Genomic insights into fine-scale recombination variation in adaptively diverging threespine stickleback fish (*Gasterosteus aculeatus*)

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Vrinda Venu
aus Koorkkenchery, India

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	02 March 2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Dr. Felicity Jones
2. Berichterstatter:	Prof. Dr. Thomas Lahaye

Contributions

In the study presented in this thesis, I designed and performed all the experiments (unless stated separately), carried out all the data analysis, generated the figures, and wrote the thesis.

My PhD supervisor Dr. Felicity Jones conceived the original idea of the project, contributed to the experimental design, guided the data analysis throughout the project, and provided suggestions for thesis writing.

My colleague Enni Harjunmaa designed and performed the ATAC-sequencing experiment and data analysis. Enni and I equally contributed in designing and performing all the ChIP sequencing related experiments and data analysis.

My colleague Andreea Dréau contributed towards preparing the crossover calling bioinformatic pipeline.

My colleagues Saad Arif, Cecilia Martinez, and Li Ying Tan collected samples of marine and freshwater sticklebacks from River Tyne, Scotland. They also performed four experimental wild crosses used for the whole genome sequencing.

A published article I coauthored has been enclosed in this thesis (Chapter 4). Author contributions are clearly stated along with the chapter.

Acknowledgements

The successful completion of this thesis would not have been possible without the help and support from a number of people. First and foremost, I express my deep sense of gratitude to my advisor Dr. Felicity Jones for her expert guidance and immense support throughout my PhD. Thank you for giving me an opportunity to work in this project, for motivating me to learn new things and helping me to gain more confidence in myself. I would like to thank Dr. Frank Chan for all the discussions, constructive suggestions, bioinformatic lessons, and help with setting up the bioinformatic pipeline used in this project. Furthermore, I wish to thank the members of my TAC committee: my university supervisor Prof. Dr. Thomas Lahaye and Prof. Dr. Ralf Sommer for their valuable suggestions and feedback during various stages of my PhD. I am grateful to Prof. Dr. Nico Michiels for his willingness to serve in my defense committee.

The Friedrich Miescher Laboratory and the Max Planck Institute for Developmental Biology provided a collaborative work environment and I greatly benefited from the expertise of numerous people. I am deeply indebted to Christa Lanz, Julia Hildebrandt, and Katrin Fritschi for performing more than 155 lanes of Illumina HiSeq sequencing for this project. I would like to thank Derek Lundberg for all his efforts in setting up the liquid handling robot for sequencing library preparation that made my life easier. I thank both Derek and Beth Rowan for all the discussions and for their help with developing a home-made multiplexing library preparation protocol. I would like to acknowledge our campus IT support for providing the wonderful cluster facility. I express my gratitude to CeGaT Tübingen for allowing us to use their Covaris® instrument and especially to Natascha Günther for making all the short notice appointments possible. Many thanks to Dagmar Sigurdardottir, PhD program coordinator and Herta Soffel, secretary to FML, for the administrative support. Also, I am indebted to the Max Planck Society and European Research Council for all the monitory and infrastructure support that made this project possible.

My life, both scientific and personal, in the past five years have been greatly influenced by my fantastic colleagues. I would like to thank: Enni Harjunmaa for being a wonderful colleague and a best buddy throughout my PhD. Thanks for supporting me through the tough times and also for proofreading the thesis; Andreea Dréau for helping me with the bioinformatic work and for being a great friend; Li Ying Tan for field sampling of the fishes used in this study and for all the help with fish-house works. I miss those days when we used to share the work bench; Elena Avdievich, for her constant support and motivation. Thank you for staying up late nights with me to finish the work on time; Muhua Wang for all the help with bioinformatics and statistics; Stanley Neufeld, Jukka-Pekka Verta, and Marek Kučka for sharing various protocols and for the fruitful discussions;

Sebastian Kick for wonderfully managing our fish facility; Layla Hiramatsu for helping to improve my thesis; Insa Hirschberg and Moritz Peters for translating the thesis abstract to German; former colleagues Saad Arif and Cecilia Martinez for collecting the wild fishes used in this study and teaching me how to handle sticklebacks. Many thanks to all the past and present members of Jones and Chan lab including Michelle Yancoskie, Bill Beluch, Domenico Scionti, Melanie Kirch, Kavita Venkataramani, Joao Castro, Stefano Lazzarano, Ludmilla Gaspar, Volker Soltys, Dingwen Su, and Min Zheng for all the helpful discussions and sharing the laughter. I have learned a lot from all of you.

I am grateful to all my teachers, friends, and family members for their constant support and love. The whole journey is fueled by three most important people in my life, my parents and my husband Ajeesh. This accomplishment would not have been possible without them. Thank you for everything!

Table of contents

List of abbreviations

AE	Axial element
ATAC-seq	Assay for transposase-accessible chromatin using sequencing
AUC	Area under the curve
bp	Base pair
BQSR	Base quality score recalibration
BWA	Burrows-Wheeler Aligner
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
cM	Centimorgan
CO	Crossover
CpG	C and G base pairs connected by a phosphodiester bond
CSS	Cluster separation score
DHJ	Double Holliday junction
D-loop	Displacement loop
DNA	Deoxyribonucleic acid
DSB	Double strand break
FDR	False discovery rate
F_{ST}	Fixation index
GATK	Genome analysis toolkit
GBGC	GC biased gene conversion
GEM	Gel bead-in-emulsion
GWAS	Genome wide association study
GWRR	Genome wide recombination rate
H3K4me3	Histone H3, lysine 4, trimethylation
H3K36me3	Histone H3, lysine 36, trimethylation
HMM	Hidden Markov model
HMW	High molecular weight
kb	Kilo base
LD	Linkage disequilibrium
Mb	Mega base
NCO	Noncrossover

N_e	Effective population size
PAR	Pseudoautosomal region
PCR	Polymerase chain reaction
QTL	Quantitative trait loci
SC	Synaptonemal complex
SD	Standard deviation
SDSA	Synthesis dependent strand annealing
SE	Standard error
SNP	Single nucleotide polymorphism
SPRI	Solid phase reverse immobilization
SRR	Standardized recombination rate
ssDNA	Single strand DNA
TES	Transcription end site
TF	Trigger factor
TSS	Transcription start site
VCF	Variant call format
Genes/proteins	
CPLX1	Complexin 1
DMC1	Disrupted meiotic cDNA1
EDA	Ectodysplasin A
HEI10	Human enhancer of invasion 10
HIM5	High incidence of males-5
MLH1	MutL homolog 1
MSH4	MutS protein homolog 4
PITX1	Pituitary homeobox transcription factor 1
PRDM9	PR Domain Zinc Finger Protein 9
RAD51	DNA repair protein RAD51 homolog 1
REC1	Recombination abnormal 1
REC8	Meiotic recombination protein REC8
RNF212	Ring Finger Protein 212
RPA	Replication protein A
SPO11	Sporulation-specific protein 11

Abstract

Meiotic recombination is one of the major molecular mechanisms generating genetic diversity and influencing genome evolution. By shuffling allelic combinations, it can directly influence the patterns and efficacy of natural selection. Studies in various organisms have shown that the rate and placement of recombination varies substantially within the genome, among individuals, between sexes and among different species. It is hypothesized that this variation plays an important role in genome evolution. In this PhD thesis, I investigated the extent and molecular basis of recombination variation in adaptively diverging threespine stickleback fish (*Gasterosteus aculeatus)* to further understand its evolutionary implications. I used both ChIP-sequencing and whole genome sequencing of pedigrees to empirically identify and quantify double strand breaks (DSBs) and meiotic crossovers (COs). Whole genome sequencing of large nuclear families was performed to identify meiotic crossovers in 36 individuals of diverging marine and freshwater ecotypes and their hybrids. This produced the first genome-wide high-resolution sex-specific and ecotype-specific map of contemporary recombination events in sticklebacks. The results show striking differences in crossover number and placement between sexes. Females recombine nearly 1.76 times more than males and their COs are distributed all over the chromosome while male COs predominantly occur near the chromosomal periphery. When compared among ecotypes a significant reduction in overall recombination rate was observed in hybrid females compared to pure forms. Even though the known loci underlying marine-freshwater adaptive divergence tend to fall in regions of low recombination, considerable female recombination is observed in the regions between adaptive loci. This suggests that the sexual dimorphism in recombination phenotype may have important evolutionary implications.

At the fine-scale, COs and male DSBs are nonrandomly distributed involving 'semi-hot' hotspots and coldspots of recombination. I report a significant association of male DSBs and COs with functionally active open chromatin regions like gene promoters, whereas female COs did not show an association more than expected by chance. However, a considerable number of COs and DSBs away from any of the tested open chromatin marks suggests possibility of additional novel mechanisms of recombination regulation in sticklebacks.

In addition, we developed a novel method for constructing individualized recombination maps from pooled gamete DNA using linked read sequencing technology by 10X Genomics®. We tested the method by contrasting recombination profiles of gametic and somatic tissue from a hybrid mouse and stickleback fish. Our pipeline faithfully detects previously described recombination hotspots in mice at high resolution and identify many novel hotspots across the genome in

both species and thereby demonstrate the efficiency of the novel method. This method could be employed for large scale QTL mapping studies to further understand the genetic basis of recombination variation reported in this thesis.

By bridging the gap between natural populations and lab organisms with large clutch sizes and tractable genetic tools, this work shows the utility of the stickleback system and provides important groundwork for further studies of heterochiasmy and divergence in recombination during adaptation to differing environments.

Zusammenfassung

Die meiotische Rekombination gehört zu den wichtigsten molekularen Mechanismen, die für genetische Vielfalt sowie die evolutionäre Entwicklung des Genoms verantwortlich sind. Von zentraler Bedeutung ist dafür die Veränderung von Allel-Kombinationen, die sich direkt auf Wirkmechanismen und Effizienz der natürlichen Selektion auswirkt. Studien an verschiedenen Organismen haben gezeigt, dass die Häufigkeit und Position von Rekombinationsereignissen innerhalb des Genoms nicht nur zwischen Individuen, sondern auch zwischen Geschlechtern und unterschiedlichen Arten variiert. Daher wird vermutet, dass ebensolche Variationen maßgeblich zur evolutionären Entwicklung des Genoms beitragen. Im Rahmen dieser Doktorarbeit habe ich das Ausmaß sowie die molekulare Basis von variablen Rekombinationsereignissen im Kontext der adaptiven Divergenz bei dreistachligen Stichlingen (*Gasterosteus aculeatus*) untersucht, um ein besseres Verständnis ihrer evolutionären Bedeutung zu erlangen. Hierzu habe ich neben der ChIP-Sequenzierung ebenfalls die Gesamtgenomsequenzierung von DNA verwandter Individuen genutzt, um Doppelstrangbrüche (DSBs) und meiotische Crossover (COs) zu identifizieren sowie zu quantifizieren. Mit Hilfe der Gesamtgenomsequnezierung wurden meiotische Crossover bei 36 Individuen einer Kernfamilie identifiziert, deren Mitglieder unterschiedlicher Meeres- und Süßwasserökotypen sowie Hybriden angehörten. Auf diese Weise wurde die bislang erste genomweite, hochauflösende, geschlechter- und ökotypspezifische Kartierung von Rekombinationsereignissen in Stichlingen erreicht. Hier wurde deutlich, dass zwischen den Geschlechtern gravierende Unterschiede bei der Anzahl und Position von Rekombinationsereignissen bestehen. Weibliche Individuen zeigen fast 1.76-mal so viele Rekombinationsereignisse wie männliche Individuen und eine Verteilung derer erstreckt über das gesamte Chromosom, wohingegen sich Crossover bei Männchen auf die chromosomale Peripherie konzentrieren. Beim Vergleich der verschiedenen Ökotypen zeigte sich bei den weiblichen Hybriden eine maßgebliche Verringerung der Rekombinationsrate verglichen mit reinerbigen Individuen. Obwohl die bekannten Loci, die der adaptiven Divergenz zwischen Meer- und Süßwasser zugrunde liegen, dazu neigen, in Regionen mit geringer Rekombination zu fallen, wird in Weibchen in den Regionen zwischen den adaptiven Loci eine beträchtliche Anzahl an Rekombinationen beobachtet. Dies legt nahe, dass der sexuelle Dimorphismus im Rekombinationsphänotyp eine wichtige evolutionäre Bedeutung haben könnte.

In männlichen Individuen sind Crossover und Doppelstrangbrüche nicht zufällig verteilt und bilden „semi-hot“ Hotspots und Coldspots. Dabei konnte ich zeigen, dass Doppelstrangbrüche und Crossover in Männchen signifikant mit funktionellen Chromatinregionen in einer offenen Konfirmation, wie beispielsweise Promotern, in Verbindung gebracht werden können. Wohingegen

weibliche Crossover nicht mehr als eine durch Zufall erwartete Assoziation zeigten. Das gehäufte Auftreten von Crossovern und Doppelstrangbrüchen, die entfernt von den getesteten offenen Chromatinregionen liegen, deutet jedoch auf die Möglichkeit zusätzlicher neuartiger Mechanismen der Rekombinationsregulation bei Stichlingen hin.

Zusätzlich wurde im Rahmen dieser Arbeit eine neue Methode zur Erstellung einer individualisierten Rekombinationskarte aus gepoolter Gameten-DNA unter Verwendung der Linked-Read-Sequenzierungstechnologie von 10X Genomics® entwicklet. Diese Methode wurde durch den Vergleich der Rekombinationsprofile von gametischem und somatischem Gewebe einer Hybridmaus und eines Stichlings getestet. Es konnten zuverlässig und in hoher Auflösung die zuvor beschriebenen Rekombinations-Hotspots in Mäusen erkannt und in beiden Spezies viele neue Hotspots im gesamten Genom identifiziert werden, was die Effizienz dieser neuen Methode bestätigt. Zukünftig könnte sie für groß angelegte QTL-Kartierungsstudien verwendet werden, um die genetische Basis der in dieser Arbeit beschriebenen Rekombinationsvariationen besser zu verstehen.

Durch die Überbrückung der Grenze zwischen natürlichen Populationen und Laborорganismen mit großem Gelege und handhabbaren genetischen Methoden zeigt diese Arbeit die Nützlichkeit des Stichlingsystems auf und liefert wichtige Grundlagen für weitere Studien zu Heterochiasmus und Divergenz bei der Rekombination während der Anpassung an unterschiedliche Umgebungen.

Chapter 1

General introduction

The genome is the blueprint of an organism, containing the heritable information that codes for cellular, tissue, organ, and phenotypic function. Variation in the genome underlies the diversity of all living organisms around us. Heritable variation with differing fitness advantages are important for natural selection to act on and facilitate adaptation. The two major mechanisms that generate stable and heritable genomic variation are mutation and recombination. Although important in the generation of variation, mutations occur mostly randomly along the genome and their effects are often unpredictable. In contrast, meiotic recombination is a well-regulated process, unique to sexually reproducing organisms, that produces novel allelic combinations during gamete production by exchanging genetic material between parental homologous chromosomes. By creating genetic variation for natural selection, recombination and its regulatory mechanisms play an important role in species adaptation. This thesis focuses on understanding meiotic recombination in the context of adaptive divergence and speciation.

Our understanding about recombination begins in the early 20th century when researchers noticed violation of Mendel's law of independent assortment during trait segregation. Studies in pea plants and *Drosophila* showed that certain combinations of traits always appear together whereas certain other combinations separate more frequently (Bateson 1905; Morgan 1910). Thomas Hunt Morgan first proposed the idea of "genetic linkage" and "crossing over" to explain the varying degrees of association between the traits (Morgan 1913). He proposed that if the genetic factors underlying these traits are physically present in the same chromosome, they could be coupled together. Based on the microscopical observation of twisting of homologous chromosomes during early meiosis, he suggested that the pairs of homologous chromosomes may cross over with each other and exchange the genetic factors. He also hypothesized that the probability of two factors present in the same chromosome to be separated is dependent on the physical distance between them. As a result, factors with physical proximity are kept in linkage whereas the ones that are further apart may separate over generations (Morgan 1911; Morgan 1913). Twenty years later, Barbara McClintock and her student Harriet Creighton observed the first physical evidence of the exchange of regions between morphologically distinguishable homologous chromosomes under a light microscope (Creighton and McClintock 1931). Since then, numerous studies in various eukaryotic organisms have been carried out to

understand the process of crossing over and its molecular and evolutionary implications.

Owing to intense research over the last few decades, we now acknowledge the vital importance of homologous recombination in the completion of meiosis. Physical attachment between homologous chromosomes during crossover is necessary for proper segregation of chromosomes (Fledel-Alon et al. 2009). Complete absence or misplacement of crossover events can lead to severe chromosomal abnormalities and result in congenital birth defects (Hassold and Hunt 2001; Inoue and Lupski 2002). On the other hand, careful placement of crossovers can produce beneficial gene combinations that increase individual fitness and facilitate natural selection (Felsenstein 1974; Rice 2002). Due to these potentially extreme consequences of recombination, the frequency and placement of recombination events are highly regulated (Webster and Hurst 2012; Wang et al. 2015a).

Despite the clear advantage of tight regulation, the rate and the placement of recombination actually vary substantially within the genome, among individuals, sexes and different species. This observation leads to several questions. Why does it vary so much? What causes the variation? Does it have any fitness consequences? To answer these questions, recombination rate variation has to be quantified at various levels in multiple species. Studying recombination landscapes in natural populations under high selection pressures could point to important insights about its fitness consequences. Towards that aim, in this thesis I present a comprehensive study of recombination landscape in an evolutionary model organism, the threespine stickleback fish. My overarching goal is to understand the extent and molecular basis of recombination rate variation in the context of adaptive divergence and speciation.

In this chapter, I review the existing knowledge about the process of meiotic recombination, variation in its rate and placement, evolutionary implications, and the unique advantages of the threespine stickleback fish model in understanding recombination in an evolutionary context.

1.1 Meiotic recombination

Meiosis is a specialized cell division occurring in sexually reproducing organisms. During meiosis, one diploid parental cell produces four genetically distinct haploid gametes. The major specialties of meiosis in contrast to mitosis are: 1) two rounds of cell divisions followed by one round of DNA replication, 2) segregation of sister chromatids into one cell during the first division, which will be then separated during the second division, and 3) exchange of genetic material between homologous chromosomes by recombination. This homologous recombination during the 1st cell division of meiosis causes genetic uniqueness of the resultant haploid cells. The ploidy level will be later restored during fertilization.

Within the specialized first cell division, the initial stage, prophase I, is the longest and most eventful stage of meiosis. It has been further divided into five stages based on chromosome appearance. In chronological order, they are named leptotene, zygotene, pachytene, diplotene, and diakinesis. Figure 1.1 shows the first three stages and major events happening during those stages. During leptotene, (derived from Greek meaning "thin threads") chromosomes start to condense and a proteinaceous structure called axial elements (AE) is established for each chromosome. Chromatin is then attached onto this AE as loops. Most importantly, it is at this stage the DNA double strand breaks (DSBs) that initiate recombination form. A topoisomerase protein called SPO11 along with accessory proteins catalyzes the DSB formation and attaches covalently to the break point (Keeney et al. 1997; Lam and Keeney 2014). This SPO11 is subsequently removed from the breaks along with an oligonucleotide. The broken short ends are further chewed to produce single strand overhangs (Neale et al. 2005). These single strand ends are initially bound by the replication protein A (RPA) and will be later replaced by RAD51 and a meiosis specific protein DMC1 (Bishop et al. 1992). This protein-DNA complex then invades the uncut homologous chromosome at the site of homology and forms a structure called displacement loop (D-loop) (Szostak et al. 1983). It is after D-loop formation that the decision is made whether the repair goes through a crossover forming pathway or alternative noncrossover pathway (SDSA: synthesis-dependent strand annealing). During these steps, meiosis progresses into the next stage, called zygotene.

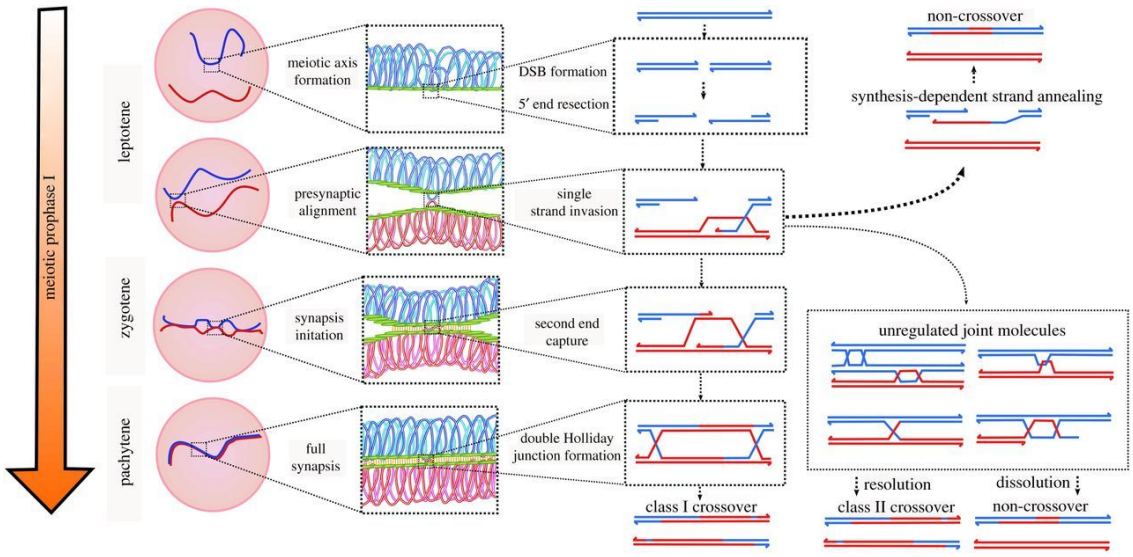


Figure 1.1: Mechanism of meiotic recombination. Major stages of prophase I along with chromatin level changes and single molecule level events at each stage is depicted. Blue and red color represent homologous chromosomes inherited from each parent. Synaptonemal complex is shown in green and orange. DNA is tightly packed into large chromatin loops. In right-most panels, molecule level events are shown. 5' end of DNA strands are marked with arrow head. Figure adapted from (Morgan et al. 2017b).

At zygotene stage ("paired threads"), homologous chromosomes are closely associated at a distance of about 100 nM, and the axial elements are connected though a central element forming the synaptonemal complex (SC) (Page and Hawley 2004). If the repair pathway did not follow SDSA, the D-loop captures the second end of the originally cut DNA and forms a double Holliday junction (DHJ) and progresses towards pachytene ("thick threads") stage. At this stage, the synapsis is complete and the homologous chromosomes along with SC exist as a packed tripartite structure. This structure holds until the next stage, diplotene ("two threads"), and during which the SC starts to degrade. In the last stage, called diakinesis ("moving through"), majority of the SC has broken down and the chromosomes are connected together solely by the chiasmata (the point of exchange between homologous chromosomes). The DHJ is resolved into crossover at this stage but contact at chiasmata remains. Meiosis then completes its first stage (prophase I) and progress into further stages such as metaphase I, anaphase I and telophase I. Homologous chromosomes segregate into two poles during anaphase I, and nuclear envelop forms around them during telophase I. This is followed by a cell division (cytokinesis) that produces two daughter cells. The first cell division of meiosis called reduction division thus completes and the daughter cells progress into Meiosis II which is similar to mitosis. At the end of meiosis II, four haploid gametes are produced. However, it has to be noted that all chromosomes in the haploid gametes are not necessarily recombinant, because a single crossover between homologous chromosomes involves only two of the four chromatids. Major steps of meiotic recombination are reviewed in Bomblies et al. (2015); Zickler and Kleckner (2015).

The steps during these initial stages of meiosis, including major steps in recombination, are found to be at least broadly conserved among species (Gray and Cohen 2016). With regards to genetic exchange, the important pathway decisions are made after D-loop formation and during double Holliday junction resolution. A majority of the D-loops are repaired through SDSA pathway that result in noncrossover. During noncrossovers (NCOs) the broken DNA is repaired using the homologous chromosome as a template. As a result, a small portion of donor chromosome is copied to the broken chromosome without altering the donor. However, a subset of DSBs resolved via double Holliday junction turn into crossovers (CO) (Zakharyevich et al. 2012). Crossovers differ from noncrossovers as they are formed by long range reciprocal exchange of genetic material between the homologs (CO and NCOs are discussed in detail in section 1.2). These crossovers are one of the major factors generating genetic diversity. In addition to that, the events that become crossovers provide physical contact between homologs and this in combination with sister chromatid cohesion enables proper segregation of chromosomes during first meiotic division (Fledel-Alon et al. 2009).

1.2 Crossover, noncrossover, and gene conversion

Crossovers and noncrossovers are two alternative outcomes of repairing a double-strand break during meiosis. In either case the double strand breaks are repaired using a chromatid from the homologous chromosome as a template. While crossovers are long range reciprocal exchanges of homologous chromosomes, noncrossovers span lengths from a few base pairs up to 2 kb (Figure 1.2). It has been observed that both crossovers and noncrossovers are often associated with short gene conversion events in which information is copied from donor to the recipient without maintaining reciprocity (Hurst et al. 1972; Cole et al. 2014). As a result, alleles within gene conversion tract segregate with 3:1 ratio deviation from the expected Mendelian transmission of 1:1 ratio of genetic information. COs always result in long range exchange of linked polymorphisms between homologous chromosomes whereas NCOs, being short, do not necessarily span sites that are heterozygous between homologs. For this reason, NCOs can only be detected using DNA sequencing if they generate a gene conversion involving a heterozygous site (Cole et al. 2012b). Consequently, even though both these processes contribute to genetic variation, noncrossovers only have a localized effect on genetic diversity.

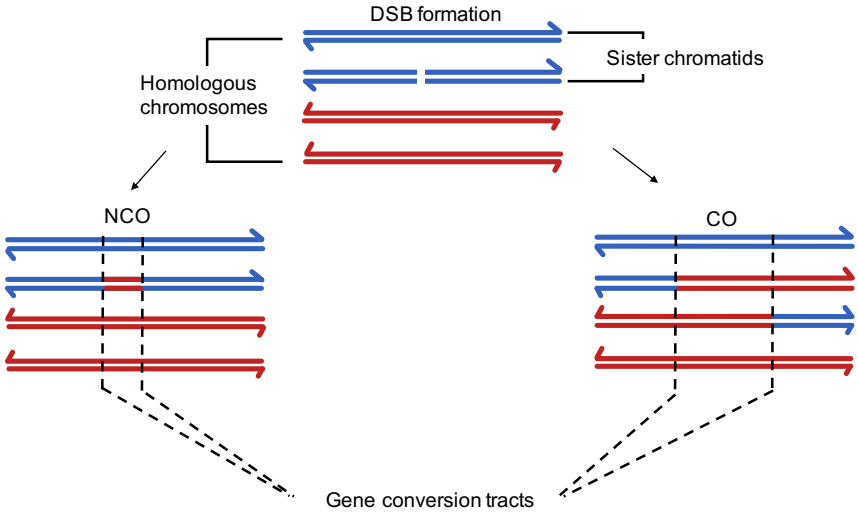


Figure 1.2: Different outcomes of double strand break repair. In a pair of homologous chromosomes (four chromatids), a double strand break is initiated in one of the chromatids and is repaired using its homologous chromosome as a template. Depending on its pathway selection for repair, either a noncrossover (NCO: patching the break by copying from homologous chromosome) or a crossover (CO: repair with reciprocal exchange between homologous chromosome) product is formed. Both NCO and CO are often accompanied by a gene conversion event in which a variant in the DSB initiated homolog is replaced by the variant in the donor homolog (region marked with dotted line in both CO and NCO product). This happens only when there is a variant present in the close vicinity of the double strand breaks. Otherwise the copying from the donor homolog will be obscure.

In most organisms, number of noncrossovers outnumbers reciprocal crossover events (Cole et al. 2012b; Stapley et al. 2017) meaning that large proportion of DSBs get repaired as a noncrossover. For example, while 10% of DSBs in mouse genome are repaired as CO (Cole et al. 2012a) and only 5% of DSBs in *Arabidopsis* turn into CO (Choi and Henderson 2015). However, in almost all studied organisms at least one crossover per chromosome pair is found to be obligatory for proper completion of meiosis. This suggests that there are crossover assurance mechanisms that may interact with earlier stages of DSB formation and CO designation (Hunter 2015). At the same time, it has been observed that 80% of the studied organisms have less than three COs per chromosome pair (Fernandes et al. 2018). One phenomenon that limits number of crossovers per chromosome is CO interference, in which a CO event at a location tends to suppress CO formation in the nearby region (Hillers 2004). This suggest that CO frequency is under selection in both directions. Even though the exact mechanism remains elusive, it has been thought that CO assurance sets the lower limit for crossover frequency and CO interference sets an upper limit (Martini et al. 2006; Hunter 2015).

Although 'recombination' is mostly used as an umbrella term that includes both crossovers and noncrossovers, in this thesis the term recombination is used interchangeably with crossovers unless otherwise stated.

1.3 Recombination rate

The frequency of crossover occurrence per physical size of the genome is called recombination rate. Recombination rate is often estimated in the unit of cM/Mb (centimorgan per megabase), where centimorgan (cM) is the measure of genetic distance between two markers and Mb is the measure of physical distance. If two markers in a chromosome have 1% chance of having a crossover between them in a generation, they are then considered to be at a genetic distance of 1 cM. Recombination rate is generally reported per individual genome or at any specific interval where crossover events have been localized.

1.4 Methods to study recombination

There are a number of different methods to study and quantify recombination that offer different strengths and weaknesses in terms of their power, cost, resolution and ability to detect contemporary recombination in individuals versus recombination in populations over evolutionary time. Popular methods being used in the field and their strengths and weaknesses are discussed below.

1.4.1 Genetic map (Linkage map)

When Thomas Hunt Morgan first proposed the idea of crossing over, he had also suggested that the number of crossing over between any two loci in a chromosome depends on the physical distance between them (Morgan 1913). This idea became

the principle of linkage mapping in which markers (or genes) across a chromosome are ordered based on the frequency with which they are coinherited. This concept has been widely used to quantify the recombination rate between markers by applying it into crosses between inbred strains or pedigrees (Broman et al. 1998; Hawken et al. 1999; Kong et al. 2010; Dumont et al. 2011). Development of DNA technologies for actual physical mapping of DNA polymorphisms enabled comparison of genetic distance versus physical distance and thereby the reporting of recombination rate in between any two markers in terms of cM/Mb.

In families, each offspring can provide information of one paternal and one maternal meiotic event. Therefore, using pedigree-based datasets, sex-specific genetic maps can be generated. However, the major limitations of genetic maps are, lack of accuracy (confidence in defining marker order) and resolution (size of the interval within which CO is identified) which is highly dependent on the marker density and the number of meioses analyzed.

1.4.2 Whole genome sequencing of pedigrees

Whole genome sequencing of pedigrees is a powerful method to directly detect crossover events. Whole genome sequencing provides high density of markers (DNA polymorphisms) for the analysis. The availability of a reference genome and high coverage sequencing enables accurate identification of the physical position of each sequence polymorphism. Applied to pedigrees, this approach enables direct identification of precise locations of recombination events. Therefore, a combination of these factors enables map construction of contemporary recombination events (as opposed to historical events in a population, discussed below) with higher accuracy and resolution. In nuclear families, deeply sequenced and haplotype phased (statistical determination of the alleles situated on the same homolog of a chromosome) parental and offspring genomes can be directly compared to detect parental crossover events. The resolution of crossover events depends on the availability of markers that can distinguish parental haplotypes. This method has been successfully employed in multi-generational pedigrees (Smeds et al. 2016) and within a family quartet (parents and their two offspring) (Roach et al. 2010) for constructing individualized high-resolution map of contemporary crossover events. The number of crossover events detected and reconstructed for a focal parent is directly dependent on the number of offspring or descendants analyzed. If applied to organisms with large nuclear families comprising tens to hundreds of offspring, it offers a powerful approach to reconstruct and quantify an individual's recombination rate across the genome, with potential to identify individual variation in recombination hot- and coldspots. Additionally, in organisms with medium to high density of DNA polymorphisms, pedigree sequencing can be used to identify short noncrossover gene conversion events. However, while powerful, this approach can be costly for organisms with

large genome size and not all organisms will be amenable due to biological limitations in obtaining large pedigrees.

In this thesis, whole genome recombination maps for individual stickleback fishes are produced using whole genome sequencing of nuclear families comprising parents and about 94 offspring (presented in chapter2).

1.4.3 Linkage disequilibrium analysis

In scenarios where obtaining large pedigree or making in vitro crosses are challenging, whole genome sequencing of unrelated individuals of a population can be used for indirect estimation of recombination rate. This method measures linkage disequilibrium between pairs of segregating polymorphisms. If alleles at a locus are segregating independent of alleles at another locus then they are considered to be in linkage equilibrium. However, more often, alleles segregate in a nonrandom fashion, with alleles in close physical proximity on the same chromosome being passed together to subsequent generations more frequently than expected by chance. As a result, blocks of linked alleles called haplotypes segregate among individuals in a population. These physically connected alleles are then considered to be in linkage disequilibrium (LD).

To study recombination using a population genetic method, unrelated individuals are sampled from a population and genotyped across the genome at high depth using array-based, reduced representation sequencing (e.g., RADseq) or whole genome sequencing. From the genotyped population, a pairwise LD between all pairwise marker comparisons is then estimated using joint genotype frequencies for each pair of markers. Statistical tools such as LDhat or LDhelmet are used to estimate population scale recombination rate rho (ρ). Per generation recombination rate, r can be then calculated using the formula $\rho = 4N_e r$, where N_e is the effective population size. By drawing on numerous historical meioses that have occurred over evolutionary time in the population sample, this strategy is a powerful method to make very high resolution recombination maps from natural populations or from species in which it is difficult to obtain large crosses (Fearnhead and Donnelly 2001; McVean et al. 2002). Even though LD-based estimations produce high resolution recombination maps and enable identification of hotspots and coldspots, the dependence on effective population size (N_e) is an unavoidable and confounding factor in this approach. Effective population size varies across the genome due to non-neutral processes such as selection, making comparisons of *rho* across genomic intervals prone to false positives and false negatives. Further, historical changes in effective population size (e.g., due to population bottlenecking, population expansion, migration and drift) are notoriously difficult to estimate and account for precisely and yet generate different effective population sizes for different populations. Combined variation in N_e across the genome and among populations badly confounds comparisons of rho across the genome and among populations leading to potential false positives

and false negatives. Further, this population genetic method of studying recombination necessarily results in a population- and sex-averaged map of historical recombination events. Therefore, it is not suitable for studying recombination rate variation among individuals or sexes.

1.4.4 Sperm typing and single sperm sequencing

Gametes carry recombined DNA. Therefore, an alternative to identifying parental recombination events from the diploid genomes of offspring, one could also obtain the same information from the haploid gametes of each parent. Sperm typing is one of the popular methods to map crossover events at a given locus with high resolution. In this method, DNA extracted from multiple sperm cells of a single donor is subjected to allele specific PCR. The PCR primers are designed in a such a way that different sized amplicons are produced from parental and recombinant genotypes. By analyzing the amount of recombinant amplicon in comparison to the parental sample, the crossover rate at that given locus could be estimated (Li et al. 1988). This method has been mostly used for validation and fine-scale characterization of known hotspots. An added advantage of this method is that from gametes we can identify even the crossovers that did not pass on to the offspring. The major limitation of this method is the difficulty to apply it in the whole genome scale. However, recent studies have obtained promising results using microfluidic platforms to separate single sperm followed by whole genome amplification and sequencing (Wang et al. 2012; Hinch et al. 2019). By sequencing large number of single sperm DNA, one could obtain high-resolution whole genome individualized recombination maps. A number of technical challenges are encountered when applying this approach: it is relatively low throughput - successful sorting and sequencing of isolated sperm has been performed for ~100 of human sperm and ~220 mouse sperm; and whole genome amplification chemistry has higher variation in performance and amplification biases than standard short read sequencing methods.

In chapter 4, I introduce a novel cost-effective method that we developed for whole genome individualized recombination map construction from pooled gamete DNA that overcomes some of the challenges mentioned above.

1.4.5 Cytology techniques

Early characterization of recombination events was heavily dependent on cytological techniques. The point of contact between homologous chromosomes called chiasmata can be visualized under light microscope by giemsa staining of meiocytes at metaphase I. The frequency of chiasmata between two physical landmarks among the total number of meiocytes analyzed is used to calculate the recombination rate between the landmarks (Rasmussen and Holm 1984). Later, development of fluorescent immunostaining techniques improved the efficiency of crossover detection. In this method, proteins involved in crossover formation

pathway (such as DMC1, RAD51, MLH1) are identified in situ using fluorescent labelled antibodies against them and the fluorescent labelled foci per nucleus are quantified (Anderson et al. 1999; Froenicke et al. 2002; Capilla et al. 2014). These methods provide opportunities to directly detect and quantify crossover events in each meiocyte and thereby enable sex-specific and individual-specific crossover analysis. However, the major limitations are the lack of resolution, low throughput, and difficulty in applying it to non-model organisms. Even though it provides a cell and chromosome specific data, the crossover events can be localized into only megabase sized regions and therefore it does not provide any fine-scale genomic feature information.

1.4.6 DSB mapping

ChIP sequencing of proteins involved in double strand break (DSB) repair pathway is used to map genome wide DSB sites with high resolution. One of the most popular methods is detection of DMC1 associated single stranded DNA. In this method, the DMC1 protein that binds to single stranded DNA intermediate of DSB repair pathway is immunoprecipitated using an anti-DMC1 antibody. Unlike standard ChIP protocols, the immunoprecipitated sample is then enriched for single strand DNA (ssDNA) following a specialized kinetic enrichment method. The ssDNA fragments are then analyzed by high throughput sequencing. Regions in the genome where these sequenced fragments pile up are identified as DSB hotspots (Smagulova et al. 2011; Khil et al. 2012). Another method to map the DSB landscape is chromatin immunoprecipitation and sequencing of SPO11 oligonucleotides. After creating the double strand break, SPO11 is removed from the DNA along with a covalently attached short oligonucleotide of 12 to 36bp in size. In this method, the oligonucleotide attached to SPO11 is captured and sequenced. Mapping of these sequenced reads provides location of DSB formation across genome with nucleotide resolution (Neale et al. 2005; Pan et al. 2011; Lam et al. 2017).

In contrast to methods discussed earlier, these strategies identify the location of DSB sites – a key initial step in meiosis. However, since only a subset of DSBs eventually resolve as crossovers, genomic maps of DSBs provide an initial picture of where crossover may later form. Therefore, comparisons of DSB landscape and CO landscape provide insight as to how much the distribution of COs across the genome is predetermined and shaped by mechanisms controlling the genomic location of DSBs. Further, DSB mapping methods are valuable for studying properties of recombination initiation at fine scale.

1.5 Recombination rate variation

Despite being constrained by molecular and evolutionary factors, recombination is found to vary across taxa, species, populations, between sexes, among

individuals and even within the genome of an individual (Ptak et al. 2005; Winckler et al. 2005; Paigen et al. 2008; Kong et al. 2010; Dumont et al. 2011). Variation in recombination quantified in terms of crossover number (rate) and location (placement) are discussed in the following subsections 1.5.1 and 1.5.2 respectively.

1.5.1 Variation in the genome-wide recombination rate

Across eukaryotes, genome-wide recombination rate (cM/Mb) varies more than 10-fold, with substantially higher recombination rates in microorganisms and fungi compared to plants and animals. Even though the pattern is not clear yet, evidence suggests that genome size, haploid chromosome number, requirement for an obligatory crossover per chromosome pair, crossover interference etc. play a role in regulating crossover count per genome (Stapley et al. 2017). Among closely related species, subspecies, and even among populations of the same species, considerable variation in overall recombination rate is observed in various organisms. For example, despite low sequence divergence (<1%), difference in crossover rates of nearly 30% was observed for closely related house mouse subspecies (Dumont et al. 2011). In natural populations, inter sub-species crossover variation is often attributed to several ecological and environmental features. Chiasma frequency per bivalent is found to be associated with latitude and population density in orthopterans. In plants, higher chiasma frequency is reported in selfers compared to out crossers (reviewed in Stapley et al. (2017)).

One of the major variations observed within species is between sexes. Given its relevance for this thesis, sex specific variation is discussed in detail in section 1.5.3. Additionally, in many species including humans, cattle, mice, soay sheep, and *Drosophila*, overall recombination rate is reported to vary even among individuals of the same sex within a population (Kong et al. 2010; Ma et al. 2015; Johnston et al. 2016). Various pedigree studies have mapped recombination rate variation as a quantitative trait, and identified trans-acting genetic loci underlying the inter-individual variation, also suggesting heritability of genome-wide recombination rate. Genetic factors underlying the heritable variation are discussed in section 1.5.4.

Furthermore, condition dependent variation in overall recombination rate has also been reported in many organisms. For example, increased recombination rate observed with increased maternal age in *Drosophila* (Bridges 1927), human (Campbell et al. 2015), and in cattle (Wang et al. 2016). Another extrinsic factor reported to influence recombination rate is temperature. Even though extreme temperatures cause meiotic recombination to fail altogether, less extreme fluctuations are shown to influence genome-wide recombination rate in different species. Studies in *Drosophila* (Plough 1917; Smith 1936) and *Arabidopsis* (Lloyd et al. 2018) have reported a U-shaped response to temperature fluctuations. Lowest recombination rate is observed at the optimum temperature and both increase and

decrease in temperature from optimum increases recombination rate. The third extrinsic factor that has been shown to influence recombination rate is parasitic infections. In organisms such as *Drosophila* (Singh et al. 2015) and *Arabidopsis* (Kovalchuk et al. 2003) recombination rate is shown to increase with increased parasitic infection. The rationale being, increased genetic diversity among offspring provide more survival chance against rapidly evolving parasites (Salathe et al. 2009).

Even though underlying molecular and evolutionary factors shaping genome-wide recombination rate are not completely understood, the extensive variation is considered as a reflection of differential selection pressure and adaptive requirement of various species.

1.5.2 Variation in recombination landscape

In almost all studied organisms, recombination is found to be distributed nonrandomly across the genome. The non-uniformity of recombination rate across chromosomes was first identified by Dobzhansky. He noticed that the distance between genes located in the middle of the chromosome is larger cytologically than it appears in the genetic map and explained it as due to lower recombination rate at the center of the chromosome in *Drosophila* (Dobzhansky 1930). Many subsequent studies have later reported that centromeres and telomeres exert very strong cis effects on recombination. While centromeres tend to suppress recombination in their vicinity, telomeric regions have higher recombination (Nachman and Churchill 1996; Choo 1998; Akhunov et al. 2003; Roesti et al. 2013). However, a recent comparison analysis across different taxa found that, in plants and animals, especially in larger chromosomes, crossovers are reduced at the chromosome center irrespective of the centromere location (Haenel et al. 2018).

The broad-scale recombination landscape variation is attributed to overall structural features of the chromosome such as chromosome condensation, centromere position, and length of the chromosome. In addition, processes like the 'telomere bouquet' formation during early meiosis are thought to play a major role; aggregation of telomeric regions at the nuclear membrane facilitate homology search and crossover initiation (Bass et al. 2000). If early crossovers are defined at the chromosomal periphery, CO interference mechanisms then prevent another crossover nearby. As a result, the second crossover is pushed to the other end of the chromosome leaving the center free of COs. It appears that these effects are conserved across taxa. Therefore, despite variation in the overall recombination rate, the recombination landscape at broad-scale (megabase scale) is found to be highly conserved among individuals within species and closely related species (Hassold et al. 2009; Garcia-Cruz et al. 2011) in a sex dependent manner. Sexes differ considerably in their recombination landscape even at the chromosome level. It is thought that differences in the above-mentioned features such as overall

chromosome packaging and CO interference may underlie the broad-scale variation. Discussed later in section 1.5.3

The fine-scale recombination landscape in most organisms is found to be heterogeneous with presence of highly active recombination 'hotspots' interspaced by recombination suppressed 'coldspots' (Jeffreys et al. 1998; McVean et al. 2004; Paigen et al. 2008; Singhal et al. 2015). Recombination hotspots are defined as small genomic regions typically 1-2kb in size with several fold higher recombination rate than surroundings (Lichten and Goldman 1995; Petes 2001). In various organisms recombination rates at individual hotspots show large variation ranging from 0.001 cM to 3 cM (Jeffreys et al. 2001; Baudat and de Massy 2007; Paigen et al. 2008; Paigen and Petkov 2010). Due to such hotspots, a large proportion of crossovers occurs in a very small proportion of the genome. Based on a recently published high-resolution recombination study in humans, nearly 75% of crossovers occur within less than 2% of the genome (Halldorsson et al. 2019). This indicates that at fine-scale, crossover placement is well-regulated and preferentially targeted towards certain narrow hotspots. However, organisms also vary in the extent of hotspot usage. The extreme examples include *Drosophila* (Smukowski Heil et al. 2015) and *C.elegans* (Kaur and Rockman 2014) in which recombination does not involve any fine-scale hotspots. Instead, a uniform distribution of crossovers is observed within highly recombining broadscale domains.

In some organisms such as yeast and birds fine-scale hotspots of recombination are found to be highly conserved over evolutionary time scale. Different species of Saccharomyces clade *S.cerevisiae* and *S.kudriavzevii* with over 15 million years of divergence share 81% of recombination hotspots with high conservation of hotspot strength (Lam and Keeney 2015). Similarly, between zebra finch and long tailed finch with 3 million years of divergence 73% hotspot sharing is observed (Singhal et al. 2015). On the other hand, in mammals including humans and mice, the fine-scale recombination landscape is rapidly evolving with limited hotspot sharing among closely related species. For example, humans and chimpanzees with ~5 million years of divergence show complete lack of hotspot sharing (Auton et al. 2012). This contrast in fine-scale recombination landscape conservation is attributed to the mechanisms underlying crossover regulation. In yeast and birds, recombination is enriched at open chromatin regions such as transcription start sites (TSS) or CpG islands. Evolutionary conservation of such functional features is thought to be the reason for their hotspot stability (Lam and Keeney 2015). Whereas in mammals, a trans-acting protein PRDM9 is shown to define crossover location (Baudat et al. 2010). PRDM9 is a zinc finger protein with histone methyl transferase activity. PRDM9 binds to its target motif and creates H3K4me3 and H3K36me3 marks at the nearest histone (Powers et al. 2016). Though the exact mechanism is yet to be understood, it has been proposed that these PRDM9 mediated histone methylation marks are targeted by the

recombination initiation machinery. In humans and mice, PRDM9 target motifs and histone methylation marks are found to be associated with almost all double strand break hotspots, and they are located away from functionally active TSSs (Brick et al. 2012; Pratto et al. 2014). PRDM9 is shown to be fast evolving (Oliver et al. 2009) and as a consequence, the PRDM9 directed recombination landscape is also rapidly updated.

These observations suggest that the fine-scale landscape of recombination is under high selection pressure. Various organisms use different mechanisms for the regulation. It is important to acknowledge the scale of recombination rate variation across the genome and understand them differently, because underlying determinants and selective forces that shape the variation at different scales are different.

1.5.3 Variation in recombination between sexes

Within species, considerable variation in recombination rate and landscape is observed between sexes. Sexual dimorphism in recombination rate, called 'heterochiasmy', is observed more or less in almost all organisms. With extreme instances of complete absence of recombination in *Drosophila* males (Morgan 1912) and *Bombyx* females (Tanaka 1914). However, when both sexes recombine, higher recombination rate is observed in females in many of the organisms studied till date (e.g., humans, mice, dogs, pigs and *Arabidopsis*) (Mikawa et al. 1999; Neff et al. 1999; Lynn et al. 2004; Drouaud et al. 2007; Paigen et al. 2008; Cox et al. 2009). However, birds, cattle, and sheep show the opposite pattern (Ma et al. 2015; Johnston et al. 2016; Smeds et al. 2016). Among fish, heterochiasmy biased towards high male recombination occurs nearly as frequently across taxa as heterochiasmy biased towards females (Brandvain and Coop 2012). These evolutionary viewpoints suggest that mechanistically heterochiasmy can evolve in either direction.

Regarding the crossover distribution, in most of organisms, irrespective of the direction of heterochiasmy, female crossovers are found to be distributed widely across the chromosome compared to male crossovers (Kong et al. 2002; Paigen et al. 2008; Smeds et al. 2016). Even though the underlying mechanism is not completely understood, potential factors that are thought to underlie heterochiasmy are listed below.

- A difference in the strength of CO interference, due to chromatin condensation: During prophase I of meiosis, female chromatids are observed to be less compacted than male chromatids (Gruhn et al. 2013). In mice and humans (organisms with female biased heterochiasmy) CO interference noticeably differs between sexes when measured in terms of physical distance (bp). Whereas, it is nearly the same when measured in microns along the chromosomal axis. (Tease and Hulten 2004; de Boer et al. 2006; Petkov et al.

2007). This suggest that CO interference acting at the same level in axis length leads to more crossover events in females with longer axis. However, this feature has to be explored in more organisms to generalize the observation.

- Difference in duration of meiosis: Even though fundamental steps in male and female meiosis are the same, their timing markedly differs (Morelli and Cohen 2005). For example: prophase I of *C.elegans* meiosis takes 20-24 hours in spermatocytes, whereas it takes about 54-60 hours in oocytes (Jaramillo-Lambert et al. 2007). Similar timing difference is also observed in mammals with considerably longer meiotic prophase arrest in females. Long meiotic arrest may benefit from increased chiasma, if it stabilizes chromosomes across the metaphase plate.
- Differences in genome methylation: A recent study in mice shows that in the early stages of meiosis DNA methylation is absent in females in contrast to males. It has been suggested that DNA methylation may change binding site preferences of recombination regulating DNA binding proteins such as PRDM9. This could lead to differential hotspot usage between sexes and consequently sex-specific landscape of recombination (Brick et al. 2018).
- Haploid selection: An evolutionary hypothesis says that male (sperm) and female (egg) gametes might experience different selection pressure in the haploid state and the sex that experiences strongest selection may recombine less (Lenormand and Dutheil 2005). In animals, females lack a haploid phase altogether as meiosis is completed only with fertilization. Whereas male gametes spend a longer time in haploid state. As a result, they may encounter high selection pressure and may benefit from reduced recombination.

Even though the evolutionary cost and benefits of heterochiasmy is yet to be clearly understood, it has been proposed that increased pericentromeric recombination in females may provide a strategy to prevent segregation of traits linked to the centromere that is preferentially selected for oocytes in contrast to polar bodies (Haig and Grafen 1991; Johnston et al. 2017). Furthermore, differences in the sex-specific recombination landscape may also confer a species the ability to both shuffle or maintain linkage. These alternative strategies may be favored when populations of a species experience heterogeneity in levels of gene flow from divergently adapted environments. This possibility is further tested and discussed in the chapter 2 of this thesis.

1.5.4 Genetic determinants of recombination rate variation

In many organisms, recombination rate is shown to be a heritable trait with multiple genes contributing towards its regulation; a polygenic trait (Fledel-Alon et al. 2011; Ma et al. 2015; Johnston et al. 2016; Dumont 2017). Genome-wide association studies (GWAS) or quantitative trait loci mapping (QTL mapping) approaches in natural populations of few species have identified several major and

minor loci affecting inter-individual variation in recombination rate. These trans acting genetic factors regulate either overall genome-wide recombination rate or recombination placement/hotspot usage with or without sex specific effect. For example, one of the major recombination modifier genes identified in mammals is PRDM9 (Baudat et al. 2010; Berg et al. 2010). PRDM9 shapes the overall landscape by differential usage of recombination hotspots. Individuals with different PRDM9 variants use different sets of hotspots (Brick et al. 2012). Besides, it has also been reported that a PRDM9 variant influences genome-wide recombination rate in human males (Kong et al. 2014) and in cattle (Ma et al. 2015). Another repeatedly mapped gene that influences genome-wide recombination rate is RNF212. In mammals RNF212 is shown to stabilize meiosis specific recombination factors in a dosage sensitive manner and thereby regulate CO formation (Reynolds et al. 2013). Another protein called HEI10 acts antagonistically with RNF212 and together they determine the genome-wide recombination rate by regulating the CO-NCO pathway (Qiao et al. 2014; Ziolkowski et al. 2017).

Table 1.1 summarizes major genetic loci repeatedly found to be associated with recombination variation in different organisms. Identification of multiple loci with varying effect sizes confirms that recombination is indeed a complex trait. These loci might be the direct targets of natural selection. Evolutionary importance of such recombination modifiers are further discussed in section 1.6. The similarities and differences in genetic architecture among organisms demand association studies in more species to elucidate the genetic underpinning of recombination rate and landscape variation.

Table 1.1: Genomic regions associated with recombination in various organisms

Genes/regions	Organisms	Associated feature	Sex effect	References
PRDM9	Human, mice, cattle	CO landscape	Both sexes	(Baudat et al. 2010; Sandor et al. 2012; Ma et al. 2015)
PRDM9	Human,	GWRR	Only in males	(Kong et al. 2014)
	Cattle	GWRR	Both sexes	(Ma et al. 2015)
RNF212	Human	GWRR	Sexually antagonistic effect	(Kong et al. 2008)
	Soay sheep	GWRR	Only in females	(Johnston et al. 2016)
	Lacaune sheep	GWRR	Reported in males	(Petit et al. 2017)
	Cattle	GWRR	Reported in males	(Sandor et al. 2012)
	Mice	GWRR	No sex-specific effect reported	(Reynolds et al. 2013)
HEI10	Human, mice, Arabidopsis	GWRR	No sex-specific effect reported	(Kong et al. 2008; Qiao et al. 2014; Ziolkowski et al. 2017)
	Lacaune sheep	GWRR	Reported in males	(Petit et al. 2017)
17q21.31 (inversion)	Human	GWRR	Only in females	(Stefansson et al. 2005)
HIM5, REC1	*C.elegans*	CO landscape	No sex-specific effect reported	(Chung et al. 2015)
CPLX1	Human, cattle	GWRR	Both sexes	(Kong et al. 2014; Ma et al. 2015)
	Soay sheep	GWRR	Only in females	(Johnston et al. 2016)
MSH4	Human, cattle	GWRR	Only in females	(Kong et al. 2014; Ma et al. 2015)
REC8	Cattle, Soay sheep,	GWRR	Both sexes	(Ma et al. 2015; Johnston et al. 2016)

*GWRR: Genome-wide recombination rate

1.6 Evolutionary costs and benefits of recombination

Evolutionary theory predicts that meiotic recombination plays an important role in adaptation because it is a key source of genetic variation. Recombination shuffles existing allelic combinations and produce novel ones, which become the raw material for natural selection (Otto 2009). However, the effect of recombination on adaptation can be different in various situations.

In the 1930's, Fisher (Fisher 1930) and Muller (Muller 1932) predicted that in a population in which new and favorable mutations occurring at many different loci, favorable mutations which arise in different individuals can ultimately be combined into the same genome by recombination (Felsenstein 1974). On the other hand, in the absence of recombination, favorable mutations that arise in different individuals have to compete with each other for fixation. Irreversible accumulation of mutations in a genome of asexually reproducing organisms is known as 'muller's ratchet'. Recombination helps to avoid that from occurring in sexually reproducing organisms (Muller 1964). Later, Robertson and Hill proposed that, in finite populations, linkage between sites under selection could reduce the overall effectiveness of selection. This is known as 'Hill-Robertson effect' (Hill and Robertson 1966). Therefore, recombination can speed up fixation of beneficial allele by breaking the linkage between it and a nearby deleterious mutation.

Recombination can also be maladaptive in some situations. For example, when divergent natural selection along an environmental gradient creates differently adapted populations living without geographic barriers, recombination during sexual reproduction could homogenize the two populations' locally adapted genomes and prevent speciation. Recombination could shuffle combination of locally adapted alleles and produce maladaptive combination which confer low fitness. Mathematical models in infinitely large populations with high gene flow show that recombination suppression is beneficial for local adaptation and divergence (Charlesworth and Charlesworth 1979). A lack of recombination will enable transmission of locally adapted loci as a linked adaptive "cassette".

In short, when recombination removes linkage between beneficial alleles at two linked adaptive loci, it can be costly (Fisher 1930). Whereas recombination might be favored if it releases a beneficial allele from linkage to a deleterious one (Hill and Robertson 1966; Felsenstein 1974). This two-sided effect of recombination in different scenarios suggests that, recombination rate and landscape are possibly evolving under strong selection pressures.

Two mechanisms have been proposed that could affect recombination rate: chromosomal rearrangements and recombination modifiers (Nei 1967). Chromosomal rearrangements such as inversions facilitate adaptive divergence and speciation by suppressing recombination (Kirkpatrick and Barton 2006).

Recombination is mediated by the sequence match between homologous chromosomes. An inversion therefore prevents recombination due to lack of sequence homology. When an inversion appears in a region carrying only selected (locally adapted) alleles, this new chromosome outcompetes other chromosomes that have mixed combinations of selected and deleterious alleles and rises rapidly in the population (Kirkpatrick and Barton 2006). In addition, inversions (and chromosome rearrangements in general) have the advantages that they are completely linked to the loci under selection and only suppress recombination in heterozygotes, maintaining the benefits of recombination within each subpopulation (Kirkpatrick and Barton 2006).

While inversions are predicted to be important in divergence at specific regions of the genome that are under selection, recombination modifiers more generally affect variation in recombination rate and placement. A recombination modifier is a genetic locus that regulates the rate of recombination between other loci (Nei 1967). A modifier locus can be linked or unlinked to the target loci and could have an effect on the local or global recombination rate and landscape. Recombination modifiers could increase or decrease recombination rate or change the recombination landscape in a way that confers fitness advantage. Therefore, modifiers provide a flexible tool for recombination rate evolution (reviewed in Ortiz-Barrientos et al. (2016)). Studies in natural populations provide empirical support for recombination regulation by both chromosomal rearrangement and modifier loci (Noor et al. 2001; Kong et al. 2002; Joron et al. 2011). A list of empirically identified recombination modifiers and its mode of action are discussed in section 1.5.4.

In addition to its role in facilitating adaptation, recombination plays a major role in the broader realm of genome evolution. Recombination is mutagenic (Arbeithuber et al. 2015); therefore, the recombination landscape in any genome may have large influence on the distribution of genetic variation across that genome. One of the major recombination associated mechanism that contribute towards genetic variation is the GC biased mismatch repair (GC biased gene conversion) which leads to higher GC content at regions of higher recombination. This mutagenic feature of recombination influences allele frequencies within a population (Duret and Galtier 2009). Therefore, variation in overall rate and placement of recombination also needs to be evaluated from the viewpoint of genome evolution.

While there is a strong theoretic framework discussing potential evolutionary costs and benefits of recombination, empirical studies testing the theoretical predictions are limited. This major gap between theoretical work and empirical studies are mainly attributed to the challenges of testing these hypotheses in evolutionary model systems. On the other hand, while empirical studies in various organisms have reported substantial variation in recombination rate and landscape at different levels, studies addressing the evolutionary

implications of such variation are also limited. Therefore, the study presented in this thesis, exploits uniqueness of an evolutionary model organism, threespine stickleback fish and lays a strong foundation towards closing that knowledge gap. In the following sections I introduce our model system and discuss the importance of studying recombination in this species and how it can be used to test the theoretical predictions.

1.7 Threespine stickleback fish

Here we choose an excellent evolutionary model organism threespine stickleback fish (*Gasterosteus aculeatus*) to investigate the role of meiotic recombination in driving adaptive divergence. Threespine sticklebacks are a teleost fish species widely distributed across the northern hemisphere. In many of its natural habitats sticklebacks exist as species pairs (marine anadromous-freshwater morphs, benthic-limnetic morphs, lake-river morphs) which are genetically, morphologically and behaviorally distinct populations living in sympatry (McKinnon and Rundle 2002). Following the latest Pleistocene glacial retreat from the Northern Hemisphere (approximately 10000-20000 years ago), ancestral marine sticklebacks colonized newly formed freshwater lakes and rivers. This led to formation of several ecologically adapted freshwater ecotypes (Bell and Foster 1994). Genetic studies have shown that, at different geographic locations freshwater sticklebacks have emerged independently and repeatedly from marine populations suggesting an example of parallel evolution. (Withler and Mcphail 1985; Taylor and McPhail 1999). Therefore, these species pairs provide great opportunity to study mechanisms underlying adaptation and speciation.

Adaptation to varying environmental conditions resulted in ecotypes with extensive differences in body size, shape, color, armor plate, spines, and including differences in several behavioral traits. A distinguishing feature of divergent stickleback ecotypes is the bony armor comprising three dorsal spines and a pelvic spine (variable), bony lateral plates (highly variable) and large and complex pelvic girdle compared to other teleost fishes (variable). Together these features provide survival advantage from vertebrate predation (Bell and Foster 1994). However, in different environmental conditions each of these features have either adaptive benefits or costs. Therefore, ecotypes have accumulated changes for better adaptation. Major features that distinguish marine sticklebacks from freshwater sticklebacks are larger body size, silver coloration in contrast to green-brown dorsal coloration in freshwater, presence of complete lateral armor plate in contrast to low number or complete lack of armor plates in freshwater, and presence of pelvic spines. Loss of armor plates and pelvic reduction in freshwater fishes are attributed to higher cost of mineralizing bones in the low calcium environment and difference in predator pressure (Reimchen 1980; Bell et al. 1993; Spence et al. 2012; Spence et al. 2013). Despite these differences marine and freshwater sticklebacks can hybridize and produce viable offspring. Therefore, this system

provides unique opportunity to understand the environmental and genetic interactions that facilitate adaptation and speciation. An example of anadromous marine (referred to as marine in this thesis) - freshwater species pairs in a natural population (River Tyne in Scotland) is shown in Figure 1.3.

Figure 1.3: Stickleback marine-freshwater species pairs in River Tyne. Geographic map of a natural stickleback habitat in River Tyne in Scotland is shown. Freshwater fishes inhabit upstream freshwater habitat whereas anadromous marine sticklebacks are found at the river opening close to the sea. Water salinity gradient is shown from blue to red where blue represent freshwater salinity and red represent near marine salinity. Example images of marine and freshwater sticklebacks collected from this site are also shown. Figure courtesy: Felicity Jones.

In the post genomic era, enormous efforts have been put into understanding genetic and genomic features underlying stickleback adaptive divergence. QTL mapping studies have identified genetic loci controlling major adaptive traits including armor plate difference (a causal mutation in the *Eda* locus) and pelvic spine reduction (deletions in the enhancer region of *Pitx1* gene) (Colosimo et al. 2004; Cresko et al. 2004; Colosimo et al. 2005; Shapiro et al. 2006; Chan et al. 2010). Moreover, a high-quality reference genome assembly was made (Jones et al. 2012) and which was further improved based on genome-wide linkage maps (Roesti et al. 2013; Glazer et al. 2015). More importantly, a comprehensive study by whole genome sequencing of 21 individuals (including both marine and freshwater) from various global populations enabled identification of genome-wide set of loci consistently divergent (divergent in parallel) between marine and freshwater ecotypes (Jones et al. 2012). Since high rates of gene flow rapidly homogenise neutral polymorphisms between ecotypes, the maintenance of parallel divergence at these loci between marine freshwater sticklebacks from across the Northern Hemisphere suggest strong divergent selection acts on the divergent alleles at these loci and that the divergent haplotypes confer a fitness benefit (have strong adaptive value) in their respective environments. The high-resolution map of 'adaptive loci' (Figure 1.4) suggests that stickleback adaptation is polygenic with nearly 242 loci that includes three large inversions. The dispersed but non-random distribution implies uneven linkage between these adaptive loci. This linked pattern together with clusters of adaptive loci captured within inversions suggest that recombination might be playing an important role in stickleback adaptive divergence.

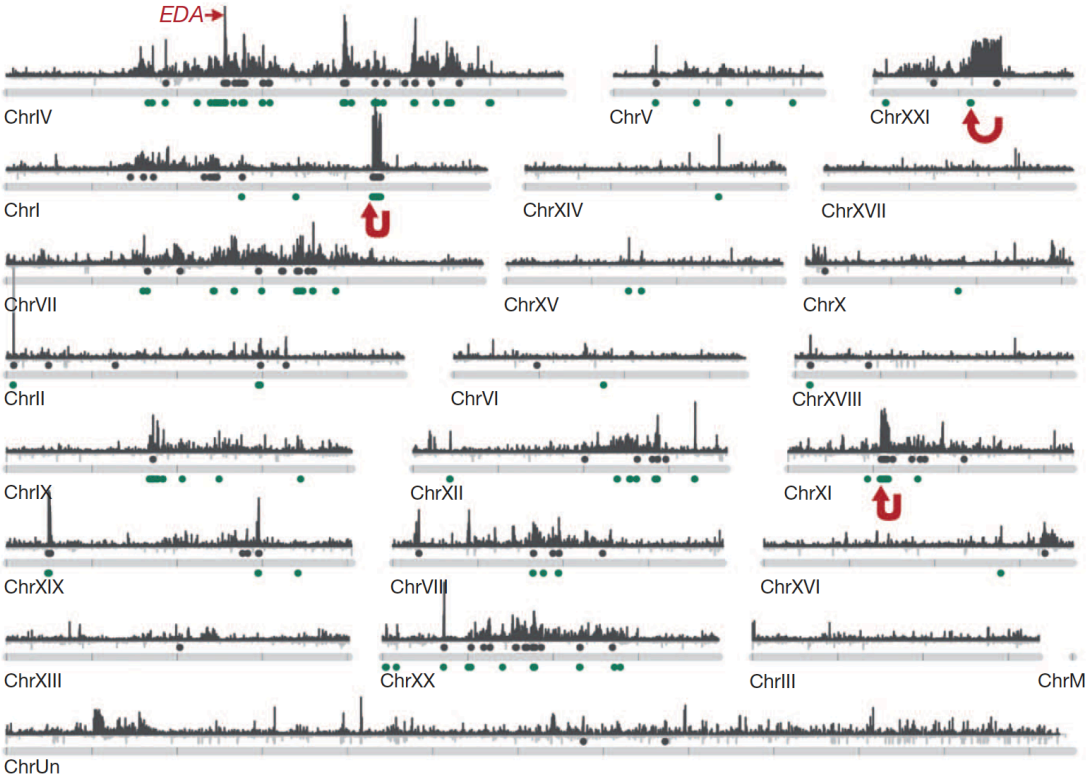


Figure 1.4: Genome-wide distribution of marine-freshwater parallel divergence loci. Marine-freshwater divergent regions detected are shown as grey peaks with grey points above chromosomes indicating regions of significant marine–freshwater divergence. Continuous regions of marine-freshwater divergence on chromosome XI, XXI, and I correspond to inversions (red arrows). – Adapted from Jones et al. (2012)

For example, freshwater individuals carry freshwater adapted versions of alleles in adaptive loci whereas marine individuals carry their marine adapted alleles. These locally adapted allelic combinations are under strong divergent selection pressure. However, since reproductive isolation between stickleback ecotypes is incomplete, many natural populations of sticklebacks experience migration and gene flow from nearby but divergently adapted population ecotypes. For example, throughout the Northern Hemisphere, contact zones between marine and freshwater sticklebacks can be found where they breed in sympatry with appreciable hybridization. Sticklebacks are therefore an excellent case study of adaptive divergence in the face of gene flow. In such a scenario, evolutionary theory predicts that locally adapted alleles may exist in tight linkage either by physical closeness or by low recombination; thereby prevent formation of maladaptive combination of alleles and disruptive effect of geneflow (Navarro and Barton 2003; Burger and Akerman 2011; Yeaman and Whitlock 2011). Therefore, this model system presents a unique opportunity to empirically study the role of recombination in adaptive divergence and test predictions from evolutionary theory.

1.7.1 Previous studies on stickleback recombination

A handful of previous studies have looked into stickleback recombination landscape in different populations. Roesti et al. generated a high-density genetic map (Roesti et al. 2013) by screening 1872 SNPs in 282 F2 individuals from an ecologically diverged lake-stream population pairs. They reported a periphery biased recombination landscape in sticklebacks with a genome-wide recombination rate of 3.11 cM/Mb. In addition, a positive correlation with genetic diversity within population, and GC content across genome was reported, suggesting a role of recombination in genome evolution. Later, Samuk et al., used this genetic map and statistically examined the theoretical prediction of recombination suppression in regions of high divergence (Samuk et al. 2017). They additionally compiled a global genomic data set of more than 1300 individuals from 52 different populations. Different stickleback populations presented different evolutionary scenarios such as populations experiencing divergent selection with gene flow, divergent selection alone, gene flow alone, or neither. With this comprehensive data set they find that in a scenario with divergent selection and geneflow, regions of higher adaptive divergence fall in areas of low recombination. However, divergent selection or gene flow alone displayed a lesser effect. Hence support the theoretical prediction that selection and gene flow interact to promote divergence in low recombining regions.

Sardell et al. produced first high-resolution sex-specific recombination map of sticklebacks using pedigree sequencing method (Sardell et al. 2018). They directly detected crossover events from 15 nuclear families (2 parents and 2 offspring per family), made by interspecies cross between *G. aculeatus* (threespine sticklebacks) and *G. nipponicus.* (Japan sea sticklebacks). They reported 1.64 times higher recombination rate in females compared to males. They also reported striking difference in crossover distribution between sexes with male crossovers clustered at the ends of the chromosomes and female crossover more evenly distributed across chromosomes. However, the sex-specific difference reported in their study could be mixed with interspecies difference in recombination. Even though they identified broad (100 kb and 10 kb) hotspots across the genome, no significant association is observed with any gene regions. Furthermore, a strong negative correlation is observed between recombination rate and population divergence between the study populations.

Most recently, Shanfelter et al. used a linkage disequilibrium-based approach to examine fine-scale recombination rate variation across stickleback genome of a freshwater and marine population (Shanfelter et al. 2019). Even though such population level studies estimate recombination rate based on historical events, it is an excellent method to identify population specific hotspots and coldspots of recombination. They identified nearly 4000 narrow hotspots and report that only ~15% of them are shared between populations. This suggest a

highly divergent landscape between closely related stickleback populations. Furthermore, they report some enrichment of hotspots to transcription starts sites (~29% of hotspots within 3 kb of TSS) and weak association with PRDM9 binding motifs, though sticklebacks does not have a functional copy of PRDM9. These results led them to the speculation of a novel mechanism for targeting recombination hotspots at fine-scale.

1.8 Overview of the thesis

The overall objective of this study is to understand how molecular mechanisms and natural selection shape and constrain recombination during adaptive divergence in natural populations. In this thesis, I present a comprehensive, empirical study of the recombination landscape in the threespine stickleback fish. Recent improvements in genome sequencing technology and bioinformatic algorithms enabled us to use the power of genomics to investigate fine-scale recombination landscape variation in a natural population. This study provides novel characterization of high-resolution, sex-specific crossover events in marine and freshwater ecotypes and their hybrids by screening large numbers of meiotic products (i.e., offspring). We also constructed the first map of DSB landscapes in sticklebacks. Furthermore, we established a novel cost- and time-effective method to quantify individual recombination rate and landscape variation at high-resolution, applicable to any organism. Our method enables future large-scale QTL mapping studies to further our understanding of recombination modifiers in sticklebacks and other natural populations.

In chapter 2, I construct high-resolution map of recombination crossovers per individual for 36 fish and identify recombination hotspots and coldspots in the stickleback genome. The 36 individuals represent marine, freshwater and hybrid ecotypes in equal numbers of males and females. This study improves upon existing knowledge of the stickleback recombination landscape by providing the first high-resolution, ecotype- and sex-specific map of contemporary crossover events in this adaptively diverging species. With this detailed map, I compare the magnitude of recombination variation across the genome, between sexes, between ecotypes and among individuals. Furthermore, I test for associations between the fine-scale crossover landscape and various genomic features attributed to stickleback adaptation. Specifically, I characterize the association with loci underlying parallel adaptive divergence, regions of freshwater-marine divergence in the study population, and chromosomal inversions. These comprehensive analyses provide estimates of recombination variation at different levels and I discuss its possible implications in a natural population under high selection pressure.

In chapter 3, I investigate various genomic features associated with recombination landscapes to gain insights about molecular regulators of

recombination and to understand how does the recombination landscape influence genome evolution. Here I complement the high-resolution crossover data from chapter 2 with a map of double strand break landscape in the stickleback genome, which we produced by ChIP sequencing of a meiotic recombination protein. Using crossover and DSB data, I investigate the genomic and epigenetic features associated with stickleback sex-specific recombination landscape. This provide insights into modes of recombination regulation in natural populations.

In chapter 4, I present a novel method we developed to make individualized recombination maps from pooled gamete sequencing. This approach enables to construct high-resolution recombination maps for organisms that are difficult to breed in laboratory conditions, which is especially useful for natural populations. The relative ease of making recombination maps per individual with this method allows QTL mapping of recombination rate variation as a trait to find recombination modifiers in the genome. This chapter was published as a research article in the Nature Communications journal in September 2019.

In chapter 5, I conclude with a discussion of the novel findings of this thesis and how they further our existing understanding of recombination rate variation, its molecular regulators, and evolutionary implications. Using the results presented in this thesis as a foundation, I also discuss the future steps required to further our understanding of adaptive importance of recombination variation.

Chapter 2

Fine-scale recombination landscape in threespine sticklebacks using nuclear family sequencing

2.1 Abstract

Recombination is a fundamental molecular mechanism generating genetic diversity and is essential for proper completion of meiosis. Therefore, it has direct influence on organismal fitness. Recombination shuffles existing allelic combinations and produce novel combinations influencing the patterns and efficacy of natural selection. Interestingly, large amount of variation in its rate and placement is detected across different taxa. Empirical studies on natural populations are required to understand the adaptive value of recombination rate variation. Therefore, in this study, we set out to empirically quantify recombination rate variation at different levels (among ecotypes, sexes, individuals and within the genome) in a natural population of adaptively diverging threespine stickleback fish. Using whole genome sequencing of 18 nuclear families (2 parents and ~94 offspring per family), we directly detected recombination crossovers in 36 individual fishes with a median crossover resolution of 3.8 kb. When individuals were compared, a major degree of variation in total recombination rate was observed between sexes. Females recombine nearly 1.76 times more than males and their crossovers are distributed widely across each chromosome. Whereas male crossovers occur predominantly near the chromosome ends. When compared between ecotypes, our empirical data shows reduced recombination in hybrids compared to pure forms. With regards to its distribution, we find that, stickleback recombination landscape is highly heterogenous across genome with several fine-scale hotspots of recombination. However, they are fewer in number and weaker in strength compared to reported mammalian hotspots. By intersecting with high-resolution adaptive loci, we find that overall recombination rate is lower within adaptive loci and at regions of higher population divergence. Adaptive loci tend to fall in regions of low recombination suggesting maintenance of linkage among adaptive alleles is important during adaptive divergence with ongoing gene-flow.

2.2 Introduction

The adaptive advantage of sexual reproduction lies in its ability to combine genetic variants originated in different individuals of a species. Sexually reproducing organisms undergo a specialized cell division called meiosis to generate haploid gametes. During meiosis, homologous recombination shuffles the paternally and maternally derived genetic material causing a novel genetic composition to be passed on to each offspring. Thus, homologous recombination during meiosis is one of the major sources of genetic diversity. It can directly influence the efficiency of natural selection by making good or breaking bad allelic combinations. Therefore, it is an important process especially in the context of adaptation to changing environments.

In this regard, there are several theoretical studies predicting the role of recombination in different evolutionary scenarios. Theory predicts that, under varying environmental conditions species often benefit from increased recombination rate (Hill and Robertson 1966; Felsenstein 1974). Recombination can break the linkage between a deleterious allele and a beneficial allele and thereby facilitate fixation of beneficial allele while purging the deleterious one. Whereas in a rather stable environment, or when individuals from adaptively diverged populations hybridize, recombination suppression between well adapted loci is preferred (Kirkpatrick and Barton 2006). Even though there are population genetic studies providing evidence for some of these predictions (Charlesworth and Charlesworth 1975; Gray and Goddard 2012; McGaugh et al. 2012; Castellano et al. 2016) empirical studies examining global picture of recombination landscape in natural populations under high selection pressure are still limited. At the same time, empirical studies in various organisms report large amount of variation in recombination rate and placement at different levels. Among species, populations, sexes, individuals, and even within genome (Coop et al. 2008; Paigen et al. 2008; Kong et al. 2010; Stapley et al. 2017). However, we are yet to clearly understand the genetic architecture and functional consequences of these variation. Theoretical studies have mostly emphasized on cost or benefits of recombination under different scenarios. Whereas there hasn't been much predictions about how much variation we might expect in a given scenario, and what would be the fitness consequences of such variation.

Empirical studies on various organisms so far have reported several different patterns of recombination rate variation. Within species, most pronounced variation is seen between sexes. A phenomenon known as heterochiasmy (Lenormand and Dutheil 2005). Males and females in many species are shown to have marked differences in average recombination rate and their recombination landscapes (Broman et al. 1998; Shifman S 2006; Sardell et al. 2018). The extent of difference is substantial such that in some species, one sex lack recombination altogether (Lenormand and Dutheil 2005). Within species,

recombination landscape on a broad scale is mostly conserved among individuals of same sex but there is substantial variation in the fine-scale recombination landscape (Paigen et al. 2008; Kong et al. 2010). In many species the fine-scale landscape variation is a result of localized hot and coldspots of recombination (Myers et al. 2005; Mancera et al. 2008; Paigen et al. 2008; Choi and Henderson 2015). Though there are exceptions such as *Drosophila* (Manzano-Winkler et al. 2013) and *C.elegans* (Kaur and Rockman 2014) lacking specific hotspots of recombination yet possess heterogeneity in fine-scale recombination landscape.

Most of our knowledge about the genomic architecture and molecular regulators of recombination rate variation comes from studies of various laboratory model organisms and few natural populations. In various organisms including human, mice, *Arabidopsis*, maize, cattle, and sheep a few candidate genes have been identified (RNF212, REC8, HEI10 etc.) that influence genome wide recombination rate variation (Bauer et al. 2013; Kong et al. 2014; Qiao et al. 2014; Ma et al. 2015; Johnston et al. 2016; Petit et al. 2017; Ziolkowski et al. 2017). This suggests a heritable genetic basis for recombination rate variation. Identification of such genetic factors are important because they probably are the direct targets of natural selection for recombination modification and consequently enable rapid adaptation. However, more empirical studies in natural populations are required in order to elucidate the interplay between evolutionary and mechanistic factors that shape recombination landscape variation and to understand its fitness consequences.

Towards that aim, here we report a detailed multidimensional empirical study of recombination in a natural population of adaptively diverging threespine stickleback fish. Over the last couple of decades, sticklebacks have been developed as an excellent model organism to study the genomic basis of adaptive divergence and speciation. As a result, a good quality genome assembly and a high-resolution map of freshwater-marine divergent loci are available (Jones et al. 2012). The high-resolution map of adaptive loci enables us to investigate how adaptive variants are shuffled and thereby to understand the impact of recombination on polygenic adaptation. Therefore, by capitalizing on these resources as well as making use of large clutch size and high density of genome variants, we produced fine-scale individualized map of contemporary crossover events.

2.3 Experimental design

In order to construct individualized high-resolution whole genome recombination maps, we used the pedigree analysis approach. Pedigree analysis enables the direct detection of crossover events in one generation. The two major requirements for high-resolution map construction are a large number of meiotic products per generation and a high-density of genomic variants differing between the parents. The stickleback fish stands as a great model organism in this regard as it has large

clutch size and a high number of heterozygous SNPs segregating in wild population. We used the following experimental design to maximally exploit the advantages of this model system (Figure 2.1).

Reproductively mature fishes were sampled from the freshwater-adapted upstream population and from the marine-adapted downstream population of River Tyne, Scotland and large clutch nuclear families were produced by in-vitro fertilization. 18 nuclear families (6 ♀MAR x ♂MAR families, 6 ♀FRESH x ♂FRESH families, 3 F1 hybrid families from ♀MAR and ♂FRESH F1 hybrids and 3 reciprocal F1 hybrid families) consisting of two parents and nearly 94 offspring per family were then whole genome sequenced.

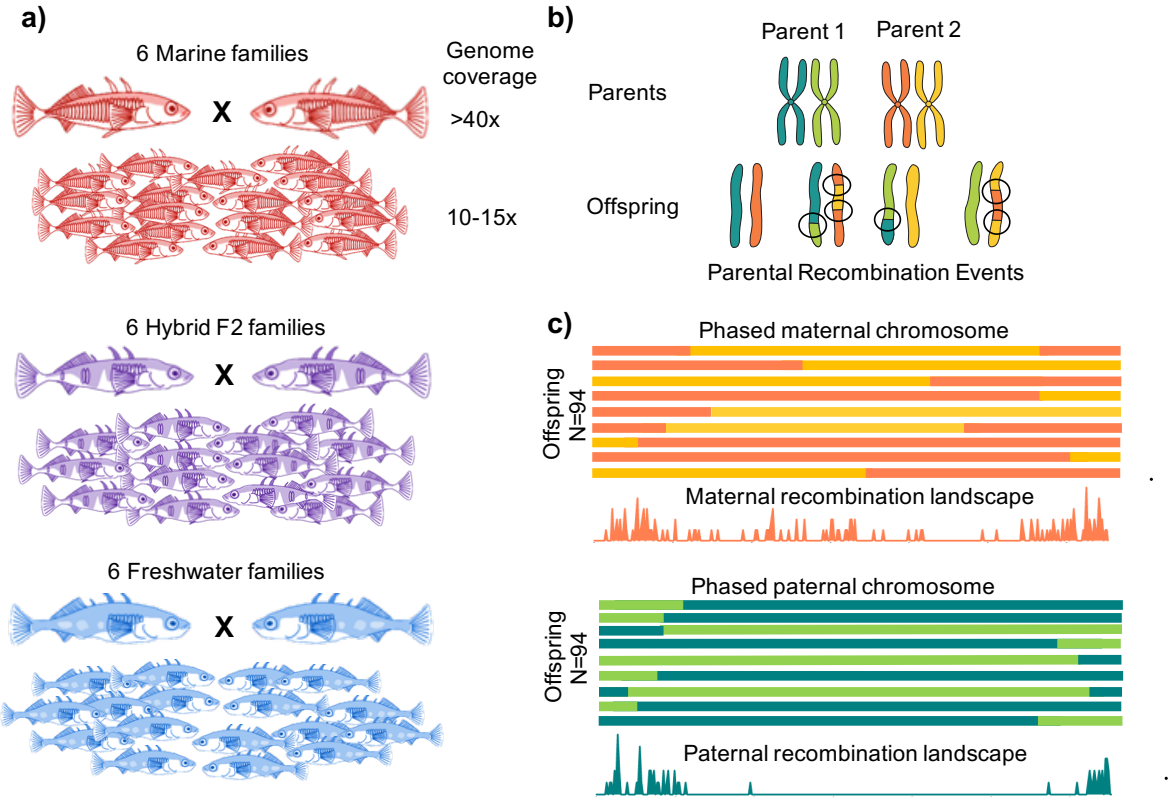


Figure 2.1: Individualized recombination map construction by nuclear family whole genome sequencing. (a) 18 large clutch nuclear families were produced by invitro fertilization of fishes collected from River Tyne. 6 ♀MAR x ♂MAR families (red), 6 ♀FRESH x ♂FRESH families (blue), 3 F1 hybrid families from ♀MAR x ♂FRESH F1 hybrids and 3 reciprocal F1 hybrid families (purple). Each family consist of two parents and nearly 94 offspring. (b) Within each nuclear family, recombined parental chromosomes are passed on to the offspring. Therefore, we can directly compare parental and offspring genomes and identify parental crossover events (highlighted with circles). (c) Within each nuclear family, paternal and maternal chromosomes are phased separately. Top panel shows a phased maternal chromosome of all offspring of a family. Maternal crossover events are detected as the switch between maternal haplotype A (orange) and haplotype B (yellow). Locations of those crossover events detected from all offspring were used to construct a high-resolution recombination map for the mother. Similarly, paternal chromosome phasing and inferred recombination map for the father are shown in the bottom panel (green).

Phased chromosomes of the parents were compared against the offspring to identify parental crossover events (shown as the switch between orange to yellow and light green to dark green in Figure 2.1c). Identified crossover events from all offspring of a parent were used to construct an individualized recombination map for each parent (Figure 2.1c). This design provides one generation fine-scale whole genome recombination map for 36 individuals (12 marine, 12 freshwater, and 12 hybrid individuals with equal representation of males and females).

Based on these high-resolution recombination crossover maps, in this chapter I discuss the general features of contemporary recombination landscape of sticklebacks, quantify recombination variation between sexes and ecotypes, and test theoretical prediction of recombination suppression in hybrids. Furthermore, I quantify recombination landscape variation across the genome at different scales and investigate the presence of hotspots and coldspots of recombination. Finally, I address the association between recombination landscape and stickleback marine-freshwater divergence.

2.4 Results

For all 18 families, parents were sequenced to 40x or more genome coverage and each offspring was sequenced to an average of 10-15x coverage. Offspring with extremely low coverage (<2x) were excluded from further analysis. Deep sequencing of large number of individuals per family enabled high density variant calling. More than 6.8 million variants were identified in each family. Following extensive and stringent filtering, an average of 348,735 informative SNPs were retained allowing us to distinguish between homologous chromosomes per parent in each family. Dad's informative SNPs only include SNPs those are heterozygous in dad at the same time homozygous in mum. Vice versa for mum's informative SNPs. For each family, cross details, post mapping read coverage, total number of informative SNPs (post filtering) and mean inter-SNP distance per parent are summarized in Table 2.1

The highly filtered informative SNP set was then phased using SHAPEIT (Delaneau et al. 2011) algorithm in combination with duoHMM (O'Connell et al. 2014). Pedigree information enabled accurate phasing of parents and offspring. Phased haplotypes of father-mother-offspring trios were used to identify parental crossover events. During recombination, double strand break repair results in reciprocal crossover events or short (100 bp to 2 kb) noncrossover gene conversion events (described in chapter 1 section 1.2). Often these gene conversion events are also associated with crossover events causing back and forth switching between haplotypes at the crossover location (complex crossovers). Even though the data is extensive enough to characterize crossover and non-crossover gene conversion events across the genome, for the scope of this thesis I have focused only on

Table 2.1: Family details and sequenced data overview

Family	Ecotype	Number of offspring	Dad coverage (x)	Mum coverage (x)	Offspring mean coverage (x) ± SD	Dad informative SNP count	Mum informative SNP count	Dad inter-SNP mean distance (bp)	Mum inter-SNP mean distance (bp)
X1	Freshwater	94	38.27	49.43	13.18 ± 3.50	259704	250250	1541	1591
X4	Freshwater	93	70.54	54.69	19.83 ± 6.21	385432	434153	1038	922
X268	Freshwater	92	236.14	188.56	14.11 ± 3.60	441892	282717	906	1406
X284	Freshwater	91	58.65	80.06	11.86 ± 2.56	320492	173594	1247	2301
X350	Freshwater	93	60.98	67.45	13.35 ± 2.44	285827	238058	1401	1679
X351	Freshwater	93	229.45	298.05	13.38 ± 2.68	350647	283794	1141	1410
X11	Marine	94	77.81	63.65	11.65 ± 2.17	385709	319840	1039	1252
X20	Marine	91	57.69	138.48	14.07 ± 5.48	325633	307571	1230	1300
X291	Marine	94	57.58	70.09	12.46 ± 2.97	465765	364930	860	1098
X294	Marine	94	119.55	117.69	14.11 ± 2.99	492870	410454	813	976
X295	Marine	93	35.04	44.24	13.24 ± 3.60	216217	111255	1851	3586
X296	Marine	93	56.15	68.44	13.62 ± 2.89	465613	246504	861	1625
X273	Hybrid	94	62.47	64.30	13.45 ± 2.71	461511	345952	868	1155
X274	Hybrid	94	87.67	72.32	13.63 ± 2.99	451754	312601	887	1282
X800	Hybrid	94	32.82	45.05	13.32 ± 2.84	269782	323189	1477	1239
X366	Hybrid	86	69.86	59.70	13.36 ± 3.68	474780	329376	844	1216
X389	Hybrid	93	62.58	69.91	14.63 ± 2.35	513093	385061	781	1041
X391	Hybrid	93	63.70	57.80	14.57 ± 2.68	501055	367392	798	1090

reciprocal crossover events. Reciprocal crossover (CO) events were identified as long (>50kb in size with ≥50 SNPs per haplotype) switches between parental haplotypes and crossover boundary was then defined with regions of uncertainty flanked by last SNP of the first haplotype and first SNP of the second haplotype as shown in Figure 2.2 offspring 1. In situations where two or more switches occurred within the minimum required block size (50 kb) of each other (complex crossovers), crossover boundaries were defined by the last SNP of first large phase block and first SNP of next large phase block (see offspring 2 in Figure 2.2). In such a scenario, crossovers have wider intervals of uncertainty. Further details of phasing and crossover identification are given in the Materials and methods section.

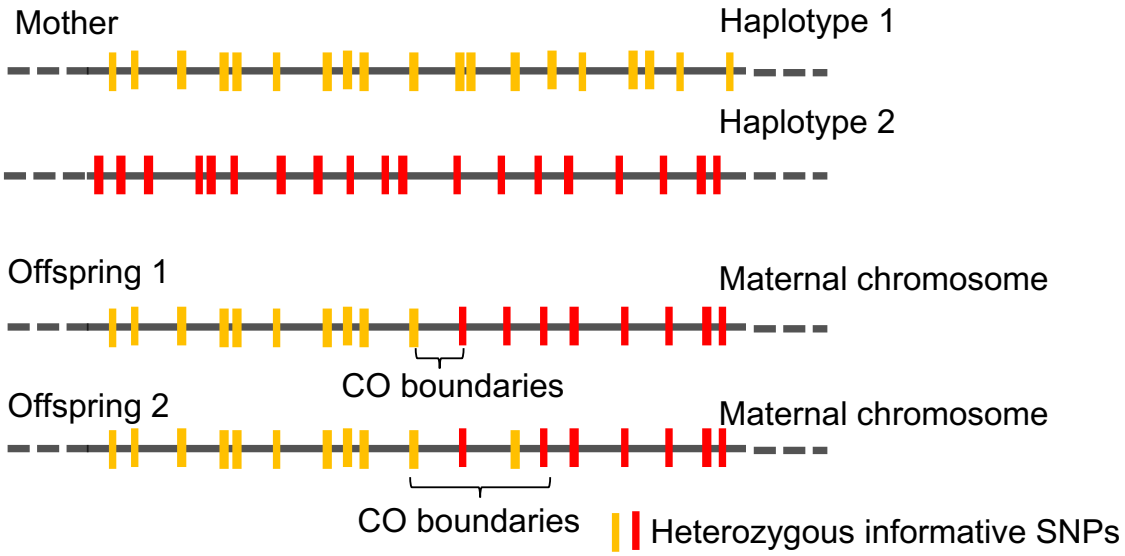


Figure 2.2: Schematic representation of crossover boundary definition. After phasing both parental and offspring genotype, parent of origin for each haplotype in the offspring have been identified. In this schematic representation, homologous chromosomes of a mother and maternal chromosome of two of her offspring are shown. The yellow and red bars represent heterozygous SNPs present in the mother. In offspring 1, crossover was defined with regions of uncertainty flanked by last SNP of the first haplotype and first SNP of the second haplotype. In scenario such as in offspring 2, where two or more switches occurred within the minimum required block size (50 kb) of each other, crossovers were defined by the last SNP of first large phase bock and first SNP of next large phase block.

Each offspring in a family corresponds to two meiotic products; one maternal and one paternal. From the 3338 meiotic products (1669 male and 1669 female meiosis) analyzed, a total of 49848 recombination crossovers were detected (18039 male and 31809 female crossovers) with a median resolution of 3845 bp (Figure 2.3a). Crossover resolution is the distance between left and right boundary SNPs of a crossover, and therefore it is dependent on the density of informative

SNPs in the parents. 74% of crossovers in our data set have resolution less than 10 kb. Median crossover resolution for each individual range between 2.5 kb and 10 kb (Figure 2.3b). Individuals with fewer informative SNPs and higher inter-SNP distance have in general lower crossover resolution.

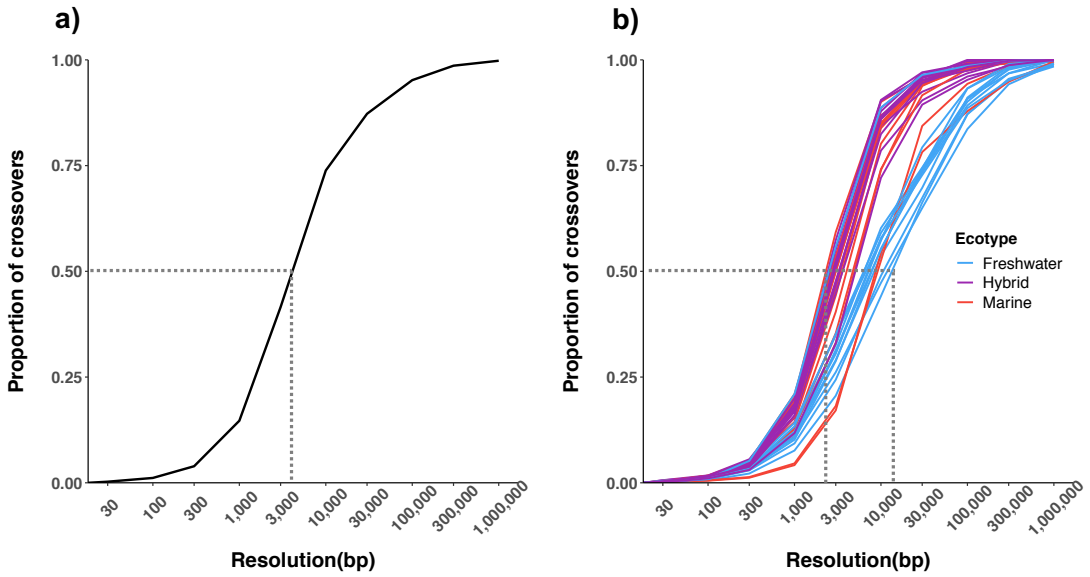


Figure 2.3 : Crossover resolution plot. Fraction of crossovers in the data set with the respective resolution are shown. (a) All individuals combined (b) Individuals plotted separately. Individuals are colored based on their ecotype. Blue: freshwater, red: marine, purple: hybrid. Median crossover resolution is marked with dotted lines.

In our data set, many freshwater individuals have fewer informative SNPs and as a consequence have lower crossover resolution (blue lines in Figure 2.3b). This is most likely due to freshwater populations having a smaller effective population size and therefore less heterozygosity than the marine populations. However, it is also possible that crossover associated gene conversion events (complex crossovers), or genotyping errors have contributed to lower crossover resolution in freshwater sticklebacks.

2.4.1 Whole genome crossover count per meiosis

Among individuals, the total number of crossover events across whole genome (21 chromosomes) per meiosis ranged from 10.14 to 22.57. When individuals were grouped based on their sex and ecotype, the highest degree of variation was observed between sexes (see Figure 2.4 a and b). We observed a significant difference in the mean number of crossovers per meiosis with females having nearly 1.76 times more crossover per meiosis than males (male mean number of crossovers per meiosis =10.81±0.09SE, N=18; female mean number of crossovers

per meiosis=19.06±0.34SE, N=18; $p=7.523x10^{-16}$, independent one tailed t-test, 19.17 degrees of freedom). In addition, variance in the number of crossovers per meiosis across individuals was higher for females than males (standard deviation of 1.45 and 0.37 respectively). This phenomenon of difference in recombination rate in one sex compared to the other, namely heterochiasmy, has been reported in various organisms including both plants and animals (Lenormand and Dutheil 2005) though the molecular mechanisms are not well understood.

When the numbers of crossover events were compared across ecotypes, a significant reduction in hybrid females was seen compared to pure forms (mean number of crossovers in hybrid females:18.169±0.4SE, N=6; pure marine or freshwater females :19.5±0.42, N= 12; p= 0.04148, Wilcoxon rank sum test, W=58) These results are consistent with theoretical predictions of recombination suppression in hybrids (Figure 2.4b). Initially by mathematical models (Charlesworth and Charlesworth 1979; Kirkpatrick and Barton 2006) and later with evidence from genetic and genomic studies (Noor et al. 2001; Wadsworth et al. 2015) it has been shown that under conditions of high gene flow in infinitely large populations, recombination suppression is beneficial for local adaptation and divergence. Here we report another empirical evidence from natural populations where reduced recombination rate is observed in hybrids with ongoing gene flow (Jones et al. 2006).

2.4.2 Genetic map length and recombination rate

Based on the observed number of crossovers, genetic map length was computed per individual. Genetic map length, represented in the units of centimorgan (cM) is essentially the number of whole genome crossover events in 100 meiosis. Mean sex averaged genetic map length in this data set is 1493 cM with mean map length of 1081 cM for males and 1906 cM for females. These numbers are in agreement with reported genetic map length in sticklebacks from previous studies (sex averaged genetic map length 1328 cM reported in Sardell et al. (2018); 1251 cM in Roesti et al. (2013); 1570 cM and 1963 cM in Glazer et al. (2015)). Our genetic map length value 1493 cM corresponds to a genome average recombination rate of 3.24 cM/Mb. However, mean recombination rate among chromosomes varies and the sex averaged value in this data set ranges from 2 cM/Mb to 5.9 cM/Mb. As we observed in the whole genome scale, recombination rate was consistently higher in females across all 21 chromosomes. The ratio of female to male recombination rate ranged between 1.09 (chrV) and 2.05 (chrXII). In general, both in males and females, recombination rate is inversely proportional to the physical size of the chromosome.

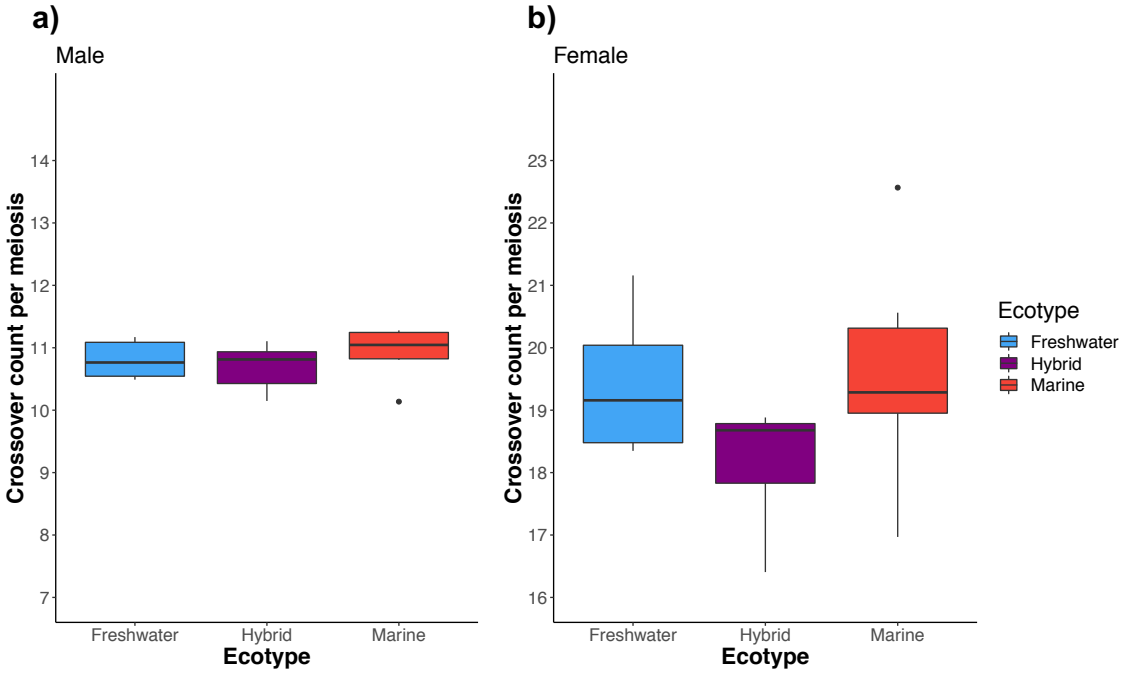


Figure 2.4: Whole genome crossover count per meiosis is higher in females compared to males. Among females, hybrids have fewer crossovers than pure forms. Crossover count per meiosis is plotted for (a) males and (b) females. Within males and females, individuals are further grouped based on their ecotype. Red: marine, blue: freshwater, Purple: hybrid.

2.4.3 Crossover count per chromosome

From 3338 meiotic products that were analyzed (1669 male and 1669 female meiosis), the number of crossover events per chromosome per meiosis ranged between 0-3 in males and 0-4 in females. It is assumed that for proper segregation of chromosomes during the first meiotic division, there has to be an obligatory crossover per chromosome (Petronczki et al. 2003). Nonetheless, 50% of the resultant gametes can be non-recombinant because, for mechanistic reasons one crossover per homolog pair (tetrad) is sufficient. Therefore, two chromatids in a homolog pair can successfully segregate without being involved in a crossing over. In agreement with that, in this data set, we observed several instances of chromosomes without any crossover. Among 21 chromosomes analyzed from 1669 male and female meiosis, 50.2% of the male meiotic products and 31.41% of the female meiotic products segregated with zero crossover events per chromosome. Higher instances of male gametes with zero crossover chromosomes is possibly a consequence of lower recombination rate in males.

Among male meiotic products only 1.7% of them had more than one crossover per chromosome. In females the incidence of multiple crossovers was twelve times higher with 20.6% meiotic products having two or more crossovers per chromosome. The median inter-crossover distance was 24 Mb in males but

only half of that, 12 Mb, in females. Both in males and in females, the incidence of multiple crossovers is depended on the physical size of the chromosome – a higher incidence of double crossovers was observed on large chromosomes (≥25 Mb) compared to small chromosomes (<25 Mb).

2.4.4 Distribution of crossovers across the genome

Next, we examined the distribution of crossover events across the genome. The distribution of all crossover events (except 3462 COs overlapping scaffold boundaries) across all 21 chromosomes is shown in Figure 2.5. The number of crossover events are plotted in 100 kb sliding windows. The vertical black dotted lines represent approximate centromere locations obtained by BLASTing centromere repeat sequence reported in Cech and Peichel (2015) against stickleback reference genome. Approximate centromere locations are identified in all but three chromosomes (chrII, chrIV and chrVIII). From this, we see that stickleback recombination follows a similar pattern as reported in most eukaryotic species studied to date: recombination rate is higher towards chromosome periphery compared to the center (Barton et al. 2008; Rockman and Kruglyak 2009; Roesti et al. 2013). Across each chromosome, the broad scale distribution of COs in our data is in agreement with previous recombination maps constructed for stickleback genome (Roesti et al. 2013; Glazer et al. 2015; Sardell et al. 2018). The pattern of increased COs towards the ends of the chromosomes is more pronounced in males than in females. In Figure 2.6 male and female crossover landscape is plotted in panel a and b respectively. More than 70% of male COs occur within first or last 15% of the chromosome whereas only 47% of female COs occur within that range. Albeit biased towards the chromosomal periphery, female COs occur more uniformly across the chromosome. In the 18 chromosomes for which approximate centromere location is known, (marked as dotted black vertical lines in Figure 2.6 a and b) in general, acrocentric chromosomes concentrate male recombination at the end of long arm except for chrXIX, the sex chromosome. In chrXIX, all observed male COs happened in the short arm that contains the pseudoautosomal region (PAR) previously identified by (Ross and Peichel 2008). Within this 3.8 Mb short arm, all male COs occurred within the first 2.67 Mb with a recombination rate of 15.10 cM/Mb. Whereas, in females, the recombination rate in PAR region is 9.20 cM/Mb and 2.6 cM/Mb recombination rate is observed in rest of the sex chromosome spanning more than 17.5 Mb outside the PAR region.

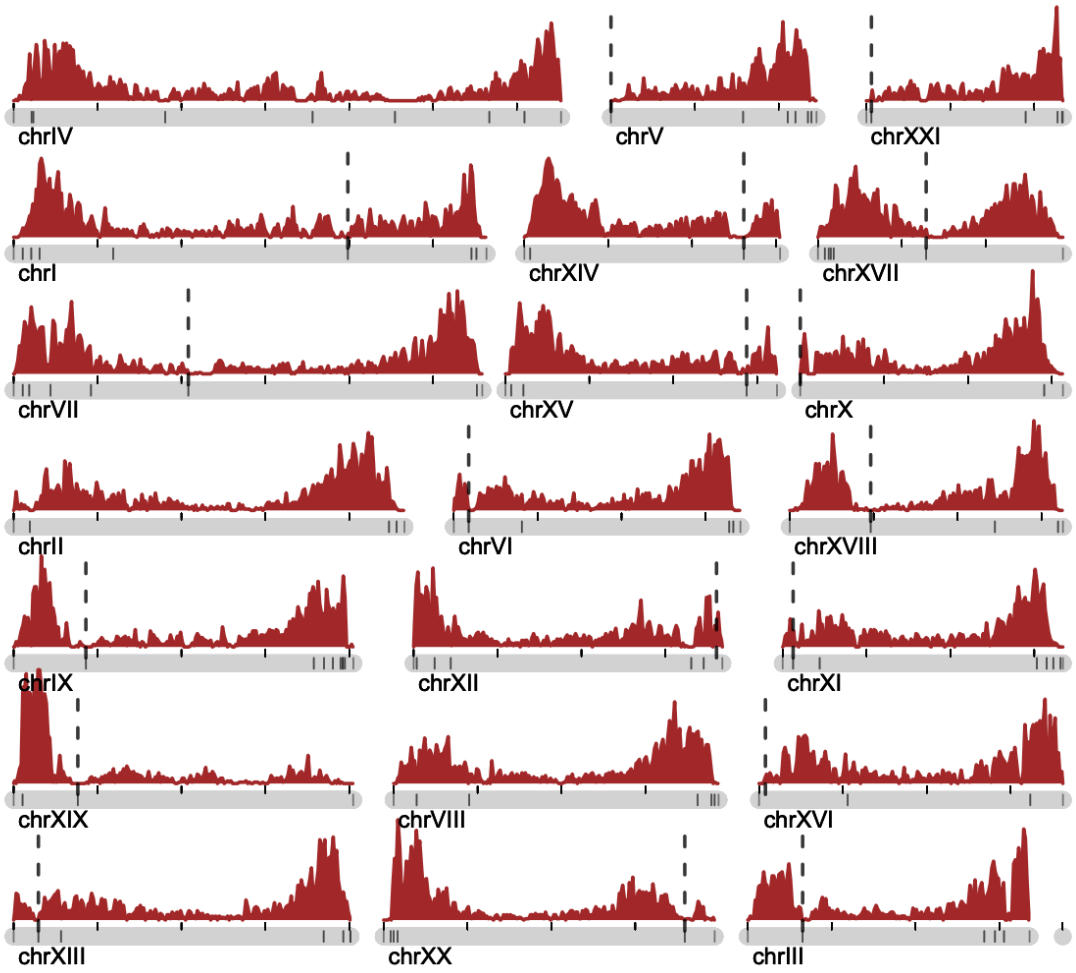


Figure 2.5: A genome-wide map of sex averaged crossovers inferred from pedigree sequencing shows that crossovers are concentrated on chromosome peripheries. Crossover events detected in all 36 individuals are plotted in 100 kb sliding windows across 21 chromosomes. Crossovers overlapping scaffold gap boundaries are excluded. Approximate position of centromere in all but three chromosomes (chrII, chrIV and, chrVIII) are plotted as vertical black dotted lines.

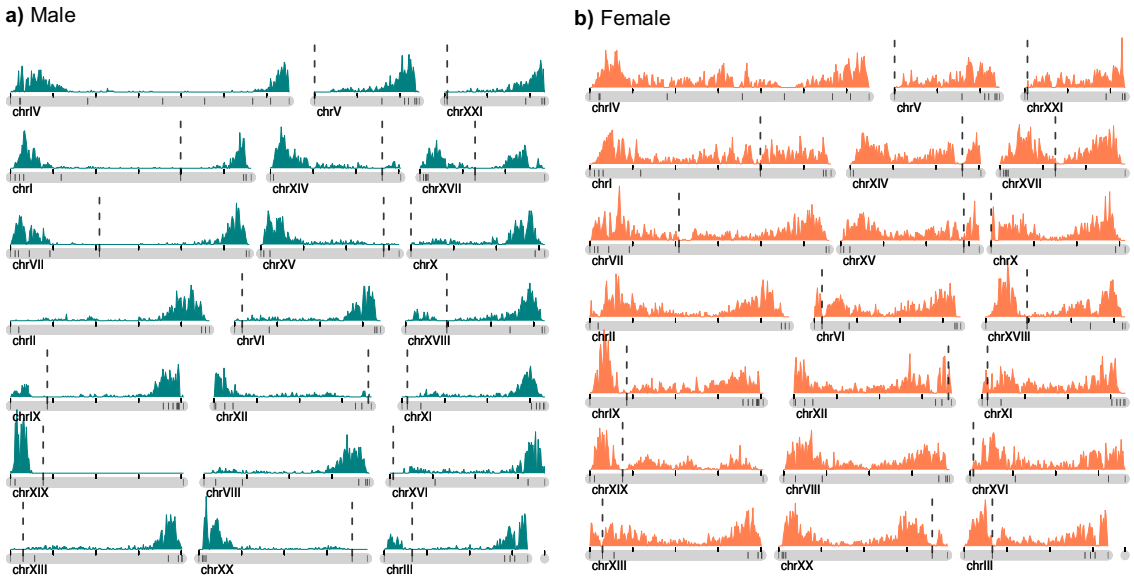


Figure 2.6: Sex specific distribution shows that male crossovers predominantly occur at chromosome ends, while female crossovers are more evenly distributed across each chromosome. Distribution of (a) male and (b) female crossover events are plotted in 100 kb sliding windows across all 21 chromosomes. Crossovers overlapping scaffold gap boundaries are excluded. Approximate position of centromere in all but three chromosomes (chrII, chrIV and, chrVIII) are plotted as vertical black dotted lines.

2.4.5 Heterogeneity in crossover distribution

Next, we examined the distribution of crossover events across the genome in order to understand the extent of variation at different scales. Already from Figure 2.5 and Figure 2.6 a and b, we know that stickleback recombination occurs nonuniformly across the genome. A large fraction of crossovers are located in the chromosomal periphery. In the whole data set, 80% of the crossovers occur in less than 35% of the genome. Using the Gini coefficient, a non-parametric measure of inequality in a distribution, we quantified the crossover count heterogeneity across the genome at different scales of resolution. The value of the Gini coefficient is bounded by 0 and 1, where 0 represents complete uniformity (crossovers are distributed uniformly across the genome) and 1 represents absolute inequality in distribution (e.g., 100% of crossovers occur in one location in the genome). Figure 2.7a shows the observed and null expectation Gini coefficient values at different scales of resolution. In each case, the null expectation Gini coefficient was estimated as the mean coefficient from 1000 random shuffles of crossover events across the genome. For all scales, standard deviation is <0.002. At all measured scales, the observed CO distribution shows higher heterogeneity than expected when assuming random distribution. The highest heterogeneity (largest magnitude Gini coefficient) in the observed data is found at fine-scale (5 kb) but a higher deviation from null expectation is seen at lower resolution scales (e.g., more than two-fold deviation of observed from expected is observed at resolution scales of 100 kb or above). Taken together these results suggest that recombination

crossover events in the stickleback genome are non-random, and therefore cluster at certain regions at broad-scale as well as fine-scale. It is possible that the observed non-random clustering of crossovers at different scales is due to regulation by different molecular and/or evolutionary mechanisms.

In order to compare heterogeneity in the stickleback recombination landscape with other organisms, the Gini coefficient was estimated based on published data sets for human (Kong et al. 2010), mouse (Paigen et al. 2008), yeast (Mancera et al. 2008), and *Drosophila* (Singh et al. 2013), and *C.elegans* (Kaur and Rockman 2014). In Figure 2.7b, the Gini coefficient estimated for different organisms are projected onto stickleback estimates at different scales. Since data sets differ in the scale of resolution applied, the estimated Gini coefficient is plotted against their respective scale. At 5 kb scale, the stickleback recombination landscape shows the highest heterogeneity with a Gini coefficient of 0.82. This value is higher than the moderate heterogeneity observed in *C.elegans, Drosophila* and yeast. However, when compared to mouse and human data at their respective scales, stickleback shows lower heterogeneity in the genomic distribution of recombination crossovers.

In this analysis, most comparable data sets (all generated by direct detection of crossover events) across species were used for the comparison. Though, some fundamental differences between different studies such as whole-genome vs small genomic intervals and sex averaged vs sex specific, inbred lab strains vs natural populations may influence the measure of heterogeneity. Therefore, this comparison and interpretation may suffer from such factors to an extent.

With the observation that the stickleback crossover distribution is highly heterogeneous across the genome at different scales, we next examined the extent of the variation in the whole data set and searched for hotspots of recombination. We also quantified variation according to sex, ecotype and also individually. The following section describes recombination landscape variation in these different categories at different scales.

2.4.6 Recombination rate variation and hotspots of recombination

2.4.6.1 Across the genome (Uncategorized data)

At 1 Mb scale, crossover count per interval ranged from 0 to 613. That corresponds to recombination rate between 0 and 18 cM/Mb. At the fine-scale (5 kb scale) crossover counts per interval ranged from 0 to 15 with recombination rate between 0 and 89.87 cM/Mb. This combined with our previous analysis of heterogeneity suggests that, there are crossover-enriched megabase-sized 'hot domains' as well as fine-scale 'hotspots' in the stickleback genome.

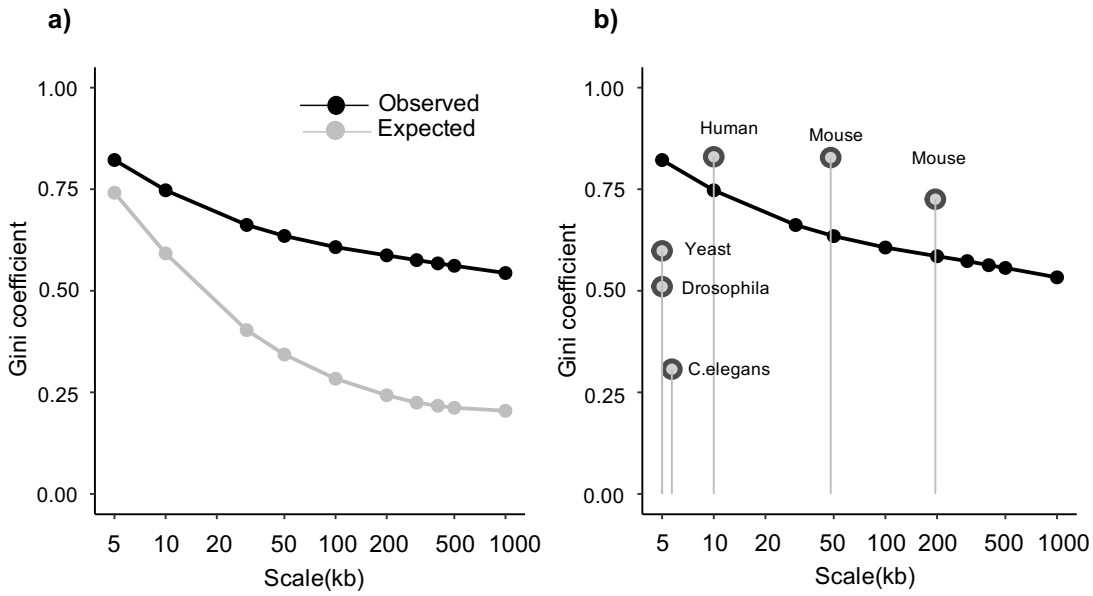


Figure 2.7: Gini coefficient reveals higher heterogeneity in the fine-scale crossover distribution. (a) Observed and expected Gini coefficient values at different scales ranging from 5 kb to 1 Mb are shown. At all tested scales, observed heterogeneity in crossover distribution is higher than null expectation assuming random crossover distribution. (b) Gini coefficient calculated for other organisms in their respective scale are projected onto stickleback Gini coefficient value at different scales (human (Kong et al. 2010), mouse (Paigen et al. 2008), yeast (Mancera et al. 2008), *Drosophila* (Singh et al. 2013), and *C.elegans* (Kaur and Rockman 2014). In Paigen et al. (2008), crossover count across mouse chr1 calculated at two different scales are used separately (see methods section for more details).

At different scales, 'hot' regions of recombination were identified as intervals containing multiple COs with a false discovery rate less than 0.05. Figure 2.8a shows the fraction of crossovers that occurred within hot bins identified at different scales. At a 1 Mb scale, 105 hot intervals were identified that spanned a total of 22.7% of the genome and contained nearly 60% of all crossovers. Most of such 1 Mb hot intervals appear to be contiguous, giving rise to large hot domains. The median size of such hot domains in our data set is 3 Mb and maximum size of 4 Mb. Subsequently in smaller scales, recombination enriched intervals were identified. We found that, most of those hot intervals fell within these megabase-sized hot domains. At 100 kb scale, 544 hot intervals were identified that span 11.8% of the genome but contain nearly 46% of all crossover events. This suggests that, 100 kb sized domains capture well the broad scale recombination landscape of sticklebacks. Whereas at the fine-scale (5 kb), 432 hot intervals identified are found to be having 7% of all crossover events which span only 0.47% percentage of the genome. The distribution of those hotspots appears to be clustered at the ends of the chromosomes with a median inter-hotspot distance of 75 kb. 85.4% of 5 kb hotspots fall within first or last 15% of chromosomes. In Figure 2.8c, crossovers across chromosome IV in 5 kb sliding windows are shown. Intervals qualified as

hotspots are marked in red. The recombination rate within 5 kb hotspots, is on average 3.7 times more than that of the flanking region (Figure 2.8b) and ten times more than the genome-wide average.

In mammals, hotspot intensity ranges from 0.001 cM to 3 cM (Paigen and Petkov 2010). In our data, the 5 kb-scale hotspots ranged from 0.15 cM to 0.45 cM. Therefore, we can say that stickleback recombination hotspots at fine-scale are 'semi-hot' and at least six times weaker in intensity than mammalian hotspots. In terms of hotspot number, a similar one generation crossover detection study (Paigen et al. 2008) have estimated about 13,670 hotspots across the mouse genome. This suggest that, even after accounting for differences in genome size, sticklebacks have nearly five times fewer hotspots than mice do. However, the hotspots we identified here have to be considered as a minimum estimate, since we have only analyzed one generation of crossover events in less than 3500 meiotic products and in addition, we used a conservative approach to bioinformatically define hotspots (by only considering crossovers less than 10 kb resolution). Therefore, it is possible that there are more hotspots in the stickleback genome which are not detected in this study. Nevertheless, we find that sticklebacks possess rather heterogeneous (based on Gini coefficient) and punctate CO landscape (with presence of 5 kb hotspots) compared to *Drosophila* (Manzano-Winkler et al. 2013) and *C.elegans* (Kaur and Rockman 2014) which are reported to have no crossover hotspots at the 5 kb scale.

2.4.6.2 Between sexes

The genomic crossover landscape differs both in terms of number and distribution between male and female sticklebacks. In section 2.4.4, we have already seen that male crossovers are concentrated at the chromosomal periphery whereas female crossovers are more evenly distributed across the chromosome. The correlation of crossover counts between sexes across the genome is highest at the 1 Mb scale (Spearman's rank correlation ρ: 0.75). In contrast, at fine-scale resolution (5 kb) the correlation in crossover counts is as low as 0.18 (Spearman's rank correlation ρ). Sex-specific crossover distribution across chromosome IV at broad as well as fine-scale is shown in Figure 2.9.

Similar to the analysis carried out with uncategorized data, after grouping individuals based on sex, we identified bins of different size with clustering of crossover events more than expected by chance (hot intervals in males and females). At 1 Mb scale, 94 hot intervals were identified across the female genome which constituted 48% of COs within 20.4% of the genome. In males 75% of COs occurred within 80 hot intervals that spanned only 17.3% of the genome. Even at the broad scale (1 Mb), there were male-specific and female-specific hot bins. There were 29 male-specific hot bins and 43 female-specific hot bins and within which only a small fraction of COs from the other sex happens.

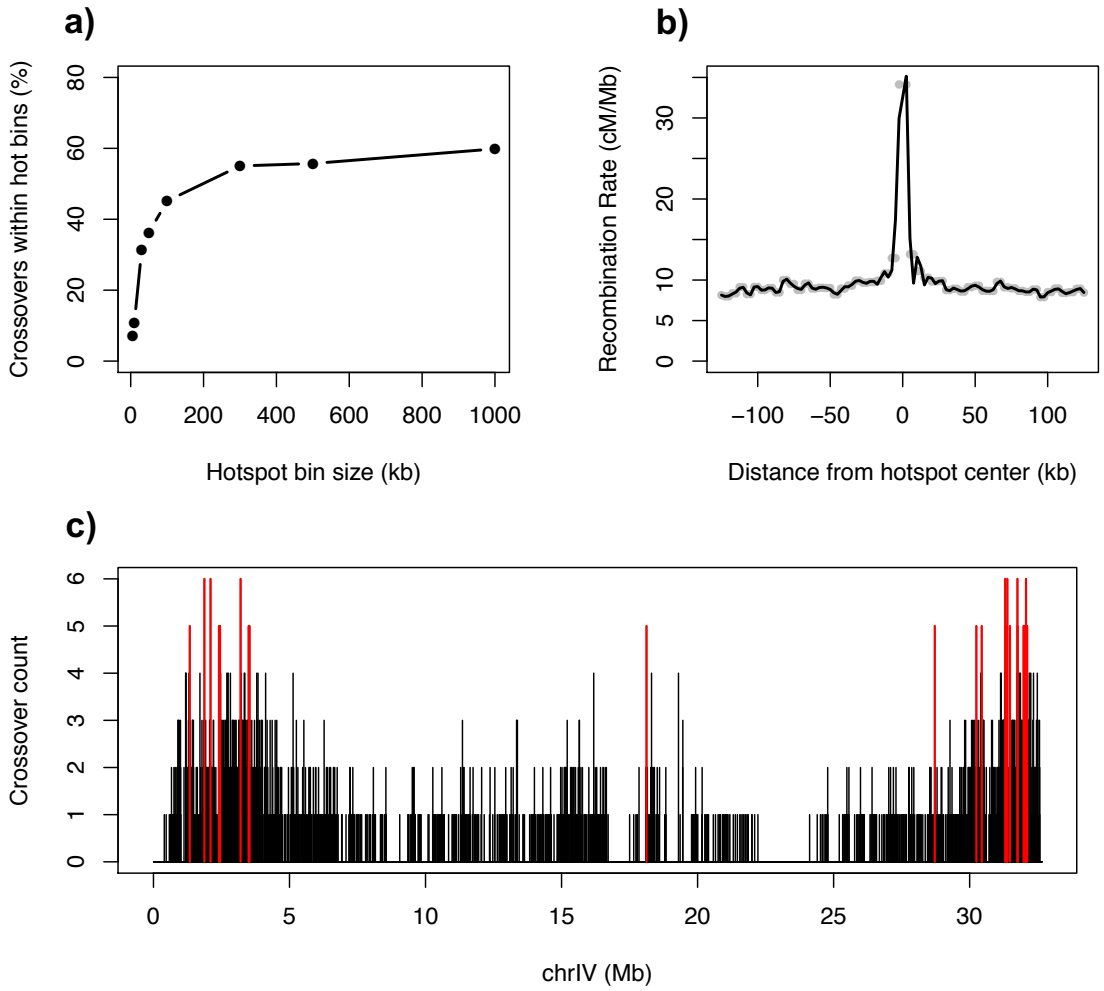


Figure 2.8: Semi-hot hotspots in the stickleback genome are clustered at chromosome ends. Across the genome, hot intervals of recombination are identified at different scales ranging from 5 kb to 1 Mb. (a) Percentage of crossovers within hot intervals identified at different scales are shown. (b) Recombination rate around fine-scale (5 kb) hotspot midpoints are shown. Recombination rate within hotspots are nearly 3.7 times more than that of the flanking region. (c) Crossovers across chromosome IV in 5 kb sliding windows are shown. Intervals qualified as hotspots (in this case intervals with 5 or more COs) are marked in red color.

At the fine-scale (5 kb), the male genome possesses 216 hotspots. Nearly 10% of male COs happen within these hotspots that span only 0.23% of the genome. On the other hand, only 2% of crossovers occur within 68 bins identified as hotspots across the female genome. This suggests that female hotspots at this scale are not only rare in number but also weaker in strength compared to male hotspots. In addition, we find that, at fine-scale, nearly 87% of female hotspots are female-specific where as 96% of male hotspots are male-specific indicating that, most of the fine-scale hotspots are unique to one sex and not shared. These exclusive hotspots hint at a possibility for sex-specific fine-scale recombination regulators in sticklebacks.

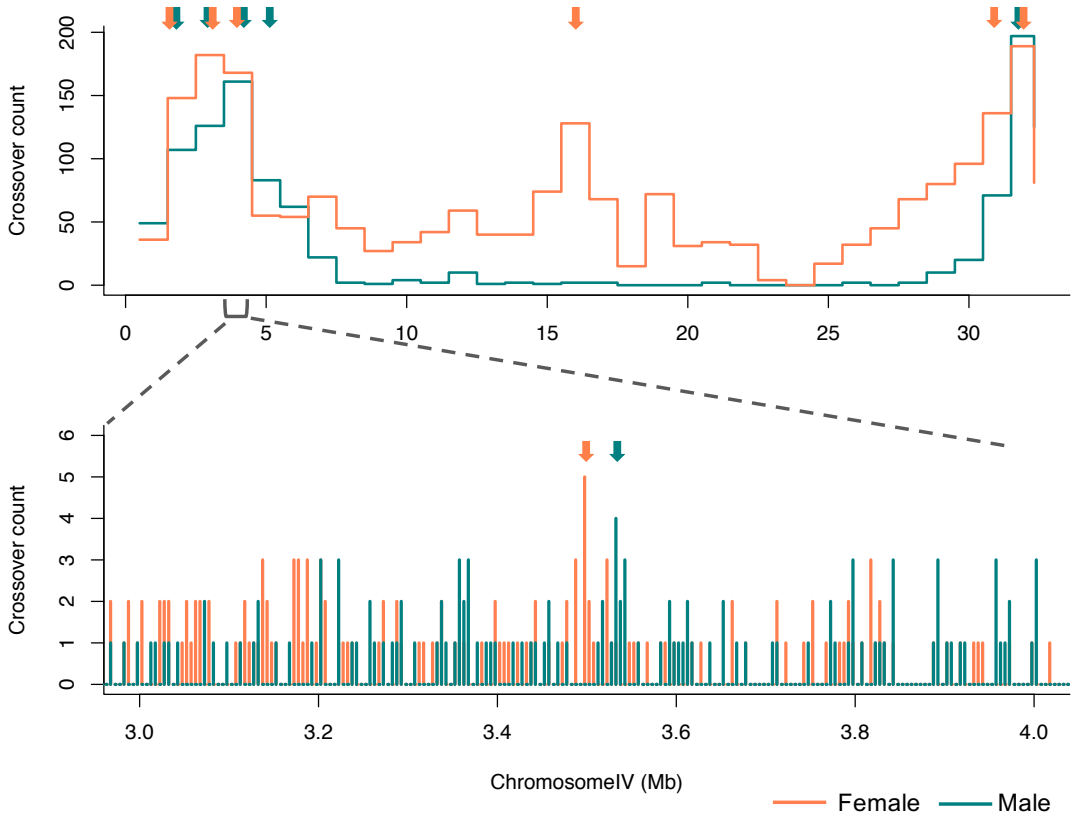


Figure 2.9: Sex-specific crossover landscape across chromosome IV shows difference in distribution over broad as well as fine-scale. Male (in green) and female (in orange) crossover counts in 1 Mb sliding windows (top panel) and in 5 kb sliding windows across a 1 Mb region (bottom panel) are shown. Windows identified as hotspots (intervals with more crossovers than expected by chance, FDR<0.05) are marked with an arrow on top. Green arrows mark male hotspots and orange arrows mark female hotspots.

The lack of high correlation between sexes even at broad-scale and the presence of sex-specific megabase-sized hot intervals suggest that in addition to fine-scale regulators, the molecular-genetic factors that influence recombination landscape at the chromosome level might act differently between sexes. Female-specific highly recombining regions, including hot intervals at the middle of the chromosomes indicates that, female recombination shuffles genes that are kept in linkage in males. This pronounced sexual dimorphism in recombination landscape probably has important evolutionary implications since they provide different choices for natural selection to act on. Further discussion about possible evolutionary implications of sexual dimorphism in recombination landscape is given in section 2.4.8 where we analyze sex-specific recombination in relation to stickleback marine-freshwater adaptive divergence loci.

2.4.6.3 Among ecotypes

Recombination rate at megabase scale is highly correlated among ecotypes in a sex depended manner. Figure 2.10 top panel shows ecotype-wise crossover counts across chromosome IV in 1 Mb sliding windows. Individuals are grouped based on their sex (panel a and b) and ecotype (overlaid). In the bottom panels, zoomed-in view across a 1 Mb region in which crossover counts in 5 kb sliding windows are shown. At the fine-scale, we can see higher variation in distribution of crossover events among ecotypes. The pairwise genome-wide correlation (Spearman's rank correlation ρ) between the ecotype-specific recombination map of males and females at broad and fine-scale are shown as a heatmap in Figure 2.10c. Correlation between female and male recombination map at 1 Mb scale is only 0.75 whereas the pairwise ecotype correlation coefficient at this scale within male recombination map and within female recombination map of different ecotypes is around 0.9. At the same time, at fine-scale, correlation among ecotypes is as low as 0.1. This indicates that recombination landscape at fine-scale is less conserved among ecotypes. A list of top 5 genomic regions (at 1 Mb and 5 kb scale) with biggest difference in recombination rate between marine and freshwater ecotypes is given in Appendix Table 3.

After grouping individuals based on their sex and ecotype, several intervals at different scales of resolution were identified as ecotype-wise hotspots of recombination. The hotspots reported here are genomic intervals with a greater number of crossovers than expected by chance in six individuals of each ecotype (around 550 meiotic products per ecotype). These hotspots are probably only a subset of all stickleback ecotype-wise hotspots due to the fact that we have only examined a single generation crossover events in a small sample size. Screening of more meiotic products per ecotype and validating hotspots using techniques such as sperm typing are required to identify more hotspots and measure their intensity. However, our study is unique in terms of analyzing contemporary crossover events in relatively large number of meiotic products in adaptively diverging ecotypes in a natural population. Compared to population level estimates of historical recombination events, our approach detects sex-specific hotspots in each ecotype with very few false positives.

Across all three ecotypes, compared to females, males have higher number of hot intervals and higher percentage of crossovers within hot intervals at every scale. However, at fine-scale (10 kb and 5 kb) hotspot sharing among ecotypes were extremely rare. Lack of ecotypes-wise hotspot sharing may suggest a possibility of ecotype specific features determining hotspots or it could just be due to higher inter-individual variation at fine-scale. Therefore, it is important to understand how much inter-individual variation is observed at different scales in this data set.

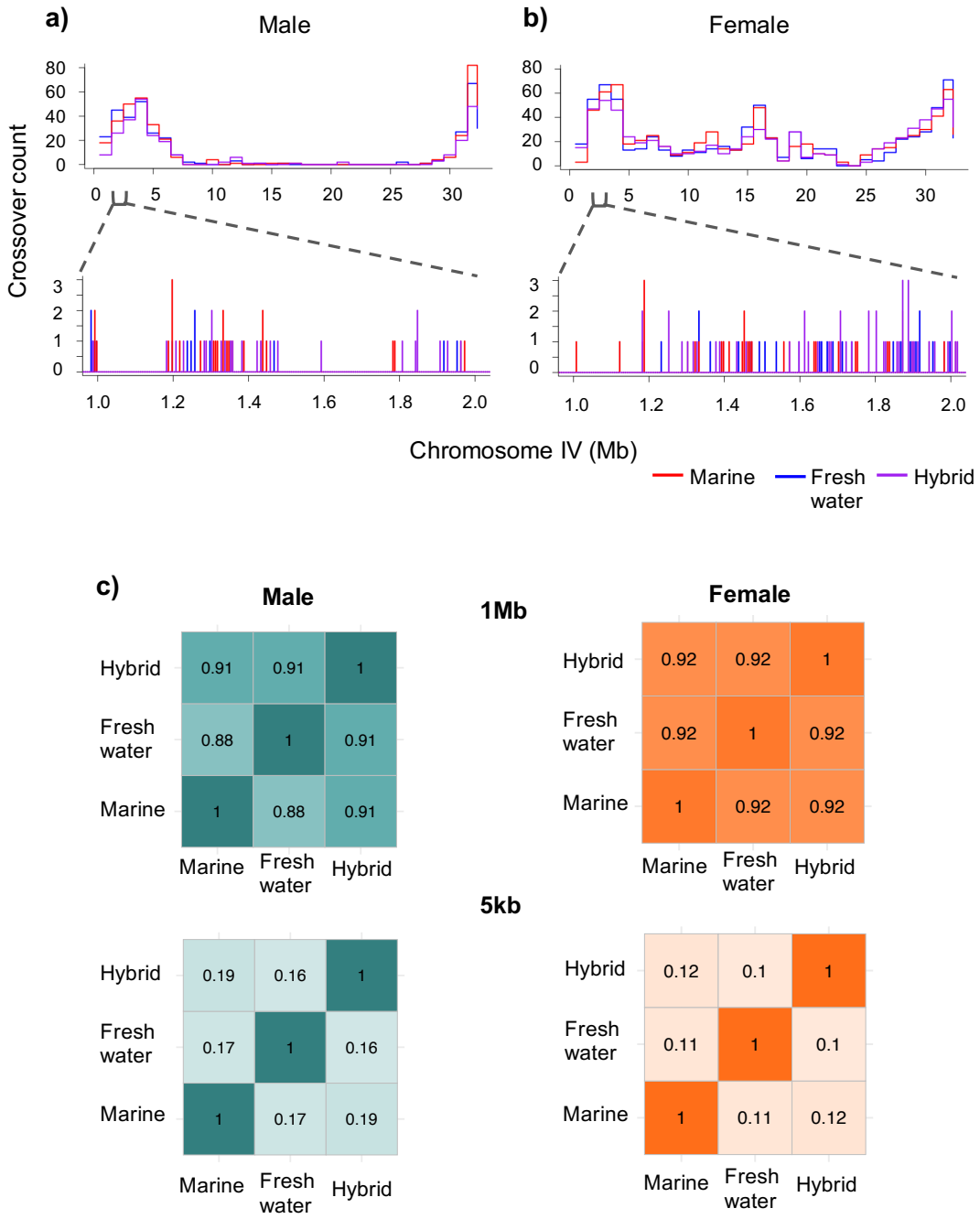


Figure 2.10: Recombination landscape among ecotypes are highly correlated at broad-scale in a sex dependent manner, whereas, low correlation is observed at fine-scale. After grouping individuals based on their sex and ecotype, crossover counts across chromosome IV in 1 Mb sliding windows are shown in top panel of (a) males and (b) females. Zoomed in view of a 1 Mb region in which crossover counts in 5 kb sliding windows are shown in their respective bottom panel. Ecotypes are overlaid. Red: marine, blue: freshwater, purple: hybrid. (c) Pair wise correlation coefficient (Spearman's rank correlation ρ) between ecotype recombination maps are shown as a heatmap. Left: male, right: female. Top panel shows correlation at 1 Mb scale. Bottom panel shows correlation at 5 kb scale.

2.4.6.1 Among individuals

Next, we quantified recombination rate variation among individuals. Correlation (Spearman's rank correlation ρ) between individual recombination maps at 1 Mb scale and at 5 kb scale are shown in Figure 2.11 a and b, respectively. At broad-scale, individuals are highly correlated with correlation values ranging from 0.33 to 0.81. Also, there is a clear grouping of individuals based on their sex. Correlation among males and among females were higher than male-female correlations. However, there was no obvious grouping of individuals based on ecotype. Despite this, there was a general trend for pure forms (freshwater or marine) to be grouped along with hybrid or its own kind more than the other pure form.

At 5 kb scale, inter-individual correlations dropped to 0.001-0.058 range suggesting high inter-individual variation in the recombination maps at this scale. However, the individuals still grouped based on their sex. No specific grouping of ecotypes was observed. This presents another line of evidence suggesting that sex specific factors regulate broad-scale and fine-scale recombination landscapes beyond the level of individual variation. In addition, at fine-scale, correlations among males were higher than that of among females, indicating possibility of more conserved fine-scale recombination regulators in males than in females. At least in this genome-wide correlation analysis no ecotype-specific effect was observed beyond the level of individual variation.

2.4.7 Recombination coldspots

Genomic regions where recombination is suppressed are called coldspots. Identifying such regions where recombination is shutdown is as important as identifying the regions with higher occurrence of crossovers (hotspots). Therefore, next we scanned through the data in search of coldspots of crossovers in the stickleback genome. In this study we operationally defined coldspots as contiguous regions that could accommodate at least five crossovers assuming a uniform distribution but contained zero crossover events in the observed data set. Coldspots were first identified from the data set as a whole, and then from sex categorized and ecotype categorized data sets separately.

In the whole data set, 499 distinct regions greater than 47 kb in size was identified as coldspots. Such coldspots spanned a total of 56.87 Mb (nearly 14% of the assembled genome). Largest coldspot identified was in the chromosome IV (chrIV:22,462,382-24,024,373) spanning nearly 1.5 megabase region with zero crossovers observed (Figure 2.12). However, when considering only males, a 17 Mb region spanning the sex determining region of the X chromosome (chrXIX) was observed to have zero crossovers. This is consistent with the fact that the sex determining (non-PAR) region in male Y chromosome lacks homology with its X chromosome homolog and therefore does not recombine.

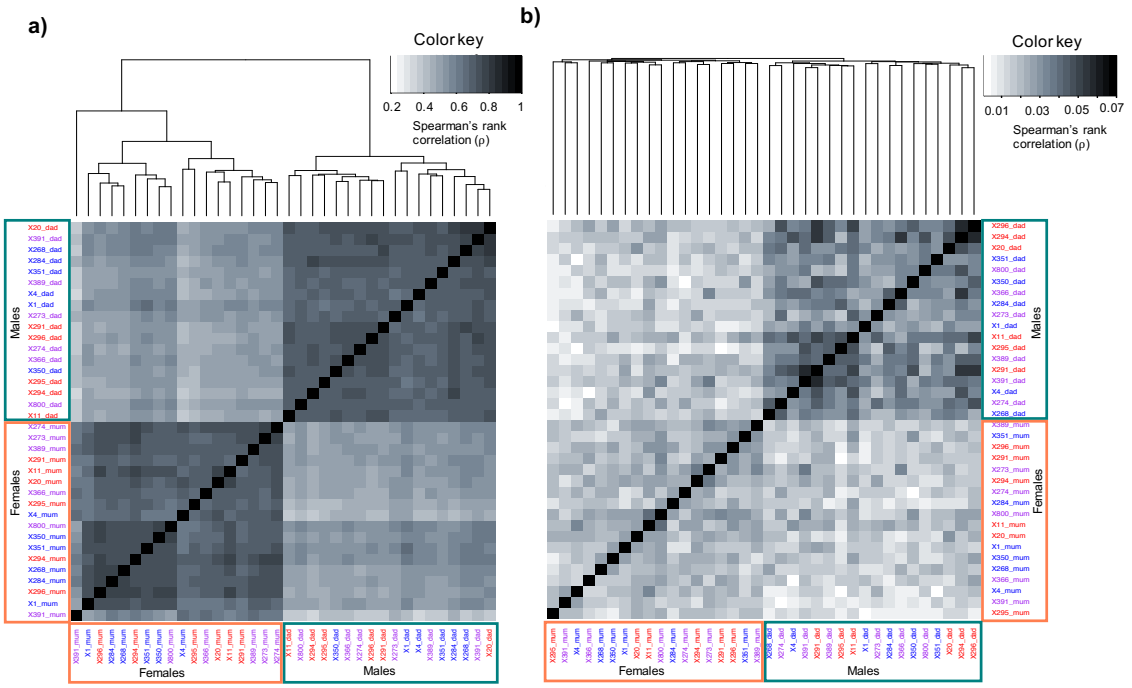


Figure 2.11: Whole genome recombination maps are correlated among individuals of same sex at broad as well as fine-scale. No specific grouping was observed based on ecotype. Pairwise inter-individual correlation (Spearman's rank correlation ρ) of the recombination rate plotted as a heatmap in a) 1 Mb sliding windows and in b) 5 kb sliding windows across the genome. Individual ids are colored based on their ecotype. Red: marine, blue: freshwater, purple: hybrid. The dendrogram plotted above the both heatmaps connect the individuals based on their correlation value. At broad- as well as fine-scale, individuals are found to be grouped according to their sex (marked by a box around individual ids). However, no specific grouping of individuals based on ecotype was observed at either scale.

As we would expect based on preferential crossover formation at the chromosome periphery, most of the coldspots were present in the center of the chromosomes and appear to be spatially clustered. However, it is important to note that, in many chromosomes (including chrIV shown in Figure 2.12) a large coldspot is identified at the very end of the chromosome. Such telomeric coldspots were especially observed for chromosomes in which scaffold assembly at the ends are nearly complete. A map of cold regions identified across all 21 chromosomes are given in Appendix figure 1. In many chromosomes, such telomeric coldspots spanned nearly 100 kb beyond the 50 kb region which we excluded while CO calling. This indicates that, in addition to centromeric regions, crossovers are also suppressed at the telomeric ends.

A coldspot search was carried out categorizing the data set based on sex and ecotype and by keeping the sexes separate for ecotype specific analysis.

Table 2.2 shows the summary of the total number and span of cold regions identified in each category. A list of top 5 coldspots identified in each category is given in Appendix Table 4. As we categorize the data, the effective sample size

gets smaller and as a result we might over estimate the number and size of coldspots. However, this one generation coldspot list would be a good starting point to further validate those ecotype and sex specific cold regions by sampling large number of meiotic products. A novel method that we developed for crossover detection from pooled gametes (Dreau et al. 2019) (also presented as the chapter 4 in this thesis) would provide a strategy to validate these cold regions from large number of meiotic products of single individuals.

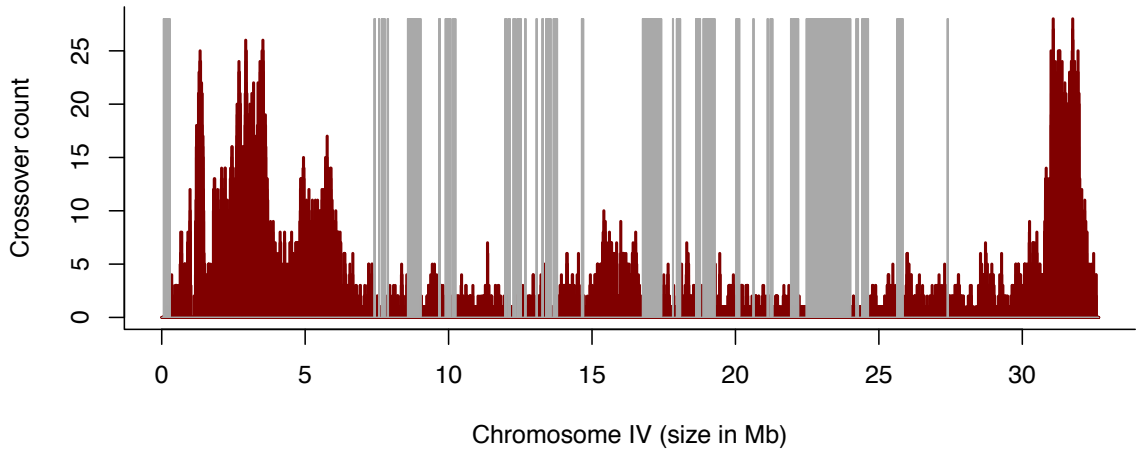


Figure 2.12: Coldspots of recombination across the largest stickleback chromosome (chrIV). Crossover events across of chromosome IV in 5 kb sliding windows are plotted in maroon. Cold spots of recombination identified in this chromosome are marked with grey vertical bars.

Table 2.2: Number and total span of coldspots detected in each data set

Ecotype	Male		Female	
	Number	Total size (Mb)	Number	Total size (Mb)
Freshwater	143	183.3	178	68.5
Marine	197	224.3	217	84.5
Hybrid	171	196.9	206	83.4
All combined	398	183	362	57.8

2.4.8 Reduced recombination rate around marine-freshwater divergent adaptive loci

Previous studies have reported that parapatric pairs of stickleback ecotypes show reduced recombination at regions of elevated divergence (Roesti et al. 2013; Marques et al. 2016; Samuk et al. 2017). Those studies were based on population

genetic estimates of recombination rate or broad-scale genetic maps. Here we have created a high-resolution sex-specific recombination map for a marine and a freshwater population and for their hybrids. Using our data, we examined how do crossover events associate with the genomic landscape of stickleback adaptive divergence.

First, we used the existing list of stickleback marine-freshwater parallel adaptive loci from Jones et al. (2012). The list contains 242 loci across 21 chromosomes of the stickleback genome. We find that the distances between adaptive loci and all observed crossover events were significantly greater than what is expected by chance (Wilcoxon rank sum test p value $< 2.2 \times 10^{-16}$), which suggests that adaptive loci fall into regions of low recombination. Male crossover events were on average significantly further away from adaptive loci than were female crossover events.

Next, we quantified the recombination rate within adaptive loci. In Figure 2.13a, mean recombination rate across all adaptive loci (scaled from 0 to 1) along with left and right flanking regions of corresponding size is shown. We find that recombination rate within adaptive loci is lower than in the surrounding region. The effect is more pronounced in female data because male recombination is on average reduced both within adaptive loci and in its flanking regions.

The importance of reduced recombination during adaptive divergence is to keep the linkage between adaptive alleles. Therefore, next we examined the recombination rate between adaptive loci that are physically present on the same chromosome. Regions of linked adaptive loci were defined by the left boundary of the first adaptive locus and right boundary of the last adaptive locus within each chromosome. Linked adaptive loci (adaptive islands) are present in all but two chromosomes (chrIII and chrXVII). Altogether they span nearly 34.4% of the genome. Average recombination rate within and outside adaptive islands was then calculated for all categories. We find that, in the uncategorized data, average recombination rate within adaptive islands is 1.6 times lower than in the rest of the chromosome. The fold difference is nearly three times in males whereas only 1.2 times in females. Figure 2.13b shows fold difference in recombination rate within and outside adaptive islands with respect to genome average recombination rate. When recombination rates within and outside adaptive islands are compared for each chromosome (excluding chromosomes without adaptive loci and an outlier chromosome XIII with too small adaptive loci), we find a significant reduction in recombination rate within adaptive islands compared to outside (Wilcoxon paired test p value for sex combined data, females, and males, were 0.0044, 0.00092, and 2.734×10^{-05} respectively). In terms of overall recombination rate within and outside adaptive loci, no vivid difference was observed among same sex individuals belonging to different ecotypes. However, we do not rule out the possibility of locus specific ecotype differences.

Even though both male and female recombination rate is lower within adaptive islands, it is important to note the difference in the extent of the reduction. Nearly twice as much female recombination occurs between the adaptive loci compared to males. This observation provides further reasons to think that, sexual dimorphism in crossover distribution might have important implications in enabling rapid adaptation in diverging natural populations. While an offspring most probably receives a cassette of alleles adapted to an environment from its father, it might receive shuffled allele combinations from its mother.

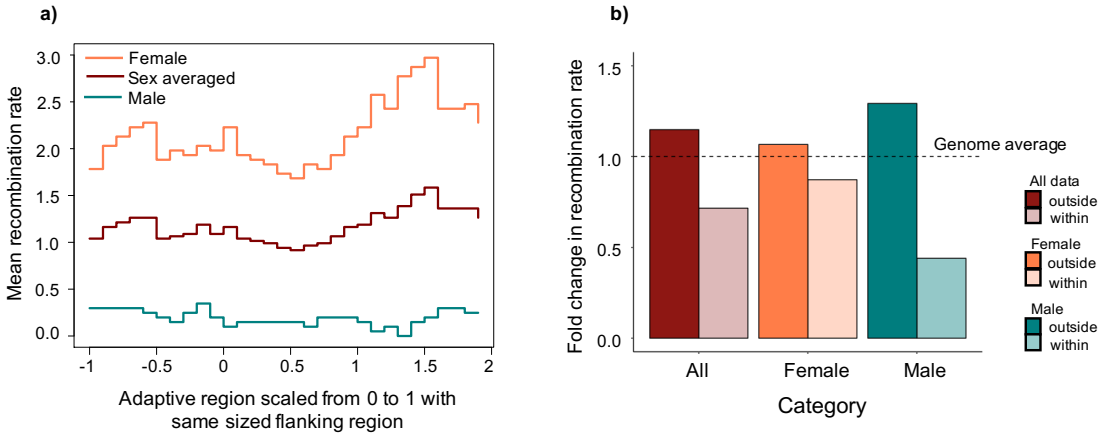


Figure 2.13: Reduced recombination rate within and between linked adaptive loci keep adaptive alleles linked. a) All 242 adaptive loci reported in Jones et al. (2012) are scaled to equal size (ranging from 0 to 1). For each adaptive locus, flanking regions of its size was included at left and right side. Mean fine-scale recombination rate across these regions estimated from all data, male and female specific data are shown. We find reduced mean recombination rate within adaptive loci compared to its surroundings. This pattern is more pronounced in females. b) Recombination rate within and outside linked adaptive loci. In all three categories analyzed, fold change in recombination rate within and outside linked adaptive loci with respect to genome average is shown. A significant difference in recombination rate within and outside adaptive loci is observed among all three categories. The most pronounced reduction is seen in males. Maroon: all data, orange: female, green: male.

In the same study, Jones et al. (2012) calculated the degree of parallel genetic divergence among marine and freshwater ecotypes using cluster separation score (CSS) across the genome. A negative correlation is observed between recombination rate and CSS score when compared across the genome in 5 kb sliding windows (Figure 2.14 a and b). We find that windows with higher CSS score occur in regions of lower recombination rate and vice versa. In addition, we also checked correlation between fine-scale recombination rate and genome wide divergence (F_{ST}) in River Tyne population. Since we have sequenced 12 marine and 12 freshwater parental individuals to a depth of more than 40x genome coverage, using this data, Weir-cockerham F_{ST} was estimated in 5 kb sliding windows across genome. At this scale, F_{ST} also shows a negative correlation with male and female recombination rate (Figure 2.14 c and d).

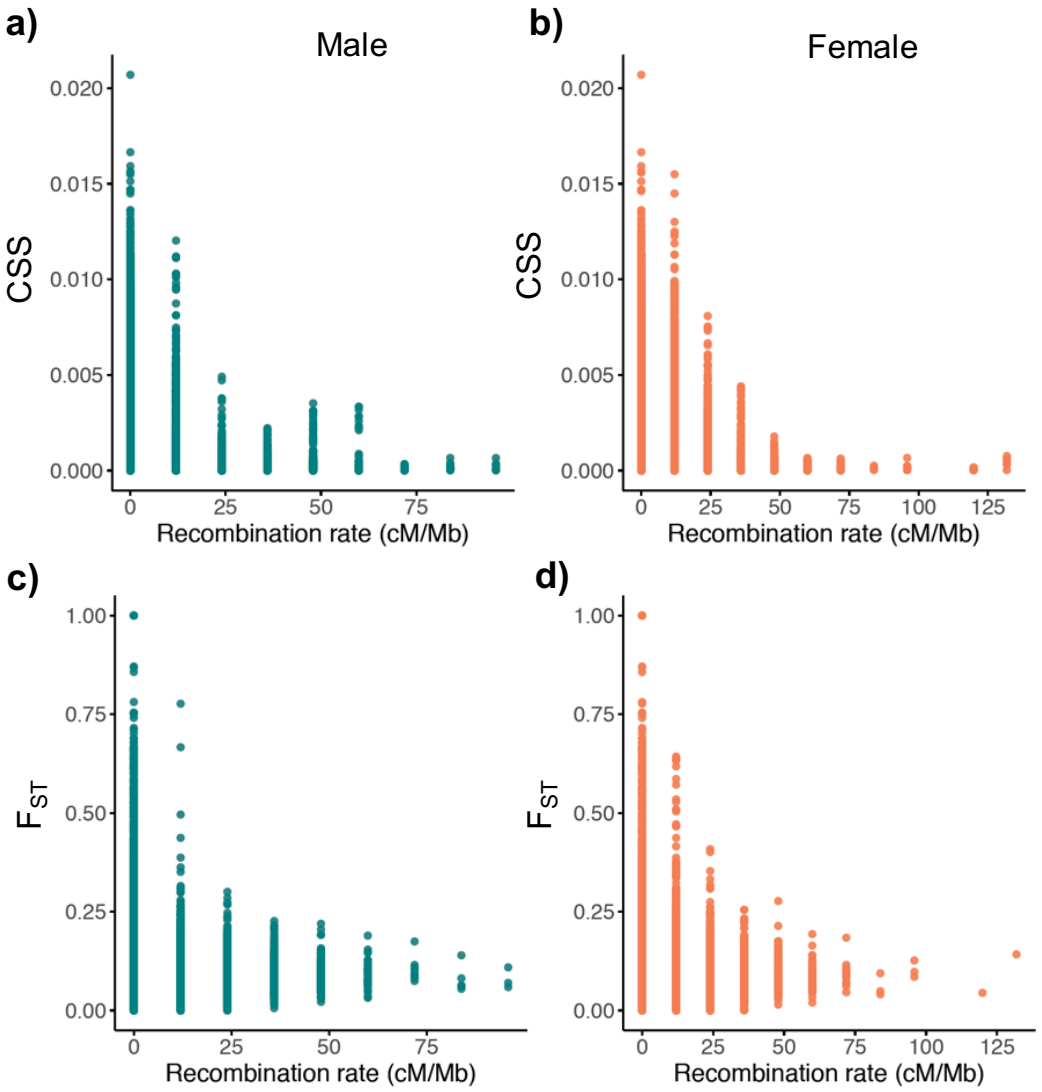


Figure 2.14: Scatter plot showing negative relationship between fine-scale recombination rate Vs CSS, F_{ST}. Sex-specific recombination rate at every 5 kb interval and freshwater-marine cluster separation score (reported in Jones et al. (2012)) in the corresponding intervals are plotted for males (a) and females (b). Similarly, sex-specific recombination rate and F_{ST} between Tyne marine-freshwater populations (estimated from 24 pure form parental individuals sequenced in this project) in corresponding 5kb intervals are plotted for males (c) and females (d). We find that at this fines-scale, regions with high F_{ST} and/or CSS score have low recombination rate.

In males among 5 kb bins with F_{ST} above 0.25, only 3.2% bins had a non-zero recombination rate. Highest recombination rate observed was 23.96 cM/Mb. Whereas among bins with F_{ST} above 0.5, only 1.5% of bins had non zero recombination rate and highest observed recombination rate was 11.98 cM/Mb. In females, among 5 kb bins with F_{ST} above 0.25, 11% bins had non zero recombination rate and highest recombination rate observed was 47.93 cM/Mb. Whereas among 5 kb bins with F_{ST} above 0.5, 9.2% bins have non zero recombination rate and highest recombination rate observed is 11.98 cM/Mb. These results suggest that, at fine-scale, regions of higher ecotype divergence fall

within regions of low recombination. As a result, intervals containing divergent alleles are mostly kept in linkage with the surroundings and therefore fitness costs due to shuffling of these divergent alleles and their putative regulatory regions are minimized. This observation is consistent for males and females, with the effect being stronger in males (small number of high divergent bins with nonzero recombination rate in males). There was no specific difference observed among ecotypes.

2.4.1 Recombination suppression within inversion heterozygotes

Finally, we explored the influence of chromosomal inversions on crossover rate. Individuals who carry chromosome homologs with differing orientations of a given genomic region are called inversion heterozygotes. The Tyne individuals we sequenced here segregate eight such inversions across different chromosomes (details given in Appendix Table 5). These include three major freshwater-marine inversions in chrI, chrXI, and chrXXI reported in Jones et al. (2012). In this analysis, we focused on the largest inversion, in chrXXI. Possibly due to a high level of admixture of the marine and freshwater ecotypes in the Tyne population, inversion heterozygotes were detected among individuals of both sexes and ecotypes.

Inversion heterozygotes were identified based on their higher heterozygous SNP density within inversion coordinates compared to outside. Individuals who have more than twice heterozygosity within the inversion compared to the genome average were considered as the ones to possess heterozygous orientation. Using heterozygosity as a measure, individuals were grouped as heterozygote or homozygote for the inversion of interest. chrXXI inversion boundaries defined in Jones et al. (2012) (chrXXI:7,486,833-9,173,772 span: 1,686,939bp) were used for the analysis. The crossover count within inversion coordinates and its flanking regions was estimated separately for each group. Figure 2.15 shows recombination rate within the inversion and the flanking region of the same size for heterozygotes (a) and homozygotes (b). Inversion heterozygotes had no crossovers within the inversion boundaries. Whereas inversion homozygotes had a recombination rate of 2.14 cM/Mb within the inversion. This suggests that in colinear orientation, the focal region normally undergoes active recombination, while in inversion heterokaryotypes, the lack of homology prevents recombination. However, nearly the same mean recombination rate was observed for the whole chromosome XXI in inversion homozygotes and heterozygotes (3.47 cM/Mb and 3.42 cM/Mb respectively). In addition, we found that, the proportion of chromosomes segregated with zero crossover events in heterozygotes was almost the same as in homozygotes. This indicates that, complete shutdown of crossovers within the inversion caused a compensatory increase in recombination elsewhere on the same chromosome. We noticed that there is an increase in recombination in the right flank of the inversion (centromere distal region) in heterozygotes.

Similarly, pronounced crossovers suppression was observed in other two relatively smaller inversions in chrI and chrXI (Appendix figure 2). When zero crossover events were observed within chrXI inversion, a single crossover event was detected within chrI inversion. In future, it would be interesting to test whether there are gene conversion events within these inversion heterozygotes.

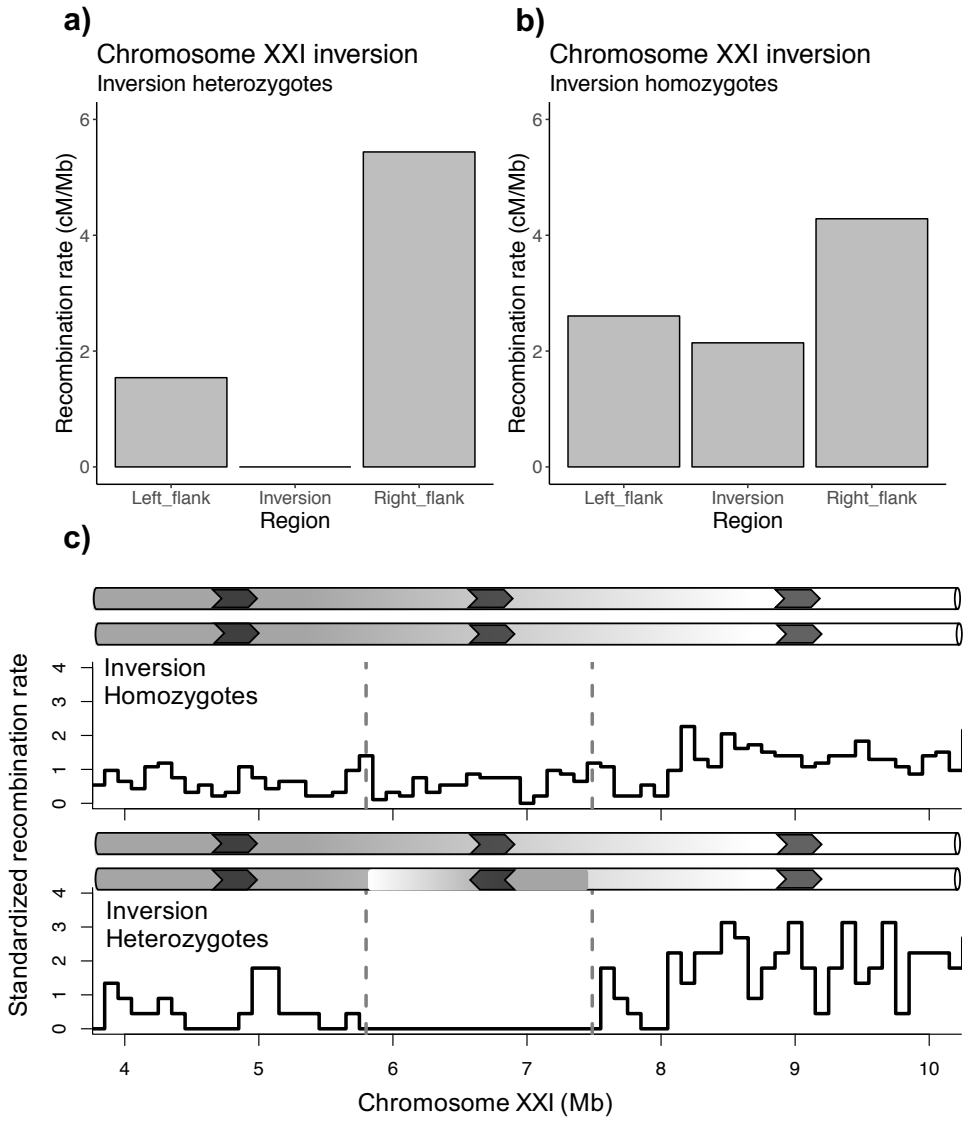


Figure 2.15:Crossover suppression within chrXXI inversion heterozygotes. Recombination rate within inversion and left and right flanking regions of the same size in inversion (a) heterozygotes (b) homozygotes are shown. (c) Standardized recombination rate (SRR) in 100 kb sliding windows across a 6 Mb region including chrXXI inversion is shown. Dotted grey vertical lines mark the inversion boundaries. SRR for inversion homozygotes (top panel) and heterozygotes (bottom panel) and are shown with schematic of the sequence orientation**.** A complete suppression of recombination within the inversion boundaries is seen in heterozygotes with compensatory increase in recombination on the right flank.

2.5 Discussion

By applying high throughput whole genome sequencing on large nuclear families, we generated the first high-resolution individualized sex-specific and ecotype-specific crossover map in threespine stickleback fish. High-resolution mapping of recombination events requires genomic data from large multi-generational pedigrees or large clutch crosses. These factors stand as a major hurdle for empirical studies in wild populations. Therefore, linkage disequilibrium (LD) based estimates of historical recombination events from unrelated individuals sampled from a population are more commonly used. Even though the LD based method is excellent for estimating population-level variation across the genome (to identify hotspots and coldspots of recombination) it is not suitable for investigating variations below population level (between sexes or among individuals). To understand the adaptive value of recombination, it is important to quantify contemporary recombination rate variation across populations and within populations. Here, we have generated 36 individualized high-resolution recombination maps that include equal numbers of individuals from divergently adapted ecotypes and sexes. In this regard our study offers a unique opportunity to document the variation in the recombination rate and landscape at different levels in a natural population and also to investigate the evolutionary implications of such variation.

Variation in the genome-wide recombination rate

We find that, in terms of overall crossover count per meiosis, females have nearly 1.76 times more crossover events than males. This female biased heterochiasmy is consistent with results presented in a previous study of sticklebacks (Sardell et al. 2018). Sardell et al. used an inter-species hybrid cross design (*G.aculeatus* females and *G.nipponicus* males) to identify crossover events. Therefore, they had lower power to differentiate sex-specific differences from species-specific differences. Our study avoids this complication and reports the full extent of sex differences in a natural population of *G.aculeatus*. It can be argued that we might have underestimated our crossover counts due to the incompleteness of the scaffold assembly of our reference genome, which would primarily affect males since their recombination events cluster more to the poorly assembled sub-telomeric regions. However, our detected percentage of zero crossover chromosomes (50.2%) is very close to the expected 50%, based on at least one obligatory crossover per chromosome pair (tetrad) per meiosis (Petronczki et al. 2003). Thus, we believe that our estimation does not deviate considerably from the real sex ratio in recombination rate.

Heterochiasmy is reported in almost all organisms studied to date. Female biased heterochiasmy, such as we observe in sticklebacks, seems to be quite common and has been reported for example in mice, dogs, and humans (Neff et al. 1999; Lynn et al. 2004; Paigen et al. 2008). Whereas male biased heterochiasmy

has been reported in organisms including birds (Smeds et al. 2016) and cattle (Ma et al. 2015). Early explanations for heterochiasmy said that recombination may be suppressed in heterogametic sex to prevent crossing over between non-homologous sex chromosomes - known as the Haldane Huxley rule (Haldane 1922; Huxley 1928). Subsequent studies revealed that this rule holds only in extreme cases of heterochiasmy, where one sex lacks recombination altogether (achiasmy). In organisms in which both sexes recombine, reduced recombination is not always observed in the heterogametic sex (e.g., cattle and sheep). Later, various mechanistic reasons, such as difference in compaction of chromosomes at the leptotene stage (Gruhn et al. 2013), difference in the extent of CO interference (Petkov et al. 2007), difference in duration of meiosis (Morelli and Cohen 2005), difference in DNA methylation (Brick et al. 2018) and evolutionary explanations such as differential selection pressure at the haploid stage (Lenormand and Dutheil 2005) were suggested to underlie the sexual dimorphism. However, patterns seen in empirical studies do not consistently support any of these hypothesized reasons and therefore demand detailed characterization of sex-specific recombination landscapes and of the underlying molecular mechanisms in more species.

In addition to the pronounced difference between sexes, we report for the first time a significant reduction in crossover count in hybrids of marine-freshwater ecotypes compared to pure forms. Evolutionary theory predicts that under divergent selection with gene flow, natural selection may prefer recombination modifiers or chromosomal rearrangements such as inversions that reduce recombination in hybrids. As a result, linkage between adaptive alleles could be maintained (Kirkpatrick and Barton 2006; Ortiz-Barrientos et al. 2016). Even though the underlying mechanism is yet to be understood, our result stands as an empirical evidence for the theoretical prediction of recombination suppression in hybrids under divergent selection. It is interesting to note that the significant difference comes from females but not from males. This indicates that recombination regulators in males and females may respond differently to selection pressure. However, the reduction in the hybrid female recombination rate was not restricted to any specific region (not limited to adaptive islands or inversions). Therefore, a female-specific recombination modifier controlling overall recombination rate may underlie this suppression.

Variation in the recombination landscape

Distribution of crossover events across the genome varies at different levels. At megabase scale, most eukaryotes show increased recombination at the sub-telomeric regions (Barton et al. 2008; Liu et al. 2014; Smeds et al. 2016) which is thought to be caused by the clustering of the telomeric ends at the nuclear envelope during early meiosis; a feature called 'telomeric bouquet' formation that facilitates homolog recognition (Da Ines and White 2015). Our data confirms that this common eukaryotic pattern prevails also in the stickleback fish and that the

clustering of crossover events to the sub-telomeric regions is more pronounced in males than in females.

The broad-scale pattern of recombination is considered to be shaped by highly conserved features such as chromosome condensation, centromere position, requirement for an obligatory crossover per chromosome per meiosis, and crossover interference (Capilla et al. 2016). Therefore, we would not expect, and do not find, variation at this broad scale between stickleback ecotypes or individuals. However, the substantial differences between sexes indicates that, the above-mentioned factors might act in a sex-dependent manner. In chromosomes with biased arm length, male crossovers cluster to the sub-telomeric region of the long arm, whereas female crossovers are more dispersed and occur also on the short arm. This suggests that female crossovers are less sensitive to the suppressive effect of centromeres. Furthermore, we report twice as much difference in inter-crossover distance in males than in females. Therefore, the strength of crossover interference seems to differ between the sexes in sticklebacks, just as it is observed in other organisms such as humans (Housworth and Stahl 2003), dogs (Campbell et al. 2016), and cattle (Wang et al. 2016). In mice, and humans it has been shown that differences in crossover interference between sexes is correlated with differences in chromosome condensation during early meiosis (Petkov et al. 2007). A detailed sex-specific study of chromosome condensation during meiosis-I is required in order to investigate the molecular underpinnings of these broad-scale patterns.

At finer scale, the recombination distribution shows substantial non-randomness across genome. In our data set, 80% of the male and female COs occurred within 27.7% and 52.7% of the genome, respectively. Whereas in dogs same amount of crossovers were observed in 17.5% of the genome in males and in 19.8% in females (Campbell et al. 2016). In humans 80% of COs occurs in less than 3.5% of the genome in both sexes (Halldorsson et al. 2019). This suggest that, non-uniformity of stickleback recombination landscape is less than that of humans and dogs. When quantified using the Gini coefficient, we find higher heterogeneity in sex averaged recombination landscape in the sticklebacks compared to that of *C.elegans*, *Drosophila*, and yeast but lower than that of mouse and humans. In accordance with this observation, we detect fine-scale recombination hotspots across the stickleback genome that are 'semi-hot' compared to reported mammalian hotspots in terms of number, and strength. The hottest hotspot we identified in the stickleback genome is approximately six times reduced in intensity compared to the hottest mammalian hotspots (Paigen et al. 2008; Paigen and Petkov 2010). Nevertheless, the presence of hotspots suggests the existence of fine-scale recombination landscape regulators in the stickleback genome. We also report sex-specific hotspots of recombination with extremely low sharing between sexes. This observation, taken together with the significant correlation of fine-scale recombination landscape among individuals of the same sex, suggests that the

fine-scale recombination regulators may act in a sex-specific manner. A detailed investigation of genomic and epigenetic features associated with the stickleback recombination landscape is given in chapter 3.

Evolutionary implications of recombination rate variation

Our analysis shows that the non-random distribution of recombination events plays a major role in shaping the genomic landscape of marine-freshwater adaptive divergence in the sticklebacks. Regions of higher divergence in parallelly adapted populations (reported in Jones et al. (2012)) fall into regions of lower sex-averaged recombination rate (Figure 2.16). This suggest that, major loci involved in freshwater versus marine adaptation are kept in linkage ('adaptive islands') and can segregate together and facilitate rapid adaptation. This strategy is especially relevant in populations, such as our study population, that are under divergent selection pressure with gene flow. Reduced recombination act as a barrier for introgression of genes adapted to a different habitat.

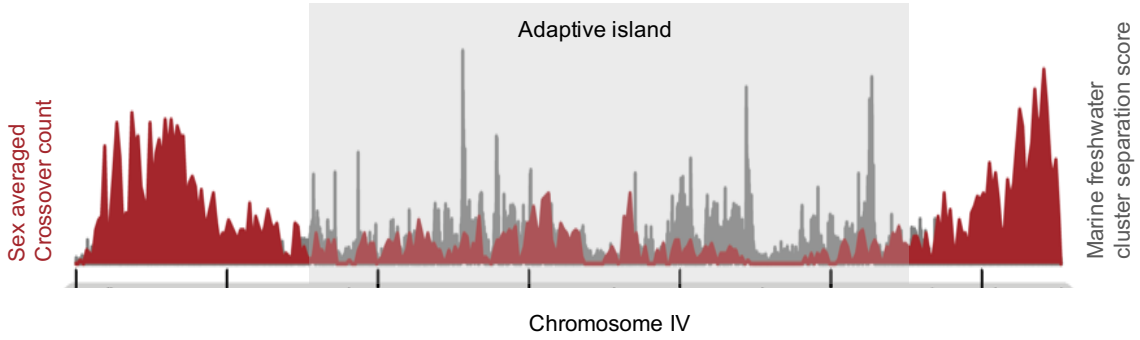


Figure 2.16: Adaptive loci are clustered at regions of reduced recombination. All crossover events identified across the largest stickleback chromosome is (maroon) overlaid on top of marine-freshwater divergence reported in terms of cluster separation score in Jones et. al., 2012 (grey). Grey rectangular box encloses all adaptive loci in this chromosome.

Interestingly, our sex-specific analysis shows that, the reduction in recombination rate in such adaptive islands are mostly driven by males. Male recombination is nearly three-fold reduced within adaptive islands compared to other regions. Whereas only a 1.2-fold reduction is observed in the females. This indicates that females more often recombine between marine-freshwater adaptive loci. As a consequence, while males tend to keep the linkage among adaptive alleles, females shuffle them more often. Even though, pure forms (marine and freshwater individuals) are not assumed to have any effect, this strategy would have important implications for hybrid fitness. Offspring of F1 hybrid females may carry shuffled combination of adaptive alleles (from mother) and a linked set of adaptive alleles (from father) in its homologous chromosomes. Our study shows that, recombination between adaptive loci in hybrid females is as much as that of pure forms. Even though we document a significant reduction in genome-wide recombination rate in hybrid females, it is not limited to the adaptive loci.

However, it is possible that hybrid recombination is suppressed between specific loci that are functionally important in this population. Investigating the recombination rate between differentially expressed genes in marine and freshwater ecotypes of this population enable us to test the adaptive significance of hybrid recombination rate reduction. One would expect to observe reduced hybrid recombination between differentially expressed genes in freshwater and marine populations. Offspring of hybrid females may experience fitness cost if they produce maladaptive gene combination via recombination. However, further studies in the field are required to investigate whether there is any fitness defect for offspring of hybrid females.

Overall, our study shows that the stickleback recombination rate and landscape variation is adaptive and evolving under natural selection pressure. Features such as inversions play a major role in suppressing recombination. Furthermore, we show that by providing semipermeable barriers to gene-flow during hybridization, heterochiasmy may have important evolutionary implications. This effect would have been obscure in population level sex-averaged recombination studies. Therefore, our results also emphasize the importance of sex-specific studies to investigate the causes and consequences of recombination rate variation in the evolutionary context.

2.6 Materials and methods

2.6.1 Ethics statement

All animal experiments were done in accordance to EU and Baden-Württemberg state regulations. The Max Planck Society holds the permits to capture and raise sticklebacks. Fish facility is maintained under Baden-Württemberg regional authority permission (Competent authority: Regierungspräsidium Tübingen, Germany; Permit and notice numbers 35/9185.82-5, 35/9185.46)

2.6.2 Stickleback samples

The stickleback fish used in this study were collected from River Tyne in Scotland. Marine and freshwater sticklebacks were caught from 4 km and 19 km respectively from the river mouth during May-June 2014. Following the standard protocol, in vitro fertilization crosses were carried out in the field and the embryos were raised in stickleback fish facility at the Max Planck campus in Tübingen. Both marine and freshwater fish were raised in 10% seawater salinity (3.5 ppt) with daily 10% water change. All fishes were fed once a day with same food that consist of both marine and freshwater invertebrate diet.

For fine-scale individualized recombination map construction, 6 marine ♀ X marine ♂ crosses, 6 freshwater ♀ X freshwater ♂ crosses, 3 (freshwater X marine) ♀ X (freshwater X marine) ♂ crosses, and 3 (marine X freshwater) ♀ X (marine X freshwater) ♂ crosses were carried out. Thereby, 18 nuclear families consist of ~94

offspring per family were generated using 36 distinct parent individuals. Out of these 18 families, two freshwater and two marine crosses were carried out in the field (Family X1, FamilyX4, FamilyX11, Family X20) whereas rest of the 14 crosses were done in the lab using wild cross offspring grown in the aquariums. For crosses that resulted in small clutch size, a second round of in vitro fertilization was carried out in the next breeding season with eggs from the same female and cryo-preserved sperm from the same male fish. Cross details are summarized in the Appendix Table 1.

2.6.3 Whole genome re-sequencing

DNA extraction

DNA was extracted either from a piece of tail fin (parents) or from the whole fish body (offspring at 1-month age) using a Solid Phase Reverse Immobilization (SPRI) bead-based protocol. In short, tissue samples were lysed in a 96 well plate over night at 55^0C with lysis buffer containing 1M Tris pH 8.0, 5M NaCl, 0.5M EDTA, 10% SDS and 20 mg/ml proteinase K. Following lysis, RNA was digested using 10mg/ml RNase A at 37^0C for one hour. 5 M KAc was added at 0.325 times lysate volume and incubated at -20^0C for 30 minutes to precipitate protein and the precipitate was then discarded after centrifugation. Homemade SPRI beads were added to the cleared lysate to bind DNA. DNA bound SPRI beads were settled using magnetic rack and washed 2 times with 80% ethanol. DNA was then eluted in 1X TE (10mM Tris pH 8.0, 1mM EDTA) buffer. Multiplexed DNA extraction protocol was carried out using TECAN® liquid handling robot.

Library preparation

This project involves high throughput sequencing of large number (~1700) of stickleback individuals. In order to reduce the cost, and to automate most of the steps using a TECAN® liquid handling robot, I optimized a low-cost high throughput library preparation protocol following the Illumina TruSeq library generation principle. This protocol is adapted and improved from (Bronner et al. 2014). Detailed protocol and materials used are given in the appendix.

Briefly, ~300-500 ng of good quality genomic DNA was sheared to an average fragment size of 300 bp using Covaris® LE220 (96 well plate mode) at CeGaT in Tübingen. Ends of the randomly sheared DNA were repaired and the 5′ ends of the fragment were phosphorylated using the end repair master mix. The reaction mix was incubated at room temperature for 30 minutes. After end repair, fragments of length between 250-500bp were size-selected using homemade SPRI beads following Ampure XP® (Beckman-Coulter) DNA size selection protocol. Subsequently, addition of a single A nucleotide to the 3′ ends of the fragment was carried out in order to enhance the efficiency of adapter ligation. Custom-made Illumina-compatible adapters with 3′ T overhangs were then added to the A-tailed fragments. Each adapter has i7, i5 duplex confirmation with a unique 6 bp or 8 bp

barcode in it. Adapter-ligated DNA library was then PCR amplified. Quantity of the amplified library was measured using TECAN PicoGreen plate reader. Equal quantity of the PCR products was then pooled in such a way that each pool contained 16 offspring libraries and 2 parental libraries. Parental libraries were added into each pool in order to give them more sequence coverage. Pooled PCR products were cleaned and selected to an average library fragment size of 420 bp. Sample clean up in between reactions, and final library size selection was carried out using homemade SPRI beads following Ampure XP (Beckman-Coulter) protocol. Final libraries were quantified using Qubit 2.0 fluorometer with high sensitivity reagent. Size distributions of the final library pools were checked using Bioanalyzer 2100 desktop system with high sensitivity reagents. Quality-checked library pools were then normalized to 2.5 nM concentration and submitted for sequencing in an Illumina HiSeq 3000 sequencer. Each pool containing 16-18 libraries was sequenced in a lane with 2 x 150 bp chemistry. With an estimate of about 360 million reads per lane, in an 18-plex pool we expected each sample to have at least 10x genome coverage. With 94 offspring and two parents, six lanes of sequencing were carried out per family. Parental libraries included in all six lanes were expected to result in a coverage of around 60x per parent. Sequencing was performed at the Genome Center in Max Planck institute for Developmental Biology, in Tübingen.

2.6.4 Data analysis

Major steps in the data analysis pipeline, from initial processing of raw sequenced reads until crossover calling, are summarized as a flowchart in Figure 2.17. The analysis pipeline involves tools that represent community standards for SNP calling and phasing, with modifications, and custom scripts to accommodate our experimental design and identify crossover events at high resolution. The pipeline scripts were put together with Unix bash wrappers to facilitate parallel processing at the MPI Tübingen computer cluster and to reduce computing times.

Initial read processing

Sequenced libraries were de-multiplexed based on individual barcodes. Next, reads from each individual were mapped to the stickleback reference genome gasAcu1 (Broad S1 assembly generated from an Alaskan freshwater female) (Jones et al. 2012) using Burrows-Wheeler Aligner (BWA) v0.7.10-r789 (Li and Durbin 2009) with bwa-mem option. Mapped reads were then sorted and indexed using SAMtools (Li et al. 2009). Parents of each family were sequenced in multiple lanes in order to get higher coverage. Therefore, their mapped reads from different lanes were merged.

Further read processing until variant calling was carried out according to the best practices recommended by Genome Analysis Tool Kit (GATK) (McKenna et al. 2010; DePristo et al. 2011). In short, optical read duplicates in each sample

were marked using Picard tools version 1.128 Markduplicate function. (Mark duplicate step was done only for five families: Family X284, FamilyX291, FamilyX295, FamilyX296 and FamilyX800. For other families, we directly proceeded to the next step). Local re-alignment of reads around indels was then carried out using IndelRealigner program of GATK v3.4. Base Quality Score Recalibration (BQSR) of reads was carried out using Recalibration program of GATK v.3.4. Since we sequenced nuclear families, all SNPs segregating among offspring of each family can be identified accurately from their parents. Therefore, an initial round of SNP calling for deep-sequenced (40x or more genome coverage) parents of each family was carried out. High quality SNPs from parents were used as known sites for BQSR of the respective family. These quality control steps helped to circumvent possible technical errors (such as PCR duplication causing SNP calling error, base pair calling error during sequencing, read mapping errors etc.) and to identify a reliable set of variants for further analysis.

Variant calling and filtering

Following initial read processing and base quality score recalibration, variant calling for each individual in a family was carried out using the HaplotypeCaller option in GATK v3.4. Each individual gVCF file from a family was then combined using GenotypeGVCF option in GATK v.3.7. This step creates a joint genotype file for a family with information of parents and all offspring at each variant position.

High quality variants for further analysis were then selected based on the following criteria: 1) SNPs (Indels are excluded), 2) biallelic, 3) heterozygous in either parent, 4) SNPs with quality score greater than first quantile of the quality score distribution of the whole family data set, 5) model-based clustering analysis of allele frequency was carried out using R package mclust version 5.4.1 (Scrucca et al. 2016). Since in each nuclear family, alleles segregate in mendelian ratio, SNPs falling in tight clusters of allele frequency around 0.25, 0.5 and 0.75 were selected, 6) SNP filtering based on GATK hard filtering criteria (QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0), 7) SNPs with coverage not differing by more than 50% of mean coverage and read bias between alleles less than 30% were selected.

Scaffold orientation correction

The stickleback reference genome, Broad S1 assembly was used for the initial mapping. However, it is known to contain scaffold orientation errors. Since the scaffold orientation can affect phasing and crossover identification, I corrected the orientation of 13 scaffolds (according to the latest stickleback assembly, Glazer et al., 2015) and correspondingly corrected the order of SNPs in this data set prior to haplotype phasing. A custom made perl script was used for correcting the SNP order. It has to be noted that, only the orientation of scaffolds in the assembled part of the genome was corrected. In contrast to the latest version of assembly, no new

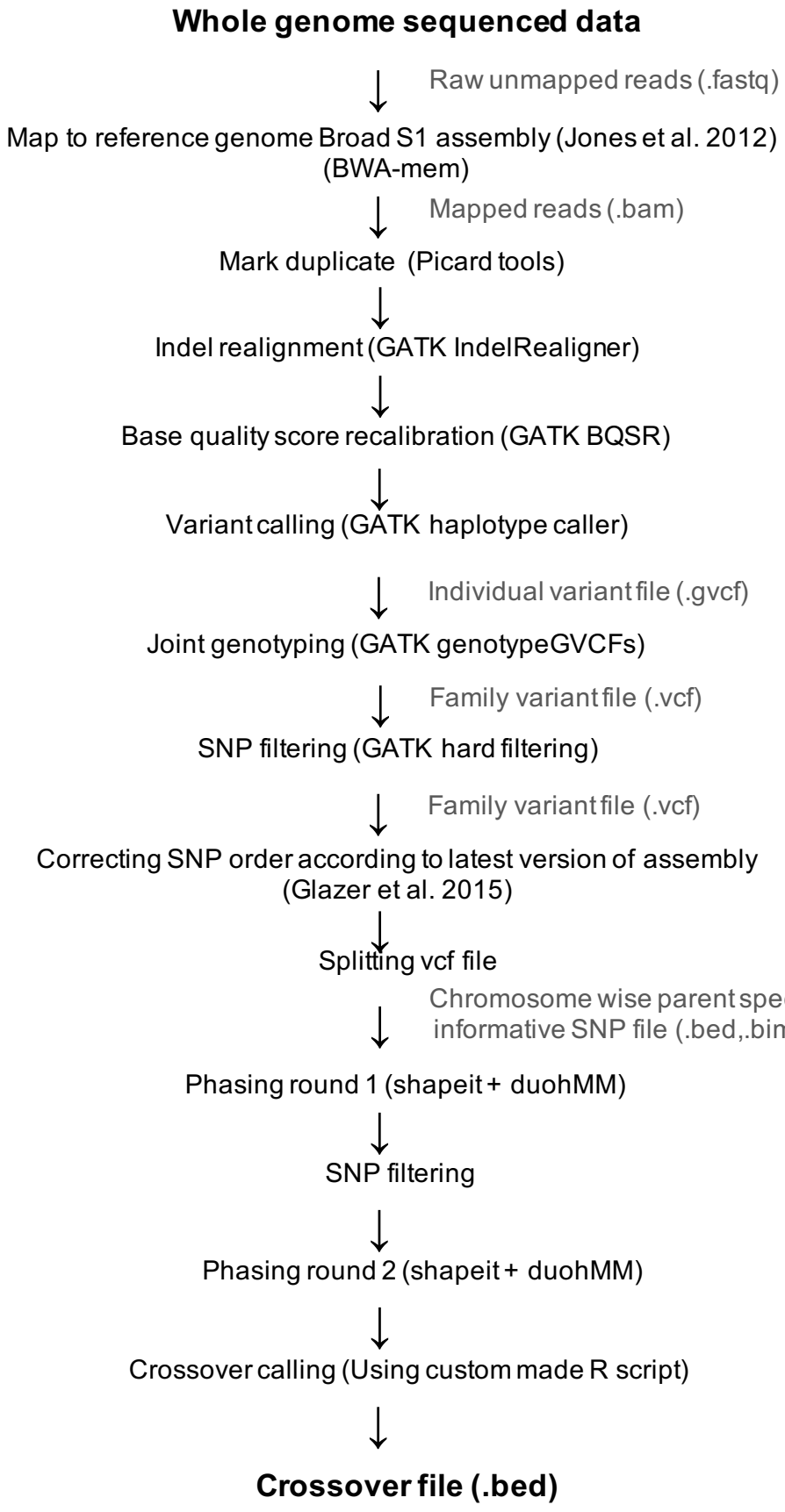


Figure 2.17: Major steps involved in sequencing data analysis pipeline

scaffold was additionally tied into the assembled chromosomes from unassembled scaffolds. For the rest of this thesis, I name this updated reference coordinates as 'modified gasAcu1' assembly.

Haplotype Phasing

The phasing algorithm named SHAPEIT (Delaneau et al. 2011) in combination with duoHMM (O'Connell et al. 2014) was used to phase SNPs within a family. SHAPEIT-duoHMM combination has been shown to produce accurate results for large data sets with pedigree information. SHAPEIT initially produces chromosome length phasing without considering the relatedness of the individuals. Then the HMM method (duoHMM) combines the haplotype with family information and identifies SNP inheritance patterns at each site. This allows the correction of switch errors and identification of genotyping errors and thereby produces phased haplotypes with high level of accuracy. In order to make the phasing and crossover detection more straightforward, the joint vcf file for a family was split into a vcf file for paternally informative SNPs and another file for maternally informative SNPs. Paternal informative SNPs are those that are heterozygous in the father while being homozygous in the mother and vice versa for maternal informative SNPs. The informative SNPs, including the pedigree information, was then converted to .bed/.bim/.fam format using plink in order to be used as input files for SHAPEIT.

SHAPEIT + duoHMM requires also a genetic map for each chromosome as an input. In order to avoid any bias in phasing due to the preexisting recombination rate per window, a linear genetic map with 3 cM/Mb recombination rate was used. This value was chosen because the reported genome-wide average recombination rate in sticklebacks is 3.11 cM/Mb (Roesti et al. 2013).

The first round of phasing was then carried out with SHAPEIT without using any pedigree information. The SHAPEIT output files were then pushed through duoHMM in order to identify and correct phasing errors by taking pedigree structure into consideration. In addition, duoHMM also generates a list of SNPs with high probability of them resulting from a genotyping error. After one round of SHAPEIT and duoHMM, problematic SNPs were then short-listed based on the two following criteria.

1) SNPs with genotyping error probability >0.9 in 20 or more offspring
2) SNPs showing biased transmission distortion (phased to one haplotype in more than 75% of offspring). While some transmission bias may be biologically real due to processes like meiotic drive, it may also be bioinformatic error and therefore these SNPs were removed from the analysis.

A second round of phasing (SHAPEIT + duoHMM) was then carried out by excluding those error-prone SNPs from the data set. This strategy produced more accurate phasing results with minimum error.

Crossover calling

In collaboration with a post-doctoral researcher in our group, Dr. Andreea Dreau, we developed an R based pipeline to identify true crossover (CO) events from chromosome-length phased haplotypes. Parental CO events can be identified as the switch between parental haplotypes in each of its offspring. In order to avoid calling gene conversion events as COs and to define the boundaries for crossovers associated with gene conversion events or genotyping errors (complex crossovers), we employed the following strategy.

Long switches in phased haplotypes were identified as true crossover events. Haplotype switches with >50 kb size on either side and 50 or more SNPs supporting each haplotype were considered as long switches. These criteria filtered out all small switches which are either genotyping error or gene conversion events. After calling crossover events from every sequenced offspring, further filters were applied on crossover list in order to remove false positive events detected as a result of phasing error or low sequencing coverage; 1) COs appearing in 50% or more offspring in a family between the same boundary SNPs (probably due to phasing error) were removed; 2) COs from offspring who had abnormally large number of COs across the genome (detected in offspring with sequencing coverage <2x) were removed; 4) COs with low resolution (>1Mb) were removed unless due to the lack of informative SNPs; 5) COs at inversion boundaries were removed (Further details of this filter can be seen in the section inversions)

Phasing followed by defining crossover boundary in all families for all 21 chromosomes (including sex chromosome) were carried out using steps described above.

Inversions

Structural rearrangements such as inversions are common in natural populations and may segregate within the nuclear families of this study. Parents may be heterozygous or homozygous for DNA sequences that show opposite orientation in the reference genome assembly. In regions where the genomic orientation of the sample is inverted compared to reference genome and if there is a crossover within that region, there will be false positive crossover called at both inversion boundaries. Such events will leave a signature distribution of triplet crossovers within a short physical distance. The list of crossovers was examined to find such triplets where first and third crossover occurred within 2 Mb physical distance. Across the genome, 8 such regions with multiple offspring having inversion triplets were detected in 7 different chromosomes (details given in Appendix Table 5). For these triplets, first and 3rd crossovers were removed and position of second crossover was corrected according to the inverted orientation.

Centromere identification

Approximate centromere location was identified for all but three chromosomes (chrII, chrIV, and chrVIII) by BLASTing a 186 bp centromere repeat sequence (Cech and Peichel 2015) against the Gasacu1 reference genome. For hits that were detected in unassembled scaffolds of the genome, if those scaffolds are tied back in the updated version of assembly (Glazer et al. 2015), assembly gaps where those scaffolds are inserted was specified as the centromere location.

Genetic map length and recombination rate

Whole genome genetic map length and recombination rate per individual was calculated using the following formula

$$\text{Genetic map length (cM)} = \left(\frac{Number\ of\ crossover\ events}{Number\ of\ meiosis}\right) \times 100$$

$$\text{Recombination rate (cM/Mb)} = \frac{Genetic\ map\ length\ (cM)}{Genome\ size\ (Mb)}$$

Similarly, for analyzing recombination rate across the genome at various scales ranging from 1 Mb to 5 kb, the genome was divided into non-overlapping sliding intervals of required size. Recombination rate within each interval was then calculated using the above-mentioned formula above. For each bin, crossovers with at least 50% overlap with the bin were counted. (Number of crossovers included in each scale is given in Appendix Table 2) This strategy avoided double counting of same events and also excluded COs that spanned more than two bins. COs overlapping scaffold boundaries were also excluded.

Standardized Recombination Rate (SRR)

Standardized recombination rate was defined by dividing the recombination rate within an interval of interest by the genome average recombination rate.

Gini coefficient estimation

The Gini coefficient is a statistical measure of inequality in distribution. In this project, Gini coefficient is used as a measure of heterogeneity in recombination landscape. Here, we used an in-built 'Gini' function in an R package called 'ineq' for estimating Gini coefficient of crossover distribution across uniform sized intervals (Figure 2.7a).

For comparison with other organisms, we also calculated the Gini coefficient for crossover distributions in published data sets. Mouse (Paigen et al. 2008), human (Kong et al. 2010), *Drosophila* (Singh et al. 2013), yeast (Mancera et al. 2008), and *C.elegans* (Kaur and Rockman 2014). The average Gini coefficient was calculated across the whole reported region for all organisms. Published data available for other organisms reports recombination rate at different resolutions for the whole genome or genomic subsets, necessitating adjustments/calculations to facilitate comparisons. For yeast, high resolution crossover interval coordinates

are reported (Mancera et al. 2008). From, which crossover counts were calculated across the whole genome in 5 kb sliding intervals prior to Gini coefficient estimation. In Kong et al. (2010), standardized recombination rate across whole genome in 10 kb sliding intervals are provided. In Singh et al. (2013), for *Drosophila melanogaster*, recombination rate across a studied 2.09 Mb region is given in 5 kb sliding intervals. In Paigen et al. (2008), crossover count between studied marker intervals across chromosome 1 is provided. Considering the whole chromosome 1, median inter marker distance is ~195 kb, but Paigen et al. have also generated high resolution data for the interval from 168.8-193.5 Mb with a median inter marker distance of 48 kb. Therefore, I estimated Gini coefficient for entire chromosome 1 and separately for the 24 Mb region with high-resolution data. In Kaur and Rockman (2014) crossovers in a 2.275 Mb region in *C.elegans* chromosome II are provided. The sex combined crossover events across this entire region were used for Gini estimation. For the mouse and *C.elegans* data sets, median inter-marker distance was considered as the scale of the study.

In some of these studies, crossovers are not quantified within uniform-sized intervals. Also, inter-marker distance varies across the region. Therefore, for comparison between sticklebacks and other organisms, the Gini coefficient for all organisms was estimated using an area under the curve approach as described in Kaur and Rockman (2014). In short, for each data set, intervals were sorted by their recombination rate/crossover count to generate a curve of, proportion of crossovers (y axis) covered in proportion of physical distance (x axis). Area under this curve (AUC) was then estimated using an inbuilt R library called "AUC" and Gini coefficient is calculated using the following formula: Gini coefficient =1-(2 x AUC).

Recombination hotspots

For each category of data, at each scale, intervals containing multiple crossovers with false discovery rate (FDR) below 0.05 were identified as hot intervals (for example, bins with x or more number of crossovers were considered as a hot interval, if the chance of x number of crossovers to occur within a bin assuming a random distribution is less than 5%). In order to find hot intervals at different scales, crossover events were partitioned into bins of increasingly smaller size (1 Mb, 500 kb,300 kb,100 kb, 50 kb, 30 kb, 10 kb, and 5 kb). For each bin, crossovers with at least 50% overlap with the bin were counted. (Number crossovers included in each scale is given in Appendix Table 2). This strategy avoided double counting of same events and also excluded COs that spanned more than two bins at each scale. COs overlapping scaffold assembly gaps were also excluded from this analysis since we are unaware of their exact position. Our estimation should thus be considered very conservative.

Recombination coldspots/regions

Coldspots were operationally defined as continuous regions that could accommodate at least five crossovers assuming a uniform distribution but contain zero CO events in the observed data set. To be conservative in defining coldspots, a list of all 49848 crossovers (including low resolution crossover and crossovers overlapping scaffold gap boundary) was used.

Tyne marine-freshwater F_{ST} estimation

Fixation index (F_{ST}) for marine versus freshwater populations of River Tyne was estimated from 12 marine and 12 freshwater parents sequenced in this project. High quality SNPs from all parents were joint genotyped using GenotypeGVCF option in GATK v.3.7. Genome wide Weir-Cockerham F_{ST} across 5 kb sliding windows was estimated using VCFtools.

Chapter 3

Genomic features associated with crossover and double strand break (DSB) landscape

3.1 Abstract

The number and placement of crossover events in sticklebacks vary substantially between sexes and across genome at different scales. In this chapter we examined the genomic and epigenetic features associated with sex-specific recombination to further our understanding of recombination regulation in adaptively diverging natural populations. A key initial step in meiosis is the formation of a small number of DNA double strand breaks (DSBs), which are then resolved as either non-crossovers or crossovers via DNA damage repair mechanisms. The genomic distribution of meiotic crossovers may therefore be determined by the location in which meiotic DSBs are formed across the genome. We carried out ChIP sequencing on DMC1, a meiosis-specific recombination protein, in stickleback male testes tissue to build a map of meiotic DSB landscape. We complimented this map with the high-resolution male crossover map, generated by nuclear family sequencing, and found that male CO distribution broadly mirrors DSB distribution. Even though both male COs and DSB hotspots appear to be promoter associated more than expected by chance, about 30% DSB hotspots and 36% of male COs do not have any of the tested functionally active open chromatin marks (promoter, H3K4me3 marks, ATAC-seq signal) within 5 kb distance. On the other hand, distribution of female crossovers does not show a significant association with any of the examined genomic features including gene promoters. However, we find a significant enrichment of GC content (proxy for conserved hotspots) within all male and female crossover intervals. This suggest that crossovers sites (including the ones away from open chromatin marks) in both males and females are probably not randomly distributed. Combined these results lead us to speculate that, in this species that lacks functional copy of PRDM9, an unknown additional mechanism may be involved in targeting DSBs and thereby crossovers to specific intervals at fine-scale.

3.2 Introduction

One of the major findings from recombination studies on various species including our study in sticklebacks is that recombination rate varies substantially across the genome at different scales. A general pattern is observed at the chromosome level, showing crossovers biased towards chromosomal periphery compared to the center (Barton et al. 2008; Chowdhury et al. 2009; Rockman and Kruglyak 2009; Roesti et al. 2013). Within large crossover enriched domains, crossovers are found to be concentrated in kilobase sized hotspots interspaced by crossover suppressed coldspots (Jeffreys et al. 1998; Petes 2001; McVean et al. 2004; Paigen et al. 2008; Singhal et al. 2015). Understanding the determinants of this non-random spatial distribution has been one of the core objectives of recombination studies. For evolutionary biologists it is of special interest as recombination landscape variation could affect efficiency of selection on different loci and thereby greatly influence the genome evolution.

From studies on different species, a number of chromatin/genomic features have shown to be associated with recombination landscape. Features such as chromatin condensation, gene distribution, and repeat content influence broad-scale patterning of recombination landscape (Schwarzacher 2003; Pan et al. 2011). Other features including nucleosome occupancy, epigenetic marks, GC content, CpG islands, heterozygosity, and sequence motifs are shown to influence crossover formation at fine-scale (Wu and Lichten 1994; Auton et al. 2013; Bernstein and Rockman 2016). However, none of these features alone are sufficient to direct crossover formation. Rather, a hierarchical combination of factors including chromatin state, epigenetic factors, and DNA sequence context are predicted to determine crossover location and frequency (Giraut et al. 2011; Pan et al. 2011; Tischfield and Keeney 2012; Shilo et al. 2015).

The real challenge in understanding recombination regulation comes from the fact that the directionality and extent of association between these features and recombination is different in different organisms. For example, GC content and gene density are positively correlated with recombination in wheat (Akhunov et al. 2003; Sidhu 2004), maize (Civardi et al. 1994), and yeast (Wu and Lichten 1994), while they have been shown to be negatively associated with recombination in *C.elegans* (Barnes et al. 1995; Rockman and Kruglyak 2009). In addition, it is also challenging to identify whether a genomic feature associated with recombination is a cause or effect of higher recombination. For example, in most of the organisms, higher GC content at crossover hotspots are found to be an effect of high recombination via GC biased gene conversion (GBGC). GBGC occurs during meiosis specific programmed DSB repair, in which mismatches at heteroduplex DNA is repaired with a bias towards strong G,C alleles over weak A,T allele (Meunier and Duret 2004; Rousselle et al. 2019). On the contrary, evidence to support the regulatory role of GC content have been reported in yeast (Marsolier-

Kergoat and Yeramian 2009). These diverse observations demand detailed investigations to characterize fine-scale recombination landscape in more taxa.

Regulation of recombination can occur either at the stage of DSB formation or during crossover designation. Large pedigree sequencing or LD based estimation can be used to produce high-resolution crossover maps. ChIP sequencing of proteins involved in meiosis specific DNA double strand repair can provide genome wide DSB maps. The DSB landscape could be substantially different from crossover landscape as majority of DSBs get repaired as a non-crossover. Despite, high-resolution crossover maps and DSB maps are valuable and complementary in regards to understanding the fine-scale features related to recombination. While, crossover landscape has been studied in variety of species representing plants, animals, and fungi DSB landscape have been characterized only in a handful of model organisms (Pan et al. 2011; Smagulova et al. 2011; Brick et al. 2012; Fowler et al. 2014).

One of the well characterized determinant of fine-scale DSB landscape in mammals including mice and humans is a zinc-finger protein called PRDM9 that binds to degenerate 13-39 bp DNA motifs (Baudat et al. 2010; Myers et al. 2010; Paigen and Petkov 2018). PRDM9 makes meiosis-specific H3K4me3 and H3K36me3 marks on nearest histone with its methyl transferase domain (Baker et al. 2014; Powers et al. 2016). Through those marks, PRDM9 directs double strand breaks to genomic regions which are generally away from functionally active regions (Brick et al. 2012). Moreover, it has been shown that PRDM9 is a fast-evolving gene (Oliver et al. 2009). PRDM9 alleles with differences in zinc-finger array can bind to different motifs and thereby leads to differential hotspot usage (Berg et al. 2010). As a consequence, PRDM9 dependent hotspots are less conserved over evolutionary time scale (Hinch et al. 2011; Auton et al. 2012).

However, many species including yeast (Lam and Keeney 2015), dogs (Auton et al. 2013), birds (Singhal et al. 2015), and most plants (Choi and Henderson 2015) lack a functional copy of PRDM9. Recombination in these organisms are found to be targeted towards functional genomic elements such as promoters. Their hotspots appear to be rather stable over evolutionary timescale compared to organisms with PRDM9 (Lam and Keeney 2015; Singhal et al. 2015). This could be due to strong selection pressure on features targeted by recombination machinery, arising from reasons unrelated to meiosis. Under high selection pressure, those regions counter act hotspot modification factors (such as GC biased gene conversion) and prevent hotspot erosion (Lam and Keeney 2015). In most of those organisms, regions with H3K4me3 marks are shown to be associated with crossover and/or programmed double strand breaks. However, in many organisms H3K4me3 is not the determinant of DSB formation. In general, crossovers are targeted towards regions of high CpG content or nucleosome depleted regions (Pan et al. 2011; Campbell et al. 2016). Mouse mutant for PRDM9 are shown to have no recombination deficiency but recombination is shifted from

PRDM9-directed H3K4me3 marks to H3K4me3 marks at the promoters (Brick et al. 2012). This suggests that, DSBs targeting functionally active open chromatin region is probably the ancestral state and PRDM9 directed DSB regulation could be the derived mechanism. Even though most of the organisms studies till date can be classified into either PRDM9 based or non-PRDM9 based open chromatin recombination regulation, there are also few exceptions such as *C.elegans*, *Drosophila* and *S.pombe*. Recombination hotspots are likely to be absent in *C.elegans* and *Drosophila,* and therefore the mechanism regulating their fine-scale recombination landscape is also unclear (Kaur and Rockman 2014; Smukowski Heil et al. 2015). Whereas in *S.pombe*, DSB hotspots are only partially associated with functional genomic regions and open chromatin marks (Fowler et al. 2014). This indicates that, recombination could have different characteristics and yet unknown mechanisms for its regulation.

Previous studies have shown that, threespine stickleback fish does not have functional copy of PRDM9. Two paralogs of PRDM9, PRDM9α and PRDM9β are identified in teleost fish. PRDM9β lacks the KRAB and SSXRD domains, which have been shown to be required for PRDM9's recombination function (Baker et al. 2017; Imai et al. 2017). A number of orders of teleost, including Percomorpha to which threespine sticklebacks belong, have lost PRDM9α but retain the incomplete PRDM9β (Baker et al. 2017; Sardell et al. 2018; Shanfelter et al. 2019). Therefore, stickleback provides a unique opportunity to understand recombination regulation mechanisms in an organism lacking PRDM9-mediated recombination and its associated rapid hotspot evolution.

Based on recombination map described in chapter 2, we find pronounced variation in recombination landscape between sexes and across genome. Therefore, to begin with unravelling regulators of stickleback recombination, in this chapter, I investigate genomic features associated with stickleback sex-specific recombination landscape using our high-resolution crossover map. We compliment this with a map of stickleback male meiosis-specific double strand breaks generated from DMC1 ChIP sequencing in order to determine how much DNA double strand break landscape shapes the genomic distribution of recombination crossovers. These two complementary data sets allow us to explore genomic and epigenetic features associated with crossover and DSB landscape at different scales.

3.3 Results

Crossover data generated by whole genome sequencing of 18 large pedigrees (described in chapter 2) is used for genomic feature association tests reported in this chapter.

3.3.1 Recombination and gene density

On a broad-scale, recombination is shown to be correlated with gene density in various organisms. In the majority of the organisms including human (Sidhu 2004), maize (Civardi et al. 1994), wheat (Akhunov et al. 2003; Sidhu 2004), and yeast (Wu and Lichten 1994), recombination is higher in gene rich regions with relatively little recombination outside. The broad-scale correlation with gene density is thought to be due to the higher degree of open chromatin at gene dense region making those regions easily accessible to recombination machinery. However, there are also exceptions such as *C.elegans,* in which the opposite pattern is observed (Barnes et al. 1995; Rockman and Kruglyak 2009). Genes in *C.elegans* genome are concentrated at the chromosome center where there is relatively little recombination.

We find that, broad scale stickleback recombination landscape resembles what is reported in other organisms with preferential crossover formation at the chromosome periphery. Then the question is whether crossovers are concentrated at gene rich regions? To investigate the potential link between recombination and gene distribution in the stickleback genome, gene density across the genome was estimated at different scales (windows of sizes 1 Mb, 500 kb, 100 kb, 10 kb, and 5 kb). There are 22456 annotated genes in the stickleback genome according to Ensembl (build 90) with median gene size of about 8 kb. Gene density at 1 Mb scale ranged between 2 to 225 with an average of 47 genes per mega base. The few outlier intervals with gene density more than 100 comes from clustering of small non protein coding genes. While genes occupy a total of 41.03% of the genome, the rest of the gene empty regions are interspaced between genes. The median intergenic spacing is about 3.8 kb and largest observed gene desert spans ~745 kb. However, there was no general trend of gene enrichment either at the chromosomal periphery or at the center. Gene distribution across all 21 chromosomes in 1 Mb sliding windows is shown in Figure 3.2.

In order to test whether recombination is biased towards gene rich intervals, 1 Mb sized genomic intervals were classified into three groups based on its gene density. As mentioned earlier, gene density at 1 Mb scale ranged between 2 to 225. The first quantile of the gene density distribution is considered as the low gene dense intervals (≤34 gene per Mb). Second and third quantiles are considered as medium gene dense intervals (35 to 60 genes per Mb). Last quantile is considered as high gene dense intervals (>60 genes per Mb). There were 130, 221 and 126 intervals in low, medium, and high categories respectively. In this analysis gene density is categorized into groups instead of considering it as a continuous variable in order to test whether there is any evident bias between gene rich regions versus gene poor regions. We find that, intervals with low gene density have significantly lower recombination rate compared to intervals with medium or high gene density (Figure 3.1 a). This suggests that, as it is reported in most of the eukaryotes,

recombination is less likely to occur in gene poor regions. It is possible that most of the gene poor regions are densely packaged heterochromatin which is less accessible for recombination machinery. However, there was no significant difference in recombination rate between regions of medium and high gene density suggesting, higher gene density might not increase recombination rate further.

Next, we analyzed the genome wide correlation between recombination rate and gene density at different scales by dividing genome into sliding windows of size ranging from 5 kb to 1 Mb (referred to as scale of the analysis). In this analysis gene density is considered as a continuous variable and the monotonic relationship (whether value of one variable increase or decrease according to the other variable) between the two features is examined. Spearman rank correlation at different scales for male, female, and sex averaged recombination map is plotted in Figure 3.1b. Overall correlation between gene density and recombination rate remained rather low. Highest correlation (Spearman ρ=0.35, p-value = 3.771×10^{-13}) was observed at 1 Mb scale between male recombination rate and gene density. At every scale analyzed male recombination rate showed higher positive correlation with gene density than female recombination. Figure 3.2 shows gene density in 1 Mb sliding windows overlaid with male recombination map.

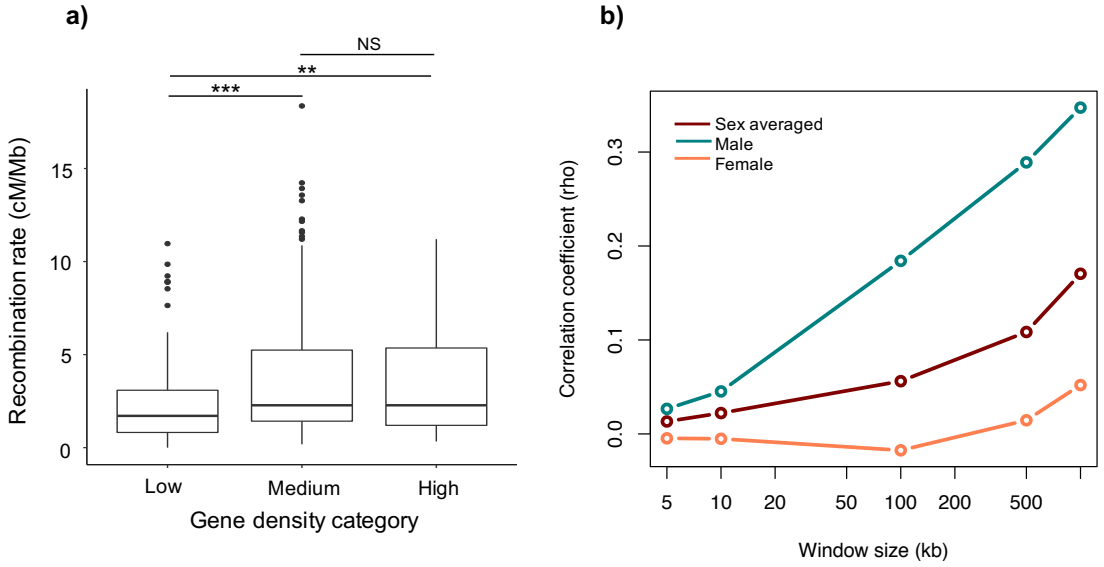


Figure 3.1: Recombination is biased towards gene rich regions. Despite the overall bias, poor linear correlation is observed between gene density and recombination rate. (a) Genomic intervals spanning 1 mega base is categorized into regions of low medium and high gene density. Number of genes per Mb in each category: low ≤34, medium 35 to 60, high ≥61. Significant difference in the recombination rate was observed between low gene-dense region versus medium and high gene-dense regions. No significant difference was observed between medium versus high gene-dense regions. Wilcoxon rank sum test one sided p values are reported. *** : p value <0.001, ** : p value <0.01, NS: non-significant. (b) Genome-wide correlation between recombination rate and gene density at different scales are plotted. Spearman correlation coefficient ρ is estimated

for sex specific and sex averaged data. Correlations for male and sex averaged data estimated at all different scales are significant with p value< 0.05. Whereas, none of the female correlations are statistically significant.

Even though genome-wide correlations are low, some chromosomes showed very high correlations with male recombination at broad scale. Spearman ρ for chrXXI, chrV and chrXV were 0.81, 0.77, 0.71 respectively. At the same time, chrVI and chrIV shows correlations as low as 0.02 and 0.04 respectively. Since chromosomes with higher correlation are acrocentric small chromosomes (Cech and Peichel 2015), it is possible that, centromere location directs both gene density and recombination rate to the distal end (see Figure 3.2). However, this difference in pattern among chromosomes may have important implications with efficiency of natural selection such that, genes in some chromosomes are more likely to be linked than genes in other chromosomes.

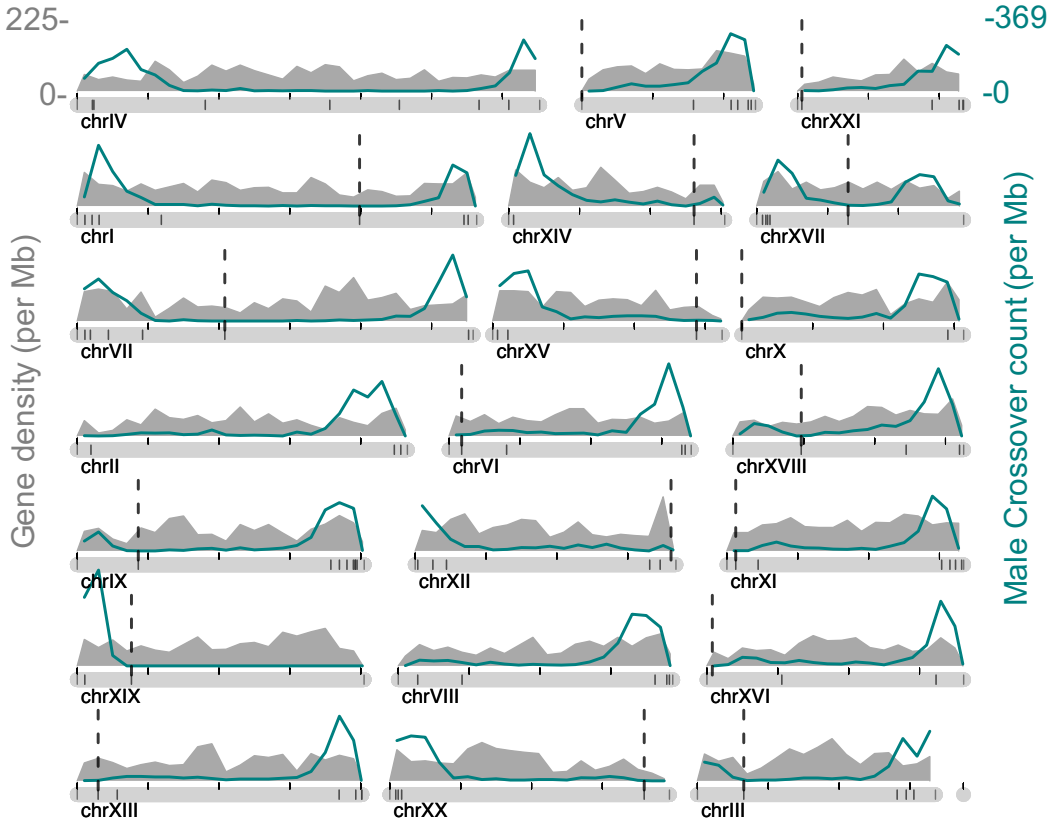


Figure 3.2: Recombination landscape is poorly correlated with gene density in sticklebacks. Density of stickleback ensembl annotated genes across the genome in 1 Mb non-overlapping sliding windows are plotted in grey, overlaid with male crossover count in 1 Mb sliding windows (green). Approximate centromere positions are marked with black vertical dotted lines. Genome-wide correlation at this scale between male crossover count and gene density is 0.35 (Spearman ρ) with p-value = 3.771×10^{-13}. However, chromosomes differ in terms correlation observed between gene density and recombination rate. Spearman ρ >0.7 is observed for chromosome XXI, V, and XV whereas <0.05 is observed for chromosome IV and VI.

With the observation that, at broad-scale, recombination mostly occurs at gene rich domains next we analyzed the fine-scale localization of crossover events within various gene features.

3.3.2 Fine-scale association with gene features

To examine the fine-scale association between crossovers and various gene features, we consider a subset of the crossover list that only includes crossover events defined with high resolution (≤5 kb) events. 27815 crossovers were present in the list that includes 9464 male crossovers and 18351 female crossovers. We estimated the percentage of crossovers overlapping various gene features, defined as follows: intergenic regions (gene-less regions more than 2 kb away from TSS or TES), genic regions (genes ± 1 kb), promoters (1 kb upstream of TSS), 1 kb downstream of TES, introns and exons. If one crossover overlapped with multiple features it was counted in all the overlapping features. To test whether any observed overlap was more than that by chance, a simulation was performed by randomly permuting crossover events (10,000 times) across the genome and estimated the overlap rate with the above-mentioned genomic features with each iteration. We find that male crossovers are significantly enriched at genic region compared to random distribution (1.05-fold more than expected by chance, empirical p value <0.0001) whereas, female crossover-overlaps with genic region did not occur more often than expected by chance. While, 58.6% of male crossovers overlapped with genic regions only 54.8% of female crossovers showed an overlap (Figure 3.3).

Applying this strategy for all the genic features, we find that, male crossover coincides with all tested genic features significantly more than expected by chance whereas female crossovers does not show any significant association compared to random. However, among genic regions, 1 kb region after transcription end sites (TES+1 kb) showed the lowest association with male crossovers. We find that 18.21% male crossovers overlap with promoter region (1 kb upstream of TSS). Among these promoters, about 56.4% of them are transcribed in testes (Jones lab unpublished data). However, it has to be noted that large number of crossovers (41.4% of male COs and 45.2% of female COs) still occur exclusively at intergenic regions those are at least 2 kb away from gene start or stop sites. In addition, among 5 kb intervals identified as hotspots, about 37% of male hotspots and 44% of female hotspots fall exclusively within intergenic region. This suggests that factors other than gene transcription may influence crossover landscape.

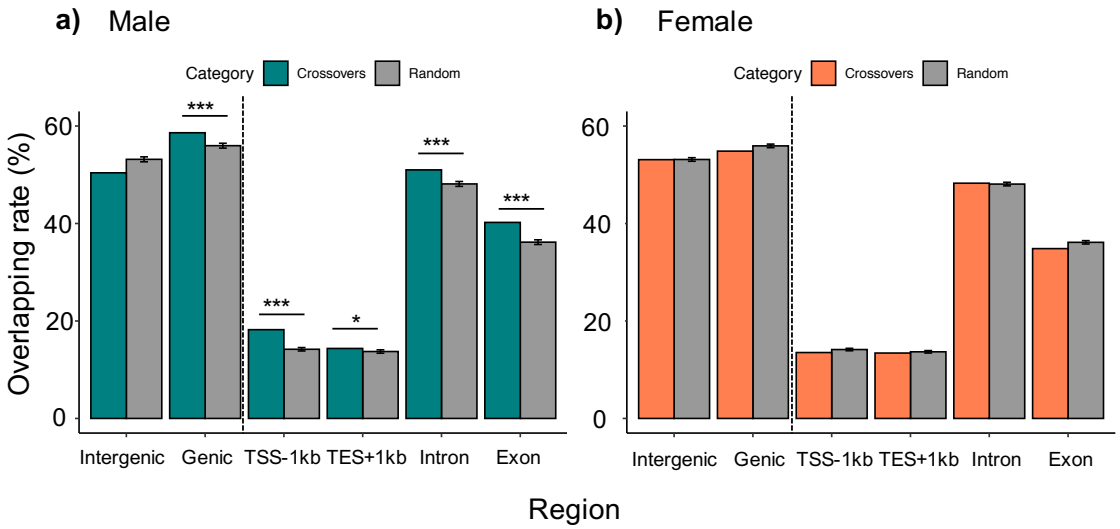


Figure 3.3: Male crossovers are significantly associated with gene features whereas female crossover association does not differ from expected by chance. High resolution crossover overlap with each genomic feature is compared against simulated random regions of matching size (grey bars) a) male crossovers b) female crossovers. The percentage of crossovers overlapping intergenic region (gene-less regions at least 2 kb away from TSS or TES), genic region (genes±1 kb region), 1 kb upstream of TSS (promoter), 1 kb downstream of TES, intron, and exon is estimated. If one crossover overlaps with multiple features it is counted in all the overlapping features. Error bars represent standard deviation of 10000 random CO simulations. Genomic feature in which crossovers overlap significantly more than expected by chance are marked with asterisks. *** : p value <0.001, ** : p value <0.01, * p value <0.05.

3.3.3 Crossover association with open chromatin region

One of the main characteristics of recombination landscape in almost all studied eukaryotic species is chromatin openness (Pan et al. 2011; Shilo et al. 2015; He et al. 2017; Kianian et al. 2018). At fine-scale resolution, crossovers are more likely to occur in nucleosome depleted regions. The only known exception so far is fission yeast *S. pombe,* in which it has been observed that DSBs are not strongly restricted to nucleosome depleted regions (Fowler et al. 2014). Another marker predictive of nearby open chromatin is H3K4me3 which has been shown to be associated with crossover hotspots in many organisms (Choi et al. 2013). While association with open chromatin region is one of the universal patterns, association with H3K4me3 marks in many organisms is incidental mostly because it coincides with promoter region, or CpG islands (Tischfield and Keeney 2012; Campbell et al. 2016).

In sticklebacks, we have seen both a statistical enrichment of crossovers around the gene promoters and a considerable amount of recombination at the intergenic region. Therefore, we next analyzed the association of crossovers with nucleosome depleted regions and H3K4me3 marks in meiotic cells. ATAC-seq (Assay for Transposase Accessible Chromatin using Sequencing) is a powerful and sensitive method to identify nucleosome depleted regions in a genome-wide scale.

A hyperactive mutant of a Tn5 transposase cleaves DNA at nucleosome depleted regions and insert sequencing adapters. This adapter tagged DNA fragments are then purified and PCR amplified to make Illumina compatible sequencing libraries. Genomic regions where short ATAC-seq reads pileup marks open chromatin regions (nucleosome depleted regions). Similarly, H3K4me3 marks in the genome can be identified employing chromatin Immunoprecipitation followed by sequencing (ChIP-seq). An antibody against H3K4me3 marks are used to immunoprecipitate the modified histone along with the DNA wound around it. The purified DNA is then converted into an Illumina compatible sequencing library. Regions where sequenced reads pileup are the regions with H3K4me3 marks.

To test the association between the stickleback crossovers and open chromatin marks, we used data generated from H3K4me3 ChIP-seq of pooled testes as well as ATAC-seq of FACS sorted primary spermatocytes. Due to the difficulty in collecting enough material from female ovaries, for the time being, genomic chromatin profile information for oocytes is not available. Therefore, here I compared female crossover locations with male ChIP-seq and ATAC-seq data (further details are given in the methods section).

By examining read depth across the genome we asked whether recombination crossovers are more likely to occur in genomic regions with H3K4me3 marks. H3K4me3 ChIP-seq read coverage across 60 kb region around high-resolution crossover midpoint in 1 kb sliding windows is plotted in Figure 3.4. We find a small enrichment of H3K4me3 reads around male crossover midpoints (Figure 3.4a). Whereas, no enrichment compared to surroundings is seen at female CO midpoints (Figure 3.4d). However, when both male and female crossovers were categorized into two groups such as promoter-associated-COs (within 2 kb of a promoter), and promoter-non-associated-COs (more than 2kb away from promoter), we find a clear pattern of H3K4me3 read enrichment around male and female promoter associated crossover midpoints. Whereas promoter-non-associated crossovers have a reduction in read count for approximately 10 kb region around their midpoints (Figure 3.4 b,c,e,f). A similar trend is also observed for ATAC-seq reads around crossover midpoint (Figure 3.5).

H3K4me3 is an epigenetic mark associated with promoters (Heintzman et al. 2007). Therefore, it is expected to have H3K4me3 marks around crossovers that are promoter associated. Lack of H3K4me3 signal around promoter-non-associated crossovers suggest that, unlike in mammals, sticklebacks might not have meiosis-specific H3K4me3 marks (laid by PRDM9 in mammals) that direct crossover events. However, lack of ATAC-seq signal around non-promoter-associated crossovers is rather unexpected. Chromatin accessibility is considered as one of the necessary requirements (even though not sufficient) for crossover formation.

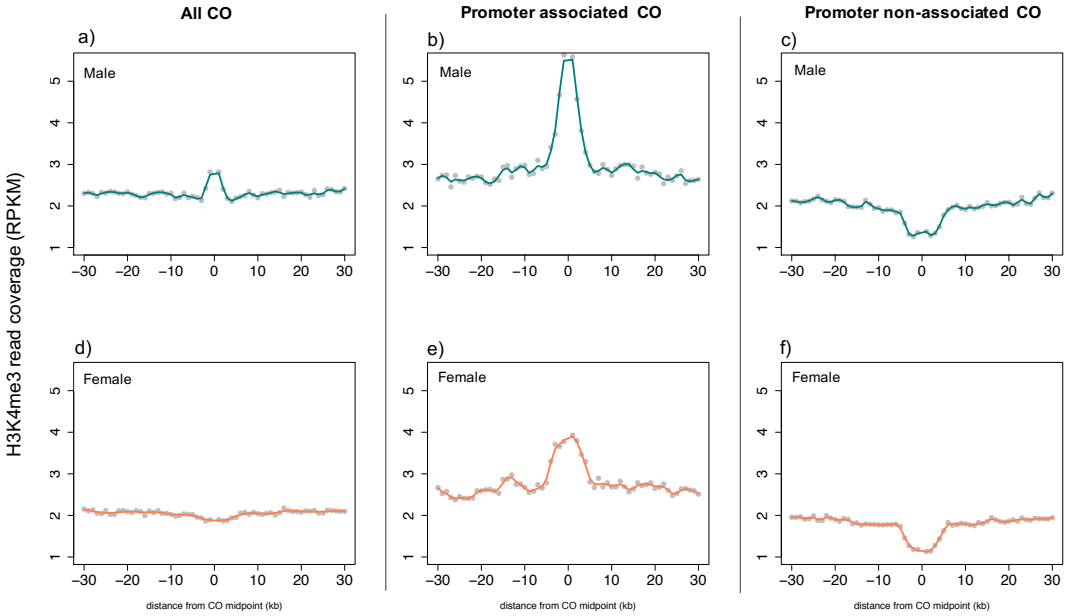


Figure 3.4: Promoter-associated but not promoter-non-associated crossovers are enriched for H3K4me3 marks. H3K4me3 normalized read coverage (RPKM) across 60 kb region around high-resolution crossover midpoint in 1 kb sliding windows are shown. Top panel: male crossovers, bottom panel: female crossovers. a,d) All CO: all high-resolution (<5 kb) crossovers. b,e) Promoter associated CO: COs within 2 kb of promoter regions, c,f) Promoter non-associated CO: COs more than 2 kb away from promoter regions.

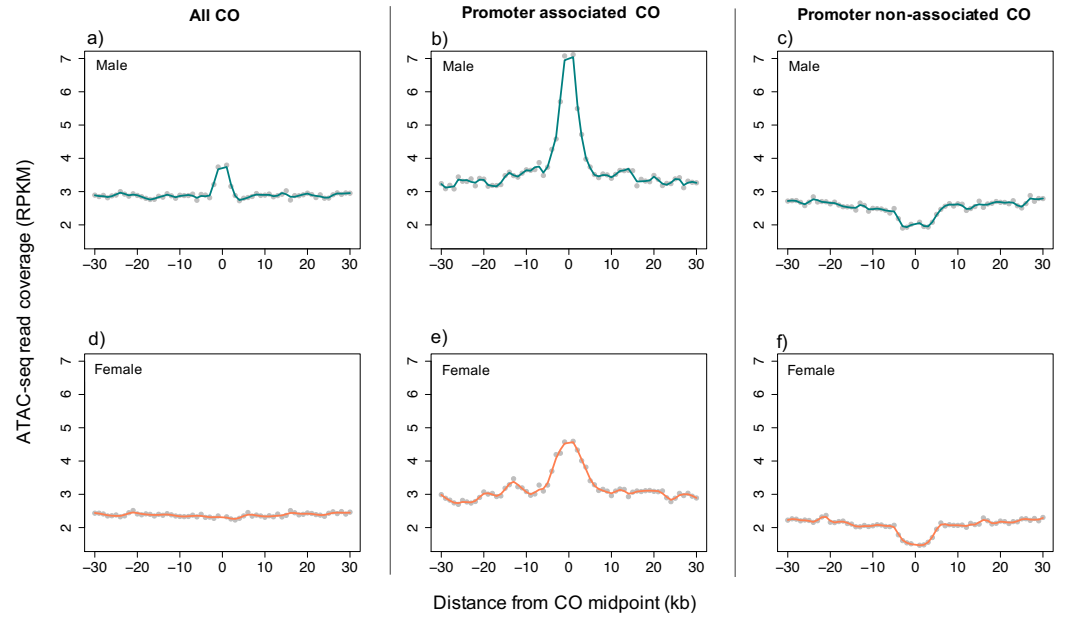


Figure 3.5: Promoter-associated but not promoter-non-associated crossovers are enriched for ATAC-seq signal. ATAC-seq normalized read coverage (RPKM) across 60 kb region around high-resolution crossover midpoint in 1 kb sliding windows are shown. Top panel: male crossovers, bottom panel: female crossovers. a,d) All CO: all high resolution (<5 kb) crossovers. b,e) Promoter-associated CO: COs within 2 kb of promoter regions, c,f) Promoter-non-associated CO: COs more than 2 kb away from promoter regions.

Therefore, to further clarify this observation, I analyzed the distance to the nearest H3K4me3 ChIP-seq and ATAC-seq peak from each promoter-non-associated crossover. In the ChIP and ATAC-seq data, regions where short sequencing reads pileup in comparison with the input DNA were called as peaks using MACS2 algorithm. Peaks called from two separate trials of freshwater and marine testes H3K4me3 ChIP-seq was combined for this analysis. Similarly, peaks called from two separate trials of primary spermatocyte ATAC sequencing using freshwater and marine testes was also combined. There was a total of 28200 H3K4me3 peaks and 35575 ATAC-seq peaks identified. For all promoter-non-associated crossovers, the distance to the nearest ATAC-seq or H3K4me3 peak is plotted in Figure 3.6.

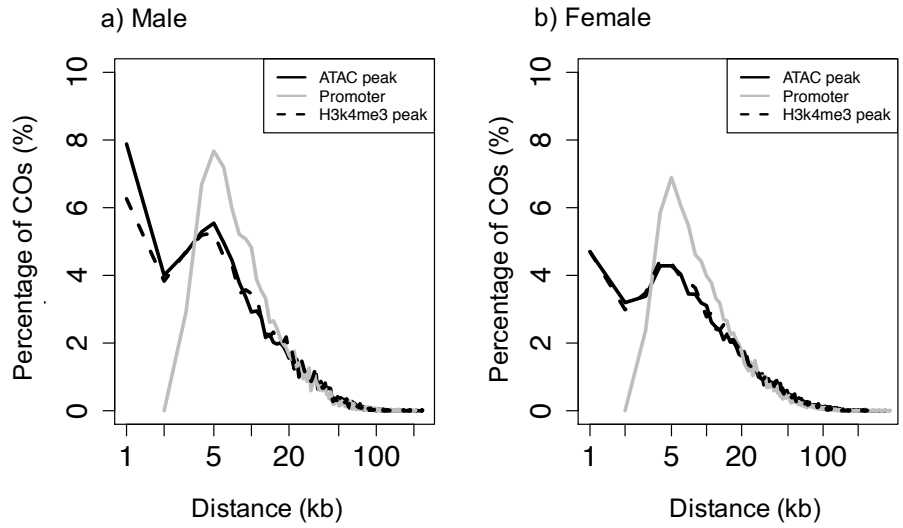


Figure 3.6:Only a small percentage of promoter-non-associated crossovers have an ATAC-seq or H3K4me3 peak in its vicinity. For each promoter-non-associated (a) male and (b) female crossover, the distance to the nearest promoter, H3K4me3 peak, and ATAC-seq peak is plotted. It shows that only a small percentage of both male and female promoter-non-associated COs have an open chromatin signature close to it.

Further confirming our previous observation, we find that only a small percentage of male and female promoter-non-associated crossovers have a nearby open chromatin mark. Considering all crossovers, about 36% of male crossovers defined with high resolution (crossover interval ≤5 kb) and 45% of female crossovers defined with high resolution (crossover interval ≤5 kb) do not have either a promoter, H3K4me3 peak, or an ATAC peak within 5 kb distance. Moreover, among 5 kb intervals identified as hotspots, 43% of male hotspots and 52.9% of female hotspots did not have an ATAC peak called within 5 kb distance. Though the lack of open chromatin signature at the crossover site was rather unexpected, it is possible that the DSB initiation sites of the crossovers may coincide with regions of chromatin accessibility. Therefore, to examine this

possibility, we mapped meiosis-specific DSB sites in the male stickleback genome and examined the association between DSB locations and open chromatin marks.

3.3.4 Stickleback male double strand break landscape

The association tests described above were performed using high-resolution crossover events detected from large pedigree sequencing. However, it is possible that the molecular level regulation of crossover location occurs prior to crossover resolution at the sites of programmed DNA double strand breaks (DSBs). Therefore, DSB sites may associate with its regulatory features than sites of crossovers. In order to understand where DSBs occur in the stickleback genome, we carried out a DMC1 ChIP sequencing on pooled testes tissue. DMC1 is a meiosis-specific protein shown to polymerize on the single strand overhang intermediate during DSB repair (Figure 3.7). DMC1 ChIP sequencing is a difficult but promising method for mapping meiotic DSBs and have been performed successfully in mice and humans (Smagulova et al. 2011; Pratto et al. 2014). The major challenges in mapping DSBs comes from the transient nature of DMC1 bound DSB repair intermediates and difficulty related to isolating the single strand DNA overhangs. Because of its transient nature, it is difficult to isolate the required number of cells in the exactly correct and same stage of meiosis.

Due to the low sequence identity of the stickleback DMC1 protein with commercially available human and mouse antibodies, we produced Guinea pig polyclonal anti-stickleback DMC1 antibodies for use in immunoprecipitation (IP, see methods). IP was followed by a kinetic enrichment protocol described in Khil et al. (2012) for enriching single-strand DNA in the sample. DMC1 ChIP-seq was carried out on pooled testes tissue (from 15 to 18 different individuals) shown to be enriched for primary spermatocytes, as well as on pooled liver tissue as a meiotic recombination negative control (Figure 3.7).

DMC1 signal track from two successful trials, 1) 15 pairs of freshwater fish testes pooled, 2) 18 pairs of marine fish testes pooled are shown in Figure 3.8 a and b respectively. From these results we find an enrichment of DMC1 signal at the ends of the chromosome which is in agreement with male crossover landscape observed from pedigree sequencing. We notice that, signal to noise ratio in our data is smaller than what is observed in other organisms (Khil et al. 2012; Pratto et al. 2014). We speculate the following three reasons for the weak signal observed here: 1) Overall DSB count per meiosis could be lower in sticklebacks compared to other model organisms studied till date. In mice nearly 200-300 DSBs are formed per nucleus per meiosis (Cole et al. 2012a; Kauppi et al. 2013) 2) Higher inter-individual variation makes more diffused rather than punctate signal in this pooled 18 individual testes ChIP. 3) experimental inefficiency due to small number of cells at the right stage in the testes that we used here

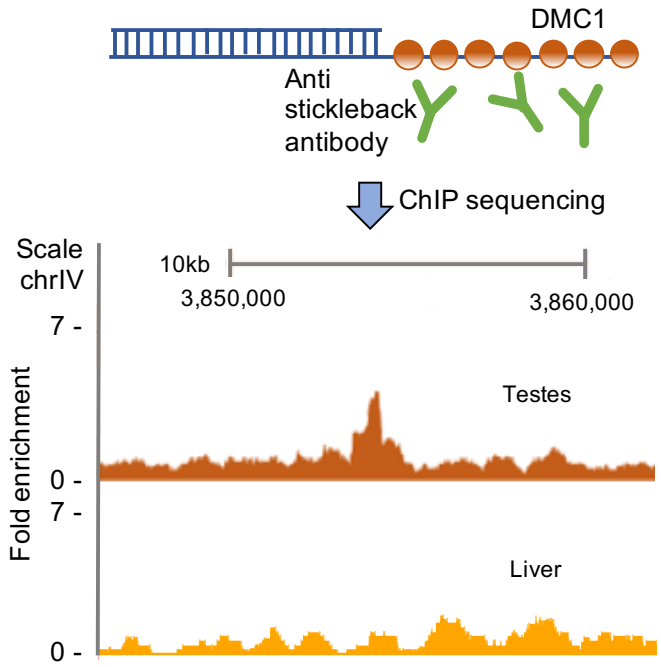


Figure 3.7:DMC1 ChIP sequencing was performed using stickleback specific anti-DMC1 antibody on pooled testes tissue and on pooled liver tissue. Mapped reads on a 12 kb region in chrIV with a testes specific ChIP peak is shown as an example.

Interestingly, we find large number of peaks with intense signal at the sex determining region (non-PAR region) of the stickleback sex chromosome (chrXIX 2.5 to 20 Mb). This could be due to the fact that, double strand breaks indeed occur at the non-PAR region of male sex chromosome but those DSBs take longer time to resolve due to the lack of homologous chromosome (Lu and Yu 2015). Prolonged DSB repair increases the probability of their detection. This also suggest that, shorter half life time of DSBs in rest of the genome makes it difficult to detect most of the autosomal hotspots. Therefore, we speculate that the DMC1 signal detected here possibly represent the most the intense DSB hotspots in the male genome.

In order to define DSB hotspots, we relied on the DMC1 ChIP peaks called using MACS2 program. Testes DMC1 ChIP peaks intersected with liver DMC1 ChIP peaks were excluded to make a list of testes-specific DSB hotspots. Hotspots identified from two of the testes ChIP trials are combined for the rest of the analysis. With this strategy, a total of 1090 DSB hotspots with median size of 1.4kb were identified across the genome. About 72.2% of those hotspots were in the first or last 15% of the individual chromosomes. This is in agreement with 71.5% of male crossovers from pedigree data observed in the first or last 15% of the individual chromosomes, and therefore suggest that, male crossover landscape broadly mirrors the DSB landscape.

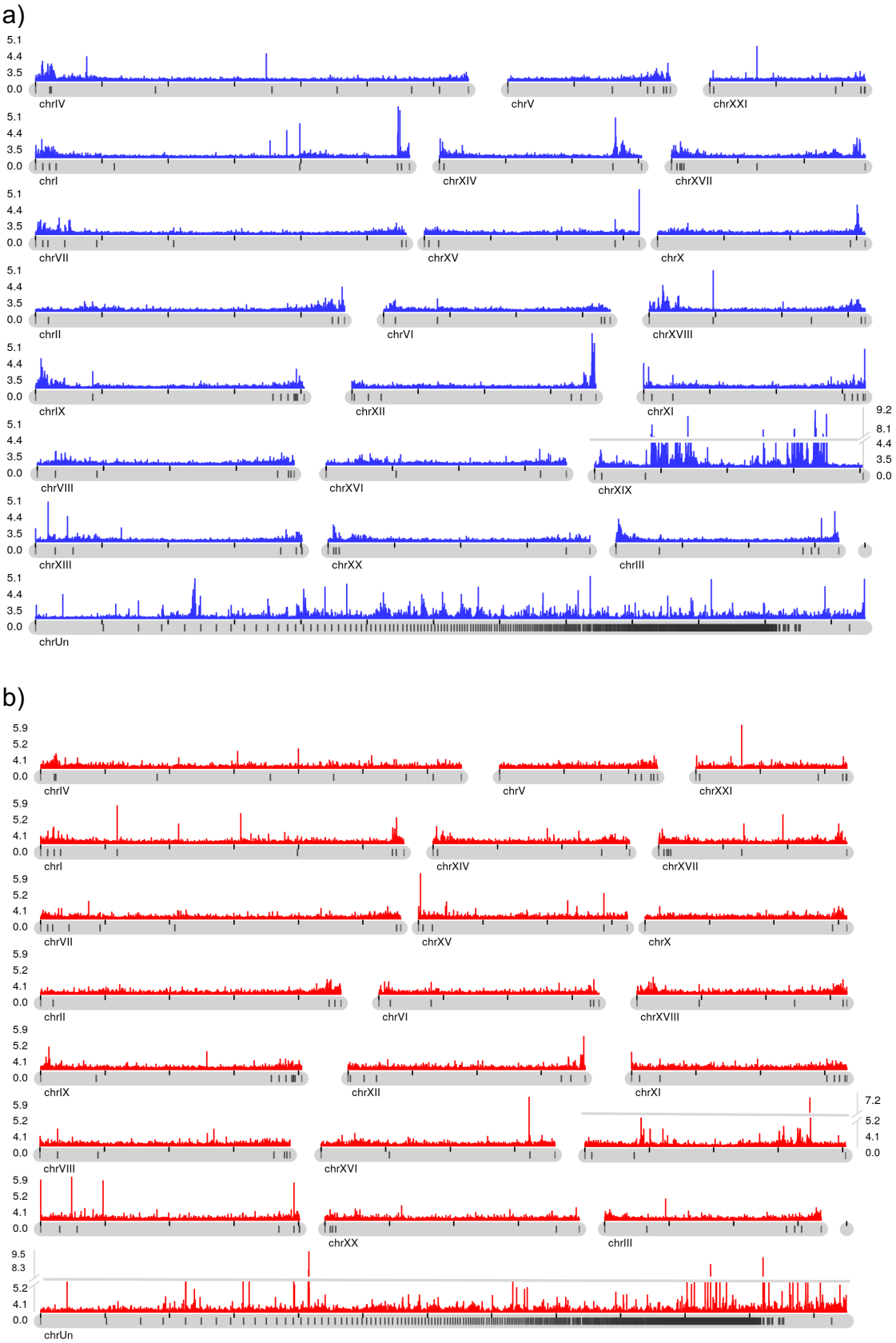


Figure 3.8: DMC1 Signal track from pooled freshwater (a) and marine (b) testes. Cubed fold enrichment of ChIP sample reads over input control reads is plotted.

However, we find that, only 38.37 % of DSB hotspots have a male crossover identified from pedigree analysis within 5 kb distance. DSB hotspots identified in non-PAR region of sex chromosome were not included in this analysis, since male recombination is absent there. Figure 3.9a shows the distance between DSB hotspot and nearest male crossover midpoint. We find that, only one third of stickleback male DSB hotspots are associated with crossovers while the remaining two thirds of DSBs may lack an association because they are repaired via the DSB non-crossover resolution pathway. However, compared to other organisms, the DSB-CO association observed in our data set is quite high. In *Arabidopsis* only 5% of DSBs are found to be resolved as a CO (Choi and Henderson 2015) where as it is nearly 10% in mice (Cole et al. 2012a) and about 58% in budding yeast *S. cerevisiae* (Mancera et al. 2008). We note that, in sticklebacks the true ratio of crossover resolved vs non-crossover resolved DSBs could be different from the number reported above, as the DSB and CO maps do not originate from the exact same biological sample. In addition, the defined DSB hotspots could be an underestimation due to the technical/biological factors mentioned above, which causes weak signal to noise ratio. In *S. pombe*, it has been reported that less intense DSB hotspots are preferentially repaired as crossovers (Fowler et al. 2014), therefore it is also possible that a good fraction of CO generating DSBs were undetected here due to the weak signal to noise ratio.

Next, we analyzed the association between the DSB hotspots identified from DMC1 ChIP-seq and open chromatin signal in stickleback testes tissue. The distance from DMC1 peaks to the nearest gene promoter, H3K4me3 peak and ATAC-seq peak is shown in Figure 3.9b. We find that, 54.2% of DSB hotspots have at least one of the open chromatin features within close proximity (<2 kb distance) which is significantly more than expected by chance (Wilcoxon Rank sum test p value < $2.2x10^{-16}$). However, this also means that, rest of 45.8% of DSB hotspots does not have an open chromatin signature proximal to them. This large proportion of DSB hotspots without an open chromatin signal nearby is intriguing because nucleosome depletion is thought of as a primary requirement for DSB formation.

However, we also note that lack of experimental efficiency in identifying all open chromatin region in a pool of cells by ATAC-seq method could also lead to a similar observation. Even though we used a cell population which is mostly enriched for primary spermatocytes (distinguished based on its larger size), it is possible that only a small fraction of them were exactly at the required stage of meiosis. If there is fast and transient chromatin remodeling in meiotic cells, reads arising from a small proportion of cells at the right stage may not show a strong read pileup, and as a result we do not detect ATAC-seq peaks at those regions. In future, single cell ATAC sequencing might enable us to investigate whether such important variation exist among cells at different stages of early meiosis. In such a

scenario, more sensitive approaches might be required to map chromatin profile of cells at the leptotene stage of meiosis I.

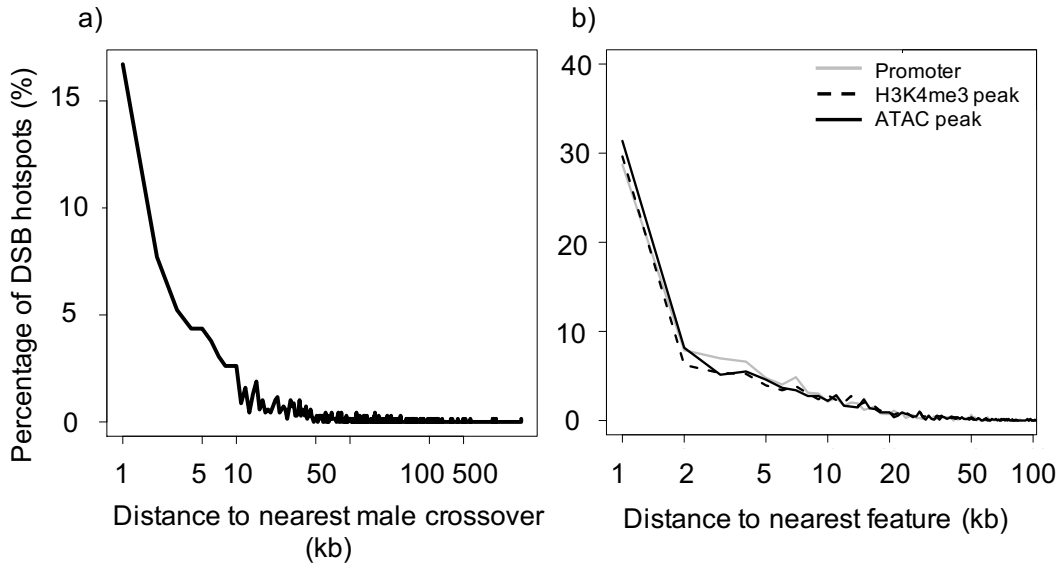


Figure 3.9: Association of DSB hotspots with crossovers and nearest open chromatin features. (a) Distance between DSB hotspot and its nearest male CO identified from the pedigree analysis is plotted. (b) Distance between DSB hotspots and nearest gene promoters, H3K4me3 peaks and ATAC-seq peaks is plotted

The DSB landscape studied in a handful of organisms show differing levels of associations with promoter, nucleosome depleted regions (NDR) and H3K4me3 marks. In budding yeast (*S.cerevisiae*), 88.2% of mapped DSB hotspots overlap with gene promoters and essentially all hotspots had low nucleosome occupancy (Pan et al. 2011). Later with the same data set it has been shown that non-promoter-associated DSB hotspots show no obvious enrichment for H3K4me3 marks (Tischfield and Keeney 2012). In contrast to *S.cerevisiae*, a relatively small proportion of DSBs are found to be associated with promoters (23%) and nucleosome depleted region (36%) in *S. pombe* (Fowler et al. 2014). In mammals with PRDM9 directed DSB landscape, more than 83% of mice DSB hotspots overlapped with H3K4me3 marks. Whereas in PRDM9 knock out mice, 94% of hotspots overlapped with H3K4me3 marks (Brick et al. 2012). In humans 57% of DSB hotspots found to coincide with H3K4me3 marks in testes (Pratto et al. 2014). Our observations suggest that sticklebacks DSB landscape shows more resemblance with the patterns observed in *S. pombe*.

3.3.5 Motif search

DNA sequence motifs have been shown to be associated with crossover or DSB landscape in many organisms. In species with PRDM9, 13-39 bp PRDM9 motifs have been identified in almost all its DSB hotspots (Pratto et al. 2014). Species that lack PRDM9 including Maize, *Arabidopsis* and even species without hotspots such as *Drosophila* are also reported to have DNA sequence motifs enriched at their

recombination sites, though their specific function and mode of action are yet unknown (Zelkowski et al. 2019). G/C rich trinucleotide repeats, and AT polymer motifs have been identified to be present in up to 86% of crossover sites in maize (Kianian et al. 2018) and are similar to motifs identified in *Arabidopsis* (Wijnker et al. 2013; Shilo et al. 2015). Various different motifs have been identified in *Drosophila* with at least one motif in 97% of the CO sites and the most abundant motif is present in 43% of the sites (Comeron et al. 2012).

Analyses from the previous sections reveal that there is considerable number of crossovers and DSBs occur away from functionally active open chromatin regions. Therefore, we next searched for DNA sequence motifs associated with crossover and DSB sites to find whether there is any sequence feature associated with stickleback recombination. A motif search was carried out separately for male, female high-resolution crossover sites, matching sized regions from low recombination cold regions, and DSB hotspots identified from DMC1 ChIP sequencing. Using motif discovery tool of the MEME suite (Bailey et al. 2009), we discovered the motifs mostly likely to be present in each category. Motifs with E-value (statistical significance of the motif) less than 10^{-19} are shown in Table 3.1 and Table 3.2. Top hits in all category were nearly identical. They were either G polymers or GT/CA di-nucleotide repeat. However, detection of these motifs also in cold regions suggest that they are not crossover specific. The top identified motifs in our analysis are nearly identical to the motifs reported in a previous sticklebacks study (Shanfelter et al. 2019).

3.3.6 GC content at CO intervals/ hotspots

GC content is one of the major features with which recombination rate is found to be highly correlated in many organisms including *S.cerevisiae* (Gerton et al. 2000), *Drosophila* (Marais et al. 2003), *C. elegans* (Marais et al. 2001), and Humans (Kong et al. 2002). It has been mostly thought as a result of increased frequency of GC biased gene conversions at the regions of high recombination (Meunier and Duret 2004; Duret and Galtier 2009). In contrary, in *S. cerevisiae*, it has been shown that the GC content is not driven by recombination, rather, higher GC content itself or other GC rich features are predicted to increase the recombination rate (Marsolier-Kergoat and Yeramian 2009).

If the GC content of the genome is driven by recombination, the conserved recombination landscape over generations increases the GC content at highly recombining domains. Therefore, correlation between GC content and recombination landscape is often considered as a measure of recombination landscape conservation. Here we estimated the GC content using the stickleback reference genome sequence. When the correlation was calculated between the GC content and crossover count at different scales, highest correlation was observed at 1 Mb scale (spearman ρ: 0.47, p value $<2.2. \times 10^{-16}$) and lowest, but significant

correlation was observed at 5 kb scale (ρ : 0.15, p value <$2.2x10^{-16}$). In agreement with results presented in chapter 2, our GC content estimate also indicate a highly conserved recombination landscape at a broad 1 Mb-scale.

Table 3.1 Motifs discovered in male and female crossover intervals

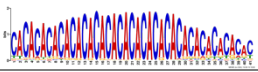
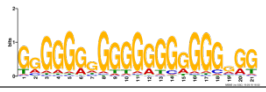
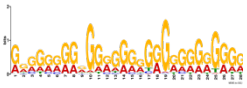
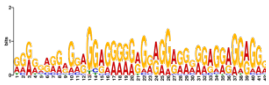
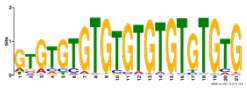
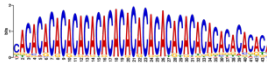
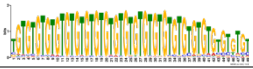
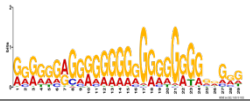
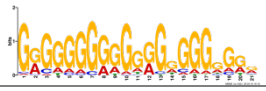
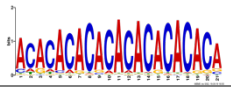
Male			Female		
Motif	**Probability**	**Frequency at CO sites**	**Motif**	**Probability**	**Frequency at CO sites**
	3.1e-090	25.5%		2.4e-139	23.2%
	9.4e-041	51.3%		6.8e-028	39.2%
	2.3e-027	1.6%		5.0e-019	33.6%

Table 3.2 : Motifs discovered in DSB hotspots and crossover cold intervals

DSB hotspots			Cold bins		
Motif	**Probability**	**Frequency at CO sites**	**Motif**	**Probability**	**Frequency at CO sites**
	9.9e-254	28.8%		8.9e-087	9.93%
	1.6e-038	41.1%		1.4e-025	28.9%
	5.8e-035	18.16%		1.0e-022	35.86%

Next, we estimated the GC content within high-resolution crossover intervals and compared it with matching cold intervals (Figure 3.10a). We find that both male and female crossover intervals and 5 kb hotspots have significantly higher GC content compared to matching cold regions and genome average GC content (Wilcoxon rank sum test p value <0.0001). Despite the genome-wide low correlation between GC content and recombination at fine-scale, the enriched GC content within CO intervals and in hotspots compared to cold regions may indicate evolutionary conserved recombination landscape also at fine-scale. To check whether high GC in CO intervals is driven by the subset of promoter associated COs, next we estimated GC content separately for promoter-associated and promoter-non-associated crossovers in males and females (Figure 3.10b). We found that, in fact both categories of COs have significantly higher GC content than genome-wide average estimate in males and females. However, in females

promoter-associated COs have significantly higher GC content than promoter-non-associated COs (Wilcoxon rank sum test p value <0.0001). It is also important to note that the high GC content within these fine-scale CO intervals are not just a reflection of overall higher GC content in recombination rich chromosomal periphery (genome average GC content is 42.98% and mean GC content within first or last 15% of each chromosome is 43.6%, whereas mean GC content within CO intervals is 45.7%). These observations suggest that both in males and females the genomic locations of CO intervals are probably stable over multiple generations irrespective of their association with gene promoter. The low genome-wide correlation observed among individuals at fine-scale is possibly because, our recombination landscape only represents one-generation events whereas the genome wide GC content is shaped over numerous generations and represent a larger set of historical crossover events.

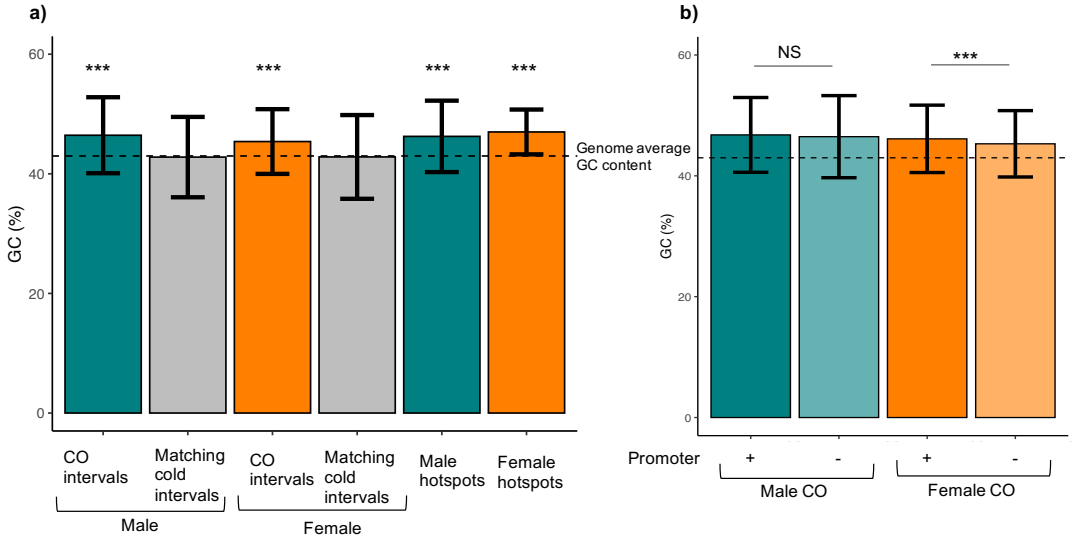


Figure 3.10: Higher GC content within crossover intervals and hotspots. a) Significantly higher mean percentage GC content is observed within male and female crossover intervals compared to matching sized regions in cold intervals (*** : p value <0.0001, Wilcoxon rank sum test). GC content within male and female hotspot intervals are also plotted. b) Mean percentage GC content in promoter associated (+) and promoter non-associated (-) male and female crossovers are shown. The difference between GC content in two categories of crossovers were nonsignificant in males but promoter associated COs have significantly high GC content in females (*** : p value <0.0001, Wilcoxon rank sum test). Dotted line represents genome average GC content. Error bars represent standard deviation.

3.4 Discussion

Our analysis of genomic features associated with stickleback crossover and double strand break landscape revealed several similarities and differences to patterns observed in other organisms. The relatively higher recombination rate observed at the chromosomal periphery is consistent with recombination landscape in almost

all organisms studied so far. In many organisms, broad-scale recombination rate across the genome found to covary with gene density (Civardi et al. 1994; Wu and Lichten 1994; Akhunov et al. 2003; Sidhu 2004). This feature is more pronounced in large plant genomes in which both recombination rate and gene density are higher at the chromosomal periphery. This can be attributed to accessible chromatin which facilitate gene transcription and recombination. We find that in sticklebacks, overall recombination is higher in gene-rich regions compared to gene poor regions but the overall effect is small and there are no specific gene clustering at the chromosomal periphery. This results in poor linear correlation between gene density and recombination rate on a broad genome-wide scale and gene density cannot explain elevated recombination at the ends of chromosomes. Despite this, some chromosomes showed significantly higher correlation between recombination rate and gene density compared to others. It is possible that chromosome-specific factors such as its structure (centromere position and arm length) content (heterochromatin vs euchromatin landscape, repetitive DNA landscape) and epigenetic modifications may influence both gene density and recombination rate. However, it is important to note that, as a result of the chromosome-specific relationship, genes in some chromosomes have more chance to get shuffled than genes in other chromosomes. Therefore, the effect of selection on genes in different chromosomes could be different. Furthermore, low recombination makes some genes inaccessible to trait mapping and selective breeding. It is also important to note that, compared to males, female crossovers show extremely low correlation with gene density suggesting that, gene density itself might not be a driving factor of recombination landscape.

At fine-scale, a fraction of crossovers is targeted towards gene features especially promoters. While this association is more than expected by chance in males, it is not different from a random association in females. However, we also note that, a considerable amount (male: 41.4%, female: 45.2%) of COs, including the intervals identified as hotspots fall within the intergenic region. This suggest that, there could be more features that regulate the stickleback recombination landscape. Therefore, next we examined the association between crossover events and other chromatin features. We find that similar to the observation in budding yeast (Tischfield and Keeney 2012), H3K4me3 correlation with crossover landscape is mostly driven by its association with promoter regions. Once the promoter-associated crossovers are removed, the association with H3K4me3 marks no longer holds. Furthermore, we also find that, majority of the promoter-non-associated crossovers also lack an association with open chromatin signal (ATAC-seq peak). More interestingly, a similar association pattern is observed for stickleback male DSB landscape with the above mentioned genomic/chromatin features. Since recombination is initiated at the DSB site, we would expect DSB site to be more closely associated with its regulatory feature. Surprisingly we find that a large proportion of (45.8%) DSB hotspots are not associated with any of the

mapped open chromatin feature. This observation is rather unexpected given the fact that chromatin openness (nucleosome depletion) is one of the universal factors found to be associated with crossovers and DSB landscape in almost all organisms studies so far. The only known exception is fission yeast *S. pombe* in which nucleosome depletion is not a necessary condition for double strand break formation (Fowler et al. 2014).

These results lead to speculate a fast spatio-temporal remodeling of nucleosome at the DSB and CO site. Earlier studies have speculated that, recombination machinery may opportunistically target an easily accessible region. But a scenario in which nucleosome remodeling is required, the question of what makes it a target site for DSB formation becomes even more relevant. Further experiments including chromatin conformation assays on meiotic cells could complement our chromatin accessibility assay to find other genomic regions associated with the DSB sites and further provide insight into the mechanisms of DSB site designation.

Absence of strong association between all crossovers and functional genomic elements, and lack of functional PRDM9 protein suggest that stickleback recombination landscape at fine-scale might be influenced by some yet unknown mechanism or feature. Higher GC content within fine-scale CO intervals suggest that, those intervals probably have harbored CO events in the past. Based on this observation, I speculate that a repertoire of crossover susceptible fine-scale intervals might exist in stickleback genome that are preferentially targeted for CO formation. The high inter-individual variation we observed could be due to random (or preferential?) choice of CO sites from these larger set of CO susceptible intervals. A preferential picking of intervals for CO formation in sexes and ecotypes could lead to sex-specific and ecotype-specific recombination landscape. However, the lack of male DSB hotspots at the middle of the chromosomes indicates that, sex-specific CO landscape is probably defined early during DSB formation. The major challenge in finding the underlying factors of sex-specific DSB landscape is the difficulty in building a DSB map for females. Due to the small number of primary oocytes present per ovary, isolation of required number of cells at the correct stage is the limiting step. However, based on our results we can say that, a subset of CO intervals in both sexes are targeted towards gene promoters. Further exploration is required to find the yet unidentified recombination modifiers in the stickleback genome.

To understand whether there are any heritable genetic factors associated with stickleback recombination and its hotspot usage, methods like QTL mapping or Genome-wide association studies (GWAS) could be employed. In chapter 4, I discuss a novel individualized crossover mapping method we developed to enable such studies in sticklebacks and in other organisms.

3.5 Materials and methods

For all the association tests described in this chapter, all genomic features were converted to the modified gasAcu1 coordinates. The conversion was carried out using a custom perl script. (described in chapter 1 methods section)

3.5.1 Association with various genomic features

Gene density

Stickleback annotated gene coordinates were downloaded from Ensembl (build 90) and lifted to modified-gasAcu1 coordinates. To estimate gene density, genome was divided into fixed size sliding windows at different scales (from 5 kb to 1 Mb). Number of genes overlapping each window was estimated using BEDtools (v2.27.1) (Quinlan and Hall 2010) coverage function.

Gene features

Gene feature coordinates were obtained from Ensembl (build 90) annotations. In this study the following definition is given to each of the gene feature. Genic region: gene coordinates ± 1 kb, intergenic region: gene less regions at least 2 kb away from transcription start sites (TSS) or transcription end sites (TES), promoter: 1 kb upstream of TSS, TES+1 kb: 1 kb down stream of transcription end sites (all transcripts of a gene are included), exon: all exon coordinates annotated by Ensembl (in the cases of multiple transcripts per gene, all regions which is an exon in at least one of the transcripts are included), intron: exon coordinates subtracted from annotated transcript coordinates. For intergenic region annotation, all gene coordinates were merged using BEDtools merge command and subtracted from the whole genome. Similarly, for intron annotation, all exon coordinates were merged and subtracted from the merged transcript coordinates. Crossovers with at least 5 kb resolution (high-resolution crossovers) were used for this analysis. Occurrence of crossover within genomic feature was quantified using BEDtools (v2.27.1) intersect function. High-resolution crossover coordinates were 10000 times randomly shuffled (using BEDtools shuffle) across the genome and quantified the intersect with gene feature with each iteration to obtain the null distribution.

GC content estimation

GC content within high-resolution (≤5 kb) sex-specific crossover intervals were estimated using BEDtools nuc function. Modified gasAcu1 assembly was used as the reference genome to compute GC content within each interval. As a control for each crossover interval, matching sized intervals from cold regions were picked by shuffling crossover coordinates within cold regions and estimated GC content for those intervals. Genome-wide average GC content was computed by calculating mean of GC content calculated for 5 kb non-overlapping sliding windows across the genome.

Motif search

For finding motifs associated with stickleback recombination, we used a program called MEME (v5.0.2) (Bailey et al. 2015). List of male and female high-resolution crossover coordinates were merged to get sex-specific crossover containing unique genomic regions. As a control, crossover coordinates were randomly shuffled across cold regions to get matching sized coldspot coordinates. In addition, double strand break hotspot coordinates identified from DMC1 ChIP sequencing was also used for motif search. For all categories, FASTA sequence taken from modified gasAcu1 reference genome was used an input in motif search. MEME search was executed with the following parameters. `–dna –oc . –nostatus –time 18000 –mod zoops –nmotifs 50 –minw 6 –maxw 50 –objfun classic –revcomp –markov_order 0 –p 10`. Within each query sequence, motifs with size in between 6bp to 50 bp were searched. MEME ignored motifs if they were present more than once in a sequence (-mod zoops). This is to avoid reporting repetitive motifs. Motif search command was executed separately for each category. In each category, MEME reported motifs in the order of their statistical significance.

3.5.2 ChIP sequencing

In this study, we carried out H3K4me3 ChIP sequencing and DMC1 ChIP sequencing on male testes tissue. Commercially available antibody was used for H3K4me3 ChIP sequencing (Rabbit polyclonal Anti-trimethyl-Histone H3 (Lys4) antibody raised against synthetic peptide from Millipore (cat#07-473)). Whereas due to the lack of suitable commercial antibody that efficiently bind stickleback DMC1 protein, we raised a Guinea pig anti-stickleback DMC1 antibody.

Antibody production and validation

To produce a stickleback specific anti-DMC1 antibody, we expressed the protein of interest in its native state to use as the antigen. In order to have an efficient expression of stickleback protein in the bacterial system we ordered codon optimized cDNA sequence of our protein of interest from GeneArt, Thermo Fisher Scientific. This fragment was then inserted into pCold™ TF-DNA vector from TAKARA clontech (cat #3365). Plasmid vector with cDNA of interest was transformed into expression strain BL21 DE3. Following transformation, single colony was inoculated in 20 ml LB broth containing 100 μg/ml ampicillin for overnight incubation at 37^0C. After incubation, 10 ml of overnight culture was inoculated into 500 ml LB without any antibiotic and incubated at 37^0C until the OD reached 0.4-0.9 range. Once the culture acquired optimal concentration, it was incubated on ice for 30 minutes (pCold™ TF-DNA vector consists of cold shock promoter). Following incubation on ice, 0.5 mM IPTG was added to the bacterial culture in order to induce protein expression. Culture was incubated at 15^0C for overnight. Large scale production of protein was carried out in multiple batches of 500 ml culture. Expressed protein was then purified using Ni-NTA column. Purified TF-DMC1 protein was sent to Eurogentec® for antibody generation.

Guinea pigs are recommended for antibody production of TF tagged antigens. Therefore an 87 days immunization program was followed on Guinea pigs. After 87 days of immunization, antisera was received from the company. Antibody was purified from the serum following affinity purification (using Affi-Gel 15 from Biorad) protocol. The specificity of the antibody was then tested using immunohistochemical staining on stickleback testes and western blot on testes lysate (Figure 3.11).

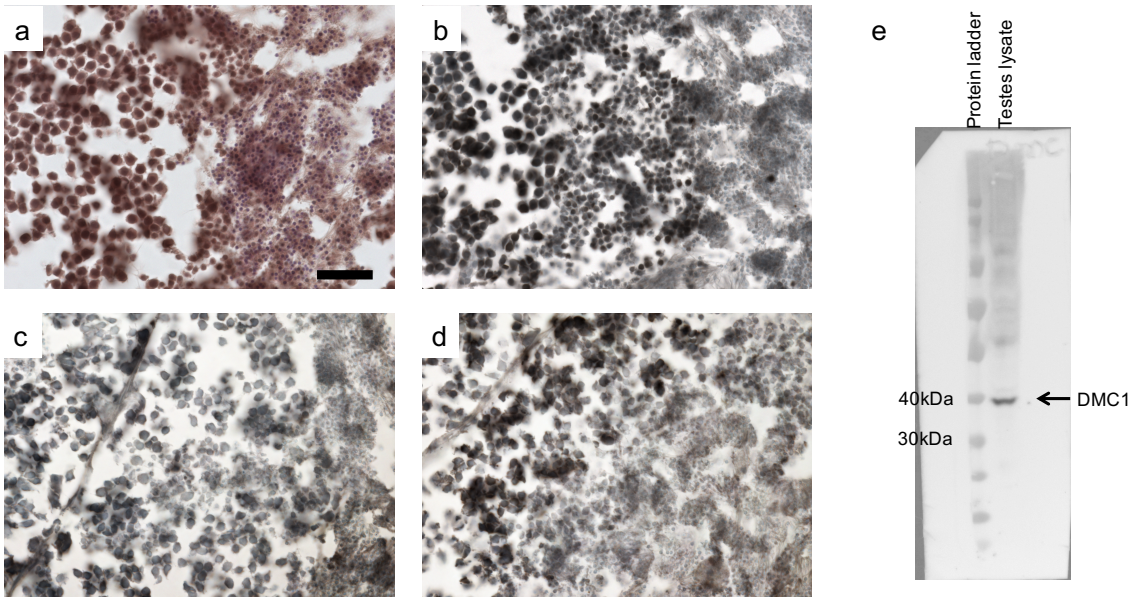


Figure 3.11: Anti-TF-DMC1 antibody validation by immunohistochemical staining and western blotting. Immunohistochemical staining on frozen sections of stickleback testes is shown. a) Haematoxylin-Eosin staining to illustrate cell morphology, primary spermatocytes on the left, secondary spermatocytes on top, matured sperm cells on the right. b) Anti-H3K4me3 staining (positive control). Staining is seen both in primary and in secondary spermatocytes. c) Unstained control (negative control). d) Anti-TF-DMC1 staining. Slight staining is seen specifically in the primary spermatocytes. Scale bar in all images is 20 µm. e) Western blot result. Stickleback primary spermatocyte enriched testes lysate was used as the sample. Antibody used in 1:1000 dilution. A band at ~37 kDa (corresponding to molecular weight of DMC1) is marked with an arrow.

Tissue samples

For all ChIP sequencing described in this chapter, lab bred marine and freshwater sticklebacks derived from River Tyne population in Scotland was used. Testes tissue was collected from males during their non-breeding season. During their breeding season, stickleback males develop bright red color on their throat. Those colored males are separated into a different tank. One month later when they lose the color (at the start of the next cycle of meiosis), fishes were sacrificed to collect testes tissue. Small tissue sample from few random fishes were collected for histology check to confirm the presence of primary spermatocytes in the testes. Once primary spermatocytes are detected in sampled individuals, testes harvest

from all colorless males in that tank was carried out. Tissue were snap frozen in liquid nitrogen and stored in -80^{0}C until use. Similarly, liver tissue was also collected from the males and stored at -80^{0}C until use.

ChIP sequencing

For chromatin immunoprecipitation (ChIP) followed by sequencing of proteins of interest, we used a pool of ~20 testes tissue. A pool of ~20 liver tissue was used as the negative control. A sequential pull down with homemade anti-DMC1 antibody followed by commercial H3K4me3 antibody was carried out with the same tissue lysate. DMC1 ChIP and library preparation were carried out following the protocol described in (Smagulova et al. 2011; Khil et al. 2012). The specialized KE treatment step in DMC1 ChIP for single stranded DNA enrichment was excluded for H3K4me3 ChIP. Library was sequenced in Illumina Hiseq 3000 with 150 bp paired end cycle. After sequencing, DMC1 ChIP reads were trimmed to 40 bp using Trimmomatic (Bolger et al. 2014) and a specialized bioinformatic pipeline described in (Khil et al. 2012) was used for processing the reads. Peaks in the ChIP data in comparison with input was called using MACS2 (v2.1.1) (Zhang et al. 2008) with the following parameters. `-q 0.1 --nomodel --slocal 5000 --llocal 10000 --extsize 800 -f BED --SPMR -g 463000000 -B.` However, no specialized processing was done for H3K4me3 ChIP reads and directly called peaks using MACS2 (using default parameters).

Chapter 4

Genome-wide recombination map construction from single individuals using linked-read sequencing

4.1 Article citation

Dreau A, **Venu V**, Avdievich E, Gaspar L, Jones FC. 2019. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun* **10**: 4309

4.2 Declaration of contributions

Author contributions: Venu V. contributed to the experimental design, prepared stickleback HMW DNA, optimized the linked read library preparation protocol, prepared all 24 sequencing libraries used in this project with the help of Avdievich E., performed supporting data analysis and equally contributed in the manuscript writing along with Dreau A. and Jones F.C. **Relevance to the collective work:** Optimization of the experimental setup and generation of the underlying data are crucial for the successful demonstration of the novel method. In addition, contribution of Venu V. was also relevant for determining various criteria and parameters used in the ReMIX bioinformatic pipeline.

Co-author contributions: Dreau A. contributed to the experimental design, wrote the ReMIX pipeline, analyzed the results, prepared figures and tables, and equally contributed in manuscript writing. Avdievich E. contributed in optimizing the linked read library preparation protocol and helped with preparation of HMW DNA and sequencing libraries of 10 samples. Gaspar L. contributed to experimental design and prepared HMW DNA from mouse gametic and somatic tissue. Jones F.C. developed the idea of using linked-read sequencing of gametes to build individualized recombination maps, contributed to the experimental design, performed additional data analyses, designed figures, and contributed in the manuscript writing.

4.3 Full article

Genome-wide recombination map construction from single individuals using linked-read sequencing

Andreea Dréau[1], Vrinda Venu[1], Elena Avdievich[1], Ludmila Gaspar[1] & Felicity C. Jones[1]

[1]Friedrich Miescher Laboratory of the Max Planck Society, Max-Planck-Ring 9, 72076 Tübingen, Germany.

Correspondence and requests for materials should be addressed to F.C.J. (email: fcjones@tuebingen.mpg.de)

Abstract

Meiotic recombination rates vary across the genome, often involving localized crossover hotspots and coldspots. Studying the molecular basis and mechanisms underlying this variation has been challenging due to the high cost and effort required to construct individualized genome-wide maps of recombination crossovers. Here we introduce a new method, called ReMIX, to detect crossovers from gamete DNA of a single individual using Illumina sequencing of 10X Genomics linked-read libraries. ReMIX reconstructs haplotypes and identifies the valuable rare molecules spanning crossover breakpoints, allowing quantification of the genomic location and intensity of meiotic recombination. Using a single mouse and stickleback fish, we demonstrate how ReMIX faithfully recovers recombination hotspots and landscapes that have previously been built using hundreds of offspring. ReMIX provides a high-resolution, high-throughput, and low-cost approach to quantify recombination variation across the genome, providing an exciting opportunity to study recombination among multiple individuals in diverse organisms.

Introduction

Recombination is an essential process during meiosis. Chromosome segregation often occurs through crossing over, which involves reciprocal exchange among homologous chromosomes and plays an essential role in meiotic chromosome segregation in sexually reproducing organisms. By shuffling parental alleles to produce novel haplotypes it is also a key source of genetic diversity that has considerable implications for the genomic landscape of variation and the evolutionary process.

In most diploid organisms, recombination is functionally constrained by the necessity for at least one recombination event per homologous chromosome pair (this ensures proper segregation during Meiosis I) (Fledel-Alon et al. 2009). Defective, excessive, or deficient recombination can cause inviable gametes and developmental abnormalities (Hassold and Hunt 2001; Inoue and Lupski 2002). For these reasons the number of crossovers and their genomic locations are thought to be tightly regulated and highly constrained (Wang et al. 2015b).

Despite this core functional constraint, recent studies have revealed remarkable variation in recombination at multiple different scales (between and along chromosomes, among individuals, sexes, populations, and species/taxa) (Koehler et al. 2002; Ptak et al. 2005; Coop et al. 2008; Paigen et al. 2008; Kong et al. 2010; Dumont et al. 2011; Comeron et al. 2012; Nachman and Payseur 2012). Crossovers are not uniformly distributed across the genome and the frequency (recombination rate), can vary by orders of magnitude and involve genomic hotspots and coldspots. For example, a well-studied recombination hotspot (Hlx1) on mouse chromosome 1 has a remarkably high recombination rate of 2.63 cM within a narrow 2.8 kb interval in F1 hybrid male mouse (C57BL/6J x CAST/EiJ), yet is relatively colder in females of the same background and among other strains (Paigen et al. 2008). This among strain variation is partly attributable to the strain genotype at the trans-acting recombination modifier protein PRDM9. Conversely recombination coldspots with a lack of crossovers in genomic regions as large as 41 Mb have also been reported (Ma et al. 2010; Fernandez et al. 2014).

Part of the extensive variation in recombination among organisms may stem from the impact of recombination on individual fitness and rates of adaptation in natural populations— in addition to its fundamental role in meiosis, recombination impacts the inheritance of linked alleles, and its modifiers may be subject to different selection pressures in different populations and taxa. Depending on the evolutionary context, recombination may be beneficial if it breaks down linkage between deleterious and beneficial alleles (known as the Hill–Robertson effect) (Hill and Robertson 1966; Felsenstein 1974), or deleterious if it breaks linkage between two adaptive alleles (Kirkpatrick and Barton 2006).

With the knowledge that number and genomic location of recombination can influence the segregation of traits, fitness of an organism, and adaptation in

natural populations, there is increasing interest in the fields of medicine, agriculture, and evolutionary genomics in the empirical quantification of fine-scale variation in recombination among individuals, populations, and species. Despite diverse approaches (linkage-maps, high density genotyping of pedigrees, and individual sperm typing/ sequencing), empirically quantifying recombination variation within and among individuals remains a challenge due to the expense and data intensity required to build numerous individualized genome-wide maps of recombination rate (Li et al. 1988; Broman et al. 1998; Kong et al. 2002; Carrington and Cullen 2004; Kauppi et al. 2004; Shifman et al. 2006; Paigen et al. 2008; Dumont et al. 2011; Wang et al. 2012; Smeds et al. 2016). Other less data intensive approaches, such as comparisons of recombination among taxa using statistical estimates of recombination from population genetic (polymorphism) data, provide population and sex-averaged historical estimates of recombination rate and can be confounded by differences in the demographic history of the taxa and differences in the effective population size of the local genomic regions being compared. Further, these averaged estimates make genetic dissection of molecular mechanisms underlying recombination variation difficult. In this study, we address these challenges by introducing a new and powerful low-cost method that quantifies empirical recombination events across the genome of a single individual using linked-read sequencing of gametes.

Linked-read libraries are generated from long (high molecular weight (HMW)) DNA molecules using a 10X Genomics Chromium controller. Numerous short reads are produced from DNA molecules encapsulated inside nanoliter-sized droplets. Using their droplet-specific barcode these short reads can be computationally reconstructed into single molecules after Illumina sequencing. This low-cost long-range information can be used to solve the problem of haplotype determination. Our pipeline called ReMIX mines the long-range information in linked-read data to identify recombination crossovers across the genome. ReMIX makes use of some parts of the 10X Genomics pipeline, Long Ranger (Zheng et al. 2016), but deviates from it in a number of important ways. Long Ranger aligns reads to a reference sequence, calls and haplotype phases SNPs, reconstructs molecules, and identifies indels and large-scale structural variants. It makes use of molecules that have a high probability of assignment to only one haplotype phase. Molecules that contain reads of mixed haplotype assignment (some reads assigned to one haplotype while others are assigned to the alternate haplotype), are considered to be errors and are discarded. However, when sequencing linked-read libraries from gamete DNA these haplotype switching molecules can also represent a valuable fraction of molecules spanning meiotic recombination crossovers. ReMIX identifies these valuable molecules and is the first method to enable reconstruction of individualized genomic recombination landscapes using linked reads.

The linked-read information is exploited by ReMIX during three steps: identification of high-quality heterozygous variants, reconstruction of molecules, and the haplotype phasing of each molecule. The molecules identified as recombinant are then used to build an individualized genomic map of recombination crossovers, enabling us to quantify recombination variation across the genome.

We demonstrate our method using gametic tissue from a hybrid mouse (*Mus musculus domesticus* × *Mus musculus castaneus*) and a stickleback fish (*Gasterosteus aculeatus*). Genetic maps, available for both organisms, allow us to evaluate the accuracy of ReMIX. To validate the precision of our pipeline, we also use samples from the somatic tissue of the tested individuals as a negative control, as well as simulated data to determine the sensitivity and specificity of our method in genomes with different levels of polymorphisms. Using data from only a single individual and without prior knowledge of polymorphic sites, ReMIX obtained results that follow the same pattern of the previously described recombination maps, but with considerably higher resolution of the detected crossovers and lower costs compared to previous methods.

Results

Linked-read sequencing of pools of gametes. The novel method and algorithm that we present in this study uses pooled gamete DNA as starting material and reliably identifies recombination landscape of an individual at the whole genome level. Here we report the complete pipeline and results obtained by applying our method to an individual C57BL/6Ncrl × CAST/EiJ hybrid mouse and freshwater stickleback fish. HMW DNA (>40 kb) was extracted from purified sperm cells and somatic tissue of both mouse and fish individuals (spleen and kidney, respectively). 10X Genomics linked-read genomic libraries were prepared on a Chromium controller and the resulting linked-read libraries were sequenced on an Illumina HiSeq3000 sequencer. Reads obtained from the sequencer were then processed through our ReMIX pipeline to identify recombinant molecules and quantify the genomic recombination landscape of each individual.

Overview of the ReMIX algorithm. ReMIX requires linked-reads generated from haploid gamete DNA as input. From meiotic division, a haploid gamete comprises of a single copy of each chromosome in the genome—products of reductional cell division that are recombinants of the diploid parental chromosomes. Of the millions of linked-read molecules sequenced, the majority will be assigned with high probability to one of the two parental haplotypes. A small fraction of molecules (those spanning recombination crossovers) will contain reads that switch between the haplotypes (Figure 4.1a). The role of ReMIX, after filtering and phasing, is to identify the rare fraction of recombinant molecules as those that switch between haplotypes (Figure 4.1b). For this, our pipeline aligns the linked-reads to a reference genome sequence in order to identify high-quality

heterozygous variants and to reconstruct the original molecules. After phasing the variants using the molecule information, the phase of each molecule is computed based on its reads spanning heterozygous-phased variants. Since the total number of sequenced gametes is high and the resulting per base coverage is high, the read coverage of each individual molecule can be considerably lower without compromising performance (<0.5x). Thus, the correct phasing of a maximum number of molecules by ReMIX is a function of the ratio between the density of heterozygous variants in the focal individual and the number of reads per molecule. In the end, the identified molecules are separated into those that are entirely non-recombinant (haplotype 1 or 2 molecules), or alternatively, recombinant (haplotype switching) molecules (full details in the "Methods" section).

A haplotype switching molecule may be generated from a true recombinant molecule or alternatively represent a false positive caused by bioinformatic errors, such as sequencing error, incorrect read mapping, structural variation, or barcode sharing among molecules from the same part of the genome. Our pipeline therefore incorporates several filtering steps to remove false positive recombinant molecules. ReMIX initially filters the linked reads based on the barcode sequence and the quality of the read. After variant calling the variants are filtered to remove polymorphisms showing allelic bias, and after molecule reconstruction, molecules with extreme high or low coverage are removed. Finally, after the haplotype phasing of molecules, genomic regions that are not covered by a similar number of molecules for each haplotype are removed. These filters allow us to remove the regions that can introduce errors in the mapping or the phasing, such as copy number variation, small deletions, inversions, translocations, etc. Finally, the ReMIX pipeline identifies molecules that have a high probability of containing a real crossover (e.g. stickleback mean probability 0.982 ± 0.068SD, source data provided as a Source Data file) along with the genomic position of that crossover.

By considering the quality of each base within a molecule, requiring at least three variants representing each haplotype, and ≥70% of reads on each side of a switch phased to the correct haplotype, ReMIX allows small erroneous switches in haplotype state within a recombinant molecule caused by single read low-quality base calls. Information on the location of recombinant molecules is then used to build an individualized genomic map of recombination crossovers (Figure 4.1c).

Identification of known hotspots in mouse. The genomic recombination landscape is well studied in various laboratory mouse strains, with one of the highest resolution sex-specific recombination maps constructed in Paigen et al. (Paigen et al. 2008). Focusing on chromosome 1, the authors genotyped 6028 progenies produced from C57BL/6J × CAST/EiJ and CAST/EiJ × C57BL/6J hybrids, mapped the locations of 5742 crossover events, and revealed the presence of a number of highly localized sex-specific recombination hotspots (Paigen et al. 2008).

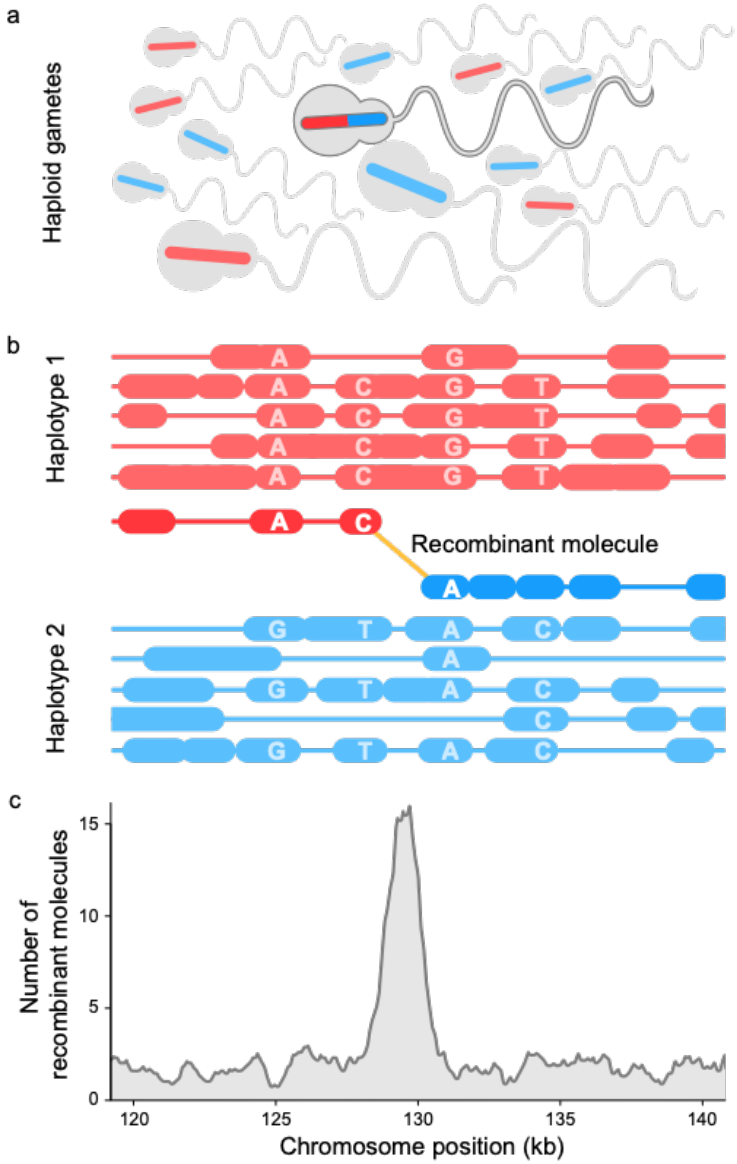


Figure 4.1: Construction of individualized genomic recombination maps using ReMIX. a) DNA is isolated from a pool of sperm where each cell represents a haploid product of a single meiotic event. Sperm with recombinant chromosomes are shown carrying bars colored both red and blue, while non-recombinant chromosomes are shown as solid red or blue. b) ReMIX identifies high-quality heterozygous variants, reconstructs molecules, then determines their haplotype phase. Three categories of molecules are identified: those belonging to haplotype 1 (red), haplotype 2 (blue), and recombinant molecules that switch from one haplotype to the other. Each contiguous line represents a molecule with the linked-reads marked by thick blocks. c) Identified recombinant molecules are used to quantify the recombination rate across the genome.

To evaluate the performance of our ReMIX pipeline, we analyzed linked-read libraries produced from the sperm, and as a negative control, somatic tissue from the spleen, of a single C57BL/6Ncrl × CAST/EiJ hybrid male. We then compared ReMIX results with the high-resolution recombination map from the 1479 C57BL/6J × CAST/EiJ male progeny (Paigen et al. 2008).

Whole genome linked-read libraries were generated from sperm and somatic cells in order to sample a similar number of recombinant molecules on chromosome 1 as reported in Paigen et al. (Paigen et al. 2008).We prepared six parallel reactions using the 10X Genomics Chromium controller—each with ~1.2 ng of DNA, approximately corresponding to a total of ~1700 haploid genomes. The final libraries were selected for an average of 600 bp insert size and sequenced at 170x coverage with 2 × 150 bp paired reads on an Illumina HiSeq3000 giving an expected read coverage per individual molecule of ~0.1x. Both sets of linked-reads were analyzed using ReMIX and the latest version of the mouse reference genome, NCBI Build 38 (mm10) [GCF_000001635.20].

A crude estimate of the expected number of recombinant versus non-recombinant molecules can be made: for linked-read libraries made from a single gamete with an average molecule size of 60 kb, sex-averaged map lengths of ~1630 cM (genomewide) and 96.55 cM (chromosome 1) (Cox et al. 2009), and assembled genome size of 2.9 Gb, we might expect to find recombinant molecules spanning crossovers at a frequency of 3.3×10^{-4} and 1.8×10^{-5}, respectively (16.3 and 0.9 recombinant molecules in a genomewide total of 48,333 molecules from a single gamete). In a pool of 1700 gametes (equivalent to the number of gametes sequenced here), we expect to uncover 27,710 recombinant molecules across the genome, with roughly 1641 of these located on chromosome 1.

After stringent filtering of the sperm sample ReMIX retained 1210M reads and reconstructed 148M molecules with an average of eight linked-reads per molecule. A total of 30,508 (0.02%) molecules were identified as recombinant (genome-wide) and 2369 of these were located on chromosome 1. Crossover positions of the recombinant molecules cluster into hotspots in a pattern closely mirroring the previously described male recombination map (Paigen et al. 2008) both in terms of position and intensity (Figure 4.2a and Supplementary Figure. 2).

Accounting for false positives (see below), we see a number of windows that have significantly more crossovers than expected by chance (Wilcox rank sum test, $p < 9.72 \times 10^{-20}$), suggesting the presence of hotspots in the mouse genome. In contrast, recombinant molecules detected in the somatic sample are less frequent, have a dispersed distribution and likely reflect false positives (discussed further below) from sequencing and/or bioinformatic errors (e.g. barcode collision) or rare mitotic recombination events. At the well-known recombination hotspot region Esrrg1 (chr1:188,078,656–188,081,229, mm10) (Paigen et al. 2008; Billings et al. 2013) ReMIX identified 33 recombinant molecules in the sperm sample (Figure 4.2b), while no recombinant molecules were identified in the corresponding genomic region in the somatic sample (Figure 4.2c). Compared with previous studies involving more than 1500 mouse offspring, our results indicate that ReMIX is a powerful method for reconstruction of the fine-scale recombination landscape using gametes from a single individual.

We used both the number of recombinant molecules detected in the somatic sample and simulations (described in detail in below) to obtain independent estimates of the false-positive rate. Adjusting ReMIX results according to these estimations, our data suggest the total number of true crossovers along chromosome 1 in 1700 sperm to be 1540, giving an average of 0.9059 crossovers per meiotic product.

This corresponds well to the sex-averaged genetic map length of mouse chromosome 1 (90.9 cM), but is 8.3 cM longer than the map length of hybrid C57BL/6J × CAST/EiJ and hybrid CAST/ EiJ × C57BL/6J males: 81 and 83.65, respectively, calculated from Table S1 of Paigen et al. (Paigen et al. 2008). This slightly higher number of observed recombinant molecules than expected based on the hybrid male map may have a biological basis (e.g. inter-individual variation (Koehler et al. 2002), inter-strain variation C57BL/6J vs. C57BL/6Ncrl (Koehler et al. 2002; Fontaine and Davis 2016), and possible differences arising from quantification of recombination from viable offspring vs. quantification of recombination from gametes) or alternatively stem from detection errors (e.g. false negatives in the Paigen study (Paigen et al. 2008) due to lack of markers in the telomeric regions).

Finally, it has previously been shown that the genomic recombination landscape in mouse is positively correlated with CpG island density (Han et al. 2008). Here, we also find that recombinant molecules recovered by ReMIX are significantly closer to CpG islands than expected by chance based on 1000 permutations (Wilcox rank sum test, $p < 2.5 \times 10^{-20}$).

Fine-scale recombination landscape in stickleback fish. We next evaluated the performance of ReMIX in an organism that has a recombination landscape with hotspots less intense than mouse. The threespine stickleback fish is an evolutionary genomics model organism with reasonably high-quality genome assembly, for which the recombination landscape has been previously described (Roesti et al. 2013; Glazer et al. 2015; Sardell et al. 2018). To match the mouse sample, we created gametic and somatic linked-read libraries each using 0.8 ng of HMW DNA (approximately equivalent to 1700 gametes) from sperm and kidney tissue of a freshwater Scottish stickleback strain (River Tyne).

The libraries were selected for a mean insert size of 600 bp and sequenced at 170x coverage on an Illumina HiSeq3000 machine. Similar to the mouse sample, the expected read coverage per molecule is ~0.1x. Both sets of linked-reads were analyzed using ReMIX and the stickleback reference genome (BROAD S1 (Jones et al. 2012), split into assembled scaffolds). 178M reads were retained post filtering and reconstructed into 21M molecules (eight linked-reads per molecule in average) of which 2639 (0.01%) were identified as recombinant by ReMIX.

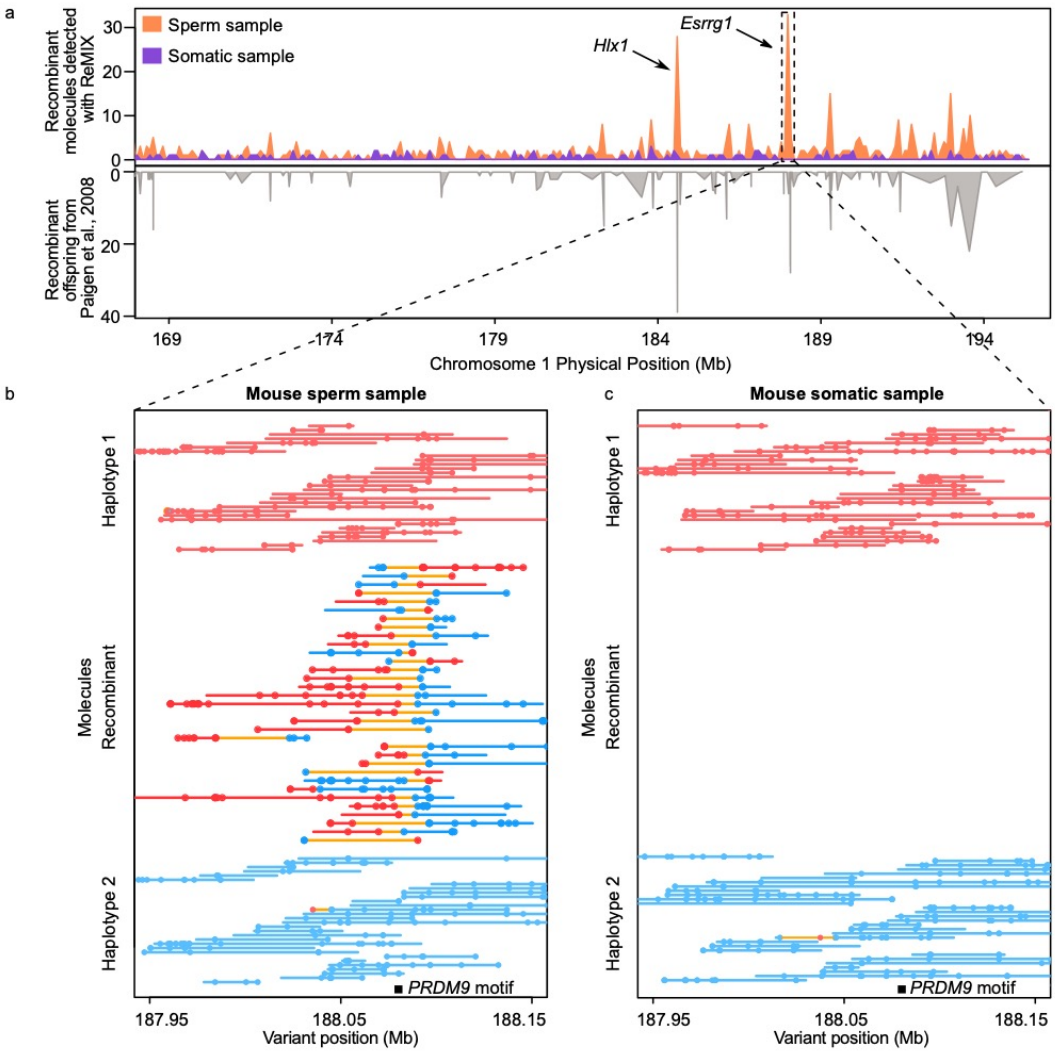


Figure 4.2: ReMIX correctly detects fine-scale recombination variation and hotspots on mouse chromosome 1. a) The recombination rate on the south end of chromosome 1 (169–195.4 Mb, mm10), determined by ReMIX (orange above axis) corresponds well to the rate described in Paigen et al. (Paigen et al. 2008) (gray below axis). As a negative control, somatic tissue (purple) shows a minimal number of dispersed recombinant molecules. b) The three types of molecules identified by ReMIX in the sperm sample in the region of a well-known recombination hotspot (Esrrg1, (Paigen et al. 2008),(Billings et al. 2013)). Each line represents a single molecule and each dot a high-quality heterozygous variant phased as haplotype 1 (red) or haplotype 2 (blue). Joining lines represent the inferred phase of the molecule with orange lines indicating a switch between haplotype states. For graphical reasons, we represented all the recombinant molecules detected by ReMIX but only 30 random (classical) molecules for each haplotype. c) The corresponding region for somatic tissue lacks recombinant molecules. PRDM9 plays a role in initiating crossovers at the Esrrg1 hotspot and has a DNA-binding motif (black bar) located near the midpoint of the detected recombinant molecules.

The stickleback recombination landscape recovered with ReMIX follows the rate inferred from the previous low resolution genetic map (Roesti et al. 2013) (Figure 4.3 and Supplementary Figure 7). Consistent with previous studies, ReMIX reveals recombination crossovers are enriched towards the distal ends of chromosomes

and are significantly clustered compared to random expectations (Wilcox rank sum test $p < 1 \times 10^{-20}$). Similar to the mouse results, ReMIX recovered a number of recombination molecules in the stickleback somatic sample providing an indication of a modest false-positive rate (Supplementary Figure 8). For most chromosomes the maximum number of these false-positive somatic recombinant molecules in 50 kb windows is 2 and we note some heterogeneity in the false-positive rates across chromosomes with elevated levels on chromosomes XIV, XIX, and XXI (as high as four molecules on chrXXI), which co-localize with scaffold ends and are likely scaffold assembly errors.

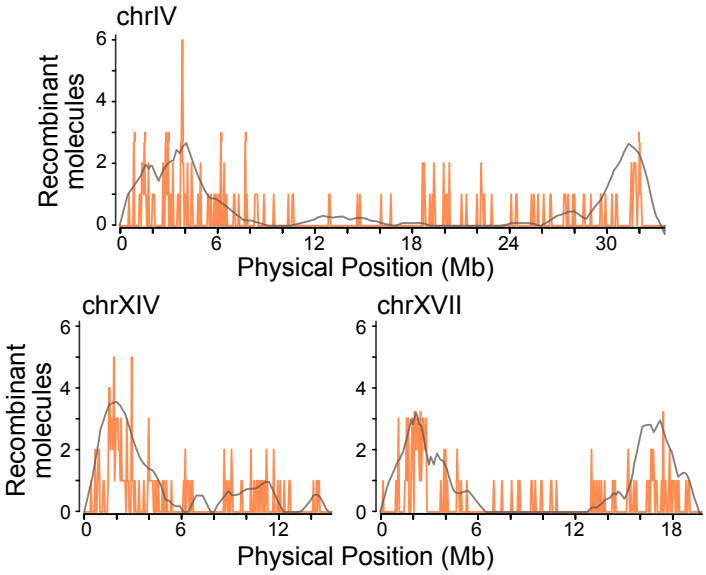


Figure 4.3: ReMIX recombination maps of example autosomes in a male freshwater stickleback. ReMIX analysis of linked-read data is plotted as the number of crossovers in 50 kb windows (orange). For comparison, recombination rate estimates obtained from a F2 lab cross (Roesti et al. 2013) of 140 males and 142 females individuals genotyped at 1872 markers are shown as gray line.

Recombination suppression in inversion heterokaryotypes. When populations adapt to divergent environments in the face of ongoing gene flow, structural rearrangements, such as inversions have the potential to play an important role facilitating and maintaining adaptive divergence. By suppressing recombination in heterozygous individuals, inversions reduce the homogenizing effects of recombination in the local genomic region, allow the maintenance of linkage among neighboring mutations and the further accumulation of genetic differences between populations (Kirkpatrick and Barton 2006; Charlesworth and Barton 2018). They therefore have the potential to act as adaptive cassettes if they harbor and maintain linkage among multiple beneficial mutations.

The recombination suppressing effects of inversions in heterozygotes is a well-known phenomena (Dobzhansky and Sturtevant 1938; Novitski and Braver 1954; Stevison et al. 2011) mediated in part by the abnormal formation of acentric and dicentric meiotic products due to improper resolution of double strand DNA

breaks within inversion loops. The effects of an inversion on recombination can be heterogeneous—varying considerably along the affected chromosome (Sturtevant 1931) with strong suppression around inversion breakpoints, partial suppression in the center of large inversions, and increased recombination in the genomic regions flanking the inversion and even on other chromosomes (Stevison et al. 2011).

Empirical quantification at the individual level of the strength and nature of recombination suppression around inversions requires analysis of a large number of meiotic products from focal individuals. Ideally, for testing the prevailing theory on the role of inversions in local adaptation, recombination should be studied in both alternative ecotypic (collinear) forms known to be undergoing adaptive divergence, as well as hybrids that are inversion heterokaryotypes. By overcoming the challenges related to expense and effort of previous methods, our ReMIX method enables us to investigate recombination variation within and among individuals and species on a scale that would previously have been difficult. A previous study (Jones et al. 2012) identified three large inversions in the stickleback genome that show consistent orientation differences among multiple independent marine and freshwater populations. We focused on >5 Mb window centered one large 1.7 Mb inversion on chromosome XXI containing 76 genes and asked how the structural rearrangement influences the fine-scale recombination landscape within and among individuals, stickleback ecotypes, and their F1 hybrids. Linked-read genomic libraries were prepared from the DNA of ~3400 sperm from each of three marine and freshwater individuals from the Little Campbell River, Canada, and four F1 hybrids. Libraries were prepared and sequenced at the same coverage as described above and ReMIX analysis performed with the parameters fine-tuned for the genomic region and applied to all 10 individuals.

We observed patterns of strong recombination suppression in F1 hybrids compared to their marine and freshwater homozygous counterparts (Figure 4.4). In contrast to the numerous crossover events that were detected within the inversion in marine and freshwater inversion homozygotes, only two recombinant molecules were detected within the inversion among the ~13,600 gametes analyzed across four F1 hybrids (an effective recombination rate < 15% of inversion homozygotes). This pattern differs in the left and right inversion flanks where recombination appears to be comparatively high in hybrids exceeding the recombination rate observed in freshwater ecotypes. Finally, we observed considerable genetic divergence between the inversion orientations indicating recombination suppression substantially reduces the homogenizing effects of gene flow between the diverging ecotypes. The alternate orientations of this inversion therefore have the potential to harbor multiple, linked, beneficial mutations conferring an adaptive advantage to marine and freshwater ecotypes in the wild.

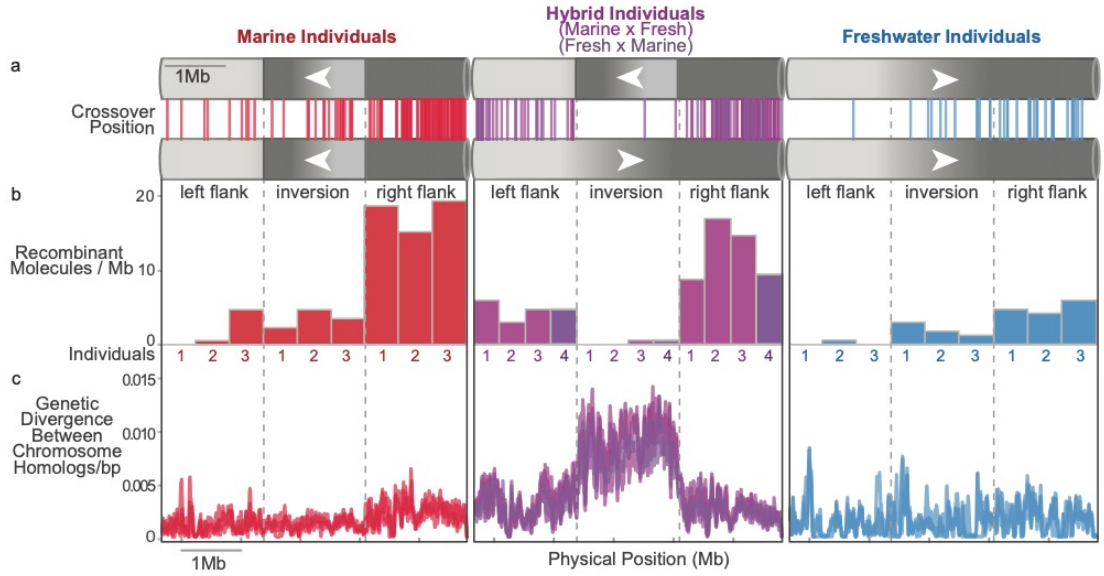


Figure 4.4: ReMIX analysis reveals recombination suppression in individuals heterozygous for a chromosomal inversion. a, b) A 1.7 Mb inversion on stickleback chromosome XXI differs in orientation between marine (red) and freshwater (blue) ecotypes. Meiotic crossovers, detected by ReMIX as recombinant molecules, occur throughout the inversion in individuals homozygous for either orientation (three marine and freshwater fish, respectively). In contrast, inversion heterokaryotypes (heterozygous for inversion orientation, hybrids, N= 4, purple) show recombination suppression within the inversion and elevated rates of recombination in the regions flanking the inversion. c) This recombination suppression allows the accumulation of linked genetic differences between the inverted haplotypes (shown for each individual as the number of heterozygous sites/bp in 20 kb windows).

ReMIX detects crossovers with high genomic resolution. A recombinant molecule is composed of two continuous sections: s_a phased to one haplotype and s_b phased to the opposite haplotype. The crossover may have occurred anywhere between the last informative variant of s_a and the first informative variant of s_b. Thus, we consider the resolution of a crossover as the physical distance between these two informative variants. By taking advantage of long-range molecular data spanning high-quality heterozygous variants segregating within a single individual, ReMIX directly identifies the recombinant molecules with high accuracy and crossover resolution (Figure 4.5). The achievable crossover resolution of our approach is limited primarily by the density of heterozygous sites within an individual (something that varies considerably across taxa), and secondarily by the sequencing coverage used to detect these informative sites. For example, based on whole genome-sequencing data, we estimate hybrid C57BL/6Ncrl × CAST/EiJ mouse and freshwater stickleback individuals used in this study will have a median distance of 44 and 63 bp between heterozygous sites, respectively.

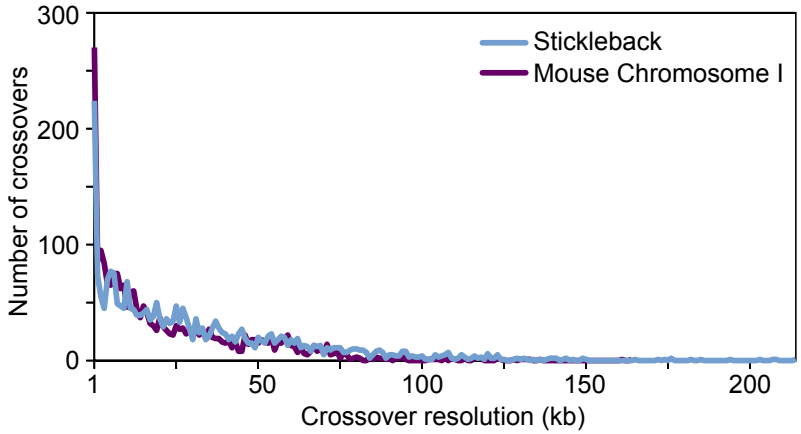


Figure 4.5: ReMIX detects recombination crossovers with high resolution in both mouse (purple) and stickleback (blue). After stringent filtering of reads ReMIX achieved a mean of 8.3 and 8.5 reads per molecule, and a median crossover resolution of 14 and 23 kb for mouse chromosome 1 and stickleback whole genome, respectively. This is considerably higher than previous studies of mice (e.g. median resolution of 225 kb in Paigen et al. (Paigen et al. 2008)) and close to the maximally achievable resolution based on the biological constraint of distance between heterozygous sites in these strains. The highest crossover resolution we achieved was 1 bp in both mouse and sticklebacks, while only 1.22% and 4% of the crossovers detected had resolution as low as 100 kb or more for mouse and stickleback, respectively. We note that if desired, further improvements to crossover resolution up to the biological limit of distances between heterozygous sites could be achieved by increasing the depth of sequencing coverage (and consequently the number of reads per molecule).

Analysis of accuracy on simulated data. Since fine-scale recombination rate can vary considerably among individuals of the same species, comparisons of our ReMIX results with previously published recombination studies provides only a qualitative assessment of the accuracy of our pipeline. To achieve a better indication of ReMIX's performance, we simulated several data sets using the linked-read simulator LRSIM (Luo et al. 2017). Starting from a reference sequence as an input, LRSIM can simulate diploid sequences with a user-specified number of heterozygous SNPs, indels and structural variants. Then the simulator extracts paired end reads from each haplotype and assigns the reads to molecules by attaching the specific 10X barcodes depending on a user-specified number of reads per molecule. In order to validate our method, we generated linked-read sets containing both non-recombinant and recombinant molecules. To achieve this, we first used LRSIM to create a set of linked-reads containing only non-recombinant molecules. Then we simulated crossovers between the two haplotypes (a switch of haplotype state) generated by LRSIM in the first run and we ran LRSIM on the recombinant haplotypes to obtain a second set of linked-reads containing recombinant molecules (those spanning the simulated crossovers). The resulting molecule sets were merged to simulate the mix of recombinant and non-recombinant molecules present in a pool of gametes. The sensitivity (or the true positive rate) is then computed as the proportion of the recombinant molecules

correctly identified by ReMIX out of the total set of simulated recombinant molecules.

Let m be a recombinant molecule with two contiguous segments s_a and s_b phased to opposite haplotypes. ReMIX is able to detect m only if reads from both s_a and s_b are spanning heterozygous variants. Thus, the heterozygosity of the organism and the sequencing coverage are two parameters that influence the sensitivity of ReMIX to detect true positive recombinants. To evaluate the sensitivity, we performed simulations with different heterozygosity levels and read density per molecule. The positions of heterozygous SNPs and reads were chosen randomly for each run. For each parameter configuration we ran the simulations 10 times and averaged the sensitivity values. We show that ReMIX is highly sensitive (with more than 90% of recombinant molecules detected at moderate to high levels of heterozygosity and moderate to high sequencing depth (Figure 4.6).

The percentage of correctly reported molecules slowly decreases with the increase of the distance between the heterozygous variants. This is caused by the lower probability of reads spanning informative variants flanking recombination crossovers when an organism has a lower level of heterozygosity. However, ReMIX sensitivity can be easily increased in those cases by using a higher sequencing coverage.

Similar to other pipelines constructed for processing linkedreads (Zheng et al. 2016; Weisenfeld et al. 2017) the performance of ReMIX is dependent on the reaction conditions during the 10X Genomics linked-read library generation. One major consideration is the probability of two independent molecules from the same locus in the genome being assigned the same barcode (barcode collision). This depends on the amount of DNA in the reaction (which influences the number of molecules per droplet (GEM)), and the genome size of the organism in question. When preparing libraries from the same weight of DNA, small genomes will have a higher molecular copy number of each genomic locus, compared to large genomes. This leads to a higher probability of barcode collision of molecules from the same genomic locus due to a higher probability of them being trapped within the same GEM. In organisms with small genomes, using less DNA in the linked-read library preparation reaction can mitigate the occurrence of barcode collision.

If barcode collision occurs among alternate haplotypes, this has the potential to lead ReMIX to identify false-positive recombinant molecules. Let m_1 and m_2 be two molecules of opposite haplotype state, from the same genomic region, that have the same barcode. The short reads are regrouped into molecules based on their barcode and a parameter specifying the maximal genomic distance separating two reads of the same molecule. Depending on the m_1 and m_2 read positions, the two molecules are detected as one molecule with two contiguous segments phased to opposite haplotypes. As a consequence, ReMIX reports the

merged molecules as a recombinant molecule. Finally, identification of false-positive recombinant molecules might also be caused by erroneous read mapping, structural variants, and reference genome assembly errors.

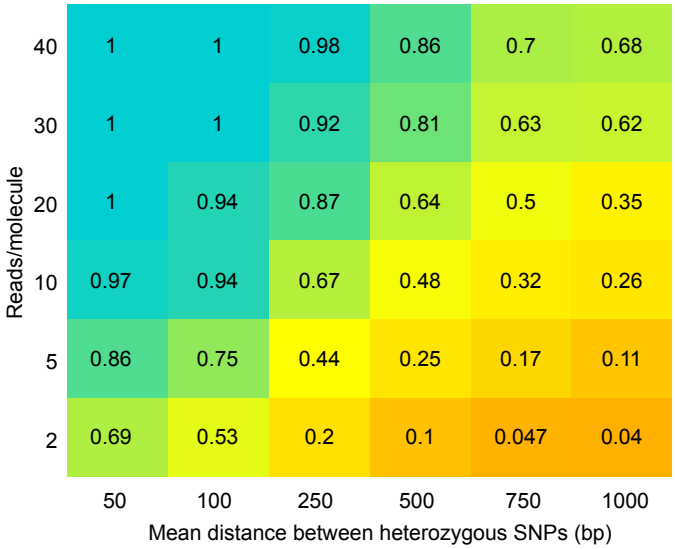


Figure 4.6: ReMIX detects recombinant molecules with high sensitivity. In simulated data, the sensitivity of ReMIX to detect recombinant molecules (low = orange; high = blue; and numbers within boxes) is high at moderate to high heterozygosity levels (x-axis) and moderate to high sequencing depth (y-axis).

To address these two different causes of false positives, we used the complete mouse reference genome (mm10) to simulate a close to real case scenario for the numbers of molecules per GEM and the density and length of structural variants. Using the method described above, we simulated seven molecules per GEM, the mean number of molecules per GEM that we obtained with our empirical mouse and stickleback data sets, and also 10 molecules per GEM, the maximum number reported by 10X Genomics. We then ran ReMIX on both sets and grouped the reported molecules in 100 kb windows (Table 4.1). Under conditions matching our empirical datasets we estimated a low recombinant molecule false-positive rate with a large majority of the intervals not containing any false-positive molecules (94.9%) and only 5% of intervals showing false positives. This increased to 10.37% of intervals when the number of molecules per GEM was simulated to be 10.

The distribution of the intervals containing false-positive molecules for mouse chromosome 1 for 7 and 10 molecules per GEM is shown in Supplementary Figures 9 and 10, respectively. Similar levels of false positives were detected on the other mouse chromosomes. We note that due to the stringent filters of our pipeline (see the "Methods" section), structural variants were filtered, and did not have an impact on the false-positive rate. Since the false-positive molecules are uniformly distributed across the genome and do not cluster in specific regions, they do not

interfere with the detection of regions with high recombination activity. And for organisms with low recombination rate the false positives detected by ReMIX can be decreased by lowering the amount of DNA used in the library preparation reaction, and the use of multiple independent reactions. This will decrease the mean number of molecules per GEM while maintaining the number of total recombinant molecules captured from the gametes.

Table 4.1: Number of false positive (FP) molecules identified by ReMIX at 7 and 10 molecules per GEM

No of FP per 100kb window	0	1	2	3	Total ratio of windows with FP
7 mol/GEM	25,889	1,335	44	1	5.06
10 mol/GEM	24,440	2,678	145	6	10.37

After splitting the mouse genome in 100kb windows (total of 27,269), we report the number of FP molecules identified by ReMIX in each window

Discussion

Understanding the extent and molecular basis of recombination variation has been challenging due to the expense of creating individualized high-resolution genome-wide recombination maps. Here we present a cost and time effective method to build individualized recombination maps from pooled gamete DNA. This method makes use of linked-read sequencing technology developed by 10X Genomics to acquire long-range haplotype information from gametes of a single individual. Our specialized bioinformatics pipeline named ReMIX then faithfully identifies recombinant molecules from the linked-read data produced. Using these recombinant molecules, crossover locations are defined as genomic intervals based on the location of the last variant of the first haplotype and first variant of the second. We demonstrate the application of our method by building fine-scale recombination maps for a male mouse, an organism with well characterized recombination hotspots, and a less traditional model organism, a male threespine stickleback fish.

We validated our method through comparisons to previously reported recombination landscapes in mouse (Paigen et al. 2008) and sticklebacks (Roesti et al. 2013), and simulations to quantify sensitivity and specificity. Our approach faithfully identified known recombination hotspots on mouse chromosome 1 with high resolution (median of 14 kb), and revealed enrichment in crossovers at the distal end of autosomes in the male mouse, and both ends of chromosomes in the male stickleback. Through simulations we show ReMIX has high sensitivity and that for organisms with low levels of heterozygosity this sensitivity can be increased by sequencing the linked-read library to higher coverage. In addition,

we used DNA extracted from somatic tissue as a control to test the specificity of our method. The use of a somatic control enabled the estimation of background noise in the data set that might be caused by bioinformatic error, reference genome assembly errors, copy number and structural variants, or rare mitotic recombination. Our results show that the true meiotic recombination signal stands out amidst the more dispersed noise from false positives, indicating ReMIX to be a reliable approach for constructing and studying variation in fine-scale recombination landscapes. Individualized genome-wide recombination maps that were previously constructed from extensive genotyping in thousands of offspring or whole genome sequencing of individual gametes (Wang et al. 2012) can now be produced with less time and effort by applying our novel method to pools of gametes from a single individual.

The whole genome recombination landscape we obtained for a male C57BL/6Ncrl × CAST/EiJ mouse (Supplementary Figure 5) is in agreement with the reported observation that male recombination activity is concentrated at the distal end of the autosomes. We also detected previously reported mouse chromosome 1 hotspots (Esrrg1 and Hlx1 (Figure 4.2b and Supplementary Figure 3b) in our data set. Using a sliding window approach by counting number of haplotypes switching molecules per 5 kb interval, we find a 9 kb interval at chr1:188,079,000–188,088,000 (mm10) region with highest recombination activity. This region spans the known Esrrg1 hotspot. Crossovers were identified from 31 haplotype switching molecules out of 1736 total mapped molecules in that 9 kb interval, suggesting a recombination rate of 1.78 cM in 9 kb. PRDM9, a protein with histone methyltransferase activity, plays an important role in recombination hotspots in many mammals including mice and humans. Consistent with previous studies showing that recombination in this region is mediated by PRDM9 (Billings et al. 2013), we find the PRDM9 motif specific to Esrrg1 located within the 9 kb hotspot (Figure 4.2). Similarly, following the criteria used by Liu et al. (Liu et al. 2014), we also detect extended genomic regions (≥500 kb) without any recombination crossovers (putative cold spots). The 167 regions detected span a total of 194 Mb (~7% of the mouse genome, similar to the proportion reported by Liu et al. (Liu et al. 2014)) and include well-characterized coldspot on chr12 at ~20 Mb that has been reported in multiple different strains (Morgan et al. 2017a).

Mouse and stickleback recombination crossovers are not distributed randomly across the genome, but are rather significantly clustered and more proximal to CpG islands than expected by chance. In stickleback, the region with the highest recombination activity is located on chromosome IV at ~3.8 Mb. Here, within a 7 kb interval, we detected six recombinant molecules out of a total of 1366 mapped molecules. This corresponds to a recombination rate of 0.44 cM, roughly one quarter the intensity of the mouse hotspot described above.

Megabase-sized genomic inversions are predicted to facilitate divergent adaptation and speciation with ongoing gene flow. Through empirical evaluation

of thousands of gametes from multiple marine, freshwater, and hybrid individuals, we have shown that a large chromosomal inversion on stickleback chrXXI causes strong recombination suppression within inversion heterokaryotypes, elevated recombination in the immediate flanking regions and harbors a high density of linked-mutations. While many studies have shown strong recombination suppression effects of inversions between species our study illustrates how recombination modifiers, such as inversions can cause strong recombination suppression even between adaptively diverging populations in the early stages of speciation.

We have demonstrated our method here using DNA extracted from sperm in organisms with high-quality genome assemblies. Considering the ease of collecting pools of gametes, and the low amount of input DNA required (e.g. 1 ng for a genome size of 3 GB genome, or <1 ng for smaller genomes), we anticipate our method can be extended to a wide range of organisms. ReMIX can detect recombination events in parts of the genome with diploid chromosome homologs that have heterozygous markers. Therefore, individualized recombination maps can be constructed for the whole genome including recombining regions of sex chromosome in the homogametic sex and pseudoautosomal regions of sex chromosomes in the heterogametic sex. While not shown here, the same principle could be expanded to study recombination in polyploids.

Our approach allows empirical quantification of fine-scale variation in recombination of both model and non-model organisms, including individuals sampled from the wild. We highlight that, for organisms whose genome assembly is lacking or of low quality, a de novo diploid assembly can be built (Weisenfeld et al. 2017) using the same linked-read data set generated from gametes. This de novo assembly can then be used as the reference genome for ReMIX analysis of recombination. By overcoming the challenges related to expense and effort of previous methods, our ReMIX pipeline, opens up numerous possibilities for investigating recombination variation within and among individuals, including the exciting potential of using forward genetic mapping to dissect and identify the molecular basis of recombination variation.

Methods

Extracting HMW genomic DNA. Stickleback genomic DNA was isolated from kidneys and sperm of a male wild-derived freshwater fish (River Tyne, Scotland). The sperm were collected via testes maceration in Hank's solution and purified to remove any potential contaminating diploid cells using a Nidacon PureSperm® gradient following the manufacturer's instructions with slight modifications. PureSperm gradient was made with 40/60/90 percentage solution and centrifuged at 300 × g for 30 min. Purified sperm cells were resuspended in 1×PBS (Thermo-Fisher, Cat. no. 10010023). Kidneys from the same male fish were dissected and rinsed in PBS prior to DNA extraction. HMW gDNA was extracted from purified

sperm cells and kidney using Qiagen Magattract HMW DNA extraction kit (Cat. no. 67563) following the protocol outlined in 10X Genomics Chromium Genome User Guide Rev B (10X Genomics 2018). We followed the Genomic DNA extraction from cell suspension protocol for the sperm sample, and the Tissue DNA extraction protocol for the kidney sample.

Mouse genomic DNA was isolated from F1 hybrid (C57BL/6Ncrl × CAST/EiJ) male spleen and sperm cells. Sperm were collected from the cauda epididymis of a 7-week-old F1 male hybrid mouse following Ijiri et al. (Ijiri et al. 2011). Extracted epididymis were finely chopped in 1 x PBS. After settling for 3–5 min at room temperature, the supernatant containing viable sperm was purified by gradient centrifugation at 300 × g for 20 min at room temperature (PureSperm 40/80; Nidacon International, Goteborg, Sweden). For somatic DNA control, excised spleen tissue was crushed between frosted glass microscope slides to make single cell suspension. Purified sperm and spleen cells were subsequently used for the isolation of HMW genomic DNA following Wu et al. (Wu et al. 1995).

The quality of extracted HMW DNA was checked by pulse field gel electrophoresis. All gametic and somatic samples showed a gradient of HMW DNA > 50 kb in size. This corresponds well to the described conditions for optimal performance of 10X Genomics linked-read library preparation (10X Genomics 2018).

Constructing linked-read sequencing libraries. We used a Chromium controller instrument (10X Genomics®) to partition input DNA into nanoliter-sized droplets and prepare linked-read libraries following the manufacturer's instructions (10X Genomics Chromium Controller User Manual) for input DNA quantification, dilution, GEM generation, and library preparation. For stickleback, we used ~0.8 ng of HMW genomic DNA as input (equivalent to ~1700 haploid genomes). To achieve the equivalent number of haploid genomes for mouse (1700), we carried out six parallel reactions with 1.2 ng input DNA for each of the sperm and somatic samples. In the Chromium Controller, input DNA was partitioned into ~1 million droplets (GEMs), each containing reagents with a unique barcode (Gemcode). The droplets were recovered from the microfluidic chip and isothermally incubated (at 30 °C) for ~3 h to produce barcoded short reads, average size ~700 bp, from each template DNA within each droplet. Following the isothermal incubation, the post GEM reads were recovered, then purified and size selected using Silane and Solid phase reverse immobilization (SPRI) beads. Illumina-compatible paired-end sequencing libraries were then prepared following 10X Genomics instructions, with 10 cycles of PCR. The final library comprises reads with a standard Illumina P5 adapter, followed by a 16 bp 10X Genomics barcode at the start of read 1, the genomic DNA insert, and an 8 bp sample index at the P7 adapter end. The final library was size selected to an average size of 600 bp. Sequencing was conducted with an Illumina HiSeq 3000 instrument with 2 × 150 bp paired-end reads. Each

library was sequenced to ~170x genome coverage. This is equivalent to ~0.1x read coverage per molecule for the ~1700 haploid gametes in the input.

ReMIX pipeline for identifying recombinant molecules. ReMIX pipeline contains three main steps: identifying high-quality heterozygous variants, reconstructing molecules, and haplotype phasing each molecule to determine the recombinant molecules and the position of their crossovers (Supplementary Figure 1). We make use of the software provided by 10X Genomics for reference guided analysis of linked-read data (Long Ranger (Long Ranger 2018)), but deviate from it in many places. After testing multiple equivalent tools for read filtering, alignment, or variant calling, we have configured ReMIX with the combination of tools for which we obtained the best results using both simulated and real data.

Identifying high-quality heterozygous variants. ReMIX's detection of recombinant molecules is based on the estimation of the two haplotypes present in the diploid individual analyzed. The accuracy of this estimation depends on the quality and frequency of heterozygous variants identified by our pipeline. Thus, in the first step of ReMIX (Supplementary Figure 1) we remove the linked-reads containing sequencing errors in their genomic sequence, align the correct linked-reads on a reference genome, call the set of variants, and apply a hard filter on this set.

In step 1 of ReMIX (Supplementary Figure 1 Step 1), the linked-reads are extracted from the Illumina's sequencer base call files (*.bcl) using Long Ranger mkfastq (Long Ranger 2018), and then filtered and trimmed with Cutadapt (Martin 2011), Trimmomatic (Bolger et al. 2014), and Long Ranger basic (Long Ranger 2018). The linked-reads with 16 bp barcode sequences matching the barcode whitelist provided by 10X Genomics are aligned with bwa mem (Li et al. 2009) to the reference genome. The duplicates are marked with Picard tools (Picard 2018) and read alignment around indels is improved using GATK's IndelRealigner (McKenna et al. 2010). ReMIX identifies variants with samtools mpileup (Li 2011) and applies a first variant filter using bcftools (Li et al. 2009) to extract high-quality heterozygous variants with low allelic bias. Specifically, we excluded variants with strand-, mapping-quality-, read-position or base-quality bias, variants with extreme low or high depth of coverage, and variants with low genotype or variant quality scores using the following thresholds: Mann–Whitney U-test of mapping quality bias (MQB) < 0.4; Mann–Whitney U-test of base quality bias (BQB) < 0.4; Mann–Whitney U-test of of mapping quality vs. strand bias MQSB < 0.8; Mann–Whitney U-test of read position bias (RPB) < 0.4; Maximum fraction of reads supporting an indel IMF < 0.1 or IMF > 0.9; Approximate read depth DP < 5 or DP > 220; genotype quality (GQ) < 30; variant quality QUAL < 100.

Reconstructing molecules. At the end of the first step of ReMIX the linked-reads are not yet organized into molecules. The purpose of the second step is to

reconstruct the molecules, so that the haplotype phasing algorithm can take advantage of the long-range information available.

The linked-reads generated from the same DNA molecule carry identical barcodes. However, since multiple molecules (e.g. 10) from diverse locations in the genome are typically trapped within the same GEM droplet and tagged with the same barcode, the molecules cannot be reconstructed based only on the barcodes of the linked-reads. From the quality control steps following HMW DNA extraction, it is possible to obtain an estimate of the expected average size of HMW DNA molecules in the reaction. Thus, we can link reads sharing an identical barcode into the same molecule if they aligned to the neighborhood of a genomic region with total molecule span similar to the expected average molecule size.

Still, this process does not always prevent linkage of reads from two or more independent molecules into a single reconstructed molecule when the original molecules share the same barcode and originate from the same genomic region. We refer to this case as *barcode collision*. For linked-read libraries constructed from organisms with large genome size using a low amount of input DNA in the library generation process, the probability of a single GEM droplet containing two HMW DNA molecules from the same genomic region is small, but non-zero. For example, the probability of barcode collision is ~3.2×10^{-3} for linked-read libraries prepared from 1 ng of mouse DNA of 60 kb average molecular weight, given a total mouse genome size of 3 Gb (meaning ~3200 of 1M GEMs will contain more than one molecule from the same region of the genome). When the original molecules are generated from opposite haplotypes, the barcode collision cases can generate recombinant-like molecules that will be identified by ReMIX as false positive. To limit the number of false positives, we introduced the following parameters: the maximum molecule length, the maximum distance between two consecutive linked-reads grouped into the same molecule and the minimum and maximum number of expected linked-reads per molecule. The values of these parameters depend on the library construction and sequencing parameters.

For this second step of ReMIX we constructed a Long Ranger sub-pipeline called Long Ranger ReportMolecules (Supplementary Figure 1 Step 2). This subpipeline is based on two parts of the Long Ranger Whole Genome Phasing and structural variant calling (SV Calling) pipeline (*Long Ranger wgs*) (Long Ranger 2018): the computational reconstruction of the molecules, and the report of the molecule information in the INFO field of the variant call format (vcf) file. Long Ranger ReportMolecules incorporates a number of changes to the original Long Ranger pipeline including the parameters mentioned above: the maximum molecule length, the minimum and maximum number of expected linked-reads per molecule. The input of this sub-pipeline is the binary sequence alignment map (bam) file with high-quality-mapped reads including a tag with their respective barcodes, and the vcf file with the filtered heterozygous variants. Long Ranger ReportMolecules outputs a file that reports for each molecule: the genomic start

and end position; the barcode; and the number of reads. This is accompanied by a modified vcf file that for each variant contains the reconstructed molecules spanning each of the alleles of this variant. Molecules with extreme low coverage (<6 reads) are excluded from further analysis.

Haplotype phasing molecules. In the last step, ReMIX estimates the two haplotypes by phasing selected variants based on the molecule information previously obtained. Then, depending on the alleles spanned by the reads of a molecule, the molecule is considered as belonging to one of the two haplotypes or as being a recombinant molecule.

Structural variants such as deletions, duplications, copy number variations, or translocations can cause errors in the read alignment, and thus variants can be incorrectly called in these regions. The false variants then interfere with the phasing process and introduce errors in the estimated haplotypes. Moreover, the structural variants can generate *barcode collision*-like cases. If misplaced reads and a real molecule share the same barcode and are aligned in the same genomic region, the algorithm used for reconstructing the molecules regroups the misplaced reads and the real molecule in a unique molecule. When the misplaced reads and the real molecule originate from opposite haplotypes, the reconstructed molecule appears as if it would span a crossover event as presented in Supplementary Figure 11. ReMIX identifies these problematic regions by removing: variants that have a notable difference between the molecular or read coverage compared to the mean values for their chromosome; and variants for which the read coverage is uneven between the alleles.

The remaining variants are then phased with HBOP (Xie et al. 2012) based on the molecules computed during the second step. HBOP is a single individual phasing algorithm that can take into account reads belonging to a longer DNA fragment and therefore capitalizes on the long-range information of the molecule during phasing.

The two haplotypes constructed by HBOP are then used to phase each molecule. For each variant spanned by a molecule with at least one read, we consider the haplotype of the covered allele and the sequencing quality score at that position. Then, based on a score function implemented in Long Ranger wgs (Long Ranger 2018), we compute for each molecule the probability of belonging to the two haplotypes or being a mix of the two. Contrary to Long Ranger wgs, we do not consider the molecules that contain reads spanning both alleles of a variant, since this behavior is likely to arise from a barcode collision. Once the probabilities are computed for each molecule, we filter again to remove variants showing an allelic bias in the number of molecules phased to each allele. Depending on the quality of the reference sequence used in the mapping process or on the copy number variation, some of the structural variants are still unidentified and can

introduce errors in the process of determining the recombinant molecule. We then recompute the haplotype probabilities for each molecule.

From the set of molecules that have a high probability of belonging to a mixture of two haplotypes states, ReMIX considers as truly recombinant the molecules for which we can identify a clear crossover position: a minimum number of variants and a minimum ratio of variants phased to the same haplotype on each side of the crossover. We then output for each recombinant molecule the genomic start and end position; the crossover positions; the barcode; and the number of reads.

All animals used in this study were housed at approved animal facilities and handled according to Baden-Wu̇rttemberg State approved protocols (Competent authority: Regierungspräsidium Tu̇bingen, Germany; Permit and notice numbers 35/9185.82-5, 35/9185.46)

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated and analyzed in the current study are available in the NCBI short read repository [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA562078]. All other relevant data is available upon request.

Code availability

ReMIX source code can be found at github [https://github.com/adreau/ReMIX] and zenodo archive https://doi.org/10.5281/zenodo.3351406.

Received: 12 December 2018 Accepted: 28 August 2019.

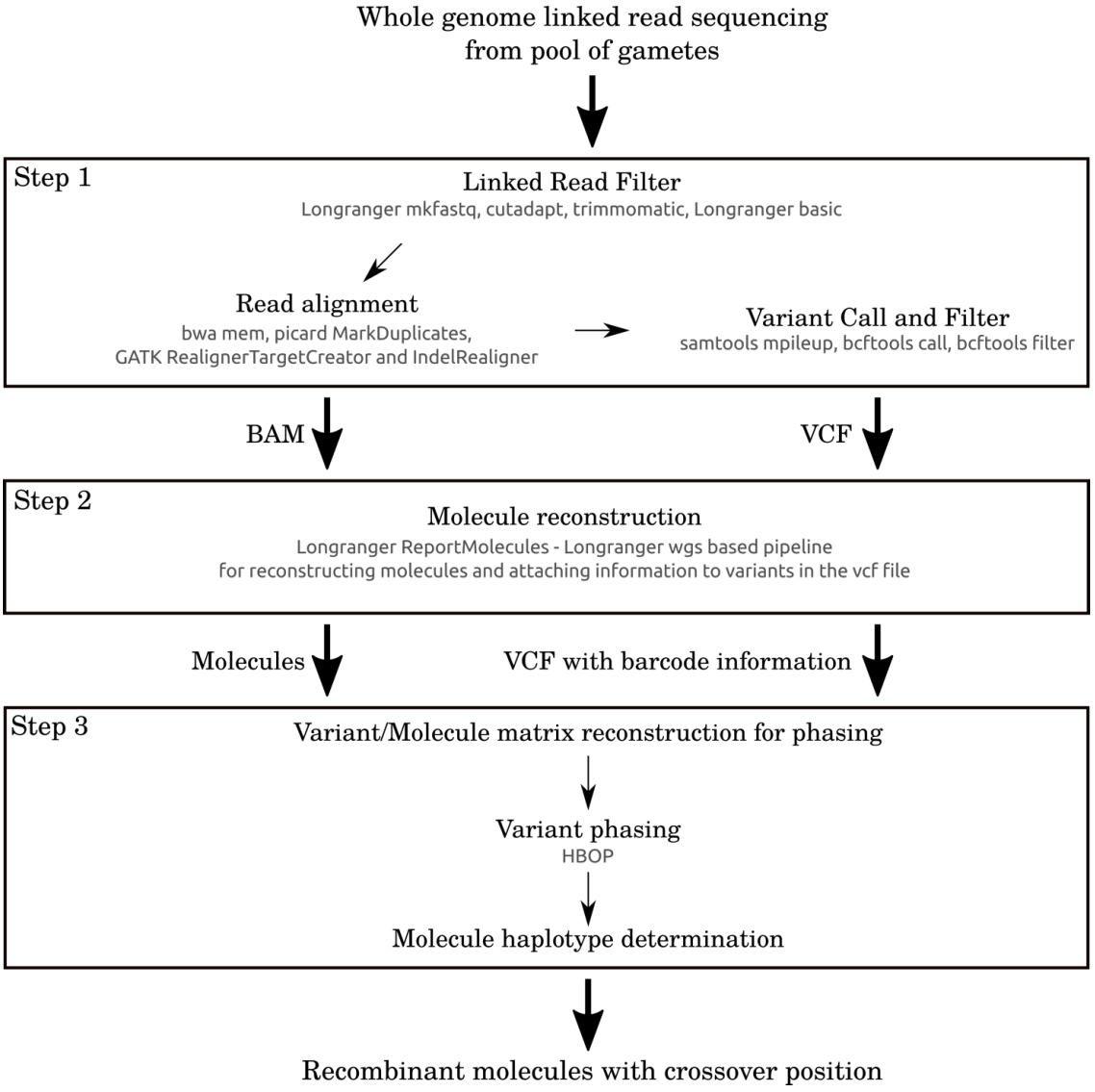
Published online: 20 September 2019

Acknowledgements

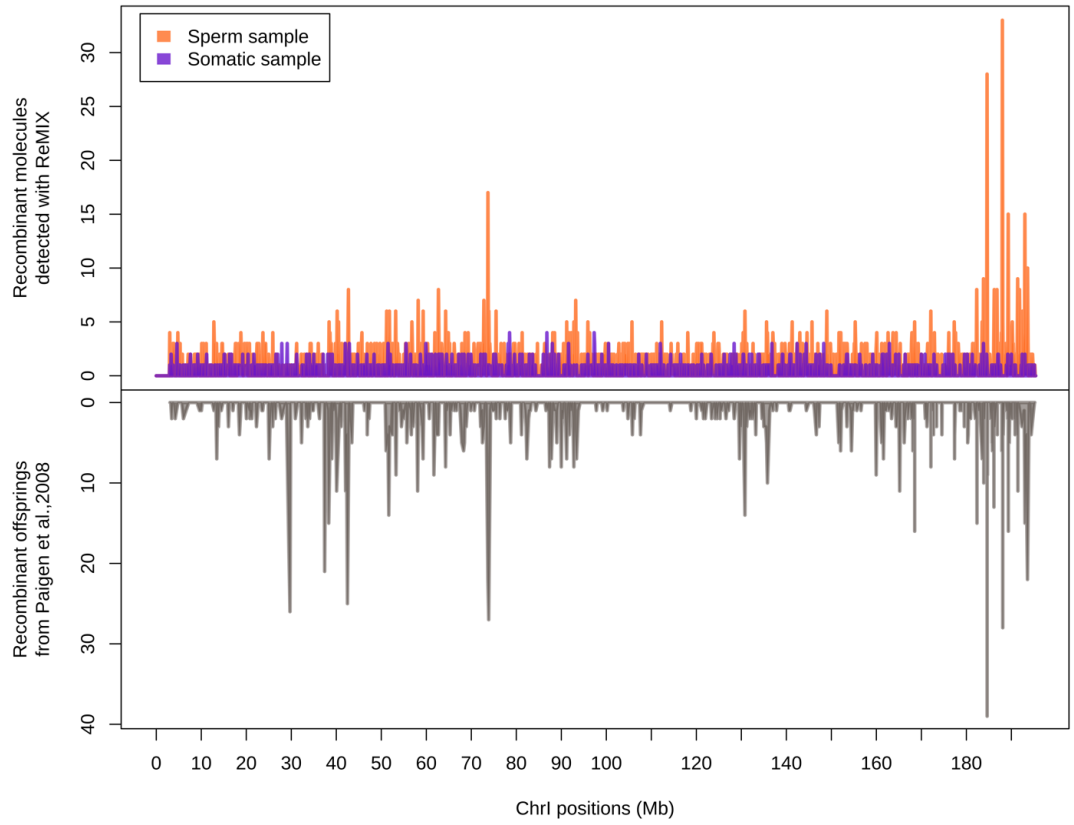
We would like to thank Frank Chan for contribution of mouse material, and both Frank Chan and Marek Kucka for ongoing discussions and insight related to further development of linked-read sequencing methods for recombination detection. We thank Enni Harjunmaa for her suggestions and support on library preparation and data analysis. We are grateful to Ruth Ley for her contribution towards the 10X Genomics Chromium controller, to Andre Noll and the Max Planck Institute for Developmental Biology Computing Core Facility for their high-performance computing support, and Christa Lanz and the Max Planck Institute for Developmental Biology Genome Core Facility for their assistance with high throughput sequencing. ReMIX uses part of the 10X Genomics Long Ranger pipeline and we are grateful to 10X Genomics for access to their open source code and their discussions in the initial stages of this project. We are grateful for the research support of a European Research Council Consolidator Grant to F.C.J. (FP7

617279). F.C.J. is also supported by the Max Planck Society. F.C.J. is also grateful for support from the Max Planck Society.

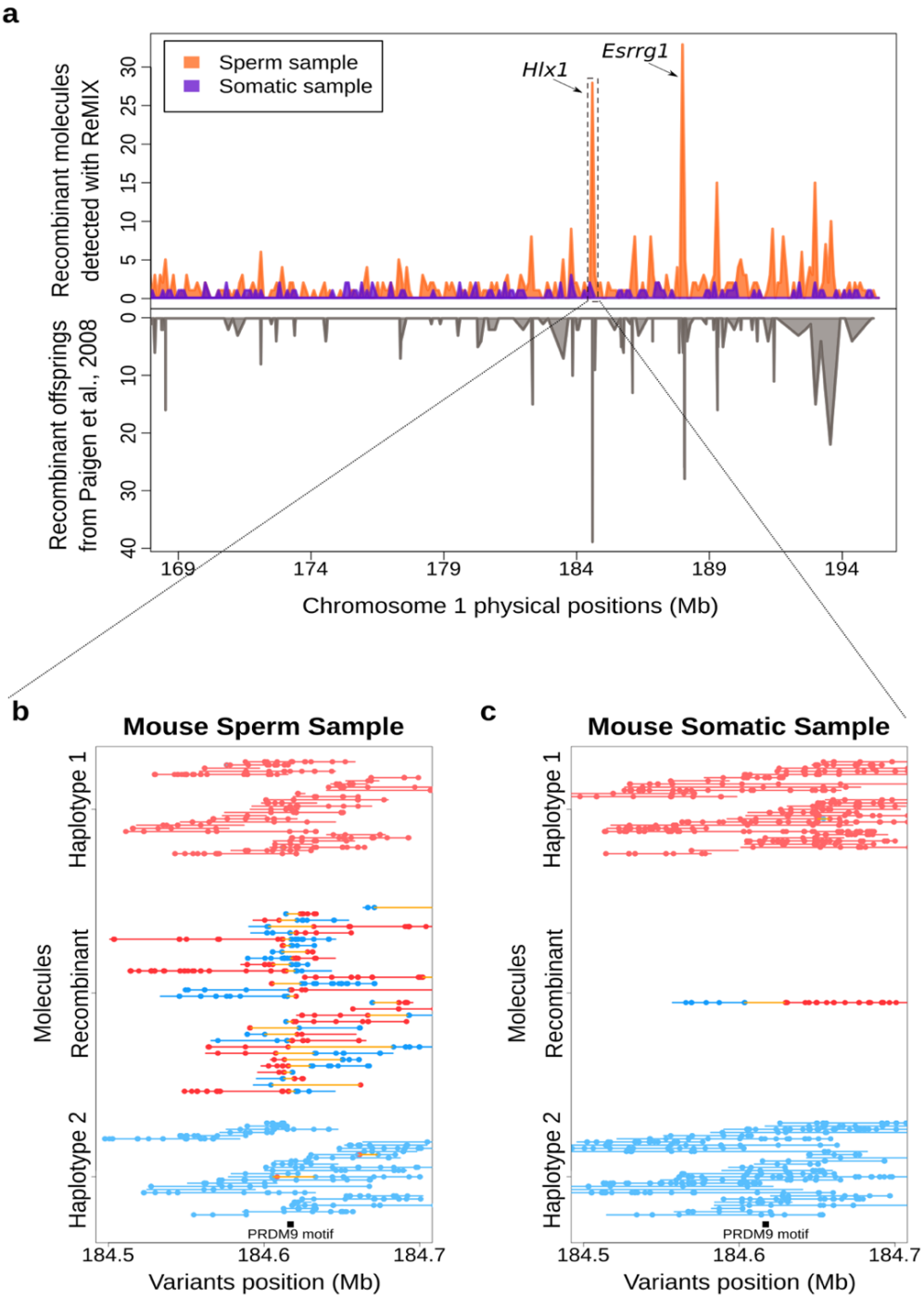
4.4 Supplementary information



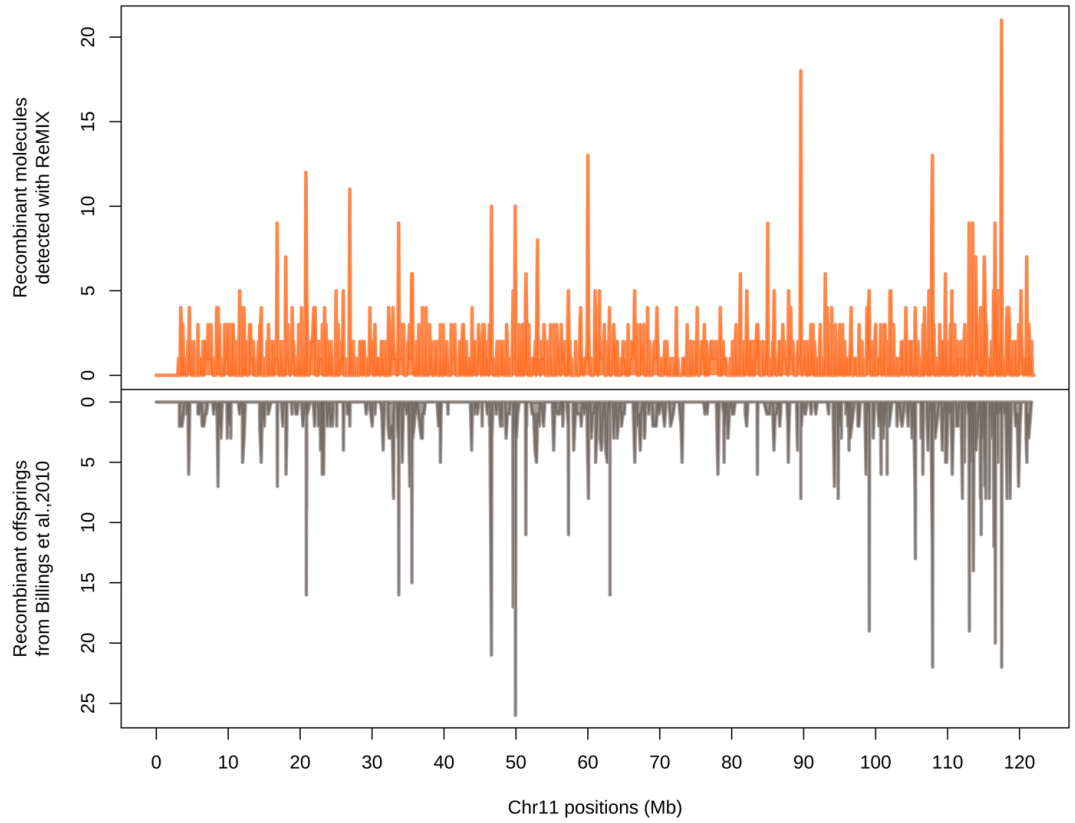
Supplementary Figure 1: ReMIX pipeline's main steps: Identification of high-quality heterozygous variants, reconstruction of molecules, and haplotype phasing each molecule. In the input our pipeline requires Illumina's base call files from sequencing pool of gametes and outputs the identified recombinant molecules and the position of their crossovers.



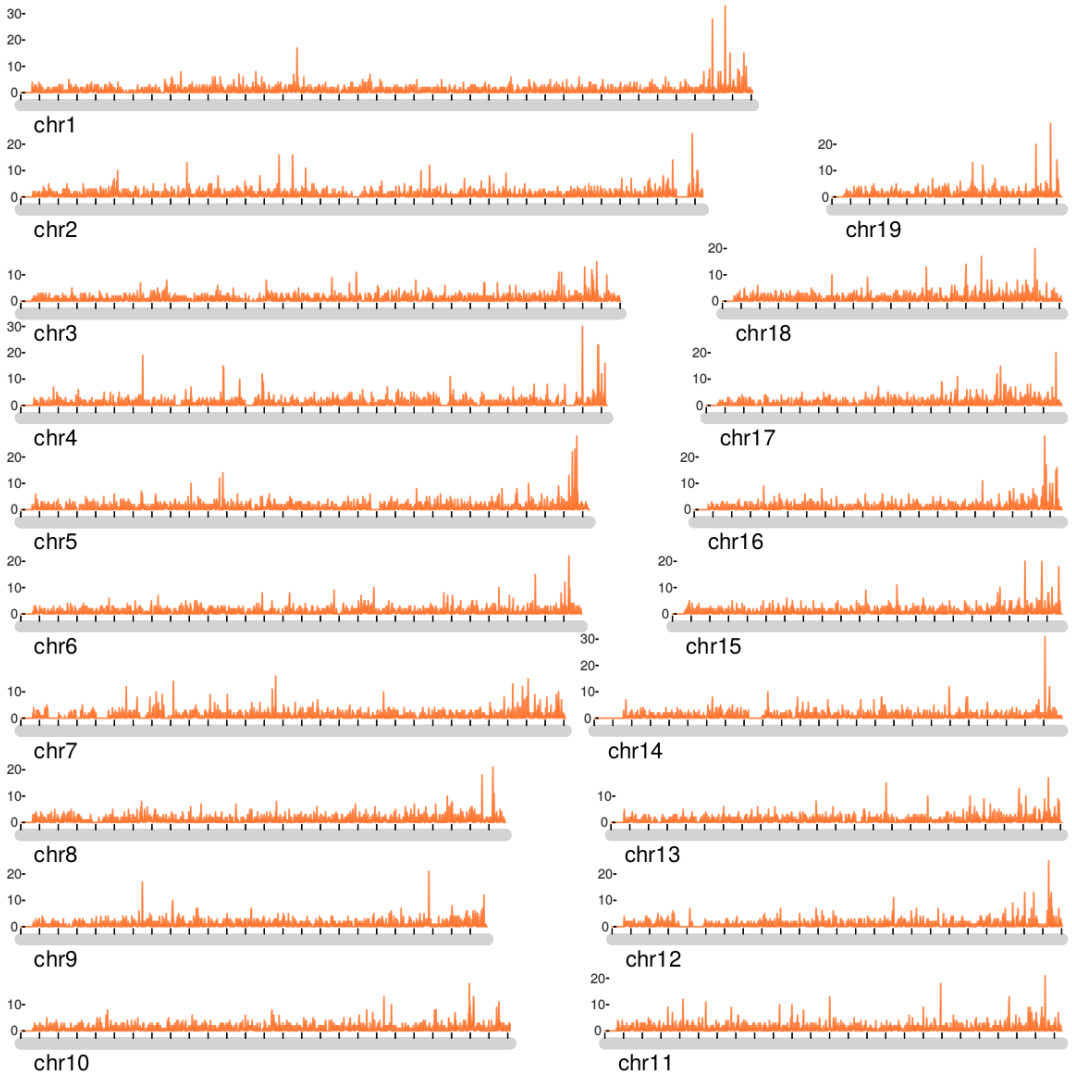
Supplementary Figure 2: ReMIX correctly detects fine-scale recombination variation and hotspots on mouse chromosome 1. The recombination rate determined by ReMIX corresponds well to the rate described in Paigen et al. (Paigen et al. 2008) . The dissimilarity observed in the northern end of the chromosome may be caused by potential sub-strain differences in recombination (Fontaine and Davis 2016) (C57BL/6Ncrl x CAST/EiJ used in our study and C57BL/6J x CAST/EiJ in Paigen et al. (Paigen et al. 2008))



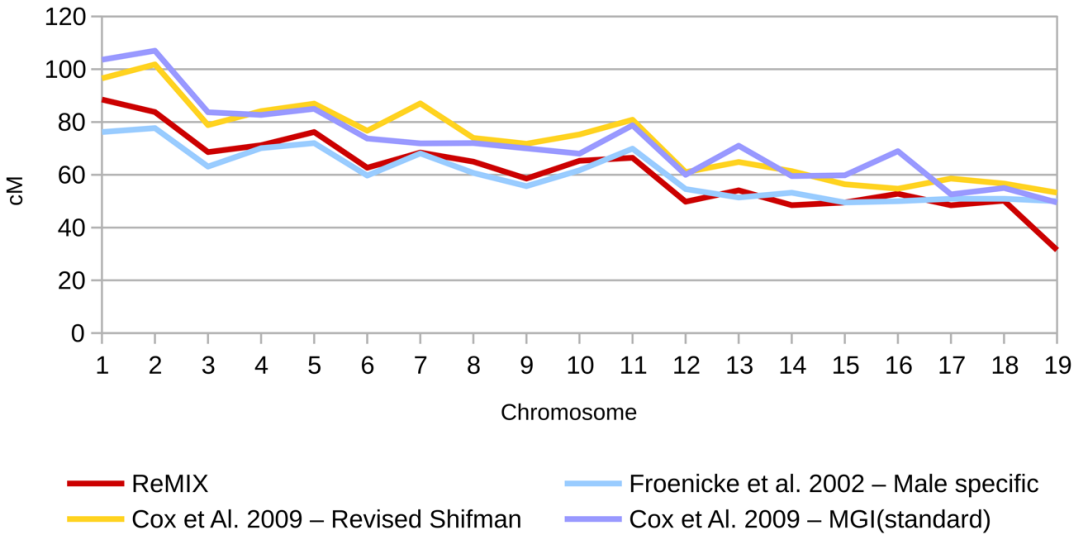
Supplementary Figure 3: ReMIX correctly detects fine-scale recombination variation and hotspots on mouse chromosome 1. (a) The recombination rate on the south end of chromosome 1 (169- 195.4Mb, mm10), determined by ReMIX corresponds well to the rate described in Paigen et al. (Paigen et al. 2008) . (b) The three types of molecules identified by ReMIX in the sperm sample in the region of a wellknown recombination hotspot (Hlx1 (Paigen et al. 2008; Billings et al. 2013)). PRDM9 plays a role in initiating crossovers at the Hlx1 hotspot and has a DNA binding motif (black bar) located near the midpoint of the detected recombinant molecules. (c) The corresponding region for somatic tissue in which ReMIX identified a recombinant molecule due to mitotic recombination or barcode collision.



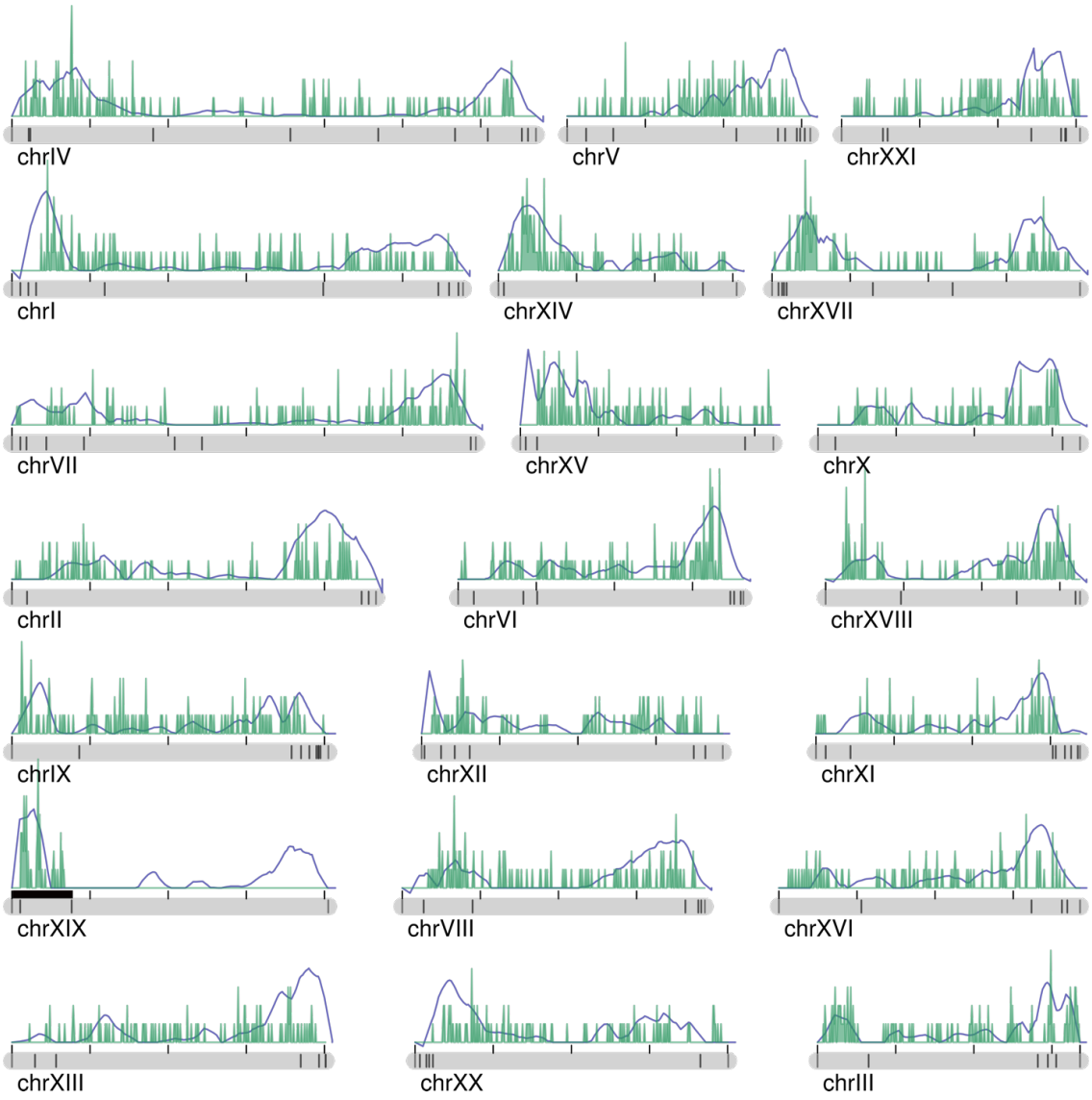
Supplementary Figure 4: ReMIX correctly detects fine-scale recombination variation and hotspots on mouse chromosome 11. The recombination rate determined by ReMIX corresponds well to the rate described in Billings et al. (Billings et al. 2010).



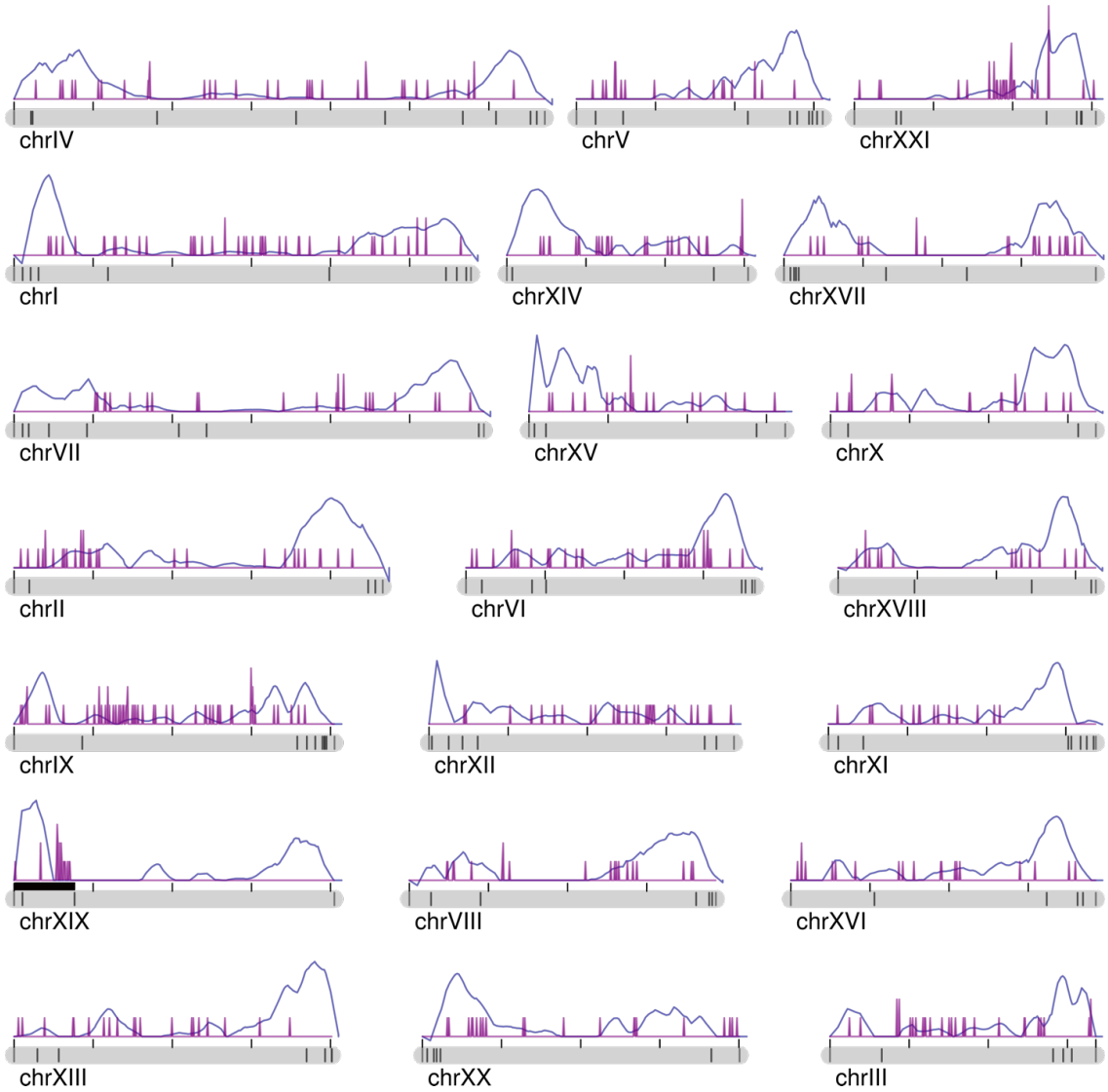
Supplementary Figure 5: ReMIX results on the mouse autosomes. Consistent with previous studies (Liu et al. 2014), ReMIX reveals recombination crossovers are enriched towards the distal ends of chromosomes in male germline.



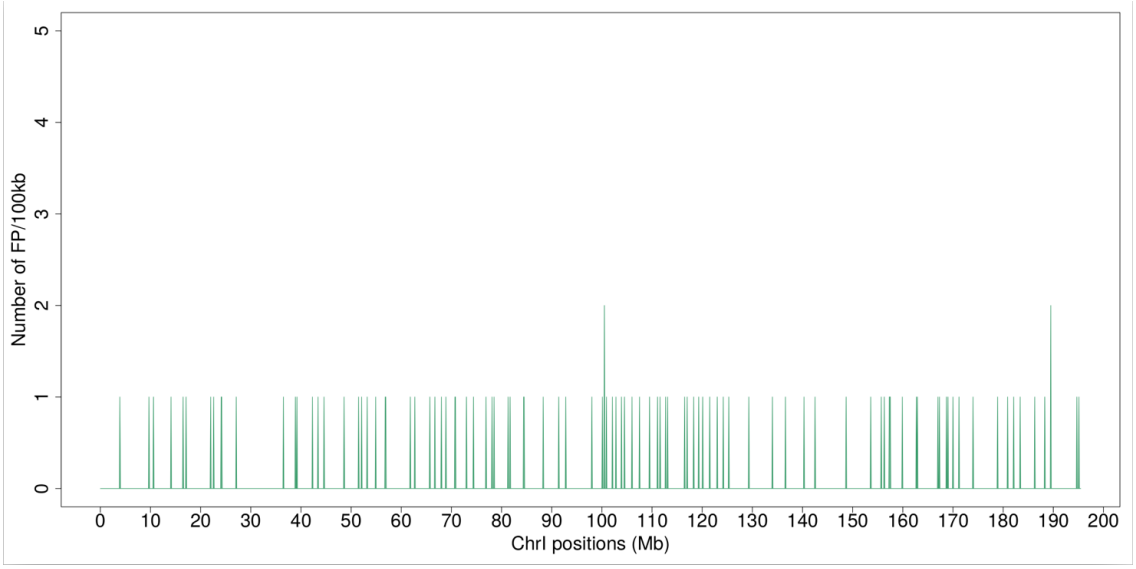
Supplementary Figure 6: Genetic map length comparison between previous studies (Froenicke et al. 2002; Cox et al. 2009) analyzing various mouse strains and ReMIX results on the mouse genome.



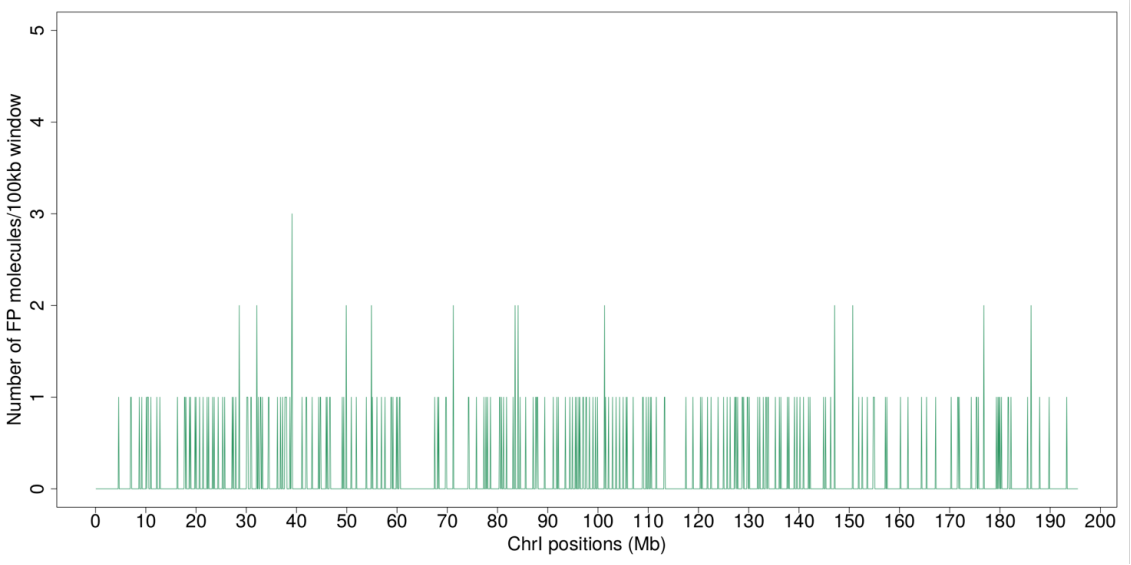
Supplementary Figure 7: Genome graph of recombination events in a male freshwater stickleback with underlying genetic map. The number of crossovers identified by our pipeline is plotted in 50 kb intervals (in green). The genetic map was previously constructed from F2 lab cross population of 282 male and female individuals and 1872 total markers (Roesti et al. 2013) (in blue). Black box on chromosome XIX represents the recombining pseudoautosomal region of the X chromosome. In an XY stickleback male, no recombination is expected in the sex determining region of the X chromosome.



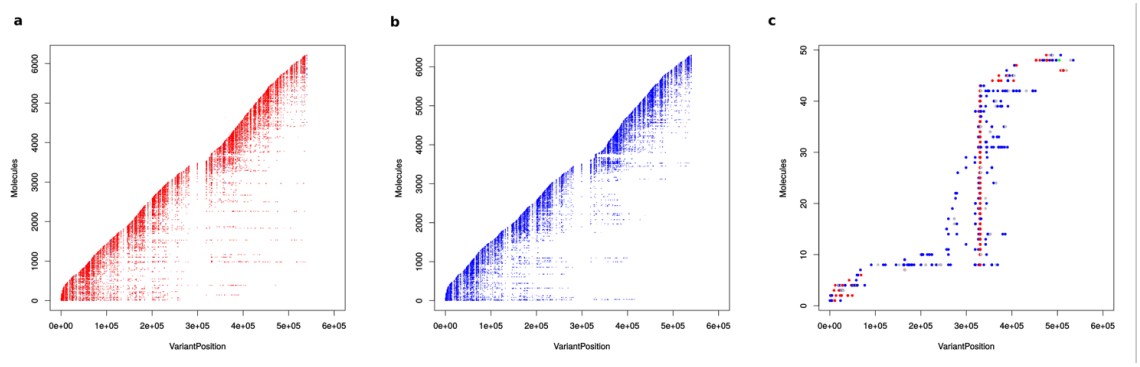
Supplementary Figure 8: Genome graph of recombination events in somatic tissue sample with underlying genetic map for negative control. The number of crossovers identified by our pipeline is plotted in 50 kb intervals (in purple). The genetic map was previously constructed from F2 lab cross population of 282 male and female individuals and 1872 total markers (Roesti et al. 2013) (in blue). For most chromosomes the maximum number of these false positive somatic recombinant molecules in 50 kb windows is 2. As expected, the moderate false positive rate is evenly distributed and does not interfere with the hotspot detection. The false positive rates across chromosomes with elevated levels co-localize with scaffold ends (chromosomes XIV, XIX, and XXI) (black lines on the gray bars) and are likely scaffold assembly errors.



Supplementary Figure 9: Frequency of false positive molecules in 100kb windows in the case of simulating 7 molecules per GEM in the mouse chromosome 1.



Supplementary Figure 10: Frequency of false positive molecules in 100kb windows in the case of simulating 10 molecules per GEM in the mouse chromosome 1.



Supplementary Figure 11: The errors in the read alignment and variant calling due to structural variants can generate recombinant-like molecules. (a) Haplotype 1 molecules. (b) Haplotype 2 molecules. (c) Recombinant-like molecules. These errors can cause incorrect variant phasing or barcode collision cases in the structural variant regions. When misplaced reads and a real molecule share the same barcode and are aligned in the same genomic region, the algorithm used for reconstructing the molecules regroups the misplaced reads and the real molecule in a unique molecule. In the case when the misplaced reads and the real molecule originate from opposite haplotypes or when variants are incorrectly phased, the reconstructed molecules appear as if it would span a crossover event and they pile-up wrongly suggest a hotspot region. ReMIX effectively identifies these problematic regions and removes them in the third step of the pipeline.

Chapter 5

Concluding remarks and future outlook

Being a fundamental molecular mechanism with crucial role in organismal fitness, homologous recombination during meiosis has been a major focus of numerous studies. It has been shown that meiotic recombination is essential for proper disjunction of homologous chromosomes during meiosis I, as well as important for generating favorable allele combinations that confer adaptive advantage. However, majority of the studies to date have focused on the molecular aspects rather than the evolutionary significance of this fundamental process. For the most part, evolutionary studies are limited to theoretical predictions of how recombination could be regulated in different evolutionary scenarios to facilitate rapid adaptation. However, empirical studies testing those predictions and exploring the molecular-genetic features fine-tuning the recombination landscape are lagging behind. This is mostly due to the difficulty in detailed characterization of the fine-scale recombination landscape in evolutionary model systems. The work presented in this thesis overcomes most of the challenges by taking advantage of the model organism threespine stickleback fish, and by exploiting the recent advancements in the field of next generation sequencing. The major improvements made with this project in terms of method development, in our understanding of stickleback recombination, its regulatory mechanisms and potential evolutionary implications are summarized in this chapter. Also, I discuss the future steps required to further understand this fundamental process in the context of adaptive divergence.

Methodological advancements

I have demonstrated the suitability of threespine stickleback fish as a model organism to improve our knowledge of recombination landscape variation in an evolutionary context. With the convenience of breeding and rearing adaptively diverged ecotypes in laboratory conditions, it is possible to carry out detailed multilevel empirical studies. This allowed us to employ next generation sequencing on large clutch nuclear families to directly detect individualized crossover events with high resolution (chapter 2). We have put together a low-cost whole genome sequencing protocol along with a bioinformatic pipeline that can be adapted to similar studies in other organisms. At present, nuclear family sequencing is the ideal strategy to construct the fine-scale recombination landscape of one generation. However, it is impractical for constructing hundreds of individualized recombination maps that enable studies such as QTL mapping. Our

novel method, recombination map construction from pooled gametes (described in chapter 4) provides an excellent alternative for such large-scale requirements. Moreover, this method can also be used to identify and measure the strength of individualized recombination hotspots and coldspots as it screens a large number of meiotic products from one single individual.

ChIP sequencing is a well-established and powerful method to map recombination initiation sites. However, successful mapping of the DNA double strand break landscape has only been carried out in a handful of organisms to date (Pan et al. 2011; Khil et al. 2012; Pratto et al. 2014). The success rate of this method is directly linked to the availability of a suitable antibody and the ability to harvest the required number of cells at the right stage. These factors pose a major hurdle for such studies in non-model natural populations. Despite these challenges, here we compiled a protocol for raising stickleback-specific anti-DMC1 antibodies, and carried out successful ChIP sequencing on pooled male testes (Chapter 3). However, I believe that this protocol is still in its infancy and recommend further efforts to develop efficient ways of harvesting meiotic cells (both from testes as well as ovaries) at the right stage. This would improve the signal to noise ratio, and thereby enable us to identify almost all DSB sites across the genome. I would like to emphasize that complementing the crossover map with the map of DSBs in evolutionary model organisms can greatly improve our understanding of recombination regulation under varying selection pressure. We can compare DSB and CO landscapes among ecotypes and investigate whether the difference due to differential selection pressures comes from the recombination initiation or CO designation.

Major findings and future directions

Based on previous studies we know that the sticklebacks have a non-uniform recombination landscape with periphery biased recombination; female biased heterochiasmy; and divergent fine-scale recombination landscapes between marine and freshwater ecotypes (Roesti et al. 2013; Sardell et al. 2018; Shanfelter et al. 2019). Our comprehensive empirical study enabled quantitative comparison of these variations at different levels. I find that the sex-specific variation is much higher than ecotype-specific variation both in terms of genome-wide recombination rate and landscape. I also report a significant reduction in overall recombination rate in hybrid females compared to pure forms. Even though overall recombination rate is reduced around regions of adaptive divergence, a clear sex-specific pattern is observed; i.e., females recombine more than twice as often in the regions between adaptive loci than males. As a result, irrespective of the ecotype, females shuffle adaptive alleles more frequently than males. Even though several previous studies in various organisms, including sticklebacks, have reported the dramatic sexual dimorphism in recombination phenotype, this study further emphasizes its evolutionary importance. Sexual dimorphism and its driving factors may provide differential barriers to gene flow. Hence, sex-specific

recombination modifiers may fine-tune the recombination landscape under varying selection pressures.

To check how generalizable these observations are, similar high-resolution ecotype and sex-specific studies are needed in other stickleback populations with varying levels of selection pressures. An advantage of the stickleback model system is that various natural populations around the northern hemisphere present replicates of independent adaptive divergence. Previous studies have characterized extent of adaptive divergence among stickleback ecotype pairs in various populations (Jones et al. 2012; Samuk et al. 2017). Therefore, comparing the extent of hybrid recombination reduction and sexual dimorphism in populations with higher and lower divergence, with and without gene flow, would provide further insights into how natural selection shapes the recombination landscape.

While the sex-specific recombination rate and landscape present the major axis of variation, high variation is also observed in the fine-scale recombination landscape among ecotypes. But, in this study, it is not distinguishable from individual variation (chapter 2). Therefore, this observation demands further studies screening large number of meiotic products per ecotype to investigate whether there is any ecotype-specific landscape of recombination at fine scale. We can employ LD based methods to estimate historical recombination rates from marine and freshwater ecotypes of this population. By complementing the one-generation ecotype-specific recombination map produced in this project with an LD based map of historical recombination events, we can test their strength of correlation and how well the stickleback crossover locations are conserved. Moreover, the high inter-individual variation observed here also require special attention since it is likely to have a heritable genetic basis. This can be addressed by employing QTL mapping studies to find genomic loci underlying the variation. If there is a genetic basis for individual variation, divergent selection on those recombination modifier loci among individuals of diverging ecotypes could eventually lead to ecotype-specific heritable recombination rate and landscape.

After quantifying the genome-wide recombination rate and landscape variation, I investigated the potential genomic/epigenetic factors that may regulate recombination landscape. In the scope of this thesis, I haven't examined the underlying factors of periphery biased broad-scale recombination landscape, rather mainly focused on the underlying features of fine-scale landscape. However, we have seen that sexes differ substantially in their broad-scale recombination distribution, and this variation may have important evolutionary implications. Therefore, future studies are required to examine genomic/chromatin features that may drive this variation in the broad-scale landscape. Studies in mice and humans have shown that, difference in genome-wide DNA methylation patterns and chromosome condensation differences might influence broad-scale, sex-specific recombination landscape (Petkov et al. 2007; Gruhn et al. 2013; Brick et al. 2018). Therefore, I suggest whole genome bi-sulphate

sequencing and chromosome spread analysis on meiotic cells to investigate these potential molecular factors that may underlie broad-scale recombination landscape variation.

Results presented in chapter 3 suggest that, at fine-scale, a combination of factors may direct DSB events to its potential target sequence. The male crossover map broadly mirrors the DSB landscape both in terms of distribution and signal intensity. In contrast to mammalian recombination landscape, diffused DSB landscape and `semi-hot` crossover hotspots are observed in sticklebacks. In males, both DSB and crossover sites are preferentially targeted to functionally active open chromatin region such as gene promoters. However, a considerable amount of DSBs as well as crossover sites are observed to be away from promoter and other tested open chromatin marks. As we do not detect any crossover-specific association or lack of association with any of the tested genomic features, I speculate that regulations at the level of DSBs, rather than at the level of crossover designation, may shape stickleback crossover landscape. Female recombination landscape also corroborates this pattern but in contrast to males, female crossover association with promoter or other open chromatin marks does not differ from random distribution. However, there are reasons to believe that the non-promoter-associated crossovers are non-random because, it includes sex-specific hotspots and possess higher GC content in those intervals (high GC content is considered as a proxy for past occurrences of recombination events). As discussed earlier, an LD based historical recombination map in this study population would provide insights into historical usage of such hotspots.

Based on the evidences collected during this project, I conclude that, the non-random but rather diffused fine-scale recombination landscape of stickleback males and females may have additional novel regulators of recombination. Nucleosome depletion might not be a necessary criterion for choosing the target sites. However, further efforts are required to find potential regulators such as trans-acting recombination modifiers and/or cis-acting genomic/epigenetic features. Unravelling rest of the recombination modifiers in sticklebacks is important as it can be the target of natural selection to confer plasticity for recombination landscape.

In short words, this empirical study provides insights into the extent of variation in the contemporary recombination landscape in an adaptively diverging natural population and emphasizes the evolutionary importance of sexual dimorphism in meiotic recombination. Our findings set a foundation for future studies and encourage further detailed investigation of the molecular regulators of adaptive recombination landscape.

References

10X Genomics. 2018. Chromium™ Genome Reagent Kit User Guide.

Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalier B, Gill BS, Miftahudin, Gustafson JP, Lazo G, Chao S et al. 2003. The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res* **13**: 753-763.

Anderson LK, Reeves A, Webb LM, Ashley T. 1999. Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of MLH1 protein. *Genetics* **151**: 1569-1579.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *P Natl Acad Sci USA* **112**: 2109-2114.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**: 193-198.

Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Greally JM, Wang J et al. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet* **9**: e1003984.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-208.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME Suite. *Nucleic Acids Res* **43**: W39-49.

Baker CL, Walker M, Kajita S, Petkov PM, Paigen K. 2014. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. *Genome Res* **24**: 724-732.

Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, Przeworski M. 2017. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *Elife* **6**.

Barnes TM, Kohara Y, Coulson A, Hekimi S. 1995. Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. *Genetics* **141**: 159-179.

Barton AB, Pekosz MR, Kurvathi RS, Kaback DB. 2008. Meiotic recombination at the ends of chromosomes in Saccharomyces cerevisiae. *Genetics* **179**: 1221-1235.

Bass HW, Riera-Lizarazu O, Ananiev EV, Bordoli SJ, Rines HW, Phillips RL, Sedat JW, Agard DA, Cande WZ. 2000. Evidence for the coincident initiation of homolog pairing and synapsis during the telomere-clustering (bouquet) stage of meiotic prophase. *J Cell Sci* **113 (Pt 6)**: 1033-1042.

Bateson W, Saunders, E.R., Punnett, R. C. . 1905. Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* **2**: 1–55,80–99.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**: 836-840.

Baudat F, de Massy B. 2007. Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot. *PLoS Genet* **3**: e100.

Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W et al. 2013. Intraspecific variation of recombination rate in maize. *Genome Biol* **14**: R103.

Bell MA, Foster SA. 1994. *The evolutionary biology of the threespine stickleback*. Oxford University Press, Oxford ; New York.

Bell MA, Orti G, Walker JA, Koenings JP. 1993. Evolution of Pelvic Reduction in Threespine Stickleback Fish: A Test of Competing Hypotheses. *Evolution* **47**: 906-914.

Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics* **42**: 859-+.

Bernstein MR, Rockman MV. 2016. Fine-Scale Crossover Rate Variation on the Caenorhabditis elegans X Chromosome. *G3 (Bethesda)* **6**: 1767-1776.

Billings T, Parvanov ED, Baker CL, Walker M, Paigen K, Petkov PM. 2013. DNA binding specificities of the long zinc-finger recombination protein PRDM9. *Genome Biol* **14**: R35.

Billings T, Sargent EE, Szatkiewicz JP, Leahy N, Kwak IY, Bektassova N, Walker M, Hassold T, Graber JH, Broman KW et al. 2010. Patterns of recombination activity on mouse chromosome 11 revealed by high resolution mapping. *PLoS One* **5**: e15340.

Bishop DK, Park D, Xu L, Kleckner N. 1992. DMC1: a meiosis-specific yeast homolog of E. coli recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell* **69**: 439-456.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Bomblies K, Higgins JD, Yant L. 2015. Meiosis evolves: adaptation to external and internal environments. *New Phytol* **208**: 306-323.

Brandvain Y, Coop G. 2012. Scrambling eggs: meiotic drive and the evolution of female recombination rates. *Genetics* **190**: 709-723.

Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485**: 642-645.

Brick K, Thibault-Sennett S, Smagulova F, Lam KWG, Pu YM, Pratto F, Camerini-Otero RD, Petukhova GV. 2018. Extensive sex differences at the initiation of genetic recombination. *Nature* **561**: 338-+.

Bridges CB. 1927. The Relation of the Age of the Female to Crossing over in the Third Chromosome of Drosophila Melanogaster. *J Gen Physiol* **8**: 689-700.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* **63**: 861-869.

Bronner IF, Quail MA, Turner DJ, Swerdlow H. 2014. Improved Protocols for Illumina Sequencing. *Curr Protoc Hum Genet* **80**: 18 12 11-42.

Burger R, Akerman A. 2011. The effects of linkage and gene flow on local adaptation: a two-locus continent-island model. *Theor Popul Biol* **80**: 272-288.

Campbell CL, Bherer C, Morrow BE, Boyko AR, Auton A. 2016. A Pedigree-Based Map of Recombination in the Domestic Dog Genome. *G3 (Bethesda)* **6**: 3517-3524.

Campbell CL, Furlotte NA, Eriksson N, Hinds D, Auton A. 2015. Escape from crossover interference increases with maternal age. *Nat Commun* **6**: 6260.

Capilla L, Garcia Caldes M, Ruiz-Herrera A. 2016. Mammalian Meiotic Recombination: A Toolbox for Genome Evolution. *Cytogenet Genome Res* **150**: 1-16.

Capilla L, Medarde N, Alemany-Schmidt A, Oliver-Bonet M, Ventura J, Ruiz-Herrera A. 2014. Genetic recombination variation in wild Robertsonian mice: on the role of chromosomal fusions and Prdm9 allelic background. *P Roy Soc B-Biol Sci* **281**.

Carrington M, Cullen M. 2004. Justified chauvinism: advances in defining meiotic recombination through sperm typing. *Trends in Genetics* **20**: 196-205.

Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016. Adaptive Evolution Is Substantially Impeded by Hill-Robertson Interference in Drosophila. *Molecular Biology and Evolution* **33**: 442-455.

Cech JN, Peichel CL. 2015. Identification of the centromeric repeat in the threespine stickleback fish (Gasterosteus aculeatus). *Chromosome Res* **23**: 767-779.

Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**: 302-305.

Charlesworth B, Barton NH. 2018. The Spread of an Inversion with Migration and Selection. *Genetics* **208**: 377-382.

Charlesworth B, Charlesworth D. 1975. An experimental on recombination load in Drosophila melanogaster. *Genet Res* **25**: 267-274.

Charlesworth D, Charlesworth B. 1979. Selection on recombination in clines. *Genetics* **91**: 581-589.

Choi K, Henderson IR. 2015. Meiotic recombination hotspots - a comparative view. *Plant J* **83**: 52-61.

Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA, Copenhaver GP, Franklin FC et al. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat Genet* **45**: 1327-1336.

Choo KHA. 1998. Why is the centromere so cold? *Genome Research* **8**: 81-82.

Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG. 2009. Genetic analysis of variation in human meiotic recombination. *PLoS Genet* **5**: e1000648.

Chung G, Rose AM, Petalcorin MI, Martin JS, Kessler Z, Sanchez-Pulido L, Ponting CP, Yanowitz JL, Boulton SJ. 2015. REC-1 and HIM-5 distribute meiotic

crossovers and function redundantly in meiotic double-strand break formation in Caenorhabditis elegans. *Genes Dev* **29**: 1969-1979.

Civardi L, Xia YJ, Edwards KJ, Schnable PS, Nikolau BJ. 1994. The Relationship between Genetic and Physical Distances in the Cloned A1-Sh2 Interval of the Zea-Mays L Genome. *P Natl Acad Sci USA* **91**: 8268-8272.

Cole F, Baudat F, Grey C, Keeney S, de Massy B, Jasin M. 2014. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nature Genetics* **46**: 1072-1080.

Cole F, Kauppi L, Lange J, Roig I, Wang R, Keeney S, Jasin M. 2012a. Homeostatic control of recombination is implemented progressively in mouse meiosis. *Nat Cell Biol* **14**: 424-+.

Cole F, Keeney S, Jasin M. 2012b. Preaching about the converted: how meiotic gene conversion influences genomic diversity. *Ann Ny Acad Sci* **1267**: 95-102.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Jr., Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**: 1928-1933.

Colosimo PF, Peichel CL, Nereng K, Blackman BK, Shapiro MD, Schluter D, Kingsley DM. 2004. The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS Biol* **2**: E109.

Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in Drosophila melanogaster. *PLoS Genet* **8**: e1002905.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**: 1395-1398.

Cox A, Ackert-Bicknell CL, Dumont BL, Ding Y, Bell JT, Brockmann GA, Wergedal JE, Bult C, Paigen B, Flint J et al. 2009. A new standard genetic map for the laboratory mouse. *Genetics* **182**: 1335-1344.

Creighton HB, McClintock B. 1931. A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proc Natl Acad Sci U S A* **17**: 492-497.

Cresko WA, Amores A, Wilson C, Murphy J, Currey M, Phillips P, Bell MA, Kimmel CB, Postlethwait JH. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc Natl Acad Sci U S A* **101**: 6050-6055.

Da Ines O, White CI. 2015. Centromere Associations in Meiotic Chromosome Pairing. *Annu Rev Genet* **49**: 95-114.

de Boer E, Stam P, Dietrich AJ, Pastink A, Heyting C. 2006. Two levels of interference in mouse meiotic recombination. *Proc Natl Acad Sci U S A* **103**: 9607-9612.

Delaneau O, Marchini J, Zagury JF. 2011. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**: 179-181.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491-498.

Dobzhansky T. 1930. Translocations Involving the Third and the Fourth Chromosomes of DROSOPHILA MELANOGASTER. *Genetics* **15**: 347-399.

Dobzhansky T, Sturtevant AH. 1938. Inversions in the Chromosomes of Drosophila Pseudoobscura. *Genetics* **23**: 28-64.

Dreau A, Venu V, Avdievich E, Gaspar L, Jones FC. 2019. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun* **10**: 4309.

Drouaud J, Mercier R, Chelysheva L, Berard A, Falque M, Martin O, Zanni V, Brunel D, Mezard C. 2007. Sex-specific crossover distributions and variations in interference level along Arabidopsis thaliana chromosome 4. *PLoS Genet* **3**: e106.

Dumont BL. 2017. Variation and Evolution of the Meiotic Requirement for Crossing Over in Mammals. *Genetics* **205**: 155-168.

Dumont BL, White MA, Steffy B, Wiltshire T, Payseur BA. 2011. Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res* **21**: 114-125.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285-311.

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. *Genetics* **159**: 1299-1318.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* **78**: 737-756.

Fernandes JB, Seguela-Arnaud M, Larcheveque C, Lloyd AH, Mercier R. 2018. Unleashing meiotic crossovers in hybrid plants. *Proc Natl Acad Sci U S A* **115**: 2431-2436.

Fernandez AI, Munoz M, Alves E, Folch JM, Noguera JL, Enciso MP, Rodriguez Mdel C, Silio L. 2014. Recombination of the porcine X chromosome: a high density linkage map. *BMC Genet* **15**: 148.

Fisher RA. 1930. *The genetical theory of natural selection*. The Clarendon press, Oxford,.

Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, Przeworski M. 2011. Variation in human recombination rates and its genetic determinants. *PLoS One* **6**: e20321.

Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C, Coop G, Przeworski M. 2009. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet* **5**: e1000658.

Fontaine DA, Davis DB. 2016. Attention to Background Strain Is Essential for Metabolic Research: C57BL/6 and the International Knockout Mouse Consortium. *Diabetes* **65**: 25-33.

Fowler KR, Sasaki M, Milman N, Keeney S, Smith GR. 2014. Evolutionarily diverse determinants of meiotic DNA break and recombination landscapes across the genome. *Genome Res* **24**: 1650-1664.

Froenicke L, Anderson LK, Wienberg J, Ashley T. 2002. Male mouse recombination maps for each autosome identified by chromosome painting. *American Journal of Human Genetics* **71**: 1353-1368.

Garcia-Cruz R, Pacheco S, Brieno MA, Steinberg ER, Mudry MD, Ruiz-Herrera A, Garcia-Caldes M. 2011. A comparative study of the recombination pattern in three species of Platyrrhini monkeys (primates). *Chromosoma* **120**: 521-530.

Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **97**: 11383-11390.

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

Glazer AM, Killingbeck EE, Mitros T, Rokhsar DS, Miller CT. 2015. Genome Assembly Improvement and Mapping Convergently Evolved Skeletal

Traits in Sticklebacks with Genotyping-by-Sequencing. *G3 (Bethesda)* **5**: 1463-1472.

Gray JC, Goddard MR. 2012. Sex enhances adaptation by unlinking beneficial from detrimental mutations in experimental yeast populations. *BMC Evol Biol* **12**: 43.

Gray S, Cohen PE. 2016. Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation. *Annu Rev Genet* **50**: 175-210.

Gruhn JR, Rubio C, Broman KW, Hunt PA, Hassold T. 2013. Cytological studies of human meiosis: sex-specific differences in recombination originate at, or prior to, establishment of double-strand breaks. *PLoS One* **8**: e85075.

Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Mol Ecol* **27**: 2477-2497.

Haig D, Grafen A. 1991. Genetic Scrambling as a Defense against Meiotic Drive. *J Theor Biol* **153**: 531-558.

Haldane JBS. 1922. Sex ratio and unisexual sterility in hybrid animals. *J Genet* **12**: 101-109.

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**.

Han L, Su B, Li WH, Zhao Z. 2008. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol* **9**: R79.

Hassold T, Hansen T, Hunt P, VandeVoort C. 2009. Cytological studies of recombination in rhesus males. *Cytogenet Genome Res* **124**: 132-138.

Hassold T, Hunt P. 2001. To ERR (meiotically) is human: The genesis of human aneuploidy. *Nature Reviews Genetics* **2**: 280-291.

Hawken RJ, Murtaugh J, Flickinger GH, Yerle M, Robic A, Milan D, Gellin J, Beattie CW, Schook LB, Alexander LJ. 1999. A first-generation porcine whole-genome radiation hybrid map. *Mamm Genome* **10**: 824-830.

He Y, Wang M, Dukowic-Schulze S, Zhou A, Tiang CL, Shilo S, Sidhu GK, Eichten S, Bradbury P, Springer NM et al. 2017. Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proc Natl Acad Sci U S A* **114**: 12231-12236.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311-318.

Hill WG, Robertson A. 1966. Effect of Linkage on Limits to Artificial Selection. *Genetics Research* **8**: 269-294.

Hillers KJ. 2004. Crossover interference. *Curr Biol* **14**: R1036-1037.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL et al. 2011. The landscape of recombination in African Americans. *Nature* **476**: 170-175.

Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, Bowden R, Donnelly P. 2019. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* **363**.

Housworth EA, Stahl FW. 2003. Crossover interference in humans. *Am J Hum Genet* **73**: 188-197.

Hunter N. 2015. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harb Perspect Biol* **7**.

Hurst DD, Fogel S, Mortimer RK. 1972. Conversion-associated recombination in yeast (hybrids-meiosis-tetrads-marker loci-models). *Proc Natl Acad Sci U S A* **69**: 101-105.

Huxley JS. 1928. Sexual difference of linkage in Gammarus chevreuxi. *J Genet* **20**: 145-156.

Ijiri TW, Merdiushev T, Cao W, Gerton GL. 2011. Identification and validation of mouse sperm proteins correlated with epididymal maturation. *Proteomics* **11**: 4047-4062.

Imai Y, Baudat F, Taillepierre M, Stanzione M, Toth A, de Massy B. 2017. The PRDM9 KRAB domain is required for meiosis and involved in protein interactions. *Chromosoma* **126**: 681-695.

Inoue K, Lupski JR. 2002. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* **3**: 199-242.

Jaramillo-Lambert A, Ellefson M, Villeneuve AM, Engebrecht J. 2007. Differential timing of S phases, X chromosome replication, and meiotic prophase in the C. elegans germ line. *Dev Biol* **308**: 206-221.

Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* **29**: 217-222.

Jeffreys AJ, Murray J, Neumann R. 1998. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell* **2**: 267-273.

Johnston SE, Berenos C, Slate J, Pemberton JM. 2016. Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis aries). *Genetics* **203**: 583-598.

Johnston SE, Huisman J, Ellis PA, Pemberton JM. 2017. A High-Density Linkage Map Reveals Sexual Dimorphism in Recombination Landscapes in Red Deer (Cervus elaphus). *G3 (Bethesda)* **7**: 2859-2870.

Jones FC, Brown C, Pemberton JM, Braithwaite VA. 2006. Reproductive isolation in a threespine stickleback hybrid zone. *J Evol Biol* **19**: 1531-1544.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55-61.

Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**: 203-206.

Kauppi L, Jasin M, Keeney S. 2013. How much is enough? Control of DNA double-strand break numbers in mouse meiosis. *Cell Cycle* **12**: 2719-2720.

Kauppi L, Jeffreys AJ, Keeney S. 2004. Where the crossovers are: Recombination distributions in mammals. *Nature Reviews Genetics* **5**: 413-424.

Kaur T, Rockman MV. 2014. Crossover heterogeneity in the absence of hotspots in Caenorhabditis elegans. *Genetics* **196**: 137-148.

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

Khil PP, Smagulova F, Brick KM, Camerini-Otero RD, Petukhova GV. 2012. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. *Genome Res* **22**: 957-965.

Kianian PMA, Wang M, Simons K, Ghavami F, He Y, Dukowic-Schulze S, Sundararajan A, Sun Q, Pillardy J, Mudge J et al. 2018. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat Commun* **9**: 2370.

Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* **173**: 419-434.

Koehler KE, Cherry JP, Lynn A, Hunt PA, Hassold TJ. 2002. Genetic control of mammalian meiotic recombination. I. Variation in exchange frequencies among males from inbred mouse strains. *Genetics* **162**: 297-306.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31**: 241-247.

Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, Villemoes R, Magnusdottir E, Olafsdottir SB, Thorsteinsdottir U, Stefansson K. 2014. Common and low-frequency variants associated with genome-wide recombination rate. *Nat Genet* **46**: 11-16.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**: 1099-1103.

Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, Jonsdottir GM, Gudjonsson SA, Sverrisson S, Thorlacius T et al. 2008. Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science* **319**: 1398-1401.

Kovalchuk I, Kovalchuk O, Kalck V, Boyko V, Filkowski J, Heinlein M, Hohn B. 2003. Pathogen-induced systemic plant signal triggers DNA rearrangements. *Nature* **423**: 760-762.

Lam I, Keeney S. 2014. Mechanism and regulation of meiotic recombination initiation. *Cold Spring Harb Perspect Biol* **7**: a016634.

Lam I, Keeney S. 2015. Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* **350**: 932-937.

Lam I, Mohibullah N, Keeney S. 2017. Sequencing Spo11 Oligonucleotides for Mapping Meiotic DNA Double-Strand Breaks in Yeast. *Methods Mol Biol* **1471**: 51-98.

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

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li HH, Gyllensten UB, Cui XF, Saiki RK, Erlich HA, Arnheim N. 1988. Amplification and Analysis of DNA-Sequences in Single Human-Sperm and Diploid-Cells. *Nature* **335**: 414-417.

Lichten M, Goldman AS. 1995. Meiotic recombination hotspots. *Annu Rev Genet* **29**: 423-444.

Liu EY, Morgan AP, Chesler EJ, Wang W, Churchill GA, Pardo-Manuel de Villena F. 2014. High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics* **197**: 91-106.

Lloyd A, Morgan C, FC HF, Bomblies K. 2018. Plasticity of Meiotic Recombination Rates in Response to Temperature in Arabidopsis. *Genetics* **208**: 1409-1420.

Long Ranger. 2018. Long Ranger. In *10X Genomics*. 10X Genomics, https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger.

Lu LY, Yu X. 2015. Double-strand break repair on sex chromosomes: challenges during male meiotic prophase. *Cell Cycle* **14**: 516-525.

Luo RB, Sedlazeck FJ, Darby CA, Kelly SM, Schatz MC. 2017. LRSim: A Linked-Reads Simulator Generating Insights for Better Genome Partitioning. *Comput Struct Biotec* **15**: 478-484.

Lynn A, Ashley T, Hassold T. 2004. Variation in human meiotic recombination. *Annu Rev Genomics Hum Genet* **5**: 317-349.

Ma J, Iannuccelli N, Duan Y, Huang W, Guo B, Riquet J, Huang L, Milan D. 2010. Recombinational landscape of porcine X chromosome and individual variation in female meiotic recombination associated with haplotypes of Chinese pigs. *Bmc Genomics* **11**: 159.

Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE et al. 2015. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS Genet* **11**: e1005387.

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

Manzano-Winkler B, McGaugh SE, Noor MA. 2013. How hot are drosophila hotspots? examining recombination rate variation and associations with nucleotide diversity, divergence, and maternal age in Drosophila pseudoobscura. *PLoS One* **8**: e71582.

Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci U S A* **98**: 5688-5692.

Marais G, Mouchiroud D, Duret L. 2003. Neutral effect of recombination on base composition in Drosophila. *Genet Res* **81**: 79-87.

Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2016. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLoS Genet* **12**: e1005887.

Marsolier-Kergoat MC, Yeramian E. 2009. GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome. *Genetics* **183**: 31-38.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **vol. 17**: pp. pp–10.

Martini E, Diaz RL, Hunter N, Keeney S. 2006. Crossover homeostasis in yeast meiosis. *Cell* **126**: 285-295.

McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MAF. 2012. Recombination Modulates How Selection Affects Linked Sites in Drosophila. *Plos Biology* **10**.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297-1303.

McKinnon JS, Rundle HD. 2002. Speciation in nature: the threespine stickleback model systems. *Trends in Ecology & Evolution* **17**: 480-488.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231-1241.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581-584.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* **21**: 984-990.

Mikawa S, Akita T, Hisamatsu N, Inage Y, Ito Y, Kobayashi E, Kusumoto H, Matsumoto T, Mikami H, Minezawa M et al. 1999. A linkage map of 243 DNA markers in an intercross of Gottingen miniature and Meishan pigs. *Anim Genet* **30**: 407-417.

Morelli MA, Cohen PE. 2005. Not all germ cells are created equal: aspects of sexual dimorphism in mammalian meiosis. *Reproduction* **130**: 761-781.

Morgan AP, Gatti DM, Najarian ML, Keane TM, Galante RJ, Pack AI, Mott R, Churchill GA, de Villena FP. 2017a. Structural Variation Shapes the Landscape of Recombination in Mouse. *Genetics* **206**: 603-619.

Morgan CH, Zhang H, Bomblies K. 2017b. Are the effects of elevated temperature on meiotic recombination and thermotolerance linked via the axis and synaptonemal complex? *Philos Trans R Soc Lond B Biol Sci* **372**.

Morgan TH. 1910. Sex Limited Inheritance in Drosophila. *Science* **32**: 120-122.

Morgan TH. 1911. Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**: 384.

Morgan TH. 1912. Complete linkage in the second chromosome of the male of *Drosophila*. *Science* **36**: 719-720.

Morgan TH. 1913. *Heredity and sex*. Columbia University Press, New York,.

Muller HJ. 1932. Some Genetic Aspects of Sex. *The American Naturalist* **66**: 118-138.

Muller HJ. 1964. The Relation of Recombination to Mutational Advance. *Mutat Res* **106**: 2-9.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321-324.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327**: 876-879.

Nachman MW, Churchill GA. 1996. Heterogeneity in rates of recombination across the mouse genome. *Genetics* **142**: 537-548.

Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos T R Soc B* **367**: 409-421.

Navarro A, Barton NH. 2003. Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* **300**: 321-324.

Neale MJ, Pan J, Keeney S. 2005. Endonucleolytic processing of covalent protein-linked DNA double-strand breaks. *Nature* **436**: 1053-1057.

Neff MW, Broman KW, Mellersh CS, Ray K, Acland GM, Aguirre GD, Ziegle JS, Ostrander EA, Rine J. 1999. A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics* **151**: 803-820.

Nei M. 1967. Modification of linkage intensity by natural selection. *Genetics* **57**: 625-641.

Noor MA, Grams KL, Bertucci LA, Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. *Proc Natl Acad Sci U S A* **98**: 12084-12088.

Novitski E, Braver G. 1954. An Analysis of Crossing over within a Heterozygous Inversion in Drosophila Melanogaster. *Genetics* **39**: 197-209.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I et al. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* **10**: e1004234.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet* **5**: e1000753.

Ortiz-Barrientos D, Engelstadter J, Rieseberg LH. 2016. Recombination Rate Evolution and the Origin of Species. *Trends Ecol Evol* **31**: 226-236.

Otto SP. 2009. The evolutionary enigma of sex. *Am Nat* **174 Suppl 1**: S1-S14.

Page SL, Hawley RS. 2004. The genetics and molecular biology of the synaptonemal complex. *Annu Rev Cell Dev Biol* **20**: 525-558.

Paigen K, Petkov P. 2010. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet* **11**: 221-233.

Paigen K, Petkov PM. 2018. PRDM9 and Its Role in Genetic Recombination. *Trends Genet* **34**: 291-300.

Paigen K, Szatkiewicz JP, Sawyer K, Leahy N, Parvanov ED, Ng SH, Graber JH, Broman KW, Petkov PM. 2008. The recombinational anatomy of a mouse chromosome. *PLoS Genet* **4**: e1000119.

Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND et al. 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144**: 719-731.

Petes TD. 2001. Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics* **2**: 360-369.

Petit M, Astruc JM, Sarry J, Drouilhet L, Fabre S, Moreno CR, Servin B. 2017. Variation in Recombination Rate and Its Genetic Determinism in Sheep Populations. *Genetics* **207**: 767-784.

Petkov PM, Broman KW, Szatkiewicz JP, Paigen K. 2007. Crossover interference underlies sex differences in recombination rates. *Trends Genet* **23**: 539-542.

Petronczki M, Siomos MF, Nasmyth K. 2003. Un menage a quatre: the molecular biology of chromosome segregation in meiosis. *Cell* **112**: 423-440.

Picard. 2018. Picard toolkit. (ed. Gr Broad Institute). Broad Institute, http://broadinstitute.github.io/picard/.

Plough HH. 1917. The effect of temperature on crossingover in Drosophila. *J Exp Zool* **24**: 147-209.

Powers NR, Parvanov ED, Baker CL, Walker M, Petkov PM, Paigen K. 2016. The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. *PLoS Genet* **12**: e1006146.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. DNA recombination. Recombination initiation maps of individual human genomes. *Science* **346**: 1256442.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans (vol 37, pg 429, 2005). *Nature Genetics* **37**: 445-445.

Qiao H, Prasada Rao HB, Yang Y, Fong JH, Cloutier JM, Deacon DC, Nagel KE, Swartz RK, Strong E, Holloway JK et al. 2014. Antagonistic roles of ubiquitin ligase HEI10 and SUMO ligase RNF212 regulate meiotic recombination. *Nat Genet* **46**: 194-199.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Rasmussen SW, Holm PB. 1984. The synaptonemal complex, recombination nodules and chiasmata in human spermatocytes. *Symp Soc Exp Biol* **38**: 271-292.

Reimchen TE. 1980. Spine deficiency and polymorphism in a population of Gasterosteus aculeatus: an adaptation to predators? . *Can J Zool* **58**: 1232-1244.

Reynolds A, Qiao HY, Yang Y, Chen JK, Jackson N, Biswas K, Holloway JK, Baudat F, de Massy B, Wang J et al. 2013. RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics* **45**: 269-278.

Rice WR. 2002. Experimental tests of the adaptive significance of sexual recombination. *Nat Rev Genet* **3**: 241-251.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**: 636-639.

Rockman MV, Kruglyak L. 2009. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet* **5**: e1000419.

Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome--patterns and consequences. *Mol Ecol* **22**: 3014-3027.

Ross JA, Peichel CL. 2008. Molecular cytogenetic evidence of rearrangements on the Y chromosome of the threespine stickleback fish. *Genetics* **179**: 2173-2182.

Rousselle M, Laverre A, Figuet E, Nabholz B, Galtier N. 2019. Influence of Recombination and GC-biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals versus Birds. *Mol Biol Evol* **36**: 458-471.

Salathe M, Kouyos RD, Bonhoeffer S. 2009. On the causes of selection for recombination underlying the red queen hypothesis. *Am Nat* **174 Suppl 1**: S31-42.

Samuk K, Owens GL, Delmore KE, Miller SE, Rennison DJ, Schluter D. 2017. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol Ecol* **26**: 4378-4390.

Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. 2012. Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genet* **8**: e1002854.

Sardell JM, Cheng CD, Dagilis AJ, Ishikawa A, Kitano J, Peichel CL, Kirkpatrick M. 2018. Sex Differences in Recombination in Sticklebacks. *G3-Genes Genom Genet* **8**: 1971-1983.

Schwarzacher T. 2003. Meiosis, recombination and chromosomes: a review of gene isolation and fluorescent in situ hybridization data in plants. *J Exp Bot* **54**: 11-23.

Scrucca L, Fop M, Murphy TB, Raftery AE. 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**: 289-317.

Shanfelter AF, Archambeault SL, White MA. 2019. Divergent Fine-Scale Recombination Landscapes between a Freshwater and Marine Population of Threespine Stickleback Fish. *Genome Biol Evol* **11**: 1573-1585.

Shapiro MD, Bell MA, Kingsley DM. 2006. Parallel genetic origins of pelvic reduction in vertebrates. *Proc Natl Acad Sci U S A* **103**: 13753-13758.

Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J. 2006. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *Plos Biology* **4**: 2227-2237.

Shifman S BJ, Copley RR, Taylor MS, Williams RW. 2006. A highresolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biology* **4**.

Shilo S, Melamed-Bessudo C, Dorone Y, Barkai N, Levy AA. 2015. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell* **27**: 2427-2436.

Sidhu D, Gill.K.S. 2004. Distribution of genes and recombination in wheat and other eukaryotes. *Plant Cell, Tissue and Organ Culture* **79**: 257–270.

Singh ND, Criscoe DR, Skolfield S, Kohl KP, Keebaugh ES, Schlenke TA. 2015. Fruit flies diversify their offspring in response to parasite infection. *Science* **349**: 747-750.

Singh ND, Stone EA, Aquadro CF, Clark AG. 2013. Fine-scale heterogeneity in crossover rate in the garnet-scalloped region of the Drosophila melanogaster X chromosome. *Genetics* **194**: 375-387.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN et al. 2015. Stable recombination hotspots in birds. *Science* **350**: 928-932.

Smagulova F, Gregoretti IV, Brick K, Khil P, Camerini-Otero RD, Petukhova GV. 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472**: 375-378.

Smeds L, Mugal CF, Qvarnstrom A, Ellegren H. 2016. High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. *PLoS Genet* **12**: e1006044.

Smith HF. 1936. Influence of temperature on crossing-over in Drosophila. *Nature* **138**: 329-330.

Smukowski Heil CS, Ellison C, Dubin M, Noor MA. 2015. Recombining without Hotspots: A Comprehensive Evolutionary Portrait of Recombination in Two Closely Related Species of Drosophila. *Genome Biol Evol* **7**: 2829-2842.

Spence R, Wootton RJ, Barber I, Przybylski M, Smith C. 2013. Ecological causes of morphological evolution in the three-spined stickleback. *Ecol Evol* **3**: 1717-1726.

Spence R, Wootton RJ, Przybylski M, Zieba G, Macdonald K, Smith C. 2012. Calcium and salinity as selective factors in plate morph evolution of the three-spined stickleback (Gasterosteus aculeatus). *J Evol Biol* **25**: 1965-1974.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philos Trans R Soc Lond B Biol Sci* **372**.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG et al. 2005. A common inversion under selection in Europeans. *Nat Genet* **37**: 129-137.

Stevison LS, Hoehn KB, Noor MAF. 2011. Effects of Inversions on Within- and Between-Species Recombination and Divergence. *Genome Biology and Evolution* **3**: 830-841.

Sturtevant AHD, T. 1931. *Contributions to the Genetics of Certain Chromosome Anomalies in Drosophila Melanogaster*. Carnegie institution of Washington.

Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW. 1983. The double-strand-break repair model for recombination. *Cell* **33**: 25-35.

Tanaka Y. 1914. A study of Mendelian factors in the silk worm *Bombx mori*. *Mol Gen Genet* **12**: 161.

Taylor EB, McPhail JD. 1999. Evolutionary history of an adaptive radiation in species pairs of threespine sticklebacks (Gasterosteus): insights from mitochondrial DNA. *Biol J Linn Soc* **66**: 271-291.

Tease C, Hulten MA. 2004. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet Genome Res* **107**: 208-215.

Tischfield SE, 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* **11**: 1496-1503.

Wadsworth CB, Li X, Dopman EB. 2015. A recombination suppressor contributes to ecological speciation in OSTRINIA moths. *Heredity (Edinb)* **114**: 593-600.

Wang F, Jiang L, Chen Y, Haelterman NA, Bellen HJ, Chen R. 2015a. FlyVar: a database for genetic variation in Drosophila melanogaster. *Database (Oxford)* **2015**.

Wang J, Fan HC, Behr B, Quake SR. 2012. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**: 402-412.

Wang S, Zickler D, Kleckner N, Zhang L. 2015b. Meiotic crossover patterns: obligatory crossover, interference and homeostasis in a single process. *Cell Cycle* **14**: 305-314.

Wang Z, Shen B, Jiang J, Li J, Ma L. 2016. Effect of sex, age and genetics on crossover interference in cattle. *Sci Rep* **6**: 37698.

Webster MT, Hurst LD. 2012. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet* **28**: 101-109.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757-767.

Wijnker E, Velikkakam James G, Ding J, Becker F, Klasen JR, Rawat V, Rowan BA, de Jong DF, de Snoo CB, Zapata L et al. 2013. The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife* **2**: e01426.

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GAT, Gabriel SB, Reich D, Donnelly P et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**: 107-111.

Withler RE, Mcphail JD. 1985. Genetic-Variability in Fresh-Water and Anadromous Sticklebacks (Gasterosteus-Aculeatus) of Southern British-Columbia. *Can J Zool* **63**: 528-533.

Wu Q, Chen M, Buchwald M, Phillips RA. 1995. A simple, rapid method for isolation of high quality genomic DNA from animal tissues. *Nucleic Acids Res* **23**: 5087-5088.

Wu TC, Lichten M. 1994. Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* **263**: 515-518.

Xie M, Wang J, Jiang T. 2012. A fast and accurate algorithm for single individual haplotyping. *BMC Syst Biol* **6 Suppl 2**: S8.

Yeaman S, Whitlock MC. 2011. The genetic architecture of adaptation under migration-selection balance. *Evolution* **65**: 1897-1911.

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

Zelkowski M, Olson MA, Wang M, Pawlowski W. 2019. Diversity and Determinants of Meiotic Recombination Landscapes. *Trends Genet* **35**: 359-370.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**: 303-311.

Zickler D, Kleckner N. 2015. Recombination, Pairing, and Synapsis of Homologs during Meiosis. *Cold Spring Harb Perspect Biol* **7**.

Ziolkowski PA, Underwood CJ, Lambing C, Martinez-Garcia M, Lawrence EJ, Ziolkowska L, Griffin C, Choi K, Franklin FC, Martienssen RA et al. 2017. Natural variation and dosage of the HEI10 meiotic E3 ligase control Arabidopsis crossover recombination. *Genes Dev* **31**: 306-317.

Appendix

Appendix Table 1: Details of 18 nuclear families used for individualized crossover map construction

Cross number	Ecotype	Category	Mother_id	Father_id	*Number of offspring
X1	Freshwater	Wild cross	Tyne8 _4	Tyne8_1	94
X4	Freshwater	Wild cross	Tyne8_7	Tyne8_2	93
X268	Freshwater	Lab cross	Tank 432_3	Tank 422_4	92
X284	Freshwater	Lab cross	Tank 533_15	Tank 432_16	92
X350	Freshwater	Lab cross	Tank 742_29	Tank 743_30	93
X351	Freshwater	Lab cross	Tank 743_31	Tank 742_32	93
X11	Marine	Wild cross	Tyne 2_16	Tyne 2_14	94
X20	Marine	Wild cross	Tyne 2_60	Tyne 2_56	91
X291	Marine	Lab cross	Tank532_21	Tank431_22	94
X294	Marine	Lab cross	Tank532_23	Tank431_24	94
X295	Marine	Lab cross	Tank532_25	Tank431_26	93
X296	Marine	Lab cross	Tank532_27	Tank431_28	93
X273	FW X Marine	Lab cross	Tank523_7	Tank822_8	94
X274	FW X Marine	Lab cross	Tank523_9	Tank822_10	94
X800	FW X Marine	Lab cross	Tank621_51	Tank822_52	94
X366	Marine X FW	Lab cross	Tank521_33	Tank954_34	86
X389	Marine X FW	Lab cross	Tank423_39	Tank823_40	93
X391	Marine X FW	Lab cross	Tank423_43	Tank823_44	93

***Number of offspring in the final data set is given**

Appendix Table 2: Number of crossovers included in the analysis at different scales of resolution

Category	Number of crossovers		
	Total	Male	Female
All	49848	18039	31809
Scaffold gap excluded	46386	16136	30250
In 1 Mb analysis	46379	16135	30244
In 100 kb analysis	45445	15803	29642
In 50 kb analysis	44492	15479	29013
In 10 kb analysis	39490	13557	25933
In 5kb analysis	35180	12035	23145

Appendix Table 3: Top five 1 Mb and 5 kb bins with highest difference in recombination rate (CO count) between marine and freshwater ecotypes are listed. CO count for all three ecotypes in those bins are given (sexes separate).

Scale	Sex	Chromosome	Start (bp)	Stop (bp)	FW	Mar	Hyb
1 Mb	Male	chrI	1000000	2000000	73	103	58
		chrVII	25000000	26000000	64	37	48
		chrXVIII	13000000	14000000	32	58	41
		chrV	10000000	11000000	61	86	74
		chrXVII	1000000	2000000	42	66	69
	Female	chrI	18000000	19000000	66	22	29
		chrXIX	17000000	18000000	11	50	37
		chrXVII	12000000	13000000	92	55	55
		chrXVIII	1000000	2000000	83	48	40
		chrXI	15000000	16000000	76	44	45
5 kb	Male	chrXVI	17340000	17345000	0	6	0
		chrXX	2195000	2200000	4	0	1
		chrIII	14780000	14785000	4	0	0
		chrI	27345000	27350000	4	0	0
		chrI	1615000	1620000	0	4	0
	Female	chrXXI	11395000	11400000	7	2	2
		chrXVIII	2580000	2585000	5	0	3
		chrXXI	11400000	11405000	6	2	2
		chrXII	1330000	1335000	4	0	0
		chrIX	19725000	19730000	4	0	0

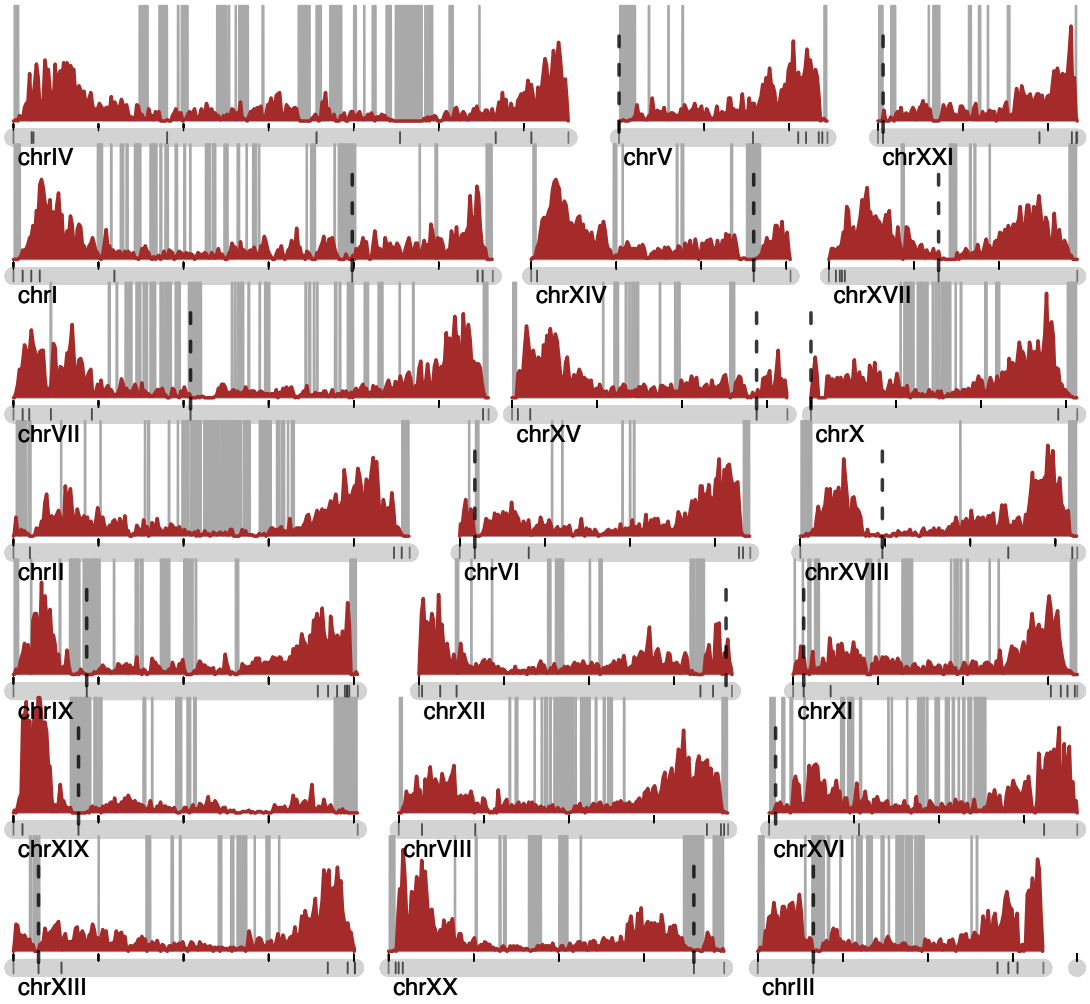
Appendix Table 4: Coordinates of five largest coldspots identified in different categories

Category	Chromosome	start	stop	Size (Mb)
All data (sex combined)	chrIV	22462382	24024373	15.62
	chrXIX	3359171	4392440	10.33
	chrIV	16752534	17430664	6.78
	chrIX	4108782	4745341	6.37
	chrXIX	19566225	20190660	6.24
All male	chrXIX	2672711	20190660	17.52
	chrVII	8511536	15243820	6.73
	chrIV	20750929	25984768	5.23
	chrIV	16503285	20410944	3.91
	chrI	18947087	22318843	3.37
All female	chrIV	22462382	24024373	1.56
	chrXIX	3359171	4392440	1.03
	chrIV	16752534	17430664	0.68
	chrIX	4108782	4745341	0.64
	chrXIX	19566225	20190660	0.62
Freshwater male	chrXIX	2419940	20190660	17.77
	chrIV	15120322	25984768	10.86
	chrVII	8511536	15439811	6.93
	chrVII	15440466	21888894	6.45
	chrII	10210378	15228900	5.02
Marine male	chrXIX	2499381	20190660	17.69
	chrVII	7527257	18347954	10.82
	chrIV	16306872	27092827	10.79
	chrI	16994643	24207760	7.21
	chrIX	2338043	6737080	4.40
Hybrid male	chrXIX	2672711	20190660	17.52
	chrVII	7270324	15243820	7.97
	chrIV	20750929	28471406	7.72
	chrI	16785195	23270078	6.48
	chrI	10850286	16779218	5.93
Freshwater female	chrIV	21915156	24587745	2.67
	chrXIX	18685599	20190660	1.51
	chrXIX	3303022	4567462	1.26
	chrXII	15755051	16884814	1.13
	chrXIV	12674899	13530291	0.86
Marine female	chrVII	10416482	12001843	1.59
	chrIV	22462382	24024373	1.56
	chrX	6967355	8042903	1.08
	chrXIX	3359171	4392440	1.03
	chrIV	8546073	9459352	0.91
Hybrid female	chrIV	22215196	24350727	2.14
	chrXIX	3203056	4539055	1.34
	chrXIV	12132909	13317434	1.18
	chrXVII	6498206	7505508	1.01
	chrVIII	9361689	10363587	1.00

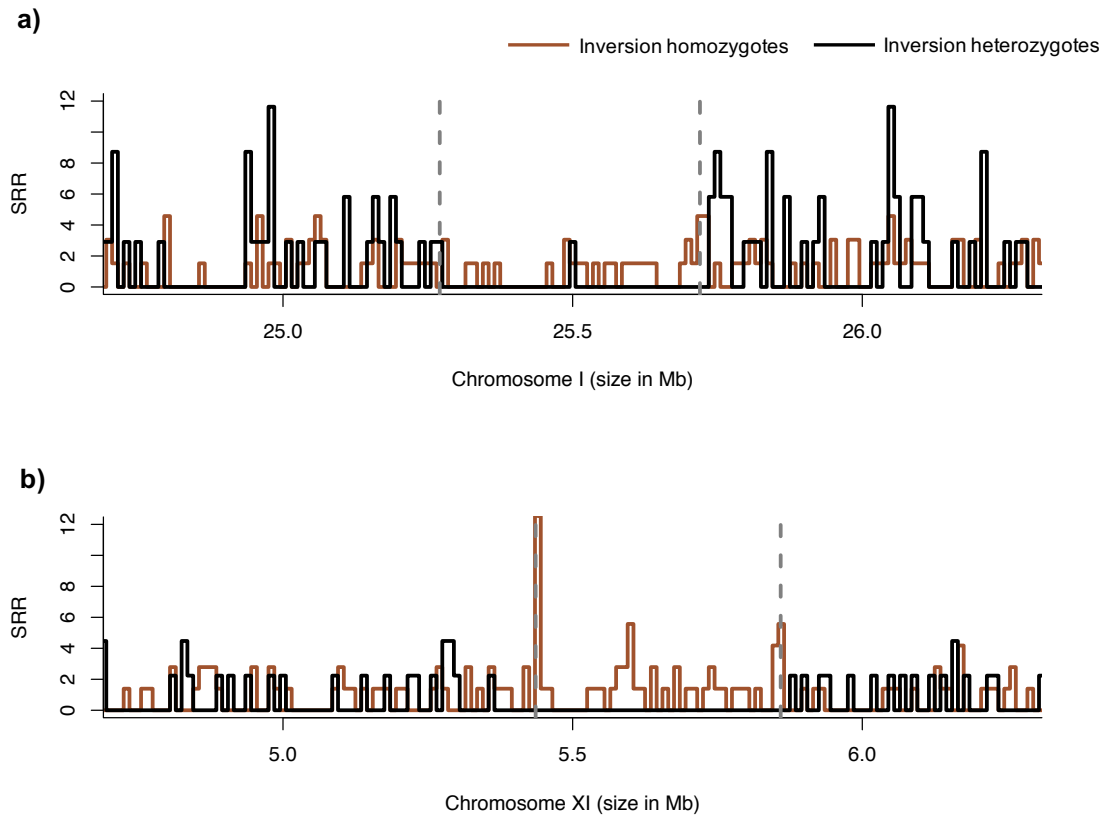
Appendix Table 5: Details of chromosomal inversions (compared to reference genome) detected in this data set.

Sex	Inversion at	Inversion detected families	Ecotypes
Male	chrI: 25,264,236 – 25,720,158	X294	Marine
	chrII : 22,372,205-23,174,871	X268, X294, X296, X350, X351, X366, X389, X391	Marine, FW, Hybrid B
	chrIX: 5,655,297 – 7,950,866	X389	Hybrid B
	chrXI : 5,431,984 – 5,868,073	X291	Marine
	chrXI : 15,730,574- 16,638,140	X11, X20, X268, X284, X291, X294, X296, X366,	Marine, FW, Hybrid B
	chrXVI : 17,195,676 – 17,968,133	X11, X20, X268, X273, X274, X284,X294, X296, X350, X351, X366, X389, X4, X800	Marine, FW, Hybrid B, Hybrid A
	chrXVII:641211-769373	X391	Hybrid B
	chrXXI : 5,681,441 – 7,787,895	X1, X11, X284, X294, X296, X350, X351,X366, X389, X4, X800	Marine, FW, Hybrid B, Hybrid A
Female	chrI: 25,267,445 – 25,725,739	X11, X291	Marine
	chrII: 22,358,302- 23,062,538	X1, X268, X366, X800	Marine, FW, Hybrid A, Hybrid B
	chrIX : 5,647,525- 7,282,910	X11, X366	Marine, Hybrid B
	chrXI : 5431082 -5858743	X11, X20, X294, X389, X800	Marine, Hybrid A, Hybrid B
	chrXI : 15,734,076 – 16,500,546	X1, X11, X273, X296, X800	Marine, FW, Hybrid A,
	chrXVI : 15,870,586- 17,836,051	X1, X11, X273, X20, X274, X284, X291, X294,	Marine, FW, Hybrid A, Hybrid B
	chrXVII:641353-769480	X350	FW
	chrXXI : 5,726,992 -7,494,035	X11, X268, X273, X284, X291, X294, X295, X296, X350,351, X366, X389, X391, X800	Marine, FW, Hybrid A, Hybrid B

***Hybrid A: (FW x Marine) F1; Hybrid B: (Mar x FW) F1**



Appendix figure 1: Coldspots identified across the genome appear to be clustered at chromosome center. Sex averaged recombination landscape of all 36 individuals combined is shown in maroon. Regions identified as coldspots are marked with grey vertical bars. Approximate centromere locations are marked with black dotted lines.



Appendix figure 2: Crossover suppression within chrI and chrXI inversion heterozygotes. Recombination rate within inversion and left and right flanking regions of the same size in (a) chromosome I and (b) chromosome XI inversions are shown. Standardized recombination rate (SRR) in 10 kb sliding windows across the region is plotted. Dotted grey vertical lines mark the inversion boundaries. SRR for inversion homozygotes (brown) and heterozygotes (black) and are overlaid**.** A complete suppression of recombination within the inversion boundaries is seen in chrXI inversion heterozygotes where as a single crossover event is identified in chrI inversion heterozygotes.

Home-made library preparation protocol

This protocol is optimized for high-throughput preparation of genomic DNA libraries compatible for sequencing on Illumina HiSeq 3000.

All DNA size selections and clean ups were performed following Ampure XP® (Beckman-Coulter) SPRI bead size selection protocol.

Step 1: DNA fragmentation using Covaris® LE220 instrument

- Dilute 300-500 ng of good quality genomic DNA in 1X TE buffer to a total volume of 130 µl and transfer to covaries 96 well microtube plate (SKU:520078)

- Insert sample filled plate into the Covaris® LE220 plate holder and run with the following settings to obtain an average fragment size of 300bp.

Sample volume	130µl
Duty factor	30%
Peak Incident Power	450
Cycles per Burst	200
Treatment time	80 sec

- Retrieve sheared samples from Covaris® plate into a 96 well plate and perform bead clean up. The following method is designed to yield fragments of size 300 bp or above.

Input DNA:	130µl
SPRI volume	104µl (0.8 times the sample volume)
Resuspension volume	30µl
Elution volume	30µl

Step 2: End repair

- Transfer 15µl of the eluate from above into a fresh plate for end repair
- Prepare the following reaction mix and add it to the DNA

Reagent	**1x**
Water	3.75 µl
10X T4 DNA ligase buffer	2.5 µl
10mM dNTP mix	1 µl
T4 DNA polymerase	1.25 µl
Klenow DNA polymerase	0.25 µl
T4 Polynucleotide kinase	1.25 µl
Total volume	**10 µl**

- Add 10 µl of reagent into 15 µl of sample. Mix well and spin down. Incubate at room temperature for 30 minutes

- Perform DNA size selection to remove fragments larger than 500 bp.

Input DNA	25 µl
SPRI volume	15 µl (0.6 times the sample volume)
PEG to save	40 µl (saving fragments < 500 bp)
SPRI volume	15 µl (add to the saved PEG to extract all smaller sized DNA)
Resuspension volume	18 µl
Final elution volume	17 µl

- Store samples in fridge if not proceeding right away

Step 3: A-tailing

- Prepare the following master mix and add it to the sample

Reagent	**1x µl**
NEB Buffer 2	2.5 µl
1 mM dATP	0.5 µl
Klenow exo-	0.5 µl
Water	4.5 µl
Total	**8 µl**

- Incubate at 37^0C for 30 minutes, then heat inactivate at 75^0C for 20 minutes
- Save 1 µl sample for running on bioanalyzer or 2 µl for checking on gel
- Directly proceed to the next step without waiting

Step 4: Adapter ligation

- Mix 1 µl 10 µM adapters with DNA sample (Illumina TruSeq single index adapter ordered from IDT)

  Note down the unique barcode used for each sample.
- Add the following ligase master mix to each sample

Reagent	**1x**
10mM rATP	3 µl
T4 DNA ligase	1.5 µl
Water	0.5 µl
Total	**5 µl**

- Incubate15 minutes at 20^0C. Then 5 minutes at 65^0C to inactivate the ligase. Allow samples to cool before proceeding

- Perform bead clean-up to remove the enzymes

DNA volume:	31 µl
SPRI volume	24.8 µl (0.8 times the sample volume)
Resuspension volume	15 µl
Elution volume	14 µl

Step 5: Library enrichment by PCR

- Add the following ligase master mix to each sample

Reagent	**1x**
Water	2.75 µl
5x Buffer	5 µl
10mM dNTP	0.5 µl
10µM Truseq PCR1	1.25 µl
10µM Truseq PCR2	1.25 µl
Phusion polymerase	0.25 µl
Total	**11 µl**

- Perform PCR reaction with the following condition

95^{0}C	30 sec	1 cycle
98^{0}C	15 sec	6 cycles
62^{0}C	30 sec	
72^{0}C	30 sec	
72^{0}C	10 min	1 cycle
4^{0}C	Hold	Infinitely

Step 6: Library validation and pooling

- Confirm adapter ligation by running the sample on bio-analyzer or checking on gel (load 2 µl sample). Successful adapter ligation will increase the fragment size by ~120 bp.
- Measure final library concentration by TECAN plate reader (using picogreen dye)
- Pool equal quantity of all libraries to be sequenced in a lane
- Perform bead clean up on pooled libraries to remove PCR reagents
  If pooled sample volume is less than 50 µl, make up to 50 µl by adding EB buffer

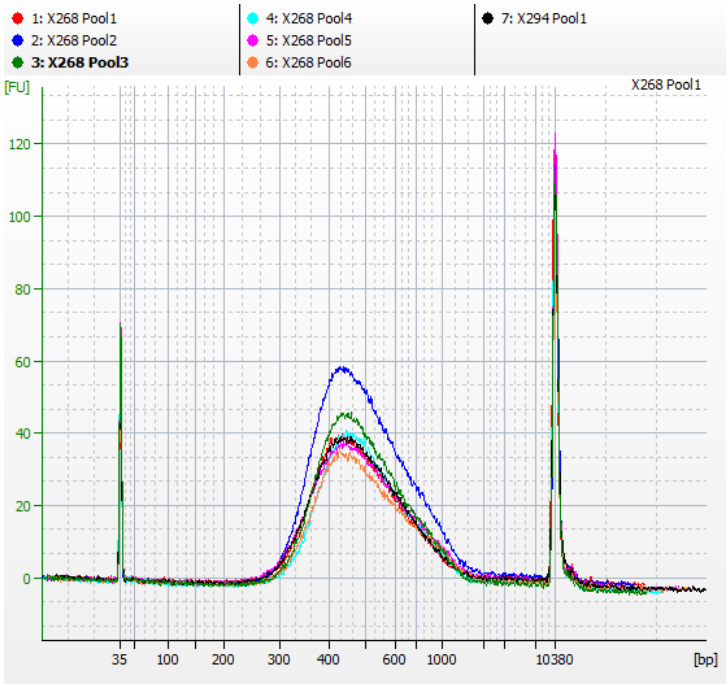
Sample volume	50 µl
SPRI volume	40 µl (0.8 times the sample volume)
Resuspension volume	61 µl
Elution volume	60 µl

- Perform DNA double size selection to remove fragments larger than 600 bp and smaller than 300 bp

Input DNA	60 µl
SPRI volume	24 µl (0.4 times the sample volume)
PEG to save	84 µl (saving fragments < 600 bp)
SPRI volume:	18 µl (add to the saved PEG to select all fragments >300 bp)
Resuspension volume	31 µl
Final elution volume	30 µl

- Check the quality of the library pools by bio-analyzer (Appendix figure 3) and measure the quantity by qubit high-sensitivity reagent.

- Submit 2.5nM library for sequencing



Appendix figure 3: Bioanalyzer profile of size selected library pools. Library size profile of 7 pools are shown. All of them have an average fragment size of about 420 bp (300 bp insert + 120 bp adapter).

Reagents used for library preparation

Reagent	NEB catalogue number
T4 DNA polymerase	M0203L
T4 Polynucleotide kinase	M0201L
Klenow exo-	M0212L
T4 DNA ligase	M0202L
Klenow DNA polymerase	M0210L or M0210S
Phusion High-Fidelity DNA polymerase	M0530L