for each individual repeat and averaged for the 1.5kb merging threshold dataset. Variability in event count amongst individual repeats is shown as a coefficient of variation (CV), calculated as the ratio between standard deviation (σ) and the mean (μ) ($CV = \sigma/\mu$). Specific changes in CO or NCO frequency are discussed, where relevant, on a per genotype basis throughout the chapter. O = Octad, T = Tetrad, N = No. of Repeats.

In contrast to most *msh2*Δ strains, wider variations in CO frequency (±22-29%) are observed within *MSH2*+, *ndt80AR* strains. Msh2 status is not expected to significantly impact upon the detection of COs, thus stochastic variation in prophase I length may account for such *ndt80*AR-dependent variability. Independently of Msh2 status, NCOs display universally high levels of variation in frequency (±21-35%) in all backgrounds, except WT. However, as expected, *msh2*Δ does improve the reproducibility of NCO frequency by ~±10-12% and bolsters detection of NCOs. For example, on average, 47.0% of all events detected in *msh2*Δ are NCOs, as opposed to 31.5% and 32.2% within WT and *ndt80AR* respectively. SNP/INDEL density (1/176bp globally) may serve as a hard coded limitation to the detection of narrow NCO events—introducing an element of chance and thus, variation. Moreover, mutants which alter gene conversion tract lengths may additionally result in changes to NCO visibility.

2.8—Single cell variability: Assessing the distributional reproducibility of repeats

Besides quantitative agreement, distributional reproducibility may also be determined through statistical goodness of fit (GoF) tests. A Kolmogorov-Smirnov (KS) GoF test, employed throughout this chapter, constitutes a non-parametric measure that compares the cumulative distribution functions (CDFs) of two samples in order to assess the null hypothesis that both samples derive from identical populations, based on their maximal difference (D_{KS}) (Massey 1951; Miller 1956). (P) values of the KS test effectively describe the probability that, if the null hypothesis is true, the observed CDFs would be as far apart as observed. (P) values may therefore constitute an indirect measure of distributional agreement. As a significance threshold, a (p) value of 0.25 is employed throughout this chapter. It is important to note that a (p) value of 0.95-1.0 does not necessarily denote an identical distribution, but rather a high probability that the test data derived from the same distribution or parental process. Two forms of the KS test exist (Massey 1951; Miller 1956): (i) a one sample KS test determines whether or not a sample derives from a *specified* theoretical distribution, such as a fitted (γ) distribution (ii) a two sample KS test determines whether or not two samples derived from the same, *unspecified* or unknown distribution.

The formation of a variable number of events (*N*) within a finite space (*lim*) (i.e. chromosome or genome length) skews CDFs—which describe the fraction (F(x)) of events below a certain IED size (x). In other words, a higher event count causes a downward shift in IED size as events become more closely spaced due to the finite space within which they can form. An IED distribution produced under identical spatial rules but with a different event count would therefore generate a significantly different CDF, failing or biasing a KS test and undermining the ability to assess distributional agreement. The impact event count has upon CDFs can be readily observed for both simulated (Figure 2.4A) (*RecombineSim* mode: Random) and experimental (Figure 2.4B) (*rad24*Δ*ndt80AR*) data. Notably, higher values of (N) cause a leftward skew. The relationship between (N) and IED size for a given *lim* is, however, linear (Batten & Beutelspacher 1993). Consequently, in order to isolate the distributional identity of any given sample (i.e. isolate $\gamma(\alpha)$ from $\gamma(\beta)$), IED data can be transformed by calculating the product of IED size and event count (IED size x event count). In the case of simulated data (Figure 2.4C), data transformation results in perfectly aligned CDFs despite varying (N), validating this approach, and experimental CDFs tighten —clustering together (Figure 2.4D).

Given the ability of data transformation to facilitate direct comparisons of event distribution and in order to assess the distributional reproducibility of biological repeats, two sample KS tests were subsequently performed in a pairwise fashion between all unique combinations of repeats for a given genotype, using transformed IED data (intra-genotype testing). Despite a wide, absolute range of results, all median (p) values for CO IED distributions are significant ($p_{MEDIAN} > 0.25$) for all genotypes, bar *rad24*Δ*msh2*Δ (Figure 2.5A)—suggesting that on average, the distributional features of COs can be reliably reproduced between single cell repeats. NCO distribution appears similarly reproducible and unaffected by Msh2 status, with all strains displaying significance ($p_{MEDIAN} > 0.25$) (Figure 2.5B). However, as (p) values describe a probability, they are inversely affected by sample size (no. of IEDs)—a given maximal difference (D_{KS}) between CDFs would therefore produce larger (p) values for progressively smaller samples.

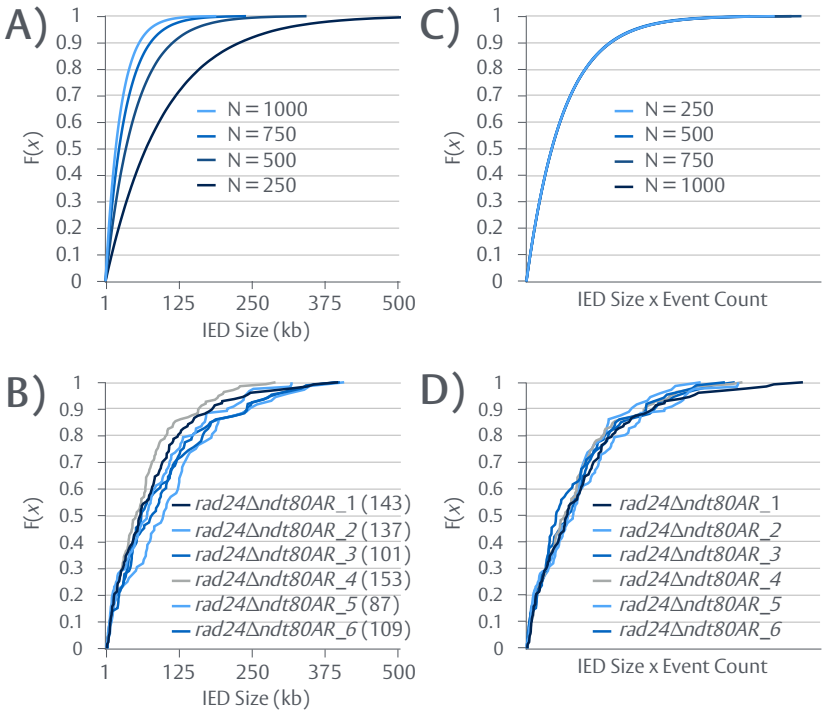


Figure 2.4. Transformation of IED data can account for differences in event count

A) Inter event distances (IEDs) were simulated (*RecombineSim* mode: Random) within a finite, constant space under identical distributional rules ($\gamma(\alpha) = 1.0$) but at varying event counts (N = 250, 500, 750, 1000). Data is visualised as cumulative distribution functions (CDFs). F(x) = Fraction of IED data. **B)** CO IED data for individual *rad24Δndt80AR* repeats (1.5kb merging threshold). The size of IED samples for each repeat are shown. **C)** To account for differential event count, simulated IED data was transformed through multiplication with the corresponding event count (IED size x event count) and replotted. Subsequent to transformation, (x) data takes on arbitrary values and thus (x) scales are omitted for clarity. **D)** Transformed *rad24Δndt80AR* CO IED data.

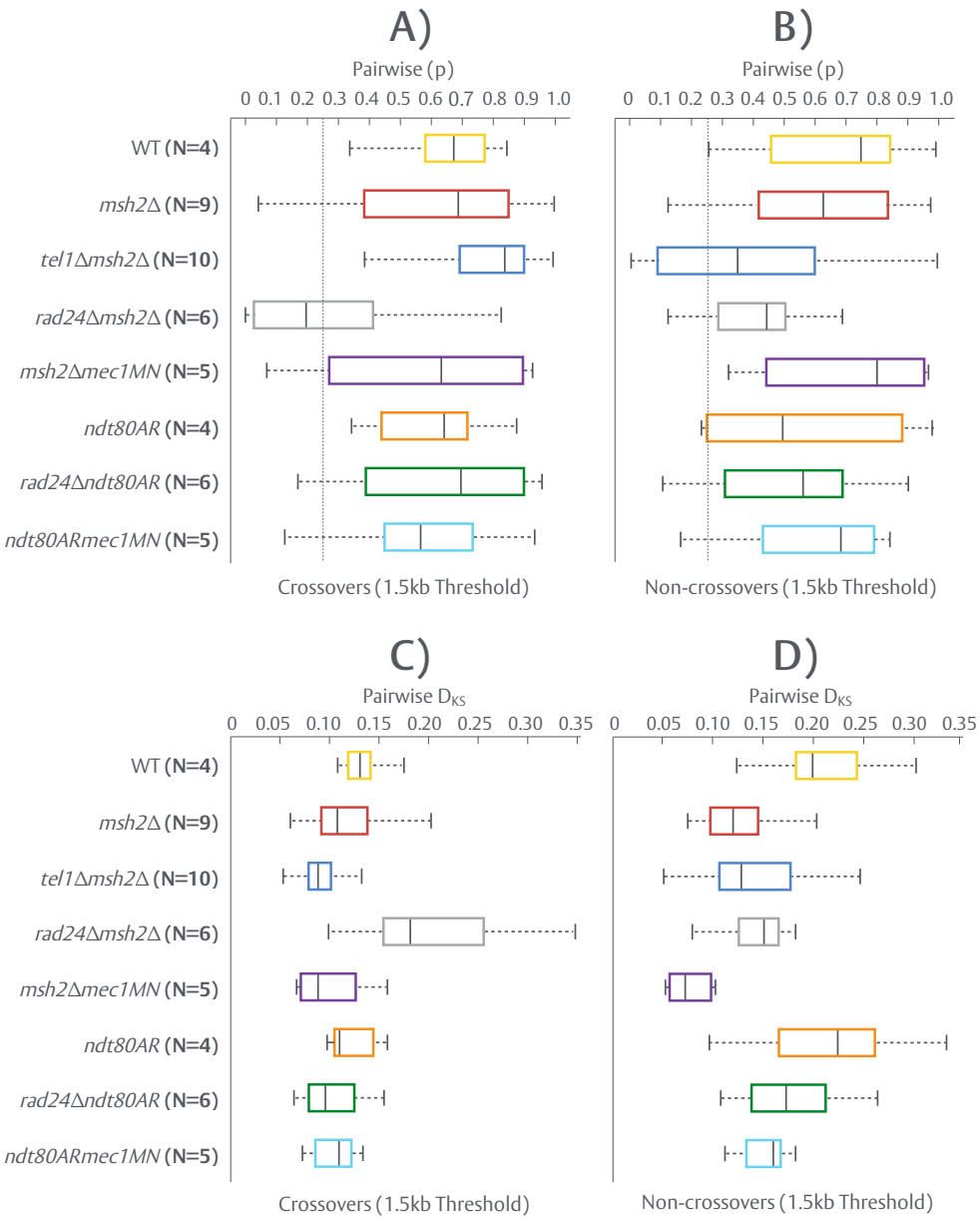


Figure 2.5. Individual repeats are distributionally well correlated (intra-genotype)

Two sample Kolmogorov–Smirnov (KS) tests were performed (MATLAB 2017a Package: *kstest2*) in a pairwise fashion, between all individual repeats of a given genotype, using transformed IED data (1.5kb merging threshold) for COs and NCOs. Resulting stats are visualised as box plots. Absolute maximum and minimum values are displayed as dotted lines. Midlines denote median values. Boxes highlight the second and third quantiles (±25% around the median). **A)** (P) values for COs. **B)** (P) Values for NCOs. As (p) values take into consideration the sample size, which varies between genotypes and event types, D_{KS} values—the maximal distance between tested CDFs—were also considered. **C)** D_{KS} values for COs **D)** D_{KS} values for NCOs. N = No. of Repeats.

In other words, the larger the sample size, the more confidence there is that the null hypothesis can be accurately assessed and therefore small D_{KS} values can have more adverse effects on (p). To account for this relationship, pairwise D_{KS} values were also considered (Figure 2.5C-D). CO IED distributions for all genotypes, bar *rad24*Δ*msh2*Δ, exhibit D_{KS} values within similar ranges of one another (~0.05-0.15), suggesting the level of distributional reproducibility within each strain is similar. In contrast, D_{KS} values for NCOs vary widely, suggesting that the observed similarity in (p) values may be artefacts of sample size. Nevertheless, collectively these observations suggest that spatial rules manifest well and reproducibly from cell-to-cell, particular for COs. Given the overall level of quantitative and distributional agreement, as demonstrated in (Sections 2.7, 2.8), individual repeats were combined into aggregated datasets per genotype to increase statistical power.

2.9—Mutations in the DDR significantly alter CO distribution

To initially assess the global impact DDR mutations may have upon CO or NCO distribution, aggregated IED data was transformed (see: Section 2.8) for all genotypes and cross compared statistically via two sample KS test (inter-genotype testing). Resulting (p) and D_{KS} values were tabulated as a comparative matrix and colour coded to denote significance ($p>0.25$) (Figure 2.6). NCO detection limits are inherently lower within *MSH2*$^{+}$ strains, which may skew distributional features relative to *msh2*Δ strains. Nevertheless, significance is observed between NCO distributions for 11 comparisons (39.2%) within and across the Msh2 status groups, including *msh2*Δ/*tel1*Δ*msh2*Δ, *tel1*Δ*msh2*/*ndt80AR* and *rad24*Δ*msh2*Δ/*ndt80AR* (Figure 2.6A). NCO distribution within WT is poorly correlated with that of all genotypes—attributable to the unusually low event count (34.5/cell) which most likely produces incomplete, variable forms of the full distribution. In contrast, significant differences in CO distribution are observed between all genotypes bar WT/*ndt80AR*, WT/*msh2*Δ*mec1MN* and *ndt80AR*/*msh2*Δ*mec1MN*—suggesting that DDR and, unexpectedly, DNA repair components influence the spatial regulation of CO formation (Figure 2.6B).

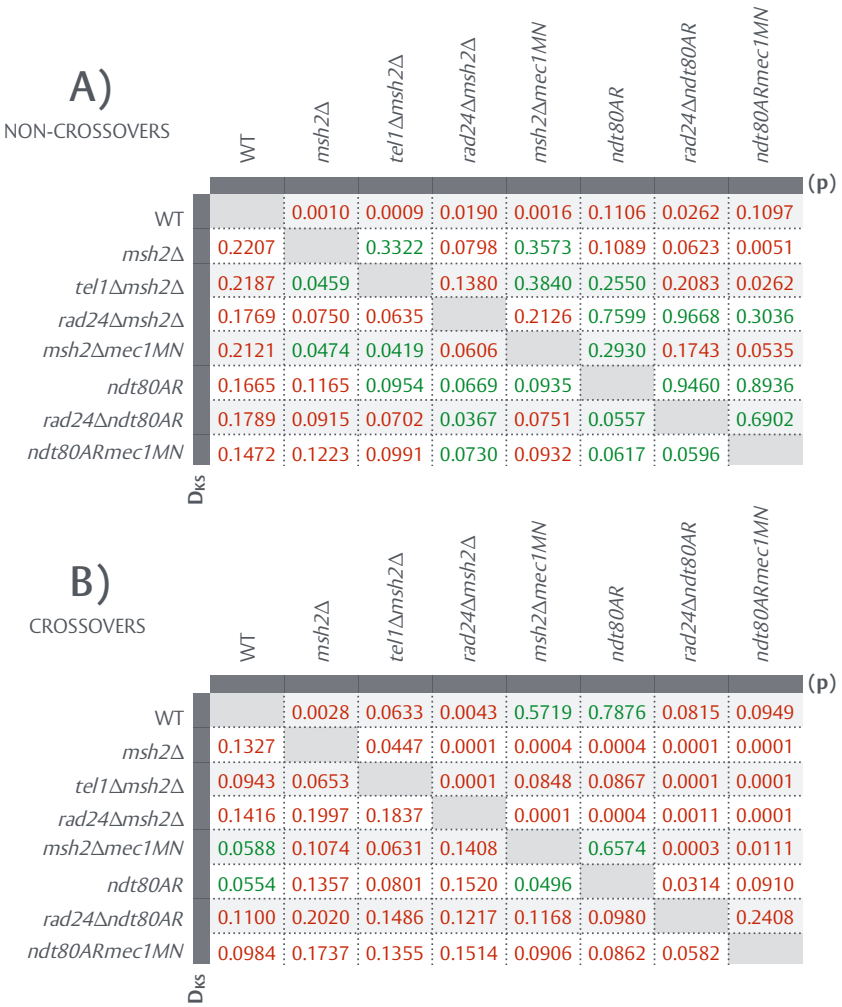


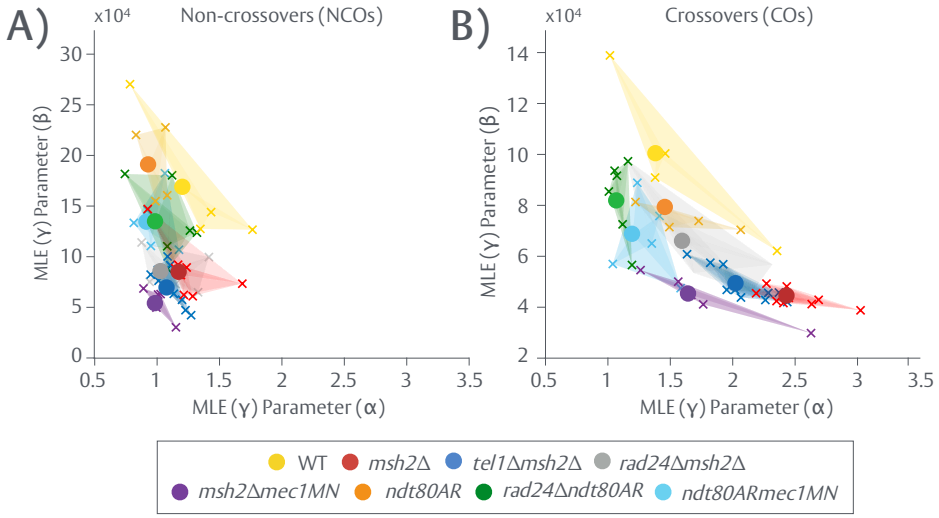
Figure 2.6. Mutations in the DDR significantly alter crossover distributions (inter-genotype)

Two sample KS tests were performed (MATLAB 2017a Package: *kstest2*) in a pairwise fashion between all genotypes, using transformed, aggregated IED data (1.5kb merging threshold) for **A)** NCOs and **B)** COs. Resulting stats are visualised as two sided matrices showing (p) (right side) and D_{KS} (left side) values. Comparisons with (p) values of >0.25, defined here as statistical similarity, are marked in green as are the corresponding D_{KS} values.

In order to further assess whether or not NCO and CO distributions are significantly altered by DDR mutation and provide values for comparison with the literature, IEDs from individual and aggregated datasets were characterised through MLE (γ) fitting to obtain best fit γ(α,β) parameters (see: Section 2.4) (Figure 2.7). Aggregated NCO IED γ(α) values range from 0.95-1.24 and 87.7% of individual repeats possess a γ(α)<1.2, suggesting that NCOs are randomly distributed in all backgrounds—consistent with a lack of spatial regulation specific to NCO events (Figure 2.7A, 2.8C). Wide variation in γ(β) (~73,000-185,000) is caused by differences in event count. In contrast and in line with previous statistical tests, aggregated CO γ(α) values broadly range from 0.97-2.44, suggesting mutations in different components of the DDR may produce distinct CO distributions (Figure 2.7B, 2.8D). While individual repeat CO γ(α) values appear to cluster around a given mean, they are more widely dispersed than for NCOs. Interestingly, CO interference may be diminished within *rad24Δndt80AR* (γ(α) 1.07) and *ndt80ARmec1MN* (γ(α) 1.20). Notably and unexpectedly, CO distributions within *msh2Δ* show no significant correlation to WT or *ndt80AR* when assessed by KS test (Figure 2.6) and MLE (γ) fitting suggests CO interference is significantly stronger within *msh2Δ* (γ(α) 2.44) relative to WT (γ(α) 1.39) and *ndt80AR* (γ(α) 1.46). In contradiction to WT data from this study (γ(α) = 1.39), previously reported γ(α) values for the strength of WT *S. cerevisiae* CO interference range from 1.9 to 2.2 (Anderson et al. 2015; Chen et al. 2008)—suggesting a weaker intensity of CO interference within the utilised S288c x SK1 cross. However, as discussed later in this chapter, (γ) distributions can be a poor comparative measure when taken in isolation.

2.10—Detectable interference between COs but not NCOs

Processes of interference generate interactions between the position of an event E(x) and all subsequent event positions E(y1…n), that manifest as deviations from randomness. Thus, in addition to MLE (γ) fitting (see: Section 2.9), evaluation of global CO interference strength may be achieved through comparison of observed IED distributions with an independent, simulated state whereby events do not generate a flanking interfering signal (see: Section 2.6).



C)

Genotype (NCO)	N	S	α	α [95% CI]		β	β [95% CI]	
WT	4	85	1.24	0.95	1.62	165657	119044	230523
msh2Δ	9	698	1.21	1.10	1.33	87998	78389	98784
tel1Δmsh2Δ	10	1057	1.11	1.03	1.20	73646	66976	80979
rad24Δmsh2Δ	6	474	1.06	0.95	1.19	88589	76817	102165
msh2Δmec1MN	5	820	1.02	0.93	1.11	59296	53170	66128
ndt80AR	4	123	0.96	0.77	1.20	186147	140151	247238
rad24Δndt80AR	6	287	1.02	0.88	1.18	134129	111554	161273
ndt80ARmec1MN	5	273	0.95	0.82	1.10	133785	110535	161925

D)

Genotype (CO)	N	S	α	α [95% CI]		β	β [95% CI]	
WT	4	237	1.39	1.18	1.63	100429	82593	122117
msh2Δ	9	799	2.44	2.23	2.68	44056	39773	48801
tel1Δmsh2Δ	10	979	2.03	1.87	2.21	48931	44570	53718
rad24Δmsh2Δ	6	408	1.60	1.41	1.82	65725	56735	76141
msh2Δmec1MN	5	654	1.65	1.49	1.82	44833	39926	50342
ndt80AR	4	319	1.46	1.27	1.68	79142	66924	93590
rad24Δndt80AR	6	636	1.07	0.97	1.18	81728	72273	92421
ndt80ARmec1MN	5	595	1.20	1.08	1.33	68504	60433	77652

Figure 2.7. Mutations in the DDR significantly alter crossover distributions (γ MLE fitting)

Best fit γ(α,β) values were obtained via maximum likelihood estimation (MLE) (MATLAB 2017a Package: *fitdist*) for aggregated and non-aggregated IED data (1.5kb merging threshold) from each genotype and visualised as 2D cluster diagrams for **A)** NCOs **B)** COs. γ(α,β) values obtained from each individual repeat are marked with 'x'. Polygons highlight the range over which individual repeat γ(α,β) values reside. Aggregated γ(α,β) values are densely marked as 'o'. **C)** Best fit γ(α,β) values for aggregated NCO IED datasets are tabulated along with 95% confidence intervals (CI) (calculated via MATLAB 2017a Package: *fitdist*), specifying the ranges within which the real γ(α,β) values are likely to reside. **D)** Best fit γ(α,β) values for aggregated CO IED datasets are tabulated along with 95% CI. N = No. of Repeats. S = aggregated IED sample size.

In order to verify and expand upon the results of statistical testing and MLE (γ) fitting (see: Section 2.9), such a comparison was performed (*RecombineSim* mode: Random) and visualised as empirical distribution functions (eCDFs). A semi-log(x) form of the eCDF is used to emphasise lower end features, where critical information is likely to reside. As previously inferred by statistical testing, NCO distribution is largely unaffected by DDR mutation—adopting a form that significantly ($p>0.25$) resembles independence in all backgrounds when assessed by a two sample KS test, consistent with the absence of spatial regulation specific to NCOs (Figure 2.8). Several strains, namely WT (Figure 2.8A), *ndt80AR* (Figure 2.8F), *rad24*Δ*ndt80AR* (Figure 2.8G) and *ndt80ARmec1MN* (Figure 2.8H) show localised deviations from randomness along each distribution curve, however, this is likely a result of their low sample size as fewer NCOs are observed within *MSH2*$^{+}$ backgrounds (see: Table 2.1).

Contrastingly, CO interference is readily apparent within WT, *ndt80AR* and *msh2*Δ—which all exhibit experimental CO distributions significantly different from randomness ($p<0.25$) (Figure 2.9). As suggested by MLE (γ) fitting, the apparent strength of CO interference is unexpectedly increased within *msh2*Δ (Figure 2.9B) relative to WT (Figure 2.9A) and *ndt80AR* (Figure 2.9C), as evidenced by a larger deviation from randomness (rightward skew).

As NCOs do not abide by any discernible spatial rules, as assessed by these methods, all subsequent sections focus on the phenomenon of CO interference and the role(s) Rad24, Mec1, Tel1 and Msh2 may play in governing the process and/or the distribution of COs.

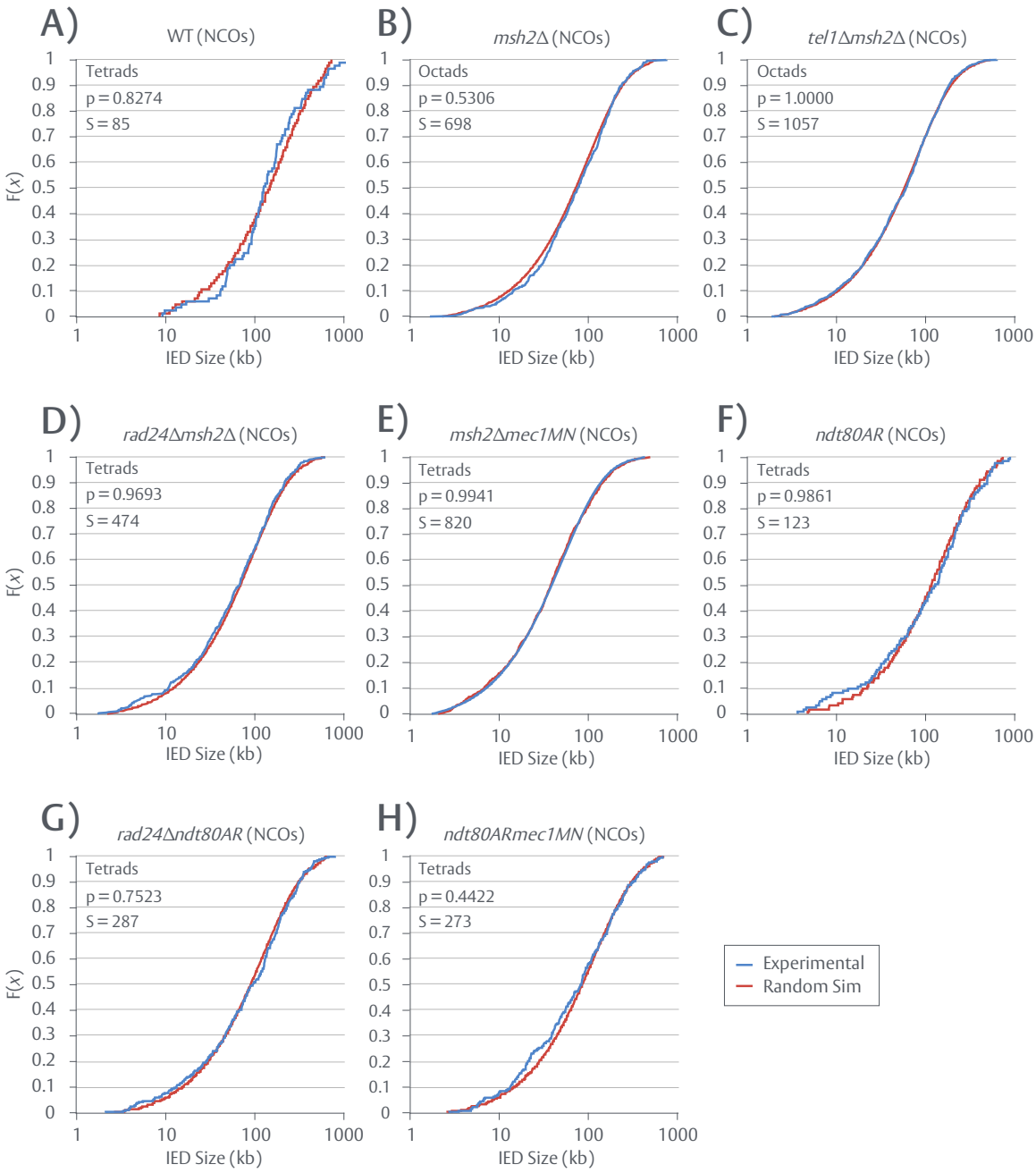


Figure 2.8. Non-crossovers (NCOs) are randomly distributed

Random NCO simulations (*RecombineSim* mode: Random)—whereby event position occurs independently of recomb(P) and no CO interference is applied—were performed for all genotypes: **A)** Mismatch repair (MMR) proficient WT. **B-E)** MMR deficient (*msh2*Δ) strains. **F-H)** MMR proficient, 8h Ndt80 arrested strains. Resulting simulated IEDs (red) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution functions (CDFs). Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). Higher p values signify a higher degree of distributional agreement between tested data. F(x) = Fraction of IED data. S = aggregated IED sample size.

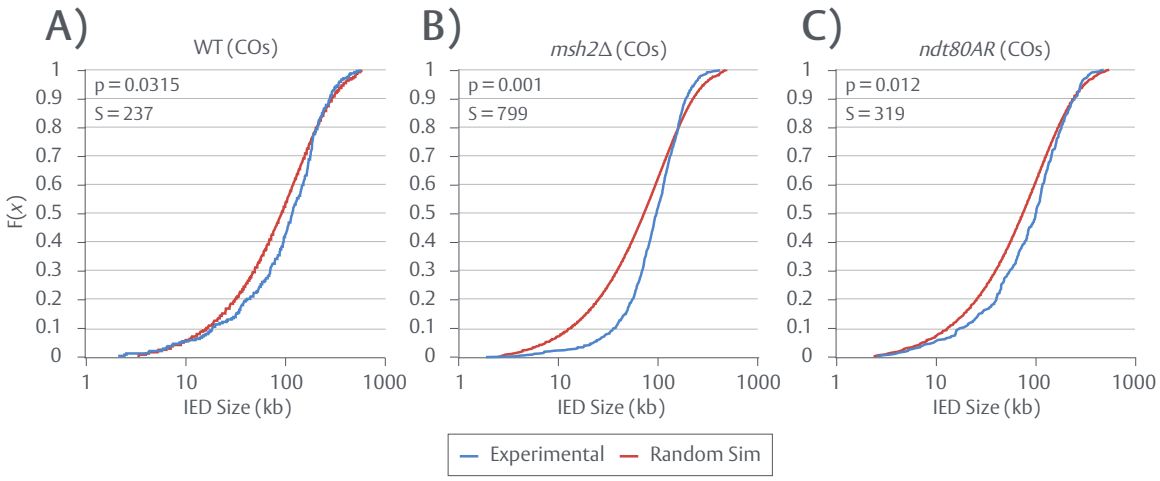


Figure 2.9. CO interference is readily detectable as a deviation from randomness

Random CO simulations (*RecombineSim* mode: Random) were performed for **A)** WT, **B)** *msh2*Δ and **C)** *ndt80AR*. Resulting simulated IEDs (red) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution functions (CDFs). Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size.

2.11—CO interference is highly reduced within Rad24 mutants

Rad24 (RAD17 in *H. sapiens*, *S. pombe*) functions to load a tripartite complex of Rad17-Mec3-Dcd1 (RAD9–RAD1–HUS1 (9–1–1) complex) onto ssDNA:dsDNA junctions as part of the $Mec1^{ATR}$ activation cascade (Lydall et al. 1996; Majka et al. 2006; Majka & Burgers 2003). Rad24 directly governs a number of meiotic processes (Cooper et al. 2014; MacQueen & Hochwagen 2011). For example, Rad24 promotes Zip3/ZMM loading within *S. cerevisiae* independently of Mec1 (Shinohara et al. 2015), suggesting Rad24 may be indirectly required for the formation of ZMM-dependent, interfering class I COs. Nevertheless, a role for Rad24 within the spatial regulation of COs has not been directly observed or characterised.

In order to assess whether or not this branch of the DDR regulates CO distribution, aggregated CO IED datasets from *rad24*Δ*msh2*Δ and *rad24*Δ*ndt80AR* were compared to simulated, random distributions (*RecombineSim* mode: *Random*). As assessed by two sample KS test, *rad24*Δ*ndt80AR* displays significant similarity to randomness ($p = 0.3878$) (Figure 2.10A), but, counterintuitively, *rad24*Δ*msh2*Δ does not ($p = 0.045$) (Figure 2.10B). This difference is further exemplified by their aggregate MLE $\gamma(\alpha)$ values—1.07 vs. 1.60 respectively (see: Figure 2.7D). Instead, *rad24*Δ*msh2*Δ appears to exhibit only a partial loss of CO interference. To visualise the extent of this loss within *rad24*Δ*msh2*Δ—as a random distribution cannot provide a suitable reference point—CO IED data from *rad24*Δ*msh2*Δ and *msh2*Δ was transformed and overlaid (Figure 2.10C). While *rad24*Δ*msh2*Δ is not fully random, loss of Rad24 activity within a *msh2*Δ background still results in a significant change in CO distribution relative to *msh2*Δ, characterised by an enrichment in smaller IEDs (leftward skew) i.e. a weakening of CO interference. Collectively, these results suggest that either (i) prolonged prophase I and the loss of Rad24 synergise to fully inactivate CO interference (*rad24*Δ*ndt80AR*) or (ii) *msh2*Δ partially rescues the loss of CO interference within *rad24*Δ. Nevertheless, CO interference is severely impaired within both *rad24*Δ*ndt80AR* and *rad24*Δ*msh2*Δ, revealing a novel meiotic role for this DDR factor.

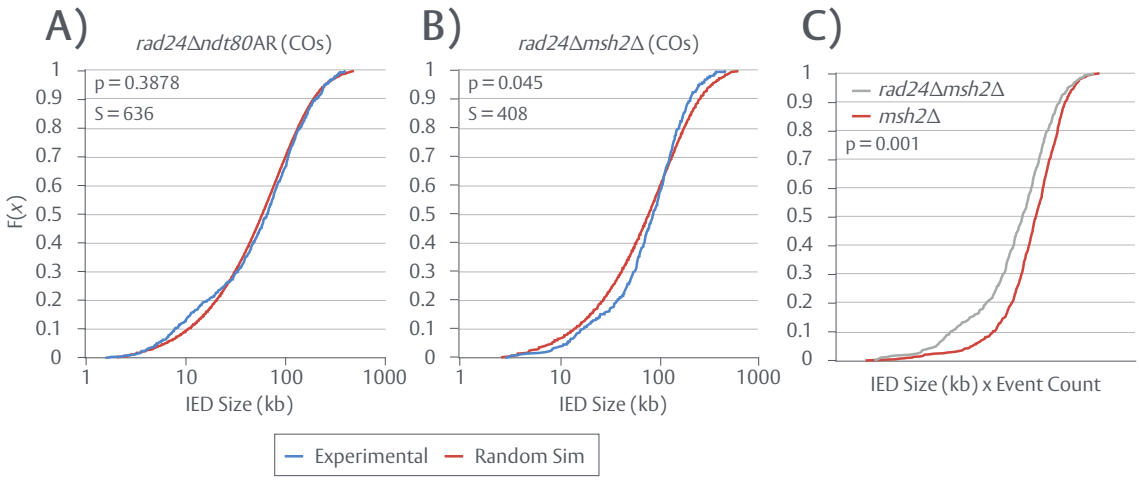


Figure 2.10. Inactivation of Rad24 diminishes the strength of CO interference

Random CO simulations (*RecombineSim* mode: Random) were performed for **A)** *rad24Δndt80AR* and **B)** *rad24Δmsh2Δ*. Resulting simulated IEDs (red) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution functions (CDFs). **C)** Aggregated CO IED data from *msh2Δ* and *rad24Δmsh2Δ* was transformed to take into account differential event count and plotted as CDFs. Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size.

2.12—Single gamma (γ) distribution models insufficiently recapture CO distributions

Of those models tested, (γ) distributions constitute the most applicable distribution in the modelling of IED data (see: Figure 2.2A). However, this does not necessarily mean that (γ) distributions accurately reflect the data. To test whether or not (γ) distributions recapture meiotic recombination, one sample KS tests were utilised to assess the distributional agreement between experimental observation and theoretical expectation according to the respective MLE (γ) best fit—revealing (γ) distributions to be an inadequate ($p<0.25$) description of *in vivo* CO distribution for all genotypes bar *rad24Δndt80AR* (Figure 2.11A). By contrast, (γ) distributions describe NCO distributions with statistical significance in all genotypes bar *tel1Δmsh2Δ*, *msh2Δmec1MN* and *ndt80ARmec1MN*—collectively suggesting that (γ) distributions are only suitable for situations where events are randomly distributed. To investigate why (γ) distributions fail to accurately describe interfering datasets, fractional ratios (f(xE)/f(xT)) between experimental eCDFs and theoretical expectations according to the respective best fit MLE (γ) distributions were taken at 1kb intervals. Because an unintended consequence of event merging (see: Section 2.2) may be responsible for the insufficiency of (γ) distributions to fit the data well, this analysis was performed on annotated (Figure 2.11B), 1.5kb (Figure 2.11C) and 5kb (Figure 2.11D) threshold datasets. (γ) distribution estimates perform well at mid-high IED ranges (>50kb, ratio ≈ 1) but universally fail at short IED ranges (<50kb, ratio ≠ 1) in all genotypes and independently of the merging threshold used. In other words, *in vivo* CO distributions primarily exhibit higher levels of shorter IEDs than theoretically expected by the respective best fit (γ) distribution. Consistent with these findings, interfering hazard function simulations (*RecombineSim* mode: *Hazard*) yield reasonable model fits when utilising single fit (γ) distributions as a basis but fail at the lower end of the IED distribution (<50kb) (Figure 2.12A-G).

A)

Genotype	Non-crossovers D_{KS}	Non-crossovers (P)	Crossovers D_{KS}	Crossovers (P)
WT	0.0334	0.3146	0.0651	0.0217
msh2Δ	0.0261	0.4011	0.0413	0.0036
tel1Δmsh2Δ	0.0280	0.0586	0.0430	0.0004
rad24Δmsh2Δ	0.0239	0.7754	0.0444	0.0580
msh2Δmec1MN	0.0530	0.0001	0.0372	0.0375
ndt80AR	0.0359	0.9732	0.0561	0.0245
rad24Δndt80AR	0.0374	0.4625	0.0283	0.3425
ndt80ARmec1MN	0.0553	0.0558	0.0376	0.0542

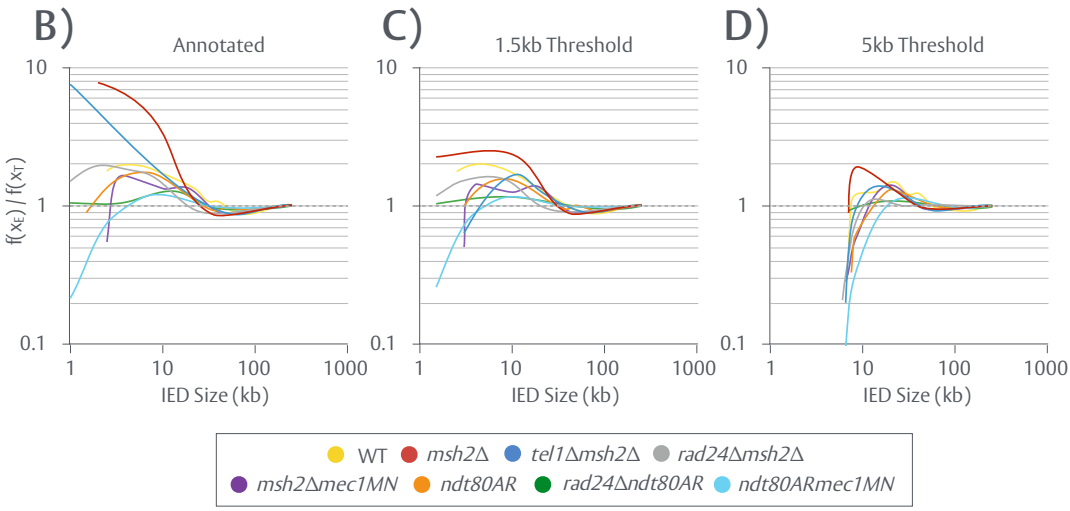
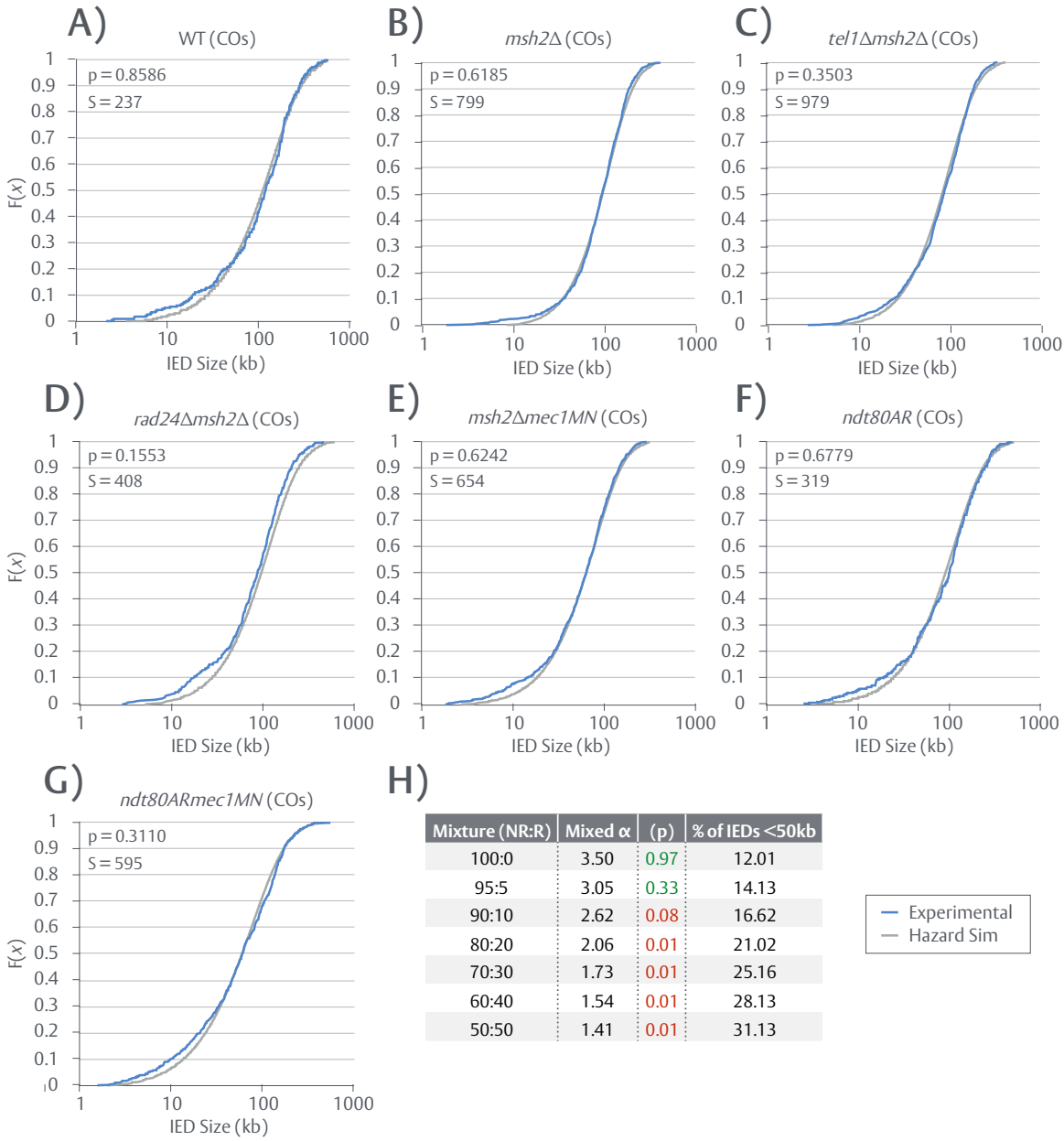


Figure 2.11. Single gamma (γ) distribution models insufficiently recapture CO distributions

A) One sample KS tests were performed (MATLAB 2017a Package: *kstest*) between aggregated CO or NCO IED data (1.5kb merging threshold) and theoretical expectations according to the respective MLE best fit (γ) distribution for all genotypes. Resulting (p) and D_{KS} values are tabulated. Entries containing (p) values of >0.25, defined here as statistical similarity, are marked in green as are the corresponding D_{KS} values. **B)** Empirical ($f(X_E)$) and theoretical ($f(X_T)$) CDFs were evaluated at 1kb (x) intervals and compared as a ratio ($f(X_E)/f(X_T)$) for manually annotated, aggregated CO IED datasets. Ratios are shown on double log plots. Ratio values of >1 denote that the experimental data contains IEDs, of a given size, at a higher frequency than theoretically expected. Ratio values of <1 denote a lower frequency. **C)** Ratio analysis repeated for aggregated CO IED datasets merged at a 1.5kb threshold. **D)** Ratio analysis repeated for aggregated CO IED datasets merged at a 5kb threshold.



H)

Mixture (NR:R)	Mixed α	(p)	% of IEDs <50kb
100:0	3.50	0.97	12.01
95:5	3.05	0.33	14.13
90:10	2.62	0.08	16.62
80:20	2.06	0.01	21.02
70:30	1.73	0.01	25.16
60:40	1.54	0.01	28.13
50:50	1.41	0.01	31.13

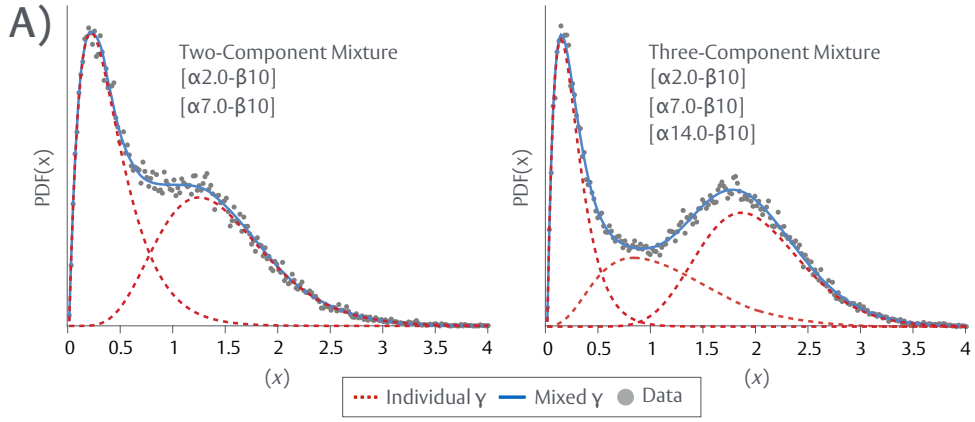
Figure 2.12. Hazard function simulations insufficiently recapture CO distributions

A-G) Interfering CO simulations (*RecombineSim* mode: Hazard)—whereby event position is dependent upon recomb(P) and CO interference is applied as a bidirectional hazard function derived from best fit γ(α,β) values of each individual repeat—were performed for all genotypes except *rad24Δndt80AR*, which is readily described through random simulation. Resulting simulated IEDs (grey) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution functions (CDFs). Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size. **H)** Interfering simulations, using *msh2Δ* CO event counts, were repeated for different mixtures of non-randomness (NR) (γ(α) = 3.5) and randomness (R) (γ(α) = 1.0)—introduced via the class II CO C_{PROB} parameter. MLE best fit γ(α,β) values were obtained for the mixed simulation (MATLAB 2017a Package: *fitdist*) and the quality of (γ) fit was assessed by one sample KS test (MATLAB 2017a Package: *kstest*). Resulting (p) values of >0.25, defined here as statistical similarity, are marked in green. The fraction of IEDs residing below 50kb was calculated for all simulated mixtures.

As previously noted, the distributional analysis of COs may be complicated by the existence of non-interfering, Mus81-Mms4-dependent class II COs. Class II COs are essentially indistinguishable from class I COs when mapping recombination via the described assay (see: Section 2.2), thus creating a mixed system with unknown quantities of each subclass. In order to ascertain how the presence of class II COs may impact upon the ability to model CO distributions, simulated, interfering datasets (*RecombineSim* mode: *Hazard*) that contain progressively increasing levels of class II COs were analysed. Simulated introduction of randomness into an otherwise non-random system has a pronounced affect upon MLE (γ) estimates and associated fit quality (p)—characterised by enrichment in smaller sized IEDs (<50kb) (Figure 2.12H). Notably, even moderate levels of class II COs (10%) are sufficient to reduce γ(α) from 3.5 to 2.62 and reduce fit quality (p) below significance ($p = 0.08$). Collectively, these observations suggest that the experimental data contains a significant amount of class II COs or an alternative, unspecified form of randomness and secondly, that the presence of a random component prevents simple modelling of interfering systems.

2.13—Gamma (γ) mixture modelling successfully identifies simulated IED subpopulations

Isolating and estimating class II frequency via genetic means (e,g, *mus81*Δ, *mms4*Δ) is complicated by the phenomenon of CO compensation—whereby class II production increases in lieu of class I COs or vice versa (Argueso et al. 2004; Gray & Cohen 2016)—and through the inviability of particular crosses (M. Crawford, M.J. Neale unpublished). However, as suggested in (Section 2.12), the presence of class II COs may prevent accurate modelling of the experimental data, necessitating a further refinement of analytical methods. Latent variables (e.g. class II COs) can be inferred through probabilistic and statistical methods, such as mixture modelling. A finite mixture model provides a natural representation of heterogeneity within a population, for a prespecified number of hidden classes. For example, a multi-component (γ) mixture model assumes all data points are generated from a mixture of (γ) distributions with unknown parameters (Figure 2.13A). Expectation maximisation (EM) is a commonly employed, iterative clustering technique used to estimate the parameters and weighted contributed of each sub class within a mixture model (Do & Batzoglou 2008).



B)

		Actual Mixture (Simulated)							Estimated Mixture (*GEM*)						
	S	α1	β1	α2	β2	W_1	W_2		α1	β1	α2	β2	W_1	W_2	N(%Δ)
Variable S	250	3.00	35000	1.00	75000	0.75	0.25		3.45	31139	0.89	84392	0.69	0.31	13.14
	500	3.00	35000	1.00	75000	0.75	0.25		3.29	32889	1.11	69448	0.72	0.28	8.16
	1000	3.00	35000	1.00	75000	0.75	0.25		2.91	35917	0.94	79650	0.78	0.22	5.79
Variable W	1000	3.00	35000	1.00	75000	0.90	0.10		2.83	37110	0.92	78595	0.91	0.09	6.28
	1000	3.00	35000	1.00	75000	0.50	0.50		3.08	36894	0.94	77998	0.48	0.52	4.24
Higher α	1000	4.00	25000	2.00	50000	0.75	0.25		4.09	23976	1.95	51704	0.77	0.23	4.21
	1000	5.00	25000	3.00	50000	0.75	0.25		5.10	25997	3.06	51874	0.78	0.22	4.55

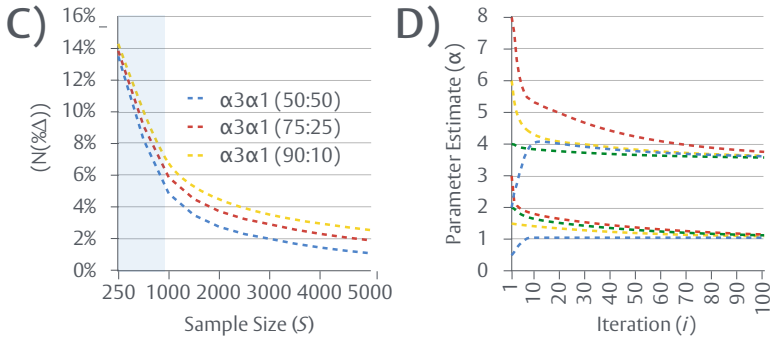


Figure 2.13. Gamma (γ) mixture modelling successfully identifies simulated IED subpopulations

A) Data was randomly sampled from a mixture of (γ) distributions (γ(α,β) values as marked) (MATLAB 2017a Package: *gamrnd*), mixed with equal weight. Probability distribution functions (PDFs) are shown for each individual (γ) distribution (red) in the mixture (blue). **B)** A gamma (γ) expectation maximisation (*GEM*) algorithm was utilised to resolve and estimate individual components of simulated two component (γ) mixtures with known parameters (α,β), at known weights (W)—generated via *RecombineSim* and direct mixed (γ) sampling. A set of representative examples are shown. Percentage differences between actual and estimated parameters, obtained via *GEM*, are calculated and averaged to estimate error rate (N(%Δ)) and algorithm accuracy. S = No. IEDs. Max no. of EM iterations (*i*) permitted = 1000. **C)** N(%Δ) values for three (γ) mixtures were calculated for varying sample size (S) **D)** *GEM* was provided with biased parameter initiations (randomised γ(α,β) values) for a two component (γ) mixture ($\alpha = 4 / \alpha = 1$) (S = 1000) and convergence toward maximum likelihood estimates was evaluated at each expectation maximisation iteration (*i*).

In order to statistically isolate class II COs, a (γ) EM algorithm (*GEM*) was developed (see: Section B2.3). Briefly, *GEM* initially segregates data into a user specified number of distributions ($n_{COMPONENTS}$) in a biased or unbiased manner and subsequently performs soft clustering—a process whereby each data point is assigned a probability reflective of how likely it is to belong to any given sub distribution. Biased initiation utilises a set of user specified γ(α,β) values to initiate the cluster. Unbiased initiation utilises a kmeans++ algorithm—an established method for the initiation of parameters (Blömer & Bujna 2013). By reiteratively cycling between an expectation step (E) and a maximisation step (M), incrementally shifting each distribution and updating data point assignment, the system repeats this process until it converges upon a maximum likelihood solution. This approach not only provides estimates of γ(α,β) values for each subpopulation (α(1..n),β(1..n), but also the relative contributions of each subpopulation (W(1..n)) to the mixture. In biological terms, for a two component system, these weights may provide an estimate of the class I:class II ratio.

In order to validate the *GEM* algorithm for use with IED data, simulated IED datasets of two component mixtures with *known* parameters, at variable (S) sample sizes, were generated by *RecombineSim* and tested (Figure 2.13B). Subpopulation I is derived from interfering COs distributed non-randomly (α1β1 values). A non-interfering subpopulation II (α2β2 values) is introduced at varying levels (W_2) via the *RecombineSim* class II frequency parameter (C_{PROB}). For mixtures devoid of a non-random component (Figure 2.13B—"higher α"), test datasets were created via direct sampling of mixed (γ) distributions. As a measure of the algorithm's accuracy and ability to resolve (γ) mixtures, the average percentage difference (N(%Δ)) between estimated parameters and actual parameters—those used to generate the test data—were calculated. Analysis was conducted on 250 test (γ) mixtures, repetitively sampled from 25 distinct sets of γ(α,β) and W_1W_2 values. Error rates, for each set, were subsequently averaged (Figure 2.13C). Importantly, *GEM* is capable of successfully resolving (γ) mixtures (see: Estimated Mixture values), however, accuracy

is strongly dependent on sample size (S) and to a lesser extent on the relative proportions of each subpopulation—and thus how likely it is to be readily observed in a mixture. For example, (γ) mixtures containing 10% or 25% class II COs exhibit average errors of 10.0% and 9.1% at (S) = 500 and 4.53% and 3.73% at (S) = 2000 respectively. Experimental CO datasets range from (S) values of 237 to 979, therefore reasonable error rates of 6.5-14.5% can be expected when estimating experimental mixtures. The method of parameter initiation employed may skew mixture modelling results toward a certain outcome (Blömer & Bujna 2013). However, despite using randomised sets of initial γ(α,β), *GEM* robustly avoids local maxima in favour of global maxima, converging on the correct γ(α,β) estimates (Figure 2.13D). Collectively, these results demonstrate the applicability of (γ) mixture modelling to IED data.

2.14—Two component gamma (γ) models significantly improve model-experimental fit

To test whether or not (γ) mixture modelling improves the ability to recapture experimental CO distributions, *GEM* was applied to aggregated CO IED data for all genotypes ($n_{COMPONENTS} = 2$) (Figure 2.14A) bar *rad24*Δ*ndt80AR*—which is readily described by a single component system (see: Section 2.11). Non-random γ(α > 1.5) (α1β1 values) and random γ(α ≈ 1) (α2β2 values) components were successfully isolated for all genotypes, except *tel1*Δ*msh2*Δ—which resolved into two, non-random components (γ(α) 4.45 and 1.97). Putative estimates of class I:class II ratios (w1w2 values) suggests the relative frequency of class II CO formation varies widely within DDR mutant backgrounds (discussed on a per genotype basis in subsequent sections). To determine whether a mixed (γ) fit improves simulation of the experimental data, simulations (*RecombineSim* mode: *UniHazard*) were conducted using the newly derived non-random (α1β1) parameters obtained from *GEM* as a basis for h(x) interference, while class II CO formation occurs at estimated frequencies (w2). As assessed by two sample KS test, model fit is greatly improved for all genotypes when considering a mixed (γ) solution (see: $D_{KS(M)}$ and $P_{(M)}$ values)—an improvement characterised by the elimination of lower end deviations from theoretical expectation, previously observed using single (γ) fit models (reshown in Figure 2.14B).

A)

	Single Fit			Estimated Mixture (*GEM*)						Single Fit		Mixed Fit		
Genotype	**S**	**α_S**	**B_S**	**α1**	**β1**	**α2**	**β2**	**W_1**	**W_2**	**$D_{KS(S)}$**	**$P_{(S)}$**	**$D_{KS(M)}$**	**$P_{(M)}$**	**N(%Δ**
WT	237	1.39	100429	3.48	37696	1.06	107930	0.671	0.329	0.055	0.859	0.046	0.956	14.2%
msh2Δ	799	2.44	44056	3.84	29877	1.21	68155	0.846	0.154	0.038	0.619	0.029	0.891	7.4%
tel1Δmsh2Δ	**979**	**2.03**	**48931**	**4.45**	**28187**	**1.97**	**23104**	**0.774**	**0.226**	**0.042**	**0.350**	-	-	**4.8%**
rad24Δmsh2Δ	408	1.60	65725	4.25	18994	1.34	75454	0.135	0.865	0.078	0.155	0.047	0.759	10.4%
msh2Δmec1MN	654	1.65	44833	3.56	35114	1.37	27035	0.587	0.414	0.041	0.624	0.034	0.847	8.4%
ndt80AR	319	1.46	79142	3.22	47109	1.40	32856	0.660	0.340	0.056	0.678	0.040	0.898	12.7%
ndt80ARmec1MN	595	1.20	68504	3.25	32718	1.29	24840	0.507	0.494	0.056	0.311	0.045	0.563	8.1%
mms4Δmsh2Δ	327	2.10	54497	3.43	35924	1.08	77128	0.959	0.041	0.024	0.927	0.040	0.955	12.6%

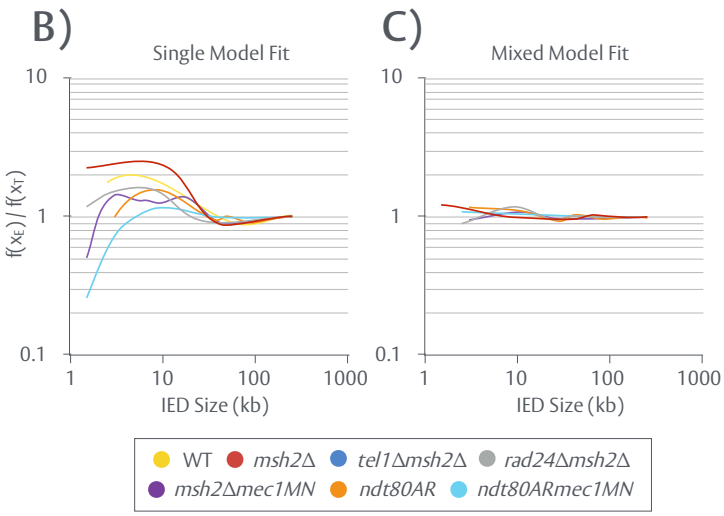


Figure 2.14. Two component gamma (γ) models significantly improve model-experimental fit

A) A gamma (γ) expectation maximisation (*GEM*) algorithm was utilised to resolve experimental, aggregated CO IED data into two components and estimate γ(α,β) and weight (W) parameters of each subcomponent for all genotypes, except *rad24Δndt80*AR. *mms4Δmsh2Δ* data, from an independent source (Oke et al. 2014) was also analysed. Non-random (α1β1w1) and random (α2β2w2) components are tabulated if detected. Genotypes which did not resolve into a random and non-random component are shown in bold. Mixed model-experimental fits were assessed via two sample KS tests (MATLAB 2017a Package: *kstest2*) between experimental data and simulated IED data (*RecombineSim* mode: UniHazard) using GEM results as a basis (see: Figure 2.16). Green values indicate an improvement in model-experimental fit as determined by (p). (N(%Δ)) values are estimated based on experimental sample size (S) using previously conducted test as a reference (see: Figure 3.14B). S = aggregated IED sample size. Max no. of EM iterations (*i*) permitted = 1000. **B-C)** Empirical ($f(X_E)$) and theoretical ($f(X_T)$) cumulative distribution function ratios were re-evaluated at 1kb (x) intervals for simulated, mixed model IED data. Ratios obtained from single (γ) fits (see: Figure 2.12) are shown for comparison.

Notably, recalculated fractional ratios (f(xE)/f(xT)) between experimental eCDFs and mixed (γ) models yield values of ≈ 1 across the full spectrum of IED sizes, however some minor deviation below 50kb still occurs (Figure 2.14C). As a further point of comparison with single (γ) simulations (see: Figure 2.12), mixed (γ) simulations—created as detailed above—were overlaid with experimental eCDFs for all modelled genotypes (Figure 2.15). Notably, the distribution of simulated COs is virtually indistinguishable from that of experimental observation within WT, *ndt80AR* and *msh2*Δ—further highlighting the improvement that (γ) mixture modelling provides (see: Figure 2.15A,B,E).

GEM was also applied to *mms4*Δ*msh2*Δ IED data from an independent source (Oke et al. 2014) (N = 4 tetrads) (S288c x YJM789 cross) (see: Figure 2.14A)—a strain within which class II CO formation is abolished. Notably, IED data from *mms4*Δ*msh2*Δ is significantly more amenable to single fit (γ) modelling than *msh2*Δ ($p = 0.619$ vs. $p = 0.927$) and is estimated to have a negligible class II frequency of ~4%. Such an observation supports the notion that class II COs are predominately responsible for the observed deviations from theoretical expectation, the inability of single fit (γ) modelling to recapture CO distributions and that *GEM* is accurately detecting class II subpopulations. Collectively, these results demonstrate that (γ) mixture modelling (via *GEM*) is an improved and effective method for the modelling of *in vivo* CO distributions over single (γ) fitting.

2.15—Gamma (γ) mixture modelling reveals a putative description of WT CO interference

To further strengthen the validity of (γ) mixture modelling results, *msh2*Δ and WT datasets from independent sources (S288c x YJM789 cross) (Fung *msh2*Δ N = 4, Fung WT N = 4, Steinmetz WT N = 49), totalling 57 additional tetrads (N), were analysed via *GEM* (Figure 2.16A) (Mancera et al. 2008; Oke et al. 2014). As before, the quality of model fit was assessed via two sample KS test for single (γ) fits and simulated mixed (γ) fits—demonstrating a similar improvement in the ability to simulate CO distributions for independent datasets when using a mixed system (Figure 2.16A).

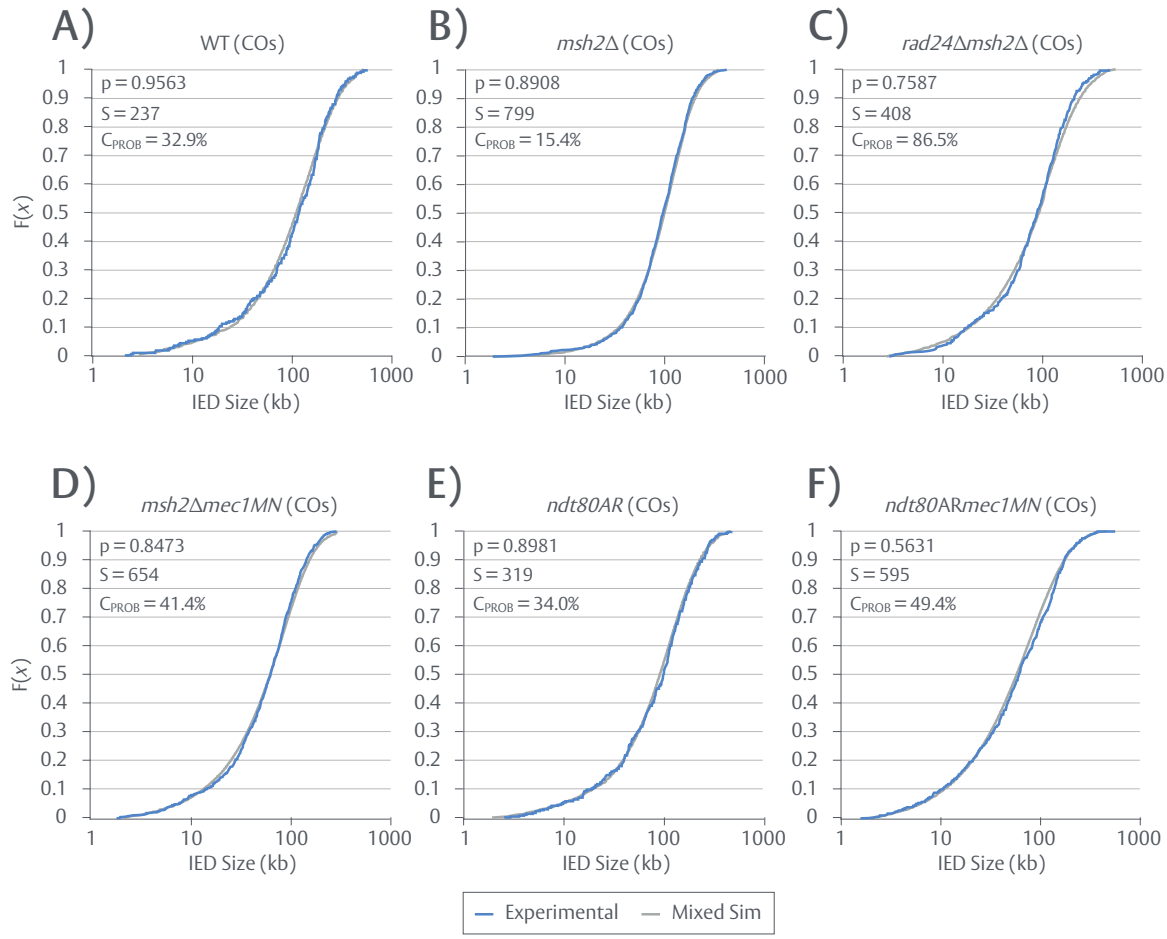


Figure 2.15. Mixed hazard function simulations significantly recapture CO distributions

A) Mixed, interfering CO simulations (*RecombineSim* mode: UniHazard)—whereby event position is dependent upon recomb(P), CO interference is applied as a bidirectional hazard function derived from user specified γ(α,β) values (obtained via *GEM*) and class II CO formation occurs at a given rate (C_{PROB})—were performed for all genotypes except *rad24*Δ*ndt80*AR and *tel1*Δ*msh2*Δ. Class II CO frequency was set based on estimated W_2 values obtained via *GEM* (see: Figure 2.15A). Resulting simulated IEDs (grey) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution functions (CDFs). Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size. F(x) = Fraction of IED data.

A)

	Single Fit			Estimated Mixture (*GEM*)						Single Fit		Mixed Fit		
Genotype	**S**	**α_S**	**B_S**	**α1**	**β1**	**α2**	**β2**	**W_1**	**W_2**	**$D_{KS\,(S)}$**	**$P_{(S)}$**	**$D_{KS\,(M)}$**	**$P_{(M)}$**	**N(%Δ**
WT	237	1.39	100429	3.48	37696	1.06	107930	0.671	0.329	0.055	0.859	0.046	0.956	14.2%
WT (Fung)	334	1.92	57941	3.12	35026	1.08	117045	0.710	0.292	0.087	0.153	0.030	0.948	12.6%
WT (Steinmetz)	5499	1.85	66173	3.17	48145	1.19	89211	0.660	0.341	0.023	0.318	0.007	0.977	1.8%
WT (Average)	-	**1.72**	**74848**	**3.26**	**40289**	**1.11**	**104729**	**0.680**	**0.321**	-	-	-	-	-
*msh2*Δ	799	2.44	44056	3.84	29877	1.21	68155	0.846	0.154	0.038	0.619	0.029	0.891	7.4%
*msh2*Δ (Fung)	403	1.89	52308	3.82	29716	1.16	27644	0.823	0.177	0.109	0.015	0.024	0.940	10.4%
*msh2*Δ (Average)	-	**2.17**	**48182**	**3.83**	**29797**	**1.19**	**47900**	**0.835**	**0.166**	-	-	-	-	-

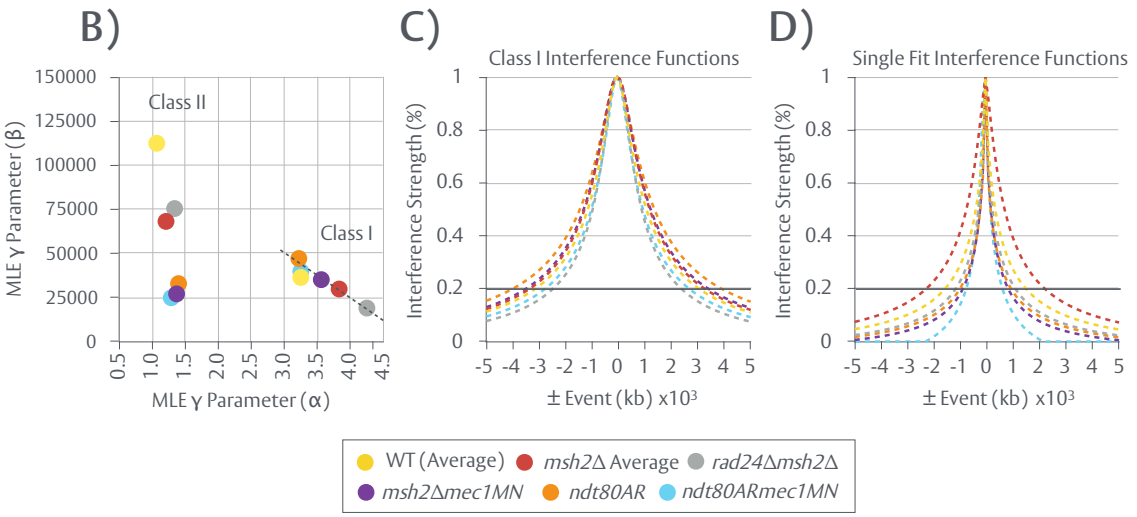


Figure 2.16. Gamma (γ) mixture modelling reveals a putative description of WT CO interference

A) A gamma (γ) expectation maximisation (*GEM*) algorithm was utilised to resolve experimental, aggregated CO IED data into two components and estimate γ(α,β) and weight (W) parameters of each subcomponent for WT and *msh2*Δ data sourced from independent datasets (WT Fung, WT Steinmetz, *msh2*Δ Fung) (Mancera et al. 2008; Oke et al. 2014). Non-random (α1β1w1) and random(α2β2w2) components are tabulated. Mixed model-experimental fits were assessed via two sample KS tests (MATLAB 2017a Package: *kstest2*) between experimental data and simulated IED data (*RecombineSim* mode: UniHazard) using *GEM* results as a basis. Green values indicate an improvement in model-experimental fit as determined by (p). (N(%Δ)) values are estimated based on experimental sample size (S) using previously conducted test as a reference. S = aggregated IED sample size. Max no. of EM iterations (*i*) permitted = 1000. **B)** Non-random (class I) and random (class II) γ(α,β) *GEM* estimates are shown on a 2D cluster diagram for each mixture modelled genotype. **C)** Interference functions were constructed using non-random class I γ(α,β) estimates obtained via *GEM* for all mixture modelled genotypes. **D)** As a comparison, interference functions were constructed using single fit γ(α,β) estimates for all mixture modelled genotypes.

Mixed γ(α,β) parameter estimates for each individual dataset are largely in agreement with those previously determined (see: Section 2.14), suggesting S288c x SK1 and S288c x YJM789 crosses behave in a similar manner. *GEM* estimates were thus combined to create averaged WT and *msh2*Δ parameters, yielding estimated class II frequencies of 32.1% and 16.6% respectively.

Interestingly, isolated non-random (class I) estimates for all mixture modelled genotypes—WT (average), *msh2*Δ (average), *rad24*Δ*msh2*Δ, *msh2*Δ*mec1MN*, *ndt80AR* and *ndt80ARmec1MN*—are quantitatively similar, clustering at γ(α) values between ~3.1-4.25 (Figure 2.16B). A clear, negative correlation between γ(α) and γ(β) is observed for class I estimates i.e. higher γ(α) values are paired with lower γ(β) values. In order to visualise CO interference in a standardised form that takes account of this γ(α,β) relationship (see: Section 2.5), interference functions—based on estimated class I γ(α,β) values and which are utilised by *RecombineSim* for mixed simulations—were plotted (Figure 2.16C). For comparison, corresponding interference functions derived from single fit γ(α,β) values are shown (Figure 2.16D). Class I windows from all genotypes adopt a highly similar form ±100kb a CO event, with moderate levels of variation in strength observed at >±100kb. In contrast and consistent with the inability of single (γ) solutions to model experimental data, single fit (γ) interference functions vary widely. For example, the single (γ) fit window within *ndt80ARmec1MN* terminates at ~±200kb, with appreciable efficacy (>20% strength) reaching ±100kb, while the corresponding mixed class I function extends beyond ±500kb, with considerable strength extending to ~±300kb.

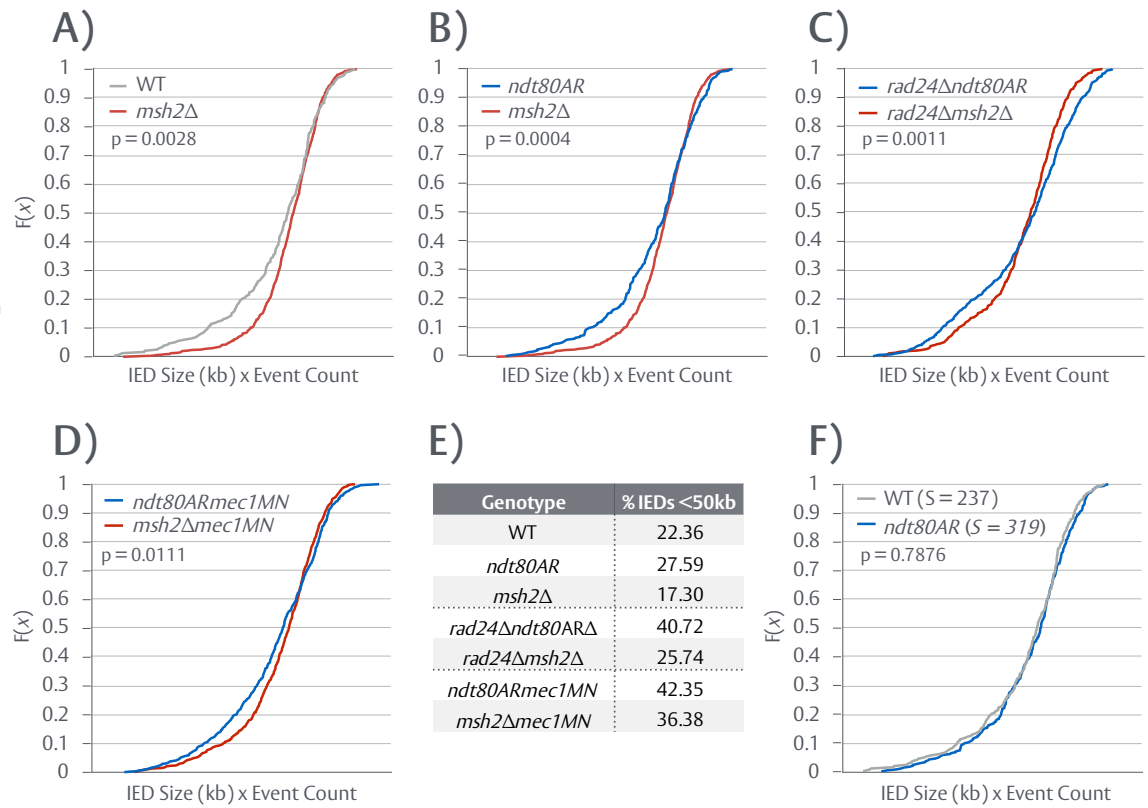
Collectively these results suggest that inter genotypic differences in CO distribution, within the mixture modelled mutant backgrounds, conceivably arise through through means other than remodelling of the class I signal and that mathematical isolation of class II COs has revealed a putative description of a universal class I interference function. Previous estimates for the range of class I CO interference within *S. cerevisiae* range from ~±100-200kb (Berchowitz & Copenhaver 2010)—in line with single (γ) fit estimates.

However, (γ) mixture modelling suggests WT meiosis, within *S. cerevisiae*, may instead employ an extensively broad 0.8-1.0Mbp zone of CO interference, with appreciable efficacy (>20% strength) extending bidirectionally outward ~±300-400kb surrounding each class I event. Moreover, WT is predicted to generate class II COs with an average frequency of 30-32%—residing at the upper boundary of previous estimates (de los Santos et al. 2003).

2.16—A novel role for the mismatch repair factor, Msh2, within the spatial regulation of COs

Resolution of class I interfering, ZMM-dependent COs relies upon the mismatch repair (MMR) family members, Msh4-Msh5 (MutSγ) and Mlh1-Mlh3 (MutLγ). Several studies have also implicated the related MMR factor, Msh2—utilised in this study to bolster detection of NCOs—in the hyper localised regulation of CO formation through modulation of recombination fidelity. Specifically, Msh2 coordinates heteroduplex rejection and the dismantling of recombination intermediates formed between imperfectly matched DNA sequences (i.e. homeologous recombination), leading to increased recombination rates within *msh2*Δ or otherwise MMR deficient strains (Borts & Haber 1987; Datta et a. 1997; Spies & Fishel 2015; Martini et al. 2011). However, as inferred by statistical testing (see: Figure 2.6B), MLE (γ) fitting (see: Figure 2.7D) and (γ) mixture modelling (see: Figure 2.14A) and as further presented here, *msh2*Δ displays a significantly distinct distribution of COs relative to WT—suggesting Msh2 may exert influence on CO formation beyond that of a hyper localised role.

In order to further characterise the Msh2-dependent impact on CO distribution, aggregated CO IED data from all *msh2*Δ strains was transformed to normalise for event number, as previously described (see: Section 2.4), and overlaid on semi log eCDF plots with transformed *MSH2*+ cognates (Figure 2.17A-D). Regardless of genotype, all *msh2*Δ strains display significantly different CO distributions to their *MSH2*+ equivalents when assessed by two sample KS test and are characterised by a depletion of smaller IEDs (<50kb) and consequently, a shift toward mid and higher range IED sizes (Figure 2.17E) i.e. a strengthening of CO interference.



Genotype	% IEDs <50kb
WT	22.36
ndt80AR	27.59
msh2Δ	17.30
rad24Δndt80ARΔ	40.72
rad24Δmsh2Δ	25.74
ndt80ARmec1MN	42.35
msh2Δmec1MN	36.38

Figure 2.17. Inactivation of Msh2 strengthens the global CO interference landscape

A-D) Aggregated CO IED data from *msh2Δ* and corresponding *MSH2*$^{+}$ strains was transformed to take into account differential event count and cognate pairs were plotted as cumulative distribution functions (CDFs). **E)** The fraction of IEDs residing below 50kb was calculated for all genotypes, except *msh2Δtel1Δ*—which has no *MSH2*$^{+}$ equivalent. **F)** Aggregated CO IED data from WT and *ndt80AR* was transformed and plotted as CDFs. Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size.

For example, 22.36% of the IEDs fall below 50kb within WT, as opposed to 17.30% within *msh2*Δ. Moreover and consistent with previous studies (Martini et al. 2011), *msh2*Δ exhibits an average ~1.4-fold increase in CO formation relative to WT (74.50 vs. 104.78 COs/cell) (see: Table 2.1). In order to account for any impact extended prophase I and the associated change in CO count (*ndt80AR*) may have on CO distribution, aggregated CO IED from *ndt80AR* was similarly transformed and compared to WT (Figure 2.17F). The distribution of COs within *ndt80AR* cells (95.75/cell) shows significant similarity ($p = 0.7876$) to WT (74.50/cell), suggesting no alteration in spatial regulation occurs as a result of extending prophase I alone.

Relative to WT, the inactivation of Msh2 therefore appears to give rise to two distinct phenotypes: (i) an increase in CO formation, as previously observed (Borts & Haber 1987; Martini et al. 2011 and (ii) a novel and global shift toward increased CO interference. Moreover, the latter phenomenon is retained in the presence of *rad24*Δ and *mec1MN* mutation, suggesting it operates independently of these DDR branches (see: Figure 2.17C,D).

2.17—Inactivation of Msh2 alters the class I:class II CO ratio

To further explore how the *msh2*Δ-dependent phenotypes may arise, previously obtained (γ) mixture modelling results were considered. Best fit models are obtained when using the previously identified universal class I function as a basis for CO interference and class II frequencies of 15.4%, 32.9% and 34.0% for *msh2*Δ, WT and *ndt80AR* respectively (see: Figure 2.14A, 2.15A/B/E)—suggesting the observed *msh2*Δ-dependent increase in global CO interference strength is a result of lower class II formation. To further strengthen the idea that class II frequency is decreased within *msh2*Δ relative to WT, *msh2*Δ aggregated CO IED data was solved (via *GEM*) for averaged WT levels of class II formation (32.1%) under the assumption that no alteration to the class I:class II ratio has actually occurred but rather a strengthening of the interference signal itself. In other words, if class II frequency (w2) is held static at 32.1%, what is the best possible combination of $\gamma(\alpha,\beta)$ that can be obtained to model *msh2*Δ IED data. As expected, CO interference is predicted to increase in

strength under these conditions ($\alpha = 4.29$ / $\beta = 32801$ vs. $\alpha = 3.84$ / $\beta = 29877$) in an attempt to offset increased class II formation, however, model-experimental fit is extremely poor ($p = 0.0155$) (Figure 2.18A). It is also possible that Msh2 skews CO position in accordance to chromosome wide patterns of SNP/INDEL density—constituting a secondary layer of spatial regulation independent of CO interference—in a manner that is misinterpreted by *GEM* as shifts in the class I:class II ratio. In order to explore this alternative possibility, random CO simulations (*RecombineSim* mode: *Random*) were repeated for *rad24*Δ*msh2*Δ event counts with site selection weighted according to smoothed S288c x SK1 polymorphism density. SNP/INDEL densities were smoothed using three different moving average window sizes: 1kb (Figure 2.18B), 5kb (Figure 2.19C) and 25kb (Figure 2.19D) to investigate the impact local or gross variation in density may have. As assessed by two sample KS test, weighting the position of CO formation by SNP/INDEL density is insufficient to produce significant shifts relative to unweighted, random simulations.

Collectively, these observations suggest that class II frequency is legitimately reduced within *msh2*Δ relative to WT, favouring a mechanism whereby Msh2 ordinarily acts as either (i) a pro class II factor or (ii) an anti class I factor. To distinguish between these possibilities, class I and class II frequency estimates, obtained from (γ) mixture modelling (via *GEM*), were converted to predicted event counts (Figure 2.18E,F). Remarkably, relative to WT, *msh2*Δ is predicted to form an additional ~38 class I COs (49.9/cell vs. 88.6/cell) (~1.77-fold increase) and ~8 less class II COs (24.5/cell vs. 16.1/cell) (~1.52-fold decrease). Elevated class I CO formation is thus predicted to account for the majority of the increased event count observed in *msh2*Δ compared to WT. A similar trend exists between *ndt80ARmec1MN* and *msh2*Δ*mec1MN*—the latter is predicted to form ~18 more class I COs over the former (68.4/cell vs. 86.17/cell) (~1.25-fold increase), with no significant difference between predicted class II formation observed. Marginal reductions in class II formation may reflect the derepression of class I formation, redirecting DSB events otherwise destined for the class II formation into the interfering pathway. Prophase I extension alone proportionately increases both predicted class I and class II formation (~1.3-fold), consistent with no change in spatial regulation.

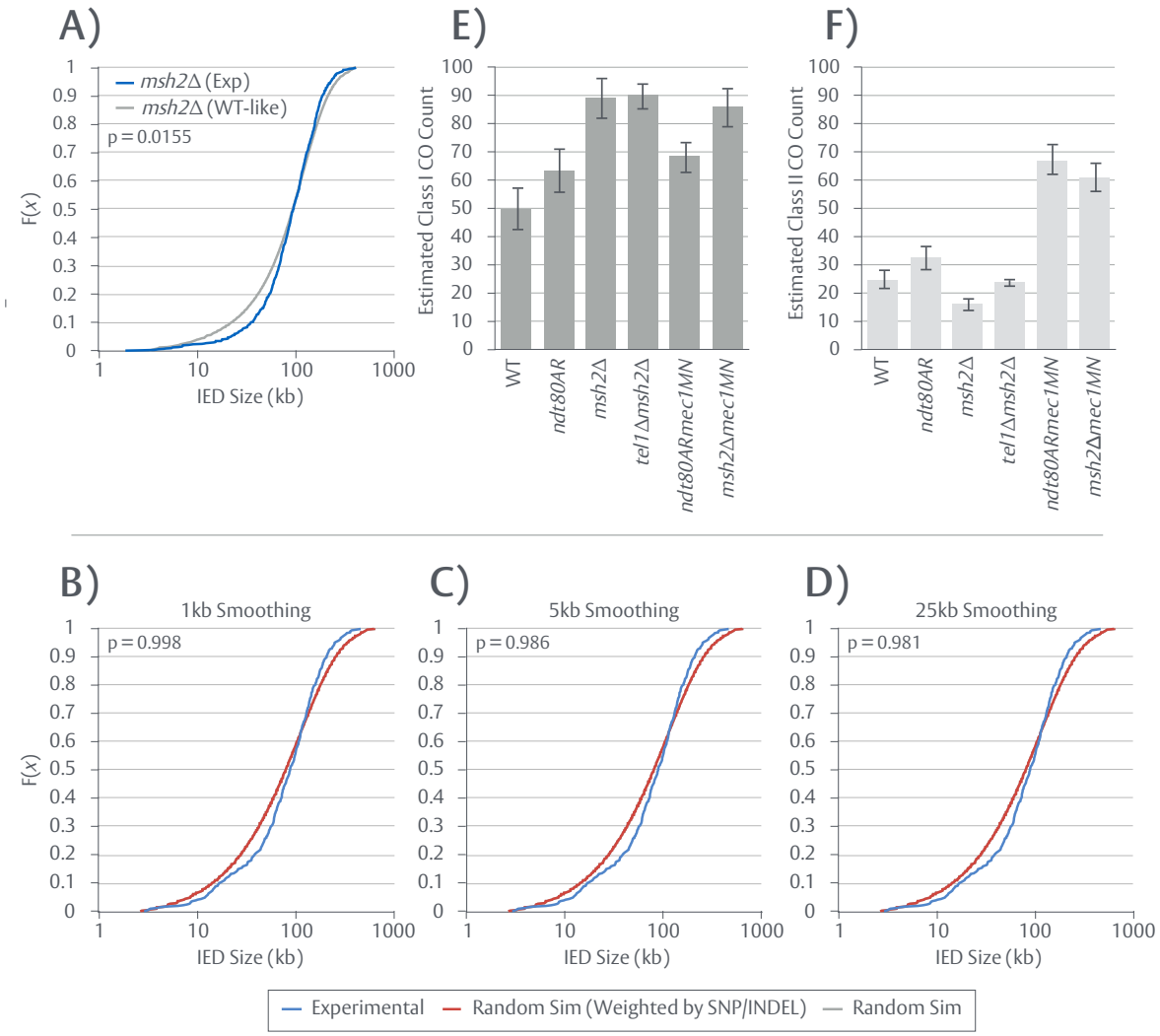


Figure 2.18. Inactivation of Msh2 alters the class I:class II CO balance

A) A mixed, interfering CO simulation (*RecombineSim* mode: UniHazard) was performed for *msh2*Δ under averaged WT conditions (universal class I window, 32.1% class II CO frequency) and plotted with aggregated, experimental CO IED *msh2*Δ data as cumulative distribution functions (CDFs). Model-experimental fit was assessed via two sample KS test (MATLAB 2017a Package: *kstest2*) (see: p value). **B-D)** Maps of S288c x SK1 SNP/INDEL density were constructed by populating empty, zeroed numerical arrays, proportional in length to *in vivo* chromosome size, with values of [1.0] at positions equivalent to each SNP/INDEL. Resulting maps were smoothed (moving average, window width as marked). Weighted, non-interfering CO simulations (*RecombineSim* mode: Random), whereby smoothed SNP/INDEL density is used in place of recom(P), were conducted using *rad24*Δ*msh2*Δ CO event counts. Resulting simulated IEDs (red) were visualised against the corresponding aggregated, experimental CO IED data and the results of a purely random simulation (see: Figure 2.11). (P) values are for weighted-random comparisons. Model-experimental fit was assessed via two sample KS test (MATLAB 2017a Package: *kstest2*) (see: p values). F(x) = Fraction of IED data. **E-F)** Class I and class II CO frequencies, obtained via *GEM*, were converted to predicted counts using averaged event counts (see: Figure 1.4). Error bars are calculated using estimated N(%Δ) values. See Section 2.20 for details on *tel1*Δ*msh2*Δ data.

Taken together, these results suggest that Msh2 specifically targets class I COs for suppression, and that the increases in CO formation and CO interference strength, as observed within *msh2*Δ, are linked phenomenon—arising through derepressed class I formation and a subsequent shift in the class I:class II ratio.

2.18—Msh2 specifically inhibits class I CO formation at sites of higher sequence divergence

Data presented thus far suggests Msh2 acts as an anti class I CO factor, whose deletion alters the global distribution of COs through increased formation of interfering, class I COs (see: Section 2.16, 2.17). Given the role of Msh2 within recombination fidelity—whereby CO formation is inhibited, in general, via heteroduplex rejection at sites of sequence divergence—specific repression of class I COs may occur through a similar mechanism. In order to explore this hypothesis and potentially elucidate any mechanistic details underpinning Msh2 anti class I activity, the density of polymorphism (SNP/INDEL) surrounding 3314 *msh2*Δ and 2086 *MSH2*$^{+}$ CO midpoints, from all genotypes, were analysed with progressively increasing window size. To account for the inclusion of INDELs and the influence large stretches of divergence may exert, polymorphisms were weighted via two methods: (i) SNPs and INDELs are considered equal (Figure 2.19A) (ii) INDELs are weighted according to the number of bases involved (Figure 2.19B). Regardless of the INDEL weighting method applied, *MSH2*$^{+}$ COs are reproducibly skewed toward regions of lower SNP/INDEL density relative to *msh2*Δ COs, with average *MSH2*$^{+}$/*msh2*Δ density ratios of ~0.65 at close range (<±100bp) (Figure 2.19C). Averaged *MSH2*$^{+}$/*msh2*Δ density ratios remain below 1.0 until ~±4.5kb but rapidly return to >0.9 at ~±1kb—suggesting this skew is primarily confined to a hyperlocal range. All *MSH2*$^{+}$ strains exhibit a skew of similar intensity except *ndt80ARmec1MN*, which displays a weakened, intermediate phenotype (see: Figure 2.19A,B).

Ascertaining whether or not such a skew is linked to class I CO suppression is complicated by an inability to identify class I and class II COs within the experimental data, excluding the possibility of individual density analysis for each subclass.

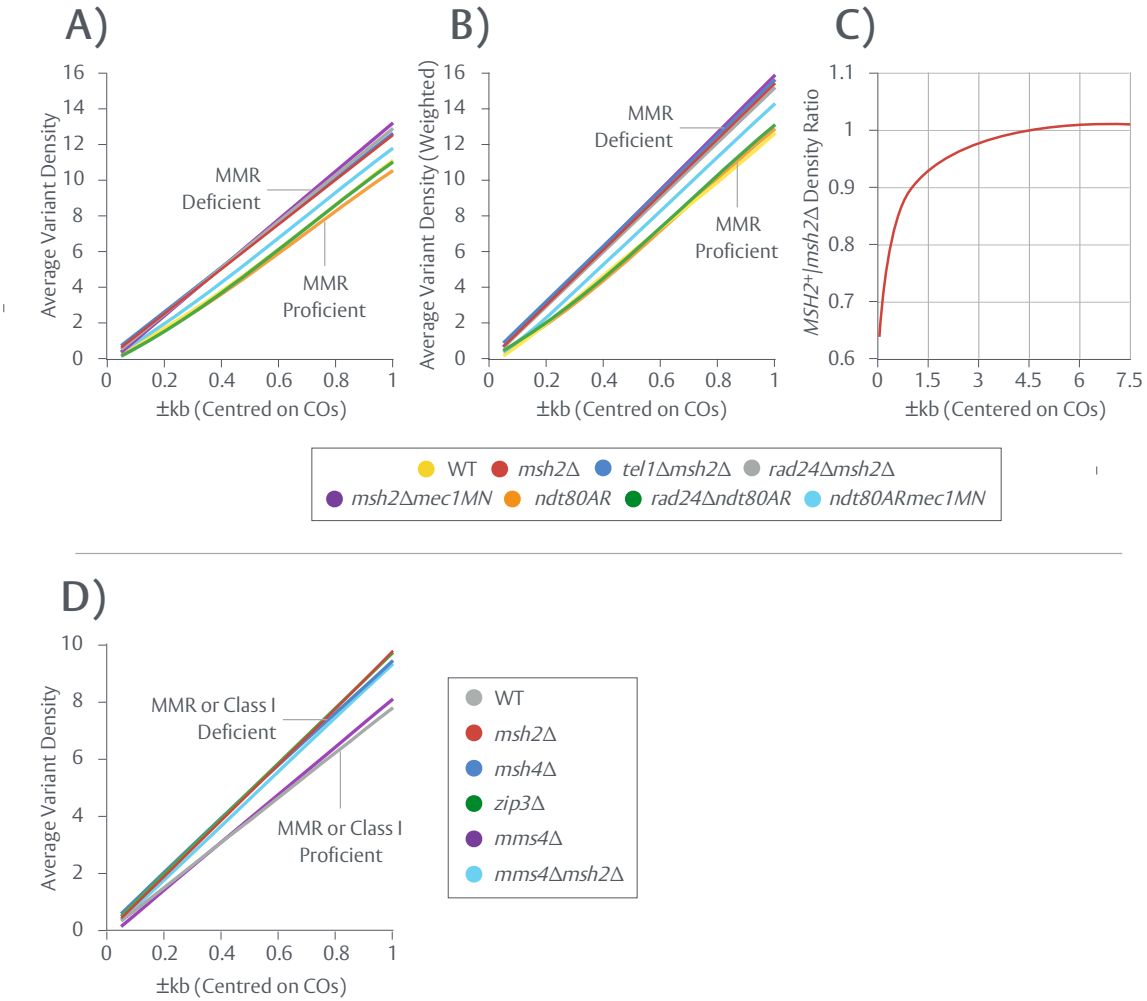


Figure 2.19. Msh2 skews class I CO formation toward regions of lower sequence divergence

A) Non-smoothed SNP/INDEL (S288c x SK1) density maps were generated by populating empty, zeroed numerical arrays, proportional in length to *in vivo* chromosome size, with values of [1.0] at positions equivalent to each SNP/INDEL. The number of SNP/INDELs surrounding each mapped CO, in all genotypes, was subsequently calculated for different window sizes (±bp) and averaged **B)** Polymorphism analysis was repeated in an identical manner using *weighted* SNP/INDEL maps, where each array value added is proportional to the number of bases involved in the polymorphism (i.e. SNP = [1.0], 3bp INDEL = [3.0]). **C)** Unweighted polymorphism density values were calculated up to ±7.5kb, averaged together for all *msh2*Δ genotypes and all *MSH2*$^{+}$ genotypes bar *ndt80ARmec1MN*, and subsequently plotted as a *MSH2*$^{+}$/*msh2*Δ ratio **D)** Non-smoothed SNP (S288c x YJM879) density maps were generated and utilised to calculate polymorphism density for independent datasets (Oke et al. 2014), in an identical manner.

As such, several mechanisms could reconcile these observations: (i) as is currently proposed, Msh2, and by extension sequence divergence, may indiscriminately suppress CO formation regardless of class (Spies & Fishel 2015), disproportionately repressing class I formation or increasing global CO interference strength through alternative means (ii) class I COs may be insensitive to polymorphism density relative to class II COs. Increased formation of class I COs within *msh2*Δ would therefore shift the system toward regions of higher polymorphism density, decreasing the influence of polymorphism density sensitive class II events or (iii) Msh2 specifically represses class I formation at regions of sequence divergence, with negligible impact on class II COs.

In order to differentiate these mechanisms, WT, *msh2*Δ, *zip3*Δ, *msh4*Δ, *mms4*Δ and *mms4*Δ*msh2*Δ datasets from an independent source (Oke et al. 2014) were subject to similar polymorphism density analyses using an appropriate S288c x YJM789 SNP list (Figure 2.19D). Notably, the disparity in density between WT and *msh2*Δ is recaptured, suggesting the effect is not limited to S288c x SK1 *S. cerevisiae* crosses. In addition, and crucially, removal of the class II pathway via *mms4*Δ does not appreciably alter the observed skew while inactivation of class I formation, via *msh4*Δ or *zip3*Δ, phenocopies *msh2*Δ. Furthermore, inactivation of class II formation within a *msh2*Δ background (*mms4*Δ*msh2*Δ) has no appreciable effect.

If Msh2 indiscriminately suppressed COs independently of class, *msh4*Δ or *zip3*Δ, within which Msh2 remains active, would not be expected to resemble the density skew of *msh2*Δ. Moreover, if class I COs were less sensitive than class II COs to polymorphism density, *mms4*Δ and *mms4*Δ*msh2*Δ should enhance the skew, rather than result in no change. These results are, however, consistent with the specific repression of class I COs at sites of higher sequence divergence by Msh2—revealing a potential and novel meiotic role for this DNA repair factor.

2.19—Loss of Mec1 function deregulates event number and alters class I:class II CO ratios

Mec1^{ATR}, a DDR kinase, is activated in response to the ssDNA formed during HR-dependent resection by a Rad24-dependent clamp system and mediates the pachytene I checkpoint as well as negative regulation of DSB formation through *trans* DSB interference (see: Section 1.4.2) (Lydall et al. 1996; Majka et al. 2006; Majka & Burgers 2003; MacQueen & Hochwagen 2011; Zhang et al. 2011). While Mec1 regulates DSB formation, how it may influence the distribution of COs remains uncharacterised.

Strains lacking Mec1 activity (*mec1MN*) display marked increases in CO and NCO formation. Relative to *msh2*Δ, *msh2*Δ*mec1MN* exhibits a ~1.65-fold increase in total event count relative to *msh2*Δ (328.00/cell vs. 197.78/cell), characterised by a disproportionate increase in NCO formation (see: Table 2.1). Specifically, *msh2*Δ displays an average CO:NCO ratio of 1.12 while this is reduced to 0.81 within *msh2*Δ*mec1MN*. An increase in both NCO and CO formation implies an excessive increase in precursor DSB formation. A similar 44.9%, albeit lower, increase in event formation is seen within *ndt80ARmec1MN* relative to *ndt80AR* (205.80/cell vs. 142.00/cell). CO:NCO ratios are unreliable within *MSH2*$^{+}$ backgrounds and are thus not assessed. Inter sister events remain invisible within recombination mapping methods dependent upon SNPs/INDELs (see: Section 2.2). Given the dependency on inter homologue (IH) interactions, it is possible that a strengthening of the IH bias may explain apparent increases in event formation. Mec1 would thus be implicated as an anti-IH and/or pro IS-factor, however, such a mechanism would be in direct contradiction to the literature (Carballo et al. 2008). Similarly, Mec1 has been established as an indirect and generalised promoter of DSB formation under conditions of suboptimal Spo11 catalysis (Gray et al. 2013; Argunhan et al. 2013), rather than a repressor. Excess formation of DSBs within *mec1MN* is thus likely due to a loss of Mec1-dependent DSB interference in *trans* (Cooper et al. 2014; Zhang et al. 2011).

Disproportionate shunting of excess DSBs into the NCO pathway suggests CO homeostasis is still operational within *mec1MN* backgrounds—buffering against excess DSB formation. However, prior

(γ) mixture modelling results obtained via *GEM* (see: Figure 2.14A) suggest excess DSBs may also be driven into the class II CO pathway. While optimal model fits are still obtained using the identified universal class I window, *ndt80ARmec1MN* and *msh2*Δ*mec1MN* are predicted to have increased class II formation relative to *ndt80AR* and *msh2*Δ—49.4% (66.7/cell) and 41.4% (60.8/cell) respectively (see: Figure 2.18F). In contrast, class I formation is not predicted to significantly change upon loss of Mec1 activity in *msh2*Δ*mec1MN* (86.2/cell) relative to *msh2*Δ (88.6/cell) or in *ndt80ARmec1MN* (68.4/cell) relative to ndt*80AR* (63.2/cell) (see: Figure 2.18E).

In order to analyse whether or not global CO interference strength is reduced in *mec1MN*—as predicted by increased class II CO formation—aggregated CO IED datasets from *ndt80ARmec1MN* and *msh2*Δ*mec1MN* were compared to simulated, random distributions (*RecombineSim* mode: Random) (Figure 2.20A,B). As assessed by two sample KS test, neither strain displays significant similarity to randomness. However, overlays of transformed CO IED data from both strains and the respective base strain (*msh2*Δ and *ndt80AR*), reveals a *mec1MN*-dependent shift toward smaller IED sizes in both strains, consistent with a weakening of global CO interference strength (Figure 2.20C,D). Although a role for Mec1 in the direct suppression of class II COs cannot be ruled out, these results collectively favour a model whereby *mec1MN*-dependent weakening of the CO interference landscape arises through increased class II CO formation as an indirect consequence of increased DSB formation.

2.20—Inactivation of Tel1 increases class II CO formation

Recent work suggests that inactivation of the DDR kinase, Tel1^{ATM}, may shift the class I:class II ratio toward class II CO formation, thus weakening the global CO landscape (Anderson et al. 2015). Unexpectedly, aggregated CO IED data from *tel1*Δ*msh2*Δ is not amenable to (γ) mixture modelling —resolving into two, non-random components (γ(α) 4.45 and 1.97) (see: Figure 2.14A)— suggesting the *tel1*Δ phenotype may be more complex than anticipated.

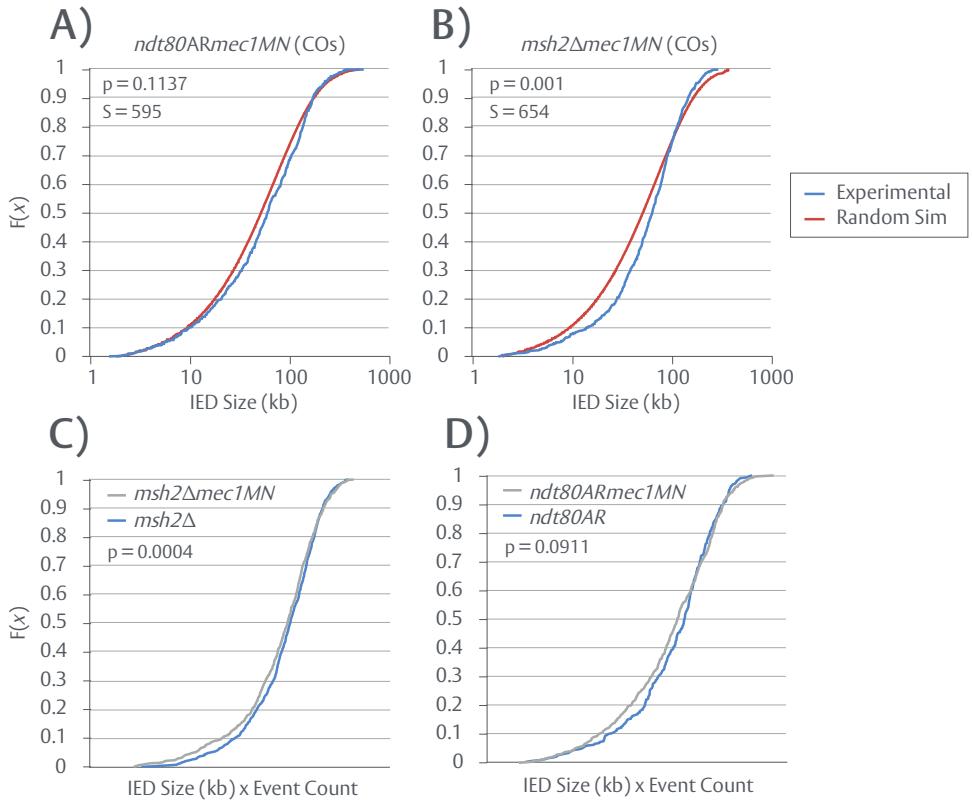
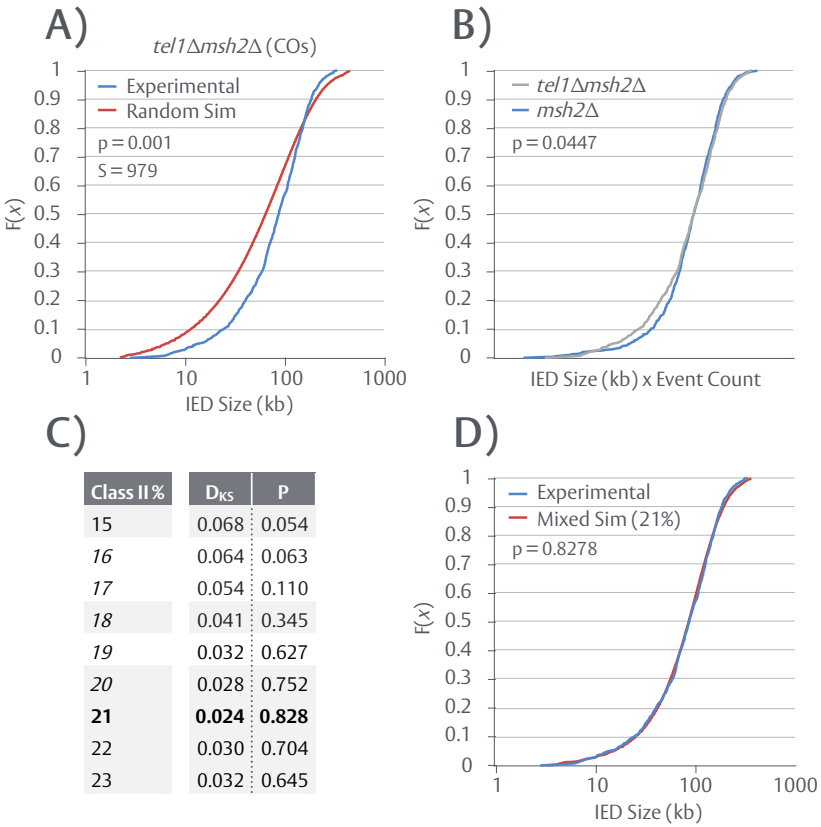


Figure 2.20. Loss of Mec1 function weakens the global CO interference landscape

Random CO simulations (*RecombineSim* mode: Random) were performed for **A)** *ndt80*AR*mec1MN* and **B)** *msh2Δmec1MN*. Resulting simulated IEDs (red) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution functions (CDFs). **C)** Aggregated CO IED data from *msh2Δ* and *msh2Δmec1MN* was transformed and plotted as CDFs. **D)** Aggregated CO IED data from *ndt80AR* and *ndt80ARmec1MN* was transformed and plotted as CDFs. Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size.

In order to investigate the distribution of COs within Tel1 deficient strains, aggregated CO IED data from *tel1*Δ*msh2*Δ was compared to a simulated, random distribution (*RecombineSim* mode: Random) (Figure 2.21A). As assessed by two sample KS test, *tel1*Δ*msh2*Δ COs do not display significant similarity to randomness—confirming the presence of CO interference. However, an overlay of transformed CO IED data from *tel1*Δ*msh2*Δ and *msh2*Δ reveals a specific, *tel1*Δ-dependent enrichment in short (<50kb) IEDs (Figure 2.21B)—indicative of increased class II formation rather than a generalised weakening of CO interference. Given the inability of *GEM* to provide a reliable measure of class II CO frequency for *tel1*Δ*msh2*Δ and in order to investigate this short range loss of CO interference strength, a bespoke model fit (*RecombineSim* mode: UniHazard) was created by simulating progressively increasing class II CO frequencies under conditions otherwise identical to *msh2*Δ (i.e. universal class I window) (Figure 2.21C). An optimal and significant model fit is obtained using a moderately increased class II frequency of 21% (Figure 2.21D)—as opposed to 15.4%, within *msh2*Δ—suggesting that, as previously proposed (Anderson et al. 2015), the *tel1*Δ phenotype may be readily and exclusively explained by shifts in the class I:class II ratio.

It therefore remains unclear why *GEM* failed to produce applicable results. Convergence of MLE $\gamma(\alpha,\beta)$ parameter estimates toward a local, rather than global, maxima may possibly skew (γ) mixture modelling results. In order to assess whether or not *GEM* can resolve *tel1*Δ*msh2*Δ data when one set of parameters is fixed, thus potentially avoiding any local maximas, *tel1*Δ*msh2*Δ CO IED data was solved under one of two conditions: (i) a fixed non-random component (*msh2*Δ class I window—$\gamma(\alpha) = 3.84$, $\beta = 29877$) and (ii) a fixed, random component (*msh2*Δ class II window—$\gamma(\alpha) = 1.21$, $\beta = 68155$). Consistent with the results of bespoke modelling, when the class I window is fixed, *GEM* successfully identifies a random component ($\gamma(\alpha) = 1.18$, $\beta = 70204$) and a class II CO frequency of 21.4%.



Class II %	D_{KS}	P
15	0.068	0.054
16	0.064	0.063
17	0.054	0.110
18	0.041	0.345
19	0.032	0.627
20	0.028	0.752
21	**0.024**	**0.828**
22	0.030	0.704
23	0.032	0.645

Figure 2.21. Inactivation of Tel1 may increase class II CO frequency

A) A random CO simulation (*RecombineSim* mode: Random) was performed for *tel1Δmsh2Δ*. Resulting simulated IEDs (red) were visualised against the corresponding, aggregated experimental IED data (blue) as semi log (x) cumulative distribution function (CDF). **B)** Aggregated CO IED data from *msh2Δ* and *tel1Δmsh2Δ* was transformed and plotted as CDFs. **C)** Mixed, interfering CO simulations (via *RecombineSim*) were performed for *tel1Δmsh2Δ* using the universal class I window and varying frequencies of class II COs. **D)** A best fit model for *tel1Δmsh2Δ* is obtained using a class II CO frequency of 21%. Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) values). F(x) = Fraction of IED data. S = aggregated IED sample size.

Likewise, when a random component is forced, *GEM* successfully identifies a class I window in line with those previously identified ($\gamma(\alpha) = 3.71$, $\beta = 32456$) and a class II CO frequency of 21.2%. Biased initiation of $\gamma(\alpha,\beta)$ parameters may therefore be an effective solution when unbiased initiation fails.

Relative to *msh2*Δ, *tel1*Δ*msh2*Δ also exhibits a modest ~1.19-fold increase in event formation (236.40/cell vs. 197.78/cell), characterised by disproportionate increases in NCO formation over CO formation. Specifically, *msh2*Δ displays an average CO:NCO ratio of 1.12, while this is reduced to 0.93 within *tel1*Δ*msh2*Δ (see: Table 2.1). Thus, akin to *mec1MN* strains (see: Section 2.19), increased class II and NCO formation may result from elevated DSB formation due to the loss of Tel1-dependent DSB interference *in cis* (see: Section 1.4.2). Consistent with the idea that excess DSBs preferentially enter the class II pathway, an estimated class II frequency of 21% for *tel1*Δ*msh2*Δ predicts the formation of ~90 class I COs on average per cell—in line with that of *msh2*Δ (~88/cell) (see: Figure 2.18E,F).

2.21—Discussion

Work presented here reveals novel consequences for the removal of DDR components (Rad24, Mec1, Tel1) on the spatial distribution of meiotic COs and details an unexpected influence of Msh2 on interfering, class I COs. Moreover, established modelling standards have been re-examined and improved upon via the introduction of (γ) mixture models, which more accurately recapture *in vivo* CO distributions.

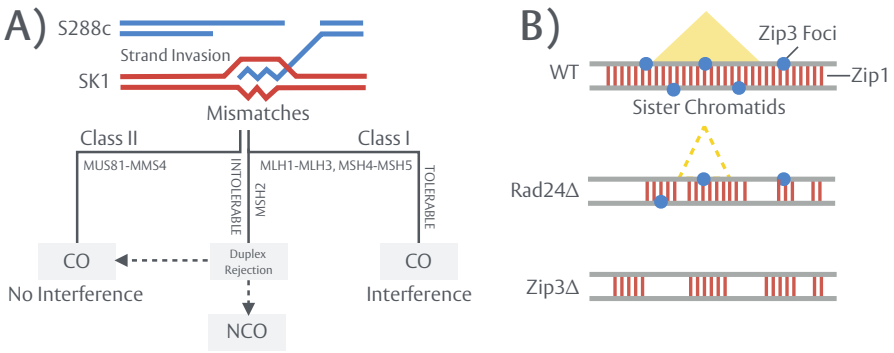
Modelling Meiotic Recombination

Historically, CO/NCO distributions have been characterised via the coefficient of coincidence (CoC), γ(α) values or deciphered through single component simulation. CoC requires extensive binning of data—obscuring short range information which is often crucial—while, as demonstrated throughout this chapter, single γ(α) fitting does not reveal a complete picture of the system and is often an insufficient description of *in vivo* CO distributions (see: Sections 2.4, 2.12). For example, single fit γ(α) values for *rad24*Δ*msh2*Δ and *msh2*Δ*mec1MN* are highly similar—1.60 and 1.65 respectively, which may lead to misinterpretation of the data and initially suggest that these two strains exhibit identical phenotypes, occurring through a common mechanism. By taking the existence of non-interfering, class II COs into account and developing a two component (γ) system for analysing and simulating CO distributions, as well as directly estimating class II frequency from the experimental data itself, a novel and effective solution to these issues has been created. A two component system readily describes most CO distributions tested, with a high degree of statistical accuracy (see: Figure 2.14A). Moreover, (γ) mixture modelling has helped to uncover a number of potential, novel meiotic roles for DDR and DNA repair components (discussed below) as well as several testable hypotheses (see: Chapter 5). It should, however, be noted that—owing to a lack of data—CO/NCO distributions have not been modelled on a per chromosome basis. Results obtained, including the putative description of WT CO interference (see: Figure 2.16C), are therefore generalised, global descriptions that may not necessarily reflect the *in vivo* situation on each chromosome.

Mismatch Repair (*Msh2*)

As detailed in sections 2.16-2.18, an unexpected and novel role for the MMR factor, Msh2, within the regulation of class I CO formation has been uncovered. Notably, results presented indicate that Msh2 acts as a specific inhibitor of class I CO formation, disproportionately forming a barrier to recombination at Mlh1-Mlh3 and ZMM-dependent class I sites over Mus81-Mms4-dependent class II sites when higher sequence divergence exists (Figure 2.22A) (see: Figure 2.19A-D). Msh2 could therefore be considered a factor involved in CO designation (see: Section 1.2.8). Such a mechanism highlights how SNP/INDEL density can govern meiotic outcome beyond a localised, indiscriminate suppression of crossing over—as was previously observed (Borts & Haber 1987; Datta et a. 1997; Spies & Fishel 2015). A model whereby Msh2 mediates the specific repression of class I COs is able to explain both the increased, global strength of CO interference observed within *msh2*Δ backgrounds (see: Figure 2.17A-E) and the elevated formation of COs (see: Table 2.1). In the presence of Msh2, mismatches—presumably at the level of strand invasion—may be detected by MMR machinery. Given the density of S288c x SK1 SNPs/INDELs (1/176bp), it is likely that most events will encounter sequence divergence of some level and therefore, class I CO formation may be permitted to occur below a certain tolerance threshold. Above this threshold, Msh2 may mediate heteroduplex rejection, redirecting the DSB toward the NCO or class II CO pathway. Within *msh2*Δ backgrounds, mismatch recognition is either weakened or fully abolished, allowing class I COs to form with higher frequencies (see: Figure 2.22A). Interestingly and consistent with the specific targeting of class I CO by Msh2, MMR does not appear to function at non-interfering, class II sites (Getz et al. 2008)—suggesting that when class I COs do successfully form within regions of lower sequence divergence, they undergo mismatch correction.

Meiotic recombination occurring with hybrid, divergent strains may closely resemble truly wild meioses. Msh2-dependent inhibition of class I formation may thus have an important role in ensuring meiotic success and maintenance of genomic integrity (see: Chapter 5). Moreover, Msh2 status will be an important consideration in future mapping studies going forward.



C)

Genotype	Class I	Interference	Mismatch Sensitivity
WT	✓	Full	✓
*msh2*Δ	✓	Increased	x
*msh2*Δ*rad24*Δ	✓	Reduced	x
*rad24*Δ*ndt80AR*	✓	None	✓
*zip3*Δ	x	None	x
*msh4*Δ	x	None	x

Figure 2.22. Models for Msh2 and Rad24 activity during CO formation

A) Msh2-dependent suppression of class I COs at regions of higher sequence divergence. Within *MSH2*$^{+}$ strains, upon strand invasion of mismatched, homeologous sequence, the level of sequence divergence is sensed by Msh2. Below a certain tolerance level, class I CO formation is successful. However, if an intolerant level of divergence exists, Msh2 mediates heteroduplex rejection and the DSB is subsequently redirected toward the non-interfering class II CO and NCO pathways. Within *msh2*Δ strains, this mechanism is dampened or abolished, resulting in increased class I CO formation. Class II COs are less sensitive or insensitive to sequence divergence and form at similar frequencies regardless of Msh2 status. **B)** Rad24-dependent generation of WT CO interference. Under WT conditions, Zip3 foci—marking the sites of class I CO formation—are dispersed along a proteinaceous Zip1-containing axis and CO interference operates as normal. Within *rad24*Δ backgrounds, Zip3 foci formation is reduced (~55%) and only short stretches of Zip1 filament form, reducing the efficacy and/or generation of CO interference. Within *zip3*Δ strains, class I CO formation and CO interference are fully removed (Berchowitz & Copenhaver 2010). **C)** Features of CO interference across analysed strains.

Nevertheless, how Msh2 specificity for class I COs arises remains unclear. Mus81-Mms4 appears to promote class II CO formation independently of Holliday junction intermediates—a substrate toward which Msh2 exhibits preferential binding activity (Alani et al. 1997; Marsischky et al. 1999). Moreover, *in vitro* data suggests that Mlh1-Mlh3 facilitates the binding of Msh2 to INDELs and that Mlh1-Mlh3 endonuclease activity is stimulated by Msh2 (Rogacheva et al. 2014). The distinct genetic requirements and mechanics of each CO class may therefore mediate their differential sensitivity to sequence divergence through class I specific recruitment of Msh2.

Mec1^{ATR}

Meiotic null strains of Mec1 (*mec1MN*) exhibit reduced global strength of CO interference and elevated event formation (see: Section 2.19). Apparent reductions in the strength of CO interference appear to occur through increased formation of class II COs, as a result of excessive DSB formation potentially caused by a loss of *trans* DSB interference. In contrast, predicted class I CO frequency remains largely unchanged within *mec1MN* strains. Consistent with a higher reliance upon class II mechanics, Msh2-dependent skew of class I COs toward regions of low sequence divergence is partially weakened within *ndt80ARmec1MN*, which is predicted to form class II COs with a higher frequency of ~50% (see: Figure 2.19A,B).

In order to facilitate a model whereby excess DSBs are preferentially shunted into the class II CO pathway, a mechanism must exist to make the distinction between subclasses. CO homeostasis (see: Section 1.3.5)—a process potentially imposed by CO interference—may primarily or solely act on class I COs, quantitatively restricting their formation while no upper limit is imposed upon class II formation. Alternatively, given the temporal distinction between class I and class II resolution within *S. cerevisiae*—defined by the late activity of Mus81-Mms4 (see: Section 1.5.1)—class I CO formation or the generation of the interfering signal may be confined to a specific window of time during prophase I and thus only a given number of events can form on average (e.g. 85-90 as observed within *msh2*Δ and *msh2*Δ*mec1MN*). Moreover, given the role Mec1 plays in modulating the length

of prophase I (see: Section 1.3.1) , it is conceivable that a lack of Mec1 signalling perturbs this phase, restricting class I formation and forcing remaining DSBs to enter NCO and class II pathways. While cells *must* repair all DSBs, it is perhaps surprising that excess DSB formation may translate into increased class II formation—a process that carries higher risk and which induces considerably more genetic reorganisation than if excess DSBs exclusively entered the NCO pathway.

Interestingly, the distribution of COs within *ndt80ARmec1MN* appears distinct from that of *rad24*Δ*ndt80AR*. Specifically, while weakened, CO interference is not fully lost within *ndt80ARmec1MN* as it is within *rad24*Δ*ndt80AR* (see: Figure 2.20A, 2.10A). The *mec1MN* allele constitutes a meiotic knockdown and residual Mec1 activity may remain, accounting for partial retention of CO interference. However, inactivation of Rad24 does not appreciably elevate NCO or CO count as loss of Mec1 activity does. Meiotic phenotypes of *rad24*Δ therefore appear distinct and partially independent from those of its downstream effector kinase, Mec1.

Tel1^{ATM}

Inactivation of Tel1 (*tel1*Δ*msh2*Δ) reduces the local strength of CO interference over a short range and elevates event formation (see: Section 2.20). Class I CO formation is, however, predicted to remain unchanged (see: Figure 2.18E). Tel1 is known to mediate the negative regulation of DSB formation through *in cis* suppression of DSBs in proximity to pre-existing breaks—a process known as Tel1-dependent DSB interference. Thus and akin to Mec1 deficient strains (see above), increases in precursor DSBs may specifically result in increased NCO and class II CO formation. Interestingly, increases in event formation are more pronounced in *ndt80ARmec1MN* (328.00/cell) than within *tel1*Δ*msh2*Δ (236.40/cell). If elevated DSB formation within *ndt80ARmec1MN* and *tel1*Δ*msh2*Δ is due to a loss of *trans* and *cis* DSB interference respectively, this observation suggests *trans* DSB interference is more important for quantitative or homeostatic control of DSB formation than *cis* DSB interference. Collectively, results from *mec1MN* and *tel1*Δ strains may highlight a novel consequence of excess DSB formation on the resulting distribution of meiotic recombination.

However, removal of DSB interference by the inactivation of Tel1 also results in concerted DSB formation over short ranges (±~7.5kb) at frequencies significantly greater than expected from independent behaviour. This phenomenon—termed negative interference—is only witnessed between DSB hotspots residing within the same chromosomal loop domain (Garcia et al. 2015) (see: Section 1.4.2). A hyper local clustering of DSBs within *tel1*Δ*msh2*Δ may "feed forward", skewing CO distributions and resulting in an enrichment of smaller IEDs. Such a mechanism may explain why (γ) mixture modelling fails to find a non-random class II component for *tel1*Δ*msh2*Δ data as well as the apparent increase in class II CO formation. This possibility is further explored in Chapter 4.

Rad24

CO interference is severely diminished within *rad24*Δ backgrounds—a loss likely reflective of the Zip3/ZMM loading, and therefore class I promoting, activities of Rad24 (Shinohara et al. 2015) (see: Figure 2.10A-C) . However, Msh2 selectivity toward class I COs has important ramifications for the interpretation of the *rad24*Δ phenotype. Importantly, ablation of CO interference within *rad24*Δ*ndt80AR* is not accompanied by a shift in polymorphism density bias (see: Figure 2.19A,B). Rather, *rad24*Δ*ndt80AR* behaves according to its Msh2 status—that is, COs within this background are skewed toward regions of lower sequence divergence akin to WT and *ndt80AR*. As Msh2 is thought to specifically act on class I COs, such an observation implies retention of class I genetic identity (resolution by Mlh1-Mlh3, Msh4-Msh5), suggesting that the near-random distribution of COs observed within *rad24*Δndt*80AR* arises through a loss of the interfering signal or its propagation, as opposed to elevated class II CO formation—which would alter the polymorphism bias akin to *ndt80ARmec1MN*. Moreover, the *msh2*Δ-dependent shift in the class I:class II ratio to favour class I formation is sufficient to reintroduce detectable CO interference into a *rad24*Δ background (*rad24*Δ*msh2*Δ) (see: Figure 2.10B). Thus, on some level, class I mechanics and CO interference still function within a *rad24*Δ background. While Zip3/ZMM loading is significantly reduced (~55%) by Rad24 inactivation, some Zip3 foci and short stretches of Zip1 still assemble

(Shinohara et al. 2015). Limited ZMM assembly may prove sufficient enough to support class I-like resolution and thus imposition of the Msh2 effect (Figure 2.22B) while reducing CO interference efficacy if propagation or interpretation of the signal requires a fully intact axis. In contrast, within *zip3*Δ or *msh4*Δ backgrounds, class I CO resolution and thus, CO interference, is fully lost—forcing all COs to form via the class II CO pathway. Consistent with these ideas, *GEM* resolves *rad24*Δ*msh2*Δ into a non-random class I component that highly resembles those of *msh2*Δ or WT (see: Figure 2.16C), but predicts the strain to have a class II frequency of 86.5% (see: Figure 2.14A). CO interference may thus adopt a WT form in areas where it is generated, however the lack of signal generation in general is misinterpreted by *GEM* as enrichment in randomly distributed class II COs. Collectively, such a model may explain how *zip3*Δ, *msh4*Δ and *rad24*Δ*ndt80AR* can simultaneously display distributional randomness (Berchowitz & Copenhaver 2010) but differential sensitivity to polymorphism density (Figure 2.22C).

2.21—Summary (Key Points)

- Developed a novel SNP/INDEL screening method (*HybridVar*) (Section 2.3)
- Constructed a novel platform for the simulation and analysis of CO and NCO distribution (*RecombineSim*) (Section 2.6)
- CO interference is severely diminished within *rad24*Δ backgrounds (Section 2.11)
- Single (γ) distributions insufficiently describe *in vivo* CO distribution (Section 2.12)
- Developed and applied a (γ) mixture model algorithm that significantly improves the ability to describe *in vivo* CO distributions (*GEM*) (Sections 2.13-2.14)
- A putative, universal description of WT CO interference can be obtained (Section 2.15)
- Msh2 specifically suppresses the formation of class I COs at regions of higher sequence divergence (Sections 2.16-2.18)
- Suppression of Mec1 activity elevates event formation and class II CO frequency (Section 2.19)
- Inactivation of Tel1 moderately elevates event formation and may increase class II CO frequencies (Section 2.20)

CHAPTER 2B

Investigating the role of DDR proteins within the spatial regulation of COs

Appendix

Appendix

B2.1—*HybridVar* (*v1.5*)

Aim: Processing of VCF files for heterogenous, hybrid spore read data

Input(s): Reference Genome (FASTA), VCF Files

Output: Dual coordinate variant tables, modified reference genome

Req(s): Perl 5.25, BioPerl

B2.1.1—VCF Processing

HybridVar is designed to work in conjunction with GATK HaplotypeCaller (v3.7), a *de novo* assembly approach to SNP/INDEL discovery, of which a typical run per sample constitutes:

java –jar GenomeAnalysisTK.jar –T HaplotypeCaller –R S288cReference.fa –I Sample_Sorted.bam –o Sample.vcf

For each sample, a Variant Call Format (VCF) file (v4.1) is produced, detailing all discrepancies (SNPs/INDELs) between read and reference. Data throughout this chapter was aligned against S288c, SGD Jan 2015 - R64-2-1. Variant miscalling rates are typically high, and thus further filtering is often required. VCF files adopt a columnar format defined by 9 sections: CHR, POS, ID, REF, ALT, QUAL, FILTER, INFO, GENOTYPE. Initial columns specify the detected variant:

CHR	POS	QUAL	REF	ALT	TYPE
I	27397	.	T	C	SNP
I	27398	.	T	C	SNP
I	27402	.	G	T	SNP
I	27405	.	G	GA	INDEL
I	27408	.	G	A,T	MULTIALLELIC SNP

The variable GENOTYPE section provides delimited information useful for assessing variant quality or confidence, namely (i) AD (Allelic Depth)—the number of reads which support each reported allele e.g. 0,19, denotes that 0 reads match REF, 19 reads match VAR (ii) DP (Read Depth—the number of reads covering this loci (i.e. coverage). Additional information is also specified including GT (genotype of the sample site), GQ (phred scaled confidence) and PL (normalised phred scale likelihood):

TAGS	VALUES
GT:AD:DP:GQ:PL	1/1:1,16:17:48:642,48,0
GT:AD:DP:GQ:PL	1/1:0,16:16:48:642,48,0
GT:AD:DP:GQ:PL	1/1:0,19:19:57:855,57,0
GT:AD:DP:GQ:PL	1/1:0,19:19:57:855,57,0
GT:AD:DP:GQ:PL	1/1:0,15:15:45:392,45,0

HybridVar exploits these scoring parameters, sequentially reading each VCF file provided and calculating (i) call frequency (CF) (% of spores (VCF files) any given allele is present within) (ii) cumulative total read depth (tRD) of each loci, calculated via DP (iii) cumulative allelic read depth (vRD) (% of reads that contain a specific allele at a specific loci), calculated via AD. A typical run of *HybridVar*, which allows user specified filtering based on CF, tRD and vRD, is:

HybridVar.pl –r <ReferenceFASTA> –lf <CallFreqLowerLimit> –uf <CallFreqUpperLim> –trd <MinReadDepth> –vrd <MinVarDepth>

HybridVar.pl –r s288c.fasta –lf 48 –uf 52 –trd 250 –vrd 0.95

Multiallelic sites, with >1 ALTs specified (e.g. A,T as shown above), are split, assessed and filtered separately. INDELs shift the relative positions of all variants. *HybridVar* therefore progressively tracks these changes in order to construct a dual coordinate tab delimited .txt variant file for all SNPs/ INDELs which pass filtering in the following format:

ID	chrom	pos_A	pos_B	seq_A	seq_B	type_A	type_B
1	1	27804	27804	C	A	s	s
2	1	27810	27810	T	C	s	s
3	1	27816	27816	G	A	s	s
4	1	27822	27822	T	C	s	s
5	1	27823	27823	C	A	s	s
6	1	27825	27825	C	T	s	s
7	1	27914	27914	T	C	s	s
8	1	27948	27948	A	G	s	s
9	1	27970	27970	T	G	s	s
10	1	27983	27983	T	A	s	s
11	1	27997	27997	T	C	s	s
12	1	28007	28007	G	C	s	s
13	1	28008	28008	C	C	d	i
13	1	-	28009	-	A	d	i
14	1	28021	28022	G	T	s	s

Pos_A specifies the reference coordinate while pos_B specifies the position of any given variant within a hypothetical genome that contains only these listed variants. Variant ID(13) denotes an insertion relative to the reference (C→CA). All subsequent pos_B positions are thus shifted by 1 bp to account for the additional A base. The terminal columns (type_A, type_B) specify the type of variant relative to each genome—s = SNP, i = insertion, d = deletion.

B2.1.2—Variant Genome

Reads heavily laden with variants relative to the base reference (i.e. S288c) may fail alignment, losing critical event information—a caveat, however, that is bypassed by a dual alignment approach against two references. To accommodate this, *HybridVar* utilises the information stored within filtered variant tables to modify a user provided FASTA file (REF), constructing a novel reference containing all detected variants (VAR), improving the alignment of variant dense reads:

TTGTTC**T**TTTTAAATT**GC**_AATTTAAAGAGCGTACCTGTAAATAAGAAG — *REF* (uIDs 11/12/13)

TTGTTC**C**TTTTAAATT**CCA**AATTTAAAGAGCGTACCTGTAAATAAGAAG — *VAR* (uIDs 11/12/13)

Variants ID(11) (T→C SNP), ID(12) (G→C SNP) and ID(13) (C→CA insertion) are marked above. The modified genome and generated tab delimited variant tables feed directly into the event assignment pipeline.

Script—HybridVar.pl

```
#!/usr/bin/env perl
#Version: 1.1

use strict;
use warnings;
use Bio::SeqIO;
use Getopt::Long;
use File::Basename qw(basename);
use List::Util qw(all);
my @files = glob("*.vcf");
my $chk   = scalar(@files);
print "\nFailed to detect any .VCF files within the current directory.\n\n"
  if $chk == 0;
exit if $chk == 0;   #Stop script if no .vcf files are found
my ( $fasta, $tRDfilt, $vRDfilt, $freqlow, $freqhigh );
my $scriptname = basename($0);   #Obtain script-name
my $usage =
"Usage: $scriptname -r <ReferenceFASTA> -lf <CallFreqLowerLim> -uf <CallFreqUpperLim> -trd
<MinReadDepth> -vrd <MinVarDepth>"; #Error/usage message
GetOptions(
   'r=s'  => \$fasta,  #Command-line arguments
   'lf=f' => \$freqlow,
   'uf=f' => \$freqhigh,
   'trd=i' => \$tRDfilt,
   'vrd=f' => \$vRDfilt
) or die("\n$usage\n");
die(
"\nError: Arguments or -flags are missing and/or incorrectly specified.\n\n$usage\n\n"
) unless all { defined } $fasta, $freqlow, $freqhigh, $tRDfilt, $vRDfilt;
print "\n\n$chk Samples Detected";
print "\n-----------------------------------------";
print "\nFiltering and Merging Variants...\n";
print "-----------------------------------------\n";
my ( @calls, @ID, @refsplit );
my ( %freq, %RD, %AD, %offset, %sequences, %refdupl, %dups );
my (
   $uID,    $varcount, $discard, $snp, $indel,
   $overlap, $rkey,    $dupkey,  $splituID
);
my $outfile  = "VariantStats.txt";
my $outfile2 = "LowQualVariants.txt";
open my $OUT,  '>', "CallStats.txt"      or die "$!";
open my $OUT2, '>', "LowQualVariants.txt" or die "$!";
print $OUT "uID\tChr\tPos\tRef\tVar\tCallFreq\ttRD\tvRD/tRD\n";
print $OUT2 "Chr\tPos\tRef\tVar\tCallFreq\ttRD\tvRD/tRD\n";

for my $file (@files) {   #For-each input file
   open my $IN, '<', $file or die "$!";
   while (<$IN>) {
      next if /^\s*#/;
      chomp $_;
      my @F     = split( "\t",    $_ );
      my @varinfo = split( m[[:,/]+], $F[9] );   #Split genotype information
      next if $varinfo[4] == 0;              #Skip false-positives
      my $check = index( $F[4], ',' );
      if ( $check == '-1' ) {   #For each unique, mono-allelic variant
```

```
        $freq{ $F[0] }{ $F[1] }{ $F[3] }
         { $F[4] }++;       #Calculate call-frequencies
        $RD{ $F[0] }{ $F[1] }{ $F[3] }{ $F[4] } +=
         $varinfo[4];       #Cumulative total of total read-depth (tRD)
        $AD{ $F[0] }{ $F[1] }{ $F[3] }{ $F[4] } +=
         $varinfo[3];       #Cumulative total of allelic read-depth (vRD)
    }
  }
}
my ( $i, $k ) = -1;

sub overlap {                   #Subroutine to identify overlapping variants
  my ( $id, $type, $value ) = @_;
  $refsplit[ ++$i ] = [ $id, $type, $value ];
  $rkey = "$type:$value";
  $refdupl{$rkey} = [] if !exists $refdupl{$rkey};
  push @{ $refdupl{$rkey} }, $i;
  return;
}
print "Identifying overlaps...\n";
print "-------------------------------------------\n";
foreach my $chrnum ( sort keys %freq ) {  #For each unique, mono-allelic variant
  foreach my $pos ( sort { $a <=> $b } keys %{ $freq{$chrnum} } ) {
    foreach my $ref ( keys %{ $freq{$chrnum}{$pos} } ) {
      foreach my $var ( keys %{ $freq{$chrnum}{$pos}{$ref} } ) {
        $varcount++;            #Total no. unique variant-count
        my $callfreq  = ( $freq{$chrnum}{$pos}{$ref}{$var} ) / $chk;
        my $readdepth = ( $RD{$chrnum}{$pos}{$ref}{$var} );
        my $vardepth  = ( $AD{$chrnum}{$pos}{$ref}{$var} ) / $readdepth;
        if (   $callfreq > $freqlow
          && $callfreq < $freqhigh
          && $readdepth > $tRDfilt
          && $vardepth > $vRDfilt )
        { #Filter variants using user-specified call-frequency, tRD and vRD thresholds
          $uID++;
          printf( $OUT "%d\t%s\t%d\t%s\t%s\t%.3f\t%d\t%.3f\n",
            $uID, $chrnum, $pos, $ref, $var, $callfreq, $readdepth,
            $vardepth );
          if (   length($ref) == length($var)
            || length($ref) < length($var) )
          {   #For SNPs or deletions (relative to reference)
            overlap( $uID, $chrnum, $pos );
          }
          elsif ( length($ref) > length($var) )
          {   #For insertions (relative to reference)
            my $del = length($ref);
            overlap( $uID, $chrnum, $pos );
            foreach my $delsplit ( 1 .. length($ref) - 1 )
            { #For each additional inserted base (within the reference)
              overlap( $uID, $chrnum, $pos + $delsplit );
            }
          }
        }
        else {
          $discard++;   #Total no. dicarded variants
          printf( $OUT2 "%s\t%d\t%s\t%s\t%.3f\t%d\t%.3f\n",
            $chrnum, $pos, $ref, $var, $callfreq, $readdepth,
            $vardepth );
        }
```

```
        }
      }
    }
}
foreach my $entries (@refsplit) {
   my $dupkey = "$entries->[1]:$entries->[2]";
   if ( @{ $refdupl{$dupkey} } > 1 )
   {   #For any non-unique chr-pos combinations (overlaps)
      $overlap++;   #Total no. overlapping variants
      $dups{ @$entries[0] } = {};   #Store uID of all overlapping variants
   }
}
my $seqio =
 Bio::SeqIO->new( -file => $fasta );   #Read and store .FASTA chromosomes
while ( my $seqobj = $seqio->next_seq ) {
   my $id  = $seqobj->display_id;
   my $seq = $seqobj->seq;
   $sequences{$id} = $seq;
}
close $OUT;
close $OUT2;
print "Constructing variant reference...\n";
print "-----------------------------------------\n";
open my $IN2, '<', "CallStats.txt" or die "$!";
<$IN2> for ( 1 .. 1 );              #Skip headline
open my $OUT3, '>', "VariantTable.txt"       or die "$!";
open my $OUT4, '>', "VariantRef.fa"          or die "$!";
open my $OUT5, '>', "VariantRefChromSizes.txt" or die "$!";
print $OUT3 "uID\tchrom\tpos_c\tpos_k\tseq_c\tseq_sk\ttype_c\ttype_k\n";

while (<$IN2>) {
   chomp $_;
   my @F2 = split( "\t", $_ );       #Split each tab-delimited field
   next if exists( $dups{ $F2[0] } );  #Skip overlapping variants
   if ( defined $offset{ $F2[1] } )
   {   #Offset counters for each chromosome (INDEL-dependent position shifts)
   }
   else {
      $offset{ $F2[1] } = 0;
   }
   $splituID++;
   if ( length( $F2[3] ) == length( $F2[4] ) ) {   #For SNPs
      substr( $sequences{ $F2[1] }, ( $F2[2] - 1 + $offset{ $F2[1] } ), 1 ) =
       $F2[4];                          #Ref->Var SNP substitution
      $snp++;                            #Total no. SNPs
      printf( $OUT3 "%d\t%s\t%d\t%d\t%s\t%s\t%s\t%s\n",
         $splituID, $F2[1], $F2[2], $F2[2] + $offset{ $F2[1] },
         $F2[3], $F2[4], "s", "s"
      );
   }
   elsif ( length( $F2[3] ) < length( $F2[4] ) )
   {   #For deletions (relative to reference)
      substr( $sequences{ $F2[1] }, ( $F2[2] - 1 + $offset{ $F2[1] } ), 1 ) =
       $F2[4];   #Insertion of additional variant bases
      $indel++;   #Total no. INDELs
      printf( $OUT3 "%d\t%s\t%d\t%d\t%s\t%s\t%s\t%s\n",
         $splituID, $F2[1], $F2[2], $F2[2] + $offset{ $F2[1] },
         $F2[3], substr( $F2[4], 0, 1 ),
         "d", "i"
```

```
      );
      foreach my $inssplit ( 1 .. length( $F2[4] ) - 1 )
      {         #Base-by-base split of insertion
        printf( $OUT3 "%d\t%s\t%s\t%d\t%s\t%s\t%s\t%s\n",
          $splituID, $F2[1], "-", $F2[2] + $offset{ $F2[1] } + $inssplit,
          "-", substr( $F2[4], $inssplit, 1 ),
          "d", "i"
        );
      }
      $offset{ $F2[1] } +=
       length( $F2[4] ) - length( $F2[3] );   #Calculate position offset
    }
    elsif ( length( $F2[3] ) > length( $F2[4] ) )
    {   #For insertions (within the reference)
      my $del = length( $F2[3] );
      substr( $sequences{ $F2[1] }, ( $F2[2] - 1 + $offset{ $F2[1] } ), $del )
       = $F2[4];   #Deletion of inserted bases
      $indel++;     #Total no. INDELs
      printf( $OUT3 "%d\t%s\t%d\t%d\t%s\t%s\t%s\t%s\n",
        $splituID, $F2[1], $F2[2],
        $F2[2] + $offset{ $F2[1] },
        substr( $F2[3], 0, 1 ),
        $F2[4], "i", "d"
      );
      foreach my $delsplit ( 1 .. length( $F2[3] ) - 1 ) {
        printf( $OUT3 "%d\t%s\t%d\t%s\t%s\t%s\t%s\t%s\n",
          $splituID, $F2[1], $F2[2] + $delsplit,
          "-", substr( $F2[3], $delsplit, 1 ),
          "-", "i", "d"
        );
      }
      $offset{ $F2[1] } -= $del - 1;   #Calculate position offset
    }
}
for my $chr ( sort keys %sequences ) {   #Construct variant .FASTA file
    print $OUT4 ">$chr\n$sequences{$chr}\n";
    print $OUT5 "$chr\t", length( $sequences{$chr} ), "\n";
}
my $run_time = time() - $^T;
print "Total Variants: $varcount\n";
print "Failed: $discard\n";
print "Overlapping: $overlap\n";
print "Passed: ", $varcount - $discard, " (SNPs: $snp, INDELs: $indel)\n";
print "-----------------------------------------\n";
print "Run Completed\n";
print "Processing Runtime: $run_time Seconds\n";
print "-----------------------------------------\n\n\n";
```

B2.2—*RecombineSim* (*v2.2*)

Aim: Processing of hybrid spore data and simulation of meiotic event distributions (CO, NCO, Total)

Input(s): Event assignment data, Variant table

Output: Event count tables, experimental IED distributions (individual/aggregated), experimental MLE (γ) fits (individual/aggregated), simulated IED distributions

Req(s): MATLAB (2017a)

RecombineSim constitutes an all inclusive data processing and simulation package specifically designed for hybrid tetrad NGS approaches to recombination mapping (see: Section 2.2). *RecombineSim*, designed within MATLAB (2017a), provides a callable function for automated job queuing:

RecombineSim(*EventAssignmentFile*, *VariantTable*, *Output Folder*, *InputGenotype*, *MergeThreshold*, *SimulatedSampleSize* (*M*), *Mode*, C_{PROB}, *customalpha*, *custombeta*)

Example Queue (M = 1000):

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*msh2*',*1500*,*1000*,'*Random*',*0*)

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*msh2tel1*',*1500*,*1000*,'*Random*',*0*)

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*ndt80AR*',*1500*,*1000*,'*Random*',*0*)

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*WT*',*1500*,*1000*,'*Hazard*',*32*)

B2.2.1—Data Processing (Event Counts)

Following event assignment and event merging (see: Figure 2.1), data is primarily specified within event assignment tables, which serve as the primary input for *RecombineSim* (*unused columns omitted):

ID	Meiosis	Threshold	Genotype	GID	Chr	CO_NCO	Midpoint
1	msh2_1	1500	msh2	1	16	NCO	63206
2	msh2_1	1500	msh2	1	10	NCO	360711
3	msh2_1	1500	msh2	1	8	NCO	399094
4	msh2_1	1500	msh2	1	13	NCO	51455
5	msh2_1	1500	msh2	1	8	CO	371642

Individual and averaged event counts are subsequently calculated on a per chromosome and per repeat basis for each event type (CO/NCO). Called midpoint values are utilised to calculate experimental IEDs as the distance between successive events of a given type. Event count and IED information for further analysis is provided to the user within tab delimited .txt files, detailing: (i) IED distributions for individual repeats, per event type (ii) aggregated IED distributions for the genotype, per event type (iii) Event counts for individual repeats, per chromosome and per event type (iv) Averaged event counts per chromosome and per event type with standard deviation values.

B2.2.2—Data Processing (MLE γ fitting)

Maximum likelihood estimation (MLE), via MATLAB's *fitdist* toolbox, is utilised to obtain best fit $\gamma(\alpha,\beta)$ parameters from calculated experimental IED distributions, on a per repeat and per genotype (aggregated IED) basis. $\gamma(\alpha,\beta)$ information is provided to the user within tab delimited .txt files, with 95% confidence interval (CI) values—detailing a range within which the real $\gamma(\alpha,\beta)$ values likely reside.

B2.2.3—Simulation (Virtual Chromosomes)

Virtual chromosomes, upon which simulated event formation occurs, are constructed at a 100bp resolution as binned, numerical arrays proportional in size to *in vivo* (chromosome length*0.01) (*S. cerevisiae*—S288c). Chromosomal lengths are further adjusted to reflect the limit of experimental detection governed by the leftmost and rightmost genetic markers (SNPs/INDELs), creating subtelomeric "dead zones". Any given 100bp bin contains values in the range of [0.0-1.0], denoting the inherent recombination potential (recom(P)) of this loci. Prior to initial event formation, all bins are populated with [1.0]—denoting an equal and full recombination potential. Under conditions of independency (random simulation), recom(P) values remain unaltered.

B2.2.4—Simulation (CO Designation, Site Selection & Event Formation)

Class II CO frequency, a user specified parameter, is set as a decimal fraction in the range of [0.0-1.0] (0-100%) via the C_{PROB} parameter. Subclass designation for any given CO event is determined via a randomly generated number (C) in the range [0.0-1.0]. If $C<C_{PROB}$, the event is designated class II and is randomly assigned a location on the chromosome independently of recom(P) values. If $C>C_{PROB}$, a class I event and subsequent CO interference is generated. Recom(P) values are sensed, in order to determine the position of an interference sensitive class I CO, via a weighted, roulette wheel selection algorithm (RWS). RWS constructs a set of arrays where lengths are proportional in length to Recom(P) values held within each chromosomal bin. These arrays are subsequently concatenated and a position along the joined array (F) is randomly chosen (R). Higher recom(P) values translate into a larger proportion of F, thus a higher probability of the corresponding array segment being selected by R. Bins containing recom(P) values of [0.0] (no recombination potential) are excluded i.e. non-selectable. No such designation check is performed during NCO simulations. During random simulations, the system effectively performs unweighted sampling without replacement—that is, the same 100bp bin cannot be chosen twice. Subsequent to the formation of each event, the potential number of IEDs that would be produced by the current array of events is assessed—taking into account simulated merging at a set threshold (e.g. 1.5kb). Event formation continues until the experimentally observed number of IEDs, as calculated by *RecombineSim*, is obtained—simplifying direct comparisons of model-experimental fit. Additional cells (e.g. N = 1000) are independently simulated and resulting simulated IED distributions, provided to the user in tab delimited .txt files, are averaged to reduce stochastic noise (via MATLAB: *downsample*).

B2.2.5—Simulation (Hazard Functions)

Under *Hazard* or *UniHazard* mode, *RecombineSim* imposes CO interference using data derived or user specified $\gamma(\alpha,\beta)$ values respectively to calculate the corresponding hazard function (h(x)) (EQN 1.1). H(x) is essentially calculated as (PDF/1-CDF)—where PDF is the $\gamma(\alpha,\beta)$ probability distribution function and 1-CDF is the inverse $\gamma(\alpha,\beta)$ cumulative distribution function. The numerator and

denominator of h(x) are differentiable functions asymptotically approaching zero with increasing values of (x). Thus, according to L'Hopital's rule (use of derivatives to evaluate limits involving indeterminate forms) the limiting, upper value of the h(x) ratio (y) can be approximated by calculating $1/\gamma(\beta)$, allowing for more rapid normalisation of any given h(x) to a scale of [0.0-1.0], as opposed to the conditional probability values naturally held by a h(x).

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x)}{S(x)}$$ **(EQN 1.1)**

Where f(x) = PDF, F(x) = CDF, S(x) = Survival Function

To reduce computational time, resulting h(x)'s are trimmed at 500kb equivalent if applicable and converted into a bidirectional interference function through inversion (1-(hx)) and horizontal concatenation of two oppositely oriented functions (see: Figure 2.2E). Upon generation of a class I CO event, this function is superimposed (through multiplication) onto the virtual chromosome array, centred on the initiating event. Recom(P) values in adjacent bins are therefore altered, reducing them in a distance-dependent manner up to ±500kb away. The bin containing the initiating event and those immediately adjacent are modified to possess values of [0.0] and no further recombination is permitted at this loci.

Script—RecombineSim.m (Data Processing, CO/NCO Simulation)

```
function = RecombineSim(eventfile,varfile,folder,genotype,threshold,samples,mode,COratio,alpha,beta)

%% Data Import & Processing
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
fid1 = fopen('ChrSizesS288cH4L2_L2HG.txt','r');
chrsizes = round(cell2mat(textscan(fid1, '%d','HeaderLines',1))/100);
data = readtable(eventfile,'Delimiter','\t');
vars = readtable(varfile,'Delimiter','\t');
indices = find(data{:,{'threshold'}}==threshold & strcmp(data{:,{'Genotype'}},genotype) & ~ismember(data{:,
{'type'}}, {'8:0','0:8','0:8_8:0','8:0_0:8'}));
eventl = table2array(data(indices,{'len_mid'}));
data = data(indices,:);
data = sortrows(data,{'Meiosis','chr','midpoint'},{'ascend','ascend','ascend'});
ID = table2array(data(:,{'GenotypeID'}));
uID = unique(ID);
chrindex = table2array(data(:,{'chr'}));
[bincounts,~] = histc(eventl,1:50:round((max(eventl)/50)*1.5)*50);
eventsizes = transpose(1:50*length(bincounts)); resec = 0;
eventw = repelem(bincounts,50);
CO_NCO = (data{:,'CO_NCO'});
eventlist = char(regexprep(CO_NCO, {'NCO','CO','U'},{'A','B','C'}));
eventcount = reshape(crosstab(chrindex(:),eventlist(:)-64,ID(:)),length(chrsizes),[],1);
fclose('all');
if ismember('C',eventlist)==0
   pad = zeros(16,length(uID));
   eventcount = InsertRows(eventcount',pad',[2:2:length(uID)*2])';
end
progress = strcat('Simulating:',genotype,'||','Mode:',mode,'||',folder);
disp(progress);

%% Interference Function & Event Distributions
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
s=0; w=0;
for m = 1:3:size(eventcount,2)
   s=s+1;
   count = eventcount(:,m:(m-1)+3);
   count(:,3) = sum(count(:,1:3),2);
   [~,col] = find(count >= 1);
   IEDnum = sum(count,1)-(histc(col, 1:size(count,2))');
   rowID = find(data{:,{'GenotypeID'}}==uID(s));
   midpoint = table2array(data(rowID,{'midpoint'}))';
   IED = zeros(length(midpoint)-length(unique(chrindex(rowID,:))),3);
   [a,~,subs] = unique([eventlist(rowID,1)-64 chrindex(rowID,1)],'rows');
   [~, I] = sort(subs);
   pos = midpoint(I);
   subs = subs(I,:);
   temp = accumarray(subs,1:numel(subs),[],@(x){abs(diff(pos(x(end:-1:1))))});
   for ii = 1:max(a(:,1))
      vals = [temp{ a(:,1) == ii }];
      IED(1:numel(vals),ii) = vals;
   end
   IED(:,3) = cell2mat(accumarray(chrindex(rowID,:),midpoint(:),[], @(x) {(diff(x))}));
   exp{s} = IED;
   for h=1:3
         w=w+1;
         rIED = IED(:,h);
```

```
    rIED(rIED==0)= [];
    gam = fitdist(rIED,'gamma');
    gamfit(w,1) = gam.a; gamfit(w,2) = gam.b;
    ci = paramci(gam,0.05);
    gamfit(w,3:6) = ci(:)';
    int = [];
  if strcmp(mode,'Hazard') == 1
    rIED = IED(:,h);
    rIED(rIED==0)= [];
    if gam.a <= 1.01
      gam.a = 1.02;
    end
    adjgam = gam.b/100;
    haz = Hazard(1:20000,gam.a,adjgam);
    indices = find(haz(1,:)>(1/adjgam)*0.95);
    haz = haz(1,1:min(indices));
    haz = normalize_var(haz,0,1);
    if length(haz) > 5000
      haz = haz(1,1:5000);
    end
    int(1:length(haz)) = fliplr(haz);
    int(length(haz)+1:length(haz)*2) = haz;
  elseif strcmp(mode,'UniHazard') == 1 || strcmp(mode,'MixModel') == 1
    adjgam = beta(h)/100;
    haz = Hazard(1:20000,alpha(h),adjgam);
    indices = find(haz(1,:)>(1/adjgam)*0.95);
    haz = haz(1,1:min(indices));
    haz = normalize_var(haz,0,1);
    if length(haz) > 5000
      haz = haz(1,1:5000);
    end
    int(1:length(haz)) = fliplr(haz);
    int(length(haz)+1:length(haz)*2) = haz;
  elseif strcmp(mode,'Random') == 1
    int = ones(1,10);
  else
    error('Invalid event distribution mode selected. Options: Random, Hazard, UniHazard or MixModel');
  end
  width = length(int)/2;
  cmcount = 0; i = 0;
  intwindows{s,h} = int;
  while cmcount~=samples
    i = i+1;
    dist = cell(1,16);
    rawdist = cell(1,16);
    for j=1:16
      num = count(j,h);
      if num==0
        continue
      end
      varID = find(vars{:,{'chrom'}}==j);
      coords = table2array(vars(varID,{'pos_c'}))';
      telomereL = round(min(coords)/100);
      telomereR = round(max(coords)/100);
      bound = []; lbound = []; rbound = [];
      model = ones(1,chrsizes(j)-telomereL-(chrsizes(j)-telomereR));
      smodel = ones(1,length(model)*10);
      edgeL = length(model)*5;
      edgeR = length(model)*6;
```

```
            for k=1:500
               pos = edgeL:edgeR;
               weight = smodel(edgeL:edgeR);
               classrnd = rand(1,1);
               if classrnd>(COratio/100)
                  ds = pos(sum(bsxfun(@ge,rand(1,1),cumsum(weight./sum(weight))),2)+1);
                  smodel(ds-width:ds+width-1)=smodel(ds-width:ds+width-1).*int;
                  bound(k,1:2) = [ds-resec,ds+resec];
               else
                  ds = randi([edgeL,edgeR]);
                  bound(k,1:2) = [ds-resec,ds+resec];
               end
               bound = sort(bound);
               lbound = bound(:,1); rbound = bound(:,2);
               matches = diff([rbound(1:end-1,:) lbound(2:end,:)],[],2)>(threshold/100);
               a = k-(sum(matches(:)==0));
               if a==num
                  break
               else
               end
            end
            stop = [matches;1];
            start = [1;stop(1:end-1)];
            merge = floor(mean([lbound(start~=0) rbound(stop~=0)],2));
            dist{:,j} = transpose(diff(merge));
            rawIED = transpose(diff(lbound+((rbound-lbound)/2)));
            rawdist{:,j} = rawIED(rawIED>0);
         end
         if sum(cellfun('length',dist))==IEDnum(h)
            cmcount = cmcount+1;
         end
         distdf{s,i,h} = [dist{:}];
         rawdistdf{s,i,h} = [rawdist{:}];
      end
      if i>samples
         idx = find(cellfun('length',distdf(:,:,h))~=IEDnum(h));
         for b=1:length(idx)
            distdf{:,idx(b),h} = [];
            rawdistdf{:,idx(b),h} = [];
         end
      end
   end
end

%% Directories
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
mkdir(strcat(pwd,'/',folder,'/','Simulations','/',genotype,'/',mode))
chk = exist(strcat(pwd,'/','Results','/','Event_Counts'),'dir');
chk2 = exist(strcat(pwd,'/','Results','/','Experimental'),'dir');
if chk ~= 7 && chk2 ~= 7
   mkdir(strcat(pwd,'/',folder,'/','Event_Counts'))
   mkdir(strcat(pwd,'/',folder,'/','Experimental'))
end

%% Population Output Labels
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
types = {'NCO' 'CO' 'Total'};
sets = {1:numel(uID),1:numel(types)};
[p1,p2] = ndgrid(sets{:});
```

```
comb = sortrows([p1(:) p2(:)],2);
for e = 1:length(comb)
  AvgPop{e} = strcat(genotype,int2str(uID(comb(e,1))),types{comb(e,2)});
end

%% Results - Population Averaging (Merged)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
resex = permute(distdf, [2 1 3]);
resex = squeeze(mat2cell(resex, size(distdf,2), ones(1, s), ones(1, h)));
stdev = cellfun(@(x) std(sort(cell2mat(x).'), 0, 2), resex, 'un', 0);
resex = cellfun(@(x) [x{:}], resex, 'uniformoutput',false);
resex = cellfun(@sort,resex,'uniformoutput',false);
dec = cellfun(@(x) decimate(x,samples),resex,'uniformoutput',false);
Lmax = max(max(cell2mat(cellfun(@numel,dec,'un',0))));
stdev = cellfun(@(x) [x; nan(max(Lmax(:)) - numel(x), 1)], stdev, 'un', 0);
stdev = cell2mat((stdev(:).'));
b = cellfun(@(c)[c(:);NaN(Lmax-numel(c),1)],dec,'uniformoutput',0);
stre = cell2mat((b(:).'))*100;
T = array2table(stre,'VariableNames',AvgPop);
filename =
strcat(pwd,'/',folder,'/','Simulations','/',genotype,'/',mode,'/',genotype,'-',mode,'-',num2str(threshold),'bp','-
PopAvg','.txt');
writetable(T,'temp.txt','Delimiter','\t');
replaceinfile('NaN',' ','temp.txt',filename);
delete('temp.txt');

%% Results - Population Averaging (Raw)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
rawresex = permute(rawdistdf, [2 1 3]);
rawresex = squeeze(mat2cell(rawresex, size(rawdistdf,2), ones(1, s), ones(1, h)));
rawresex = cellfun(@(x) [x{:}], rawresex, 'uniformoutput', false);
rawresex = cellfun(@sort,rawresex,'uniformoutput',false);
rawdec = cellfun(@(x) decimate(x,samples),rawresex,'uniformoutput',false);
rawLmax = max(max(cell2mat(cellfun(@numel,rawdec,'un',0))));
rawb = cellfun(@(c)[c(:);NaN(rawLmax-numel(c),1)],rawdec,'uniformoutput',0);
rawstre = cell2mat((rawb(:).'))*100;
Tr = array2table(rawstre,'VariableNames',AvgPop);
filename =
strcat(pwd,'/',folder,'/','Simulations','/',genotype,'/',mode,'/',genotype,'-',mode,'-',num2str(threshold),'bp','-
RawPopAvg','.txt');
writetable(Tr,'temp.txt','Delimiter','\t');
replaceinfile('NaN',' ','temp.txt',filename);
delete('temp.txt');

%% Experimental Output Labels
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for e=1:length(uID)
  for r=1:length(types)
    ExpDat{(e-1)*length(types)+r} = strcat(genotype,int2str(uID(e)),types{r});
  end
end

%% Results - Experimental Data
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
filename = strcat(pwd,'/',folder,'/','Experimental','/',genotype,'-',num2str(threshold),'bp','-ExpIED','.txt');
chk = exist(filename,'file');
if chk ~= 2
  Lmax = max(cellfun('size',exp,1));
  exp = cellfun(@(c) [c;NaN(Lmax-size(c,1),3)],exp,'uniformoutput',0);
```

```
   exp = horzcat(exp{:});
   exp(exp==0)=NaN;
   exp = sort(exp);
   T = array2table(exp,'VariableNames',ExpDat);
   writetable(T,'temp.txt','Delimiter','\t');
   replaceinfile('NaN',' ','temp.txt',filename);
   delete('temp.txt');
end

%% Results - Experimental Data (Aggregate)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if chk ~= 2
   filename = strcat(pwd,'/',folder,'/','Experimental','/',genotype,'-',num2str(threshold),'bp','-
AggregateExpIED','.txt');
   agg = reshape(permute(reshape(exp,size(exp,1),3,[]),[1,3,2]),[],3);
   agg = sort(agg);
   for y=1:size(agg,2)
      w=w+1;
      idx = ~isnan(agg(:,y));
      gam = fitdist(agg(idx,y),'gamma');
      gamfit(w,1) = gam.a; gamfit(w,2) = gam.b;
      ci = paramci(gam,0.05);
      gamfit(w,3:6) = ci(:)';
   end
   TAgg = array2table(agg,'VariableNames',{'NCOAvg' 'COAvg' 'TotalAvg'});
   writetable(TAgg,'temp.txt','Delimiter','\t');
   replaceinfile('NaN',' ','temp.txt',filename);
   delete('temp.txt');
end
%% Results - Gamma-Fitting
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
filename = strcat(pwd,'/',folder,'/','Experimental','/',genotype,'-',num2str(threshold),'bp','-GammaFit','.txt');
chk = exist(filename,'file');
if chk ~= 2
   GamDat = horzcat(ExpDat,'NCOAggregate','COAggregate','TotalAggregate');
   fit = array2table(gamfit,'RowNames',GamDat,'VariableNames',
{'Alpha','Beta','LowerLimA','UpperLimA','LowerLimB','UpperLimB'});
   writetable(fit,'temp.txt','Delimiter','\t','WriteRowNames',true);
   replaceinfile('Row',' ','temp.txt',filename);
   delete('temp.txt');
end

%% Results - Genotype Aggregation (Merged)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
agg = sort(reshape(stre, size(stre,1)*(numel(resex)/3), 3));
filename =
strcat(pwd,'/',folder,'/','Simulations','/',genotype,'/',mode,'/',genotype,'-',mode,'-',num2str(threshold),'bp','-
AggregateSim','.txt');
TAgg = array2table(agg,'VariableNames',{'NCOAvg' 'COAvg' 'TotalAvg'});
writetable(TAgg,'temp.txt','Delimiter','\t');
replaceinfile('NaN',' ','temp.txt',filename);
delete('temp.txt');

%% Results - Genotype Aggregation (Raw)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
agg = sort(reshape(rawstre, size(rawstre,1)*(numel(rawresex)/3), 3));
filename =
strcat(pwd,'/',folder,'/','Simulations','/',genotype,'/',mode,'/',genotype,'-',mode,'-',num2str(threshold),'bp','-
RawAggregateSim','.txt');
```

```
TAgg = array2table(agg,'VariableNames',{'NCOAvg' 'COAvg' 'TotalAvg'});
writetable(TAgg,'temp.txt','Delimiter','\t');
replaceinfile('NaN',' ','temp.txt',filename);
delete('temp.txt');

%% Event-count Output Labels
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
types = {'NCO' 'CO' 'NA' 'Total'};
for e=1:length(uID)
  for r=1:length(types)
    ExpDat{(e-1)*length(types)+r} = strcat(genotype,int2str(uID(e)),types{r});
  end
end

%% Results - Event-count
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
filename = strcat(pwd,'/',folder,'/','Event_Counts','/',genotype,'-',num2str(threshold),'bp','-EventCount','.txt');
chk = exist(filename,'file');
if chk ~= 2
  ecsum = [];
  for q = 1:3:size(eventcount,2)
    ectmp = eventcount(:,q:(q-1)+3);
    ectmp(:,4) = sum(ectmp(:,1:3),2);
    ecsum = horzcat(ecsum,ectmp);
  end
  NCOavg = (round((mean(ecsum(:,1:4:size(ecsum,2)),2))*100)/100);
  NCOstd = (round((std(ecsum(:,1:4:size(ecsum,2)),0,2))*100)/100);
  COavg = (round((mean(ecsum(:,2:4:size(ecsum,2)),2))*100)/100);
  COstd = (round((std(ecsum(:,2:4:size(ecsum,2)),0,2))*100)/100);
  totavg = (round((mean(ecsum(:,4:4:size(ecsum,2)),2))*100)/100);
  totstd = (round((std(ecsum(:,4:4:size(ecsum,2)),0,2))*100)/100);
  ecsum = horzcat(ecsum,NCOavg,NCOstd,COavg,COstd,totavg,totstd);
  sum(ecsum(:,size(ecsum,2)-1));
  for u=1:16
    chrlabels{u,:} = strcat('chr',num2str(u));
  end
  ExpDat = horzcat(ExpDat,'NCOAverage','NCO_STD','COAverage','CO_STD','TotAverage','Tot_STD');
  EC = array2table(ecsum,'RowNames',chrlabels,'VariableNames',ExpDat);
  writetable(EC,'temp.txt','Delimiter','\t','WriteRowNames',true);
  replaceinfile('Row',' ','temp.txt',filename);
  delete('temp.txt');
end

%% Results - Store Database
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
filename =
strcat(pwd,'/',folder,'/','Simulations','/',genotype,'/',mode,'/',genotype,'-',mode,'-',num2str(threshold),'bp','.mat');
save(filename)
```

Script—HazardFunction.m

```
function y = Hazard(t,A,B)
A(A <= 0) = NaN;
B(B <= 0) = NaN;
y = gampdf(t,A,B)./(1-gamcdf(t,A,B));
y(t < 0) = 0;
```

B2.3—*Gamma Expectation Maximisation* (*GEM*) (*v1.1*)

Aim: Cluster and (γ) mixture modelling for IED data

Input(s): IED distributions (aggregated)

Output: γ(α,β) and weight parameter estimates, log-likelihood trace

Req(s): MATLAB (2017a)

B2.3.1—(γ) Parameter Estimation (MLE)

GEM constitutes a (γ) distribution specific application of the expectation maximisation (EM) algorithm. EM is an iterative method to obtain maximum likelihood estimates (MLE) for statistical, mixed models that contain latent values (e.g. class II CO frequency) (Do & Batzoglou 2008). *GEM* is designed as a callable function and a standalone package:

GEM(*IEDdata*,$n_{COMPONENTS}$,*maxIter*,*error_thresh*,*init_mode*,*init_alpha*,*init_beta*,*init_weight*)

MLE derivation of mixed γ(α,β) parameters is well established (Webb 2000; Destrempes et al. 2011). The likelihood function L(X|α|β) is key to statistical inference, describing the collective likelihood that the observed data (X_I) arose from the probability density function f(α,β) (EQN 1.2) of the proposed or fitted quantitative model. Maximum likelihood estimation seeks to maximise likelihood by obtaining γ(α,β) values that best fit (X_I). A more mathematically convenient form is the log likelihood function, the natural logarithmic transformation of EQN 1.2 (EQN 1.3).

$$L(X|\alpha,\beta) = \prod_{i=1}^{N} f(x_i;\alpha,\beta) \quad \textbf{(EQN 1.2)}$$

For N independently, identically distributed (i.i.d) variables (x1….., xN)

$$logL(x_i|\alpha_j,\beta_j) =$$

$$(\alpha_j - 1)\sum_{i=1}^{N} log(x_i) - \sum_{i=1}^{N}\frac{x_i}{\beta_j} - N\alpha_j \log(\beta_j) - Nlog(\Gamma(\alpha_j)) \quad \textbf{(EQN 1.3)}$$

By obtaining the derivative, setting the equation to equal zero and finding the maximum with respect to γ(β), it can be shown that γ(β) estimation can be fully expressed in terms of (X_I) and γ(α) (EQN 1.4). Substituting EQN1.4 into EQN1.2 and subsequently taking the derivative, setting the

equation to equal zero and finding the maximum with respect to γ(α), the equation for γ(α) MLE is obtained (EQN 1.5, 1.6)—where ψ equals the digamma function (EQN 1.7). No closed form solution for γ(α) exists, however, f(x) = log(x) -ψ(x) is numerically well behaved and therefore γ(α) can be estimated through numerical means. MLE γ(β) values are subsequently obtained using the obtained maximised γ(α) value.

$$\beta_j = \frac{1}{\alpha_j} \frac{\sum_{i=1}^{N} \gamma_{i,j} x_i}{\sum_{i=1}^{N} \gamma_{i,j}}$$ **(EQN 1.4)**

Where γ(i,j) denotes the probability of X(i) (data) belonging to cluster j

$$log(\alpha) + \psi(\alpha) = log\left(\frac{\Sigma_i^N \gamma_{i,j} x_i}{\Sigma_i^N \gamma_{i,j}}\right) - \left(\frac{\Sigma_i^N \gamma_{i,j} log x_i}{\Sigma_i^N \gamma_{i,j}}\right)$$ **(EQN 1.5)**

$$log\left(\frac{\Sigma_i^N \gamma_{i,j} x_i}{\Sigma_i^N \gamma_{i,j}}\right) - \left(\frac{\Sigma_i^N \gamma_{i,j} log x_i}{\Sigma_i^N \gamma_{i,j}}\right) - log(\alpha_j) + \psi(\alpha_j) = 0$$ **(EQN 1.6)**

$$\psi(x) = \frac{d}{dx} ln\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$ **(EQN 1.7)**

B2.3.2—Cluster Analysis

GEM initially segregates data into ($n_{COMPONENTS}$) number of soft clusters (e.g. 2) and subsequently utilises MLE to reiteratively improve the fit of each sub distribution and the overall model, recalculating the log likelihood in a cyclical fashion until a termination criteria is met such as an error threshold or maximum allowed iterations (Figure 2.23A). A useful property of EQN1.6 is that the solution adopts a (-) value if the γ(α) estimate is below the maximised value, and correspondingly a (+) value if above (Figure 2.23B). Via MATLAB function *fzero*, which attempts to find a point (x) where fun(x) = 0 based on sign change, EQN 1.6 is numerically evaluated over a given range of γ(α) values for each subpopulation—a range periodically shifted based on the evaluative outcome. Such a process allows *GEM* to narrow in on the best fit γ(α) value. The relative contribution of each sub distribution (i.e. weight), is estimated by approximating the number of data points which are likely belong to each cluster.

B2.3.3—Parameter Initiation

GEM provides two parameter initiation methods: (i) A kmeans++ algorithm—an established method of parameter initiation (Blömer & Bujna 2013). Kmeans++ initially assigns a centre point to a given number of clusters (nC), assigns data an identity denoting which cluster it belongs to and directly approximates $\gamma(\alpha,\beta)$ parameters of each cluster via method of moments (EQN 1.8) (ii) A biased, non-automated approach whereby the user specifies initial $\gamma(\alpha,\beta)$ and/or weight values for a given number of clusters (nC). The choice of initiation method is context dependent, as shown in (Section 2.20).

$$\alpha = \left(\frac{x}{s}\right)^2 \qquad \beta = \frac{s^2}{x}$$

Where S = Standard Deviation, x = Sample Mean

(EQN 1.8)

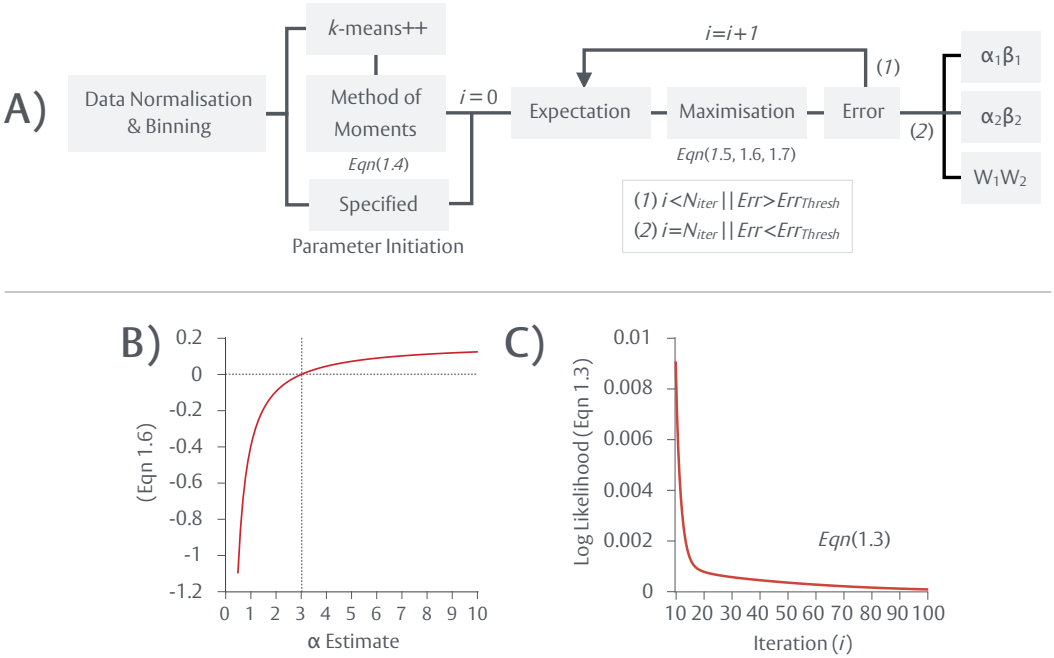


Figure 2.23. Gamma (γ) expectation-maximisation—An overview

A) During a typical run of the gamma (γ) expectation maximisation (*GEM*) algorithm, IED data is initially normalised and binned at 250bp intervals to reduce noise and the influence of outliers. Parameter initiations—initial γ(α,β) and/or weight (W) values—are either determined via an implementation of the kmeans++ algorithm, which segregates data into a preset number of cluster ($n_{COMPONENTS}$) and calculates γ(α,β) via method of moments, or specified by the user for each subpopulation. Subsequent to this, the system cycles between an expectation and a maximisation step as it converges on MLE γ(α,β) values for each subpopulation and the mixed model as a whole—a process that continues until a maximum number of iterations (*i*) is reached (e.g. 1000) or a certain error rate is met. Error rate is calculated as the cumulative, standard error of MLE. For a two component (γ) mixture, GEM provides (α,β) and W values both subpopulations (α1-2,β1-2,W1-2). **B)** Numerical evaluation of EQN 1.6 for a γ(α) = 3 distribution demonstrates how the equation equals zero when the MLE is reached. Underestimates are characterised by (-) values. Overestimates are characterised by (+) values. **C)** During each iteration (*i*), *GEM* seeks to maximise the log likelihood.

Script—Gamma Expectation Maximisation (GEM)

```
function [alpha, beta, weight, trace,logLH] =
GEM(x,nC,maxIter,error_thresh,init_mode,init_alpha,init_beta,init_weight)

%% Data Normalisation & Binning
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
x = x(:)';
s_factor = sum(x)/length(x);
norm = x./s_factor;
[N,M] = hist(norm,linspace(min(norm(:)),max(norm(:)),250));
binned_data = N/(sum(N*(M(2)-M(1))));

%% Parameter & Distribution Initialisation (Pre-EM)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if strcmp(init_mode,'static') == 1
   alpha = init_alpha; beta = init_beta/s_factor; weight = init_weight;
elseif strcmp(init_mode,'kmeans') == 1
   idx = kmeans(x',nC,'Replicates',10);
   beta = zeros(1,nC); alpha = zeros(1,nC); weight = zeros(1,nC);
   for k=1:nC
      beta(1,k) = std(norm(idx==k))^2/mean(norm(idx==k));    %Method of Moments Estimation
      alpha(1,k) = (mean(norm(idx==k))/std(norm(idx==k)))^2;
      weight(1,k) = sum(idx==k)/sum(N);
   end
end
dist = zeros(nC,length(norm));
m_dist = zeros(1,length(norm));
for j=1:nC
   dist(j,:) = gampdf(norm,alpha(j),beta(j));
   m_dist = m_dist+dist(j,:).*weight(j);
end
for p=1:nC
   w(p,:) = dist(p,:).*weight(p)./m_dist;
end
alpha_trace{1} = alpha; beta_trace{1} = beta;


%% Expectation Maximisation (EM) Algorithm
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
error = inf; i=1; iter=i; c=0; options = []; options = statset(statset('gamfit'),options);
MLE = @(x_MLE,y_MLE) y_MLE-log(x_MLE)+psi(x_MLE); %Log-likelihood function
while (error>error_thresh && i<maxIter)
   m_dist = zeros(1,length(norm));
   c=c+1;
   for n=1:nC
      weight(n) = sum(w(n,:))/(sum(w(:)));
      A = log(sum(w(n,:).*norm)/sum(w(n,:)));
      B = sum(w(n,:).*log(norm+eps))/(sum(w(n,:))+eps);
      data_term = A-B;
      logLH(c,n) = MLE(alpha(n),data_term);
      if MLE(alpha(n),data_term) > 0
         upper = alpha(n); lower = upper/2;
         while MLE(lower,data_term) > 0
            upper = lower; lower = upper/2;
         end
      else
         lower = alpha(n); upper = lower*2;
```

```
        while MLE(upper,data_term) < 0
          lower = upper; upper = lower*2;
        end
      end
      boundaries = [lower upper];
      [ahat, ~, ~] = fzero(MLE,boundaries,options,data_term);
      alpha(1,n) = ahat;
      beta(n) = sum(w(n,:).*norm)/(sum(w(n,:))*alpha(n)+eps);
      dist(n,:) = gampdf(norm,alpha(n),beta(n));
      m_dist = m_dist+dist(n,:).*weight(n);
    end
    alpha_trace{i+1} = alpha;
    beta_trace{i+1} = beta;
    w = zeros(1,length(norm));
    for b=1:nC
      w(b,:) = dist(b,:).*weight(b)./m_dist;
      w(isnan(w)) = 1;
    end
    error = 0;
    for r=1:nC
      error = error+max(abs(weight - sum(w,2)'/(sum(w(:)))));
    end
    i=i+1;
    iter = [iter,i];
end
for t=1:nC
    weight(1,n) = sum(w(n,:))/(sum(w(:)));
end

%% Results & Plots
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure('position',[0,0,950,700])
fit_final = zeros(size(M));
for g=1:nC
    it_final = fit_final+weight(g)*gampdf(M,alpha(g),beta(g));
    plot(M,weight(g)*gampdf(M,alpha(g),beta(g)),'r');hold on
    capture(:,g) = weight(g)*gampdf(M,alpha(g),beta(g));
end
plot(M,binned_data,'x','Color',[0.33 0.33 0.34]);
plot(M,fit_final,'Color',[0.14 0.24 0.62],'linewidth',3);
beta = beta*s_factor;
[val,idx2] = sort(weight,'descend');
weight = val;
beta = beta(idx2);
alpha = alpha(idx2);
trace(:,1:nC) = cell2mat(alpha_trace');
trace(:,nC+1:nC*2) = (cell2mat(beta_trace')*s_factor);
```

Strain	Entry	Background	Mat	Genotype
WT	MJ513	SK1	a	*ho::LYS2 lys2Δ leu2Δ arg4Δ*
	MJ600	S288c	α	*ho::LYS2 lys2Δ ade8Δ*
msh2Δ	MC26	SK1	α	*ho::LYS2 lys2Δ ura3Δ arg4Δ leu2Δ msh2Δ::Kan*
	MC49	S288c	a	*ho::LYS2 lys2Δ ade8Δ msh2Δ::Kan*
tel1Δmsh2Δ	MC29	SK1	a	*ho::LYS2 lys2Δ arg4Δ leu2Δ tel1Δ::HphMX4 msh2Δ::Kan*
	MC55	S288c	α	*ho::LYS2 lys2Δ ade8Δ msh2Δ::Kan tel1Δ::HphMX4*
rad24Δmsh2Δ	MC105	SK1	a	*ho::LYS2 lys2Δ ura3Δ arg4Δ leu2Δ rad24Δ::HphMX4 msh2Δ::Kan*
	MC203	S288c	α	*ho::LYS2 ade8Δ rad24Δ::HphMX4 msh2Δ::Kan*
msh2Δmec1MN	MC163	SK1	a	*ho::LYS2 lys2Δ ura3Δ arg4Δ leu2Δ::hisG nuc1Δ::LEU2 PCLB2–MEC1::Kan msh2Δ::Kan*
	MC172	S288c	α	*ho::LYS2 lys2Δ ade8Δ PCLB2–MEC1::Kan msh2::HphMX4*
ndt80AR	MJ43	SK1	α	*ho::LYS2 lys2Δ arg4Δ leu2Δ::hisG trp1Δ::hisG his4XΔ::LEU2 nuc1Δ::LEU2 PGAL1–NDT80::TRP1 ura3::pGPD1–GAL4(848)–ER::URA3*
	MC42	S288c	a	*ho::LYS2 lys2Δ ade8Δ ndt80Δ::Kan*
rad24Δndt80AR	MJ835	SK1	α	*ho::LYS2 lys2Δ arg4Δ leu2Δ::hisG trp1Δ::hisG his4XΔ::LEU2 nuc1Δ::LEU2 ura3Δ::PGPD1–GAL4(848)–ER::URA3 PGAL–NDT80::TRP1 rad24Δ::hphMX*
	MC89	S288c	a	*ho::LYS2 ade8Δ rad24Δ::HphMX4 ndt80Δ::Kan*
ndt80ARmec1MN	MC3	SK1	a	*ho::LYS2 lys2Δ arg4Δ leu2Δ::hisG his4XΔ::LEU2 nuc1Δ::LEU2 trp1Δ::hisG ura3Δ::PGPD1–GAL4(848)–ER::URA3 PCLB2–MEC1::Kan PGAL–NDT80::TRP1*
	MC198	S288c	α	*ho::LYS2 lys2Δ ade8Δ PCLB2–MEC1::Kan ndt80Δ::Kan*

Table 2.2. Strain Table—Genome-wide mapping of recombination

CHAPTER 3

Genome-wide mapping of Spo11 DSBs

3.1—Introduction

The inherently dangerous but essential act of meiotic DSB formation is subject to multiple forms of regulation that help to counteract the risks involved. Notably, and akin to COs (Chapter 2), DSBs are subject to comparable processes of spatial regulation. For example, the DNA damage response (DDR) kinase Tel1^{ATM} mediates the negative regulation of DSB formation through *in cis* suppression of DSBs in proximity to pre-existing breaks (Garcia et al. 2015). This process, known as DSB interference, results in a non-random distribution of DSBs across each chromatid (see: Chapter 4 for further analysis of DSB interference).

In order to understand how DSB formation is regulated, it is often necessary to map Spo11-dependent DSBs (Spo11 DSBs) on a genome-wide level. Mapping of Spo11 DSBs has been primarily achieved through the Spo11-oligonucleotide assay—a protocol involving the immunoprecipitation and subsequent sequencing of Spo11 associated DNA molecules released by Mre11 nucleolytic activity (Figure 3.1A) (Neale et al. 2005; Pan et al. 2011) (see: Section 1.2.4). While the Spo11-oligonucleotide assay provides near base pair resolution, it has several caveats: (i) poly(G) tailing of Spo11-oligonucleotides, a required step of the technique, produces base pair discrepancies and coordinate ambiguity when a reference 5' cytosine (C) is present (ii) short ~10-15bp Spo11-oligonucleotides are lost owing to poor alignment, recovery or multi-mapping—resulting in an incomplete picture of DSB formation (iii) immunoprecipitation of Spo11 requires affinity tags (Pan et al. 2011). Tagged *spo11–HA*, previously employed in mapping studies, is a known hypomorph that exhibits only ~10-50% WT activity and may alter DSB distribution (Gray et al. 2013; Martini et al. 2006). Recent studies have instead employed *spo11–FLAG* or *spo11–ProA* constructs to overcome this caveat (Mohibullah & Keeney 2017; Thacker et al. 2014). Nevertheless, *spo11–FLAG* or *spo11–ProA* may influence Spo11 activity in as of yet unknown ways. A further refinement of Spo11 mapping technologies is thus required to address these issues. Moreover, knowledge of how Tel1 may otherwise impact upon the formation and distribution of Spo11 DSBs is limited and a full understanding of how a loss of DSB interference or Tel1 activity may manifest itself has not yet been

realised. Work presented throughout this chapter therefore details the development, subsequent validation and usage of a novel Spo11 DSB analysis pipeline—designed alongside a newly devised *sae2*Δ-dependent mapping technique—to investigate the role Tel1 may play within the localised regulation of DSB formation.

3.2—High resolution, genome-wide mapping of Spo11 DSBs

During DSB formation, Spo11 generates a 5' phosphotyrosyl linkage—remaining covalently attached to both sides of the DSB as a protein:DNA complex (Bergerat et al. 1997; Keeney et al. 1997; Keeney & Kleckner 1995; Liu et al. 1995; Neale et al. 2005). Removal of Spo11, so that repair may proceed, ordinarily requires the concerted effort of Mre11-Rad50-Xrs/Nbs1 (MRX/N complex) and Sae2 (CtIP), releasing short ~24-40bp or ~10-15bp Spo11-oligonucleotides (Garcia et al. 2011; Symington et al. 2014). However, recombinant human tyrosyl-DNA phosphodiesterase 2 (TDP2) protein can also directly hydrolyse Spo11-DNA covalent bonds without nucleotide loss, freeing a 5' phosphate (D. Johnson, M.J. Neale unpublished). TDP2 chemistry may thus be exploited to map Spo11 DSBs with precise and unambiguous nucleotide resolution. In order to prevent Mre11-dependent removal of Spo11 moieties, mapping is conducted within a *sae2*Δ background—an end processing deficient mutant within which Spo11:DNA species accumulate. Mapping of Spo11 DSBs via the *sae2*Δ-dependent technique is based on isolation of protein bound DNA and encompasses several key steps (Figure 3.1B): (i) meiosis is induced within *sae2*Δ *S. cerevisiae* SK1 strains (ii) unproteolysed genomic DNA is isolated from 6h time course samples and sonicated (iii) a column based purification step removes non-protein bound DNA and isolated molecules of interest are prepped for next-generation sequencing (NGS). In order to impart polarity—thus identifying which end of the molecule corresponded to the Spo11 DSB—Illumina Read-2 adaptors are ligated on prior to TDP2-dependent removal of Spo11. Ligation of adaptors to the DSB end is blocked by the Spo11 moiety. Partial proteolysis by Proteinase K and full removal of Spo11 by TDP permits selective ligation of Read-1 adaptors, identified via unique, embedded sequences, to the DSB end—generating polar molecules. Size selection is subsequently performed, enriching for ~250bp molecules, and obtained libraries undergo paired end sequencing (2 x 75bp reads, Illumina MiSeq).

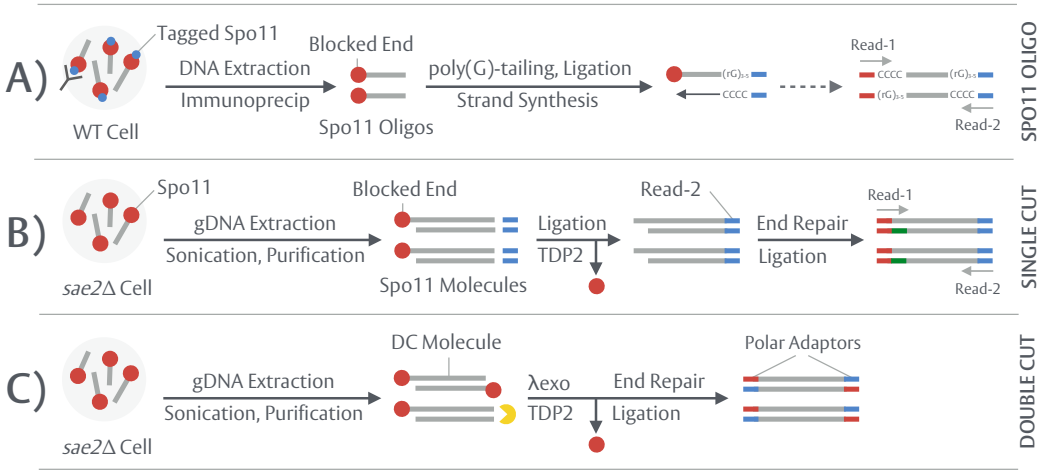


Figure 3.1. High resolution, genome-wide mapping of Spo11 DSBs

A) Spo11-oligonucleotide prep (Pan et al. 2011). Denatured nuclear extracts are prepared from WT SK1 *S. cerevisiae* cells, harbouring a tagged Spo11 allele (e.g. Spo11-HA), undergoing meiosis. Immunoprecipitation subsequently isolates Spo11-oligonucleotides for next-generation sequencing (NGS) library prep. Purified oligonucleotides are poly(G)-tailed and Read-2 illumina adaptors (blue) are ligated on. Adaptor ligation at the DSB end is blocked by the Spo11 moiety. Following additional denaturing and purification steps, Read-1 adaptors are ligated to the cleaned DSB end—imparting identifiable polarity to the captured oligos. **B)** Single cut library prep. Single cut molecules are defined as those harbouring a 5', covalently linked Spo11 moiety at one end. Unproteolysed genomic DNA (gDNA) is extracted from end processing deficient *sae2*Δ SK1 *S. cerevisiae* cells undergoing meiosis, and sonicated. A column based purification step subsequently isolates protein bound DNA (e.g. Spo11 molecules) for NGS library prep. Read-2 illumina adaptors are ligated on prior to TDP2-dependent removal of Spo11. Adaptor ligation at the DSB end is blocked by the Spo11 moiety. Partial proteolysis by Proteinase K and full removal of Spo11 by TDP2 permits selective ligation of Read-1 adaptors to the DSB end—imparting identifiable polarity to the captured molecules. **C)** Double cut library prep. Double cut molecules are defined as those harbouring 5', covalently linked Spo11 moieties at both ends. Unproteolysed, protein bound gDNA fragments are isolated as above. Lambda exonuclease (λexo), whose activity is blocked by covalently attached, terminal Spo11 molecules, degrades free 5' ended single cut molecules. Subsequent to λexo treatment, Spo11 is removed at both ends by TDP2 and polar adaptors are ligated on. The protocols detailed in (B-C) were devised by (D. Johnson, M.J. Neale unpublished).

The described protocol generates "single cut" libraries, pertaining to molecules where only a single end originated from a Spo11 DSB while the other constitutes a sonication shear point. A "double cut" variant was also developed, enriching for molecules bound by Spo11 at both ends (Figure 3.1C). Lambda exonuclease (λexo) activity is blocked by covalently attached, terminal Spo11 molecules (D. Johnson, M.J. Neale unpublished) and thus can be used to degrade free 5' ended molecules (i.e. single cuts), isolating an enriched fraction of double cuts. Following λexo treatment, Spo11 is removed from both ends via TDP2-dependent hydrolysis and polar adaptors are ligated on. It is important to note that, as opposed to the mapping of COs/NCOs (see: Chapter 2) which occurs on a per cell basis, mapping of Spo11 DSBs creates population averaged datasets. All experimental work and NGS library prep was performed by (D. Johnson, unpublished). The following sections detail the downstream computational work done to align, process and analyse the resulting data.

3.3—Spo11Mapper: A novel mapping pipeline

In order to facilitate *sae2*Δ-dependent mapping of Spo11 DSBs, a novel alignment, mapping and analysis package (*Spo11Mapper*) was developed (see: Section B3.1). *Spo11Mapper* constitutes a low memory and efficient pipeline for the batch processing of single cut and double cut Spo11 DSB libraries. Separate modes, specified by the user, for each library type exist. A typical run for *S. cerevisiae* data, under default settings, comprises several key steps (Figure 3.2A): (i) *Alignment*: Read-1 and Read-2 FASTQ files—the primary output of NGS detailing raw, paired reads for each sample—are aligned. Alignment is performed by *Bowtie2* (v2.2.6) under *end-to-end* mode. End-to-end mode aligns reads "as is", without internal read trimming to optimise map quality scores—thus preserving exact coordinate information. Data throughout this chapter was aligned against a custom version of the *S. cerevisiae* S288c reference genome (S288c—SGD Jan 2015, R64-2-1), designated Cer3H4L2, which incorporates sequences of the exogenous DSB hotspots *HIS4*::*LEU2* and *LEU2*::*HISG.* (ii) *Processing and Filtering*: *Bowtie2* produces tab delimited SAM files as a primary output, detailing key information for each mapped, unmapped or partially mapped read pair. *Spo11Mapper* subsequently assesses the quality of each read pair using the contained information.

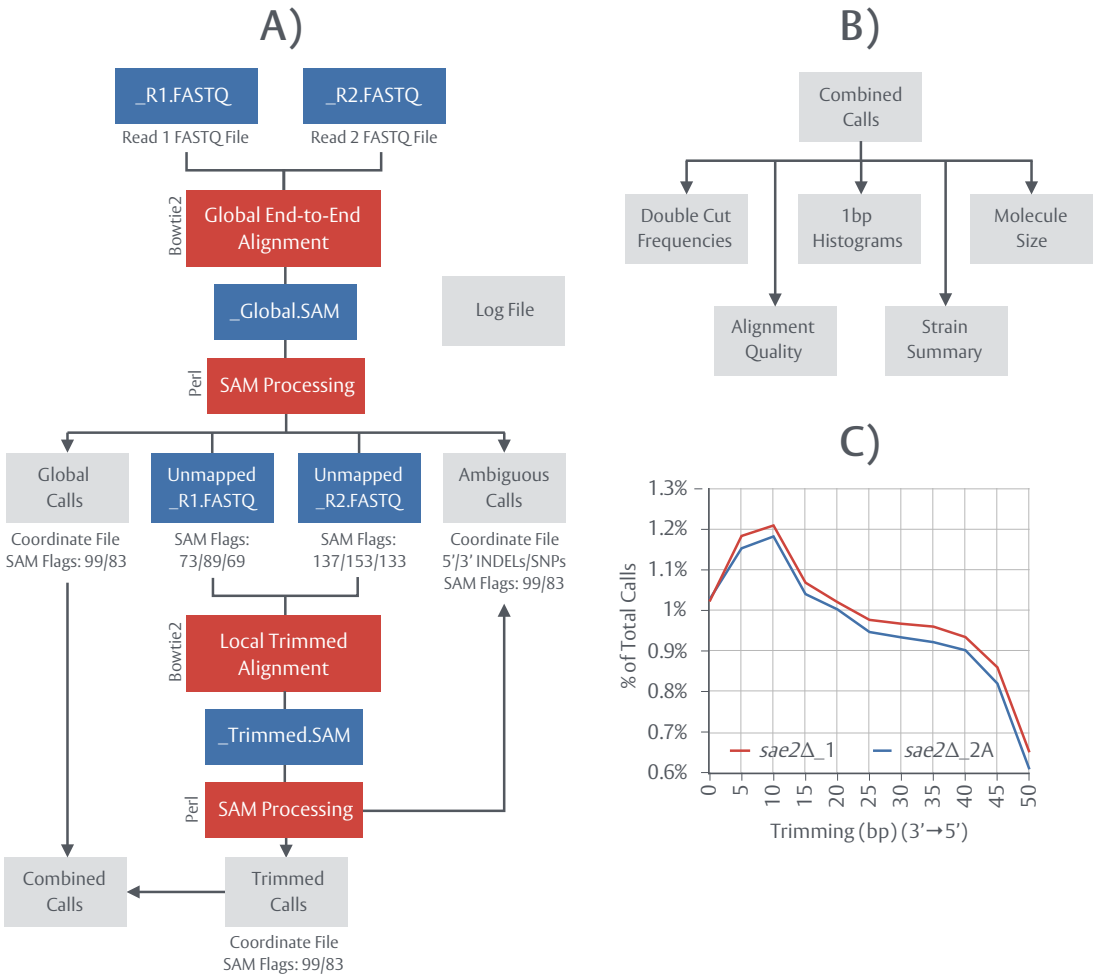


Figure 3.2. Spo11Mapper—An overview

A) *Spo11Mapper* schematic. Paired end FASTQ Read-1 and Read-2 files are initially aligned via *Bowtie2* under strict, global mode—disallowing internal read trimming. *Spo11Mapper* subsequently calculates 5' coordinates for fully aligned, properly paired reads (SAM flags 99-144 and 83-163) (SAM Processing). Ambiguous calls, defined by the presence of >1SNP or >1bp of INDEL at the 5' informative end, are detected and separated from the main dataset. Under two step alignment mode, *Spo11Mapper* trims unmapped mates (3'→5') (SAM flags 133, 69) by a user specified amount and reconstructs "unmapped" FASTQ files along with untrimmed, mapped mates (SAM flags 73, 89, 137 and 153). Trimmed, unmapped pairs are aligned via *Bowtie2* under a less strict, local mode and reprocessed as above. Default alignment parameters: -X 1000 --no-discordant --very-sensitive --mp 5,1. **B)** *Spo11Mapper* output files. From high quality, unambiguous calls, *Spo11Mapper* calculates a number of datasets (see: in-text and Section B3.1). **C)** *sae2*Δ_1 and *sae2*Δ_2A repeat libraries were aligned and processed by *Spo11Mapper* at progressively increasing levels of 3'→5' trimming. For a given trimming level, the number of high quality, unambiguous calls specifically obtained from second round alignment were assessed as a fraction of the total number of calls.

In single cut mode, 5' Read-1 coordinates (Spo11 end) are called and recorded for all high quality pairs where both individual reads, referred to as mates, successfully aligned. Use of the SK1 *S. cerevisiae* strain, which diverges from the S288c reference (~65,000 SNPs, ~4000 INDELs) (see: Section 2.3) necessitates further filtering considerations. On occasion, the 5' Read-1 end of a Spo11 molecule may map within an S288c x SK1 SNP or INDEL. In order to preserve the absolute integrity of coordinate calling, any read pair where the informative end is ambiguous is subsequently removed from the main dataset and recorded separately. In double cut mode, the 5' coordinates of both Read-1 and Read-2 are called. Ambiguity at either end, or both, results in the entire pair being disqualified from the main dataset. (iii) *Analysis*: In addition to coordinate calling, *Spo11Mapper* performs several analyses and provides a number of output files (Figure 3.2B). Using filtered, called coordinates, the pipeline generates 1bp, sparsely formatted histograms for Read-1 5' ends (single cut mode) or Read-1 and Read-2 5' ends (double cut mode)—tallying the number of Watson (W) and Crick (C) hits for any given base pair across each chromosome. 1bp histograms are primarily used to visualise Spo11 DSB maps and perform several downstream analyses, including the determination of sequence bias. Moreover, in double cut mode, *Spo11Mapper* calculates molecule size—that is, the distance between both called 5' Spo11 ends—and the frequency with which any given pair of 5' coordinates is observed. (iv) *Logs and Summary*: Throughout processing, central log files are collated for each sample—detailing *Bowtie2* alignment quality, number of called 5' ends (strict and ambiguous) and a set of stats for inter sample comparisons.

3.4 — Limited 3'→5' trimming of reads improves mapping within polymorphic regions

Reads overlapping with SNP/INDEL dense stretches of the genome can suffer significant hits to map quality scores, disqualifying a read pair from being considered successfully aligned by *Bowtie2* even if the informative, 5' end is unambiguous and intact. Information about Spo11 DSB formation in proximity to SNP/INDEL rich areas may thus be incomplete. In an attempt to improve mapping within any such region and salvage additional reads, *Spo11Mapper* was modified to include a 3'→5' trimming and two step alignment regime (see: Figure 3.2A). During step (ii) (Processing and

Filtering) (see: Section 3.3), any read pair for which only a single mate successfully mapped undergoes 3'→5' trimming of the unmapped mate to a user specified amount. Trimming is performed under the theory that, should an INDEL, for example, reside toward the middle or end of a read, 3'→5' trimming may either remove the corresponding read segment or lessen map quality penalties. Trimmed Read-1 and Read-2 FASTQ files are subsequently reconstructed to contain both mates for secondary alignment. Secondary alignment is performed by *Bowtie2* under *local* mode—a less strict process than *end-to-end*—and resulting reads are reprocessed. Any read pairs where both mates now fully map are called and appended onto the main dataset. Ambiguous end filtering still operates, ensuring coordinate integrity is not compromised.

In order to assess the impact of two step alignment and determine an optimal level of 3'→5' trimming, two *sae2*Δ single cut libraries were processed by *Spo11Mapper* at progressively increased levels of trimming (Figure 3.2C). *Local* alignment (0bp trimming) alone is sufficient to produce an additional ~1% of data owing to internal read trimming performed by *Bowtie2*. Trimming of 10bp from the 3' end maximises the amount of salvaged information (~1.2%) while further trimming progressively reduces mappability. The majority of extra reads obtained (97.4%) fall within known, annotated hotspots and collectively cluster at ~22 genomic loci, all of which contain at least one INDEL. As visualised, trimmed reads fill in strand specific gaps present in untrimmed datasets (Figure 3.3A-C) or add hits to pre-existing peaks (Figure 3.3D-F). These observations collectively suggest that moderate read trimming of 10bp from the 3' end in conjunction with a two step alignment regime constructs a more complete picture of DSB formation.

3.5—Sampling & Processing

In order to assess the validity of *sae2*Δ mapping and investigate the impact Tel1^{ATM} inactivation may have upon DSB formation, single cut Spo11 DSBs libraries, generated using the previously described assay (see: Section 3.2), from *sae2*Δ, *sae2*Δ*tel1*Δ, *sae2*Δ*tel1KD* (kinase dead), *sae2*Δ*ndt80*Δ and *sae2*Δ*ndt80*Δ*tel1*Δ cells were aligned under the newly devised default settings of *Spo11Mapper* (two step alignment, 10bp 3'→5' trim) and processed (Table 3.1).

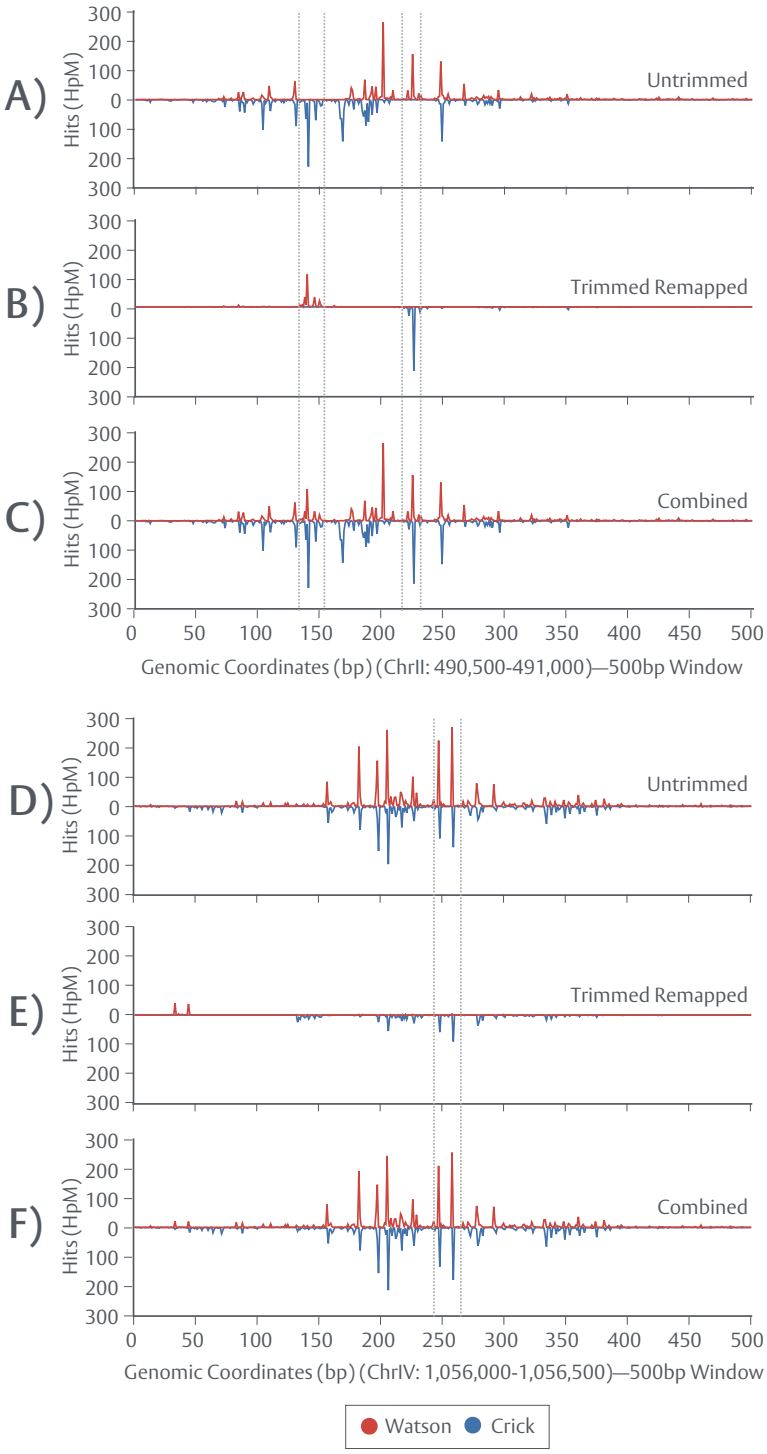


Figure 3.3. Limited 3'→5' trimming of reads improves mapping within polymorphic regions

Data from *sae2*Δ_2A was processed via *Spo11Mapper*, using an optimal 3'→5' trimming level of 10bp, and 1bp histograms were subsequently visualised on both strands for two hotspot regions (on ChrII and ChrIV). Coordinate (x) values are shown relative to the extracted region (500bp window). **A/D)** Untrimmed, high quality and unambiguous reads obtained during first round, global alignment **B/E)** Trimmed, high quality and unambiguous reads obtained during second round, local alignment. **C/F)** Combined datasets. Lines highlight regions complemented by trimming.

Genotype	Pairs (A)	Mapped Pairs (B)	% of (A)	MM Pairs	% of (B)	Valid Hits	% of (B)	Ambig Hits	% of (B)
*sae2*Δ_1	3,512,089	3,402,853	96.89	184,676	5.427	3,376,630	99.229	25,234	0.742
*sae2*Δ_2A	10,101,423	8,849,152	87.60	275,888	3.118	8,796,998	99.411	50,583	0.572
*sae2*Δ_3	4,957,477	4,781,381	96.45	253,989	5.312	4,745,192	99.243	34,889	0.73
*sae2*Δ_5	4,012,972	3,798,655	94.66	163,952	4.316	3,705,100	97.537	92,604	2.438
*sae2*Δ*tel1*Δ_1A	4,425,101	4,207,321	95.08	237,355	5.641	4,164,249	98.976	41,798	0.993
*sae2*Δ*tel1*Δ_1B	6,053,495	5,768,099	95.29	324,034	5.618	5,708,444	98.966	57,860	1.003
*sae2*Δ*tel1*Δ_2	4,957,735	4,783,672	96.49	297,322	6.215	4,738,165	99.049	43,904	0.918
*sae2*Δ*tel1*Δ_3	4,175,752	4,032,206	96.56	252,741	6.268	4,001,366	99.235	29,401	0.729
*sae2*Δ*tel1*Δ_4	4,921,201	4,746,862	96.46	315,574	6.648	4,709,175	99.206	35,484	0.748
*sae2*Δ*tel1KD*_1A	3,744,355	3,563,609	95.17	216,168	6.066	3,530,601	99.074	31,767	0.891
*sae2*Δ*tel1KD*_1B	5,168,646	4,934,110	95.46	300,059	6.081	4,888,107	99.068	44,179	0.895
*sae2*Δ*tel1KD*_2	4,689,990	4,506,569	96.09	421,543	9.354	4,444,032	98.612	60,064	1.333
*sae2*Δ*ndt80*Δ_1	6,582,256	6,358,954	96.61	555,249	8.732	6,232,107	98.005	122,821	1.931
*sae2*Δ*ndt80*Δ_2	5,566,483	5,414,016	97.26	186,939	3.453	5,379,781	99.368	32,266	0.596
*sae2*Δ*ndt80*Δ*tel1*Δ_1	5,813,562	5,637,125	96.97	174,019	3.087	5,577,623	98.944	58,715	1.042
*sae2*Δ*ndt80*Δ*tel1*Δ_2	4,888,938	4,752,911	97.22	157,338	3.31	4,726,808	99.451	25,151	0.529

Table 3.1. Spo11 DSB libraries processed by Spo11Mapper

Paired end, single cut libraries were sequenced (MiSeq, 2 x 75bp reads) for *sae2*Δ (4 repeats), *sae2*Δ*tel1*Δ (5 repeats), *sae2*Δ*tel1KD* (kinase dead) (3 repeats), *sae2*Δ*ndt80*Δ (2 repeats) and *sae2*Δ*ndt80*Δ*tel1*Δ (2 repeats), aligned against *Cer3H4L2* and processed via *Spo11Mapper* (3'→5' trimming = 10bp). All samples were taken at a 6h meiotic time point. Alignment and call stats, generated by *Spo11Mapper*, are tabulated. Pairs (A) denotes the raw number of read pairs present in FASTQ files. Mapped Pairs (B) details the total number of reads aligned, including multi-mappers. Multi-mapping (MM) reads are defined, by *Bowtie2*, as reads which may map to >1 loci with equal probability. Under such conditions, *Bowtie2* portions these reads between the mappable loci at random. Valid Hits denotes the number of combined, unambiguous 5' ends called from untrimmed and trimmed alignment. Ambig Hits details the number of ambiguous reads that contained >1SNP or >1bp of INDEL at the informative, 5' end.

An average library produces 5.19m read pairs—of which 4.94m (95.2%) pairs successfully align. Of these mapped pairs: (i) 5.67% constitute multi-mappers (ii) 1.03% fail the ambiguous end filter. Due to the inclusion of *HIS4*::*LEU2* and *LEU2*::*HISG* alongside the endogenous LEU2 locus in the reference genome, as well as the presence of other repeat regions, multi-mapping reads are retained in the main dataset and shared equally amongst the possible loci. Overall, an average library of 5.19m read pairs produces 4.89m (94.2%) valid 5' Read-1 Spo11 hits, signifying high alignment rates. A full strain table is available (see: Table 3.3—Appendix).

3.6—Positional correlation with Spo11-oligo mapping

In order to compare the Spo11 DSB maps produced by the *sae2*Δ and Spo11-oligo (Pan et al. 2011) methods, 1bp histogram data from each method was visualised on a per chromosome basis (Figure 3.4A-H). Despite quantitative differences between the datasets, all significant signal resides at similar genomic locations, suggesting *sae2*Δ mapping detects legitimate Spo11 DSBs and that *Spo11Mapper* accurately processes *sae2*Δ data. Inspection of higher resolution data (200bp window) reveals several key similarities and differences between the methods: (i) the positional agreement between *sae2*Δ (Figure 3.4I) and Spo11-oligo (Figure 3.4J) data extends to the 1bp level, as evidenced by the precise alignment of major peaks (ii) a 2bp Watson (+)-Crick (-) offset, previously observed and predicted from the homodimeric activity of Spo11 (Liu et al. 1995; Pan et al. 2011), is recaptured by *sae2*Δ mapping (iii) *sae2*Δ mapping appears to result in broader domains of signal relative to Spo11-oligo mapping, possibly reflecting the loss of short ~10-15bp oligonucleotides in the latter technique (see: Section 3.1).

Spo11-oligo mapping previously identified 3599 regions of clustered DSB formation, termed DSB hotspots (Pan et al. 2011). Of these, *sae2*Δ mapping detects 3576 (99.36%) and 3384 (94.02%) hotspots with 2-fold and 5-fold enrichment over background respectively. Given the high degree of positional correlation between mapping techniques, all *sae2*Δ hits residing outside of hotspot regions were excluded for any direct comparative, quantitative analyses (as in Section 3.7).

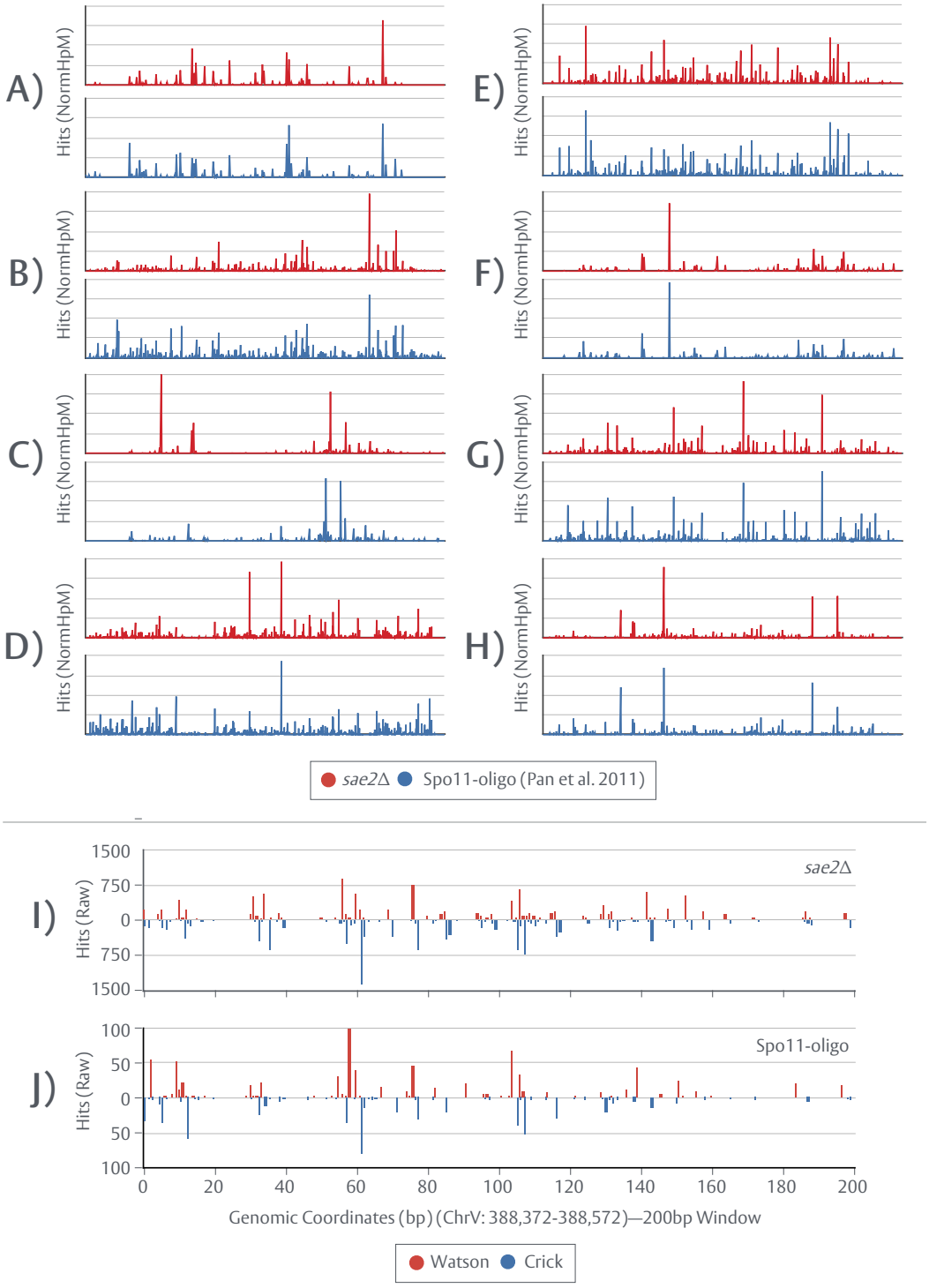


Figure 3.4. Positional correlation with Spo11-oligo mapping

1bp histogram data for *sae2*Δ_2A, generated via *Spo11Mapper*, was visualised (red) alongside equivalent WT Spo11-oligo data (blue) (Pan et al. 2011) for **A)** ChrI **B)** ChrII **C)** ChrIII **D)** ChrIV **E)** ChrV **F)** ChrVI **G)** ChrVII **H)** ChrVIII. (x) and (y) values along with ChrIX-XVI are omitted for clarity. Data from each strand was combined to form singular peaks. **I-J)** 1bp histogram data for *sae2*Δ_2A was visualised for both strands within a high resolution, 200bp window on ChrV and compared to equivalent WT Spo11-oligo data (Pan et al. 2011).

3.7—Quantitative reproducibility and correlation with Spo11-oligo mapping

As with any genome-wide mapping technique, contaminating non-specific DNA may introduce background into *sae2*Δ Spo11 libraries. Background reads decrease signal:noise ratios and preclude accurate cross comparisons between samples. Intragenic Spo11 hits are extremely rare (Pan et al. 2011) and thus, in order to estimate and correct for background, read density across 47 >5.5kb ORFs was calculated, excluding hits within 1kb of the ORF start or end. A typical *sae2*Δ library is estimated to possess background levels of 0.0043-0.0086 hits/bp/million reads (5-10% of reads)—a moderately higher level than Spo11 oligo-nucleotide mapping (~4% of reads), suggesting *sae2*Δ mapping has a relatively lower specificity for Spo11 breaks. Calculated background levels were subsequently used to correct the collective strength of each annotated hotspot by estimating the proportion of background reads contained within. Normalisation, to account for differential sample size, was conducted on a per sample basis, producing normalised values per million reads (NormHpM).

To determine the quantitative reproducibility of *sae2*Δ biological repeats, the collective strength of each annotated hotspot was compared and plotted for all *sae2*Δ samples (Figure 3.5A-F). Pearson's rho (ρ), a measure of linear correlation, is employed throughout this chapter. The degree of correlation is reflected in the value of (ρ) on a scale [0-1.0], while the sign of (ρ) denotes either a positive (+) or negative correlation (-). Pairwise comparisons of all *sae2*Δ repeats yield (ρ) values of >0.96, signifying strong, quantitative correlation between each dataset and demonstrating a high degree of reproducibility amongst repeat samples. Strong correlations are observed across the spectrum of hotspot strength, including at the lower end—indicating a high dynamic range for *sae2*Δ mapping. Given this level of reproducibility, *sae2*Δ repeat data was combined into an averaged dataset and compared to equivalent WT Spo11-oligo data (Figure 3.6A). While a medium-strong correlation ($\rho = 0.8409$) is observed between *sae2*Δ and Spo11-oligo data, significant differences appear to exist—particularly at the lower end where hotspots may differ ~5-10-fold in strength.

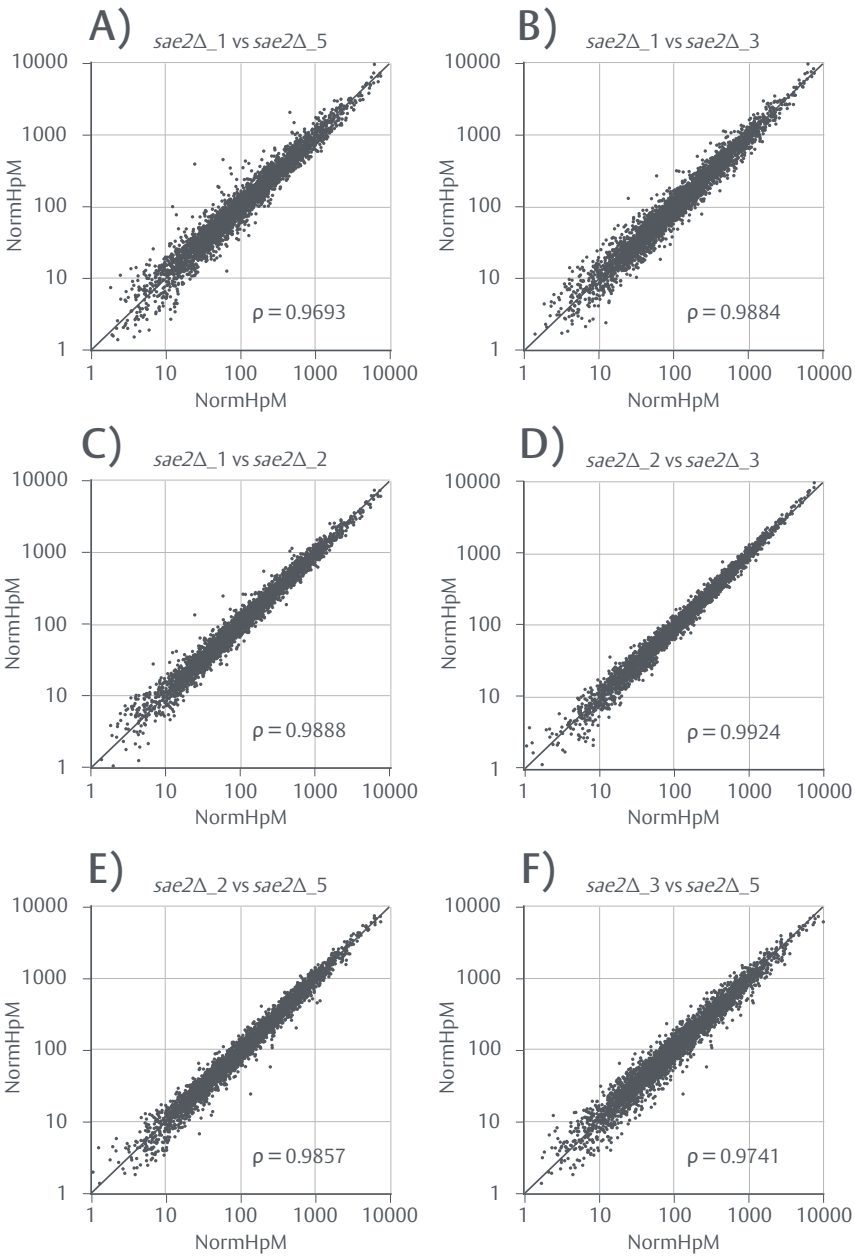


Figure 3.5. Individual repeat Spo11 DSB libraries are quantitatively well correlated

Spo11 5' hits, from *sae2*Δ libraries, were tallied across 3599 previously annotated *S. cerevisiae* hotspots (Pan et al. 2011) and normalised to account for calculated background levels and read count on a per library basis (NormHpM—hits per million). All non-hotspot hits were discarded. The quantitative strength of each hotspot was compared for all unique *sae2*Δ combinations: **A)** *sae2*Δ_1, *sae2*Δ_5 **B)** *sae2*Δ_1, *sae2*Δ_3 **C)** *sae2*Δ_1, *sae2*Δ_2 **D)** *sae2*Δ_2, *sae2*Δ_3 **E)** *sae2*Δ_2, *sae2*Δ_5 and **F)** *sae2*Δ_3, *sae2*Δ_5. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values).

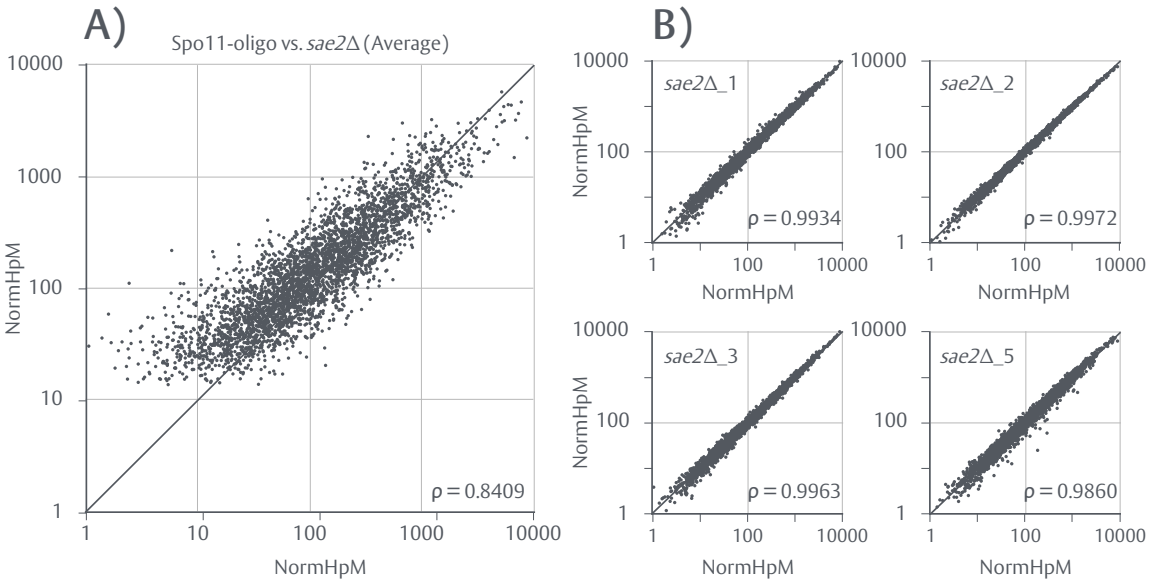


Figure 3.6. Quantitative correlation with Spo11-oligo mapping

Normalised data (NormHpM) from *sae2*Δ libraries was averaged for each annotated hotspot (4 repeats), yielding a single, aggregated dataset. **A)** The quantitative strength of each averaged *sae2*Δ hotspot was compared to equivalent WT data from Spo11-oligo mapping (Pan et al. 2011). **B)** The quantitative strength of each hotspot was compared for all individual *sae2*Δ repeats (as marked) against the *sae2*Δ average. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values).

Any quantitative discrepancies between the methods may reflect variations in the filter thresholds applied however a biological difference between *sae2*Δ and truly WT strains cannot be ruled out. Notably, the *sae2*Δ average represents each individual repeat well ($\rho > 0.98$) (Figure 3.6B) and was thus taken forward for all further analyses involving annotated hotspot data.

3.8—Disproportionate formation of DSBs on smaller chromosomes is HR-independent

Historically, Spo11 DSB mapping studies have observed a negative correlation between hit density (hits/bp) and chromosomal size (Blitzblau et al. 2007; Gerton et al. 2000; Pan et al. 2001; Martini et al. 2006). Synapsis-dependent and Zip3-mediated shutdown (SDS) of DSB formation has been proposed to account for this relationship (Thacker et al. 2014)—whereby larger chromosomes have an increased chance of engaging the correct homolog earlier in meiosis. In contrast, smaller chromosomes are afforded extra opportunity to form additional DSBs before shutdown occurs. Synapsis is, however, dependent upon homologous recombination (HR) within *S. cerevisiae* and therefore should be absent in the end processing deficient mutant, *sae2*Δ.

In order to assess the relationship between signal and chromosomal length, total hits (NormHpM) and hit density (NormHpM/bp) were calculated for *sae2*Δ and Spo11-oligo WT datasets and plotted against chromosome size (Figure 3.7). As expected, a strong, positive correlation is observed between total hits and chromosome size for both datasets ($\rho > 0.93$) (Figure 3.7A). However, for unknown reasons, ChrXII forms significantly fewer hits than expected specifically within a *sae2*Δ background, and was thus excluded from consideration. While a negative correlation between hit density and chromosomal size is observed for Spo11-oligo data ($\rho = -0.6344$), it is noticeably weakened or absent within *sae2*Δ ($\rho = -0.1283$) (Figure 3.7B)—consistent with the suggestion that an end processing or HR-dependent process, such as SDS, gives rise to this negative correlation. Nevertheless, disproportionate formation of DSBs on the smaller chromosomes is retained in a *sae2*Δ background and occurs at similar levels to that observed by Spo11-oligo mapping (see: Figure 3.7B).

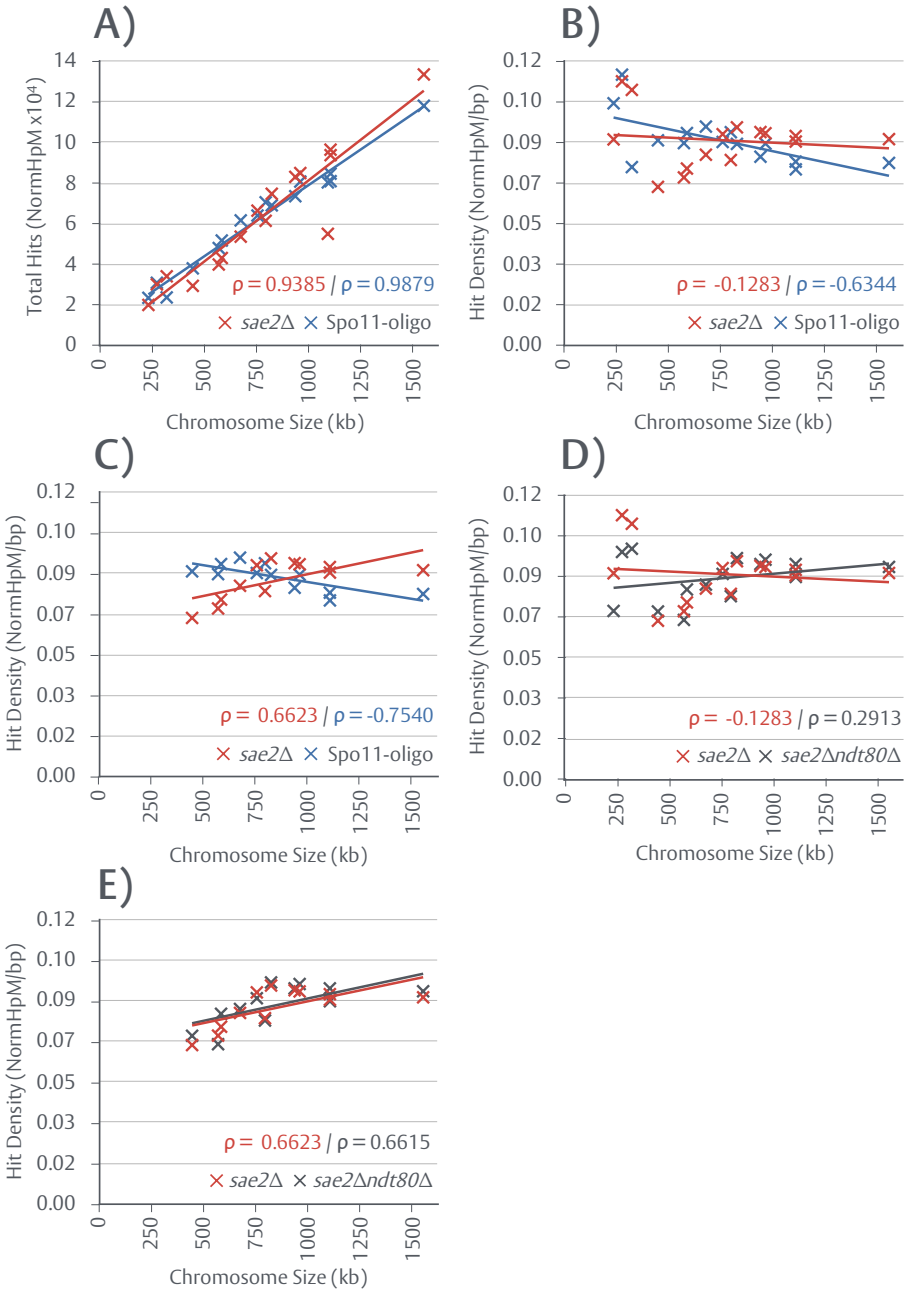


Figure 3.7. Distribution of Spo11 DSBs at the chromosomal level

A) Normalised data (NormHpM) was summed across each chromosome (total hits) for *sae2*Δ and Spo11-oligo WT data (Pan et al. 2011) and plotted against *S. cerevisiae* chromosome size. **B)** NormHpM values, summed across each chromosome, were converted to hit densities (NormHpM/bp) **C)** Hit densities for smaller <400kb chromosomes (ChrI, ChrIII, ChrVI) were omitted (*sae2*Δ vs. Spo11-oligo comparison) **D)** Equivalent hit density data for *sae2*Δ*ndt80*Δ was calculated and compared to *sae2*Δ **E)** Hit densities for smaller <400kb chromosomes (ChrI, ChrIII, ChrVI) were omitted (*sae2*Δ vs. *sae2*Δ*ndt80*Δ comparison). Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values). Linear trendlines are marked onto each plot.

In order to remove the influence smaller chromosomes may have on the overall trend, hit density was replotted for all >400kb chromosomes (Figure 3.7C). Unexpectedly and in direct contrast to Spo11-oligo data, a medium-strong *positive* correlation between hit density and >400kb chromosomal length is observed for *sae2*Δ data ($\rho = 0.6623$). Within *sae2*Δ backgrounds, the length of prophase I may be altered owing to a weakened ssDNA-dependent checkpoint signal. A shortened or lengthened window of DSB formation may in turn impact the distribution of DSBs at the chromosomal level. In order to assess this possibility and account for any impact prophase I length may have, *sae2*Δ*ndt80*Δ data—within which prophase I exit is abolished (Xu et al. 1995; Winter 2012; Allers & Lichten 2001)—was compared to *sae2*Δ in an identical manner. While the disproportionate formation of DSBs on smaller chromosomes is less apparent within *sae2*Δ*ndt80*Δ (Figure 3.7D), the positive correlation between hit density and >400kb chromosome size is near identical (Figure 3.7E).

Collectively, these observations suggest that: (i) as expected, synapsis-dependent shutdown (SDS) of DSB formation is lost within *sae2*Δ backgrounds (ii) disproportionate formation of DSBs on smaller chromosomes (ChrI, III and VI) occurs via mechanisms independent of SDS and (iii) in the absence of SDS, the larger chromosomes unexpectedly form DSBs at densities higher than expected from their size alone.

3.9—Spo11 DSBs preferentially form within nucleosome depleted promoter regions

As per previous studies, Spo11-dependent DSB formation exhibits a strong preference toward nucleosome depleted, promoter regions (Baudat & Nicolas 1997; Gerton et al. 2000; Pan et al. 2011). In order to determine whether or not this preference is recaptured by *sae2*Δ mapping, hit densities (hits/kb) were calculated for several types of intergenic region (IGR) (tandem, convergent, divergent) and all annotated ORFs (genic) (Figure 3.8A). As a point of comparison, Spo11-oligo WT data was similarly analysed (Figure 3.8B). Tandem IGRs reside between identically oriented genes, and harbour a single promoter.

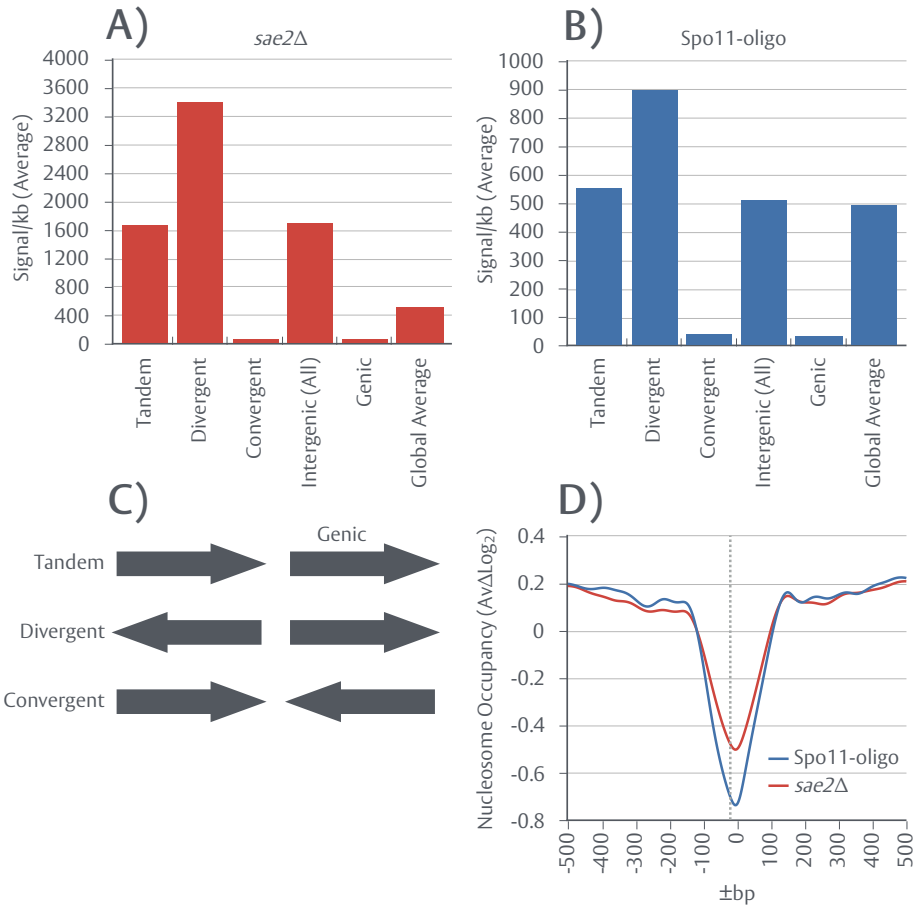


Figure 3.8. Spo11 DSBs preferentially form within nucleosome depleted promoter regions

Normalised Spo11 5' hits were summed across varying types of intergenic region (IGR) (tandem, divergent, convergent) and each annotated ORF (genic). Hit counts were subsequently converted to averaged densities (signal/kb) using the total, cumulative length of each region type for **A)** *sae2*Δ data and **B)** Spo11-oligo data. Global averages are calculated under the assumption that all detected signal is evenly spread across the genome. **C)** Schematic showing the different types of IGR analysed. **D)** Normalised nucleosomal occupancy, expressed as a logarithmic deviation from the genome-wide average ($Av\Delta Log_2$) and derived from (Kaplan et al. 2008), flanking each significant Spo11 DSB signal peak (>0.5HpM), for *sae2*Δ and Spo11-oligo data, was extracted, piled up, subsequently averaged and smoothed (moving average, n=5).

Divergent IGRs reside between oppositely oriented, outward facing genes and harbour two promoters. Convergent IGRs reside between oppositely oriented, inward facing genes and contain terminator regions (Figure 3.8C).

Spo11 DSBs show a near identical distribution regardless of mapping method, characterised by (i) a significant enrichment of signal within promoter-containing intergenic regions (IGRs) (divergent, tandem) proportional to the number of promoters present (ii) a depletion of signal within genic regions and promoter-less IGRs (convergent). To investigate the local chromatin environment of DSB formation, normalised nucleosomal occupancy, derived from (Kaplan et al. 2009), was piled up centred on all major Spo11 DSB peaks (>0.5HpM) for both datasets (Figure 3.8D). Nucleosomal occupancy is expressed as a logarithmic deviation from the genome-wide average ($Av\Delta Log_2$). Regardless of mapping method, Spo11 DSBs form within ~180-200bp regions exhibiting nucleosomal occupancy levels significantly below the genome-wide average (-0.5-0.8 $Av\Delta Log_2$). Given the widespread presence of nucleosomal depletion within terminator regions (convergent IGRs), these findings corroborate previous observations that an open chromatin structure, in isolation, is insufficient for DSB formation and that additional, promoter associated factors are also required (see: Section 1.4.1) (Cooper et al. 2016; de Massy 2013).

3.10—Mapping of Spo11 DSBs reveals a weak sequence bias with rotational symmetry

A weak, preferential sequence bias for Spo11 cleavage has been previously observed (Pan et al. 2011). However, this was calculated using Spo11-oligo mapping data. Due to the poly(G)-tailing of Spo11-oligonucleotides during library prep and the subsequent ambiguity in 5' C-residues, a blurred and asymmetric bias was obtained. Moreover, the obtained bias may prove incomplete given the loss of shorter oligos (see: Section 3.1) (Pan et al. 2011). To facilitate investigations of sequence bias, a novel utility script (*SeqBias*) was developed to be compatible with 1bp histograms, produced by *Spo11Mapper* (see: Section B3.1.8). *SeqBias* directly samples a user provided FASTA genomic reference file (e.g. Cer3H4L2) to pileup and calculate per base frequencies for A/G/C/T, centred on all listed coordinates (e.g. 5' Spo11-hits). Such an approach, as opposed to direct read pileup, more

accurately takes into account the presence of SNPs/INDELs and benefits from filters applied during alignment or processing.

In order to determine the sequence bias of *sae2*Δ Spo11 DSBs, *SeqBias* was utilised to calculate the base frequencies ±20bp surrounding around all *sae2*Δ 5' coordinates—revealing an extremely weak, asymmetrical bias (Figure 3.9A). Base pair frequencies are expressed as logarithmic deviations from the expected global average (i.e. G 0.19 or A 0.31). A relative bias of [0.0] therefore represents the expected frequency for any given base. Filtering of signal (>0.5HpM), to enrich for legitimate peaks, markedly improves the sequence bias (Figure 3.9B), however, partial asymmetry still remains—a feature not expected for an enzyme functioning as a homodimer (Sasanuma et al. 2007). Upon close inspection of the data, it was discovered that a subset of non-cognate peaks exist—that is, significant signal on Watson (+) or Crick (-) without a corresponding peak on the opposing strand, 2bp away. To further refine the obtained sequence bias, *sae2*Δ libraries (>0.5HpM) were thus filtered into cognate and non-cognate subpopulations and re-analysed via *SeqBias*. Cognate peaks, defined as peaks with corresponding, offset signal and which are expected to reflect legitimate DSBs, display a perfectly symmetrical , rotational (palindromic) bias (Figure 3.9C).

By contrast, non-cognate peaks exhibit a Spo11-like but heavily perturbed bias (Figure 3.9D). While non-cognate peaks constitute, on average, ~30% of the mapped data, their origin remains unclear. As the *sae2*Δ mapping method is technically specific to *any* protein:DNA complex, contamination may explain the presence of non-cognate signal. However, 54.2% of filtered (>0.5HpM) non-cognate peaks have corresponding peaks within Spo11-oligo datasets—a method that utilises immunoprecipitation of Spo11—suggesting non-cognate peaks may, instead, reflect an alternative chemistry or activity of Spo11. With the dyad-axis set to position (0), a Spo11 DSB bias half site thus comprises significant C enrichment at position -1, with flanking A-enrichment at (-3,-4). In full, the top strand sequence bias may be read as AAGC*A|TGCTT with the dyad axis and cleavage site denoted as | and * respectively.

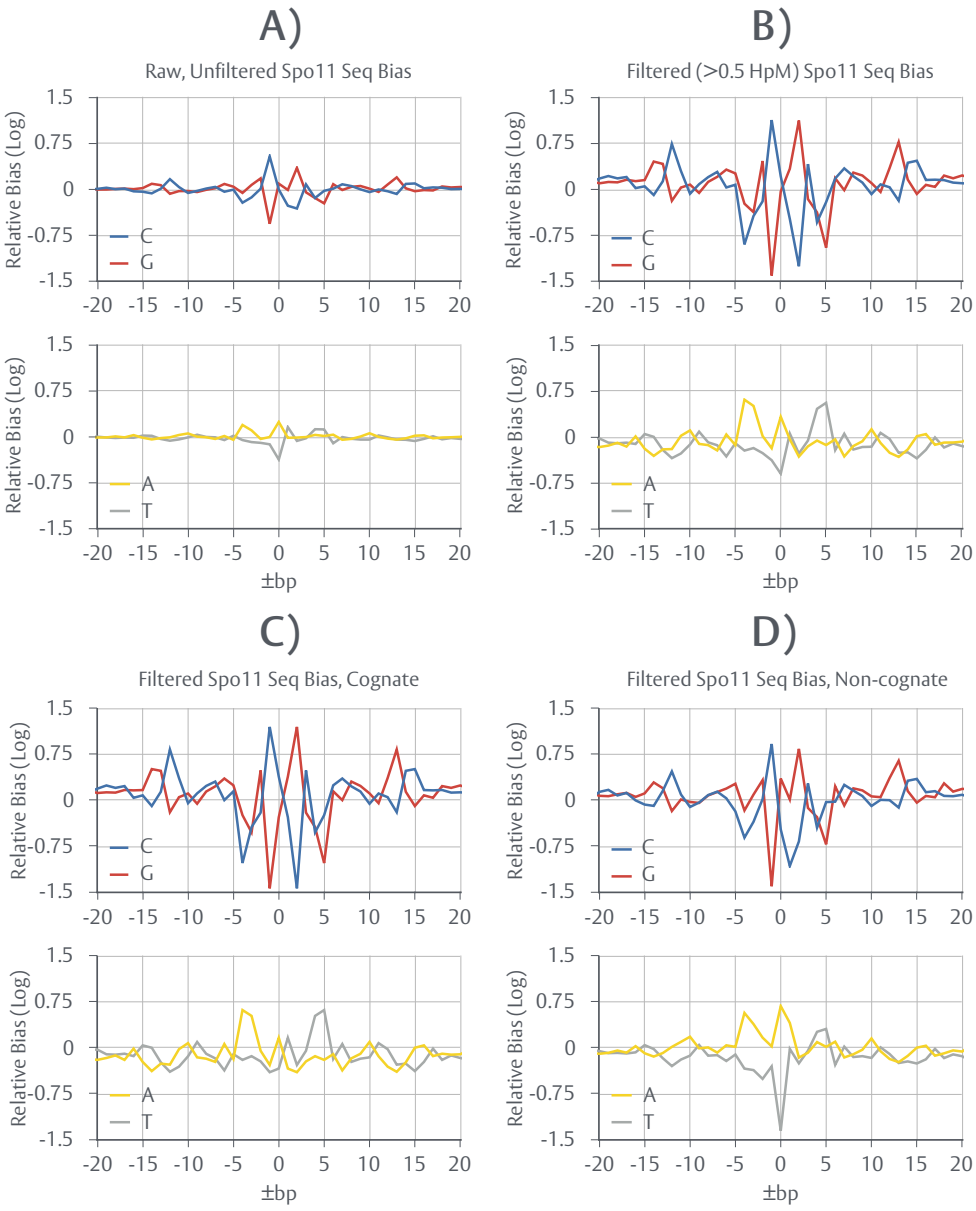


Figure 3.9. A weak, symmetrical Spo11 sequence bias for the generation of DSBs

A) Base frequencies (% of A/G/C/T) were calculated, via *SeqBias*, flanking (±20bp) all *sae2*Δ Spo11 5' hits on both strands. Averaged base compositions are expressed as a logarithmic (log_2 base) bias relative to expected frequencies for *S. cerevisiae* (38.1% GC richness). A relative bias of [0.0] represents the expected frequency for a given base. A relative bias of [1.0] would represent a two-fold enrichment over expectation. **B)** Base frequencies were calculated, via *SeqBias*, flanking (±20bp) filtered (>0.5HpM) *sae2*Δ Spo11 5' hits on both strands. **C)** Filtered *sae2*Δ data was further partitioned into cognate and non-cognate subpopulations. Non-cognate peaks were defined as any >0.5HpM peak without a >0.5HpM peak 2bp away on the opposing strand. Base frequencies were calculated, via *SeqBias*, flanking (±20bp) filtered, cognate (>0.5HpM) *sae2*Δ peaks. **D)** Base frequencies were calculated, via *SeqBias*, flanking (±20bp) filtered, non-cognate (>0.5HpM) *sae2*Δ peaks.

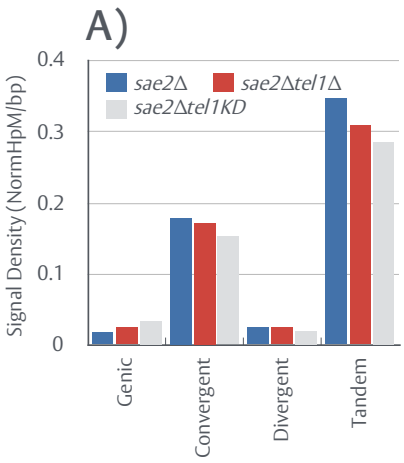
3.11—Tel1^{ATM} is required for WT-like suppression of DSB formation within genic regions

As previously noted (see: Section 3.1), the DNA damage response (DDR) kinase, Tel1^{ATM}, is closely involved in the regulation of DSB formation (Cooper et al. 2014). In order to determine whether or not Tel1 inactivation alters DSB distributions in a generalised sense, *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* (kinase dead) hit densities (hits/kb) were calculated for several types of genomic loci (tandem, convergent, divergent and genic) and compared to *sae2*Δ (Figure 3.10A) (see: Figure 3.8C). Unexpectedly, suppression of intragenic DSB formation is weakened by Tel1 inactivation. Specifically, intragenic signal density is increased 38.81% and 81.73% within *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* relative to *sae2*Δ respectively (Figure 3.10B)—a shift that appears to indiscriminately occur at the expense of tandem and divergent signal density. Moreover, loss of suppression is considerably more pronounced within *sae2*Δ*tel1KD* than *sae2*Δ*tel1*Δ, suggesting a dominant negative effect of the *tel1KD* allele.

3.12—Inactivation of Tel1^{ATM} causes a "spreading" of Spo11 DSB signal

Introduction of the *tel1KD* allele into a *sae2*Δ background has previously been shown to cause a "spreading" of southern blot signal—as primarily assessed at the *HIS4*::*LEU2* hotspot—specifically in the direction of the adjacent ORF (D. Johnson, V. Garcia, M.J. Neale unpublished). An equivalent smear is not observed within *sae2*Δ*tel1*Δ. While smearing of DSB bands may be caused by resection, *sae2*Δ strains are resection deficient and thus the cause of this smear remains unclear.

In order to assess whether or not this phenomenon is related to the derepression of intragenic DSB formation, observed within genome-wide data (see: Section 3.11), averaged data from *sae2*Δ, *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* single cut libraries was visualised and compared for two strong hotspots (*ARE1* and *ERG25*) (Figure 3.11A,B respectively). Consistent with observations at *HIS4*::*LEU2*, spreading of the signal, primarily in the direction of the flanking ORFs, is readily apparent in *sae2*Δ*tel1KD* but also to a lesser extent within *sae2*Δ*tel1*Δ.



B)

Type	*sae2Δtel1Δ*	*sae2Δtel1KD*
Genic	+38.81%	+81.73%
Convergent	+3.80%	-12.03%
Tandem	-4.00%	-14.46%
Divergent	-11.40%	-17.73%

Figure 3.10. Tel1^{ATM} is required for WT-like suppression of DSB formation within genic regions

A) Normalised Spo11 5' hits were summed across varying types of intergenic region (IGR) (tandem, divergent, convergent) and each annotated ORF (genic). Hit counts were subsequently converted to averaged densities (signal/kb) using the total, cumulative length of each region type for *sae2*Δ, *sae2*Δ*tel1*Δ and *sae2*Δ*telKD*.

B) Percentages differences, relative to *sae2*Δ densities, were calculated for each region type.

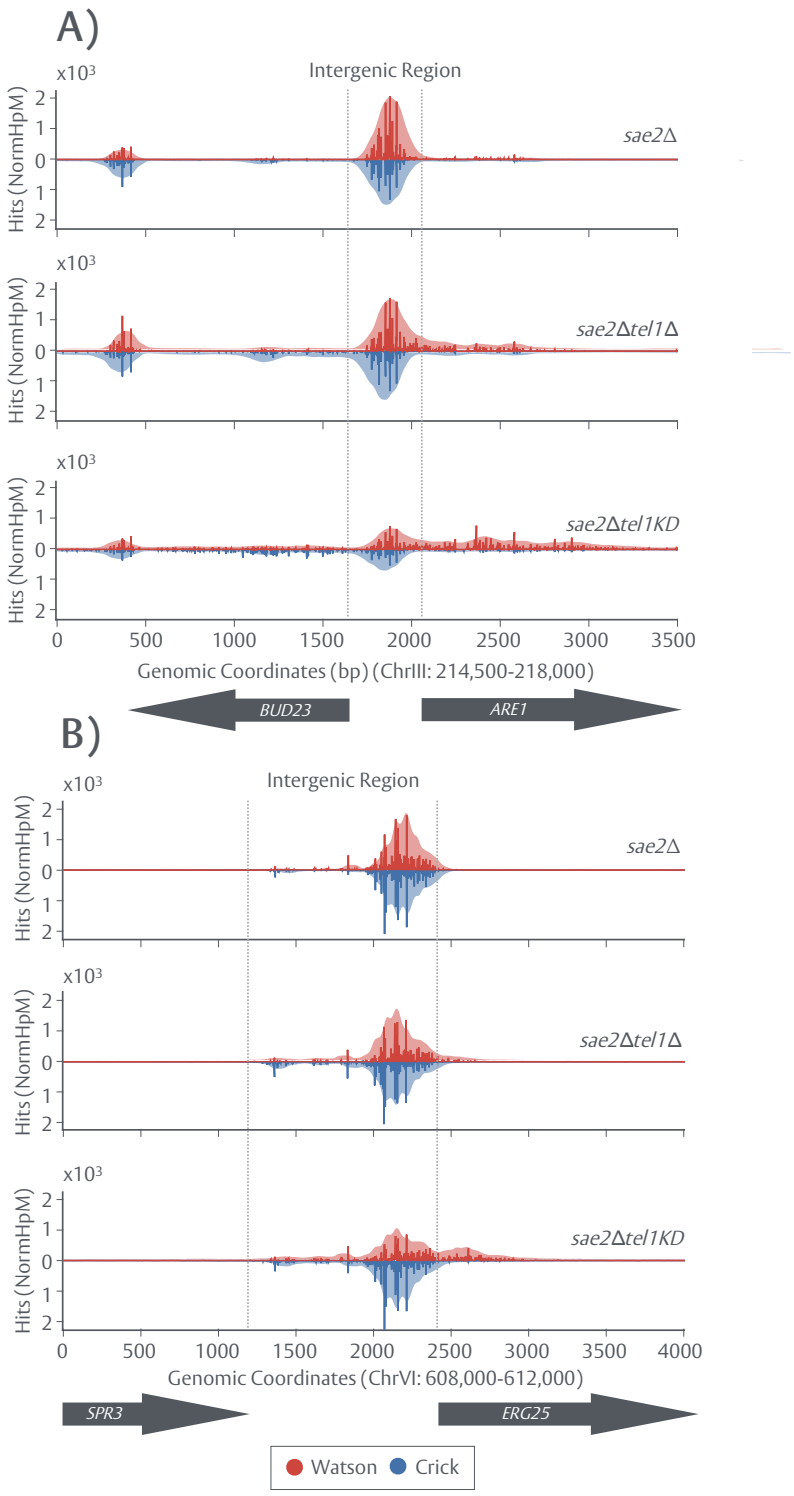


Figure 3.11. Inactivation of Tel1^{ATM} causes a "spreading" of Spo11 DSB signal

Normalised 1bp histogram datasets for *sae2*Δ, *sae2*Δ*tel1*Δ and *sae2*Δ*telKD*, generated via *Spo11Mapper*, were visualised on both strands for two strong hotspots: **A)** *ARE1* (3500bp window) and **B)** *ERG25* (4000bp window). Coordinate (x) values are shown relative to the extracted region. Data was additionally smoothed (moving average, n = 50bp) and overlaid (transparent areas). The orientation of adjacent genes are shown below each plot.

To determine if this spreading occurs genome-wide, Spo11 DSB signal from all three backgrounds was piled up ±1000kb around the start position of every annotated *S. cerevisiae* ORF and subsequently averaged (Figure 3.12). While the majority of signal resides in the upstream promoter region as expected (see: Figure 3.10A), some level of spreading into the ORF is observed in all three backgrounds—primarily confined to the 5' end of the gene. Importantly, and reminiscent of increases in genic DSB formation, spreading is increased approximately ~2-fold and ~3.4-fold within *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* respectively, relative to *sae2*Δ.

Collectively these observations suggest that (i) increased intragenic signal within *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* occurs via a spreading of DSB formation into the 5' portion of adjacent ORFs, as opposed to ORF wide derepression and (ii) Tel1 activity is ordinarily required to repress spreading. Moreover, given an enhancement of these phenotypes within *sae2*Δ*tel1KD*, Tel1 may act alongside redundant pathways to suppress genic DSB formation—pathways now blocked by the presence of a dominant negative, kinase dead protein.

3.13—Long range (>100bp) 10bp periodicity within Tel1 mutants

Within *tel1*Δ mutants, Spo11-oligos display a prominent ~10bp periodicity in size, occurring at discrete bands (33/43/53/63bp), above those of the standard Spo11 oligos, when assessed by gel (Mohibullah & Keeney 2017). 10bp periodicity is also visible, to a lesser extent, within WT and *sae2*Δ backgrounds (Mohibullah & Keeney 2017; D. Johnson, M.J. Neale unpublished). Crucially, these periodic upper bands are Mre11-independent (D. Johnson, M.J. Neale unpublished), suggesting they may constitute molecules released solely by Spo11 activity. In other words, such periodicity may be the result of Spo11 double cutting—whereby Spo11, rather than Mre11, cleaves the 3' end.

In order to further investigate this hypothesis, double cut *sae2*Δ, *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* libraries, as prepared by (D. Johnson), were processed via *Spo11Mapper* (see Section 3.2 and Figure 3.1C).

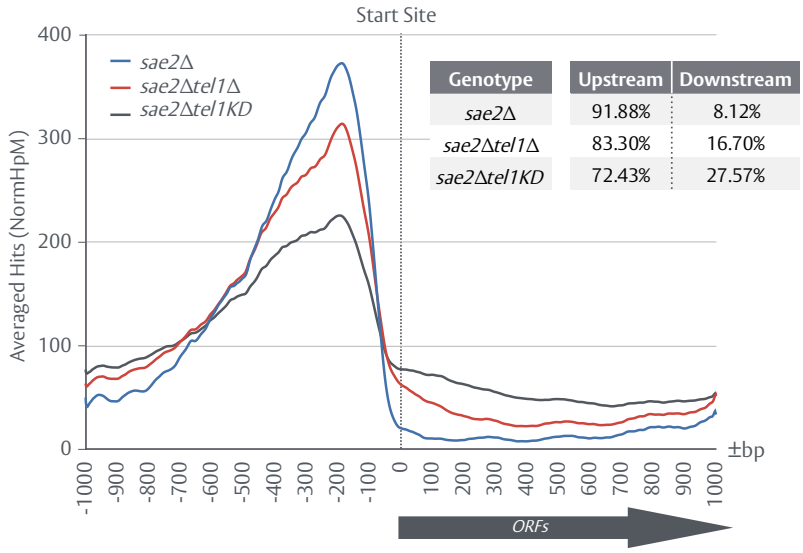


Figure 3.12. Genome-wide spreading of Spo11 DSBs occurs in the direction of transcription

Normalised Spo11 5' hits for *sae2*Δ, *sae2*Δ*tel1*Δ and *sae2*Δ*telKD* were piled up centred on the start site of all annotated *S. cerevisiae* ORFs, averaged and smoothed (moving average, n = 50bp). The orientation of the collective, piled up ORF is marked below the plot. All (-) strand ORFs were reversed in orientation for visual clarity. The fraction of averaged signal residing upstream and downstream of the start site was additionally quantified for all backgrounds.

Molecule sizes—the distance between both, called 5' ends—were tallied, smoothed (moving average) and visualised at a zoomed in range of 0-225bp for single cut (Figure 3.13A-C) and double cut libraries (Figure 3.13D-F). WT samples appear devoid of any obvious periodicity regardless of the library type. In contrast, a clear 10bp periodicity is observed within *sae2*Δ*tel1*Δ and, to higher frequencies, within *sae2*Δ*tel1KD* above molecule sizes of ~100-125bp. Importantly, this periodicity is dependent upon use of the double cut protocol and is not present in single cut libraries. Furthermore, exacerbation of this phenotype within *sae2*Δ*tel1KD* over *sae2*Δ*tel1*Δ implies that spreading of DSB formation into genic regions (see: Section 3.12) may occur through long range double cutting. Despite the presence of 10bp periodic molecules, no periodicity in the size ranges (33-63bp) previously observed is apparent—potentially owing to size selection or technical issues of the *sae2*Δ method, resulting in very little <100bp data.

3.14—Short range (<75bp) 10bp periodicity is a WT phenomenon

To further characterise the phenomenon of 10bp periodicity in the lower 33-63bp range, previously published WT, *tel1*Δ and *tel1KD* Spo11-oligo libraries (Mohibullah & Keeney 2017) were reprocessed via a modified double cut *Spo11Mapper* pipeline (see: Section B3.1) to produce comparable datasets to those of *sae2*Δ mapping (Table 3.2). In order to improve the mappability of short Spo11-oligos, FASTQ entries are trimmed prior to alignment to remove any poly(G/C) tails and Illumina adaptor sequences present in overlapping paired end reads (see: Section B3.1.6). As with *sae2*Δ libraries, the Spo11 end is defined by Read-1 adaptors. An average library contains 8.96m read pairs—of which 6.88m (76.18%) successfully align. Of these mapped pairs: (i) 7.02% constitute multi-mappers and (ii) 2.48% fail the ambiguous end filter. Overall, an average library of 8.96m read pairs produces 6.71 valid 5' Read-1 hits and correspondingly, 6.71 valid Read-2 hits (74.95%). Lower average alignment rates for Spo11-oligo libraries relative to *sae2*Δ data (76.18% vs. 95.2%) likely reflects the difficulty in uniquely mapping shorter molecules.

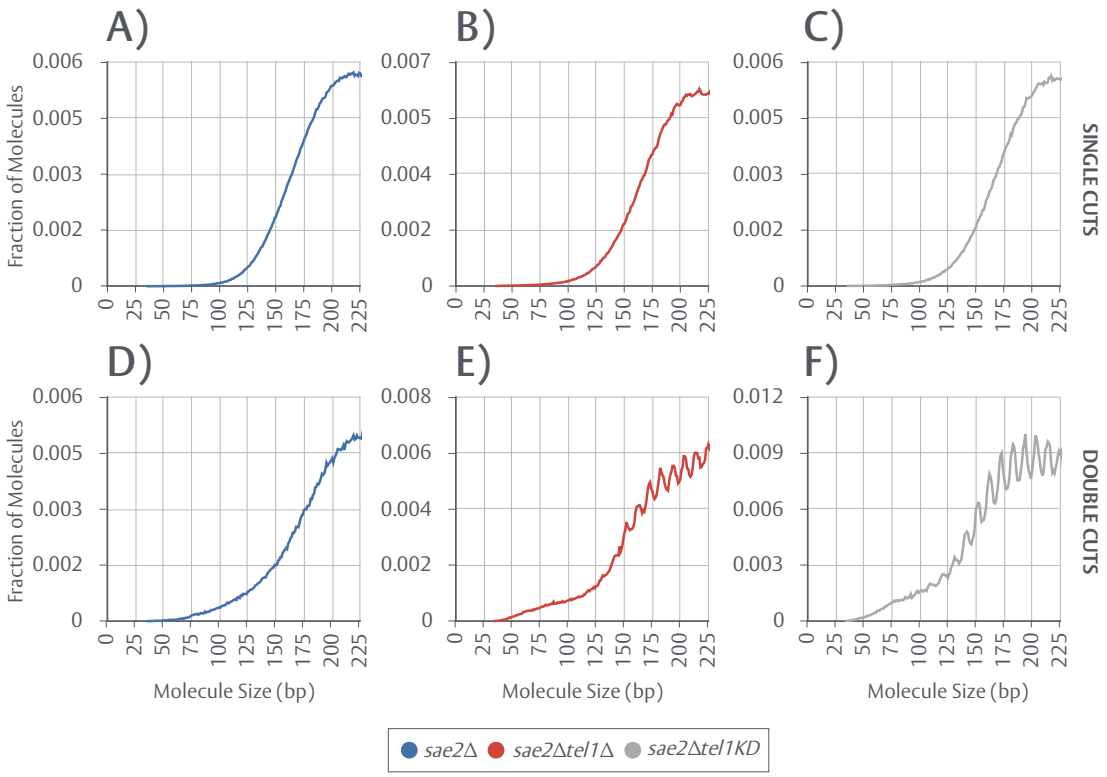


Figure 3.13. Long range (>100bp) 10bp periodicity within Tel1 mutants

Paired end, double cut libraries were sequenced (MiSeq, 2 x 75bp reads) for *sae2*Δ (2 repeats), *sae2*Δ*tel1*Δ (2 repeats) and *sae2*Δ*tel1KD* (kinase dead) (2 repeats), aligned against *Cer3H4L2* and processed via *Spo11Mapper* (3'→5' trimming = 10bp). Molecule sizes, defined as the absolute distance between the 5' Read-1 end and the 5' Read-2 end, were calculated via *Spo11Mapper*, aggregated and plotted as fractions of total based on their respective sizes (0-225bp) for **A)** *sae2*Δ (single cut) **B)** *sae2*Δ*tel1*Δ (single cut) **C)** *sae2*Δ*tel1KD* (single cut) **D)** *sae2*Δ (double cut) **E)** *sae2*Δ*tel1*Δ (double cut) and **F)** *sae2*Δ*tel1KD* (double cut).

Genotype	Pairs (A)	Mapped Pairs (B)	% of (A)	MM Pairs	% of (B)	Valid Hits	% of (B)	Ambig Hits	% of (B)
WT_4h_1	4,780,906	2,453,919	51.33	167,402	6.822	4,789,404	97.587	118,434	2.413
WT_4h_2	9,370,434	6,850,152	73.10	771,871	11.268	13,194,656	96.309	505,648	3.691
WT_6h_1	11,775,326	7,765,855	65.95	1,216,624	15.666	14,797,044	95.27	734,666	4.73
WT_6h_2	7,331,757	4,965,877	67.73	544,684	10.969	9,548,778	96.144	382,976	3.856
*tel1*Δ_4h_1	15,107,826	12,069,698	79.89	928,102	7.69	23,570,958	97.645	568,438	2.355
*tel1*Δ_4h_2	9,481,848	8,112,232	85.56	332,142	4.094	15,956,364	98.348	268,100	1.652
*tel1*Δ_6h_1	12,034,357	9,959,338	82.76	398,592	4.002	19,683,306	98.818	235,370	1.182
*tel1*Δ_6h_2	8,523,723	7,334,146	86.04	290,199	3.957	14,441,432	98.453	226,860	1.547
*tel1KD*_4h	10,416,828	8,443,364	81.06	276,928	3.28	16,670,562	98.72	216,166	1.28
*tel1KD*_6h	10,212,421	8,147,063	79.78	312,293	3.833	16,035,926	98.415	258,200	1.585

Table 3.2. Spo11-oligo libraries processed by Spo11Mapper

WT (4 repeats), *tel1*Δ (5 repeats) and *tel1KD* (kinase dead) (3 repeats) Spo11-oligo libraries, obtained from a previously published source (Mohibullah & Keeney 2017), were realigned against *Cer3H4L2* and processed via *Spo11Mapper* (no 3'→5' trimming) to generate comparable datasets. Prior to alignment, all FASTQ sequences are trimmed to remove poly(G/C)-tailing, added during Spo11-oligo library prep, and illumina adaptor sequence. The time point at which each sample was taken during meiosis is specified (4h or 6h). See: Figure 3.4 for an explanation of each field.

Oligo sizes, as calculated by *Spo11Mapper*, were visualised at a range of 0-100bp for WT, *tel1*Δ and *tel1KD* samples taken at 4h (Figure 3.14A) and 6h (Figure 3.14B). As observed by gel, a 10bp periodicity, that accumulates over time, is clearly visible within all backgrounds including WT. Importantly, periodic peaks are consistent in size with those previously observed for *tel1*Δ mutants (33-43-53-63bp) (Mohibullah & Keeney 2017). Spo11-oligo libraries predominately consist of molecules released by Mre11-dependent nucleolytic cleavage at the 3' end—which appear to exhibit a preferred size (~27bp) (see: Figure 3.14A,B). Despite this, some Mre11 end points are assumed to be random and therefore present at low, contaminating frequencies across the size spectrum. As previously outlined, 10bp periodicity may arise through Spo11 double cutting. An expected, distinguishing feature of Spo11-oligos released by double cutting are frequently occurring, reciprocal coordinates whereby the 3' end of one molecule aligns with the 5' of another (Figure 3.14C). In contrast, the 3' end of any Spo11-oligo released by Mre11, will only align with a 5' Spo11 end by chance—thus generating reciprocal coordinates at much lower frequencies. Therefore, in order to enrich for Spo11-Spo11 species over Spo11-Mre11 background, mapped 6h libraries were filtered for reciprocity. Reciprocal oligo sizes were recalculated and plotted (Figure 3.14D). Consequently, a significant enrichment in periodicity is observed—collectively suggesting, along with results from *sae2*Δ mapping (see: Section 3.13), that 10bp periodicity is a hallmark of legitimate short range (<75bp) and long range (>100bp) Spo11 double cuts.

3.15—Short range (<75bp) periodic molecules have Spo11-like sequence bias at both ends

Legitimate Spo11 double cuts may occur between sites that, on average, exhibit preferential Spo11 sequence bias. In order to investigate this possibility and the identity of <75bp periodic molecules, WT sequence bias was calculated, via *SeqBias* (see: Section B3.1.8), for several subpopulations of *unfiltered* oligo sizes (Figure 3.15). Analysis of 27bp molecules reveals a 5' Spo11-like bias, consistent with that previously determined for *sae2*Δ data (see: Figure 3.9) and a non-Spo11, presumably Mre11-like, bias precisely 27bp downstream at the 3' end (Figure 3.15A). Similar results are obtained for non-periodic, 39bp molecules which reside in a trough between 33-43bp peaks.

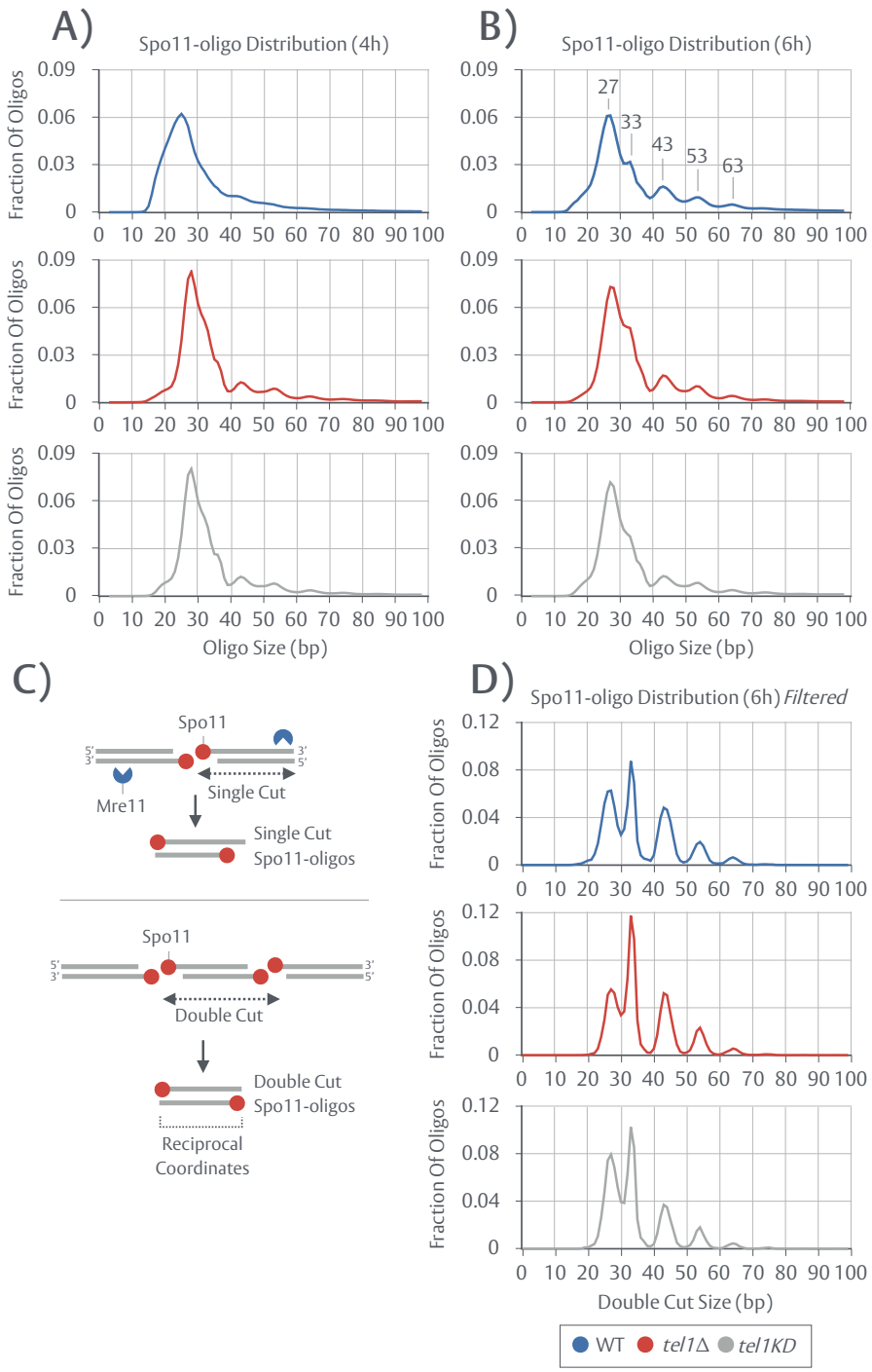


Figure 3.14. Short range (<75bp) 10bp periodicity is a WT phenomenon

Oligo sizes, defined as the absolute distance between the 5' Read-1 end and the 5' Read-2 end, were calculated via *Spo11Mapper*, aggregated and plotted as fractions of total based on their respective sizes (0-100bp) for **A)** WT, *tel1*Δ and *tel1KD* Spo11-oligo 4h libraries **B)** WT, *tel1*Δ and *tel1KD* Spo11-oligo 6h libraries. **C)** Spo11-oligos are ordinarily released via Mre11 endonucleolytic cleavage—defined as a canonical single cut. Hypothetical double cuts may arise through Spo11 double cutting, releasing Spo11-oligos independently of Mre11. An expected, distinguishing feature of double cut Spo11-oligos are frequently occurring, reciprocal coordinates whereby the 3' end of one molecule aligns with the 5' of another. **D)** Oligo sizes were recalculated using filtered, reciprocal molecules for WT, *tel1*Δ and *tel1KD* Spo11-oligo 6h libraries.

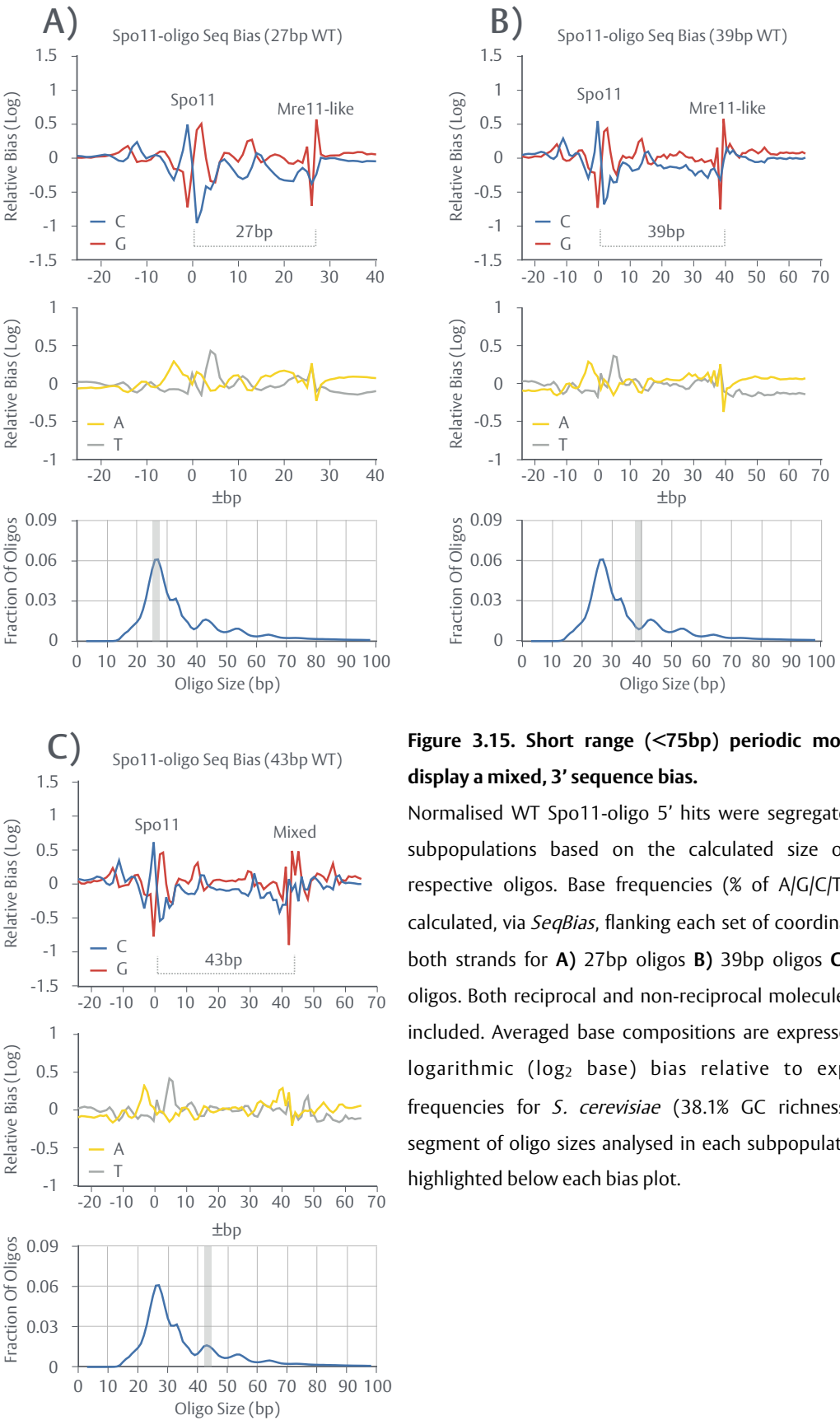


Figure 3.15. Short range (<75bp) periodic molecules display a mixed, 3' sequence bias.

Normalised WT Spo11-oligo 5' hits were segregated into subpopulations based on the calculated size of their respective oligos. Base frequencies (% of A/G/C/T) were calculated, via *SeqBias*, flanking each set of coordinates on both strands for **A)** 27bp oligos **B)** 39bp oligos **C)** 43bp oligos. Both reciprocal and non-reciprocal molecules were included. Averaged base compositions are expressed as a logarithmic (log_2 base) bias relative to expected frequencies for *S. cerevisiae* (38.1% GC richness). The segment of oligo sizes analysed in each subpopulation are highlighted below each bias plot.

In the case of 39bp molecules, a Mre11-like bias is observed precisely 39bp downstream of the 5' Spo11 end (Figure 3.15B). In contrast, analysis of 43bp periodic molecules reveals a perturbed 3' pattern that resembles a mixture of Spo11 and Mre11 bias features (Figure 3.15C). To further refine the obtained sequence bias, reciprocally *filtered* subpopulations (43bp, 53bp) were analysed, via *SeqBias*, for WT, *tel1*Δ and *tel1KD* data. Consistent with the idea that molecules which exhibit frequently occurring reciprocal coordinates represent legitimate double cuts, strong Spo11-like biases are observed at both the 5' and the 3' end of 43bp (Figure 3.16) and 53bp (Figure 3.17) molecules in all genotypes. No obvious differences are notable between WT, *tel1*Δ and *tel1KD*. Crucially, the distance between the cleavage axes is precisely equivalent to the oligo size analysed (43 or 53bp). Collectively, these results further strengthen the idea that short range 10bp periodicity is a WT phenomenon, that arises through hyper localised, Spo11 double cutting.

3.16—Estimating the frequency of short range Spo11 double cuts

Despite enrichment of periodic molecules via reciprocal filtering (see: Section 3.14), considerable levels of ~27bp Spo11-Mre11 signal remains—suggesting Mre11 contamination is not fully removed. Retention of Spo11-Mre11 molecules suggests that, by chance, Mre11 sometimes cleaves at sites also utilised by Spo11 for double cutting. Direct quantitative estimate of double cut frequency thus remains difficult. However, the mixed 3' bias observed for *unfiltered* periodic molecules (see: Figure 3.15C) should contain information regarding the Mre11:Spo11 (M:S) ratio. Therefore, in order to determine what proportion of this bias may be ascribed to Spo11, thereby approximating the M:S ratio, sequence biases were mixture modelled for WT and *tel1*Δ data in an attempt to match the observed mixed, 43bp 3' biases (Figure 3.18A,B). 3' biases for 27bp *unfiltered* molecules—assumed to be a near pure representation of Mre11—were centred and mixed with the 3' bias of 43bp *filtered* molecules—a strong, Spo11-like pattern—to varying amounts. The average, absolute difference between simulated and observed bias was utilised to obtain optimal mixtures of 55%:45% M:S and 48%:52% M:S for WT and *tel1*Δ respectively.

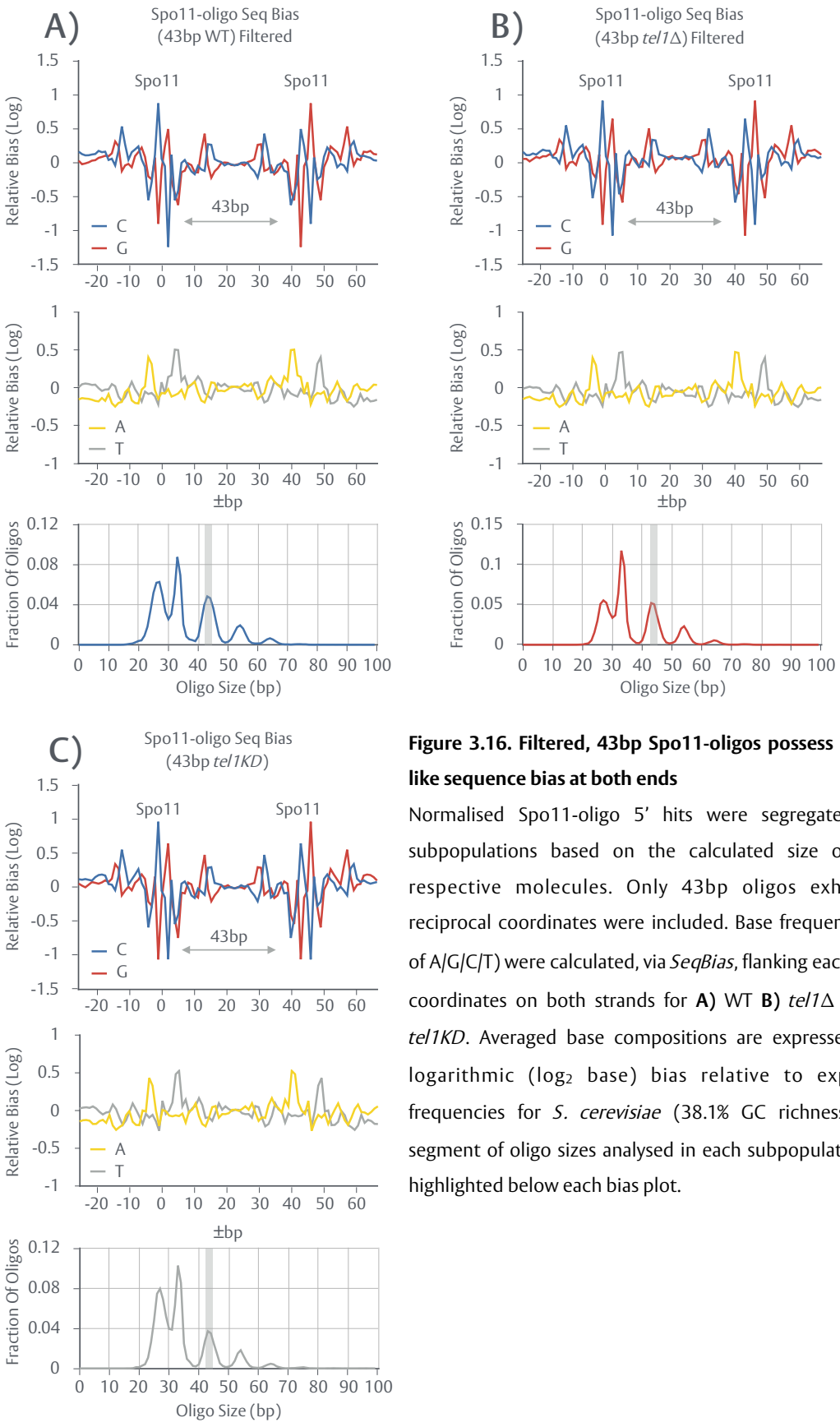


Figure 3.16. Filtered, 43bp Spo11-oligos possess Spo11-like sequence bias at both ends

Normalised Spo11-oligo 5' hits were segregated into subpopulations based on the calculated size of their respective molecules. Only 43bp oligos exhibiting reciprocal coordinates were included. Base frequencies (% of A/G/C/T) were calculated, via *SeqBias*, flanking each set of coordinates on both strands for **A)** WT **B)** *tel1*Δ and **C)** *tel1KD*. Averaged base compositions are expressed as a logarithmic (log_2 base) bias relative to expected frequencies for *S. cerevisiae* (38.1% GC richness). The segment of oligo sizes analysed in each subpopulation are highlighted below each bias plot.

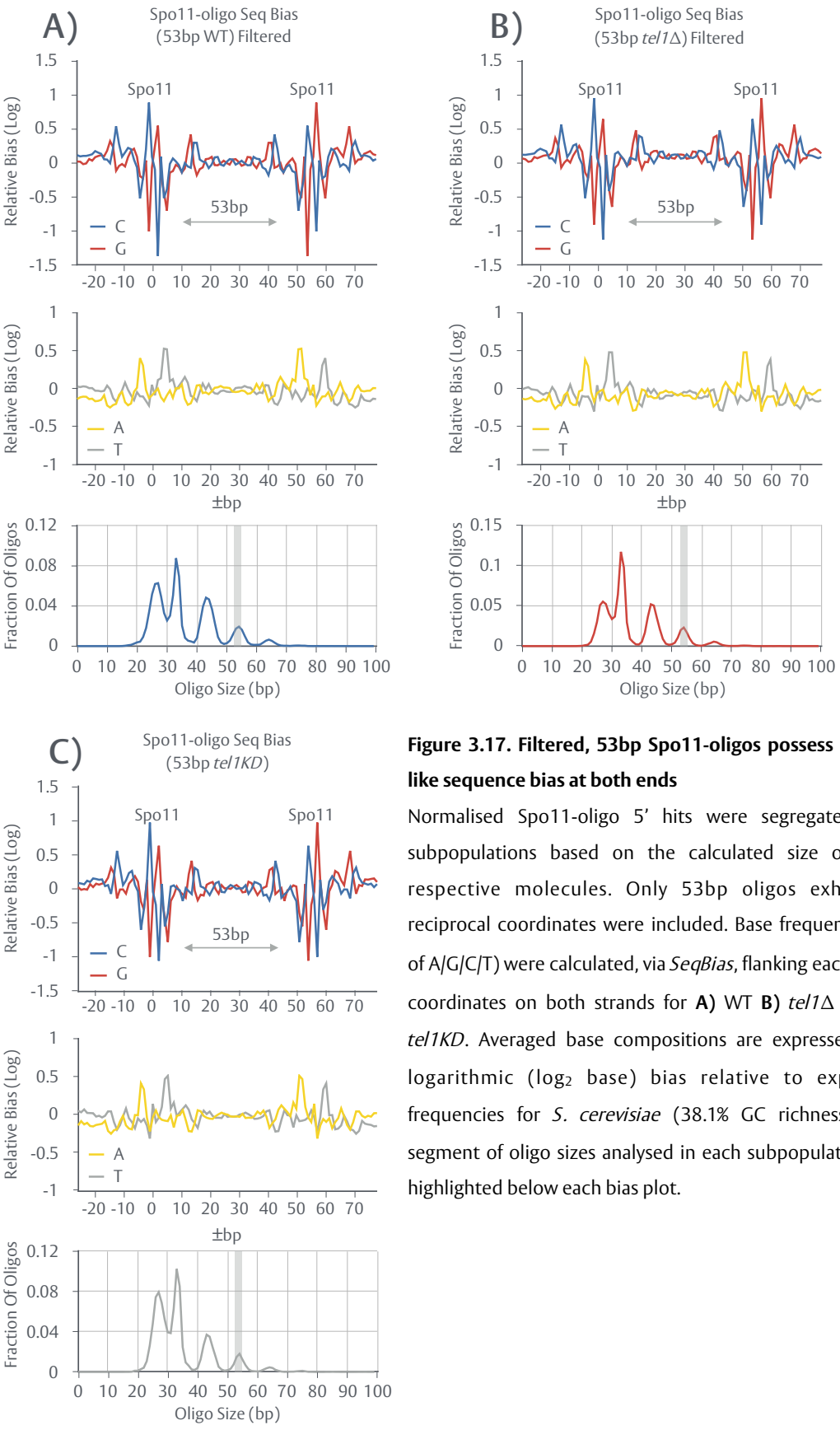


Figure 3.17. Filtered, 53bp Spo11-oligos possess Spo11-like sequence bias at both ends

Normalised Spo11-oligo 5' hits were segregated into subpopulations based on the calculated size of their respective molecules. Only 53bp oligos exhibiting reciprocal coordinates were included. Base frequencies (% of A/G/C/T) were calculated, via *SeqBias*, flanking each set of coordinates on both strands for **A)** WT **B)** *tel1*Δ and **C)** *tel1KD*. Averaged base compositions are expressed as a logarithmic (log_2 base) bias relative to expected frequencies for *S. cerevisiae* (38.1% GC richness). The segment of oligo sizes analysed in each subpopulation are highlighted below each bias plot.

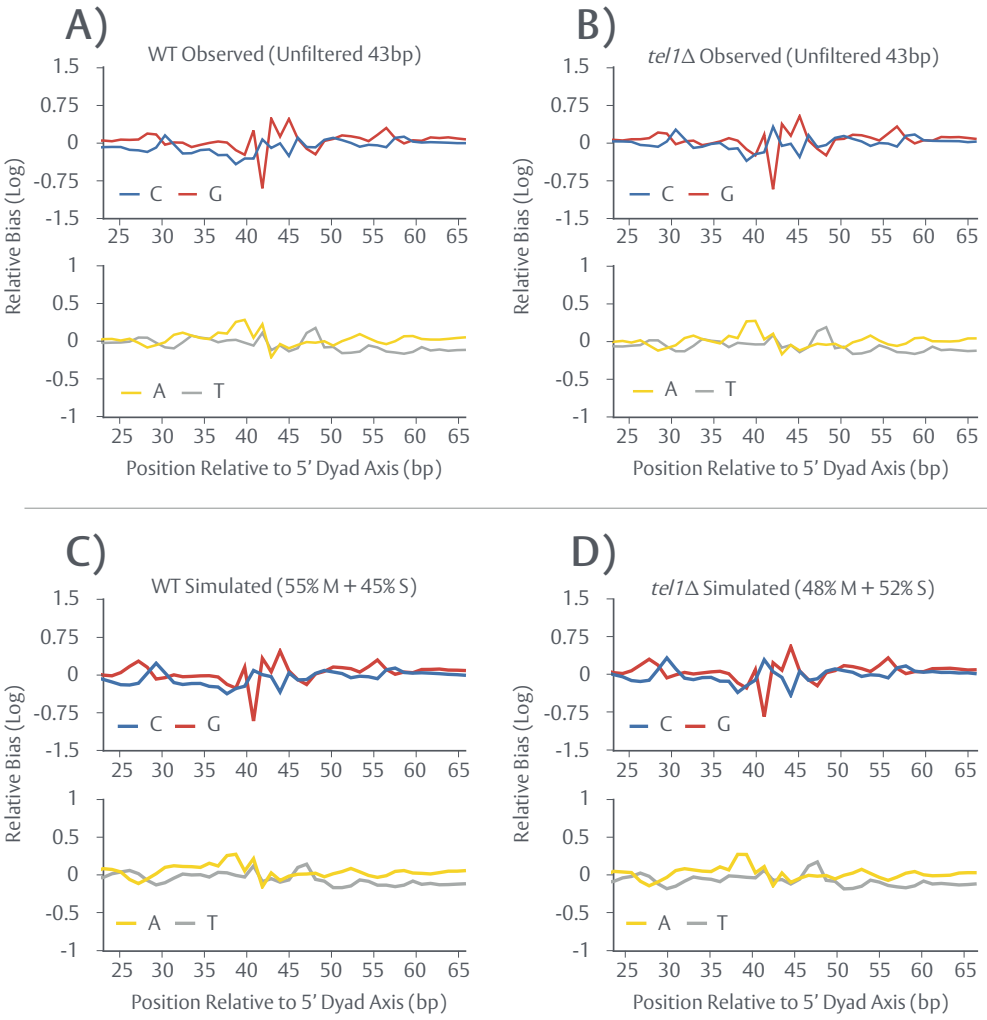


Figure 3.18. Estimating the frequency of short range Spo11 double cuts

3' sequence biases, derived via *SeqBias*, from unfiltered 43bp Spo11-oligos (see: Figure 3.17C) were centred and shown at (x) coordinates relative to the 5' dyad axis for **A)** WT and **B)** *tel1*Δ. **C-D)** 3' biases for 27bp unfiltered molecules (Mre11 bias) were centred and mixed, to varying extents, with the 3' bias for 43bp filtered molecules (Spo11 bias) to simulate approximations of the observed 3' 43bp unfiltered biases. Mre11:Spo11 ratios were assessed by calculating point by point percentage differences between observed and simulated biases. Ratios yielding the lowest averaged difference were considered optimal.

By applying these optimal values to the oligo frequencies present in each periodic peak (e.g. 0.45*33bp oligo count for WT), it can be determined that ~2.77% (WT) and ~4.04% (*tel1*Δ) of the molecules within each respective library may represent Spo11 double cuts. Given the 5' C ambiguity that arises in the alignment of Spo11-oligos, less conservative estimates can be obtained by considering periodic peaks (33-43-53-63bp) ±2bp—yielding estimates of ~12.38% (WT) and ~17.81% (*tel1*Δ). Collectively, these results suggest that while canonical Spo11-Mre11 oligos are the dominant species in Spo11-oligo libraries, Spo11 double cuts appear to form at an appreciable, albeit low rate that may moderately increase within a *tel1*Δ background.

3.17—Short range double cut molecules occur across the genome

Spo11 double cuts may cluster within specific regions of the genome. Alternatively, they may prove a common feature to all hotspots. Furthermore, it is unclear whether or not any given region gives rise to a single length of double cut or a mixed set. In order to initially assess and distinguish these possibilities, frequently occurring 33-43-53-63bp periodic molecules were visualised within two hotspots regions (*ERV15*, *ARE1*) (100bp window) as colour coded arcs denoting the sites between which cleavage occurred (Figure 3.19). Within *ERV15* (Figure 3.19A) and *ARE1* (Figure 3.19B), a mixed set of oligo sizes are observed across the hotspot, occurring between established single cut library peaks. To expand these findings to a genome-wide level, the occurrence of periodic molecules within each known hotspot was quantified. Of the 3599 annotated hotspots present in *S. cerevisiae*, 1257 (34.93%) contain at least 5 double cut molecules 33/43/53/63bp in length, while 1980 (55.01%) contain at least 5 double cut molecules 33-43-53-63bp ±2bp in length. Hyper localised double cutting thus appears to be a widely observed phenomenon across the genome. Those hotspots lacking molecules may produce double cut molecules below the detection threshold of the Spo11-oligo mapping assay.

Figure 3.19. Hotspots exhibit a range of double cut sizes

Based on Read-1 5' and Read-2 5' coordinates, the positions of putative double cut Spo11-oligos 33, 43, 53 or 63bp in length were visualised as arcs across 100bp windows for two hotspots: **A)** *ERV15* (ChrII) **B)** *ARE1* (ChrIII). Each arc is colour coded based on the size of oligo. (x) coordinates are omitted for clarity.

3.18—Etoposide-dependent genome-wide formation of Topo II lesions

Topoisomerase II resolves DNA superhelical tension, knots and catenanes through the transient formation of DSBs with a 4bp overhang (Burden & Osheroff 1998; Nitiss 2009; Schoeffler & Berger 2005). Catalytically, Topo II DSBs and Spo11 DSBs are thought to form via similar catalytic mechanisms (Bergerat et al. 1997; Keeney et al. 1997; Keeney & Kleckner 1995; Liu et al. 1995). Topo II DSBs can, however, be stabilised through the use of cellular toxins. Etoposide, one such toxin, binds to the protein:DNA interface—misaligning the DNA strands, inhibiting religation and stalling covalently bound Topo II complexes (Pommier & Marchand 2011). Akin to Spo11, Topo II lesions require end processing prior to repair via NHEJ or HR (Gómez-Herreros et al. 2013; Cruz-García et al. 2014). Within mammalian cells, TDP2 directly hydrolyses the 5' phosphotyrosine linkage—releasing Topo II (Ledesma et al. 2009; Gao et al. 2014). Alternatively, CtIP (Sae2) and Mre11 may endonucleolytically process the DSB (Aprarício et al. 2016; Nakamura et al. 2010). *S. cerevisiae* relies solely upon the latter Sae2/Mre11-dependent pathway and no TDP2 ortholog has been identified in this organism (Hartsuiker et al. 2009).

Akin to Spo11 DSBs (see: Section 3.9), genome-wide mapping of stalled, mammalian Topo II complexes revealed a preference for Topo II lesions to form within promoter regions with an anti correlation to nucleosomal occupancy (Baranello et al. 2014). However, such mapping was low resolution. In order to generate nucleotide resolution maps of Topo II lesions and expand *Spo11Mapper* to novel datasets as a proof of principle, the *sae2*Δ-dependent method of Spo11 DSB mapping (see: Section 3.2) was adapted and applied to cycling WT, *sae2*Δ and *mre11*Δ haploid *S. cerevisiae* cells (D. Johnson, M.J. Neale unpublished). Increased sensitivity to etoposide treatment was conferred through repression of pleiotropic drug resistance (PDR) extrusion pumps, which otherwise export etoposide (D. Johnson, M.J. Neale, unpublished). FASTQ data was processed via *Spo11Mapper* (see: Section 3.3), extracting 5' Read-1 Topo II hits and generating 1bp histograms used for map visualisation across two representative chromosomes (ChrI and ChrV) (Figure 3.20).

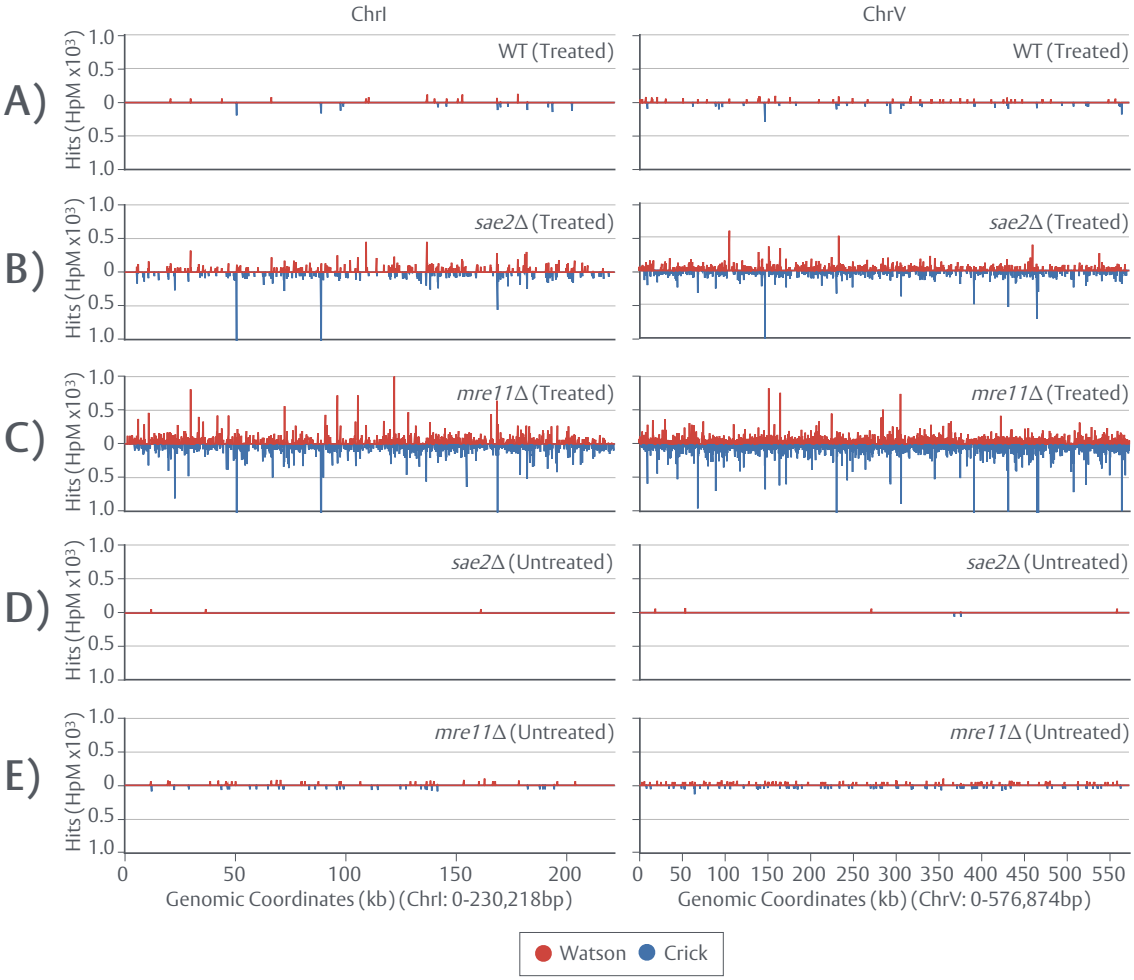


Figure 3.20. Etoposide-dependent genome-wide formation of Topo II lesions

1bp Topo II histogram data, generated via *Spo11Mapper*, was normalised (HpM) and visualised on both strands across ChrI and ChrV for **A)** WT (T) **B)** *sae2*Δ (T) **C)** *mre11*Δ (T) **D)** *sae2*Δ (U) **E)** *mre11*Δ (U). U = untreated, T = etoposide treated.

Samples treated with etoposide exhibit a non-random distribution of Topo II lesions characterised by distinct peaks, residing above background level (>1HpM) (Figure 3.20A-C). Signal within WT is considerably lower than that of *sae2*Δ and *mre11*Δ, confirming the involvement of the Sae2-Mre11 pathway within repair of etoposide-dependent Topo II lesions. Furthermore, relative to *sae2*Δ, signal within *mre11*Δ is further enriched suggesting a heavier reliance upon Mre11 for lesion repair. In contrast, untreated samples exhibit wide ranging coverage but a lack distinct peaks—indicative of non-specific background (Figure 3.20D,E). However, cross correlation of samples reveals a high degree of similarity between the nucleotide positions of significant, etoposide treated *mre11*Δ (>1HpM) peaks and all positions containing >1HpM signal within WT, *sae2*Δ and untreated samples (~98-99% commonality) (Figure 3.21A). Such a similarity suggests that untreated signal is not background, that Topo II exhibits a high degree of base pair specificity for lesion formation and that etoposide-independent, naturally occurring lesions are detected by genome-wide mapping at identical locations to those within treated libraries. Consistent with greater enrichment of signal within treated *mre11*Δ cells, significant (>1HpM) signal occupies ~80,000 unique sites across the genome, as opposed to ~40,000 and ~30,000 within *sae2*Δ and WT respectively (Figure 3.21A). Spo11 DSBs display specificity at both the base pair and chromosomal levels (see: Sections 3.7-3.8). In order to determine whether or not chromosomal length has an impact on Topo II lesion formation, hit densities (HpM/bp) were calculated for treated *mre11*Δ and *sae2*Δ samples (Figure 3.21B). However, no clear correlation is apparent ($\rho = <0.3$), suggesting Topo II lesions form with equal density across all chromosomes.

3.19—Topo II lesions preferentially form within nucleosome depleted regions (NDRs)

As per previous studies (Baranello et al. 2014), Topo II lesions exhibit a preference toward nucleosomally depleted, promoter regions. In order to determine whether or not this preference is recaptured, hit densities (hits/kb) were calculated for several types of genomic loci (tandem, convergent, divergent and genic) using treated, *mre11*Δ and *sae2*Δ data (Figure 3.21C) (see: Figure 3.8C).

A)

Sample	Watson	Crick	Shared
WT (U)	16,455	16,691	98.959%
WT (T)	17,654	18,089	99.603%
*sae2*Δ (U)	15,327	15,685	99.994%
*sae2*Δ (T)	24,653	24,789	99.996%
*mre11*Δ (U)	14,608	14,709	99.894%
*mre11*Δ (T)	42,525	42,974	N/A

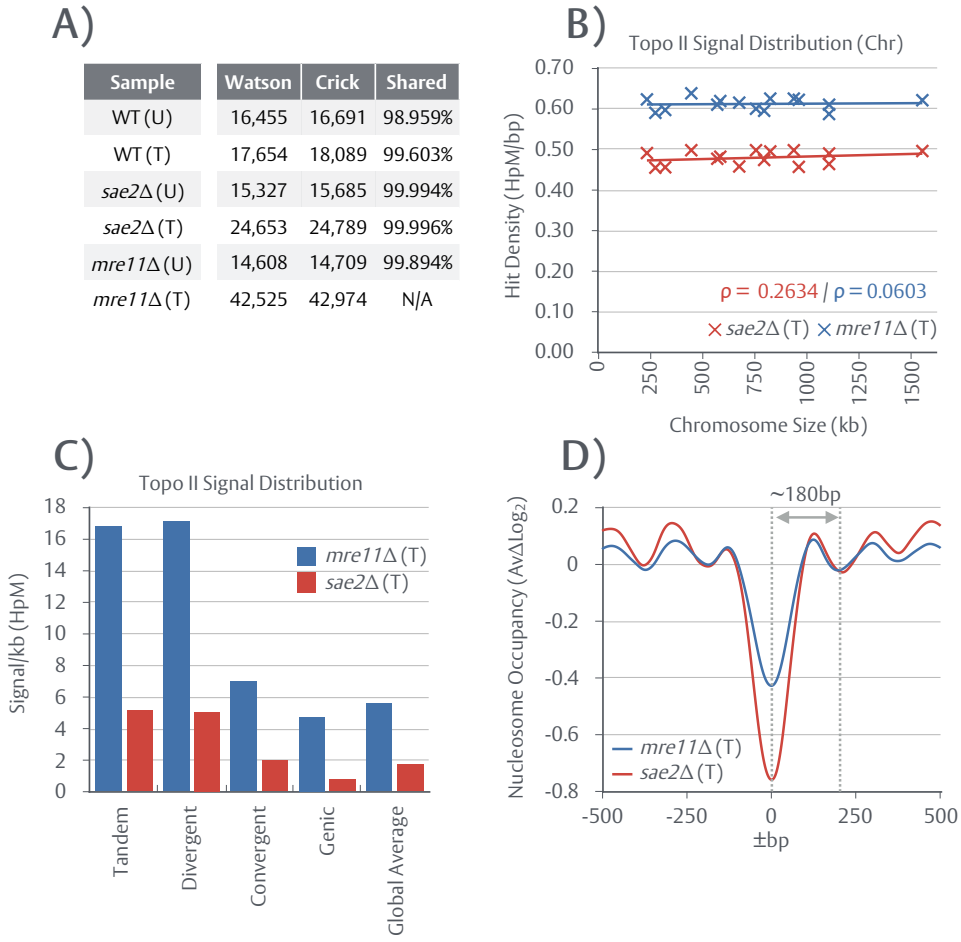


Figure 3.21. Topo II lesions preferentially form within nucleosome depleted regions (NDRs)

A) Normalised 1bp Topo II histogram data, generated via *Spo11Mapper*, was cross correlated between each strain, tallying the number unique and shared nucleotide positions containing >1HpM signal relative to treated, *mre11*Δ samples. U = untreated, T = treated. **B)** Normalised data (HpM) was summed across each chromosome, converted to hit densities (HpM/bp) and plotted against *S. cerevisiae* chromosome size. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values). Linear trendlines are marked. **C)** Normalised Topo II 5' hits were summed across varying types of intergenic region (IGR) (tandem, divergent, convergent) and each annotated ORF (genic) for *sae2*Δ and *mre11*Δ data. Hit counts were subsequently converted to averaged densities (signal/kb) using the total, cumulative length of each region type. **D)** Normalised nucleosomal occupancy, expressed as a logarithmic deviation from the genome-wide average ($Av\Delta Log_2$) and derived from (Kaplan et al. 2008), flanking each significant Topo II peak (>1HpM), for *sae2*Δ and *mre11*Δ data, was extracted, piled up, subsequently averaged and smoothed (moving average, n=5). T = etoposide treated.

Regardless of background, Topo II lesions exhibit a clear distribution characterised by: (i) a significant enrichment of signal within promoter-containing intergenic regions (IGRs) (divergent, tandem) (ii) a moderate enrichment of signal within convergent regions and (iii) moderate intragenic formation (~10% of Topo II lesions). In contrast to Spo11, which exhibits strong preference for divergent regions, Topo II signal density is not proportional to the number of promoters present. Topo II lesions may therefore form independently of promoter associated factors and chromatin accessibility or presence of transcriptionally induced torsional stress may prove sufficient to guide formation. In order to investigate the local chromatin environment of Topo II lesion formation, normalised nucleosomal occupancy, derived from (Kaplan et al. 2009), was piled up centred on all major peaks (>1HpM) for *mre11*Δ and *sae2*Δ datasets (Figure 3.21D). Topo II lesions in both mutant backgrounds form within ~180-200bp regions exhibiting nucleosomal occupancy levels significantly below the genome-wide average (-0.5-0.8 AvΔLog_2)—corroborating previous observations that Topo II lesions form within NDRs (Baranello et al. 2014). Interestingly, flanking signal with a ~150-180bp periodicity is clearly visible—suggesting Topo II lesions also form within regions containing highly ordered nucleosomes such as ORFs. Consistent with this, ~20-30% of Topo II signal (>1HpM) within treated *mre11*Δ and *sae2*Δ samples resides within non-NDR regions (Figure 3.22A). Interestingly, the fraction of non-NDR lesions increases within *mre11*Δ, relative to *sae2*Δ and WT—suggesting non-NDR events are either low frequency, thus more readily detected within repair deficient strains, or a novel consequence of repair inhibition. To further investigate the relationship between lesion formation and NDRs, the hit count of all major, treated *mre11*Δ peaks (>1 HpM) was plotted against the average nucleosome occupancy ±5bp flanking each peak (Figure 3.22B). No clear correlation is observed (ρ = -0.1124) and a significant amount of formation occurs at occupancy levels above the genome-wide average (>0 AvΔLog_2). Thus, while formation of lesions within NDRs is preferential, the absence of a nucleosome may not be essential. Alternatively, transient regions of nucleosome depletion, not detected in the population average, may account for these apparent non-NDR lesions.

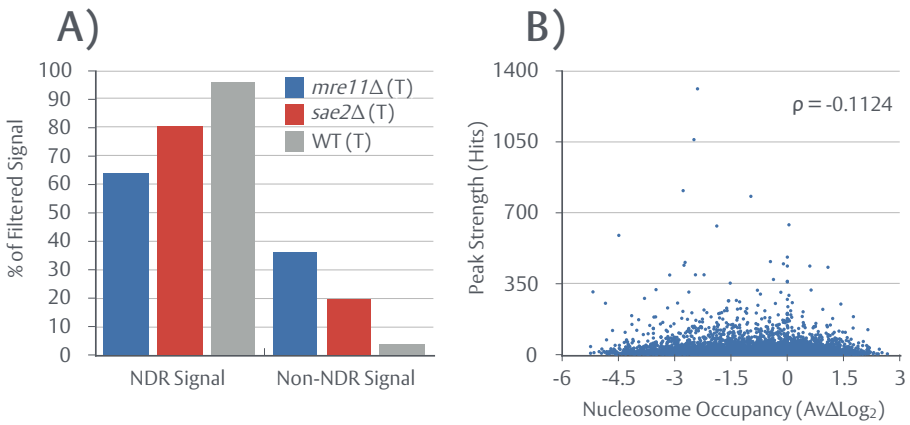


Figure 3.22. NDRs may not be essential for Topo II lesion formation

A) Normalised Topo II peaks (>1HpM), from WT, *sae2*Δ and *mre11*Δ, were segregated into subpopulations based on the immediately adjacent (±5bp), averaged nucleosome occupancy levels. Non-NDRs are defined as regions containing values of >0 AvΔLog$_2$. NDRs are defined as regions containing values of <0 AvΔLog$_2$. T = etoposide treated. **B)** Local nucleosome environments (averaged ±5bp occupancy) were plotted against the total number of hits contained within each Topo II peak using treated *mre11*Δ data.

3.20—Intragenic Topo II lesions primarily form at the 5' end of genes

In contrast to Spo11 DSBs—which only form within genic regions at significant frequencies in *tel1*Δ or *tel1KD* backgrounds (see: Sections 3.11-3.12)—~10% of Topo II signal is intragenic. In order to investigate the distribution of intragenic Topo II lesions, all major signal (>1HpM) was piled up around the start and stop position of every annotated *S. cerevisiae* ORF and subsequently averaged (Figure 3.23). As expected, Topo II signal primarily resides within the upstream promoter regions (Figure 3.23A), and to a lesser extent downstream of the 3' gene end (Figure 3.23B). Interestingly, and consistent with nucleosomal pileups (see: Figure 3.21D), periodically repeating ~150-180bp peaks of signal are observed stretching into the flanking ORF. Such a periodicity is reminiscent of MNase sensitive linker regions between nucleosomes (Axel 1975), further suggesting that chromatin accessibility is a major factor governing Topo II lesion formation. Moreover, the intensity of both signal and periodicity appears to diminish across the ORF. The vast majority of intragenic Topo II signal thus resides at the 5' end of genes, suggesting Topo II may be more active or present at the initial portion of each ORF.

3.21—Topo II exhibits a weak, symmetrical sequence bias for the generation of DSBs

High resolution nucleotide mapping of Topo II permits the generation of accurate sequence bias profiles, previously impossible with low resolution data. In order to determine the sequence bias of Topo II lesions, *SeqBias* (*see*: *Section B3.1.8*) was utilised to calculate the base frequencies ±20bp surrounding treated and filtered (>1HpM) *mre11*Δ peaks—revealing a weak, asymmetrical bias (Figure 3.24A). As with Spo11 DSBs, inspection of the data revealed a substantial subpopulation of non-cognate peaks—that is, significant hits on Watson (+) or Crick (-) without a corresponding peak on the opposing strand, 4bp away. To further refine the obtained sequence bias, *mre11*Δ data was further filtered into cognate and non-cognate subpopulations and re-analysed via *SeqBias*. Cognate peaks, defined as peaks with corresponding, offset signal and which are expected to reflect legitimate Topo II DSBs, display a perfectly symmetrical, rotational (palindromic) bias (Figure 3.24B).

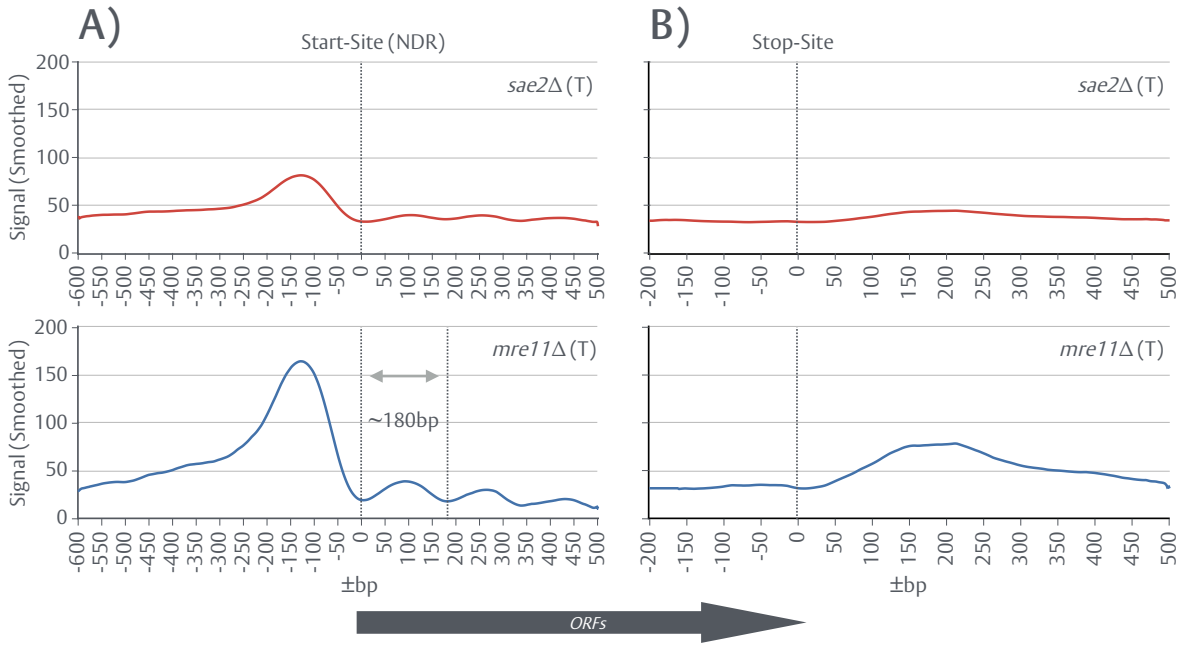


Figure 3.23. Intragenic Topo II lesions primarily form at the 5' end of genes

Normalised *sae2*Δ and *mre11*Δ 5' Topo II hits were piled up centred on the start and stop sites of all annotated S. cerevisiae ORFs, averaged and smoothed (moving average, n = 50bp). **A)** Start sites **B)** Stop sites. The orientation of the collective, piled up ORF is marked below the plot. All (-) strand ORFs were reversed in orientation for visual clarity. T = etoposide treated.

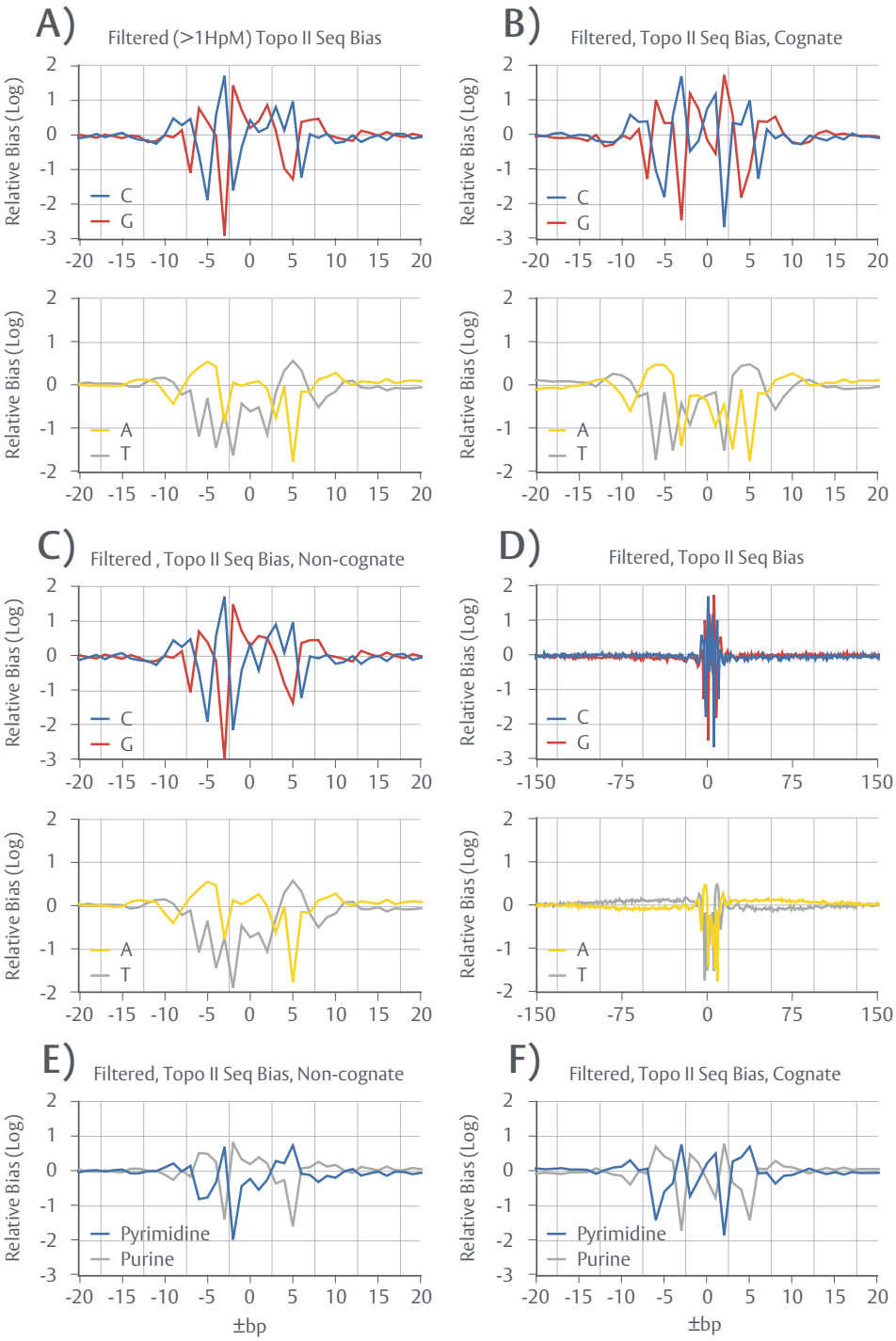


Figure 3.24. Topo II exhibits a weak, symmetrical sequence bias for the generation of DSBs

A) Base frequencies (% of A/G/C/T) were calculated, via *SeqBias*, flanking (±20bp) all filtered (>1HpM) etoposide treated *mre11*Δ 5' hits. Filtered *mre11*Δ data was further partitioned into cognate and non-cognate subpopulations. Non-cognate peaks were defined as any >1HpM peak without a >1HpM 4bp away on the opposing strand. **B)** Base frequencies were calculated, via *SeqBias*, flanking (±20bp) filtered, cognate peaks **C)** Base frequencies were calculated, via *SeqBias*, flanking (±20bp) filtered, non-cognate peaks. **D)** Base frequencies were calculated, via *SeqBias*, flanking (±150bp) all filtered (>1HpM) etoposide treated *mre11*Δ 5' hits **E-F)** Base frequencies were combined for pyrimidines (T+C) and purines (A+G) and replotted for filtered, non-cognate and cognate peaks respectively.

With the dyad-axis set to position (0), a Topo II DSB bias half site from the top strand may thus be read as GAAC*GG|CCGTTC with the dyad-axis and cleavage site denoted as | and * respectively.

By contrast, non-cognate peaks—perhaps reflective of SSBs—display an asymmetric bias (Figure 3.24C). Non-cognate A-T patterns are perturbed but remain reasonably similar. In contrast, G-C patterns resemble an incomplete, half-site. Notably, and in agreement with previous studies which noted the predominance of SSBs over DSBs in etoposide treated samples (Bromberg et al. 2003), ~90% of filtered (>1HpM) Topo II peaks are non-cognate. While Topo II SSB formation has been ascribed to low etoposide concentrations—where only a single monomer is inhibited—this study utilised a high (1μM) concentration of etoposide. Therefore, formation of SSBs may be guided by alternative factors, such as incomplete, half site sequence biases which, on average, restrict cleavage to a single monomer.

Interestingly, when base frequencies are calculated ±150bp surrounding both cognate and non cognate filtered (>1HpM) *mre11*Δ peaks, a long range AT skew is observed extending over ~±75bp (Figure 3.24D). No such skew is observed for GC base pairs. Within *S. cerevisiae*, poly(dA:dT) tracts—characterised by 10-20bp homopolymeric stretches of A/T—correlate with and generate regions of nucleosomal depletion, particularly at promoters (Segal & Widom 2009). Such an AT skew is therefore likely to reflect the preferential formation of Topo II lesions within NDRs and/or MNase-sensitive linker regions (see: Section 3.19).

Importantly, the obtained sequence biases for Topo II SSBs and DSBs appear consistent with previous observations that Topo II preferentially cleaves purine-pyrimidine (RY) repeats (Figure 3.24E,F) (Spitzner et al. 1990; Burden & Osheroff 1999). RY periodicity is more apparent and symmetrical for cognate peaks (see: Figure 3.24F) than it is for the more abundant, non-cognate peaks (see: Figure 3.24E).

3.22—Discussion

Work presented here details development and usage of a novel analytical package for the alignment and analysis of genome wide Spo11 DSBs or Topo II lesions, and reveals novel consequences for the removal of Tel1^{ATM} activity on the hyperlocal regulation of DSB formation. Moreover, distributional features of Topo II etoposide-dependent DSB/SSB formation were also analysed.

Mapping Spo11 DSBs

In conjunction with *Spo11Mapper*, *sae2*Δ-dependent mapping of Spo11 DSBs is able to generate single nucleotide resolution, genome-wide maps comparable to those previously obtained through Spo11-oligo mapping—validating the accuracy of *Spo11Mapper* (see: Section 3.6-3.7). In principle and as demonstrated for Topo II DSB/SSBs (see: Section 3.19), *Spo11Mapper* is applicable to any dataset containing 5' Read-1 or Read-2 information. Spo11 DSBs, mapped via *Spo11Mapper*, are observed to form preferentially within nucleosomally depleted regions present at promoters (see: Figure 3.8C) with a distinct, albeit weak, symmetrical sequence bias (see: Figure 3.9C)—in line with previous observations (Baudat & Nicolas 1997; Gerton et al. 2000; Pan et al. 2011). While a similar sequence bias had been previously obtained (Pan et al. 2011), 5' cytosine ambiguities inherent in the Spo11-oligo method skewed the data. Mapping via the *sae2*Δ protocol however eliminates this caveat, and thus presents a more complete picture of how DNA sequence may influence Spo11 cleavage at the base pair level. An improved bias may aid pre-existing hotspot designation models, which predict hotspot position via parameters including genomic sequence (Champeimont & Carbone 2014). As noted (Blitzblau et al. 2007; Gerton et al. 2000; Pan et al. 2001; Martini et al. 2006), a negative correlation is observed between Spo11-oligo hit density and chromosomal size when excluding smaller <400kb chromosomes from consideration (see: Figure 3.7C)—previously attributed to the phenomenon of synapsis-dependent shutdown (SDS) of DSB formation (Thacker et al. 2014). Unexpectedly, this trend is not only abolished within *sae2*Δ but also reversed, in a manner independent of prophase I length (see: Figure 3.7C,E)—that is, within resection deficient strains, larger chromosomes form DSBs at higher densities than expected by size (in kb) alone.

Interestingly, *zip3*Δ mutants—within which homologue pairing is severely delayed—only exhibit a weakening of the negative correlation between hit density and chromosomal size, rather than a reversal (Thacker et al. 2014). Collectively, such observations may reveal a novel, inherent property of larger chromosomes that occurs through unknown mechanisms but which is specific to *sae2*Δ backgrounds.

Spo11 Double Cutting

Short range (<75bp) molecules occurring at discrete sizes with a 10bp periodicity and which exhibit distinct Spo11 sequence biases at either end are unexpectedly and readily observable within fully WT meioses when Spo11-oligo data is reanalysed (see: Figure 3.14, 3.19). The frequency of these molecules may increase in a *tel1*Δ background (see: Figure 3.18C,D). Moreover, data from *sae2*Δ-dependent mapping of Spo11 DSBs suggests the size of these 10bp periodic molecules increases upon loss of Tel1^{ATM} activity beyond 100-150bp (see: Figure 3.13). Long range molecules are however, contrastingly absent within *sae2*Δ samples. While the evidence presented in this chapter can not conclusively prove that these molecules represent legitimate Spo11-dependent double cuts, it is perhaps the most probable mechanism—implicating Tel1 within the confinement and suppression of long range Spo11 double cutting (Figure 3.25A). Loss of Tel1 activity also results in derepression of intragenic DSB formation at a genome-wide level (see: Figure 3.12), which may occur through long range double cutting that spreads in the direction of transcription through an unknown mechanism. As previously outlined, Tel1 may act redundantly with other factors, such as Mec1^{ATR}, to limit double cutting—accounting for the exacerbation of the long range double cut and intragenic derepression phenotypes observed in *sae2*Δ*tel1KD* compared to *sae2*Δ*tel1*Δ. A kinase dead allele may thus constitute a dominant negative mutant, whereby the presence of the inactive Tel1 protein blocks redundant pathways from acting. Whether or not Tel1 accomplishes suppression of double cut formation through mechanisms similar to those employed for DSB interference remains unclear.

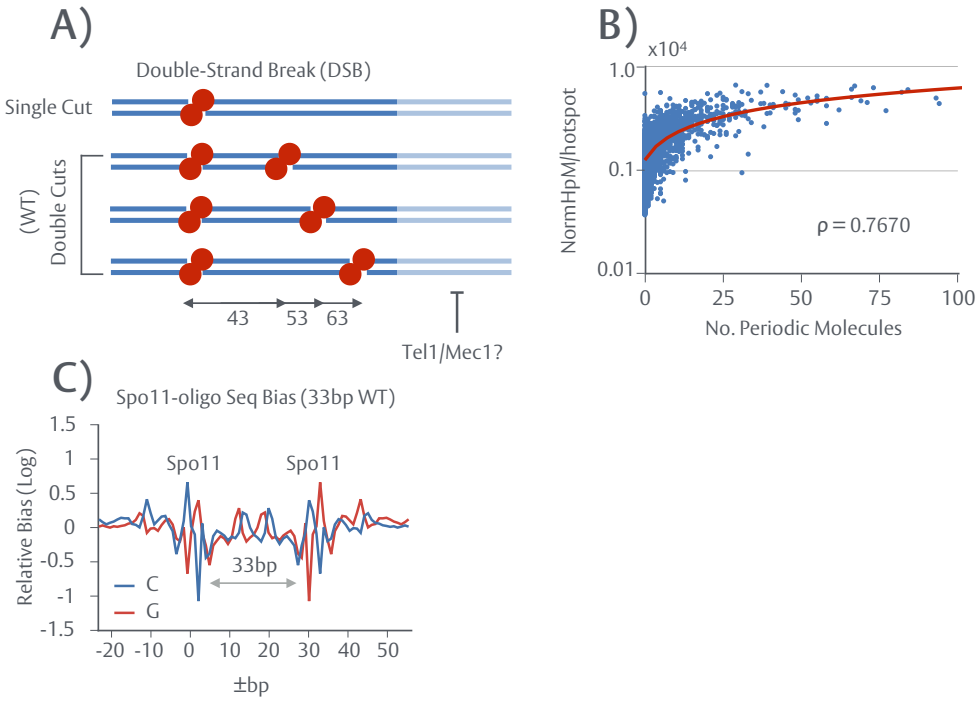


Figure 3.25. A model for Spo11 double cutting

A) Within WT cells, clustered, hyper localised Spo11 cleavage generates double cuts at discrete sizes with a 10bp periodicity (33, 43, 53, 63bp). Upon inactivation of Tel1 (*sae2*Δ*tel1*Δ), the range of double cutting increases beyond 100-125bp. Loss of Tel1 inactivity, but retention of the protein (*sae2*Δ*tel1KD*), appears to exacerbate this phenotype suggesting Tel1 ordinarily acts redundantly to confine and suppress the extent to which Spo11 double cutting occurs. **B)** The number of periodic molecules, displaying reciprocal coordinates, were tallied across each of the 3599 annotated *S. cerevisiae* hotspots and compared to the quantitative strength of each respective hotspot. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values). **C)** Base frequencies (% of A/G/C/T) were calculated, via *SeqBias*, flanking each set of coordinates on both strands for 33bp, reciprocal molecules.

Two models may explain why Spo11 double cutting occurs with a distinct periodicity. Notably, 10bp periodicity is reminiscent of the ~10.4bp helical rise of DNA, which exposes DNase-sensitive sites on the surface of nucleosomes (Brogaard et al. 2012; Cockell et al. 1983). Any exposed region, containing a secondary Spo11 compatible bias, may therefore be susceptible to double cut formation. However, it remains unclear whether or not Spo11 could access nucleosome-bound DNA for cleavage. Alternatively, Spo11 homodimers may be attached to a surface—such as the chromatin axis—with fixed orientation, as opposed to freely diffusible. Within this model, a DNA molecule is initially captured by a singular homodimeric unit within the accessible groove of the helix. Over short ~10-100bp ranges, DNA is notably rigid as described by its persistence length (~50nm, 150bp) (Manning 2006)—a mechanical property of polymers that describes the distance at which correlations in direction are lost i.e. bending may occur. Thus—due to the inflexibility of DNA over short ranges—if a secondary Spo11 homodimer captures the same DNA molecule, it will do so within the same groove at a precise distance from the first capture, governed by the ~10.4bp helical rise of DNA, thereby generating the observed periodicity. However, such a model may not account for the longer >100-150bp double cuts observed within *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* backgrounds.

Double cuts appear to be a common feature of DSB hotspots across the genome (see: Section 3.17). Indeed, the frequency of short range double cut molecules within any given hotspot is well correlated with the overall hotspot strength (NormHpM/hotspot) ($\rho = 0.7670$) (Figure 3.25B). A lack of detectable <33bp double cut molecules may reflect mappability issues of shorter molecules. However, sequence bias analysis of 33bp *filtered* molecules reveals a merging of the Spo11 5' and 3' bias patterns—suggesting a sterical limit to the size of double cut molecules, imposed by the DNA footprint of the Spo11 homodimer (Figure 3.25C). Consistent with a lower size limit, periodic bands below 33bp are not observed experimentally (D. Johnson, M.J. Neale, unpublished).

The seemingly irrepressible formation of short range (<75bp) double cuts within fully WT cells raises important implications for the canonical model of homologous recombination (HR) which

traditionally depicts a single initiating DSB (see: Figure 1.2). Hyperlocal double cutting would instead generate a double-stranded gap (DSG) concomitant with release of short, Spo11-capped double stranded molecules. Interestingly, conversion of singular DSBs into DSGs, via unspecified endo or exonucleases, was a previously proposed model for meiotic recombination (Szostak et al. 1983), that may now have to be re-examined closely. Within this model, both 5' and 3' ssDNA ends invade the homologous template and gap repair is accomplished through two rounds of single stranded repair synthesis, using the D-loop as a template (Szostak et al. 1983). Exactly why Tel1 is unable to repress short range <75bp double cuts—thereby suppressing gap formation—remains unclear. However, Spo11 double cuts at short range may arise coincidentally (i.e. at the same time), therefore precluding the ability to repress either event.

Mapping Topo II SSBs and DSBs

By adapting the *sae2*Δ-dependent method for Spo11 DSB mapping, and in conjunction with *Spo11Mapper*, SSBs and DSBs generated by topoisomerase II were mapped genome-wide with nucleotide resolution. Dependency upon Sae2, and to a larger extent, Mre11 for Topo II lesion repair—when exposed to etoposide—was demonstrated (see: Figure 3.20). As previously observed, Topo II lesions are observed to preferentially form within nucleosome depleted regions (NDRs) (Baranello et al. 2014). A previously undeterminable, symmetrical sequence bias was also observed for the formation of Topo II DSBs that is distinct from that of Spo11—despite a similar catalytic mechanism. Unlike Spo11 DSBs, promoter associated factors do not appear to be essential for Topo II SSB/DSB formation as evidenced by considerable levels of signal within promoter-less convergent regions (see: Figure 3.21C). Nevertheless, promoter regions are still substantially enriched for signal over terminators (see: Figure 3.22). Strains deficient in Topo II activity exhibit a global downregulation in gene expression, at the level of transcription (Pedersen et al. 2012). Efficient transcription requires the relaxation of negative and positive supercoils, which flank RNA polymerase during transcriptional elongation, by topoisomerases (Pedersen et al. 2012). Moreover, topoisomerases are required for transcription initiation complex assembly (Roedgaard et al. 2015).

Early recruitment of Topo II to transcriptional units may therefore account for the disparity of signal strength at promoters, relative to terminators.

3.23—Summary (Key Points)

- Developed and validated a novel analytical package for the alignment and processing of genome-wide, *sae2*Δ-dependent Spo11 DSB data (*Spo11Mapper*) (Section 3.3-3.4)
- Spo11 DSBs preferentially form within nucleosomally depleted, promoter regions (Section 3.9)
- A symmetrical, improved Spo11 DSB sequence bias can be obtained (Section 3.10)
- Tel1^{ATM} is required for suppression of intragenic DSB formation (Section 3.11)
- Loss of Tel1^{ATM} activity results in a spreading of DSB signal in the direction of transcription and formation of long range (>100bp) molecules displaying a 10bp periodicity (Section 3.12-3.13)
- Short range (<75bp) Spo11 double cutting may occur within WT cells (Section 3.14-3.15)
- *Spo11Mapper* was expanded to the mapping of Topo II DSBs and SSBs (Section 3.18)
- Topo II lesions preferentially form within nucleosomally depleted regions (Section 3.19)
- Topo II exhibits a weak, symmetrical sequence bias for the generation of DSBs but not SSBs (Section 3.21)

CHAPTER 3B

Genome-wide mapping of Spo11 DSBs

Appendix

Appendix

B3.1—*Spo11Mapper* (*v2.7*)

Aim: Alignment, filtering and analysis of genome-wide Spo11-mapping data

Input(s): Paired-end _R1 and _R1 FASTQs, indexed reference genome (FASTA), user configuration

Output(s): 1bp Histograms, molecule sizes, molecule frequencies, alignment logs

Req(s): Perl 5.25.9, BioPerl, Bash 4.1

Spo11Mapper constitutes a novel, low memory software package for the automated batch processing of *sae2*Δ Spo11 DSBs, Topo II DSB/SSBs, Spo11-oligos or any library containing informative Read-1/2 5' ends. *Spo11Mapper*, run on the command line, requires two main arguments:

```
Usage: Spo11Mapper -i [INPUT FOLDER] -c [CONFIGURATION FILE]
-i INPUT: Input data folder containing paired-end FASTQ files
-c CONFIG: Configuration file specifying user-parameters (Spo11Mapper.config)
```

B3.1.1—Configuration

Spo11Mapper is initially configured via an external .config file specifying key variables:

(i) *CALL_MODE* (*SINGLE*/*DOUBLE*/*OLIGO*)—specification of library type which in turn differentiates the type of coordinates called and the analyses performed (single cuts, double cuts, oligos)

(ii) *SPACE_SAVER* (*Y*/*N*)—when enabled, *Spo11Mapper* will reduce the disk footprint of the pipeline, progressively deleting non-essential files including .SAM and .FASTQ files

(iii) *CORE*—no. of CPU cores available

(iv) *READ1/2_EXT*—FASTQ file extension for automated detection of paired end samples

(v) *GLOBAL_OPTIONS—s*pecification of Bowtie2 parameters for end-to-end alignment. Default settings, utilised throughout this chapter, are (-X 1000 --no-discordant --very-sensitive --mp 5,1)

(vi) *LOCAL_OPTIONS*—specification of Bowtie2 parameters for local alignment (default as above)

(vii) *GENOME_DIR*, *GENOME–NAME*—directory and filename of a Bowtie2 indexed FASTA reference genome. Data throughout this chapter was aligned against a modified S288c reference (SGD Jan 2015 - R64-2-1) containing sequence for the exogenous hotspots, *HIS4*::*LEU2* and *LEU2*::*HISG*.

(viii) *TRIM*, *TRIM_LEN*—when enabled, *Spo11Mapper* will perform a two step alignment (see: Section 3.4), 3'→5' trimming unmapped mates by a length specified by TRIM_LEN.

Spo11Mapper.config - Example

```
################################################################
## Program Settings
################################################################
CALL_MODE = DOUBLE
SPACE_SAVER = N
CORE = 4
################################################################
## Input Data
################################################################
READ1_EXT = _R1
READ2_EXT = _R2
################################################################
## Alignment Options
################################################################
GLOBAL_OPTIONS = -X 1000 --no-discordant --very-sensitive --mp 5,1
LOCAL_OPTIONS = -X 1000 --no-discordant --very-sensitive --mp 5,1
GENOME_DIR = /usr/local/Genomes/Cer3H4L2/
GENOME_NAME = Cer3H4L2
TRIM = Y
TRIM_LEN = 10
```

B3.1.2—SAM files

Tab delimited SAM files, the primary output of *Bowtie2* alignment, specify (i) 1-based leftmost coordinates (by Chr, Position) for all mapped or partially mapped read pairs (ii) the mapped or unmapped read sequence with associated quality scores (iii) Numerical "flags" denoting the aligned identity of each read pair:

Paired, *fully aligned* (*SAM Flags*)

99 - Read-1, Watson | 147 - Read-2, Crick

83 - Read-1 Crick | 163 - Read-2, Watson

Paired, *partially aligned* (*one–mate*)

73 - Read-1, Watson, Mapped | 133 - Read-2, Unmapped

89 - Read-1, Crick, Mapped | 133 - Read-2, Unmapped

69 - Read-1, Unmapped | 137 - Read-2, Watson, Mapped

69 - Read-1, Unmapped | 153 - Read-2, Crick, Mapped

(iv) Alpha-numeric CIGAR codes describing base by base alignment (5'->3') and detailing the presence of INDELs. SNPs/mismatches are not included. For example, a CIGAR code of *5M2I30M1D25M* denotes:

- 5bp reference match (5M) ("M" may contain unspecified mismatches)
- 2bp insertion in the read (relative to the reference) (2I)
- 30bp reference match (30M)
- 1bp deletion in the read (relative to the reference) (1D)
- 25bp reference match (25M)

(v) Alpha-numeric MD:Z tags denoting the position and base composition of any SNPs/deletions present in the read relative to the reference. Insertions (relative to the reference) are not specified. For example, an MD:Z-tag of 0T0C25A5^T10 denotes (from left to right):

- An initial 2bp mismatch (reference specifies TC, read contains alternative bases) (0T, 0C)
- 25bp of precise reference:read match (25) followed by a 1bp mismatch (25A)
- 5bp of precise reference:read match (5) followed by a 1bp deletion in the read (reference contains a T) (5^T)
- 10bp of precise reference:read match (10)

ATGAGCGTACCTGTAAATAAGAAGATC**G**ATCGA_GGTACATACT — *READ* (*0T0C25A5^T10*)

TCGAGCGTACCTGTAAATAAGAAGATC**A**ATCGA**T**GGTACATACT — *REF*

Spo11Mapper collectively reads and interprets this information read-by-read for each .SAM file to accurately filter and extract high quality coordinate lists and facilitate read trimming.

B3.1.3—Orientation and ambiguous end filtering

Coordinate calling is only performed for properly paired, fully aligned 99-147 (Read-1 Watson—Read-2 Crick) or 83-163 (Read-1 Crick—Read-2 Watson) read pairs that pass ambiguous end and orientation checks. Atypical read orientations can arise through the sequencing of self circles or incomplete sequencing runs, generating asymmetrically overlapping read pairs. SAM "flags", MD:Z-tags or CIGAR codes contain no information pertaining to this phenomenon—thus, *Spo11Mapper* employs a custom filter using the leftmost positional information held in .SAM files. Any 99/147 or 83/163 read pairs where the 3' end of the (-) Crick read is to the left (upstream) of the 5' end of the (+) Watson read—signifying asymmetric overlap—is discarded. Ambiguous ends are determined through parsing and interpretation of MD:Z-tags, detecting mismatches at any informative end. In SINGLE mode, 2bp of mismatch or more at the Read-1 5' end disqualifies a read and enters it into a separate, ambiguous coordinate dataset. A single terminal base of mismatch is permitted to accommodate alignment of data from divergent strains (e.g. SK1 to S288c). In DOUBLE mode, the same threshold is applied to both Read-1 and Read-2 5' ends. If either end is ambiguous, the entire read pair is disqualified and separated. Non-informative 3' ends are not considered. A MD:Z-tag of *0T0C73* (2bp 5' mismatch) would thus fail the ambiguous end check, while a tag of *0C74* (1bp 5' mismatch) or 74*T* (1bp 3' mismatch) would pass.

B3.1.4—Coordinate calling

For properly oriented, unambiguous 99-147 and 83-163 read pairs, coordinate positions of the informative (e.g. Spo11) ends are calculated. SAM files specify 1-based leftmost coordinates—thus, for Watson (+) reads (99 or 163) the 5' end is readily called by *Bowtie2*. In contrast, for Crick (-) reads (83 and 163), the leftmost base is the 3' end of the read. To call 5' Crick (-) coordinates, CIGAR codes are parsed and scored to determine the mapped read length—according to the following rules: (M = 1, D = 1, I = 0)—which is then added to the 3' coordinate. Insertions (I) (in the read) are ignored in order to call coordinates accurate to the utilised reference. As an example, a Crick (-) read with a CIGAR code of 75M and a leftmost coordinate is 10200 is called as 102074

(10200+75-1). A 1bp adjustment is made as the leftmost base is included as part of 75M. A more complex Crick (-) read with a CIGAR code of 35M2D10M3I30M and a leftmost coordinate of 10200 is called as 10276 (10200 + 35 + 2 + 10 + 30 -1).

B3.1.5 —3'->5' Trimming

If enabled (TRIM=Y), *Spo11Mapper* additionally handles SAM flags 73-133, 89-133, 69-137 and 69-153—all of which denote pairs of reads where only a single mate mapped. During first round processing of SAM files, trimmed FASTQ files are reconstructed by *Spo11Mapper* in a standard, four line format:

@M00561:9:000000000-ALBWJ:1:1101:15097:1775 1:N:0:1 - Read Header
TAATGAATTAATCAACTTCAACTCATCACTGCCCAATGATTCGTCGGGTTTCACTATTTTTAGATAATCTTCCCT - Seq
+
@—A---CE,,CC,,;EEE,,;CC,C;,;C,<;;,,,;,,<,<,;@+++,886,<C,<@CEF,,,<,,<<@C,;CC - Quality

Mapped mates (73, 89, 137 and 153) are sorted into their respective Read-1 or Read-2 trimmed FASTQ files "as is", without trimming. SAM files store mapped read sequences based on the top (+) strand, regardless of which strand the read aligned to. Sequences for mapped Crick (-) mates (89, 153) are thus reverse complemented before addition to a FASTQ file. Read sequences for unmapped mates (133, 69) are trimmed from the 3' end as are the associated quality lines. SAM files store the actual read sequence for unmapped mates, thus no further processing is required. Untrimmed FASTQ files are subsequently auto detected for all samples and entered into *-local Bowtie2* alignment. Local alignment mode is less strict, however it also permits Bowtie2 to trim reads from *either* end if it improves MAPQ (quality) scores. Any *-local* trimming is recorded into the CIGAR code as 'S'. As *-local* trimming at the 5'-end lowers the integrity of coordinate calling, ambiguous end filtering is expanded to disqualify any reads with 'S' CIGAR entries at an informative end (e.g. 4S71M). Any read pairs that now sufficiently align will be marked by 99-147 or 83-163 flags, undergo coordinate calling if unambiguous and properly oriented as above and are appended onto the main dataset.

B3.1.6—Aligning Spo11-oligos

Alignment of ~25-65bp Spo11-oligos requires additional pre-alignment processing. Any oligo or double cut molecule shorter than the sequencing run length (i.e. 75bp) will contain poly(G/C) tails —added during Spo11-oligo library prep—and portions of illumina adaptor at both ends of the associated read. In order to improve mappability of Spo11-oligos, *Spo11Mapper* includes a utility script (*OligoTrim.pl*) capable of trimming FASTQ files according to user specified sequence patterns:

perl OligoTrim.pl –1 <Read 1 FASTQ> –2 <Read 2 FASTQ> –U <5' Upstream Pattern> –L <3' Downstream Pattern>

Multiple patterns may be specified. All Spo11-oligo libraries analysed throughout this chapter were trimmed using -U CCCC -U CCC -L GGGG -L GGG, to accommodate situations where poly(G) tailing was incomplete.

B3.1.7—Log Files

Spo11Mapper records several log files: (i) Individual alignment reports per sample, including a summary of coordinate calling (ii) Batch, strain summary, detailing stats for each sample processed (see: Figure 3.4) (iii) System log, detailing errors and recording the command-line output.

Alignment Report (Individual Sample) - Example

```
MJ315_WT_2A_6h
----------------------
GLOBAL
----------------------
3512089 reads; of these:
  3512089 (100.00%) were paired; of these:
    164323 (4.68%) aligned concordantly 0 times
    3190282 (90.84%) aligned concordantly exactly 1 time
    157484 (4.48%) aligned concordantly >1 times
    164323 pairs aligned 0 times concordantly or discordantly; of these:
      328646 mates make up the pairs; of these:
        215163 (65.47%) aligned 0 times
        89083 (27.11%) aligned exactly 1 time
        24400 (7.42%) aligned >1 times
96.94% overall alignment rate
```

```
----------
TRIMMED
----------
68175 reads; of these:
  68175 (100.00%) were paired; of these:
    14152 (20.76%) aligned concordantly 0 times
    27312 (40.06%) aligned concordantly exactly 1 time
    26711 (39.18%) aligned concordantly >1 times
    ----
    14152 pairs aligned 0 times concordantly or discordantly; of these:
      28304 mates make up the pairs; of these:
        10822 (38.23%) aligned 0 times
        10900 (38.51%) aligned exactly 1 time
        6582 (23.25%) aligned >1 times
92.06% overall alignment rate

----------
CALL STATS
----------
Total Hits:         3401787
----------
Valid Hits:         3377298
  Global:           3341250
  Trimmed:          36048
----------
Ambig Hits:         24489
```

System Log - Example

```
-------------------------------------------
FASTQ Alignment (Global --end-to-end)
-------------------------------------------
Currently aligning:
MJ315_WT_2A_6h
---------------------------------------
Calculating Coordinates....
---------------------------------------
Currently processing:
MJ315_WT_2A_6h
----------
Completed
Runtime: 0:05:50
```

```
-------------------------------------------
FASTQ Alignment (Trimmed --local)
-------------------------------------------
Currently realigning:
MJ315_WT_2A_6h
---------------------------------------
Calculating Coordinates....
---------------------------------------
Currently processing:
MJ315_WT_2A_6h
----------
Completed
Runtime: 0:00:14
----------
```

B3.1.8—Analysis and output files

Spo11Mapper generates several key output files: (i) all raw unambiguous, coordinate calls are recorded into tab delimited .txt files, which are in turn utilised for downstream analysis:

Single Cut Library:

Strand	Chr	Pos	ReadLength	CIGAR	Adjustment	Read-Flag
w	9	204982	75	75M	0	99
c	15	1059148	75	75M	74	83
c	6	36193	75	75M	74	83
c	6	221802	73	73M	72	83
c	6	226858	75	43M1D32M	75	83

Double Cut Library (*Paired W–C Lines*, *Molecule Size Added*):

PairID	Strand	Chr	Pos	ReadLength	CIGAR	Adjustment	Read-Flag	mSize
1	w	7	307254	19	19M	0	99	21
1	c	7	307274	19	19M	18	147	21
2	w	13	577501	23	23M	0	99	25
2	c	13	577525	23	23M	22	147	25
3	w	15	770237	25	25M	0	163	27
3	c	15	770263	25	25M	24	83	27

Ambiguous, filtered calls are stored in separate but identically formatted files (ii) Sparsely-formatted, 1bp histograms are produced for Read-1 5' ends, and separately, Read-2 5' ends (if detected) and stored as tab delimited .txt files, detailing the total number of Watson (+) and Crick (-) per base pair on each chromosome (iii) In DOUBLE mode, *Spo11Mapper* calculates the frequency of molecule sizes as the absolute distance between Read-1 5'- and Read-2 5' ends for each unambiguous read pair. Moreover, the system records the frequency with which any given pair of 5' coordinates is observed. Molecule sizes and frequencies are stored as tab delimited .txt files:

1bp Histogram:

Chr	Pos	Watson	Crick
1	6457	32	0
1	6458	0	32
1	6463	0	1
1	6468	2	10
1	6469	0	1

Molecule Sizes:

Size	Freq
20	42104
21	53755
22	62272
23	69923
24	82517

Molecule Frequencies:

Chr	Coord-A	Coord-B	R1W Freq	R1C Freq
1	2434	2460	1	5
1	2434	2461	0	6
1	2434	2462	5	12
1	2434	2464	0	2
1	2434	2465	2	0

B3.1.9—Sequence bias

Spo11Mapper includes a utility script (*SeqBias.pl*) capable of determining per base (A/G/C/T) frequencies flanking a given set of coordinates. *SeqBias* directly samples a user provided FASTA reference genome file (e.g. Cer3H4L2) to pileup and calculate per base frequencies (A/G/C/T) for a given ±bp width, centred on Watson (+) and Crick (-) coordinates listed within a 1bp histogram file generated by *Spo11Mapper*:

perl SeqBias.pl –i <Histogram File> –r <Reference FASTA> –w <Width> –m <Mode> –o <Output File>

For Watson (+) hits, *SeqBias* samples the reference FASTA "as is", without further processing. As FASTA files specify the (+) strand, *SeqBias* reverse complements sequences flanking Crick (-) hits. *SeqBias* provides two analysis modes: (i) *Pos*—under this mode, *SeqBias* calculates biases flanking all specified nucleotides in an unweighted manner (ii) *Freq*—under this mode, *SeqBias* weights biases by the number of hits present at any given nucleotide. Sites with stronger signal will therefore be more heavily represented within the resulting bias. All biases produced throughout this chapter were generated under *Pos* mode. A population averaged bias is calculated proceeding 1bp histogram processing, and provided in a tab delimited file.

Script—Spo11Mapper.sh (Command-line tool, automation, logs)

```
#!/usr/bin/env bash
#Package Version: 2.0

################################################################################################
# Author(s): T.J.Cooper
# Updated: 15/2/2017
# Automates batch-processing of FASTQ/SAM files for genome-wide mapping
################################################################################################

function usage {
  echo -e "\nUsage: Spo11Mapper -i [INPUT FOLDER] -c [CONFIGURATION FILE]\n";
  echo -e "-i INPUT: Input data folder containing paired-end FASTQ files\n"
  echo -e "-c CONFIG: Configuration file specifying user-parameters (Spo11Mapper.config)\n"; exit 1;
}

# Command-line Options
CONF=""
INPUT=""
while getopts ":c:i:" FLAG; do
  case $FLAG in
    i) INPUT=$OPTARG;;
    c) CONF=$OPTARG;;
    \?) echo -e "\nInvalid option: -$OPTARG"
      usage;;
    :) echo -e "\nOption -$OPTARG requires an argument."
      usage;;
  esac
done
if [ "$#" -eq 0 ]; then
   usage
fi

# Read Config File
BASEDIR=$( cd "$( dirname "${BASH_SOURCE[0]}" )" && pwd )
cd "$INPUT" || exit
mkdir -p "$INPUT/Logs" || exit
exec > >(tee "$INPUT/Logs/System.log")
shopt -s extglob
declare -A config=()
while IFS='=' read -r k v; do
  [[ $v ]] || continue
  [[ $k = "#"* ]] && continue
  k=${k%+([[:space:]])}
  v=${v#+([[:space:]])}
  config[$k]=$v
done < "$CONF"
if [ ${config[CALL_MODE]} = SINGLE ]; then
  PERLDIR=$BASEDIR/Scripts/SingleEndExtract.pl
  ALIGN_MODE="--end-to-end"
elif [ ${config[CALL_MODE]} = DOUBLE ]; then
  PERLDIR=$BASEDIR/Scripts/DualEndExtract.pl
  ALIGN_MODE="--end-to-end"
elif [ ${config[CALL_MODE]} = OLIGO ]; then
  PERLDIR=$BASEDIR/Scripts/DualEndExtractOligo.pl
  ALIGN_MODE="--end-to-end"
fi
```

```
if [[ ! -f $PERLDIR || ! -d ${config[GENOME_DIR]} || -z ${config[GENOME_NAME]} || -z ${config[READ1_EXT]} || -z
${config[READ2_EXT]} || -z ${config[TRIM]}
|| -z ${config[TRIM_LEN]} || -z ${config[CORE]} || -z ${config[SPACE_SAVER]} ]]; then
   echo -e "\nError: User parameters or script files are missing and/or incorrectly specified.\n\n"; exit -1
fi
if [ ${config[TRIM]} = Y ] && [ ${config[CALL_MODE]} = OLIGO ]; then
  echo -e "\nTrimmed alignment is not available in OLIGO mode.\n"; exit -1
fi

# Input Files
declare -A STR
shopt -s nullglob
for file in *${config[READ1_EXT]}.*; do
  STR["${file%*${config[READ1_EXT]}.*}"]=1
done
if [ ${#STR[@]} = 0 ]; then
  echo -e "\nNo valid FASTQ files were found."
  usage
fi

# Global Alignment
S=$(date +%s)
echo "-----------------------------------------------"
echo "FASTQ Alignment (Global $ALIGN_MODE)"
echo "-----------------------------------------------"
echo "Currently aligning:"
for i in "${!STR[@]}"; do
  echo ${i}
  printf "%s\n%s\n%s\n%s\n" "$i" "-----------------------" "GLOBAL" "-----------------------" > "$INPUT/Logs/$i.txt"
     bowtie2 "$ALIGN_MODE" ${config[GLOBAL_OPTIONS]} -p ${config[CORE]} -x ${config[GENOME_DIR]}/$
{config[GENOME_NAME]} -1 $i${config[READ1_EXT]}.fastq -2 $i${config[READ2_EXT]}.fastq -S $i"_Global".SAM
2>> "$INPUT/Logs/$i.txt"
  if [ ${config[SPACE_SAVER]} = Y ]; then
    rm $i${config[READ1_EXT]}.fastq
    rm $i${config[READ2_EXT]}.fastq
  fi
done
perl "$PERLDIR" "Global.SAM" ${config[SPACE_SAVER]} ${config[GENOME_NAME]} ${config[TRIM]} $
{config[TRIM_LEN]} ${config[READ1_EXT]} ${config[READ2_EXT]}
E=$(date +%s)
i=$((E-S)); ((sec=i%60, i/=60, min=i%60, hrs=i/60))
runtime=$(printf "%d:%02d:%02d" $hrs $min $sec)
echo "-----------------------";
echo -e "Completed\nRuntime: $runtime";
echo -e "-----------------------\n\n";

# Trimmed Alignment
if [ ${config[TRIM]} = Y ]; then
  echo "-----------------------------------------------"
  echo "FASTQ Alignment (Trimmed --local)"
  echo "-----------------------------------------------"
  echo "Currently realigning:"
  for n in "${!STR[@]}"; do
    echo ${n}
    printf "\n%s\n%s\n%s\n" "-----------------------" "TRIMMED" "-----------------------" >> "$INPUT/Logs/${n}.txt"
              bowtie2 --local ${config[LOCAL_OPTIONS]} -p ${config[CORE]} -x ${config[GENOME_DIR]}/$
{config[GENOME_NAME]} -1 ${n}${config[READ1_EXT]}"_unmapped_trimmed.fastq" -2 ${n}$
{config[READ2_EXT]}"_unmapped_trimmed.fastq" -S $n"_Trimmed".SAM 2>> "$INPUT/Logs/${n}.txt"
    if [ ${config[SPACE_SAVER]} = Y ]; then
```

```
      rm ${n}${config[READ1_EXT]}"_unmapped_trimmed.fastq"
      rm ${n}${config[READ2_EXT]}"_unmapped_trimmed.fastq"
    fi
  done
  perl "$PERLDIR" "Trimmed.SAM" ${config[SPACE_SAVER]} ${config[GENOME_NAME]}
  ET=$(date +%s)
  it=$((ET-E)); ((sec=it%60, it/=60, min=it%60, hrs=it/60))
  runtime=$(printf "%d:%02d:%02d" $hrs $min $sec)
  echo "----------------------";
  echo -e "Completed\nRuntime: $runtime";
  echo -e "----------------------\n\n";
fi

# Directory Organisation
if [ ${config[SPACE_SAVER]} = N ]; then
  mkdir -p "$INPUT/FASTQ"
  mkdir -p "$INPUT/SAM"
  mv *.fastq ./FASTQ
  mv *.SAM ./SAM
fi

# Histogram Maps & Statistics
echo "-------------------------------------------"
echo "Generating Maps and Statistics...."
echo -e "-------------------------------------------\n"
printf "%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\n" "Strain" "Total Read Pairs (A)" "Total Mapped Pairs (B)" "% of (A)"
"Multimapping Pairs" "% of (B)" "Valid Hits" "% of (B)" "Ambiguous Hits" "% of (B)" > "$INPUT/Logs/
StrainSummary.txt"
cd "$INPUT/Coordinates" || exit
for k in "${!STR[@]}"; do
  printf "\n%s\n%s\n%s\n" "----------------------" "CALL STATS" "----------------------" >> "$INPUT/Logs/${k}.txt"
  TotalRead=$(awk '/reads; of these/ && ! seen {print $1; seen=1}' < "$INPUT/Logs/${k}.txt")
  MultiMap=$(awk '/concordantly >1/ {sum+=$1} END{print sum}' < "$INPUT/Logs/${k}.txt")
    MappedRead=$(($(awk '/concordantly exactly/ {sum+=$1} END{print sum}' < "$INPUT/Logs/${k}.txt") +
$MultiMap))
  Ambig=$(($(wc -l < Coordinates.${config[GENOME_NAME]}_${k}_Ambiguous.txt)-1))
  Global=$(($(wc -l < Coordinates.${config[GENOME_NAME]}_${k}_Global.txt)-1))
  if [ ${config[TRIM]} = Y ]; then
    awk 'FNR==1 && NR!=1{next;}{print}' Coordinates.${config[GENOME_NAME]}_${k}_Global.txt Coordinates.
${config[GENOME_NAME]}_${k}_Trimmed.txt > Coordinates.${config[GENOME_NAME]}_${k}_Combined.txt
    Trimmed=$(($(wc -l < Coordinates.${config[GENOME_NAME]}_${k}_Trimmed.txt)-1))
    ValidHits=$(($Global + $Trimmed))
        printf "%s\t%d\n%s\n%s\t%d\n%11s\t%d\n%11s\t%d\n%s\n%s\t%d\n" "Total Hits:" "$(($Global + $Trimmed +
$Ambig))" "----------------------" "Valid Hits:" "$ValidHits" "Global:" "$Global" "Trimmed:" "$Trimmed"
"----------------------" "Ambig Hits:" "$Ambig" >> "$INPUT/Logs/${k}.txt"
  elif [ ${config[TRIM]} = N ]; then
    ValidHits=$Global
          printf "%s\t%d\n%11s\t%d\n%s\n%s\t%d\n" "Total Hits:" "$(($Global + $Ambig))" "Valid Hits:" "$Global"
"----------------------" "Ambig Hits:" "$Ambig" >> "$INPUT/Logs/${k}.txt"
  fi
  if [ ${config[CALL_MODE]} = DOUBLE ] || [ ${config[CALL_MODE]} = OLIGO ]; then
    ValidHitsPerc=$(bc -l <<< "(($ValidHits/2)/$MappedRead)*100")
    AmbigPerc=$(bc -l <<< "(($Ambig/2)/$MappedRead)*100")
  elif [ ${config[CALL_MODE]} = SINGLE ]; then
    ValidHitsPerc=$(bc -l <<< "($ValidHits/$MappedRead)*100")
    AmbigPerc=$(bc -l <<< "($Ambig/$MappedRead)*100")
  fi
```

```
    printf "%s\t%d\t%d\t%.3f\t%d\t%.3f\t%d\t%d\t%.3f\t%d\t%.3f\n" "${k}" "$TotalRead" "$MappedRead" "$(bc -l <<< "($MappedRead/$TotalRead)*100")" "$MultiMap" "$(bc -l <<< "($MultiMap/$MappedRead)*100")" "$ValidHits" "$ValidHitsPerc" "$Ambig" "$AmbigPerc" >> "$INPUT/Logs/StrainSummary.txt"
done
cd "$INPUT" || exit
mkdir -p "$INPUT/Histograms"
perl "$BASEDIR/Scripts/HistogramMap.pl" "-i" "$INPUT/Coordinates" "-g" ${config[GENOME_NAME]}
```

Script—SingleEndExtract.pl (Single cut processing)

```
#!/usr/bin/env perl
#Package Version: 2.0

##############################################################################################
# Author(s): T.J.Cooper
# Updated: 15/2/2017
# Processes paired-end .SAM files, extracting Watson + Crick coordinate information for single-cut Spo11 and Topo-II libraries
# Quality-control and filtering (atypical read-orientation, dubious ends)
# Two-step alignment (unmapped mate read-trimming, --local alignment)
##############################################################################################

use strict;
use warnings;
use Cwd;
use List::Util qw(first);
local $| = 1;
my $outext = '.txt';          #Output .file-extension
my $inext  = $ARGV[0];        #Input .file-extension
my @files  = glob("*$inext");
my $chk    = scalar(@files);
print "\nFailed to detect any .SAM files within the current directory.\n\n"
  if $chk == 0;
exit if $chk == 0;            #Stop script if no .SAM files are found
my $sub = cwd() . "/Coordinates";
mkdir("$sub") unless $chk == 0;
my $genome    = $ARGV[2];
my $trimmode  = $ARGV[3];
my $trimlength = $ARGV[4];
my ( $ghits, $ahits );
print "---------------------------------------";
print "\nCalculating Coordinates....\n";
print "---------------------------------------\n";
print "Currently processing:\n";

for my $file (@files) {       #For-each input file
   open my $IN, '<', $file or die "$!";   #Open and read input .SAM file(s)
   ( my $strain = $file ) =~ s/_[^_]+$//;  #Strain-name
   ( my $mode   = $ARGV[0] ) =~ s/\.SAM//; #Alignment-mode
   print "$strain\n";
   my $outfile =
      "Coordinates."
    . $genome . "_"
    . $strain . "_"
    . $mode
    . $outext;                    #Output files
   my $outfile2 =
     "Coordinates." . $genome . "_" . $strain . "_Ambiguous" . $outext;
```

```
  my ( $OUT, $OUT2, $OUT3, $OUT4 );
  open $OUT,  '>',  "$sub/$outfile"  or die "$!";
  open $OUT2, '>>', "$sub/$outfile2" or die "$!";
  print $OUT "Strand\tChr\tPos\tReadLength\tCIGAR\tAdjustment\tRead-Flag\n";

  if ( $inext eq "Global.SAM" && $trimmode eq "Y" ) {
    print $OUT2
"Strand\tChr\tPos\tReadLength\tCIGAR\tAdjustment\tRead-Flag\tMD-Tag\n";
    my $outfile3 =
      $strain
     . $ARGV[5]
     . "_unmapped_trimmed.fastq";   #Unmapped R1 FASTQ file
    my $outfile4 =
      $strain
     . $ARGV[6]
     . "_unmapped_trimmed.fastq";   #Unmapped R2 FASTQ file
    open $OUT3, '>', "$outfile3" or die "$!";
    open $OUT4, '>', "$outfile4" or die "$!";
  }
  while (<$IN>) {                  #For-each .SAM record
    chomp $_;
    next if /^\s*@/;               #Skip .SAM headerlines
    my @F = split( "\t", $_ );     #Split each tab-delimited field
    my $orientation = $F[3] - $F[7];  #Discard atypical read-orientations
    if (   $F[1] == 99 && $orientation > 0
      || $F[1] == 83 && $orientation < 0 )
    {
      my $skipline = <$IN>;
      next;
    }
    if ( $inext eq "Global.SAM" && $trimmode eq "Y" )
    {   #Populate unmapped R1/R2 FASTQ files mapped-unmapped pairs
      if ( grep { $_ == $F[1] } 73, 137 ) {
        print $OUT3 "\@$F[0] 1:N:0:1\n$F[9]\n+\n$F[10]\n"
          if $F[1] == 73;
        print $OUT4 "\@$F[0] 1:N:0:1\n$F[9]\n+\n$F[10]\n"
          if $F[1] == 137;
      }
      if ( grep { $_ == $F[1] } 89, 153 ) {
        $F[9] =~ tr/GATC/CTAG/;
        my $revseq  = reverse( $F[9] );
        my $revqual = reverse( $F[10] );
        print $OUT3 "\@$F[0] 1:N:0:1\n$revseq\n+\n$revqual\n"
          if $F[1] == 89;
        print $OUT4 "\@$F[0] 1:N:0:1\n$revseq\n+\n$revqual\n"
          if $F[1] == 153;
      }
      if ( grep { $_ == $F[1] } 69, 133 ) {
        my $trimseq  = substr( $F[9],  0, $trimlength );
        my $trimqual = substr( $F[10], 0, $trimlength );
        print $OUT3 "\@$F[0] 1:N:0:1\n$trimseq\n+\n$trimqual\n"
          if $F[1] == 69;
        print $OUT4 "\@$F[0] 1:N:0:1\n$trimseq\n+\n$trimqual\n"
          if $F[1] == 133;
      }
    }
    if ( grep { $_ == $F[1] } 99, 83 ) {
      my $index = first { /MD:Z/ } @F;  #Obtain variable-column MD:Z: tag
      my @MDtag = $index =~ /\d+/g;     #Remove non-numeric characters
```

```
        my @revMDtag = reverse(@MDtag);
        my %rules = ( M => 1, D => 1, I => 0, S => 1 )
         ;   #Rules to handle insertion/deletions/matches/soft-clipping
        my ( $s, $LS, $RS ) = (0) x 3;
        while ( $F[5] =~ /(\d+)([MDIS])/g )
        {     #Parse and interpret CIGAR code
           my ( $n, $op ) = ( $1, $2 );
           $s += $n * $rules{$op}
            unless $op eq
            'S';   #Calculate POS adjustment (insertions/deletions)
           $LS += $n * $rules{$op}
            if $op eq 'S' && $-[0] == 0;   #(upstream soft-clip)
           $RS += $n * $rules{$op}
            if $op eq 'S'
            && $+[0] == length( $F[5] );   #(downstream soft-clip)
        }
        my $l  = length( $F[9] );   #Read-length
        my $wp = $F[3] - $LS;       #Adjusted 5' coordinate (Watson strand)
        my $cp =
         $F[3] + ( $RS + $s ) - 1;  #Adjusted 5' coordinate (Crick strand)
        if (   $MDtag[0] == 0 && $MDtag[1] == 0 && $F[1] == 99
           || $LS > 1 && $F[1] == 99 )
        {                    #Detect ambigious ends (Watson strand)
           printf( $OUT2 "%s\t%d\t%d\t%d\t%s\t%d\t%s\t%s\n",
              "w", $F[2], $wp, $l, $F[5], 0 - $LS, "99", $index );
        }
        elsif ( $F[1] == 99 && $wp > 0 ) {
           printf( $OUT "%s\t%d\t%d\t%d\t%s\t%d\t%s\n",
              "w", $F[2], $wp, $l, $F[5], 0 - $LS, "99" );
        }
        if (   $revMDtag[0] == 0 && $revMDtag[1] == 0 && $F[1] == 83
           || $RS > 1 && $F[1] == 83 )
        {                    #Detect ambigious ends (Crick strand)
           printf( $OUT2 "%s\t%d\t%d\t%d\t%s\t%d\t%s\t%s\n",
              "c", $F[2], $cp, $l, $F[5], $RS + $s - 1,
              "83", $index
           );
        }
        elsif ( $F[1] == 83 && $cp > 0 ) {
           printf( $OUT "%s\t%d\t%d\t%d\t%s\t%d\t%s\n",
              "c", $F[2], $cp, $l, $F[5], $RS + $s - 1, "83" );
        }
     }
  }
  if ( $ARGV[1] eq "Y" ) {
     chomp($file);
     unlink($file);
  }
}
```

Script—DualEndExtract.pl (Double cut processing, molecule sizes, molecule frequencies)

```
#!/usr/bin/env perl
#Package Version: 2.0

###############################################################################################
# Author(s): T.J.Cooper
# Updated: 15/2/2017
# Processes paired-end .SAM files, extracting Watson + Crick coordinate information for double-cut Spo11 libraries
# Quality-control and filtering (atypical read-orientation, dubious ends)
# Two-step alignment (unmapped mate read-trimming, --local alignment)
# Calculates inter-event distances (between double-cut DSBs) and tallies instances of specific double-cuts
###############################################################################################

use strict;
use warnings;
use Cwd;
use List::Util qw(first);
use Sort::Naturally;
local $| = 1;
my $outext = '.txt';          #Output .file-extension
my $inext  = $ARGV[0];        #Input .file-extension
my @files  = glob("*$inext");
my $chk    = scalar(@files);
print "\nFailed to detect any .SAM files within the current directory.\n\n"
  if $chk == 0;
exit if $chk == 0;            #Stop script if no .SAM files are found
my $sub = cwd() . "/Coordinates";
mkdir("$sub") unless $chk == 0;
my $sub2 = cwd() . "/Analysis";
mkdir("$sub2") unless $chk == 0;
my $genome    = $ARGV[2];
my $trimmode  = $ARGV[3];
my $trimlength = $ARGV[4];
print "---------------------------------------";
print "\nCalculating Coordinates....\n";
print "---------------------------------------\n";
print "Currently processing:\n";

for my $file (@files) {   #For-each input file
   open my $IN, '<', $file or die "$!";   #Open and read input .SAM file(s)
   ( my $strain = $file ) =~ s/_[^_]+$//;  #Strain-name
   ( my $mode   = $ARGV[0] ) =~ s/\.SAM//; #Alignment-mode
   print "$strain\n";
   my $outfile =
      "Coordinates."
    . $genome . "_"
    . $strain . "_"
    . $mode
    . $outext;                    #Output files
   my $outfile2 =
     "Coordinates." . $genome . "_" . $strain . "_Ambiguous" . $outext;
   my ( $OUT, $OUT2, $OUT3, $OUT4 );
   open $OUT,  '>',  "$sub/$outfile"  or die "$!";
   open $OUT2, '>>', "$sub/$outfile2" or die "$!";
   print $OUT
"PairID\tStrand\tChr\tPos\tReadLength\tCIGAR\tAdjustment\tRead-Flag\tmSize\n";
```

```
  if ( $inext eq "Global.SAM" && $trimmode eq "Y" ) {
    print $OUT2
"PairID\tStrand\tChr\tPos\tReadLength\tCIGAR\tAdjustment\tRead-Flag\tmSize\tMD-Tag\t\n";
    my $outfile3 =
      $strain
     . $ARGV[5]
     . "_unmapped_trimmed.fastq";   #Unmapped R1 FASTQ file
    my $outfile4 =
      $strain
     . $ARGV[6]
     . "_unmapped_trimmed.fastq";   #Unmapped R2 FASTQ file
    open $OUT3, '>', "$outfile3" or die "$!";
    open $OUT4, '>', "$outfile4" or die "$!";
  }
  my ( $ID, $AID, $A, $B, $R1var, $R2var, $Wflag, $Cflag, %IED, %DoubleCut );
  while (<$IN>) {                  #For-each .SAM record
    chomp $_;
    next if /^\s*@/;               #Skip .SAM headerlines
    my @F = split( "\t", $_ );      #Split each tab-delimited field
    my $orientation = $F[3] - $F[7];  #Discard atypical read-orientations
    if (   $F[1] == 99 && $orientation > 0
      || $F[1] == 83 && $orientation < 0 )
    {
      my $skipline = <$IN>;
      next;
    }
    if ( $inext eq "Global.SAM" && $trimmode eq "Y" )
    {   #Populate unmapped R1/R2 FASTQ files mapped-unmapped pairs
      if ( grep { $_ == $F[1] } 73, 137 ) {
        print $OUT3 "\@$F[0] 1:N:0:1\n$F[9]\n+\n$F[10]\n"
          if $F[1] == 73;
        print $OUT4 "\@$F[0] 1:N:0:1\n$F[9]\n+\n$F[10]\n"
          if $F[1] == 137;
      }
      if ( grep { $_ == $F[1] } 89, 153 ) {
        $F[9] =~ tr/GATC/CTAG/;
        my $revseq  = reverse( $F[9] );
        my $revqual = reverse( $F[10] );
        print $OUT3 "\@$F[0] 1:N:0:1\n$revseq\n+\n$revqual\n"
          if $F[1] == 89;
        print $OUT4 "\@$F[0] 1:N:0:1\n$revseq\n+\n$revqual\n"
          if $F[1] == 153;
      }
      if ( grep { $_ == $F[1] } 69, 133 ) {
        my $trimseq  = substr( $F[9],  0, $trimlength );
        my $trimqual = substr( $F[10], 0, $trimlength );
        print $OUT3 "\@$F[0] 1:N:0:1\n$trimseq\n+\n$trimqual\n"
          if $F[1] == 69;
        print $OUT4 "\@$F[0] 1:N:0:1\n$trimseq\n+\n$trimqual\n"
          if $F[1] == 133;
      }
    }

    sub parseSAM {   #Subroutine to interpret SAM-field data
      my @rcd = @_;
      my @read;
      my $index = first { /MD:Z/ } @rcd;  #Obtain variable-column MD:Z tag
      my @MDtag = $index =~ /\d+/g;      #Remove non-numeric characters
```

```
  my %rules = ( M => 1, D => 1, I => 0, S => 1 )
   ;   #Rules to handle insertion/deletions/matches/soft-clipping
  my ( $s, $LS, $RS ) = (0) x 3;
  while ( $rcd[5] =~ /(\d+)([MDIS])/g )
  {     #Parse and interpret CIGAR code
    my ( $n, $op ) = ( $1, $2 );
    $s += $n * $rules{$op}
     unless $op eq
     'S';   #Calculate POS adjustment (insertions/deletions)
    $LS += $n * $rules{$op}
     if $op eq 'S' && $-[0] == 0;   #(upstream soft-clip)
    $RS += $n * $rules{$op}
     if $op eq 'S'
     && $+[0] == length( $rcd[5] );   #(downstream soft-clip)
  }
  my $l  = length( $rcd[9] );  #Read-length
  my $wp = $rcd[3] - $LS;     #Adjusted 5' coordinate (Watson strand)
  my $cp =
   $rcd[3] + ( $RS + $s ) - 1; #Adjusted 5' coordinate (Crick strand)
  push( @read, $rcd[2], $wp, $cp, $l, $rcd[5], $s, $LS, $RS, $index );
  return ( \@read, \@MDtag );
}
if ( grep { $_ == $F[1] } 99, 83 ) {   #For 99/147 or 83/163 read-pairs
  my $partner = <$IN>;
  my @F2 = split( "\t", $partner );   #Split each tab-delimited field
  if ( $F[1] == 99 ) {
    ( $A, $R1var ) = parseSAM(@F);
    ( $B, $R2var ) = parseSAM(@F2);
  }
  else {
    ( $A, $R1var ) = parseSAM(@F2);
    ( $B, $R2var ) = parseSAM(@F);
  }
  my @revMDtag = reverse( @{$R2var} );
  my @Wat    = @{$A};
  my @Cri    = @{$B};
  my ( $chr, $pos, $rl, $cigar, $Lclip, $vtag ) =
   @Wat[ 0, 1, 3, 4, 6, 8 ];
  my ( $chrp, $posp, $rlp, $cigarp, $cc, $Rclip, $vtagp ) =
   @Cri[ 0, 2, 3, 4, 5, 7, 8 ];
  my $mlength = abs( $pos - $posp );
  if ( $F[1] == 99 ) {
    $Wflag = 99;
    $Cflag = 147;
    if ( $ARGV[0] eq "Global.SAM" ) {
      $IED{$mlength}++;
      $DoubleCut{$chr}{$pos}{$posp}{"99"}++;
    }
  }
  elsif ( $F[1] == 83 ) {
    $Wflag = 163;
    $Cflag = 83;
    if ( $ARGV[0] eq "Global.SAM" ) {
      $IED{$mlength}++;
      $DoubleCut{$chr}{$pos}{$posp}{"163"}++;
    }
  }
  if (   $R1var->[0] == 0 && $R1var->[1] == 0 && $pos > 0
    || $revMDtag[0] == 0 && $revMDtag[1] == 0 && $pos > 0
```

```
            || $Lclip > 1       && $pos > 0
            || $Rclip > 1       && $pos > 0 )
          {   #Detect ambigious ends (99/147 or 83/163 pairs)
            $AID++;
            printf( $OUT2
"%d\t%s\t%s\t%d\t%d\t%s\t%d\t%d\t%d\t%s\n%d\t%s\t%s\t%d\t%d\t%s\t%d\t%d\t%d\t%s\n",
              $AID,          "w",   $chr,      $pos,
              $rl,           $cigar, 0 - $Lclip, $Wflag,
              $mlength,      $vtag,  $AID,      "c",
              $chrp,         $posp,  $rlp,      $cigarp,
              $Rclip + $cc - 1, $Cflag, $mlength,   $vtagp
            );
          }
          else {
            $ID++;
            printf( $OUT
"%d\t%s\t%s\t%d\t%d\t%s\t%d\t%d\t%d\n%d\t%s\t%s\t%d\t%d\t%s\t%d\t%d\t%d\n",
              $ID,           "w",   $chr,
              $pos,          $rl,   $cigar,
              0 - $Lclip,    $Wflag, $mlength,
              $ID,           "c",   $chrp,
              $posp,         $rlp,  $cigarp,
              $Rclip + $cc - 1, $Cflag, $mlength
            );
          }
        }
      }
      close $IN;
      close $OUT;
      close $OUT2;
      if ( $ARGV[0] eq "Global.SAM" ) {
        my $outfile5 = "IED." . $genome . "_" . $strain . $outext;
        my $outfile6 = "DoubleCuts." . $genome . "_" . $strain . $outext;
        open my $OUT5, '>', "$sub2/$outfile5" or die "$!";
        open my $OUT6, '>', "$sub2/$outfile6" or die "$!";
        print $OUT5 "IED\tFreq\n";
        print $OUT6 "Chr\tCoord-A\tCoord-B\tR1W Freq\tR1C Freq\n";
        foreach my $key ( sort { $a <=> $b } keys %IED ) {
          print $OUT5 "$key\t$IED{$key}\n";
        }
        foreach my $chrn ( nsort keys %DoubleCut ) {
          foreach
           my $coord1 ( sort { $a <=> $b } keys %{ $DoubleCut{$chrn} } )
          {
            foreach my $coord2 ( sort { $a <=> $b }
              keys %{ $DoubleCut{$chrn}{$coord1} } )
            {
              print $OUT6 join( "\t",
                $chrn, $coord1, $coord2,
                map $_ // 0,
                @{ $DoubleCut{$chrn}{$coord1}{$coord2} }{qw{99 163}} );
              print $OUT6 "\n";
            }
          }
        }
      }
      if ( $ARGV[1] eq "Y" ) {
        chomp($file);
        unlink($file); } }
```

Script—HistogramMap.pl (1bp histograms)

```
#!/usr/bin/env perl
#Package Version: 2.0

###############################################################################################
# Author(s): T.J.Cooper
# Updated: 6/12/2016
# Generates 1bp-histogram (sparse format) from combined coordinate files
###############################################################################################
use strict;
use warnings;
use Cwd;
use Getopt::Long;
use File::Basename qw(basename);
use List::Util qw(all);
use Sort::Naturally;
my $scriptname = basename($0);   #Obtain script-name
my ( $coord, $genome, $target );
my $usage = "Usage: $scriptname -i <Input Folder> -g <Genome Name>";
GetOptions(
  'i=s' => \$coord,
  'g=s' => \$genome
) or die("\n$usage\n");
die(
"\nError: Arguments or -flags are missing and/or incorrectly specified.\n\n$usage\n\n"
) unless all { defined } $coord, $genome;
my @files = glob( $coord . "/*_Combined.txt" );
my $chk  = scalar(@files);

if ( $chk == 0 ) {
  @files = glob( $coord . "/*_Global.txt" );
  $chk  = scalar(@files);
}
$genome =~ s/_//g;
print
"\nFailed to detect any valid coordinate files within the specified directory.\n\n"
 if $chk == 0;
exit if $chk == 0;
for my $file (@files) {
  open my $IN, '<', $file or die "$!";
  ( my $strain = basename($file) ) =~ s/_[^_]+$//;   #Strain-name
  $strain =~ s/^[^_]*_//;
  ( my $wd = $coord ) =~ s/[^\/]+$//;
  my ( %hits, %R1hits, %R2hits );
  my @headers = split( "\t", <$IN> );
  my %headidx;
  @headidx{@headers} = 0 .. $#headers;
  my $index = $headidx{Strand};

  while (<$IN>) {
    chomp $_;
    my @F = split( "\t", $_ );
    $hits{ $F[ $index + 1 ] }{ $F[ $index + 2 ] }{ $F[$index] }++;
    if ( $F[ $index + 6 ] == 99 || $F[ $index + 6 ] == 83 ) {
      $R1hits{ $F[ $index + 1 ] }{ $F[ $index + 2 ] }{ $F[$index] }++;
    }
    if ( $F[ $index + 6 ] == 147 || $F[ $index + 6 ] == 163 ) {
```

```
        $R2hits{ $F[ $index + 1 ] }{ $F[ $index + 2 ] }{ $F[$index] }++;
      }
    }
    my $outfile =
     $wd . "/Histograms/" . "FullMap." . $genome . "_" . $strain . ".txt";
    open my $OUT, '>', "$outfile" or die "$!";
    print $OUT "Chr\tPos\tWatson\tCrick\n";
    foreach my $chr ( nsort keys %hits ) {
      foreach my $pos ( sort { $a <=> $b } keys %{ $hits{$chr} } ) {
        print $OUT join( "\t",
          $chr, $pos,
          map $_ // 0,
          @{ $hits{$chr}{$pos} }{qw{w c}} );
        print $OUT "\n";
      }
    }
    if (%R2hits) {
      my $outfile2 =
        $wd
       . "/Histograms/"
       . "R1_FullMap."
       . $genome . "_"
       . $strain . ".txt";
      my $outfile3 =
        $wd
       . "/Histograms/"
       . "R2_FullMap."
       . $genome . "_"
       . $strain . ".txt";
      open my $OUT2, '>', "$outfile2" or die "$!";
      open my $OUT3, '>', "$outfile3" or die "$!";
      for my $OF ( $OUT2, $OUT3 ) { print $OF "Chr\tPos\tWatson\tCrick\n"; }
      foreach my $chr ( nsort keys %R1hits ) {
        foreach my $pos ( sort { $a <=> $b } keys %{ $R1hits{$chr} } ) {
          print $OUT2 join( "\t",
            $chr, $pos,
            map $_ // 0,
            @{ $R1hits{$chr}{$pos} }{qw{w c}} );
          print $OUT2 "\n";
        }
      }
      foreach my $chr ( nsort keys %R2hits ) {
        foreach my $pos ( sort { $a <=> $b } keys %{ $R2hits{$chr} } ) {
          print $OUT3 join( "\t",
            $chr, $pos,
            map $_ // 0,
            @{ $R2hits{$chr}{$pos} }{qw{w c}} );
          print $OUT3 "\n";
        }
      }
    }
}
```

Script—OligoTrim.pl (Spo11-oligo FASTQ trimming)

```
#!/usr/bin/env perl
#Package Version: 2.0

###############################################################################################
# Author(s): T.J.Cooper
# Updated: 15/2/2017
# Trims FASTQ-entries from the 5'-end using user-specified patterns
###############################################################################################
use strict;
use warnings;
use Getopt::Long;
use File::Basename qw(basename);
use List::Util qw(all);
my $scriptname = basename($0);   #Obtain script-name
my ( $R1, $R2, @UTrim, @LTrim );
my $usage =
"Usage: $scriptname -1 <Read-1 FASTQ> -2 <Read-2 FASTQ> -U <5' Upstream Pattern> -L <3' Downstream
Pattern>";
GetOptions(
  '1=s' => \$R1,
  '2=s' => \$R2,
  'U=s' => \@UTrim,
  'L=s' => \@LTrim
) or die("\n$usage\n");
die(
"\nError: Arguments or -flags are missing and/or incorrectly specified.\n\n$usage\n\n"
) unless all { defined } $R1, $R2, @UTrim, @LTrim;
my $outfile = "Trim_" . basename($R1);
open my $OUT, '>', "$outfile" or die "$!";
open my $IN,  '<', $R1       or die "$!";
my ( $trimseq, $R1trimU, $R1trimL, $match, $R2trim, $platform );

while (<$IN>) {
  chomp $_;
  if ( $. % 4 == 2 ) {
    $trimseq = $_;
    $platform = substr( $trimseq, 0, 9 );
    my $Uregex = join "|", @UTrim;
    if ( $platform =~ m/^.*?$Uregex/ ) {
      $R1trimU = $+[0];
      $trimseq = substr( $_, $R1trimU );
    }
    else {
      $R1trimU = 0;
    }
    my $Lregex = join "|", @LTrim;
    my $revseq = reverse($trimseq);
    if ( $revseq =~ m/^.*?$Lregex/ ) {
      $R1trimL = $+[0];
      $revseq = substr( $revseq, $R1trimL );
    }
    else {
      $revseq = substr( $revseq, 2 );
      $R1trimL = 2;
    }
    my $finalseq = reverse($revseq);
```

```
      print $OUT "$finalseq\n";
    }
    elsif ( $. % 4 == 0 ) {
      my $revqual = reverse($_);
      my $trimqual = substr( $revqual, $R1trimL );
      $trimqual = reverse($trimqual);
      $trimqual = substr( $trimqual, $R1trimU );
      print $OUT "$trimqual\n";
    }
    else {
      print $OUT "$_\n";
    }
}
my $outfile2 = "Trim_" . basename($R2);
open my $OUT2, '>', "$outfile2" or die "$!";
open my $IN2,  '<', $R2        or die "$!";
while (<$IN2>) {
    chomp $_;
    if ( $. % 4 == 2 ) {
      $trimseq = $_;
      $platform = substr( $trimseq, 0, 4 );
      my $Uregex = join "|", @UTrim;
      if ( $platform =~ m/^.*?$Uregex/ ) {
        $R1trimU = $+[0];
        $trimseq = substr( $_, $R1trimU );
      }
      else {
        $R1trimU = 0;
      }
      my $Lregex = join "|", @LTrim;
      my $revseq = reverse($trimseq);
      if ( $revseq =~ m/^.*?$Lregex/ ) {
        $R1trimL = $+[0];
        $revseq = substr( $revseq, $R1trimL );
      }
      else {
        $revseq = substr( $revseq, 2 );
        $R1trimL = 2;
      }
      my $finalseq = reverse($revseq);
      print $OUT2 "$finalseq\n";
    }
    elsif ( $. % 4 == 0 ) {
      my $revqual = reverse($_);
      my $trimqual = substr( $revqual, $R1trimL );
      $trimqual = reverse($trimqual);
      $trimqual = substr( $trimqual, $R1trimU );
      print $OUT2 "$trimqual\n";
    }
    else {
      print $OUT2 "$_\n";
    }
}
```

Script—SeqBias.pl (Sequence bias)

```
#!/usr/bin/env perl
#Package Version: 2.0

################################################################################################
# Author(s): T.J.Cooper
# Updated: 15/2/2017
# Extracts genomic sequence surrounding user-specific peaks/features for sequence composition analysis
################################################################################################
use strict;
use warnings;
use Bio::SeqIO;
use Getopt::Long;
use File::Basename qw(basename);
use List::Util qw(all);
my $scriptname = basename($0);   #Obtain script-name
my ( $peaks, $fasta, $width, $mode, $outfile );
my $usage =
"Usage: $scriptname -i <HistogramFile> -r <ReferenceFASTA> -w <Width> -m <Mode: Pos/Freq> -o
<Output>"
 ;                      #Error/usage message
GetOptions(
  'i=s' => \$peaks,
  'r=s' => \$fasta,
  'w=i' => \$width,
  'm=s' => \$mode,
  'o=s' => \$outfile
) or die("\n$usage\n");
die(
"\nError: Arguments or -flags are missing and/or incorrectly specified.\n\n$usage\n\n"
) unless all { defined } $peaks, $fasta, $width, $mode, $outfile;
open my $IN,  '<', $peaks   or die "$!";
open my $OUT, '>', $outfile or die "$!";  #Create mapping-coordinate output file
my ( @flankseq, $seqEX, %sequences );
my $seqio = Bio::SeqIO->new( -file => $fasta );

while ( my $seqobj = $seqio->next_seq ) {
  my $id  = $seqobj->display_id;
  my $seq = $seqobj->seq;
  $sequences{$id} = $seq;
}

sub seqstore {
  my @rcd = @_;
  if ( $mode eq "Pos" ) {
    push( @flankseq, $rcd[0] );
  }
  elsif ( $mode eq "Freq" ) {
    push( @flankseq, $rcd[0] ) for ( 1 .. $rcd[1] );
  }
  return;
}
<$IN> for ( 1 .. 1 );
while (<$IN>) {
  chomp $_;
  my (@F) = split( "\t", $_ );
  next
```

```
    if ( $F[0] == 12 && $F[1] > 451000 && $F[1] < 470000 )
    ;   #Skip rDNA coordinates
   next if $F[1] - $width < 1;
   if ( $F[3] > 0 ) {
      $seqEX = substr( $sequences{ $F[0] }, $F[1] - 1 - $width,
         $width )   #Extracts Xbp upstream (not incl. 5' end)
       . substr( $sequences{ $F[0] }, $F[1] - 1, $width + 1 );
      $seqEX =~ tr/GATC/CTAG/;
      $seqEX = reverse($seqEX);
      seqstore( $seqEX, $F[3] );
   }
   if ( $F[2] > 0 ) {
      $seqEX = substr( $sequences{ $F[0] }, $F[1] - 1 - $width,
         $width )   #Extracts Xbp upstream (not incl. 5' end)
       . substr( $sequences{ $F[0] }, $F[1] - 1, $width + 1 );
      seqstore( $seqEX, $F[2] );
   }
}
my %freq;
my $inc = 1 / @flankseq;
for my $FS (@flankseq) {
   $freq{ substr $FS, $_, 1 }[$_] += $inc for 0 .. length($FS) - 1;
}
printf $OUT " " . "%5s  " x ( keys %freq ) . "\n", sort keys %freq;
my $label = -$width - 1;
for my $pos ( 1 .. length $flankseq[0] ) {
   $label++;
   printf $OUT "%1d" . "%7.4f" x ( keys %freq ) . "\n",
    $label, map { $freq{$_}[ $pos - 1 ] // 0 } sort keys %freq;
}
my $run_time = time() - $^T;
print "\n---------------------------------------";
print "\nAnalysis Complete\n";
print "Execution Time: $run_time Seconds\n";
print "---------------------------------------\n\n";
```

	Strain	Entry	Background	Genotype
Spo11 DSB Mapping	Parent	MJ6	SK1	*ho*::*LYS2*/", *lys2*/", *ura3*/", *arg4–nsp*/", *leu2*::*hisG*/", *his4X*::*LEU2*/", *nuc1*::*LEU2*/"
	*sae2*Δ	MJ315	SK1	*sae2*Δ::*KanMX6*/'
	*sae2*Δ*tel1*Δ	VG402	SK1	*sae2*Δ::*KanMX4*/", *tel1*Δ::*HphMX4*/"
	*sae2*Δ*tel1KD*	VG431	SK1	*sae2*Δ::*KanMX6*/", *tel1–D2612A*, *N2617A*, *D2631A*/"
	*sae2*Δ*ndt80*Δ	MJ962	SK1	*sae2*Δ::*KanMX4*/', *ndt80*Δ::*LEU2*/'
	*sae2*Δ*ndt80*Δ*tel1*Δ	MJ965	SK1	*sae2*Δ::*KanMX4*/', *tel1*Δ::*HphMX4*/', *ndt80*Δ::*LEU2*/'
Topo II Mapping	Parent	MJ429	SK1	*ura3*Δ*0*, *leu2*Δ*0*, *his3*Δ, *met15*Δ*0*, *pdr1*Δ::*PDR1–DBD–CYC8*::*LEU2*
	WT	MJ429	SK1	*As above*
	*sae2*Δ	MJ475	SK1	*sae2*Δ::*KanMX6*/'
	*mre11*Δ	MJ551	SK1	*mre11*Δ::*KanMX4*

Table 3.3. Strain Table—Genome-wide mapping of Spo11 DSBs

All strains contain the genotype of their respective parent.

CHAPTER 4

Investigating Tel1^{ATM}-dependent DSB interference

4.1—Introduction

Mechanisms influencing the position of CO formation (e.g. CO interference) (see: Chapter 2) are layered upon the preceding decision: the distribution of DSBs. Importantly, and akin to crossovers (COs), the precursor array of DSBs is also subject to several processes of spatial regulation (see: Section 1.4) (Cooper et al. 2016). Tel1^{ATM}, a DNA damage response (DDR) kinase, mediates the phenomenon of DSB interference—one such process (Garcia et al. 2015). DSB interference manifests as a non-random distribution of DSBs—locally suppressing DSB formation in a distance-dependent manner. However, inactivation of Tel1^{ATM} unexpectedly results in two distinct outcomes: (i) over mid-long range distances (±20-100kb), DSBs form independently of one another at frequencies similar to those expected by chance (ii) by contrast, at short range (~±7.5kb) DSBs exhibit concerted activity, arising coincidentally at frequencies significantly greater than expected from independent behaviour (Garcia et al. 2015)—an observation termed negative interference. Meiotic chromosomes display a unique architecture, organising into linear arrays of protruding chromatin loops—10-15kb in size (12.1kb average) within *S. cerevisiae*—each basally attached to a proteinaceous axis (see: Section 1.2.10) (Blat et al. 2002; Kleckner 2006; Borde & de Massy 2013; Ito et al. 2014). Rec8, a meiosis-specific component of the cohesin complex, along with other axial components including Rec114-Mer2-Mei4 (RMM-complex), demarcate loop boundaries (Novak et al. 2008; Glynn et al. 2004; Panizza et al. 2011). Interestingly, chromatin architecture is intimately linked to meiotic recombination. The tethered-loop axis model (see: Section 1.2.10) proposes that tethering of chromatin loops to the axis, by the PHD finger domain protein, Spp1, bridges together DSB hotspots and the necessary axis bound machinery (e.g. RMM-complex) to facilitate DSB formation (Sommermeyer et al. 2013; Acquaviva et al. 2013; Borde et al. 2009; Tischfield & Keeney 2012). Remarkably, negative interference is only observed *between* DSB hotspots residing within the same chromosomal loop domain (Garcia et al. 2015). In contrast, Spo11 double cutting (see: Chapter 3), exacerbated upon loss of Tel1 activity, predominately occurs *within* hotspots—suggesting negative interference is a distinct phenomenon. Confinement of concerted activity to within singular loop domains may conceivably arise if a process upstream of DSB formation—such as tethering—"activates" the contained hotspots and where activation of any given loop only occurs in

a subset of the population (Garcia et al. 2015; Cooper et al. 2016). Within any given cell, DSB formation may thus be channeled toward narrow windows (activated loops) scattered across each chromosome, resulting in clustered DSB formation. However, this hypothesis has not yet been tested. Moreover, no model to investigate the features of Tel1^{ATM}-dependent DSB interference has been established. Work presented throughout this chapter thus seeks to utilise computational and mathematical methods to probe the mechanisms of DSB and negative interference.

4.2—Inactivation of Tel1 results in a genome-wide redistribution of DSBs

Tel1-dependent DSB interference appears to moderately influence the population average landscape of DSB formation—as assessed by genome-wide mapping of Spo11-oligos (Mohibullah & Keeney 2017). In order to initially determine whether or not such a redistribution of DSBs also occurs within *sae2*Δ Spo11 DSB data (see: Section 3.2), the collective strength of each annotated hotspot was compared for averaged *sae2*Δ, *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* (kinase dead) samples (Figure 4.1). Loss of Tel1 activity results in a spreading of DSB formation outside the confines of established hotspot boundaries (see: Section 3.12). Therefore, in order to account for this effect when quantitatively comparing *TEL1*$^{+}$/*tel1*Δ samples, hotspot boundaries were widened by 300bp to calculate expanded, normalised values per million reads (NormHpM300). Pearson's rho (ρ), a measure of linear correlation, is employed throughout this chapter (see: Section 3.7). While a strong overall correlation ($\rho = 0.9064$) is observed between *sae2*Δ and *sae2*Δ*tel1*Δ data, significant differences appear to exist across the full spectrum with hotspots differing up to ~5-10-fold in strength (Figure 4.1A). This trend is similarly recaptured when comparing *sae2*Δ and *sae2*Δ*tel1KD* ($\rho = 0.9028$) (Figure 4.1B). Within *sae2*Δ backgrounds, the length of prophase I may be altered and inactivation of Tel1 may further perturb the phase owing to a lack of checkpoint signalling in response to DSB formation. A shortened or lengthened window of DSB formation may alter the population average distribution of DSBs in unexpected ways, accounting for the observed *TEL1*$^{+}$/*tel1*Δ variation.

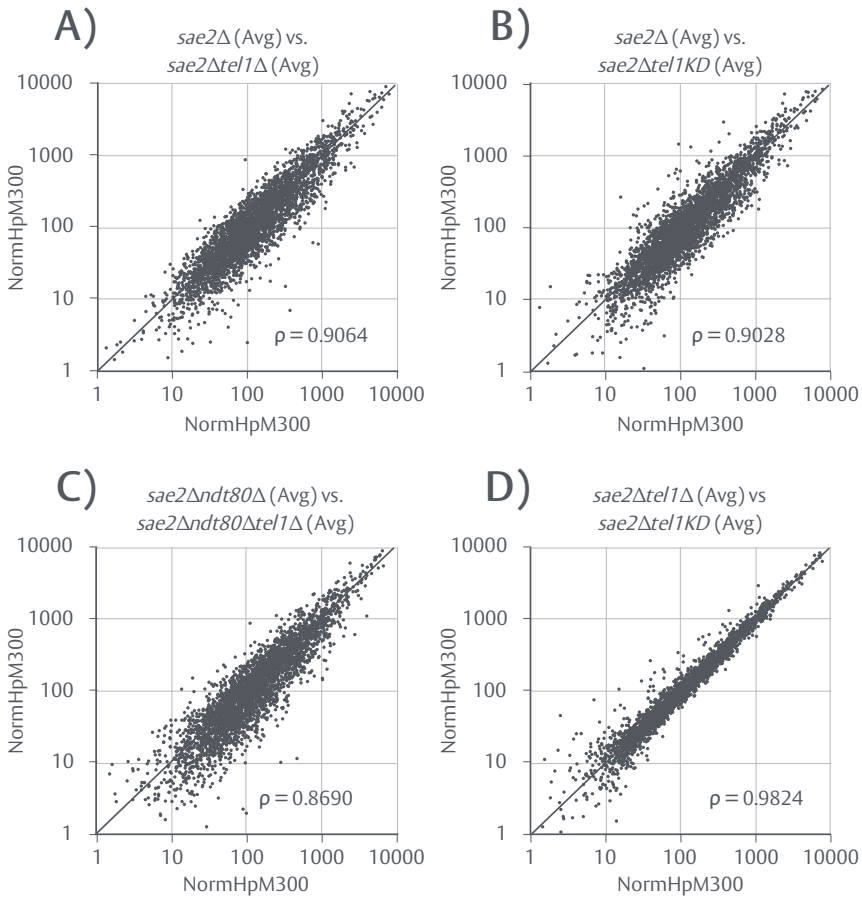


Figure 4.1. Inactivation of Tel1 results in a genome-wide redistribution of DSBs

Spo11 5' hits were tallied across 3599 previously annotated *S. cerevisiae* hotspots (Pan et al. 2011) and averaged for all repeats. An additional 300bp was included either side of each defined hotspot region (expanded hotspots) and data was normalised to account for calculated background levels and read count on a per library basis (NormHpM300—hits per million). All non-hotspot hits were discarded. The quantitative strength of each hotspot was compared for **A)** *sae2*Δ, *sae2*Δ*tel1*Δ **B)** *sae2*Δ, *sae2*Δ*tel1KD* **C)** *sae2*Δ*ndt80*Δ, *sae2*Δ*ndt80*Δ*tel1*Δ and **D)** *sae2*Δ*tel1*Δ, *sae2*Δ*telKD*. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values).

In order to assess this possibility and account for any impact prophase I length may have, *sae2*Δ*ndt80*Δ data—within which prophase I exit is abolished (Xu et al. 1995; Winter 2012; Allers & Lichten 2001)—was compared to *sae2*Δ*ndt80*Δ*tel1*Δ data in an identical manner (Figure 4.1C). While the strength of quantitative correlation is slightly reduced ($\rho = 0.8690$) relative to *NDT80*$^{+}$ comparisons, the overall trend remains suggesting that altered prophase I length is not responsible for the observed *tel1*Δ-dependent deviations from *TEL1*$^{+}$ DSB distributions. Interestingly, *sae2*Δ*tel1*Δ and *sae2*Δ*tel1KD* datasets are highly correlated ($\rho = 0.9824$) and exhibit much less variation (Figure 4.1D). Difficulty in quantifying *sae2*Δ*tel1KD* hotspot strength due to the extensive spreading of Spo11 DSB signal (see: Section 3.12) may account for any remaining variation. Collectively, these results suggest that the loss of Tel1 activity within a *sae2*Δ background results in a genome-wide redistribution of DSBs at the population level—consistent with previous observations (Mohibullah & Keeney 2017)—and that this redistribution is independent of end processing and HR, which are absent within *sae2*Δ.

4.3—Tel1-dependent redistribution of DSBs occurs within domains of concerted change

According to previous studies, the redistribution of DSBs within *tel1*Δ mutants occurs in a non-random manner across the length of each chromosome, confined to specific chromosomal domains (Mohibullah & Keeney 2017). In order to further explore the features of *tel1*Δ-dependent change, averaged 1bp histogram data—generated by *Spo11Mapper*—for *sae2*Δ, *sae2*Δ*tel1*Δ, *sae2*Δ*ndt80*Δ and *sae2*Δ*ndt80*Δ*tel1*Δ samples were visualised for two chromosomes (ChrXI and ChrXIV) (Figure 4.2A-D respectively). Subtle, quantitative differences are observed between *tel1*Δ and *TEL1*$^{+}$ datasets in both directions—that is, DSB frequency increases at a subset of hotspots in *tel1*Δ relative to *TEL1*$^{+}$, while decreasing or remaining unchanged at others. To characterise these changes, *sae2*Δ/*sae2*Δ*tel1*Δ and *sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ ratios were calculated between the collective strength of all annotated hotspots and smoothed (moving average) (Figure 4.2E,F respectively). Consistent with previous observations (Mohibullah & Keeney 2017), *tel1*Δ-dependent changes—in

the relative usage of hotspots—appear to cluster within gross chromosomal domains of concerted change. In other words, hotspots within a given region tend to exhibit a decrease or increase in usage, but not both. Interestingly, while *sae2*Δ|*sae2*Δ*tel1*Δ ratios moderately resemble the generalised features of *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ, they appear distinct. Upon close inspection of the data, it was noticed that *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ ratios for ChrXI and ChrXIV exhibit an anti correlation to smoothed hotspot strength (Figure 4.2G). Specifically, regions of higher hotspot strength are utilised less often in *sae2*Δ*ndt80*Δ, relative to *sae2*Δ*ndt80*Δ*tel1*Δ (see: Figure 4.2G, Segment Y-Z). Similarly, regions of lower hotspot strength are utilised more often in *sae2*Δ*ndt80*Δ, relative to *sae2*Δ*ndt80*Δ*tel1*Δ (see: Figure 4.2G, Segment X). This pattern appears to be exacerbated at subtelomeric hotspots, which exhibit considerably higher frequencies within *sae2*Δ*ndt80*Δ than in *sae2*Δ*ndt80*Δ*tel1*Δ, despite possessing some of the lowest hotspot strengths.

In order to assess whether or not these patterns are observed at a genome-wide level, normalised hotspot strength was piled up ±25kb flanking the midpoints of all *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ and *sae2*Δ|*sae2*Δ*tel1*Δ non-subtelomeric ratio domains that display either positive (>1.1) or negative (<0.9) changes. Strong negative correlations (ρ = -0.88-0.92) are observed for both domain types within *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ data (Figure 4.3A,B), suggesting this relationship is readily observed within across the genome. However, no clear correlation is observed for *sae2*Δ|*sae2*Δ*tel1*Δ ratios (Figure 4.3C,D), suggesting the Tel1-dependent process giving rise to this pattern requires a WT-like or lengthened prophase I to fully establish itself.

Crucially, such *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ patterns are consistent with the imposition of DSB interference through a bidirectionally spreading signal. Intra-arm induction of DSBs will "push" DSB formation toward the periphery i.e. subtelomeric regions. Moreover, subtelomeric regions only experience interference from a singular direction and thus are, on average, suppressed for DSB formation less often. Similarly, regions of lower hotspot density induce interference less frequently, and may experience interference from neighbouring regions with lower intensity.

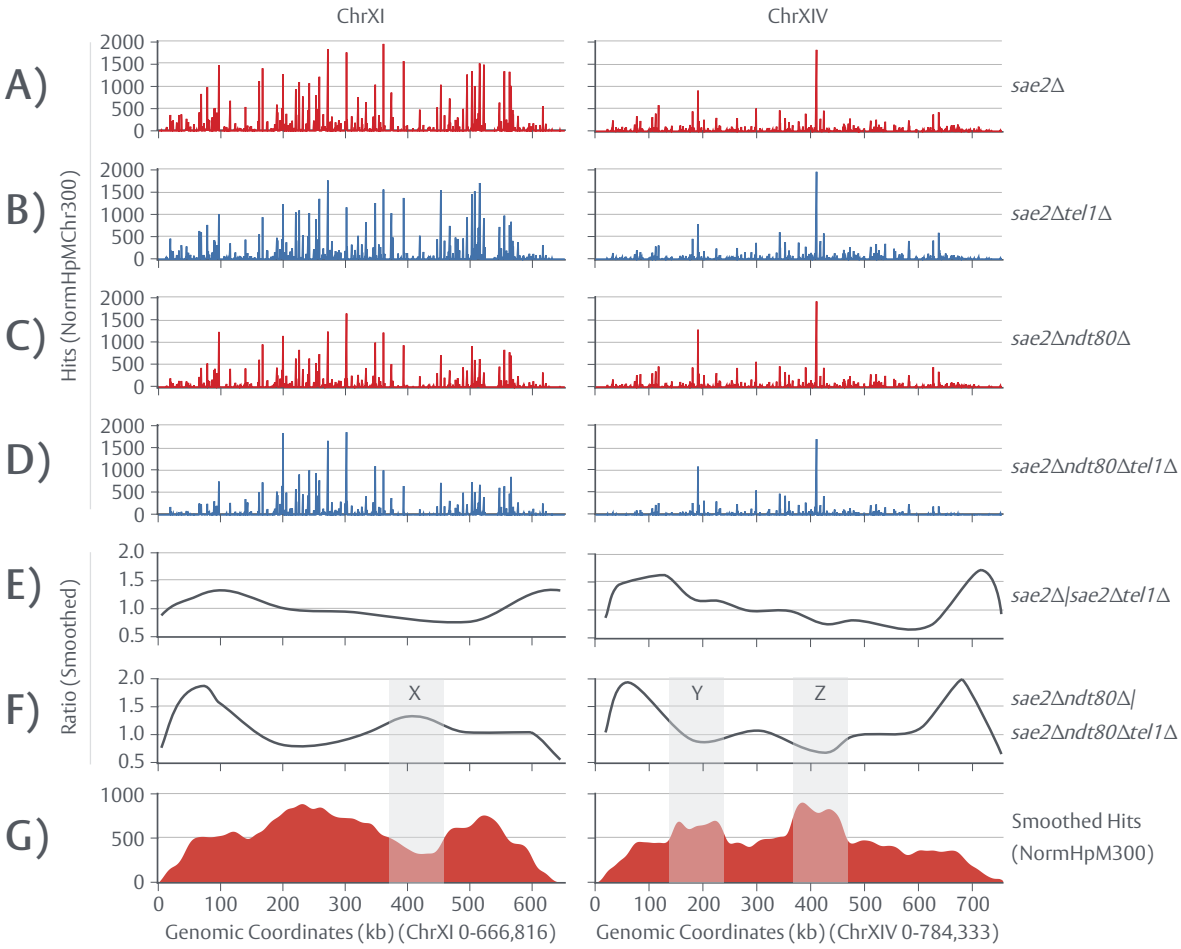


Figure 4.2. Tel1-dependent redistribution of DSBs occurs within domains of concerted change

Spo11 5' hits were tallied across 3599 previously annotated *S. cerevisiae*, expanded hotspots (+300bp) (Pan et al. 2011), normalised and visualised on ChrXI and ChrXIV for **A)** *sae2*Δ **B)** *sae2*Δ*tel1*Δ **C)** *sae2*Δ*ndt80*Δ and **D)** *sae2*Δ*ndt80*Δ*tel1*Δ data. Ratios, calculated between the strength of corresponding pairs of hotspots, were calculated and smoothed (moving average, n = 3) for **E)** *sae2*Δ/*sae2*Δ*tel1*Δ and **F)** *sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ. **G)** Hotspot density, derived from *sae2*Δ*ndt80*Δ data, was smoothed (moving average, n = 5kb). Highlighted regions (X/Y/Z) correspond to example domains of concerted change.

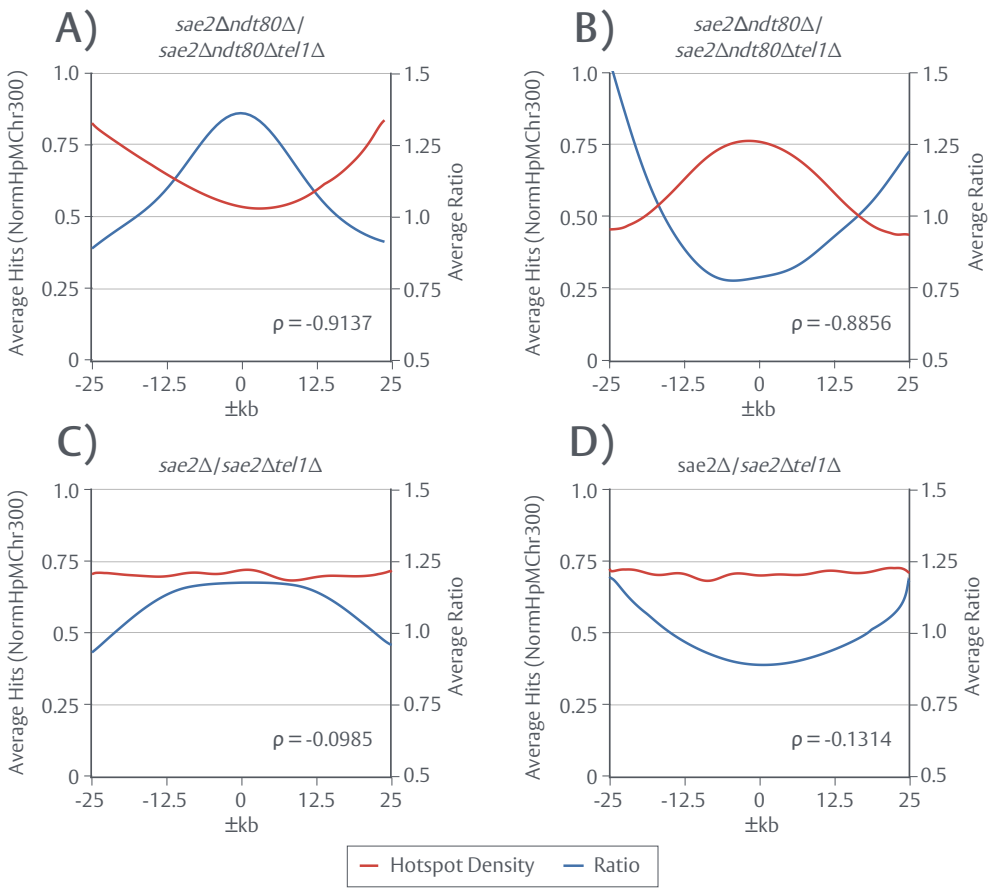


Figure 4.3. Domains of concerted change anti correlate with hotspot density

Midpoints of *sae2Δ|tel1Δ* and *sae2Δndt80Δ|sae2Δndt80Δtel1Δ* ratio domains were calculated for each chromosome using a peak-calling algorithm (MATLAB 2017a Package: *findpeaks*). Any domain containing maximal values of >0.9, <1.1 or which reside within 75kb of either chromosomal end were discarded. Hotspot density, derived from *sae2Δndt80Δ* data, was normalised between [0.0]-[1.0] on a per chromosome basis (NormHpMChr300), smoothed (moving average, n = 5kb) and piled up flanking each identified domain midpoint for **A)** *sae2Δndt80Δ|sae2Δndt80Δtel1Δ* positive (>1.1) domains **B)** *sae2Δndt80Δ|sae2Δndt80Δtel1Δ* negative (<0.9) domains **C)** *sae2Δ|sae2Δtel1Δ* positive domains **D)** *sae2Δ|sae2Δtel1Δ* negative domains. Averaged ratio values, piled up flanking each domain midpoint, are also shown.

Contrastingly, regions of higher hotspot density are the focal point of DSB interference and are therefore suppressed more often and utilised less often in *sae2*Δ*ndt80*Δ. Collectively, these results suggest that the *tel1*Δ-dependent redistribution of DSBs across each chromosome may be a population average hallmark of DSB interference that in turn generates a distinct but quantitatively weak change in the hotspot landscape. Given the lack of clear and explainable pattern to *sae2*Δ/*sae2*Δ*tel1*Δ ratios and difficulty quantifying *sae2*Δ*tel1KD* hotspots, all following sections primarily concern *ndt80*Δ/*ndt80*Δ*tel1*Δ data.

4.4—DSBSim: A novel simulation platform

In lieu of an established model for DSB formation and to dissect the mechanics of negative interference and Tel1-dependent DSB interference, a novel simulation platform (*DSBSim*) was established (see: Section B4.1). (γ)-distribution and hazard function (h(x)) based modelling, as applied to COs and NCOs (see: Chapter 2), is primarily useful for global, genome wide descriptions of a process when handling limited, per cell datasets which must be pooled together to increase statistical power. Meiotic recombination is, however, ideally modelled on a per chromosome basis—taking into account and/or revealing any chromosome-specific differences. In contrast to the mapping of CO/NCOs within individual cells (see: Section 2.2), genome-wide mapping of Spo11 DSBs yields a population averaged picture of DSB formation and provides an opportunity to investigate spatial regulation on each chromosome. *DSBSim* was therefore tailored toward simulations at the population level, using *Spo11Mapper* data as a basis (see: Table 3.1).

A typical simulation run, as depicted in (Figure 4.4A), is split into several key processes: (i) Virtual chromosomes are constructed as binned, numerical arrays at a 100bp resolution based on *S. cerevisiae* (S288c) chromosomal length (see: Section B4.1.1). Any given 100bp bin contains a value that describes the relative probability of that bin being selected for DSB formation (DSB(P)). DSB(P) values are pre-populated using *sae2*Δ*ndt80*Δ*tel1*Δ Spo11 DSB data, normalised on a per chromosome basis (NormHpMChr300), at positions corresponding to annotated hotspots. By

employing *ndt80*Δ*tel1*Δ data, *DSBSim* seeks to re-introduce DSB interference into the system in an attempt to mimic *sae2*Δ*ndt80*Δ data. In order to assess the relationship(s) between chromatin architecture and DSB formation, Rec8 ChIP-seq data, derived from (Ito et al. 2014), is utilised to establish chromatin loop domains (see: Section B4.1.2). (ii) To a user specified frequency (e.g. 30%), a subset of loops—chosen at random—are "activated", greatly boosting the contained DSB(P) values. Boosting is performed either uniformly across the loop (Figure 4.4B) or non-uniformly, using a Gaussian function—where the intensity of boosting increases toward the centre of the window (Figure 4.5C) (see: Section 4.5). (iii) DSB formation is subsequently simulated, for a specific chromosome and a given number of DSBs (N). DSB formation is a two step process. Initially, the position of a potential DSB is determined directly by DSB(P) values, heavily skewing site selection toward strong, activated hotspots. The success of DSB formation at this chosen site, however, depends on a second array containing interference values (Int(P)) (see: Section B4.1.3). (iv) DSB interference is imposed centred on each formed DSB by altering flanking Int(P) values. If a strong hotspot is site selected, but is otherwise heavily interfered with—as described by Int(P) values—DSB formation may fail. A system whereby DSB formation can fail allows for DSB frequencies to be suppressed below that intended—an expected consequence of Tel1^{ATM}-dependent interference (Garcia et al. 2015). During NCO/CO simulations (via *RecombineSim*) the shape and range of CO interference is determined by best fit $\gamma(\alpha,\beta)$ parameters (see: Section 2.6). No such derivation is available for population average data. Instead, *DSBSim* applies bidirectional DSB interference as either an exponential decay—defined by a slope factor (μ)—or a Hann window—defined by a set with (±kb) (see: Section 4.6). (v) *DSBSim* supports *cis* interference (single chromatid), *trans* interference (dual chromatid) and random simulation modes. Under *cis* mode, interference is solely imposed across the initiating chromosome. In contrast, under *trans* mode, interference is imposed on both the initiating and non-initiating chromosome to the same extent and intensity (see: Figure 4.4A). During random simulations, no interference is applied.

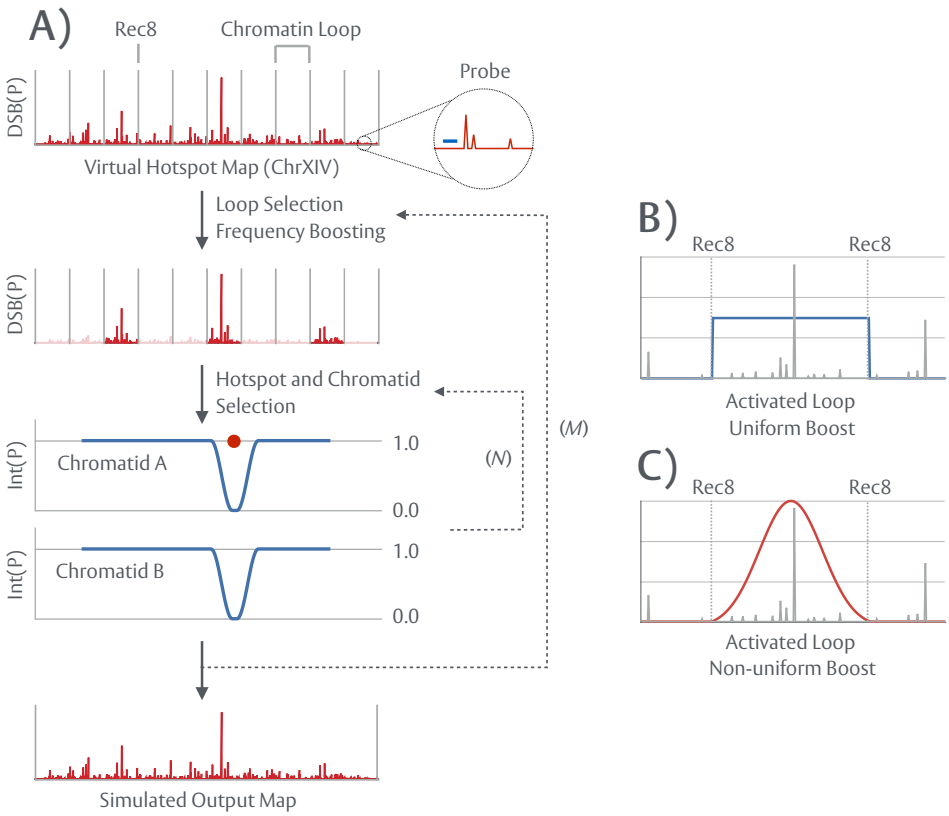


Figure 4.4. DSBSim—An overview

A) Virtual chromosomes are constructed at a 100bp resolution as binned, numerical arrays proportional in size to *in vivo* chromosome length x 0.01 (S. cerevisiae—S288c). Any given 100bp bin contains a value, based on normalised hotspot data, in the range of [0.0-1.0]x10^6 that describes the relative probability of that bin being selected for DSB formation (DSB(P)). Peak maximas of Rec8 ChIP-seq, derived from (Ito et al. 2014), are utilised to establish probable loop boundary locations along each chromosome. Prior to event formation, a given proportion of loop domains (e.g. 30%) are activated by application of a boost window through multiplication of DSB(P) values, largely disqualifying DSB formation from occurring within non-activated regions. Proceeding the formation of an initial DSB, DSB interference is applied as a bidirectional signal by multiplying values (Int(P)) held in secondary arrays. Int(P) arrays are initially populated with values of [1.0]—denoting a 100% chance of successful DSB formation. Interference is imposed in *cis*, along the initiating chromosome (chromatid A)—which is chosen at random—or also in *trans* on the non-initiating chromosome (chromatid B). All subsequent DSBs (N) undergo site selection based on DSB(P) values and may or may not successfully form based on Int(P) values. Finalised datasets are generated by repeating this procedure for additional chromosomes (*M*) (e.g. 10,000) and combining results to generate a simulated, population average hotspot map. **B)** A uniform, flat boost window (blue). **C)** A non-uniform, Gaussian boost window (red). Boost windows are shown across a representative hotspot region.

(vi) Processes outlined above are repeated for a specific number of (M) of independently simulated chromosomes. For each simulated chromosome, a novel set of loops are chosen at random for activation. (vii) As a primary output, *DSBSim* provides simulated hotspot maps, calculated based on the frequency at which any given 100bp bin successfully formed a DSB within the simulated population. Moreover, *DSBSim* also permits placement of virtual probes at a given loci—mimicking pulse field gel electrophoresis or southern blotting and calculating probed molecule frequencies in order to assess phenomena such as negative interference.

4.5—Non-uniform loop activation necessitates inversion of hotspot maps

In order to explore the mechanics and consequences of loop activation, non-interfering simulations (*DSBSim* mode: Random) were conducted for an idealised, 500kb chromosome and ChrXI (Figure 4.5A) using either a flat, uniform boost or a non-uniform, Gaussian boost—both at an activation frequency of 30%. An idealised chromosome is devoid of hotspots—rather, each 100bp contains an equal probability (DSB(P)) of being selected for DSB formation. Resulting, population averaged output maps for each simulation were compiled, via *DSBSim*, normalised and compared to input maps by calculating output/input ratios. Ratio values of ~1.0 signify no population average change as a result of loop activation. Application of a flat boost across each activated loop has no appreciable impact on the population average distribution of DSBs for either input model—as evidenced by ratio values of ~1.0 across the chromosome length (Figure 4.5B,D). Such an observation is expected as the boost is uniformly applied to each hotspot regardless of position (see: Figure 4.4B) and the selection of loops for activation is randomised, thus any effects of flat boosting on a per cell basis are averaged out across a large population of simulated cells. In contrast, non-uniform boosting—which modifies the relative strength of any given hotspot in an unequal manner (see: Figure 4.4C)—alters the population distribution of DSBs, as evidence by widely varying ratio values of [0.0-1.0], and imparts the shape of the boost applied (Figure 4.5C,E). Importantly, the way in which any *in vivo* process influences the distribution of DSBs, for a given genotype, will be inherently present within the respective genome-wide map of Spo11 DSBs.

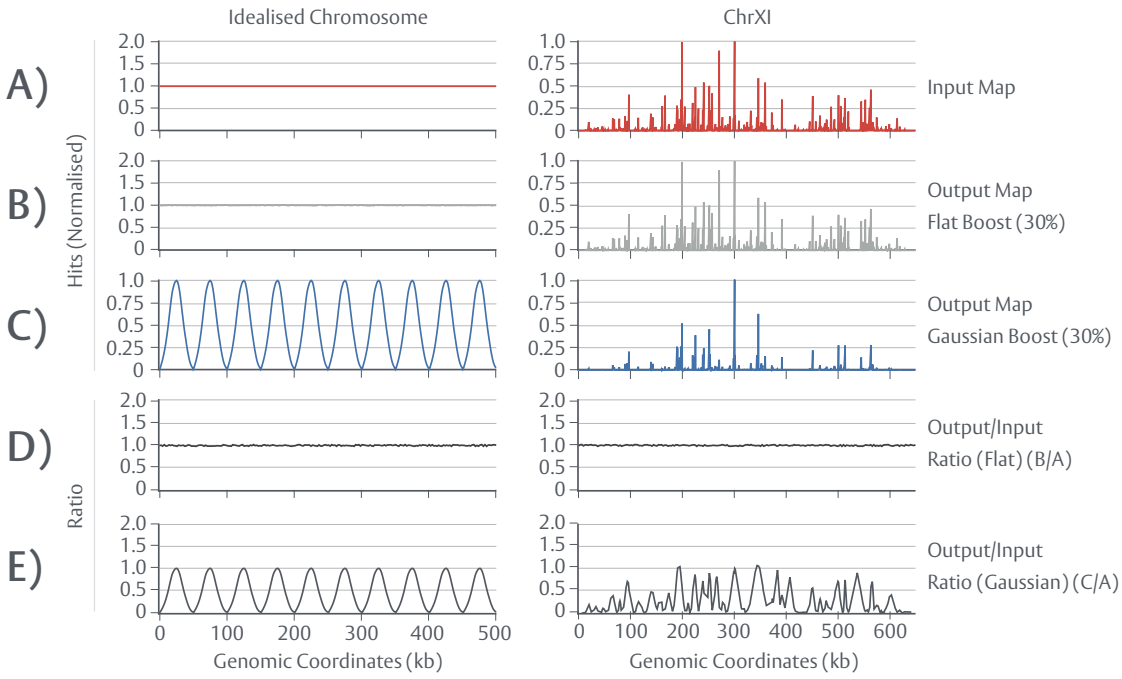


Figure 4.5. Non-uniform loop activation alters the population average distribution of DSBs

A) *Input maps*. A 500kb idealised model was constructed at a 100bp resolution as an array populated with values of [1.0]. A ChrXI model—based on normalised *sae2*Δ*ndt80*Δ*tel1*Δ Spo11 5' hits tallied across annotated *S. cerevisiae*, expanded hotspots (+300bp)—was constructed at a 100bp resolution. **B-C)** Non-interfering simulations (*DSBSim* mode: Random) using either a flat, uniform boost or a Gaussian, non-uniform boost and a loop activation frequency of 30% were conducted for both input maps respectively. **D-E)** Ratios, on a hotspot by hotspot basis, were calculated between input maps and the resulting simulated, output maps (B/A and C/A). All input and simulated hotspot maps are normalised to a range of [0.0-1.0] for clarity.

While simulated DSB interference may alter alter the population average, non-interfering simulations (*DSBSim* mode: Random), as conducted above, impose no interference and simply seek to recapture the input model. Therefore, under this simulation mode, any output map must strictly reproduce the input—a requirement not met when loops are boosted non-uniformly (see: Figure 4.5E).

In order to accommodate non-uniform boosting, *DSBSim* was modified to include a map inversion mechanic. Map inversion effectively approximates a "pre-boost" hotspot map by applying inverted, non-uniform boosts across each loop region for a given loop activation frequency (see: Section B4.1.3). To validate map inversion, non-interfering simulations (*DSBSim* mode: Random) were repeated for an idealised, 500kb chromosome and ChrXI using inverted input maps (Figure 4.6B), non-uniform Gaussian boosting and a loop activation frequency of 30%. As a point of comparison, non-inverted maps are reshown (Figure 4.6A). Resulting, population averaged output maps were normalised and compared to the *non–inverted* input model by recalculating output/input ratios. Crucially, resulting ratios adopt values of ~1.0 across the length of each chromosome, signifying that inversion successfully cancels out the population average effects of non-uniform boosting—allowing *DSBSim* to reproduce the experimentally observed distribution of DSBs, while still activating loops in this manner.

4.6—Simulated activation of chromatin loops generates regions of negative interference

As previously detailed (see: Section 4.1), concerted formation of DSBs above expected levels—the phenomenon of negative interference—is confined to singular loop domains as primarily assessed at the *ARE1* locus (ChrIII) (Figure 4.7A) (Garcia et al. 2015). Interference is calculated using the standard formula (1-OBS/EXP) i.e. the ratio between the observed number of inter-hotspot double cuts and the expected amount based on the individual frequency of each DSB in the population. A value of [0.0] thus denotes expected behaviour while negative values signify that observed frequencies exceed those expected under conditions of independence.

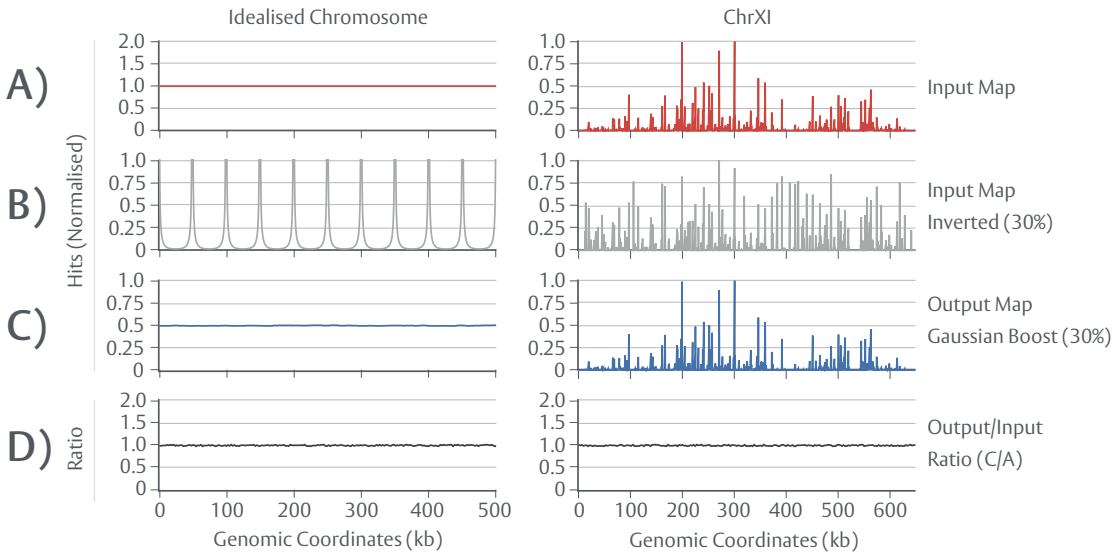


Figure 4.6. Non-uniform loop activation necessitates inversion of hotspot maps

A) *Input maps*. A 500kb idealised model was constructed at a 100bp resolution as an array populated with values of [1.0]. A ChrXI model—based on normalised *sae2*Δ*ndt80*Δ*tel1*Δ Spo11 5' hits tallied across annotated *S. cerevisiae*, expanded hotspots (+300bp)—was constructed at a 100bp resolution. **B)** Inverted idealised and ChrXI input maps were created by application of inverted (1-x) Gaussian boost windows across each loop region for a loop activation frequency of 30%. **C)** Non-interfering simulations (*DSBSim* mode: Random) using a Gaussian, non-uniform boost and a loop activation frequency of 30% were conducted using inverted input maps. **D)** Ratios, on a hotspot by hotspot basis, were calculated between non-inverted input maps and the resulting simulated, output maps (C/A). All input and simulated hotspot maps are normalised to a range of [0.0-1.0] for clarity.

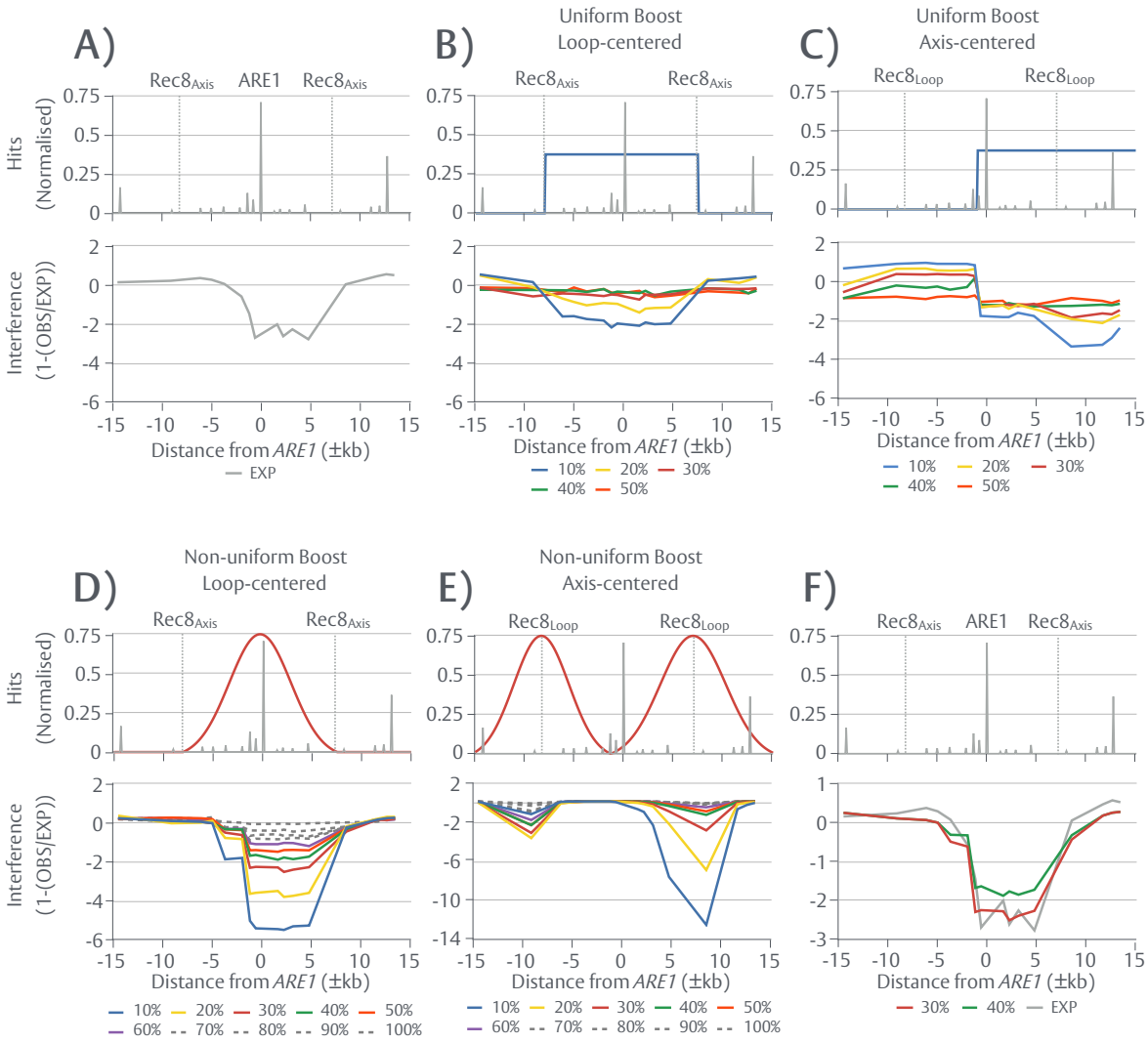


Figure 4.7. Simulated activation of chromatin loops generates regions of negative interference

A) Experimentally observed negative interference for *sae2*Δ*tel1*Δ, across the *ARE1* locus (shown above each plot), was calculated using the standard interference formula (1-OBS/EXP)—an inverted ratio between the observed frequency of inter-hotspot double cuts and the expected frequency. Observed data was obtained from (Garcia et al. 2015). Expected data was recalculated, based on *sae2*Δ*tel1*Δ hotspot data normalised to an *ARE1* frequency of 13.2% by multiplying the quantitative strength of each hotspot pair. **B-C)** Simulated negative interference using a uniform, flat boost centred on the loop or on axial elements respectively. **D-E)** Simulated negative interference using a non-uniform, Gaussian boost centred on the loop or on axial elements respectively. **F)** Best fit negative interference models. In (B-E), the shape and position of the boost applied is shown above each plot. All simulations were conducted on non-inverted (flat boost) or inverted (gaussian boost) ChrIII *sae2*Δ*tel1*Δ input maps, using variable loop activation frequencies as marked (*DSBSim* mode: Random). Simulated OBS values were obtained using virtual probes placed either side of the main *ARE1* peak.

Experimental data (OBS) is obtained via southern blot using probes placed either side of the central, major *ARE1* hotspot peak (see: Figure 4.7A)—approximately measuring the frequency of inter-hotspot double cuts between *ARE1* and any flanking peak (Garcia et al. 2015). Expected (EXP) double cut frequencies were recalculated, under the assumption that all hotspots behave independently, using *sae2*Δ*tel1*Δ Spo11 DSB hotspot data normalised to an *ARE1* DSB frequency of 13.2% as determined by southern blot for this genotype (Garcia et al. 2015). As observed (see Figure 4.7A), negative interference increases toward the centre of the loop region. The origin of such a skew is unknown. The proposed "loop activation" model (see: Section 4.1) of negative interference may entail a non-uniformly imposed boost that disproportionately favours centrally located hotspots for breakage. Moreover, it remains unclear whether or not any such boost process would be applied centred on the loop, or on the axial bound DSB machinery.

In order to investigate how such a pattern of negative interference is generated, non-interfering simulations (*DSBSim* mode: Random) were conducted for ChrIII using a flat, uniform boost at varying loop activation frequencies (10-50%). Virtual probes were placed either side of *ARE1*, calculating the frequencies of each inter-hotspot double cut. Boosts were either centred on the loop (Figure 4.7B) or on axial, Rec8 elements (Figure 4.7C). Loop-centred, flat boosting results in a relatively uniform region of negative interference across the *ARE1* loop region (see: Figure 4.7B). Interestingly, only low frequency activation (10-20%) produces an appreciable deviation from expectation (ratios ≠ 0) (see: Figure 4.7B). By contrast, axis-centred, flat boosting shifts the region of negative interference in accordance to the new position of the boost (see: Figure 4.7C), suggesting activation must be applied across the loop region itself in order to match the experimentally observed pattern.

While these results demonstrate the ability of upstream loop activation to generate localised regions of negative interference, they do not recapture the experimental data. Non-interfering simulations (*DSBSim* mode: *Random*) were thus repeated for ChrIII using inverted maps (see:

Section 4.5) and non-uniform Gaussian boosts at varying loop activation frequencies (10-100%). Boosts were either centred on the loop (Figure 4.7D), or on axial, Rec8 elements (Figure 4.7E). Loop-centred, Gaussian boosting results in non-uniform regions of negative interference that increase toward the centre of the *ARE1* loop region (see: Figure 4.7D). Notably, a wider range of activation frequencies (10-60%) produce appreciable deviations from expectation relative to flat boosting (ratios ≠ 0). Moreover, the intensity of negative interference is observed to weaken as loop activation frequency increases. Such an observation is expected; as the number of regions competent for DSB formation across the chromosome increases, the less DSBs are forcibly channeled into narrow regions—lowering the chance of concerted formation. As expected, an activation frequency of 100% produces interference values of ~0, further validating map inversion and the site selection mechanics of *DSBSim* (see: Figure 4.7D). Interestingly, the resulting patterns of negative interference are off-centre and do not fully align with the boost shape applied. Such an observation may reflect the inherent layout of the hotspots within the loop region. As before, axis-centred Gaussian boosting shifts the region of negative interference in accordance to the boost position (see: Figure 4.7E). Importantly, loop activation frequencies of 30-40% produce negative interference patterns that highly resemble the experimental data both in intensity, and shape (Figure 4.7F).

Collectively these results suggest that, at the *ARE1* locus, a loop activation process—applied non-uniformly with a frequency of 30-40%—may be responsible for the observed *tel1*Δ-dependent pattern of negative interference, providing evidence for this proposed model. These results, however, do not infer that all loops across the genome exhibit average activation frequencies of 30-40%, nor that the *in vivo* process precisely matches the mechanics employed here (see: Section 4.12).

4.7—Simulated interference is able to generate domains of concerted change

As previously outlined, *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ ratios reveal domains of concerted change whereby clustered areas of hotspots are under- or over-utilised in a manner dependent upon Tel1 status. Moreover, such patterns appear to be consistent with the imposition of DSB interference (see: Section 4.3). In order to initially test whether or not a process of interference can generate such domains, interfering (INT) simulations (*DSBSim* modes: *Cis*, *Trans*) were conducted on an idealised, 500kb chromosome with or without added, simplistic hotspots (Figure 4.8). DSB interference was applied as a short range (±100kb) dense region of inhibition via a Hann window (Figure 4.8A), or a broad (±500kb) exponentially decaying window (Figure 4.8B)—reminiscent of a diffusive, kinase signal—and output (INT+)/input (INT-) ratios were subsequently calculated. Notably, both forms of interference are able to generate substantial subtelomeric enrichment on fully idealised models (Figure 4.8C-D). Interestingly, *trans* interference exacerbates the intensity of ratio changes. Such an observation is perhaps expected; the first chromatid to incur a DSB will pattern both chromatids and therefore, all subsequent breaks on the secondary chromatid will be predisposed toward formation at the periphery.

Introduction of simplistic hotspots, however, results in additional complexity. Consistent with the observed anti correlation between the direction of ratio change and hotspot strength for *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ data (see: Figure 4.3), application of short range (±100kb) Hann window interference generates an internal domain of concerted increase within a region of lower, average hotspot strength (Figure 4.8E). This domain is largely ablated when wider, exponential interference is applied, presumably because regions of lower hotspot density are still sufficiently interfered with, although a minor skew is still observed (Figure 4.8F). As previously observed, ratio changes are intensified under the application of *trans* interference.

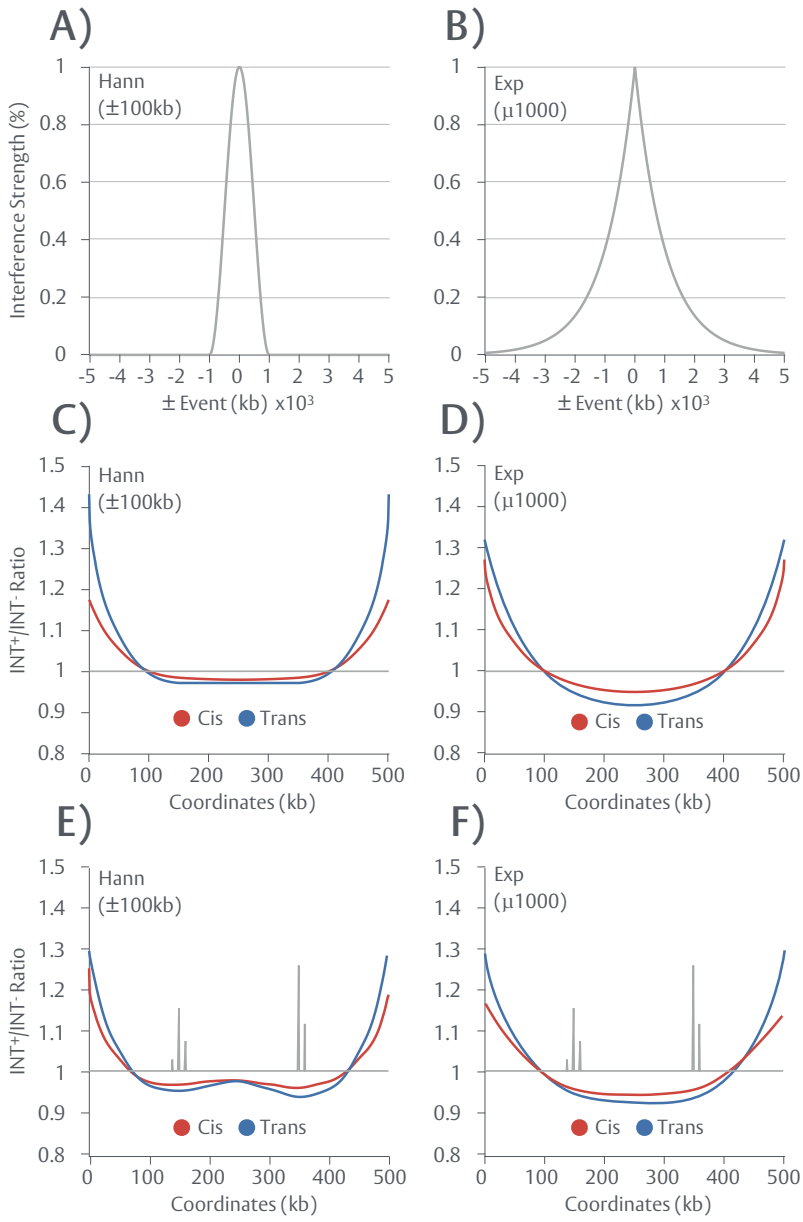


Figure 4.8. Simulated interference is able to generate domains of concerted change

A) A narrow, short range ±100kb Hann interference window. **B)** A broad, long range ~±500kb Exponential interference window (μ1000). **C-D)** Interfering simulations (*DSBSim* modes: *Cis*, *Trans*) were conducted on 500kb idealised chromosomes, populated with values of [1.0], using Hann and Exponential interference respectively in *cis* or in *trans* (as marked). Simulated ratios were calculated on a hotspot by hotspot basis between simulated output maps containing the effects of interference (INT+) and input maps (INT-). **E-F)** Interfering simulations (*DSBSim* modes: *Cis*, *Trans*) were repeated on 500kb idealised chromosomes containing two simplistic hotspots using Hann and Exponential interference respectively in *cis* or in *trans* (as marked). Input maps used for each simulation are overlaid on each plot (grey). All ratios are smoothed (moving average, n = 3).

Collectively, these results provide further evidence that DSB interference manifests within the population average and does so with a predictable pattern. Moreover, the distinctive patterns produced by *cis*, *trans*, long and short range forms of interference suggest that information regarding the form in which DSB interference adopts *in vivo* may be encoded within *sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ ratios.

4.8—Genome-wide maps of Spo11 DSBs contain evidence of DSB interference in trans

In order to determine whether or not the experimentally observed ratios (*sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ) may be explained by a process of DSB interference, a set of "model training" chromosomes (ChrXI, ChrXIV and ChrXVI) were analysed via in *cis* interfering simulations (*DSBSim* mode: *Cis*) using progressively increasing strengths of interference. These training chromosomes were specifically chosen as each exhibits several regions of distinct and strong concerted change. As before, DSB interference is imposed as either a broad, exponential window (Figure 4.9A)—defined by a slope factor (μ)—or a Hann window—defined by a set width (W) (in kb) (Figure 4.9B). By incrementally altering μ or W—thus the strength of interference—the effect each iteration has on the resulting ratio can be monitored (Figure 4.9C,D) and statistically screened to obtain optimal model fits. Pearson's rho (ρ), a measure of linear correlation, is utilised to assess shape fit while the average difference (AvgD)—calculated as the average, absolute difference in value between each simulated and experimental ratio point—is utilised to assess intensity fit. A statistically similar match is defined here as combined values of $\rho > 0.75$ and $AvD < 0.15$.

As suggested by idealised testing (see: Section 4.7), narrow width (±50kb) Hann interference is able to recapture the generalised shape features of *sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ ratios for all three chromosomes, but not the intensity as evidenced by large AvD values (0.18-0.3) (Figure 4.10A,C,E). Notably, subtelomeric enrichment is either weak or fully absent. By contrast, wider (±250kb) Hann interference enhances subtelomeric enrichment but does so at the expense of internal domains of concerted change (Figure 4.10B,D,F).

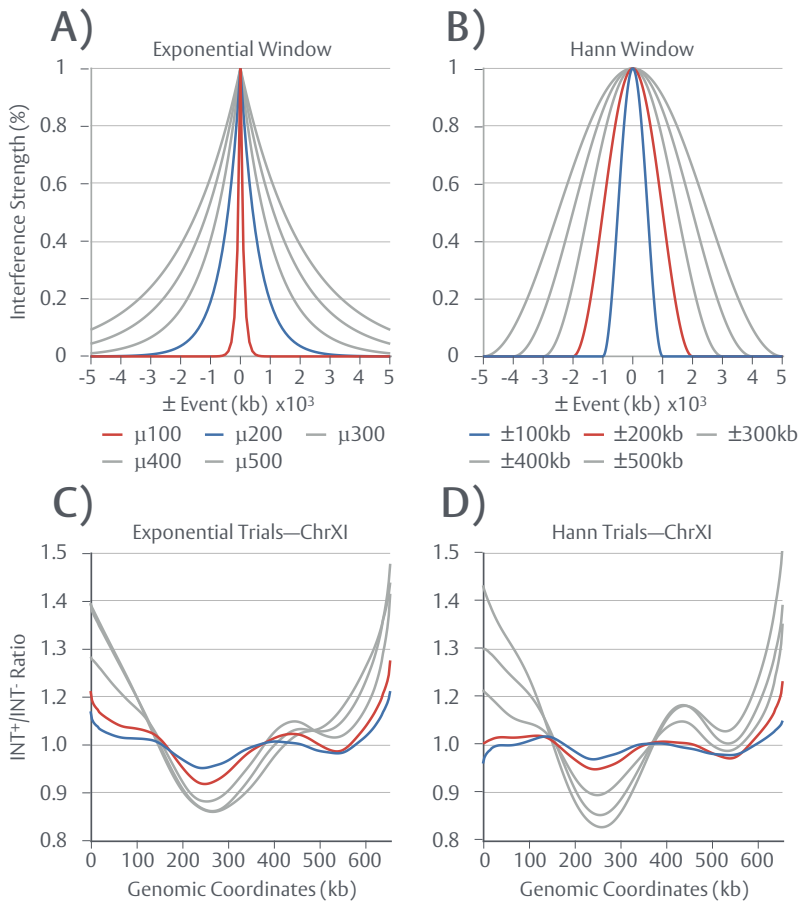


Figure 4.9. Iterative screening of DSB interference

A) Broad, long range Exponential interference windows at varying (μ) values (100-500). **B)** Narrow, short range Hann interference windows at varying widths (±100-500kb). **C-D)** *Example trial output*. Interfering simulations (*DSBSim* mode: *Cis*) were conducted on *sae2Δndt80Δtel1Δ* ChrXI input models using varying widths of Exponential or Hann interference respectively as shown in (A-B). Simulated ratios were calculated on a hotspot by hotspot basis between simulated output maps containing the effects of interference (INT+) and input maps (INT-). All ratios are smoothed (moving average, n = 3).

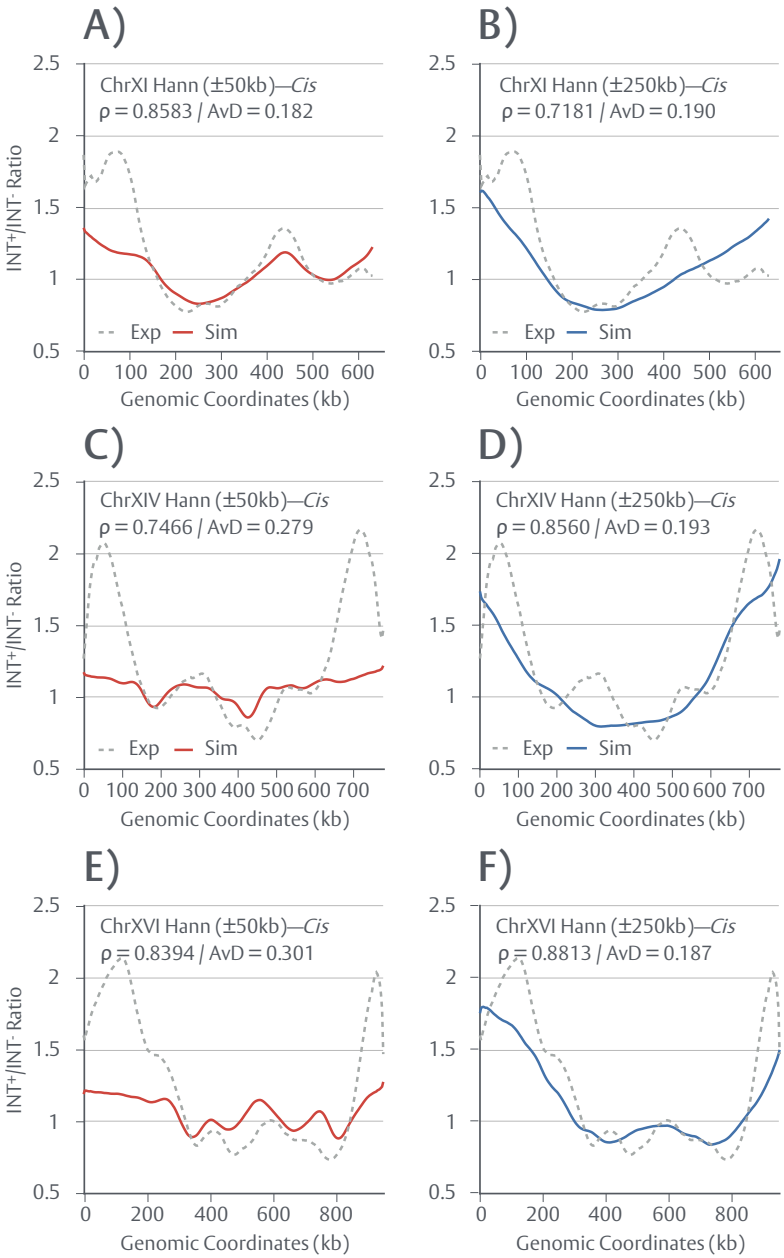


Figure 4.10. In cis interference insufficiently recaptures experimental data (Hann Window)

Interfering simulations (*DSBSim* mode: *Cis*) were conducted on *sae2*Δ*ndt80*Δ*tel1*Δ input models using varying widths of Hann window interference. Resulting simulated INT^+/INT^- ratios were smoothed (moving average, n = 3) and statistically screened. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values) to determine shape match. Average deviations were calculated as the average point by point difference to assess intensity fit (see: marked AvD values). No models display statistical similarity to the experimental data, defined as combined values of ($\rho > 0.75$, $AvD < 0.15$). Example results are shown for **A)** ChrXI Hann ±50kb **B)** ChrXI Hann ±250kb **C)** ChrXIV Hann ±50kb **D)** ChrXIV Hann ±250kb **E)** ChrXVI Hann ±50kb and **F)** ChrXI Hann ±250kb. Experimental ratios (*sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ) are overlaid on each plot (grey).

No statistically similar match was found for any simulated Hann iteration. Ratio patterns produced by both narrow and broad range Hann interference are similarly replicated when applying DSB interference as an exponential window (Figure 4.11). However, detail of internal ratio domains is less well captured relative to Hann windows. Collectively, these results suggest that a process of *in cis* interference, as simulated and assessed by these methods, is unable to reconcile together the requirement for an interference window strong enough to produce intense shifts in DSB distribution at the subtelomeres, but which can also generate internal domains of concerted change.

It is possible that *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ ratios are a composite of two distinct processes: (i) Tel1-dependent DSB interference and (ii) an independent, separate mechanism of Tel1-dependent regulation specific to subtelomeric regions. Nevertheless, for the latter to account for increased subtelomeric DSB formation within *sae2*Δ*ndt80*Δ relative to *sae2*Δ*ndt80*Δ*tel1*Δ, Tel1 would have to occupy a dual role as both a repressor and a promoter of DSB formation and no such subtelomeric-specific mechanism has been noted in the literature. Moreover, it is perhaps counterintuitive to propose a system that directly promotes cleavage within regions normally repressed for DSB and CO formation (see: Section 1.4.6) (Blitzblau et al. 2007; Buhler et al. 2007; Pan et al. 2011). Tel1 has, however, been shown to function in a process of *trans* DSB interference (Zhang et al. 2011). As illustrated by idealised testing (see: Section 4.7), *trans* interference intensifies ratio changes at the periphery without loss of concerted domains—consistent with the apparent requirements of this model. Interfering simulations (*DSBSim* mode: *Trans*) were thus repeated in *trans* for ChrXI, ChrXIV and ChrXVI, using progressively increasing widths of interference as previously described. Remarkably, experimental *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ ratios are recaptured with a high degree of statistical similarity ($\rho > 0.75$, AvD < 0.15) for all three chromosomes when applying interference as a narrow ±70-90kb Hann window (Figure 4.12A,C,E).

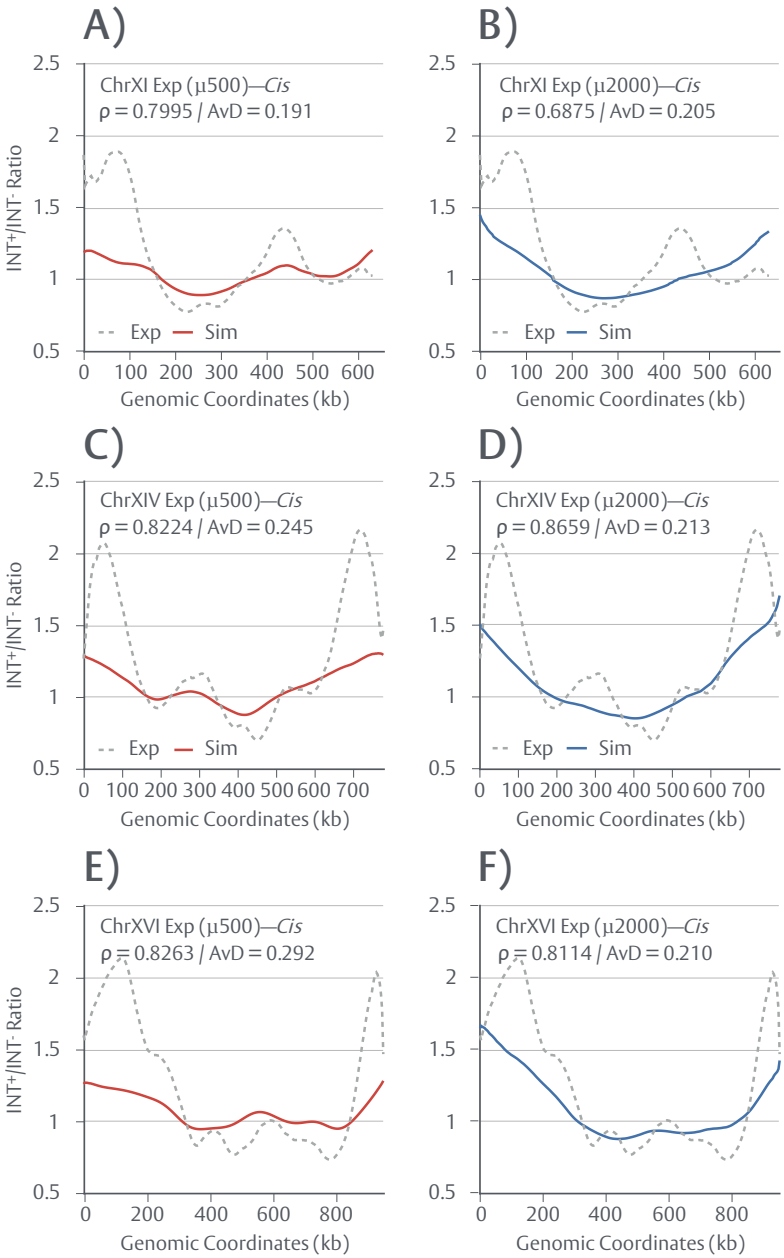


Figure 4.11. In cis interference insufficiently recaptures experimental data (Exp Window)

Interfering simulations (DSBSim mode: *Cis*) were conducted on *sae2*Δ*ndt80*Δ*tel1*Δ input models using varying widths of Exponential window interference. Resulting simulated INT+/INT- ratios were smoothed (moving average, n = 3) and statistically screened. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values) to determine shape match. Average deviations were calculated as the average point by point difference to assess intensity fit (see: marked AvD values). No models display statistical similarity to the experimental data, defined as combined values of ($\rho > 0.75$, AvD < 0.15). Example results are shown for **A)** ChrXI Exp µ500 **B)** ChrXI Exp µ2000 **C)** ChrXIV Exp µ500 **D)** ChrXIV Exp µ2500 **E)** ChrXVI Exp µ500 and **F)** ChrXI Exp µ2500. Experimental ratios (*sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ) are overlaid on each plot (grey).

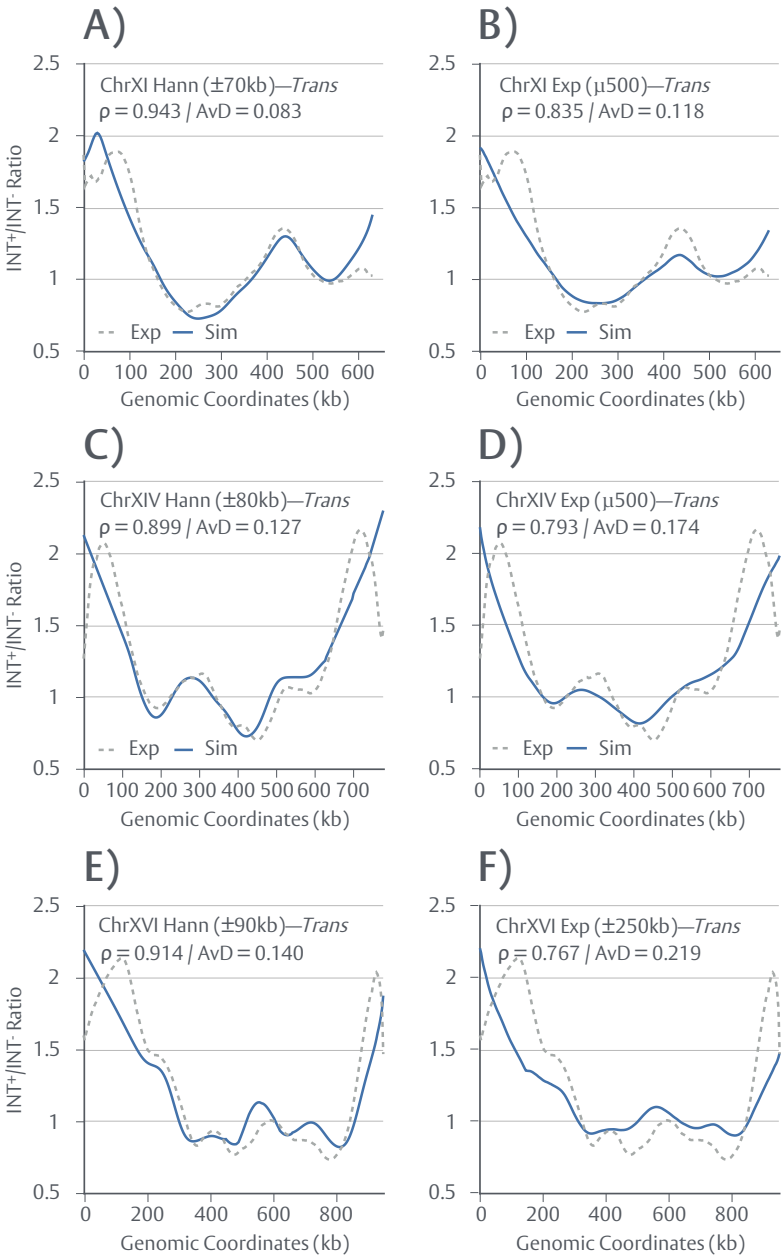


Figure 4.12. Genome-wide maps of Spo11 DSBs contain evidence of trans DSB interference

Interfering simulations (*DSBSim* mode: *Trans*) were conducted on *sae2*Δ*ndt80*Δ*tel1*Δ input models using varying widths of Hann or Exponential window interference. Resulting simulated INT+/INT- ratios were smoothed (moving average, n = 3) and statistically screened against the corresponding experimental ratio. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values) to determine shape match. Average deviations were calculated as the average point by point difference to assess intensity fit (see: marked AvD values). Best fit models, as determined by ρ and AvD values and a statistical similarity threshold of ($\rho > 0.75$, AvD < 0.15), are shown for **A-B)** ChrXI **C-D)** ChrXIV and **E-F)** ChrXVI. Experimental ratios (*sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ) are overlaid on each plot (grey).

As previously noted, experimental ratios are less accurately captured using an exponential window, suggesting DSB interference may not involve a diffusive signal but rather a more localised, dense mode of inhibition (Figure 4.12B,D,F).

4.9—DSB interference is predicted to act over a fixed spatial distance

Collectively, results presented thus far suggest that *sae2*Δ*ndt80*Δ|*sae2*Δ*ndt80*Δ*tel1*Δ ratios may be readily explained by a process of *cis* and *trans* interference on ChrXI, ChrXIV and ChrXVI, implicating Tel1 in both branches of DSB interference—as previously proposed (Zhang et al. 2011; Cooper et al. 2014; Cooper et al. 2016). In order to expand these findings to all remaining chromosomes, bar ChrII (which contains the *tel1*Δ allele) and ChrXII (which contains rDNA), interfering simulations (*DSBSim* mode: *Trans*) were conducted and statistically screened as previously described to identify best fit solutions (Figure 4.13A). In line with the results obtained for model training chromosomes, statistically similar fits ($\rho > 0.75$, AvD < 0.15) are obtained for all chromosomes using both Hann and exponential windows. As before, Hann windows (average stats—$\rho = 0.90$, AvD = 0.12) universally outperform exponential windows (average stats—$\rho = 0.82$, AvD = 0.142) for all chromosomes. Best and worst fit solutions for each chromosome, as defined by ρ and AvD values, have been shown for Hann window simulations against the respective experimental ratio (Figure 4.13B-O).

Remarkably, the width of interference required to yield a best fit solution widely varies. For example, ChrXV is best described by a Hann window ±100kb in width, as opposed to ±70kb for ChrVIII or ±140kb for ChrIII. DSB interference may thus operate differently on a per chromosome basis, governed by inherent traits such as chromosomal size. In order to further assess this possibility, best fit parameters for Hann windows were plotted against S. cerevisiae chromosome size (S288c) (Figure 4.14A). A weak correlation ($\rho = 0.2274$) is observed, however, chromosomes toward the lower end heavily skew the overall trend.

A)

Chr	Chr Size	Hann (±kb)	ρ	AvgD	Exp (μ)	ρ	AvgD
I	230,208	110	0.926	0.138	800	0.848	0.149
VI	270,148	120	0.894	0.128	800	0.829	0.148
III	316,617	140	0.911	0.140	1000	0.857	0.111
IX	439,885	100	0.912	0.106	500	0.811	0.143
VIII	562,643	70	0.880	0.135	500	0.821	0.136
V	576,869	70	0.915	0.094	500	0.835	0.151
XI	666,454	70	0.943	0.083	500	0.835	0.118
X	745,742	80	0.872	0.133	600	0.849	0.116
XIV	784,333	80	0.899	0.127	500	0.793	0.174
XIII	924,429	100	0.887	0.119	700	0.820	0.130
XVI	948,062	90	0.914	0.140	700	0.767	0.219
VII	1,090,947	120	0.909	0.115	800	0.802	0.128
XV	1,091,289	110	0.927	0.102	700	0.828	0.121
IV	1,531,919	150	0.915	0.120	1100	0.809	0.143
Avg	-	-	0.907	0.120	-	0.822	0.142

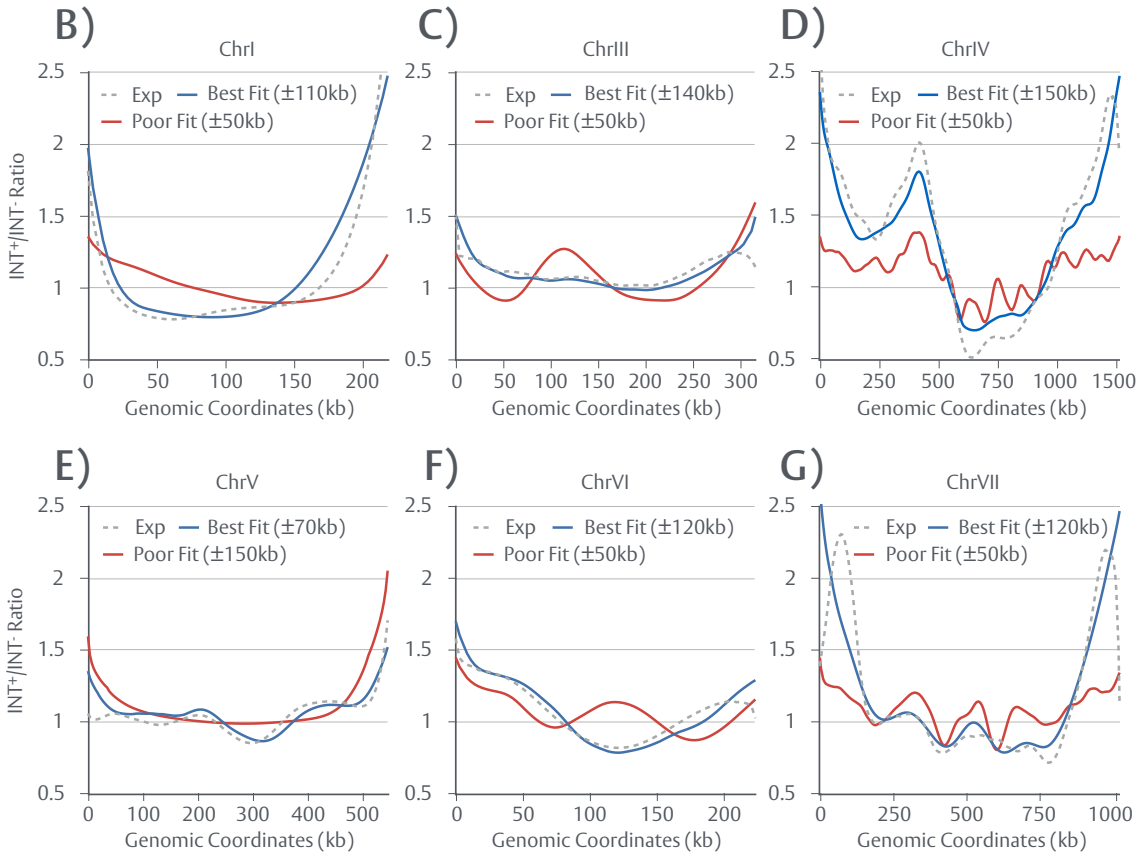
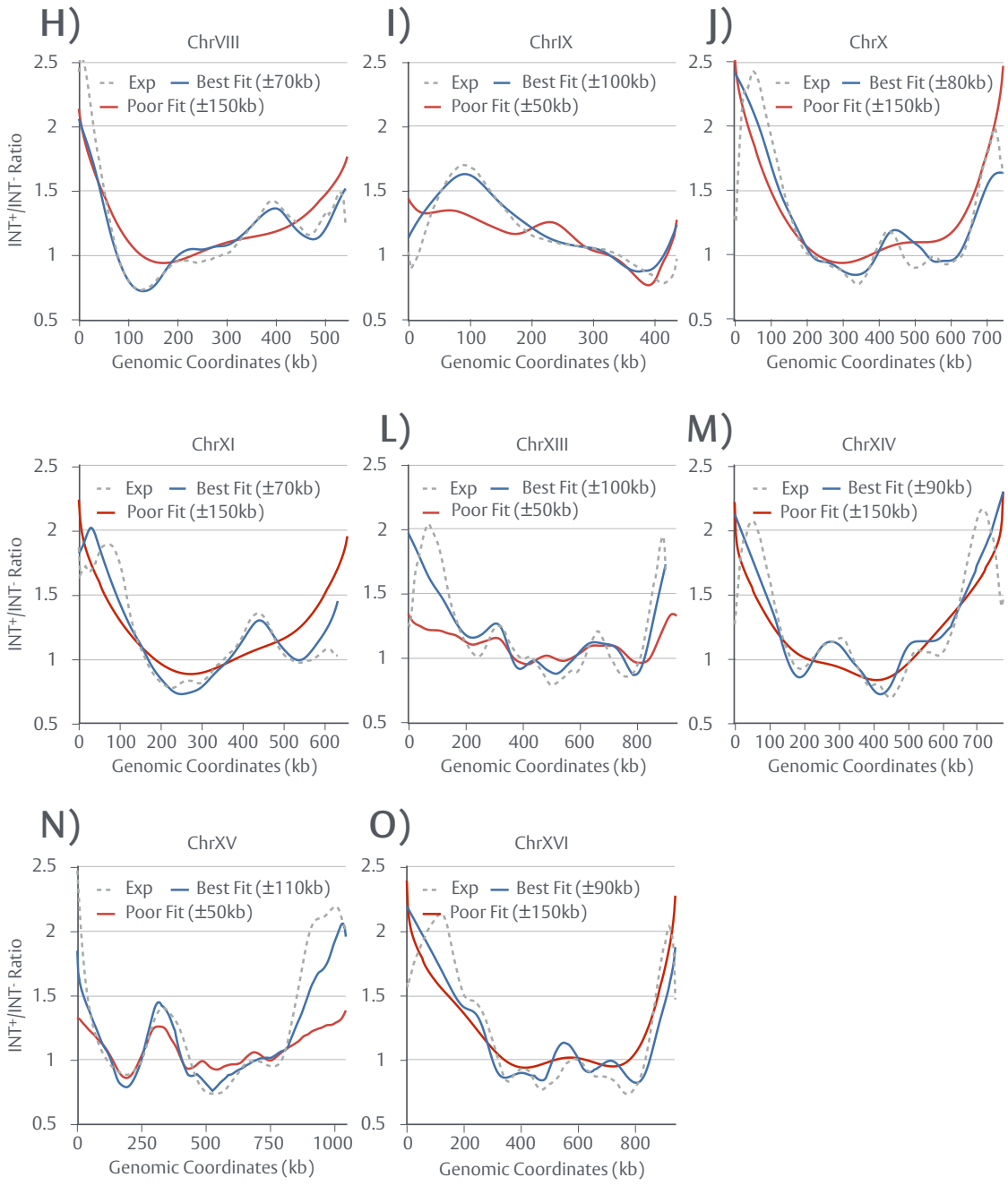


Figure 4.13. Best fit DSBSim models

Interfering simulations (*DSBSim* mode: *Trans*) were conducted on *sae2*Δ*ndt80*Δ*tel1*Δ input models using varying widths of Hann or Exponential window interference. Resulting simulated INT+/INT- ratios were smoothed (moving average, n = 3) and statistically screened against the corresponding experimental ratio as previously described. **A)** Best fit parameters, as determined by ρ (MATLAB 2017a Package: *corr*) and AvD values and a statistical similarity threshold of ($\rho > 0.75$, $AvD < 0.15$) are tabulated for all chromosomes except ChrII and ChrXII. Data is sorted in descending order based on chromosomal size. **B-O) (see: Next Page)** Best and worst fit simulated ratios, as determined by ρ and AvD values, are overlaid with experimental ratios (*sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ) for each tabulated chromosome.


H)
ChrVIII
Exp
Best Fit (±70kb)
Poor Fit (±150kb)
INT+/INT- Ratio
Genomic Coordinates (kb)
I)
ChrIX
Exp
Best Fit (±100kb)
Poor Fit (±50kb)
Genomic Coordinates (kb)
J)
ChrX
Exp
Best Fit (±80kb)
Poor Fit (±150kb)
Genomic Coordinates (kb)
ChrXI
Exp
Best Fit (±70kb)
Poor Fit (±150kb)
INT+/INT- Ratio
Genomic Coordinates (kb)
L)
ChrXIII
Exp
Best Fit (±100kb)
Poor Fit (±50kb)
Genomic Coordinates (kb)
M)
ChrXIV
Exp
Best Fit (±90kb)
Poor Fit (±150kb)
Genomic Coordinates (kb)
N)
ChrXV
Exp
Best Fit (±110kb)
Poor Fit (±50kb)
INT+/INT- Ratio
Genomic Coordinates (kb)
O)
ChrXVI
Exp
Best Fit (±90kb)
Poor Fit (±150kb)
Genomic Coordinates (kb)

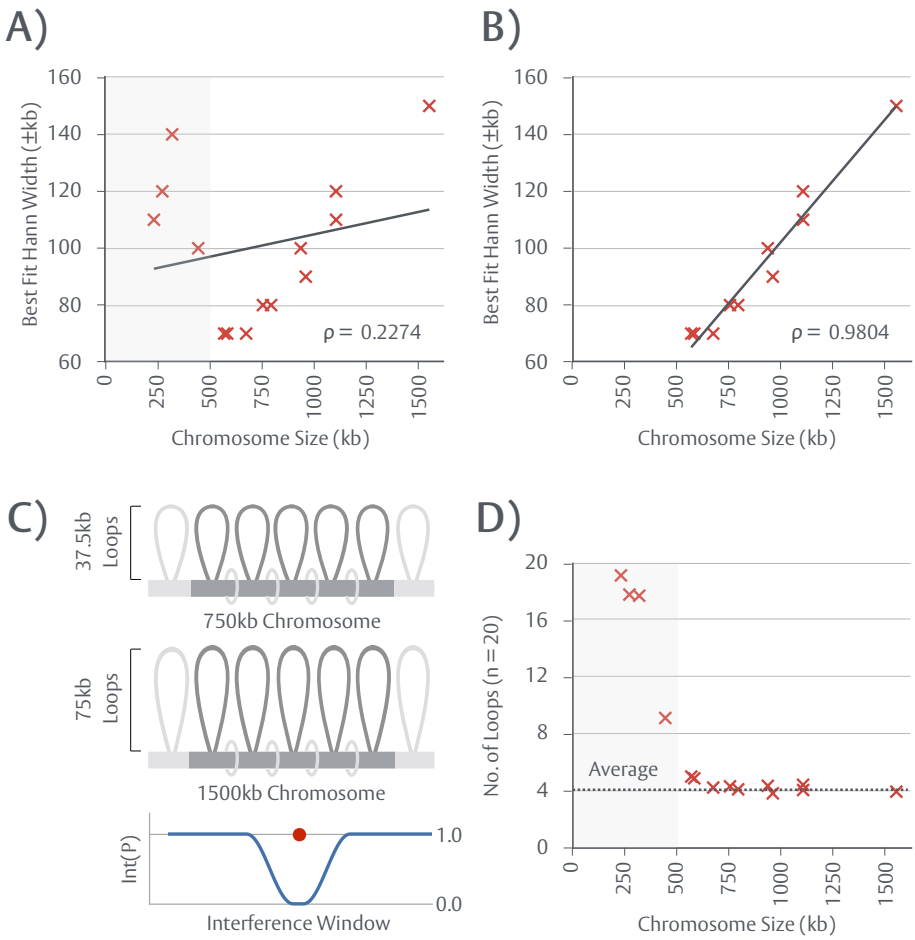


Figure 4.14. DSB interference is predicted to act over a fixed spatial distance

A) Best fit model parameters (width ±kb), for Hann interference simulations (*DSBSim* mode: *Trans*), were plotted against *S. cerevisiae* chromosome size. **B)** Parameters for smaller <500kb chromosomes (ChrI, ChrIII, ChrVI, ChrIX) were omitted. **C)** Differential amounts of DNA on each chromosome may be accommodated through variable loop size, while total loop count (n) remains, on average, constant. Average loop sizes, as marked, are calculated for (n = 20). Such a model may allow DSB interference (shown below) to act over a fixed, spatial range in three dimensional space, while suppressing DSB formation over different amounts of DNA (in kb) proportional to the length of the chromosome. **D)** Each chromosome was segregated into a fixed number of loops (n = 20) of differential size (in kb), based on chromosomal size. Best fit model parameters were converted from (kb) to the number of loops the signal is predicted to span and replotted. Linear correlations were assessed via Pearson's rho (MATLAB 2017a Package: *corr*) (see: marked ρ values). Linear trendlines are marked onto each plot.

Specifically, smaller, <500kb chromosomes appear to require interference functions considerably wider than anticipated by length alone. Consistent with a requirement for wider interference, ChrI, ChrIII, ChrVI and ChrIX experimental ratios are devoid of internal domains of concerted change (see: Figure 4.13B,C,F,I) (see: Section 4.8). By removing these chromosomes (ChrI, ChrIII, ChrVI, ChrIX) from consideration, a strong, linear correlation is observed ($\rho = 0.9804$) as if the range DSB interference extends is directly proportional to the length of the chromosome (Figure 4.14B). However, a more plausible explanation for this observation is that DSB interference acts over a fixed distance in three dimensional space. To facilitate such a model, differing amounts of DNA would have to be packaged into similar or identical areas of space—a requirement that may be met by differently sized chromatin loops (Figure 4.14C). While the distance between axial Rec8 or Rec114-Mer2-Mei4 (RMM) ChIP peaks remains, on average, similar between all chromosomes (Panizza et al. 2011; Ito et al. 2014) (~12.1kb average), ChIP data represents a population average. It thus remains unknown whether or not every single Rec8/RMM axial site is occupied within a given, individual cell. Under the assumption that, due to differential levels of Rec8/RMM occupancy, the number of loops formed per chromosome is fixed (n), the amount of DNA (in kb) that must be packaged into each loop, on average, can be calculated per chromosome. For example, at (n = 20), the average loop size on ChrI is 11.5kb (230.2kb/20), as opposed to 76.5kb on ChrIV (1531.9kb/20). In simpler terms, the larger the chromosome is, the larger the loops need to be in order to accommodate the full length of the chromosome when (n) remains constant. Larger loops may not, however, necessarily occupy a considerably increased zone of three dimensional space and/or portion of the chromosomal axis—allowing an interfering signal, static in range, to traverse more base pairs than it would for smaller loops (see: Figure 4.14C). Notably and consistent with the ability of loops to facilitate differing amounts of DNA, loop size is markedly increased within mammals—which possess considerably larger genomes—while predicted, total loop count is not (Blat et al. 2002; Kleckner 2006; Novak et al. 2008; Kauppi et al. 2011). The width of best fit interference may thus be expressed in alternative terms instead of (kb), based on the number of loops the signal is predicted to encompass for a given value of (n).

Remarkably, all >500kb chromosomes are predicted to employ interference windows that span a near identical number of loops (~3.8-4.2) (n = 20) (Figure 4.14D). In contrast, chromosomes <500kb in size are predicted to employ interference windows that span a substantially increased ~9-20 loops (n = 20) (see: Figure 4.14D). While the choice of (n) value is arbitrary, these relationships would be maintained for any value—collectively suggesting DSB interference acts over a fixed, spatial distance, perhaps defined by structural units such as chromatin loops.

4.10—Simulated interference is sufficient to negate the effects of loop activation

Negative interference is revealed by inactivation of Tel1 (*tel1*Δ), otherwise masked within interfering *TEL1*$^{+}$ backgrounds (Garcia et al. 2015) (Figure 4.15). Interference values for *sae2*Δ*TEL1*$^{+}$ were calculated using the standard formula (1-OBS/EXP) as before. Experimental data (OBS) is obtained via southern blot using probes placed either side of the central, major *ARE1* hotspot peak (Garcia et al. 2015). Expected (EXP) double cut frequencies were recalculated using *sae2*Δ Spo11 DSB hotspot data normalised to an *ARE1* DSB frequency of 8.37% as determined by southern blot for this genotype (Garcia et al. 2015). As negative interference is no longer apparent in this background, a model with an accurate implementation of DSB interference must suppress the effects of loop activation. Interfering simulations (*DSBSim* mode: *Trans*) were thus conducted for ChrIII, using a non-uniform Gaussian boost at loop activation frequencies of 30-40% and the best fit Hann interference window identified for ChrIII (±140kb). As before, virtual probes were placed either side of *ARE1*, calculating the frequencies of any given inter-hotspot double cut. DSB interference, as simulated, is capable of suppressing concerted DSB formation to a similar extent to that observed experimentally, at both activation frequencies (see: Figure 4.15), further validating the way in which *DSBSim* models interference. Notably, the simulated interference values highly resemble those obtained when using experimentally, observed data (see: Figure 4.15).

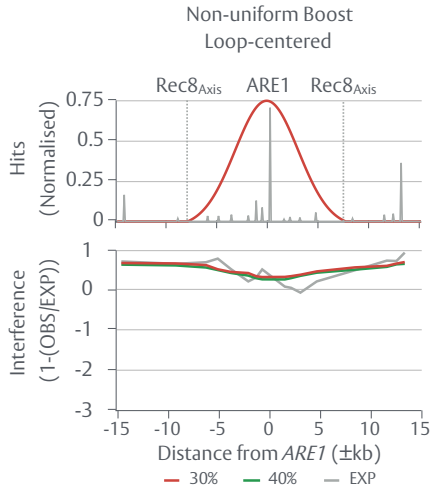


Figure 4.15. Simulated DSB interference is sufficient to negate the effects of loop activation

Experimentally observed interference values for *sae2*Δ*TEL1*$^{+}$, across the *ARE1* locus (shown above), were calculated using the standard interference formula (1-OBS/EXP)—an inverted ratio between the observed frequency of inter-hotspot double cuts and the expected frequency. Observed data was obtained from (Garcia et al. 2015). Expected data was recalculated, based on *sae2*Δ hotspot data normalised to an *ARE1* frequency of 8.37% by multiplying the quantitative strength of each hotspot pair. Interfering simulations (*DSBSim* mode: *Trans*) were conducted on an inverted ChrIII *sae2*Δ*tel1*Δ input map, using the identified best fit Hann interference window (±140kb) and loop activation frequencies of 30% and 40%. Simulated OBS values were obtained using virtual probes placed either side of the main *ARE1* peak.

4.11—DSB clustering, as a result of loop activation, may skew the distribution of COs

Recent work, presented in (Section 2.20) and (Anderson et al. 2015), suggests inactivation of Tel1 may shift the class I CO:class II CO ratio toward class II CO formation, thus weakening the global CO landscape as events form with a higher degree of randomness. A weakening of CO interference is characterised by an enrichment in smaller IEDs within *tel1*Δ*msh2*Δ relative to *msh2*Δ (see: Figure 2.22B,C). However, local clustering of DSBs, via loop activation, as opposed to increased class II CO formation may provide an alternative explanation for the formation of smaller IEDs.

In order to investigate how clustered DSB formation may influence recombination outcomes, random CO simulations (*RecombineSim* mode: *Random*) (see: Section 2.6) were repeated for *tel1*Δ*msh2*Δ with site selection weighted according to per cell DSB maps produced under *DSBSim*, using variable loop activation frequencies (10-40%). Formation of COs at sites that did not form a DSB was fully disallowed. Resulting non-weighted (fully random) and weighted IEDs were compared by evaluating their respective cumulative distribution functions (CDFs) at 1kb (x) intervals and calculating a ratio ($f(X_{NW})/f(X_W)$) (Figure 4.16A)—as previously described (see: Figure 2.11). Loop activation frequencies of 10-30% are sufficient to produce a global skew across the full IED range toward smaller IED sizes—suggesting inactivation of Tel1 within a CO interference deficient strain may have a profound "feedforward" effect on the distribution of recombination. The impact of loop activation is considerably diminished at a frequency of 40%, which only exhibits a minor enrichment in smaller IEDs <30kb.

CO interference may, however, overcome the impact of DSB clustering. To assess how additional, downstream processes of spatial regulation (e.g. CO interference) may interact with DSB clustering, CO interference was reintroduced into this unified model using the universal, class I window previously identified (*RecombineSim* mode: UniHazard) (see: Figure 2.16C). Recalculated fractional ratios ($f(X_{NW})/f(X_W)$) between interfering simulations without and with DSB clustering yield values of ≈ 1 across the full spectrum of IED size (Figure 4.16B).

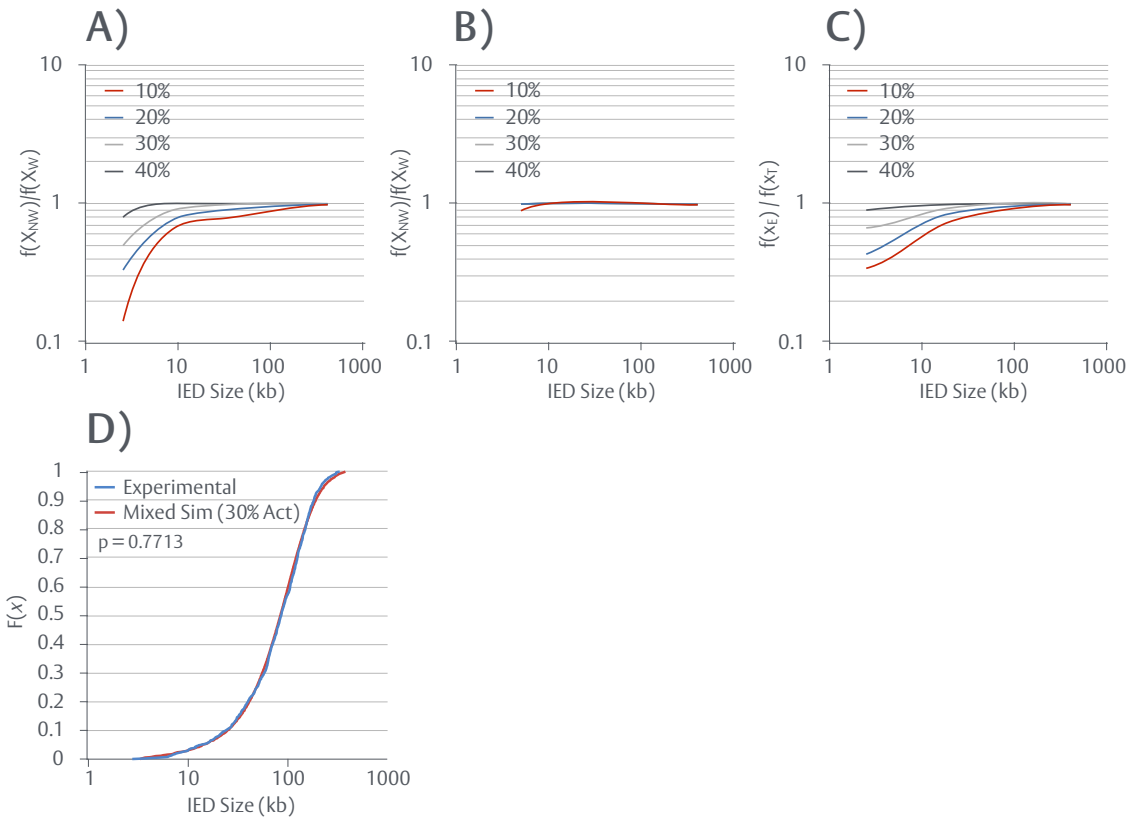


Figure 4.16. DSB clustering, as a result of loop activation, may skew the distribution of COs

A) Maps of DSB distribution were generated on a per cell basis, via *DSBSim*, for all chromosomes at variable loop activation frequencies. Weighted, non-interfering CO simulations (*RecombineSim* mode: Random), whereby DSB position is used in place of recom(P), were conducted using *tel1Δmsh2Δ* CO event counts. Resulting non-weighted (fully random) and weighted IEDs were analysed by evaluating their respective cumulative distribution functions (CDFs) as 1kb (x) intervals and comparing them as a ratio ($f(X_{NW})/f(X_W)$). Ratios are shown on double log plots. Ratio values of <1 denote that the weighted data contains IEDs, of a given size, at a higher frequency than the non-weighted dataset. **B)** Weighted, interfering CO simulations (*RecombineSim* mode: UniHazard) were conducted using *tel1Δmsh2Δ* CO event counts and the previously identified, universal class I CO interference window (see: Figure 2.17C). Resulting IEDs, without and with weighting, were compared as before. **C)** Mixed, weighted, interfering CO simulations (*RecombineSim* mode: UniHazard) were conducted using *tel1Δmsh2Δ* CO event counts, the previously identified, universal class I CO interference window and a class II frequency of 15.4%. Resulting IEDs, with and without weighting, were compared as before. **D)** A best fit model for *tel1Δmsh2Δ*, overlaid with experimental data as a CDF, is obtained using a loop activation frequency of 30% and a mixed, interfering simulation. Model-experimental fits were assessed via two sample KS (MATLAB 2017a Package: *kstest2*) (see: (p) value).

Minor deviation, toward smaller IEDs, is only observed at 10% loop activation—suggesting that CO interference is sufficient to negate the effects of DSB clustering, evenly spreading class I COs as normal.

As previously shown (see: Figure 2.14A), WT cells contain a mixture of interfering class I and non-interfering class II COs (~32% class II). DSB clustering may thus specifically skew the placement of class II COs closer to one another or to class I sites, accounting for the enrichment of smaller IEDs within *tel1*Δ*msh2*Δ without a need to increase class II frequency. In order to test this hypothesis, interfering CO simulations (*RecombineSim* mode: UniHazard) were repeated for *tel1*Δ*msh2*Δ under *msh2*Δ conditions (15.4% class II's, universal class I window), with DSB clustering present. In other words, this simulation assesses the idea that nothing else has changed within *tel1*Δ*msh2*Δ relative to *tel1*Δ*msh2*Δ bar the placement of DSBs. Fractional ratios ($f(X_{NW})/f(X_W)$) were recalculated as before, demonstrating a more appreciable effect of clustered DSB formation on the downstream distributions of COs when class II COs are included in the system (Figure 4.16C). As before, any effect is largely confined to loop activation frequencies of 10-30%. Resulting IED distributions—from mixed simulations containing DSB clusters—were statistically screened via two sample KS test to obtain a best fit to the experimentally observed distribution of COs within *tel1*Δ*msh2*Δ (Figure 4.16D). A statistically significant model fit ($p = 0.771$), comparable to that generated by increased class II frequency ($p = 0.827$) (see: Figure 2.21D), is obtained for a loop activation frequency of 30% —a frequency in line with that required to model negative interference at the *ARE1* locus, suggesting such a value may apply genome-wide. Collectively, these results suggest that three distinct mechanisms may explain the distribution of COs within *tel1*Δ*msh2*Δ: (i) an increased class II CO frequency (ii) DSB clustering as a result of loop activation or (iii) a combination of both to varying extents.

4.12—Discussion

Work presented here details development and usage of a novel simulation platform for the analysis of DSB distributions. The impact Tel1^{ATM} activity exerts over the population average distribution of DSBs has been further characterised and evidence implicating Tel1 in a branch of *trans* interference was strengthened. Moreover, a proposed model for the phenomenon of negative interference has been examined.

DSB interference

Tel1^{ATM}-dependent DSB interference manifests within the population average distribution of DSBs as quantitatively modest domains of concerted change readily detected by genome-wide mapping of Spo11 DSBs and accurately modelled by simulated processes of *cis* and *trans* interference between two chromatids (see: Section 4.8). Consistent with these findings, Tel1 has been previously implicated within a form of *trans* interference (Zhang et al. 2011). *Trans* interference produces a distinct population average outcome to that of *cis* interference alone, characterised by intensified ratio changes—perhaps a previously under-appreciated hallmark of the process. Interestingly, while *sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ ratios exhibit patterns explainable by DSB interference, *sae2*Δ/*sae2*Δ*tel1*Δ ratios do not (see: Section 4.3). Such an observation may reveal how DSBs are patterned by alternative, underlying processes earlier in prophase I—such as DNA replication (see: Section 1.4.5)—prior to the imposition of DSB interference. However, given the partial similarities between *sae2*Δ*ndt80*Δ/*sae2*Δ*ndt80*Δ*tel1*Δ and *sae2*Δ/*sae2*Δ*tel1*Δ ratios—characterised by retention of subtelomeric enrichment (see: Figure 4.2E,F)—a more plausible explanation is that the DSB landscape must fully mature, reaching maximal levels of formation, before DSB interference is fully visible in population data. Therefore, early in prophase I, the generation of an interfering signal within the chromosomal arms appears to favour DSB formation toward the subtelomeres. As DSB formation progresses and interfering signals progressively layer upon one another, creating complex patterns of inhibition, internal domains of concerted change are shaped and fully constructed as all remaining DSBs form.

Unexpectedly, narrow, denser regions of inhibition (Hann windows) more closely recapture experimental ratios than a broader, diffusive signal (Exponential windows) (see: Figure 4.12). An exponential decay is perhaps expected for a spreading, distance-dependent process emanating from a single focal point which appears to act over a fixed spatial distance. Moreover, the width of DSB interference, when modelled via *DSBSim*, exhibits a linear relationship with chromosomal size on all >500kb chromosomes. As outlined in (Section 4.9), such observations support a model whereby DSB interference acts over a fixed, spatial distance defined by something other than the mechanics of diffusion—such as a loop or a cluster of loops—and where loop size, but not total loop count, differs between chromosomes (see: Figure 4.14C). In contrast to >500kb chromosomes, simulation of smaller chromosomes (ChrI, ChrIII, ChrVI, ChrIX) requires interference windows much wider than predicted by chromosome length alone. Should smaller chromosomes exhibit disproportionately lower levels of Rec8/RMM loading, and thus disproportionately larger loops within any given cell, the predicted spatial range of interference would be more in line with other chromosomes (see: Figure 4.14D). Interestingly, previous studies have observed a disproportionately *increased* loading of Rec8/RMM onto ChrI, III and VI when assessed by ChIP-chip (Panizza et al. 2011)—suggesting smaller chromosomes, instead, exhibit higher occupancy of axis sites and therefore construct smaller loops. However, enrichment of ChIP-chip signal may, alterantively, arise if Rec8/RMM prove to be disproportionately stable on smaller chromosomes. Importantly, cohesion complexes, such as Smc1-Smc3-Rec8-Scc3, are thought to act as extrusion factors, directly generating chromatin loops (Barrington et al. 2017). Stabilised association of Rec8 with smaller chromosomes may therefore result in higher rates of extrusion and thus, larger loops—explaining the apparent requirement for extensively broad interference on these chromosomes.

Trans interference may occur between sister chromatids (inter sister), or homologues (inter homologue). While results presented in this chapter provide evidence against a purely in *cis* model of Tel1^{ATM}-dependent DSB interference, exactly what form of *trans* interference Tel1 is involved in is not discernible by these methods. Numerous, unique combinations of model parameters may

achieve model fits of similar accuracy. For example, expansion of *DSBSim* to include four chromatids may permit a narrower or less intense window of *trans* interference to recapture experimental results—even though Tel1 may only mediate inter sister interference *in vivo* as previously suggested (Cooper et al. 2014; Cooper et al. 2016).

Negative Interference

Localised, concerted formation of DSBs—confined to singular loop domains—may be readily explained by an upstream process of loop activation that is applied in a non-uniform manner (see: Section 4.6). Appreciable levels of negative interference, as simulated, are observed for loop activation frequencies of 10-60%, as opposed to 10-20% when applying a uniform boost. Such an observation may be readily explained by an effective narrowing of the DSB competent region. Specifically, during non-uniform boosting, hotspots toward the loop periphery experience only relatively minor boosts such that a reduced genomic region toward the centre of the loop must accommodate a fixed level of DSBs—increasing the chance of concerted formation (see: Figure 4.7). However, it remains unclear how a non-uniform activation process may arise. The inherent, biomechanical properties of DNA may impose constraints on the tethering of chromatin loops to the axis such that centralised tether points are favoured. Within such a model, the probability of DSB formation (DSB(P)) increases toward the tether point in a distance-dependent manner (Figure 4.17A). However, induction of DSBs proximal to Rec8 binding sites is notably inefficient (Ito et al. 2014). An otherwise uniform activate process may therefore, alternatively, interact with repression emanating from the flanking axis sites—producing a composite, non-uniform effect (Figure 4.17B).

By combining *RecombineSim* and *DSBSim* (see: Section 4.11), thus generating a unified model, potential and novel consequences for the clustering of DSBs upon the ultimate outcome of meiotic recombination may have been uncovered. Notably, under non-interfering conditions, DSB clustering has a profound impact upon the distribution of COs (see: Figure 4.16A). Such findings also apply to NCOs, which do not exhibit any detectable spatial regulation (see: Section 2.10).

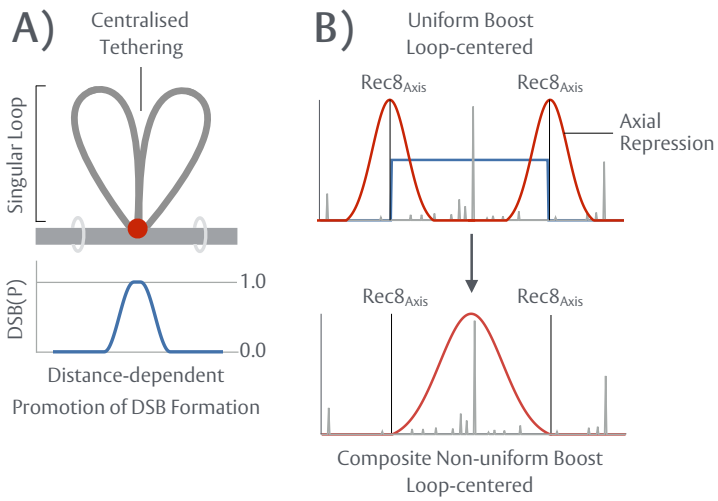


Figure 4.17. Models for non-uniform loop activation

Non-uniform loop activation may potentially arise through two, distinct mechanisms: **A)** Biomechanical constraints, perhaps imposed by the rigidity of the DNA polymer, may favour centralised tethering of any given chromatin loop. As tethering is primarily mediated by machinery required for DSB formation (e.g. Mer2) (see: Section 1.2.10), DSBs may more readily form in proximity to the tether point—disproportionately favouring centralised hotspots for breakage. **B)** Rec8, bound to the axis, suppresses proximal DSB formation (Ito et al. 2014). An otherwise uniform activation process may therefore interact with a distance-dependent, repressive signal emanating from the flanking axis sites—producing a composite, non-uniform boost.

The impact of loop activation (i.e. DSB clustering) may be fully negated by secondary layers of spatial regulation—as observed for simulated CO interference (see: Figure 4.16B). Thus, even at low loop activation frequencies (e.g. 10%), enough regions scattered across each chromosome exist, on average, to support to even spacing of COs. However, inclusion of simulated non-interfering, class II COs partially weakens the ability of CO interference to offset DSB clustering as class II COs are free to form within clustered regions (see: Figure 4.16C). DSB clustering thus has the potential to influence CO distributions, increasing the frequency of closely spaced double CO events, without a necessary increase in class II CO frequency. Nevertheless, genetic evidence still favours a model whereby class II CO frequency is increased within *tel1*Δ (Anderson et al. 2015). Such an increase may, however, still occur as a result of DSB clustering. For example, formation of COs in close proximity to one another may preclude normal, simultaneous repair—forcing one or both events to enter the class II CO pathway, which appears more adept at handling abnormal joint molecules (Jessop & Lichten 2008).

4.13—Summary (Key Points)

- Inactivation of Tel1^{ATM} results in a non-random redistribution of DSBs at the population level (Sections 4.2-4.3)
- Constructed a novel platform for the simulation and analysis of DSB distribution (*DSBSim*) (Section 4.4)
- Localised regions of concerted DSB formation can be generated via simulated loop activation (Section 4.6)
- Tel1^{ATM} may mediate both *cis* and *trans* interference (Sections 4.8-4.9)
- DSB interference appears to act over a fixed spatial distance (Section 4.10)
- Simulated clustering of DSBs is able to skew the downstream distribution of COs (Section 4.11)

CHAPTER 4B

Genome-wide mapping of Spo11 DSBs

Appendix

Appendix

B4.1—*DSBSim* (*v1.8*)

Aim: Simulation of DSB distributions, DSB clustering

Input(s): 1bp histogram, annotated hotspots, inverted maps

Output: Simulated hotspot maps, probed fragment sizes

Req(s): MATLAB (2017a)

DSBSim constitutes a novel simulation platform specifically designed to handle *Spo11Mapper* data formats (see: Section B3.1). *DSBSim*, designed within MATLAB (2017a), provides a callable function for automated job queuing:

DSBSim(*Chromosome*, *SimulatedSampleSize* (*M*), *DSBNumber*, *WindowIterations*, *ActV*, *IntMode*, *IntParameters*)

Exampled Queue (M = 1000)

DSBSim(*11*,*100000*,*250*,*25*,*30*,'*Hann*',[*10*:*10*:*300*])

DSBSim(*12*,*100000*,*250*,*25*,*30*,'*Random*',[*10*:*10*:*300*])

DSBSim(*13*,*100000*,*250*,*25*,*30*,'*Exp*',[*100*:*100*:*2500*])

B4.1.1—Simulation (Virtual Chromosomes)

Virtual hotspot maps, upon which simulated DSB formation occurs, are constructed at a 100bp resolution as binned, numerical arrays proportional in size to *in vivo* (chromosome length*0.01) (*S. cerevisiae*—S288c). Any given 100bp bin contains a value, based on normalised hotspot data, in the range of [0.0-1.0]x10^6 that describes the relative probability of that bin being selected for DSB formation (DSB(P)). Bin values, in the units of NormHpMCHr300, are derived from *sae2*Δ*tel1*Δ*ndt80*Δ data—mapped and processed via *Spo11Mapper*. NormHpMChr300 is calculated as the sum of Spo11 5' hits across annotated hotspots, as previously defined (Pan et al. 2011), with an additional 300bp upstream and downstream included to account for signal spreading within

*tel1*Δ strains (see: Section 3.12). Values are subsequently normalised on a per chromosome basis, expressing hotspot strength as a fraction of the total number of hits on each chromosome.

B4.1.2—Simulation (Loop Boundaries)

Any given chromosome is segregated into numerous, structural subdomains using Rec8 ChIP-seq data (Ito et al. 2014). A loop boundary is defined as any singular 100bp bin whose position is defined by Rec8 peak maximas, as determined by a pre-existing peak calling algorithm (MATLAB 2017a Package: *findpeaks*), residing 2-fold above the background level as specified (Ito et al. 2014). In total, 919 loops were identified across all 16 chromosomes, with an average size of 12.3kb—consistent with previous estimates (Ito et al. 2014).

B4.1.3—Simulation (Loop Activation & Map Inversion)

Loop activation, a user specified parameter, is set as a decimal fraction in the range of [0.0-1.0] (0-100%) via the ActV parameter. For any given chromosome, the number of loops to activate (nloops) is rounded to the nearest integer based on AtcV. For example, for ChrII and ActV = 0.3, 20 loops will be activated (66 total loops*0.3 = 19.8). During any given round of simulation, loop selection is performed randomly by sampling the possible loops without replacement (MATLAB 2017a Package: *datasample*). Loops are activated using uniform or non-uniform boosts. Uniform boosts are applied as a "flat" array containing values of [100,000]—a value chosen to sufficiently disallow DSB formation from occurring in non-activated loops. Non-uniform boosts are applied as a Gaussian window (MATLAB 2017a Package: *gausswin*), normalised between values of [1.0] and [100,000]. All boost arrays are adjusted in length to match the width of the particular loop being activated, and applied through multiplication of DSB(P) values. To accommodate non-uniform boosting, inverted maps are generated by applying inverted (1-boost) boost windows across all identified loop domains. The extent of inversion is governed by ActV and thus the frequency at which any given loop experiences a boost during simulation.

B4.1.4—Simulation (Site Selection, Event Formation & Interference)

DSB(P) values are sensed, in order to determine the position of potential DSB formation, via a weighted, roulette wheel selection algorithm (RWS). As previously described (see: Section B2.2.4), RWS constructs a set of array segments where lengths are proportional in length to each DSB(P) value. A position along the full, combined array is subsequently chosen at random. Higher DSB(P) values, which create larger array segments, thus translate into higher probabilities of being chosen for DSB formation. Average DSB count per chromosome is calculated based on three parameters: (i) a user specified, cell wide DSB count (e.g. 200DSBs/cell) (ii) the length of a given chromosome (iiii) the number of chromatids being simulated. For example, ChrII constitutes 6.74% of the full genome length thus at 200DSBs/cell, ChrII should collectively incur—on average—13.48 DSBs (200*6.74%) across all four chromatids. For a single chromatid simulation, this is reduced to 3.37 DSBs on average (13.48/4). To allow for variation in DSB count and the use of fractional averages, DSB count for any given simulation is determined by a Poisson distribution—as previously proposed for DSB formation (Toyoizumi & Tsubouchi 2012). All simulations throughout this chapter were conducted at 250DSBs/cell. Importantly, DSB frequency does not appreciably alter the outcome of *cis* or *trans* interfering simulation trials (Figure 4.18A,B respectively). As DSB frequency is lowered, the size of the population simulated must increase to maintain signal:noise ratios.

Unlike *RecombineSim*, where interference directly modifies Recom(P) values (see: Section B2.2.4), *DSBSim* employs a different approach to avoid complicating loop activation mechanics and to permit DSB formation to fail. A secondary array, initially populated with [1.0] values and identical in length to the simulated, virtual chromosome, is constructed. Each 100bp bin within this secondary array contains values denoting the probability that a DSB, at the selected site, will successfully form (Int(P)). Upon formation of an initial DSB, *DSBSim* imposes interference centred on the event as an exponential window—described by a slope factor (μ) (MATLAB 2017a Package: *exppdf*)—or as a hann window of a given width (MATLAB 2017a Package: *hann*), by altering Int(P) values.

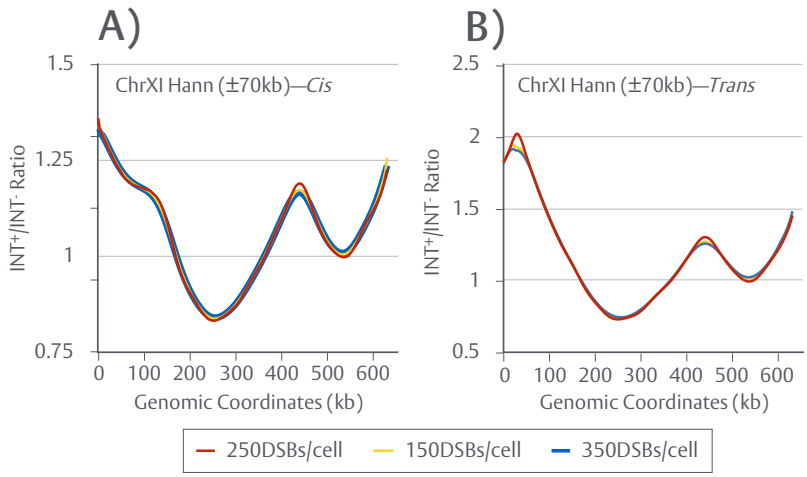


Figure 4.18. DSB frequency does not appreciably alter the population average

Interfering simulations (*DSBSim* mode: *Cis*, *Trans*) were conducted on a *sae2Δndt80Δtel1Δ* ChrXI input model using a ±70kb Hann window and varying DSB frequencies (150, 250, 350 DSBs/cell). Resulting simulated INT^+/INT^- ratios were smoothed (moving average, n = 3) and plotted for **A)** *Cis* and **B)** *Trans*.

The success of all subsequent DSBs is determined by a randomly generated number (D) in the range of [0.0-1.0] which is checked against the Int(P) for the corresponding bin. If D<Int(P), the DSB forms and interference is imposed. If D>Int(P) the DSB fails to form and no further attempt is made to place it at an alternative position. Thus, with interference applied, the user specified DSB frequency (e.g. 250 DSBs/cell) is likely to be reduced through event failure. All interference windows are normalised between values of [0.0-1.0], such that formation of a secondary DSB in the immediate vicinity of another is fully disallowed.

Script—DSBSim.m (DSB Simulation)

```
DSBSim(ChrModel,Samples,DSB,Iterations,Actv,IntMode,wInt,iInt,TukeyFactor,Slope,VFactor)

%% Data Import & Processing
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%=%%%%%%%%%%
load('WTMapsNormHpMChr.mat')
invfilename = strcat('Chr',num2str(ChrModel),'Inversion30','.mat');
invmodel = cell2mat(struct2cell(load(invfilename, 'invmodel')));
loopweight = cell2mat(struct2cell(load(invfilename, 'weight')));
fid1 = fopen('Rec8-S288c.txt','r');
data = textscan(fid1, '%f %f %*[^\r\n]');
genome = sum(cellfun('length',WTMaps));
data = cell2mat(data(:,1:2));
index = find(data(:,1)==ChrModel);
rec8 = sort(data(index,:));
rec8(:,2) = round(rec8(:,2)/100);
nloops = length(rec8)-1;
actloop = round(Actv/100*nloops);
regions = zeros(length(rec8)-1,2);
for i=1:nloops
  regions(i,1) = rec8(i,2);
  regions(i,2) = (rec8(i+1,2)-1);
end
lsize = diff(rec8(:,2));
boosts = cell(1,length(lsize));
for n=1:length(lsize)
  boostwin = gausswin(lsize(n));
  boosts{1,n} = normalize_var(boostwin,1,100000);
end
Top = repmat(TukeyFactor,25,1);
FlatTop = repmat(VFactor,Iterations,1);
TRstr = 0.3;
Save = 'N';
DC = 0;

%% DSBS/cell
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
fid2 = fopen('AdjustedDSB.txt','r');
DSBmean = cell2mat(textscan(fid2, '%f %f %*[^\r\n]'));
DSBmean = (DSBmean(ChrModel,2)*DSB)/4;
DSBn = poissrnd(DSBmean,Samples,1);
%% Distributions & Interference
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for v=1:Iterations
  model = cell2mat(struct2cell(load(invfilename, 'model')));
  disp('Currently Simulating Iteration:');
  disp(v);
  allcuts = cell(sum(DSBn),1);
  dsst = zeros(max(DSBn),Samples);
  if strcmp(IntMode,'Exp') == 1
    dist = fliplr(1-(normalize_var(exppdf(1:20000,Slope(v)),0,iInt)));
    dist(20001:40000) = fliplr(dist);
    dist = transpose(dist(15000:25000));
    buffer = zeros(5000*10,1);
    r = 5000;
    tr = normalize_var(dist,1-TRstr,1);
  elseif strcmp(IntMode,'Tukey') == 1
```

```
    dist = 1-(normalize_var(tukeywin((wInt(v)*20)+1,Top(v)),0,iInt));
    buffer = zeros(wInt(v)*10,1);
    r = wInt(v)*10;
    tr = normalize_var(dist,1-TRstr,1);
  elseif strcmp(IntMode,'Random') == 1
    dist = ones(101,1);
    buffer = zeros(100*10,1);
    r = 50;
    tr = ones(r+1,1);
  elseif strcmp(IntMode,'Volcano') == 1
    dist = fliplr(1-(normalize_var(exppdf(1:20000,Slope(v)),0,iInt)));
    dist(20001:20001+(FlatTop(v)/100)) = 0;
    dist(20001+(FlatTop(v)/100):20001+(FlatTop(v)/100)+19999) = fliplr(dist(1:20000));
    dist = transpose(dist(15000:25000+FlatTop(v)/100));
    buffer = zeros(5000*10,1);
    r = 5000+(FlatTop(v)/200);
    tr = normalize_var(dist,1-TRstr,1);
  elseif strcmp(IntMode,'DeadZone') == 1
  end
  for s=1:Samples
    smodel = [buffer;invmodel;buffer];
    intf = [buffer;ones(length(invmodel),1);buffer];
    prime = datasample(1:nloops,actloop,'Replace',false,'Weight',loopweight);
    cutp = [];
    DCpl = [];
    DCpr = [];
    for z=1:actloop
      sel = prime(z);
      smodel(regions(sel,1)+length(buffer):regions(sel,2)+length(buffer)) = smodel(regions(sel,
1)+length(buffer):regions(sel,2)+length(buffer)).*boosts{sel};
    end
    for q=1:DSBn(s)
      weight = smodel(min(regions(:,1))+length(buffer):max(regions(:,2))+length(buffer));
      pos = min(regions(:,1))+length(buffer):max(regions(:,2))+length(buffer);
      ds = pos(sum((rand(1) >= cumsum(weight./sum(weight))))+1);
      if ismember(2162+length(buffer),ds)
        count(s) = 1;
      end
      check = rand();
      if check < intf(ds)
        intf(ds-r:ds+r) = intf(ds-r:ds+r).*dist;
        cutp(q,1) = ds-length(buffer);
      else
        continue
      end
    end
    cutp(cutp==0) = [];
    allcuts{s,1} = cutp;
    DSBpc(s,1) = length(cutp);
    cutp = sort(cutp);
    if ismember(2162,cutp)==1
      cutp(cutp<2000) = [];
      cutp(cutp>2300) = [];
      DCpl = max(find(cutp<2162));
      DCpr = min(find(cutp>2162));
      if isempty(DCpl)==0
        DC=DC+1;
        DoubleCuts(DC,:) = cutp(DCpl);
      end
```

```
        if isempty(DCpr)==0
          DC=DC+1;
          DoubleCuts(DC,:) = cutp(DCpr);
        end
      end
    end
    %% Results
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    dsbFreq = numel(cell2mat(allcuts(:)))/sum(DSBn);
    out = histc(cell2mat(allcuts(:)),1:length(invmodel));
    model = model/sum(model);
    out = out/sum(out);
    outmap{v} = out;
    modelchk = find(model>0);
    outchk = find(out>0);
    chk = setdiff(modelchk,outchk);
    model(chk) = 0;
    model(model==0) = [];
    out(out==0) = [];
    outfiltered{v} = out;
    ratio = out./model;
    rawratio{v} = ratio;
    intwindows{v} = dist;
end
%% Output
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if strcmp(Save,'Y') == 1
    folder = 'RatioTrials/';
    if strcmp(IntMode,'Exp') == 1
        filename =
strcat(folder,'Cis_Chr',num2str(ChrModel),IntMode,num2str(min(Slope)),'-',num2str(max(Slope)));
    elseif strcmp(IntMode,'Tukey') == 1
        filename =
strcat(folder,'Cis_Chr',num2str(ChrModel),IntMode,'-',num2str(Top(1)*100),'-',num2str(min(wInt)),'-',num2str
(max(wInt)));
    elseif strcmp(IntMode,'Random') == 1
        filename = strcat(folder,'Cis_Chr',num2str(ChrModel),IntMode);
    elseif strcmp(IntMode,'Volcano') == 1
        filename =
strcat(folder,'Cis_Chr',num2str(ChrModel),IntMode,num2str(min(Slope)),'-',num2str(max(Slope)),'-',num2str(
FlatTop(1)));
    end
    save(filename)
end
DSBCount = sum(DSBn);
disp(DSBmean)
a = sum(count)/Samples;
disp(a)
```

Script—MapInversion.m

```
ChrModel = 11;    %Set model chromosome (1-16)
Actv = 30;       %Proportion (in %) of loops to activate/cell (rounded)
load('WTMapsNormHpMChr.mat');  %Loads resampled + smoothed SK1 chromosomes
fid1 = fopen('Rec8-S288c.txt','r');
data = textscan(fid1, '%f %f %*[^\r\n]');
genome = sum(cellfun('length',Tel1Maps));
data = cell2mat(data(:,1:2));
```

```
index = find(data(:,1)==ChrModel);
rec8 = sort(data(index,:));
rec8(:,2) = round(rec8(:,2)/100);
nloops = length(rec8)-1;
actloop = round(Actv/100*nloops);
for i=1:nloops
  regions(i,1) = rec8(i,2);
  regions(i,2) = (rec8(i+1,2)-1);
end
lsize = diff(rec8(:,2));
model = Tel1Maps{ChrModel};
model(1:regions(1,1)) = 0;
model(regions(end,2):end) = 0;
plat = zeros(1,length(model)*3);
plat(length(model):length(model)*2-1) = model;
total = sum(model(min(regions(:,1)):max(regions(:,2))));
for j=1:nloops
  weight(j) = sum(model(regions(j,1):regions(j,2)))/total;
end
if Actv==100
  weight(weight==0) = 0.00001;
end
res = zeros(500000,actloop);
finalprob = zeros(length(weight),1);
for k=1:500000
  res(k,:) = datasample(1:length(weight),actloop,'replace',false,'weights',weight);
end
for v=1:length(weight)
  finalprob(v) = sum(sum(res==v))/500000;
end
invmodel = model;
for h=1:nloops
  win = gausswin(lsize(h));
  invboost = normalize_var(win,1,100000*finalprob(h));
  invmodel(regions(h,1):regions(h,2)) = invmodel(regions(h,1):regions(h,2))./invboost;
end
invmodel(isinf(invmodel)) = 0;
invmodel(isnan(invmodel)) = 0;
filename = strcat('Chr',num2str(ChrModel),'Inversion',num2str(Actv),'.mat');
```

CHAPTER 5

Discussion

5.1—Summary

Work presented throughout this thesis sought to further characterise and elucidate the mechanisms underpinning the spatial regulation of meiosis within *S. cerevisiae*. Collectively, such regulation has been probed at multiple levels across two key stages of meiosis—DSB and CO formation—using a variety of computational and mathematical approaches. Notably, a novel simulation platform (*RecombineSim*) was developed and utilised to dissect how the DNA damage response (DDR) and DNA repair factors Rad24, Mec1^{ATR}, Tel1^{ATM} and Msh2 influence the subclass identity (class I, class II) and/or positioning of crossovers (COs) on a genome-wide basis **(Chapter 2)**. Moreover, alongside creation of a software package for the analysis of genome-wide Spo11 DSBs, generalised and hyperlocal features of Spo11 DSB distribution were investigated within WT and Tel1-deficient strains **(Chapter 3)**. Lastly, a secondary simulation platform (*DSBSim*) was developed and utilised to investigate the mechanics of Tel1-dependent DSB interference and examine how chromatin architecture may shape the localised formation of DSBs **(Chapter 4)**.

5.2—Modelling, analysing and interpreting the distribution of meiotic COs

As evidenced throughout much of this thesis, CO distributions may be influenced and altered in a multitude of ways (see: Chapter 2, 4). Namely: (i) a loss of the interfering signal or its propagation (*rad24*Δ) (see: Figure 2.22B) (ii) a change in the class I:class II balance (*msh2*Δ) (see: Figure 2.22A) or (iii) interplay between the distribution or frequency of precursor DSBs and the downstream distribution of COs (*mec1MN* and *tel1*Δ) (see: Figure 2.20, 2.21, 4.16). As emphasised in Chapter 2, it is not always immediately obvious what has happened *in vivo* when approaching analysis using unsuitable tools and findings may be easily misinterpreted. For example, despite considerable differences in their proposed mechanisms, CO interference, when assessed globally via a single (γ) distribution (see: Section 2.13), is weakened within Rad24, Mec1 and Tel1. Interpreting the distribution of recombination in any given mutant must therefore be done carefully—considering a wide range of possible mechanisms, including those outlined above. Importantly, (γ) mixture modelling (*GEM*) or simulation of NCO and CO distribution via *RecombineSim*, only requires a list of

accurate event positions (see: Section B2.2)—easily obtainable from a wide variety of mapping protocols. *RecombineSim* or *GEM* may therefore be readily applied to pre-existing datasets (see: Section 2.15) or incorporated into future studies to aid in the interpretation of data. Msh2, a mismatch repair (MMR) component proposed to act as an anti class I CO factor at sites of higher sequence divergence (see: Sections 2.16-2.18), may further complicate the analysis of recombination by obscuring the distributional phenotypes of other factors or misleading interpretation of the results. The extent to which parental strains diverge and the inclusion of Msh2 mutation will thus be important considerations going forward for all mapping studies.

5.3—Tel1^{ATM}—A master controller of spatial regulation

Tel1^{ATM}, a DDR kinase, is rapidly emerging as a key and central factor in the spatial regulation of meiotic recombination. Remarkably, Tel1^{ATM} acts at both a hyperlocal and long range scale to shape the distributions of DSBs. At short range, Tel1 appears to suppress intragenic DSB formation by spatially confining *intra*-hotspot Spo11 double cuts—an otherwise seemingly unavoidable consequence of DSB formation in WT cells—from spreading beyond a distance of ~75bp (see: Sections 3.13-3.15). At mid range distances, across singular 10-15kb loop domains, Tel1 inhibits *inter*-hotspot Spo11 double cuts which otherwise arise concertedly (negative interference) (Garcia et al. 2015) as a potential consequence of loop tethering (see: Section 4.6). At long range distances (~50-150kb), Tel1 mediates the process of DSB interference both in *cis* (along a chromatid) and possibly also in *trans* (between chromatids) over a fixed spatial distance that may correspond to a set number of structural units (see: Sections 4.8-4.9). At the level of COs, Tel1 appears to suppress class II CO formation (see: Section 2.20)—corroborating previous observations (Anderson et al. 2015). Tel1 therefore spatially guides recombination at both key phases in meiotic recombination (DSB and CO formation). Ultimately, removal of Tel1 activity peels back the regulatory mask—revealing the inherently chaotic and stochastic process that unregulated DSB formation is—and reinforces the need for such regulation in the first place.

Mechanisms of Tel1-dependent regulation, as uncovered or further elucidated throughout each chapter, may also have important implications for the study of meiosis in other organisms. Notably, ATM mediates the negative regulation of DSB formation within *M. musculus* and *D. melanogaster* (see: Section 1.3.4) (Lange et al. 2011; Joyce et al. 2011). Moreover, $ATM^{-/-}$ null mice display severe meiotic defects, that ultimately result in infertility (Barlow et al. 1996; Xu et al. 1996; Barlow et al. 1998)—a phenotype ascribed to the excessive formation of DSBs (Lange et al. 2011)—while Ataxia telangiectasia (AT) patients, who contain a mutated form of ATM, also display infertility (Boder 1975). Understanding how $Tel1^{ATM}$ sculpts the meiotic landscape is thus of great importance. However, specific phenomena, such as DSB interference or Spo11 double cutting, are either poorly characterised or as of yet unobserved within higher organisms. Work presented in this thesis may therefore serve as a framework to guide future investigations. Moreover, exactly how—in terms of target factors—Tel1 mediates and regulates the spatial distribution of DSBs and/or COs within *S. cerevisiae* is yet to be determined. Given the intimate involvement of Tel1 within the spatial regulation of recombination, finding the elusive pathway(s) will be a crucial endeavour to undertake.

5.4—Evolutionary pressures of sequence divergence

As previously noted, Msh2 appears to exhibit considerable anti class I CO activity—disproportionately suppressing the formation of interfering, Mlh1-Mlh3 and ZMM-dependent COs at sites of higher sequence divergence (see: Sections 2.16-2.18). Notably, inactivation of Msh2 (*msh2*Δ) is predicted to result in ~38-39 additional class I COs relative to WT, with no appreciable increase in class II formation (see: Figure 2.19E,F). Specificity for class I COs, as observed in this thesis, expands upon previous findings that sequence heterozygosity indiscriminately suppresses CO formation during meiosis. Importantly, such a mechanism may prove widely conserved. SNP/INDEL density has been directly observed to inhibit CO formation within a wide range of organisms, including *S. cerevisiae* (Borts & Haber 1987), *M. musculus* (Baudat & de Massy 2007; Cole et al. 2010), *H. sapiens* (Jeffreys & Neumann 2005), *A. thaliana* (Ziolkowski et al. 2015) and *Z. mays*

(Dooner 1986). However, no distinction between CO subclasses has yet been fully characterised and/or observed in any organism, bar *S. cerevisiae* (see: Chapter 2).

Class I COs are the predominant subclass within WT meioses (~65-70% within *S. cerevisiae*) and may be specifically required for correct meiosis I disjunction (Getz et al. 2008)—ensuring the success of meiosis. Therefore, if conserved in higher organisms, a process of Msh2-dependent and divergence-based inhibition of class I CO formation may have far-reaching implications. Specifically, Msh2 may help determine the compatibility of mates within and between sexually reproducing populations, thereby shaping the evolutionary landscape and modulating the process of speciation. For example, class I CO frequency could conceivably fall below a critical, required level within offspring of distantly related individuals or strains—jeopardising normal meiotic completion, rendering the offspring infertile and contributing to the sexual isolation of the organism.

Interestingly, MMR incompatibilities—defined by an inactive or hypomorphic mismatch correction system—have been observed within certain diverse crosses of *S. cerevisiae* caused by variant forms of MMR machinery (Demogines et al. 2008; Bui et al. 2017). As heteroduplex rejection appears to rely upon the full MMR pathway (Spies & Fishel 2015), as opposed to Msh2 alone, suppression of class I COs may be highly cross-specific. For example, within hybrid backgrounds lacking a functional MMR pathway, class I CO formation would, presumably, occur unhindered.

Regardless of how class I CO suppression is mechanistically achieved, Msh2 activity is likely to constrain the level of genetic diversity achievable per meiosis by limiting the amount of genetic exchange (crossing over) that may occur between divergent homologues—in turn influencing the rates of evolution within a population and favouring more gradual changes to the gene pool—and result in a less evenly spaced array of CO events within divergent cells.

5.5—Future work

Mathematical and computational modelling of meiotic recombination within *S. cerevisiae* DDR and DNA repair mutants has yielded a plethora of testable hypotheses and much is left to understand about processes of spatial regulation in general. Furthermore, there is opportunity to expand the analytical methods, developed throughout each chapter, to different biological systems. Below, a number of future questions and directions have been discussed.

Expanding genome–wide mapping of Spo11 DSBs

As evidenced by the capture of Topo II lesions (see: Sections 3.18-3.21), the protocol devised for the nucleotide resolution mapping of Spo11 DSBs—and by extension *Spo11Mapper*—may be expanded to any biological pathway that involves a 5' covalently linked DNA:protein intermediate and where a method to stabilise this intermediate, either genetically (e.g. *sae2*Δ) or chemically (e.g. etoposide) and remove the attached moiety (e.g. TDP2), exists. For example, mapping of Rec12^{Spo11}-oligos, released by Mre11 and Ctp1^{Sae2}, within *S. pombe* is subject to similar caveats to those observed within *S. cerevisiae*—that is, loss of short <15bp oligonucleotides and a need for affinity tagging (Fowler et al. 2014). Generation of a *ctp1*Δ mutant would however allow for full, genome-wide mapping of Spo11 DSBs, rapidly processed and analysed by *Spo11Mapper*. Additionally, analyses conducted by *Spo11Mapper* can reveal novel phenotypes (e.g. periodic, double cutting) that may be otherwise missed—as was the case within WT Spo11-oligo libraries.

Mechanics of class I CO suppression

Elucidating how Msh2 specifically targets class I COs and mediates event redirection toward alternative pathways (i.e. class II COs or NCOs) will be a fundamental question to address in future studies as will determining whether or not this inhibitory mechanism is present in other organisms. Importantly, using the tools and investigative frameworks developed in this thesis (e.g. *RecombineSim*, *GEM*), high resolution, genome-wide data of CO position from other organisms may now be analysed in a similar fashion in order to search for comparable findings.

Within *S. cerevisiae*, the observed *msh2*Δ-dependent phenotypes may become more or less exacerbated proportional to the level of sequence divergence between a given cross—an hypothesis testable via the use of different hybrids. While SK1 exhibits a divergence rate of ~0.7% relative to S288c (64,591 SNPs, 3973 INDELs) (see: Section 2.3), other commonly employed strains are more S288c-like. For example, W303, previously utilised to study meiosis (Primig et al. 2000; Moreau et al. 1999), exhibits a reduced divergence rate of ~0.08% relative to S288c (Schacherer et al. 2007). Suppression of class I COs, by Msh2, may therefore be largely absent within W303 x S288c crosses if the proposed model for Msh2-dependent inhibition is correct. However, changes in culture synchronicity, hybrid viability and MMR compatibility will have to be tightly controlled before comparisons can be drawn. Interestingly, MMR machinery within *S. cerevisiae* exhibits differential ability to detect certain mismatches (Lichten et al. 1990). Specifically, C-C mismatches, relative to G-G mismatches, often remain undetected and unrepaired (Lichten et al. 1990). Subgrouping of S288c x SK1 SNP/INDEL lists and reanalysis of polymorphism density surrounding meiotic COs on a per group basis may therefore be warranted and may further assess whether or not the full MMR pathway is required for CO suppression in *S. cerevisiae*.

Modelling recombination: *Where to next*?

RecombineSim constitutes a reductionist model, designed to minimise the number of involved parameters and reduce NCO or CO distributions to their fundamental constituents. Nevertheless, the inclusion of an additional step—namely DSB formation—proved effective in the analysis of *tel1*Δ*msh2*Δ data (see: Section 4.11) and provided a means to assess how changes in the underlying DSB distribution influence downstream processes. A more comprehensive unified model, for example including site selection weighted by DSB hotspot strength, may therefore be warranted—removing the need for two separate platforms (e.g. *RecombineSim*, *DSBSim*).

The exact level of sequence divergence at which class I CO formation becomes untenable remains unclear and may vary between crosses and species, thus requiring further examination. Within, *S.*

cerevisiae, a single mismatch within a ~300bp region of recombination (0.3% divergence) was sufficient to reduce CO frequencies 3-fold (Datta et al. 1997). Moreover, CO frequencies logarithmically decreased up until ~1% divergence (Datta et al. 1997). Importantly, the way in which SNP/INDEL density regulates recombination outcome may be further explored via *RecombineSim*, at an expanded, genome-wide level. For example, class II CO frequency (1-100%) is currently determined by the model parameter, C_{PROB} and class II COs may indiscriminately occur at any site. *RecombineSim* may, however, be adapted to calculate the probability of class II formation in accordance to the local SNP/INDEL density at any given site selected. By varying the density threshold at which class II COs become favoured, it may be possible to elucidate the SNP/INDEL density range over which suppression begins to occur, on average, *in vivo* across the genome. Moreover, differential levels of suppression may occur in response to specific patterns of SNP/INDELs. Using (γ) mixture modelling results, it may be possible to assign a statistical likelihood to each detected CO, describing its class II-likeness—therefore narrowing down the number of loci to be analysed at high resolution and aiding in discovery of any such pattern.

A number of predictions made by *RecombineSim* may also be examined experimentally. For example, within Mec1 or Tel1 mutants, excess DSBs are predicted to preferentially enter the class II CO and NCO pathways, while class I frequency remains static—potentially held at an upper, homeostatic limit imposed by CO interference (see: Figure 2.4, Figure 2.19E,F). Inactivation of class II CO formation, via *mus81*Δ or *mms4*Δ, within *tel1*Δ or *mec1MN* mutants should therefore result in further elevations in NCO formation without change in the class I frequency—a phenotype readily detectable by genome-wide mapping of recombination. Alternatively, the introduction of Spo11 hypomorphs, such as s*po11–HA* (Gray et al. 2013; Argunhan et al. 2013), into *tel1*Δ or *mec1MN* backgrounds may rescue the observed changes to CO distribution, increase the global strength of CO interference and help validate this proposed mechanism by reducing overall DSB frequencies.

References

Acquaviva, L. et al., 2013. The COMPASS subunit Spp1 links histone methylation to initiation of meiotic recombination. Science (New York, N.Y.), 339(6116), pp.215–8.

Agarwal, S. & Roeder, G.S., 2000. Zip3 provides a link between recombination enzymes and synaptonemal complex proteins. Cell, 102(2), pp.245–55.

Alani, E. et al., 1997. Saccharomyces cerevisiae MSH2, a mispaired base recognition protein, also recognizes Holliday junctions in DNA. Journal of Molecular Biology, 265(3), pp.289–301.

Allers, T. & Lichten, M., 2001. Differential timing and control of noncrossover and crossover recombination during meiosis. Cell, 106(1), pp.47–57.

Anderson, C.M. et al., 2015. Reduced Crossover Interference and Increased ZMM-Independent Recombination in the Absence of Tel1/ATM M. Lichten, ed. PLOS Genetics, 11(8), p.e1005478.

Aparicio, T. et al., 2016. MRN, CtIP, and BRCA1 mediate repair of topoisomerase II–DNA adducts. The Journal of Cell Biology, 212(4).

Aplan, P.D., 2006. Causes of oncogenic chromosomal translocation. Trends in genetics : TIG, 22(1), pp.46–55.

Argueso, J.L. et al., 2004. Competing Crossover Pathways Act During Meiosis in Saccharomyces cerevisiae. Genetics, 168(4), pp.1805–1816.

Argunhan, B. et al., 2013. Direct and indirect control of the initiation of meiotic recombination by DNA damage checkpoint mechanisms in budding yeast. PloS one, 8(6), p.e65875.

Arora, C. et al., 2004. Antiviral protein Ski8 is a direct partner of Spo11 in meiotic DNA break formation, independent of its cytoplasmic role in RNA metabolism. Molecular cell, 13(4), pp. 549–59.

Auton, A. et al., 2013. Genetic recombination is targeted towards gene promoter regions in dogs. PLoS genetics, 9(12), p.e1003984.

Axel, R., 1975. Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. Biochemistry, 14(13), pp.2921–5.

Axelsson, E. et al., 2012. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. Genome research, 22(1), pp.51–63.

Bakkenist, C.J. & Kastan, M.B., 2003. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. Nature, 421(6922), pp.499–506.

Baranello, L. et al., 2014. DNA Break Mapping Reveals Topoisomerase II Activity Genome-Wide. International Journal of Molecular Sciences, 15(7), pp.13111–13122.

Barchi, M. et al., 2008. ATM promotes the obligate XY crossover and both crossover control and chromosome axis integrity on autosomes. PLoS genetics, 4(5), p.e1000076.

Barlow, A.L. & Hultén, M.A., 1998. Crossing over analysis at pachytene in man. European Journal of Human Genetics, 6(4), pp.350–358.

Barlow, C. et al., 1996. Atm-deficient mice: a paradigm of ataxia telangiectasia. Cell, 86(1), pp.159–71.

Barlow, C. et al., 1998. Atm deficiency results in severe meiotic disruption as early as leptonema of prophase I. Development (Cambridge, England), 125(20), pp.4007–17.

Barnes, T.M. et al., 1995. Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. Genetics, 141(1), pp.159–79.

Barrington, C., Finn, R. & Hadjur, S., 2017. Cohesin biology meets the loop extrusion model. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology, 25(1), pp.51–60.

Bartek, J. & Lukas, J., 2007. DNA damage checkpoints: from initiation to recovery or adaptation. Current Opinion in Cell Biology, 19(2), pp.238–245.

Batten, L.M. & Beutelspacher, A., 1993. The theory of finite linear spaces : combinatorics of points and lines, Cambridge University Press.

Baudat, F. et al., 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science (New York, N.Y.), 327(5967), pp.836–40.

Baudat, F. & de Massy, B., 2007. Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot. PLoS genetics, 3(6), p.e100.

Baudat, F. & Nicolas, A., 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. Proceedings of the National Academy of Sciences, 94(10), pp.5213–5218.

Berchowitz, L.E. et al., 2007. The role of AtMUS81 in interference-insensitive crossovers in A. thaliana. PLoS genetics, 3(8), p.e132.

Berchowitz, L.E. & Copenhaver, G.P., 2010. Genetic interference: don't stand so close to me. Current genomics, 11(2), pp.91–102.

Berg, I.L. et al., 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature genetics, 42(10), pp.859–63.

Bergerat, A. et al., 1997. An atypical topoisomerase II from archaea with implications for meiotic recombination. Nature, 386(6623), pp.414–417.

Bewick, V., Cheek, L. & Ball, J., 2004. Statistics review 12: survival analysis. Critical care (London, England), 8(5), pp.389–94.

Bhalla, N. et al., 2008. ZHP-3 acts at crossovers to couple meiotic recombination with synaptonemal complex disassembly and bivalent formation in C. elegans. PLoS genetics, 4(10), p.e1000235.

Bishop, D.K. et al., 1992. DMC1: a meiosis-specific yeast homolog of E. coli recA required for recombination, synaptonemal complex formation, and cell cycle progression. Cell, 69(3), pp. 439–56.

Bishop, D.K. & Zickler, D., 2004. Early decision; meiotic crossover interference prior to stable strand exchange and synapsis. Cell, 117(1), pp.9–15.

Blat, Y. et al., 2002. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. Cell, 111(6), pp.791–802.

Blitzblau, H.G. et al., 2007. Mapping of Meiotic Single-Stranded DNA Reveals Double-Strand-Break Hotspots near Centromeres and Telomeres. Current Biology, 17(23), pp.2003–2012.

Blitzblau, H.G. & Hochwagen, A., 2013. ATR/Mec1 prevents lethal meiotic recombination initiation on partially replicated chromosomes in budding yeast. eLife, 2, p.e00844.

Blömer, J. & Bujna, K., 2013. Adaptive Seeding for Gaussian Mixture Models.

Boddy, M.N. et al., 2001. Mus81-Eme1 are essential components of a Holliday junction resolvase. Cell, 107(4), pp.537–48.

Boder, E., 1975. Ataxia-telangiectasia: some historic, clinical and pathologic observations. Birth defects original article series, 11(1), pp.255–70.

de Boer, E. et al., 2006. Two levels of interference in mouse meiotic recombination. Proceedings of the National Academy of Sciences, 103(25), pp.9607–9612.

Borde, V. et al., 2009. Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. The EMBO journal, 28(2), pp.99–111.

Borde, V., Goldman, A.S. & Lichten, M., 2000. Direct coupling between meiotic DNA replication and recombination initiation. Science (New York, N.Y.), 290(5492), pp.806–9.

Borde, V. & de Massy, B., 2013. Programmed induction of DNA double strand breaks during meiosis: setting up communication between DNA and the chromosome structure. Current opinion in genetics & development, 23(2), pp.147–55.

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(1), pp.29–45.

Borts, R.H. & Haber, J.E., 1987. Meiotic recombination in yeast: alteration by multiple heterozygosities. Science (New York, N.Y.), 237(4821), pp.1459–65.

Brick, K. et al., 2012. Genetic recombination is directed away from functional genomic elements in mice. Nature, 485(7400), pp.642–5.

Brogaard, K. et al., 2012. A map of nucleosome positions in yeast at base-pair resolution. Nature, 486(7404), pp.496–501.

Broman, K.W. et al., 2002. Crossover interference in the mouse. Genetics, 160(3), pp.1123–31.

Bromberg, K.D., Burgin, A.B. & Osheroff, N., 2003. A Two-drug Model for Etoposide Action against Human Topoisomerase IIα. Journal of Biological Chemistry, 278(9), pp.7406–7412.

Buard, J. et al., 2009. Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. The EMBO journal, 28(17), pp.2616–24.

Buard, J. et al., 2014. Diversity of Prdm9 zinc finger array in wild mice unravels new facets of the evolutionary turnover of this coding minisatellite. PloS one, 9(1), p.e85021.

Buhler, C., Borde, V. & Lichten, M., 2007. Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae. J. E. Haber, ed. PLoS biology, 5(12), p.e324.

Bui, D.T. et al., 2017. Mismatch Repair Incompatibilities in Diverse Yeast Populations. Genetics, 205(4), pp.1459–1471.

Burden, D.A. & Osheroff, N., 1999. In vitro evolution of preferred topoisomerase II DNA cleavage sites. The Journal of biological chemistry, 274(8), pp.5227–35.

Burden, D.A. & Osheroff, N., 1998. Mechanism of action of eukaryotic topoisomerase II and drugs targeted to the enzyme. Biochimica et biophysica acta, 1400(1–3), pp.139–54.

Bzymek, M. et al., 2010. Double Holliday junctions are intermediates of DNA break repair. Nature, 464(7290), pp.937–41.

Callender, T.L. & Hollingsworth, N.M., 2010. Mek1 suppression of meiotic double-strand break repair is specific to sister chromatids, chromosome autonomous and independent of Rec8 cohesin complexes. Genetics, 185(3), pp.771–82.

Carballo, J.A. et al., 2013. Budding yeast ATM/ATR control meiotic double-strand break (DSB) levels by down-regulating Rec114, an essential component of the DSB-machinery. PLoS genetics, 9(6), p.e1003545.

Carballo, J.A. et al., 2008. Phosphorylation of the axial element protein Hop1 by Mec1/Tel1 ensures meiotic interhomolog recombination. Cell, 132(5), pp.758–70.

Carpenter, A.T., 1975. Electron microscopy of meiosis in Drosophila melanogaster females: II. The recombination nodule--a recombination-associated structure at pachytene? Proceedings of the National Academy of Sciences of the United States of America, 72(8), pp.3186–9.

Cejka, P. et al., 2010. Rmi1 stimulates decatenation of double Holliday junctions during dissolution by Sgs1–Top3. Nature Structural & Molecular Biology, 17(11), pp.1377–1382.

Champeimont, R. & Carbone, A., 2014. SPoRE: a mathematical model to predict double strand breaks and axis protein sites in meiosis. BMC bioinformatics, 15(1), p.391.

Chang, H.H.Y. et al., 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nature Reviews Molecular Cell Biology.

Chapman, J. ?Ros. et al., 2012. Playing the End Game: DNA Double-Strand Break Repair Pathway Choice. Molecular Cell, 47(4), pp.497–510.

Chen, L. et al., 2001. Promotion of Dnl4-catalyzed DNA end-joining by the Rad50/Mre11/Xrs2 and Hdf1/Hdf2 complexes. Molecular cell, 8(5), pp.1105–15.

Chen, S. -h. et al., 2010. A Proteome-wide Analysis of Kinase-Substrate Network in the DNA Damage Response. Journal of Biological Chemistry, 285(17), pp.12803–12812.

Chen, S.Y. et al., 2008. Global Analysis of the Meiotic Crossover Landscape. Developmental Cell, 15(3), pp.401–415.

Cheng, Y.-H. et al., 2013. Three distinct modes of Mec1/ATR and Tel1/ATM activation illustrate differential checkpoint targeting during budding yeast early meiosis. Molecular and cellular biology, 33(16), pp.3365–76.

Choi, K. et al., 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nature genetics, 45(11), pp.1327–36.

Choi, K. & Henderson, I.R., 2015. Meiotic Recombination Hotspots - a Comparative View. The Plant journal : for cell and molecular biology, 83(1), pp.52–61.

Ciccia, A. & Elledge, S.J., 2010. The DNA Damage Response: Making It Safe to Play with Knives. Molecular Cell, 40(2), pp.179–204.

Cockell, M., Rhodes, D. & Klug, A., 1983. Location of the primary sites of micrococcal nuclease cleavage on the nucleosome core. Journal of molecular biology, 170(2), pp.423–46.

Cole, F. et al., 2012. Homeostatic control of recombination is implemented progressively in mouse meiosis. Nature Cell Biology, 14(4), pp.424–430.

Cole, F., Keeney, S. & Jasin, M., 2010. Comprehensive, Fine-Scale Dissection of Homologous Recombination Outcomes at a Hot Spot in Mouse Meiosis. Molecular Cell, 39(5), pp.700–710.

Comeron, J.M., Ratnappan, R. & Bailin, S., 2012. The many landscapes of recombination in Drosophila melanogaster. PLoS genetics, 8(10), p.e1002905.

Cooper, T.J. et al., 2014. Homeostatic regulation of meiotic DSB formation by ATM/ATR. Experimental cell research, 329(1), pp.124–131.

Cooper, T.J., Garcia, V. & Neale, M.J., 2016. Meiotic DSB patterning: A multifaceted process. Cell Cycle, (1), pp.13–21.

Copenhaver, G.P., Housworth, E.A. & Stahl, F.W., 2002. Crossover interference in Arabidopsis. Genetics, 160(4), pp.1631–9.

Cromie, G. a et al., 2006. Single Holliday junctions are intermediates of meiotic recombination. Cell, 127(6), pp.1167–78.

Cruz-García, A., López-Saavedra, A. & Huertas, P., 2014. BRCA1 Accelerates CtIP-Mediated DNA-End Resection. Cell Reports, 9(2), pp.451–459.

Datta, A. et al., 1997. Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proceedings of the National Academy of Sciences of the United States of America, 94(18), pp.9757–62.

Demogines, A. et al., 2008. Incompatibilities Involving Yeast Mismatch Repair Genes: A Role for Genetic Modifiers and Implications for Disease Penetrance and Variation in Genomic Mutation Rates O. Cohen-Fix, ed. PLoS Genetics, 4(6), p.e1000103.

Destrempes, F. et al., 2011. Segmentation of Plaques in Sequences of Ultrasonic B-Mode Images of Carotid Arteries Based on Motion Estimation and a Bayesian Model. IEEE Transactions on Biomedical Engineering, 58(8), pp.2202–2211.

Do, C.B. & Batzoglou, S., 2008. What is the expectation maximization algorithm? Nature Biotechnology, 26(8), pp.897–899.

Dooner, H.K., 1986. Genetic Fine Structure of the BRONZE Locus in Maize. Genetics, 113(4), pp. 1021–36.

Drouaud, J. et al., 2013. Contrasted Patterns of Crossover and Non-crossover at Arabidopsis thaliana Meiotic Recombination Hotspots H. Ma, ed. PLoS Genetics, 9(11), p.e1003922.

Egel, R., 1978. Synaptonemal complex and crossing-over: structural support or interference? Heredity, 41(2), pp.233–37.

Fan, Q.Q. et al., 1997. Competition between adjacent meiotic recombination hotspots in the yeast Saccharomyces cerevisiae. Genetics, 145(3), pp.661–70.

Fan, Q.Q. & Petes, T.D., 1996. Relationship between nuclease-hypersensitive sites and meiotic recombination hot spot activity at the HIS4 locus of Saccharomyces cerevisiae. Molecular and cellular biology, 16(5), pp.2037–43.

Foss, E. et al., 1993. Chiasma interference as a function of genetic distance. Genetics, 133(3).

Foss, E.J. & Stahl, F.W., 1995. A test of a counting model for chiasma interference. Genetics, 139(3), pp.1201–9.

Fowler, K.R. et al., 2014. Evolutionarily diverse determinants of meiotic DNA break and recombination landscapes across the genome. Genome research, 24(10), pp.1650–64.

Franco, S., Alt, F.W. & Manis, J.P., 2006. Pathways that suppress programmed DNA breaks from progressing to chromosomal breaks and translocations. DNA Repair, 5(9–10), pp.1030–1041.

Fukuda, T. et al., 2008. Targeted induction of meiotic double-strand breaks reveals chromosomal domain-dependent regulation of Spo11 and interactions among potential sites of meiotic recombination. Nucleic acids research, 36(3), pp.984–97.

Fung, J.C. et al., 2004. Imposition of crossover interference through the nonrandom distribution of synapsis initiation complexes. Cell, 116(6), pp.795–802.

Gao, R. et al., 2014. Proteolytic Degradation of Topoisomerase II (Top2) Enables the Processing of Top2·DNA and Top2·RNA Covalent Complexes by Tyrosyl-DNA-Phosphodiesterase 2 (TDP2). Journal of Biological Chemistry, 289(26), pp.17960–17969.

Garcia, V. et al., 2011. Bidirectional resection of DNA double-strand breaks by Mre11 and Exo1. Nature, 479(7372), pp.241–244.

Garcia, V. et al., 2015. Tel1(ATM)-mediated interference suppresses clustered meiotic double-strand-break formation. Nature, 520(7545), pp.114–8.

Gerton, J.L. et al., 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America, 97(21), pp.11383–90.

Getz, T.J. et al., 2008. Reduced mismatch repair of heteroduplexes reveals "non"-interfering crossing over in wild-type Saccharomyces cerevisiae. Genetics, 178(3), pp.1251–69.

Glynn, E.F. et al., 2004. Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae. PLoS biology, 2(9), p.E259.

Goffeau, A. et al., 1996. Life with 6000 genes. Science (New York, N.Y.), 274(5287), pp.546, 563–7.

Goldfarb, T. & Lichten, M., 2010. Frequent and efficient use of the sister chromatid for DNA double-strand break repair during budding yeast meiosis. R. S. Hawley, ed. PLoS biology, 8(10), p.e1000520.

Gómez-Herreros, F. et al., 2013. TDP2–Dependent Non-Homologous End-Joining Protects against Topoisomerase II–Induced DNA Breaks and Genome Instability in Cells and In Vivo N. Maizels, ed. PLoS Genetics, 9(3), p.e1003226.

Gray, S. et al., 2013. Positive regulation of meiotic DNA double-strand break formation by activation of the DNA damage checkpoint kinase Mec1(ATR). Open biology, 3(7), p.130019.

Gray, S. & Cohen, P.E., 2016. Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation. Annual Review of Genetics, 50(1), pp.175–210.

Grey, C. et al., 2011. Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. PLoS biology, 9(10), p.e1001176.

Haber, J.E., 2012. Mating-type genes and MAT switching in Saccharomyces cerevisiae. Genetics, 191(1), pp.33–64.

Haldane, 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. J Genet, 8, pp.299–309.

Hartsuiker, E., Neale, M.J. & Carr, A.M., 2009. Distinct Requirements for the Rad32Mre11 Nuclease and Ctp1CtIP in the Removal of Covalently Bound Topoisomerase I and II from DNA. Molecular Cell, 33(1), pp.117–123.

Hastings, P.J., 2010. Mechanisms of ectopic gene conversion. Genes, 1(3), pp.427–39.

Heyting, C. et al., 1999. Mre11 and Ku70 interact in somatic cells, but are differentially expressed in early meiosis. Nature Genetics, 23(2), pp.194–198.

Higgins, J.D. et al., 2008. Expression and functional analysis of AtMUS81 in Arabidopsis meiosis reveals a role in the second pathway of crossing-over. The Plant Journal, 54(1), pp.152–162.

Hillers, K.J. & Villeneuve, A.M., 2003. Chromosome-wide control of meiotic crossing over in C. elegans. Current biology : CB, 13(18), pp.1641–7.

Hollingsworth, N.M., Ponte, L. & Halsey, C., 1995. MSH5, a novel MutS homolog, facilitates meiotic reciprocal recombination between homologs in Saccharomyces cerevisiae but not mismatch repair. Genes & development, 9(14), pp.1728–39.

Holloway, J.K. et al., 2008. MUS81 Generates a Subset of MLH1-MLH3?Independent Crossovers in Mammalian Meiosis R. S. Hawley, ed. PLoS Genetics, 4(9), p.e1000186.

Hou, Y. et al., 2013. Genome Analyses of Single Human Oocytes. Cell, 155(7), pp.1492–1506.

Housworth, E.A. & Stahl, F.W., 2003. Crossover interference in humans. American journal of human genetics, 73(1), pp.188–97.

Hunt Morgan, T., 1916. A critique of the theory of evolution. The Eugenics Review, 10(4), p.231.

Ito, M. et al., 2014. Meiotic recombination cold spots in chromosomal cohesion sites. Genes to cells : devoted to molecular & cellular mechanisms, 19(5), pp.359–73.

Jantsch, V. et al., 2004. Targeted Gene Knockout Reveals a Role in Meiotic Recombination for ZHP-3, a Zip3-Related Protein in Caenorhabditis elegans. Molecular and Cellular Biology, 24(18), pp. 7998–8006.

Jeffreys, A.J. & Neumann, R., 2005. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Human Molecular Genetics, 14(15), pp.2277–2287.

Jeggo, P.A., Pearl, L.H. & Carr, A.M., 2015. DNA repair, genome stability and cancer: a historical perspective. Nature Reviews Cancer, 16(1), pp.35–42.

Jessop, L. & Lichten, M., 2008. Mus81/Mms4 Endonuclease and Sgs1 Helicase Collaborate to Ensure Proper Recombination Intermediate Metabolism during Meiosis. Molecular Cell, 31(3), pp. 313–323.

Jones, G.H. & Franklin, F.C.H., 2006. Meiotic crossing-over: obligation and interference. Cell, 126(2), pp.246–8.

Joshi, N. et al., 2009. Pch2 Links Chromosome Axis Remodeling at Future Crossover Sites and Crossover Distribution during Yeast Meiosis G. P. Copenhaver, ed. PLoS Genetics, 5(7), p.e1000557.

Joyce, E.F. et al., 2011. Drosophila ATM and ATR have distinct activities in the regulation of meiotic DNA damage and repair. The Journal of cell biology, 195(3), pp.359–67.

Kadyk, L.C. & Hartwell, L.H., 1992. Sister chromatids are preferred over homologs as substrates for recombinational repair in Saccharomyces cerevisiae. Genetics, 132(2), pp.387–402.

Kan, F., Davidson, M.K. & Wahls, W.P., 2011. Meiotic recombination protein Rec12: functional conservation, crossover homeostasis and early crossover/non-crossover decision. Nucleic Acids Research, 39(4), pp.1460–1472.

Kaplan, N. et al., 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature, 458(7236), pp.362–6.

Kauppi, L. et al., 2011. Distinct properties of the XY pseudoautosomal region crucial for male meiosis. Science (New York, N.Y.), 331(6019), pp.916–20.

Kauppi, L. et al., 2013. Numerical constraints and feedback control of double-strand breaks in mouse meiosis. Genes & development, 27(8), pp.873–86.

Kaur, T. & Rockman, M. V, 2014. Crossover heterogeneity in the absence of hotspots in Caenorhabditis elegans. Genetics, 196(1), pp.137–48.

Kee, K. et al., 2004. Spatial organization and dynamics of the association of Rec102 and Rec104 with meiotic chromosomes. The EMBO journal, 23(8), pp.1815–24.

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(3), pp.375–84.

Keeney, S. & Kleckner, N., 1995. Covalent protein-DNA complexes at the 5' strand termini of meiosis-specific double-strand breaks in yeast. Proceedings of the National Academy of Sciences of the United States of America, 92(24), pp.11274–8.

Keeney, S., Lange, J. & Mohibullah, N., 2014. Self-organization of meiotic recombination initiation: general principles and molecular pathways. Annual review of genetics, 48, pp.187–214.

Keeney, S. & Neale, M.J., 2006. Initiation of meiotic recombination by formation of DNA double-strand breaks: mechanism and regulation. Biochemical Society transactions, 34(Pt 4), pp.523–5.

Khil, P.P. et al., 2012. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. Genome research, 22(5), pp.957–65.

Kim, K.P. et al., 2010. Sister cohesion and structural axis components mediate homolog bias of meiotic recombination. Cell, 143(6), pp.924–37.

Kim, S.T. et al., 1999. Substrate specificities and identification of putative substrates of ATM kinase family members. The Journal of biological chemistry, 274(53), pp.37538–43.

King, J.S. & Mortimer, R.K., 1990. A polymerization model of chiasma interference and corresponding computer simulation. Genetics, 126(4), pp.1127–38.

Kleckner, N. et al., 2004. A mechanical basis for chromosome function. Proceedings of the National Academy of Sciences of the United States of America, 101(34), pp.12592–7.

Kleckner, N., 2006. Chiasma formation: chromatin/axis interplay and the role(s) of the synaptonemal complex. Chromosoma, 115(3), pp.175–94.

Koehler, K.E. et al., 2002. Genetic control of mammalian meiotic recombination. I. Variation in exchange frequencies among males from inbred mouse strains. Genetics, 162(1), pp.297–306.

Kowalczykowski, S.C. et al., 1998. Rad52 protein stimulates DNA strand exchange by Rad51 and replication protein A. Nature, 391(6665), pp.407–410.

Krogh, B.O. & Symington, L.S., 2004. Recombination Proteins in Yeast. Annual Review of Genetics, 38(1), pp.233–271.

Kugou, K. et al., 2009. Rec8 guides canonical Spo11 distribution along yeast meiotic chromosomes. Molecular biology of the cell, 20(13), pp.3064–76.

Kumar, R., Bourbon, H.-M. & de Massy, B., 2010. Functional conservation of Mei4 for meiotic DNA double-strand break formation from yeasts to mice. Genes & development, 24(12), pp.1266–80.

Lam, I. & Keeney, S., 2015. Mechanism and regulation of meiotic recombination initiation. Cold Spring Harbor perspectives in biology, 7(1), p.a016634.

Lam, S.Y. et al., 2005. Crossover interference on nucleolus organizing region-bearing chromosomes in Arabidopsis. Genetics, 170(2), pp.807–12.

Lange, J. et al., 2011. ATM controls meiotic double-strand-break formation. Nature, 479(7372), pp. 237–40.

Lange, J. et al., 2016. The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. Cell, 167(3), p.695–708.e16.

Lao, J.P. & Hunter, N., 2010. Trying to avoid your sister. PLoS biology, 8(10), p.e1000519.

Ledesma, F.C. et al., 2009. A human 5′-tyrosyl DNA phosphodiesterase that repairs topoisomerase-mediated DNA damage. Nature, 461(7264), pp.674–678.

Lee, C.-S. et al., 2014. Dynamics of yeast histone H2A and H2B phosphorylation in response to a double-strand break. Nature structural & molecular biology, 21(1), pp.103–9.

Li, J., Hooker, G.W. & Roeder, G.S., 2006. Saccharomyces cerevisiae Mer2, Mei4 and Rec114 form a complex required for meiotic double-strand break formation. Genetics, 173(4), pp.1969–81.

Li, X., Li, L. & Yan, J., 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. Nature communications, 6, p.6648.

Lichten, M. et al., 1990. Detection of Heteroduplex DNA Molecules Among the Products of Saccharomyces cerevisiae Meiosis. Proceedings of the National Academy of Sciences of the United States of America, 87, pp.7653–7657.

Lichten, M., 2014. Tetrad, Random Spore, and Molecular Analysis of Meiotic Segregation and Recombination. In Methods in molecular biology (Clifton, N.J.). pp. 13–28.

Lisby, M. et al., 2004. Choreography of the DNA Damage Response. Cell, 118(6), pp.699–713.

Liu, J., Wu, T.C. & Lichten, M., 1995. The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate. The EMBO journal, 14(18), pp.4599–608.

de los Santos, T. et al., 2001. A role for MMS4 in the processing of recombination intermediates during meiosis in Saccharomyces cerevisiae. Genetics, 159(4), pp.1511–25.

de los Santos, T. 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(1), pp.81–94.

Lu, S. et al., 2012. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. Science (New York, N.Y.), 338(6114), pp.1627–30.

Lydall, D. et al., 1996. A meiotic recombination checkpoint controlled by mitotic checkpoint genes. Nature, 383(6603), pp.840–3.

Lynn, A., Soucek, R. & B?rner, G.V., 2007. ZMM proteins during meiosis: Crossover artists at work. Chromosome Research, 15(5), pp.591–605.

MacQueen, A.J. & Hochwagen, A., 2011. Checkpoint mechanisms: the puppet masters of meiotic prophase. Trends in cell biology, 21(7), pp.393–400.

Madigan, J.P., Chotkowski, H.L. & Glaser, R.L., 2002. DNA double-strand break-induced phosphorylation of Drosophila histone variant H2Av helps prevent radiation-induced apoptosis. Nucleic acids research, 30(17), pp.3698–705.

Maguire, M.P., 1988. Crossover site determination and interference. Journal of theoretical biology, 134(4), pp.565–70.

Majka, J. & Burgers, P.M.J., 2003. Yeast Rad17/Mec3/Ddc1: a sliding clamp for the DNA damage checkpoint. Proceedings of the National Academy of Sciences of the United States of America, 100(5), pp.2249–54.

Majka, J., Niedziela-Majka, A. & Burgers, P.M.J., 2006. The checkpoint clamp activates Mec1 kinase during initiation of the DNA damage checkpoint. Molecular cell, 24(6), pp.891–901.

Maleki, S. et al., 2007. Interactions between Mei4, Rec114, and other proteins required for meiotic DNA double-strand break formation in Saccharomyces cerevisiae. Chromosoma, 116(5), pp. 471–486.

Mancera, E. et al., 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature, 454(7203), pp.479–85.

Manhart, C.M. & Alani, E., 2016. Roles for mismatch repair family proteins in promoting meiotic crossing over. DNA Repair, 38, pp.84–93.

Manning, G.S., 2006. The persistence length of DNA is reached from the persistence length of its null isomer through an internal electrostatic stretching force. Biophysical journal, 91(10), pp. 3607–16.

Mantiero, D. et al., 2007. Dual role for Saccharomyces cerevisiae Tel1 in the checkpoint response to double-strand breaks. EMBO reports, 8(4), pp.380–7.

Mao-Draayer, Y. et al., 1996. Analysis of meiotic recombination pathways in the yeast Saccharomyces cerevisiae. Genetics, 144(1), pp.71–86.

Maréchal, A. & Zou, L., 2013. DNA damage sensing by the ATM and ATR kinases. Cold Spring Harbor perspectives in biology, 5(9).

Mari, P.-O. et al., 2006. Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. Proceedings of the National Academy of Sciences, 103(49), pp.18597–18602.

Marsischky, G.T. et al., 1999. 'Saccharomyces cerevisiae MSH2/6 complex interacts with Holliday junctions and facilitates their cleavage by phage resolution enzymes. The Journal of biological chemistry, 274(11), pp.7200–6.

Marsolier-Kergoat, M.-C. et al., 2017. Mechanistic view and genetic control of DNA recombination during meiosis. bioRxiv.

Marston, A.L., 2009. Meiosis: DDK is not just for replication. Current biology : CB, 19(2), pp.R74-6.

Martini, E. et al., 2006. Crossover Homeostasis in Yeast Meiosis. Cell, 126(2), pp.285–295.

Martini, E. et al., 2011. Genome-wide analysis of heteroduplex DNA in mismatch repair-deficient yeast cells reveals novel properties of meiotic recombination pathways. M. Lichten, ed. PLoS genetics, 7(9), p.e1002305.

Massey, F.J. & Jr., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association, 46(253), p.68.

de Massy, B., 2013. Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. Annual review of genetics, 47, pp.563–99.

Matsuoka, S. et al., 2007. ATM and ATR Substrate Analysis Reveals Extensive Protein Networks Responsive to DNA Damage. Science, 316(5828), pp.1160–1166.

McMahill, M.S. et al., 2007. Synthesis-Dependent Strand Annealing in Meiosis M. Lichten, ed. PLoS Biology, 5(11), p.e299.

McPeek, M.S. & Speed, T.P., 1995. Modeling interference in genetic recombination. Genetics, 139(2), pp.1031–44.

Mehta, A. & Haber, J.E., 2014. Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. Cold Spring Harbor Perspectives in Biology, 6(9), pp.a016428–a016428.

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(3), pp.1169–77.

Mieczkowski, P.A. et al., 2007. Loss of a histone deacetylase dramatically alters the genomic distribution of Spo11p-catalyzed DNA breaks in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America, 104(10), pp.3955–60.

Miller, L.H., 1956. Table of Percentage Points of Kolmogorov Statistics. Journal of the American Statistical Association, 51(273), p.111.

Mills, J.A. & Prasad, K., 1992. A comparison of model selection criteria. Econometric Reviews, 11(2), pp.201–234.

Mimitou, E.P. & Symington, L.S., 2009. DNA end resection: Many nucleases make light work. DNA Repair, 8(9), pp.983–995.

Mirkin, E. V & Mirkin, S.M., 2007. Replication fork stalling at natural impediments. Microbiology and molecular biology reviews : MMBR, 71(1), pp.13–35.

Mohibullah, N. & Keeney, S., 2017. Numerical and spatial patterning of yeast meiotic DNA breaks by Tel1. Genome Research, 27(2), pp.278–288.

Moreau, S., Ferguson, J.R. & Symington, L.S., 1999. The nuclease activity of Mre11 is required for meiosis but not for mating type switching, end joining, or telomere maintenance. Molecular and cellular biology, 19(1), pp.556–66.

Muller, H.J., 1916. The Mechanism of Crossing-Over. The American Naturalist, 50(592), pp.193–221.

Muñoz-Fuentes, V., Di Rienzo, A. & Vilà, C., 2011. Prdm9, a major determinant of meiotic recombination hotspots, is not functional in dogs and their wild relatives, wolves and coyotes. PloS one, 6(11), p.e25498.

Munz, P., 1994. An analysis of interference in the fission yeast Schizosaccharomyces pombe. Genetics, 137(3), pp.701–7.

Murakami, H. et al., 2003. Correlation between premeiotic DNA replication and chromatin transition at yeast recombination initiation sites. Nucleic acids research, 31(14), pp.4085–90.

Murakami, H. & Keeney, S., 2014. Temporospatial Coordination of Meiotic DNA Replication and Recombination via DDK Recruitment to Replisomes. Cell, 158(4), pp.861–873.

Murakami, H. & Nicolas, A., 2009. Locally, meiotic double-strand breaks targeted by Gal4BD-Spo11 occur at discrete sites with a sequence preference. Molecular and cellular biology, 29(13), pp. 3500–16.

Nakada, D., Matsumoto, K. & Sugimoto, K., 2003. ATM-related Tel1 associates with double-strand breaks through an Xrs2-dependent mechanism. Genes & development, 17(16), pp.1957–62.

Nakamura, K. et al., 2010. Collaborative Action of Brca1 and CtIP in Elimination of Covalent Modifications from Double-Strand Breaks to Facilitate Subsequent Break Repair J. E. Haber, ed. PLoS Genetics, 6(1), p.e1000828.

Neale, M.J., 2010. PRDM9 points the zinc finger at meiotic recombination hotspots. Genome biology, 11(2), p.104.

Neale, M.J., Pan, J. & Keeney, S., 2005. Endonucleolytic processing of covalent protein-linked DNA double-strand breaks. Nature, 436(7053), pp.1053–7.

Nishant, K.T. & Rao, M.R.S., 2006. Molecular features of meiotic recombination hot spots. BioEssays : news and reviews in molecular, cellular and developmental biology, 28(1), pp.45–56.

Nitiss, J.L., 2009. DNA topoisomerase II and its growing repertoire of biological functions. Nature reviews. Cancer, 9(5), pp.327–37.

Niu, H. et al., 2007. Mek1 kinase is regulated to suppress double-strand break repair between sister chromatids during budding yeast meiosis. Molecular and cellular biology, 27(15), pp.5456–67.

Niu, H. et al., 2005. Partner choice during meiosis is regulated by Hop1-promoted dimerization of Mek1. Molecular biology of the cell, 16(12), pp.5804–18.

Novak, I. et al., 2008. Cohesin Smc1beta determines meiotic chromatin axis loop organization. The Journal of cell biology, 180(1), pp.83–90.

Oh, S.D. et al., 2007. BLM Ortholog, Sgs1, Prevents Aberrant Crossing-over by Suppressing Formation of Multichromatid Joint Molecules. Cell, 130(2), pp.259–272.

Oh, S.D. et al., 2008. RecQ Helicase, Sgs1, and XPF Family Endonuclease, Mus81-Mms4, Resolve Aberrant Joint Molecules during Meiotic Recombination. Molecular Cell, 31(3), pp.324–336.

Oke, A. et al., 2014. Controlling Meiotic Recombinational Repair – Specifying the Roles of ZMMs, Sgs1 and Mus81/Mms4 in Crossover Formation M. Lichten, ed. PLoS Genetics, 10(10), p.e1004690.

Page, S.L. & Hawley, R.S., 2001. c(3)G encodes a Drosophila synaptonemal complex protein. Genes & Development, 15(23), pp.3130–3143.

Page, S.L. & Hawley, R.S., 2004. The genetics and molecular biology of the synaptonemal complex. Annual review of cell and developmental biology, 20, pp.525–58.

Pan, J. et al., 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. Cell, 144(5), pp.719–31.

Panizza, S. et al., 2011. Spo11-accessory proteins link double-strand break sites to the chromosome axis in early meiotic recombination. Cell, 146(3), pp.372–83.

Pâques, F. & Haber, J.E., 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. Microbiology and molecular biology reviews : MMBR, 63(2), pp. 349–404.

Parvanov, E.D., Petkov, P.M. & Paigen, K., 2010. Prdm9 controls activation of mammalian recombination hotspots. Science (New York, N.Y.), 327(5967), p.835.

Paull, T.T., 2015. Mechanisms of ATM Activation. Annual review of biochemistry, 84, pp.711–38.

Pedersen, J.M. et al., 2012. DNA Topoisomerases maintain promoters in a state competent for transcriptional activation in Saccharomyces cerevisiae. PLoS genetics, 8(12), p.e1003128.

Pommier, Y. & Marchand, C., 2011. Interfacial inhibitors: targeting macromolecular complexes. Nature Reviews Drug Discovery, 11(1), pp.25–36.

Pratto, F. et al., 2014. Recombination initiation maps of individual human genomes. Science, 346(6211), pp.1256442–1256442.

Price, B.D. & D'Andrea, A.D., 2013. Chromatin remodeling at DNA double-strand breaks. Cell, 152(6), pp.1344–54.

Prieler, S. et al., 2005. The control of Spo11's interaction with meiotic recombination hotspots. Genes & development, 19(2), pp.255–69.

Primig, M. et al., 2000. The core meiotic transcriptome in budding yeasts. Nature Genetics, 26(4), pp.415–423.

Rasmussen, S.W. & Holm, P.B., 1984. The synaptonemal complex, recombination nodules and chiasmata in human spermatocytes. Symposia of the Society for Experimental Biology, 38, pp. 271–92.

Refolio, E. et al., 2011. The Ddc2/ATRIP checkpoint protein monitors meiotic recombination intermediates. Journal of cell science, 124(Pt 14), pp.2488–500.

Renkawitz, J. et al., 2013. Monitoring homology search during DNA double-strand break repair in vivo. Molecular cell, 50(2), pp.261–72.

Reynolds, A. et al., 2013. RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. Nature Genetics, 45(3), pp.269–278.

Robert, T. et al., 2016. The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. Science, 351(6276), pp.943-949.

Robine, N. et al., 2007. Genome-wide redistribution of meiotic double-strand breaks in Saccharomyces cerevisiae. Molecular and cellular biology, 27(5), pp.1868–80.

Rockmill, B. et al., 2013. High throughput sequencing reveals alterations in the recombination signatures with diminishing Spo11 activity. PLoS genetics, 9(10), p.e1003932.

Rockmill, B. & Roeder, G.S., 1991. A meiosis-specific protein kinase homolog required for chromosome synapsis and recombination. Genes & development, 5(12B), pp.2392–404.

Roeder, G.S. & Bailis, J.M., 2000. The pachytene checkpoint. Trends in genetics : TIG, 16(9), pp.395–403.

Roedgaard, M. et al., 2015. DNA Topoisomerases Are Required for Preinitiation Complex Assembly during GAL Gene Activation. PloS one, 10(7), p.e0132739.

Rogacheva, M. V. et al., 2014. Mlh1-Mlh3, a Meiotic Crossover and DNA Mismatch Repair Factor, Is a Msh2-Msh3-stimulated Endonuclease. Journal of Biological Chemistry, 289(9), pp.5664–5673.

San Filippo, J., Sung, P. & Klein, H., 2008. Mechanism of Eukaryotic Homologous Recombination. Annual Review of Biochemistry, 77(1), pp.229–257.

Sasaki, M. et al., 2013. Meiotic recombination initiation in and around retrotransposable elements in Saccharomyces cerevisiae. PLoS genetics, 9(8), p.e1003732.

Sasanuma, H. et al., 2007. Meiotic association between Spo11 regulated by Rec102, Rec104 and Rec114. Nucleic Acids Research, 35(4), pp.1119–1133.

Schacherer, J. et al., 2007. Genome-wide analysis of nucleotide-level variation in commonly used Saccharomyces cerevisiae strains. PloS one, 2(3), p.e322.

Schoeffler, A.J. & Berger, J.M., 2005. Recent advances in understanding structure–function relationships in the type II topoisomerase mechanism. Biochemical Society Transactions, 33(6), p.1465.

Schwacha, A. & Kleckner, N., 1994. Identification of joint molecules that form frequently between homologs but rarely between sister chromatids during yeast meiosis. Cell, 76(1), pp.51–63.

Schwacha, A. & Kleckner, N., 1997. Interhomolog Bias during Meiotic Recombination: Meiotic Functions Promote a Highly Differentiated Interhomolog-Only Pathway. Cell, 90(6), pp.1123–1135.

Schwartz, E.K. & Heyer, W.-D., 2011. Processing of joint molecule intermediates by structure-selective endonucleases during homologous recombination in eukaryotes. Chromosoma, 120(2), pp.109–127.

Segal, E. & Widom, J., 2009. Poly(dA:dT) tracts: major determinants of nucleosome organization. Current Opinion in Structural Biology, 19(1), pp.65–71.

Shiloh, Y. & Ziv, Y., 2013. The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. Nature reviews. Molecular cell biology, 14(4), pp.197–210.

Shinohara, M. et al., 2008. Crossover assurance and crossover interference are distinctly regulated by the ZMM proteins during yeast meiosis. Nature Genetics, 40(3), pp.299–309.

Shinohara, M. et al., 2003. Crossover interference in Saccharomyces cerevisiae requires a TID1/RDH54- and DMC1-dependent pathway. Genetics, 163(4), pp.1273–86.

Shinohara, M. et al., 2015. DNA damage response clamp 9-1-1 promotes assembly of ZMM proteins for formation of crossovers and synaptonemal complex. Journal of Cell Science, 128(8), pp. 1494–1506.

Shroff, R. et al., 2004. Distribution and Dynamics of Chromatin Modification Induced by a Defined DNA Double-Strand Break. Current Biology, 14(19), pp.1703–1711.

Sidhu, G.K. et al., 2015. Recombination patterns in maize reveal limits to crossover homeostasis. Proceedings of the National Academy of Sciences, 112(52), pp.15982–15987.

Smagulova, F. et al., 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature, 472(7343), pp.375–8.

Smith, G.R. et al., 2003. Fission yeast Mus81.Eme1 Holliday junction resolvase is required for meiotic crossing over but not for gene conversion. Genetics, 165(4), pp.2289–93.

Smith, J. et al., 2010. The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. Advances in cancer research, 108, pp.73–112.

Smolka, M.B. et al., 2007. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. Proceedings of the National Academy of Sciences of the United States of America, 104(25), pp.10364–9.

Sommermeyer, V. et al., 2013. Spp1, a member of the Set1 Complex, promotes meiotic DSB formation in promoters by tethering histone H3K4 methylation sites to chromosome axes. Molecular cell, 49(1), pp.43–54.

Spies, M. & Fishel, R., 2015. Mismatch Repair during Homologous and Homeologous Recombination. Cold Spring Harbor Perspectives in Biology, 7(3), p.a022657.

Spitzner, J.R., Chung, I.K. & Muller, M.T., 1990. Eukaryotic topoisomerase II preferentially cleaves alternating purine-pyrimidine repeats. Nucleic acids research, 18(1), pp.1–11.

Stahl, F., 1996. Meiotic recombination in yeast: coronation of the double-strand-break repair model. Cell, 87(6), pp.965–8.

Stahl, F.W. et al., 2004. Does crossover interference count in Saccharomyces cerevisiae? Genetics, 168(1), pp.35–48.

Stahl, F.W. & Foss, H.M., 2009. On Spo16 and the coefficient of coincidence. Genetics, 181(1), pp. 327–30.

Steiner, W.W. et al., 2009. Novel nucleotide sequence motifs that produce hotspots of meiotic recombination in Schizosaccharomyces pombe. Genetics, 182(2), pp.459–69.

Sun, X. et al., 2015. Transcription dynamically patterns the meiotic chromosome-axis interface. eLife, 4, p.e07424.

Sym, M., Engebrecht, J.A. & Roeder, G.S., 1993. ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis. Cell, 72(3), pp.365–78.

Sym, M. & Roeder, G.S., 1994. Crossover interference is abolished in the absence of a synaptonemal complex protein. Cell, 79(2), pp.283–92.

Symington, L.S., Rothstein, R. & Lisby, M., 2014. Mechanisms and regulation of mitotic recombination in Saccharomyces cerevisiae. Genetics, 198(3), pp.795–835.

Szostak, J.W. et al., 1983. The double-strand-break repair model for recombination. Cell, 33(1), pp. 25–35.

Tease, C., Hartshorne, G. & Hultén, M., 2006. Altered patterns of meiotic recombination in human fetal oocytes with asynapsis and/or synaptonemal complex fragmentation at pachytene. Reproductive biomedicine online, 13(1), pp.88–95.

Thacker, D. et al., 2014. Homologue engagement controls meiotic DNA break number and distribution. Nature, 510(7504), pp.241–6.

Tischfield, S.E. & Keeney, S., 2012. Scale matters: the spatial correlation of yeast meiotic DNA breaks with histone H3 trimethylation is driven largely by independent colocalization at promoters. Cell cycle (Georgetown, Tex.), 11(8), pp.1496–503.

Toyoizumi, H. & Tsubouchi, H., 2012. Estimating the Number of Double-Strand Breaks Formed During Meiosis from Partial Observation. Journal of Computational Biology, 19(12), pp.1277–1283.

Traven, A. & Heierhorst, J., 2005. SQ/TQ cluster domains: concentrated ATM/ATR kinase phosphorylation site regions in DNA-damage-response proteins. BioEssays, 27(4), pp.397–407.

Tsai, C.J. et al., 2008. Meiotic crossover number and distribution are regulated by a dosage compensation protein that resembles a condensin subunit. Genes & Development, 22(2), pp. 194–211.

Tsubouchi, T., Zhao, H. & Roeder, G.S., 2006. The Meiosis-Specific Zip4 Protein Regulates Crossover Distribution by Promoting Synaptonemal Complex Formation Together with Zip2. Developmental Cell, 10(6), pp.809–819.

Tung, K.S., Hong, E.J. & Roeder, G.S., 2000. The pachytene checkpoint prevents accumulation and phosphorylation of the meiosis-specific transcription factor Ndt80. Proceedings of the National Academy of Sciences of the United States of America, 97(22), pp.12187–92.

Usui, T., Ogawa, H. & Petrini, J.H., 2001. A DNA damage response pathway controlled by Tel1 and the Mre11 complex. Molecular cell, 7(6), pp.1255–66.

Vader, G. et al., 2011. Protection of repetitive DNA borders from self-induced meiotic instability. Nature, 477(7362), pp.115–9.

Valencia, M. et al., 2001. NEJ1 controls non-homologous end joining in Saccharomyces cerevisiae. Nature, 414(6864), pp.666–669.

Vallente, R.U., Cheng, E.Y. & Hassold, T.J., 2006. The synaptonemal complex and meiotic recombination in humans: new approaches to old questions. Chromosoma, 115(3), pp.241–249.

Veuger, S.J. et al., 2003. Radiosensitization and DNA repair inhibition by the combined use of novel inhibitors of DNA-dependent protein kinase and poly(ADP-ribose) polymerase-1. Cancer research, 63(18), pp.6008–15.

Vincenten, N. et al., 2015. The kinetochore prevents centromere-proximal crossover recombination during meiosis. eLife, 4, pp1-25.

Vrielynck, N. et al., 2016. A DNA topoisomerase VI—like complex initiates meiotic recombination. Science, 351(6276), pp.939-944.

Wan, L. et al., 2004. Mek1 kinase activity functions downstream of RED1 in the regulation of meiotic double strand break repair in budding yeast. Molecular biology of the cell, 15(1), pp.11–23.

Webb, A.R., 2000. Gamma mixture models for target recognition. Pattern Recognition, 33(12), pp. 2045–2054.

Wilson, T.E. & Lieber, M.R., 1999. Efficient processing of DNA ends during yeast nonhomologous end joining. Evidence for a DNA polymerase beta (Pol4)-dependent pathway. The Journal of biological chemistry, 274(33), pp.23599–609.

Winter, E., 2012. The Sum1/Ndt80 transcriptional switch and commitment to meiosis in Saccharomyces cerevisiae. Microbiology and molecular biology reviews : MMBR, 76(1), pp.1–15.

Wojtasz, L. et al., 2009. Mouse HORMAD1 and HORMAD2, two conserved meiotic chromosomal proteins, are depleted from synapsed chromosome axes with the help of TRIP13 AAA-ATPase. M. Lichten, ed. PLoS genetics, 5(10), p.e1000702.

Woltering, D. et al., 2000. Meiotic segregation, synapsis, and recombination checkpoint functions require physical interaction between the chromosomal proteins Red1p and Hop1p. Molecular and cellular biology, 20(18), pp.6646–58.

Wood, V. et al., 2002. The genome sequence of Schizosaccharomyces pombe. Nature, 415(6874), pp.871–80.

Wu, L. & Hickson, I.D., 2003. The Bloom's syndrome helicase suppresses crossing over during homologous recombination. Nature, 426(6968), pp.870–874.

Wu, T.C. & Lichten, M., 1995. Factors that affect the location and frequency of meiosis-induced double-strand breaks in Saccharomyces cerevisiae. Genetics, 140(1), pp.55–66.

Wu, X., Wilson, T.E. & Lieber, M.R., 1999. A role for FEN-1 in nonhomologous DNA end joining: the order of strand annealing and nucleolytic processing events. Proceedings of the National Academy of Sciences of the United States of America, 96(4), pp.1303–8.

Xu, L. et al., 1995. NDT80, a meiosis-specific gene required for exit from pachytene in Saccharomyces cerevisiae. Molecular and cellular biology, 15(12), pp.6572–81.

Xu, Y. et al., 1996. Targeted disruption of ATM leads to growth retardation, chromosomal fragmentation during meiosis, immune defects, and thymic lymphoma. Genes & development, 10(19), pp.2411–22.

Yamada, S., Ohta, K. & Yamada, T., 2013. Acetylated Histone H3K9 is associated with meiotic recombination hotspots, and plays a role in recombination redundantly with other factors including the H3K4 methylase Set1 in fission yeast. Nucleic acids research, 41(6), pp.3504–17.

Yokoo, R. et al., 2012. COSA-1 Reveals Robust Homeostasis and Separable Licensing and Reinforcement Steps Governing Meiotic Crossovers. Cell, 149(1), pp.75–87.

You, Z. et al., 2005. ATM Activation and Its Recruitment to Damaged DNA Require Binding to the C Terminus of Nbs1. Molecular and Cellular Biology, 25(13), pp.5363–5379.

Zakharyevich, K. et al., 2012. Delineation of Joint Molecule Resolution Pathways in Meiosis Identifies a Crossover-Specific Resolvase. Cell, 149(2), pp.334–347.

Zanders, S. & Alani, E., 2009. The pch2? Mutation in Baker's Yeast Alters Meiotic Crossover Levels and Confers a Defect in Crossover Interference G. P. Copenhaver, ed. PLoS Genetics, 5(7), p.e1000571.

Zhang, L., Liang, Z., et al., 2014. Crossover Patterning by the Beam-Film Model: Analysis and Implications R. S. Hawley, ed. PLoS Genetics, 10(1), p.e1004042.

Zhang, L. et al., 2011. Meiotic double-strand breaks occur once per pair of (sister) chromatids and, via Mec1/ATR and Tel1/ATM, once per quartet of chromatids. Proceedings of the National Academy of Sciences of the United States of America, 108(50), pp.20036–41.

Zhang, L., Wang, S., et al., 2014. Topoisomerase II mediates meiotic crossover interference. Nature, 511(7511), pp.551–556.

Zhao, H., Speed, T.P. & McPeek, M.S., 1995. Statistical analysis of crossover interference using the chi-square model. Genetics, 139(2).

Zierhut, C. et al., 2004. Mnd1 is required for meiotic interhomolog repair. Current biology : CB, 14(9), pp.752–62.

Ziolkowski, P.A. et al., 2015. Juxtaposition of heterozygous and homozygous regions causes reciprocal crossover remodelling via interference during Arabidopsis meiosis. eLife, 4.