Title
Four Statistical Explorations of Genetic Variation

Author
Brandt, Alexander Joseph

Publication Date
2018

Peer reviewed|Thesis/dissertation

Four Statistical Explorations of Genetic Variation

by

Alexander Joseph Brandt

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Daniel Rokhsar, Co-chair
Professor David Wemmer, Co-chair
Professor Phillip Geissler
Professor Rasmus Nielsen

Fall 2018

Four Statistical Explorations of Genetic Variation

Abstract

Four Statistical Explorations of Genetic Variation

by

Alexander Joseph Brandt

Doctor of Philosophy in Chemistry

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Daniel Rokhsar, Co-chair

Professor David Wemmer, Co-chair

Genetic variation is the result of DNA sequence differences between individuals or populations. Our understanding of genetic variation has benefited greatly from advances in DNA sequencing, computational power, and statistical tools. I will present four statistical analyses of genetic variation related to genetic mapping, genome editing, evolutionary biology, and *de novo* genome assembly and annotation.

First, I will present a novel methodology for genotype imputation and composite genetic marker construction. This technique, minimum spanning tree imputation, is designed to improve linkage maps in outbred F1 crosses, particularly from genetic markers derived from low confidence sequencing. I then use minimum spanning tree imputation to construct sex-specific genetic maps for *Xenopus laevis*, *Branchiostoma floridae*, and *Miscanthus sinensis*.

Second, I will present a case of three mouse lines edited with the CRISPR/Cas9 system in order to delete an enhancer locus in the IL2RA gene. All three lines were confirmed to have the intended deletion. Curiously, one line displayed a severe immune deficient phenotype that persistently bred true. By resequencing mice from these lines, I was able to identify the occurrence of a tandem duplication as an off-target consequence of the editing in the immune compromised individuals that was absent in the other lines. The tandem duplication was then confirmed experimentally. I close by proposing a repair mechanism mediated by microhomology that might have caused the tandem duplication to form.

Third, I will compare the chromosomal position of orthologous genes between lancelet amphioxus, *Branchiostoma floridae* (an early-branching living chordate) and five vertebrates. These comparisons will offer support for the "2R hypothesis," that early vertebrate organisms underwent two rounds of whole genome duplication. This analysis takes advantage of a novel way of computing and visualizing mutual best hits between previously identified chordate linkage groups (CLGs) and segments of vertebrate chromosomes.

Finally, I will present my contributions to the assembly and annotation of the genome of a regenerative model organism, *Hofstenia miamia*.

For Mom

Contents

Acknowledgments

This work could not have happened with the help and kindness of so many. First and foremost, I would like to express my profound thanks to Professor Daniel Rokhsar for his tireless and steadfast support. Our explorations into statistical mechanics and genetics will always remain of critical importance to my development as a person, investigator, and student. You are one of the rare individuals who understands so many topics in phenomenal detail, and teaches with genius, kindness, and zeal. You are also one of the greatest mentors with whom I've ever had the privilege of working. I'd also like to thank my labmates – Jessen Bredeson, Austin Mudd, Jess Lyons, Adam Session, Sofía Medina-Ruiz, Therese Mitros and Sanjit Batra for their teamwork and mentorship.

I have had the great privilege to collaborate with so many amazing individuals, and would like to thank Emma Farley and Dimitre Simeonov for including me on such exciting projects.

Mom and Dad, your continued and enthusiastic support for my educational pursuits has made it so easy and exciting to follow my dreams. Thank you for always supporting me during times of triumph and challenge, and for always making sure I felt close to home 2,800 miles away. I love you more than words could ever say.

Patricia and Courtney, you are the greatest cheering section a person could ask for. Thank you for always being the most incredible people in the world that I am fortunate enough to call "sisters." Thank you for being the greatest friends on the planet.

Jeff, thank you for hundreds of miles in the Falcon, and millions of laughs along the way. I love you like a brother, brother.

Alan, Eduardo, Newton, and all my teammates and coaches at Ralph Gracie Berkeley, thank you for keeping me sound of mind and body.

Stephen, Alexandra, Nathan, Matthew, and Tamara, thank you for the weekly adventures. I will never forget them.

And Charlotte, thank you for your constant kindness, support, and love.

Chapter 1

Introduction

1.1 Basics and history of DNA sequencing

Insights in molecular biology, developmental biology, and genetics have been greatly aided by improvements in nucleotide sequencing and analysis. A fairly consistent trend in the field is that increasing volumes of DNA sequencing data has enabled discoveries, and new technologies have arisen to meet ever increasing demand. The Sanger DNA sequencing methodology was and remains a reliable method for sequencing short (usually under 1 kb), amplified regions. Originally, four separate reactions were run. DNA polymerase moves along the template, until one of the four nucleotide bases is added as a dideoxynucleotide (ddNTP), terminating further elongation of the template reaction. The four reactions are run on a gel, and their position juxtaposed with a ladder determines which base is positioned along the template. In more modern incarnations of Sanger sequencing, ddNTPs are fluorescently labeled and all 4 reactions are run together. As the DNA fragments are separated by size through capillary electrophoresis, and as incoming fragments encounter fluorescent stimulation, a chromatogram trace is produced, where each color peak corresponds to an adenine, thymine, cytosine, or guanine moiety. Though most of this dissertation deals with more modern sequencing technologies, Sanger sequencing is still used frequently to this day as confirmation of novel findings such as mutations or structural variations. Indeed, in chapter 3, we will make use of Sanger traces to confirm findings that were initially discovered by more modern sequencing methods.

The resequencing of human DNA to study genetic variation became more routine after the completion of the Human Genome Project ("HGP") and the development of High Throughput Sequencing ("HTS," also known as Next-Generation Sequencing, "NGS"). HTS can produce millions of short sequence reads per sample for a fraction of the cost of Sanger sequencing. These sequencing data can be aligned and analyzed for differences with respect to the reference sequences produced by the HGP. These single nucleotide polymorphisms (SNPs) and short insertions or deletions (INDELs) have created an entirely new understanding of human genetic variation that is used for ascertaining the basis of phenotypic

traits and diseases. Methods for *de novo* assembly of these short sequences have also been developed, making it routine to produce the genome of any organism without the need for massive, multi-center efforts.

HTS as implemented by Illumina offers far larger volumes of data at a significantly reduced price than Sanger sequencing methods. Furthermore, it can be used to perform both amplified and amplification-free sequencing. There are considerable advantages to this method that we will see in later chapters. Illumina sequencing operates by use of a microfluidic chip with a "lawn" of evenly spaced single stranded DNA adapters. Complementary adapters can be ligated to the ends of size-selected DNA fragments of interest. Then, clonal clusters are produced through a process known as "bridge amplification," where all resulting clonal molecules are located in close proximity to the original DNA fragment on the array [1]. This clustering steps allows for a sufficient signal of fluorescently labeled nucleotides, which are visualized via 2 or 4 channel laser stimulation and captured in real time. The localized sequence of flashing colors represents the sequence content of that DNA cluster. A flow cell can contain hundreds of millions of such clusters.

This methodology can be applied to myriad other sequencing assays, such as RNA sequencing. In this case, RNA is reverse transcribed to cDNA before being eventually sequenced in a similar fashion. Other assays utilize similar transformations, though the specific chemistry is highly assay dependent.

Illumina paired-end library preparation ligates adapters at both ends of the sequence fragment. In this way, the end-sequences of each DNA fragment can be determined. For such "paired ends" (also called "mate pairs,") both the sequencing and orientation (i.e., strand) are known. There are several advantages to this information. One is that reads with similarity to several places in the genome can be uniquely mapped if their "mate" maps to a less repetitive sequence nearby. Another benefit is that reads mapped with unexpected orientations can signal structural differences between the sample and the reference haplotype. For example, if the read pairs map to different chromosomes, it could signal a chromosomal rearrangement. I will make use of read-pair orientation identify an interesting structural variant in chapter 3.

Single molecule sequencing technologies, sometimes referred to in the literature as "third-generation sequencing," attempt to circumvent the needed for clonal cluster generation. These technologies (e.g. Pacific Biosciences SMRT sequencing, Oxford Nanopore Technologies [2, 3]) focus on creating signal from single molecules, which allows for longer read lengths, and while none of these technologies are focused on in this dissertation, they remain and important and transformative technology for genomics and biological sequence analysis.

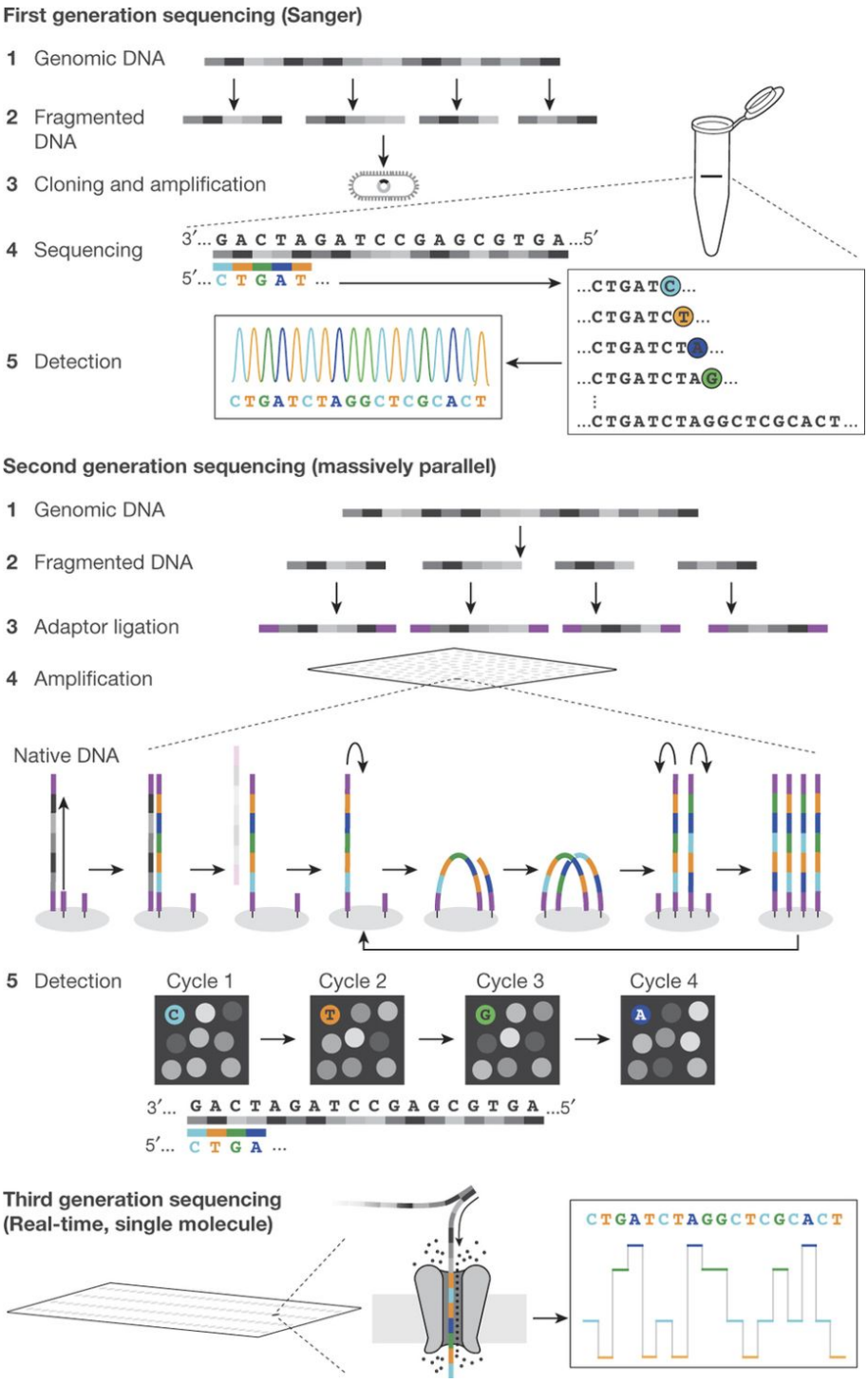


Figure 1.1: Sequencing technologies throughout history. Reprinted with permission from Springer Nature.

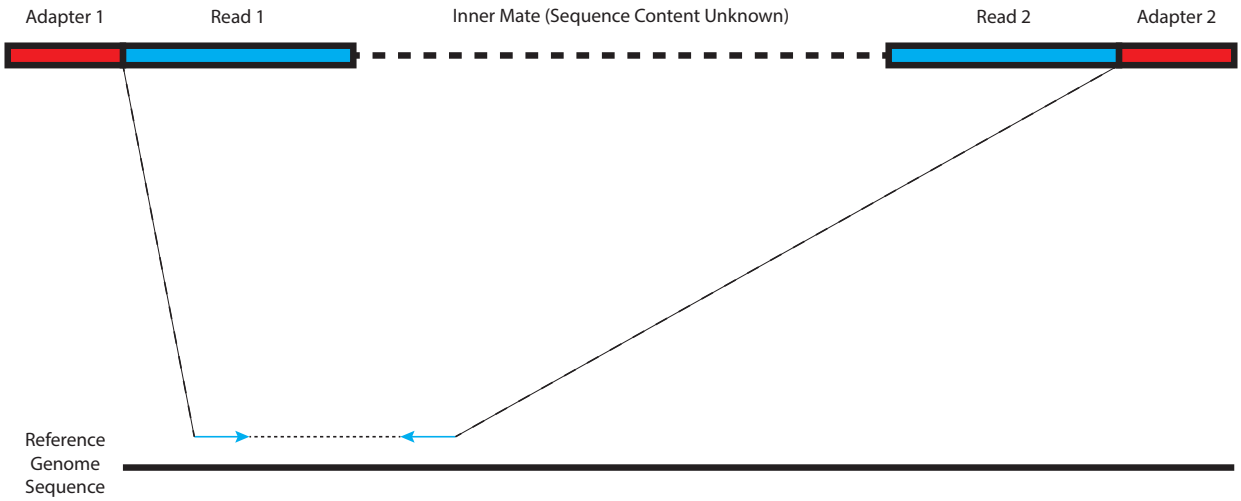


Figure 1.2: Paired-end reads improve mapping fidelity and provide structural information.

Quality scores

In most sequencing assays, a quality score is typically associated with each base, corresponding to the probability of correct moiety identification, which is associated with each nucleotide that is read. These qualities, known as a Phred score, are logarithmically scaled to the probability of correctly reading the base:

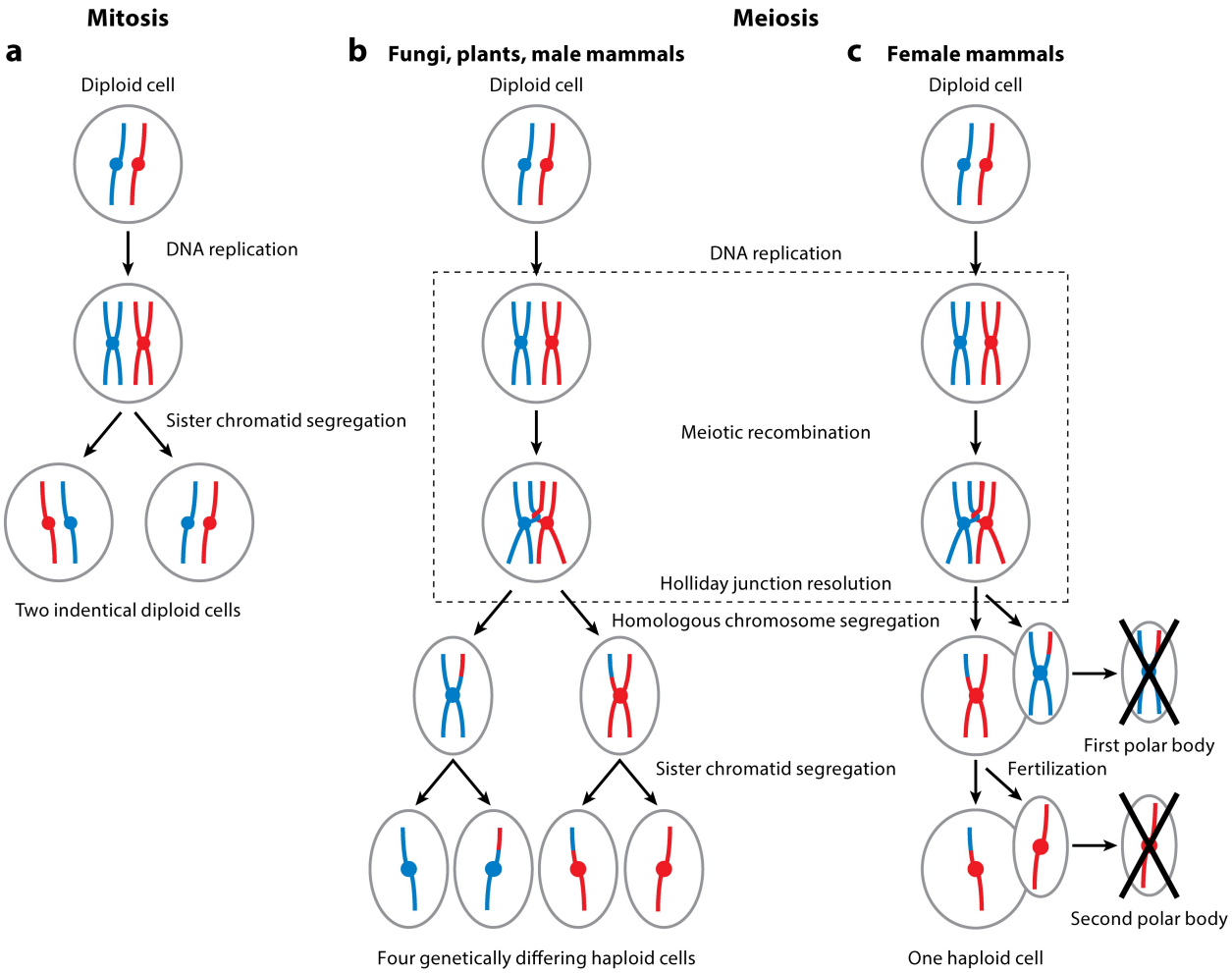
$$Q = -10 \log_{10} P$$

and are a useful quantity to propagate through various genomic analyses. Here P represents the probability the nucleotide is incorrect, estimated based on signal processing.

1.2 Genetic mapping and linkage analysis

Background

A haplotype is the nucleotide sequence of the chromosome an individual inherited from a parent. Contrary to Mendel's Second "Law," sexual reproduction does not produce simple assortments of parental chromosomes. When gametes are created, homologous regions of parental chromosomes can pair and recombine to produce new haplotype structures via meiotic homologous recombination. These recombinations increase genetic diversity as well as help orient segregating homologous chromosomes pairs. Recombination is a random process that is not uniform along the chromosome, and quantifying the relationship between physical distance, in base pair or genomic coordinates, and the propensity for recombination in a given section of DNA is the goal of a genetic map.



AR Gray S, Cohen PE. 2016.
Annu. Rev. Genet. 50:175–210

Figure 1.3: Genetic recombination occurs in meiosis. Reprinted with permission from Annual Reviews.

Molecular markers

Studying phenotypic differences of offspring in model organisms led biologists to understand that some traits do not assort independently. Syntenic genes, that is, genes located on the same chromosome, were the reason for these linked traits. We now use molecular or genetic markers, DNA sequence polymorphisms that can be definitively associated with a known region of a genome, to track linkage. In the context of the three genetic maps I will present in chapter 3, these genetic markers are derived from genetic variants discovered from whole genome shotgun (WGS) sequencing.

Pseudo-testcrosses

Traditional genetic maps have relied on inbred lines in which the parental linkages (i.e. haplotypes) are known. As assembling new genomes becomes an increasingly routine task, it is also becoming easier to understand the recombination landscape of an organism's genome, mapping populations through only one round of breeding. Multigenerational inbred lines are more costly and tedious to develop than a simple outbred cross, but can be replaced with outbred individuals if a sufficiently large number of genetic markers are sequenced.

By analyzing markers where one parent is heterozygous but the other parent is homozygous, we can use the genotypic state of that heterozygous site in the progeny as an indicator of which haplotype was inherited along that stretch of the genome from the heterozygous parent. These sites will show a Mendelian 1:1 ratio of inherited alleles for the heterozygous parent. This formulation of the F1 cross is described in the literature as a "two-way pseudo-testcross" [4], since it mimics Mendel's testcross methodology. With a sufficient number of progeny, a single F1 outbred cross can be used to develop a genetic map. Ideally, as many progeny as possible should be used in the creation of the map, as more progeny offer more opportunities to observe crossovers in the mapping population. Note that runs of homozygosity can foil this approach.

My imputation method described later will take advantage of the pseudo-testcross structure for segregating loci (D class markers in Wu et al. 2001's table of marker configurations taxonomy that has become standard in the field). In the pseudo-testcross, we ignore non-informative markers (*aa* x *bb* or *ab* x *ab*) and focus on informative markers which have the form:

$$ab \text{ x } aa$$

or

$$ab \text{ x } oo$$

and produce individuals of *aa* or *ab* in the first notation, or *ao* and *bo* in the second notation.

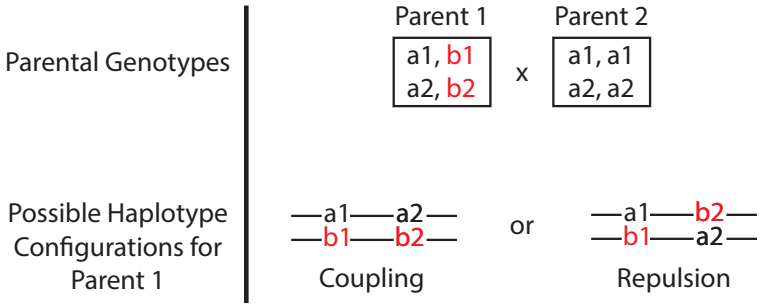


Figure 1.4: Markers used in the pseudo-testcross are either coupled or repulsive in the heterozygous parent.

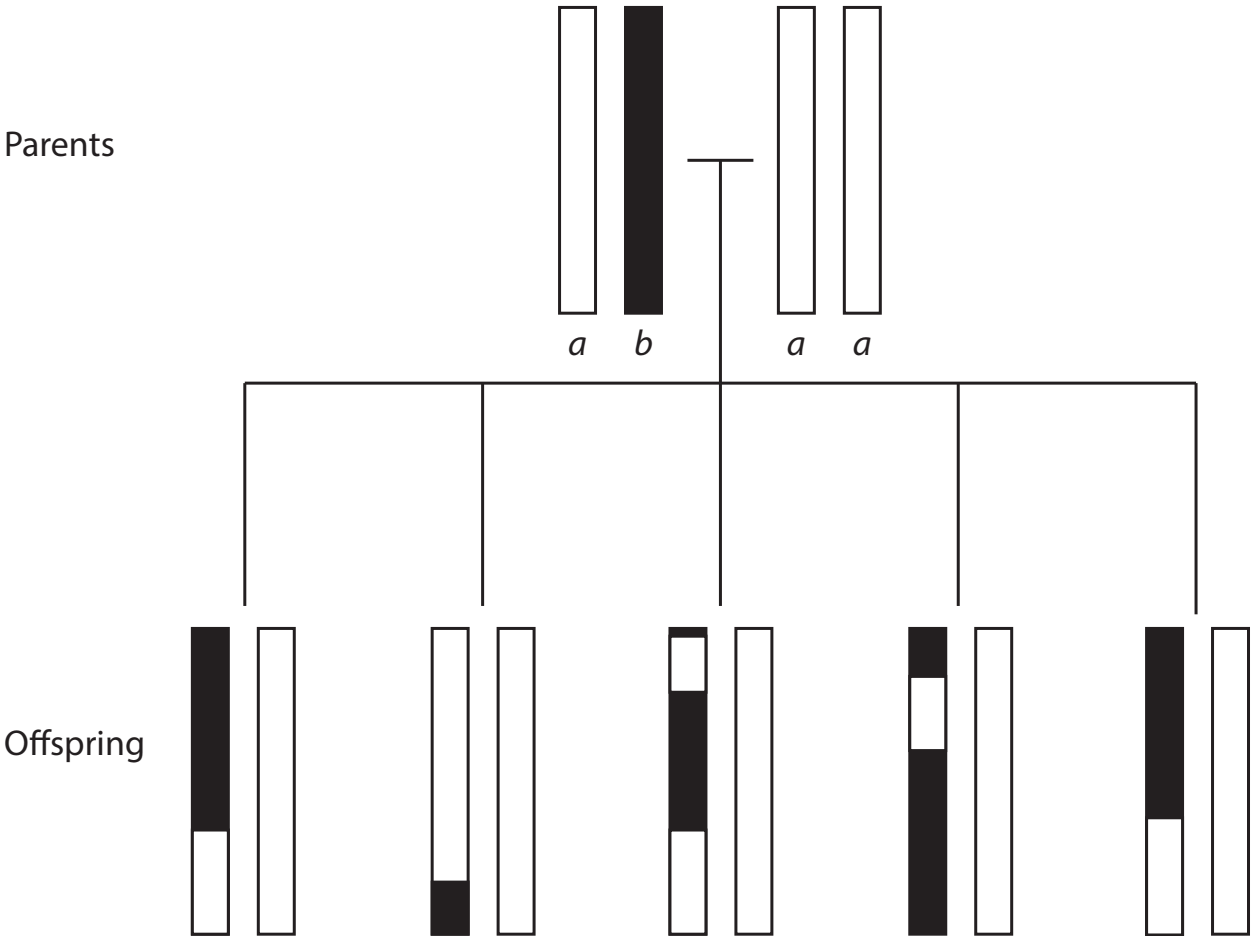


Figure 1.5: Illustrating recombinations and identification of crossovers with a theoretical F1 pseudo-testcross and perfectly dense markers

Linkage groups

Before we can order genetic markers, we must first group them into sets of linked markers. Markers on the same chromosomes should have sufficiently high association with one another, and be essentially unrelated with markers on other chromosomes. Ideally, each group corresponds to markers derived from the same chromosome, and the total number of linkage groups would be equal to the number of haploid chromosomes in a given organism's karyotype. A common method for generating these linkage groups is the two-point recombination fraction calculation via LOD (Log-of-odds) Scores between sites i, j. For a pedigree of N offspring with R recombinant individuals, and a recombination fraction of $r = R/N$,

this value:

$$\begin{aligned} LOD &= \log_{10} \frac{L_{H_A}}{L_{H_0}} \\ &= \log_{10} \frac{(1-r)^{N-R}(r)^R}{0.5^N} \end{aligned}$$

represents the log odds ratio of the likelihood of the test hypothesis that the two sites are linked over the likelihood of the null hypothesis that two genetic sites are independent (implying $r = 0.5$). In the linkage mapping community, a LOD score of 3 is commonly accepted as significant linkage between sites, which represents 1000x more certainty of linkage than not [5].

Current state of the art for grouping markers in outbred crosses is the maximum likelihood approach for simultaneous estimation of linkage and linkage phases [6].

Marker ordering

Brute-force calculation of all pairwise interactions and between markers becomes computational prohibitive as the number of markers increases ($\frac{n(n-1)}{2}$ calculations and pairwise combinations for n markers, which then must be ordered exhaustively in $n!/2$ ways). This level of complexity constitutes an NP-hard problem (and is in fact a special case of the "traveling salesman" [7]). Unlike the cities in the travelling salesman problem, however, we have prior belief that markers are ordered along a line (chromosome) so global optima are also likely local optima. Approximations and heuristics must be employed in order to produce maps with hundreds or thousands of markers required for accurate, ultra-dense maps. One common set of methods center around the two-point recombination frequency. These include:

- TRY (TRY) [8]
- Seriation (SER) [9]
- Rapid chain deletion (RCD) [10]
- Recombination counting and ordering (RECORD) [11]
- Unidirectional growth (UG) [12]
- Product of adjacent recombination fractions (PARF) [13]
- Sum of adjacent recombination fractions (SARF) [14]
- Sum of adjacent LOD scores (SALOD) [15]
- Likelihood through Markov chains (LMHC) [16]

Reference [17] provides an excellent review of the advantages and disadvantages of each of these methods. For the purposes of brevity and focus, I will describe the RECORD algorithm. The RECORD algorithm has been natively implemented in `onemap`, a popular open-source library for the `R` programming language that facilitates the creation of linkage maps [18], and is highlighted as a favored method in subsequent attempts to parallelize and improve that algorithm [19]. RECORD is known to have advantages over other methods when working with sparsely-genotyped data, which frustrates map estimation – exactly the type of data set imputation seeks to improve.

These recombination fractions can be mapped to a centimorgan measurement based on the assumptions of the presence or absence of crossover interference by utilizing the Kosambi or Haldane map function, respectively.

Recombination counting and ordering (RECORD) algorithm

Now that our markers have been collected into linkage groups, we must determine their ordering along their chromosome. Here I describe Van Os's RECORD algorithm [11] which will be used later in conjunction with my imputation protocol to produce three genetic maps. First, define x as the number of observed crossovers. A given genotype $A_1/A_1, B/-$ can resolve as either $A_1/A_1, B/B$ or $A_1/A_1, B/b$ in an F_2 or pseudo-testcross F_1. Incomplete genotypes can be used by conditioning on the probability of the unknown genotypes:

$$E(x|A_1/A_1, B/-) = 0 \cdot \frac{(1-r)^2}{1-r^2} + 1 \cdot \frac{2(1-r)}{1-r^2} = \frac{2}{1+r}$$

where x is again the expectation of crossovers given a $A_1/A_1, B/-$ genotype and recombination frequency r. Matrix X_{ij} then represents all x between loci i and j. The criteria for our minimization, COUNT, is then defined as:

$$\text{COUNT} = \sum_{i=1}^{n-1} X_{seq(i),seq(i+1)}$$

where $seq(i)$ is the ith element in the sequence. Authors of the method recommended that markers with no recombinations be binned, a suggestion I followed when computing the final maps in chapter 2. To search for the optimal ordering of the markers, an random pair of markers is chosen in X_{ij}. Each marker is added in a branch-and-bound procedure [20]. Markers are added in this way randomly until an initial ordering is produced.

To optimize in order, the map is then permuted by examining windows of increasing size from 2 to $n-1$ markers across the sequence, testing improvements to the map quality by making inversions in the window subsequence. Inversions are accepted if they lower the overall COUNT value. The procedure is repeated until convergence (i.e., no more improvements are detected) [21].

Brief derivation of the Haldane and Kosambi map functions

A map function M defines the relationship between genetic distance and recombinant events seen in a testcross in the form $r = M(d)$. The units of d are most commonly given as centimorgans (cM). A genetic distance of 1 cM between two sites corresponds to a 1% probability that a crossover event will occur between those two markers.

For any given meiosis, more than one crossover event per chromosome is possible, and at larger separations, even numbers of recombinations ("invisible" crossovers) become possible if not likely. Let's assume crossovers are independent and Poisson distributed with rate d, where $r = R/N$ is again our recombination fraction for R recombinant offspring in N total offspring. Our observed recombinations are then:

$$\begin{aligned} E[r] = r &= de^{-d} + \frac{d^3}{3!}e^{-d} + ... \\ &= e^{-d}\left[1 + \frac{d^3}{3!} + ...\right] \\ &= e^{-d}\left[\frac{e^d - e^{-d}}{2}\right] \\ &= \frac{1}{2}(1 - e^{-2d}). \end{aligned}$$

This is Haldane's mapping formula, with $r \approx d$ for short distances. We now examine mapping functions in the context of crossover interference, describing previously published derivations from Speed, Zhao, and Dudoit [22–24]. For three markers, let us define the coincidence coefficient C between two consecutive intervals, I_1 and I_2, with π_{i_1,i_2} denoting the joint probability of i_1 and i_2 recombination events in the intervals I_1 and I_2.

$$C = C(I_1, I_2) = \frac{\pi_{11}}{(\pi_{10} + \pi_{11})(\pi_{01} + \pi_{11})}.$$

$C = 1, C < 1$, and $C > 1$ correspond to no, positive, and negative crossover interference, respectively. The coincidence coefficient for connected distances is then:

$$C(d, \delta) = \frac{M(d) + M(\delta) - M(d + \delta)}{2M(d)M(\delta)}.$$

We take $\lim \delta \to 0$, and assume both $M(0) = 0$ (two markers 0 M away have 0 chance of recombination) and $M'(0) = 1$ (two very close markers can be approximated by $r \approx M$). We can use these boundary conditions to find:

$$M'(d) = 1 - 2C(d)M(d),$$

where $C(d)$ is now the three-point semi-infinitesimal coincidence function. Since Haldane's mapping function made an assumption of no crossover interference, we set $C(d) = 1$, as crossovers nearby do not affect one another:

$$M'(d) = 1 - 2M(d),$$

which we can solve to recover Haldane's formula [25] from above:

$$M_H(d) = \frac{1}{2}(1 - e^{-2d}).$$

Kosambi took $C = 2M(d)$ as it offered a simple solution to the differential equation, and fit well with existing *D. melanogaster* genetic map data [26]. The recombination data suggested there was a positive crossover interference, as crossovers were farther spaced than a Poisson distribution would suggest. By solving with this three-point semi-infinitesimal coincidence function, we can derive Kosambi's map function:

$$M_K(d) = \frac{1}{2}\tanh(2d).$$

For short distances ($d << 1$), $M(d) \approx d$ for both.

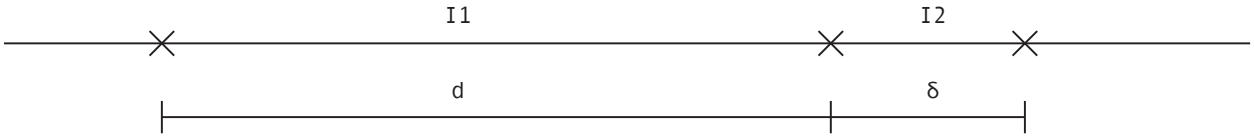


Figure 1.6: Illustrating the three-point coincidence function between two intervals I_1, I_2 with distances d, δ

1.3 Genome resequencing and disease gene discovery

Modern genetic sequencing methods have revolutionized our understanding of inherited diseases. The field's modern renaissance is generally understood to have begun with mapping of Huntington's Disease to repetitive changes in the HTT gene was on the p arm of chromosome 4 [27]. Other notable landmarks include identifying the specific genetic locus that caused cystic fibrosis in the CFTR gene (cystic fibrosis transmembrane conductance regulator, which bears the name of the disease with which it is associated) [28]. Diseases identified were fairly prevalent in the general population. After the human genome project was completed, resequencing specific individuals was a routine task. The cost of resequencing a human genome, using NGS methods, famously outpaced Moore's law starting around 2007.

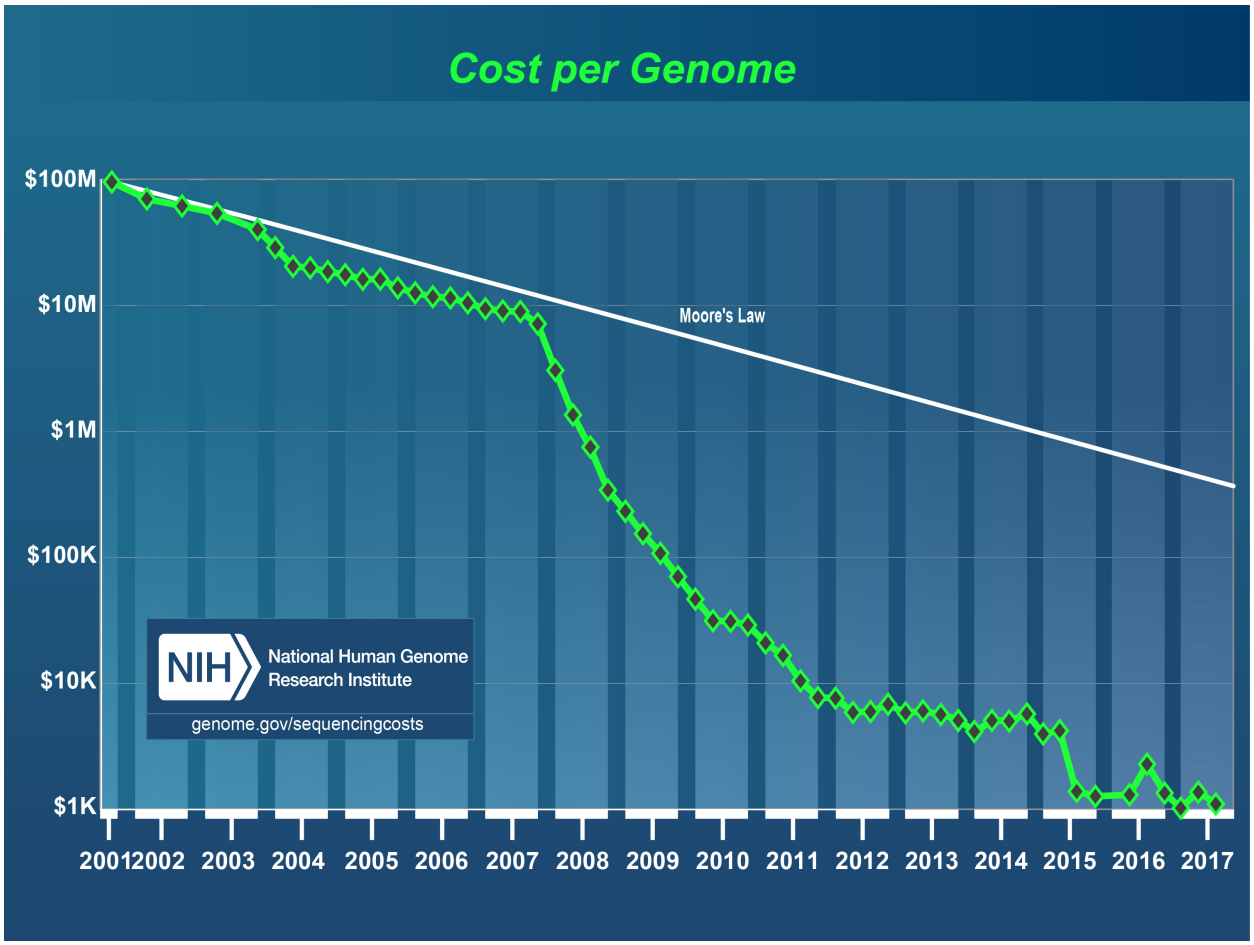


Figure 1.7: Cost of Resequencing a Human Genome. Reprinted from the National Institute of Health National Human Genome Research Institute.

Perhaps not surprisingly, the study genetic diseases recalcitrant to linkage analysis (i.e., those which are studied in the context of large cohorts) benefits greatly from the decrease in sequencing cost. One of the great moments in the history of personalized medicine was the correct identification of the genetic basis of Miller Syndrome in 2009 [29]. Although it had been described in the medical literature as many as thirty years before with a hallmark set of craniofacial and developmental abnormalities, its 1 per million live birth occurrence prevented the same sort of systematic investigation as Huntington's disease or cystic fibrosis. A group of investigators used NGS methods to interrogate the exome of two affected siblings and two unrelated affected individuals against a background of healthy individuals. They were able to identify a causative variant in the DHODH (dihydroorotate dehydrogenase) gene.

Using NGS to solve rare diseases via resequencing is becoming routine, and physicians and researchers are identifying causative variants with increasing rates of success. As sequencing costs drop even further, they are looking beyond exomic sequence and into intergenic regions

of the human genome. Though certainly less understood than protein-coding regions, they are increasingly important areas of investigation, especially as more and more exonic-related diseases are identified. Later in chapter 3 will use inheritance filters to identify a new mutation in a mouse line that was modified with the CRISPR/Cas9 system. The reasoning behind its discovery is highly similar to the reasoning that is employed by rare disease physicians.

1.4 CRISPR/Cas9

In genetic diseases, treatments usually consist of a therapeutic agent, such as the exogenous administration of any enzyme that the body cannot produce, or a small molecule drug that remediates an implicated metabolic pathway. But theoretically, there is also another way to fix a patient's disease state: by altering the affected somatic cells' genomes with a corrective mutation. Though these therapies are still in their infancy, recent advances in genome editing suggest that they may soon represent a powerful tool in a physician's arsenal.

In an ideal world, gene editing systems allow scientists and clinicians to produce targeted mutations at the base-pair level with genome-wide specificity. Recombinant DNA methods have been around for decades, starting with the discovery of restriction endonucleases and their use for transfecting genes into *E. coli*. To modify eukaryotic cells, however, more sophisticated enzymes and vectors would be required. The field has tried several systems to achieve this goal, moving to Transcription Activator-Like Effector Nucleases (TALENs) and Zinc-Finger Nucleases (ZFNs). The most recent and lauded system in the field is the CRISPR/Cas9 system [30], whose function will be the focus of part of chapter 3 of this dissertation.

The CRISPR/Cas9 system achieves its specificity by way of an RNA sequence called a guide RNA (gRNA) [31]. Guide RNAs are short ($\sim$100 bp) sequences of RNA that mimic the CRISPR RNA (crRNAs) and trans-activating crRNAs (tracrRNAs) that are required by Cas9's native role in bacterial immunity. The gRNA also has a $\sim$20 bp sequence that binds with high fidelity to complementary regions of DNA in the targeted genomic region. Initially, the Cas9 complex searches for a three nucleotide sequence that matches its protospacer-adjacent motif (PAM) sequence (an NGG or CCN sequence). If its neighboring 20 base pair guide sequence also matches the proximal sequence, it will create a double stranded break exactly 3 base pairs away from the gRNA's PAM sequence. This breakage will then be repaired by the cell's endogenous repair pathways.

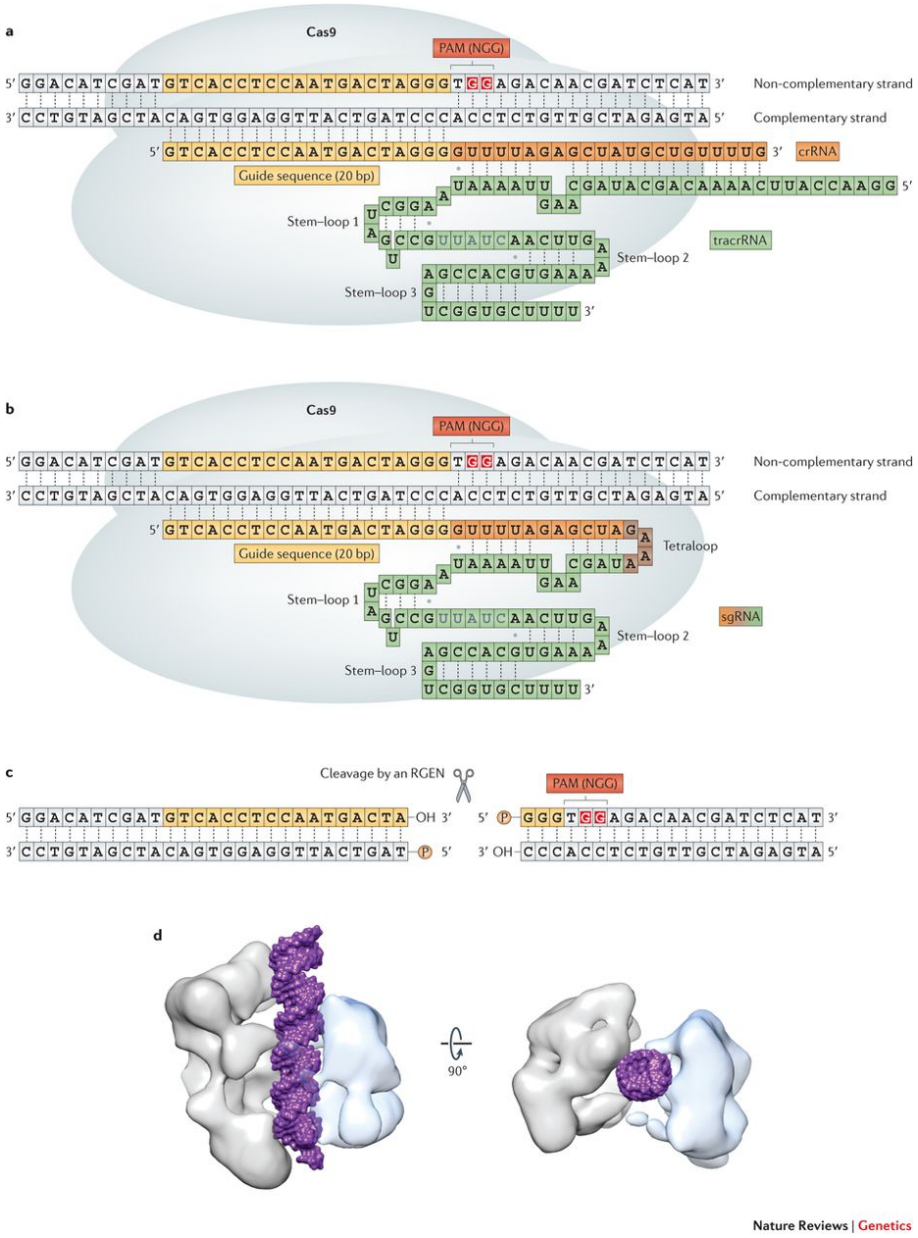


Figure 1.8: Mechanisms of CRISPR/Cas9 cleavage. Reprinted with permission from Springer Nature.

After CRISPR/Cas9 induces a double stranded break (DSB), a cell's endogenous DNA repair pathways determines how its genome will be repaired. The fundamental difference is between how the DSB is repaired. Three main classes of repair are highlighted in the literature: homology directed repair (HDR), non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ).

HDR uses template strands of DNA, either the cells own, or a synthetically derived DNA

template with a complementary sequence to the intended genome edit. Several pathways exist, but all fundamentally occur when the 5' end is trimmed to expose a 3' end. A new strand invades the now open site and DNA repair pathways fill in a complementary sequence as well as correct the recombinant complex [32]. NHEJ occurs by first annealing one side of the double stranded break and then repairing from the newly repaired template. MMEJ is the least understood and most error-prone repair pathway. What is known is that it utilizes short sequences of homology (< 25 bp) and results in deletions near the DSB [33].

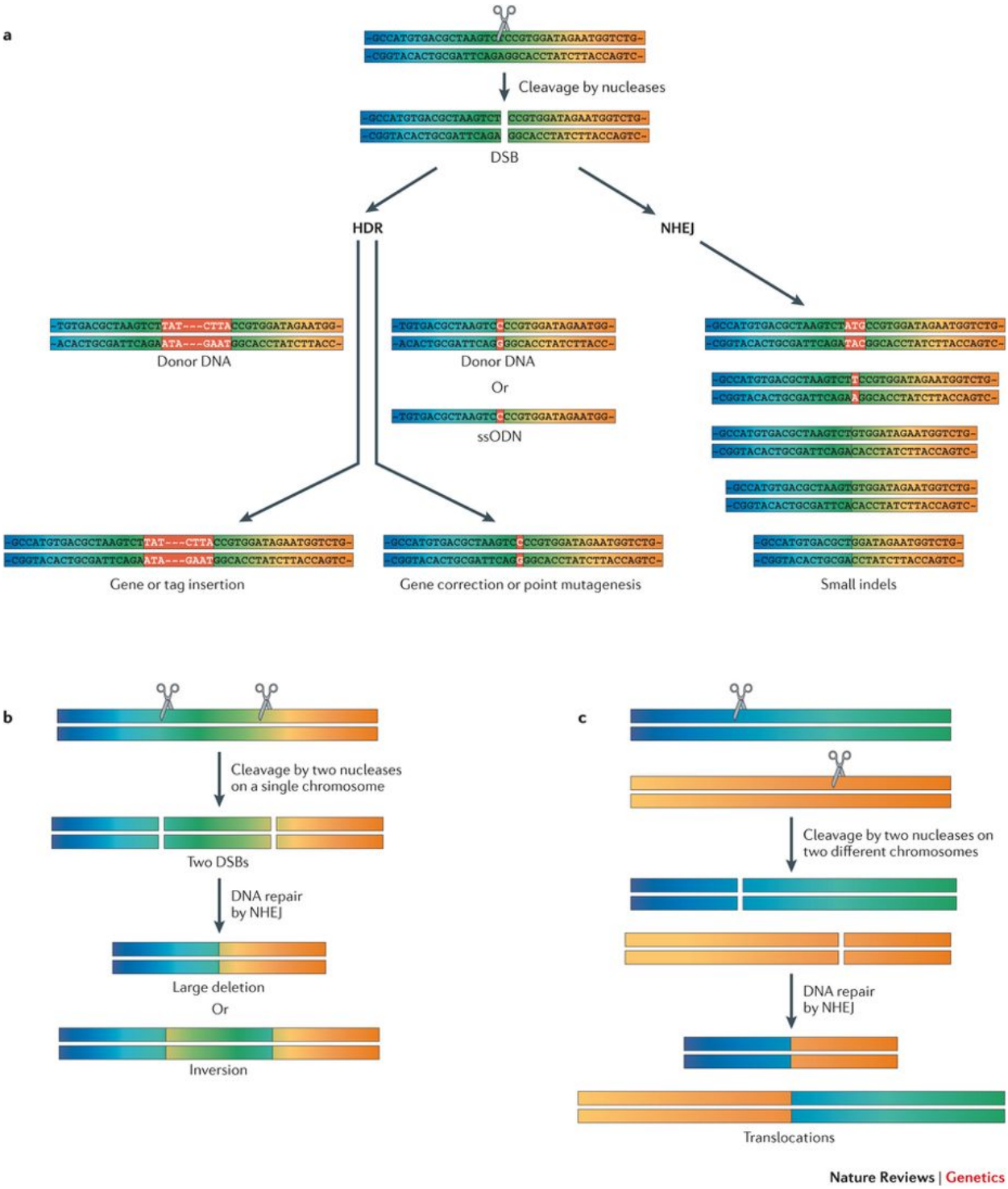


Figure 1.9: Mechanisms of endogenous repair for double stranded breaks. Reprinted with permission from Springer Nature.

1.5 Amphioxus and chordate evolution

With chromosome scale genome sequencing increasingly accessible, evolution at genome scale can be studied. Genomic duplications in the proto-vertebrate lineage were originally suspected by Ohno [34], based on old estimates of genome size. Interest in this hypothesis was revived by the characterization of the Hox cluster of the invertebrate chordate amphioxus [35], along with the discovery of numerous unlinked multi-copy gene families in vertebrates [36, 37]. Further support for genome-wide duplications came from the discovery of paralogous (i.e., duplicated) blocks of linked genes on multiple chromosomes within the human genome [38–42]. Multiply conserved synteny of these blocks compared with an initial assembly of amphioxus, albeit at a sub-chromosomal scale, demonstrated a genome-wide quadruplication in jawed vertebrates [43]. Based on these genome-wide analyses of paralogy and linkage, and studies of numerous gene families, the "2R" scenario, in which early vertebrate genomes experienced two rounds of "whole genome duplication," is now widely accepted [44]. The relative nature of the duplications (i.e., whether by allo- or auto-tetraploidy, or a combination), and their timing, however, remain unclear [36, 45]. Other hypotheses are still discussed, including a single whole genome duplication plus extensive segmental duplication [46], or a series of segmental duplications [47, 48].

1.6 *Hofstenia miamia* and the genomics of whole body regeneration

Since antiquity humans have suffered irreparable injury as a result of disease, trauma, and aging, and that struggle continues today. Though humans lack the ability to regenerate major organs or tissues, many other animals can regenerate extensively. Studying the process of regeneration in these species is essential for understanding the mechanisms that control this process, and may open up avenues for directed regeneration in humans as another tool in fighting disease and aging. In order to further our understanding of these species, we sequenced and analyzed the genome of a highly regenerative new model organism, the three-banded panther worm, *Hofstenia miamia*.

Hofstenia is an acoel flatworm, a group of worms that were previously classified with other flatworms, including the iconic planarian. Molecular analyses now suggest, however, that they are likely the earliest-diverging lineage of animals with bilateral symmetry. *Hofstenia miamia* is easily cultured in lab, and was used by Srivastava et al. to show the importance of Wnt and Bmp-Admp signaling in whole body regeneration, with a transcriptome assembly as an integral part of the study. Since acoel flatworms and planarian flatworms diverged around 500 million years ago, comparisons of regenerative mechanisms between *Hofstenia* and planarians (a well-established regeneration model), are actually comparisons of very anciently diverged systems [49]. Their comparative analysis will allow us to infer features of regeneration common to all animals. *Hofstenia* presents many advantages as a model

regenerative organism - it can be cultured easily, gene function can be studied by *in situ* hybridization and RNAi, and adult stem cells can be isolated.

1.7 Organization of this dissertation

This dissertation is organized into 5 chapters. Chapters 2 - 5 represent the primary focus of the text. Since most of these projects were collaborative, I have attempted to present attributions throughout. My contribution to each project was as follows:

- Chapter 2 - Minimum Spanning Tree Imputation and a Genetic Map of *Xenopus laevis*, *Branchiostoma floridae*, and *Miscanthus sinensis*. I have developed an efficient new algorithm for imputing genotypes in an F1 outbred cross, and used this technique to create genetic maps for three model organisms.

- Chapter 3 - A Large CRISPR-Induced Bystander Mutation Causes Immune Dysregulation. A team led by Dimitre Simeonov and Alexander Marson at University of California, San Francisco, attempting to understand an enhancer locus in the IL2RA gene, inadvertently developed an individual with a severe, unexplained immune phenotype in one of their mouse lines while using CRISPR/Cas9 to knock out the enhancer. Upon several rounds of backcrossing, the phenotype bred true. I investigated and discovered the tightly linked off-target mutation. I also found signatures of microhomology close to the duplication, and proposed a mechanism for the repair pathway that formed the duplication.

- Chapter 4 - Deeply Conserved Synteny Between Amphioxus and Five Vertebrates. I developed a novel method of visualizing and computing the statistical of significance conserved synteny between *Branchiostoma floridae* and vertebrate genomes.

- Chapter 5 - Assembling and annotating the Genome of *Hofstenia miamia*. I assisted with the *de novo* assembly and scaffolding of the genome of *Hofstenia miamia*, an emerging model organism. I also assisted with an annotation of the genome to help understand how this organism is able to perform whole body regeneration.

Chapter 2

Genetic Mapping by Minimum Spanning Tree Imputation and Genetic Maps of *Xenopus laevis*, *Branchiostoma floridae*, and *Miscanthus sinensis*

2.1 Introduction

Linkage mapping seeks to associate the probability of recombination events along the chromosome with genomic position. Linkage mapping influences or informs many endeavors in genetics, including: genome assembly construction and validation, recombination frequency calculation, and quantitative trait locus (QTL) mapping. Given the current ease with which next generation sequencing data is produced, myriad computational techniques have been developed for the construction of linkage maps.

High density maps are facilitated by cohorts with large numbers of offspring, as larger numbers of offspring increase the probability of observing diverse recombination events in parental chromosomes. If variants are known, we can simply use microarrays, but for wild parents, we need to deduce the variant sites and the genotypes together. To sequence numerous individuals at a high depth is costly, and, potentially difficult for genome sequencing projects involving emerging model organisms. One way to circumvent this is to sequence a cohort's parents with high confidence, and the progeny at a lower confidence. Rather than attempt to solve the problem at the level of linkage map construction itself, we report a novel form of imputation and composite marker creation based on low-depth genotypes by taking advantage of an outbred (F1) pedigree used in the construction of model organism linkage maps. By splitting our variant calls into loci which demonstrate heterozygosity in one parent, and homozygosity in the other, we can construct sex-specific linkage maps as well as a

sex-averaged linkage map. Our method associates individual variants (SNPs) by statistical significance within predefined windows, and then phases the haplotypes with these windows by constructing a minimum spanning tree. These composite markers can then be exported for later genetic map construction using any standard software, which we will also demonstrate. This method also has the benefit of a trivially parallel formulation, allowing it to be run efficiently on multiprocessor cluster environments as well as individual machines. This method is compatible with data from reduced representation sequencing or whole genome sequencing.

2.2 Methodology

Determining binary genotype

Genotypes were subset according to a pseudo-testcross F1 structure by collecting loci called heterozygous in the father and homozygous in the mother in one set, and homozygous in the father and heterozygous in the mother in another set. Each site was counted for occurrences of the major allele, c_{major}, and occurrences of the minor (heterozygous variant) allele, c_{minor}. Note that major and minor alleles are defined in the context of the familial pedigree, not based on larger population allele frequencies. Each pair of counts is then used to produce a binary genotype, $g_{i,\alpha}$, for the αth individual at the ith site.

$$g_{i,\alpha}(c_{\text{minor}}, c_{\text{major}}) = \begin{cases} 1 & \text{if } c_{\text{minor}} > C \\ -1 & \text{else if } c_{\text{major}} > D \\ 0 & \text{else} \end{cases}$$

Where C, and D are allele depth cutoffs. $g_{i,\alpha}$ can indicate the definitive presence of the minor allele (g = 1), likely absence of the minor allele (g = -1), or an indeterminate genotype (g = 0). For the purposes of the *Xenopus laevis*, *Branchiostoma floridae*, and *Miscanthus sinensis* map, I set $C = 1$ (to prevent mismapped reads or base call errors from erroneously reporting the presence of the minor allele). The major allele cutoff was set to $D = 3$, which gives a lower bound confidence of 93.75% for $g_{i,\alpha} = -1$ sites, as $0.5^{D+1} = 0.9375$).

Filtering

Confident genotypes with good representation of both binary genotype states are critical to the imputation and phasing steps. I filter all sites that reject the hypothesis of independence, since, for large numbers of progeny, we expect 50% of the progeny from one haplotype (showing evidence of the minor allele), and 50% of the progeny from the other haplotype (showing no evidence of the minor allele). I define the indicator:

$$I[g_{i,\alpha} = b] = \begin{cases} 1 & \text{if } g_{i,\alpha} = b \\ 0 & \text{else} \end{cases}$$

and exclude sites that violate the independence hypothesis for the chi-squared test with zero degrees of freedom ($p < 0.05$), given $\sum_\alpha I[g_{i,\alpha} = 1]$ individuals with the minor allele and $\sum_\alpha I[g_{i,\alpha} = -1]$ individuals presumed homozygous for the major allele (pre-imputation) at a given site.

In *Xenopus laevis*, our filtering methods reduced the number of viable biallelic SNP markers from 558,928 sites to 448,200 sites segregating from the father and from 778,811 to 627,071 sites in the mother.

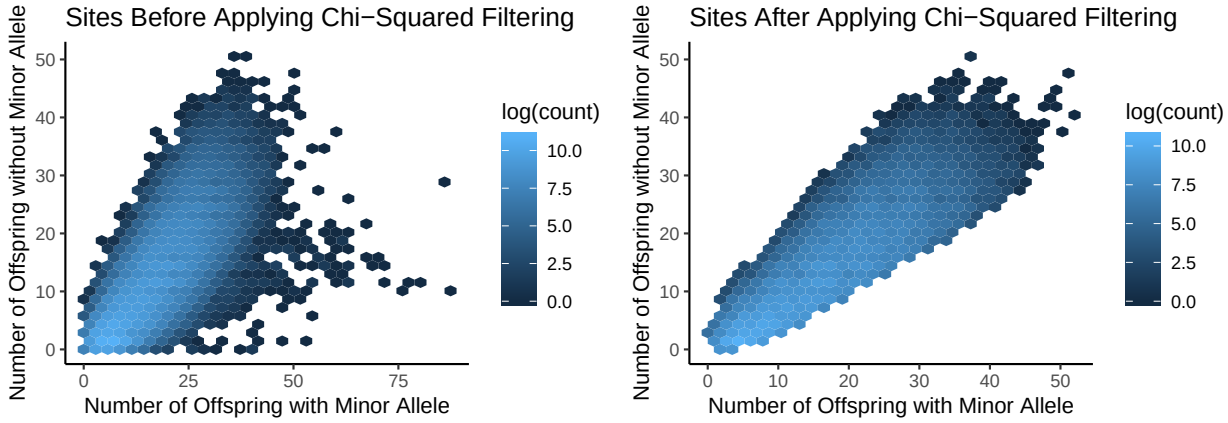


Figure 2.1: Sites before and after chi-squared filtering

Haplotype phasing by composite marker construction

In order to perform a local imputation and create composite markers representing segments of the genome, we first segment the genome into non-overlapping windows. For any contig with viable loci of size L, and a window of size w_s, I partition each contig into $\left\lfloor \frac{L}{w_s} \right\rfloor$ non-overlapping windows of size w_s, and an additional window of size $L \bmod w_s$ if $L \bmod w_s \neq 0$. The process is repeated for every contig. I chose $w_s = 1$ Mb for our imputation/haplotype reconstruction procedure in *Xenopus laevis*.

To determine whether markers are in coupling or repulsion, I wish to associate neighboring sites within window k by the significance of their coherence (Figure 2.2). Sites that are coherently phased with respect to one another would possess a large numbers of sites exclusively in coupling or in repulsion to one another. To quantify this, I first define two indicator matrices of window k with N sites and A individuals:

$$M_{>0,k} = \begin{bmatrix} I[g_{i=1,\alpha=1,k} = 1] & \dots & I[g_{i=1,\alpha=A,k} = 1] \\ \vdots & \ddots & \vdots \\ I[g_{i=N,\alpha=1,k} = 1] & \dots & I[g_{i=N,\alpha=A,k} = 1] \end{bmatrix}$$

and

$$M_{<0,k} = \begin{bmatrix} I[g_{i=1,\alpha=1,k} = -1] & \dots & I[g_{i=1,\alpha=A,k} = -1] \\ \vdots & \ddots & \vdots \\ I[g_{i=N,\alpha=1,k} = -1] & \dots & I[g_{i=N,\alpha=A,k} = -1] \end{bmatrix}.$$

I then use simple matrix algebra to find the indices of the four possibilities of site-to-site genotype transitions for an N x N matrix $S_{i,j,k}$ which counts the four possible informative genotype transitions between sites i and j in block k:

$$\begin{aligned} S_{1\to1,k} &= M_{>0,k} M_{>0,k}^T \\ S_{-1\to-1,k} &= M_{<0,k} M_{<0,k}^T \\ S_{1\to-1,k} &= M_{>0,k} M_{<0,k}^T \\ S_{-1\to1,k} &= M_{<0,k} M_{>0,k}^T, \end{aligned}$$

which I can group into counts of coupling or repulsion between sites:

$$\begin{aligned} S_{\text{coupling},k} &= M_{1\to1,k} + M_{-1\to-1,k} \\ S_{\text{repulsion},k} &= M_{1\to-1,k} + M_{-1\to1,k}. \end{aligned}$$

Using each marker/marker (i.e., row/column) entry in $S_{\text{coupling},k}$ and $S_{\text{repulsion},k}$, I can construct a graph which discerns the local structure of the parental haplotypes. Let G_k be an undirected weighted graph, where i, j, represent markers, and their edge weight corresponds to their p-value based on the two-sided binomial test for $\max((S_{\text{coupling}})_{i,j,k}, (S_{\text{repulsion}})_{i,j,k})$ successes and $\min((S_{\text{coupling}})_{i,j,k}, (S_{\text{repulsion}})_{i,j,k})$ failures (Figure 2.2). This graph will only allow edges if this weight is below a given p-value threshold. For our *Xenopus laevis* maps, it is $p <$ 1e-3.

Haplotype phasing refers to the assignment of markers to specific chromosomal structures. In our population, we would like to know the major and minor allele states of the heterozygous parent's two haplotypes. We can phase the haplotype of block k by computing the minimum spanning tree (MST) through G_k, which we will define as $G_{\text{MST},k}$. There may be more than one minimum spanning tree for the graph given that the minimum p-value required to connect i, j can result in removal of edges. In this case, the spanning tree with the largest number of nodes will be selected to represent the haplotype that can be phased for the block (Figure 2.3). All sites not in the block will be discarded, and are not used for either haplotype determination or imputing. A minimum spanning tree structure has several benefits, including that it prevents the formation of loops or ambiguous connections that would frustrate the haplotype estimation.

To resolve the phased haplotype sites, we arbitrarily choose a node in $G_{\text{MST},k}$ and label it with a 1. By traversing along the tree, assigning relative phased based on the chosen node, neighboring a genomic loci will be assigned 1 or -1, according to the excess of coupling (the same phase) or repulsion (the opposite phase) counts. The sequence of genotypes along the

resolved $G_{\mathrm{MST},k}$ represents one heterozygous parental haplotype $H_{A,k}$. By construction, as all sites are heterozygous, the two haplotypes are compliments of one another, so the other, $H_{B,k}$, is given as:

$$(H_A)_{i,k} = -(H_B)_{i,k}$$

for each site i in all k blocks of the genome.

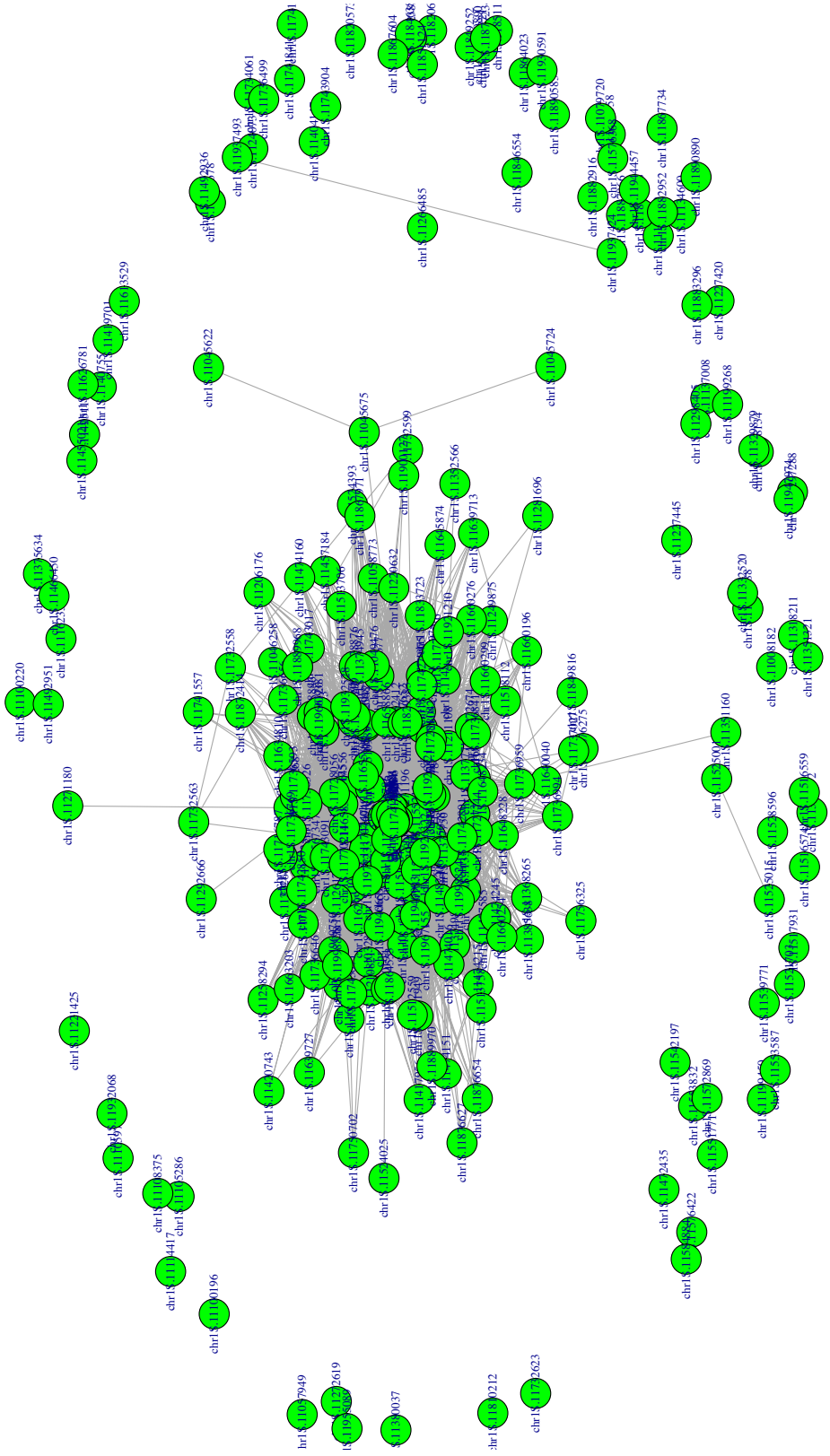


Figure 2.2: Kamada-Kawai visualization [50] of all potential markers in the haplotype in a 1 Mb (chr1S:11 Mb - 12 Mb) block after setting a minimum connection thresholding. Note the main set of nodes with extensive connections which form the focus of our minimum spanning tree search. Unconnected components form the periphery of the graph and will not be used for haplotype phasing.

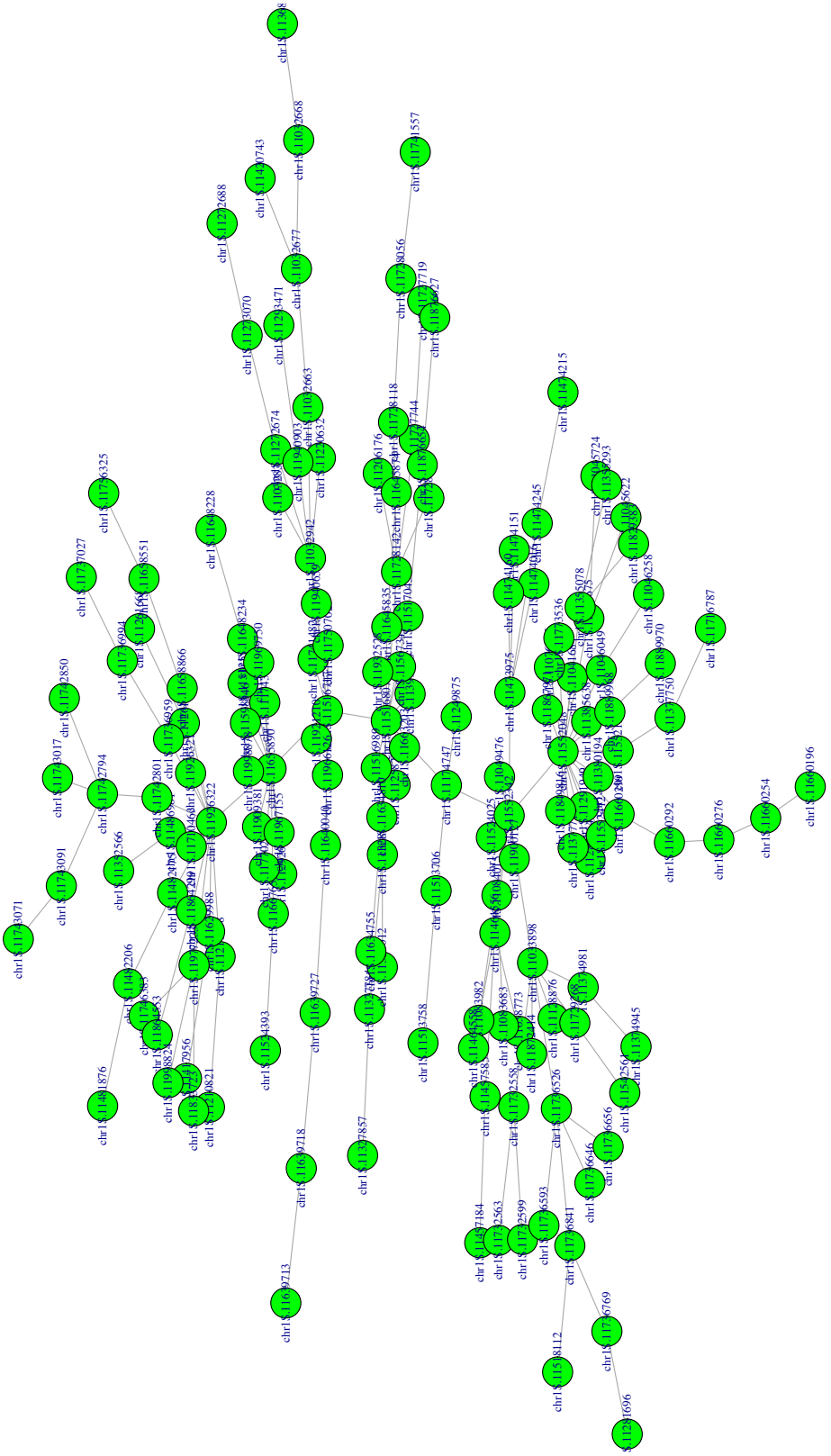


Figure 2.3: Kamada-Kawai visualization [50] of final markers used in assigning haplotype in a 1 Mb (chr1S:11 Mb - 12 Mb) block after a minimum spanning tree is found weighted by p-value. Note that nearest neighbor nodes in genomic might not be the nodes that have the strongest association.

Assigning haplotypes states to offspring and imputation

Once I have constructed the two parental haplotypes, I can infer membership in either haplotype A or B for each progeny, and impute missing or incorrect genotypes from that assignment. Given that my protocol is designed for low-depth sequencing, it cannot be assumed that the initial binary genotype calls are accurate. Furthermore, if a crossover has occurred within the block for a given individual, neither parental haplotype is appropriate (Figures 2.4, 2.5 and 2.6). By summing an indicator function of agreement with either haplotype for the N sites in the 1 Mb region, a confidence can be assigned to their inheritance of the haplotype called within a window of the genome.

To perform imputation, I first attempt assign to every offspring α in a cohort with a phased haplotype that is N_H sites long (the size of the MST). If the individual is assigned to a haplotype, every site in window k in that individual's variant call table is changed to match that haplotype. This has the benefit of not only imputing missing variants, but correcting variants that may have been initially miscalled due to low depth. For the sake of these genetic maps, I filtered sites where less than 60% for *X. laevis* or 80% for *B. floridae* and *M. sinensis* of individuals could be assigned to haplotype. The bounds of this parameter is suggested, but ultimately left up to the user. I define $(c_A)_{\alpha,k}$ or $(c_B)_{\alpha,k}$ as the sum of coincidences for an individual α between either haplotype along its i sites in window k:

$$(c_A)_{\alpha,k} = \sum_i I[g_{i,\alpha,k} = (H_A)_{i,k}]$$
$$(c_B)_{\alpha,k} = \sum_i I[g_{i,\alpha,k} = (H_B)_{i,k}]$$

A two-tailed binomial test statistic is computed with $\max((c_A)_{\alpha,k}, (c_B)_{\alpha,k})$ successes and $\min((c_A)_{\alpha,k}, (c_B)_{\alpha,k})$ failures, and an individual is only assigned if the resulting p-value is below a given threshold (for *X. laevis*, the threshold is $p < 1e-3$). We can visualize the quality of these assignments to the A or B haplotype within block k graphically by plotting $((c_A)_{\alpha,k}, (c_B)_{\alpha,k})$ for all α in block (Figure 2.4). In situation with perfect phasing, all individuals would appear completely along the x or y axes.

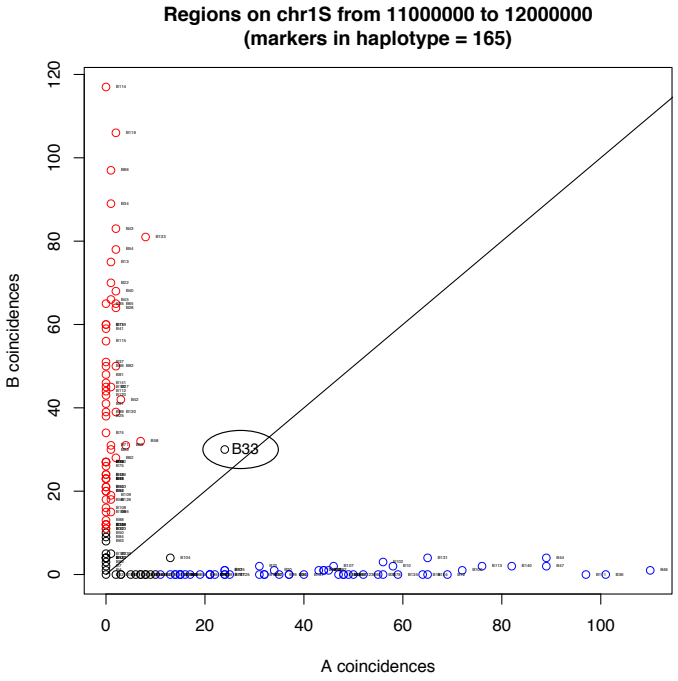


Figure 2.4: Assigning the A or B haplotype to individuals in a 1 Mb block for N3M meiotic map on chr1L from 11 Mb - 12 Mb. x = y is shown as a black line. Individuals in red are assigned to the B haplotype. Individuals in blue are assigned to the A haplotype. Individuals in black are unassigned. One individual shows a large number of markers from both haplotypes. I will examine this block in the larger chr1S map to show that its ambiguity is the result of a crossover event somewhere the block.

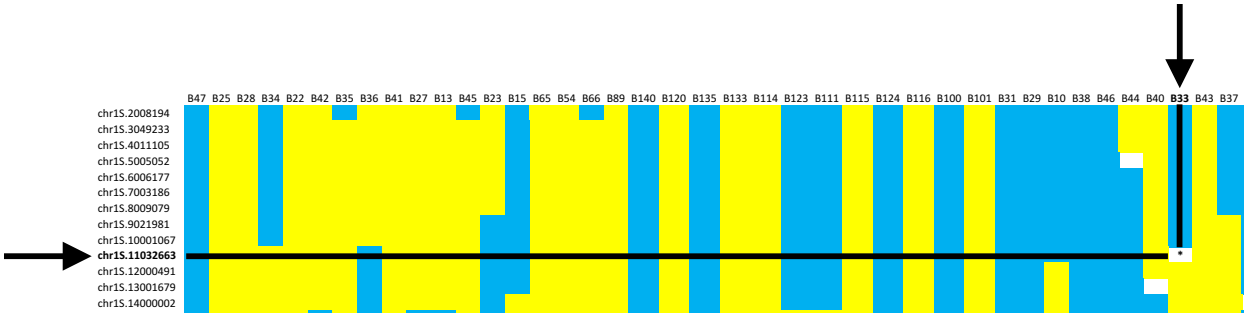


Figure 2.5: 1 Mb blocks grouped into larger haplotype structure and used to visualize crossovers. Here is subset of the entire chromosome (chr1S), which confirms that individual B33, who had a large number of markers from both haplotypes, crossed over in the block between 11Mb - 12Mb. Further work might allow for greater localization of the crossover position.

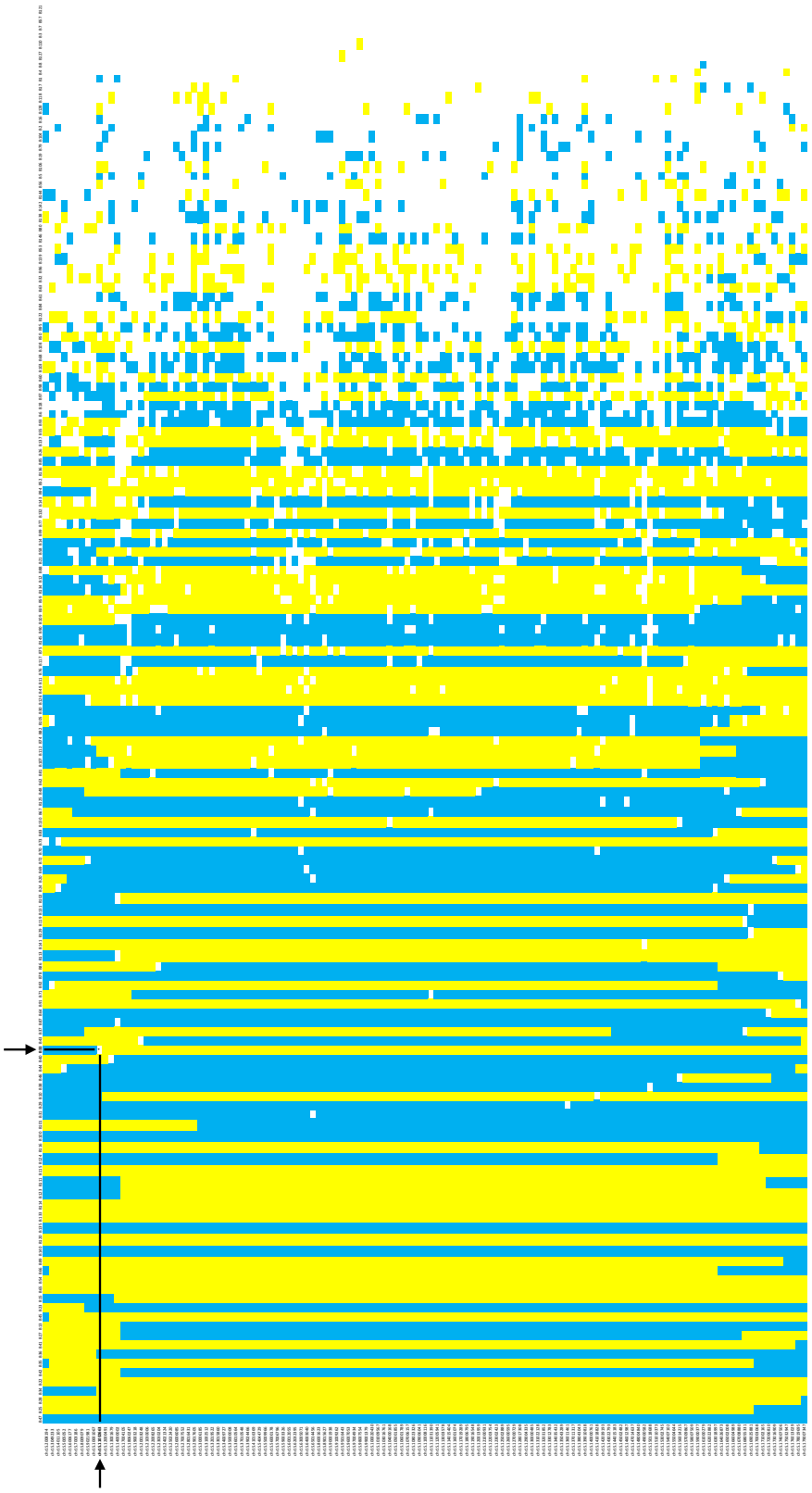


Figure 2.6: After reconstructing the haplotype structures for N3M, children can then be assigned one of the two stats, and colored to easily visualize recombination events.

2.3 Results

Paternal imputation

For the male *X. laevis* map, 448,200 sites passed chi-squared filtering and were used as the basis of the imputation. After running MISTI, 375,362 sites (83.7%) could be phased coherently within 1 Mb blocks. Of the remaining loci, 692,017 genotypes (1.3%) were discarded/removed. 30,052,207 genotypes (55.6%) were added/created. 192,369 (0.3%) were changed from one genotype to another during imputation, implying an error in initial call $g_{i,\alpha}$. 23,115,535 genotypes (42.8%) remained the same. Of the viable 375,362 sites, genotype density increased from 20.9% of sites with initial binary genotypes to 75.2% sites with imputed binary genotypes.

N3M		After		
		1	0	-1
Before	1	5,894,370 (10.90%)	456,678 (0.84%)	3,220 (0.01%)
	0	14,385,709 (26.61%)	12,690,686 (23.48%)	15,666,498 (28.98%)
	-1	189,149 (0.35%)	235,339 (0.44%)	4,530,479 (8.38%)

Table 2.1: Binary genotype changes as a result of imputation (male *X. laevis*)

Maternal imputation

For the female *X. laevis* map, 627,071 sites passed chi-squared filtering and were used as the basis of the imputation. After running MISTI, 267,384 sites (42.6%) could be phased coherently within 1 Mb blocks. Of the remaining loci, 521,947 genotypes (1.3%) were discarded/removed. 21,388,087 genotypes (55.5%) were added/created. 136,950 (0.3%) were changed from one genotype to another during imputation, implying an error in initial call $g_{i,\alpha}$. 16,456,312 genotypes (42.7%) remained the same. Of the viable 267,384 sites, genotype density increased from 20.9% of sites with initial binary genotypes to 75.1% sites with imputed binary genotypes.

WC3F		After		
		1	0	-1
Before	1	4,161,541 (10.81%)	351,747 (0.91%)	2,843 (0.01%)
	0	10,336,310 (26.84%)	9,077,203 (23.57%)	11,051,777 (28.70%)
	-1	134,107 (0.35%)	170,200 (0.44%)	3,217,568 (8.36%)

Table 2.2: Binary genotype changes as a result of imputation (female *X. laevis*)

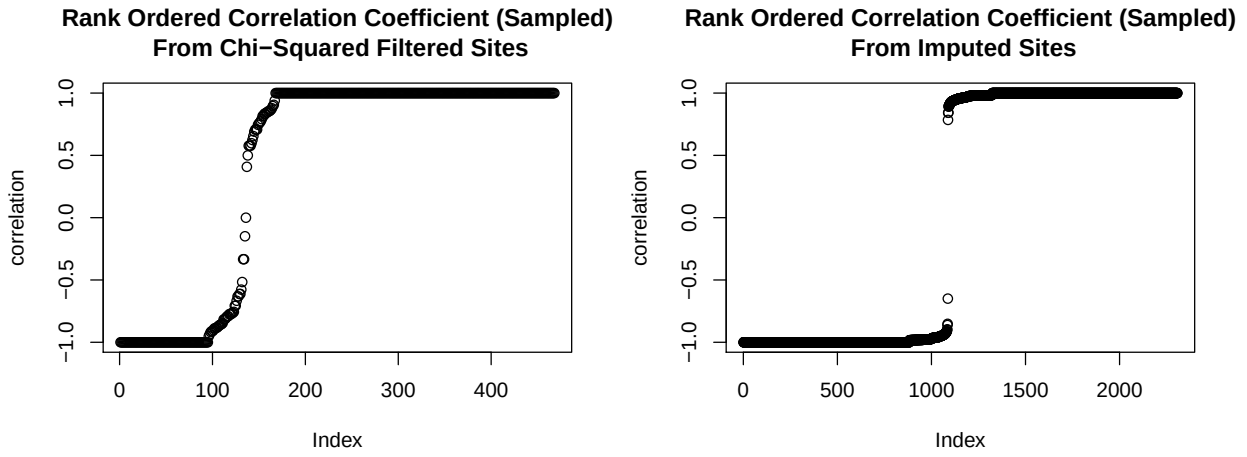


Figure 2.7: Pearson correlation coefficient of neighboring genotypes before, and composite markers after, imputation. The balancing of both coupling ($\rho \approx 1$) and repulsion ($\rho \approx -1$) correlation coefficients, as well as a sharper transition between the two, shows a more coherently genotyped cohort.

2.4 Discussion

The problem of missing and erroneous genotypes in linkage mapping is well posed. Through construction of composite markers and imputation, we are able to recover sites that would otherwise be excluded from the analysis. Minimum spanning trees have been used in map construction, albeit at the whole chromosome level [51]. Our attempt to use MSTmap [51] to construct a map for the paternal chr1S with chi-squared filtered binary genotype markers were unsuccessful (yielding maps with over 1000 cM and grossly incorrect ordering). This should not be taken as a shortcoming of their method since it was never designed for outbred crosses or data sets with such sparse data. Other protocols that are designed for low-depth coverage with missing data, such as Lep-MAP3 [52] have attempted to deal with the difficulties in the data by use of Viterbi-like algorithms, searching for optimal paths through the markers. In our attempts to use Lep-MAP3 in *Branchiostoma floridae*, we found the protocol to be prohibitively slow. The user guide suggests constructing linkage maps independently with markers subset by chromosome in order to speed up computation. However, because one of our main goals with the *Branchiostoma floridae* map was to validate the assembly, we did not pursue this option, as mapping chromosomes independently would prevent identification of translocation assembly errors.

2.5 Constructing the maps

For each organism, slightly different settings for the MISTI imputation and `onemap` package were used. These are detailed below. Other mapping programs could be used with the imputed genotype block markers from MISTI. Chromosomal position/genetic distance figures were generated using genetic-mapper vectorial genetic map drawer [53].

2.6 A genetic map of *Xenopus laevis*

Sequencing and variant calling

Our MISTI methodology was applied to the *X. laevis*, also known as the African clawed frog. *X. laevis* is a model organism, and the resulting composite markers linked using the `R` package `onemap`.

A Nasco *X. laevis* male and a wild (feral) caught *X. laevis* female imported from Chile were crossed at Woods Hole in summer 2014. The tadpoles were killed and stored in lysis buffer (1% SDS, 20 mM EDTA, 100 mM NaCl, 20 mM Tris pH 7.5–8.0) at 4°C. Proteinase K was added to a final concentration of 200 μg/mL, and the tadpoles were lysed overnight at 55°C. The DNA was purified using Phenol:Chloroform:IAA, followed by NH_4OAc and isopropanol precipitation of the DNA. The DNA was washed with 70% EtOH and resuspended in 10 mM Tris pH 8.5. The tadpole DNA was then digested with NlaIII, and barcoded adapters were ligated with T4 DNA Ligase. The resulting libraries were pooled and then cleaned and concentrated using the QIAgen MinElute PCR Purification Kit. The pooled libraries were size selected on a 1% agarose gel, and DNA in the cut gel was extracted using the QIAgen MinElute Gel Extraction Kit. The extracted DNA was then enriched for ligated sequence with 5 cycles of PCR. The DNA extraction and GBS library preparation were completed by Jessica Lyons and Austin Mudd.

For the two parents, GBS libraries were prepared using the aforementioned GBS library preparation protocol, starting with NlaIII digestion and ending with PCR enrichment. Additional libraries for each parent were prepared by the HudsonAlpha Institute for Biotechnology using Illumina's TruSeq DNA PCR-Free Library Prep Protocol. The prepared GBS libraries were sequenced on Illumina HiSeq 2500 and Illumina HiSeq 4000 by the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley. The prepared TruSeq libraries were sequenced on the HiSeq X10 by the HudsonAlpha Institute for Biotechnology.

The sequenced reads were aligned to the v9.1 assembly of *X. laevis* using the GBS preprocessing PE pipeline for the GBS libraries (https://bitbucket.org/rokhsar-lab/gbs-analysis) and the WGS preprocessing pipeline for the TruSeq libraries (https://bitbucket.org/rokhsar-lab/wgs-analysis) [54]. Using bamtools and SAMtools, respectively, the resulting BAM files were merged, and scaffolds smaller than 10 kb were excluded [55, 56]. SNPs were called using freebayes and then filtered with a custom script that selects biallelic pseudo-testcross SNPs

based on Hardy-Weinberg statistical expectations [57]. The custom script uses a Fisher's exact test with a p value threshold of 0.1 and writes out an allele depth file that is used as input for MISTI.

The wild caught mother was originally imported from Chile, which has only had a wild population of *Xenopus laevis* since the 1980s. Corroborating existing literature about the low genetic variation and bottlenecks in the Chilean wild population [58], we have identified large regions of the mother's genome lacking heterozygous SNPs, while still maintaining expected read depth.

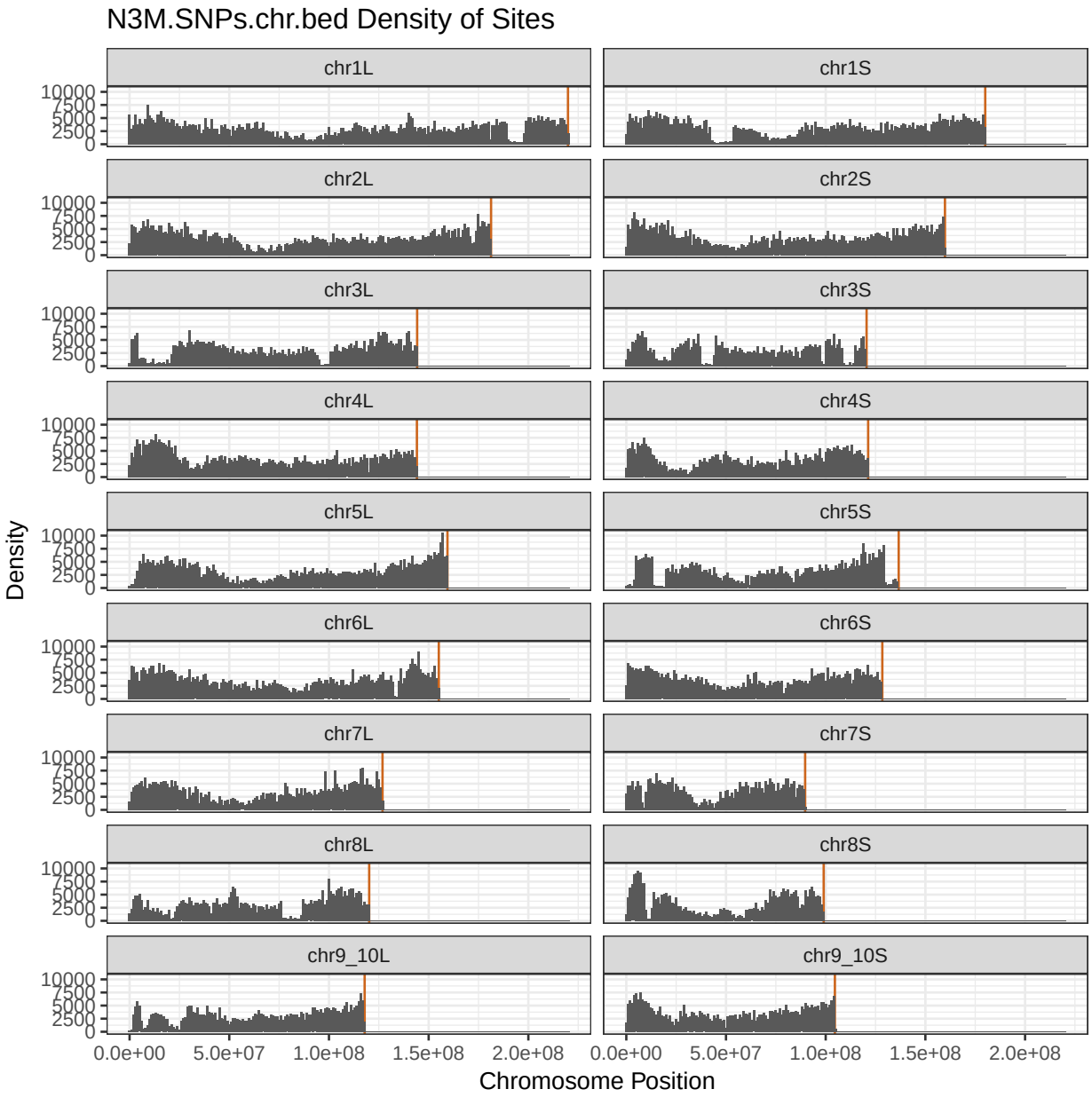


Figure 2.8: Male (N3M) *X. laevis* heterozygous variant density.

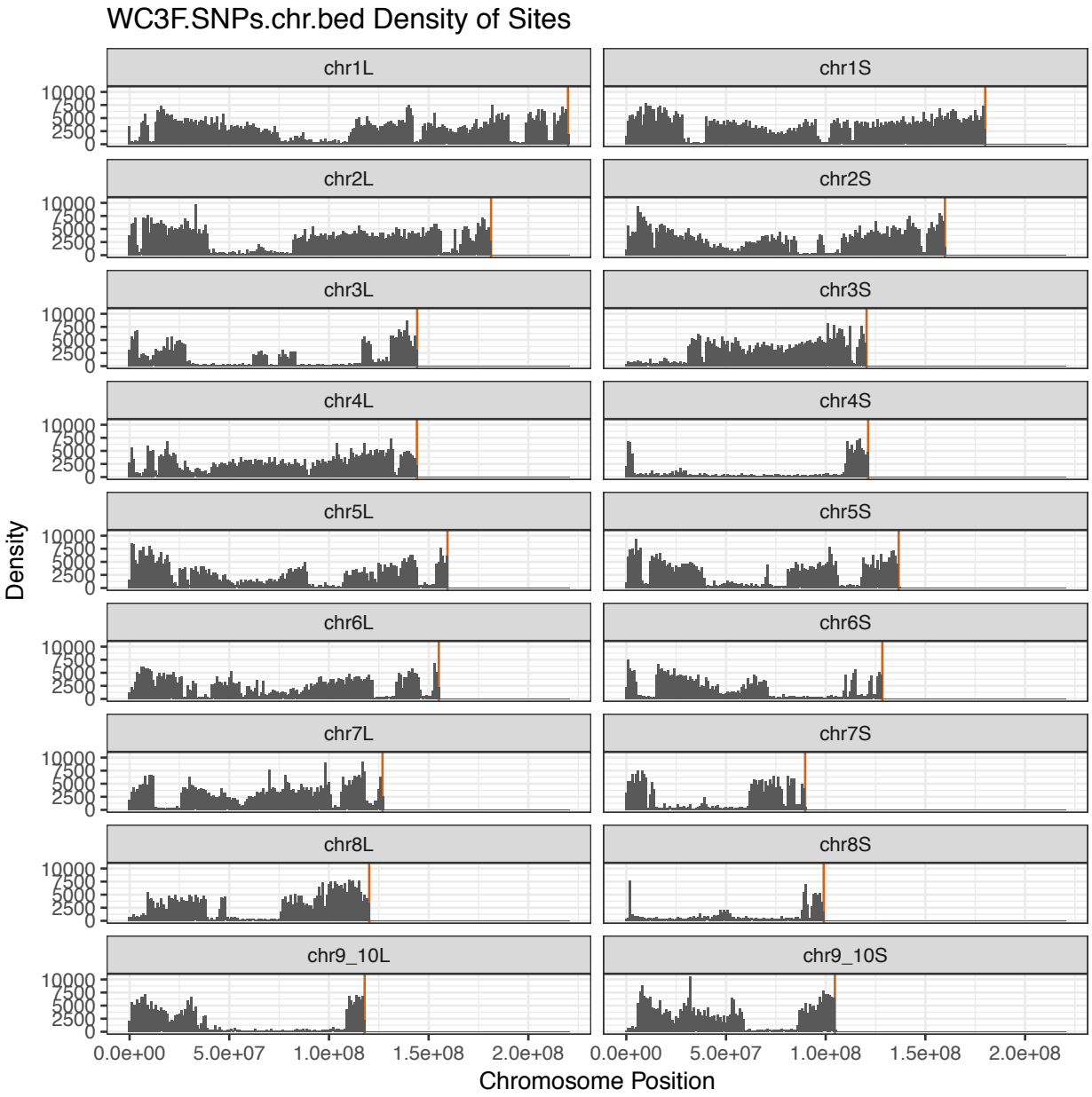


Figure 2.9: Female (WC3F) *X. laevis* heterozygous variant density, large gaps illustrate the difficulties in constructing the maternal map.

Imputation and map construction

We constructed a minimum spanning tree of edges among nodes in the window, minimizing the sum of the negative absolute value of two-sided binomial test p-value along the spanning tree $p <$ 1e-3. The sign of the correlations along the tree allows us to determine

the relative phase of each SNP inherited from the heterozygous parent. This analysis determines the two haplotypes of the heterozygous parent within each 1 Mb window. Progeny are assigned one of the two haplotypes if the haplotypes could be assigned with $p <$ 1e-3 confidence.

Markers for paternal meiotic map: Out of 2,590 non-overlapping full 1 Mb genomic windows, 2,429 (93.8%) contained sufficiently many correlated paternal SNPs. Of these, 2,029 could be called in more than 60% of progeny, and were used for paternal map construction. In addition, 328 windows were shorter than 1 Mb that had at least 1 SNP, (representing ends of contigs/chromosomes, and sub-1 Mb scaffolds). Of these windows, 105 (32.0%) had sufficiently many correlated paternal SNPs. Of these, 22 could be called in more than 60% of progeny.

Markers for maternal meiotic map: Out of 2,582 non-overlapping full 1 Mb genomic windows that had at least 1 SNP, 1,661 (64.3%) contained sufficiently many correlated maternal SNPs. Of these, 1424 could be called in more than 60% of progeny, and were used for maternal map construction. Of the 253 windows shorter than 1 Mb that had at least 1 SNP, 82 (32.4%) contained sufficiently many correlated maternal SNPs. Of these, 22 could be called in more than 60% of progeny.

I constructed separate male and female linkage maps with the `onemap` package (v2.0-3) in R [18], constructing each map using the "F1 cross" setting and the recommended LOD score offered by `suggest_lod` providing the genotype calls for non-overlapping 1 Mb windows as described above. Markers were further grouped into bins with 0 cM recombination according to best practices from Schiffthaler et al. [59]. We found 19 major linkage groups in the father and 30 major linkage groups in the mother (log odds (LOD) threshold 5.886182 for male map and 5.992417 for female map).

The largest 18 male and female linkage groups are almost in 1:1 correspondence with the 18 chromosome-scale scaffolds assembled using HiC chromatin linkages, confirming the accuracy of these chromosomes. The total length of the 18 linkage groups with the largest number of markers from each of the 18 *X. laevis* chromosomes for the male and female maps were 1,639.2 and 1,446.9 cM, respectively. The final male and female linkage groups that represent each chromosome comprised 2,051 and 1,446 composite markers for the male and female maps, respectively. Although we genetically confirmed the chromosomal linkages of our HiC-based assembly, the ordering of markers was not perfectly concordant with their order along the assembly. Several small misassemblies can be visualized along the chromosomes for both maps, largely at the chromosome ends. An example of this can be seen in as an inversion in p-arm of the chr1L. Because of issues creating the maternal map due to an absence of large regions of heterozygosity in the maternal map, some chromosomal diagram plots will be omitted.

Discussion

Complete chromosomes from the female map and the entire male map show long stretches of minimal recombination in the centers of the chromosomes. Furthermore, chr1L, chr2L,

and chr7L show signs of possible inversion errors in assembly near the chromosome ends. In chr8S, there is a large discontinuity that could be the result of errors in mapping, imputation, or assembly.

I will next apply my methodology in two other model organisms. Any difference of parameters will be noted below. The first is *Branchiostoma floridae* (amphioxus), whose genetic map I will use to validate a new, chromosome-sized assembly that will later used in chapter 4. The second is *Miscanthus sinensis*, a perennial grass whose study has implications for bioenergy production.

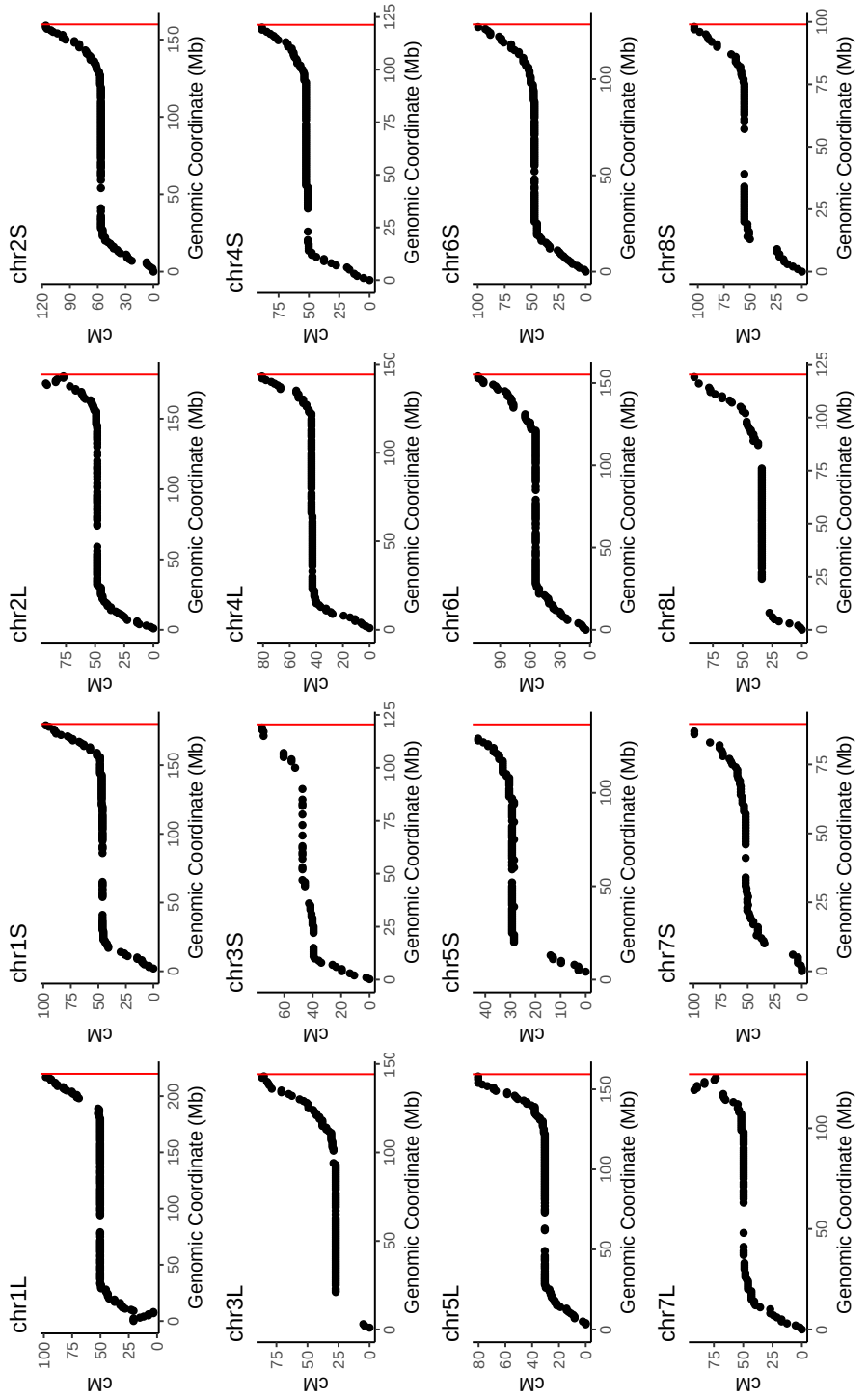


Figure 2.10: A genetic map of male (N3M) *Xenopus laevis*

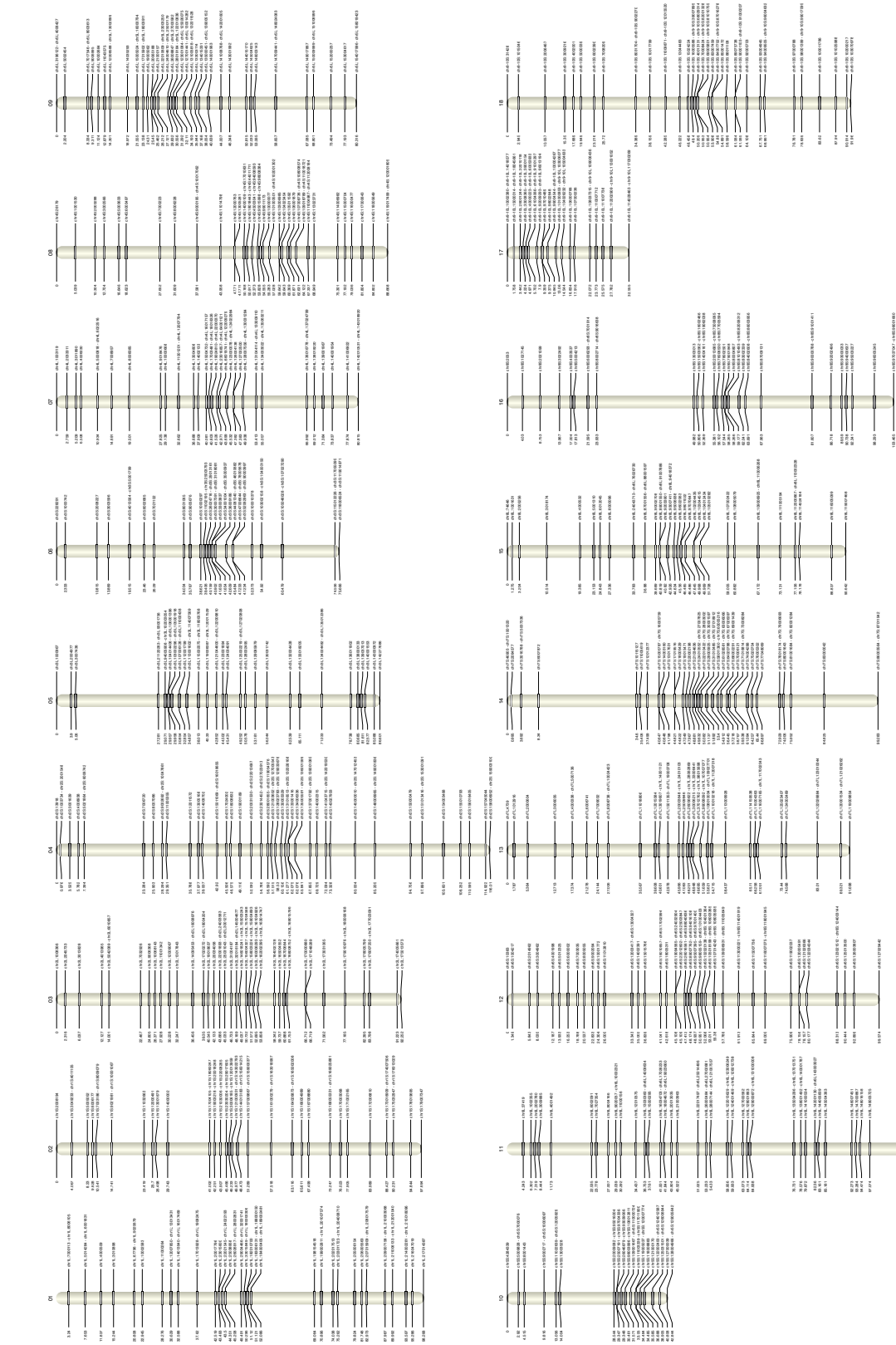


Figure 2.11: A genetic map of female (WC3F) *Xenopus laevis (Part 2).*

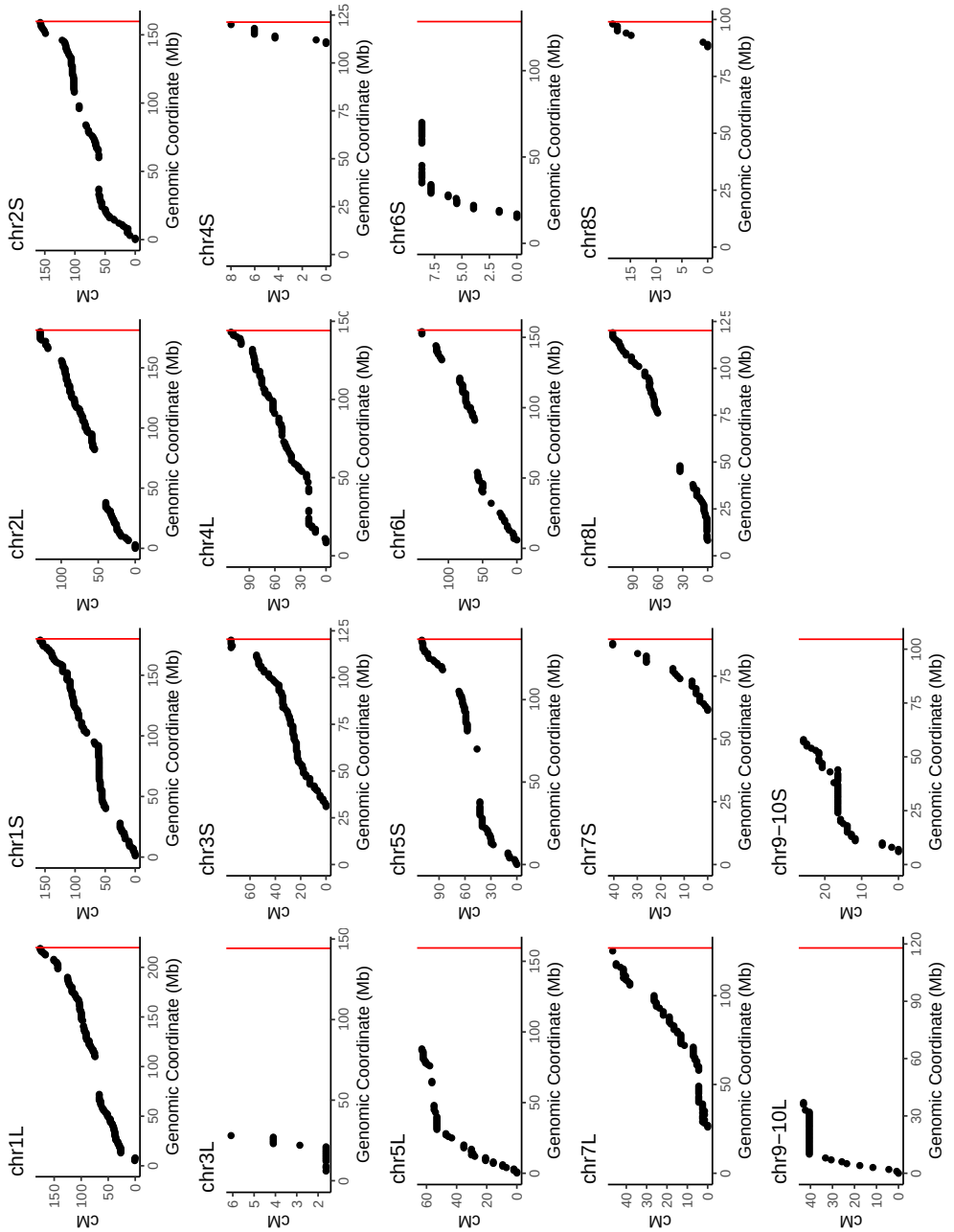


Figure 2.12: A partial genetic map of female (WC3F) *Xenopus laevis* (chromosome and linkage group). Given the paucity of heterozygous SNP data for WC3F, a chromosomal figure should be considered incomplete for chr3L, chr4S, chr5L, chr6S, chr7S, chr8S, chr9-10L, and chr9-10S

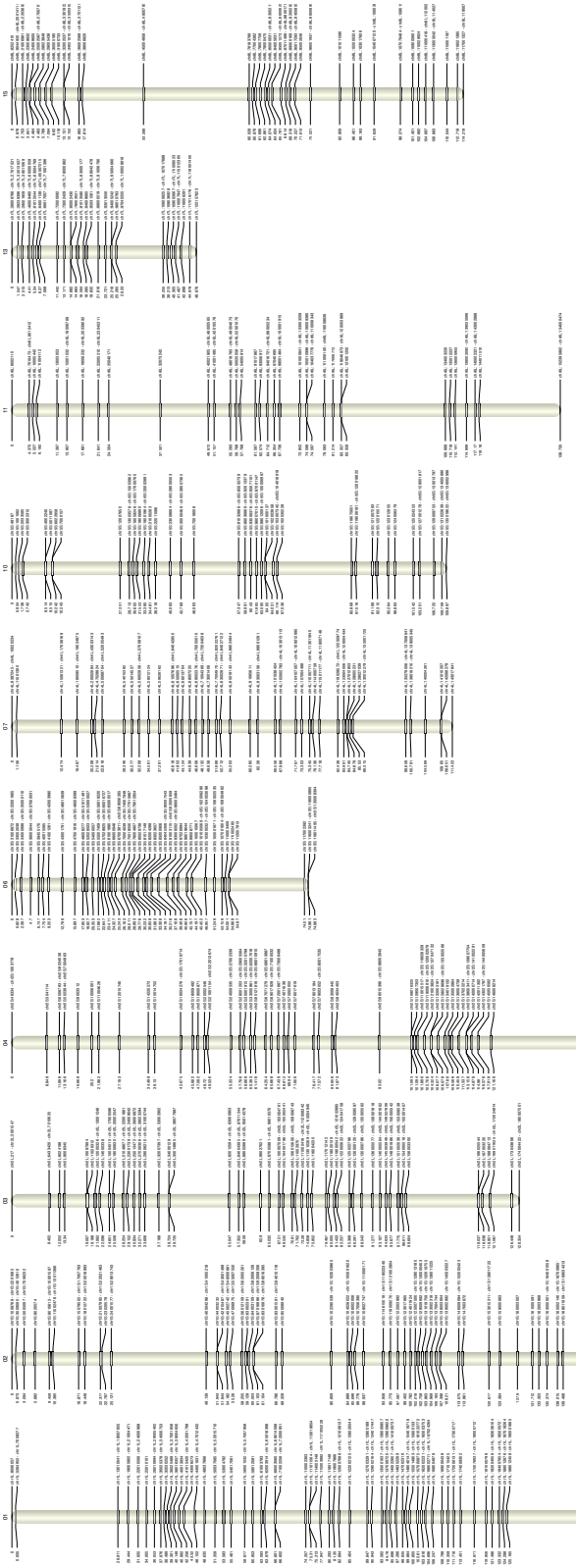


Figure 2.13: A partial genetic map of female (WC3F) *Xenopus laevis* (chromosome and linkage group). 10 Chromosomes from the Female (WC3F) Genetic Map.

2.7 A genetic map of *Branchiostoma floridae*

Sequencing and variant calling

To construct a genetic map for *B. floridae*, Jr-Kai Sky Yu and colleagues crossed two unrelated parents and raised 96 of their F1 progeny until young adult stage. The parents of the cross were selected from a laboratory colony of amphioxus adults that were raised from embryos at the Institute of Cellular and Organismic Biology, Academia Sinica, Taiwan. This colony was derived from ∼100 wild caught amphioxus adults from Tampa Bay, Florida, provided by Linda Holland and Nicholas Holland at the Scripps Institution of Oceanography, University of California at San Diego, and Daniel Meulemans Medeiros at the University of Colorado, Boulder.

DNA from the 96 progeny samples was prepared using the Maxwell Tissue DNA purification kit on a Maxwell Instrument (Promega). Illumina libraries were prepared and sequenced on four HiSeq4000 lanes in 2x150 mode, yielding an average 12 M paired-end reads per individual (nominal ∼7x coverage). DNA from both parents of the cross was extracted using Phenol-Chloroform protocol and sequenced at nominal 70x depth on a HiSeq2500 instrument with rapid run mode enable 2x250 paired-end reads.

Parents and progeny were genotyped by whole genome shotgun sequencing. All reads were aligned to the chromosome-scale reference *B. floridae* assembly using BWA-MEM [56], alignments were merged, and sorted, and duplicates were marked using novosort (Novocraft). Per site allele depths were extracted from the population from a VCF file of biallelic single nucleotide polymorphisms (SNPs) generated by FreeBayes v1.1.0 with a minimal alternate count of 3 reads (-C) and a minimal alternate fraction of 0.2 and filtered on quality (>20) [57]. Genotyped sites in the male and female parents had median depth of 66x and 67x, respectively. Progeny were sequenced to an average depth of 6.9x base coverage, which after mapping and filtering yields a 3x depth for SNPs.

We constructed separate male and female meiotic linkage maps using the pseudo-testcross method [4]. For the male map, we used biallelic SNPs that were heterozygous in the father and homozygous in the mother, allowing transmission of paternal alleles to be tracked; conversely, for the female map we used biallelic SNPs that were heterozygous in the mother and homozygous in the father. Sites that were heterozygous in both parents were ignored. This filter resulted in 459,883 raw biallelic SNP markers segregating from the father, and 423,361 from the mother.

Imputation and map construction

At an average progeny depth of 3x, we could reliably positively detect heterozygous progeny (i.e., progeny that inherited the minor allele in the cross from the heterozygous parent) but progeny with only a few observed major allele-bearing reads could be either homozygotes, or heterozygotes for which the minor allele was simply not sampled. These

"rough" genotypes are not themselves directly suitable for linkage map construction, but can be used to impute high confidence genotypes using linkage.

To robustly call genotypes, we used the fact that nearby sites (e.g., within 500 kb) are strongly linked, and are therefore either strongly correlated (minor alleles "coupling" on the same haplotype in the heterozygous parent) or anti-correlated (minor alleles "in repulsion" found on opposite haplotypes in that parents). We constructed a minimum spanning tree of edges among nodes in the window, minimizing the sum of the negative absolute value of two-sided binomial test p-value along the spanning tree $p <$ 1e-3. The sign of the correlations along the tree allows us to determine the relative phase of each SNP inherited from the heterozygous parent. This analysis determines the two haplotypes of the heterozygous parent within each 500 kb window. Progeny are assigned one of the two haplotypes if the haplotypes could be assigned with $p <$ 1e-3 confidence.

Markers for paternal meiotic map: Out of 970 non-overlapping full 500 kb genomic windows that had at least 1 SNP, 957 (98.6%) contained sufficiently many correlated paternal SNPs. Of these, 898 could be called in more than 80% of progeny, and were used for paternal map construction. In addition, 273 windows were shorter than 500 kb that had at least 1 SNP, (representing ends of contigs/chromosomes, and sub-500kb scaffolds). Of these windows, 104 (38.1%) had sufficiently many correlated paternal SNPs. Of these, 35 could be called in more than 80% of progeny.

Markers for maternal meiotic map: Out of 970 non-overlapping full 500 kb genomic windows that had at least 1 SNP, 956 (98.5%) contained sufficiently many correlated maternal SNPs. Of these, 896 could be called in more than 80% of progeny, and were used for maternal map construction. Of the 272 windows shorter than 500 kb that had at least 1 SNP, 100 (36.8%) contained sufficiently many correlated maternal SNPs. Of these, 40 could be called in more than 80% of progeny.

I constructed separate male and female linkage maps with the `onemap` package (v2.0-3) in R [18], constructing each map using the "F1 cross" setting and the recommended LOD score offered by `suggest_lod` providing the genotype calls for non-overlapping 500 kb windows as described above. Markers were further grouped into bins with 0 cM recombination according to best practices from Schiffthaler et al. [59]. We found 19 major linkage groups in the mother, and 20 major linkage groups in the father (log odds (LOD) threshold 4.845264 for male map and 5.427699 for female map).

The 19 major male and female linkage groups are almost in 1:1 correspondence with the 19 chromosome-scale scaffolds assembled using HiC chromatin linkages, confirming the accuracy of these chromosomes. The total length of the 19 linkage groups with the largest number of markers from each of the 19 BFL chromosomes for the male and female maps were 848.4 and 945.5 cM, respectively. The final male and female linkage groups that represent each chromosome comprised 929 and 936 composite markers for the male and female maps, respectively.

Discussion

Amphioxus chromosomes are mostly acrocentric, that is, their centromeres, typically regions of very little recombination, are located at or near the chromosome end. This explains the lack of central plateau in the map distance vs. genomic coordinate plots generally found in metacentric chromosomes, like those of *X. laevis*. Although we genetically confirmed the chromosomal linkages of our HiC-based assembly, the ordering of markers was not perfectly concordant with their order along the assembly. In particular, both male and female maps show that the assembly of chromosome 19 appears to have an intrachromosomal rearrangement relative to the genetic map, suggesting need for further refinement of assembly. Since our major results depend on amphioxus linkage but are insensitive to the intrachromosomal gene order, we used the HiC-based assembly of *B. floridae* for all comparative analyses presented in chapter 4.

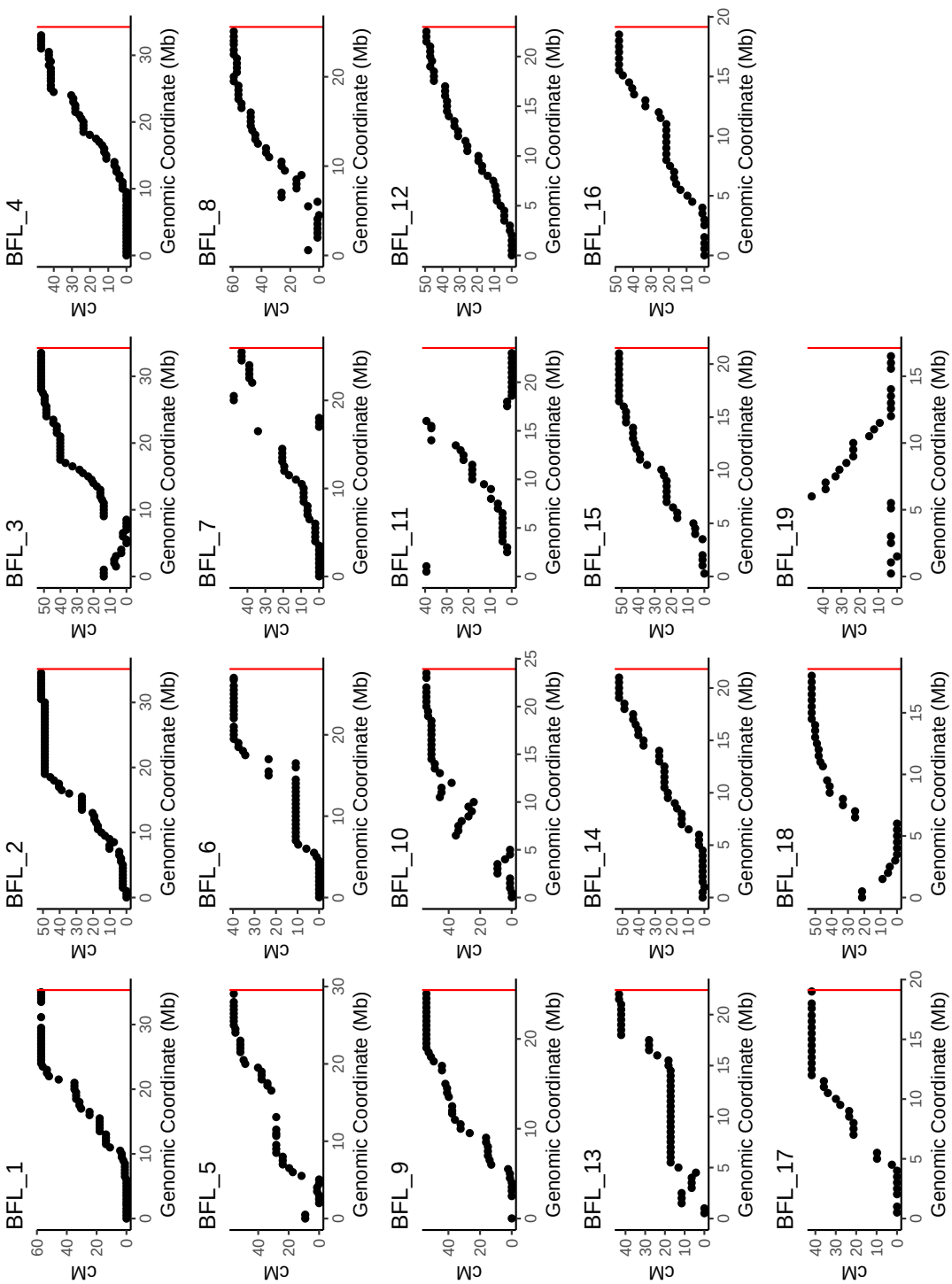


Figure 2.14: A genetic map of B. floridae

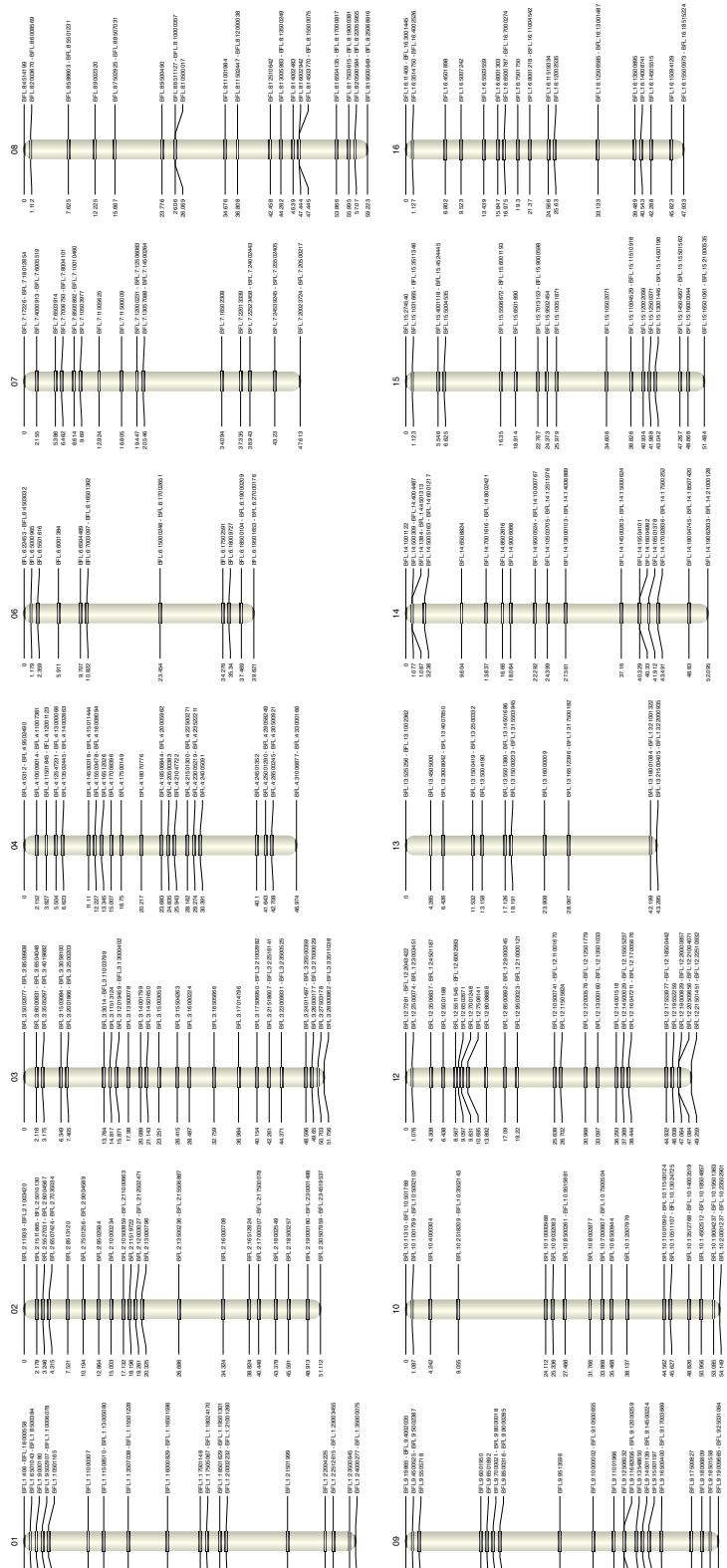


Figure 2.15: A genetic map of B. floridae.

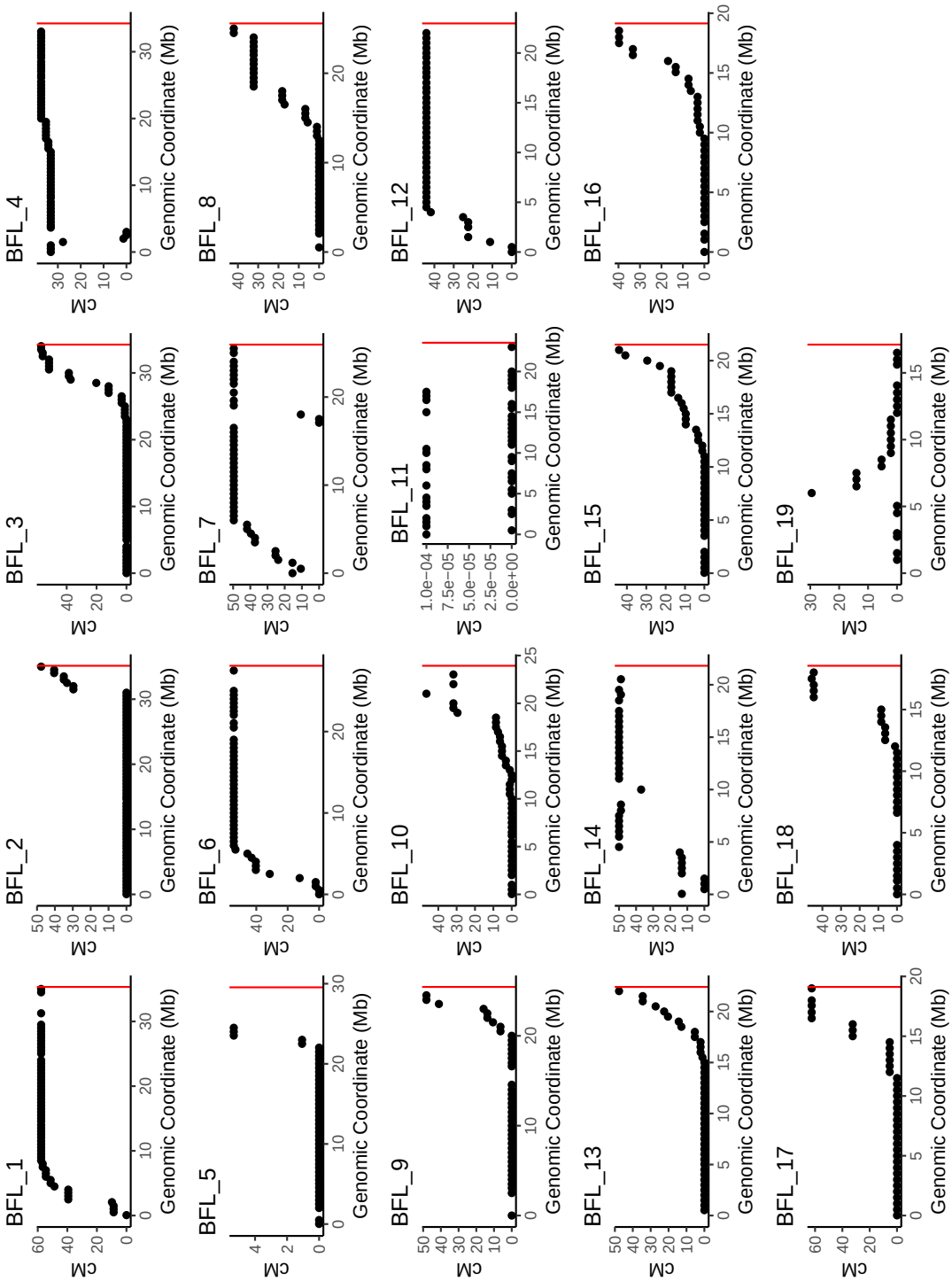


Figure 2.16: A genetic map of B. floridae

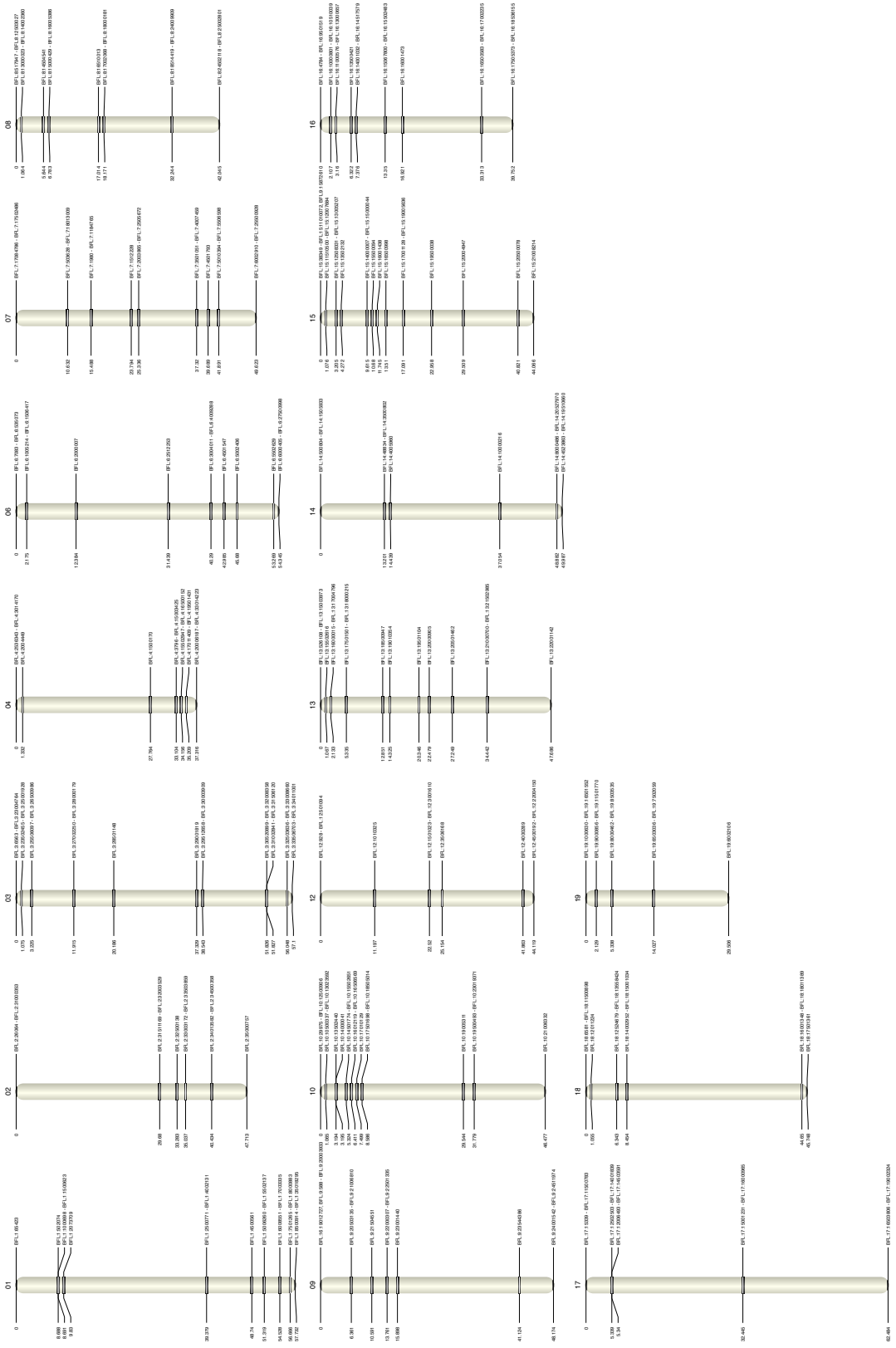


Figure 2.17: A genetic map of B. floridae

2.8 A genetic map of *Miscanthus sinensis*

Sequencing and variant calling

Miscanthus sinensis is a giant grass with a paleotetraploid genome. Considerable interest exists in creating accurate genetic maps for *Miscanthus sinensis*, especially to study the structure of its two sub-genomes in the context of whole genome duplication [60–62]. While *Miscanthus sinensis* is used as an ornamental grass in landscaping, its interest to researchers is as a bioenergy crop.

2-3 young leaf rolls were sampled from a full-sib biparental population (181 lines) as well as the two *M. sinensis* parents (Grosse Fontaine and Undine). All lines were grown in the field at the University of Illinois Energy Farm. The leaf rolls were flash frozen in liquid nitrogen and ground frozen into a fine powder. Genomic DNA was isolated using a CTAB extraction protocol [63, 64]. The DNA was arrayed into a 96 well plate and submitted to the JGI for Illumina sequencing. They were sequenced on an Illumina HiSeq-2000,2x150.

Reads were aligned to the genome with bwa mem v.0.7.15. SNPs were called using the Genome Analysis Toolkit (GATK) version 3.7-0-gcfedb67 using HaplotypeCaller in GVCF mode with a minimum mapping quality score of 25 followed by GenotypeGVCFs with useNewAFCalculator.

Imputation and map construction

We constructed a minimum spanning tree of edges among nodes in the window, minimizing the sum of the negative absolute value of two-sided binomial test p-value along the spanning tree $p <$ 1e-3. The sign of the correlations along the tree allows us to determine the relative phase of each SNP inherited from the heterozygous parent. This analysis determines the two haplotypes of the heterozygous parent within each 250kb window. Progeny are assigned one of the two haplotypes if the haplotypes could be assigned with $p <$ 1e-3 confidence.

Markers for paternal meiotic map: Out of 7,566 non-overlapping full 250 kb genomic windows, 6,992 (92.4%) contained sufficiently many correlated paternal SNPs. Of these, 6,054 could be called in more than 80% of progeny, and were used for paternal map construction. In addition, 4,811 windows were shorter than 250 kb that had at least 1 SNP, (representing ends of contigs/chromosomes, and sub-250kb scaffolds). Of these windows, 1,656 (34.4%) had sufficiently many correlated paternal SNPs. Of these, 523 could be called in more than 80% of progeny.

Markers for maternal meiotic map: Out of 7,568 non-overlapping full 250 kb genomic windows that had at least 1 SNP, 7031 (92.9%) contained sufficiently many correlated maternal SNPs. Of these, 6298 could be called in more than 80% of progeny, and were used for maternal map construction. Of the 4,735 windows shorter than 250 kb that had at least 1 SNP, 1,728 (36.5%) contained sufficiently many correlated maternal SNPs. Of these, 546 could be called in more than 80% of progeny.

I constructed separate male and female linkage maps with the `onemap` package (v2.0-3) in R [18], constructing each map using the "F1 cross" setting and the recommended LOD score offered by `suggest_lod` providing the genotype calls for non-overlapping 250 kb windows as described above. Markers were further grouped into bins with 0 cM recombination according to best pratices from Schiffthaler et al. [59]. We found 19 major linkage groups in the mother and father (log odds (LOD) threshold 6.634203 for male map and 6.591244 for female map).

The 19 major male and female linkage groups are almost in 1:1 correspondence with the 19 chromosome-scale scaffolds assembled using HiC chromatin linkages, confirming the accuracy of these chromosomes. The total length of the 19 linkage groups with the largest number of markers from each of the 19 *M. sinensis* chromosomes for the male and female maps were 1,308.28 and 1,138.17 cM, respectively. The final male and female linkage groups that represent each chromosome comprised 6,577 and 6,844 composite markers for the male and female maps, respectively.

Discussion

The genome assembly this map built from is still in development. It was made using HiC chromatin confirmation linkages, which can be ambiguous. Although we genetically confirmed the chromosomal linkages of our HiC-based assembly, the ordering of markers was not perfectly concordant with their order along the assembly. Several small misassemblies can be visualized along the chromosomes for both male and female. Small discontinuities with inverted ordering represent sequences in the genome that were likely inverted during contig assembly or scaffolding. An example of this can be seen as an inversion in Chr06 centered around 12.5 Mb or in the first 10 Mb of Chr03. The genetic distance of our map of *Miscanthus sinensis* map is the smallest reported in the literature by ∼500-700 cM, though we had we had a denser and more uniformly distributed set of markers. Because it is far easier to erroneously inflate maps than it is to erroneously contract them, we believe this to represent an improvement in existing maps. Ultimately, the primary goal of this map is to assist with the genome assembly.

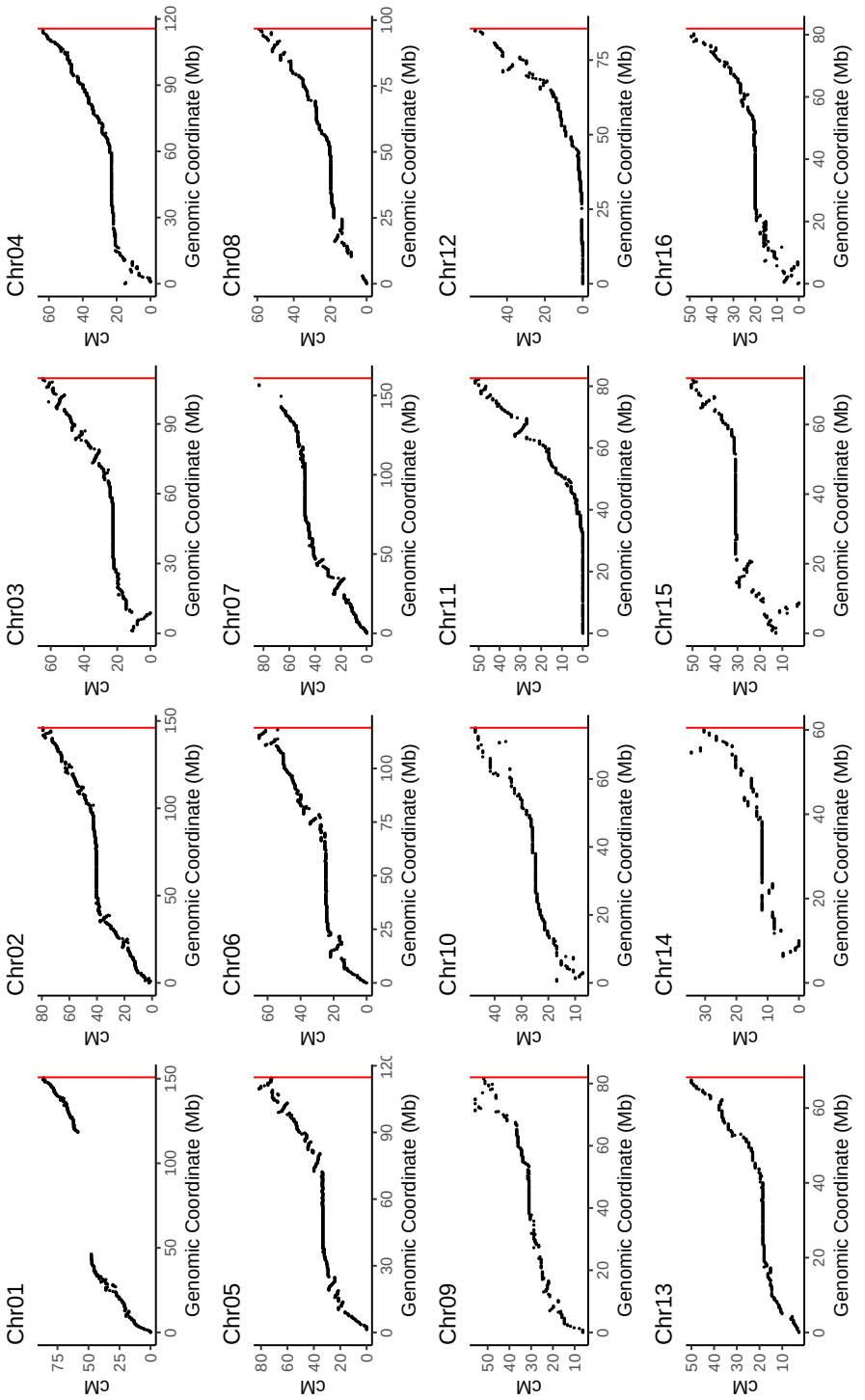


Figure 2.18: A genetic map of *M. sinensis* (Female)

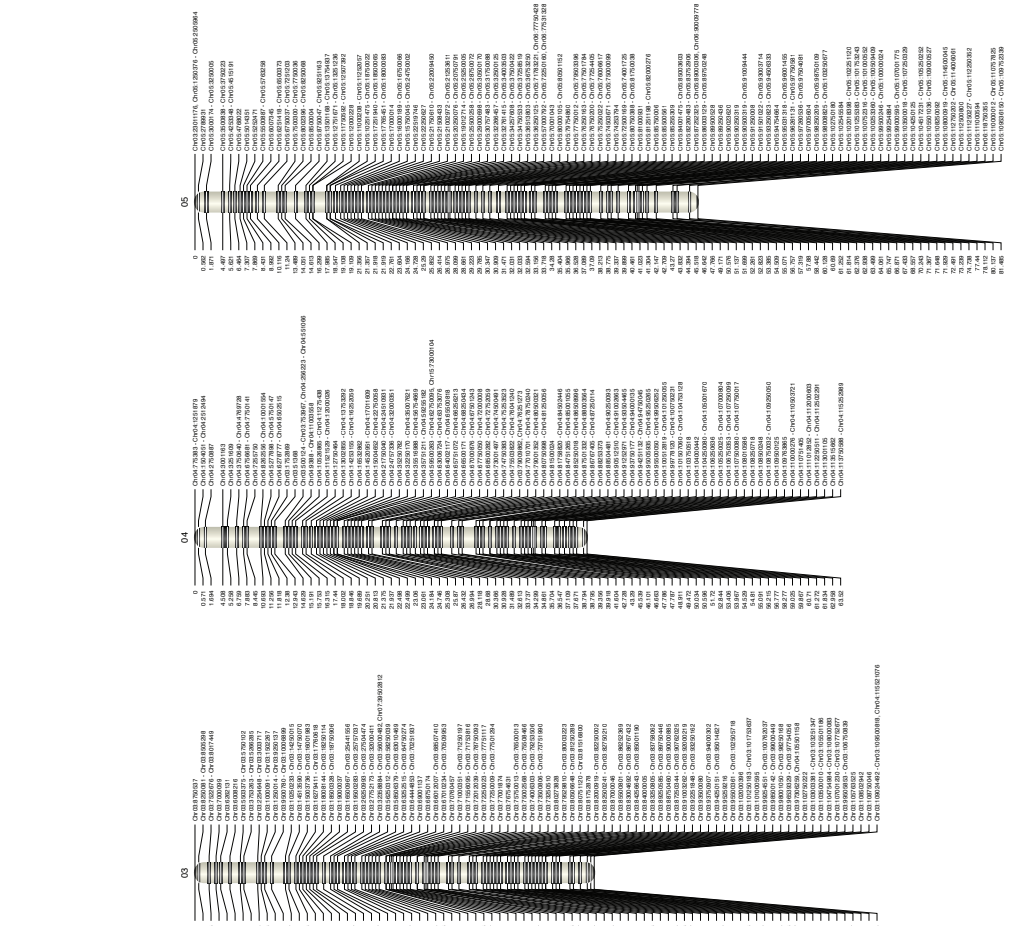


Figure 2.19: A genetic map of *M. sinensis* (Female pt. 1)

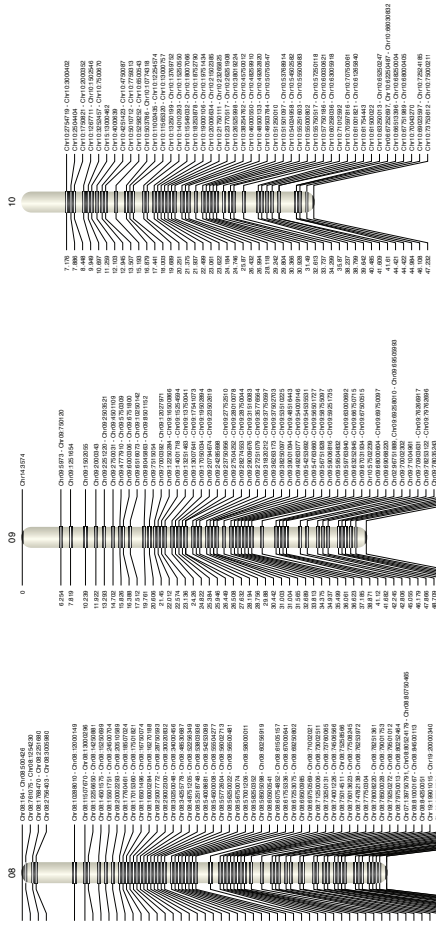


Figure 2.20: A genetic map of *M. sinensis* (Female pt. 2)

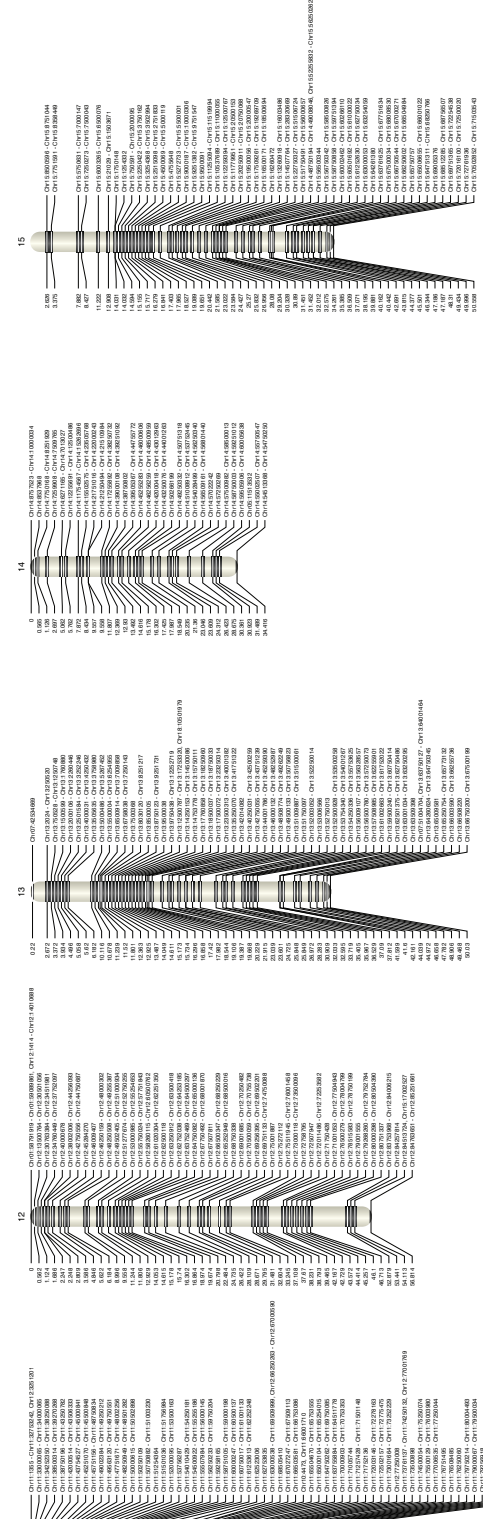


Figure 2.21: A genetic map of *M. sinensis* (Female pt. 3)

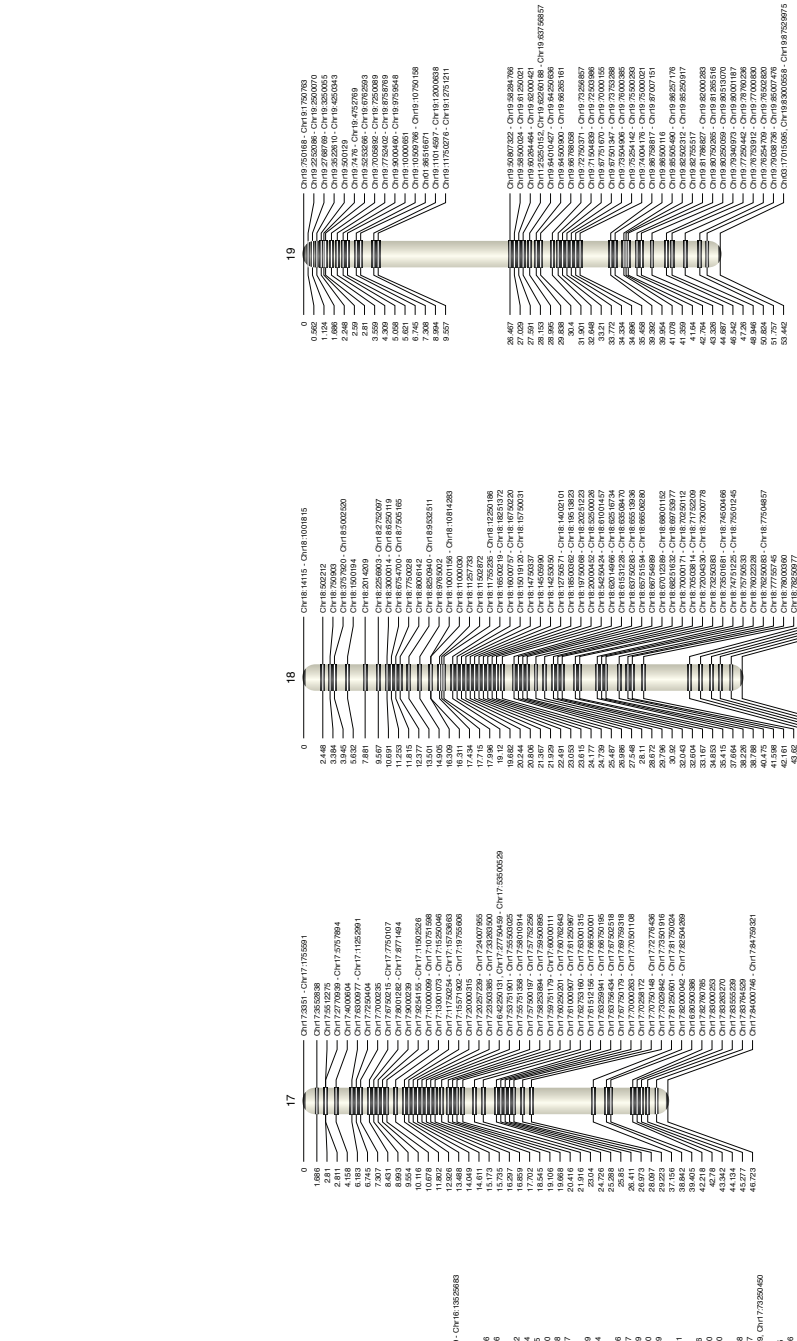


Figure 2.22: A genetic map of *M. sinensis* (Female pt. 4)

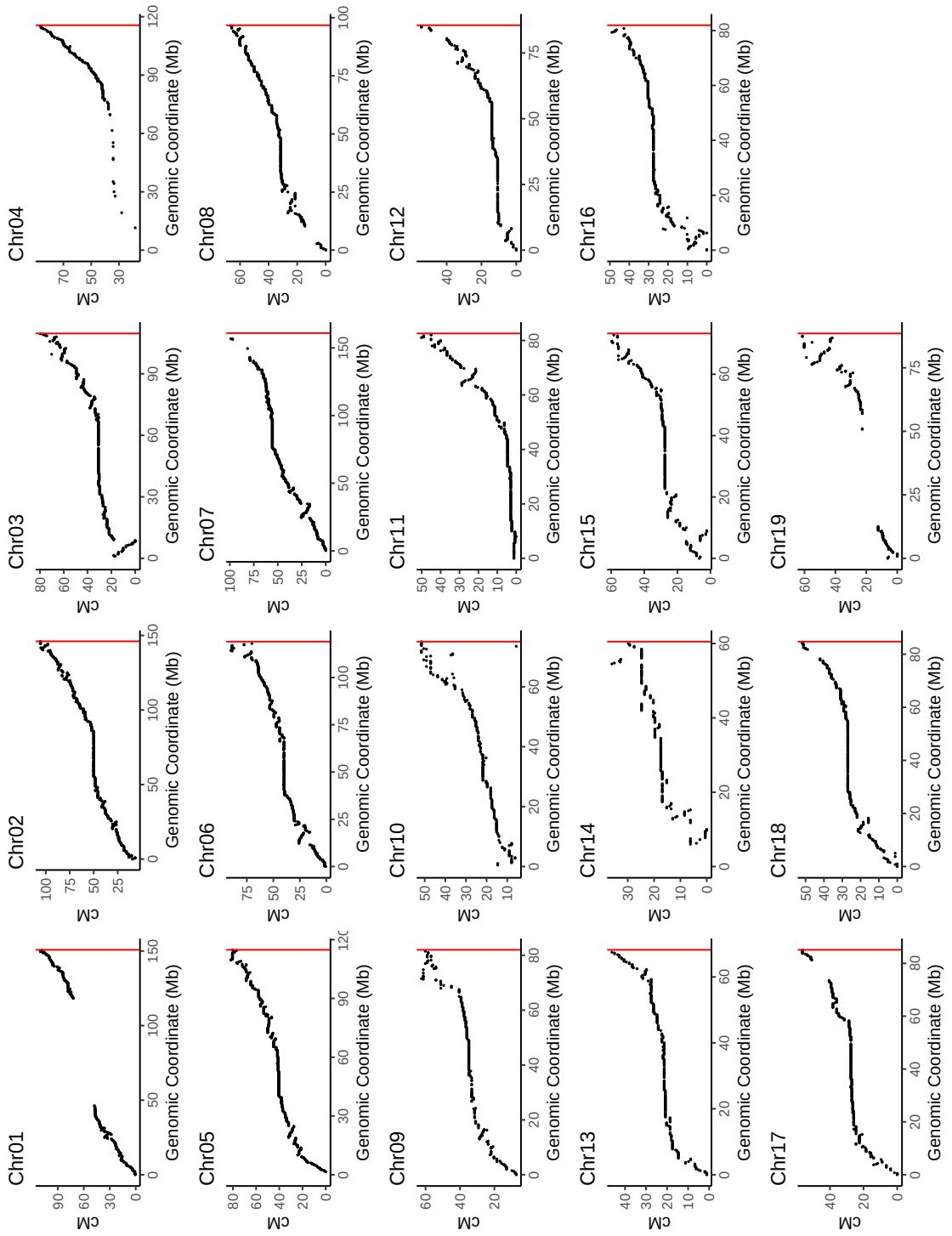


Figure 2.23: A genetic map of *M. sinensis* (Male)

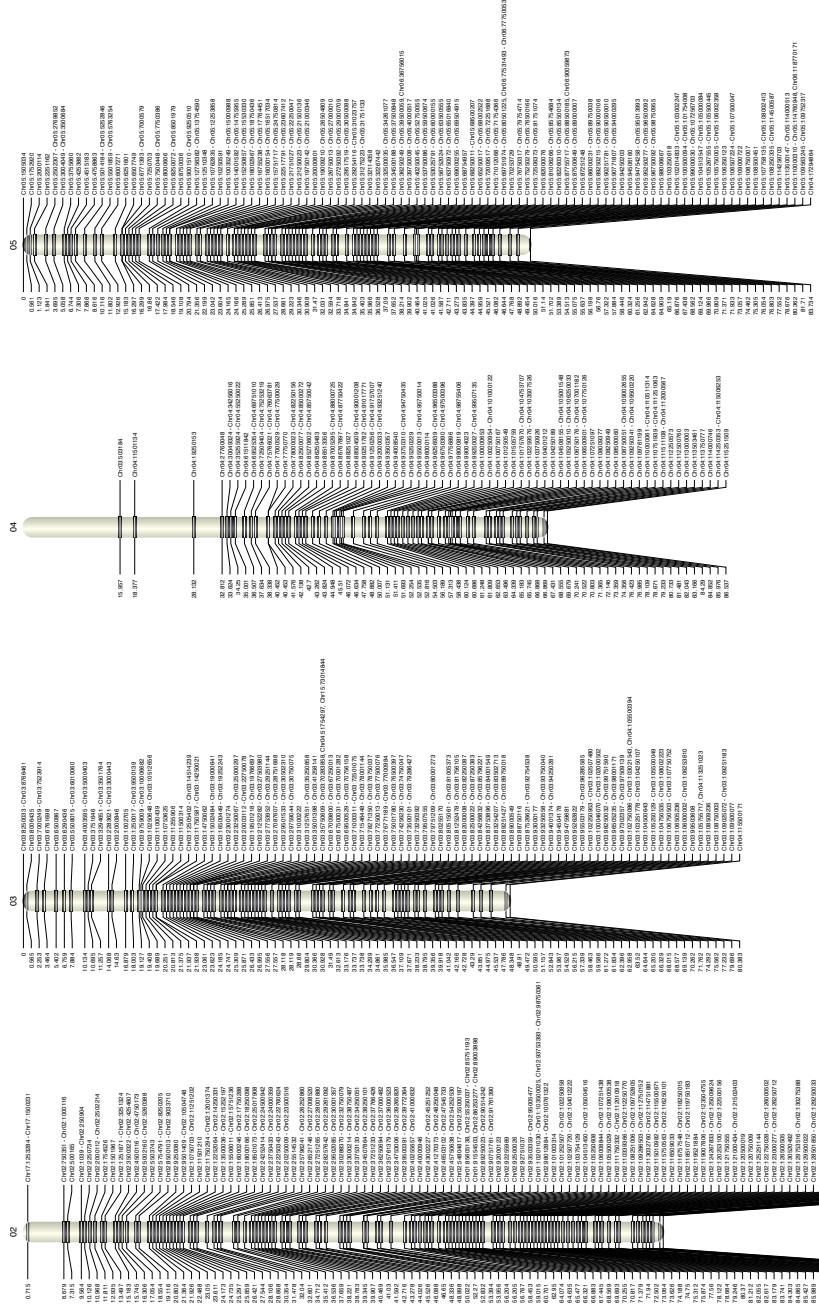


Figure 2.24: A genetic map of *M. simensis* (Male pt. 1)

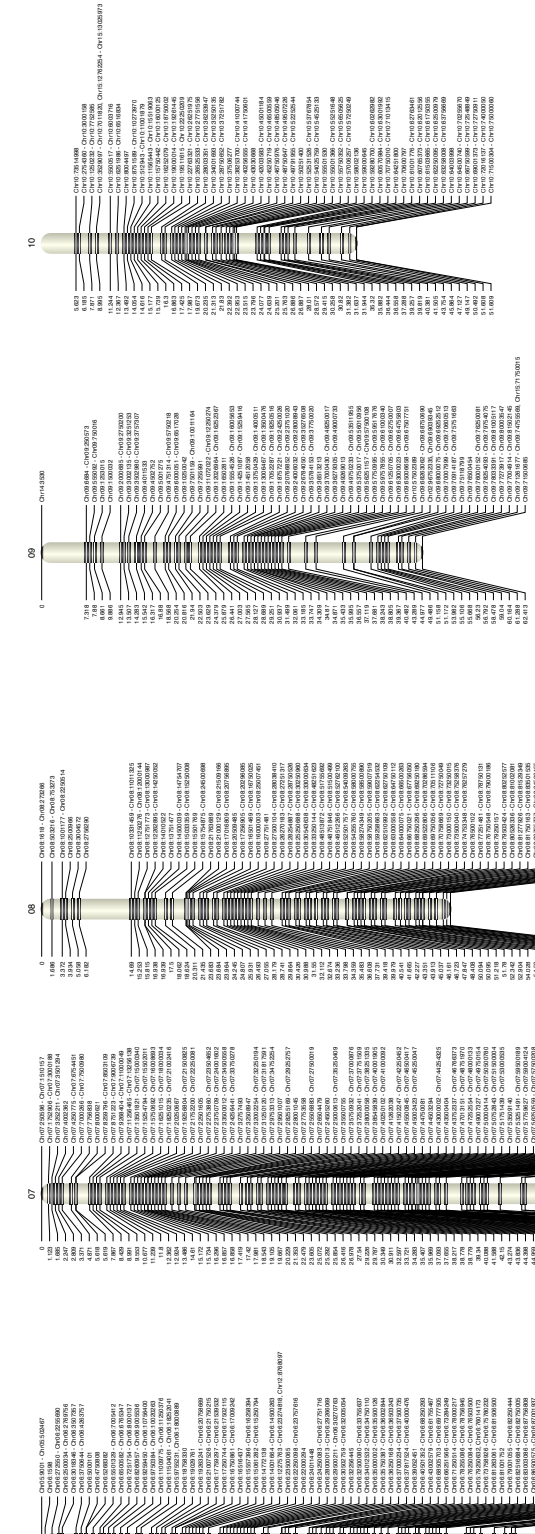


Figure 2.25: A genetic map of *M. sinensis* (Male pt. 2)

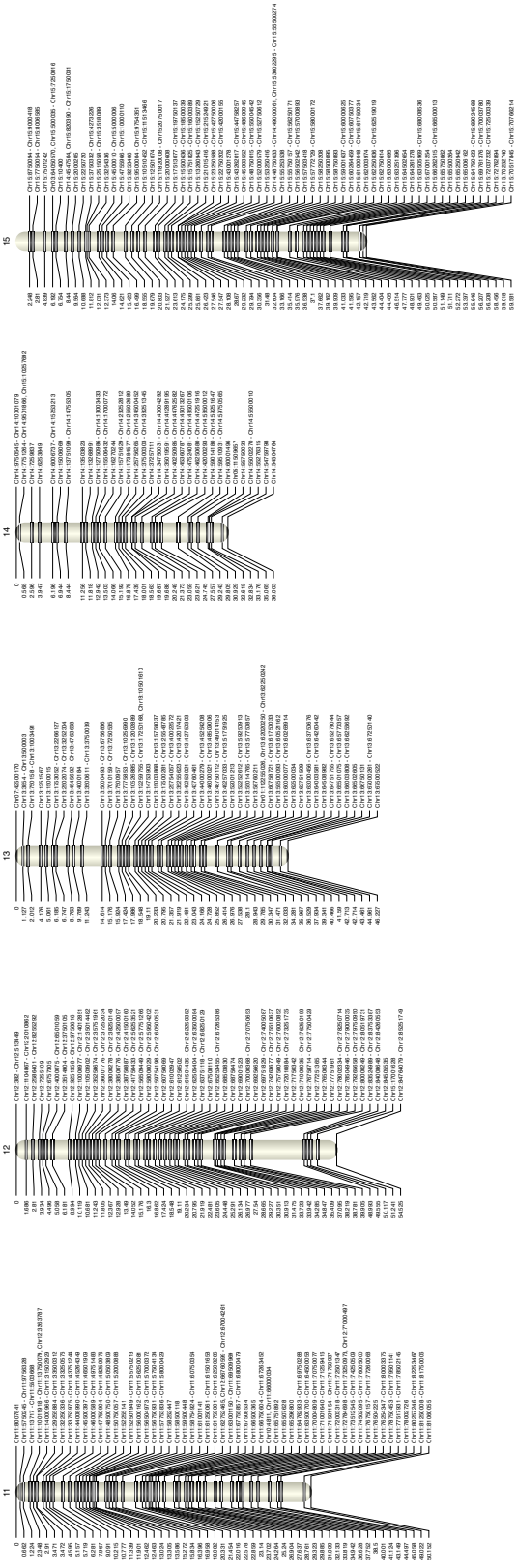


Figure 2.26: A genetic map of *M. simensis* (Male pt. 3)

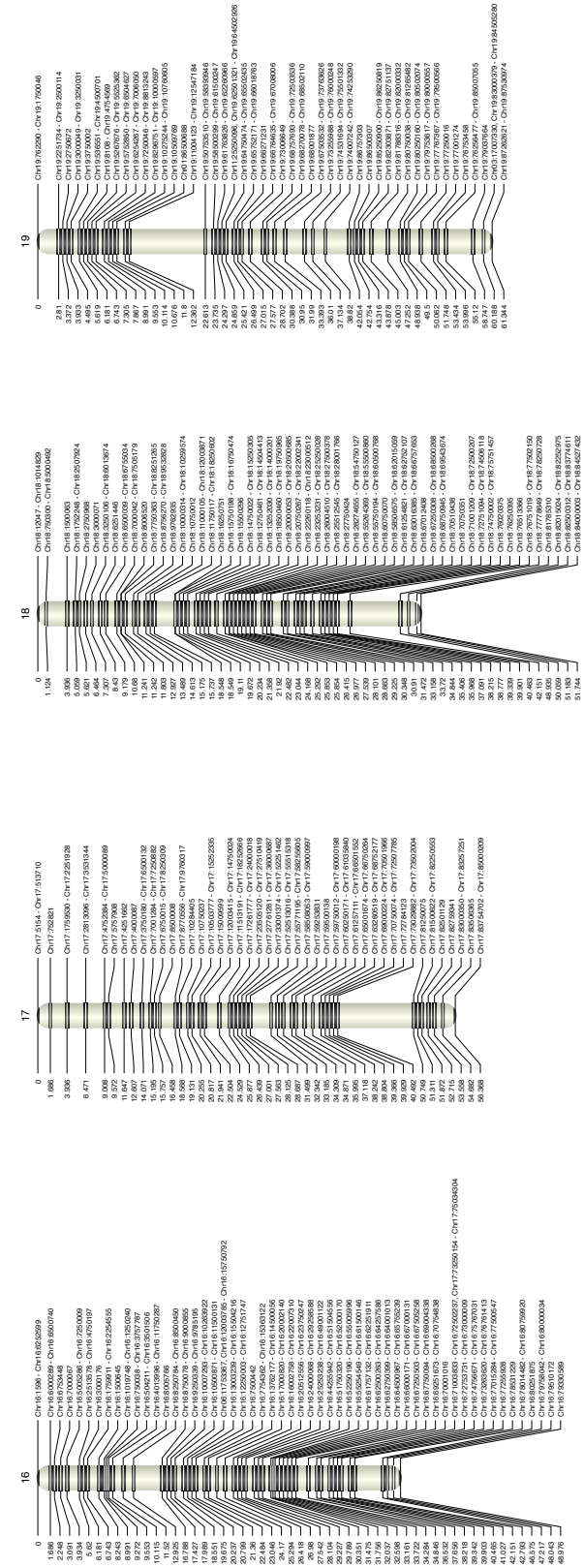


Figure 2.27: A genetic map of *M. simensis* (Male pt. 4)

Chapter 3

A Large CRISPR-Induced Bystander Mutation Causes Immune Dysregulation[†]

3.1 Introduction

CRISPR-Cas9 genome engineering is employed widely to generate targeted *in vitro* and *in vivo* genetic modifications [65]. The Cas9 nuclease can be programmed to target specific genome sequences via a short guide RNA. Although unintended genome alterations have been mitigated by recent technical advances [66–70], they remain a significant concern, especially for therapeutic applications of CRISPR. To date, attention has been focused on "off target" editing, in which Cas9 nuclease activity is directed towards genomic sites, other than the target, with varying degrees of homology to the guide RNA. Here we demonstrate that "bystander" mutations – unintended mutations neighboring the "on-target" cut site - must also be considered.

One advantage of genome editing over RNA knock-down approaches is that non-coding sequences can be modified, which enables studies of non-coding variants commonly associated with human disease risk. Our collaborators lead by Dimitre Simeonov and Dr. Alexander Marson recently identified a conserved autoimmunity-associated IL2RA intronic enhancer that controls the timing of gene expression in response to T cell stimulation [71]. To study its *in vivo* function, we used CRISPR to engineer NOD mice with deletion of this enhancer (EDEL). We successfully generated four EDEL founder lines by targeting Cas9 to cut on either side of the 360 bp enhancer (Figure 3.1a). Genomic PCR and targeted Sanger sequencing confirmed that approximately 360-370bp was deleted at the enhancer site in multiple founders (Figure 3.1b). Three of the founders were backcrossed to wildtype NOD animals at least one generation before breeding the enhancer deletion to homozygosity for

[†]This chapter is based on the manuscript "A Large CRISPR-Induced Bystander Mutation Causes Immune Dysregulation" by Simeonov and Brandt et al.

experimentation.

Surprisingly, immunophenotyping revealed a marked systemic difference in one line of mice. Unlike the other two characterized lines, homozygous EDEL progeny from the third founder line had hallmark features of a lymphoproliferative disorder, including variable splenomegaly, increased cellularity and memory T cell expansion (Figures 3.1c,d, 3.5, and 3.6). Despite the phenotypic differences among the lines, the on-target enhancer deletion only differed by a few nucleotides at the margins of the deletion. The evolutionarily conserved DNA sequence at the site was deleted in all three lines and the line with more severe phenotype had a slightly smaller deletion, suggesting that the genotyped sequence differences directly at the deletion site did not explain observed differences in immune regulation (Figures 3.5 and 3.6). The more severe immune phenotype persisted in progeny with the enhancer deletion from the affected line, even after an additional round of backcrossing and multiple generations of breeding, suggesting a mutation in close genomic proximity to the on-target deletion site rather than an unlinked off-target effect. Taken together our data suggested the presence of an additional mutation linked to the Il2ra enhancer deletion in this immune dysregulated founder line (IDFL).

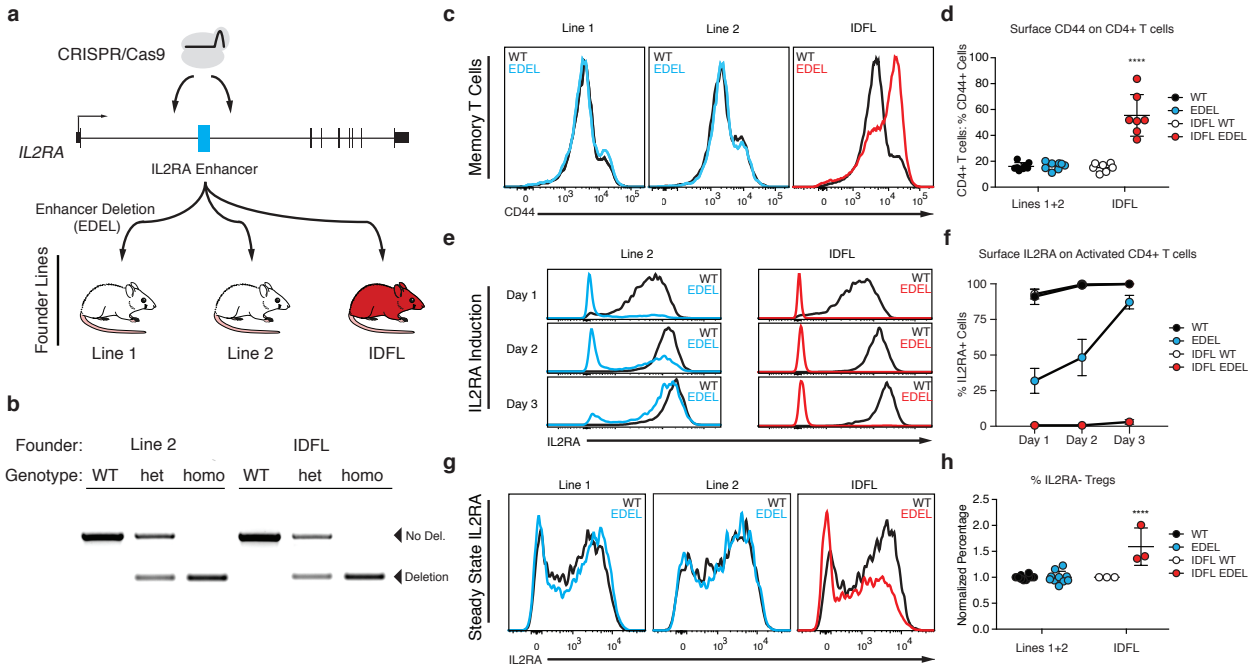


Figure 3.1: Figure 1. Immune Dysregulation in a Founder Line of CRISPR-Engineered Il2ra Enhancer Deletion Mice. a, CRISPR-engineered Il2ra enhancer deletion (EDEL) founder lines that were bred for immunophenotyping. b, Genomic DNA PCR to genotype the Il2ra enhancer deletion in animals from Line 2 and the immune dysregulated founder line (IDFL). c, Representative CD44 surface staining on CD4+ T cells isolated from spleens of wild-type (WT) and EDEL mice from different founder lines. d, Quantification of CD44 surface staining from (c) (Lines 1 and 2: WT n=8, EDEL n=7; IDFL: WT n=8, EDEL n=7). e, Representative induction of IL2RA surface expression on naïve CD4+ T cells (CD4+/IL2RA-/CD44-) activated with anti-CD3/CD28 antibodies. f, Quantification of percent IL2RA+ cells from (e) (Line 2: WT n=4, EDEL n=4; IDFL: WT n=4, EDEL n=4). g, Representative IL2RA surface expression on FOXP3+/CD4+/T cells (Tregs) from spleen of different founders. h, Quantification of normalized IL2RA surface expression on FOXP3+/IL2RA+ Tregs from (g) (Lines 1 and 2: WT n=10, EDEL n=10; IDFL: WT n=3, EDEL n=3). Panels (d) and (h) include data from Line 1 and 2 animals previously published [71]

3.2 Searching for possible off-target mutations

Given the segregation of the novel immune deficiency, I strongly suspected the trait was closely linked, and therefore close in physical space, to initial (intended) deletion. When SNP, short INDEL, and longer structural variants were called, I attempted to search for a mutation or group of mutations that segregated homozygously in GT_05105 (the affected IDFL mouse), and heterozygously in GT_05102 (an unaffected sibling of the IDFL mouse, assumed to be heterozygous for the unintended mutation), and absent in GT_05111 (an unaffected mouse

from a line that never showed the IDFL phenotype). No potential candidates were found in this way.

However, by manually inspecting the alignments for the three individuals, I found an sharp increase in coverage next to the intended genetic deletion in the IDFL mouse and its sibling, but not in the unaffected individual from the other line. When I subset the reads based on mapping pair orientation, I found fragments mapping in the forward-reverse (FR) orientation rather than the expected reverse-forward (RF) orientation (Figure 3.2). These reads had a consistent pattern that suggested the presence of a large segmental duplication, and represented reads that were derived from the of the duplication junction (Figure 3.3).

Sample ID	Phenotype	Lineage
GT_05105	Immune Dysregulation	IDFL
GT_05102	No Immune Dysregulation	IDFL
GT_05111	No Immune Dysregulation	Line 2 (Non-IDFL Line)

Table 3.1: Sample key for IGV

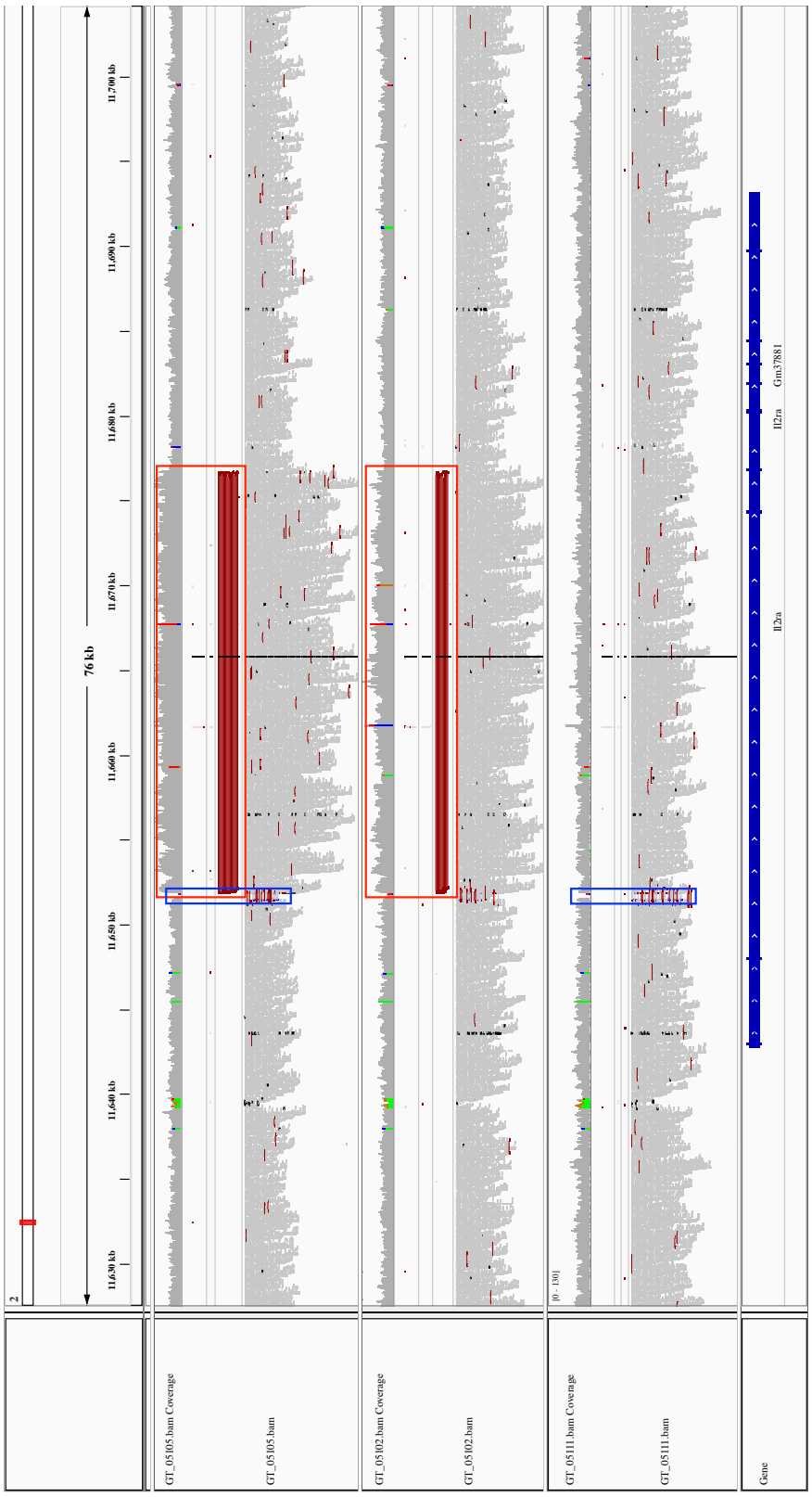


Figure 3.2: An annotated IGV screenshot supporting evidence of a large tandem duplication in the IL2RA locus. The intended deletion is boxed in blue. The fragments supporting the presence of the tandem duplication are boxed in red.

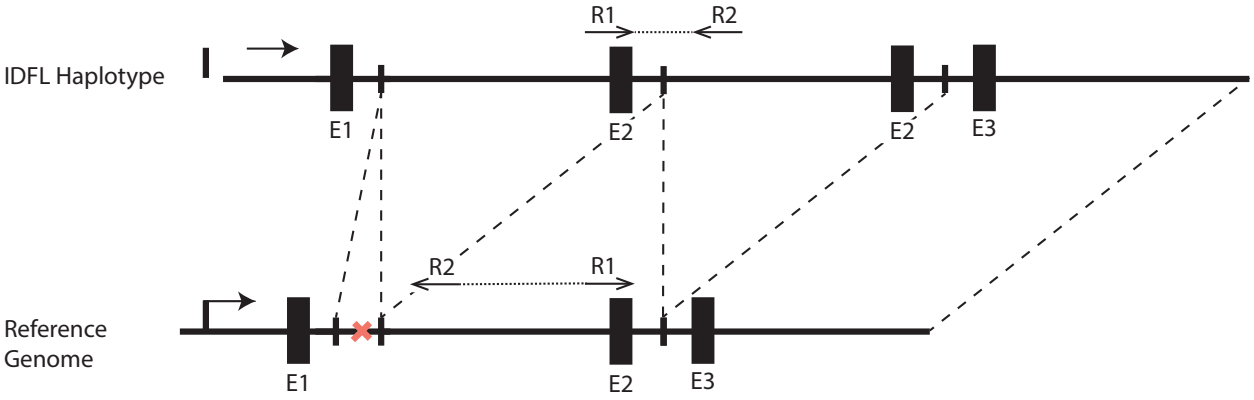


Figure 3.3: Read pair mapping consistent with duplication.

3.3 Results

To determine the molecular and cellular effects of the linked mutation in the IDFL mice, we analyzed IL2RA expression. T cell development was relatively spared, with the exception of double negative (DN) thymocytes, which had marked loss of IL2RA expression (Figure 3.5 and 3.6). Mature CD4+ effector T cells (Teff) normally upregulate IL2RA to their surface after activation. Strikingly, activated IDFL Teff failed to express IL2RA on their surface and were devoid of the receptor (Figure 3.5 e, f). This was in contrast to the other EDEL lines, which showed delayed but not ablated induction of IL2RA following stimulation of naive T cells [71]. We also examined FOXP3+ regulatory T cells (Tregs), which constitutively express high levels of IL2RA and require it for their survival. Across lymphoid tissues there was an increased percentage of FOXP3+/IL2RA- Tregs in IDFL mice but not other EDEL lines (Figure 3.5g,h). *In vitro* and *in vivo* Treg differentiation was profoundly impaired. Interestingly, a subset of Tregs did express high levels of IL2RA. An Il2ra null mutation would be expected to ablate expression across cell types. Instead we find that the linked mutation has cell type specific effects on IL2RA expression, with a subset of T cells selectively maintaining IL2RA expression.

To identify the mutation causing marked immune dysregulation, we sequenced the whole genomes of EDEL mice from the immune dysregulated founder line and from one of the other founder lines (Figure 3.3a). We systematically looked for a causative IDFL mutation both at the IL2ra locus and throughout the genome. Consistent with the observed genetic linkage with the enhancer deletion, we discovered a large structural mutation in the Il2ra locus that was unique to the IDFL genome. Careful analysis of the read pileups revealed a 24kb block of DNA with elevated coverage in the IDFL genome compared to adjacent sequences, consistent with an increase in copy number (Figures 3.2 and 3.4a). Paired end reads at the breakpoint implied a tandem duplication, which we confirmed by genomic PCR and Sanger sequencing (Figures 3.2 and 3.4b). The duplicated sequence starts immediately downstream of the deleted Il2ra enhancer, spans the remainder of the first intron, the second Il2ra exon

and most of the second intron. This unexpected structural mutation tightly linked to the intended on-target edit is a "bystander" mutation that causes marked immune dysregulation.

We next interrogated how the duplication formed. Previous work showed that paired CRISPR-induced DNA breaks can result in tandem duplication of the intervening sequence (Figure 3.11d), however no predicted off-target cutting sequences were identified in the Il2ra locus to explain this duplication event (Figure 3.9a). Although we cannot rule out spontaneous DNA breaks from DNA replication, our data suggests that the duplication is more likely an unintended product of repair from on-target editing. Sequence homology could contribute to a duplication event. We did not find extended sequence homology between the cut site and the duplication junction, but we did discover microhomology at the breakpoint junction (Figure 3.9c). However, three nucleotides of microhomology are found commonly in the genome, raising a question of why this distal site may have been used for repair. Chromatin looping can bring distal genomic sites into close proximity. Indeed, published high-resolution chromatin conformation capture data revealed three-dimensional proximity between the Il2ra enhancer and the site of the duplication junction (Figure 3.9b) [72]. We then tested whether microhomology and looping were sufficient to drive recurrent duplications at this site (Figure 3.11d). We repeated CRISPR microinjections into single cell NOD/ShiLtJ zygotes, cultured them *in vitro* and analyzed >50 blastocysts and failed to observe the recurrence of this particular duplication event despite efficient deletion (∼80%) of the enhancer at the on-target site. Note, there were two blastocysts with PCR bands of the expected duplication size, but the sequence could not be confirmed. Taken together our analysis of the duplication is consistent with a complex unintended repair consequence that occurs less frequently than the intended enhancer deletion.

The duplication of exon 2 is predicted to generate a novel splice junction that would result in a premature stop codon in the Il2ra mRNA (Figure 3.4c). While the expected effect of the homozygous duplication would be to ablate protein expression in every cell, we were nevertheless able to detect CD4+ T cells with near normal levels of IL2RA expression. To understand how these cells expressed IL2RA despite the predicted premature stop codon generated by the duplication we performed RNA-seq on IL2RA+ T cells from spleen. As expected, we could identify reads with the aberrant exon2-exon2 splice junction only in the IFDL/EDEL cells (Figure 3.4c). However, IL2RA+/IDFL cells had 10-fold more reads that contained the wild-type exon2-exon3 junction than the aberrant exon2-exon2 junction (Figure 3.4c). PCR and Sanger sequencing of Il2ra cDNA confirmed that these cells predominantly generate Il2ra transcript with a single exon 2 between exons 1 and 3 (Figure 3.7a and 3.7b). In contrast, we did not detect transcripts with a single exon2 in IL2RA- T cells induced to express Il2ra (Figure 3.7a and 3.7b). These isoform differences suggest that some T cells, including a subset of Foxp3+ Tregs, are able to correctly splice the aberrant Il2ra genetic structure and productively translate IL2RA.

This chapter links a CRISPR-induced bystander mutation to *in vivo* pathology. Further work is needed to understand the frequency with which such mutations occur, as well as the DNA repair rules that underlie these events. Identification of the bystander mutation depended on having multiple independent founder pedigrees to demonstrate an aberrant

phenotype in one line and genetically link it to the on-target edit. New methods and analytical tools are needed to detect both unintended CRISPR-induced bystander and off-target mutations. Genome engineering not only allows gene knockout, but also permits targeted alterations of non-coding cis-regulatory sequences for mechanistic study of human variants and for cell therapies. The bystander mutation allele observed here was introduced by murine zygote editing by Cas9 mRNA and gRNA microinjection. The marked immune phenotype was revealed by breeding the rare allele to homozygosity. The functional consequences, if any, of rare unintended alleles in a population of human primary somatic cells edited by various CRISPR delivery strategies remain largely untested. Bystander editing effects – which can be easily missed with conventional genotyping methods – must be carefully assessed for research and clinical CRISPR applications, especially for mounting therapeutic efforts to fine tune gene regulatory programs.

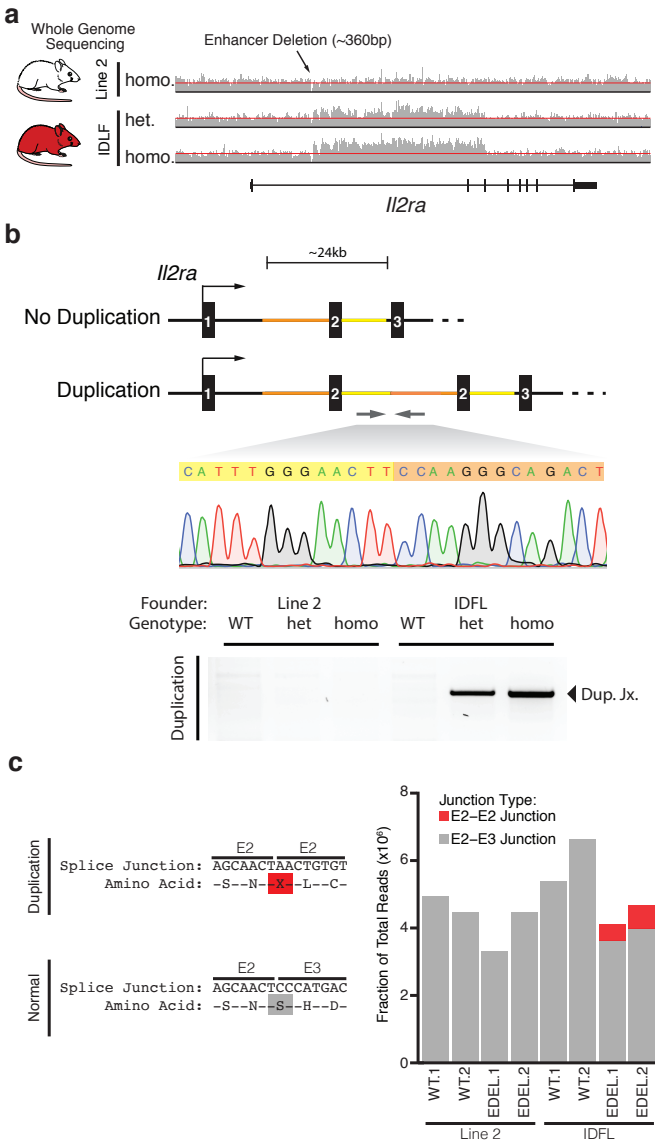


Figure 3.4: Identifying a Large Tandem Duplication in the Il2ra Locus. a, Read pileups at the Il2ra locus from genome sequencing of a homozygous enhancer deletion (EDEL) mouse (Line 2) and homozygous and heterozygous EDEL mice from the immune dysregulated founder line (IDFL). Red lines were added to highlight the elevated read counts in IDFL mice. b, Schematic of the Il2ra locus with the large tandem duplication in IDFL mice. PCR and Sanger sequencing across the novel junction sequence created by the duplication. c, Read counts from RNA sequencing of IL2RA+/CD4+ T cells showing reads that span the aberrant exon 2-exon 2 junction in red and the normal exon 2-exon 3 junction in grey. Data in (c) are from biological replicates derived from two independent experiments.

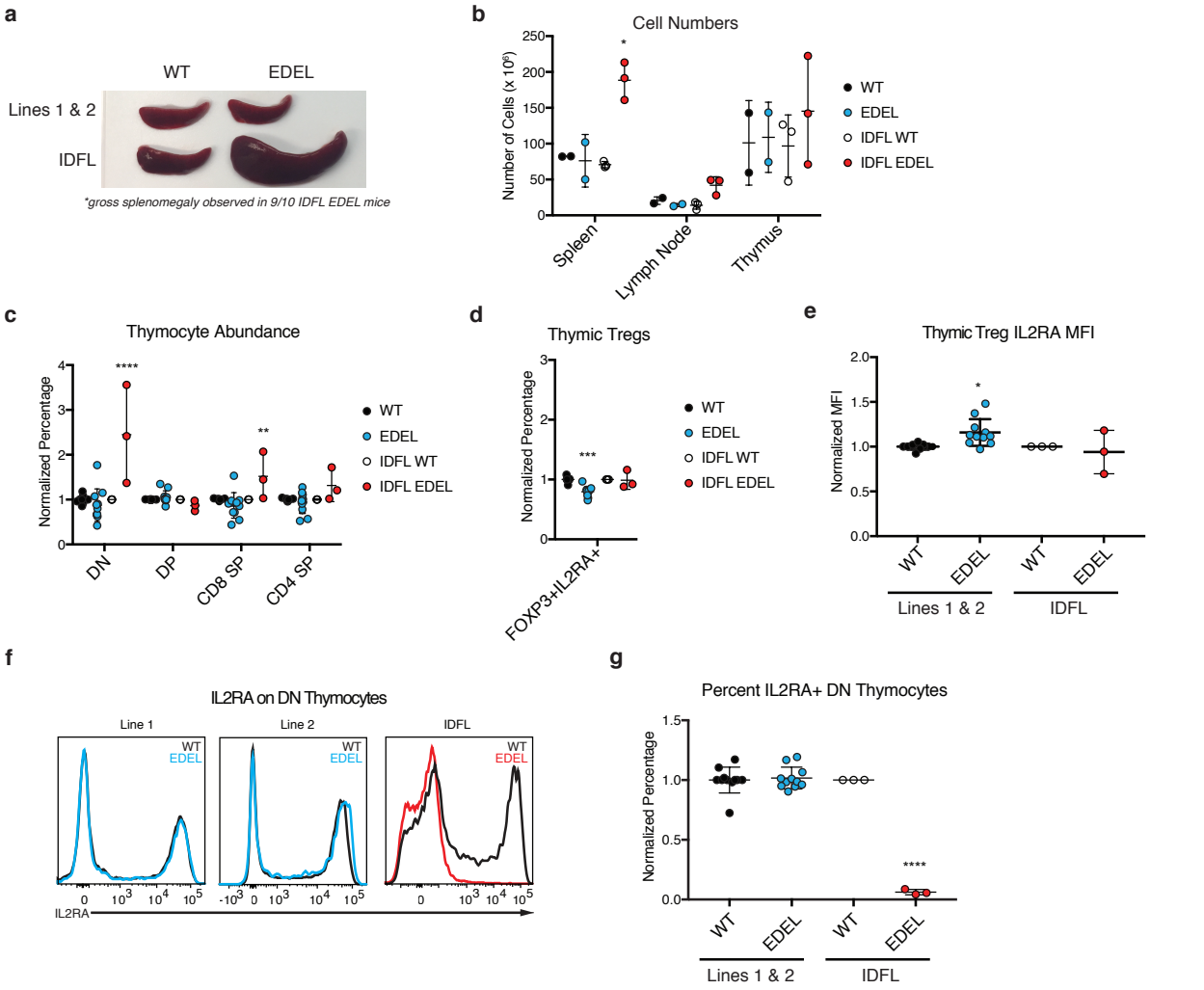


Figure 3.5: Characterization of Enhancer Deletion Founder Lines. a, Representative spleens from WT and EDEL mice derived from different founder lines. b, Cell counts from spleen, inguinal lymph nodes and thymus. c, Normalized percentage of double negative (DN), double positive (DP), CD8 SP and CD4 SP thymocytes of live CD45+ thymocytes (Lines 1 and 2: WT n=11, EDEL n=11; IDFL: WT n=3, EDEL n=3). d, Normalized percentage of FOXP3+/IL2RA+ mature regulatory T cells (Tregs) of CD4 SP thymocytes (Lines 1 and 2: WT n=11, EDEL n=11; IDFL: WT n=3, EDEL n=3). e, Normalized IL2RA MFI on mature FOXP3+/IL2RA+ Tregs (Lines 1 and 2: WT n=11, EDEL n=11; IDFL: WT n=3, EDEL n=3). f, Representative IL2RA surface expression on DN thymocytes from different founders. g, Normalized percentage of IL2RA+ DN cells of live CD45+ thymocytes.

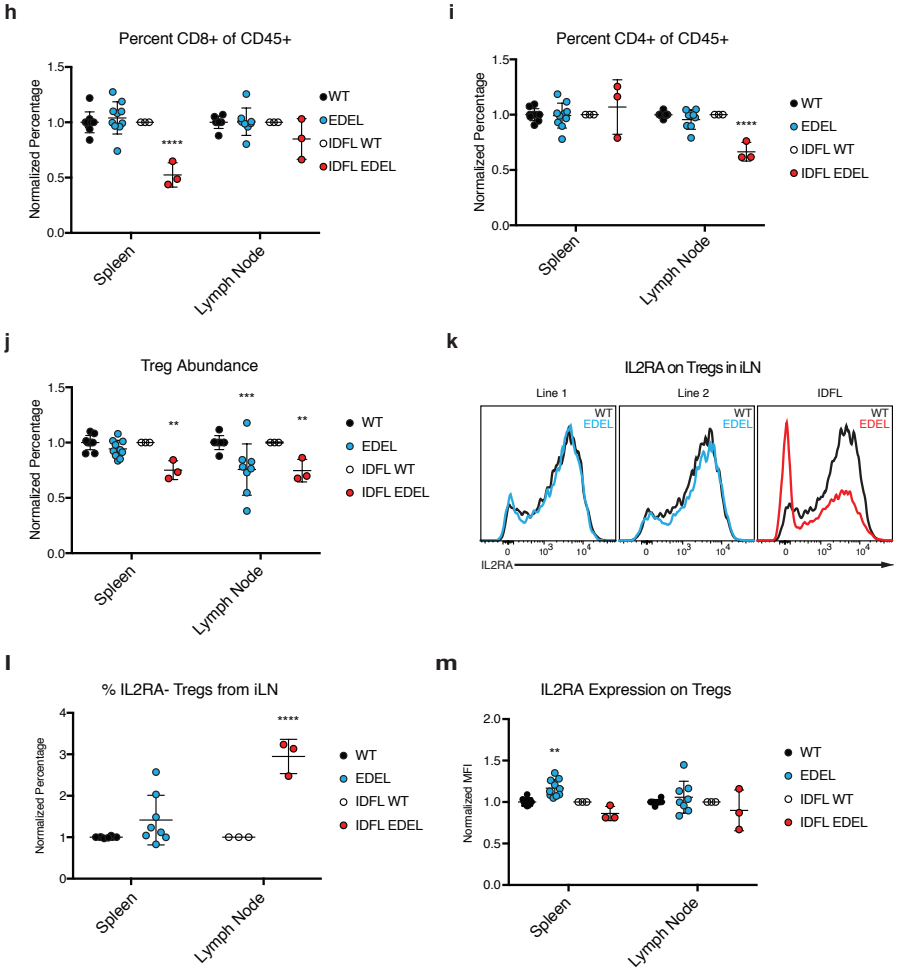


Figure 3.6: Characterization of Enhancer Deletion Founder Lines. h, Normalized CD8+ percentage of live CD45+ cells in spleen and inguinal lymph nodes. i, Normalized CD4+ percentage of live CD45+ cells in peripheral lymphoid organs. j, Normalized percentage of FOXP3+ Tregs of CD4+ T cells in peripheral lymphoid organs. k, Representative IL2RA surface expression of FOXP3+ Tregs. l, Normalized percentage of IL2RA-/FOXP3+ Tregs in peripheral lymphoid organs. m, Normalized IL2RA MFI on FOXP3+/IL2RA+ Tregs in peripheral lymphoid organs. Data for Lines 1 and 2 includes animals for which immunophenotyping was previously published [71]

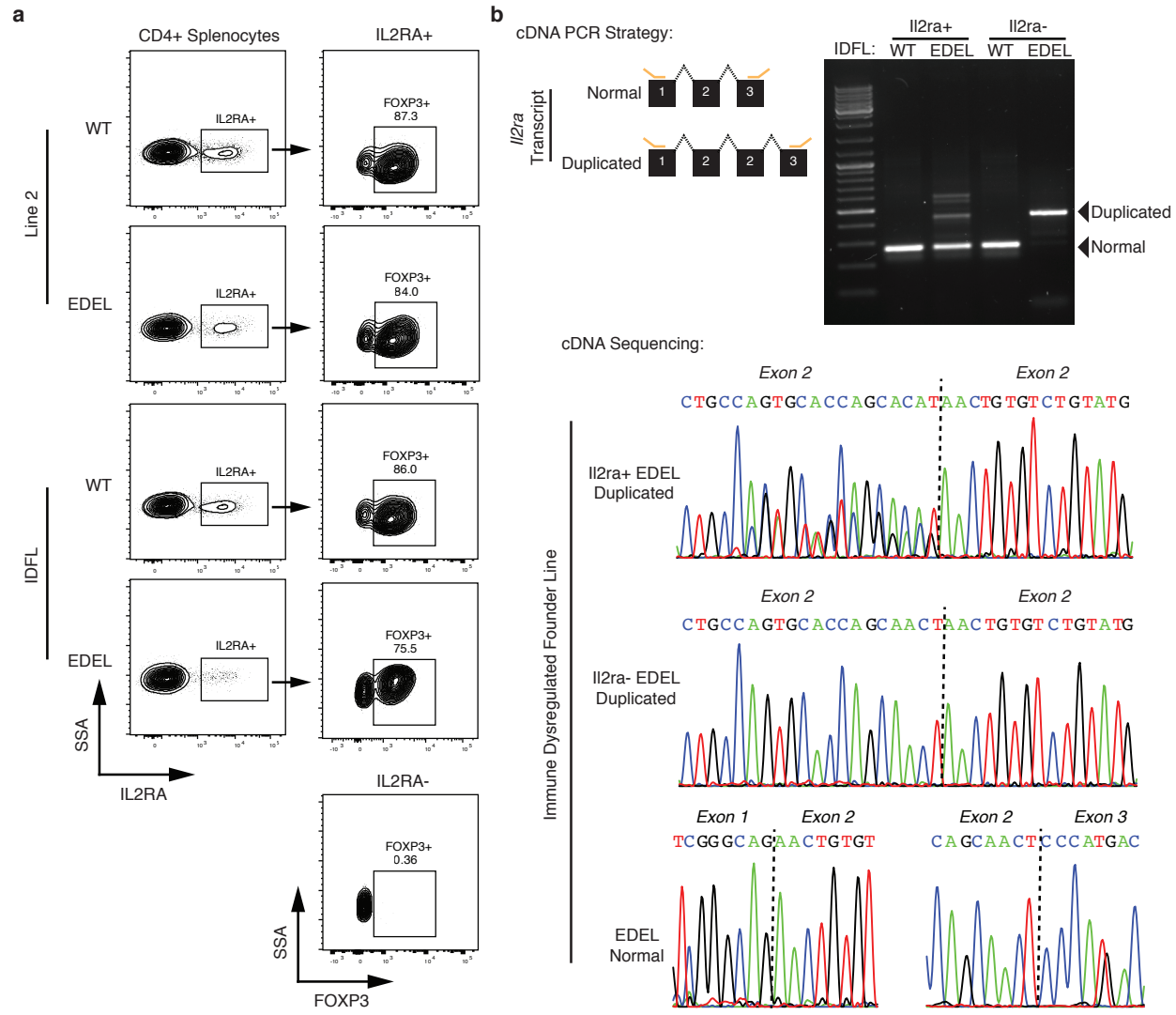


Figure 3.7: Il2ra Splicing Analysis. a, Gating scheme used to sort IL2RA+ and IL2RA- cells. Foxp3 staining in sorted IL2RA+ cells. b, Il2ra exon 2 splicing analysis on IL2RA+ and stimulated IL2RA-/CD44- cells in WT and EDEL CD4+ T cells from the immune dysregulated founder line. PCR amplicons across exon 2 were generated from cDNA and run on a 1% agarose gel. The lower ("normal") band is consistent a single exon 2 Il2ra isoform, whereas the higher ("duplicated") band is consistent with an Il2ra isoform with two exon 2s. The amplicons were isolated and Sanger sequenced for verification. We observed that IL2RA+ IDFL EDEL cells showed variable larger amplicons of unknown significance. Sequencing these larger amplicons showed Il2ra transcript with two exon 2s, but did not explain the variable sizes observed. In these cells, we noted mixed Sanger sequencing peaks adjacent to the exon2-exon2 junction potentially consistent with an aberrant splice isoform or isoforms for which we do not know the significance.

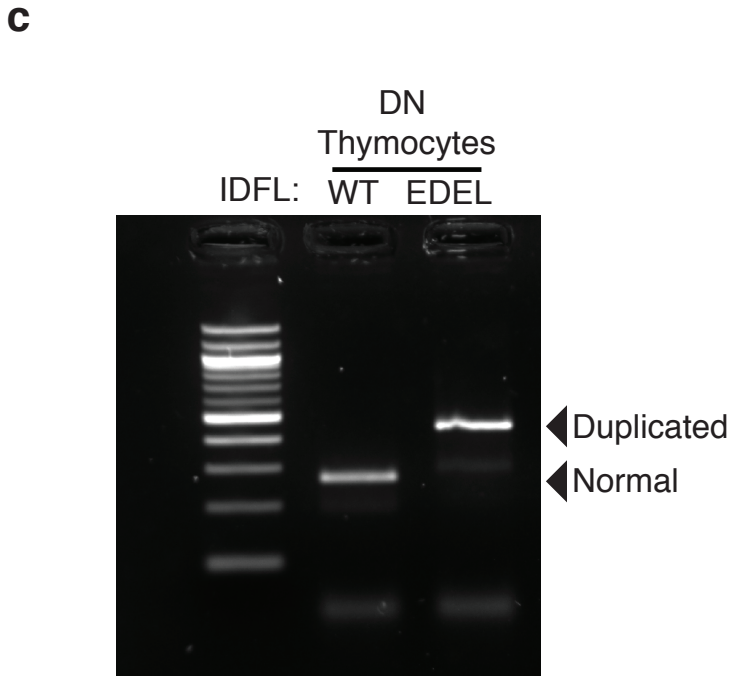


Figure 3.8: Il2ra Splicing Analysis. c, Il2ra exon 2 splicing analysis of IDFL WT and EDEL double negative (DN) thymocytes. The amplicons were isolated and Sanger sequenced for verification.We observed variable smaller amplicons of unknown significance. Sequencing these amplicons showed Il2ra transcript with normal splicing (WT) or two exon 2s (EDEL), but did not explain the variable sizes observed.

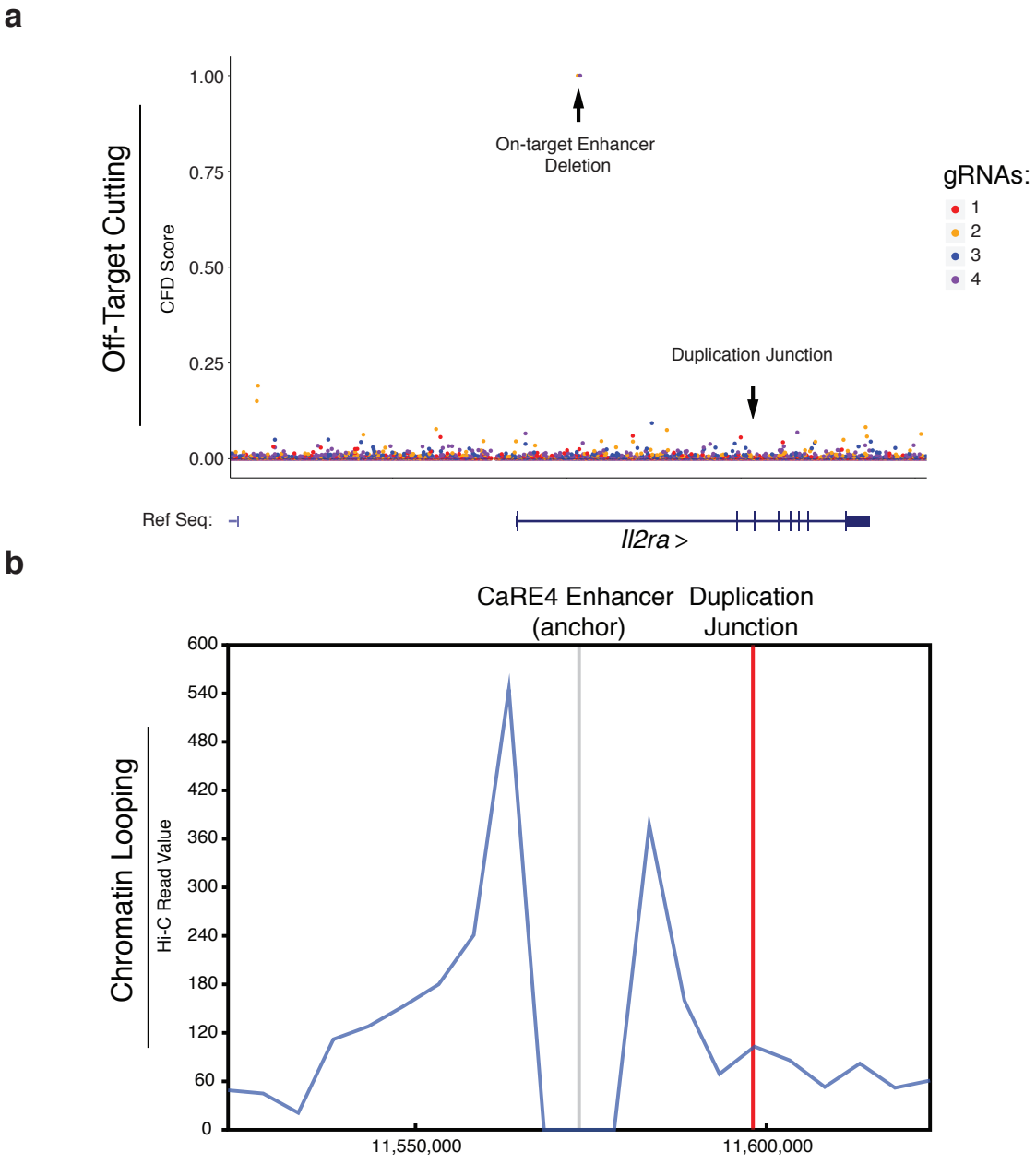


Figure 3.9: Characterization of Il2ra bystander duplication. a, Computational prediction of off-target sites for the Il2ra enhancer gRNAs throughout the Il2ra gene body assessed by Continuous Frequency Determination (CFD) scoring [67]. We highlight the position of the duplication breakpoint. b, Overlap of the Il2ra locus with 5kb-resolution Hi-C data from mouse embryonic stem cells anchored at the Il2ra enhancer [72]

c

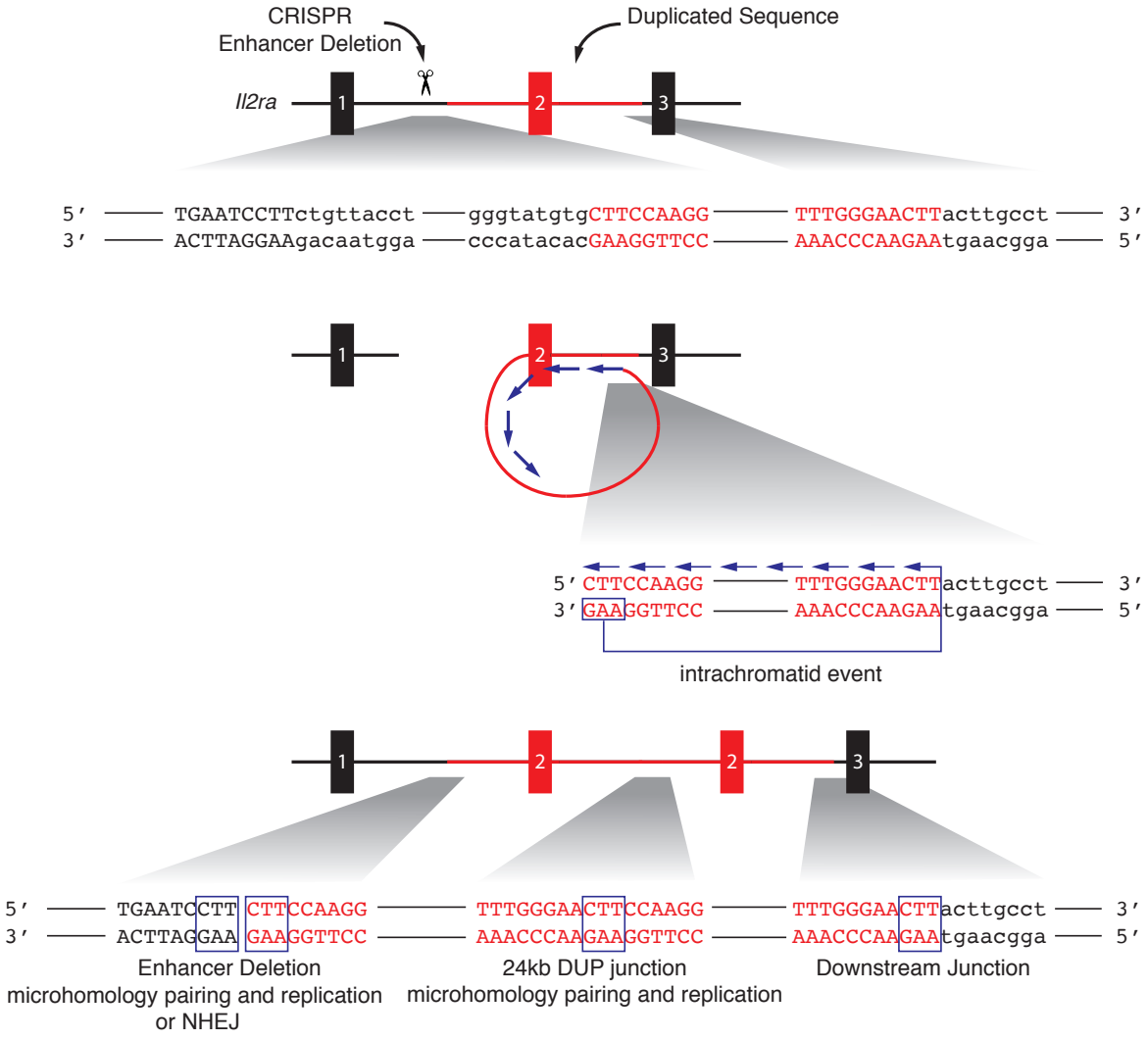


Figure 3.10: Characterization of Il2ra bystander duplication. c, Schematic of proposed microhomology-mediated repair at the Il2ra locus that could generate the observed duplication.

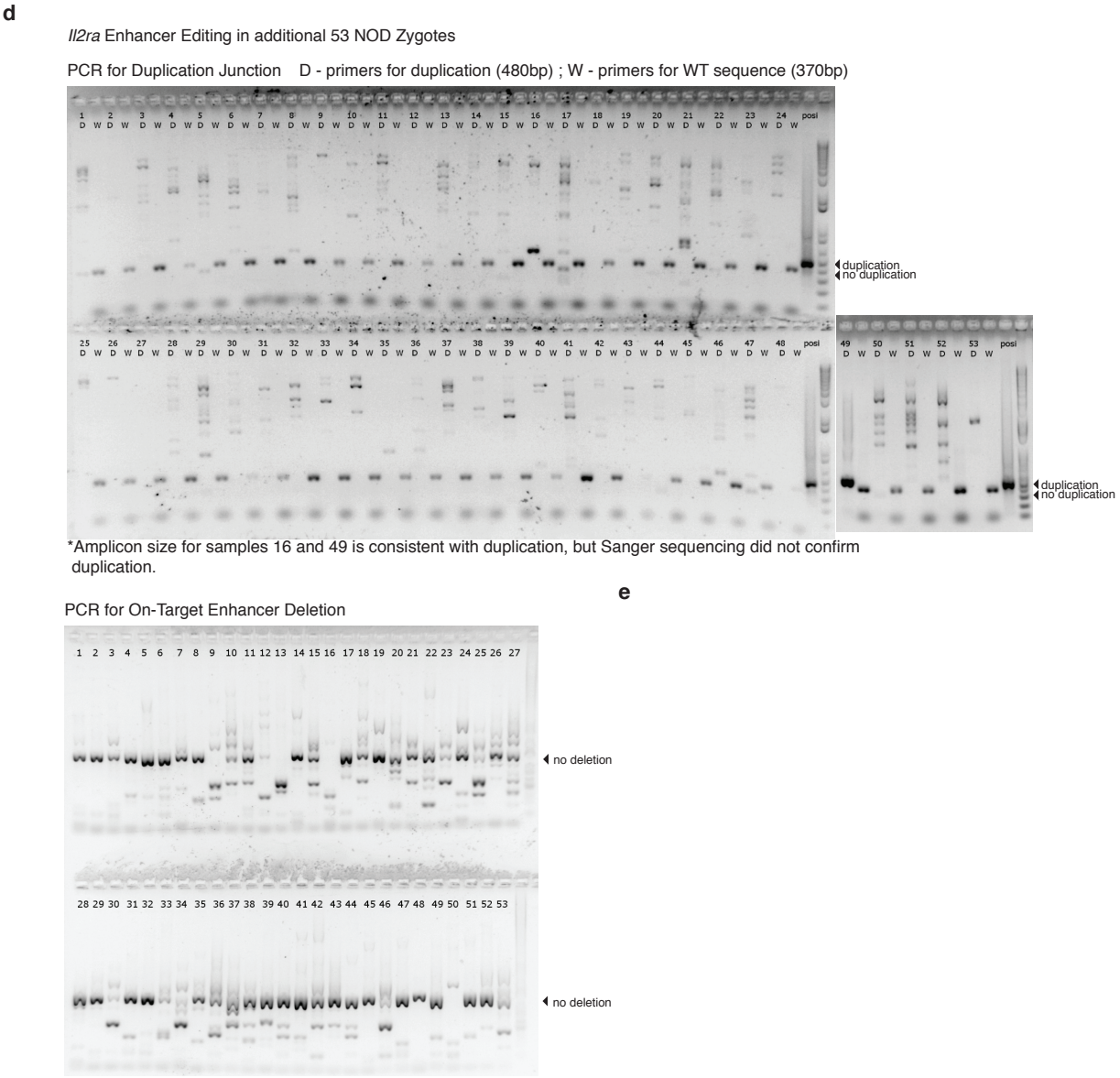


Figure 3.11: Characterization of Il2ra bystander duplication. d, PCR genotyping for Il2ra enhancer editing experiment in additional 53 microinjected NOD zygotes. Top, PCR results for the duplication breakpoint. Amplification was carried out for the duplication junction (D) or WT sequence (W). Bottom, PCR results for the on-target enhancer deletion.

3.4 Methods

Mouse generation

Details of mouse generation were previously reported [71]. Briefly, NOD/ShiLtJ enhancer deletion mice were generated by the Jackson Laboratory (Bar Harbor, ME, USA) by microinjection of gRNA and Cas9 mRNA. Four founders with the enhancer deletion were identified by PCR amplicon size and confirmed by sequencing of TOPO-cloned PCR products. We immunophenotyped three founders. The EDEL mouse lines were established by backcrossing founders for at least one generation before breeding to homozygosity. Protospacer sequence in gRNAs used for the production of the founder lines used in this study can be found in table 3.2.

Mouse genotyping

Enhancer Deletion. All founders were initially genotyped by Sanger sequencing genomic DNA from proteinase K digested tail tissue. PCR amplification of the CaRE4 enhancer was carried out using HotStart Taq (Bioline USA Inc.) and primers (mIl2ra-EDEL-F, mIl2ra-EDEL-R) that span the edited site. PCR amplicons were then sequenced with the mIl2ra-EDEL-F primer. The primers used are in table 3.3. Duplication Junction. The duplication junction was confirmed by PCR amplification of the junction followed by gel electrophoresis and Sanger sequencing. The primers used are in table 3.3.

Genome sequencing

Sample Preparation and Sequencing. DNA was isolated from kidney tissue by phenol-chloroform extraction. PCR-free whole genome libraries were constructed by the Genome Technologies Core at The Jackson Laboratory using the KAPA Hyper Prep Kit (KAPA Biosystems), targeting an insert size of 400 base pairs. Libraries were checked for quality and concentration using the Bioanalyzer High Sensitivity DNA Assay (Agilent), Qubit dsDNA BR Assay (ThermoFisher), and quantitative PCR (KAPA Biosystems), according to the manufacturers' instructions. Libraries were sequenced at Novogene, 150 base pairs paired-end on the HiSeq X (Illumina) to a target mean coverage depth of 30X. Alignment. Bwa mem was used for alignment to the mouse mm10 reference sequence. Reads at the identified tandem duplication junctions were then assembled using Velvet [73] in order to confirm the exact base pair sequence as well as assist in picking primers for confirmation of the duplication via PCR. Variant Calls. The Picard Software Suite and GATK 4.0 pipeline [74] with default settings was used for variant analysis. Base recalibration was performed with NOD specific variants (both SNPs and INDELs) obtained from the Wellcome Sanger Institute Mouse Genome Project.

Off-Target Analysis. We first performed a biased off-target analysis looking for variants 5bp on either side of predicted off-target cut sites. This was done for the top 49 predicted

off-target sites for each of four gRNAs that were used to make the enhancer deletion lines. In total 6 variants were found with this analysis, all of which were present in the NOD background variant panel. We also performed an unbiased variant analysis to examine potentially confounding mutations that fit a likely inheritance model. The cohort variant call file was subset for biallelic SNPs and INDELs where all individuals were assigned a genotype with a sufficient average coverage ($>$ 10 reads). Variants that were unique or in excess in the immune dysregulated mouse as compared to the other two mice in the cohort. The resulting alleles were further subset by removing NOD specific SNP and INDEL variants and selecting only variants that fell within exonic regions. This revealed 2407 variants. The remaining variants are likely specific to the NOD mice used to generate our founder lines. Other than the 24kb duplication, we found no evidence for coding mutations near the enhancer deletion that might contribute to the observed phenotypes. Identifiers GT_05102 and GT_05105 refer to heterozygous and homozygous EDEL mice from the IDFL founder line. GT_05111 is homozygous EDEL mouse from Line 2 (Table 3.1).

Off-target analysis

We first performed a biased off-target analysis looking for variants 5bp on either side of predicted off-target cut sites. This was done for the top 49 predicted off-target sites for each of four gRNAs that were used to make the enhancer deletion lines. In total 6 variants were found with this analysis, all of which were present in the NOD background variant panel. We also performed an unbiased variant analysis to examine potentially confounding mutations that fit a likely inheritance model. The cohort variant call file was subset for biallelic SNPs and INDELs where all individuals were assigned a genotype with a sufficient average coverage ($>$ 10 reads). Variants that were unique or in excess in the immune dysregulated mouse as compared to the other two mice in the cohort. The resulting alleles were further subset by removing NOD specific SNP and INDEL variants and selecting only variants that fell within exonic regions. This revealed 2407 variants. The remaining variants are likely specific to the NOD mice used to generate our founder lines. Other than the 24kb duplication, we found no evidence for coding mutations near the enhancer deletion that might contribute to the observed phenotypes. Identifiers GT_05102 and GT_05105 refer to heterozygous and homozygous EDEL mice from the IDFL founder line. GT_05111 is homozygous EDEL mouse from Line 2 (Table 3.1).

RNA sequencing

Briefly, approximately 500,000 CD4+/IL2RA+ cells were sorted and total RNA was isolated from samples using the RNeasy Micro Kit (QIAGEN) according to the manufacturer's instructions with the following options: cells were pelleted and re-suspended in RLT buffer with β-mercaptoethanol and homogenized using QIAshredder (QIAGEN). DNA removal was performed with gDNA Eliminator Columns (QIAGEN). RNA samples were analyzed with a NanoDrop spectrophotometer and all samples had a 260/280 ratio of 1.80 or higher. RNA

integrity was measured by Bioanalyzer and all samples had an RNA Integrity score (RIN) of 8.0 or more. RNA-seq libraries were prepared by the Functional Genomics Laboratory at Berkeley. RNA samples were poly-A selected and then converted into sequencing libraries with the ultra-low input SMART-seq kit. The samples were pooled and sequenced on one lane of the Illumina HiSeq4000. Il2ra isoform analysis was done using the UNIX grep command to identify reads in raw fastq that contained sequences for the E2-E2 junction (ACCAGCAACTAACTGTGTCT) or E2-E3 junction (ACCAGCAACTCCCATGACAA). Read counts were normalized to the total number of reads for a given sample. Short reads were also aligned with STAR to the mouse mm10 reference. Differential expression analysis was performed using EdgeR from Bioconductor Package for R (9). Pairwise comparisons (quantile-adjusted conditional maximum likelihood) were performed between wild type (WT) and IDFL EDEL or between WT and EDEL RNA-seq samples, after filtering out lowly expressed genes.

Il2ra cDNA isoform analysis

500,000 IL2RA+ (CD4+/IL2RA+) and IL2RA- (CD4+/IL2RA+/CD44-) CD4+ T cells were sorted from CD4 enriched splenocytes. IL2RA-/CD4+ T cells were stimulated *in vitro* for 10hrs with 2 μg/ml plate bound anti-CD3/CD28 antibodies (Biolegend). RNA was extracted using the RNA Micro Kit (QIAGEN) as described above. RNA was reverse transcribed into cDNA using SuperScript VILO MasterMix as per manufacturer's protocol (Thermo Scientific). PCR of the cDNA to assess exon 2 splicing was performed with forward and reverse primers that sit in Il2ra exon 1 and exon 3, respectively (Supplementary Table 1). PCR was carried out with Bioline Taq 2x MasterMix as per manufacturer's protocol. Amplicons were cut out of the agarose gel and purified using QIAGEN's Gel Extraction Kit. Amplicons were sequenced in the forward and reverse directions using the primers from the initial PCR amplification.

Zygote Il2ra enhancer editing

NOD/ShiLtJ zygotes were microinjected with gRNAs and Cas9 mRNA, identical to the generation of Il2ra EDEL mice. PCR amplification and Sanger sequencing were used to check gDNA for the duplication junction and WT Il2ra sequence upstream of Il2ra exon 3, the genomic site of the IDFL duplication junction (Supplementary Fig. 5). Blastocyst gDNA was also checked for on-target enhancer deletion by PCR amplification and Sanger sequencing (Supplementary Fig. 5). A second test for the duplication junction was performed using a nested PCR. Briefly, duplication junction PCR samples were diluted (1 μl sample in 10 μl water) and 1μl of dilution was used as the template. A non-specific band was observed in a majority of the blastocysts (data not shown). Nine amplicons were extracted and sent for Sanger sequencing, of which six were successfully sequenced. The primers used for these assays are listed in table 3.3. Takara PrimeSTAR polymerase was used for amplification with 30 second extension for 35 cycles. PCR products were run on agarose gel for size separation and visualization.

Mechanism of duplication formation

By scoring the cutting frequency determination (CFD score) of all possible CRISPR-Cas9 cut sites in the IL2RA locus for the 4 gRNAs. I've shown that no credible off-target cut sites exist (Figure 3.9). In collaboration with James Lupski's group we have analyzed the duplication and identified microhomology at the breakpoints, which could mediate repair. We also include chromatin conformation data showing looping between the site of the targeted enhancer and the duplication breakpoint. Taken together, these analyses suggest a mechanism for the formation of the bystander tandem duplication we identified in our mice. Furthermore, we checked the downstream site in the genome sequencing data from the different lines. We find no INDELs in either IDFL or non-IDFL mice consistent with no Cas9 off-target activity at the downstream site and suggesting that the duplication was formed as an unintended repair consequence of the on-target deletion. Although we did not identify long stretches of homology between the enhancer and the downstream breakpoint that could have explained the duplication, we do find trinucleotide microhomology proximal to the junction sequences that could potentially explain the breakpoint. We also include high resolution Hi-C data that remarkably shows chromatin looping between the two sites. We suspect chromatin looping to be involved in bringing the microhomologies into proximity to one another (Figures 3.9b,c).

3.5 Tables

gRNA Name	Targeting Sequence
mIl2ra-EDEL-up1	TGCTCTTTGAAGGTAACAGA
mIl2ra-EDEL-up2	GTTACCTTCAAAGAGCAGCC
mIl2ra-EDEL-down1	AAGATGGGTATGTGCTTCCA
mIl2ra-EDEL-down2	AGATGGGTATGTGCTTCCAA

Table 3.2: CRISPR gRNA Sequences

Primer Name	Sequence	Purpose
mIl2ra-EDEL-F	TCCTCAGGACCCTGCTAGTC	gDNA PCR to genotype enhancer deletion
mIl2ra-EDEL-R	GAGAAGCAAAGCAGCAGACA	gDNA PCR to genotype enhancer deletion
Dup_jx_F	CTGAGAAGCAAAGCAGCAGA	gDNA PCR to confirm duplication junction
Dup_jx_R	TGGCTGATGGCTAAGGGATA	gDNA PCR to confirm duplication junction
mIl2ra-cDNA-E1-F	CCAGTTGTCGGGCAGAAC	mIl2ra cDNA PCR to check splicing of exon 2
mIl2ra-cDNA-E3-R	TGCATGTCTGTTGTGGTTTG	mIl2ra cDNA PCR to check splicing of exon 2
7720-NOD-comF	CCGGGATTTAAGCTCATTCA	zygote editing experiment, duplication genotyping
7721-NOD-dupR	GTGGAGTGTGTGTCCACCAG	zygote editing experiment, duplication genotyping
7722-NOD-wtR	GGCTAGAGGATGGTTGCTGA	zygote editing experiment, duplication genotyping
7764-Il2ra-tgt-f	cctcttgctctcccagacag	zygote editing experiment, enhancer deletion genotyping
7764-Il2ra-tgt-r	aaccttgcgtgaagttgctc	zygote editing experiment, enhancer deletion genotyping
Il2raNstdF	GCCATTTCTCATGCCTGTCT	zygote editing experiment, nested primers for duplication genotyping
Il2raNstdR	CTCAGCCCTTAGCTTGGGTA	zygote editing experiment, nested primers for duplication genotyping

Table 3.3: DNA Primers

Chapter 4

Deeply Conserved Synteny Between Amphioxus and Five Vertebrates[†]

The lancelet amphioxus, an early-branching living chordate, is an invaluable comparative tool for deciphering ancient events in chordate evolution [35]. To this end we reconstructed the $n = 19$ chromosomes of amphioxus (the Florida lancelet, *Branchiostoma floridae*) by combining *in vitro* [75] and *in vivo* [76] chromatin conformation capture sequencing with an improved assembly of previously generated whole genome shotgun data [43]. The long-range accuracy of this chromosome-scale assembly was confirmed by a dense genetic map generated from a biparental cross (detailed in Chapter 3). Comparison with diverse invertebrate draft genomes confirms the deep conservation of synteny between amphioxus and sea anemone, [77] limpet, [78] acorn worm, [79] and sea star (Figure 4.1).

A simple dot plot comparing the chromosomal position of orthologous genes in amphioxus and multiple jawed vertebrates reveals striking patterns of conserved synteny vs. gar and chicken (Figure 4.2, see Figure 4.3 for frog and human). Orthologs of genes on each amphioxus chromosome are clearly restricted to specific chromosomal regions in vertebrates, indicating ancient conserved linkages. For example, orthologs of genes on amphioxus chromosome 1 are largely confined to the first (i.e., lower) half of chicken 3 and the second (i.e., upper) half of chicken 5, with sharp boundaries between ortholog-dense and ortholog-poor regions. Weaker concentrations are found in the middle of chicken 1 and chicken 25. A similar concentration of orthologs on a limited number of gar chromosomes is also found. No other amphioxus chromosomes show the pattern of conserved synteny found for amphioxus 1.

[†]This chapter is based on my contributions to the manuscript "Deeply conserved synteny and the origins of vertebrate chromosomes" by Simakov et al., currently in review at *Nature Evolution & Ecology*.

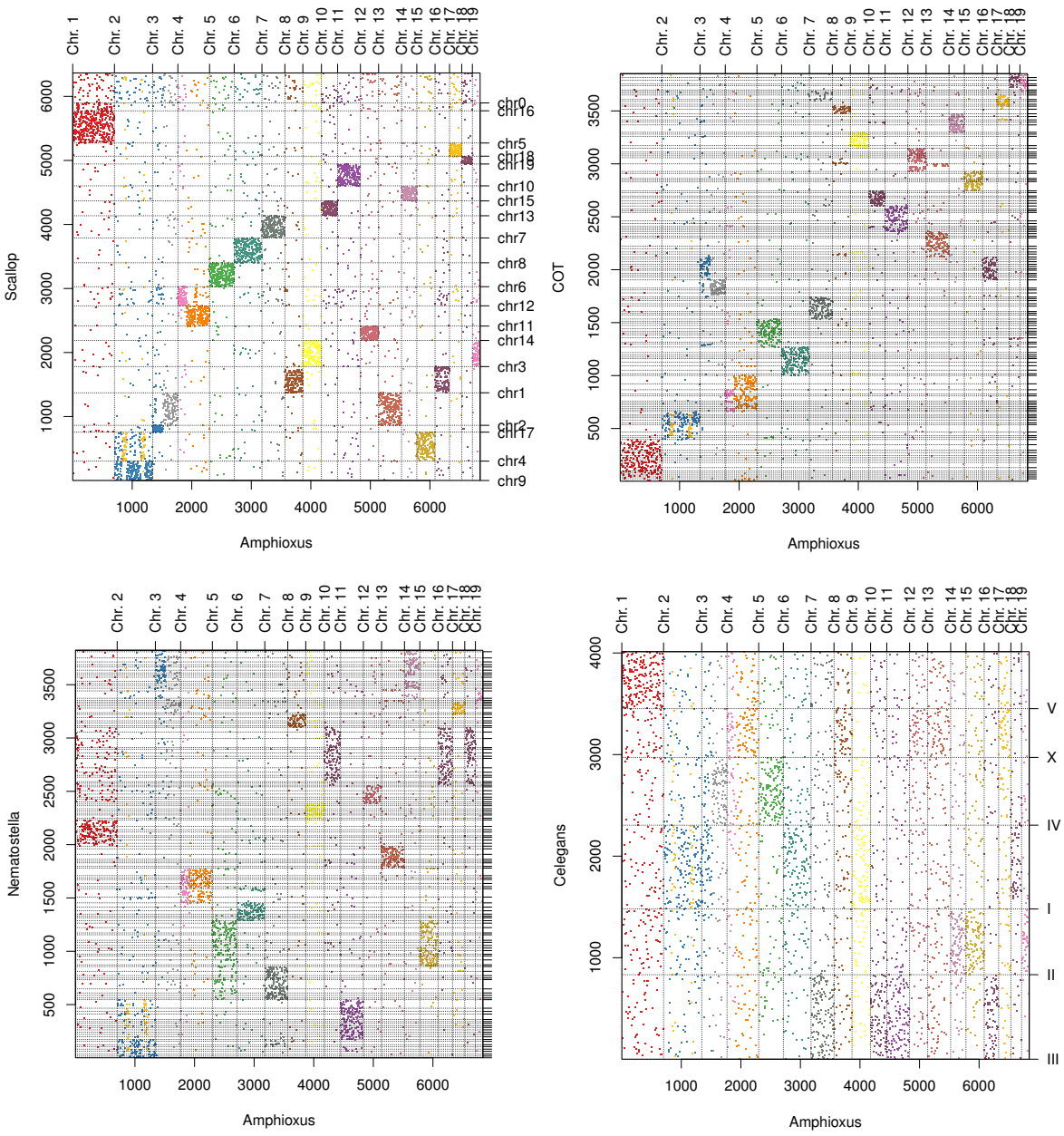


Figure 4.1: Invertebrate mutual best hit pairs

4.1 Mutual best hits (MBH)

After examining several possible methods for pairwise ortholog detection, we found that the simplest method of computing mutual best hits of all genomes to amphioxus provides the least-assumptive and still complete dataset. We required all vertebrate genomes and amphioxus to be represented to yield a mutual best hit group, resulting in 6,151 orthologous groups. It is assumed that in the case of 1:many (e.g., due to whole genome duplications) one of the many paralogs will be chosen for the single amphioxus gene. We then mapped the outgroup genomes (such as *Drosophila*) to the 6,151 groups, yielding 4,334 groups for *Drosophila*, 5,421 for *Lottia*, 3,624 for *C. elegans*, 5,725 for *Saccoglossus*.

4.2 Identifying paralogs in vertebrate MBH clusters

As an additional control for the "conventional" clustering methods, we implemented a simple method to identify paralogs in vertebrate genomes, based on the phylogenetically informed clustering described in Simakov et al., 2013. In the first step, we constructed paralog groups for each of the vertebrate genomes by clustering all sequences that have better BLAST hits to each other than to their nearest amphioxus sequence. We then assigned those paralogous clusters to the MBH clusters based on presence of the MBH cluster sequences in each of the paralogous group. Only cases where all vertebrate sequences from the MBH cluster could be assigned to their corresponding paralogous groups and each of those paralogous groups had the best hit to the same MBH cluster were considered. In total, there were 4,892 such gene families. To complement this approach, we also implemented the previously described gene family clustering method that utilizes hierarchical, phylogenetically-fixed, construction of orthologous groups, as described in Simakov et al. 2013. We compared the dot plot distributions between MBH and both clustering-derived data and found them largely consistent.

4.3 Significance testing of blocks of conserved synteny

Previously, Smith and Keinath (2015) and Smith et al. (2018) have argued that relatively few chromosomal comparisons between sea lamprey and bony vertebrates (e.g., chicken (Smith and Keinath, 2015) and chicken and spotted gar (Smith et al., 2018)) are significantly enriched for shared orthologs when compared with a null model, leading to their rejection of the "2R" hypothesis and development of a model in which jawed vertebrate chromosomes arose through a combination of individual chromosome-scale duplications preceding a single genome-wide event.

From our main Figure 4.2, however, it is evident that, especially for chicken and spotted gar macro-chromosomes, orthologs are enriched across sharply defined segments of the larger vertebrate chromosomes. Significance tests based on chromosome-chromosome comparisons

could therefore be under-powered, since enrichments confined to a portion of a large chromosome will be diluted when considered on a chromosome scale. Smith-and-Keinath-style analyses may also be underpowered due to the use of lamprey chromosomes as the units of comparison, as the lamprey orthologs of jawed-vertebrate genes appear to be distributed over multiple lamprey chromosomes (Figures 4.7 and Figures 4.8).

To test for segmental enrichments taking into account the apparent structure of vertebrate chromosomes relative to the chordate ancestor, we used sliding windows of m = 50 (or $<$ 100) genes across vertebrate chromosomes. The number of windows per chromosome is $[2g/m + 1/2]$, where the sliding distance is determined so that the last window aligns with the last gene in a chromosome. Importantly, the boundaries of the tested windows are chosen without regard to the boundaries seen in Figure 4.2, to avoid biasing the calculation.

Unlike Smith and Keinath, we consider only 6,823 mutual best hits between amphioxus and chicken/gar. We note that the use of mutual best hits may reduce the power of our calculation, but significance found using mutual-best-hits is an upper bound on significance calculated using more complex definitions of ortholog.

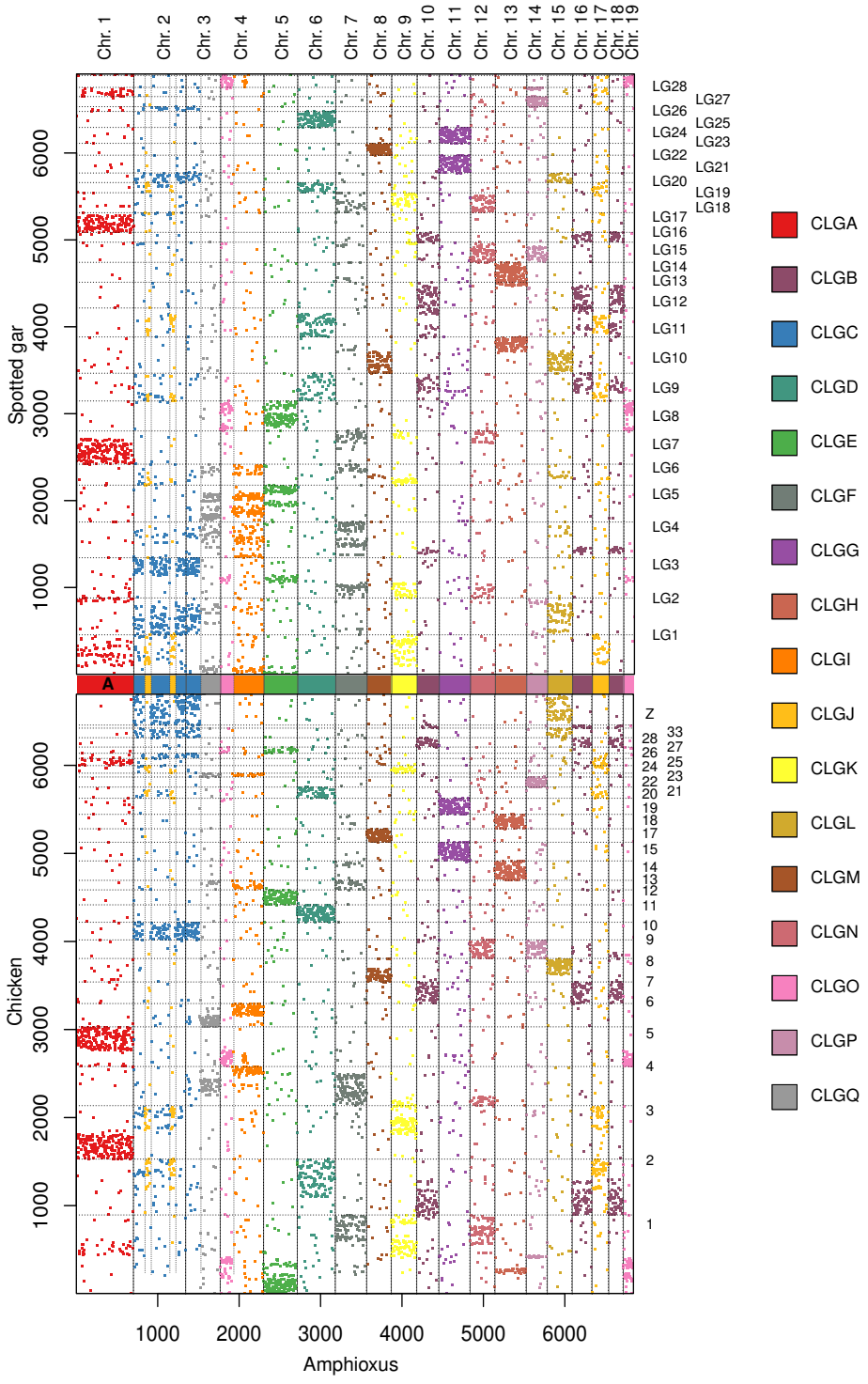


Figure 4.2: Chicken and gar mutual best hit pairs

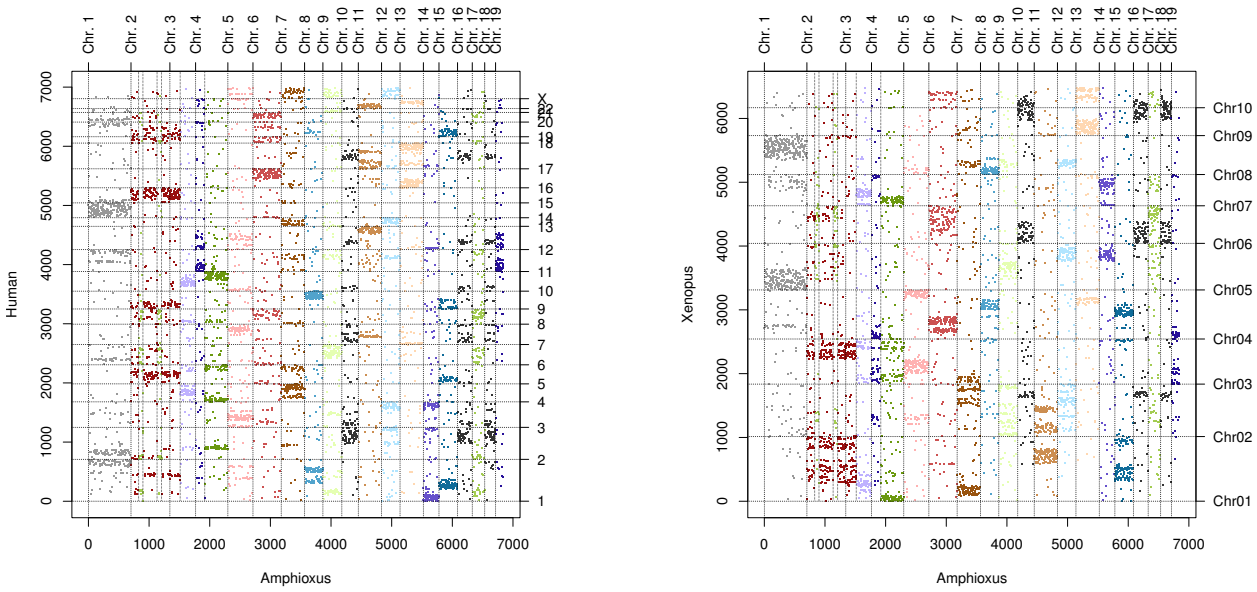


Figure 4.3: Human and frog mutual best hit pairs

An advantage of using mutual best hits is that these are unique in both amphioxus and the comparator vertebrate genome, so the relevant distribution of shared orthologs is given by the hypergeometric distribution. For any given window i of some predefined length, and a chordate linkage group j, the number of genes found within these two regions, relative to those found outside these regions can be computed using the one-tailed test for the hypergeometric distribution. This is calculated as follows:

$$p_{i,j} = \sum_{k=obs}^{w_s} \frac{\binom{K}{k}\binom{N-K}{w_s-k}}{\binom{N}{w_s}}$$

Where $p_{i,j}$ is the significance for window i and CLG j of window size w_s, with obs mutual best hits, N possible best hits within the ith window of organism, and K possible best hits for the jth CLG. The p-values computed in this manner must be scaled by a Bonferroni correction to account for the total number of windows tested.

Figures 4.4, 4.5, and 4.5 (multiple panels) show the number of shared mutual best hits per window as circles of the appropriate area, with significant windows (Bonferroni-corrected $p < 0.05$) shown in blue or yellow. This analysis clearly shows that all CSGs have three or more significant windows of conserved synteny, contrary to Smith and Keinath's chromosome-chromosome comparisons. This analysis shows that Smith and Keinath's rejection of the 2R scenario based on the scarcity of evidence for three and higher paralogous blocks is flawed. Our further analyses in the main text demonstrate than an "auto-then-allo" 2R model is consistent with comparisons between the amphioxus chromosomes and those of bony vertebrates (chicken, spotted gar, frog, and human).

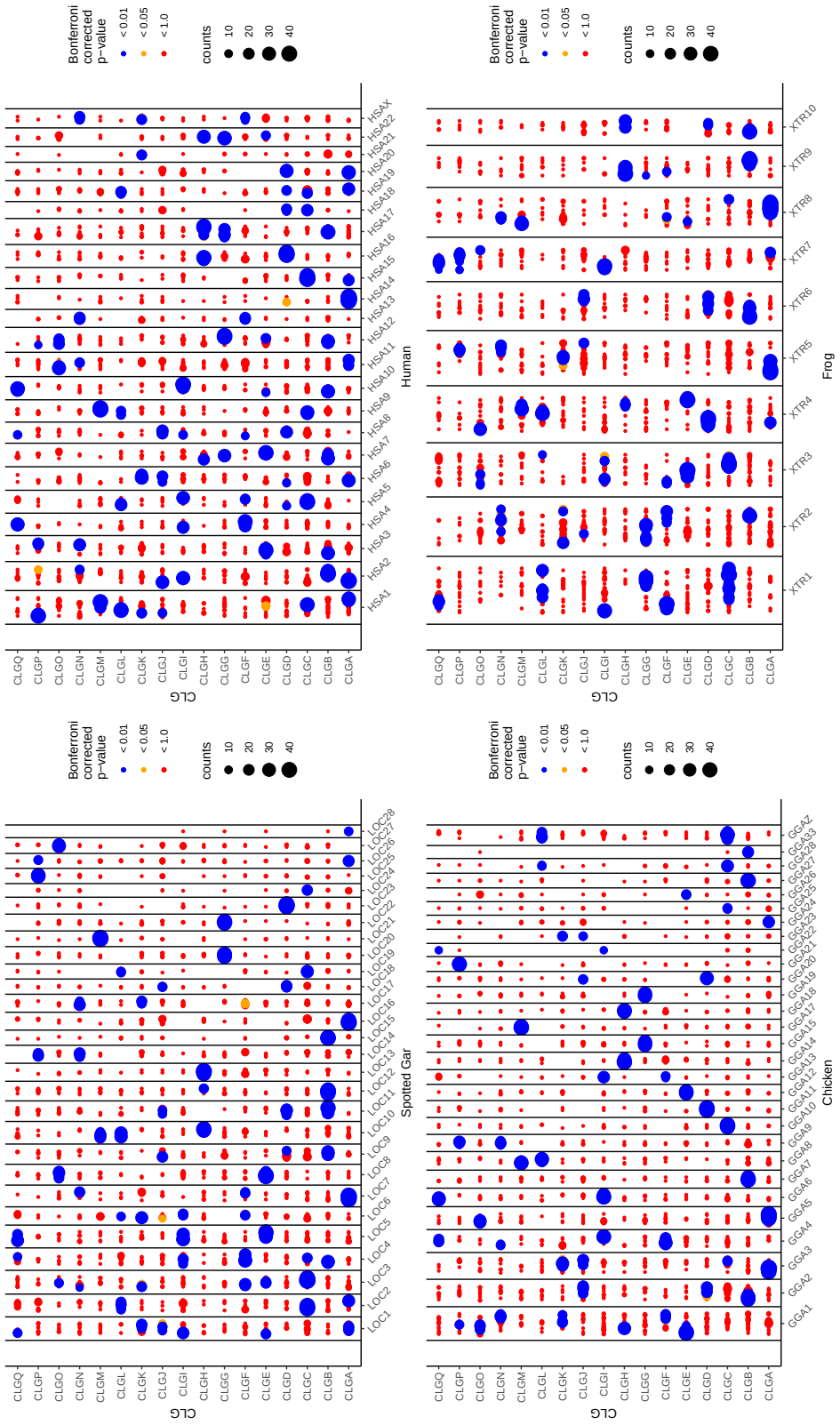


Figure 4.4: 50 gene windows

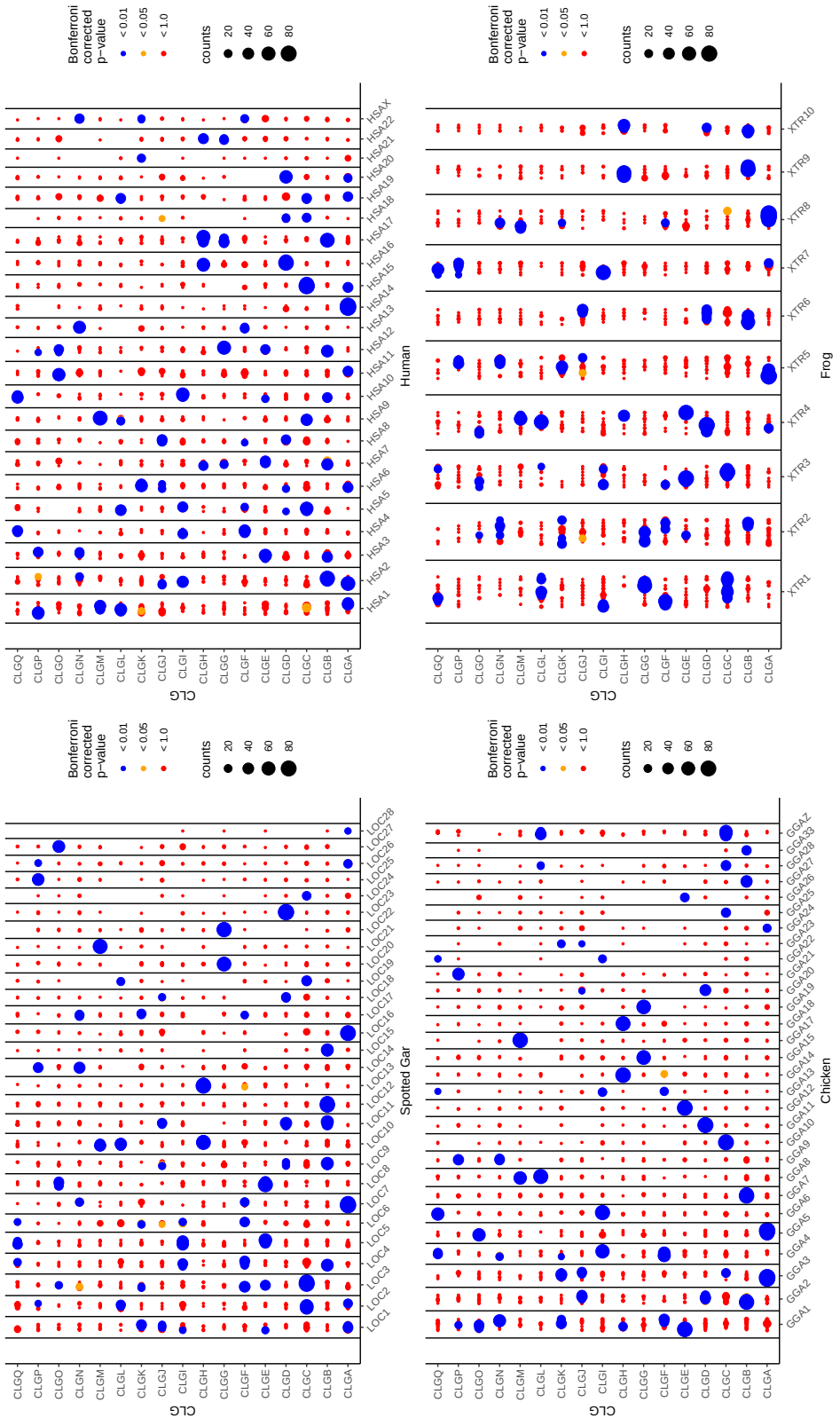


Figure 4.5: 100 gene windows

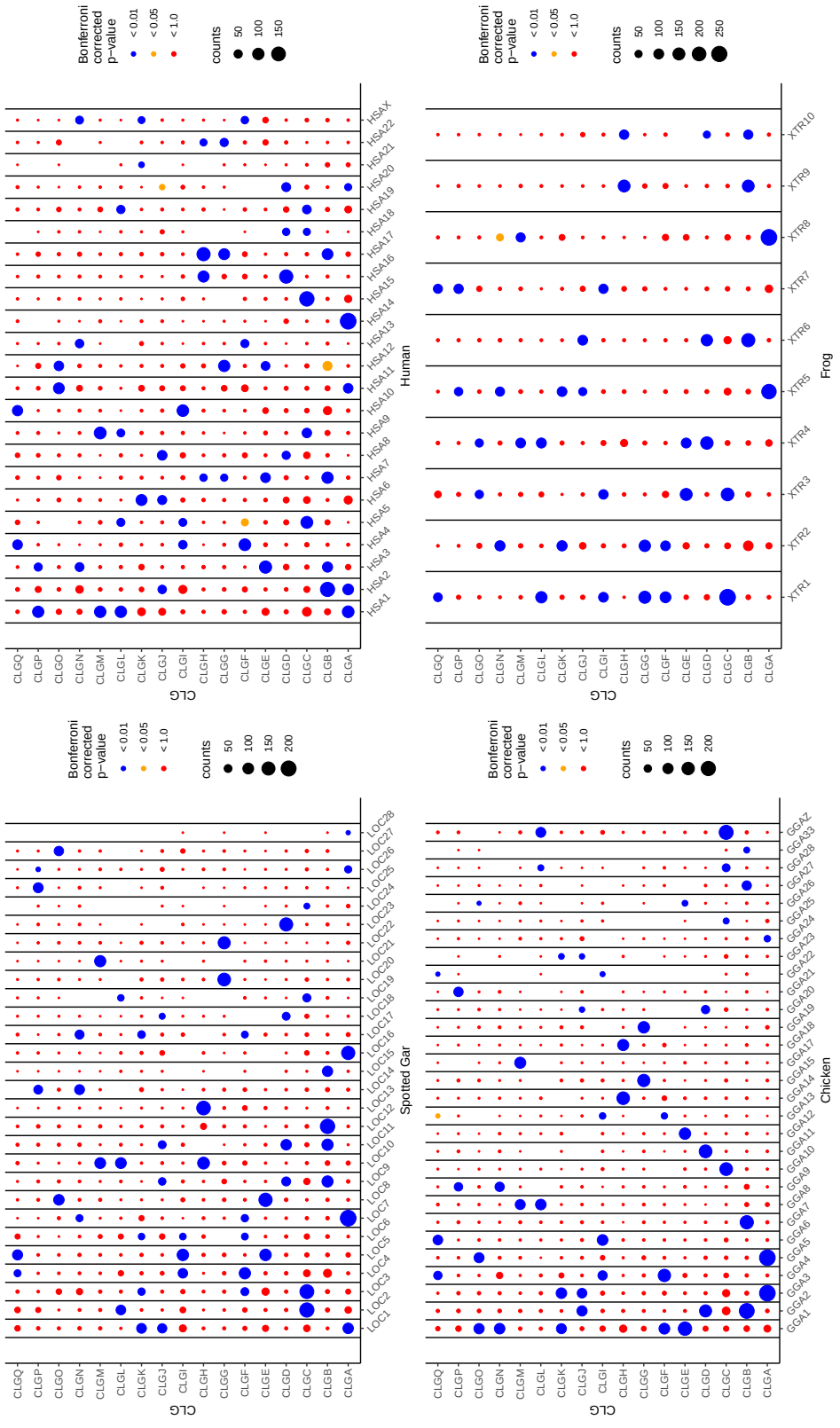


Figure 4.6: Chromosome-sized windows

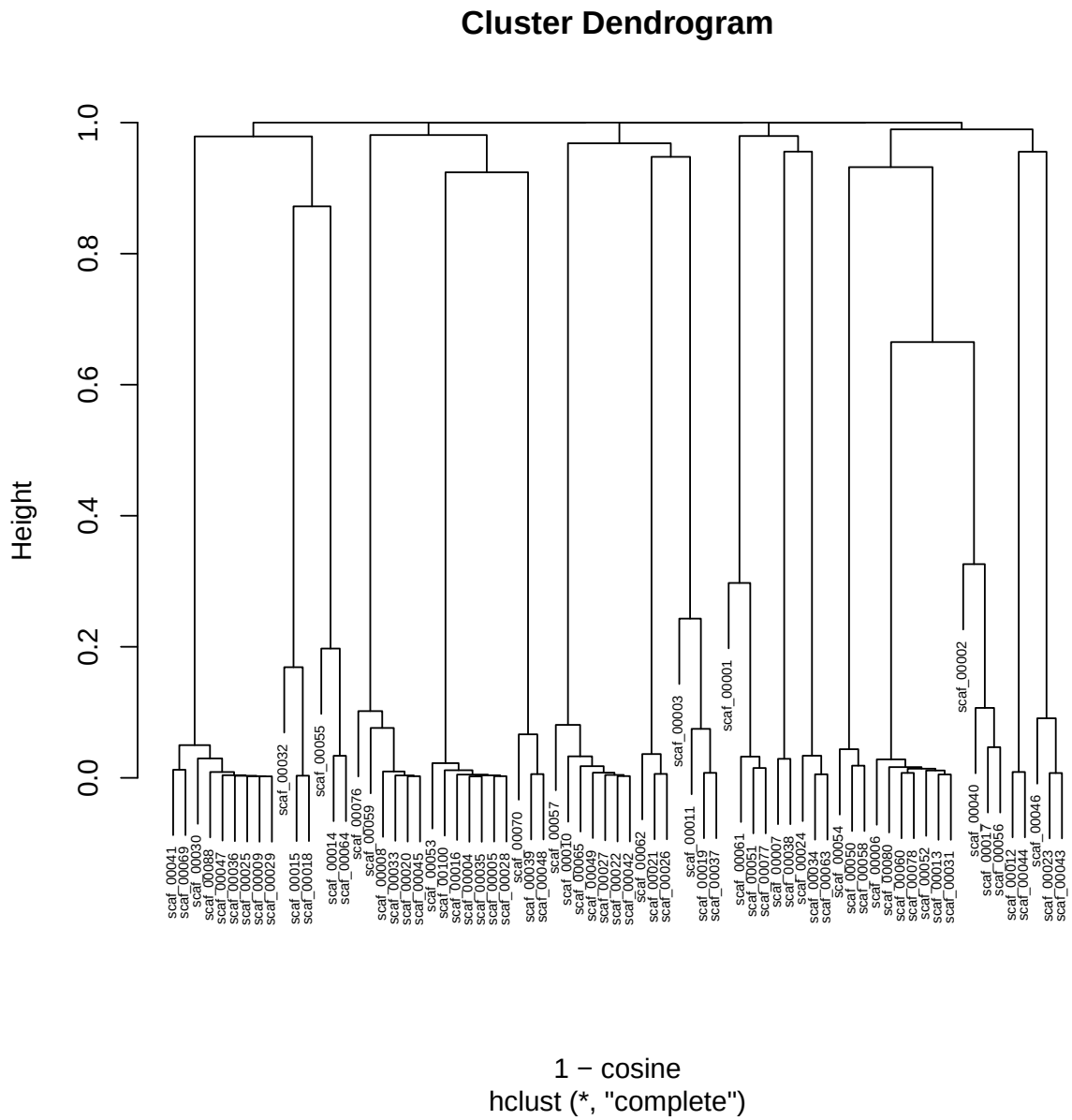


Figure 4.7: Clustering of lamprey chromosomes with cosine dissimilarity distance matrix

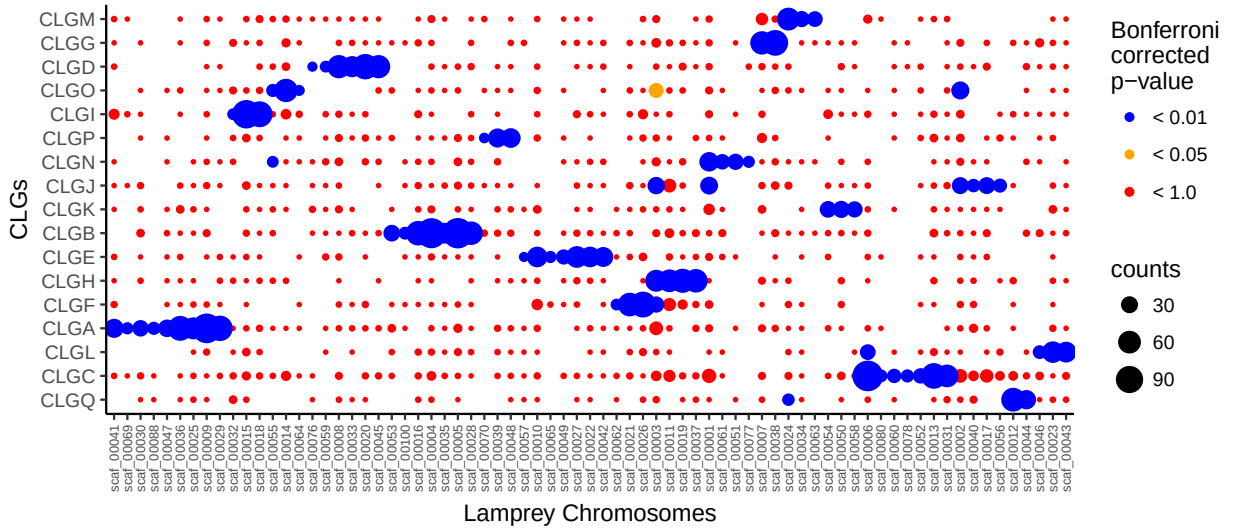


Figure 4.8: Lamprey mutual best hit pairs

4.4 Assignment of amphioxus chromosomes and reconstruction of chordate synteny groups

Based on examination of the dot plots (both MBH and clustering-based) between amphioxus and other species, we could identify consistent breakpoints of conserved synteny on three amphioxus chromosomes (2, 3, and 4). Additionally, several chromosomes show almost full overlapping gene repertoire and can be merged into a single CLG. In total, we can infer 17 CLGs, consistent with the previous estimates in Putnam et al. (2008).

4.5 Auto-then-allotetraploidy

Asymmetric gene loss across chromosomes is a hallmark of allotetraploidy – interspecific hybridization followed by genome doubling – as described for multiple paleotetraploid flowering plants [80] and demonstrated in the African clawed frog *Xenopus laevis* [81]. In contrast to the inherently symmetrical process of autotetraploidy (genome doubling within a single species), allo-tetraploidy brings together a pair of diverged progenitor genomes with distinct epigenetic landscapes, cytonuclear interactions, transposable element histories. An asymmetric response is therefore expected for allotetraploids, but not autotetraploids.

Here we propose a 1R/2R scenario of "auto-then-allotetraploidy" (Figure 4.9). In this model, the first "1R" event was an autotetraploidization. Autotetraploidy is common in fish lineages including salmonids, cyprinids (carps and their relatives), and sturgeons (reviewed in Braasch and Postlethwait 2012 [82]). In salmonids, polysomic inheritance (i.e., ongoing homeologous recombination) has persisted for tens of millions of years [83]. We hypothesize

that autotetraploidy similarly occurred early in the vertebrate lineage, followed by a radiation of tetraploid proto-vertebrates. The merger of two of these already tetraploidy lineages would produce an auto-then-allo-octaploid, leading to asymmetric gene loss as observed.

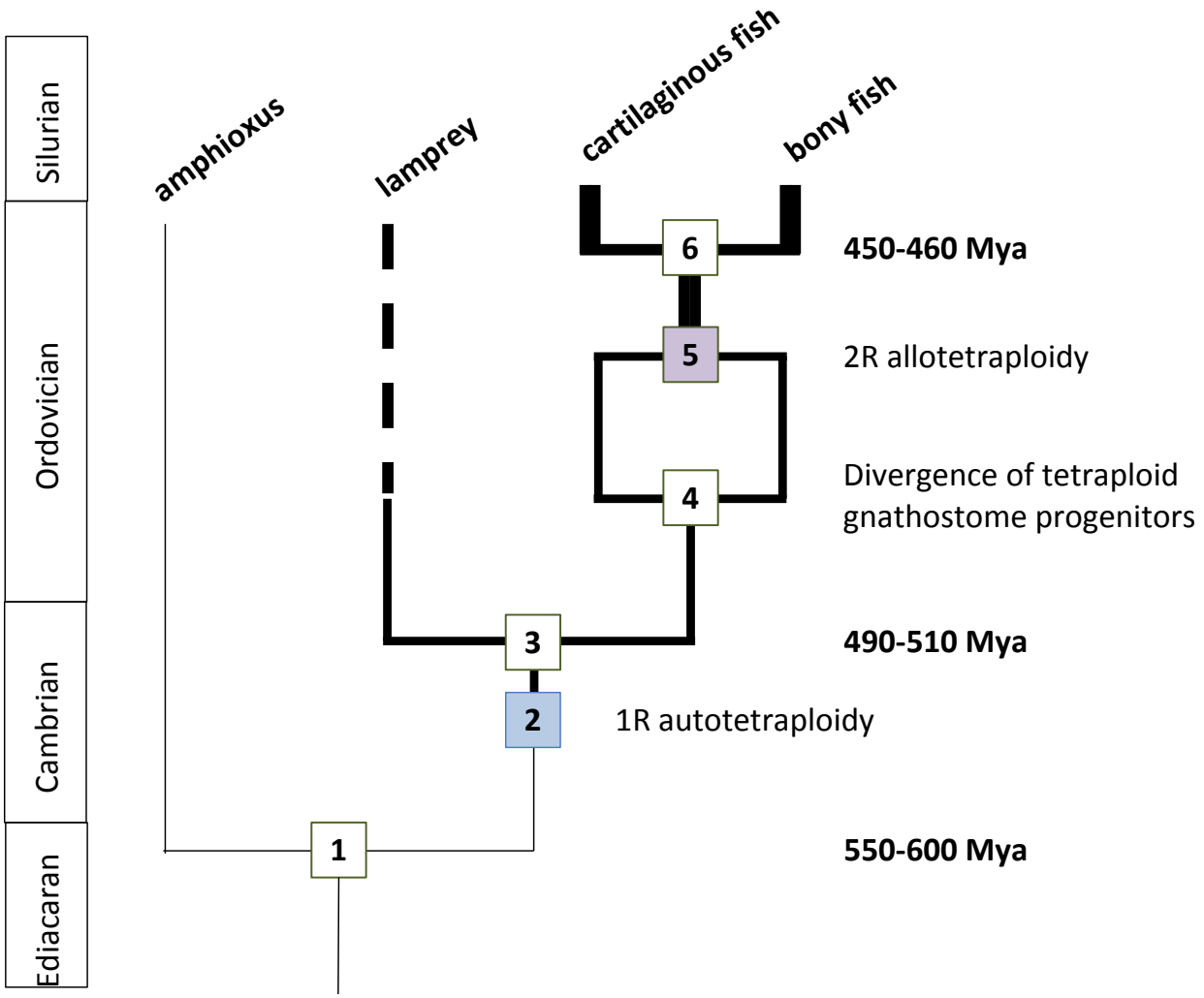


Figure 4.9: The emergence of allotetraploidy

4.6 Asymmetric retention in sub-genomes

We compute the retention rate of a vertebrate chromosome relative to each of its component chordate linkage groups (CLGs) as follows: the number of CLG gene families on that chromosome divided by the total number of gene families within that CLG. To do that, we discarded families with more than one paralog on a single chromosome. Retention rate generally match the patterns observed on the dot plot. We found that the fraction of CLG genes recovered on chromosomes assigned to a different CLG is 0.05, this number likely cor-

responds to the average orthology mis-assigment and/or translocation rate of genes among chromosomes.

Plotting the cumulative distributions of retention rates (Figure 4.11), we observe two peaks (around 0.1-0.2 and around 0.3-0.4). Those two populations represent two linkage groups that we call "high" and "low" retention groups.

We tested our hypothesis that paralogs are retained asymmetrically among sub-genomes against a null hypothesis in which paralogs are retained at random. To this end we simulated a null model in which the retention rate of each of four paralogous segments was chosen from a normal distribution with mean ($\mu = 25\%$) and standard deviation ($\sigma = 10\%$). In these simulations, negative retention rates were reset to zero. The mean and standard deviation were determined from the retention rates observed across all CSGs.

Pairs of retention values were taken from this distribution at random. For each pair, the larger value was designated "HR" and the smaller value "LR." For CSGs for which no fusions have been documented, we cannot relate a specific "HR" segment to a particular "LR" counterpart. To simulate these, we take four retention values from the normal distribution, and assign the top two to "HR" segments and the bottom two to "LR" segments. In this way, we construct a simulated version the table of these values from an explicitly symmetrical model in which the difference between "HR" and "LR" do not arise from any inherent asymmetry in the retention process.

For a test statistic we used the difference between high retention ("HR") and low retention ("LR") rates averaged over all pairs of chromosomal descendants of CSG. To determine the distribution of this test statistic under the null model, we computed it for one million simulations. This empirical bootstrap distribution is shown in Figure 4.10, along with the observed test statistic based on chicken and gar. We therefore reject the null hypothesis with p < 10e-6, and conclude that the retention rates are asymmetrically distributed, supporting an allotetraploid model for the second duplication in the 2R scenario.

Hypothesis testing using a min/max normal distribution

Given the null model of two normally distributed random variables r_i and r_j, with random variables $r_{\min}$ and $r_{\max}$:

$$r_{\min} = \min(r_i, r_j)$$
$$r_{\max} = \max(r_i, r_j)$$

As our gene retention data set had N entries, we draw N/2 of the min/max pairs and compute the average distance $E[r_{max} - r_{min}]$. We sample this average gene retention distance 1,000,000 times to achieve an empirical p-value bound of 1e-6, as our empirical gene distance of 0.2875738 is well outside all values for the distribution (Figure 4.10).

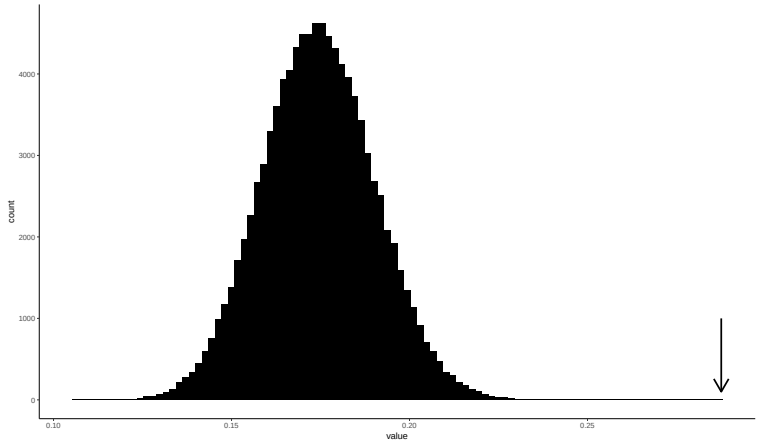


Figure 4.10: Bootstrapping $E[r_{max} - r_{min}]$ (true value marked with arrow)

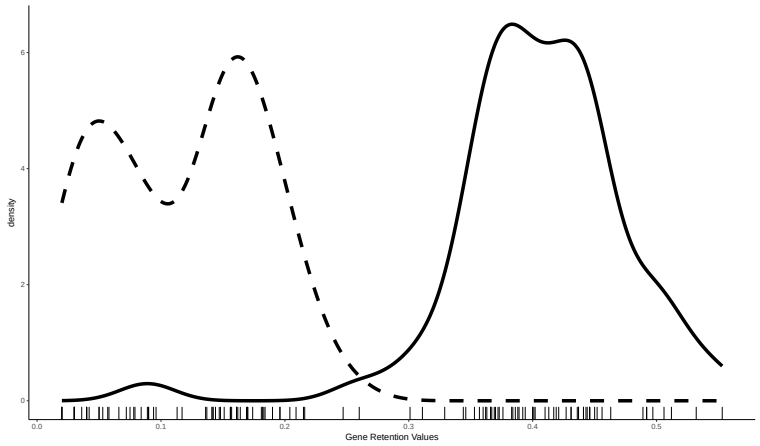


Figure 4.11: Bimodal distribution of HR/LR with fit density

Theoretical considerations

Though it is sufficient to simulate the data with a large number of bootstrap repetitions, the explicit distribution for $r_{\max}$ and $r_{\min}$ is known, and its probability distribution function can be shown to be:

$$f_{r_{\max}}(x) = 2\phi(x;\mu,\sigma)\Phi(x;\mu,\sigma)$$
$$f_{r_{\max}}(x) = 2\phi(x;\mu,\sigma)\left(1 - \Phi(x;\mu,\sigma)\right)$$

4.7 Discussion

Our results support a novel scenario for the chromosome-scale events that shaped jawed vertebrate genomes. First, we find ancient chordate linkage groups are stable, and relicts of them are readily detectable in both amphioxus and vertebrate genomes, accounting for extensive genome-wide duplication. Second, two distinct tetraploidizations occurred in the jawed-vertebrate lineage. The first genome-wide duplication ("1R") preceded the divergence of lampreys and jawed vertebrates, and was followed by a period dominated by chromosomal fusions on the jawed vertebrate stem, although many ancient chromosomal units retained their identities and persist as microchromosomes. A second genome-wide duplication ("2R") followed. The observation of asymmetric gene retention indicates that this second whole genome duplications was an allotetraploidization. Taken together, these observations provide new insights into the genomic events on the vertebrate stem. Our analysis provides a foundation for study of gene- and gene-regulatory level changes that occurred to facilitate the emergence of vertebrates from protochordate ancestors.

Chapter 5

Assembling and Annotating the Genome of *Hofstenia miamia*[†]

5.1 Introduction

The capacity to replace all missing tissues, i.e., whole-body regeneration, is present in nearly all animal phyla but notably absent in vertebrates [84]. Whereas studies of candidate genes and the advent of high-throughput transcriptomics have yielded insight into genes involved in wound response [85, 86], stem cell dynamics [87], and repatterning of tissue identities [88–90], how these genes are connected into networks is largely unknown. A detailed understanding of the DNA binding site logic of regulatory transcription factors is needed to distinguish direct versus indirect regulation of a particular locus and to elucidate connections between genes that are deployed as a cascade. Recent work has shown that "tissue regeneration enhancer elements" are active during zebrafish heart and fin regeneration" [91, 92], but how the epigenome responds to the process of whole-body regeneration is unknown.

To investigate whole-body regeneration from a regulatory genomics perspective, we focused our studies on *Hofstenia miamia*, an acoel worm that possesses the ability to regenerate its entire body and has recently been established as a model system with easy laboratory culture and systemic RNAi for functional studies [49]. Besides regenerative capacity, we chose *Hofstenia* for two main reasons: 1) sexually mature *Hofstenia* lay plentiful and accessible embryos that can be utilized for functional analysis of genomic predictions (Fig. 1a) acoels belong to the likely sister group (*Xenacoelomorpha*) to all other bilaterians (*Nephrozoa*), a phylogenetic position that can be leveraged to understand the evolution of regeneration and development [49, 93, 94] (Fig. 1b). We have sequenced the genome of *Hofstenia* at an average coverage of 89.6x via Illumina sequencing of paired-end (300-bp and 500-bp inserts) and mate pair (2-kb, 3-kb and 7-kb inserts) libraries derived from a single individual and a pool of five animals (Tables 5.1 and 5.3). The sequenced animals were descendants of a wild

[†]This chapter is based on my contributions to the manuscript "Acoel genome reveals the regulatory landscape of whole-body regeneration." by Gehrke et al.

population of diploid individuals and we estimated the polymorphism level at 0.426% (or 1 single nucleotide polymorphism every 235 bases). The draft assembly, achieved by the sequential application of PRICE [95], SSPACE [96], and the Chicago [75] method, totals 976.5 Mb of sequence (11.5% gaps). The assembly is of high quality, with half of the sequence (N50) contained in 294 scaffolds longer than 1 Mb.

To predict protein-coding genes, we generated a new transcriptome assembly to train Augustus [97] and recovered 22,632 gene models, 97% of which were supported by transcriptome data. BUSCO [98] analysis determined the *Hofstenia* genome annotation to be 84% complete. We performed principal components analysis on gene content in 36 metazoan genomes and found the *Hofstenia* genome nested among non-bilaterian and protostome genomes, indicating that it contains a standard complement of animal genes (Figure 4.1).

5.2 Genome assembly

Genomic sequence datasets and library construction

Genome sequencing efforts resulted in five datasets (Table 5.1) that were used for assembly and scaffolding (dsA-E), and three datasets that were used for scaffolding (Table 5.3). The first (dsA) consisted of three Illumina flowcell lanes of Nextera-generated paired-end libraries. The second (dsB) consisted of Illumina-sequenced mate-pair libraries with an intended insert size of 4kb. The third (dsC) consisted of additional Illumina-sequenced paired-end data. The fourth and fifth consisted of Illumina-sequenced mate-pair libraries with attended insert sizes of 3kb (dsD) and 7kb (dsE).

All libraries were filtered to remove low-quality sequences prior to use for assembly. A read pair was removed if, according to the fastq quality scores, fewer than 95% of nucleotides in either read had at least a 99% chance of having been called correctly. Mate-pair libraries were additionally trimmed at their 3' ends if >90% identity matches to linker sequences were found using an ungapped alignment, and pairs were removed if either read was cut to <80nt (dsB) or <60nt (dsD,E). Truncation was performed for partial matches to the 3' ends of reads or the 5' ends of adapters.

Genome assembly

Initial contig synthesis using SOAPdenovo

Initial contigs were generated using SOAPdenovo v1.0553 with all of the above datasets provided. The SOAPdenovo GapCloser program [99] was run next, using the scaffold output of the SOAPdenovo run and the same config file. The output was 1,653,197 scaffolds totaling 1.17Gb of sequence. Scaffolds <200nt long were culled, leaving 435,003 scaffolds of 1.02Gb total sequence, with a scaffold N50 of ∼13kb. Those scaffolds were split into 10 sets of roughly even nucleotide length. Thus, each seed set contained ∼102Mb of sequence. The scaffolds in each set were then cut at any stretch of 5 or more consecutive uncalled nucleotides

(N's), with terminal N's trimmed from the split-apart contigs, and with contigs shorter than 200nt culled. The scaffold N50 of ∼13kb was reduced to a contig N50 of ∼5.3kb, and the total sequence length of 1.02Gb was reduced to 778Mb.

Contig extension and collapse using PRICE

The contigs from each set described above were extended and made more coherent using the PRICE assembler (v1.0.1, http://derisilab.ucsf.edu/software/price/) [95]. Each contig set was supplied to a job as seeds, then extended for two cycles. The "repeat element fasta file" was manually constructed to contain potentially repetitive sequence elements based on elevated coverage of mapped reads.

The products of the two-cycle PRICE extend-and-join jobs above were combined using single-cycle runs of PRICE (v1.0.3), this time only using 5 million randomly-selected read pairs from the 3kb mate-pair data (dsD 5M) to allow nearby, overlapping contigs to be joined. The result was 235,962 contigs with a total length of 946Mb and contig N50 of 10.6kb.

Further collapse of the extended contigs was obtained using 3 million randomly-selected pairs from the 7kb mate-pair library (dsE) and PRICE v1.2. The result was 218,406 contigs with a total length of 901Mb and contig N50 of 11.2kb. Further collapse was achieved through a similar job using both the 5M 3kb and 3M 7kb mate-pair data sets simultaneously, yielding 206,917 contigs with a total length of 876Mb and contig N50 of 11.6kb.

Two shortcomings of PRICE assembly were addressed during post-analysis: a) the occasional production of tandemly and palindromically duplicated contigs, and b) the inability of PRICE to collapse highly-overlapping contigs if they are not linked by paired-end or mate-pair reads.

Source Material	Type of library	Type of Sequencing	Number of Reads	Library name
One worm (head3)	Nextera paired-end	100X100	46,540,003 pairs	dsA
One worm (head3)	Nextera paired-end	100X100	44,499,473 pairs	dsA
One worm (head3)	Nextera paired-end	100X100	37,331,548 pairs	dsA
multiple worms (3 or 6)	Nextera paired-end; aim was to have >500bp fragments	100X100	85,509,593 pairs	dsC
multiple worms	4kb Illumina mate pair	80X80	about 33 million read pairs	dsB
multiple worms (3 or 6)	3kb Nextera mate pair made by Macrogen	100X100	38,174,852 pairs	dsD
multiple worms (3 or 6)	7kb Nextera mate pair made by Macrogen	100X100	46,172,172 pairs	dsE

Table 5.1: Sequencing libraries used in initial contig assembly

Scaffolding using SSPACE

The resulting PRICE contigs were assembled using SSPACE STANDARD 3.0 and SSPACE GAPFILLER 1.1014 [96, 100] using libraries dsA, dsC, dsD and dsE from the contig assembly, as well as three others (Table 5.3). SSPACE operates by using an aligner (bwa [56]), and searching for a minimum depth of overlap for discordant pairs between two contigs, which allows a gap (sequence of "N" nucleotides linking and attaching the two contigs) to be formed. SSPACE GAPFILLER similarly looks to find reads that partially map to this gap, and using a minimum consensus to correct the unknown nucleotides covered by the reads. 5 paired end libraries (of insert size 300-500 bp: dsA, dsA, dsA, dsC, and dsF) and 4 mate pair libraries (of insert sizes 2000, 2000, 3000, and 7000 bp: dsG, dsH, dsD, and dsE, respectively) were used for both scaffolding and gap filling of the contigs, after being trimmed for adapter content. Results from SSPACE STANDARD and SSPACE GAPFILLER are presented in Table 5.1. SPACE GAPFILLER was able to close 27411 out of 159857 gaps (17.14%) and 9888508 out of 86596334 (11.42%) of nucleotides. This scaffolded assembly was then used for further refinement through the CHICAGO scaffolding protocol provided by Dovetail Genomics.

	Before Scaffolding	After Scaffolding/Closing
Main genome scaffold total	195763	94260
Main genome contig total	195763	172558
Main genome scaffold sequence total	855.0 MB	938.3 MB
Main genome contig sequence total	855.0 MB (0% gap)	854.9 MB (8.9% gap)
Main genome scaffold N/L50	19,927/11.8 KB	4,615/55.8 KB
Main genome contig N/L50	19,927/11.8 KB	16,275/14.6 KB
Number of scaffolds > 50 KB	506	5265
Main genome in scaffolds > 50 KB	3.6%	53.7%

Table 5.2: Improvements to the assembly as a result of initial scaffolding and gap filling with SSPACE and GAPFILLER.

Source Material	Type of library	Type of Sequencing	Number of Reads	Library name
One worm (head3)	Nextera paired-end	80X80	62,558,530 pairs	dsF
multiple worms	2kb Illumina mate pair	80X80	38,350,384 pairs	dsG
multiple worms	2kb Illumina mate pair	80X80	53,520,603 pairs	dsH

Table 5.3: Additional libraries used for scaffolding and gap closing

Assembly improvement via Dovetail's Chicago method

Chicago library preparation and sequencing: A Chicago library was prepared as described previously [75]. Briefly, 500ng of HMW gDNA (mean fragment length = 45 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, cross-links were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to 350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq platform to produce 463 million 2x151 bp paired-end reads, which provided 349.2-fold physical coverage of the genome (1-50 kb pairs).

Scaffolding the assembly with HiRise: The input *de novo* assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies [75]. Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs.

Genome coverage and polymorphism rate

Mate-pair and paired-end libraries dsA, dsB, dsC, dsE, dsF, dsG, and dsH, used for assembly and scaffolding (tables 3.1 and 3.3), were aligned to the reference using bwa mem (0.7.17) [101]. Coverage was calculated by using bedtools' (v2.27.0) genomecov module examining depth of sequence at all non-N bases in the reference [102]. Summary statistics were computed by examining the resulting table in R. For detecting polymorphism, variants were called using GATK (v4.0.0.0) HaplotypeCaller and GenotypeGVCF by examining bases with coverage between 10 - 3000 reads, and selecting for biallelic SNPs GATK [74]. The resulting number of SNPs was divided by the number of bases that fell within our coverage range in order to discover the per-base polymorphism rate (0.426%).

5.3 Genome annotation

Generation of a new transcriptome

To generate a new transcriptome assembly with the focus on regeneration and stem cell population, additional paired-end reads from regeneration time course fragments (6 and 48

Assembled genome size:	976.5 Mb
Number of scaffolds	18,348
N50 scaffold size	1.04 Mb
N50 contig size	15.8 kb
GC content	31.81%
Repeats	49.39%
Number of genes	22,632
Gene density	23.8 per Mb
Mean gene size	4,796 bp
Mean exon per gene	4.98
Mean intron per gene	3.98
Mean exon size	239 bp
Mean intron size	3,313 bp

Table 5.4: Draft genome statistics

hpa) and FACS-sorted X1 cells (158 M reads, 2014-02-11) were incorporated. Together with the original paired-end reads from embryonic mix and regeneration mix (87 M reads, 2011-09-19), 245 M reads were *de novo* assembled with Trinity (v2.4.0) using a genome-free approach. To remove redundancy, the new assembly (294,502 transcripts) was merged with the original assembly (19,860 transcripts) using EviGene (v2016.07.11). All antisense transcripts in the merged assembly were reverse complemented based on the longest open reading frame (ORF) prediction from TransDecoder (v5.0.1) and best-hits to Swiss-Prot database. The final new transcriptome (hmi_transcriptome_20140211_filtered, 30,056 transcripts) was generated after a functional filtering based on gene expression levels (TPM $>$ 5).

Repeat and protein-coding gene annotation

Repetitive elements were *de novo* identified with RepeatScout (v1.0.5) with the standard criteria (repeat length $>$ 50 bp and present $>$ 10 times). Repetitive elements (i.e. repeat hints) were then masked with RepeatMasker (v4.0.7) and annotated with known repeats using BLASTN and TBLASTX searches against RepBase (v20170127). To better identify the exon-intron boundary, paired-end RNA-seq raw reads were mapped to the genome assembly to generate intron hints. The transcriptome assembly was spliced aligned to the genome assembly with BLAT (v0.35) to generate exon hints. For training the gene prediction program, Augustus (v3.3), gene structures extracted from the transcriptome assembly were generated

with PASA (v2.2.0). After optimization (gene level sensitivity: 0.245), the *Hofstenia* genome was annotated with optimized Augustus using evidence from repeat, intron, and exon hints. Fragmented scaffolds smaller than 1 kb were removed. The mitochondrial genome (scaffold2257, 15,661 bp) is retained in the final *Hofstenia* genome assembly. In total, 14,990 gene models were annotated based on the best-hits using BLAST searches (e-value < 5e-1) against humans, mice, zebrafish, *D. melanogaster*, *C. elegans*, and Swiss-Prot databases in a stepwise manner. Further gene orthology annotation was performed using the KEGG Automatic Annotation Server [103].

5.4 Assessment of assembly quality

Genome completeness analysis

To assess the assembly quality, 395 *bona fide* genes under study in our lab were searched against the gene models using BLAST. Of these, 92% (365/395) cloned genes were found in our gene models. In addition, genome completeness was evaluated by searching 978 core metazoan genes in the gene models using BUSCO [98] (Benchmarking Universal Single-Copy Orthologs). To capture active transcripts in the genome, the transcriptome assembly was aligned to the gene models with BLAST (97% of predicted genes were supported by transcriptome data; e-value < 1e-5). The transcripts were also spliced aligned to the genome assembly with BLAT (89% of transcripts mapped to the genome).

Gene family analysis

Orthologous genes among 36 selected metazoan genomes were identified using all-to-all BLASTP searches and clustered with OrthoMCL (v2.0.9). Gene family annotation was performed with PANTHER (v12.0) [104] using the PANTHER HMM scoring tool (pantherScore2.1.pl). To gain insights into the overall composition of *Hofstenia* gene families among metazoans, a gene family size matrix with selected metazons was generated based on PANTHER gene family annotation. Principal component analysis (PCA) of gene family sizes was performed using the `R` package, `prcomp`.

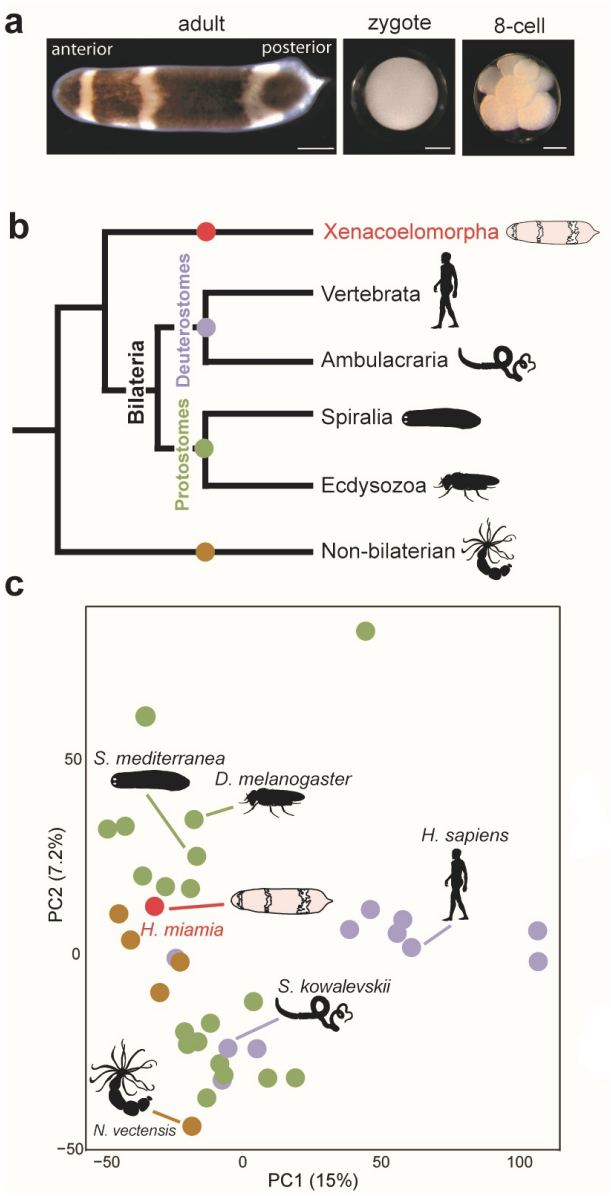


Figure 5.1: a, *Hofstenia miamia* adult, zygote, and 8-cell embryo. Scale bars represent 500 μm for adult, 100 μm for embryos. b, Phylogenetic tree showing the placement of acoels as the likely sister group to the rest of bilaterians. c, Principal components analysis of metazoan genomes showing that *Hofstenia* contains a standard complement of animal genes. Each dot represents a species with a sequenced genome and is colored based on membership to the major clades indicated in b.

5.5 Discussion

This genome assembly offers a valuable tool in the study of the dynamics of chromatin regulation in *Hofstenia miamia* and presents a useful resource for those studying both acoel flatworms and regenerative organisms alike. Though we have not achieved a chromosomal-scale assembly like in other genomes presented in this dissertation, half of our genome is contained in scaffolds of 1 Mb or greater (N50 of 1.04 Mb), which suggests that our assembly can still offer some understanding the underlying structure of the genome. Also useful is the presence of the 365 of 395 (92%) experimentally verified cDNAs from previous studies, as this genome was sequenced with the intention of being used for functional genomics experiments. Given 97% of annotated gene models are supported by transcriptomic data, we believe our annotation to accurately capture *bona fide Hofstenia* genes. This assembly allows for the design and analysis of functional genomic experiments, and eventually a starting point for further improvements that might lead to a chromosomal scale assembly.

Bibliography

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17,** 333–351. ISSN: 1471-0056 (June 2016).
2. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)* **323,** 133–8. ISSN: 1095-9203 (Jan. 2009).
3. Madoui, M.-A. *et al.* Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16,** 327. ISSN: 1471-2164 (Dec. 2015).
4. Grattapaglia, D. & Sederoff, R. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* **137,** 1121–37. ISSN: 0016-6731 (Aug. 1994).
5. Nyholt, D. R. All LODs are not created equal. *American journal of human genetics* **67,** 282–8. ISSN: 0002-9297 (Aug. 2000).
6. Wu, R., Ma, C.-X., Painter, I. & Zeng, Z.-B. Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in Outcrossing Species. *Theoretical Population Biology* **61,** 349–363. ISSN: 0040-5809 (May 2002).
7. Monroe, J. G. *et al.* TSPmap, a tool making use of traveling salesperson problem solvers in the efficient and accurate construction of high-density genetic linkage maps. *BioData mining* **10,** 38. ISSN: 1756-0381 (2017).
8. Lander, E. S. *et al.* MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1,** 174–81. ISSN: 0888-7543 (Oct. 1987).
9. Buetow, K. H. & Chakravarti, A. Multipoint gene mapping using seriation. I. General methods. *American journal of human genetics* **41,** 180–8. ISSN: 0002-9297 (Aug. 1987).
10. Doerge, R. *Constructing genetic maps by rapid chain delineation* 1996. <`http://agris.fao.org/agris-search/search.do?recordID=US9701956`>.
11. Van Os, H., Stam, P., Visser, R. G. F. & Van Eck, H. J. RECORD: a novel method for ordering loci on a genetic linkage map. *Theoretical and Applied Genetics* **112,** 30–40. ISSN: 0040-5752 (Dec. 2005).

12. Tan, Y.-D. & Fu, Y.-X. A Novel Method for Estimating Linkage Maps. *Genetics* **173,** 2383–2390. ISSN: 0016-6731 (May 2006).

13. Wilson, S. R. & Rao, D. C. A major simplification in the preliminary ordering of linked loci. *Genetic Epidemiology* **5,** 75–80. ISSN: 07410395 (1988).

14. Falk, C. T. A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. *Progress in clinical and biological research* **329,** 17–22. ISSN: 0361-7742 (1989).

15. Weeks, D. E. & Lange, K. Preliminary ranking procedures for multilocus ordering. *Genomics* **1,** 236–42. ISSN: 0888-7543 (Nov. 1987).

16. Lander, E. S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84,** 2363–7. ISSN: 0027-8424 (Apr. 1987).

17. Mollinari, M., Margarido, G. R. A., Vencovsky, R. & Garcia, A. A. F. Evaluation of algorithms used to order markers on genetic maps. *Heredity* **103,** 494–502. ISSN: 0018-067X (Dec. 2009).

18. Margarido, G. R. A., Souza, A. P. & Garcia, A. A. F. OneMap: software for genetic mapping in outcrossing species. *Hereditas* **144,** 78–79. ISSN: 00180661 (June 2007).

19. Schiffthaler, B., Bernhardsson, C., Ingvarsson, P. K. & Street, N. R. BatchMap: A parallel implementation of the OneMap R package for fast computation of F1 linkage maps in outcrossing species. *PLOS ONE* **12** (ed Candela, H.) e0189256. ISSN: 1932-6203 (Dec. 2017).

20. Land, A. H. & Doig, A. G. An Automatic Method of Solving Discrete Programming Problems. *Econometrica* **28,** 497. ISSN: 00129682 (July 1960).

21. Van Os, H. *et al.* Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* **173,** 1075–87. ISSN: 0016-6731 (June 2006).

22. Speed, T. P. in *Genetic Mapping and DNA Sequencing* 65–88 (Springer New York, New York, NY, 1996). doi:10.1007/978-1-4612-0751-1_5. <http://link.springer.com/10.1007/978-1-4612-0751-1%7B%5C_%7D5>.

23. Zhao, H. & Speed, T. P. On genetic map functions. *Genetics* **142,** 1369–77. ISSN: 0016-6731 (Apr. 1996).

24. Dudoit, S. *Statistical Analysis of Meiosis* Berkeley, CA, 2015.

25. Haldane, J. B. S. The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8,** 299–309 (1919).

26. Kosambi, D. D. The estimation of map distances from recombination values. *Annals of Eugenics* **12,** 172–175. ISSN: 20501420 (Jan. 1943).

27. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–238. ISSN: 0028-0836 (Nov. 1983).

28. Tsui, L. C. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science (New York, N.Y.)* **230,** 1054–7. ISSN: 0036-8075 (Nov. 1985).

29. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics* **42,** 30–35. ISSN: 1061-4036 (Jan. 2010).

30. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nature Communications* **9,** 1911. ISSN: 2041-1723 (Dec. 2018).

31. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821. ISSN: 0036-8075 (Aug. 2012).

32. Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harbor perspectives in biology* **5,** a012740. ISSN: 1943-0264 (Nov. 2013).

33. Rodgers, K. & McVey, M. Error-Prone Repair of DNA Double-Strand Breaks. *Journal of cellular physiology* **231,** 15–24. ISSN: 1097-4652 (Jan. 2016).

34. Ohno, S. *Evolution by Gene Duplication* ISBN: 978-3-642-86661-6. doi:10.1007/978-3-642-86659-3. <http://link.springer.com/10.1007/978-3-642-86659-3> (Springer Berlin Heidelberg, Berlin, Heidelberg, 1970).

35. Garcia-Fernàndez, J. & Holland, P. W. H. Archetypal organization of the amphioxus Hox gene cluster. *Nature* **370,** 563–566. ISSN: 0028-0836 (Aug. 1994).

36. Spring, J. Vertebrate evolution by interspecific hybridisation–are we polyploid? *FEBS letters* **400,** 2–8. ISSN: 0014-5793 (Jan. 1997).

37. Escriva, H., Manzon, L., Youson, J. & Laudet, V. Analysis of Lamprey and Hagfish Genes Reveals a Complex History of Gene Duplications During Early Vertebrate Evolution. *Molecular Biology and Evolution* **19,** 1440–1450. ISSN: 0737-4038 (Sept. 2002).

38. Pebusque, M. J., Coulier, F., Birnbaum, D. & Pontarotti, P. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Molecular Biology and Evolution* **15,** 1145–1159. ISSN: 0737-4038 (Sept. 1998).

39. Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P. & Inoko, H. Evidence of en bloc duplication in vertebrate genomes. *Nature Genetics* **31,** 100–105. ISSN: 1061-4036 (May 2002).

40. Lundin, L.-G., Larhammar, D. & Hallböök, F. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. *Journal of structural and functional genomics* **3,** 53–63. ISSN: 1345-711X (2003).

41. Hokamp, K., McLysaght, A. & Wolfe, K. H. The 2R hypothesis and the human genome sequence. *Journal of structural and functional genomics* **3,** 95–110. ISSN: 1345-711X (2003).

42. Dehal, P. & Boore, J. L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology* **3** (ed Holland, P.) e314. ISSN: 1545-7885 (Sept. 2005).

43. Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453,** 1064–1071. ISSN: 0028-0836 (June 2008).

44. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10,** 725–732. ISSN: 1471-0056 (Oct. 2009).

45. Furlong, R. F. & Holland, P. W. H. Were vertebrates octoploid? *Philosophical Transactions of the Royal Society B: Biological Sciences* **357,** 531–544. ISSN: 0962-8436 (Apr. 2002).

46. Smith, J. J. & Keinath, M. C. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Research* **25,** 1081–1090. ISSN: 1088-9051 (Aug. 2015).

47. Friedman, R. & Hughes, A. L. Pattern and Timing of Gene Duplication in Animal Genomes. *Genome Research* **11,** 1842–1847. ISSN: 1088-9051 (Nov. 2001).

48. Naz, R., Tahir, S. & Abbasi, A. A. An insight into the evolutionary history of human MHC paralogon. *Molecular Phylogenetics and Evolution* **110,** 1–6. ISSN: 10557903 (May 2017).

49. Srivastava, M., Mazza-Curll, K. L., van Wolfswinkel, J. C. & Reddien, P. W. Whole-Body Acoel Regeneration Is Controlled by Wnt and Bmp-Admp Signaling. *Current Biology* **24,** 1107–1113. ISSN: 09609822 (May 2014).

50. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Information Processing Letters* **31,** 7–15. ISSN: 0020-0190 (Apr. 1989).

51. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS genetics* **4,** e1000212. ISSN: 1553-7404 (Oct. 2008).

52. Rastas, P. & Berger, B. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **33** (ed Berger, B.) 3726–3732. ISSN: 1367-4803 (Dec. 2017).

53. Bekaert, M. *Genetic-Mapper: vectorial genetic map drawer* in *F1000Research* **5** (June 2016). doi:10.7490/F1000RESEARCH.1112266.1. <https://f1000research.com/posters/5-1301>.

54. Bredeson, J. V. *et al.* Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology* **34,** 562–570. ISSN: 1087-0156 (May 2016).

55. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27,** 1691–1692. ISSN: 1367-4803 (June 2011).

56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760. ISSN: 1367-4803 (July 2009).

57. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv: 1207.3907. <http://arxiv.org/abs/1207.3907> (July 2012).

58. Lobos, G., Mendez, M. A., Cattan, P. & Jaksic, F. Low genetic diversity of the successful invasive African clawed frog *Xenopus laevis* (Pipidae) in Chile. *Studies on Neotropical Fauna and Environment* **49,** 50–60. ISSN: 0165-0521 (Jan. 2014).

59. Schiffthaler, B., Bernhardsson, C., Ingvarsson, P. K. & Street, N. R. BatchMap: A parallel implementation of the OneMap R package for fast computation of F1 linkage maps in outcrossing species. *PLOS ONE* **12** (ed Candela, H.) e0189256. ISSN: 1932-6203 (Dec. 2017).

60. Swaminathan, K. *et al.* A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. *BMC Genomics* **13,** 142. ISSN: 1471-2164 (Apr. 2012).

61. Ma, X.-F. *et al.* High Resolution Genetic Mapping by Genome Sequencing Reveals Genome Duplication and Tetraploid Genetic Structure of the Diploid *Miscanthus sinensis*. *PLoS ONE* **7** (ed Hazen, S. P.) e33821. ISSN: 1932-6203 (Mar. 2012).

62. Liu, S. *et al.* High-density genetic map of *Miscanthus sinensis* reveals inheritance of zebra stripe. *GCB Bioenergy* **8,** 616–630. ISSN: 17571693 (May 2016).

63. Doyle, J. in *Molecular Techniques in Taxonomy* 283–293 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1991). doi:10.1007/978-3-642-83962-7_18. <http://www.springerlink.com/index/10.1007/978-3-642-83962-7%7B%5C_%7D18>.

64. Cullings, K. Design and testing of a plant-specific PCR primer for ecological and evolutionary studies. *Molecular Ecology* **1,** 233–240. ISSN: 0962-1083 (Dec. 1992).

65. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* **32,** 347–355. ISSN: 1087-0156 (Apr. 2014).

66. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495. ISSN: 0028-0836 (Jan. 2016).

67. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34,** 184–191. ISSN: 1087-0156 (Feb. 2016).

68. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–88. ISSN: 0036-8075 (Jan. 2016).

69. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nature Biotechnology* **32,** 577–582. ISSN: 1087-0156 (June 2014).

70. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339,** 819–823. ISSN: 0036-8075 (Feb. 2013).

71. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549,** 111–115. ISSN: 0028-0836 (Aug. 2017).

72. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171,** 557–572.e24. ISSN: 1097-4172 (Oct. 2017).

73. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18,** 821–829. ISSN: 1088-9051 (Feb. 2008).

74. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–1303. ISSN: 1088-9051 (Sept. 2010).

75. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* **26,** 342–350. ISSN: 1088-9051 (Mar. 2016).

76. Burton, R. S., Pereira, R. J. & Barreto, F. S. Cytonuclear Genomic Interactions and Hybrid Breakdown. *Annual Review of Ecology, Evolution, and Systematics* **44,** 281–302. ISSN: 1543-592X (Nov. 2013).

77. Putnam, N. H. *et al.* Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science* **317,** 86–94. ISSN: 0036-8075 (July 2007).

78. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493,** 526–531. ISSN: 0028-0836 (Dec. 2012).

79. Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* **527,** 459–465. ISSN: 0028-0836 (Nov. 2015).

80. Garsmeur, O. *et al.* Two Evolutionarily Distinct Classes of Paleopolyploidy. *Molecular Biology and Evolution* **31,** 448–454. ISSN: 1537-1719 (Feb. 2014).

81. Session, A. M. *et al.* Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538,** 336–343. ISSN: 0028-0836 (Oct. 2016).

82. Braasch, I. & Postlethwait, J. H. in *Polyploidy and Genome Evolution* 341–383 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012). doi:10.1007/978-3-642-31442-1_17. <http://link.springer.com/10.1007/978-3-642-31442-1%7B%5C_%7D17>.

83. Allendorf, F. W. & Thorgaard, G. H. in *Evolutionary Genetics of Fishes* 1–53 (Springer US, Boston, MA, 1984). doi:10.1007/978-1-4684-4652-4_1. <http://link.springer.com/10.1007/978-1-4684-4652-4%7B%5C_%7D1>.

84. Bely, A. E. & Nyberg, K. G. Evolution of animal regeneration: re-emergence of a field. *Trends in Ecology & Evolution* **25,** 161–170. ISSN: 01695347 (Mar. 2010).

85. Wenemoser, D., Lapan, S. W., Wilkinson, A. W., Bell, G. W. & Reddien, P. W. A molecular wound response program associated with regeneration initiation in planarians. *Genes & Development* **26,** 988–1002. ISSN: 0890-9369 (May 2012).

86. Wurtzel, O. *et al.* A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians. *Developmental cell* **35,** 632–645. ISSN: 1878-1551 (Dec. 2015).

87. Zhu, S. J. & Pearson, B. J. (Neo)blast from the past: new insights into planarian stem cell lineages. *Current Opinion in Genetics & Development* **40,** 74–80. ISSN: 0959437X (Oct. 2016).

88. Petersen, C. P. & Reddien, P. W. Smed-betacatenin-1 Is Required for Anteroposterior Blastema Polarity in Planarian Regeneration. *Science* **319,** 327–330. ISSN: 0036-8075 (Jan. 2008).

89. Gurley, K. A., Rink, J. C. & Alvarado, A. S. Beta-Catenin Defines Head Versus Tail Identity During Planarian Regeneration and Homeostasis. *Science* **319,** 323–327. ISSN: 0036-8075 (Jan. 2008).

90. Petersen, C. P. & Reddien, P. W. Polarized notum Activation at Wounds Inhibits Wnt Function to Promote Planarian Head Regeneration. *Science* **332,** 852–855. ISSN: 0036-8075 (May 2011).

91. Kang, J. *et al.* Modulation of tissue repair by regeneration enhancer elements. *Nature* **532,** 201–206. ISSN: 0028-0836 (Apr. 2016).

92. Goldman, J. A. *et al.* Resolving Heart Regeneration by Replacement Histone Profiling. *Developmental Cell* **40,** 392–404.e5. ISSN: 15345807 (Feb. 2017).

93. Cannon, J. T. *et al.* Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530,** 89–93. ISSN: 0028-0836 (Feb. 2016).

94. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. *Nature* **530,** 94–97. ISSN: 0028-0836 (Feb. 2016).

95. Ruby, J. G., Bellare, P. & DeRisi, J. L. PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3: Genes, Genomes, Genetics* **3,** 865–880. ISSN: 2160-1836 (May 2013).

96. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding preassembled contigs using SSPACE. *Bioinformatics* **27,** 578–579. ISSN: 1460-2059 (Feb. 2011).

97. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24,** 637–644. ISSN: 1460-2059 (Mar. 2008).

98. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212. ISSN: 1367-4803 (Oct. 2015).

99. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20,** 265–72. ISSN: 1549-5469 (Feb. 2010).

100. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biology* **13,** R56. ISSN: 1465-6906 (June 2012).

101. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714. ISSN: 1367-4803 (Mar. 2008).

102. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842. ISSN: 1460-2059 (Mar. 2010).

103. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35,** W182–W185. ISSN: 0305-1048 (May 2007).

104. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* **41,** D377–D386. ISSN: 0305-1048 (Nov. 2012).

Appendix A

Code Appendix

A.1 Minimum Spanning Tree Imputation Code

Main Code (R)

```
library(ggfortify)
library(cluster)
library(plyr)
library(igraph)
library(foreach)
library(doParallel)

args = commandArgs(trailingOnly=TRUE)
source(args[1])

setwd(MY_DIR)

count_coindicidences <- function(l1, l2) {
  return(sum(na.omit(l1 == l2)))
}

binomial_haplotype_phasing <- function(datf) {

  p_values = matrix(1, ncol = nrow(datf),
                    nrow = nrow(datf))

  offspring.datf = datf
  offspring.datf[offspring.datf == 0] = -1
  offspring.datf[is.na(offspring.datf)] = 0

  datf1to1 = as.matrix(offspring.datf > 0) %*% as.matrix(t(offspring.datf >
     0))
  datfN1toN1 = as.matrix(offspring.datf < 0) %*% as.matrix(t(offspring.datf
     < 0))
  datfN1to1 = as.matrix(offspring.datf > 0) %*% as.matrix(t(offspring.datf <
      0))
  datf1toN1 = as.matrix(offspring.datf < 0) %*% as.matrix(t(offspring.datf >
      0))

  stayMatrix = datf1to1 + datfN1toN1
  switchMatrix = datfN1to1 + datf1toN1

  as.matrix(t(datf)) %*% as.matrix(datf)
```

```
for (i in seq(nrow(datf)-1)) {
  for (j in seq(i+1,nrow(datf),1)) {
    stay = stayMatrix[i,j]
    switch = switchMatrix[i,j]
    if ((stay > 0) || (switch > 0)) {
      pvalue = binom.test(x=c(stay,switch))$p.value
      p_values[i,j] = pvalue
      p_values[j,i] = pvalue
    }
  }
}

pairs <- as.data.frame(t(combn(row.names(datf), m = 2)))
colnames(pairs) <- c("p1", "p2")

p_values = as.data.frame(p_values)
names(p_values) = row.names(datf)
row.names(p_values) = row.names(datf)

datf[datf==0] <- -1
datf[is.na(datf)] <- 0
datf <- as.matrix(datf)

cont_correlation_matrix <- cor(t(datf), use="pairwise.complete.obs")
cont_correlation_matrix[is.na(cont_correlation_matrix)] = 0
if (ncol(as.matrix(cont_correlation_matrix)) < 2) { return(data.frame(A=
   rep(NA, length(ncol(correlation_matrix))),
    B=rep(NA, length(ncol(correlation_matrix))),
    row.names = names(correlation_matrix))) }

pairs <- as.data.frame(t(combn(row.names(cont_correlation_matrix), m = 2))
   )
colnames(pairs) <- c("p1", "p2")

pairs$cor <- mapply(function(x, y) {cont_correlation_matrix[toString(x),
   toString(y)]}, pairs$p1, pairs$p2)
pairs$p_values <- mapply(function(x, y) {p_values[toString(x), toString(y)
   ]}, pairs$p1, pairs$p2)

valid_pairs <- subset(pairs, p_values < MAX_PVALUE)
```

```
  my_graph <- graph_from_data_frame(d = valid_pairs ,
                                    directed = FALSE,
                                    vertices = row.names(cont_correlation_
                                       matrix))
  my_path <- mst(graph = my_graph, weights = valid_pairs$p_values)

  my_sub_path <- decompose(my_path)[[which.max(clusters(my_path)$csize)]]

  # plot(my_sub_path)

  return_haplotype = rep(NA, length(names(V(my_sub_path))))
  names(return_haplotype) <- names(V(my_sub_path))
  return_haplotype[1] <- 1

  while( anyNA(return_haplotype) ) {
    phase_from <- names(return_haplotype[!is.na(return_haplotype)])
    for (i in seq(length(phase_from))) {
      to_phase <- names(neighbors(my_path, phase_from[i]))

      for (j in seq(length(to_phase))) {
        return_haplotype[to_phase[j]] <- return_haplotype[phase_from[i]] *
            sign(cont_correlation_matrix[phase_from[i], to_phase[j]])
      }
    }
  }
  return_haplotypes <- data.frame(A=rep(NA, length(return_haplotype)),
                                  B=rep(NA, length(return_haplotype)),
                                  row.names = names(return_haplotype))
  return_haplotypes$A <- lapply(return_haplotype, function(x) { if (x == -1)
      0 else x })
  return_haplotypes$B <- lapply(-return_haplotype, function(x) { if (x ==
     -1) 0 else x })
  return(return_haplotypes)
}

data <- read.table(paste(MY_DIR, FILENAME, sep="/"), header=TRUE)

row.names(data) <- paste(data$CHROM, data$POS, sep = '.')
chrom_names <- unique(data$CHROM)

if (DEBUG) { chrom_names <- chrom_names[grepl(TEST_CHR, chrom_names)] }
```

```
offspring <- unlist(lapply(X = colnames(data), FUN = grepl, pattern = paste(
   OFFSPRING_STR,"CHROM|POS", sep="|")))
just.offspring.data <- data[offspring]
offspring_names <- colnames(data)[unlist(lapply(X = colnames(data), FUN =
   grepl, pattern = OFFSPRING_STR))]

final_genotype_calls <- data.frame(matrix(ncol = length(offspring_names),
   nrow = 0))
final_genotype_calls_all <- data.frame(matrix(ncol = length(offspring_names)
   , nrow = 0))
preimpute_genotype_calls <- data.frame(matrix(ncol = length(offspring_names)
   , nrow = 0))
imputation_statistics <- data.frame(matrix(ncol = 8, nrow = 0))

colnames(imputation_statistics) <- c("CHROM", "BLOCK_START",
                                     "BLOCK_END", "HAPLO_LENGTH", "SITES_W_GENO
                                        _BEFORE_IMP",
                                     "SITES_W_GENO_AFTER_IMP", "INDVID_GENO", "
                                        INDVID_NOT_GENO")

colnames(final_genotype_calls) <- offspring_names
block_coordinates = data.frame(matrix(ncol=2,nrow=0))

for (chrom in chrom_names) {
  biggest_position <- max(subset(data, CHROM == chrom)$POS)
  blocks <- c(seq(1, biggest_position, BLOCK_SIZE), biggest_position)

  block_indices = seq(1, length(blocks)-1)

  a = do.call(rbind, Map(data.frame, chrom=rep(chrom, length(block_indices))
     ,
                  block_i=block_indices,
                  biggest_chr_pos=rep(biggest_position, length(block_
                     indices))))
  row.names(a) = NULL
  block_coordinates = rbind(block_coordinates, a)
  print(a)
}

registerDoParallel(cores=NCORES)
return_vals = foreach(j = seq(nrow(block_coordinates)),
                    .inorder=TRUE,
```

```
                  .combine=rbind,
                  .export = c("binomial_haplotype_phasing",
                              "count_coindicidences")) %dopar% {
my_current_chrom <- as.character(block_coordinates[j,]$chrom)
biggest_position <- block_coordinates[j,]$biggest_chr_pos
i = block_coordinates[j,]$block_i

  length_of_haplotypes_for_block <- 0
  sites_with_genotypes_before_imputation <- 0
  sites_without_genotypes_before_imputation <- 0
  individuals_genotyped <- 0
  individuals_not_genotyped <- length(offspring_names)

  just.offspring.data.subset <- subset(just.offspring.data, CHROM == my_
     current_chrom &
                                (POS < ((BLOCK_SIZE * i))) &
                                (POS > ((BLOCK_SIZE * (i-1)))) )[seq(3,ncol
                                   (just.offspring.data),1)]
  t.just.offspring.data.subset <- t(just.offspring.data.subset)
  sites_with_genotypes_before_imputation <- sum(colSums(!is.na(just.
     offspring.data.subset)))
  sites_without_genotypes_before_imputation <- sum(colSums(is.na(just.
     offspring.data.subset)))

  just.offspring.data.preimpute = just.offspring.data.subset

  imputation_calls <- rep(NA, ncol(just.offspring.data.subset))

  adding_genotype_calls <- data.frame(matrix(NA, ncol = length(offspring_
     names),
                                             nrow = nrow(just.offspring.data.
                                                preimpute)))
  names(adding_genotype_calls) = offspring_names
  row.names(adding_genotype_calls) = row.names(just.offspring.data.
     preimpute)

  full_imputation = adding_genotype_calls
  names(full_imputation) = data.frame(NA, ncol = length(offspring_names),
                                             nrow = nrow(preimpute_genotype_
                                                calls))
  names(full_imputation) = offspring_names
```

```
if (REDUCE_DATA) {
  adding_genotype_calls <- full_imputation[1,]
}

if (nrow(just.offspring.data.subset) > 2) {

haplotypes <- binomial_haplotype_phasing(just.offspring.data.subset)
length_of_haplotypes_for_block <- length(haplotypes$A)
if (dim(haplotypes)[1] > 1) {

include_list <- rownames(haplotypes)
just.offspring.data.subset <- just.offspring.data.subset[include_list,]
just.offspring.data.preimpute = just.offspring.data.subset

A_co <- sapply(X = just.offspring.data.subset, FUN = count_
   coindicidences, l2 = haplotypes$A)
B_co <- sapply(X = just.offspring.data.subset, FUN = count_
   coindicidences, l2 = haplotypes$B)

AB_table <- table(A_co, B_co)
graph_lab <- rep(NA, length(A_co))

for (l in seq(length(A_co))) {
  graph_lab[l] <- AB_table[toString(unname(A_co[l])), toString(unname(B_
     co[l]))]
}

A_co <- sapply(X = just.offspring.data.subset, FUN = count_
   coindicidences, l2 = haplotypes$A)
B_co <- sapply(X = just.offspring.data.subset, FUN = count_
   coindicidences, l2 = haplotypes$B)

imputation_calls <- mapply(function(a_counts, b_counts, total) {
  if (is.na(a_counts) || is.na(b_counts) || is.na(total)) { NA; }
  else if ( a_counts == 0 && b_counts == 0 ) { NA; }
  else if ( binom.test(c(a_counts, b_counts))$p.value < ASSINGMENT_
     PVALUE ) {
    if (a_counts > b_counts) { "A"; }
    else { "B"; }
  }
  else {
    NA;
```

```
  }
}, A_co, B_co, rep(nrow(just.offspring.data.subset), length(A_co)))

color_calls = imputation_calls
color_calls[is.na(color_calls)] = 'black'
color_calls[color_calls == "A"] = 'blue'
color_calls[color_calls == "B"] = 'red'

pdf(sprintf("%s.%i.pdf", my_current_chrom, BLOCK_SIZE *(i-1)))
title = sprintf("Regions␣on␣%s␣from␣%i␣to␣%i\n(markers␣in␣haplotype␣=␣%i
   )",
              my_current_chrom, BLOCK_SIZE *(i-1),
              BLOCK_SIZE *(i), length(haplotypes$A))
plot(A_co, B_co, xlab = "A␣coincidences", ylab = "B␣coincidences",
     main = title, col = color_calls)
lines(x = seq(0,max(max(A_co),max(B_co))), y = seq(0,max(max(A_co),max(B
   _co))))
text(A_co, B_co, labels = offspring_names, pos = 4, cex = .25)
dev.off()

adding_genotype_calls <- data.frame(matrix(ncol = length(offspring_names
   ), nrow = length(haplotypes$A)))
colnames(adding_genotype_calls) <- colnames(just.offspring.data.subset)
rownames(adding_genotype_calls) <- rownames(just.offspring.data.subset)

for (y in seq(length(colnames(just.offspring.data.subset)))) {
  this_offspring <- colnames(just.offspring.data.subset)[y]
  if (!is.na(imputation_calls[this_offspring])) {
  adding_genotype_calls[,this_offspring] <- unlist(haplotypes[,toString(
     imputation_calls[this_offspring])])
  }
  else {
  adding_genotype_calls[,this_offspring] <- rep(NA, length(nrow(adding_
     genotype_calls)))
  }
}

full_imputation = adding_genotype_calls

if (REDUCE_DATA) {
  z <- min(round(nrow(adding_genotype_calls)/2), 1)
  adding_genotype_calls <- adding_genotype_calls[z,]
```

```
    }

    individuals_genotyped <- sum(colSums(is.na(adding_genotype_calls))==0)
    individuals_not_genotyped <- sum(colSums(is.na(adding_genotype_calls))
       >0)

    }
    }

    data_to_add <- unlist(list(my_current_chrom, BLOCK_SIZE * (i-1), min(
       BLOCK_SIZE * (i), biggest_position),
                  length_of_haplotypes_for_block,
                  sites_with_genotypes_before_imputation,
                  sites_without_genotypes_before_imputation,
                  individuals_genotyped,
                  individuals_not_genotyped))

    names(data_to_add) = c("CHROM", "BLOCK_START", "BLOCK_END", "HAPLO_
       LENGTH",
                         "SITES_W_GENO_BEFORE_IMP", "SITES_W_GENO_AFTER_IMP"
                            ,
                         "INDVID_GENO", "INDVID_NOT_GENO")

    return(list(adding_genotype_calls, data_to_add, full_imputation, just.
       offspring.data.preimpute))
}
stopImplicitCluster()

for (k in seq(nrow(return_vals))) {
  final_genotype_calls = rbind(final_genotype_calls, return_vals[,1][[k]])
  imputation_statistics = rbind(imputation_statistics, return_vals[,2][[k]],
                                stringsAsFactors = FALSE)
  final_genotype_calls_all = rbind(final_genotype_calls_all, return_vals
     [,3][[k]])
  preimpute_genotype_calls = rbind(preimpute_genotype_calls, return_vals
     [,4][[k]])
}

colnames(imputation_statistics) = c("CHROM", "BLOCK_START",
                                    "BLOCK_END", "HAPLO_LENGTH",
                                   "SITES_W_GENO_BEFORE_IMP",
                                    "SITES_W_GENO_AFTER_IMP",
```

```
                                "INDVID_GENO", "INDVID_NOT_GENO")


tmp.final_genotype_calls <- final_genotype_calls
final_genotype_calls[final_genotype_calls == 0] <- "0/0"
final_genotype_calls[final_genotype_calls == 1] <- "0/1"

write.table(preimpute_genotype_calls,
           paste(MY_DIR, paste(OUTPREFIX, "preimputed.full.gt", sep="."),
              sep="/"),
           na = ".",
           col.names = TRUE,
           append = FALSE,
           sep = "\t",
           quote = FALSE)

write.table(final_genotype_calls_all,
           paste(MY_DIR, paste(OUTPREFIX, "imputed.full.gt", sep="."), sep="
              /"),
           na = ".",
           col.names = TRUE,
           append = FALSE,
           sep = "\t",
           quote = FALSE)

write.table(final_genotype_calls,
           paste(MY_DIR, paste(OUTPREFIX, "imputed.strict.gt", sep="."), sep
              ="/"),
           na = ".",
           col.names = TRUE,
           append = FALSE,
           sep = "\t",
           quote = FALSE)

write.table(final_genotype_calls[colSums(!is.na(t(final_genotype_calls)))/
   length(offspring_names) > .8,],
           paste(MY_DIR, paste(OUTPREFIX, "imputed.80per.strict.gt", sep="."
              ), sep="/"),
           na = ".",
           col.names = TRUE,
           append = FALSE,
           sep = "\t",
```

```
          quote = FALSE)

write.table(final_genotype_calls[colSums(!is.na(t(final_genotype_calls)))/
   length(offspring_names) > .6,],
          paste(MY_DIR, paste(OUTPREFIX, "imputed.60per.strict.gt", sep="."
             ), sep="/"),
          na = ".",
          col.names = TRUE,
          append = FALSE,
          sep = "\t",
          quote = FALSE)

write.table(final_genotype_calls[colSums(!is.na(t(final_genotype_calls)))/
   length(offspring_names) > .4,],
          paste(MY_DIR, paste(OUTPREFIX, "imputed.40per.strict.gt", sep="."
             ), sep="/"),
          na = ".",
          col.names = TRUE,
          append = FALSE,
          sep = "\t",
          quote = FALSE)

write.table(imputation_statistics,
          paste(MY_DIR, paste(OUTPREFIX, "impute_stats.tsv", sep="."), sep
             = "/"),
            na = ".",
            col.names = TRUE,
            append = FALSE,
            sep = "\t",
            quote = FALSE)
```

Example Configuration File (R)

```
MY_DIR="~/amphioxus_map/"
FILENAME = "female.het.chisquared.gt"
OUTPREFIX = "female.het.chisquared.500k.1e-3"
TEST_CHR = "Sc7u5tJ_1590"
DEBUG = FALSE
OFFSPRING_STR = "G|N|L"
BLOCK_SIZE = 500000
NCORES = 3
MAX_PVALUE = 1e-3
```

```
ASSINGMENT_PVALUE = 1e-3
REDUCE_DATA=TRUE
```

Run Command (SH)

```
Rscript Parallel_Imputation.R CONFIG.R
```