UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DEVELOPING AND APPLYING CHROMATIN PROXIMITY LIGATION METHODS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Brendan L. O'Connell

December 2017

The Dissertation of Brendan L. O'Connell
is approved:

Professor David Haussler, Chair

Professor Richard E. Green

Professor Beth Shapiro

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Developing and Applying Chromatin Proximity Ligation Methods

by

Brendan L. O'Connell

Proximity ligation methods are a means of capturing long-range spatial information about DNA sequences with short-read sequencing compacted and concatenated DNA molecules. HiC, a method of gathering genome spatial information by the process of chromatin proximity ligation and capture, represents a powerful and versatile tool for modern genomics. Over the course of this dissertation, I demonstrate improvements to the HiC method and novel uses for the HiC data. An early novel use of chromatin proximity data was in the form of Chicago, an *in vitro* assembled chromatin version of HiC primarily used for *de novo* genome assembly scaffolding. This method was used to scaffold nearly one hundred new, high contiguity genome assemblies over the last two years. I next describe my own improvements to the HiC method, resulting in a faster, more economical version of HiC, and apply the improvements to explore the rapid-aging ICE mouse model and to scaffold a high-contiguity assembly of the Atlantic herring genome. I also show that the haplotype informative nature of HiC, when combined with recombined germline samples, allows for individual, personalized recombination maps, using both the Atlantic herring and human samples.

This work is dedicated to

My grandparents,

who got me into engineering very, very early.

Acknowledgments

I would like to thank my committee, Professor David Haussler, Professor Richard 'Ed' Green, and Professor Beth Shapiro, for their time and support. The text of this dissertation includes a reprint of the following previously published material: "Chromosome-scale shotgun assembly using an in vitro method for long-range linkage", *Genome Research* 26.3 (2016): 342-350. Richard E. Green directed and supervised the research which forms the basis for the published work.

Dr. Chris Troll at Dovetail Genomics discovered the SPRI-based Chicago method. Furthermore, my discussions with Dr. Troll provided the impetus to try many of the methodological improvements described in this dissertation.

I would like to thank Professor David Sinclair, Dr. Motoshi Hayano, and Jae-Hyun Yang, for providing the ICE mice and the previous research used as the basis for Chapter 3.

For Chapter 4, I would like to acknowledge the voulenteers who provided samples, as well as Professor Russ Corbett-Deitig, who provided invaulable advice concerning the recombination mapping method.

For Chapter 5, I would like to thank Professor Leif Andersson of Uppsala University. Professor Andersson provided the herring samples as well as the draft genome assembly.

Acknowledgment also goes to my readers and editors: Elizabeth Lewicki, Kathleen Patterson, and my lovely and talented fiancee Ruth Nichols. This would

be a pretty rough read without them.

Portions of my PhD were funded by Mr. Ed Schulak. Your support has been much appreciated!

I would like to acknowledge the members of the UCSC Paleogenomics lab, in appreciation of their support over the last five years. Finally, I thank my family for their support (moral and material), without which this dissertation would never have been written.

Chapter 1

Introduction

Over the past several decades, DNA sequencing technologies have revolutionized the biological sciences [43, 33]. Generating huge sequencing datasets has evolved from an impossibly expensive dream to a regular part of genomics[43]. In the last decade, many researchers have focused on developing biological methods to better realize the promise of high-throughput sequencing [33]. There have been initiatives to sequence thousands of species, genotype a multitude of humans, and probe the mystery of how a 10 micron diameter nucleus manages to transcribe anything with three meters of DNA packed inside [29]. HiC is a biological method which has proven to be of great utility[51].

HiC is a method which assays the proximity of DNA sequences in three dimensional space via chromatin proximity ligation. It was originally developed by Erez Lieberman-Aiden, *et al.* in 2009 [51] for examining chromosome architecture and localization. But HiC is not simply a method for probing genome architecture. Rather, it represents a versatile, powerful, group of methods for answering genetic, genomic, and

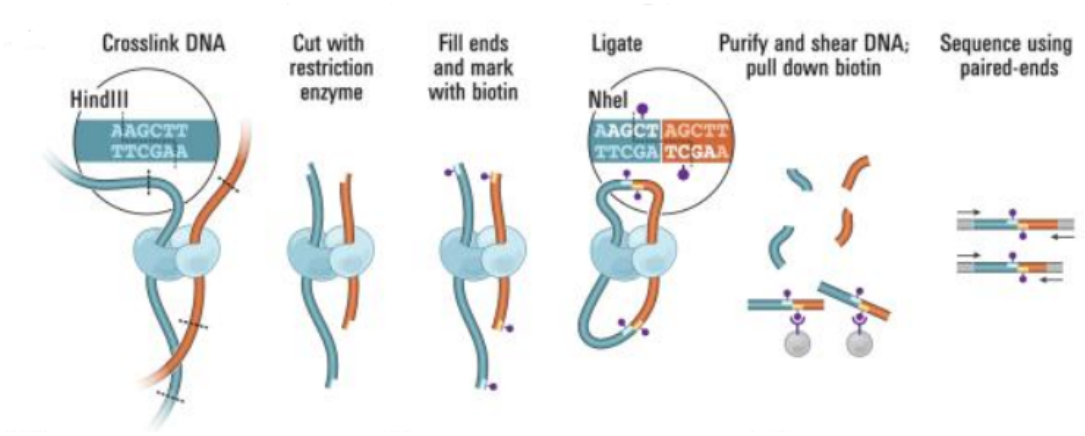

Figure 1.1: Figure 1A from Lieberman-Aiden *et al.* 2009, showing the HiC process.

other biological questions. Over the course of this dissertation, I will present multiple examples of the versatility of HiC applied to numerous organisms.

Methodologically, HiC borrows from "Chromatin Conformation Capture" (3C), "Circularized Chromosome Conformation Capture" (4C), and "Carbon-Copy Chromosome Conformation Capture" (5C) [21, 96, 24]. HiC takes advantage of restriction-digested DNA bound to histones which have been chemically crosslinked, usually with formaldehyde (See Figure 1.1). Unlike the methods mentioned above, however, HiC captures genome-wide interactions, as opposed to specific loci. To accomplish this, the HiC protocol uses a restriction endonuclease that leave 5' single stranded DNA overhangs. The overhangs are then filled in with biotinylated nucleotides (Figure 1.1). The proximity and chromatin conformation information is captured by ligating the biotinylated fragments, then purifying them on streptavidin coated beads, before sequencing with paired end Illumina. As originally conceived, HiC had several important

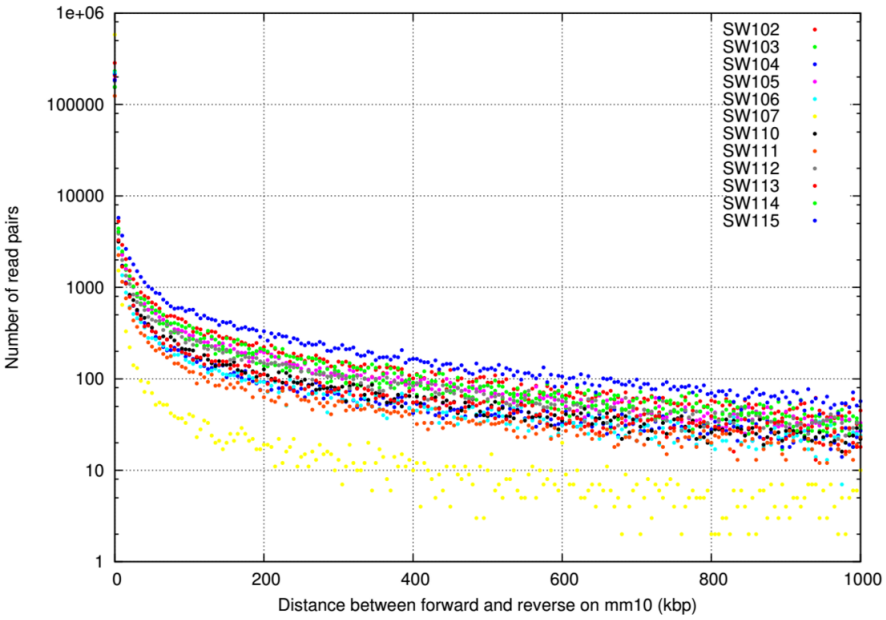

Figure 1.2: Hi-C library insert distributions.

practical limitations. HiC required large numbers of cells grown in culture, and could not be performed on tissue samples. Furthermore, as originally published, the protocol required geometrically increasing volumes for each step. This results in a very expensive bill of reagents, over $6000 before taking into account sequencing [51]. Finally, the increasing volumes required at each step made sample handling difficult and time-consuming. Subsequent research refined the methodology to decrease the sample size required for HiC, the time required to make a HiC library, and the error rates of the HiC library [39, 22, 62, 37].

When the sequencing data from HiC are mapped to a known reference genome, they show an insert distribution unique to HiC libraries. The frequency of reads at a given insert distance declines exponentially with increasing insert distance (Figure **??**).

This characteristic of the data can predict the insert distance between two loci even if the absolute distance is unknown. In Figure **??**, for example, if the frequency with which one sees a pair of loci is $\approx 1.2 \times 10^{10.5}$, then the loci are around 50 Mbp apart.

The overarching theme of my doctoral work has been the development, optimization, and application of chromatin proximity ligation methods. In the remainder of this dissertation, I will detail some of the novel uses for chromatin proximity ligation data. Additionaly, I will report on the optimizations and modifications I have made to the HiC method during the course of my doctoral work.

First, I will show one of the early successes in using the HiC concept of chromatin proximity ligation with *in vitro* reassembled chromatin for genome scaffolding, also known as 'Chicago', as detailed in our 2016 paper: "Chromosome-scale shotgun assembly using an in vitro method for long-range linkage" (Chapter 2) [69, 28]. In this paper, we demonstrated the applicability of HiC-type data (specifically Chicago, or *in vitro* assembled chromatin proximity ligation data) for genome scaffolding. This method has resulted in dozens of new *de novo* genome assemblies and, as shown both in Chapter 2 and in Chapter 3, allows one to achieve levels of genome contiguity quickly, and far more easily than was previously possible.

In Chapter 3, I will detail my own applications of the HiC method to various sample types. I review the HiC methodological improvements, as well as some of my analyses, and the implications of the work. Specifically, I will focus on a) the methodological improvements I have made to HiC, and b) a collaboration with the Sinclair lab at Harvard University focusing on finding changes in the genomic architecture of mouse

model for rapid aging, and c) using the new methods to scaffold the Atlantic herring genome to increase the contiguity of the assembly 15-fold over the currently published assembly, using 2 lanes of Illumina HiSeq and a draft contig assembly generated at Uppsala University, with herring liver samples provided by Professor Leif Andersson.

In Chapter 4, I will show how HiC may be used for generating meiotic recombination rate maps. Based on previous evidence that HiC data has a high degree of haplotype concordance within the read pairs, it follows that HiC data may be used to haplotype phase genomic data. By making HiC libraries from somatic tissue and from sperm, and then comparing the rate at which the germline haplotype information differs from the somatic dataset, it is possible to estimate the recombination rate for a single individual, across the entirety of the genome. This method is limited, however, by the heterozygosity of the sample, the rate of chimeric ligation in the datasets, and by the accuracy of the somatic hapltotype phasing. I will show the results in both the Atlantic herring and in a human sample.

Appendix A includes the most recent version of the HiC protocol (used to generate the Herring data in Chapter 3). Appendix B will include supplementary figures for Chapters 3 and 4.

Chapter 2

Chromosome-scale shotgun assembly using an in vitro method for long-range linkage

Note: this chapter is a reproduction of a previously published paper for which I was co-first author [69]. I was primarily responsible for the laboratory methods, while Nick Putnam, my co-first author, wrote the majority of the scaffolding software.

2.1 Abstract

Long-range and highly accurate de novo assembly from short-read data is one of the most pressing challenges in genomics. Recently, it has been shown that read pairs generated by proximity ligation of DNA in chromatin of living tissue can address this problem, dramatically increasing the scaffold contiguity of assemblies. Here, we describe

a simpler approach (Chicago) based on in vitro reconstituted chromatin. We generated two Chicago data sets with human DNA and developed a statistical model and a new software pipeline (HiRise) that can identify poor quality joins and produce accurate, long-range sequence scaffolds. We used these to construct a highly accurate de novo assembly and scaffolding of a human genome with scaffold N50 of 20 Mbp. We also demonstrated the utility of Chicago for improving existing assemblies by reassembling and scaffolding the genome of the American alligator. With a single library and one lane of Illumina HiSeq sequencing, we increased the scaffold N50 of the American alligator from 508 kbp to 10 Mbp.

2.2 Introduction

A "holy grail" of genomics is the accurate reconstruction of full-length haplotype-resolved chromosome sequences with low effort and cost. High-throughput sequencing methods have sparked a revolution in the field of genomics. By generating data from millions of short fragments of DNA at once, the cost of resequencing genomes has fallen dramatically, rapidly approaching $1000 per human genome [78]. Substantial obstacles remain, however, in transforming short read sequences into long, contiguous genomic assemblies.

Currently accessible and affordable high-throughput sequencing methods are best suited to the characterization of short-range sequence contiguity and genomic variation. Achieving long-range linkage and haplotype phasing requires either the ability

to directly and accurately read long (i.e., tens of kilobase) sequences or the capture of linkage and phase relationships through paired or grouped sequence reads.

A number of methods for increasing the contiguity and accuracy of de novo assemblies have recently been developed. Broadly, they attempt either to increase the read lengths generated from sequencing or to increase the insert size between paired short reads that can subsequently be used to scaffold genome assemblies. For example, the PacBio RS II chemistry updated in 2014 is advertised as producing raw reads with mean lengths of 15 kbp but suffers from error rates as high as 15% and remains about 100-fold more expensive than high-throughput short reads [46, 70]. Commercially available long-reads from Oxford Nanopore are promising but have even higher error rates and lower throughput [32]. These long-read technologies greatly simplify the process of assembly since, in many cases, repetitive or otherwise ambiguous regions of a genome are traversed in single reads. Illumina's TruSeq synthetic long-read technology (formerly Moleculo) is limited to 10-kbp reads maximum [**?**]. CPT-seq is somewhat similar in approach but does not rely on long-range PCR amplification [2, 4]. Despite a number of improvements, fosmid library creation [92, 94] remains time-consuming and expensive. To date, the community has not settled on a consistently superior technology for large inserts or long reads that is available at the scale and cost needed for large-scale projects like the sequencing of thousands of vertebrate species [29] or hundreds of thousands of humans [84].

The challenge of creating reference-quality assemblies from low-cost sequence data is evident in the comparison of the quality of assemblies generated with today's

technologies and the human reference assembly . Many techniques, including BAC clone sequencing, physical maps, and Sanger sequencing, were used to create the high-quality and highly contiguous human reference standard with an 38.5-Mbp N50 length (the size of the scaffold at which at least half of the genome assembly can be found on scaffolds at least that large) and error rate of one per 100,000 bases [36] . In contrast, a recent comparison of the performance of whole-genome shotgun (WGS) assembly software pipelines, each run by their developers on very high coverage data sets from libraries with multiple insert sizes, produced assemblies with N50 scaffold length ranging up to 4.5 Mbp on a fish genome and 4.0 Mbp on a snake genome [**?**].

High coverage of sequence with short reads is rarely enough to attain a high-quality and highly contiguous assembly. This is due primarily to repetitive content on both large and small scales, including the repetitive structure near centromeres and telomeres, large paralogous gene families like zinc finger genes, and the distribution of interspersed nuclear elements such as LINEs and SINEs. Such difficult-to-assemble content composes large portions of many eukaryotic genomes, for example, 60% - 70% of the human genome [19]. When such repeats cannot be spanned by the input sequence data, fragmented and incorrect assemblies result. In general, the starting point for de novo assembly combines deep-coverage (50200 minimum), short-range (300500 bp) paired-end shotgun data with intermediate range mate-pair libraries with insert sizes between 2 and 8 kbp and longer range (35-kbp) fosmid end pairs [31, 74]. However, even mate-pair data spanning these distances is often not completely adequate for generating megabase scale assembles.

Recently, high-throughput short-read sequencing has been used to characterize the three-dimensional structure of chromosomes in living cells. Proximity ligation–based methods like Hi-C [51] and other chromatin capture–based methods [23, 39] rely on the fact that, after fixation, segments of DNA in close proximity in the nucleus are more likely to be ligated together, and thus sequenced as pairs, than are distant regions. As a result, the number of read pairs between intrachromosomal regions is a slowly decreasing function of the genomic distance between them. Several approaches have been developed that exploit this information for the purpose of genome assembly scaffolding and haplotype phasing [12, 40, 75, 54].

While Hi-C and related methods can identify biologically mediated long-range chromatin contacts at multi-megabase length scales, most of the data describe DNA–DNA proximity on the scale of tens or hundreds of kilobases. These contacts arise from the polymer physics of the nucleosome-wound DNA fiber rather than from chromatin biology. In fact, the large-scale organization of chromosomes in nuclei provides a confounding signal for assembly since, for example, telomeres of different chromosomes are often associated in cells.

We demonstrate here that DNA linkages up to several hundred kilobases can be produced in vitro using reconstituted chromatin rather than living chromosomes as the substrate for the production of proximity ligation libraries. The resulting libraries share many of the characteristics of Hi-C data that are useful for long-range genome assembly, including a regular relationship between within–read pair distance and read count. By combining this in vitro long-range linking library with standard WGS and

jumping libraries, we generated a de novo human genome assembly with long-range accuracy and contiguity comparable to more expensive methods for a fraction of the cost and effort. This method, called Chicago, depends only on the availability of modest amounts of high-molecular-weight DNA and is generally applicable to any species. Here we demonstrate the value of this Chicago data not only for de novo genome assembly using human and alligator but also as an efficient tool for the identification and phasing of structural variants.

2.3 Results

2.3.1 Libraries and Sequencing

We extracted 5.5 g of high-molecular-weight DNA for Chicago libraries (in fragments of 150 kbp using the Qiagen HMW DNA kit and in fragments of 500 kb with agarose gel plug extraction) from the human cell line GM12878 and from the blood of a wild-caught American alligator (Supplemental Fig. S1). We reconstituted chromatin by combining the DNA with purified histones and chromatin assembly factors. Ordered chromatin assembly was confirmed by partial MNase digestion and gel electrophoresis (Supplemental Fig. S2). The reconstituted chromatin was then fixed with formaldehyde, and Chicago libraries were generated (Fig. 2.1 and Methods). For the human GM12878 sample, we generated three Chicago libraries. Two libraries were generated from DNA with an average size of 150 kb and using either the restriction enzyme MboI (library L1) or MluCI (L2). The ultra-high-molecular-weight (500 kb) library (L3) was created with

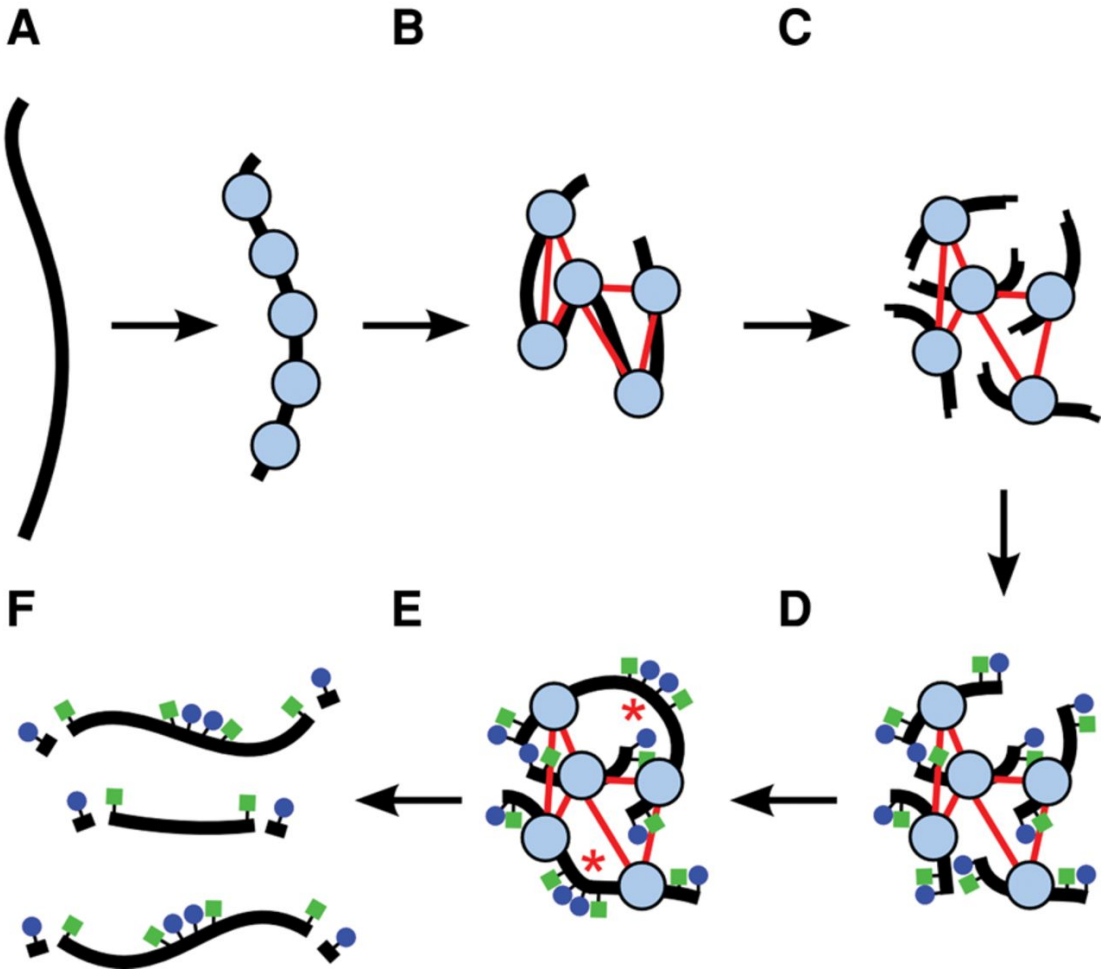

Figure 2.1: A diagram of a Chicago library generation protocol. (A) Chromatin (nucleosomes in blue) is reconstituted in vitro upon naked DNA (black strand). (B) Chromatin is fixed with formaldehyde (thin red lines are crosslinks). (C) Fixed chromatin is cut with a restriction enzyme, generating free sticky ends (performed on streptavidin-coated beads; data not shown). (D) Sticky ends are filled in with biotinylated (blue circles) and thiolated (green squares) nucleotides. (E) Free blunt ends are ligated (ligations indicated by red asterisks). (F) Crosslinks are reversed and proteins removed to yield library fragments, which are then digested with an exonuclease to remove the terminal biotinylated nucleotides. The thiolated nucleotides protect the interior of the library fragments from digestion.

MboI. These libraries were sheared to an average of 300500 bp in size and ligated to adapters for sequencing on the Illumina HiSeq 2500 as paired 100-bp reads, generating 46 million pairs for L1, 52 million for L2, and 165 million read pairs for L3. For the American alligator (Alligator mississipiensis), we similarly constructed a single MboI Chicago library and sequenced it on a single lane, yielding 210 million read pairs.

To determine the utility of these data for genome assembly and haplotype phasing, we aligned the GM12878 Chicago data to the reference human assembly, hg19 (Fig. 2.2). The Chicago libraries provided useful linking information for separations up to 150 kbp for L1 and L2 and up to 500 kbp for L3, consistent with the expected maximum size of input DNA fragments. By mapping these read pairs back to the reference human genome, we assessed the rate of background noise, defined for libraries L1 and L2 as reads pairs that map reliably to the genome but span distances >500 kbp or map to different chromosomes. For these libraries, we estimated the noise rate to be approximately one spurious link between unrelated 500-kbp genomic windows (mean of 0.97 such links). The linkage data span various size ranges. For illustration purposes, these data can be conceptually partitioned into various size bins based on the observed genomic distance between reliably mapped read pairs. Considered in this way, the single lane of sequencing from the GM12878 libraries provides linking information equivalent to 3.8, 8.4, 8.6, 18.6, 13.5, and 6.5 physical coverage in 0- to 1-kbp, 1- to 5-kbp, 5- to 10-kbp, 10- to 25-kbp, 25- to 50-kbp, and 50- to 200-kbp bins, respectively, while for alligator the comparable coverage estimates were 5.4, 16.7, 16.7, 42.2, 36.1, and 16.5 respectively (Fig. 2.3).

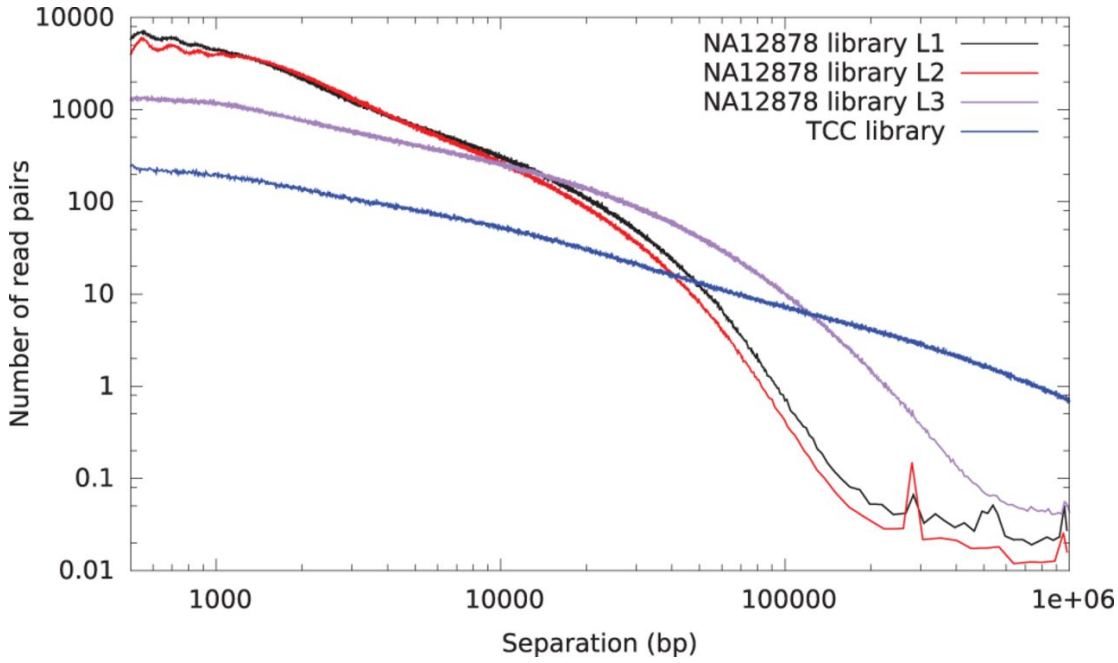

Figure 2.2: Histogram of read pair separations for several sequencing libraries mapped to hg19. (Black) Chicago library L1, prepared with MboI and 150-kbp input DNA; (red) Chicago library L2, prepared with MluCI and 150-kbp input DNA; and (violet) Chicago library L3, prepared with 500-kbp input DNA. A human Hi-C library (Kalhor et al. 2012) is shown in dark blue for comparison.

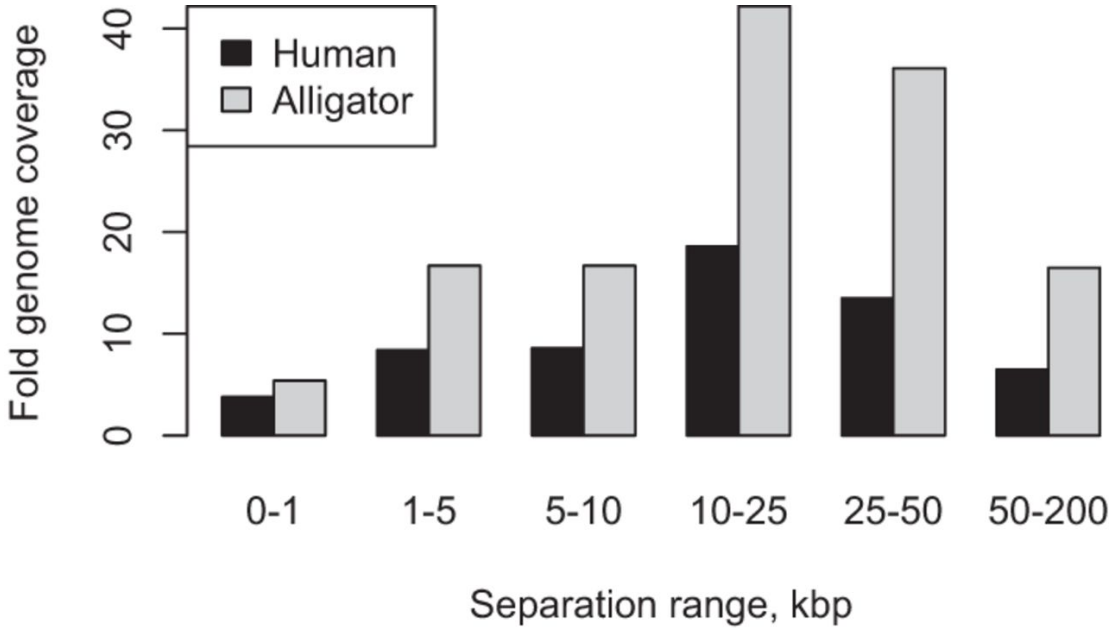

Figure 2.3: Genome coverage (sum of read pair separations divided by estimated genome size) in various read pair separation bins.

2.3.2 Chicago data for genome scaffolding

We next determined the capability for Chicago data to aid the scaffolding of a previously described meraculous assembly of GM12878 that used 101-bp paired-end Illumina reads to yield 84 genomic coverage and a N50 of 33 kbp [15, 79]. First, we mapped the L1 and L2 Chicago read pairs to this initial assembly as described in the Methods. We found that 68.1% of read pairs mapped such that both forward and reverse reads had map-quality scores of 20 or greater and were thus considered uniquely mapping within the assembly and were not duplicates. Of these read pairs, 35.4% had forward and reverse reads that mapped to different contigs and were thus potentially informative for further scaffolding of the assembly. We also used the same Chicago data to scaffold a DISCOVAR assembly of 50 coverage in 250-bp paired-end reads

(ftp://ftp.broadinstitute.org/pub/crd/Discovar/assemblies) with an initial scaffold N50 of 178 kbp (Weisenfeld et al. 2014). We found that 67.3% of L1 and L2 read pairs mapped to the DISCOVAR assembly with both forward and reverse reads having map quality scores of 20 or greater and were not duplicates. Of these reads, 26.5% mapped to different contigs.

We developed a likelihood model describing how Chicago libraries sample genomic DNA and integrated it with a software pipeline called HiRise for iteratively identifying and breaking misassemblies and for rescaffolding contigs based on Chicago links (Methods). We compared the completeness, contiguity and correctness at local and global scales of the resulting assembly to assemblies of rich WGS data sets, including extensive coverage in fosmid end pairs created by two of the leading WGS de novo assemblers: meraculous (MERAC) [15] and ALLPATHS-LG (APLG) (Table 2.1; Supplemental Table S1; [31]). To avoid the arbitrary choices involved in constructing alignment-based comparisons of assembly quality, we based our comparison on the locations in the assembly of 25.4 million 101-bp marker sequences. Because the de novo assemblers report only a single haplotype at each locus, to avoid ambiguity we selected marker sequences that are a randomly selected subset of all distinct 101-bp sequences that occur exactly once in each haplotype of a diploid reconstruction of the GM12878 genotype [**?**]. In this way, these markers are likely single-copy, unique segments of the human genome that are homozygous in the individual we sequenced (GM12878). We then assessed each assembly by gauging the completeness and accuracy of these markers in each assembly versus the well-assembled human reference genome [16].

2.3.3 Long-range scaffolding accuracy

The genomic scaffolds that the HiRise pipeline produced were longer and had a lower rate of global misassemblies than the published meraculous and APLG assemblies, both of which rely on deep coverage in paired fosmid end reads. Table 2.3.3 shows the fraction of the total assembly found in scaffolds containing a misjoin. Misjoins were identified at three thresholds, as follows: A scaffold s is anchored to a chromosome c when all the marker 101-mers within some 5-, 10-, or 50-kb intervals on s are found on c in the reference. Scaffolds anchored to two or more chromosomes are classified as misjoined. To assess the completeness of each assembly, we computed the fraction of marker 101-mers present.

Because the DNA ligation events that create Chicago pairs are not constrained to produce read pairs of defined relative strandedness, contig relative orientations during scaffolding must be inferred from read density information. As a result, the Chicago HiRise scaffolds have a higher rate of scaffolding orientation errors. For each of the four human genome assemblies compared in Table 2.3.3, we counted the number of pairs of consecutive 101-mers along the scaffold that map to the same reference chromosome but with incongruent orientation, indicating a strand switch in the assembly, and report the mean density of such errors on the genome. Similarly, the broad range of read pair separations in the Chicago library can lead to more uncertainty in the estimation of gap sizes. To assess the impact of this on the assemblies, we identified pairs of marker 101-mers that were separated by s_a between 49.5 and 50 kb in each assembly, and examined

their separations s_r in the reference genome; we report in Table 2.3.3 the minimum separation discrepancy x such that $|s_a - s_r| < x$ for 95% of the sample. The sample sizes were 458,966 and 478,494 marker pairs for the MERAC/HiRise and DISCOVAR/HiRise assemblies, respectively. The Supplemental Material includes a graphical depiction of all MERAC/HiRise scaffold misjoins.

To assess the effect on scaffolding quality on the quantity of Chicago data generated, we used a 50% subsample of the L1 and L2 libraries to scaffold the 30-kb N50 meraculous assembly and found that this reduced the scaffold N50 to 7.1 Mb (a 53% reduction) with a comparable number of misjoins. When we increased the coverage in Chicago data to 1.7 the original physical coverage with the addition of the L3 library, the scaffold N50 increased nearly threefold to 43 Mb, while the number of misjoins counted at the most sensitive of the thresholds that we used increased by 38% to 94 (Table 2.1).

Assembler	N Misjoins			N50(Mbp)	95%-CI	%C	Orientation Errors
	5kbp	10kbp	50kbp				
MERAC PE+MP+Fos	20	13	5	9.1	1.3 kbp	94.8	1/601 kb
APLG PE+MP+Fos	111	67	33	12.1	6.4 kbp	92.2	1/1013 kbp
MERAC PE+HiSRise 1.0	68	38	4	15.1	7.7 kbp	95.3	1/131kb
DISCOVAR PE+HiRise 1.0	39	20	2	20.9	3.8 kbp	98.8	1/307 kb
MERAC PE+HiRise 1.0 (50%)	70	37	4	7.1	8.0 kbp	95.3	1/111 kb
MERAC PE+HiRise 1.0 (+L3)	94	50	12	43.0	9.2 kbp	95.3	1/110 kb

Table 2.1: The number of global misjoins computed at three different thresholds for anchoring scaffolds to the reference. (N50) Scaffold (95% CI 50 kbp Δ) 50-kbp separation discrepancy 95% confidence interval, 95% CI=x, or given a pair of unique 101-mer tags in the assembly, 95% of them are within 50kbp $\pm$x of each other in the reference. (%C) Completeness, mean distance between 101-mer strand switches relative to the reference.

2.3.4 Improving the alligator assembly with Chicago data

To further assess the utility of Chicago data for improving existing assemblies, we generated a single Chicago library for the American alligator and mapped these data to a de novo assembly (N50 81 kbp) created using publicly available data [34], and applied the HiRise scaffolding pipeline. The resulting assembly had a scaffold N50 of 10.3 Mbp. To assess the accuracy of these scaffolds, we aligned a collection of 1485 previously generated [76] bacterial artificial chromosome (BAC) end sequences to the assembly. Of those, 1298 pairs were uniquely aligned by GMAP (Wu2005) with 90% coverage and 95% identity to the genome assembly and the HiRise scaffolded version. In the input assembly, 12.5% of the BAC end pairs were captured in the same scaffold with the expected orientation and separation. In the HiRise assembly, 96.5% of the BAC end pairs were aligned in the same scaffold with 98.1% of the BAC end pairs on the same scaffold in correct relative orientation. Five (0.39%) BAC end pairs were placed on the same scaffold but at a distance significantly larger than the insert size, and 14 (1.08%) were placed on separate scaffolds but far enough from the edge of the scaffold that the distance would be larger than the insert size, suggesting a global density of misjoins of fewer than one per 8.36 Mbp of assembly.

2.3.5 Identification of structural variants

Mapping paired sequence reads from one individual against a reference is the most commonly used sequence-based method for identifying differences in genome structure like inversions, deletions, and duplications [85]. Figure 2.4 shows how Chicago

read pairs from GM12878 mapped to the human reference genome GRCh38 reveal two previously identified structural differences, and illustrates how the variant haplotype phase can be inferred. Supplemental Figures S3 and S4 show schematically the expected read mapping distributions. Because GM12878 derives from an individual that has been trio-sequenced, gold-standard haplotype phase information is available to check the accuracy of Chicago phasing information. Read pairs that are haplotype informative and that span between 10 and 150 kbp are 99.83% in agreement with the known haplotype phase for GM12878. This allows confident assignment of variant allele phase based on read mapping. To estimate the sensitivity and specificity of Chicago data for

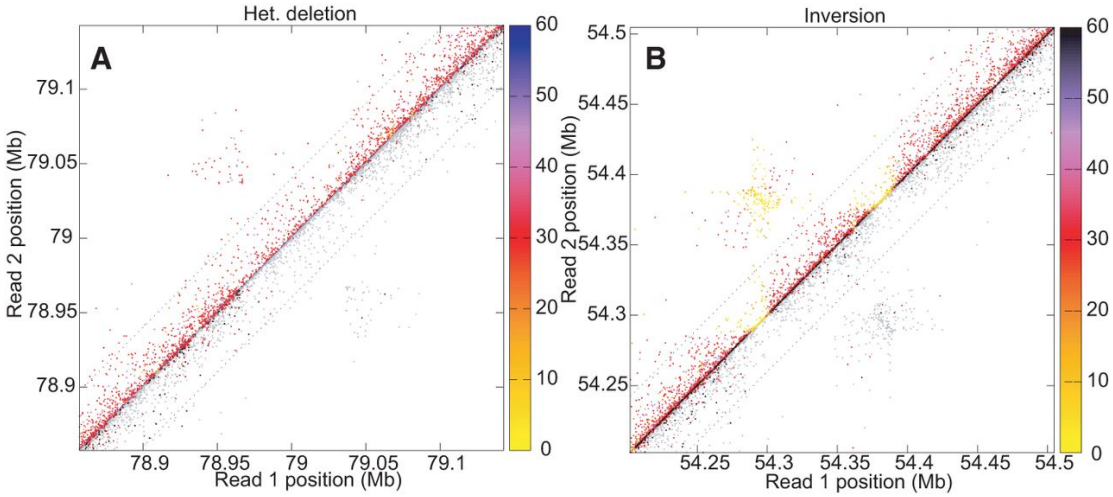

Figure 2.4: The mapped locations on the GRCh38 reference sequence of Chicago read pairs are plotted in the vicinity of structural differences between GM12878 and the reference (A, deletion; B, inversion). Each Chicago pair is represented both above and below the diagonal. Above the diagonal, color indicates map quality score on the scale shown; below the diagonal, colors indicate the inferred haplotype phase of Chicago pairs based on overlap with phased SNPs, with read pairs of unknown haplotype origin shown in gray.

identifying structural differences, we tested a simple maximum likelihood discriminator

(Methods) on simulated data sets constructed to simulate the effect of heterozygous inversions. We constructed the test data by randomly selecting intervals of a defined length L from the mapping of our Chicago GM12878 reads to the GRCh38 reference sequence, assigning each Chicago read pair independently at random to the inverted or reference haplotype, and editing the mapped coordinates accordingly. Nonallelic homologous recombination is responsible for much of the structural variation observed in human genomes, resulting in many variant breakpoints that occur in long blocks of repeated sequence [42]. We simulated the effect of varying lengths of repetitive sequence surrounding the inversion breakpoints by removing all reads mapped to within a distance W of them. In the absence of repetitive sequences at the inversion breakpoints, we found that for 1-, 2-, and 5-kbp inversions, respectively, the sensitivities (specificities) were 0.76 (0.88), 0.89 (0.89), and 0.97 (0.94), respectively. Simulating 1-kbp regions of repetitive (unmappable) sequence at the inversion breakpoints, the sensitivity (specificity) for 5-kbp inversions was 0.81 (0.76).

2.4 Discussion

We have described an in vitro method for generating long-range mate-pair data that improves the scaffolding of de novo assembled genomes from high-throughput sequencing data. This approach has several advantages over existing methods.

First, Chicago library construction requires no living biological material, namely, no primary or transformed tissue culture or living organism. The libraries described here

were each generated from 500 ng to 5 g of input DNA. Furthermore, although the in vitro chromatin reconstitution is based on human histones and chromatin assembly factors, DNA samples from a wide variety of plants, animals, and microbes can be substrates for in vitro chromatin assembly using the protocol described. Our production facility has successfully generated Chicago libraries from several plants, prokaryotes, and vertebrate and invertebrate animals. As expected for histones that indiscriminately bind DNA, the chief considerations for successful in vitro chromatin assembly are the purity of the DNA and not its biological source.

Second, because Chicago data are generated from proximity ligation of chromatin assembled in vitro rather than chromatin obtained from in vivo sources, there is no confounding biological signal (e.g., telomeric clustering or chromatin looping) to potentially confuse the assembly. As expected for in vitro assembled chromatin, we find a low background rate of noise and a virtual absence of persistent and spurious read pairs. Supplemental Figure S3 shows a comparison of the distribution of equal numbers of Chicago and Hi-C pairs in a 4-Mb region of the human genome.

Third, in contrast to in vivo Hi-C methods, the maximum separation of the read pairs generated is limited only by the molecular weight of the input DNA. This has allowed us to generate contiguous scaffolding of vertebrate genomes using just short fragment Illumina sequence plus Chicago libraries. To date, high-quality scaffolding based on in vivo Hi-C libraries has started from assemblies with an order of magnitude more scaffold contiguity than the 30-kbp N50 input contigs successfully scaffolded by Chicago HiRise. Nevertheless, it remains the case that the Chicago libraries we have

generated do not span all difficult-to-assemble regions. Centromeres, for example, are typically >1 Mb in size in the human genome. The smallest centromere in the human genome is on the Y Chromosome and is estimated to span 300 kb [59]. In our experience, we can reliably prepare DNA spanning up to 150 kb from commercially available high-molecular-weight kits. DNA extraction and preparation methods that recover clean DNA of larger sizes have been described. However, we find that the high-molecular-weight kits provide DNA that allows for an attractive combination of speed, reliability, flexibility in input sample requirements, and performance in the Chicago protocol.

Fourth, these libraries eliminate the need for creating and sequencing a combination of long-range mate-pair and fosmid libraries and do not require the use of expensive, specialized equipment for shearing or size-selecting high-molecular-weight DNA that is normally required to create such libraries. Our approach thus greatly simplifies genome assembly as a single library is generated that spans short, medium, and long-range connectivity–up to the size of the input DNA.

In summary, we have presented simple DNA library construction and associated bioinformatic methods that generate significantly longer-range genome assembly scaffolds than existing methods. Furthermore, we have demonstrated the usefulness of our data for the discovery of structural genome variation. Our methods and results mark a substantial step toward the goal of accurate reconstruction of full-length haplotype-resolved chromosome sequences with low effort and cost.

2.5 Methods

2.5.1 DNA preparation

DNA was extracted with Qiagen blood and cell midi kits according to the manufacturer's instructions. Briefly, cells were lysed and centrifuged to isolate the nuclei. The nuclei were further digested with a combination of Proteinase K and RNase A. The DNA was bound to a Qiagen genomic column, washed, eluted and precipitated in isopropanol, and pelleted by centrifugation. After drying, the pellet was resuspended in 200μl TE (Qiagen).

2.5.2 Chromatin assembly

Chromatin was assembled overnight at 27C from genomic DNA using the Active Motif in vitro chromatin assembly kit. Following incubation, 10% of the sample was used for MNase digestion to confirm successful chromatin assembly.

2.5.3 Biotinylation and restriction digestion

Chromatin was biotinylated with iodoacetyl-PEG-2-biotin (IPB). Following biotinylation, the chromatin was fixed in 1% formaldehyde for 15 min at room temperature (RT) , followed by a quench with twofold molar excess of 2.5 M glycine. Excess IPB and cross-linked glycine were removed by dialyzing chromatin in a Slide-A-Lyzer 20-KDa MWCO dialysis cassette (Pierce) against 1 liter of dialysis buffer (10 mm Tris-Cl at pH 8.0, 1 mM EDTA) for a minimum of 3 h at 4C . Subsequently, the chromatin

was digested with either MboI or MluCI in 1 CutSmart for 4 h at 37C. The chromatin was again dialyzed in a 50-KDa MWCO dialysis Flex tube (IBI Scientific no. IB48262) for 2 h at 4C and then again with fresh buffer overnight to remove enzyme as well as short, free DNA fragments.

Dynabead MyOne C1 streptavidin beads were prepared by washing and re-suspending in PBS + 0.1% Tween-20, before adding to chromatin and incubating for 1 h at RT. The beads were then concentrated on a magnetic concentrator rack, before being washed, reconcentrated, and resuspended in 100μl 1 NEBuffer 2.

2.5.4 dNTP fill-in

To prevent the labeled dNTPs (Fig. 2.1) from being captured during the fill-in reaction, unbound streptavidin sites were occupied by incubating beads in the presence of free biotin for 15 min at RT. Subsequently, the beads were washed twice before being resuspended in 100μl 1 NEBuffer 2.

Sticky ends were filled in by incubating with dNTPs, including a-S-dGTP and biotinylated dCTP along with 25 U of Klenow (no. M0210M, NEB) in 165μl total volume at 25C for 40 min. The fill-in reaction was stopped by adding 7μl of 0.5 M EDTA. The beads were then washed twice in preligation wash buffer (PLWB; 50 mM Tris at pH 7.4, 0.4% Triton X-100, 0.1 mM EDTA), before being resuspended in 100μl PLWB.

2.5.5 Ligation

Ligation was performed in at least 1 mL of T4 ligation buffer for a minimum of 4 h at 16C . A large ligation volume was used to minimize cross-ligation between different chromatin aggregates. The ligation reaction was stopped by adding 40μl of 0.5 M EDTA. The beads were concentrated and resuspended in 100μl extraction buffer (50 mM Tris-cl at pH 8.0, 1 mM EDTA, 0.2% SDS). After adding 400 g Proteinase K (no. P8102S, NEB), the beads were incubated overnight at 55C, followed by a 2-h digestion with an additional 200 g Proteinase K at 55C. DNA was recovered with SPRI beads at a 2:1 ratio, with a column purification kit, or with a phenol:chloroform extraction. DNA was eluted into low TE (10 mM Tris-Cl at pH 8.0, 0.5 mM EDTA).

2.5.6 Exonuclease digestion

DNA was next digested for 40 min at 37C with 100 U Exonuclease III (no. M0206S, NEB) to remove biotinylated free ends, followed by SPRI cleanup and elution into 101μl low TE.

2.5.7 Shearing and library prep

DNA was sheared using a Diagenode Bioruptor set to low for 60 cycles of 30 sec on/30 sec off. After shearing, the DNA was filled in with Klenow polymerase and T4 PNK (no. EK0032, Thermo Scientific) for 30 min at 20C. Following the fill-in reaction, DNA was pulled down on C1 beads that had been prepared by washing twice with Tween wash buffer before being resuspended in 200μl 2 NTB (2 M NaCl, 10 mM Tris

at pH 8.0, 0.1 mM EDTA at pH 8.0, 0.2% Triton X-100). Once the sample was added, the beads were incubated for 20 min at RT with rocking. Subsequently, unbiotinylated DNA fragments were removed by washing the beads three times before resuspending in low TE. Sequencing libraries were generated using established protocols [58].

2.5.8 Read mapping

Sequence reads were aligned with a modified version of SNAP (http://snap.cs.berkeley.edu/). Our modifications included masking out the base pairs that follow a restriction-enzyme junction (GATCGATC for MboI, AATTAATT for MluCI). Additionally, we removed the map quality penalty for read pairs that mapped to different scaffolds. PCR duplicates were marked using Novosort (http://www.novocraft.com/products/novosort/). Nonduplicate read pairs were used in analysis if both reads mapped and had a map quality score of 20 or greater.

2.5.9 Ultra-high-molecular-weight Chicago library

Human GM12878 cells (Coriell) were grown in RPMI 1640 medium supplemented with 2 mM L-glutamine and 15% FBS using recommend growth conditions to a density of 5×10^6 cells/mL. Cells were centrifuged and washed once with PBS and resuspended in ice-cold PBS at 1×10^8 cells/mL. Cells were quickly warmed up to 37C and then embedded in agarose by mixing 0.5 mL of the PBS suspension with 0.5 mL of 1.5% SeqKem LE agarose (Lonza) that had been first melted at 95C followed by cooling and maintaining at 50C. The agarose-cell suspension was rapidly aspirated in a

1-mL syringe and allowed to solidify for 60 min at 4C. The agarose plug was unmolded from the syringe and incubated twice with 50 mL of lysis solution (2% sodium lauryl sarcosine, 0.4 M EDTA at pH 8.0, 0.5 mg/mL Proteinase K [recombinant PCR grade, Roche]) for 24 h at 55C. The Proteinase K was then inactivated by incubating twice for 2 h with 50 mL of 0.1 mM PMSF at 4C followed by at least 2 h in TE50 (10 mM Tris-Cl at pH 8.0, 50 mM EDTA at pH 8.0). The agarose plug was then incubated twice for 1 h with 50 mL 0.5 KBB buffer (Sage Sciences). The small DNA fragments and contaminants were removed by performing a 16-h electrophoresis using the 5- to 80-kb waveform type using the Pippin pulse electrophoresis system (Sage Science) by loading the agarose plug in a large preparative well. The DNA-embedded agarose plug was then cut in 1-mm slices (about six to 10 slices), and each slice was incubated twice for 1 h at 4C with 400μl Mg-free MboI buffer (10 mM Tris-Cl at pH 8.0, 100 mM NaCl) and for 1 h with 400μl Mg-free MboI buffer containing 1 U MboI (Neb). Following the incubation with the MboI restriction enzyme, 5μl of 1 M MgCl2 was added to each tube and incubated for 15 min at 4C, then transferred for 30 min to 37C, and then immediately transferred on ice and supplemented with 150μl of 0.5 M EDTA (pH 8.0). The restriction enzyme was digested by adding 75μl of 10% sodium lauryl sarcosine and 15μl Proteinase K 20 mg/mL for 1 h at 37C. The Proteinase K was inactivated by replacing the solution with 500μl of 0.1 mM PMSF twice for 1 h at 4C. The agarose slices were then transferred to a 15-kDa dialysis tube with a minimum amount of 0.5 KBB and subjected to 16 h of electrophoresis using the 5- to 430-kb waveform type on the Pippin pulse electrophoresis system followed by 10 min of electrophoresis with the

opposite current direction. The dialysis tube was dialyzed three times for 1 h at 4C against 1 liter of TE. The electroeluted DNA solution was recovered from the dialysis tube and stored at 4C prior to chromatin assembly.

We generated a Chicago library from this very high-molecular-weight DNA and sequenced it on an Illumina HiSeq 2500 platform. Read processing and mapping were performed as described above.

2.5.10 De novo assemblies

The human and alligator de novo shotgun assemblies were generated with meraculous 2.0.3 [15] using publicly available short-insert and mate-pair reads [79, 34]. The alligator mate-pair reads were adapter-trimmed with Trimmomatic [10]. Some overlapping alligator short-insert reads had been merged. These were unmerged back into forward and reverse reads. The NA12878 APLG PE + MP + Fos assembly was downloaded from NCBI (BioProject accession PRJNA59877).

2.5.11 Chicago HighRise (HiRiSE) scaffolder

2.5.11.1 Input preprocessing

To exclude Chicago reads that map to highly repetitive genomic regions likely to provide misleading links, we used the depth of aligned shotgun reads to identify problematic intervals. We used a double threshold strategy: Identify all intervals of the starting assembly with mapped shotgun read depth exceeding t1 that contain at least one base with a mapped read depth exceeding t2. In practice, we set t1 and t2 such that

0.5% of the assembly was masked. We also excluded all Chicago links falling within a 1-kbp window on the genome that is linked to more than four other input contigs by at least two Chicago links.

2.5.11.2 Estimation of likelihood model parameters

Several steps of the HiRise pipeline use a likelihood model of the Chicago data to guide assembly decisions or to optimize contig order and orientation within scaffolds. The likelihood function

$$L(l1, l2, g, o) = \frac{N!}{(N-n)!}(1-P_0)^{N-n}\prod_{i=1}^{n} f(d_i) \tag{2.1}$$

gives the probability of observing the number n and implied separations of spanning Chicago pairs d_i between contigs 1 and 2, assuming the contigs have relative orientations $0 \in ++, +-, -+, --$ and are separated by a gap of length g. The function $f(x)$ is the normalized probability distribution over genomic separation distances of Chicago read pairs and is assumed to have a contribution from noise pairs that sample the genome independently.

$$f(x) = \frac{p_n}{G} + (1-p_n)f'(x) \tag{2.2}$$

is represented as a sum of exponential distributions.

To obtain robust estimates of N, p_n, G, and $f(x)$ when the available starting

assembly has limited contiguity, we first fixed an estimate of the product Np_n, the total number of noise pairs by tabulating the densities of links (defined as n/l_1l_2) for a sample of contig pairs, excluding the highest and lowest 1% of densities, and setting $N_n = G^2 \sum n_{ij} / \sum l_i l_j$, using the sum of the lengths of input contigs as the value of G. We then fit the remaining parameters in $Nf(x)$ by least squares to a histogram of observed separations of Chicago read pairs mapped to starting assembly contigs after applying a multiplicative correction factor of $G \left(\sum_{i=1}^{N_o} min(0, l_i - x) \right)^{-1}$ to the smoothed counts at separation x.

2.5.11.3 Contig–contig linking graph construction

During the assembly process, the Chicago linking data were represented as a graph in which (broken) contigs of the starting assembly are nodes and edges are labeled with a list of ordered pairs of integers, each representing the positions in the two contigs of the reads from a mapped Chicago pair. The initial steps of scaffolding were carried out in parallel on subsets of the data created by partitioning the graph into connected components by excluding edges with fewer than a threshold tL number of Chicago links. We chose tL to be the lowest integer threshold that did not lead to any clusters comprising >5% of the input contigs.

2.5.11.4 Seed scaffold construction

The iterative phase of scaffold construction was seeded by filtering the edges of the contigcontig graph and decomposing it into high-confidence linear subgraphs.

First, the contigcontig edges were filtered, and the minimum spanning forest of the filtered graph was found (see Edge Filtering below). The graph was linearized by three successive rounds of removing nodes of degree 1 followed by removal of nodes with a degree greater than 2. Each of the connected components of the resulting graph had a linear topology and defined an ordering of a subset of the input contigs. The final step in the creation of the initial scaffolds was to find the maximum likelihood choice of the contig orientations for each linear component.

2.5.11.5 Edge filtering

The following filters were applied to the edges of the contigcontig graph before linearization. Edges from promiscuous contigs were excluded. Promiscuous contigs were those for which the ratio of the degree in the graph of the corresponding node to the contig length in base pairs exceeds tp, or have links with at least tL links to more than dm other contigs. The thresholds tp and dm were selected to exclude 5% of the upper tail of the distribution of the corresponding value.

2.5.11.6 Contig orienting

Each input scaffold can have one of two orientations in the final assembly, corresponding to the base sequences of the forward and reverse, or Watson and Crick, DNA strands. The optimal orientations for the scaffolds in each linear string were found by dynamic programming using the following recursion relationship: In an ordered list of scaffolds of length n, the score of the highest-scoring sequence of orientation choices

for the scaffolds up to scaffold i, such that scaffolds $i-k-k$ to i have particular orientations o_{i-k}, o_{i-k+1}, o_i, is given by

$$S_m(i, o_{i-k}, o_{i-k+1}, o_i) = \max_{o_{i-1-k} \in [+,-]} \left(S_M(i-1, o_{i-1-k}, o_{i-k}, o_{i-1}) + \sum_{j=i-i-k}^{j=i-1} logp(o_j, o_i) \right) \tag{2.3}$$

Including links from contigs k steps back provided a significant improvement in orientation accuracy because small intercalated scaffolds might only have linking and therefore orientation information on one side, with important orientation information for the flanking scaffolds coming from links that jump over it.

2.5.11.7 Merge scaffolds within components

Contig ends were classified as free if they lie at the end of a scaffold or as buried if they were internal to a scaffold. For all pairs of contig ends within each connected component, the log likelihood ratio (LLR) score for joining them was computed with a standard gap size of g_o. These candidate joins were sorted in decreasing order of score and evaluated according to the following criteria. If both ends are free and from different scaffolds, we tested linking the two scaffolds end-to-end. If one end is buried and the other is free and if the ends are from different scaffolds, we tested inserting the scaffold of the free end into the gap adjacent into the buried end. If one or both ends is buried and if the ends are on the same scaffold, we tested inverting the portion of the scaffold between the two ends. If both ends are buried and from different scaffolds, we

tested all four ways of joining the scaffolds end-to-end. In all cases, the possible joins, insertions, and inversions were tested by computing the total change in LLR score by summing the LLR scores between all pairs of contigs affected by the change. If the change increased the LLR score, the best move was accepted.

2.5.11.8 Local order and orientation refinement

To refine both the local ordering and orientations of contigs in each scaffold, a dynamic programming algorithm was applied that slides a window of size w across the ordered and oriented contigs of each scaffold. At each position i, all the $w!2^w$ ways of ordering and orienting the contigs within the window were considered, and a score representing the optimal ordering and orientation of all the contigs up to the end of the current window position that ends with the current O&O of the contigs in the window was stored. The scores of all compatible O&Os in windows at positions $i1, i2, iw$, and the scores of the extension of their orderings with the current O&O were used. Since $w!2^w$ is such a steep function, the method is limited in practice to small values of w.

2.5.11.9 Iterative joining

After the initial scaffolds had been constructed within each connected component, the resulting scaffolds were returned to a single pool, and multiple rounds of end-to-end and intercalating scaffold joins were carried out. In each round, all pairs of scaffolds were compared, and likelihood scores were computed in parallel for end-to-end and intercalating joins. The candidate joins were then sorted, and nonconflicting joins

were accepted in decreasing order of likelihood score increase.

2.5.11.10 Break low-support joins within scaffolds

To identify and break candidate misjoins in the assembly, we used the likelihood model to compute the log likelihood change gained by joining the left and right sides of each position i of each contig in the starting assembly (i.e., the LLR,

$$L_i = ln \frac{L(g = 0)}{L(g = \infty)}, \tag{2.4}$$

for the two contigs that would be created by breaking at position i). A robust version of the support score is made by virtually masking up to n bins of size w to the left or right of candidate breakpoints, such that the bins contributing the most to the score are excluded. This score is less susceptible to misjoins mediated by repetitive sequences. When the resulting support scores fell below threshold values over a maximal internal segment of an input contig, we defined the segment as a low support segment. After merging low support segments lying within 300 bp of one another and excluding those within 1 kbp of a contig end, we either (1) introduced a break in the contig at the midpoint of the segment or (2) introduced, if the segment is longer than 1000 bp, breaks at each end of the segment.

2.5.11.11 Gap closing

HiRise can use paired-end shotgun reads to close some of the gaps of unknown sequence created when scaffolds are joined based on Chicago read pairs. Groups of reads localized by SNAP alignment to the vicinity of each such gap are passed to marauder, the gap-closing module of meraculous, which returns a gap-closing sequence when a unique closure can be inferred by local k-mer walking.

2.6 Data Access

The HiRise scaffolder source code used here (version 0.75) is available in the Supplemental Material and hosted on GitHub at https://github.com/DovetailGenomics/HiRise_July2015_GR. The Chicago reads for human L1, L2, and L3 have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession numbers SRR2911057, SRR2911058, and SRR2911066, respectively; the Chicago reads for alligator, under accession number SRR2911055. Genome assemblies have been submitted to BioProject under the following accession numbers: PRJNA306147 (NA12878 MERAC PE + MP + Fos), PRJNA305645 (NA12878 MERAC PE), PRJNA301471 (MERAC PE + HiRise 1.0), PRJNA305644 (NA12878 DISCOVAR PE + HiRise 1.0), PRJNA305314 (MERAC PE + HiRise 1.0%50%), PRJNA305315 (MERAC PE + HiRise 1.0 + L3), PRJNA305633 (DISCOVAR PE), PRJNA305630 (Alligator MERAC PE + HiRise 1.0), and PRJNA301461 (Alligator MERAC PE).

Chapter 3

Optimizing and Applying HiC

3.1 Introduction

HiC has emerged as one of the most powerful and versatile tools of the past decade for genome assembly and analysis [51, 22, 69]. Initially developed to examine the spatial organization of the genome in the nucleus, HiC and its variants have been extremely useful for a wide range of applications [9]. For example, the short-to-medium range information in the characteristic HiC insert distribution (e.g., Fig. 3.2), with or without Chicago data, works very well for genome scaffolding [73]. Furthermore, using the long-range, intrachromosomal information allows clustering of scaffolds into chromosomes (e.g., the LACHESIS described in Burton, *et al.* 2014) [13].

Outside of the genome assembly sphere, development of applications for HiC data has also progressed rapidly [53, 82]. One of the major fields that has developed from the use of HiC data is the characterization and analysis of Topologically Associating

Domains (TADs). Within these domains, interactions between genes and regulatory elements seem to occur at greater frequency than between two adjacent TADs [64]. Additionally, at least some TADs contain multiple genes with similar function and coordinated gene expression [23]. Further, many of these TADs appear to be conserved across tissue types, individuals, and even multiple species [23]. Consequently, changes in the TAD structure are often informative about changes in gene expression, even in cases where methods such as ChipSeq would be uninformative. Finally, the same concept of physical proximity (due to folding in the genome allowing for regulation across long genomic distances) can also be applied to the long-range and interchromosomal interactions.

One major improvement to HiC came in 2012, when Reza Kalhor, *et al.* , published an updated version of the protocol called Tethered Chromatin Capture (TCC) [39]. This method dramatically improved upon the original protocol, primarily through the added process of biotinylating the chromatin with a sulfhydral-reactive biotinylating reagent before immobilizing the chromatin on streptavidin coated beads (see Figure 3.1). This change allows the entire protocol to be carried out under much smaller reaction volumes, as buffers can be exchanged between reaction steps. In the original HiC protocol, each new step required that the buffer from the previous step be diluted about five-fold, geometrically increasing the reaction volumes over time. Smaller volumes ease handling, reduce the required amount of input material by 75%, and decrease reagent costs by about 60%. Smaller reaction volumes also decrease the rate of spurious ligations. Spurious ligations, those between DNA

associated with histones not actually cross-linked to each other, are the most invidious form of noise in the protocol.

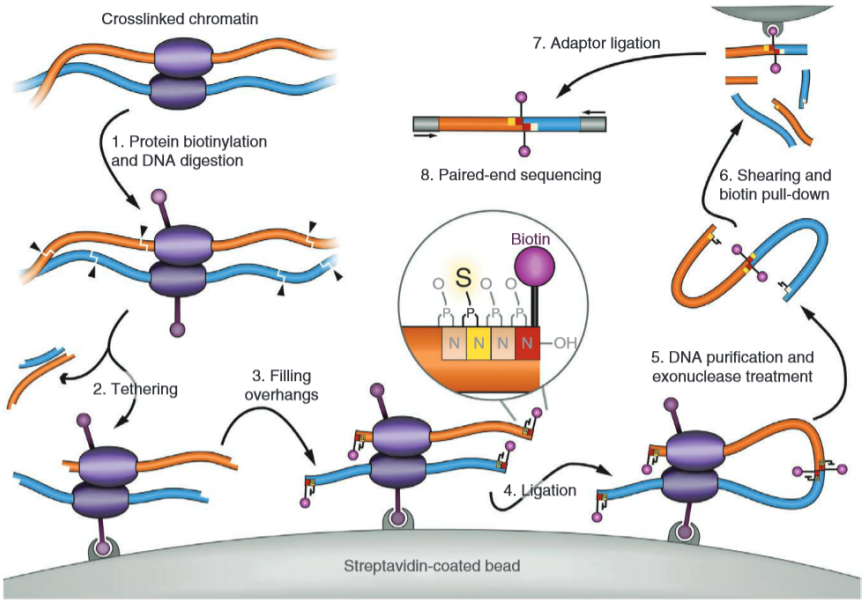

Figure 3.1: Figure 1 from Kalhor, *et al.* , 2012, showing the Tethered Chromatin Capture process.

Another recent improvement to the HiC method came in 2014 from Dr. Jay Shendure's group at the University of Washington [22]. Deng, *et al.* , switched from a restriction-enzyme based HiC to DNAse-based HiC. The change from a sequence-specific restriction enzyme to a randomly-cutting DNAse increases the amount of the genome which can be surveyed. Also, rather than using streptavidin-coated magnetic beads for immobilization of the chromosomes, Deng, *et al.* , used solid phase reversible immobilization ("SPRI") beads. Previously, HiC required dialyzing the chromatin for long periods of time to remove the excess biotinylating reagent. This extra step

is unnecessary with the SPRI method because the SPRI beads bind to DNA and, consequently, do not require any biotinylation. Furthermore, the earlier version of HiC required carefully occupying the excess streptavidin sites on the paramagnetic beads with a precise amount of biotin, prior to filling in the restriction enzyme overhangs with biotinylated nucleotides (otherwise the biotinylated nucleotides would be captured by the streptavidin). Both of these steps also contributed to a very high DNA loss rate in earlier versions of the HiC protocol. Much smaller samples may be used in the improved version of HiC. Deng, *et al.* , use only 2×10^6 to 5×10^6 cells for their version, compared to the 1×10^8 cells used by Lieberman-Aiden, *et al.* , and Kalhor, *et al.* [22, 51, 39].

It is important to note, however, that the novel uses for HiC data have outpaced the development of the method itself. The original method requires 2.5×10^7 cells per library, and uses multi-milliliter enzymatic reactions, topped off with a 300mL phenol chloroform extraction [51]. Furthermore, the protocol takes almost one week and, due to the previously mentioned logistical requirements, precludes processing more than one or two replicates or samples at a time. The method published by Kalhor, et al., in 2012 was a major improvement in terms of reducing both the input requirements, as well as reducing reaction volumes, by tethering the chromatin to a magnetic-bead solid phase via biotin-streptavidin interactions [39]. However, the added time required to biotin-label the chromatin (and clean up the biotinylation by dialysis) still effectively limits the sample throughput of this method. Consequently, my doctoral work has focused on optimizing and improving the HiC protocol, so that HiC can be efficiently and economically applied to a wide range of biological questions.

In 2014, Deng, et al., proposed the use of carboxylate modified SPRI beads to clean chromatin between enzymatic steps, instead of using biotin-streptavidin interactions. Working with Dovetail Genomics, we developed an improved method for Chicago [22]. Dr. Chris Troll at Dovetail observed that formaldehyde-fixed chromatin, once bound to SPRI beads, can only be removed by digesting the histones and also reversing the crosslinks in contrast to the Deng *et al.* [1]. I hypothesized that the new, SPRI-based Chicago method could be adapted to HiC. The method proved readily adaptable, with most of the changes focused on the preparation and quantification of the chromatin from cells, as opposed to the *in vitro* chromatin of Chicago (see Section 3.2 for the results). Quantifying the DNA concentration is necessary to avoid overloading the SPRI beads, as that would cause a major increase in the rate of chimeric read pairs in the final library and, at worst, result in complete failure of the library.[2]

I continued to optimize the HiC version of the Chicago protocol developed in conjunction with Dovetail Genomics and, when Dovetail began developing their own commercial HiC method, I collaborated with Dr. Troll to continue improving the method. Specifically, we worked to generalize the protocol to facilitate making HiC libraries from tissue, as opposed to being limited to cell culture samples.[3] Consequently,

[1] An implication of this discovery was that DNA that was not associated with chromatin could, in fact, be separated from the chromatin simply by washing the SPRI beads with an aqueous buffer. This allows for less than ideally preserved samples to still be processed into HiC libraries. At the time of Dr. Troll's observation, Dovetail was concentrating on developing Chicago, rather than HiC.

[2] Absolute quantification of the DNA concentration is not necessary, nor is it feasible since the histones mask fluorophore binding and affect spectrophotometery in unpredictable ways. However, from some testing I performed in 2014 using *in vitro* chromatin from a known quantity of DNA, it appears that in most cases using fluorometric measurement (e.g., Qubit Fluorometer) gives a result for chromatin that is somewhere between 1/4 and 1/5 the actual DNA concentration. This has proven to be sufficiently accurate so as to not overload the SPRI beads, both in my own work and based on my personal communications with the researchers at Dovetail.

[3] When adapting HiC to different sample types, nearly all of the changes happen during the first few

I first adapted the protocol to use liver (for example, in Puma, Rhesus Macaque, Burmese python, etc.) and later other tissue types including muscle, blood (e.g., Nicobar Pigeon), and even sperm (see Chapters 4 and 5 for more details). Several of these adaptations are described in Appendix A. Dovetail has also developed its own proprietary protocol, which uses relatively expensive equipment. My protocol for chromatin extraction from tissue uses 1mm diameter garnet beads in a bead-beating step to homogenize the tissue in lysis buffer. The method has the advantage of being easy and can be performed without the need for expensive equipment.

Most recently, I needed human somatic HiC data for haplotype phasing (described in more detail in Chapter 4). I hypothesized that I could obtain the human somatic HiC data by optimizing the HiC protocol for saliva. Spit samples: a) are easier to obtain than blood or tissue, and b) were already approved by the IRB for the project. While saliva was not an obvious choice, it turns out that saliva yields as much or more chromatin as blood samples from a healthy patient. Specifically, 200 μl of saliva was sufficient to make 4 HiC library replicates, at $>$ 100M unique reads each, compared to 100μl of blood making only 1 HiC library of similar complexity. This is due to the large number of white blood cells present in saliva [83]. One necessary addition to the protocol was to centrifuge the sample at low speed to fractionate out any particulate, including food detritus and plaque. The libraries prepared in this fashion work better than blood samples, while the test libraries prepared without centrifugation worked so poorly as to be unsequenceable. Currently, I would highly recommend using saliva for

steps, when the chromatin is extracted and bound to the solid phase.

any non-tissue specific HiC sequencing, both for ease of sampling and the high quality of the sample.

The latest developments that I have introduced to our in-lab HiC method, while primarily logistical, are very useful. First, I modified the proximity ligation reaction to reduce the volume to 200 μL from 250 μL. This is 1:10 the volume used in the original HiC protocol[51]. Since the relative positions of the chromatin aggregates are fixed as soon as they are first bound to the solid phase in my HiC protocol, the large volume has little effect on the overall chimerism rate. The reduced volume allows the entire protocol to be performed on a PCR plate using multichannel pipettes, with the exception of sonicating to fragment the DNA after proximity ligation and crosslink reversal. This modification allows for vastly increased sample and replicate throughput. Replicates are very important in HiC due to the high degree of stochasticity between replicates. Thus, the use of plate-format and multichannel pipettes are both quality of life improvements and also allowed me to make 6-8 replicates on some of my more recent libraries. As a result, I could choose the highest quality (and best complexity) replicates, rather than sequencing all the libraries that I made.

The other major new improvement has been adapting the Tn5 Transposase library preparation protocol to function with HiC [67]. This has been an elusive goal for almost two years, and I was only recently able to produce working HiC libraries with transposase. There are several reasons why using Tn5 or some other hyperactive transposase is preferable to sonication for HiC library preparation. First, Tn5 is much faster, with the library preparation method I used taking about two hours less than the

NEB Ultra II library preparation kit – the previous method of choice for both Prof. Green's lab and Dovetail Genomics. Transposase is known to more efficiently convert template DNA into sequencing libraries compared to sonication-and-ligation methods [67]. Since I was able to use transposase produced at UCSC, furthermore, the cost of the library is cut nearly in half compared to using the Ultra II method. Given that the Ultra II kit costs as much per sample as the rest of the reagents combined (about $25 per sample for the Ultra II kit), large numbers of technical replicates would quickly become both cost- and labor- prohibitive. One additional useful feature of transposase for library preparation is that it does not attach sequencing adapters to the free ends of the template DNA. In HiC library preparation, free ends are one of the major potential sources of noise, so the transposase actually yields a higher quality HiC library. Finally, some sequence bias has been reported in Tn5-based libraries. However, this bias is of minor concern because restriction enzyme-based HiC already has far more bias from the choice of restriction enzyme.

Overall, the recent logistical improvements permit my current version of HiC to be run in a PCR plate and does not require transfer to sonicator tubes during the library preparation phase of the protocol. As a result, the protocol can easily be completed in two days at a sharply reduced cost, even compared to the earlier version developed in collaboration with Dovetail Genomics. Finally, the protocol can easily be automated on low-cost robotic platforms (e.g., the $3000 Opentrons robot). This optimized protocol will, hopefully, lead to a profusion of new HiC data sets. Indeed, after changing the volume to allow for plate-format HiC libraries, in two weeks in late May 2017, I nearly

doubled the total number of HIC libraries that I had made in the previous two *years* of work.

3.2 Using HiC data to examine aging with the ICE-mouse model

3.2.1 Introduction

One of the long-held theories of aging posits that part of the aging process is due to DNA damage and the accumulation of novel somatic mutations. More recently, it has been suggested that chromatin organization changes with age. Dr. Motoshi Hayano, while at Professor David Sinclair's lab at Harvard University, developed an inducible, accelerated-aging mouse model called "ICE" (for Inducible Changes in the Epigenome). This model uses the addition of I-PpoI, a homing endonuclease from *Physarum polycephalum*. I-PpoI cleaves at a 15bp semi-palindromic DNA site and leaves a short, 4bp overhang. This serves to introduce mild DNA damage (in the case of the mm10 mouse genome assembly, there are 18 I-PpoI sites). The I-PpoI expression can then be turned off, allowing the cells to recover from the DNA damage. Any epigenetic changes can then be quantified.

Previous work by Jae-Hyun Yang (a graduate student in the Sinclair lab) on Mouse Embryonic Fibroblasts ("MEFs") established some important information about the ICE model. First, the I-PpoI mediated DNA damage does not activate cell-cycle checkpoints. Furthermore, it does not lead to cellular senescence, nor to apoptosis.

Additionally, the DNA damage machinery remains functional, with most cells retaining stable genomes after DSB repair. However, Jae found that gene-expression patterns and the chromatin landscape (as assayed by an MNase sensitivity assay) do change after the I-PpoI cycle, along with post-translational modification of histones. Some of these post-translational modifications are associated with aging, premature aging, and cellular senescence. Most importantly, ICE mice show all signs of accelerated aging alongside the MEF cultures.

Our goal was to create a HiC dataset comparing the ICE mice with the wild-type mice to ascertain which parts of the genome would be affected by the modifications to the epigenetic landscape. One of the features of interest were Topologically Associating Domains ("TADs")[30, 72]. Previous studies have shown a linkage between shifting the boundaries of TADs (or breaking them) and genetic diseases including cancers. Vietri Rudan, *et al.* , 2015 compared HiC libraries across four mammalian species and observed a high degree of stability across lineages, despite genomic rearrangements [87, 25]. Consequently, changes in the murine TAD landscape might have implications in human health. More generally, the chromatin contacts have been shown to have regulatory effects [6]

I created a set of HiC libraries for the I-PpoI ("ICE") mice cell lines and for control lines, using cells provided by the Sinclair lab. After sequencing, I then analyzed the results to see if there were detectable changes in the genome architecture. I found an overwhelming number of changes, both in the location of interactions in the nucleus, and in changes to the strength of shared interacting loci. While analysing the results, I

Sequencing Results

Sample	Replicate	Reads Collected	Unique Mapping	% Containing Junction
1 (‘A’)	1	63253847	25251604	29.50%
	2	41136273	24183842	49.30%
2 (‘B’)	1	57332243	22434744	31.50%
	2	60010723	30667134	40.24%
3 (‘C’)	1	60937660	0.0%	35.12%
	2	40399818	0.0	43.18%
4 (‘D’)	1	46015587	0.0	50.61%
	2	39252495	0.0	42.02%
5 (‘E’)	1	102242659	0.0	31.80%
6 (‘F’)	1	136142189	0.0	44.99%

Table 3.1: Replicates and sequencing statistics for ICE and control cell-line HiC libraries.

was fortunate to have some data from previous experiments (ChIP-Seq, RNA-Seq, etc.) previously performed by the Sinclair lab.

3.2.2 Library preparation and sequencing

The Sinclair lab sent four tubes each of six MEF cell lines: 3 ‘Cre’ and 3 ‘Cre/I-PpoI’. Two replicates of each of the six samples received were prepared as SPRI-C libraries. The libraries were sequenced for QC purposes on Professor Shapiro’s MiSeq, and then mapped to the mm10 reference genome to check complexity and the ‘HiC-ness’ of the insert distribution (Figure 3.2). Two of the replicates were removed, one because it had a very low proportion of proximity ligation junctions, and the other because it was very low complexity. The remaining 10 libraries were pooled and sent for sequencing on the NextSeq for paired–end sequencing (see Table 3.1).

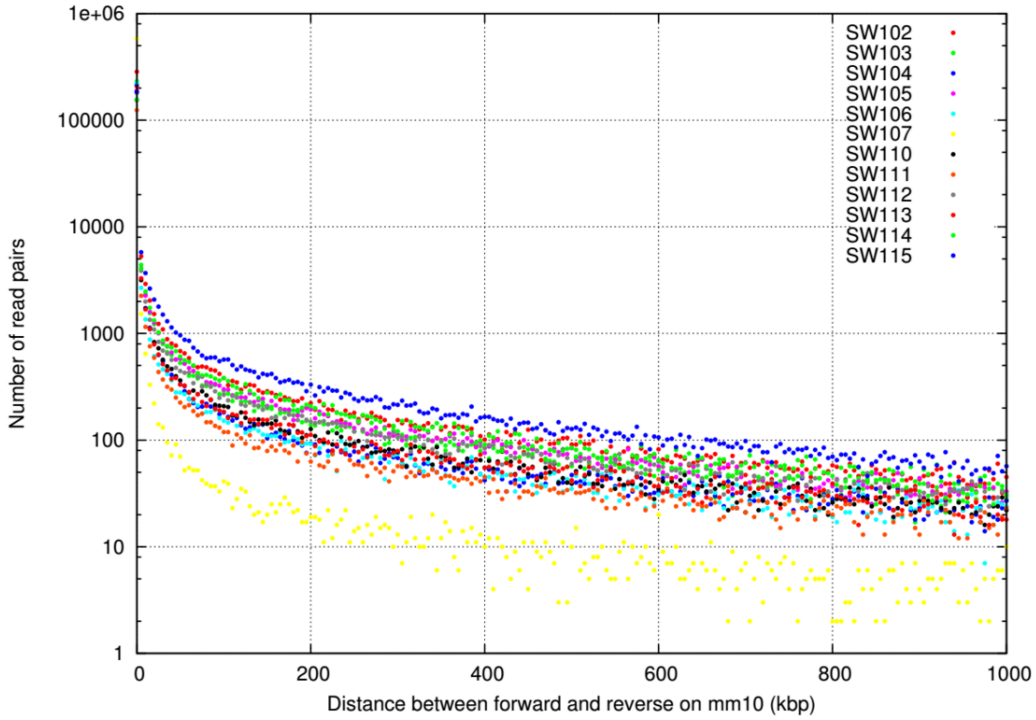

Figure 3.2: SPRI-C library insert distributions from MiSeq QC sequencing.

3.2.3 Data processing with HOMER

3.2.3.1 Individual Replicates

The sequencing data were mapped using BWA to the mm10 reference genome [49]. The resulting BAM files were further filtered to remove all read pairs with one or both reads mapping at less than Q20. The data were then input into HOMER, a software package for, among other purposes, HiC data analysis. Each of the replicates had a separate 'Tag Directory' (HOMER's method for storing interaction information). Each tag directory was further processed to remove paired-end background reads, self-

ligation reads, and reads from regions where there were an extremely high number of reads mapping to that region. The parameters for all of these steps were those suggested by the HOMER manual. After determining which parameters seemed to work with our data, an undergraduate researcher (Raquel Figeroua) ran HOMER to look for significant interactions at 10k, 100kb and 500kb resolution sizes to look for significant interactions for all pair-wise comparisons (i.e., A-D, A-E, A-F, B-D, etc.). Ms. Figeroua then cross-referenced the significant interactions files for all of the comparisons to ascertain which interactions were common across the data. I observed that while the replicates with the most reads had the most significant interactions, they also contained all the significant interactions of the lower-coverage replicates. These observations suggest the primary reason that some of the samples had fewer detected significant interactions was a lack of coverage for the individual replicates.

3.2.3.2 Combined analysis

To ascertain if combining all the samples could yield more informative results, I combined all of the unfiltered tag directories for the group containing the wild-type mice, 'A', 'B', and 'C', and the group containing the ICE mice, 'D', 'E', and 'F'. I re-filtered the combined data sets. For the 'ABC' data set, 74% of the tags remained after filtering. For the DEF data set, 72% of the tags remained after filtering. Thus, for the ABC set, there were a total of 106 million distinct tags after filtering, versus 100 million for the 'DEF' set. Next, I used HOMER's analyzeHiC feature to compare the

DEF (ICE mice) versus the ABC control set, for 500kbp, 100kbp, and 10kbp.[4]

Window Resolution	Significant Interactions
500kbp	58951
100kbp	44699
10kbp	3470

Table 3.2: Significant interactions found at different resolution levels with HOMER.

I looked in the significant interactions to determine the presence of regions corresponding to the specific loci which Jae-Hyun Yang in Professor Sinclair's lab had previously found to potentially be affected in the ICE mice. See Table 3.3 Specific Regional Interactions for a summary of the results.

I used three different resolutions for my examination: 10kb, 100kb, and 500kb. The reason for using several resolutions is that different features are visible at different resolutions. The examination revealed several noteworthy pieces of information. First, there were only a few 10kb resolution significant interactions. More sequencing data might reveal more significant interactions, but there were only 3470 total significant interactions at 10kb resolution across the whole genome. Second, I calculated the frequency of significant interactions in the desired regions, relative to the frequency of interactions in the genome as a whole. For the 10kb resolution, 0.288% of the interactions are in the specified regions, which only make up 0.039% of the genome, or

[4]Actually, 100kb windows with a 10kbp sliding overlap for speed purposes.

7.43X enrichment over what one would expect. For 100kb regions, the specific regions are enriched 2.48X over the rest of the genome. For 500kb regions, the specific regions are enriched 17.5X over the rest of the genome. Third, a large portion of the 500kb interactions were interchromosomal interactions.

Region	Chromosome:Position	Resolution (kb)	number
Igf2	chr7:142,648,581-142,672,421	100	2
		500	15
Igfbp2	chr1:72,822,584-72,854,575	10	1
		100	7
		500	21
Icam1	chr9:21,013,960-21,030,796	10	1
		500	36
HIST1	chr13:23,506,168-23,784,373	10	3
		100	6
		500	16
HIST2	chr3:96,203,376-96,284,397	500	77
HIST3	chr11:58,938,783-58,959,390	500	23
HoxA	chr6:52,139,351-52,292,262	10	1
		100	9
		500	40
HoxB	chr11:96,242,459-96,384,331	100	8
		500	99
HoxC	chr15:102,898,684-103,059,210	10	3
		100	5
		500	64
HoxD	chr2:176,931,459-177,069,717	100	6
		500	9

Table 3.3: Significant regional interactions found using HOMER. The Hox clusters are of particular interest, since there is evicence that the clusters have their own functional chromatin domains during cellular differentiation[63]

[H]

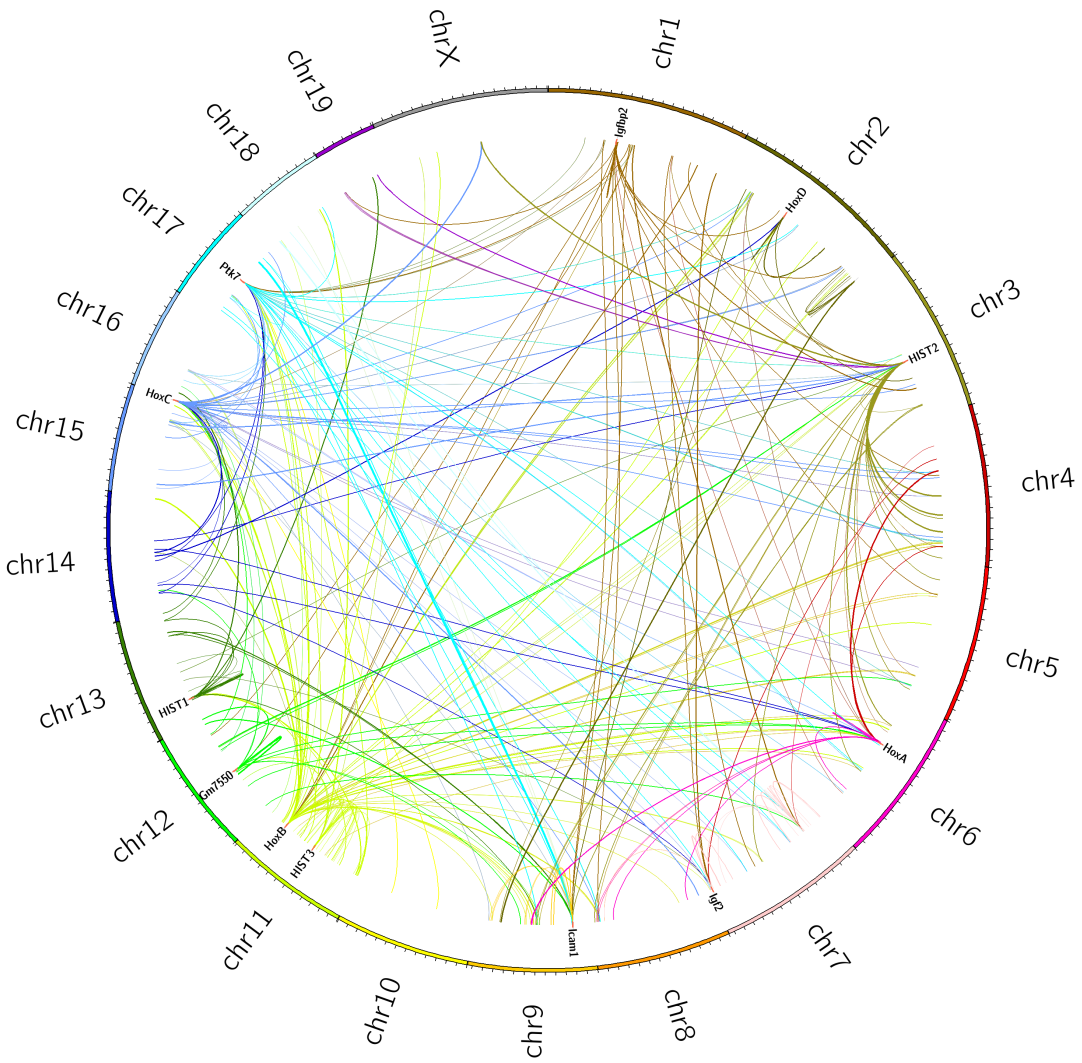

Figure 3.3: Circos diagram of all the significant interactions corresponding to the sites listed in Table 3.4. Chromosome Y, and the alternate assemblies from m10, have not been plotted. The thickness of the lines corresponds to the difference between the interaction strength in the ICE mice vs. the wild type.

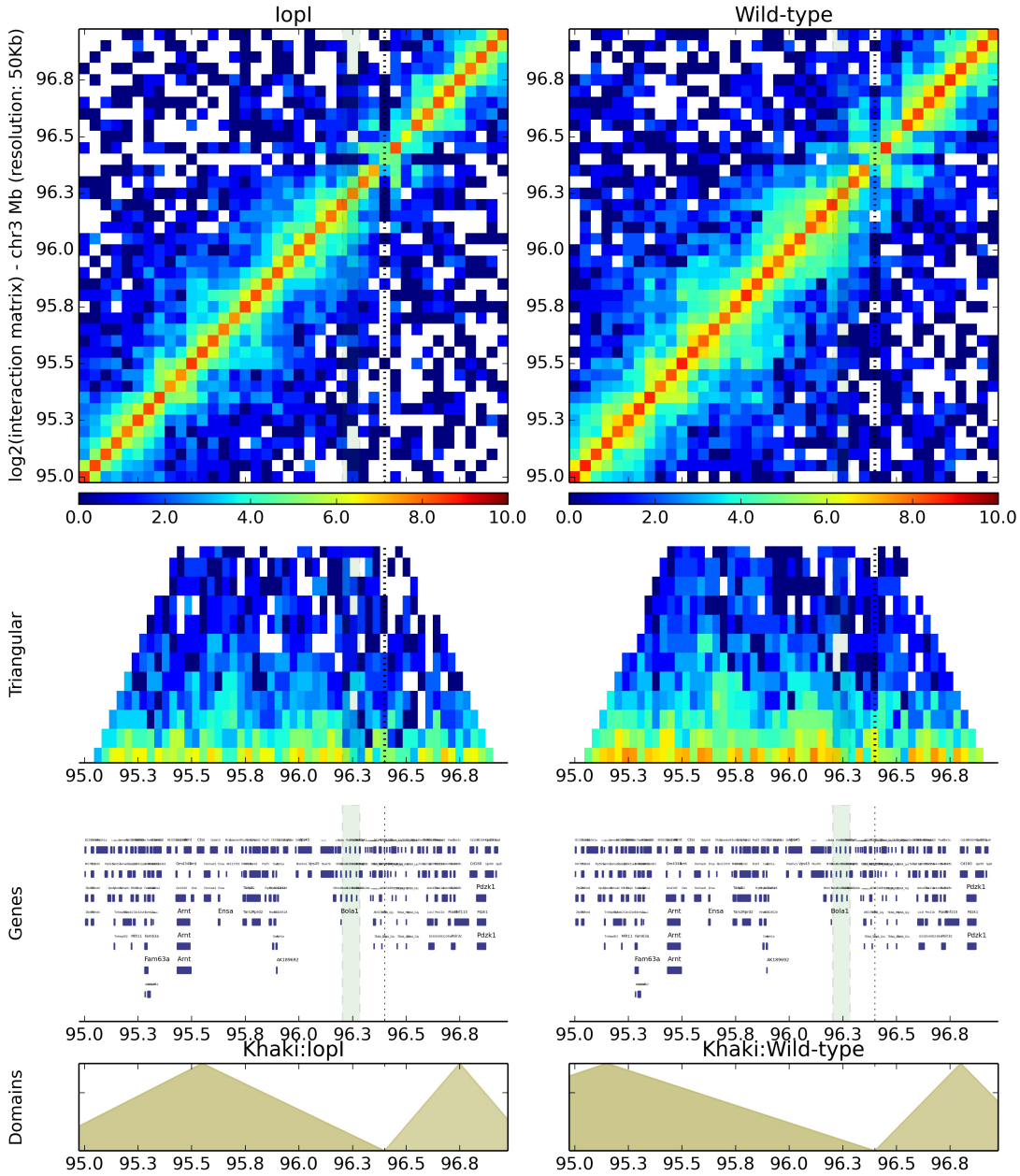

Figure 3.4: 50kb resolution heat map of the regions surrounding Hist2 (highlighted). The TAD domains are plotted in tan in the bottom boxes.

3.3 Analysis of contact maps

I used HiCPlotter to make heat maps at 50kb, 100kb, and 500kb resolutions. I used 50kb instead of 10 kb because the 10kb heat maps repeatedly crashed the software as a result of too much memory usage. I made separate heat maps for each chromosome, as well as for the previously indicated genes of interest. Each heat map includes: a) the Cre/I-PpoI combined data, b) the control ('Wild-type') data, c) a heat map showing the comparison between the two, and d) a track showing the genes for that chromosome. The chromosome-scale heat maps indicate the placement of the genes of interest. For the gene-scale heat maps, I tried to create versions with the TADs plotted. Figure 3.4 shows the HiC contact map in the region surrounding HIST2. One of the overall observations in the ICE mice versus the wild type is that there are fewer medium- and long- range interactions in the ICE mice. Furthermore, the TAD boundaries seem to be shifted in a large number of the regions surveyed. In some cases, the shift in TAD boundary puts genes previously indicated to be diffentialy expressed by Chip-Seq or other methods into different TADs, which would be consistent with regulatory changes to those genes, and overall expression changes in the pathways in which those genes participate.

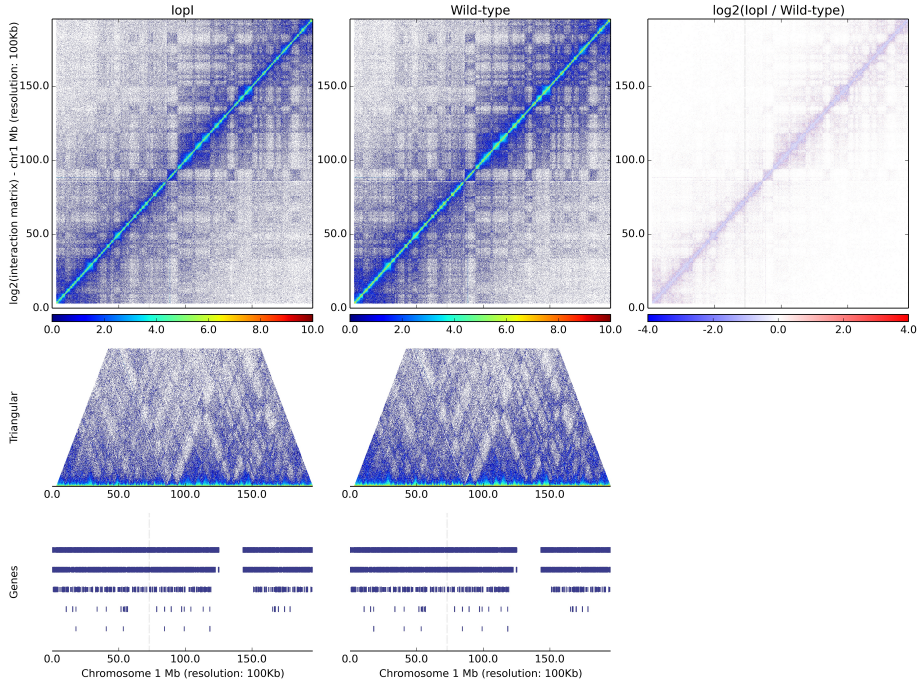

Figure 3.5: 100kb resolution heatmap of Chromosome 1. Note the major difference is that the I-PpoI has fewer long-range chromatin interactions than the wild-type cell-lines.

3.4 Other Analyses

3.4.1 10kb interchromosomal interactions

I also checked the set of significant interactions in the ICE mice at 10k resolution, by using 50kb windows with a 10kb step between windows. There were only 6 interchromosomal interactions at 10kb (see Table 3.4, and Figure 3.6). The 'LogP vs. Background' is a value that HOMER calculates. In Table 4, the instances of low values for the LogP are due to differences of only a single read, e.g., 10 vs. 11 reads between

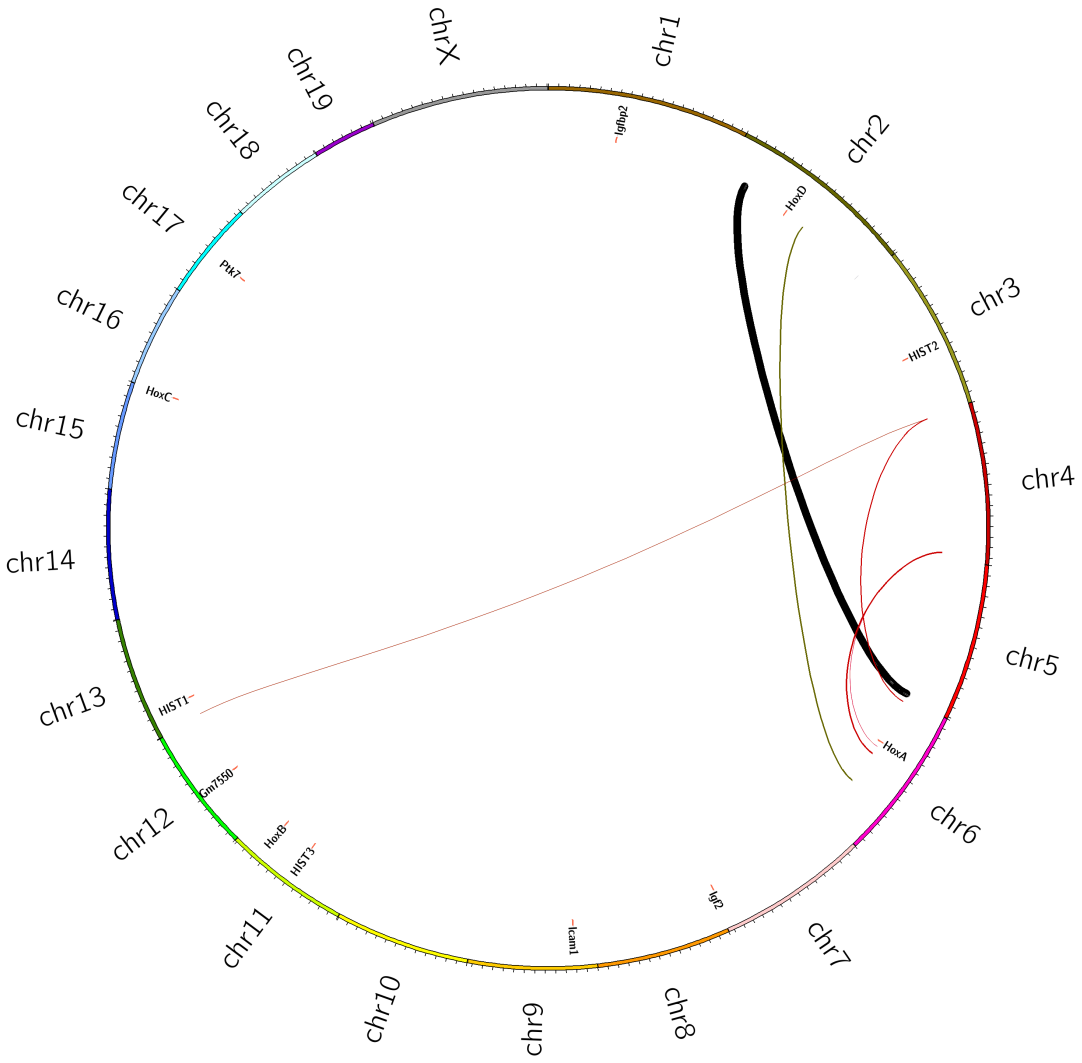

Figure 3.6: Circos plot of interchromosomal interactions at 10kb resolution. The weight of the line corresponds to the log-likelihood of the interaction.

the two data-sets. Some of these differences may be a result of structural variations or assembly errors (not shown in Table 3.4 was an interaction between an unplaced contig and chromosome 4). However, the likelihood of a 10kb interchromosomal interaction occurring at random is incredibly low (also, these data have had PCR duplicates removed during the HOMER processing).

	Position 1	Position 2		'LogP vs. Bg'	Position 1 Genes	Position 2 Genes
chr2	22740000	chr5	146260000	-24.5	Mir3967 & Apbb1ip	Apbb1ip & Cdk8
chr2	98660000	chr6	103640000	3.781	None	(In) Chl1
chr4	3240000	chr6	3190000	-2.281497	(In) Bach2	None
chr4	3240000	chr13	3000000	0.058454	(In) Bach2	None
chr4	147410000	chr6	58590000	0.039002	AB341588	(In) Abcg2
chr4	147410000	chr6	67680000	3.751408	AB341588	None

Table 3.4: 10kb Interchromosomal Interactions. 'logP vs. Bg' is the measure by which HOMER scores the confidence of interactions when given a pair of libraries to compare. Note the first interaction (between chr2 and chr5) corresponds to the extremely dark line in Figure 3.6

3.5 Discussion

In the ICE mice, the limited DNA damage caused by I-PpoI resulted in a genome-wide change in nuclear architecture. This included changes in the architecture near genes such as the Histone protein genes, and the Hox clusters. As can be seen in Figure 3.3, there are a host of interactions which change in strength between the wild-type mice and the ICE mice. Furthermore, some of the TAD changes found in the ICE mice shift genes from one domain to another. This in turn will affect the regulation and expression of these genes. One of the gratifying aspects of this project was the ability to cross-check the HiC results with Chip-Seq, RNA-Seq, etc., experiments performed by the Sinclair lab and their other collaborators.

HiC was also able to suggest several interacting regions (e.g., Table 3.4) which had not previously appeared in the Sinclair lab's research. Interaction 1 on Table 3.4 is an excellent example of an *extremely* strong new interchromosomal interaction which is between two coding regions, and yet was not detected by any of the multiple assays previously used.

3.6 Scaffolding the genome of the Atlantic herring using HiC

3.6.1 Introduction

The Atlantic Herring (*Clupea Harengus*) is one of the most abundant fish in the world. With an estimated census population of 10^{12} [27], schools number in the billions [71] and they are heavily fished. Oddly, despite their extremely high census population size, herring have a relatively modest rate of heterozygosity ($\pi = 0.3\%$) compared to terrestrial mammals with much smaller census populations [55, 27]. This makes herring an excellent species for testing theories of ecological adaptation, as even the smaller subpopulations of herring have effective population sizes (N_e) in the hundreds of thousands. Consequently, the effects of genetic drift are minimized, and genetic differentiation between the subpopulations is unambiguously the result of selection [55].

Recent work has demonstrated that herring have an extremely low mutation rate, compared to most other species. Feng, *et al.* , report a mutation rate of 2.0×10^{-9} per site per generation. This is six-fold lower than humans, and the lowest reported rate for any vertebrate [27]. A recombination map of the Atlantic herring would be very useful for future studies of population differentiation, due to the role of recombination in maintaining genetic variation and the importance of having an accurate assessment of the recombination rate for population-biology.

While Atlantic herring have been extensively studied in non-whole-genomic-scale biology, the current herring genome assembly is not very contiguous. Herring

are an ideal model for testing recombination mapping with HiC because it is relatively easy to procure herring sperm and tissue samples and, despite their low mutation rate, $\pi = 0.3\%$ is still three times higher than the human heterozygosity. As described in Chapter 4, this implies that nine times more sequencing reads will be informative for the herring than for human samples.

Using Atlantic herring samples as my study organism, I show how my improved HiC protocol allows rapid scaffolding, variant calling, haplotype phasing, and recombinatino mapping, all without recourse to other sequencing library types (the variant calling, phasing, and recombination mapping results are shown in Chapter 4). Futhermore, I show that the Atlantic herring has a lower than expected genome wide recombination rate, in addition to it's low genetic diversity for a species with such a large census population size.

3.6.2 Scaffolding the Atlantic Herring genome

As a precursor to using the herring for recombination mapping, I first needed to improve the assembly to chromosome-length. Professor Leif Andersson at Uppsala University provided matched liver and sperm samples from three male herring, and also a draft assembly (assembled using PacBio sequences). The HiC libraries were some of the first libraries I made using the multi-channel, plate-format protocol. Three to six technical replicates were produced for each sample. These libraries were quality control sequenced in-lab at 2x75bp PE on the Illumina MiSeq, and all libraries passing quality control requirements were sequenced at 2x150bp PE on the Illumina HiSeq 4000 at

UCSD IGM. The sequencing statistics are shown in Table 3.5. The initial assembly had

Table 3.5: Sequencing Statistics for Herring Somatic HiC. Sample A7 was not used due to the low coverage.

Sample	No. reads (M)	%Mapped Uniquely (M)	Read Coverage
A5L	174	62.3	32x
A6L	179	115	59.06x
A7L	35	N/A	N/A

a contig N50 of 0.5Mbp. To check the quality of the assembly prior to running HiRise, I mapped 2.0×10^7 reads to the draft assembly, and then plotted the resulting links as a heatmap (Figure 3.7). There was definitely an off-diagonal signal suggesting that the HiC data would be capable of joining scaffolds in the assembly.

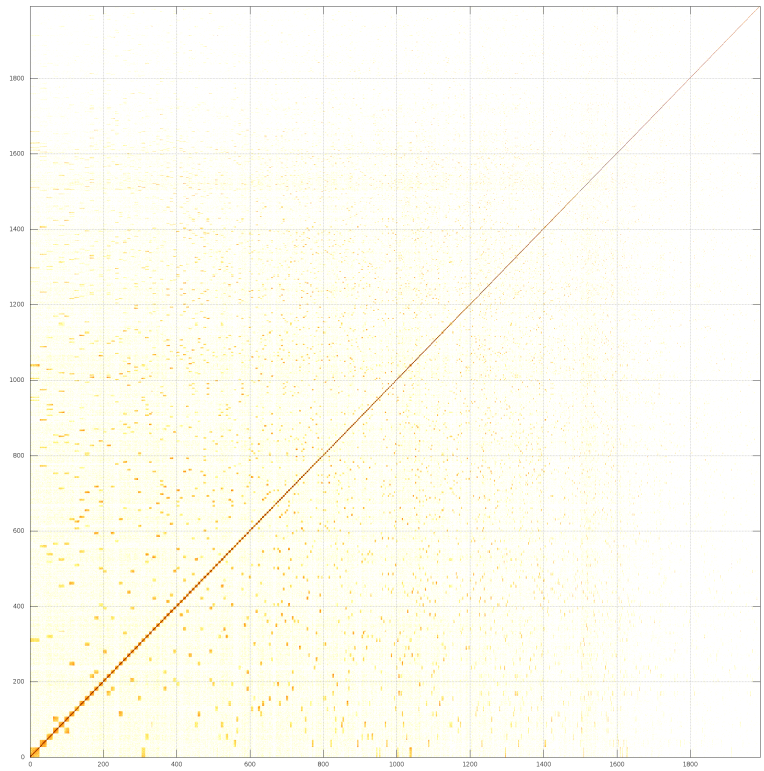

Figure 3.7: Heatmap showing link density for library A5L (liver HiC from Herring #5). Note the strong diagonal signal in the data. The off-diagonal spots are primarily indicative of joins.

After running HiRise, I looked at the link density histogram produced by HiRise during its report generation (Figure 3.9). One of the very useful improvements in the HiRise assembly was the appearance of chromosome-scale scaffolds (specifically, 26 scaffolds) corresponding to the previously reported number of chromosomes in herring.

Additionally, Professor Ed Green compared the HiRise assembly to the linkage map provided by Leif Andersson and Mats Petersson at Uppsala University. While most of the linkage groups correspond to scaffolds, there are some puzzling results, especially in linkage group 5 (Figure 3.8). This linkage group primarily comprises scaffolds 5, 11, 18, and 19, for a total of 116.1 Mbp of sequence, or just over 1/8 the entire genome. However, there are huge genetic distances between each of the scaffolds, occasionally interspersed with positions from other large scaffolds at low frequency. However, there is no evidence for linking these scaffolds in the contact plot (Figure 3.9). The other odd result is linkage group 1, which corresponds to a single scaffold (Scaffold 11), there appears to be no relationship between physical distance and genetic distance within that scaffold.

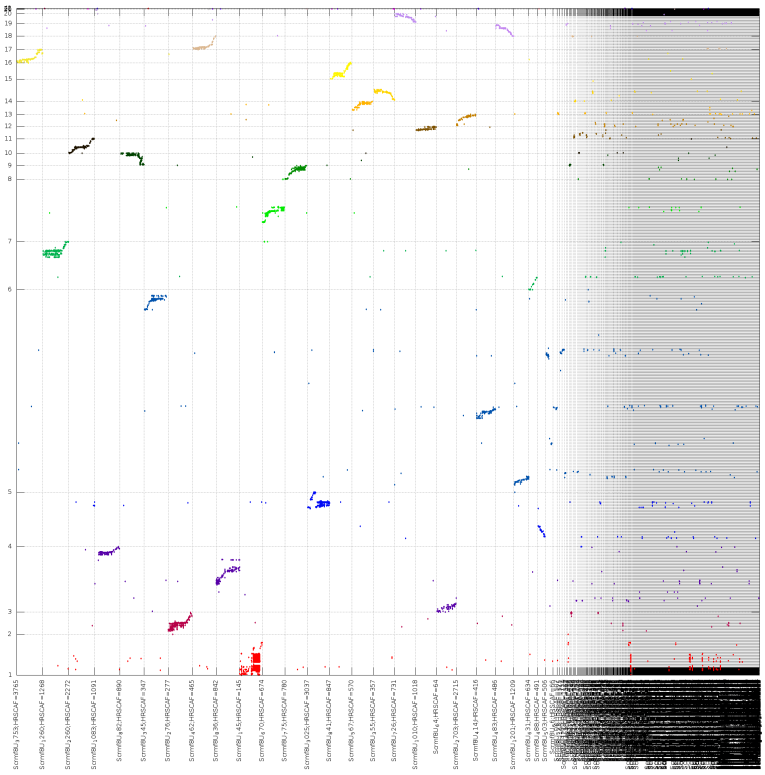

Figure 3.8: Comparison between the linkage map from Uppsala University and the HiRise assembly. The X-axis is the HiRise assembly (scaled to length in base pairs). The Y-axis are the linkage groups (scaled by genetic distance).Chromosome 10 / Linkage group 1 is known to have non-standard recombination behavior.

3.6.3 Scaffolding Results

After running two iterations of the HiRise assembler/scaffolder, the final scaffold N50 was 28.2Mbp. 50% of the genome assembly was on just 16 scaffolds. Figure 4.2 shows the length distribution of the 26 longest scaffolds. As can be seen in Figure 3.8, the linkage map generally does agree with the scaffolding results, with the exception of linkage groups 1 and 5. Based on the results shown in Figure 3.9, however, there is no support to join all of the scaffolds as suggested by linkage group 5. In the case

of linkage group 1, Professor Andersson indicated that it is known to exhibit abnormal recombination.

Assembly	L50 (Mbp)	N50
Published	1.86	113
Uppsala Draft	0.75	297
HiRise Round 1	26.1	17
HiRise Round 2	28.2	16

Table 3.6: Herring genome assembly statistics. The HiRise round 2 assembly was used for downstream variant calling, recombination rate mapping, etc.

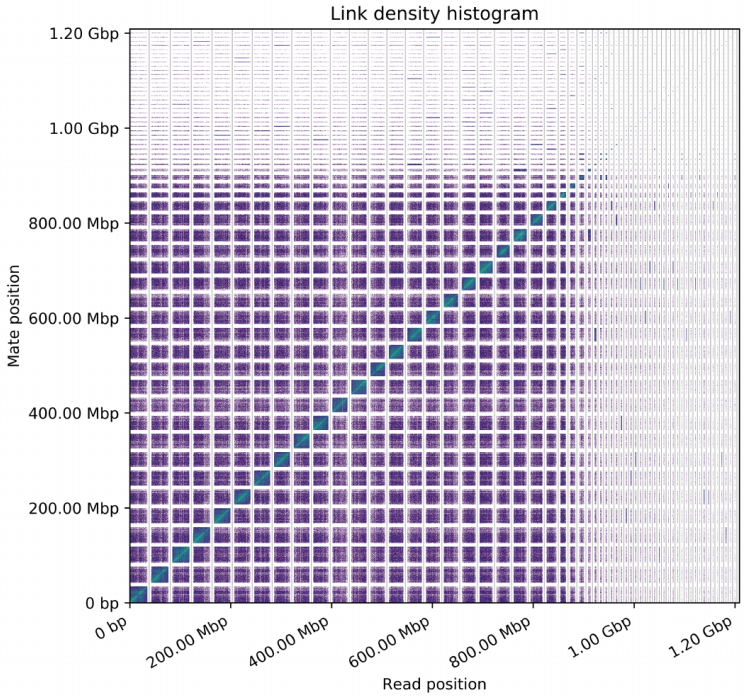

Figure 3.9: Heatmap showing link density for the combined Herring somatic HiC libraries, mapped to the HiRise v.2 assembly. There are 26 large scaffolds, corresponding to the *1n=26* chromosome number in Atlantic Herring.

The newly scaffolded Atlantic herring genome is nearly full chromosome scale. Future joins could be made using the linkage map, especially in conjunction with the HiC

data. This represents a 15-fold improvement in the median contiguity of the genome assembly, and an even larger improvement in terms of scaffolding the small contigs. It might be useful to go back and gap-fill the current draft assembly with shotgun data. This was not done in the current HiRise runs because the shotgun data were not provided.

Chapter 4

Recombination Mapping with HiC

4.1 Introduction

Sexual reproduction in diploids is only possible through meiosis. Meiosis, or the division of cells into haploid gametes, is the defining attribute of eukaryotes. Meiosis would be nearly impossible, however, without genetic recombination [48]. Genetic recombination is the process whereby homologous chromosomes exchange sections of their DNA. Recombination has two basic functions in sexually reproducing species. First, recombination is biologically necessary to ensure proper alignment of homologous chromosomes during meiosis. For sexual reproduction to be possible, a zygote (fertilized egg) must have two copies of the genome in its nucleus (assuming that it is diploid). Even haploid fungi, which would not normally be able to undergo recombination, temporarily diplodize so as to facilitate meiotic recombination [20]. Second, recombination plays a valuable role in generating genetic variation by increasing the combinations of genetic

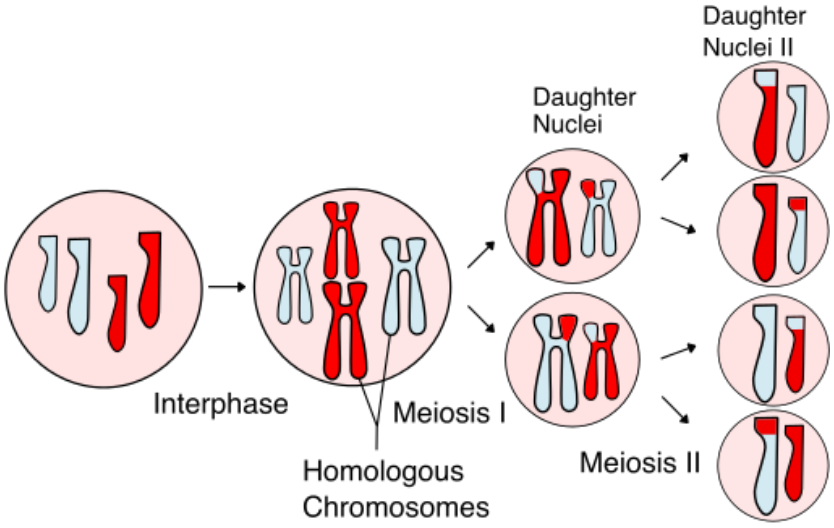

Figure 4.1: Meiosis, including recombination. Adapted from *Molecular Biology of the Genome*[90].

haplotypes between generations [90]. Consequently, the mechanisms of recombination have been extensively studied [59, 65, 5, 44, 8, 45].

The mechanisms behind meiotic recombination can be considered a specific subset of the double-stranded DNA break repair pathways [66]. The pairing of homologous chromosomes during meiosis, called synapsis, is necessary to ensure a complete haploid copy of the genome in each daughter cell. During meiosis, the diploid (2N) genome content of a cell is reduced to haploid (1N) form [90]. When the cell is in the S-phase of meiosis, the two copies of each chromosome are both replicated, resulting in a temporarily tetraploid (4N) cell (Figure 4.1). The new copies of each chromosome, called sister chromatids, remain next to each other.

Prior to the first nuclear division of the germline cell, the duplicated homologous chromosomes must pair with each other to ensure proper alignment and

assortment [90]. While paired, the chromosomes tend to wind around each other in regions of shared homology, allowing for the formation of a tetrahedral structure called a Holliday junction. The DNA of one chromosome is then broken, and re-ligated to its sister chromatid. This process of chromosomal crossover comprises the majority of recombinations.

There are currently two main methods for mapping where recombination occurs. The first method examines patterns of linkage disequilibrium, "LD", (a measurement of coinheritance of genetic markers within a population) across the genome using population sequencing data or genotype data. Historical recombinations are inferred from places where the linkage disequilibrium decays [60]. Using LD to map recombination has the advantages of extreme sensitivity and the ability to obtain sex-specific recombination maps for the population assayed. However, LD-based methods have several major limitations. First, genotype data for a large *population* must exist. Second, the recombinations that one can detect are the subset of non-lethal recombinations. Finally, the recombination map is general for the population, and may vary greatly in a specific individual[89, 8]. For many species, a population-based approach to mapping recombination is not practicable.

The other approach for mapping recombination is to genotype individual gametes, generally via single-cell sequencing [17, 89]. Sperm is normally used as the source of recombined DNA because of the relative ease of sample collection and the power of genotyping to find recombination is heavily dependent on sample size (See Fig 4.2)[5]. Until recently, single-cell genotyping more than a few hundred cells was not feasible,

Table 1

The Number of Meioses Necessary to Reject the Genome Average Recombination Rate of an Interval with 75% Power[Note]

	Hot-Spot Recombination Rate			
IntervalWidth(kb)	2×	10×	100×	500×
1	>100,000	39,500	1,500	500
10	>100,000	>100,000	4,000	1,000
20	>100,000	>100,000	4,500	1,000
100	>100,000	>100,000	6,500	1,000
1,000	>100,000	>100,000	10,500	1,000

Note.— The hot spot is assumed to be localized to a 1-kb portion of the interval. The columns correspond to the hot-spot recombination rate given as multiples of the genome average rate.

Figure 4.2: Table 1 from Arnheim, *et al.* , 2003, showing the number of meioses that must be sampled to map recombinations at a given resolution.

and even current methods are limited to not more than 10,000 cells [88]. Thus, a better method is needed for generating a recombination map from a single individual.

We hypothesized that the haplotype-informative nature of Chicago and HiC libraries could be used to infer recombination rates. We based this hypothesis on several qualities of HiC libraries. First, regardless of tissue type, both the forward and reverse reads in a HiC library read-pair originate from the same cell, and usually from the same chromosome. Consequently, the majority of read pairs are haplotype phased. Second, the insert distribution between the reads of a read-pair is non-constant, and the insert size distribution follows a predictable pattern. Empirically, the probability of any given insert distance is approximately $P(x) \approx \frac{1}{x}$ where x is an insert length. The limit for maximum HiC insert length is the length of the chromosomes. As a result, there are a fairly large proportion of long-insert reads: generally 1-3% of reads have an insert length greater than 100kbp in a "good" HiC library.

If we sequence somatic and recombined (e.g., spermatocyte) HiC libraries from a single individual, we can use the somatic library to haplotype phase the individual at

all variant sites within several hundred base-pairs of the restriction sites in their genome. Then, by comparing the haplotype phase information from the recombined library, we can use the discordance between the two libraries to estimate the rate of recombination (Figure 4.3). One of our goals was to see if it was possible to identify "hotspots" of recombination [61]

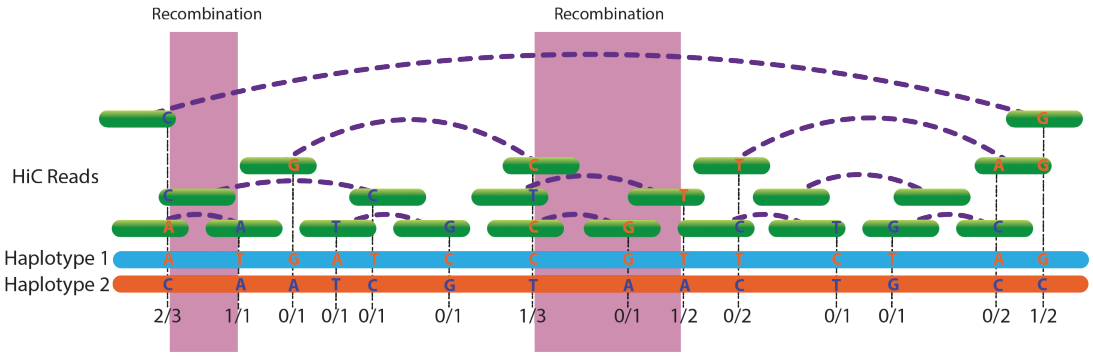

Figure 4.3: Calculating recombination rate using germline HiC data. The blue line represents 4kbp of shotgun sequence aligned to the reference genome. The green lines are 250bp HiC reads. The orange points are SNPs that are not recombined. The purple X's are reads with a recombined SNP. In this case, the total recombination rate for the region would be 714 cM/Kbp.

The probability of recombination between two sites, furthermore, is directly proportional to the distance between the sites. Thus, the long-range HiC read pairs (i.e., >10kbp) may be used to infer recombinations.

In this chapter, I will present the results of attempting to recombination map both the Atlantic herring, using the genome assembly and samples described in Chapter 3.6, and also a human sample. For the sake of clarity, the work will not be presented in the original chronological order.

4.2 Atlantic Herring Recombination mapping

4.2.1 Methods

4.2.1.1 Variant calling Atlantic Herring samples

I called variants for two of the samples (A5 and A6) which had sufficient coverage to expect that a reasonable number of variants would be detectable. I re-mapped all the reads to the new draft assembly using BWA *mem* [49]. Variant calling for the Atlantic herring samples was performed using two methods. First, I applied a script written by Prof. Ed Green, which called heterozygous SNV sites based on at least two observations of variants in each strand, of each variant (i.e., 8 reads supporting the variant call) at that site. Second, since recent versions of the popular samtools package include a simple variant caller, I performed variant calling with samtools to evaluate the accuracy of Professor Ed Green's script. I used both the somatic and the sperm HiC libraries for variant calling. I split the reference assembly into scaffolds, using the longest 26 scaffolds which roughly correspond to the herring chromosomes for variant calling and subsequent analyses. Finally, as with the human sample described in Chapter 4, splitting the reference assembly into scaffolds made haplotype phasing much easier.

4.2.1.2 Haplotype phasing Atlantic Herring samples

The Atlantic herring samples were haplotype phased using the most recent version of HapCUT2 as of Sept. 19, 2017. While both the germline and somatic data were used for variant calling to increase coverage, only the somatic data were used for

Sample	Tissue	Fraction Mitochondria
A5	Liver	0.0022
	Sperm	0.00074
A6	Liver	0.0047
	Sperm	0.00059

Table 4.1: Relative mitochondiral content of the different HiC libraries.

the haplotype phasing step. A maximum insert size of 10Mbp was applied due to the relatively small size of the herring chromosomes. A minimum base quality of 15 was required, and a threshold cutoff of 10 (i.e., 90% confidence) was applied to the output. Finally, the software was set to remove falsely-called heterozygotes. The results of this analysis are described in Section 4.2.2.1.

After haplotype phasing, I checked the concordance between the somatic libraries and the HapCUT2 results. The results are shown in Figure 4.5 (below).

4.2.1.3 Recombination mapping

I checked for possible somatic contamination of the sperm HiC libraries by comparing the mitochondrial content of the libraries. To do this, I mapped the libraries to the mitochondrial reference KC193777 (from NCBI), then removed duplicates with *samtools* and looked at the fraction of reads mapping to the mitochondrion. Ideally, there should be very few mitochondrial reads in either somatic or sperm HiC libraries, since the SPRI beads filter out non-chromatin-bound DNA. However, since sperm have the mitochondria in the tail, physically separated from the nucleus, one would expect even fewer mitochondrial reads in the sperm data**??**[11]. Table 4.1 shows that there are far fewer mitochondrial reads in the sperm HiC data than in the somatic HiC.

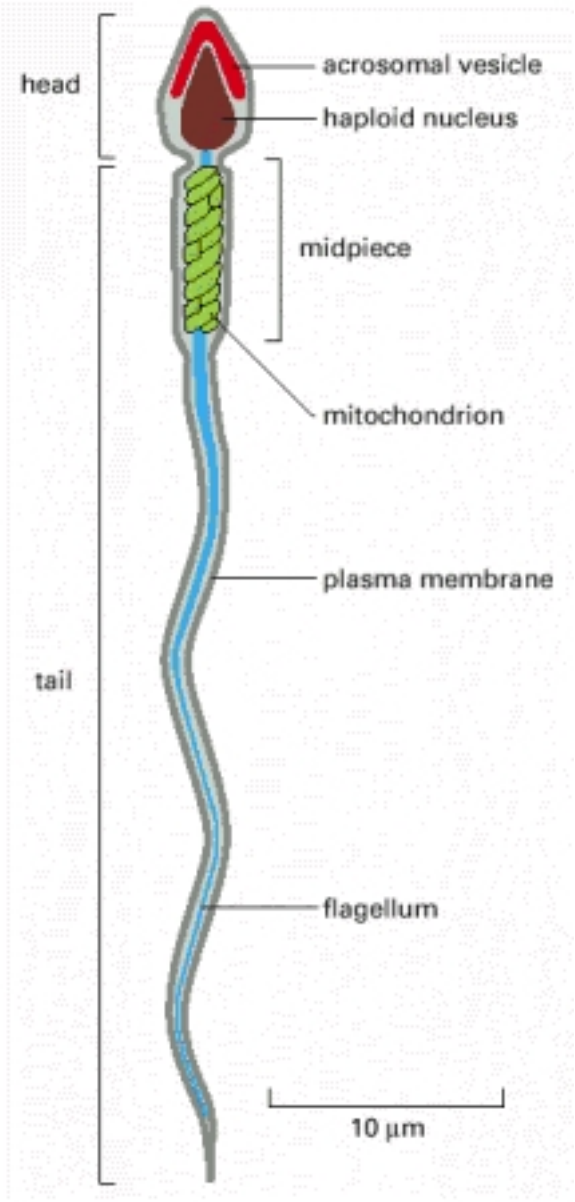

Figure 4.4: Diagram of sperm cell, showing the positioning of the mitochondria versus the nucleus. From Alberts *et al.* 2002 [11].

I ran the recombination mapping software on the 26 longest scaffolds (corresponding to the 26 chromosome pairs). I used 100kbp window sizes, and a Q50 haplotype phasing quality cutoff. The output was a bedgraph formatted file for each scaffold, with

no regions explicitly excluded (since I don't know the location of the centromeres). Of note during the mapping was that a much lower proportion of haplotype edges required pruning/repair as compared to the human data. This may be because there was only HiC data for both haplotype phasing and recombination mapping.

I added a function to the software which performs a two-tailed t-test for each window, comparing the distribution of haplotype discordance in the somatic data for that window to the proportion of discordance (presumably due to recombinations) in the germline data. The software then filters out windows with $p > 0.1$ (this parameter can be changed in the options). Under the null hypothesis (i.e., no recombination within the window), the discordance in the germline data should be approximately the same as in the somatic data, and should result from the same combination of phasing errors and chimeric read pairs. The t-test should filter out regions where a high degree of phasing errors or other unknown genomic features causes a false recombination rate. This feature is optional, since it removes most non-hotspots from the final output.

4.2.2 Results

4.2.2.1 Variant calling Atlantic Herring

Overall, the number of variant sites called was higher in the A6 sample, due to higher read coverage (Figure 4.2). The GATK in-del realignment did reduce the number of variant sites seen by around 50% (this varied between samples and scaffolds). Overall, sample A5C had 64% the number of variants as sample A6. In both cases, the number of variants found after in-del realignment was smaller than would be expected

from 0.3% heterozygosity. This discrepancy can be attributed to two factors. First, the program written by Professor Green for variant calling only considers SNPs, not insertions and deletions. When I compared the two individual samples to the reference using the Samtools v1.5 /BCFtools variant calling pipeline, I found a very large number of in-del polymorphisms relative to the reference (and indeed, internal to each of the samples). Second, I required very high coverage for calling the internally heterozygous sites because false-heterozygous sites are extremely detrimental to the recombination mapping process. For future experiments, I would incorporate the in-del polymorphisms, despite the difficulty of calling them. Both HapCUT2 and my recombination mapping software (based on the HapCUT2 file structure) are quite capable of using the in-del polymorphisms.

Most of the scaffolds had a single large haplotype block containing between 20 and 25% of the variant sites. Sample A5 had fewer total variant sites called, although it had on average about 20% more somatic read coverage. As can be seen in Figure 4.5, the somatic concordance for the herring libraries is nearly 100%. However, the germline and somatic libraries are easily differentiable in both samples. This indicates that there is recombination information in the germline datasets.

4.2.2.2 Recombination mapping the Atlantic herring

The two herring samples showed different average recombination rates, with A6 generally having a higher estimated rate than A5. This is most likely a result of more data in the case of A6, allowing for both more variant sites to be called and

Scaffold	A5 Variants	A6 Variants	A5 % Het	A6 % Het	Size (Mbp)
1	26985	43202	0.08%	0.12%	34.7
2	21817	32677	0.06%	0.09%	34.1
3	25845	40307	0.07%	0.12%	34.0
4	28624	41260	0.08%	0.12%	33.5
5	21350	33601	0.06%	0.10%	32.4
6	20768	32210	0.06%	0.09%	31.9
7	21184	33430	0.06%	0.10%	31.8
8	21903	35548	0.06%	0.10%	31.4
9	23237	34539	0.07%	0.10%	30.9
10	18052	27086	0.05%	0.08%	30.1
11	28685	47135	0.08%	0.14%	30.1
12	21296	32234	0.06%	0.09%	30.1
13	24997	39762	0.07%	0.11%	29.1
14	25200	39545	0.07%	0.11%	29.1
15	19046	30067	0.05%	0.09%	28.5
16	18732	31059	0.05%	0.09%	28.3
17	21621	34376	0.06%	0.10%	28.1
18	19842	30812	0.06%	0.09%	27.3
19	17925	26774	0.05%	0.08%	26.3
20	18942	31122	0.05%	0.09%	26.2
21	20767	29639	0.06%	0.09%	25.8
22	22900	35276	0.07%	0.10%	24.0
23	12743	18836	0.04%	0.05%	19.7
24	17316	24919	0.05%	0.07%	11.6
25	15930	25558	0.05%	0.07%	10.4
26	11694	19701	0.03%	0.06%	10.1
Average/Total	547401	850675	0.06%	0.09%	709.6

Table 4.2: Herring variant calling results.

also better ability to detect recombinations. Both of the two samples were fairly close to the expected chromosome-wide recombination rate for most of the scaffolds, generally between 0.5X and 1.5X the expected rate of 1 recombination per chromosome per meiosis. Furthermore, it was possible to detect both low-recombination and high-recombination regions when looking at the high-confidence windows. Full recombination rate maps are included as Figures B.3 and B.2.

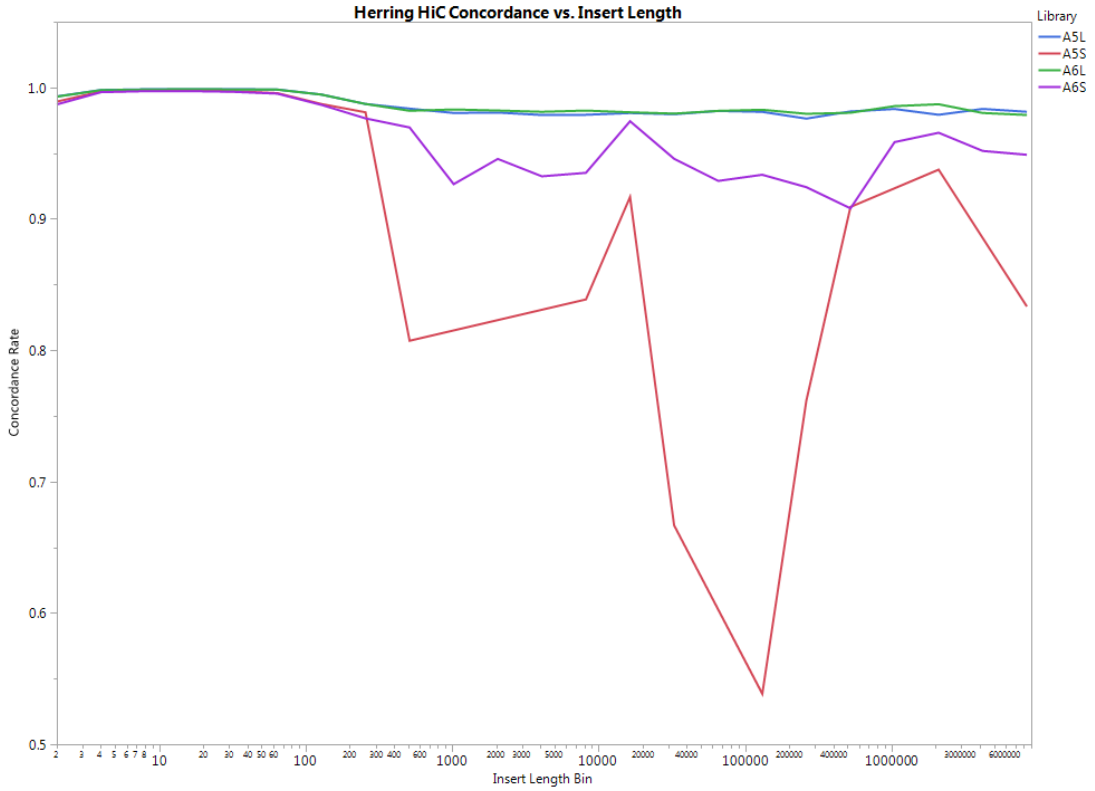

Figure 4.5: Comparing concodance versus insert length for the two herring libraries. The two lines near 1.0 are the somatic libraries. Note that the A5 germline library (orange) has lower coverage than the A6 germline library. Both libraries show lower concordance with increasing insert length when compared to the somatic libraries.

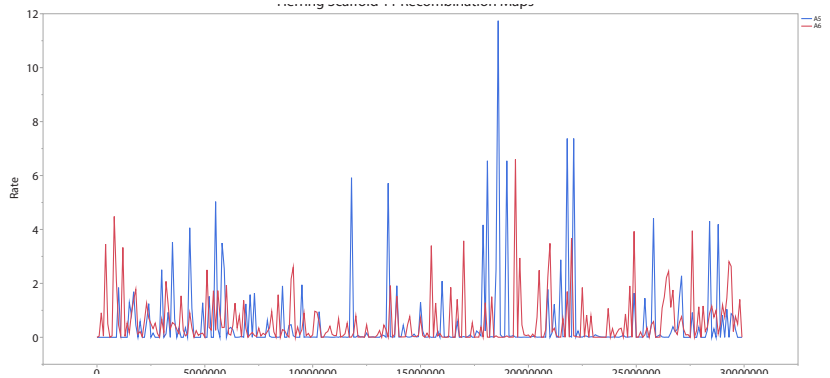

Figure 4.6: Recombination rate map for Atlantic herring Scaffold 11. Window size was 100kb, and only windows significantly different from the background noise ($p < 0.1$) are shown. Sample A5 is in blue, sample A6 is in red. The expected recombination in this case would be around 0.4cM/window.

In the case of scaffold 10 (Figure 4.7, corresponding to linkage group 1 in the linkage map (Figure 3.8), the results show some very high spikes in recombination rate (upwards of 100X the genome average). However, further exploration of the data is necessary, especially to see if there is some difference in the 3D structure of scaffold 10 compared to the rest of the genome.

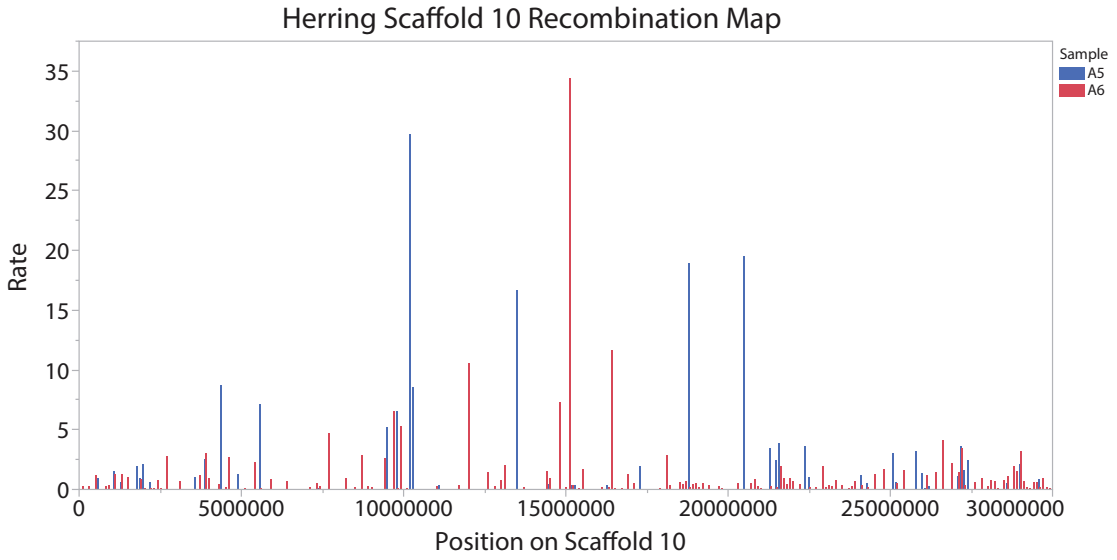

Figure 4.7: Recombination rate map for Atlantic herring Scaffold 10. Window size was 100kb, and only windows significantly different from the background noise ($p < 0.1$) are shown. Sample A5 is in blue, sample A6 is in red. This scaffold definitely shows higher peaks for the recombination rate than most of the other scaffolds.

Figure 4.8 shows one of the more typical results, with samples A5 and A6 compared. The much lower amount of germline data (about 1/3 as much) for sample A5 can be clearly seen in the lower number of significantly differentiable windows in Figure 4.8.

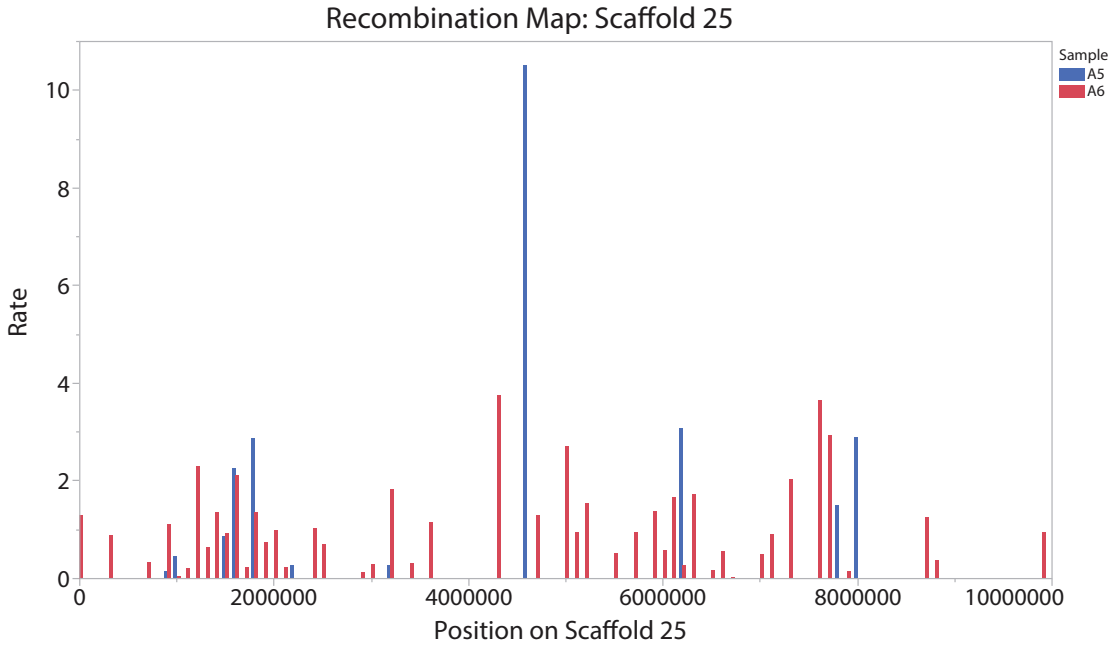

Figure 4.8: Recombination rate map for Atlantic herring Scaffold 25. Window size was 100kb, and only windows significantly different from the background noise ($p < 0.1$) are shown. Sample A5 is in blue, sample A6 is in red. Note the correlation in recombination rate between position 1000000 and 2000000.

4.2.3 Discussion

The recombination mapping results indicated a lower-than-expected CWAR (chromosome-wide-average-recombination rate). This may be a biologically relevant result in the herring, compounding the previously reported low mutation rate [27]. Alternatively, the low CWAR may be due to the lower detectable heterozygosity in these samples. Further work might include shotgun sequencing the samples to high depth, and calling variants relative to the previously published SNP datasets [55, 27].

Based on the results, it is apparent that the herring recombination map is composed of discrete hot- and cold- spots. Furthermore, unlike the human recombina-

tion data shown in Chapter 4, there is no obvious block of high discordance windows corresponding to the centromere. Additionally, the t-test removed a number of windows which had a high 'recombination' rate but relatively low read-coverage, suggesting that it is an effective addition for removing potentially erroneous results.

Future work on the herring recombination can be divided into four points. First, I would like to re-do the genotyping for the current herring samples with a genotype caller which can make use of in-del polymorphisms. Second, I would like to collect more data for samples A5 and A7 (which was not included in the above analyses due to insufficient coverage to call variants). These steps would allow improved haplotype phasing and should, consequently, result in higher resolution recombination maps. It should also permit sample A7 to be genotyped and recombination mapped. Next, I would create new HiC libraries from the herring sperm samples after first separating out any epithelial contamination by sucrose gradient centrifugation. This would ensure that only sperm cells are incorporated. It might also be useful to loosen the chromatin structure in the sperm by a lithium diiodosalicylate treatment to remove some of the histones. Finally, I would like to use the new herring genome assembly to create a better linkage map, so that I have an independent population-based method of validating the results.

4.3 Recombination Mapping in *homo sapiens*

4.3.1 Methods

4.3.1.1 Collecting human samples

After obtaining IRB approval, we advertised for volunteers at UCSC. We collected matched saliva and semen samples from five healthy male volunteers between the (approximate) ages of 25 and 40. All samples were anonymized with random 4-digit codes. We prepared genomic DNA samples from the saliva, and subsampled the semen to collect recombined DNA samples, retaining the rest for future HiC library use [50, 35].

4.3.1.2 Generating Chicago libraries

Somatic and recombined Chicago libraries were constructed using the method described in [69], with some of the early improvements described in Section 3.1. Specifically, the samples were bound to SPRI beads rather than biotinylating the chromatin and then immobilizing on streptavidin coated beads. The libraries were quality-control sequenced on Illumina MiSeq 2x75 PE in-lab, and then sequenced at UCSD IGM sequencing core facility with Illumina 2x125bp PE.

4.3.1.3 Generating HiC libraries

Two somatic HiC libraries were generated for one sample, UCSC1989, using saliva as the cell-source according to the method described in Appendix A. The sample

was initially quality control sequenced with Illumina MiSeq 2x75bp PE, then 1.5x10^7 additional reads were sequenced with MiSeq 2x300PE at UCSC. The libraries were later sequenced to approximately 3x10^8 reads with 2x150PE on the Illumina HiSeq 4000 at UCSD IGM.

Additionally, six HiC library replicates were prepared from UCSC1989's sperm sample according to the method described in Appendix A. These libraries were quality control sequenced in-lab at 2x75bp PE on the Illumina MiSeq, and the top 3 libraries (in terms of complexity, insert distribution, etc.) were sequenced at 2x150bp PE on the Illumina HiSeq 4000 at UCSD IGM.

The somatic and recombined libraries were sequenced on different lanes to prevent 'spreading-of-signals' from cross-contaminating the data [80].

4.3.1.4 Calling variants

We sent genomic DNA samples from all five human volunteers to the New York Genome Center (NYGC) for high-coverage shotgun sequencing. The samples were converted to sequencing libraries at NYGC, and sequenced at 2x125bp PE on an Illumina X-10 Sequencer. Variant calling was performed at NYGC, and confirmed via SNP-chip. The resulting .vcf files and mapped reads were downloaded into local storage at UCSC.

4.3.1.5 Haplotype Phasing with Chicago

Since Chicago reads were the only somatic data available until winter 2017, I proceeded to use the somatic Chicago data, in conjunction with the high-coverage NYGC somatic shotgun dataset, for somatic haplotype phasing. Initially, I tried HapCUT version 1.0, and later my own version of a haplotype phasing script. I eventually used HapCUT2, which is specifically intended to be HiC-type data compatible. HapCUT2 provided the best results, but even 200 million Chicago reads do not provide enough clone coverage to sufficiently haplotype phase a genome for recombination rate mapping, as will be shown below.

Several factors contributed to the problems with haplotype phasing. First, the general problem of relatively short Illumina read-lengths (150bp in this case) meant that only 2.25% of read pairs were actually haplotype informative (based on 1 SNP/kb $\times$150bp= 15% of r1 and r2 each having SNPs = 2.25%). Second, Chicago reads generally follow a distribution where P(insert distance) = 1/distance with an upper bound of 200kb. This means that only a few percent of reads (X % in the case of the main Chicago library I was testing) will actually have an insert distance >10kb, which is necessary to surpass the results of simply using shotgun reads for phasing. As a result, only about 0.1% of all the sequencing reads used will be both haplotype informative and also long enough to actually improve the phasing (e.g., 0.1% (useful reads) $\times$ 2000bp (average length) * 200M reads /3gbp genome = 13X clone coverage).

4.3.1.6 Phasing with combined Chicago and HiC

We primarily focused on samples UCSC1989 and UCSC0035, as these two samples had the most data collected during the initial sequencing. Initial haplotype phasing was performed using a combination of shotgun and Chicago data, using HapCUT2 [26]. HapCUT2 works in two stages. In the first stage, the mapped reads are processed to extract the haplotype-informative bases. In the second stage, the haplotype information is used to create a raw haplotype graph for each sequence, which is then processed into a final output consisting of phased haplotype blocks. I ran HapCUT2 on the UCSC1989 sample data using the combination of HiC and Chicago somatic libraries, with a phasing quality cutoff of 10 (90%) and with HapCUT2 set to remove from the output those sites which appear to be falsely called heterozygous. Ultimately, UCSC1989 was chosen for further analysis, due to have the most data (including somatic HiC libraries).

4.3.1.7 Quantifying and reducing haplotype phasing errors using somatic data

I decided to check the haplotype phasing results because the recombination mapping requires extremely accurate haplotype phasing. To do so, I compared the raw mapped sequencing data to the HapCUT2 phasing results. I then looked at the rate of concordance in each edge in the haplotype graph produced by HapCUT2 (Figure 4.9).

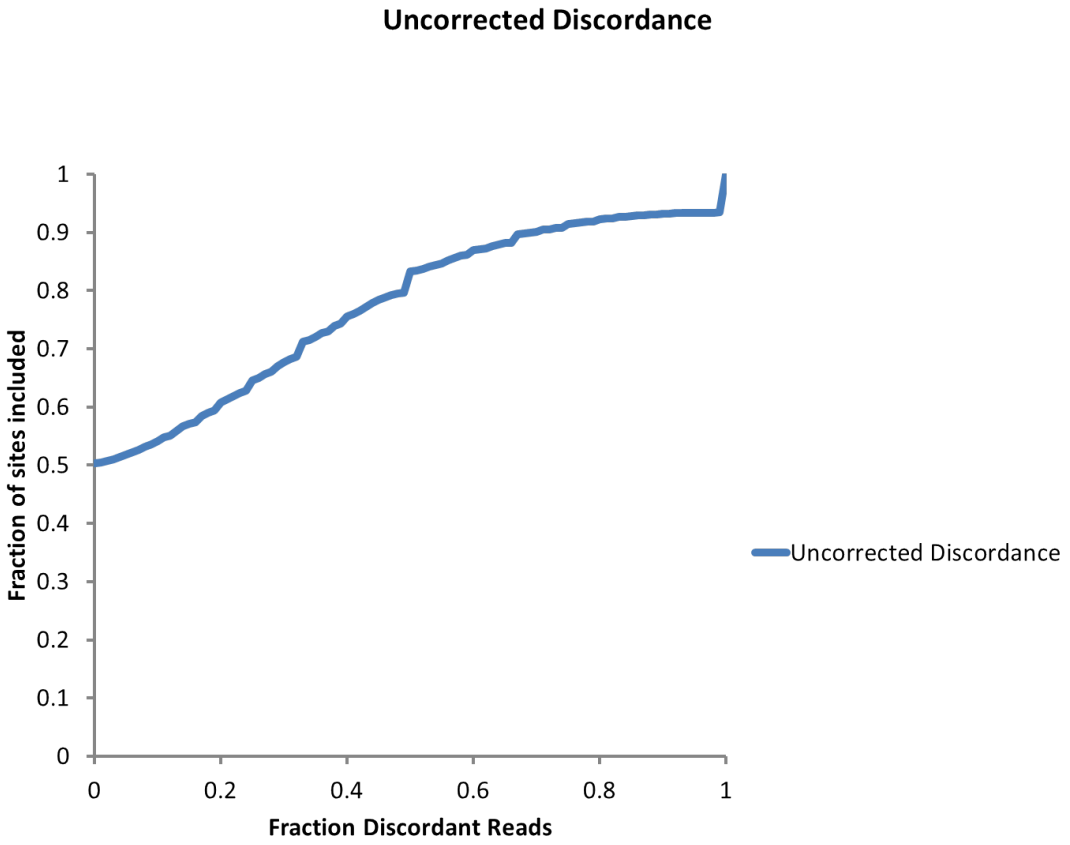

Figure 4.9: Discordance in the UCSC1989 Somatic data prior to haplotype phasing correction and pruning. Uncorrected discordance referrs to the discordance in the UCSC1989 data prior to the haplotype pruning and correction steps detailed in Section 4.3.1.7.

The results were surprising. I had expected to see very accurate results in the short range edges due to the high shotgun coverage. I would expect the long range edges to be more discordant, due to the sparser coverage of the Chicago and HiC libraries at long ranges lowering the true phasing signal relative to the chimeric noise rate. However, of the 24% of haplograph edges which had more discordant reads

than concordant reads, 8% were incorrectly phased (i.e., 100% of the observed reads between the two SNPs disagreed with the HapCUT2 output).

To ascertain which component of the data was causing these errors, I plotted the rate of discordant reads versus the edge lengths (Figure 4.10). Contrary to my expectations, the long–range edges were not the source of error. Rather, the short–range edges (i.e., the ones primarily phased by the shotgun reads) caused the bulk of the errors. I tried re-phasing with only the Chicago and HiC data to determine if the shotgun reads were responsible for the bad phasing results. To further check the phasing results, I relied on a metric reported by HapCUT2, namely the Phred-based confidence scores, that HapCUT2 reports for every variant in the phasing results. This score indicates the probability that the variant site is correctly phased relative to the rest of the haplotype phase block. While the completeness of the phasing did decrease slightly when I removed the 35X shotgun library (see Fig 4.11), the confidence scores also increased dramatically.

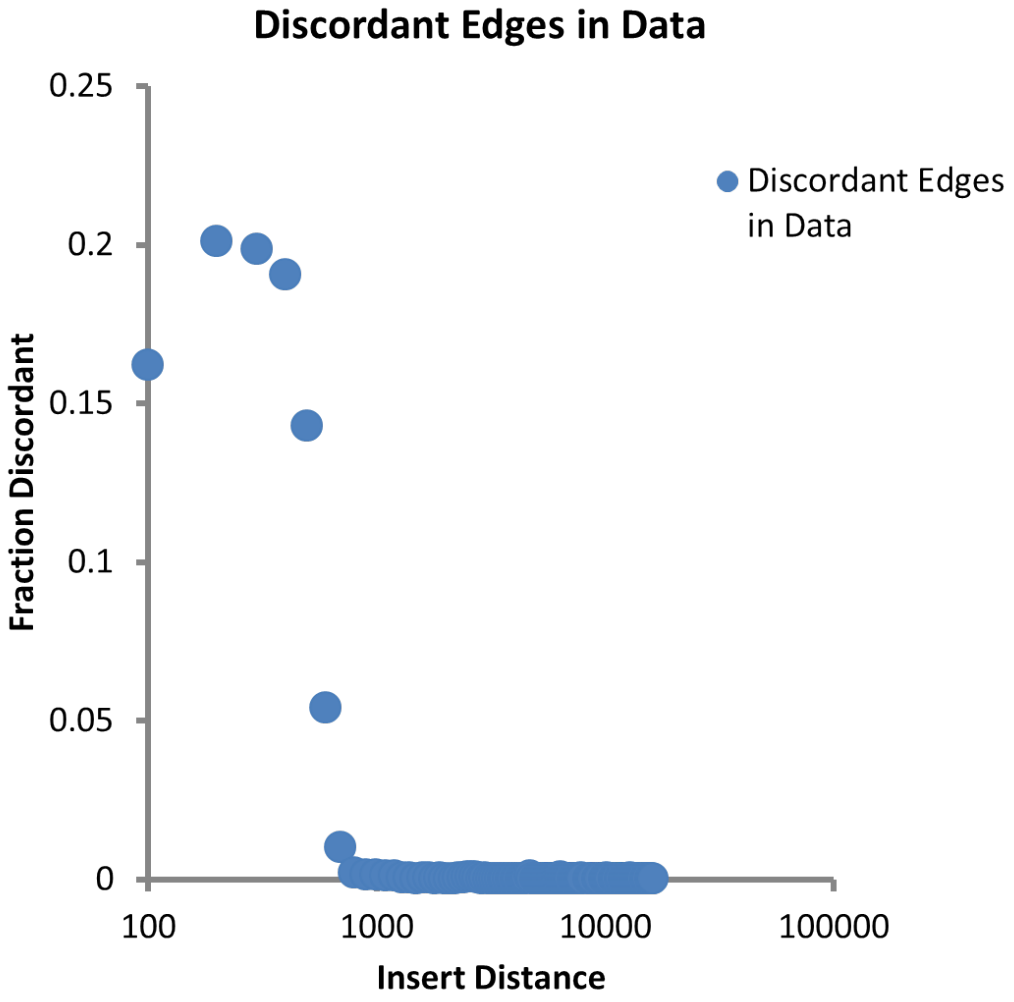

Figure 4.10: The insert length distribution of discordant edges in the UCSC1989 somatic Chicago/HiC library.

There were still some errors, based on the edges retaining 100% discordance relative to the data. I decided to add a phasing correction step to the recombination mapping software. The software then filters out any edges with either low coverage or high discordance. While these steps remove about 20% of the edges from the final results, recombination mapping is a case where accurate phasing is more important than

completeness (Figure 4.11) .

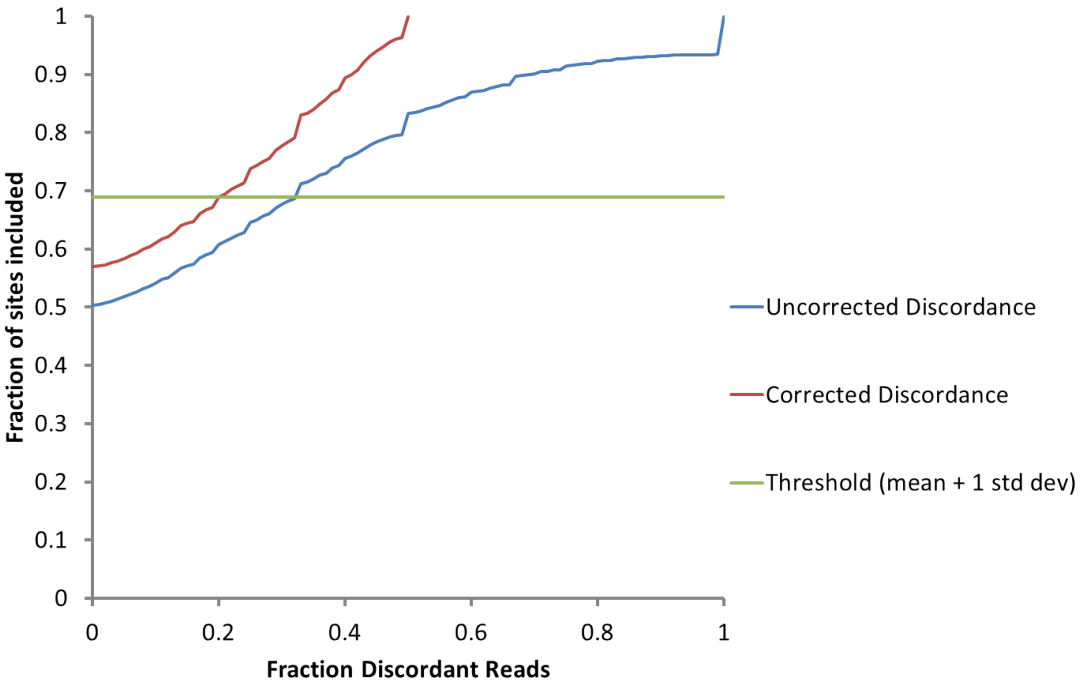

Figure 4.11: The result of adding the haplotype phasing/correction steps to the pipeline with the UCSC1989 somatic data.

4.3.1.8 Improving somatic haplotype phasing

As previously described in Section 4.3.1.7, I removed the shotgun data from the haplotype phasing step. I also added the HiC library from spit. The improvement in concordance (see Figures 4.11 and 4.13) was dramatic. I then filtered the somatic Chicago data to remove all read pairs over 100kbp in an effort to remove spurious or

chimeric read pairs from the somatic phasing results. This proved to be important, as there were thousands of "long–insert" Chicago read pairs in the Chicago data sets which had random haplotype information.

4.3.1.9 Implementing phasing quality filtering

To further improve the quality of the somatic haplotype reference, I added the option to impose a phasing confidence cutoff based on the confidence score that HapCUT2 assigns to each site in the output. HapCUT2 also has the option to set a cutoff. However, my recombination mapping software runs much faster than HapCUT2, so duplicating this feature in the recombination mapping software allows me to save time when tweaking cutoff to find a good balance between correctness and completeness. While this feature did not dramatically affect the overall chromosome-wide averave recombination rate ("CWAR"), it did in several cases remove very puzzling "hotspots" that were actually due to low confidence phasing errors that had passed the HapCUT2 filter. Most of the HapCUT2 errors seem to be "swap" errors (i.e., a single site is incorrectly haplotype phased relative to the rest of the haplotype phase block) as opposed to "switch" errors (at some point a large section of the block is correctly phased internally, but switched to the rest of the block). I believe that this is due to variant sites which are far enough from other sites that they must a) be phased with the Chicago and HiC data and b) are at sufficient distance that the chimeric noise is difficult to differentiate from the signal. As we know from the nature of chromatin proximity ligation insert distributions, these sites will also tend to be covered at low frequency, so

the result is a great deal more variability in the signal to noise ratio for phasing these sites.

4.3.1.10 Recombination mapping

Initially, I tried a simple method of recombination rate mapping. I used a windowed approach to tile over the genome, and counted the number of germ-line reads where the haplotype phase information disagreed with the haplotype phase from HapCUT2. Unfortunately, this method resulted in a mean recombination rate of 24% across all reads (when using 1Mbp windows). One would expect a recombination rate in humans of between 5-3 cM/Mbp, corresponding to about 0.3% of reads showing recombination. Consequently, I investigated what could be causing this discrepancy. I discovered that the shotgun data was causing phasing errors, as described in Section 4.3.1.7 (above). Also, some of the haplotype phasing calls were probably incorrect. As a result, I implemented several filters and correction steps (described below) using the difference between the expected and the reported chromosome wide average recombination rate ("CWAR") to gauge the efficacy of my additions.

To map recombinations, I wrote custom Python code (part of the same pipeline as I used for correcting the haplotype phasing) to process the germline data. First, the variants are read in to memory, along with the HapCUT2 phase-blocks. Next, as described in the previous sections, the somatic data are used to make a pruned and corrected version of the phased haplotypes output by HapCUT2. Specifically, a haplotype graph is constructed from the hapCUT2 output, and compared to the raw somatic

data. Each edge in the HapCUT2 haplotype graph is compared with the physical read support, and then run through the correction steps described in Section 4.3.1.9. Any edges which cannot be corrected (i.e., they have a large proportion of discordant reads) are added to a blacklist of pruned edges. The pruned edges are completely excluded from all further analyses to reduce the likelihood of haplotype phasing error suggesting spurious recombination events. At this point, the recombined data are processed in the form of a HapCUT2 "hairs" file. This choice allows for more efficient processing of the data. Any haplotype pairs which correspond to blacklisted edges are discarded, while the remainder of the information is used to fill in a recombination graph. Finally, the software traverses across the genome in user-defined-size window and totals up the rate of recombination for each window. Recombination graph edges which span multiple windows contribute a fraction of their values to each spanned window equal to $\frac{1}{L} \times n$, where L =the length of the edge in base-pairs, and n =the number of bases of the bin that the edge spans (i.e., for an edge spanning the whole window, n =window length). This allows the software to leverage longer-insert distance reads, which have a higher likelihood of recombination (see Figure 4.3 for an illustration of the concept). The resulting recombination map can then be output as either a simple histogram or as a bedGraph file for use on the UCSC genome browser [41].

While the recombination rates are being calculated, the software also estimates a background average rate based on the length of the reference sequence under consideration. The software then reports the relative difference between this rate and the calculated CWAR from the data. This can be a useful debugging tool, e.g., if the two

rates are an order of magnitude different, the reported recombination rate map may not be valid.

4.3.1.11 Comparing phasing concordance versus insert distance in germline HiC data

I plotted the rate of phasing concordance as a function of insert distance to see how well my analysis was able to recover the recombination information in the germline data. From first principles, the concordance should decrease with increasing recombination rate, which in turn increases based on the physical distance between sites. Figure 4.13 shows that phasing concordance did indeed decrease with distance as predicted in UCSC1989. Interestingly, the concordance decreased faster than I would have predicted, with 60kb reads decreasing to 70% concordance. This may be explained by the extremely non-uniform nature of the recombination rate across the chromosome, as can be seen in the deCODE map (Figure 4.12).

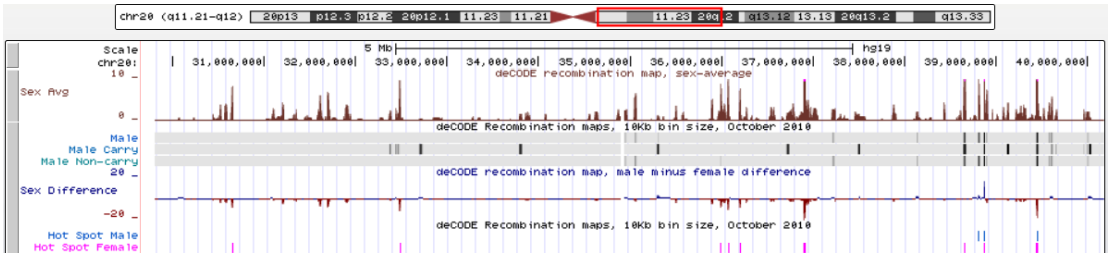

Figure 4.12: A screenshot from the UCSC Genome Browser, showing the deCODE recombination map for hg19, chr20. Note the highly variable recombination rate in the Sex Averaged track. The track scale is in cM/Mb, and the track resolution is 10kb.

Additionally, I checked the rate of human somatic cell/DNA contamination by looking at the rates of reads mapping to the mitochondria in both the combined somatic

and combined HiC data sets. Ideally, there should be no miotochondrial reads in the germline data. I calculated the rate of mitochondrial reads in both libraries, to account for the different coverage levels. The somatic mitochondrial rate was 0.02665%, while the germline rate was 0.0037%, or about 18-fold less. This indicates that the germline library is relatively free from somatic contamination.

4.3.2 Results

4.3.2.1 Haplotype phasing

The original heplotype phasing of UCSC1989 with only the Chicago reads resulted in very non-contiguous results, as previously discussed in Section 4.3.1.5. Consequently, I focused on the combined somatic HiC and Chicago data for UCSC1989. After initial haplotype phasing using the combined data, I plotted the haplotype phasing concordance between the reads versus the binned insert length of the read pairs (Figure 4.13)[1]. From previous examination of HiC libraries, one would expect high concordance out to at least 1Mbp. In the case of Chicago libraries, the concordance should be high out to the average limit of the input DNA, and then decrease to 60% at the point of the longest sampled input molecule [69]. My results were rather different. As can be seen in Figure 4.13, the concordance decreases dramatically out to about 200kbp, after which it recovers. However, I saw an improvement when I applied a quality cutoff to the input sites to remove low-confidence sites from the haplotype blocks. The explanation for this

[1]The data used for Figure 4.13 were produced after the haplotype pruning and fixing described in section 4.3.1.8 above. Consequently, they represent the best possible results achievable with the phasing output.

is that the Chicago reads in the 50-200kb range were of decreasing veracity (i.e., most of the reads >50kb were simply chimeric ligation products). Given that I had relatively equal amounts of sequencing data for HiC and Chicago, and nearly 50% of the Chicago reads were incorrect, the concordance results make sense.

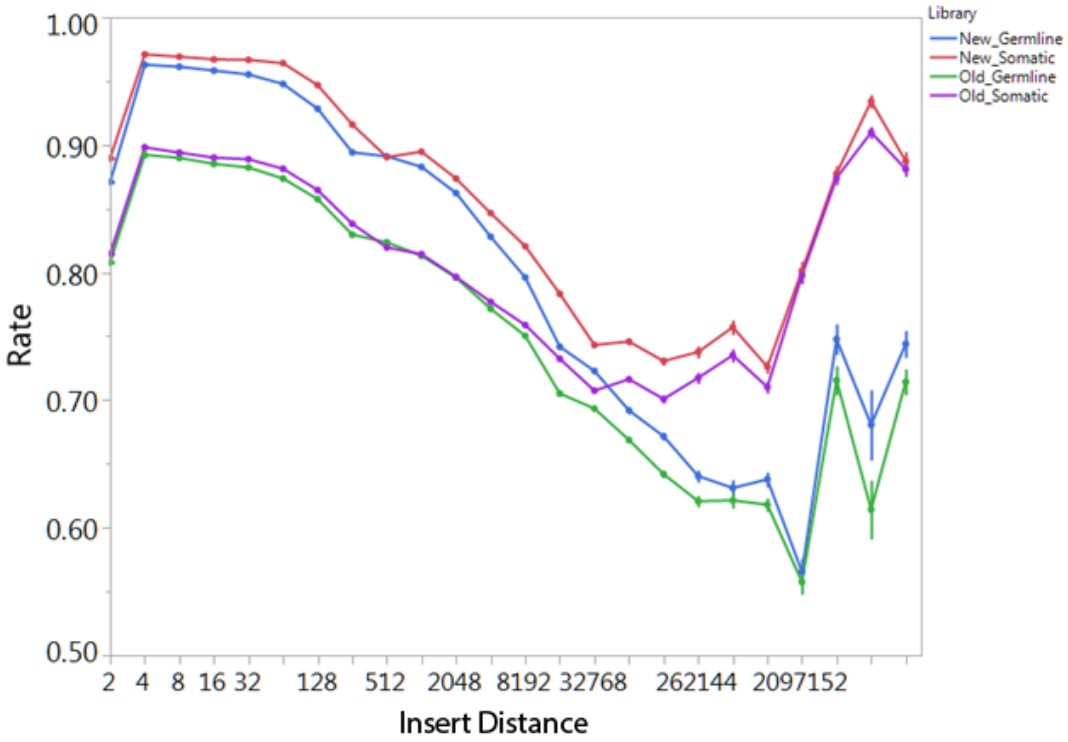

Figure 4.13: Comparing concordance versus insert length for the UCSC1989 libraries before and after trimming the Chicago reads. The blue and green lines are the new and old versions of the germline library, the red and purple are the new and old version of the somatic library, respectively. The points are the mean of the rates for all haplotype edges in that bin. The error bars represent the standard error. The germline libraries are significantly differentiable ($p<0.001$) from the somatic libraries in all places except for 512bp in the case of the new results ($p=0.48$) and 1kb ($p=0.11$) and 2kb ($p=0.46$) in the old results. The pruned version of the somatic data is always significantly more concordant than the old version, indicating that the haplotype correction and pruning did produce cleaner results.

To solve this issue, I wrote a small Python script which filtered only the Chicago

reads for insert length < 50Kbp. I then re-ran the HapCUT2 pipeline and added the concordance results to Figure 4.13. As can be seen in Figure 4.13, the concordance decreases with increasing insert length; however, it is much higher after the filtering than before.

4.3.2.2 Recombination mapping

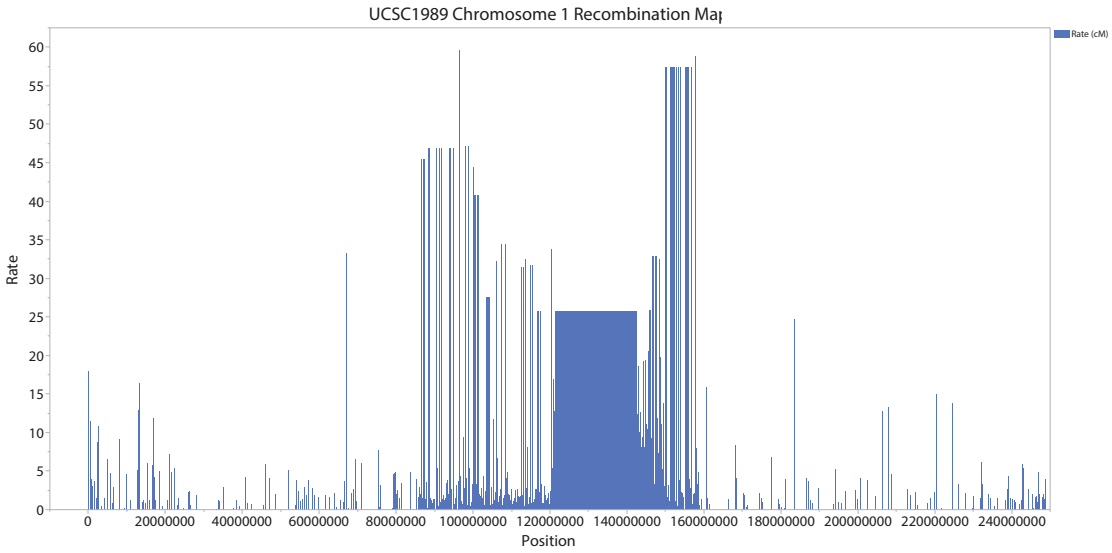

Figure 4.14: Recombination rate map for UCSC1989 Chromosome. Window size was 100kb, and only windows significantly different from the background noise ($p < 0.1$) are shown. Note the plateau corresponding to the centromere, as well as the oddly high results on either side.

Aside from the concordance plot (Figure 4.13) which indicates that there is recombination informative data in the libraries, but also a consistent problem with the somatic haplotype phasing, the recombination mapping results for UCSC1989 show some interesting patterns. First, as can be very clearly seen in Figure 4.14, the region

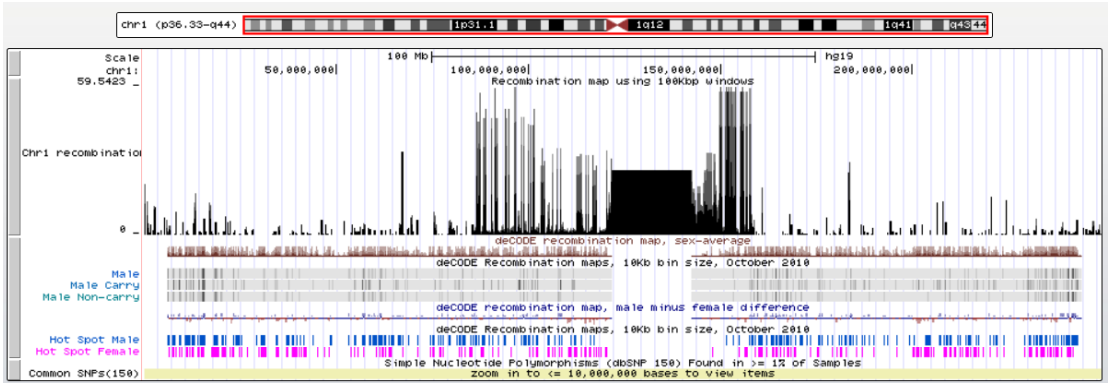

Figure 4.15: UCSC1989 Recombination map for Chromosome 1 compared to the deCODE recombination map on the UCSC Genome Browser. Note the very odd results near the centromeric region. Also note the correlation between the UCSC1989 map and the deCODE map vis a vis hotspots near the ends of the chromosome arms.

corresponding to the centromere is very clearly differentiable[2]. Additionally, the effects of the centromeric region on the haplotype phasing, and thus recombination mapping, seem to propagate outwards along the chromosome arms. This may be due to higher co-localization of the centromeres within the nucleus causing a higher rate of chimerism in the regions nearer to the centromere. Interestingly, this seems to fairly definitely end, and the map changes to something much more closely resembling the deCODE map. Additionally, the UCSC1989 map does extend very close to the ends of the chromosome, while the deCODE maps end about 6Mbp away from the ends.

4.3.3 Discussion

From the concordance results shown in Figure 4.13, it is clear that retrieving recombination data in the sperm HiC libraries is a more difficult task for the human

[2]This does not correspond to any real recombination rate, but does correspond very closely to the blank region in the deCODE map when the two are compared on the UCSC genome browser (Figure 4.15).

UCSC1989 sample than for the herring. As can be seen in the recombination map example in Figure 4.14, however, there is room for improvement in the algorithm. The centromeres are a major cause of erroneous recombination results, probably due to phasing errors. This effect appears to propagate out as far as a megabase away from the centromere in some cases[3]. In the short term, however, the best solution to this problem is most likely to simply break the haplotype blocks across the centromere, since the chromosome arms should largely be recombining independently of each other. Not withstanding these concerns, this work, in conjunction with the herring data presented above, demonstrates the feasibility of using chromatin proximity ligation for individual recombination mapping.

There are several future directions in which the human recombination mapping project should be taken. First, it would make sense to trio-phase the UCSC1989 sample (if possible), or else procure a sample more amenable to trio phasing. This would allow me to start from a best case scenario for the recombination rate mapping, without correcting or pruning haplotype phasing results[4]. Second, the software should be further refined to allow it to better differentiate between spurious 'recombinations' caused by phasing errors and chimeric reads, and actual recombination-informative sites.

[3]Oddly, some preleminary re-running of the human data with the t-test feature described in Chapter 5 shows that the high-discordance region across the centromere is significantly different in the germline data compared to the somatic data.

[4]Prof. Haussler suggested this point during my advancement presentation. I did not proceed with trio phasing due to 1) cost, 2) privacy, and 3) generalization concerns. The cost concern proved to be a false economy. It would have definitely been less expensive to trio-phase than to try and achieve comparable accuracy and completeness with Chicago and HiC data (at least for humans). The privacy and generalization issues still remain.

Bibliography

[1] Opentrons.

[2] Andrew Adey, Jacob O Kitzman, Joshua N Burton, Riza Daza, Akash Kumar, Lena Christiansen, Mostafa Ronaghi, Sasan Amini, Kevin L Gunderson, Frank J Steemers, and Jay Shendure. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome research*, 24(12):2041–9, dec 2014.

[3] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, Inanç Birol, Sébastien Boisvert, Jarrod A Chapman, Guillaume Chapuis, Rayan Chikhi, Hamidreza Chitsaz, Wen-Chi Chou, Jacques Corbeil, Cristian Del Fabbro, T Roderick Docking, Richard Durbin, Dent Earl, Scott Emrich, Pavel Fedotov, Nuno A Fonseca, Ganeshkumar Ganapathy, Richard A Gibbs, Sante Gnerre, Élénie Godzaridis, Steve Goldstein, Matthias Haimel, Giles Hall, David Haussler, Joseph B Hiatt, Isaac Y Ho, Jason Howard, Martin Hunt, Shaun D Jackman, David B Jaffe, Erich D Jarvis, Huaiyang Jiang, Sergey Kazakov, Paul J Kersey, Jacob O Kitzman, James R Knight, Sergey Koren, Tak-Wah Lam, Dominique Lavenier, François Laviolette, Yingrui Li, Zhenyu Li, Binghang Liu, Yue Liu, Ruibang Luo, Iain MacCallum, Matthew D MacManes, Nicolas Maillet, Sergey Melnikov, Delphine Naquin, Zemin Ning, Thomas D Otto, Benedict Paten, Octávio S Paulo, Adam M Phillippy, Francisco Pina-Martins, Michael Place, Dariusz Przybylski, Xiang Qin, Carson Qu, Filipe J Ribeiro, Stephen Richards, Daniel S Rokhsar, J Graham Ruby, Simone Scalabrin, Michael C Schatz, David C Schwartz, Alexey Sergushichev, Ted Sharpe, Timothy I Shaw, Jay Shendure, Yujian Shi, Jared T Simpson, Henry Song, Fedor Tsarev, Francesco Vezzi, Riccardo Vicedomini, Bruno M Vieira, Jun Wang, Kim C Worley, Shuangye Yin, Siu-Ming Yiu, Jianying Yuan, Guojie Zhang, Hao Zhang, Shiguo Zhou, and Ian F Korf. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990.

[4] Sasan Amini, Dmitry Pushkarev, Lena Christiansen, Emrah Kostem, Tom Royce, Casey Turk, Natasha Pignatelli, Andrew Adey, Jacob O Kitzman, Kandaswamy Vijayan, Mostafa Ronaghi, Jay Shendure, Kevin L Gunderson, and Frank J Steemers. Haplotype-resolved whole-genome sequencing by contiguity-preserving

transposition and combinatorial indexing. *Nature genetics*, 46(12):1343–9, dec 2014.

[5] Norman Arnheim, Peter Calabrese, and Magnus Nordborg. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *American Journal of Human Genetics*, 73(1):5–16, jul 2003.

[6] F Ay, T L Bailey, and W S Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, 24, 2014.

[7] Zachary Baker, Molly Schumer, Yuki Haba, Lisa Bashkirova, Chris Holland, Gil G Rosenthal, and Molly Przeworski. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. *eLife*, 6, jun 2017.

[8] F Baudat, J Buard, C Grey, A Fledel-Alon, C Ober, M Przeworski, G Coop, and B de Massy. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science (New York, N.Y.)*, 327(5967):836–40, feb 2010.

[9] W A Bickmore. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet.*, 14, 2013.

[10] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014.

[11] Julian Lewis Martin Raff Keith Roberts Bruce Alberts, Alexander Johnson and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 2002.

[12] Joshua N Burton, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman, and Jay Shendure. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12):1119–1125, nov 2013.

[13] Joshua N Burton, Ivan Liachko, Maitreya J Dunham, and Jay Shendure. Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps. *G3 (Bethesda, Md.)*, 4:1339–1346, 2014.

[14] Shai Carmi, Ken Y Hui, Ethan Kochav, Xinmin Liu, James Xue, Fillan Grady, Saurav Guha, Kinnari Upadhyay, Dan Ben-Avraham, Semanti Mukherjee, B Monica Bowen, Tinu Thomas, Joseph Vijai, Marc Cruts, Guy Froyen, Diether Lambrechts, Stéphane Plaisance, Christine Van Broeckhoven, Philip Van Damme, Herwig Van Marck, Nir Barzilai, Ariel Darvasi, Kenneth Offit, Susan Bressman, Laurie J Ozelius, Inga Peter, Judy H Cho, Harry Ostrer, Gil Atzmon, Lorraine N Clark, Todd Lencz, and Itsik Pe'er. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nature Communications*, pages 1–9, 2014.

[15] Jarrod A. Chapman, Isaac Ho, Sirisha Sunkara, Shujun Luo, Gary P. Schroth, and Daniel S. Rokhsar. Meraculous: De Novo Genome Assembly with Short Paired-End Reads. *PLoS ONE*, 6(8):e23501, aug 2011.

[16] Deanna M. Church, Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M. McLaren, Graham R.S. Ritchie, Derek Albracht, Milinn Kremitzki, Susan Rock, Holland Kotkiewicz, Colin Kremitzki, Aye Wollam, Lee Trani, Lucinda Fulton, Robert Fulton, Lucy Matthews, Siobhan Whitehead, Will Chow, James Torrance, Matthew Dunn, Glenn Harden, Glen Threadgold, Jonathan Wood, Joanna Collins, Paul Heath, Guy Griffiths, Sarah Pelan, Darren Grafham, Evan E. Eichler, George Weinstock, Elaine R. Mardis, Richard K. Wilson, Kerstin Howe, Paul Flicek, and Tim Hubbard. Modernizing Reference Genome Assemblies. *PLoS Biology*, 9(7):e1001091, jul 2011.

[17] X F Cui, H H Li, T M Goradia, K Lange, H H Kazazian, D Galas, and N Arnheim. Single-sperm typing: determination of genetic distance between the G gamma-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. *Proceedings of the National Academy of Sciences of the United States of America*, 86(23):9389–93, dec 1989.

[18] Asher D Cutter and Bret A Payseur. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature reviews. Genetics*, 14(4):262–74, apr 2013.

[19] A. P. Jason de Koning, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, and David D. Pollock. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics*, 7(12):e1002384, dec 2011.

[20] J W Deacon. *Fungal biology*. Blackwell Pub, Malden, MA, 2006.

[21] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295(5558):1306–11, feb 2002.

[22] Xinxian Deng, Wenxiu Ma, Vijay Ramani, Andrew Hill, Fan Yang, Ferhat Ay, Joel B Berletch, Carl Anthony Blau, Jay Shendure, Zhijun Duan, William S Noble, and Christine M Disteche. Bipartite structure of the inactive mouse X chromosome. *Genome Biology*, 16(1):152, aug 2015.

[23] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, apr 2012.

[24] Josée Dostie and Job Dekker. Mapping networks of physical interactions between genomic elements using 5C technology. *Nature Protocols*, 2(4):988–1002, jan 2007.

[25] Z Duan, M Andronescu, K Schutz, S McIlwain, Y J Kim, and C Lee. A three-dimensional model of the yeast genome. *Nature.*, 465, 2010.

[26] Peter Edge, Vineet Bafna, and Vikas Bansal. HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 27(5):801–812, may 2017.

[27] Chungang Feng, Mats Pettersson, Sangeet Lamichhaney, Carl-Johan Rubin, Nima Rafati, Michele Casini, Arild Folkvord, and Leif Andersson. Moderate nucleotide diversity in the Atlantic herring is associated with a low mutation rate. *eLife*, 6, jun 2017.

[28] Dmitry V Fyodorov and James T Kadonaga. Chromatin assembly in vitro with purified recombinant ACF and NAP-1. *Methods in enzymology*, 371:499–515, jan 2003.

[29] Genome.gov. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *Journal of Heredity*, 100(6):659–674, nov 2009.

[30] J H Gibcus and J Dekker. The hierarchy of the 3D genome. *Mol Cell.*, 49, 2013.

[31] Sante Gnerre, Iain Maccallum, Dariusz Przybylski, Filipe J Ribeiro, Joshua N Burton, Bruce J Walker, Ted Sharpe, Giles Hall, Terrance P Shea, Sean Sykes, Aaron M Berlin, Daniel Aird, Maura Costello, Riza Daza, Louise Williams, Robert Nicol, Andreas Gnirke, Chad Nusbaum, Eric S Lander, and David B Jaffe. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1513–8, jan 2011.

[32] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–6, nov 2015.

[33] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature*, 17, 2016.

[34] Richard E. Green, Edward L. Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W. Vandewege, John A. St. John, Salvador Capella-Gutiérrez, Todd A. Castoe, Colin Kern, Matthew K. Fujita, Juan C. Opazo, Jerzy Jurka, Kenji K. Kojima, Juan Caballero, Robert M. Hubley, Arian F. Smit, Roy N. Platt, Christine A. Lavoie, Meganathan P. Ramakodi, John W. Finger, Alexander Suh, Sally R. Isberg, Lee Miles, Amanda Y. Chong, Weerachai Jaratlerdsiri, Jaime Gongora, Christopher Moran, Andrés Iriarte, John McCormack, Shane C. Burgess, Scott V. Edwards, Eric Lyons, Christina Williams, Matthew Breen, Jason T. Howard, Cathy R. Gresham, Daniel G. Peterson, Jürgen Schmitz, David D. Pollock,

David Haussler, Eric W. Triplett, Guojie Zhang, Naoki Irie, Erich D. Jarvis, Christopher A. Brochu, Carl J. Schmidt, Fiona M. McCarthy, Brant C. Faircloth, Federico G. Hoffmann, Travis C. Glenn, Toni Gabaldón, Benedict Paten, and David A. Ray. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215), 2014.

[35] A. Hossain. Modified guanidinium thiocyanate method for human sperm DNA isolation. *Molecular Human Reproduction*, 3(11):953–956, nov 1997.

[36] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, oct 2004.

[37] M Imakaev, G Fudenberg, R P McCord, N Naumova, A Goloborodko, and B R Lajoie. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods.*, 9, 2012.

[38] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(July):90–98, 2011.

[39] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–8, jan 2012.

[40] Noam Kaplan and Job Dekker. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nature Biotechnology*, 31(12):1143–1147, nov 2013.

[41] W J Kent, C W Sugnet, T S Furey, K M Roskin, T H Pringle, and A M Zahler. The human genome browser at UCSC. *Genome Res.*, 12, 2002.

[42] Jeffrey M. Kidd, Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, Brian Teague, Can Alkan, Francesca Antonacci, Eric Haugen, Troy Zerr, N. Alice Yamada, Peter Tsang, Tera L. Newman, Eray Tüzün, Ze Cheng, Heather M. Ebling, Nadeem Tusneem, Robert David, Will Gillett, Karen A. Phelps, Molly Weaver, David Saranga, Adrianne Brand, Wei Tao, Erik Gustafson, Kevin McKernan, Lin Chen, Maika Malig, Joshua D. Smith, Joshua M. Korn, Steven A. McCarroll, David A. Altshuler, Daniel A. Peiffer, Michael Dorschner, John Stamatoyannopoulos, David Schwartz, Deborah A. Nickerson, James C. Mullikin, Richard K. Wilson, Laurakay Bruhn, Maynard V. Olson, Rajinder Kaul, Douglas R. Smith, and Evan E. Eichler. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, may 2008.

[43] DanielC. Koboldt, Larson Steinberg, Karyn andMeltz, DavidE., RichardK. Wilson, and Elaine R. Mardis. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell*, 155, 2013.

[44] Augustine Kong, Daniel F Gudbjartsson, Jesus Sainz, Gudrun M Jonsdottir, Sigurjon A Gudjonsson, Bjorgvin Richardsson, Sigrun Sigurdardottir, John Barnard, Bjorn Hallbeck, Gisli Masson, Adam Shlien, Stefan T Palsson, Michael L Frigge, Thorgeir E Thorgeirsson, Jeffrey R Gulcher, and Kari Stefansson. A high-resolution recombination map of the human genome. *Nature Genetics*, 31(3):241–7, jul 2002.

[45] Augustine Kong, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G. Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th. Kristinsson, Sigurjon A. Gudjonsson, Michael L. Frigge, Agnar Helgason, Unnur Thorsteinsdottir, and Kari Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, oct 2010.

[46] Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, and Adam M Phillippy. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, 30(7):693–700, jul 2012.

[47] S Kurukuti, V K Tiwari, G Tavoosidana, E Pugacheva, A Murrell, and Z Zhao. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc Natl Acad Sci U S A.*, 103, 2006.

[48] Marie-Therese Kurzbauer, Clemens Uanschou, Doris Chen, and Peter Schlögelhofer. The recombinases DMC1 and RAD51 are functionally and spatially separated during meiosis in Arabidopsis. *The Plant cell*, 24(5):2058–70, may 2012.

[49] H Li and R Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.*, 26, 2010.

[50] H H Li, U B Gyllensten, X F Cui, R K Saiki, H A Erlich, and N Arnheim. Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature*, 335(6189):414–7, sep 1988.

[51] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, 326(5950):289–93, oct 2009.

[52] Audrey Lynn, Kara E Koehler, LuAnn Judis, Ernest R Chan, Jonathan P Cherry, Stuart Schwartz, Allen Seftel, Patricia A Hunt, and Terry J Hassold. Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science (New York, N.Y.)*, 296(5576):2222–5, jun 2002.

[53] Wenxiu Ma, Ferhat Ay, Choli Lee, Gunhan Gulsoy, Xinxian Deng, Savannah Cook, Jennifer Hesson, Christopher Cavanaugh, Carol B Ware, Anton Krumm, Jay Shendure, Carl Anthony Blau, Christine M Disteche, William S Noble, and Zhijun Duan. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Methods.*, 12(1):71–8, jan 2015.

[54] Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Jean-François Flot, Gianni Liti, Dante Poggi Parodi, Sylvie Syan, Nancy Guillén, Antoine Margeot, Christophe Zimmer, and Romain Koszul. High-quality genome (re)assembly using chromosomal contact data. *Nature Communications*, 5:5695, dec 2014.

[55] Alvaro Martinez Barrio, Sangeet Lamichhaney, Guangyi Fan, Nima Rafati, Mats Pettersson, He Zhang, Jacques Dainat, Diana Ekman, Marc Höppner, Patric Jern, Marcel Martin, Björn Nystedt, Xin Liu, Wenbin Chen, Xinming Liang, Chengcheng Shi, Yuanyuan Fu, Kailong Ma, Xiao Zhan, Chungang Feng, Ulla Gustafson, Carl-Johan Rubin, Markus Sällman Almén, Martina Blass, Michele Casini, Arild Folkvord, Linda Laikre, Nils Ryman, Simon Ming-Yuen Lee, Xun Xu, and Leif Andersson. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, 5, may 2016.

[56] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–303, sep 2010.

[57] Gilean A T McVean, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science (New York, N.Y.)*, 304(5670):581–4, apr 2004.

[58] Matthias Meyer and Martin Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, 2010(6):pdb.prot5448, jun 2010.

[59] Karen H Miga, Yulia Newton, Miten Jain, Nicolas Altemose, Huntington F Willard, and W James Kent. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome research*, 24(4):697–707, apr 2014.

[60] Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, and Peter Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)*, 310:321–324, 2005.

[61] Simon Myers, Rory Bowden, Afidalina Tumian, Ronald E Bontrop, Colin Freeman, Tammie S MacFie, Gil McVean, and Peter Donnelly. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science (New York, N.Y.)*, 327(5967):876–9, feb 2010.

[62] T Nagano, Y Lubling, T J Stevens, S Schoenfelder, E Yaffe, and W Dean. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature.*, 502, 2013.

[63] V Narendra, P P Rocha, D An, R Raviram, J A Skok, and E O Mazzoni. Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, 347, 2015.

[64] E P Nora, B R Lajoie, E G Schulz, L Giorgetti, I Okamoto, and N Servant. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature.*, 485, 2012.

[65] S V Nuzhdin, E G Pasyukova, C L Dilda, Z B Zeng, and T F Mackay. Sex-specific quantitative trait loci affecting longevity in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 94(18):9734–9, sep 1997.

[66] F Pâques and J E Haber. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. *Microbiology and molecular biology reviews : MMBR*, 63(2):349–404, jun 1999.

[67] Simone Picelli, Asa K. Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research*, 24(12):2033–2040, 2014.

[68] F. Pratto, K. Brick, P. Khil, F. Smagulova, G. V. Petukhova, and R. D. Camerini-Otero. Recombination initiation maps of individual human genomes. *Science*, 346(6211):1256442–1256442, nov 2014.

[69] Nicholas H Putnam, Brendan L O'Connell, Jonathan C Stites, Brandon J Rice, Marco Blanchette, Robert Calef, Christopher J Troll, Andrew Fields, Paul D Hartley, Charles W Sugnet, David Haussler, Daniel S Rokhsar, and Richard E Green. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research*, 26(3):342–50, mar 2016.

[70] Michael Quail, Miriam E Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, jul 2012.

[71] D. V. (Dmitrii Viktorovich) Radakov. *Schooling in the ecology of fish.* J. Wiley, 1973.

[72] S S Rao, M H Huntley, N C Durand, E K Stamenova, I D Bochkov, and J T Robinson. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.*, 159, 2014.

[73] Edward S Rice, Satomi Kohno, John St John, Son Pham, Jonathan Howard, Liana F Lareau, Brendan L O'Connell, Glenn Hickey, Joel Armstrong, Alden Deran, Ian Fiddes, Roy N Platt, Cathy Gresham, Fiona McCarthy, Colin Kern, David Haan, Tan Phan, Carl Schmidt, Jeremy R Sanford, David A Ray, Benedict Paten, Louis J Guillette, and Richard E Green. Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome research*, 27(5):686–696, may 2017.

[74] Steven L Salzberg, Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, Michael C Schatz, Arthur L Delcher, Michael Roberts, Guillaume Marçais, Mihai Pop, and James A Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research*, 22(3):557–67, mar 2012.

[75] Siddarth Selvaraj, Jesse R Dixon, Vikas Bansal, and Bing Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, 31(12):1111–1118, nov 2013.

[76] Andrew M Shedlock, Christopher W Botka, Shaying Zhao, Jyoti Shetty, Tingting Zhang, Jun S Liu, Patrick J Deschavanne, and Scott V Edwards. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8):2767–72, feb 2007.

[77] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

[78] Cormac Sheridan. Milestone approval lifts Illumina's NGS from research into clinic. *Nature Biotechnology*, 32(2):111–112, feb 2014.

[79] Jared T Simpson and Richard Durbin. Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, 22(3):549–56, mar 2012.

[80] Rahul Sinha, Geoff Stanley, Gunsagar Singh Gulati, Camille Ezran, Kyle Joseph Travaglini, Eric Wei, Charles Kwok Fai Chan, Ahmad N Nabhan, Tianying Su, Rachel Marie Morganti, Stephanie Diana Conley, Hassan Chaib, Kristy Red-Horse, Michael T Longaker, Michael P Snyder, Mark A Krasnow, and Irving L Weissman. Index Switching Causes Spreading-Of-Signal Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*, 2017.

[81] Brian Smith. Rinse, Swab or Spit – What's the Real Source of DNA in Saliva?, 2010.

[82] Michelle M. Thiaville, Hana Kim, Wesley D. Frey, Joomyeong Kim, Fan Yang, Ferhat Ay, Joel B. Berletch, Carl Anthony Blau, Jay Shendure, Zhijun Duan, William S. Noble, and Christine M. Disteche. Identification of an evolutionarily

conserved cis-regulatory element controlling the Peg3 imprinted domain. *PLoS One.*, 8(9):e75417, sep 2013.

[83] C Thiede, G Prange-Krex, J Freiberg-Richter, M Bornhäuser, and G Ehninger. Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants. *Bone Marrow Transplantation*, 25(5):575–577, mar 2000.

[84] Ingrid Torjesen. Genomes of 100000 people will be sequenced to create an open access research resource. *BMJ*, 347, 2013.

[85] Eray Tuzun, Andrew J Sharp, Jeffrey A Bailey, Rajinder Kaul, V Anne Morrison, Lisa M Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V Olson, and Evan E Eichler. Fine-scale structural variation of the human genome. *Nature Genetics*, 37(7):727–732, jul 2005.

[86] N Varoquaux, F Ay, W S Noble, and J P Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics.*, 30, 2014.

[87] Matteo VietriRudan, Christopher Barrington, Stephen Henderson, Christina Ernst, DuncanT. Odom, Amos Tanay, and Suzana Hadjur. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports*, 10(8):1297–1309, mar 2015.

[88] Sarah A Vitak, Kristof A Torkenczy, Jimi L Rosenkrantz, Andrew J Fields, Lena Christiansen, Melissa H Wong, Lucia Carbone, Frank J Steemers, and Andrew Adey. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, 14(3):302–308, jan 2017.

[89] Jianbin Wang, H.Christina Fan, Barry Behr, and StephenR. Quake. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell*, 150(2):402–412, jul 2012.

[90] James D Watson, Tina A Baker, Stephen P Bell, Alexander Gann, Michael Levine, Richard Losick, and Stephen C Harrison. *Molecular Biology of the Gene*. Cold Spring Harbor Laboratory Press, 7th edition, 2014.

[91] Neil I Weisenfeld, Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, Diana Tabbaa, Louise Williams, Carsten Russ, Chad Nusbaum, Eric S Lander, Iain MacCallum, and David B Jaffe. Comprehensive variation discovery in single human genomes. *Nature Genetics*, 46(12):1350–1355, oct 2014.

[92] Louise J S Williams, Diana G Tabbaa, Na Li, Aaron M Berlin, Terrance P Shea, Iain Maccallum, Michael S Lawrence, Yotam Drier, Gad Getz, Sarah K Young, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Paired-end sequencing of Fosmid libraries by Illumina. *Genome research*, 22(11):2241–9, nov 2012.

[93] Gösta Winberg and Rickard Sandberg. Preparation of NGS libraries using in-house Tn5 Gene expression, Next-generation sequencing, genomics, transposon, RNA-seq, DNA-seq, tagmentation.

[94] Cheng-Cang Wu, Rosa Ye, Svetlana Jasinovica, Megan Wagner, Ronald Godiska, Amy Hin-Yan Tong, Si Lok, Amanda Krerowicz, Curtis Knox, David Mead, and Michael Lodes. Long-span, mate-pair scaffolding and other methods for faster next-generation sequencing library creation. *Nature Methods*, 9(9), aug 2012.

[95] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, may 2005.

[96] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11):1341–7, nov 2006.

Appendix A

Additional HiC Protocols

Hi-C protocol

Prepare buffers:

1X PBS (chill on ice)

Formaldehyde

2.5M Glycine

0.1M DTT (keep on ice)

7X Protease Inhibitors (cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail)

- Dissolve 1 tablet in 1.5mls H20 by vortexing. Save any unused at -20°C.

Lysis Buffer (make fresh; 5mls per sample; add DTT just before use); chill on ice

Stock		**Final**
1M HEPES pH 7.6	50 ul	10mM
5M NaCl	10 ul	10mM
10% IGEPAL CA-630	100 ul	0.2%
7X Protease Inhibitors	714 ul	1X
0.1M DTT	50 ul	1mM
dH20	4.076 mls	

For each sample, chill 4.236mls Lysis buffer and add protease inhibitors and DTT above before use.

Lysate Wash Buffer; chill on ice

Stock		**Final**
1M TRIS pH 8	2.5 mls	50mM
5M NaCl	500 ul	50mM
0.5M EDTA pH 8	100 ul	1mM
dH20	46.9 mls	

HiC SPRI Wash Buffer

Stock		Final
1M TRIS pH 8	500 ul	10mM
5M NaCl	500 ul	50mM
10% Tween	250 ul	0.05%
dH20	48.75 mls	

Crosslink Tissue

Prepare small mortar and pestle in liquid nitrogen. Try to keep tissues frozen/cold until formaldehyde.

PBS/formaldehyde can be premixed before adding to samples if timing of crosslinking can be kept relatively consistent.

For animal tissue, use no more than 50mg.

1. Measure out a chunk of frozen tissue. Coarsely chunk in mortar or mince with a razor blade (on petri dish surface on dry ice).
2. Transfer to a 1.5ml tube containing 1ml cold PBS. Vortex to mix/disperse.
3. Add and vortex, then rotate for 15 minutes at room temp:
    - 40.5 ul formaldehyde (1.5% final)
4. Add and vortex, then rotate for another 5 minutes at room temp:
    - 50 ul 2.5M glycine
5. Spin the sample at max speed for 1 minute in microfuge at max speed to pellet. Repeat spin if debris still floating.
6. Carefully aspirate supernatant.
7. Wash sample with 1 ml of 1X PBS by vortexing, then pellet as above.
8. Aspirate 1X PBS wash completely from pellet (careful of pellet loss!).
9. Resuspend the fixed sample in 1 ml of chilled Lysis Buffer (gently, to avoid losing sample in the lid). Keep on ice.

For sperm, use no more than 100 ul.

1. Pipette 100 ul of semen, centrifuge at 2500 rcf for 5 minutes, and discard supernatant.
2. Transfer to a 1.5ml tube containing 1ml cold PBS. Vortex to mix/disperse.
3. Add and vortex, then rotate for 15 minutes at room temp:
    - 40.5 ul formaldehyde (1.5% final)
4. Add and vortex, then rotate for another 5 minutes at room temp:
    - 50 ul 2.5M glycine

5. Spin the sample at max speed for 1 minute in microfuge at max speed to pellet. Repeat spin if debris still floating.
6. Carefully aspirate supernatant.
7. Wash sample with 1 ml of 1X PBS by vortexing, then pellet as above.
8. Aspirate 1X PBS wash completely from pellet (careful of pellet loss!).
9. Resuspend the fixed sample in 0.9 ml of chilled Lysis Buffer (gently, to avoid losing sample in the lid) and add 100 ul of 1M DTT or BME. Keep on ice.

For plants, measure out 250mg of leaves (exclude stems).

1. Measure quickly onto a weigh boat, then transfer to small mortar in liquid nitrogen.
2. Grind leaves to a coarse powder. Transfer to a labeled 5ml tube.
3. Add and vortex, then rotate for 15 minutes at room temp:
    - 2 mls 1X PBS
    - 81 ul formaldehyde (1.5% final)
4. Add and vortex, then rotate for another 5 minutes at room temp:
    - 100 ul 2.5M glycine
5. Spin the sample at max speed (5000xg) for 5 minutes in tabletop at 4°C to pellet. Repeat spin if debris still floating.
6. Carefully aspirate supernatant.
7. Wash sample with 2 mls of 1X PBS by vortexing, then pellet as above.
8. Aspirate 1X PBS wash completely from pellet (careful of pellet loss!).
9. Resuspend the fixed sample in 1 ml of chilled Lysis Buffer (gently, to avoid losing sample in the lid). Keep on ice.

Extract Chromatin

For Tissue, Sperm, etc.

1. Transfer resuspended sample into a tube containing 100-200 ul of 0.5 mm garnet beads (MoBio).
2. Place tubes sideways on a vortexer, and vortex at max speed for 2 minutes, or until the sample has been completely homegonized.
3. Using a pipette, remove the homegonate from the beads, including any foam and transfer into a 1.5 ml centrifuge tube. Heat at 37°C for 15 minutes.
4. Centrifuge at 3500 rcf for 5 minutes to pellet chromatin. Remove supernatant and wash twice with **Lysate Wash Buffer**, re-centrifuging if necessary to adhere the pellet to the tube.
5. Resuspend in 100 ul **Lysate Wash Buffer**, and add 2.5 ul of 20% SDS.

6. Incubate at 37°C for 15 minutes, with shaking.
7. Qubit HS quantitate 1 ul of SN.
    - If measurement is 15ng/ul or less, use entire volume for SPRI bead binding.
    - If measurement is >15ng/ul, use an amount equivalent to <800ng total. For liver/testes, will probably use 1/4 of input.
    - Save any remainder at 4°C for up to a week.

Crosslink Cells/nuclei

For cell culture, use 0.5 x 10^6 cells; scale volumes if necessary.

Pellet cells at 2500 x g in a 1.5ml silanized tube.

Resuspend in 1 ml 1X PBS.

For blood, use 150-300 ul and process through step 5 of the Qiagen blood protocol.

Resuspend pelleted nuclei in 1 ml 1X PBS in 1.5ml silanized tube.

1. Add 27 ul formaldehyde (1% final) and incubate for 15 minutes at room temp.
2. Add 54 ul 2.5M glycine and incubate on ice for 10 minutes.
3. Pellet nuclei at 2500 x g for 5 minutes 4°C.
4. Wash sample with 1 ml of 1X PBS, pellet as above.
5. Aspirate 1X PBS wash completely from pellet (careful of pellet loss!).

Extract Chromatin

6. Pipet in 50 ul **Lysate Wash Buffer**. Add 2.5 ul of 20% SDS.
7. Vortex to resuspend pellet for >30 seconds. Pipet to break up clumps if necessary.
8. Incubate at 37°C for 15 minutes, with shaking.
9. Qubit HS quantitate 1 ul of SN.
    - If measurement is 15ng/ul or less, use entire volume for SPRI bead binding.
    - If measurement is >15ng/ul, use an amount equivalent to <800ng total. For liver/testes, will probably use 1/4 of input.
    - Save any remainder at 4°C for up to a week.

Bind Chromatin to SPRI beads

For samples where less than the entire 50 ul chromatin sample will be bound to SPRI beads (to prevent overloading), bring up the sample volume to 50 ul with **Lysate Wash Buffer**.

The remaining prep proceeds essentially like a Chicago Prep, except: 1) aliquots are taken of the Input and Digest; 2) beads are vortexed to resuspend to avoid pipetting loss; and 3) DTT is added to 1mM to the digest and end fill-in.

1. Add 100 ul SPRI beads to 50 ul chromatin/cell debris in **Lysate Wash Buffer** with 1% SDS. Vortex to mix, then quick spin down. Bind 5-10 minutes.
2. Magnet 5 minutes. Very carefully pipet off and discard the supernatant; the pellet may be loose—switch to a 10 ul pipet to remove as much liquid as possible.
3. Wash pellet with 200 ul **HiC SPRI Wash Buffer** by vortexing to resuspend.
4. Quick spin, then magnet. Carefully pipet off supernatant.
5. Repeat wash, twice more.
6. After vortexing and quick spin for the third wash, take a 10 ul aliquot of the 200 ul wash into a PCR tube as an **input sample**, before the final magnet. Store aliquots on ice.
7. SPRI-bound samples may be stored in the final wash overnight at 4°C.

Dpn II Digest

Vortex to resuspend beads in 50 ul DpnII digestion mix; quick spin down.

H20	42.5 ul
10X DpnII Buffer	5 ul
100 mM DTT	0.5 ul
DpnII (10 U/ul, NEB)	2 ul

1. Digest for 1 hour at 37°C with shaking.
2. Wash twice with 200 ul **HiC SPRI Wash Buffer** by vortexing to resuspend.
3. Quick spin, then magnet. Carefully pipet off supernatant.
4. Repeat wash, once more.
5. After vortexing and quick spin for the second wash, take a 10 ul aliquot of the 200 ul wash into a PCR tube as a **digest sample**, before the final magnet. Store aliquots on ice.

End Fill-In

Vortex to resuspend beads in 50 ul End Fill-In mix; quick spin down.

H20	37 ul
10X NEB Buffer #2	5 ul
1mM Biotin-dCTP	4 ul
10mM dATP,dTTP,dGTP	1.5 ul
100 mM DTT	0.5 ul
Klenow (5 U/ul, NEB)	2 ul

1. Fill-in for 30 minutes at 25°C with shaking.
2. Wash twice with 200 ul **HiC SPRI Wash Buffer** by vortexing to resuspend.
3. Quick spin, then magnet. Carefully pipet off supernatant.
4. Repeat wash, once more.

Intra-Aggregate DNA End Ligation

Vortex to resuspend beads in 250 ul Intra-Aggregate Ligation mix; quick spin down.

H20	215.5 ul
10X NEB T4 DNA Ligase Buffer	25 ul
BSA (20 mg/ml, Thermo)	1.25 ul
10% Triton X-100	6.25 ul
T4 DNA Ligase (4000 U/ul, NEB)	2 ul

Ligate for at least 1 hour at 16°C with shaking.

Crosslink Reversal

For each HiC sample and their aliquots, make complete Crosslink Reversal Buffer:

Crosslink Reversal Mix	48.5 ul
Proteinase K (20 mg/ml, NEB)	1.5 ul

1. Magnet ligation reactions; carefully remove supernatant.
2. Add 50 ul Crosslink Reversal Buffer to pellets. Vortex to resuspend; quick spin down.

3. Also add 50 ul Crosslink Reversal Buffer to the **input sample** and **digest sample** aliquots.
4. Digest 15 minutes at 55°C, the 45 minutes at 68°C with shaking.

SPRI Purification

1. Magnet reactions; transfer the **SUPERNATANT** to a clean 1.5ml tube.
2. Add 100 ul SPRI beads to each; pipet to mix ~10 times. Bind 5-10 minutes.
3. Magnet 5 minutes. Remove and discard supernatant.
4. Wash by pipetting 250 ul 80% ethanol onto the pellet while on the magnet; wait 30 seconds, then remove the wash.
5. Repeat wash.
6. Quick spin tubes, then place back onto the magnet and remove the last traces of ethanol with a 10 ul pipet tip.
7. Air dry on the magnet for 5-7 minutes.
8. Resuspend the HiC pellets in 52 ul TE; resuspend **input sample** and **digest sample** aliquots in 6 ul TE.
9. Elute off the magnet for 3 minutes.
10. Magnet, and transfer the eluted sample to 1.5 ml tubes (or PCR strip for aliquots).
11. Qubit HS 1 ul of the HiC crosslink reversal samples.
12. QC samples on a Genomic TapeStation tape: 1 ul sample + 10 ul buffer. Dilute HiC samples if necessary.

Shear at 4-10 ng/ul (in 78 or 100 ul) in 0.6ml Bioruptor tubes; use 200 ng per library prep; run 11 cycles of index PCR.

Otherwise, shear less than 200ng in 50 ul in 0.1ml Bioruptor tubes and use entire 50ul in library prep. Scale PCR cycles up if necessary.

Alternatively, prepare using Tn5 transposase with up to 200ng total DNA, or as little as 20ng. Scale PCR levels down if using more than 50ng DNA.

Appendix B

Supplementary Figures for Chapter 4

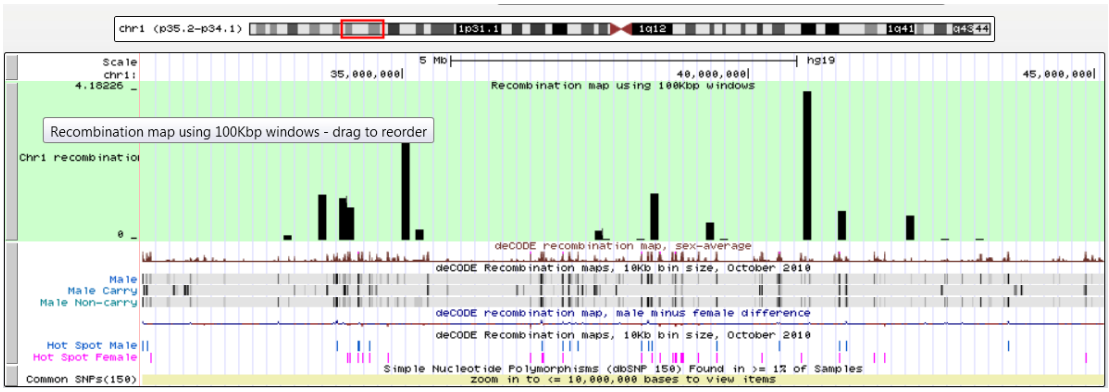

Figure B.1: UCSC1989 Recombination map for Chromosome 1 compared to the deCODE recombination map on the UCSC Genome Browser. The view is chr1:31,252,788-45,142,914. Note the correspondence between the Male deCode track and the UCSC1989 map.

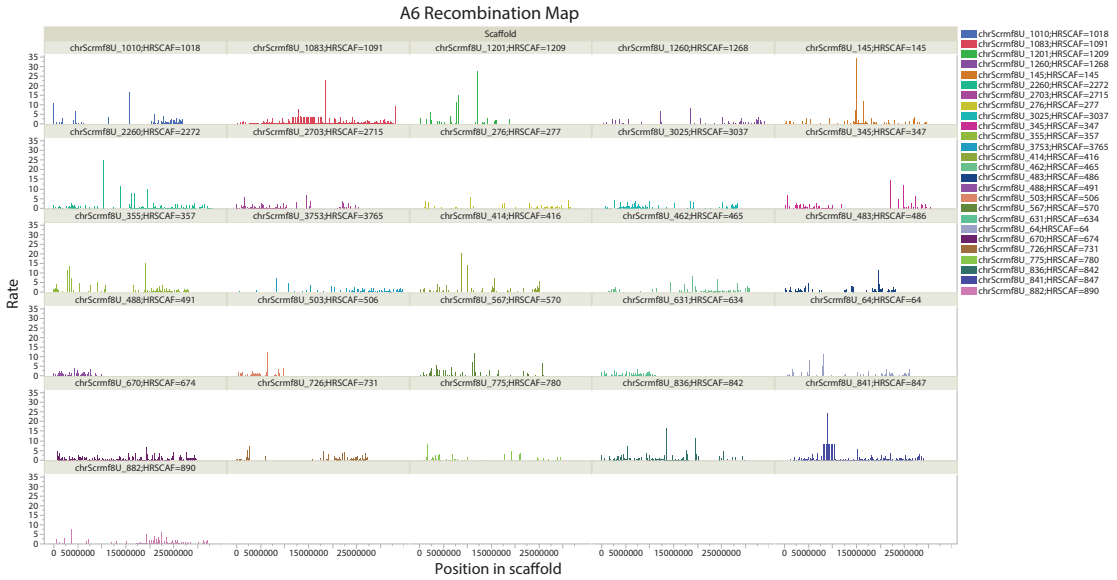

Figure B.2: Recombination map for the 26 longest Atlantic herring scaffolds, using sample A6. The map is produced at 100Kbp resolution, with only the windows differentiable from the background error rate shown.

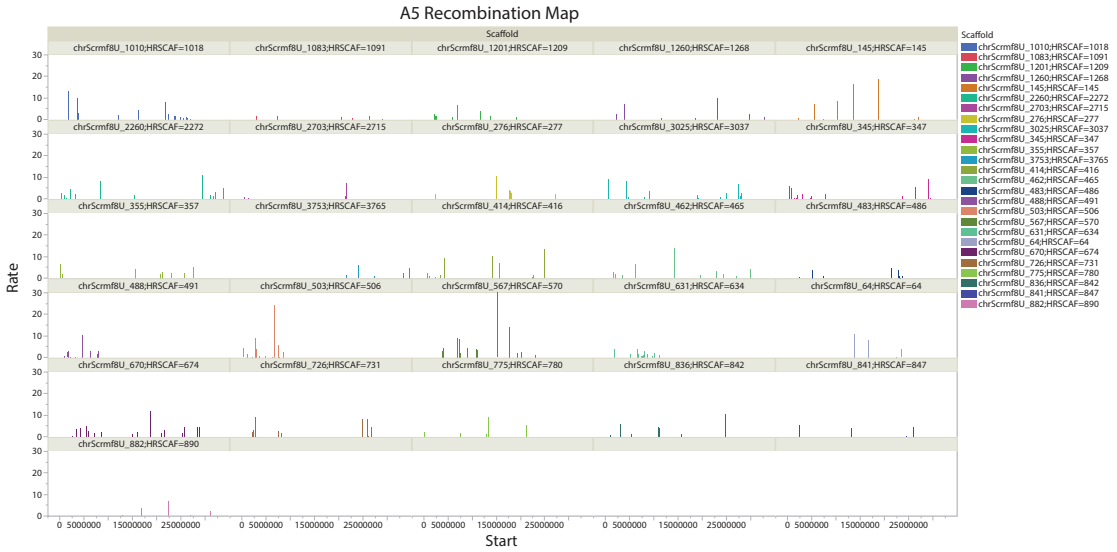

Figure B.3: Recombination map for the 26 longest Atlantic herring scaffolds, using sample A5. The map is produced at 100Kbp resolution, with only the windows differentiable from the background error rate shown.