Massively parallel analysis of nucleic acid structure

Vijay Ramani

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Jay Shendure

Steven Henikoff

Cole Trapnell

Program Authorized to Offer Degree:

Department of Genome Sciences

University of Washington

ABSTRACT

Massively parallel analysis of nucleic acid structure

Vijay Ramani

Chair of the Supervisory Committee:

Professor Jay Shendure

Department of Genome Sciences

A goal amongst modern biologists is to “compute” cellular and organismal function. This “computation” necessitates a holistic understanding of the biochemical mechanisms underlying different cellular states. Methods that pair massively parallel sequencing and biochemical analyses of cells have brought us much closer to this ultimate goal, enabling genome-wide mapping and quantification of molecules responsible for cellular function. My thesis work has focused on developing methods to measure the structure of RNA and DNA molecules in cells, a parameter thought to play a critical role in regulating the expression of genes, and thus cellular state. In this thesis, I describe novel methods for studying i.) intramolecular RNA structure, ii.) genome-wide chromosomal structure in populations of mammalian cells, and iii.) genome-wide chromosomal

structure in single mammalian cells. Importantly, these cutting-edge approaches enable scalable, genome-wide, and high-resolution analyses of intramolecular nucleic acid structure in populations of cells, and within single cells.

TABLE OF CONTENTS

List of Figures

List of Tables

ACKNOWLEDGEMENTS

This work would not have been possible without the support of family and friends. To Dad, Mom, and Shraya, thank you for your unyielding familial support. To SK, SS, FW, DB, and ML, many thanks for your friendship, and for providing me with the closest thing to a "family away from home" one could ask for. Many thanks to Ruolan Qiu—an incredibly talented and gracious teacher, without whom much of the work presented here would not have been possible, and without whom I would be utterly lost as a molecular biologist. Finally, I must thank my cohort of graduate students in the UW Genome Sciences department, my lab mates in the Shendure Lab, and, last but certainly not least, my fantastic PhD advisor Jay Shendure, for providing what can only be described as the best possible environment for carrying out research in molecular biological technology development.

Chapter 1. INTRODUCTION

This thesis describes three molecular methods I have developed over the course of completing my PhD. These methods are unified by a common theme: they are high-throughput tools to study a critical biological parameter—the three-dimensional organization of nucleic acids within cells. The *in vivo* structure of RNAs, and the *in vivo* structure of chromatin—the protein-DNA nucleoprotein complex that all higher eukaryotic chromosomes are packaged into—both play key roles in regulating biology's central dogma. Chromosomal architecture is thought to be integral to regulating the transcription of DNA into RNA, and the faithful replication of DNA prior to cell division; intramolecular RNA structures are known to critically regulate the manner by which RNAs are translated into protein.

In **Chapter 1**, I introduce methods to study chromosomal architecture, and review the field's current state-of-the-art. In **Chapter 2**, I briefly review methodological progress in studying RNA structure, and describe a novel method of my design—RNA Proximity Ligation, that improves upon existing high-throughput methods for spatially resolving RNA molecular structure. In **Chapter 3**, I return to the study of chromosomal architecture, and describe a novel method for studying chromatin structure in populations of cells at high-resolution, termed *in situ* DNase Hi-C. In **Chapter 4,** I describe a novel paradigm for studying chromosomal architecture in each of thousands of single-cells. This method is termed single-cell combinatorially indexed Hi-C (sciHi-C). Finally in **Chapter 5**, I close with a prospective look towards the future of massively parallel methods development.

Note: The following sections of **Chapter 1** have been adapted from work published in the February 2016 issue of *Genomics, Proteomics, & Bioinformatics* as:

Ramani V., Shendure J, Duan Z. "Understanding spatial genome organization: methods and insights**."** *Genomics, Proteomics, & Bioinformatics* (2016)

1.1 ABSTRACT

The manner by which a eukaryotic genome is packaged into the nucleus, while still enabling nuclear function, remains one of biology's fundamental mysteries. Over the last ten years, we have witnessed rapid advances in both microscopic and nucleic acid-based approaches to map genome architecture. These techniques have made it clear that interphase chromosomes are hierarchically organized. At the finest scale, both transient and long-lived chromosomal loops bring distinct genomic loci together. At a larger scale, self-interacting domains (*i.e.* "topologically associating domains," or TADs) offer another level of genomic compaction. Zooming out even further, higher-order genomic "compartmentalization" exists, as megabase scale sequence elements preferentially self-organize in a manner that coordinates critical nuclear functions like transcription, replication, and DNA repair. Finally, at the coarsest resolution possible—that of entire chromosomes, specific chromosomes preferentially associate to form chromosome territories that differ in positioning with respect to nuclear features like the lamina, nucleolus, and nuclear speckles. This review describes features of these broadly-defined hierarchical structures, insights into the mechanisms underlying their formation, our current understanding of how interactions in the nuclear space are linked to gene regulation, and potential future directions for the field.

1.2 INTRODUCTION

The human body is comprised of trillions cells harboring nearly identical genetic genomes, yet subsets of these cells are distinct both functionally and morphologically. It is widely accepted that "epigenetic" mechanisms are responsible for the differential regulation of shared genetic information, and thus for the generation of a diverse array of terminal cell types during zygotic development. Importantly, for the purposes of this chapter, we use the term "epigenetic" to describe non-genetic, biochemical phenomena that enable cellular function. We do not necessarily demand that these phenomena propagate faithfully across cellular divisions or generations, though many of the examples described here do adhere to this canon as well.

The physical organization of eukaryotic chromosomes within a nucleus is intertwined with the reading, interpretation, and propagation of genetic information by epigenetic mechanisms. Metazoan cells package genomic DNA up to 2 meters long into a tiny nuclear space ~10 μm in diameter through a hierarchy of organizational structures (Felsenfeld and Groudine, 2003). This compaction begins with the wrapping of 147 base pairs (bp) of DNA around a histone octamer to form the nucleosome; this nucleoprotein complex serves as the basic repeating unit of chromatin. The histone octamer itself is composed of eight subunits that assemble as one histone H3–H4 tetramer and two histone H2A–H2B dimers. Both DNA and protein components of the nucleosome particle are subject to a diversity of chemical modifications (*e.g.* CpG methylation, lysine tail acetylation) (Zhou et al., 2011), and several histone variants exist (*e.g.* H2A.Z, CENPA) (Talbert and Henikoff, 2010). The combinatorial diversity of possible chromatin states made possible by these diverse modifications has long been posited to "encode" specific biological functions (Jenuwein and Allis, 2001), though evidence for a causal link between epigenetic modification and cell state is lacking.

The next level of compaction is commonly believed to be the organization of nucleosomes into a 10 nm "beads-on-string" chromatin fiber (Oudet et al., 1975). Additional nucleosomal organization into higher-order structures on the order of 30 nm or 100 nm has been hotly debated; the existence of the native structure beyond the 10 nm fiber *in vivo* has been questioned (Fussner et al., 2012; Joti et al., 2012). A recent study has proposed that native chromatin fibers in *Saccharomyces cerevisiae* are formed by heterogeneous clutches of nucleosomes interspersed with nucleosome-depleted regions, arguing against the existence of highly ordered structures such as the 30 nm fiber (Ricci et al., 2015). This result (*i.e.* the absence of a 30-nm fibre *in vivo*) has been explicitly confirmed by electron tomography of nuclei, using a novel method termed ChromEMT (Ou et al., 2017).

Though we still know little of the dynamics of *in vivo* chromatin folding, we have gained important insights into the higher-order spatial organization of eukaryotic genomes thanks to significant advances in DNA imaging technology and high-throughput biochemical techniques (Gibcus and Dekker, 2013; Gorkin et al., 2014; Misteli, 2007; Ou et al., 2017; Pombo and Dillon, 2015). In mammalian genomes, individual chromosomes preferentially occupy a distinct nuclear area, called chromosome territories (CTs) (Cremer and Cremer, 2010). Transcriptionally silent regions generally localize near the nuclear envelope and peri-nucleolar space, whereas transcriptionally active regions occupy the remaining nuclear space (Finlan et al., 2008; Gonzalez-Sandoval et al., 2015). At the cytological level, the eukaryotic genome is partitioned into euchromatin and heterochromatin (Felsenfeld and Groudine, 2003). At the molecular level, the nucleus is geometrically compartmentalized in mammalian cells to contain morphologically and molecularly distinct sub-structures (*e.g.* nuclear bodies), suggesting that nuclear activities are also spatially organized (Dundr and Misteli, 2010; Schneider and Grosschedl, 2007). Individual

chromosomes are partitioned into various compartments and domains (Bouwman and de Laat, 2015; Duan and Blau, 2012; Gibcus and Dekker, 2013; Sexton and Cavalli, 2015). Given the strong link between these common organizational features and cellular function, it is tempting to speculate that modulation of chromatin organization itself is a basic mechanism by which cellular functions are enacted. Critical experiments, however, are still required to elucidate whether these observed structural features play some general causal role, or are simply correlative in nature. Regardless, we can be reasonably certain that, as the fundamental units of chromosomal architecture and compaction, features like chromatin loops, clustered highly-transcribed genomic loci, and large-scale chromosome domains are basic elements of chromatin folding worth considering within the context of developmental and environmental cues.

Here, we first review well-established and emerging technologies that are revolutionizing our understanding of higher-order genome architecture. We then discuss our current understanding of spatial genome organization in greater detail, covering insights into the mechanisms underlying the formation of organizational structures, as well as the links between chromatin folding and gene regulation. Finally, we propose a handful of pressing questions we believe to be central to ultimately understanding of the spatiotemporal organization of and function of nucleome.

1.3 TOOLS FOR EXPLORING THE 3D GENOME

Three largely orthogonal approaches are commonly used to study the structure and function of the three-dimensional (3D) genome. Microscopy-based DNA imaging techniques and high-throughput genomic mapping tools based on massively parallel sequencing have been used to delineate higher-order genomic architecture, while genome perturbation tools (e.g., genome-editing) have then been used to ascertain the functional significance of specific architectural elements. Over the last ten years, the field has witnessed tremendous methodological advances in

all three areas (de Wit and de Laat, 2012; Gaj et al., 2013; Risca and Greenleaf, 2015; van Steensel and Dekker, 2010).

Traditionally, chromosome and nuclear structure have viewed through DNA imaging technologies, which can be based on electron microscopy (Daban, 2011; Ou et al., 2017; Rapkin et al., 2012), or light microscopy (Huang et al., 2010; Rapkin et al., 2012). Electron microscopic techniques, including transmission electron microscopy (TEM) and cryo-electron microscopy (Cryo-EM), have typically been used to characterize cell-free systems, though this is rapidly changing. Cryo-EM, in particular, has become an increasingly popular structural biological tool, owing in part to dramatic improvements in resolution and ease of sample preparation (Callaway, 2015). Recently, Cryo-EM was used to determine an 11 Å–resolution structure of 30-nm chromatin fibers assembled from arrays of 12 nucleosomes (Song et al., 2014). Most recently, however, methodological approaches have enabled *in vivo* marking of chromatin using an EM-sensitive DNA stain termed ChromEM; this advance enabled fine scale analysis of chromosomal packaging in the contexts of interphase and mitotic chromatin (Ou et al., 2017).

Before the advent of massively parallel analyses by microarray and later high-throughput sequencing, our knowledge of 3D genome organization largely derived from studies using fluorescence labeling followed by light microscopy, such as DNA fluorescence *in situ* hybridization (Langer-Safer et al., 1982) (FISH) and live-cell imaging (Tsukamoto et al., 2000). FISH and live-cell imaging can directly measure physical distances between DNA loci and visualize the nuclear position of loci and/or whole chromosomes within single cells. Today, many variants of the FISH technique exist, including conventional two-dimensional FISH (2D-FISH), 3D-FISH (Cremer et al., 2008), and cryo-FISH (Branco et al., 2008), with the resolution of such assays dropping below 100 kb (Wang et al., 2016). More recently, a high-throughput imaging

position mapping platform (HIPmap) has been implemented (Shachar et al., 2015), presenting a breakthrough in overcoming the scalability- and throughput-limitations associated with conventional FISH techniques. While FISH assays are typically used to characterize only a few loci at a time, HIPmap enables large-scale (384-well format), automated, high-resolution localization of 3D gene positions in single cells. In addition to HIPmap, a quantitative high-resolution imaging approach, which combines FISH, array tomography imaging (AT) and multiplexed immunostaining, has also been implemented for investigating 3D chromatin organization in complex tissues (Linhoff et al., 2015). The development of automated image analysis toolkits such as these is likely to be critical as the field moves toward visualizing chromatin architecture in a large number of diverse contexts.

One commonly cited limitation of light microscopic techniques, despite their versatility, is the resolution limit owing to the wavelength of light. To overcome this, several super-resolution fluorescence microscopy approaches, such as structured illumination microscopy (SIM), stimulated emission depletion (STED), and photoactivation localization microscopy/stochastic optical reconstruction microscopy (PALM/STORM), have been developed during the last decade (reviewed in Toomre and Bewersdorf (2010)). These techniques have been applied to study higher-order nuclear architecture (Lakadamyali and Cosma, 2015; Markaki et al., 2012; Schermelleh et al., 2008; Wang et al., 2011b), and have been combined with more advanced fluorescent labeling techniques to image chromosome dynamics with unprecedented spatiotemporal resolution in both fixed and live cells at the single-molecule level (Boettiger et al., 2016; Liu et al., 2015).

Complementary to microscopy-based DNA imaging tools, biochemical tools decipher nuclear organization by measuring physical contacts between different genomic regions or between genomic DNA and other nuclear components. Initially coupled with oligonucleotide-

decorated array (microarray) technology, and now typically paired with massively parallel DNA sequencing (Shendure and Aiden, 2012), these biochemical tools enable genome-wide characterization of myriad aspects of higher-order chromosome structure and organization, and can also allow for reconstruction and modeling of 3D genome architecture with the aid of sophisticated computational algorithms (Ay and Noble, 2015).

Broadly, the current state-of-the-art for these biochemical techniques can be classified into three groups, based on the biological origin of the chromatin contacts being assayed (**Figure 1.1**). Methods that can detect protein-DNA interactions, such as chromatin immunoprecipitation (ChIP-seq), DNA adenine methyltransferase identification (DamID), and sedimentation fractionation, have been used for probing physical contacts between genomic loci and nuclear landmarks such as the nuclear envelope or nucleolus, providing information about where particular genomic loci localize within the nucleus. In ChIP techniques (Solomon and Varshavsky, 1985), antibodies specific to a nuclear complex of interest are used to immunoprecipitate chemically crosslinked or native chromatin, and the associated DNA is used to create a high-throughput sequencing library. In DamID (van Steensel and Henikoff, 2000), bacterial adenine methyltransferase is fused to a protein of interest and is allowed to interact with physically proximal DNA. Sequences containing methylated adenine are enriched through digestion with Dam-specific restriction enzymes, and the products are then sequenced. In sedimentation fractionation (Nemeth et al., 2010), chromatin is subjected to ultracentrifugation and fractionation, and the DNA present in desired fractions is sequenced. Both ChIP and DamID have been used to identify genomic regions associated with nuclear pore complexes (NPCs) (van Steensel and Dekker, 2010) while sedimentation fractionation has been used to isolate nucleolus-associated domains (NADs) (Nemeth et al., 2010). Most commonly, the DamID approach has been used to catalog genomic regions that interact with

the inner face of nuclear membrane, so-called lamina-associated domains (LADs) (Guelen et al., 2008; Kind et al., 2015; 2013).

The second class of methods includes those that probe chromatin-RNA interactions, a hotly debated class of interactions that may eventually be used to define chromatin domains or sub-nuclear bodies. Currently, there are three different methods for identifying chromatin-RNA interactions: ChIRP (Chu et al., 2011), CHART (Simon et al., 2013) and RAP (Engreitz et al., 2013). All three of these techniques follow the same basic schema: crosslinked chromatin is sheared and then hybridized to biotinylated anti-sense oligonucleotides specific to a transcript or transcripts of interest. Following a streptavidin enrichment step, the DNA that is co-enriched with targeted RNA is subjected to deep sequencing. All three of these techniques have been used to study the dynamics of the long noncoding RNAs, including the *Drosophila melanogaster* RoX transcript (Quinn et al., 2014), and the murine long noncoding RNA Xist (Engreitz et al., 2013; Simon et al., 2013), both of which play critical roles in each species' respective dosage compensation mechanism.

The third group of techniques covers the chromosome conformation capture (3C) family of methods (de Wit and de Laat, 2012; Dekker et al., 2002), which measure the relative spatial proximity between individual genomic loci through digestion and re-ligation of physically proximal chemically crosslinked fragments of chromatin. 3C techniques are probably the most popular tools for mapping chromatin interactions, and a diversity of methods based on 3C have been developed during the past decade. 3C derivatives themselves can be classified into two groups (**Table 1.1**): those for globally mapping genome-scale chromatin interactions occurring in a nucleus, including Hi-C (Lieberman-Aiden et al., 2009), TCC (Kalhor et al., 2012), single-cell Hi-C (Cusanovich et al., 2015; Nagano et al., 2013), DNase Hi-C (Ma et al., 2015; Ramani et al.,

2016a), *in situ* Hi-C (Rao et al., 2014), and Micro-C (Hsieh et al., 2015); and those for targeted detection of a subset of chromatin interactions, such as 3C (Dekker et al., 2002), ChIP-loop (Horike et al., 2005), 4C (Simonis et al., 2006; Zhao et al., 2006), e4C (Schoenfelder et al., 2010), 5C (Dostie et al., 2006), ChIA-PET (Fullwood et al., 2009; Tang et al., 2015), Capture-C (Hughes et al., 2014), Capture-Hi-C (Mifsud et al., 2015), and targeted DNase Hi-C (Ma et al., 2015). Since 2009, Hi-C and its variants have been used to generate whole-genome contact probability maps in bacteria (Burton et al., 2014; Le et al., 2013; Marbouty et al., 2015), budding and fission yeast (Duan et al., 2010; Hsieh et al., 2015; Mizuguchi et al., 2014), a pathogenic eukyarote (Ay et al., 2014), plants (Feng et al., 2014; Grob et al., 2014), worm (Crane et al., 2015), fly (Hou et al., 2012; Sexton et al., 2012), mouse (Deng et al., 2015; Dixon et al., 2012), and human (Dixon et al., 2012; 2015; Lieberman-Aiden et al., 2009; Rao et al., 2014).

Depending on the protocol and depth of high-throughput sequencing used, the resolution of Hi-C-derived contact probability maps can be multiple orders of magnitude lower than that of base-pair level genomic annotations. In the absence of incredibly high sequencing depth, Hi-C and its variants are most suitable for identifying conformational signatures at the sub-megabase or megabase scale, such as chromosome territories (CTs), chromatin compartments and TADs. Other chromatin conformation signatures, including so-called “loops” between promoters and other cis-elements, or pairs of binding sites for the transcription factor CTCF, are best carried out using the second group of approaches, which are each designed to map a specific set of chromatin interactions and thus allow for considerably higher resolution at a given sequencing depth. Indeed, 4C, 5C, ChIA-PET, Capture-C, Capture-Hi-C and targeted DNase Hi-C have all successfully been used to map specific regulatory interactions.

The biochemical methods discussed above offer detailed molecular views of chromosome structure. However, these assays are all performed on many thousands to millions of cells per experiment, thus masking the variability inherent between individual cells. Single-cell versions of ChIP-seq (Rotem et al., 2015), Dam-ID (Kind et al., 2015), and Hi-C (Nagano et al., 2013; Ramani et al., 2017) have all been recently described, though in all cases the sensitivity of the assay is markedly low due to the difficulty in sensitively amplifying the small amounts of DNA present in a single cell. Still, this field of "single-cell" chromatin profiling by high-throughput biochemical methods is nascent, and offers an interesting complement to traditional single-cell assays carried out through microscopy. The most accurate models for the spatiotemporal organization of eukaryotic genome architecture will likely be derived using a combination of high-resolution microscopy-based imaging technologies (FISH and live-cell imaging) and high-throughput, genome-wide single-cell biochemical approaches.

A major goal in the field of chromatin biology concerns characterizing the functional significance of the genomic regions identified using the techniques described above. Several genome editing tools are currently available, including the zinc-finger nuclease (ZFNs) (Urnov et al., 2010), transcription activator-like effector nucleases (TALENs), and the RNA-guided CRISPR/Cas9 system (Gaj et al., 2013). All of these tools have been used to perturb higher-order chromatin architecture through genome and epigenome editing (Guo et al., 2015; Hilton et al., 2015; Lupianez et al., 2015; Maeder et al., 2013; Mendenhall et al., 2013; Sanborn et al., 2015; Thakore et al., 2015; Therizols et al., 2014). Due to limited space, this review will not cover these tools and their applications, which have been reviewed elsewhere (Deng and Blobel, 2014).

1.4 Organizational Features of Eukaryotic Genomes & Their Nuclear Functions

Microscopy-based and high-throughput biochemical studies have revealed common organizational structures in eukaryotic genomes, including CTs, chromatin and nuclear compartments, various types of chromatin domains (*e.g.*, NADs, LADs, and TADs), and chromatin loops. In this section, we discuss their respective biophysical characteristics, and discuss links between these structural features.

1.4.1 *Chromosomal Territories*

The non-randomness of genome organization in the nuclear space at chromosome level was observed more than a century ago. The 'Rabl' configuration, which postulated that centromeres and telomeres occupy opposite poles of the nucleus, was proposed by Carl Rabl in 1885 (Cremer et al., 2015) and was later confirmed by both microscopic and molecular studies in yeast and some plants (Duan et al., 2010; Feng et al., 2014; Grob et al., 2014). In 1909, Theodor Boveri suggested that animal interphase chromosomes occupied distinct regions within the nucleus, for which Boveri introduced the term chromosome territories (CTs). Since then, microscopy studies and genome-wide chromatin interaction mapping have revealed several features of CTs. First, although the existence of CTs in yeast and some plants is debatable, CTs have been observed in a wide range of animals, particularly in mammals (Dixon et al., 2012; 2015). Second, each CT is predominantly a self-interacting entity that still harbors interactions with other CTs (Cremer and Cremer, 2010). The physical clustering of centromeres, rDNA genes, and tRNA genes located on different chromosomes, which can be seen in species as divergent as *S. cerevisiae* and humans, is a prime example of contacts occurring between different CTs. Third, although the position of each CT is stochastic in a cell population (*i.e.*, different positioning in each cell),

individual CTs show preferences for nuclear positioning in mammalian cells, which correlate with known genomic properties (*e.g.*, GC content, gene density, chromosome size) and genomic functions (*e.g.*, transcriptional activity and replication timing) (Boyle et al., 2001; Croft et al., 1999; Grasser et al., 2008; Kosak and Groudine, 2004; Takizawa et al., 2008). In general, large and gene-poor chromosomes tend to be located near the nuclear periphery, whereas small and gene-rich chromosomes group together near the center of the nucleus. For example, human chromosomes 18 (gene poor) and 19 (gene rich) localize preferentially to the periphery and center of the nucleus in human lymphocytes, respectively (Croft et al., 1999). Interestingly, homologous chromosomes in diploid cells are generally found to be far apart from each other in the interphase (Heride et al., 2010). Fourth, in each cell, the relative position of CTs is stably maintained from mid G1 to late G2/early prophase during the cell cycle; this has been demonstrated in both HeLa cells and normal rat kidney (NRK) cells (Gerlich et al., 2003; Walter et al., 2003). Whether these global chromosomal arrangements are transmitted through mitosis, however, remains unknown. In NRK cells, this is believed to be the case (Gerlich et al., 2003), while in HeLa and HT1080 fibrosarcoma cells, this appears to not be the case (Thomson et al., 2004; Walter et al., 2003). Fifth and last, while the functional significance of a given CT's positional preference remains unknown, the spatial configurations of chromosomes relative to one another are tissue-specific (Parada et al., 2004) and may even be evolutionarily conserved (Tanabe et al., 2002). As an example of tissue-specificity, X chromosomes localize more peripherally in liver cells compared to kidney cells (Parada et al., 2004).

1.4.2 *Chromatin folding and compartmentalization of nuclear activities*

At any given time within a living cell's interphase chromosomes, certain genomic loci may be embedded in a constitutive heterochromatin region, some may associate with the nuclear

lamina, some may be attached to the nucleolus, and others may be embedded in various sub-nuclear bodies, engaging in specific nuclear activities. One widely-held model for transcription postulates that active genes may colocalize into discrete "transcription factories" where high local concentrations of RNA polymerase II and basal transcriptional machinery enforce gene expression (Eskiw et al., 2010). This supports the notion that chromatin folding is somehow influenced by various nuclear processes (*e.g.*, transcription, DNA replication/repair) and is constrained by nuclear context (e.g., geometrical / volume constraints). Microscopy and molecular studies have identified several chromatin domains, with each representing some aspect of chromatin folding. Here we summarize the characteristics of the most commonly discussed chromatin domains and review how they relate among each other.

1.4.3 *A/B compartments*

Hi-C studies have observed that within CTs, chromosomes are partitioned into large compartments at the multi-Mb scale, containing either the active and open ("A" compartments) or inactive and closed chromatin ("B" compartments) (Lieberman-Aiden et al., 2009). The open "A" compartments contain high GC-content regions, are gene-rich, are generally highly transcribed, and are enriched in DNase I hypersensitivity and histone modifications associated with active chromatin. In contrast, B compartments are gene-poor, less transcriptionally active and enriched in high levels of the silencing H3K9me3 mark (Lieberman-Aiden et al., 2009). It is interesting to consider the extent to which A/B compartments are correlated with cytogenetically-defined euchromatin/heterochromatin. A-compartments preferentially cluster with other A compartments throughout the genome, as do B compartments. B compartments are also highly correlated with late replication timing and LADs, suggesting that their nuclear position might be close to the nuclear periphery (Ryba et al., 2010). A recent high-resolution Hi-C study found that the two

compartments can be further subdivided into six subcompartments (A1, A2, and B1-B4) (Rao et al., 2014). A/B compartments and subcompartments have also been found to be cell-type specific and are each associated with distinct chromatin patterns (Lieberman-Aiden et al., 2009; Rao et al., 2014), a sensible finding given that different cell types express gene sets driven by distinct groups of regulatory elements. Thus, the compartmentalization of CTs into distinct A/B compartments and sub-compartments is directly correlated with cell-type-specific gene expression and chromatin state. Indeed, recent computational work has shown that A/B compartments can be reconstructed using genome-wide DNA methylation or chromatin accessibility data types alone (Fortin and Hansen, 2015).

1.4.4 *Self-interacting Domains (TADs, CIDs)*

With increases in resolution provided by greater depth of sequencing, recent Hi-C and 5C studies have revealed that CTs and A/B compartments may be broken down further into smaller self-interacting domains, which have been identified in the genomes of a wide range of species from bacteria to human (Ciabrelli and Cavalli, 2015; Dekker and Heard, 2015). In metazoan genomes, these chromatin-folding modules are called physical domains in flies (Sexton et al., 2012) or Topologically Associating Domains (TADs) (Dixon et al., 2012; Nora et al., 2012) in mammalian cells, while in bacteria and yeast, these domains are typically referred to as chromosomal interacting domains (CIDs) (Hsieh et al., 2015; Le et al., 2013). TADs in mammalian genomes range in size between several hundred kilobases (kb), to 1–2 megabases (Mb) (with a median size of about 800 kb in mouse) (Dixon et al., 2012; Nora et al., 2012). Physical domains in fly are smaller (60 kb) (Hou et al., 2012; Sexton et al., 2012), while CIDs are typically even smaller (Hsieh et al., 2015; Le et al., 2013).

While the formal definition for these self-interacting domains is quite broad, they all share common core properties. First, they are characterized by a greater frequency of within-domain interactions as compared to external interactions. This is in fact how TADs are identified in Hi-C data, through a measure of the directionality index (DI) of ligation pairs across a chromosome (Dixon et al., 2012). Identification of self-interacting domains is thus strongly dependent on the resolution of the Hi-C data set analyzed. This is evidenced by the much smaller self-interacting domains (median length 185 kb), identified in both mouse and human cells in a recent high-resolution Hi-C study (Rao et al., 2014). Second, domain boundary regions are generally enriched in transcription start sites, active transcription, active chromatin marks, housekeeping genes, tRNA genes and short interspersed nuclear elements (SINEs), as well as binding sites for architectural proteins like CTCF and cohesin (Dixon et al., 2012). A recent study also highlighted the role of histone acetylation in the formation of TADs, suggesting that TADs are primarily built from nonacetylated nucleosomes and that TAD boundaries are composed of acetylated nucleosomes (Ulianov et al., 2015). Third, TADs are evolutionarily conserved and cell-type independent (Ciabrelli and Cavalli, 2015; Dekker and Heard, 2015), a feature that is expected, given the presence of housekeeping genes at TAD boundaries. Fourth, self-interacting domains represent basic units of chromatin folding. This is supported by early microscopic studies that revealed that CTs consist of chromosomal domains (CDs) spanning 100 kb – 1 Mb in size (Cook and Brazell, 1975), the same length scale as recently defined self-interacting domains. This suggests that TADs and similar domains may represent the same structures as microscopy-defined CDs.

Recent work has directly linked TADs and cytologically-defined chromosome domains in fly (Eagen et al., 2015). In this study, Hi-C on Drosophila polytene chromosomes revealed equivalence between polytene bands/inter-bands and TAD/TAD boundaries, suggesting that

different types of TADs correspond to distinct packing states. For example, inactive TADs, which contain fully condensed chromatin at the nuclear periphery, correspond to classical heterochromatin, whereas active TADs (partially packaged) and TAD boundaries (fully extended chromatin fibers) correspond to classic euchromatin (less dense chromatin in the nuclear interior). Since these polytene bands are observed in single salivary gland cells, the correspondence of TADs to polytene bands also suggests that TADs are unlikely to be a statistical feature of population-level Hi-C experiments, but rather exist at the level of single cells. Recently, a super-resolution microscopy study on human and mouse cells using STORM revealed that nucleosomes are grouped into discrete clutches along the fiber, with areas of relative depletion between them (Ricci et al., 2015). The relationship between these "clutches" of nucleosomes and self-interacting domains in metazoans remains unknown, though the recently published Micro-C method in yeast hints at a strong linkage between the two (Hsieh et al., 2015).

Though the definition of self-interacting domains has greatly helped our understanding of how chromatin might be organized in the nucleus, the functional relevance of these domains and the mechanisms underlying their formation remain poorly understood. To get at the function of particular domain boundaries, recent studies have employed genome-editing to edit out or invert CTCF sites (Guo et al., 2015; Lupianez et al., 2015; Sanborn et al., 2015). In some cases, this editing led to drastic changes in gene expression, particularly when single nucleotide polymorphisms in these CTCF sites were already implicated in genome-wide association studies for a particular syndrome. Another naturally occurring example of this was recently shown in the context of brain cancer, where hypermethylation at particular CTCF sites in low-grade *IDH1*-mutant gliomas leads to differential CTCF binding, changes in genome topology, and consequent dysregulation of proto-oncogenes (Flavahan et al., 2015). In other cases, however, inversion or

deletion led to only slight changes in gene expression. The results of such experiments hint at the underlying complexity of gene regulation, perhaps suggesting that genome architecture alone is not the master regulator of gene expression.

1.4.5 *Gene clustering in transcription factories*

One common model for transcription posits the existence of transcription factories, discrete nuclear foci in eukaryotic nuclei where transcription occurs (Cook, 2010). Biochemical purification of transcription factories associated with RNA polymerase I, II, or III has demonstrated that transcription factories consist of nascent RNAs, genomic templates and regulatory DNA elements (*e.g.*, enhancers), and a variety of proteins involved in transcription initiation, elongation and regulation (Melnik et al., 2011). Several features of transcription factories have been revealed: i) >95% of all nuclear transcription takes place within transcription factories (Buckley and Lis, 2014); ii) each transcription factory contains only one type of RNA polymerase and the number of the RNA polymerase molecules in a factory is variable among different cell types (Buckley and Lis, 2014); iii) genes sharing the same factory can be on the same chromosome or on different chromosomes, and may be co-regulated or functionally unrelated (Fraser et al., 2015); iv) the number of transcription factories found per nucleus appears is largely dependent on the species studied and the imaging technique used to detect them, ranging from a few hundreds to a few thousands (Buckley and Lis, 2014); and v) the size of the factory varies depending on both the RNA polymerase featured and cell type (Buckley and Lis, 2014). Given this model, the question of whether transcription factory formation is a byproduct of the process of transcription, or whether these are stable structures whose formation, in fact, precedes and/or drives transcription itself, remains unanswered. What is clear, however, is that the colocalization

of genomic loci into these "factories" is a strongly tissue-specific mark of both chromatin folding and 3D genome organization in the nucleus.

1.4.6 *Nucleolar associating domains (NADs)*

The nucleolus is the largest subnuclear organelle in the nucleus of eukaryotic cells and is the prototype for transcription factories, as it serves as the primary site of ribosomal RNA (rRNA) biogenesis. In addition to its primary role as the site of rRNA transcription and maturation, the nucleolus also hosts several other biological processes, including viral replication, signal recognition particle biosynthesis, and sequestration of proteins (reviewed by (Matheson and Kaufman, 2015)). Nucleoli assemble around the ribosomal DNA (rDNA) genes clustered from different chromosomes, where the genes are transcribed by RNA polymerase I. In a given nucleus, only a subset of rDNA loci are transcribed at once, where they are looped into the inside of the nucleolus. The remaining rDNA loci are associated with the periphery of the nucleolus and accumulate marks of constitutive heterochromatin. Genomic regions that interact frequently with the nucleolus are called nucleolar associating domains (NADs) (Nemeth et al., 2010; van Koningsbruggen et al., 2010). NADs are characterized by repetitive DNA elements, mostly from centromeric and pericentromeric regions, are gene poor, and typically contain silent chromatin (*e.g.* regions of the inactive X chromosome (Xi), repressed olfactory receptor genes, tissue-specifically repressed RNA polymerase II genes), and several RNA polymerase III-transcribed genes. NADs cover about 4% of the human genome and are significantly overlapped with LADs (discussed in further detail below), indicating that a certain amount of redistribution occurs between the nuclear lamina and nucleolar periphery after mitosis (Nemeth et al., 2010; van Koningsbruggen et al., 2010). Mechanisms for this redistribution remain poorly understood, though it has been suggested that nucleolous tethering is mediated by *trans* acting factors such as

CTCF, chromatin assembly factor (CAF)-1, nucleolar proteins, and long noncoding RNAs (Matheson and Kaufman, 2015).

1.4.7 *Lamina-associated domains (LADs)*

Regions of the genome that interact with the nuclear lamina at the interior of the nuclear envelope are called lamina-associated domains, or LADs. LADs were first characterized using the DamID technique, which revealed that mammalian LADs are large, gene-poor domains spanning 40 kb to 30 Mb and covering ~40% of the genome (Guelen et al., 2008; Peric-Hupkes et al., 2010). LADs are enriched for heterochromatic silencing marks, largely overlap with previously identified H3K9me2 "locks," and show very sharp borders that are significantly enriched for bidirectional transcription, CpG islands and CTCF binding sites (Guelen et al., 2008), features reminiscent of the borders found at self-associating domains. As with NADs, the mechanisms underlying tethering of LADs to the nuclear periphery largely remain unclear. However, a recent single-cell study has revealed that LADs showing stable contact (*i.e.* contact across many single cells) with the nuclear lamina (NL) are extremely gene poor, suggesting a structural role, whereas LADs with variable NL contacts tend to be cell-type specific (Kind et al., 2015). Moreover, the consistency of NL contacts is inversely linked to gene activity in single cells and correlates positively with the heterochromatic histone modification H3K9me3 (Kind et al., 2015), suggesting that the tethering of LADs to the NL is important for physically and functionally compartmentalizing eukaryotic genomes.

1.4.8 *Chromatin loops and gene regulation*

Looping is an intrinsic property of chromatin fibers and serves as a basic mechanism of chromatin folding. In as early as 1878, Walther Flemming observed large chromosomal loops in

the so-called “lampbrush” chromosomes of amphibian oocytes (Fraser et al., 2015). Ptashne and others have since posited that long-range looping interactions may be key effectors of gene expression (Griffith et al., 1986), a hypothesis that has gained credence thanks to recent mapping efforts via 3C-based methods. The chromatin loop is likely tightly related to the formation of self-associating domains. For example, recent work has shown that the stability of a TAD is determined by specific long-range loops within it (Giorgetti et al., 2014). The best-studied chromatin loops are those between genes and their distal regulatory elements, such as enhancers. One example of this is the observation of an active chromatin hub (ACH) at the active beta- and alpha-globin loci. The ACH configuration is formed when multiple regulatory elements are juxtaposed against one another in 3D space via looping to coordinate gene expression (Tolhuis et al., 2002).

Recent genome-wide mapping of chromatin interactions has uncovered general features of this type of loop. First, ~50% of active genes are engaged in long-range chromatin interactions in the cell-types examined (Jin et al., 2013; Li et al., 2012), with those active genes not found to interact with a distal enhancer being enriched in housekeeping genes (Jin et al., 2013). Second, in addition to promoter-enhancer interactions, promoter-promoter and enhancer-enhancer loops have also been detected, and there is extensive colocalization among multiple promoters and/or multiple distal-acting enhancers (Li et al., 2012; Ma et al., 2015; Sanyal et al., 2011; Zhang et al., 2013). Given that 3C-based methods are designed to detect second-order interactions (*i.e.* pairs of interacting loci), the question remains whether an element interacts with multiple other elements simultaneously within the same nuclear environment, or whether these interactions actually occur within different single cells. As discussed in further detail below, arriving at an answer to these questions may become possible through the further development of “single-cell epigenomic” technologies. Third, promoter–enhancer interactions generally show high cell-type specificity and

are correlated with cell type-specific transcription, though it has been argued that promoter-enhancer loops are generally unchanged across tissue contexts and across development (Ghavi-Helm et al., 2014; Jin et al., 2013). Collectively, these findings underscore that chromatin looping is an important mechanism by which long-range interaction between distal regulatory elements and genes may be achieved.

Building upon these findings, recent functional studies using gene-editing tools have further suggested a causal link between chromatin looping and gene regulation. It has long remained unclear whether looped interactions are a prerequisite for or merely a consequence of gene regulation. Direct evidence that chromatin looping between a gene promoter and strong enhance can lead to transcriptional activation has recently been obtained (Deng et al., 2012; 2014). The Blobel group, in collaboration with synthetic ZFN pioneers Sangamo Biosciences, showed that chromatin loops may be induced between the globin locus control region (LCR) and the beta-globin promoter in ΔGATA1 murine cells using synthetic zinc-finger proteins tethered to the self-association domain of Ldb1. These induced chromatin loops led to substantial activation of ß-globin transcription in the absence of GATA1 (Deng et al., 2012). Using the same approach, the group also more recently demonstrated that forced LCR-promoter looping could lead to transcriptional reactivation of the developmentally silenced fetal ϒ-globin gene in adult murine erythroblasts (Deng et al., 2014). These new insights argue that, in the proper context, forced chromatin looping can directly guide transcriptional activity (Dekker and Misteli, 2015; Deng and Blobel, 2014).

Many factors, including transcription factors (*e.g.*, CTCF, YY1, NRSF), co-activators (e.g., Mediator), chromatin structural proteins (*e.g.* cohesin), and noncoding RNAs (*e.g.*, Xist (Lee, 2012), Firre (Hacisuleyman et al., 2014; Yang et al., 2015). HOTTIP (Wang et al., 2011a)) have

been shown to play roles in mediating chromatin looping; the roles of CTCF and cohesin in spatial genome organization are by far the best characterized. Both CTCF and cohesin have been found to bind thousands to tens of thousands genomic sites, a significant portion of which are co-occupied by both proteins in mammalian cells (Bouwman and de Laat, 2015; Fraser et al., 2015; Phillips and Corces, 2009). Early studies also established CTCF as a transcription factor with versatile roles in transcription activation and repression, as well as a global insulator protein (Bell et al., 1999; Lobanenkov et al., 1990). Cohesin is best known for its role in sister chromatid cohesion, chromosome segregation, and DNA repair (Wood et al., 2010).

Insights obtained from recent studies have also suggested that CTCF and cohesin play important roles in the hierarchical folding of the interphase chromosome, from chromatin looping to establishment of chromatin domains. It has been found that CTCF mediates thousands of chromatin loops in mouse and human genomes, which account for a substantial portion of all the loops detected in a genome (Dowen et al., 2014; Handoko et al., 2011; Ji et al., 2015; Rao et al., 2014; Takizawa et al., 2008; Tang et al., 2015). The formation of CTCF-mediated loops requires cohesin, which also colocalizes with mediators to facilitate tissue-specific promoter-enhancer looping (Kagey et al., 2010). Moreover, it has been revealed that the orientation of CTCF binding guides directional chromatin looping (de Wit et al., 2015; Guo et al., 2015; Rao et al., 2014; Sanborn et al., 2015). This is in agreement with an extrusion model of loop formation (Alipour and Marko, 2012; Sanborn et al., 2015).

At the chromatin-domain level, it is believed that CTCF and cohesin also play important roles. CTCF binding has been found enriched at LAD boundaries, suggesting involvement in the formation of LADs (Guelen et al., 2008). CTCF and cohesin are also enriched at the boundaries of TADs, and depletion of CTCF results in the elimination of TADs and loops, with minimal

transcriptional effects (Nora et al., 2017). As mentioned briefly above, this was also shown recently by a study demonstrating that IDH mutations promote gliomagenesis by disrupting CTCF binding via hypermethylation, in turn disrupting TAD boundaries and allowing aberrant enhancer-promoter interactions to activate normally insulated oncogenes (Flavahan et al., 2015). These results suggest a general role for CTCF and cohesin in chromatin folding and genome compartmentalization.

1.4.9 *Biochemically Defined Domains: concentration gradients, residence times, and genomic measurement*

Genomic and microscopic techniques have provided us with an invaluable, descriptive understanding of how chromosomes hierarchically fold into domains. Still, our ultimate understanding of what these domains are, how these domains are formed, and their functional significance rests entirely on our ability to define the biophysical and biochemical parameters governing their formation. To properly do so, we must first understand the limitations of both genomic and microscopic assays, and synthesize our descriptive knowledge of chromatin state with existing biophysical and biochemical paradigms.

A common assumption made by chemical engineering students studying for their first fluid mechanics exam is the assumption that a reaction volume is well-mixed. This assumption simplifies calculations, allows for facile modeling of processes like heat and mass transfer, and for many engineering applications, holds nicely. Since the earliest observations of nuclear speckles, however, it has been well-accepted that the nuclear volume is not, in any sense, well-mixed. Local concentration gradients are, in fact, essential to biology. The aforementioned transcription factories serve as volumetric hubs for Polymerase II, Mediator, and other essential transcriptional components (Cisse et al., 2013). The nucleolus, a membraneless organelle, serves as a conserved

depot for RNA Polymerase I and the rDNA locus, supporting the rRNA synthesis necessary to support cellular function and translation. Mammalian limb development, largely defined by the coordinated expression of a conserved cluster of homeobox *Hox* genes, is thought to be driven by dynamic changes in the local concentration of *Hox* transcription factors at the gene cluster (Crocker et al., 2015). The mechanisms underlying the formation of these local concentration gradients, however, remain poorly understood.

Recent work has raised the possibility that a common biophysical phenomenon, liquid-liquid phase separation (LLPS) may contribute to the formation of local concentration gradients within the nucleus. Briefly, LLPS (reviewed in detail by Shin and Brangwynne (2017)) occurs when sufficiently high local concentrations of proteins capable of making multivalent protein-nucleic acid and protein-protein interactions (*e.g.* through intrinsically disordered regions) is achieved; when such conditions are met, the thermodynamically favorable outcome is phase separation of the factors in this sub-volume from their surroundings, leading to the formation of a novel compartment. While phase separation has long been associated with the formation of membraneless organelles like the nucleolus, P-bodies, and stress granules, it has more recently been directly linked to transcriptional regulation. HP1α is a factor associated with the formation of heterochromatin; recent work has demonstrated that phosphorylated forms of the protein phase separate *in vitro*, *in vivo*, and in living fly embryos (Larson et al., 2017; Strom et al., 2017). Phase separation has also been touted as a model for understanding the regulation of promoters by enhancers. These observations raise the possibility that the above-described domains may, in fact, be maintained through this biophysical phenomenon. Importantly, however, observations to date have simply associated the phenomenon of LLPS with transcriptional regulation. Future work must, of course, focus on causally linking the formation of membraneless compartments with

proper deployment of the regulatory processes that define health and disease. Still, the possibility of linking the observations of nuclear phase separation with the domains defined through genomics remains tantalizing.

Concepts tangentially related to the maintenance of local concentrations within the nuclear milieu are those of transcription factor residence time and transcription factor occupancy. Sequence-specific transcription factors vary greatly in their relative affinity and specificity for DNA (Jolma et al., 2013). Certain factors, like the architectural protein CTCF, display high specificity for particular motifs, and remain bound to DNA for timescales on the order of minutes (Hansen et al., 2017). Other factors, however, including the Yamanaka factors C-MYC and SOX2, demonstrate markedly lower "information content" in their predicted sequence specificities, and far shorter dwell times (8 – 12 seconds) on chromatin (Chen et al., 2014; Phair et al., 2004). Transcriptional programs (*e.g.* promoter/enhancer activation) critically depend on the occupancy of factors like CTCF, C-MYC, and SOX2, which is a function of the residence time of the factor at a given site, and the local concentration of the factor in a given nuclear volume. Thus, critical molecular assemblies, like loops between the SOX2 enhanceosome complex at pluriplotency enhancers, or C-MYC dependent transcription factories, may exist for tens of seconds, and likely assemble in an asynchronous fashion across single cells.

Genomic assays, and genomic analyses, inherently rely on population averaging. Conventional sequencing assays perform biochemical measurement on hundreds of thousands to millions of cells, and as such any measurements made, like transcription factor occupancy, are averaged over the biochemical state of all of those cells at the time of cell harvest. This issue of averaging is not ameliorated by single-cell genomic analysis; even in this context, large numbers of cellular observations are critical to ensure that any biological observations are not simply the

consequence of noise. This can be problematic, as these assays are prone to a massive amount of ascertainment bias. CTCF-CTCF loops are the strongest signal in high-resolution Hi-C data, but CTCF also has one of the longest documented residence times out of surveyed transcription factors, with its binding partner cohesin displaying a residence time of 20 minutes (Hansen et al., 2017). Conversely, enhancer-promoter contacts are weak, or often altogether absent from high-resolution contact mapping datasets. The question must at some point be asked: to what extent are Hi-C and other genomic measurements driven by biochemically-rooted ascertainment biases? Biochemical definition of loops, domains, compartments, and territories will require a careful understanding of what, precisely, these features are at a biophysical level, how these features arise from often transient interactions between DNAs and proteins, and ultimately which high-throughput sequencing assays are best suited to defining and characterizing these features in an unbiased way.

1.5 Future Directions

The synthesis of classical microscopy-based approaches and more recent high-throughput biochemical techniques has led to an explosion in our knowledge of the physical organization of eukaryotic genomes. Through a diverse array of techniques including electron microscopy, FISH, ChIP-seq, DamID and 3C and its derivatives, we are generating increasingly fine-scale catalogs of the chromatin loops, self-associating domains, and chromosomal territories that comprise the eukaryotic nuclear genome. Given this dense catalog of structural elements, then, we believe that the field will eventually move into two primary directions: i.) functional dissection of this vast catalog of structural elements, and ii.) large-scale characterization of the dynamics and mechanisms of chromatin folding both across biological processes such as differentiation, and across homogenous and heterogeneous cell populations.

1.5.1 *Functional Dissection of Structural Elements*

The advent of CRISPR/Cas9 as an easy to use, highly multiplexable system for perturbing primary sequence has opened up considerable avenues for testing the functional significance of genomic elements. We predict the continued use of genome-editing reagents in validating key structural elements (*e.g.* CTCF binding sites), with respect to various phenotypes of interest (*e.g.,* pathogenicity, dysregulation of global and local gene expression). Already, several groups have successfully utilized Cas9-mediated genome editing to generate clonal populations harboring inverted or deleted transcription factor binding sites, and have performed assays like Hi-C and RNA-seq to link structural and functional changes (Guo et al., 2015; Lupianez et al., 2015; Sanborn et al., 2015).

As low-throughput (*e.g.* testing of single edited clones) approaches become more popular, we anticipate the eventual development of high-throughput screens for large-scale characterization of structural elements. Already, genome-editing-based lentiviral and *in vivo* saturation mutagenesis screens have been employed, to dissect the functional significance of genes (Shalem et al., 2014; Wang et al., 2015; 2014), codons (Findlay et al., 2014), small insertions/deletions (indels) (Canver et al., 2015; Vierstra et al., 2015), and single nucleotide polymorphisms (Findlay et al., 2014). A key next step in determining the functional significance of catalogued elements will be employing such approaches to perturb key structural features in a variety of biological contexts; these experiments may be critical to eventually understanding the link between human disease phenotypes (*e.g.* cancer) and dysregulation of chromatin architecture.

1.5.2 *Characterizing Structural Dynamics Across Time and Space*

regarding the dynamics of chromatin—the processes by which chromatin architecture and state change as a function of a given biological process—remain largely unanswered. The nascent

field of single-cell epigenomics (Schwartzman and Tanay, 2015), however, has offered a key set of tools that may finally be able to address such questions. While traditional epigenomic assays must be performed on populations of cells, single-cell epigenomics provide an opportunity to characterize heterogeneity within populations—an invaluable tool for both defining novel cell types from a heterogenous population (*e.g.* an organ system), and for characterizing transitory states in biological processes such as differentiation. Recently published approaches such as single-cell DamID (Kind et al., 2015), single-cell ChIP-seq (Rotem et al., 2015), and single-cell Hi-C (Nagano et al., 2013; Ramani et al., 2017) all provide valuable proofs-of-concept for such assays. The next step, then, is to scale these approaches to easily process hundreds of thousands of single-cells. We recently described a method that leverages combinatorial DNA barcoding of single-cells to provide chromatin accessibility, Hi-C, or RNA-seq information from thousands of cells in a single experiment (Cao et al., 2017; Cusanovich et al., 2015; Ramani et al., 2017). Such approaches may be adapted to other epigenomic assays, including DamID, and ChIP, thus providing a way forward to achieving the required throughput to confidently define new cell types, or organize populations of cells going through some biological process into some sort of "pseudotime."

It may also be useful to consider the marriage of single-cell biochemical techniques with complimentary microscope-acquired "*in situ*" transcriptomic datasets (Chen et al., 2015; Lee et al., 2014; Lubeck and Cai, 2012; Lubeck et al., 2014). *In situ* transcriptomics may, for example, be necessary to properly spatially organize large populations of tissue-derived nuclei in some biologically meaningful way. Furthermore, by matching *in situ* transcriptomic data with replicate single-cell epigenomic experiments in this way, one may be able to link differential genome

architectural features with gene regulatory phenomena, in the process furthering our progress towards ultimately understanding the links between 3D genome architecture and gene regulation.

Of course, the application and development of any of these techniques is intertwined with the development of data analytical techniques. While algorithmic development will have to keep pace with the development of these technologies, we believe the incredible strides already made in the relatively young field of single-cell RNA sequencing (Trapnell, 2015) are a positive indicator that analytical methods will be able to keep pace with this exploding field.

1.6 Closing Remarks

There are many fundamental and long-standing biological questions linked to 3D genome architecture: Does genome architecture itself define cellular identity? How does chromatin state (*i.e.* histone modifications, DNA methylation) impact higher-order chromatin structure? How might defects or differences in 3D genome architecture lead to human disease? Obtaining the knowledge necessary to answer these questions requires a multi-pronged approach employing creative microscopic, biochemical, and computational tools. As reviewed here, these are thankfully requirements that the field is actively addressing, suggesting that we will be well-positioned to answer many if not all of these pressing questions in the years to come.

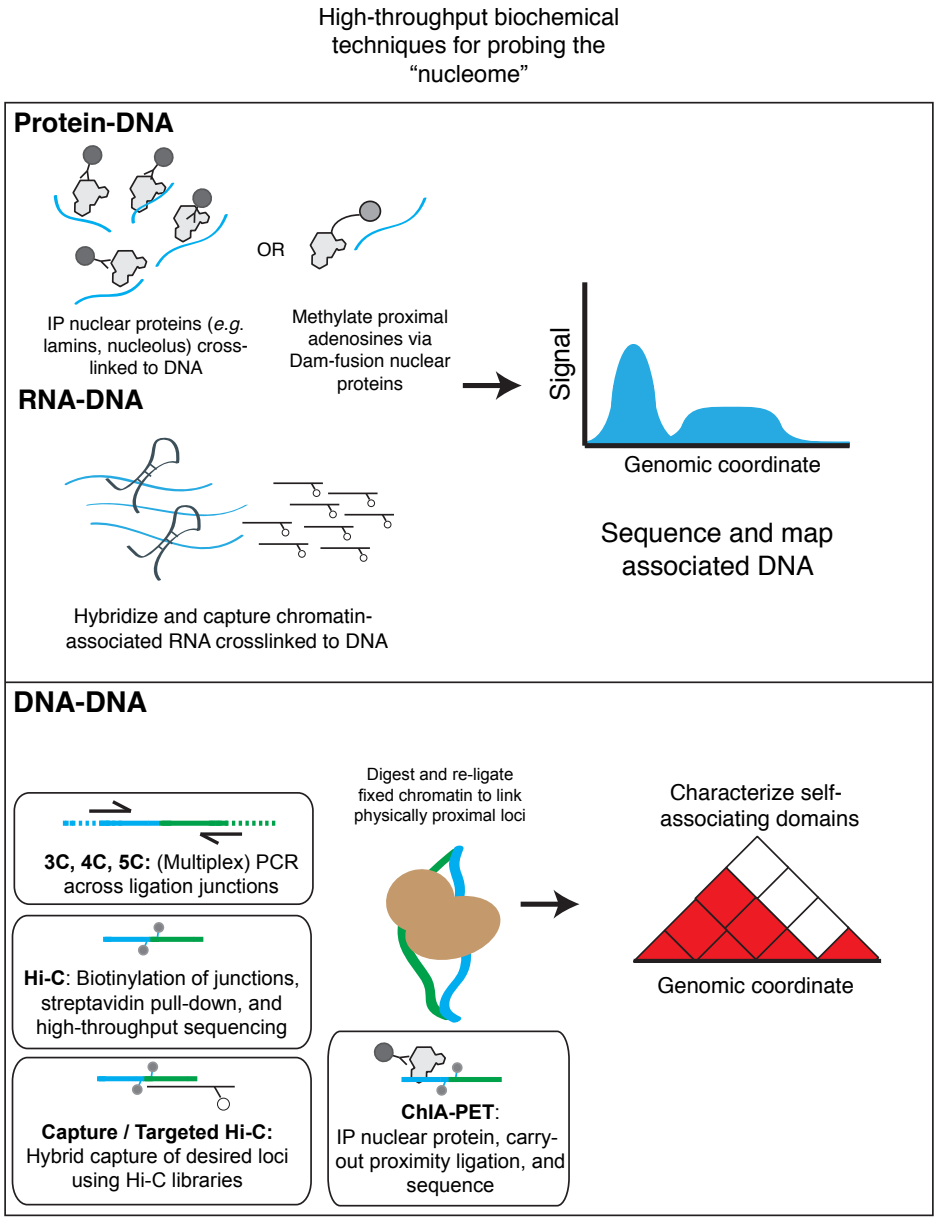


Figure 1.1. Schematic summary of high-throughput techniques for probing three-dimensional genome architecture.

Table 1.1. Table summarizing methods in the "3C" family

Scale	**Method**
Whole-genome	Hi-C
	TCC
	Single-Cell Hi-C
	DNase Hi-C
	In situ Hi-C
	Micro-C
	In situ DNase Hi-C
Targeted	3C
	ChIP-loop
	4C
	e4C
	5C
	ChIA-PET
	Capture-C
	Capture-Hi-C
	Targeted DNase Hi-C

Chapter 2. HIGH-THROUGHPUT DETERMINATION OF RNA STRUCTURE BY PROXIMITY LIGATION

Note: Chapter 2 was published in the September 2015 issue of *Nature Biotechnology* as:

Ramani V., Qiu R., Shendure J. "High-throughput determination of RNA structure by proximity ligation**."** *Nature Biotechnology* (2015)

2.1 ABSTRACT

We present an unbiased method to globally resolve RNA structures through pairwise contact measurements between interacting regions. RNA Proximity Ligation (RPL) uses proximity ligation of native RNA followed by deep sequencing to yield chimeric reads with ligation junctions in the vicinity of structurally proximate bases. We apply RPL in both baker's yeast (*Saccharomyces cerevisiae*) and human cells and generate contact probability maps for ribosomal and other abundant RNAs, including yeast snoRNAs, the RNA subunit of the signal recognition particle, and the yeast U2 spliceosomal RNA homolog. RPL measurements correlate with established secondary structures for these RNA molecules, including stem-loop structures and long-range pseudoknots. We anticipate that RPL will complement the current repertoire of computational and experimental approaches in enabling the high-throughput determination of secondary and tertiary RNA structures.

2.2 INTRODUCTION

The folding of RNA species into complex secondary and tertiary structures is central to RNA's catalytic, regulatory, and information-carrying roles (Doudna et al., 2014). Pioneering approaches for elucidating RNA structure—including crystallography (Cate et al., 1996), electron

microscopy (Wang et al., 1994), and spectroscopy (Latham et al., 2005)—are technically complex and difficult to scale, motivating the development of computational algorithms for RNA structure prediction (Lorenz et al., 2011; Reuter and Mathews, 2010; Zuker and Zuker, 2003). Current algorithms have limited predictive power, particularly for long-range interactions such as pseudoknots (secondary structures involving intercalated stem loops).With the advent of massively parallel sequencing (Shendure and Lieberman-Aiden, 2012), less laborious experimental techniques have been developed for the global interrogation of RNA secondary structures. These include methods relying on structure-specific chemical modifications (Ding et al., 2014; Lucks et al., 2011; Rouskin et al., 2014), such as DMS-seq and SHAPE-seq, as well as methods involving digestion with structure-specific RNases (Kertesz et al., 2010; Underwood et al., 2010; Wan et al., 2014), like PARS-seq and Frag-seq. Although these methods probe the extent to which individual bases participate in secondary structures, they do not directly query which specific pairs of bases or regions interact to form these structures. To address this, recent efforts have combined systematic mutagenesis and structure-specific probing to generate pairwise information for inferring RNA folds (Kladwang et al., 2011; Siegfried et al., 2014). However, despite considerable progress, the high-throughput determination of RNA secondary and tertiary structures remains a challenging problem.

Here we show that proximity ligation is a straightforward means of generating global pairwise data about RNA secondary and tertiary structure. Proximity ligation records the physical proximity of two nucleic acid termini through their ligation, and has been applied to detect DNA aptamer-bound proteins (Fredriksson et al., 2002), to probe protein-protein interactions via antibody-bound oligonucleotides (Soderberg et al., 2006), and for targeted or global chromosome conformation capture (3C) (Dekker et al., 2002; Lieberman-Aiden et al., 2009). Proximity ligation

has also been applied in conjunction with crosslinking and either affinity purification or immunoprecipitation to characterize snoRNA-rRNA interactions (Kudla et al., 2011) and Ago-mediated miRNA-target interactions (Helwak et al., 2013). However, these efforts have primarily focused on assessing specific *trans* interactions, rely on low-efficiency 254 nanometer UV crosslinking, and require time-consuming purification steps.

2.3 RESULTS

RPL ('ripple') globally assesses which pairs of regions are interacting to form intramolecular RNA structure (**Figure 2.1**). Similar to 3C methods for DNA conformation, RPL uses digestion and re-ligation of RNA, but omits crosslinking, relying instead on the inherent spatial proximity of RNA nucleobases in secondary structural features (i.e. stem-loops). To generate RPL libraries, we performed RNase digestion *in situ* (or, for yeast, took advantage of endogenous single-stranded RNases), followed by treatment with exogenous T4 RNA Ligase I under non-denaturing conditions. These steps result in chimeric molecules formed from RNA strands intra-molecularly ligating across digested loops (**Figure 2.1a**, inset). By deeply sequencing these resulting fragments and quantifying the relative abundance of specific intramolecular ligation junctions, we are able to create pairwise contact maps that reflect the short- and long-range stem-loop and pseudoknot interactions of intramolecular RNA secondary structures.

First we tested RPL in the budding yeast *S. cerevisiae*. To create libraries, we spheroplasted whole yeast cells for 1 h with zymolyase (dissolved in 1X PBS without DTT to allow endogenous RNases to remain active). We then treated the resulting slurries with T4 polynucleotide kinase (PNK) to convert 5′-hydroxyl to 5′-phosphate termini, and diluted and incubated these mixtures overnight in the presence of a single-stranded RNA ligase (T4 RNA Ligase I) under non-denaturing conditions. We then purified total RNA using acid guanidinium-phenol, and carried

out a standard RNA-seq library preparation. Sequencing (Illumina) yielded 304 million (M) concatenated reads for a (+) ligase sample, and 342M concatenated reads for a (-) ligase control sample (**Methods**).

To identify candidate ligation junctions in these sequencing reads, we adapted an algorithm for identifying novel RNA isoforms from RNA-seq data (Dobin et al., 2012), relaxing constraints on splice-site composition to more generally recognize intramolecular chimeric reads that map discontinuously to a single RNA sequence (**Methods**). To quantify the enrichment of candidate ligations in our samples, we first examined the distribution of spanned distances of intramolecular chimeric reads (i.e. gap sizes), per million reads, in both (+) and (-) ligase samples. Although the overall fraction of reads corresponding to candidate intramolecular ligation junctions is low, the (+) ligase sample is enriched for these across a broad range of spanned distances (0.28% in (+) ligase sample vs. 0.011% in (-) ligase sample; **Figure 2.5**).

Potential sources of technical artifacts in these data include the formation of chimeric molecules by reverse transcriptase (RT) template switching, systematic mapping artifacts, PCR-mediated duplicates and non-specific ligation events. To reduce the impact of RT template switching, we discarded candidate ligation junctions with >5 nucleotides (nt) microhomology, as well as those mapping to opposite strands. To remove PCR-mediated duplicates, we collapsed all reads with identical mapping coordinates and CIGAR alignment strings. To reduce the impact of systematic mapping artifacts caused by errors within our reference transcriptome (for example, gross deletions, un-annotated splice junctions), we conservatively discarded candidate ligation junctions containing the highest 1% of ligation counts, for each RNA species analyzed. Finally, to quantify the extent of nonspecific ligation, we performed an experiment in duplicate wherein human cells were taken through a modified version of the RPL protocol (**Methods**) and spiked

into yeast slurries immediately before proximity ligation. The resulting data demonstrate marked enrichment for intraspecies, intramolecular chimeric reads (**Figure 2.6**).

We first analyzed RPL data in the context of the complex but extensively validated secondary structures of the yeast ribosomal RNAs (rRNAs). The yeast ribosome is comprised of the 60S large subunit (LSU), which includes the 3.4 kb 25S rRNA and short 5.8S and 5S rRNAs, and the 40S small subunit (SSU), which includes the 1.8 kb 18S rRNA. To assess whether RPL captures the proximity implied by secondary structure base-pairing, we tallied candidate ligation junctions in a 500 base-pair window centered on known base pairs of the established rRNA structures, effectively quantifying ligation probability as a function of distance (in linear sequence) from known base pairs (in secondary structure). We observe an enrichment of candidate ligation junctions immediately proximal (i.e. within 10 nt) to known base pairs in both the 5.8S/25S rRNAs (~9-fold; Fig. 1b) and 18S rRNA (~6-fold; Fig. 1c). Furthermore, in the case of the 5.8S/25S rRNAs, which contain many long-range base-pairing interactions, this enrichment is maintained even if we restrict analysis to candidate ligation junctions that span >100 bases in the linear sequence (**Figure 2.7**).

The observed signal is entirely dependent on the inclusion of ligase, and is not explained by sequencing errors, mapping artifacts or by proximity in sequence space (as opposed to structure space). As such, we conclude that it primarily derives from intramolecular ligation events between structurally proximal bases. Nonetheless, the signal shown in **Figure 2.1b,c** is "noise-averaged" over all base pairs in these rRNA structures. Consistent with the stochastic nature of individual ligation events, we observe weaker enrichment when repeating our analysis with a randomly selected subset of 10, 25, or 50 paired bases in either the LSU or SSU rRNAs (**Figure 2.8**). The ligation junctions that we observe are also clearly affected by other biases, including the bias

against G/C extremes routinely seen with Illumina sequencing, as well as more subtle base composition preferences at the ligation junction (**Figure 2.9**). We also observe that ligation junctions are enriched for single-stranded bases in the LSU and SSU rRNAs (Odds Ratio (OR) = 2.24; $P < 2.2E-16$, Fisher's Exact Test). This bias, and the noisiness of the raw data, is evident when ligation junctions are overlaid onto a known secondary structure (**Figure 2.2a**).

Given these observations, we concluded that the signal of RPL likely arises from the combinatorial digestion and ligation of predominantly unpaired ribonucleotides across broken loop structures. Considering this, along with the stochastic, biased nature of individual ligation events, we speculated that our ability to resolve secondary structure would improve by calculating the frequency of ligation events between pairs of sliding windows (21 nt each), effectively capturing a combinatorial diversity of ligation events surrounding secondary structural elements. Concurrent with this, we adapted normalization methods developed for Hi-C matrices (Rao et al., 2014) to account for other one-dimensional biases (for example, sequence biases of RNA ligase and PCR) (**Methods**). We then visualized these normalized RPL scores, calculated for pairwise windows, by directly overlaying them onto known secondary structures. RPL scores broadly mirror the secondary structures of the 5.8S/25S LSU rRNAs (**Figure 2.1d**, **Figure 2.2b**; **Figure 2.10a**) as well as the SSU 18S rRNA (**Figure 2.10b**). Furthermore, we observe signal corresponding to distal tertiary structures, including long-range "pseudo-knots" in the LSU rRNAs (**Figure 2.1d**, right inset) (Ben-Shem et al., 2011).

We next sought to evaluate the correspondence between proximity ligation events and the structures of non-ribosomal RNA transcripts. Because we are limited by sampling depth, we focused on well-characterized, abundant RNAs; specifically, the snoRNA *snR86* (**Figure 2.3a**), which guides uridylation of the LSU rRNA, the U1 spliceosomal RNA (*snR19*) (**Figure 2.3b**), the

RNA component of the signal recognition particle (*SCR1*) (**Figure 2.3c**), and the U2 spliceosomal RNA homolog (*LSR1*) (**Figure 2.11**). In "contact probability maps" for these RNAs (based on the normalized RPL scores described above), we observe a striking anti-diagonal pattern, reminiscent of signal observed at known stems in the 5.8S/25S and 18S rRNAs. When comparing our contact probability maps to secondary structure predictions generated with INFERNAL (Nawrocki et al., 2013) using covariance models taken from Rfam (Burge et al., 2012) our observations are consistent with conserved stems in both *snR86* and *snR19* (**Figure 2.3a,b**). In RPL measurements for *snR19*, we also observed signal indicative of stem formation in the region comprising bases 320-510—MFE predictions suggest that this region can form a helix, raising the possibility that this structure is present endogenously.

We also analyzed RPL measurements in the context of a non-ribosomal RNA with a solved structure, the RNA subunit of the signal recognition particle (*SCR1*). Again, we observed broad agreement between RPL scores and regions containing paired bases (**Figure 2.3c**), though we do find that certain expected long-range interactions (for example, folding between the molecule termini) are not seen. Further work will be needed to determine whether this is simply an artifact of insufficient depth-of-coverage, or is symptomatic of some other bias with respect to the classes of structural elements that proximity ligation can resolve.

Finally, our observations for *LSR1* (**Figure 2.11**) are consistent with previous work employing cross-linking, affinity-purification, and proximity ligation of RNA (Kudla et al., 2011), which found ligation products supporting stem-formation between the two termini. In agreement with this cross-linking based approach, our data support the formation of both proximal (for example, stem formation at bases 1100 – 1150), and distal folds.

We next explored the value of RPL scores as a predictive tool for classifying pairs of interacting regions within a structured RNA. To show that RPL scores can be used in this manner, we examined their positive predictive value (PPV) at varying quantile thresholds for the gold-standard 5.8S/25S and 18S rRNAs (**Figure 2.4a,b**). This is a challenging classification problem (92,392 true positive interacting windows out of 6,317,235 possible interacting windows for the LSU rRNAs (1.5%); 41,981 true positive interacting windows out of 1,620,900 possible interacting windows for the SSU rRNA (2.6%)). The highest RPL scores are strongly enriched for true positive interacting windows (LSU rRNA: PPV of 54% using the top 1% of RPL scores; SSU rRNA: PPV of 61% using the top 1% of RPL scores). Plotting PPV as a function of threshold illustrates the tradeoff with sensitivity (**Figure 2.4c,d**). For example, at a sensitivity of 50%, RPL scores have a PPV of 43% for the LSU rRNA and 27% for the SSU rRNA, for predicting structurally interacting pairs of regions.

2.4 DISCUSSION

The high-throughput, unbiased identification of intermolecular RNA-RNA interactions is of strong interest in the RNA biology field. Recent work has shown that psoralen-mediated crosslinking may be used in tandem with anti-sense purification to capture *trans* RNA-RNA interactions (Engreitz et al., 2014). In principle, RPL should be able to provide complementary information, as interacting RNAs may form ligation products at a higher rate than non-interacting RNAs. Although we observed a modest enrichment for intermolecular yeast ligation junctions in the species mixing experiment (**Figure 2.6**), this enrichment in our yeast RPL experiment derives primarily from ligation products between the small and large ribosomal subunits (**Figure 2.12**). While no inter-subunit RPL scores approached those of strongly interacting intramolecular windows, it remains possible that a combination of methodological improvements to reduce

background and deeper sequencing of RPL libraries may enable global surveys of *trans* RNA-RNA interactions (for example, the signal recognition particle-ribosome interaction; subunit interactions in the translating ribosome).

We next sought to adapt RPL to generate secondary structure information corresponding to RNAs in human cells. Most notably, we replaced the zymolyase treatment with a limited *in situ* digestion with exogenous single-stranded RNases A and T1. In analyzing the resulting data in the context of the well-studied human ribosomal RNAs, we again observed correlation of high RPL scores with known interacting regions (**Figure 2.13**). However, an RNase (-), ligase (-) control also demonstrated signal that correlated with secondary structure, albeit much more weakly and possibly reflecting endogenous nuclease and ligase activity (**Figure 2.14**). The possibility that endogenous enzymatic activity may contribute to the formation of chimeric RNAs is not novel; recent work using a cross-linking approach to characterize the miRNA interactome of *C. elegans* curiously found that expected ligation products could form in the absence of exogenous T4 RNA Ligase I (Grosswendt et al., 2014).

We anticipate several directions for improving RPL. First, RPL libraries require deep sequencing to reliably map interacting regions, even for highly abundant RNA species. The sufficient sampling of lower-abundance RNA species of interest (for example, mRNAs) might be achieved by optimizing the enzymatic steps of the protocol, by adopting hybrid capture enrichment or subtraction, or simply by brute force deep sequencing.

Second, given the high predictive value (Das et al., 2012; Kladwang et al., 2011; Rouskin et al., 2014; Siegfried et al., 2014) of *in vivo* structure probing methods (for example, DMS-seq, SHAPE-seq) in determining the pairedness of individual bases in secondary structures, a framework that integrates two-dimensional, lower-resolution RPL data with one-dimensional,

higher-resolution structure probing data seems highly attractive. Ideally, computational predictions would be integrated at the same time, thereby taking advantage of three largely orthogonal approaches to maximize the accuracy of RNA structural predictions.

The current repertoire of high-throughput empirical assays for RNA secondary structure provides us with a deep, but ultimately one-dimensional window into the structural landscape of RNA molecules. In contrast, RPL globally captures information with respect to pairwise interactions within RNA secondary structures. Through its integration with complementary computational and experimental approaches, we anticipate that RPL will facilitate the high-throughput elucidation of RNA secondary structures in diverse organisms.

2.5 METHODS

2.5.1 *Cell culture.*

S. cerevisiae strain FY3 was struck out on YPD plates and grown at 30 °C. Mammalian cells (lymphoblastoid cell line GM12878; Coriell) were cultured at 37 °C, 5% CO_2 in RPMI-1640 supplemented with 1X Anti-Anti (Gibco), 1X Plasmocin (Invivogen), and 15% FBS (Gibco).

2.5.2 *RNA Proximity Ligation (RPL).*

Individual yeast colonies were added directly to 0.5 U Zymolyase in 10 uL 1X phosphate buffered saline (PBS) (Gibco) w/ 0.2% IGEPAL (Sigma) and incubated at 37 °C for 60 min to spheroplast while maintaining endogenous RNase activity. Spheroplasted yeast were immediately transferred to ice, and mixed with 0.5 uL SuperASE-In (Ambion), 2.5 uL T4 PNK (NEB), 5 uL 10X T4 DNA Ligase Buffer w/ 10 mM ATP (NEB), and 32 uL 1X PBS w/ 0.2% IGEPAL, after which the slurry was incubated at 37 °C for 30 min. Following end-repair, complexes were immediately transferred to 450 uL ligation reaction mix (50 uL 10X T4 DNA Ligase Buffer w/ 10

mM ATP (NEB); 5 uL SuperASE-In (Ambion), 12.5 uL T4 RNA Ligase I (NEB), 382.5 uL 1X PBS w/ 0.2% IGEPAL), and incubated overnight in a 16 °C water bath, after which complexes were added to 1.5 mL TriZOL (Ambion). Samples were then purified using Direct-ZOL spin columns (Zymo) according to manufacturer's protocols. For mammalian experiments a modified version of RPL was performed wherein 2E6 whole human lymphoblastoid cells (GM12878, Coriell) were treated *in situ* with 0.2 uL of RNace-IT (Agilent) diluted in 9.8 uL 1X PBS w/ 0.2% IGEPAL for 10 min at 22 °C, after which the RPL protocol was followed, beginning with PNK treatment.

T4 PNK is known to have minimal 3′ phosphatase activity under the buffer conditions we use during our end-repair step (Cameron and Uhlenbeck, 1977). To ensure that phosphatase activity was not limiting ligation efficiency, we also repeated our yeast RPL experiments using a low pH imidazole buffer (50 mM imidazole-HCl, pH 6.0, 10 mM $MgCl_2$, 1 mM ATP, and 10 mM DTT) for our PNK reactions. We observed comparable ligation efficiencies independent of the use of low pH buffer (0.28% of analyzed reads in our sample compared to 0.21% and 0.14% in imidazole experiments performed in duplicate).

For spike-in experiments, an individual yeast colony and 5E5 human lymphoblastoid cells were treated with respective RPL treatments described above. Following PNK treatment, the two slurries were mixed and treated with T4 RNA Ligase I overnight, after which complexes were purified as described above.

To quantify the extent of RNA degradation during the yeast RPL protocol, we repeated the yeast RPL experiment, isolating RNA after PNK treatment, as well as after overnight incubations both in the presence and absence of T4 RNA Ligase I. We then analyzed the integrity of these RNA products using an RNA 6000 Nano Lab-on-Chip (Agilent), finding our products were mildly

degraded following PNK treatment (RIN Score of ~7), though this degradation appears to have been halted before ligation (**Figure 2.15**).

2.5.3 *Library Preparation.*

Libraries were prepared according to standard Illumina TruSeq RNA guidelines, with minor changes. Notably, polyA-selection steps were skipped, RNA fragmentation (Elute, Prime, Fragment) was carried out for 2.5 min, and PCR amplification of the final library was carried out using qPCR for 8-12 cycles on a BioRad OpticonMini to prevent library overamplification. Two biological replicate libraries were generated and sequenced for (+) ligase yeast experiments, one of which was selected for deep sequencing and analyzed further in this paper. Two biological replicate libraries each were generated for imidazole and species-mixing experiments, for both (+) and (-) ligase samples.

2.5.4 *Sequencing and sequence alignment.*

Sequencing of libraries was carried out using the Illumina MiSeq, NextSeq 500, and HiSeq 2000 instruments, generating paired-end 80 bp and 101 bp reads.

FASTQ Post-processing: Raw paired-end FASTQ files were adaptor-trimmed and merged with SeqPrep (https://github.com/jstjohn/SeqPrep) to account for all read pairs that contain redundant information (i.e. sequence) content. We then took the resulting "singleton" forward and reverse reads (i.e. those that did not contain sufficient overlap to be fused) and concatenated them along with fused reads to yield 304M (for the treated sample) and 342M (for the negative control) concatenated reads, which were then analyzed.

Alignment: These resulting FASTQ files were aligned to references generated from either a manually curated list of yeast transcripts with duplicated transcripts removed, taken from the Saccharomyces Genome Database (http://yeastgenome.org), or a selected list of deduplicated RefSeq human transcripts, using the STAR aligner with the following parameters:

–outSJfilterOverhangMin 6 6 6 6

–outSJfilterCountTotalMin 1 1 1 1

–outSJfilterDistToOtherSJmin 0 0 0 0

–alignIntronMin 10

–chimSegmentMin 15

–chimScoreJunctionNonGTAG 0

–chimJunctionOverhangMin 6

2.5.5 *Bioinformatic Analyses.*

Secondary structures in BPSEQ format for *S. cerevisiae* were downloaded from the Comparative RNA Website (Cannone et al., 2002) and RNA structures were visualized through a modified version of VARNA. *H. sapiens* rRNA structures were inferred from a published cryo-EM structure (Anger et al., 2013), using 3DNA (Lu et al., 2003). STAR-generated output was analyzed with custom Python and R scripts to generate contact probability maps. First, STAR alignments were deduplicated by collapsing all alignments with identical start coordinates and CIGAR strings. These deduplicated alignments were then converted to "splice junction" and "chimer" files using awk, and ligation junctions were parsed from these files. For specific species of interest, these ligation counts were then filtered further to remove the highest 1% of counts between individual pairs of bases. To calculate the distribution of ligations around known base-pairs, we looked at all pairs of bases (*i,j*) in our secondary structure BPSEQ files, and calculated

the abundance of ligation events between ($i, j - 250$) to ($i, j + 250$) for each base. For sub-sampling experiments, we randomly sampled 10, 25, or 50 paired-bases and repeated these calculations.

To compute RPL scores, which measure the extent of ligation between two regions of a molecule, we first considered the sparse matrix M where M_{ij} is the ligation count between base i and base j. To generate the RPL score matrix M^*, we compute the coverage at each base i and j (c_i; c_j) and generate a normalized matrix M_{norm} such that:

$$M_{ij}^{norm} = \frac{M_{ij}}{\sqrt{c_i c_j}}$$

We then use this normalized matrix to generate M^* by binning all normalized scores:

$$M_{ij}^{*} = \sum_{a=i-10}^{i+10} \sum_{b=j-10}^{j+10} M_{ab}^{norm}$$

Classification analyses were performed as follows: we thresholded the RPL scores resulting from the above smoothing by quantiles, with a quantile step size of 0.001, and classified true positive interacting windows as those interacting 21 nt windows with RPL scores greater than our specified threshold, that also contain at least 1 set of paired bases.

To generate secondary structures for *snR86* and *snR19*, we downloaded covariance models from Rfam (*snR86* Accession: RF01272; *snR19* Accession: RF00488), aligned respective yeast sequences to their covariance models using the cmalign method from INFERNAL v1.1.1, and converted the resulting Stockholm alignment files to BPSEQ format using VARNA.

Structures of the yeast ribosome (PDB Accession: 4V88) were visualized using PyMol (http://www.pymol.org/).

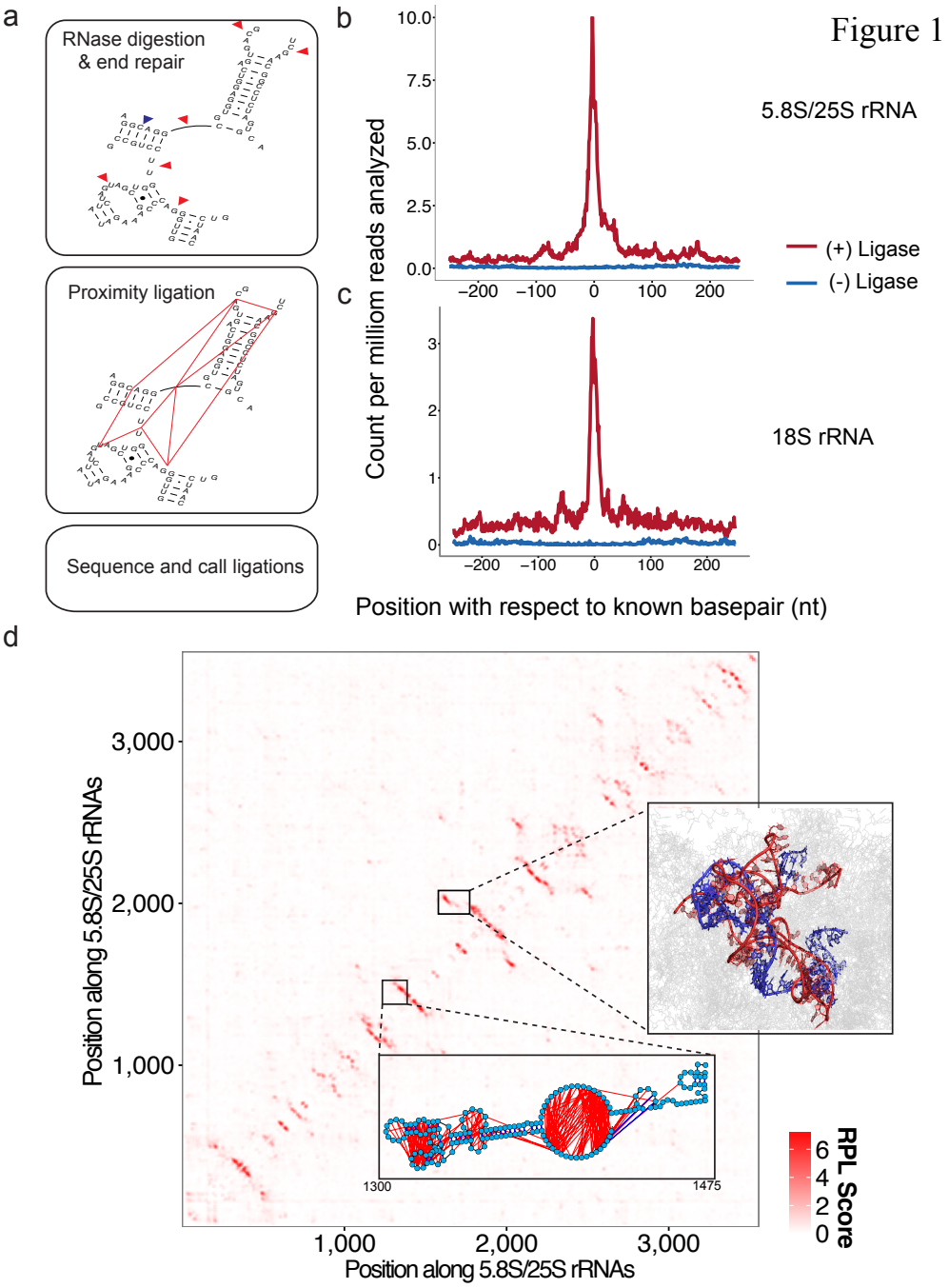


Figure 2.1. RNA Proximity Ligation identifies structurally proximate regions within the complex secondary structures of *S. cerevisiae* ribosomal RNAs.

a.) A schematic representation of the RPL method. Whole cells are spheroplasted with zymolyase and RNA is allowed to react with endogenous RNases. RNA ends are repaired *in situ* via T4 PNK to yield 5'-phosphate termini. Complexes are ligated overnight in the presence of T4 RNA Ligase I. Ligation products are cleaned up via acid guanidinium-phenol and subsequent DNase treatment,

and subjected to Illumina TruSeq RNA-seq library preparation. These libraries are sequenced to map and count ligation junctions; b.-c.) We examined the distribution of ligation junctions as a function of distance from known base-pair partners in the 25S/5.8S rRNA and 18S rRNAs. Ligation products capture the structural proximity implied by base-pairing relationships, as evidenced by the enrichment for ligation junctions immediately near paired bases. Y-axes are shown as ligation counts per million reads analyzed. d.) Contact probability map for the eukaryotic 5.8S/25S rRNA based on RPL scores, which are calculated from the frequencies of ligation events between pairs of 21 nt windows (Methods). Lower inset: Ligation events, shown for bases 1300 to 1475 of the LSU rRNA in orange, primarily occur across digested single-stranded loops. RPL scores effectively smooth this noisy signal and are enriched for pairs of interacting regions. Plotted here are the 8,463 ligation events where both nucleotides fall within the displayed domain (compared to 17,029 ligation events where one nucleotide falls within the displayed domain and one does not, not shown). Right inset: RPL scores localize known pseudo-knots in the LSU rRNA structure, such as the interaction between bases 1727-1812 (shown in red) and bases 1941 – 2038 (shown in blue).

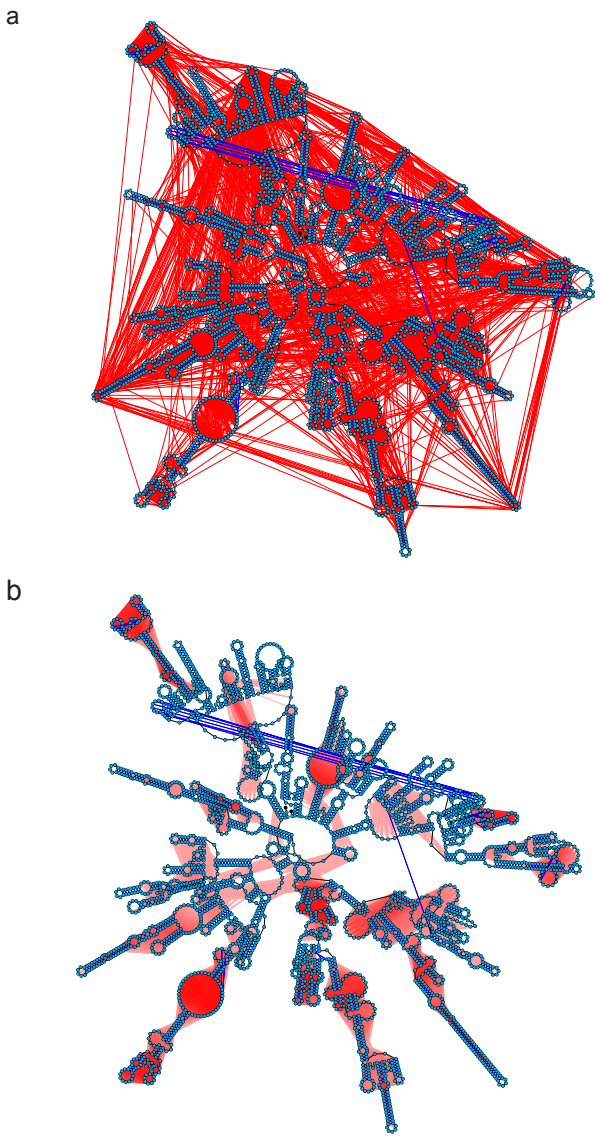


Figure 2.2. Smoothing of ligation junction data results in ligase-dependent signal around known stem-loop formations.

a.) The 10,000 most abundant ligation pairs for the LSU rRNA (red) overlaid onto the known secondary structure (blue). While signal across stem-loops is evident, there is considerable noise. b.) Top 25,000 interacting windows based on RPL scores, which are calculated from the frequencies of ligation between pairs of 21 nt windows (Methods), for the LSU rRNA in the (+) ligase sample (red), again overlaid onto the known secondary structure (blue). Lines are drawn between the central bases of two interacting 21 nt windows. For b.), the shading of the red lines is proportional to the ligation frequency.

Figure 3

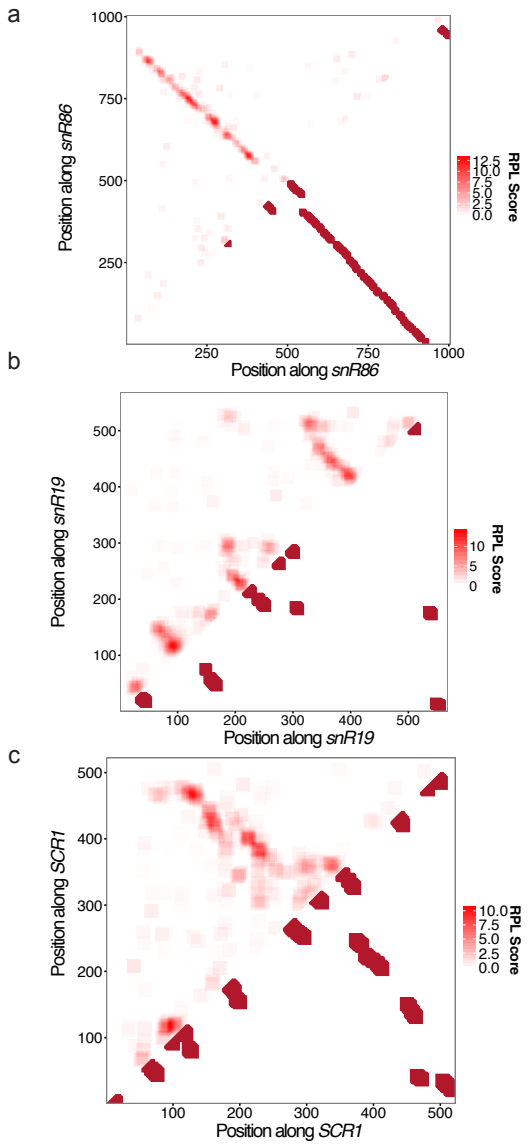


Figure 2.3. 2D RPL contact probability maps recapitulate known and predicted non-ribosomal RNA structures.

a.) Contact probability map for *snR86* mirrored against interacting windows containing paired bases, based on conserved secondary structure. b.) Contact probability map for *snR19* mirrored against interacting windows containing paired bases, based on conserved secondary structure. RPL signal indicating the formation of a stem-loop in bases 320-510 within the molecule is supported by MFE predictions, but not conservation. c.) Contact probability map for *SCR1* mirrored against interacting windows containing paired bases, based on the known structure of *SCR1*. For all analyses shown here, RPL scores were calculating using a window size of 21 nt.

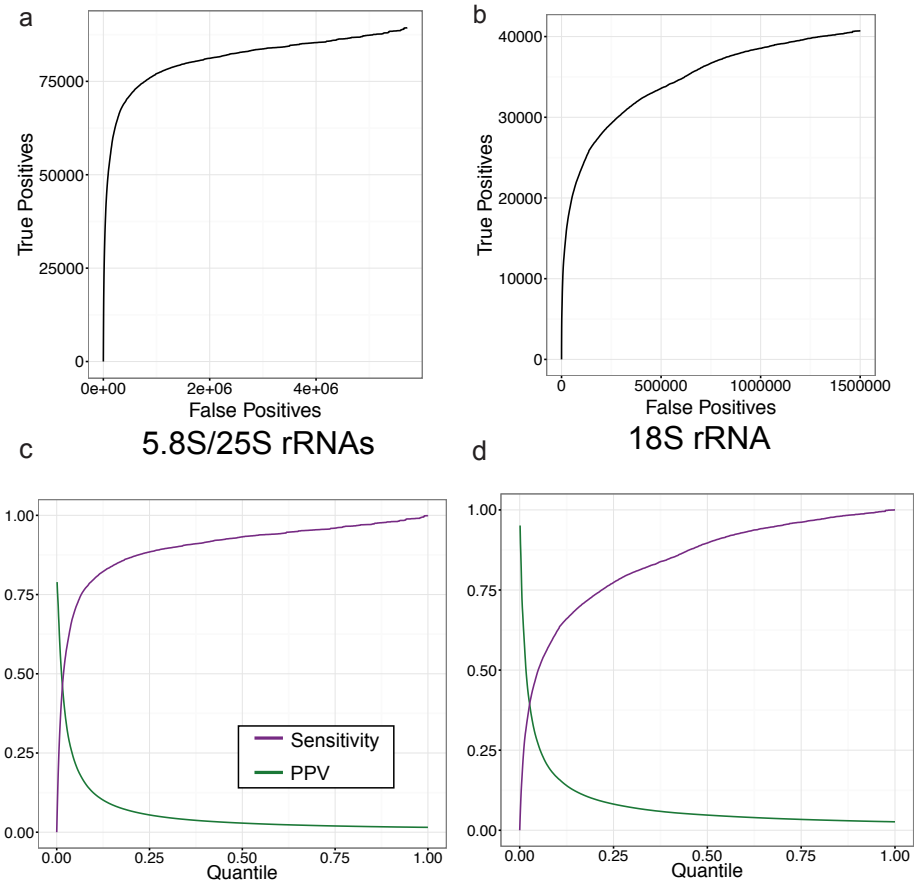


Figure 2.4. RPL scores demonstrate modest positive predictive value for pairs of interacting windows in RNA secondary structure.

a-b.) Plots of number of true positive interacting windows versus number of false positive interacting windows for the (a) 5.8SS/25S rRNAs and (b) 18S rRNA, at various quantile thresholds on RPL scores. This analysis shows that RPL scores have predictive value in classifying interacting regions containing at least one set of paired bases within RNA secondary structure. c-d.) Plots of the positive predictive value (green) and sensitivity (purple) of RPL-based classification of interacting regions, as a function of quantile threshold used for (c) 5.8S/25S and (d) 18S rRNAs. The quantile step size used for all analyses shown in this figure was 0.001.

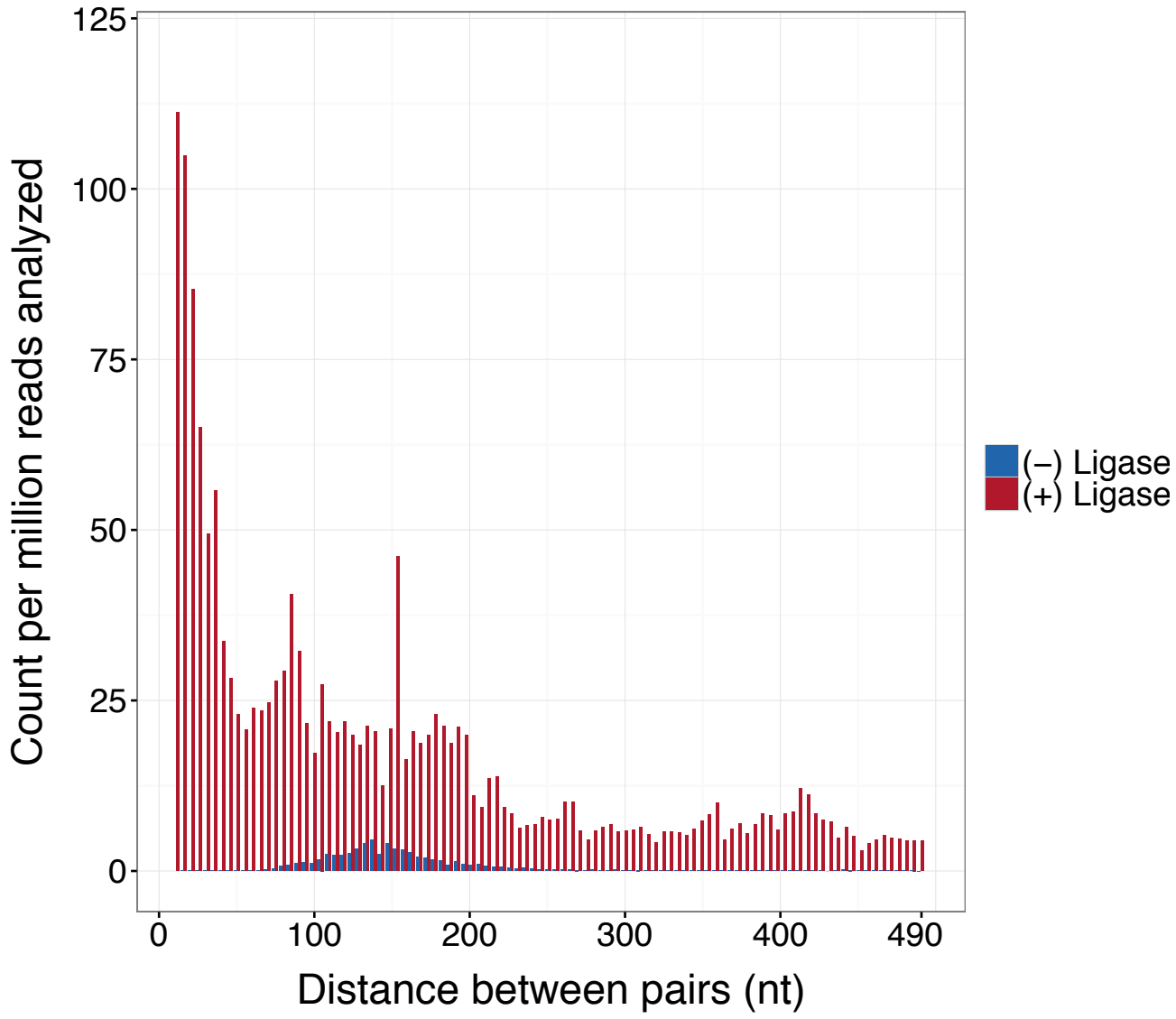


Figure 2.5. Samples treated with exogenous ligase are enriched for "gapped," or intramolecular chimeric, reads.

We observe enrichment for gapped, or intramolecular chimeric reads, over most gap sizes in our RPL sample compared to a control sample in which no T4 RNA Ligase was added. For gap sizes > 495 bases, we observed 627 reads per million analyzed in the (+) ligase sample, versus 25 reads per million analyzed in the (-) ligase sample. This ligase-dependent enrichment for long gap sizes suggests that the long-distance ligation products generated RPL are neither a result of gross mapping artifacts, nor the result of biological artifacts such as RT template switching.

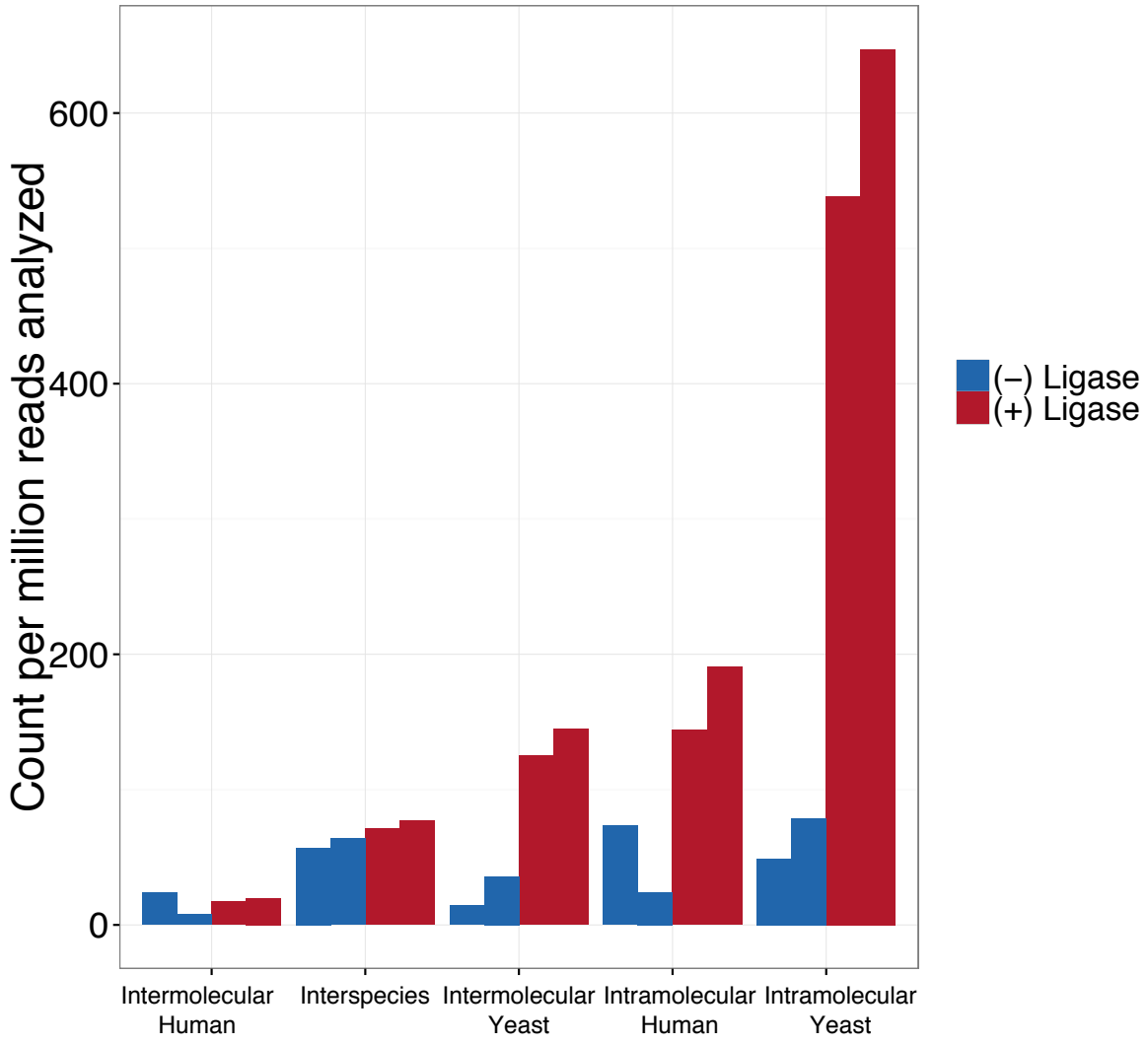


Figure 2.6. Mixing of RNA from two species during an RPL experiment to quantify the extent of non-specific product generation during the RPL protocol.

We carried out, in duplicate and with matched (-) ligase controls, the RPL protocols for yeast and human cells separately (1 colony picked for yeast; 5E5 GM12878 cells used for human) and then mixed the two slurries together prior to overnight proximity ligation by T4 RNA Ligase I. Comparing the (+) ligase and (-) ligase samples, we observe the strongest enrichments for intraspecies, intramolecular ligations.

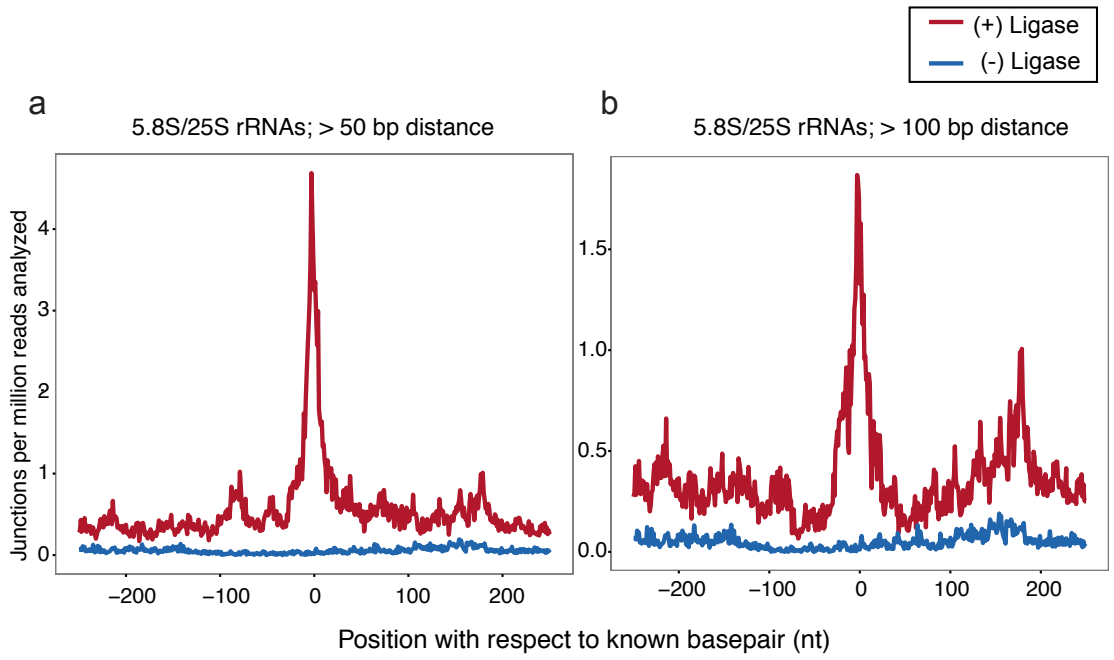


Figure 2.7. RPL signal recapitulates known long-range base-pairing interactions, and is dependent on exogenous ligase.

a.-b) The distribution of ligation junctions shows enrichment centered at known secondary structure base-pairs even when only considering long-range ligation junctions. Shown here are distributions for the 5.8S/25S rRNA, (a) excluding all ligation events ≤ 50 bp or (b) excluding all ligation events ≤ 100 bp. Results from a matched (-) ligase control are shown in blue.

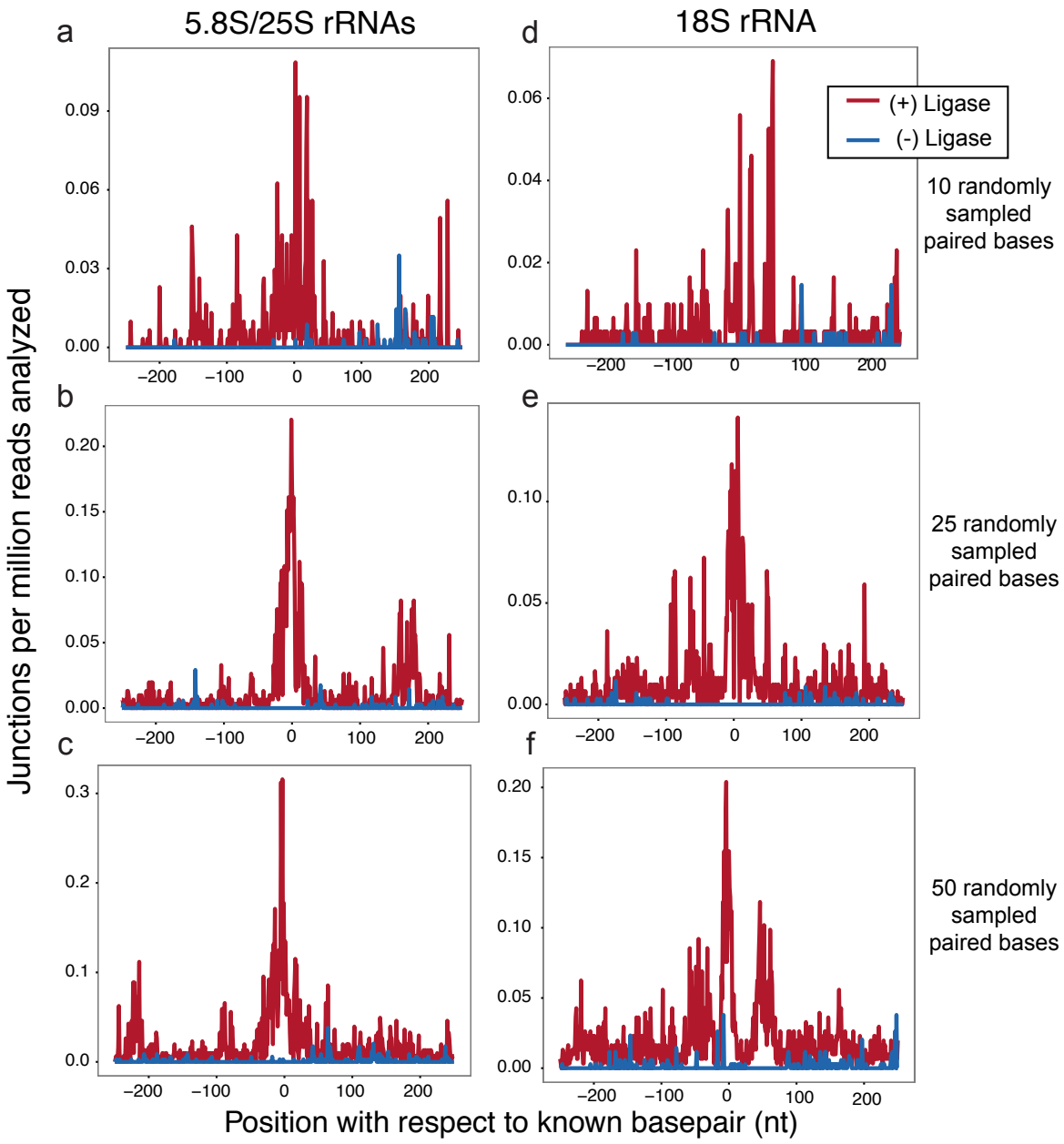


Figure 2.8. The raw ligation count data is noisy.

a.-c) We randomly sampled 10 (a), 25 (b), and 50 (c) paired bases from the 5.8S/25S rRNA and plotted the distribution of ligation junctions as a function distance to pairing partner. d-f.) Same as above, but for the 18S rRNA. We find that the enrichment evident when averaging over all base-pairs in the molecule (Figure 1b,c) is apparent but much nosier.

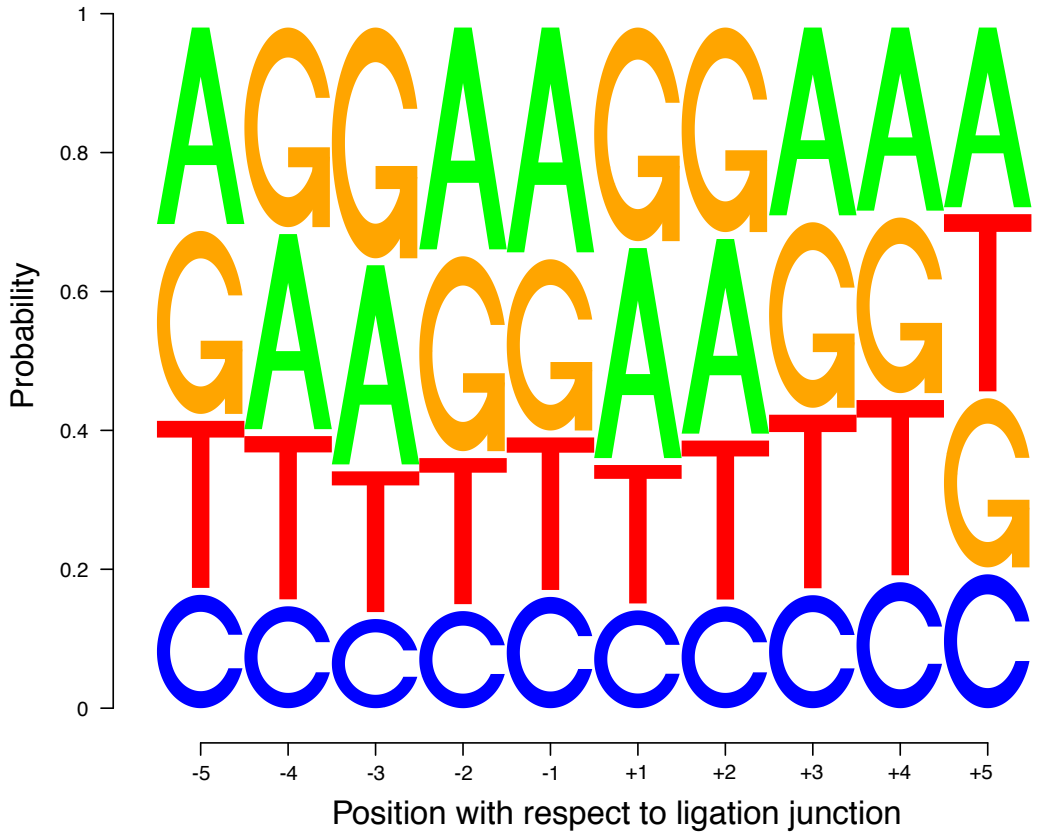


Figure 2.9. RPL ligation junctions demonstrate a slight sequence composition bias.

Shown here is a sequence-logo representation for the five bases upstream and downstream of observed ligation products. We observe a slight enrichment for A/T (~1.2 fold with respect to background frequency) immediately proximal to the ligation junction in our sequencing products, consistent with kinase and/or T4 RNA Ligase I bias during proximity ligation. We compute the background distribution by first calculating the individual nucleotide frequencies of all transcripts with at least one alignment, then calculating a weighted background frequency from these based on the number of reads aligning to each transcript.

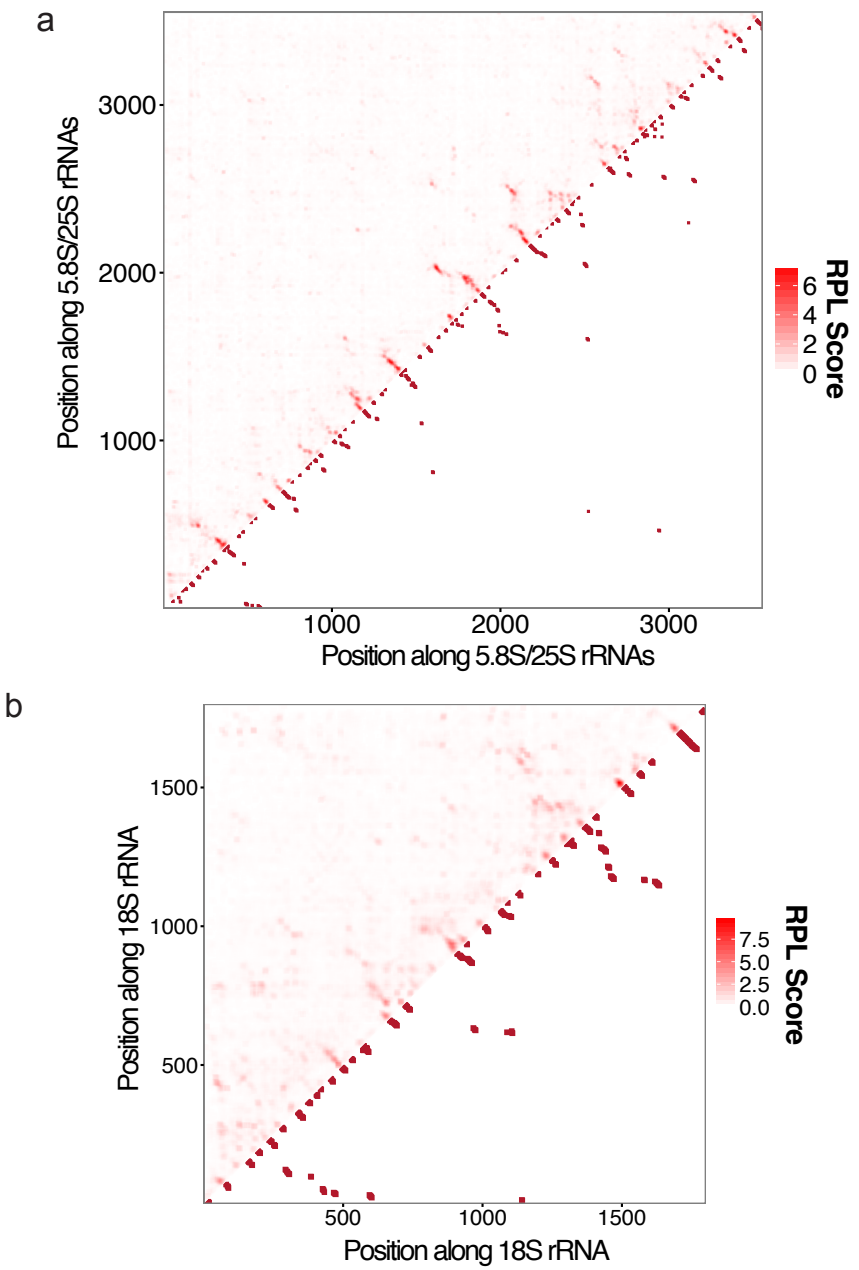


Figure 2.10. RPL contact probability maps broadly recapitulate the proximity implied by base-pairing relationships in structurally complex yeast ribosomal RNAs.

a.) RPL contact probability map for the 5.8S/25S rRNAs mirrored against all interacting 21 nt windows that contain paired bases in the known structures. b.) Same as above, but for the 18S rRNA. In both cases, high RPL scores broadly agree with the interacting windows in the known RNA secondary structures.

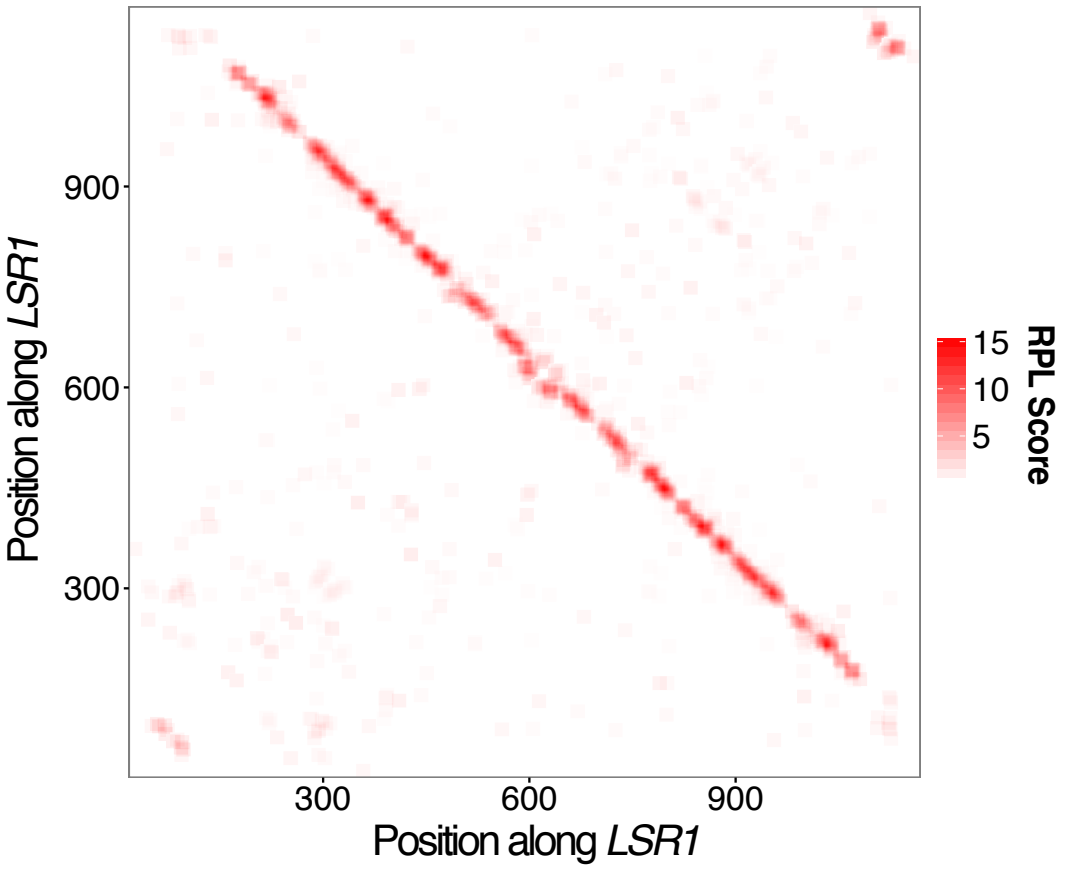
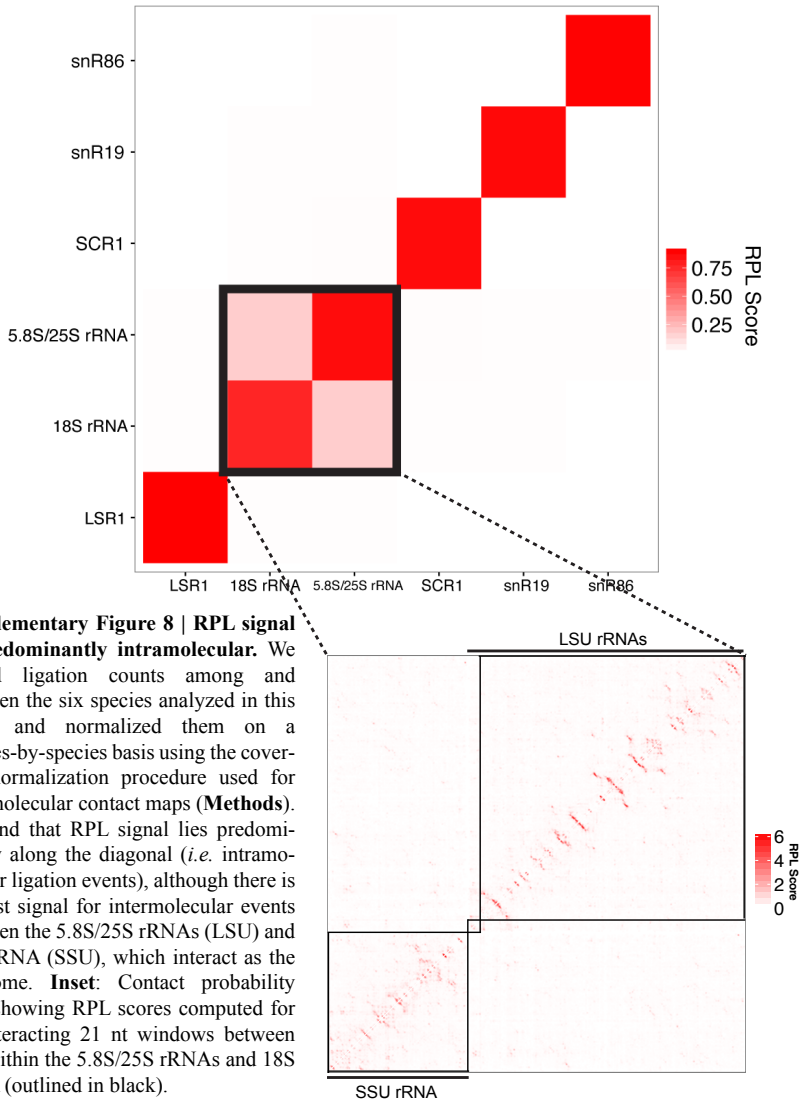


Figure 2.11. 2D RPL contact probability map for the S. cerevisiae U2 spliceosomal RNA homolog LSR1.

Anti-diagonal RPL scores imply the formation of a long stem in this molecule. This analysis was carried out using 21 nt window-based RPL scores.



Supplementary Figure 8 | RPL signal is predominantly intramolecular. We tallied ligation counts among and between the six species analyzed in this study and normalized them on a species-by-species basis using the coverage normalization procedure used for intramolecular contact maps (**Methods**). We find that RPL signal lies predominantly along the diagonal (*i.e.* intramolecular ligation events), although there is modest signal for intermolecular events between the 5.8S/25S rRNAs (LSU) and 18S rRNA (SSU), which interact as the ribosome. **Inset**: Contact probability map showing RPL scores computed for all interacting 21 nt windows between and within the 5.8S/25S rRNAs and 18S rRNA (outlined in black).

Figure 2.12. RPL signal is predominantly intramolecular.

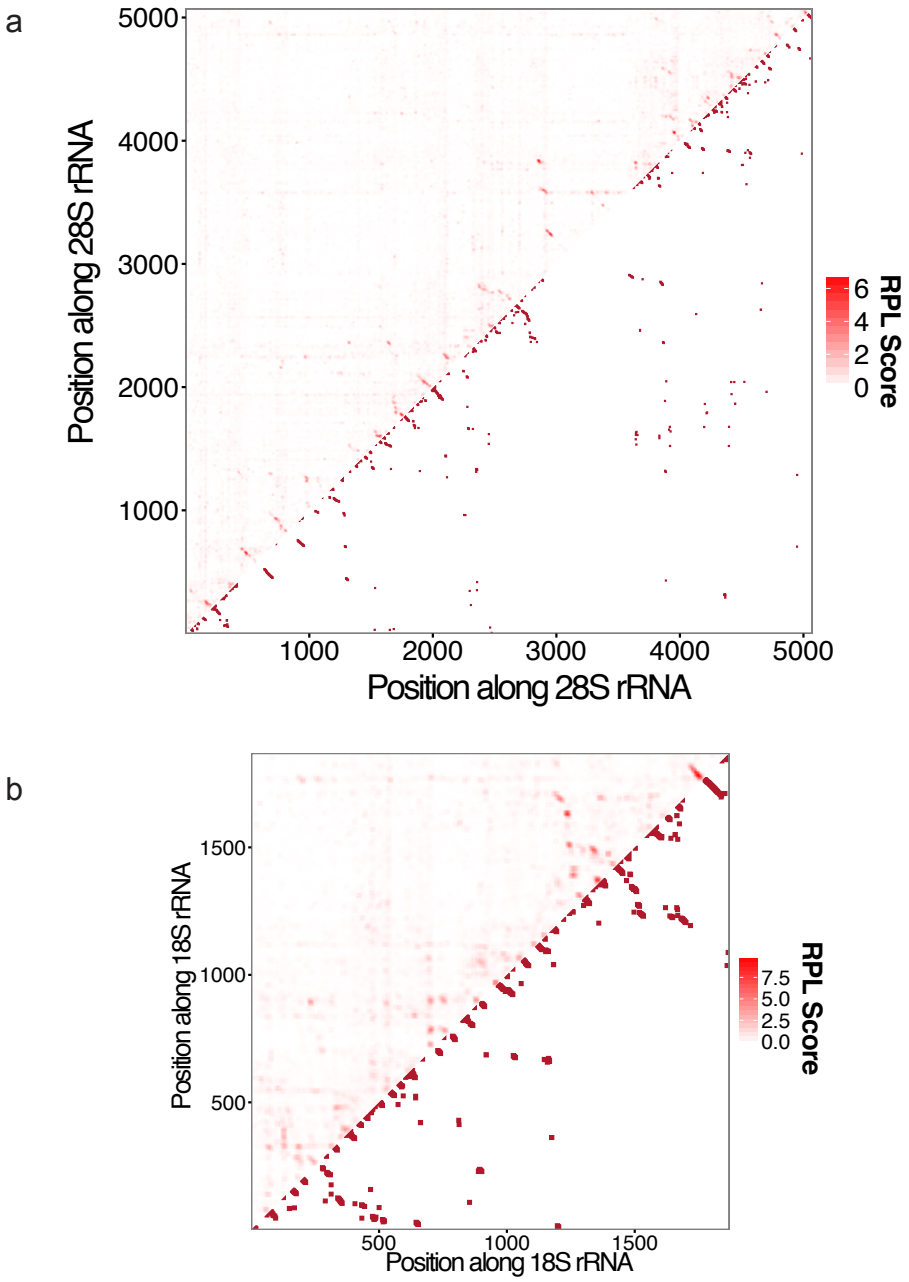


Figure 2.13. Extension of RPL to RNA secondary structures in mammalian cell culture.

a.) Contact probability map for the LSU 28S rRNA mirrored against interacting windows containing paired bases. b.) Contact probability map of the SSU 18S rRNA mirrored against interacting windows containing paired bases. Secondary structures for these molecules were derived from a cryo-EM structure of the human ribosome. All RPL scores shown here were calculated using 21 nt windows.

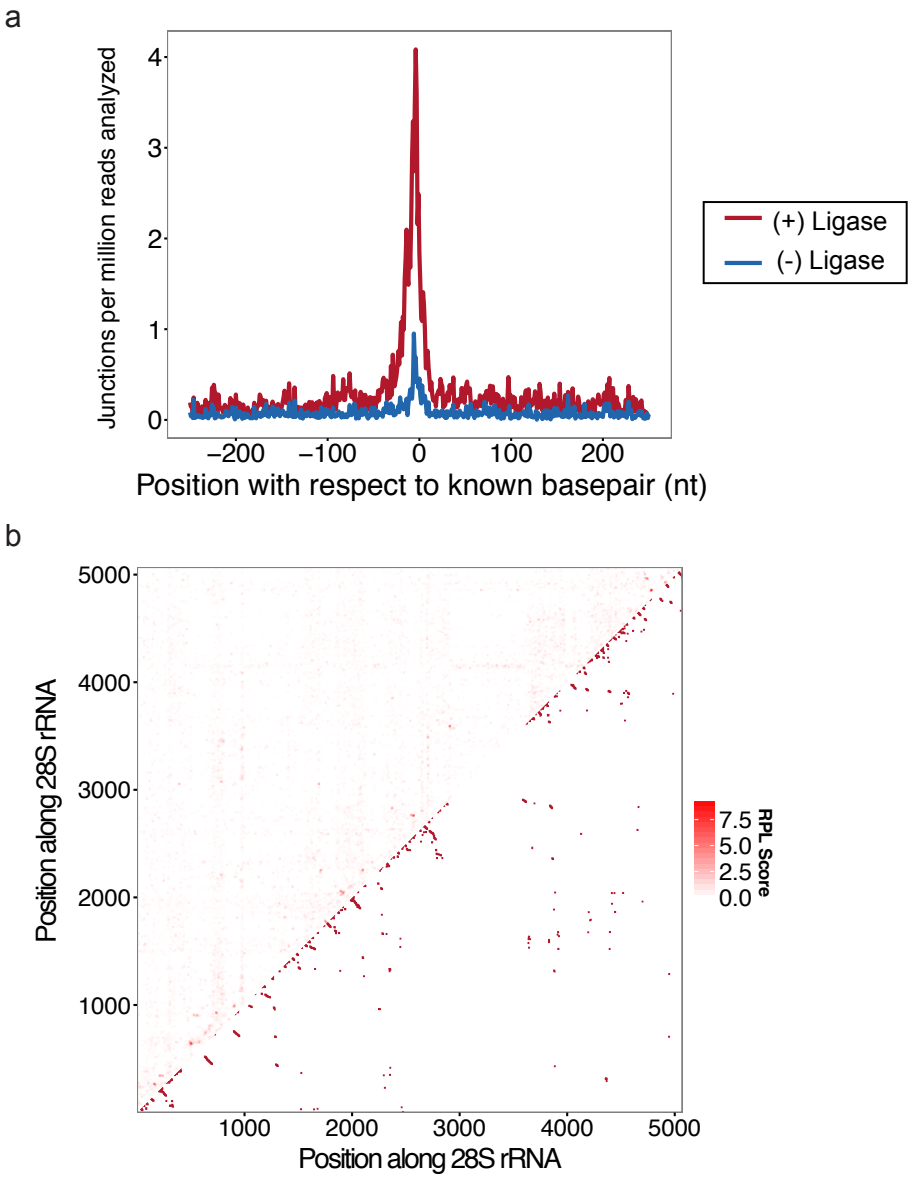


Figure 2.14. Mammalian RPL (-) RNase, (-) Ligase control demonstrates weak signal for structure-related ligation junctions.

a.) A distribution of ligation junctions centered at known base-pairing partners for the mammalian LSU rRNA displays a weak enrichment in signal at known pairing partners, in a (-) ligase sample also untreated by RNases. This suggests that endogenous ligases/RNases may be active to a small degree in lymphoblastoid cell lines. However, the signal is much stronger in the (+) RNase, (+) ligase sample. b.) A contact probability map illustrates the extent of noise in these potentially endogenous ligations, though certain highly scoring regions do appear consistent with known structures within the molecule (shown mirrored). RPL scores were calculated using 21 nt windows.

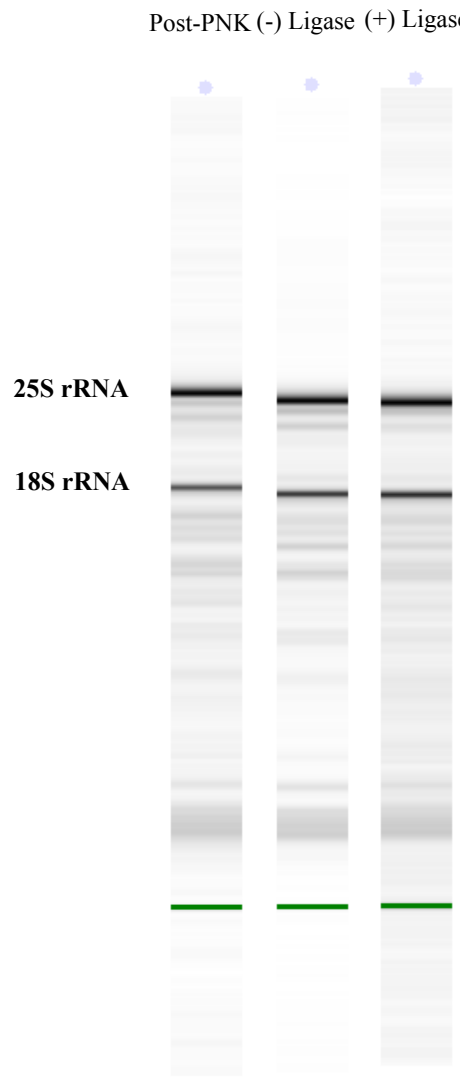


Figure 2.15. The Yeast RPL protocol demonstrates limited degradation of RNA products following PNK treatment.

Bioanalyzer gel representation of purified RNA at three conditions: 1.) Following PNK treatment; 2.) Following a negative control incubation at 16°C overnight in the absence of ligase; 3.) Following incubation at 16°C overnight in the presence of T4 RNA Ligase I. RNA Integrity Numbers (RIN) for the three samples were 7.0, 7.2, and 7.0, respectively.

Chapter 3. MAPPING 3D GENOME ARCHITECTURE WITH *IN SITU* DNASE-HI-C

Note: Chapter 3 was published in the November 2016 issue of *Nature Protocols* as:

Ramani V., Cusanovich D., Hause R.J., Ma W., Qiu R., Deng X., Blau C.A., Disteche C.M., Noble W.S., Shendure J, Duan Z. "Mapping 3D Genome Architecture with *in situ* DNase Hi-C." *Nature Protocols* (2016).

3.1 ABSTRACT

With the advent of massively parallel sequencing, considerable work has gone into adapting chromosome conformation capture (3C) techniques to study chromosomal architecture at genome-scale. We recently demonstrated that the inactive murine X chromosome adopts a bipartite structure using a novel 3C protocol, termed *in situ* DNase Hi-C. Like traditional Hi-C protocols, during *in situ* DNase Hi-C chromatin is chemically crosslinked, digested, end-repaired, and proximity ligated with a biotinylated bridge adaptor. The resulting ligation products are optionally sheared, affinity-purified via streptavidin bead immobilization, and subjected to traditional next-generation library preparation for Illumina paired-end sequencing. Importantly, *in situ* DNase Hi-C obviates the dependence on a restriction enzyme to digest chromatin, instead relying on the endonuclease DNase I. Libraries generated by *in situ* DNase Hi-C have a higher effective resolution than traditional Hi-C libraries, making them valuable in cases where high sequencing depth is allowed for, or when hybrid capture technologies are expected to be used. The protocol described here, which involves approximately four days of bench work, is optimized for the study of mammalian cells but can be broadly applicable to any cell or tissue of interest given experimental parameter optimization.

3.2 INTRODUCTION

The manner in which an incredibly long DNA polymer topologically organizes itself within a cell or nucleus is crucially linked to higher-order cellular function (Cremer and Cremer, 2001; Fraser and Bickmore, 2007). This form-function relationship, first realized through early light microscopic studies of higher-order structures like mitotic chromosomes (Rieder and Khodjakov, 2003), the inactive X Barr body (BARR and BERTRAM, 1949), and polytene chromosomes (Hochstrasser and Sedat, 1987), has only become clearer in the face of advancing technologies. Techniques such as fluorescence *in situ* hybridization (FISH) of chromatin (Manuelidis, 1985; Pinkel et al., 1986; Schardin et al., 1985), have provided clear evidence that chromosomes occupy compartments within the nucleus, ultimately leading to the development of correlative models associating biological function (i.e. transcription, splicing, silencing) with particular nuclear locales (Lawrence et al., 1989; Zirbel et al., 1993).

With the advent of genome-scale technologies, high-throughput assays have been developed to characterize nuclear architecture at both increasing scale and resolution. Techniques like DNA adenine methyltransferase identication (DamID) (van Steensel and Henikoff, 2000), typically used to map protein-DNA interactions (Orian et al., 2003), have been modified to map genome-wide associations between primary sequence and the nuclear lamina (Guelen et al., 2008) (i.e.lamina associated domains, or LADs), where silenced domains typically reside. Methods involving the "proximity ligation" of chromatin, now termed chromosome conformation capture (3C) (Dekker et al., 2002), have also gained popularity. 3C techniques represent matured versions of early methods that used T4 DNA ligase to quantify the physical proximity of DNA sequences brought together by proteins (Cullen et al., 1993; Mukherjee et al., 1988), and all share

a common paradigm: fixation of chromatin within the nucleus via formaldehyde, endonucleolytic digestion of chromatin (normally via restriction enzyme digestion), and re-ligation of physically proximal fragments. The first 3C variants (e.g. 4C, 5C) used specific primers or sets of primers to determine contact frequencies between predefined sites in the genome (de Wit et al., 2006; Dostie et al., 2006). Later, massively-parallel versions of 3C, generally termed "Hi-C", were developed (Duan et al., 2010; Kalhor et al., 2012; Lieberman-Aiden et al., 2009), which leverage paired-end sequencing to generate contact frequency estimates between sequence windows across entire genomes.

Since the advent of 3C techniques, much work has gone into characterizing 3D genome architecture in a wide-variety of biological contexts (Dixon et al., 2012; Le et al., 2013; Mizuguchi et al., 2014; Sexton et al., 2012; Zhang et al., 2012), including mitotic cell division (Naumova et al., 2013), the life cycle of a parasite (Ay et al., 2014), and in mammalian dosage compensation (Deng et al., 2015; Giorgetti et al., 2016).

The vast amount of available Hi-C data has also enabled the discovery of novel "units" of genome topology, including topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012) and chromosomal interacting domains (CIDs) (Hsieh et al., 2015; Le et al., 2013), genomic domains that predominantly self-associate in three-dimensional space. Although the ultimate significance of these domains remains unknown, strong correlations between one-dimensional epigenomic features (e.g., histone marks, DNA methylation, transcription factor binding) and sequence both within and at the borders of these domains suggest that they may play a gene regulatory role.

Although current Hi-C techniques generally allow us to visualize genome-scale chromosome architecture at the resolution of 100 kb to 1 Mb, methodological resolution

limitations imposed by incomplete sequencing depth and genome-wide restriction site density have typically precluded identification of topological units at smaller scales, in which local interactions may play crucial gene regulatory roles. The need for fine-scale resolution of these higher-order interactions has only become clearer in the wake of the immense amount of high-resolution, one-dimensional epigenomic data generated by consortia such as ENCODE (ENCODE Project Consortium, 2012) and Roadmap Epigenomics (Sabo et al., 2015).

Given the availability of such data, one crucial interest of the gene regulatory field is the potential link between complex gene regulatory programs and dynamic long-range "looping" interactions between distal regulatory elements, features at a scale even smaller than that of TADs and LADs (Levine et al., 2014). Since the earliest realizations that long-range interactions are effectors of gene expression (Griffith et al., 1986; Müller et al., 1989), the gene regulatory field has worked towards completely cataloguing functional DNA looping interactions. In the realm of proximity ligation protocol development, considerable work has gone towards improving the resolution of the Hi-C protocol to the scale of kilobases, where specific regulatory contacts (i.e. enhancer-promoter interactions, CCCTC-binding factor (CTCF)-mediated loops) might be identified.

The protocol presented here complements existing high-resolution Hi-C approaches by providing another flexible, convenient, and scalable methodology that eschews the use of restriction enzymes. Our approach therefore avoids the theoretical limit in resolution of the standard Hi-C protocol imposed by the occurrence of restriction sites in the genome, given enough sequencing depth and library complexity.

3.3 Moving Towards Fine-scale Resolution of 3D Contacts

Core methodological improvements to the Hi-C protocol to improve resolution have broadly spanned three primary areas: deeper sequencing, simplified library preparation protocols, and the use of hybridization capture to enrich for sets of desired loci in a massively parallel fashion. We recently developed a method that unites many of these improvements with additional empirical changes to further increase the effective resolution of Hi-C libraries (Ma et al., 2015). Our method, termed DNase Hi-C, eliminates the reliance on restriction enzymes associated with Hi-C by digesting fixed chromatin with the endonuclease DNase I in the presence of divalent manganese. We demonstrated that DNase Hi-C libraries mitigate many of the biases associated with traditional Hi-C, reducing the effective distance between fragments imposed by 4- and 6-cutter restriction enzymes while improving robustness with respect to G-C content, mappability, and genomic coverage. Furthermore, we also showed that DNase Hi-C may be paired with commercially available hybridization capture kits to visualize long intergenic noncoding RNA (lincRNA) promoters at a previously unprecedented scale of 1 kb without the gross sequencing depth requirements typically associated with high-resolution contact maps.

Motivated by the observation that the vast majority of proximity ligations occur in insoluble chromatin (Gavrilov et al., 2013), we recently published an improved version of our previously published DNase Hi-C termed *in situ* DNase Hi-C (Deng et al., 2015). We applied this simplified and robust Hi-C protocol to study the inactive X chromosome in primary mouse brain tissue and an immortalized mouse embryonic kidney cell line, demonstrating for the first time that the murine inactive X chromosome adopts a bipartite conformation. *In situ* DNase Hi-C represents a considerable improvement over its parent protocol, requiring considerably less hands-on time and lower cellular input requirements.

3.4 OVERVIEW OF *IN SITU* DNASE HI-C

A schematic of the *in situ* DNase Hi-C protocol is illustrated in **Figure 3.1**. Anywhere from $5x10^5$ to $1x10^7$ cells are fixed in formaldehyde to reversibly crosslink in vivo protein-DNA interactions. Fixed cells are then lysed to liberate nuclei, which are treated with the endonuclease DNase I to digest chromatin. Digested chromatin ends are end-repaired and dA-tailed, facilitating the ligation of an exogenous, dT-tailed "bridge" adaptor containing a single biotinylated thymidine, half BamHI restriction site, and 4-base overhang. After clearing out excess adaptors, the free ends of chromatin (now capped with bridge adaptors) are phosphorylated with T4 Polynucleotide Kinase (T4 PNK) and proximity ligated *in situ* with T4 DNA Ligase I. During all of these steps, nuclei are immobilized against carboxylated paramagnetic beads (commonly referred to as Solid Phase Reversible Immobilization (SPRI) beads), both providing a scaffold to prevent loss of nuclei during enzymatic reactions and allowing for the simple removal of free DNA and excess bridge adaptor, which adversely affect downstream library preparation.

Following proximity ligation, nuclei are lysed and crosslinks are reversed with Proteinase K treatment. DNA is then isolated with an isopropanol precipitation, after which fragments are optionally sheared. Ligated DNA fragments harboring the biotinylated bridge adapter are then affinity-purified using streptavidin beads, end-repaired, dA-tailed, and ligated to standard Illumina sequencing adaptors. Finally, ligation products are PCR amplified to generate sequencing libraries. Prior to sequencing, libraries may be treated with a simple BamHI digestion to assess the efficiency of proximity ligation.

3.5 TRADITIONAL HI-C VS. *IN SITU* DNASE HI-C

In situ DNase Hi-C can be used in any situation where traditional Hi-C would be used. Thanks to a reliance on the endonuclease DNase I, *in situ* DNase Hi-C eliminates the characteristic restriction enzyme biases that limit resolution in traditional Hi-C libraries while lowering the input cell requirements for library construction. Unlike other Hi-C protocols, *in situ* DNase Hi-C is the only protocol, to our knowledge, to use paramagnetic carboxylated beads as a tool to immobilize nuclei during *in situ* enzymatic treatments. This immobilization step not only reduces nuclei loss during the protocol, aiding low-input experiments, but also facilitates the removal of contaminating adaptors and free DNA. Finally, like traditional *in situ* Hi-C, *in situ* DNase Hi-C requires considerably less hands-on time for library prep, and more efficiently generates cis (*i.e.* intrachromosomal) ligation products compared to trans (*i.e.* interchromosomal) ligation products.

Considering the high sequencing depth required to generate high-resolution genome-wide contact maps, we note that at low resolution, maps generated using *in situ* DNase Hi-C are practically very similar to those generated using other Hi-C protocols (except in cases where loci may have particularly low restriction site density). In cases where high-resolution (i.e. 1 kb resolution) maps are desired, however, we strongly believe that the relatively unbiased ligation junctions generated through DNase Hi-C present an important alternative to existing methods. This point is particularly relevant when hybrid capture techniques may be applied, as high-resolution, RE independent maps can be generated for a fraction of the cost of genome-scale library sequencing.

Still, we acknowledge that in many cases cost may preclude the use of deep sequencing or hybrid capture. In cases such as these, we suggest more cost-effective solutions using more focused

techniques (e.g. 3C, 4C, 5C), albeit at the price of only interrogating interactions among a set number of loci.

In situ DNase Hi-C is broadly applicable to any situation where high-resolution chromatin conformation data or 3D maps are required. We have successfully carried out *in situ* DNase Hi-C in several immortalized cell lines and primary tissues, generating libraries for the human cell lines K562 and GM12878, as well as mouse embryonic kidney cells and homogenized mouse brain tissue.

3.5.1 *Limitations of the protocol*

In situ DNase Hi-C is subject to the same limitations as any bulk Hi-C protocol. First, the protocol requires $5x10^5$ to $1x10^7$ cells to generate sequenceable libraries. Thus, in cases where input might be particularly limited, or where small populations of cells sorted by fluorescence activated cell sorting (FACS), *in situ*DNase Hi-C may not be appropriate. Second, it is also important to note that while the DNase enzyme is nonspecific when compared to restriction enzymes, it has been shown to exhibit mild sequence bias at cleavage sites (He et al., 2014). This must be considered when applying *in situ* DNase Hi-C to organisms with radical nucleotide content (i.e. low GC content), and when considering the inherent biases within *in situ* DNase Hi-C maps (as would be done with any Hi-C contact map).

3.6 Experimental Design Considerations

The *in situ* DNase Hi-C protocol described here is relatively straightforward, and can be completed over four days, allotting 3 – 6 hours of bench work per day. Still there are several experimental design parameters that should be considered before applying *in situ* DNase Hi-C to a new cell type of interest. These considerations primarily concern maintaining intact nuclei during

the various *in situ* enzymatic treatments in the protocol. The *in situ* DNase Hi-C protocol also allows for sequencing-free quality control of libraries, thanks to the integration of half BamH1 sites in the bridge adapter. As discussed below, this allows for easy quantification of the efficiency of proximity ligation in the final *in situ* DNase Hi-C library.

Although the protocol presented here is robust to many different cell types, different immortalized cell lines may require optimization of formaldehyde crosslinking, DNase I digestion and SDS concentration during digestion. Below we detail our process for optimizing these various parameters:

3.6.1 *Formaldehyde concentration*

As with other 3C methods and ChIP-seq protocols, formaldehyde fixation is an important component of the *in situ* DNase Hi-C protocol, promoting proximity ligation of long-range genomic contacts while maintaining the integrity of nuclei during *in situ* enzymatic steps. Incomplete crosslinking can lead to an underrepresentation of proximity ligation products in Hi-C libraries, and excessive breakage of nuclei can lead to considerable decreases in the ultimate molecular complexity of libraries, and at worst can increase the degree of "spurious" ligations formed. The guidelines for formaldehyde fixation of cells for *in situ*DNase Hi-C are the same as those for the other 3C-based techniques and ChIP-seq methods. In general, for single-cell suspension cultures (e.g. GM12878 and K562 cells) and monolayer adherent cells (e.g. Hela cells) a standard condition of cross-linking, such as 1% formaldehyde for 10 min at room temperature (RT, 25°C), can be employed. For other cell cultures (e.g. mouse and human embryonic stem cells (ESCs)) and primary tissue cells (e.g. mouse brain cells and plant leaves), for which single-cell suspensions are difficult to obtain, increased formaldehyde concentrations or longer fixation times may be required to ensure efficient crosslinking. For example, both human and mouse ESCs often

aggregate to form large clumps in culture. As such, higher concentrations of formaldehyde are generally used in these situations[48,54].

3.6.2 *Cell lysis and DNase I digestion*

After crosslinking chromatin interactions with formaldehyde, one must render fixed chromatin accessible to enable chromatin fragmentation and other downstream enzymatic reactions. As with restriction digestion-based 3C methods, cell lysis in *in situ* DNase Hi-C is achieved primarily through SDS treatment. To ensure that nuclei remain intact throughout the multiple enzymatic reactions through the end of nuclear ligation (step 48), the *in situ* DNase Hi-C protocol employs a relative mild condition (0.3–0.5% SDS treatment for 45 min. at 37°C). During this step, it is crucial to avoid overly lysing nuclei. Expected results are shown in **Figure 3.2a**. We also note that overly lysed nuclei become apparent during any of the many centrifugation steps in the *in situ* DNase Hi-C protocol, as no pellet forms. Nuclei should remain intact through proximity ligation, as shown in **Figure 3.2b**.

We stress that the required SDS concentration for cell lysis and the amount of DNase I used during the DNase I digestion step can vary depending on the cell type being studied, and the number of nuclei being processed. We recommend carrying out a DNase I and SDS optimization experiment using varying units of DNase I and varying concentrations of SDS when attempting the protocol on new cell types, and examining the DNase I fragmentation pattern following digestion. An example fragmentation pattern is shown in **Figure 3.3a**.

3.6.3 *The role of paramagnetic carboxylated beads*

Paramagnetic carboxylated beads (i.e. AMPure XP beads) have been used in both our standard and *in situ*DNase Hi-C protocols. As demonstrated in **Figure 3.2**, these beads appear to bind to

intact nuclei and serve as carriers to pellet the nuclei by low-speed centrifugation. Here, we employ these beads to efficiently remove DNase I and low molecular weight DNA that might escape the nucleus following chromatin digestion, as well as free unligated internal bridge adaptor following bridge adaptor ligation. Furthermore, the beads also aid with visualization of the nuclei pellet throughout the protocol when starting the protocol with fewer than a million cells.

3.6.4 *Nuclei treatment*

It is crucial that the fixed nuclei remain intact over the course of the DNase Hi-C protocol. To this end, pipetting should be carried out gently to minimize shear forces that may burst nuclei.

3.6.5 *BamH1 Digestion Control*

A BamH1 digestion test on the final PCR-amplified library can be used to quantify ligation efficiency of the reaction. Lack of a library "shift" (properly digested products shown in **Figure 3.3b**) suggests inefficiency in the formation of proximity ligation products, and can be indicative of suboptimal fixation conditions or defective reagents.

3.7 PROCEDURE

Steps 1-2: Adaptor Annealing (Timing: 1h + overnight incubation)

1) Set up the following reactions:

Blunt Bridge Adaptor (40 μM final)

Component	**Amount (μL)**	**Final Concentration**
100 μM Bridge Adaptor 5'	80	40 μM
100 μM Blunt Bridge Adaptor 3'	80	40 μM
10X NEBuffer 2	20	1X
ddH_2O	20	
Total Volume	200	

Biotinylated Bridge Adaptor (40 μM final)

Component	**Amount (μL)**	**Final Concentration**

100 µM Biotinylated Bridge Adaptor 5'	80	40 µM
100 µM Bridge Adaptor 3'T	80	40 µM
10X NEBuffer 2	20	1X
ddH_2O	20	
Total Volume	200	

Y-Adaptor (25 µM final)

Component	**Amount (µL)**	**Final Concentration**
100 µM SeqAdapt_F	50	25 µM
100 µM SeqAdapt_R	50	25 µM
10X NEBuffer 2	20	1X
ddH_2O	80	
Total Volume	200	

2) Anneal mixtures by heating to 98°C for 6 minutes, then allow the tubes to naturally cool to RT overnight.
PAUSE POINT Annealed adapters can be kept at -20°C indefinitely.

Step 3: Cross-linking of cells (Timing: 2h)
[TROUBLESHOOTING 3]

3) Cells should be grown in appropriate culture medium. 2-5 x 10^6 cells are sufficient for making one DNase Hi-C library. However, we suggest growing, crosslinking, and aliquoting many cells (i.e. 1-5 x 10^7 cells) to provide replicates if necessary. Below are protocols for handling adherent monolayer cells (option A) or suspension cells (option B):
   a) ***Adherent monolayer cells:***
      i) Aspirate out media and add 10 ml of serum-free media per 10 cm plate.
      ii) Crosslink the cells by adding 280 µl of 37% formaldehyde to obtain 1% final concentration. Mix gently, immediately after addition of formaldehyde.
      iii) Incubate cells at RT for exactly 10 min, gently rocking the plates every 2 min.
      iv) Quench reaction by adding 0.5 ml of 2.5 M glycine and mixing well.
      v) Incubate for 5 min at RT, then on ice for 15 min to stop cross-linking completely.
      vi) Wash cells once with cold 1X PBS.
      vii) Treat the cells with 3-5 ml per dish 0.25% trypsin at 37°C for 5 min.
      viii) Add 5 ml fresh medium with serum.
      ix) Scrape the cells from the plates with a cell scraper and transfer to a 50 ml tube (combine all the cells from all the dishes to one tube).
      x) Centrifuge the cross-linked cells at 800x***g*** for 10 min.
      xi) Discard the supernatant by aspiration and wash the cross-linked cells with 1 x PBS once.
      xii) Aliquot the cells into 1.5 ml microtubes (2.5 million cells per tube).

      PAUSE POINT: Cells can be snap-frozen in liquid nitrogen and stored for at least one year at -80°C, or one can continue with cell lysis.

b) Suspension cells:

i) Gently pellet the cells by spinning at 300x***g*** for 10 min at RT.
ii) Discard the supernatant.
iii) Resuspend the pellet in 10 ml of fresh culture medium without serum. Break cell clumps by pipetting up and down.
iv) Crosslink the cells by adding 280 µl of 37% formaldehyde (1% final concentration). Mix quickly by inverting the tube several times.
v) Incubate at RT for exactly 10 min. Gently invert the tube every 2 min.
vi) Add 0.5 ml of 2.5 M glycine to quench the cross-linking reaction, mix well.
vii) Incubate for 5 min at RT, then on ice for 15 min to stop cross-linking completely.
viii) Centrifuge the cross-linked cells at 800x***g*** for 10 min at 4°C.
ix) Discard the supernatant by aspiration and wash the cross-linked cells with 1X PBS once.
x) Split the cross-linked cell suspension into aliquots of 2.5 x 10^6 cells (in 1.5 ml microtubes).
xi) Centrifuge the cross-linked cells at 800x***g*** for 10 min at RT.
xii) Discard the supernatant by aspiration.

PAUSE POINT: Cells can be snap-frozen in liquid nitrogen and stored for up to 1.5 years at -80°C, or one can continue with cell lysis.

Steps 4 – 20: Cell lysis and chromatin digestion with DNase I (Timing: 1.5 h)

4) Resuspend one cross-linked cell aliquot (0.5-2.5 x 10^6 cells) in 0.4 ml of ice-cold cell lysis buffer containing protease inhibitor.
CRITICAL STEP: Add 1 tablet protease inhibitor to 10 mL of ice-cold lysis buffer immediately prior to lysis. We recommend using lysis buffer with freshly added protease inhibitor for all experiments.
5) Incubate on ice for 10 min.
6) Centrifuge for 60 seconds at 2,500x***g*** at RT.

[TROUBLESHOOTING 7]

7) Discard the supernatant and resuspend the pellet in 100 µl of 0.5X DNase I digestion buffer containing 0.2% SDS.
CRITICAL STEP: For larger cell inputs (*i.e.* 3-5 x 10^6), we recommend using 200 µl 0.5X DNase I digestion buffer instead.
8) Incubate at 37°C for 30 min.
9) Add 100 µl of 0.5X DNase I digestion buffer containing 2% Triton X-100 and 4 µl RNase A, mix well.
10) Incubate at 37°C for 10 min.
11) Add 1.5 units of DNase I and mix well.
12) Incubate at RT for 4 min.
13) Add 40 µl of 6X Stop Solution, mix well.

[**TROUBLESHOOTING 14**]

14) (Optional) To determine the efficacy of DNase I digestion, take 20 µl of lysed cells from the previous step and add to a new tube. Add 70 µl 1X TE lysis buffer and 10 µl Proteinase K (20 mg/ml). Incubate for 30 minutes at 65°C. Purify DNA using a Qiaquick PCR purification kit.

Check the quality of chromatin digestion by running the samples out on a 6% TBE-PAGE gel. The sample is properly digested if one sees a large smear of DNA fragments between ~100 bp and 1 kb (see **Figure 3a**). We recommend characterizing DNase I digestion efficiency when performing the protocol on a new cell type. In the event of over-digestion or under-digestion of chromatin, we recommend optimizing the concentration of SDS in the digestion reaction, amount of DNase I used, or digestion time.

15) Centrifuge for 60 seconds at 2,500x***g*** at RT.
16) Discard the supernatant and resuspend the pellet in 150 µl water.
17) Add 300 µl AMPure XP beads; mix thoroughly by pipetting up and down.
18) Incubate at RT for 5 min and place the tube in a DynaMag magnet for 2 min.
19) Discard the supernatant and wash the beads twice with 1 ml of freshly prepared 80% (vol/vol) ethanol. Briefly spin down the beads and remove the residual ethanol.
20) Resuspend the beads in 169 µl of water, and proceed immediately to the next step.

Steps 21 – 30: Chromatin End Repair and dA-tailing (Timing: 2.5 h)

21) Prepare the End-Repair reaction as follows:

Reagents (add in this order)	**Volume (µL)**	**Final Concentration**
Nuclei w/ beads	169	
10X T4 ligase buffer w/ ATP	20	1X
10 mM dNTPs	5	0.25 mM
T4 DNA Polymerase (3U/ µl)	3	0.045 U / µL
Klenow (10U/ µl)	3	0.15U / µL
Total Volume	200	

22) Incubate at RT for 1 h.
23) Add 5 µl 10% SDS to stop the reaction.
24) Centrifuge for 60 seconds at 2,500x***g*** at RT.
25) Aspirate and resuspend the pellet in 135 µl water.
26) Prepare the dA-Tailing reaction as follows:

Reagents (add in this order)	**Volume (µL)**	**Final Concentration**
Nuclei w/ beads	135	
10X NEBuffer 2	20	1X
10 mM dATP	10	0.5 mM
10% Triton X-100	20	1%
Klenow (exo$^-$) (5U/ µl)	15	0.375U / µL
Total Volume	200	

27) Incubate the resulting mixture at 37^{o}C for 1 hr.
28) Add 5 µl 10% SDS to stop reaction.
29) Centrifuge for 60 seconds at 2,500x***g*** at RT.
30) Aspirate and resuspend the pellet in 30 µl nuclease-free water.

Steps 31 – 44: Ligation of Biotin-labeled Bridge adaptors (Timing: Overnight, followed by 0.5 h)

31) Prepare the adaptor ligation reaction as follows:

Reagents (add in this order)	**Volume (µl)**	**Final Concentration**
Nuclei w/ beads	30	
Annealed Bridge Adaptor w/ Biotin (40 µM)	20	8 µM
Annealed Blunt Adaptor w/o Biotin (40 µM)	20	8 µM
10X T4 ligase buffer w/ ATP	10	1X
PEG-4000 (50%)	10	5%
10% Triton X-100	5	0.5%
T4 DNA Ligase (5 U/ µl)	5	0.25U / µL
Total Volume	100	

32) Incubate at 16^{o}C overnight.

PAUSE POINT: Reaction should be allowed to incubate overnight.

33) (Optional) To examine the efficacy of the above end-repair, dA-tailing and adaptor ligation reactions, take 3 µl of nuclei from the step 30 to perform a control ligaion reaction with the Illumina Y-adaptor as below:

Reagents (add in this order)	**Volume (µl)**	**Final Concentration**
Nuclei w/ beads	3	
Illumina Y-adaptor (50 µM)	1	2.5 µM
Water	10	
10X T4 ligase buffer w/ ATP	2	1X
PEG-4000 (50%)	2	5%
10% Triton X-100	1	0.5%
T4 DNA Ligase (5 U/ µl)	1	0.25U / µL
Total Volume	20	

After incubation at 16^{o}C for overnight, add 70 µl 1X TE lysis buffer and 10 µl Proteinase K (20 mg/ml). Incubate for 30-60 min at 65°C. Purify genomic DNA using a QiaQuick PCR purification kit. Check the ligation efficiency by carrying out qPCR with Illumina PCR primers. If upstream end-repair and dA-tailing steps are efficient, one should see amplification before 10 PCR cycles using 10 ng genomic DNA as template. We recommend this quality control step when performing the protocol on a new cell type. In the event of inefficiency of these steps, we recommend optimizing the concentration of SDS in the cell lysis step, or the amount of DNase I digestion used.

34) Add 5 µl of 10% SDS to stop the reaction.
35) Centrifuge for 60 seconds at 2,500x***g*** at RT.
36) Resuspend the pellet in 200 µl nuclease-free water.

37) Add 165 µl AMPure buffer; mix thoroughly by pipetting up and down.
38) Incubate at RT for 5 min, and place the tube in a DynaMag magnet for 2 min.
39) Discard the supernatant and wash the beads once with 1 ml of freshly prepared 80% (vol/vol) ethanol. Briefly spin down the beads and remove the residual ethanol.
CRITICAL STEP: We recommend diluting fresh 80% (vol/vol) ethanol before every experiment.
40) Resuspend the pellet in 200 µl water.
41) Add 165 µl of AMPure bead buffer; mix thoroughly by pipetting up and down.
42) Incubate mixture at RT for 5 min, then place tube in DynaMag magnet for 2 min.
43) Discard the supernatant and wash the beads twice with 500 µl of 80% (vol/vol) ethanol. Briefly spin down the beads and remove residual ethanol as completely as possible, then air-dry the beads for no more than 2 min.
44) Resuspend the nuclei-bead mixture in 80 µl of nuclease-free water.

Steps 45 – 46: *In situ* phosphorylation (Timing: 1.25 h)
45) Prepare the PNK reaction as follows:

Reagents (add in this order)	**Volume (µL)**	**Final Concentration**
Nuclei w/ beads	80	
10X T4 ligase buffer w/ ATP	10	1X
PNK (10 U/ µl)	10	1U / µL
Total Volume	100	

46) Incubate at 37^{o}C for 1 hr.

Steps 47 – 48: *In situ* ligation (Timing: 4.25 h)
47) Add the following reaction to the above tube:

Reagents (add in this order)	**Volume (µl)**	**Final Concentration**
H_2O	794	
10X T4 ligase buffer	100	1X
T4 DNA ligase (5 U/ µl)	6	0.03U / µL
Total Volume	1 mL	

48) Incubate at RT for 4 hr. For a micrograph of nuclei after this stage, see **Figure 3.2b**.

Steps 49 – 62: Cross-linking reversal, isopropanol precipitation and DNA purification (Timing: Overnight, followed by 2.5 h)
49) Centrifuge for 60 seconds at 2,500x***g*** at RT.
50) Resuspend the pellet in 400 µl 1X NEBuffer 2.
51) Add 40 µl 10% SDS.
52) Add 40 µl of 20 mg/ml Proteinase K.
53) Incubate overnight at 60^{o}C.
54) Add 3 µl GlycoBlue, 50 µl 3M sodium acetate, pH 5.2 and 550 µl of isopropanol.

55) Incubate mixture at -80°C for 2 hours.
56) Centrifuge mixture for 30 min. at 4°C at maximum speed in a microcentrifuge.
57) Resuspend the DNA pellets in each tube with 100 µl nuclease-free water.
58) Add 100 µl AMPure XP beads, mix well.
59) Incubate mixture at RT for 5 min, and place the tube in a DynaMag magnet for 2 min.
60) Discard the supernatant and wash the beads twice with 1 ml of 80% (vol/vol) ethanol. Briefly spin down the beads and remove residual ethanol as completely as possible, then air-dry the beads for no more than 2 min.
61) Resuspend the beads in 130 µl nuclease-free water.
62) Incubate beads at RT for 1 min. Collect beads via DynaMag magnet and transfer eluent to fresh 1.5 mL tube. At this point, determine the concentration of the recovered DNA with a spectrophotometer. A typical yield is 3-5 µg if starting with $2.5x10^6$ cells.
PAUSE POINT: Purified DNA can be stored indefinitely at -20°C.

Steps 63-65: DNA Sonication (Timing: 0.5 h)
CRITICAL At this point, purified DNA may be sonicated to shear large fragments to the 100-500 bp range or taken directly to sequencing library prep. Sonication promotes a less biased representation of fragment ends at the cost of additional prep time and loss of material. The protocol here is suitable for Covaris sonicators. If sonication is not desired, skip to step 66.

63) Transfer DNA to Covaris microtube.
64) Shear the DNA to a size of 100 – 500 bp using a sonicator. For a Covaris instrument use the following parameters:

Duty Cycle	15%
Peak Incident Power	450
Cycles per Burst	200
Set Mode	Frequency sweeping
Continuous degassing	
Process time	80 s
Number of cycles	5

65) Transfer 130 µL sonicated DNA to a 1.5 ml tube.
PAUSE POINT: Eluted DNA may be stored indefinitely at -20°C.

Steps 66 – 74: Biotin pull-down (Timing: 0.5 h)
66) Wash 30 µl of MyOne C1 beads twice with 100 µl 1X B&W buffer, once with 100 µl 2X B&W buffer, then resuspend in 100 µl 2X B&W buffer.
67) Add 100 µl eluted DNA to resuspended streptavidin beads and mix well.
68) Incubate the sample for 20 min at RT on a rotator.
69) Place tube in DynaMag magnet for 1 min and discard the supernatant.
70) Wash beads once with 300 µl 0.5X TE Lysis Buffer plus 300 µl 0.5X B&W buffer.
71) Wash beads twice with 600 µl 1X B&W buffer.
72) Wash beads once with 600 µl 1X NEBuffer 2.
73) Wash beads once with 600 µl EB buffer.
74) Resuspend beads in 170 µl of EB buffer.

PAUSE POINT: Resuspended beads may be stored at -20°C indefinitely or 4°C for short-term storage.

Steps 75 – 84: End Repair and dA-tailing (Timing: 1.5 h)

75) Set up the end-repair reaction with the Fast DNA End Repair Kit as follows:

Reagents (add in this order)	Volume (µl)	Final Concentration
Purified DNA	170	
10X Reaction buffer	20	1X
End-repair enzyme mix	10	
Total Volume	200	

76) Incubate at 18°C for 10 min.
77) Add 200 µl of Ampure buffer, mix thoroughly by pipetting up and down.
78) Incubate at RT for 5 min and place the tube in a DynaMag-Spin magnet for 2 min.
79) Discard the supernatant and wash the beads twice with 500 µl of 80% (vol/vol) ethanol. Briefly spin down the beads, remove the residual ethanol as completely as possible, and air-dry the beads for 5 min.
80) Resuspend beads in 21.5 µl water.
81) Set up the dA-tailing reaction as follows:

Reagents (add in this order)	Volume (µl)	Final Concentration
End-repaired DNA w/ beads	21.5	
10X NEBuffer 2	3	1X
10 mM dATP	3	1 mM
Klenow (exo$^-$) (5 U/ µl)	2.5	0.42U / µL
Total Volume	30	

82) Incubate at 37°C for 30 min.
83) Wash beads twice with 400 µl 1X B&W buffer.
84) Wash beads twice with 400 µl EB buffer and resuspend in 30 µl EB buffer.
CRITICAL STEP: Proceed immediately to adaptor ligation.

Steps 85 – 95: Ligation of sequencing adaptors (Timing: 1 h)

85) Immediately resuspend beads in the following reaction mixture:

Reagents (add in this order)	Volume (µl)	Final Concentration
dA-tailed DNA w/ beads	30	
5X Thermo Rapid Ligation Buffer	10	1X
Y-Adaptor (2.5 µM)	6	0.3 µM
T4 DNA ligase (5 U/ µl)	4	0.4U / µL
Total Volume	50	

86) Incubate at RT for 30 min.
PAUSE POINT: The ligation reaction in step 86 can also be performed at 16°C overnight.

87) Add 5 µl of 0.5 M EDTA to stop the reaction. Add 145 µl of ddH_2O to bring up the volume to 200 µl and mix thoroughly by pipetting up and down.
88) Add 200 µl of AMPure buffer to each tube and mix thoroughly by pipetting up and down.
89) Incubate at RT for 5 min and then place the tubes in a DynaMag magnet for 2 min.
90) Discard the supernatant and wash the beads twice with 500 µl of 80% (vol/vol) ethanol. Briefly spin down the beads and remove residual ethanol as completely as possible, then air-dry the beads for no more than 2 min.
91) Resuspend beads in 200 µl ddH_2O and add 165 µl of AMPure buffer
92) Mix thoroughly by pipetting up and down.
93) Incubate at RT for 5 min, and place the three tubes in a DynaMag magnet for 2 min.
94) Discard the supernatant and wash the beads twice with 0.5 ml of 80% (vol/vol) ethanol. Briefly spin down the beads and remove the residual ethanol as completely as possible, then air dry the beads for 5 min.
95) Resuspend the beads in each tube with 50 µl of EB.

Steps 96 – 104: Library amplification (Timing: 2.5 h)
CRITICAL: Optimization of input amount and PCR cycle number is integral to obtaining a sufficiently diverse *in situ* DNase Hi-C library. We recommend running several "pilot" PCR reactions with various bead input amounts and various cycle numbers and running these "pilot" libraries on a 6% TBE-PAGE gel to ensure that library overamplification is not occurring.

96) To determine the number of PCR cycles necessary to generate ample PCR products for sequencing—importantly, without over-amplification—set up trial PCR reactions with 10, 12, or 14 cycles, and 2.5 or 5 µl of DNA-bound streptavidin beads as follows:

Reagents (add in this order)	**Volume (µl)**	
End-repaired DNA w/ beads	2.5 / 5	
2X HotStart ReadyMix	10	1X
10 µM SeqPrimer_F	1	1 µM
10 µM SeqPrimer_R	1	1 µM
ddH_2O	up to 20	
Total Volume	20	

Using the following PCR program:

Cycle number	**Denature**	**Anneal**	**Extend**
1	95°C, 3 min.		
2-6	98°C, 20 sec.	60°C, 20 sec.	72°C, 1 min.
7-17*	98°C, 20 sec.	65°C, 20 sec.	72°C, 1 min.

*: Use optimized cycle number

97) Run 2 µl of each PCR reaction on a 6% TBE-PAGE gel to determine the appropriate number of cycles and amount of input beads for each PCR reaction. PCR products should run from ~200 bp to ~1 kbp, with the majority of product running from 300-600 bp, as shown in **Figure**

3.3b. Presence of products much larger than 1 kbp (*i.e.* those that do not migrate on a 6% TBE-PAGE gel) indicates overamplification, and should be avoided by reducing PCR cycle number or volume of beads used.

98) Aliquot remaining beads into 20 µl PCR reaction and amplify the remaining beads using multiple PCR reactions at the optimized cycle and input parameters.

99) Pool all PCR reactions into one 1.5 mL microcentrifuge tube.

100) Purify library by adding 0.8X volumes of AMPure XP beads.

101) Incubate mixture at RT for 5 min and place tube in a DynaMag magnet for 2 min.

102) Discard the supernatant and wash the beads twice with 1 ml of 80% (vol/vol) ethanol. Briefly spin down the beads and remove residual ethanol as completely as possible, then air-dry the beads for no more than 2 min.

103) Resuspend beads in 25 µl EB buffer and incubate at RT for 1 min.

104) Place resuspended beads on DynaMag magnet and transfer supernatant containing eluted DNA to fresh 1.5 mL tube.

Steps 105 – 109: Quality control of DNase Hi-C library by BamHI digestion (Timing: 1.25 h)

105) Quantitate amount of dsDNA in library using Qubit dsDNA HS kit as per manufacturer's protocols.

106) Digest a small aliquot of the final DNase Hi-C library (50 – 100 ng) with BamHI to estimate the portion of molecules with valid biotinylated junctions as follows:

Reagents (add in this order)	**Digest**	**(-) Control**
10X Fast digestion buffer	1 µl	1 µl
DNase Hi-C product	1-2 µl (50-100 ng)	1-2 µl (50-100 ng)
Fast digestion BamHI	1 µl	0 µl
Water	to 10 µl	to 10 µl

107) Incubate at 37^{o}C for 30 min.

108) Run the entire volume of the reaction on a 6% TBE-PAGE gel. Digested libraries should demonstrate a marked shift in library size distribution, as shown in **Figure 3.3b**. If libraries pass this QC metric, proceed to Illumina sequencing.

109) (Optional) Hybrid capture experiments may be carried out according to manufacturer's protocols provided with the Agilent SureSelect system.

Steps 110 – 120: Mapping, Normalization, and Visualization of Hi-C Contact Maps. TIMING dependent on volume of data)

110) Copy the output fastq sequencing files generated by the Illumina sequencer to the storage on the Linux computer.

[TROUBLESHOOTING 111]

111) Open a terminal on the computer and enter after the $ sign the commands described in the following steps. First, run FastQC to investigate the sequencing qualities, in which "L1_1"

and "L1_2" correspond to the fastq sequence files for read 1 and read 2, respectively.

```
$ fastqc --extract -f fastq L1_1.fq L1_2.fq
```

112) Obtain reference genome sequences. For instance, the mouse mm9 reference sequences can be downloaded from the UCSC Genome browser using the command below.

```
$ wget "http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/
chromFa.tar.gz"
$ tar -xzvf chromFa.tar.gz
$ gunzip -c chr*.fa.gz > mm9.fa
```

CRITICAL STEP: If the *in situ* DNase Hi-C data are from female cells, do not include chrY.

113) Run BWA to generate index files for the reference genome.

```
$ bwa index -a bwtsw -p mm9 mm9.fa
```

114) Run BWA to map each end of the pair-ended reads to the reference genome separately.

```
$ bwa aln mm9 L1_1.fq > L1_1.sai
$ bwa samse mm9 L1_1.sai mm9.fa > L1_1.sam
$ bwa aln mm9 L1_2.fq > L1_2.sai
$ bwa samse mm9 L1_2.sai mm9.fa > L1_2.sam
```

CRITICAL STEP: The two ends of the reads should be mapped separately.

115) Run samtools to exact high-quality (MAPQ>=30) and uniquely mapped reads.

```
$ samtools view -S -F 4 L1_1.sam | awk '$5>=30 && $12=="XT:A:U"'
| cut –f 1-4 | sort -k1,1 > L1_1.mapped
$ samtools view -S -F 4 L1_2.sam | awk '$5>=30 && $12=="XT:A:U"'
| cut –f 1-4 | sort -k1,1 > L1_2.mapped
```

116) Join mapped loci pairs if both ends are successfully mapped.

```
$ join L1_1.mapped L1_2.mapped > L1.mapped
```

117) Remove PCR duplicates.

```
$ cut –f 2-7 L1.mapped | awk 'BEGIN{OFS="\t";} {if($2<$5){print
$0;} else if($2>$5){print $4,$5,$6,$1,$2,$3;} else if($3<=$6){
print $0;} else{print $1,$2,$6,$2,$5,$3;}}' | sort –u > L1.unique
```

118) Parse the mapped contacts loci pairs to generate the Hi-C contact map at a given resolution.
119) Run ICE(Imakaev et al., 2012) to normalize the contact matrix using the Mirny lab's hiclib library (https://bitbucket.org/mirnylab/hiclib).
120) Visualize the contact map.

3.8 TROUBLESHOOTING

Step	Problem	Possible reasons	Solution
3	Low percentage of long-range contacts in sequencing library or BamHI digest does not shift library	Inefficient or incomplete formaldehyde crosslinking	For new cell types, optimizing the amount of formaldehyde used for crosslinking may be necessary.
7	Nuclear pellet disappears during *in situ* enzymatic treatments	Overtreatment of fixed nuclei with SDS	Reduce the amount of SDS used in the cell lysis.
14	gDNA digestion efficiency is poor	Undertreatment of fixed nuclei with SDS; inadequate amount of DNase I used for digestion	Optimization of the appropriate SDS and DNase I amounts may be necessary. We recommend performing the protocol through Step 38 for a variety of SDS concentrations (*i.e.* 0.1% - 0.5%) and DNase I amounts (*i.e.* 1U – 8U).
111	FastQC metrics are poor	High duplication rate in library (*e.g.* Fewer than 60% unique sequences); low quality sequencing run (*e.g.* total percentage of bases with q > 30 is less than 85%)	To maximize library complexity, make sure to set up several PCR reactions in Step 114. Issues with sequencing runs themselves may be difficult to diagnose and may require outside help.

3.9 TIMING

Day 1: Steps 1 – 32: Fixation; cell lysis; chromatin digestion, end repair, and adaptor ligation; ~6 h

Day 2: Steps 33 – 53: Adapter cleanup; *in situ* phosphorylation and ligation; crosslink reversal; ~6.5 h

Day 3: Steps 54 – 65: DNA purification and sonication; 2.5 – 3.5 h

Day 4: Steps 66 – 86: Biotin pulldown, end repair/dA tailing, and adaptor ligation of Hi-C fragments; ~3 h

Day 5: Steps 87 – 109: Library amplification, Bam HI quality check, and sequencing; ~4 h for amplification and quality check; up to several days / weeks for sequencing, instrumentation depending.

Day 6 and beyond: Steps 110 – 120: Data analysis time depends on sequencing depth and available compute resources.

3.10 ANTICIPATED RESULTS

We recommend QCing all libraries that pass the BamHI digestion test (typical results, including a negative control EcoRI digest, shown in **Figure 3.3b**) by sequencing at low depth first to ensure that the libraries are sufficiently complex for your desired application. We also recommend quantifying the length-classes of sequenced ligation pairs in libraries; *in situ* DNase Hi-C libraries should demonstrate an enrichment for pairs mapping with long-range (*i.e.* > 1 kb) distances between them (example distributions shown in **Figure 3.4a**). Furthermore, we recommend quantifying the relative numbers of different ligation pairs (*i.e.* "in-facing," "out-facing," "left," and "right") in libraries (a typical example is shown in **Figure 3.4b**). Corrected matrices generated from valid *in situ* DHC libraries should be analogous to the example shown in **Figure 3.4c**, with large scale structures (*i.e.* TADs) clearly visible even at 100 kb resolution.

We have observed that the relative fraction of interchromosomal ligation pairs in *in situ* DNase Hi-C libraries is largely cell-type specific, but highly reproducible—in line with previously published *in situ* results (Nagano et al., 2015; Rao et al., 2014). This is evident in **Figure 3.5**, which compares fractions of various ligation pairs between the Patski cell line, and three replicates of the human lymphoblastoid cell line GM12878. When considering gold-standards for *in situ* DNase Hi-C experiments, we typically look to the abundance of "long-range" ligation pairs in our libraries, which typically make up > 40% of uniquely mapped read pairs.

Using this modified DHC protocol, we have shown that the inactive murine X chromosome adopts a bipartite structure, consistent with results obtained using traditional Hi-C both in an analogous murine system (Minajigi et al., 2015) and human lymphoblastoid cells (Rao et al., 2014). These results suggest that the *in situ* DHC protocol produces signal comparable to existing Hi-C protocols while ultimately providing a less-biased empirical method for generating higher-resolution 3D maps of chromatin structure.

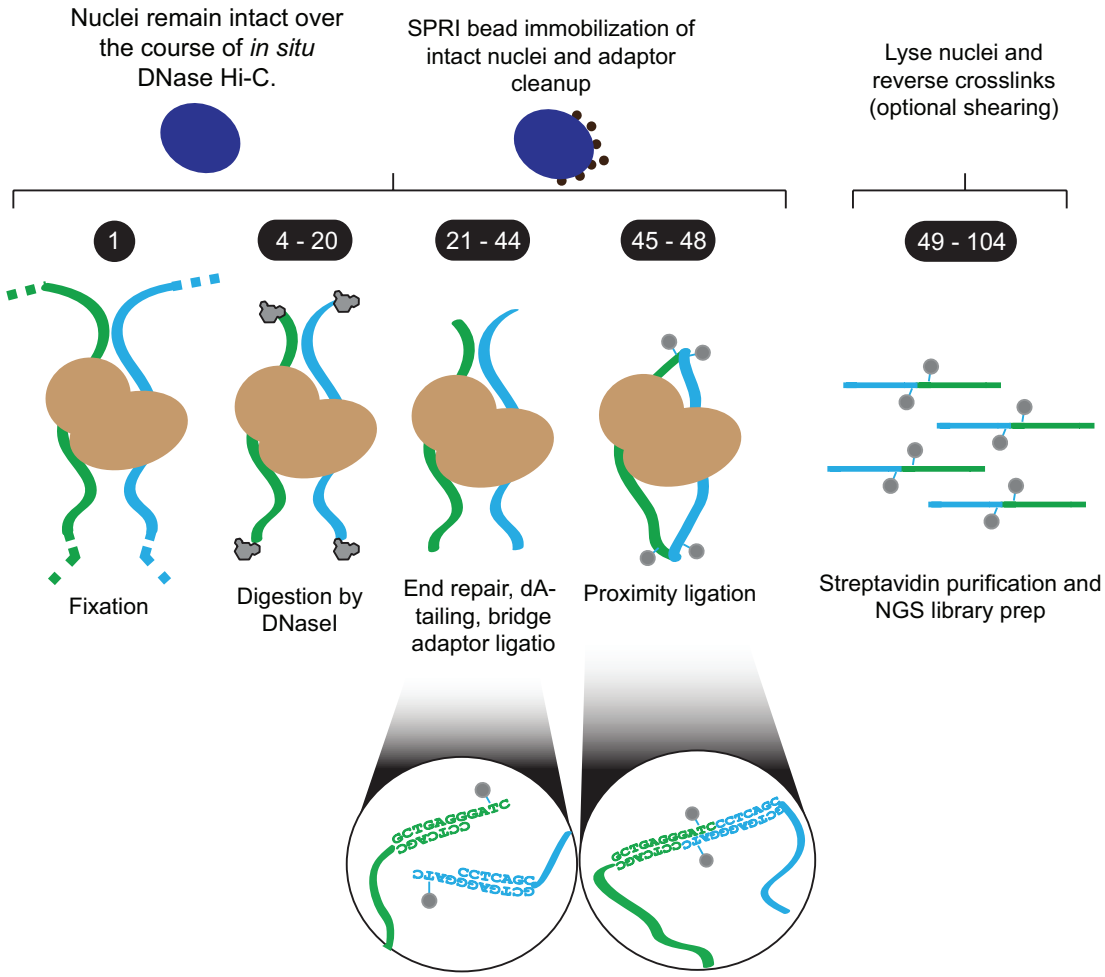


Figure 3.1. A schematic overview of *in situ* DNase Hi-C.

First, fixed cells are lysed and digested with the endonuclease DNase I in the presence of divalent manganese—yielding double stranded breaks. Nuclei are then immobilized on carboxylated paramagnetic beads (*i.e.* 'AMPure' beads) to purify intact nuclei and remove free digested DNA fragments. Chromatin is then end-repaired and dA-tailed *in situ*, and a biotinylated 'bridge adaptor' containing a half BamHI site is ligated onto free chromatin ends. Nuclei are then subjected to phosphorylation and *in situ* proximity ligation, after which DNA is purified and fragments containing ligation junctions are enriched for via streptavidin beads and on-bead Illumina library prep (optionally following sonication).

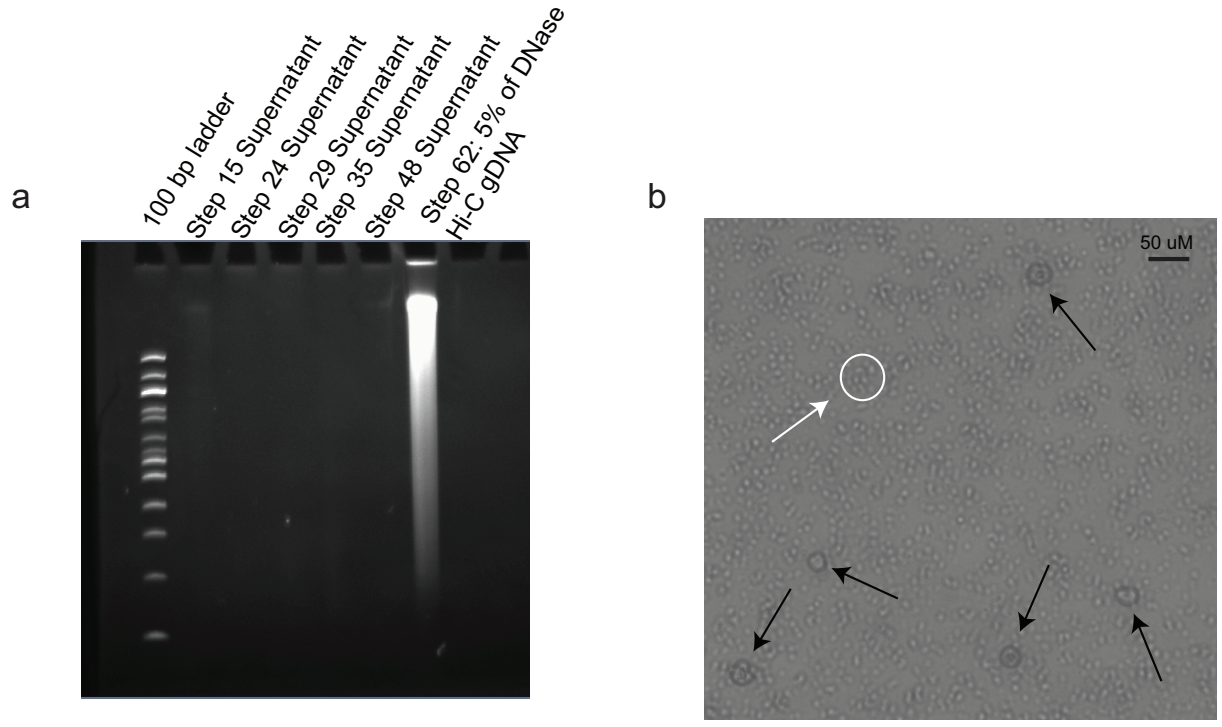


Figure 3.2. Nuclei remain intact during the *in situ* DNase Hi-C protocol.

a.) Purified supernatant DNA (see **Box 1)** from 6 different steps of the DNase Hi-C protocol. Minimal DNA is purified after each enzymatic purification, compared to a large amount of DNA, taken from 5% of the total gDNA yield following nuclear lysis. b.) Phase contrast micrograph (20X magnification) of GM12878 nuclei bound to beads, following proximity ligation (Step 46). Nuclei are highlighted using black arrows, and an example of a clump of carboxylated beads, which are found scattered across the image, is shown circled in white, with an accompanying white arrow.

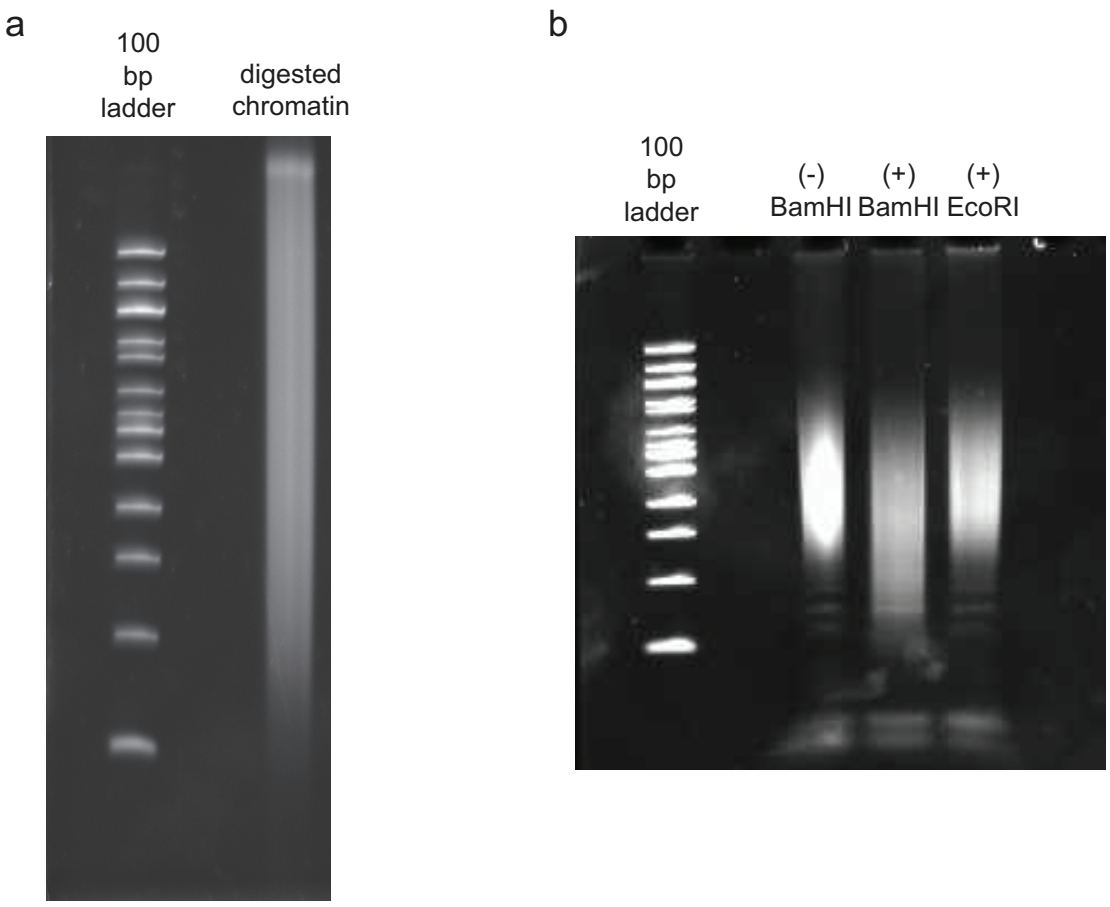
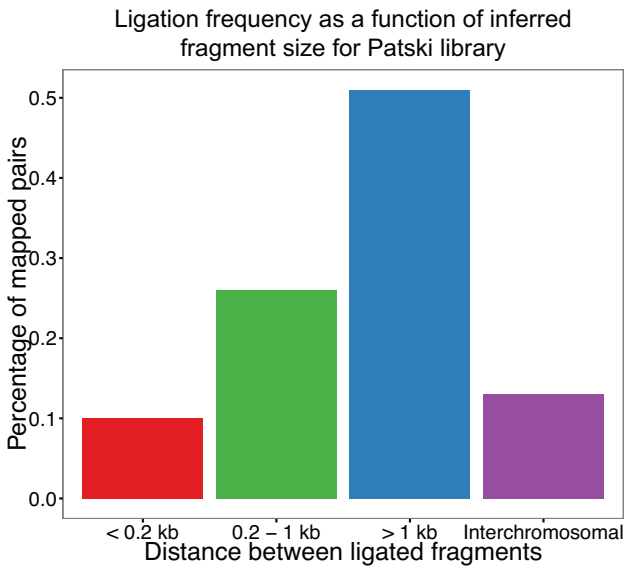


Figure 3.3. Digestion quality controls throughout the *in situ* DNase Hi-C protocol.

a.) A typical digestion pattern for DNase I-digested fixed chromatin prior to proximity ligation, run on a 6% TBE-PAGE gel. b.) Example of the BamH1 quality control experiment performed on GM12878 *in situ* DNase Hi-C libraries; in this example, BamH1 shifts the *in situ* DNase Hi-C library by digesting the reconstituted BamH1 site that forms following proximity ligation of the biotinylated bridge adaptors. Crucially, digestion with another 6-cutter (EcoR1), does not recapitulate this pattern, proving that the BamH1 digestion is specific to proximity ligated fragments. All reactions were run on one 6% TBE-PAGE gel.

a



b

	Left - Left	Right - Right	Inner	Outer
Fragments > 1 kb	120252	120010	121724	120262
Fragments <= 1 kb	13259	13298	293999	21621

c

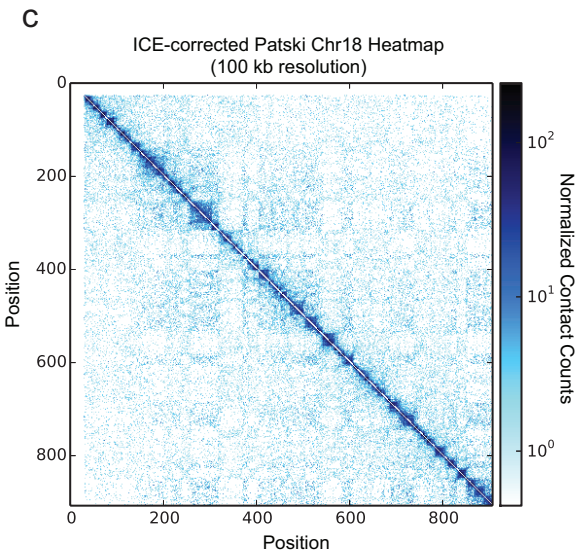


Figure 3.4. *In situ* DNase Hi-C results for the mouse embryonic kidney Patski cell line.

a.) *In situ* DNase Hi-C reads (950,206 downsampled reads from data published in Deng, Ma *et al* (2015) (using the mouse Patski cell line, rather than GM12878) demonstrate an enrichment for long-range (*i.e.* > 1 kb) intrachromosomal read pairs expected of Hi-C libraries. b.) Expected breakdown of mate orientations for read pairs in *in situ* DNase Hi-C data. For intrafragment distances > 1 kb, a roughly 25% split should be observed for each orientation class. c.) Normalized heat map generated from data published in Deng, Ma *et al* (GEO Accession: GSE68992) for mouse chromosome 18 at 100 kb resolution. The dataset used to generate this heatmap contained 60,666,200 uniquely mapped, high-quality read pairs.

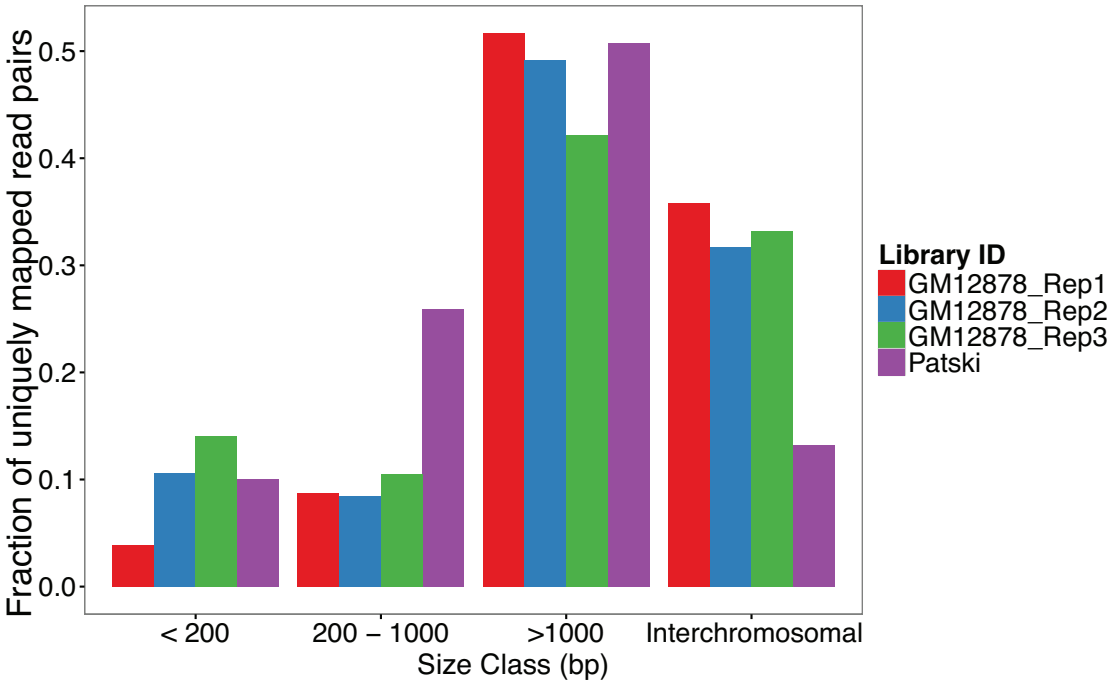


Figure 3.5. Relative Abundances of ligation types in 3 biological replicate GM12878 libraries, vs. a Patski library.

In situ DNase Hi-C reads demonstrate some cell-type specificity for the relative breakdown in long-range intrachromosomal read-pairs with respect to interchromosomal and short-range read-pairs, but these differences are reproducible, as shown when comparing three biological replicate libraries derived from the immortalized lymphoblastoid GM12878 cell line.

Chapter 4. MASSIVELY MULTIPLEX SINGLE-CELL HI-C

Note: Chapter 4 was published in the February 2017 issue of *Nature Methods* as:

Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, Shendure J. "Massively multiplex single-cell Hi-C." *Nature Methods* (2017).

4.1 ABSTRACT

We present single-cell combinatorial indexed Hi-C (sciHi-C), which applies the concept of combinatorial cellular indexing to chromosome conformation capture. In this proof-of-concept, we generate and sequence six sciHi-C libraries comprising a total of 10,696 single cells. We use sciHi-C data to separate cells by karytoypic and cell-cycle state differences and identify cell-to-cell heterogeneity in mammalian chromosomal conformation. Our results demonstrate that combinatorial indexing is a generalizable strategy for single-cell genomics.

4.2 MAIN TEXT

Our understanding of genome architecture has largely progressed through the successive development of new technologies (Ramani et al., 2016b). Advances in microscopy revealed the presence of "chromosome territories," nuclear regions that preferentially self-associate (Cremer and Cremer, 2001). The invention of Chromosome Conformation Capture (3C) and its derivatives (van Steensel and Dekker, 2010) resulted in a proliferation of data measuring genome architecture and its relation to other aspects of nuclear biology at increasing resolution.

3C assays rely on the concept of proximity ligation, a technique that has been used to measure local protein-protein (Soderberg et al., 2006), RNA-RNA (Ramani et al., 2015), and DNA-DNA interactions (Dekker et al., 2002). By coupling an "all-vs-all" 3C assay with massively

parallel sequencing (Duan et al., 2010; Lieberman-Aiden et al., 2009) (*e.g.* "Hi-C"), one is able to query relative contact probabilities genome-wide. However, contact probabilities generated by these assays represent ensemble averages of the respective conformations of the millions of nuclei used as input, and scalable techniques characterizing the variance underlying these population averages remain largely underdeveloped. A pioneering study in 2013 demonstrated proof-of-concept that Hi-C could be performed on single isolated mouse nuclei, but relied on the physical separation and processing of single murine cells in independent reaction volumes, with consequent low-throughput (Nagano et al., 2013).

The repertoire of high-throughput single-cell techniques for other biochemical assays has expanded rapidly as of late (Cusanovich et al., 2015; Rotem et al., 2015). Single-cell RNA-seq (scRNA-seq) was recently paired with droplet-based microfluidics to markedly increase its throughput (Klein et al., 2015; Macosko et al., 2015). Orthogonally, we introduced the concept of combinatorial cellular indexing (Cusanovich et al., 2015), a method that eschews microfluidic manipulation and instead tags the DNA within intact nuclei with successive (combinatorial) rounds of nucleic acid barcodes, to measure chromatin accessibility in thousands of single cells without physically isolating each single cell (single-cell combinatorial indexed ATAC-seq, or sciATAC-seq). Such throughput-boosting strategies have yet to be successfully adapted for single-cell chromosome conformation analysis.

To address this gap, we developed a high-throughput single-cell Hi-C protocol, termed single-cell combinatorial indexed Hi-C, or sciHi-C (**Figure 4.1a**), based on the concept of combinatorial indexing and also building on recent improvements to the Hi-C protocol (Deng et al., 2015; Rao et al., 2014). A population of 5 to 10 million cells is fixed, lysed to generate nuclei, and restriction digested *in situ* with the enzyme DpnII. Nuclei are then distributed to 96 wells,

wherein the first barcode is introduced through ligation of barcoded biotinylated double-stranded bridge-adaptors. Intact nuclei are then pooled and proximity ligated all together, followed by dilution and redistribution to a second 96-well plate. Importantly, this dilution is carried out such that each well in this second plate contains at most 25 nuclei. Following lysis, a second barcode is introduced through ligation of barcoded Y-adapters.

As the number of barcode combinations (96 x 96) exceeds the number of nuclei (96 x 25), the vast majority of single nuclei are tagged by a unique combination of barcodes. All material is once again pooled, and biotinylated junctions are purified with streptavidin beads, restriction digested, and further processed to Illumina sequencing libraries. Sequencing these molecules with relatively long paired-end reads (*i.e.* 2 x 250 base pair (bp)) allows one to identify not only the genome-derived fragments of conventional Hi-C, but also external and internal barcodes (each combination of which is hereafter referred to as a 'cellular index') which enable decomposition of the Hi-C data into single-cell contact probability maps (**Figure 4.1b**). Like sciATAC-seq (Cusanovich et al., 2015), this protocol can process hundreds to thousands of cells per experiment without requiring the physical isolation of each cell.

As a proof-of-concept, we applied sciHi-C to synthetic mixtures of cell lines derived from mouse (primary mouse embryonic fibroblasts (MEFs), and the 'Patski' embryonic fibroblast line) and human. All experiments were carried out such that subsets of cell types received specific barcodes during the first round of barcoding (*e.g.* in ML1 and ML2, each well during the first round of barcoding contained either HeLa S3 + Patski cells or HAP1 + MEF cells; see **Methods**).

Before deconvolving the resulting data to single cells, we examined the overall distribution of ligation junctions (*i.e.* contacts). Encouragingly, there were very few contacts between mouse and human (ML1: 0.006%; ML2: 0.008%), demonstrating minimal cross-talk between cellular

indices, and that nuclei remain intact through all ligation steps (confirmed through phase-contrast microscopy; **Figure 4.4**). We also examined the *cis:trans* ratio, defined here as the ratio of long-range (*i.e.* >20 kb) intrachromosomal contacts to interchromosomal contacts (**Figure 4.1c**), and found it to be on par with expectation for high-quality Hi-C datasets (ML1: 4.41; ML2: 4.38).

We next split the Hi-C data by cellular index and characterized the number of unique read-pairs associated with each, the vast majority of which should correspond to single cells. When examining a histogram of unique index occurrences as a function of read depth, we noted a bimodal distribution, reminiscent of patterns seen in sciATAC-seq datasets (Cusanovich et al., 2015), where low-coverage indices likely represent 'noise' consequent to tags from free DNA in solution (**Figure 4.5**). After discarding these, we infer 1,081 cellular indices in ML1, with a median of 9,274 unique read-pairs per index (ML2: 841 cellular indices; median of 8,335 unique read-pairs per index). Importantly, we also observe minimal barcode bias across replicate experiments (**Figure 4.6**), as well as similar median *cis*:*trans* ratios per cell (ML1: 4.43 with median absolute deviation (MAD) of 1.66; ML2: 4.34 with MAD of 1.66) (**Figure 4.1d, Figure 4.7**).

The only previously published example of single-cell Hi-C data suggests that high single cell *cis*:*trans* ratios are a hallmark of high-quality single-cell data (Nagano et al., 2013). The high *cis:trans* ratios that we observe are comparable to those of the 10 single-cell maps generated in that study, which reported a median value of 6.26 (MAD = 0.74), calculated as the ratio of *all* intrachromosomal contacts to interchromosomal contacts (*i.e.* with no cutoff for minimal intrachromosomal distance). Reanalyzing our own data using this more liberal criterion yielded similar ratios of 6.17 (ML1; MAD = 1.99) and 5.96 (ML2; MAD = 1.94). Of note, our ratios are calculated over 1,922 cellular indices (ML1 and ML2 combined), 857 of which have more than 10,000 unique contacts, compared to the 10 previously reported single cells each with at least

10,000 unique contacts. This comparison illustrates the scalability of combinatorial methods, as compared with methods relying on the physical isolation and serial processing of each single cell.

We designed our experiments to facilitate validation of the single-cell origin of each cellular index. Uniquely tagged cells should be associated with species-specific cellular indices in mixture experiments, with a collision rate broadly defined by a formulation of the "birthday problem (Cusanovich et al., 2015)." Consistent with the expected collision rate, we observed that 4.53% of all ML1 cellular indices (4.40% in ML2) were "collisions" (*i.e.* had less than 95% of reads mapping to either the mouse or human genome) (**Figure 4.2a,b**). For further analyses we filtered out any cellular indices failing this criterion, while accepting that we remain blind to "within species" collisions. We also filtered out indices where the associated *cis*:*trans* ratio was less than 1 (1.94% of indices in ML1; 1.62% in ML2), which could suggest broken nuclei.

Before continuing, we combined filtered data from ML1 and ML2 with equivalently filtered data from secondary experiments (PL1 and PL2) (**Figure 4.8**). We then employed a conservative genotype filter(2013a) which removed 20.4% of human cellular indices (**Figure 4.9**), leaving us with a combined dataset of 3,609 human single cell Hi-C maps. Together with mouse data (which were filtered for coverage, *cis*:*trans* ratio, and species purity), a total of 8,141 single cell Hi-C maps were generated across these four experiments.

We next explored whether cell types could be separated *in silico* on the basis of single-cell Hi-C signal. We generated matrices where rows represent single cells, and columns represent the number of contacts between pairs of chromosomes (**Figure 4.10a**). Principal components analysis (PCA) on this matrix resulted in separation of single HeLa S3 and HAP1 cells (**Figure 4.2c**), which was validated by our programmed barcode associations. Principal component 1 (PC1), which strongly correlated with coverage (**Figure 4.11**), accounted for the majority of the variance

(52.1%), while the combination of PC1 and principal component 2 (PC2; 1.07% of the variance) separated HeLa S3 and HAP1 cells. We then analyzed the "loadings" of our features in PC2, the axis separating HeLa S3 and HAP1 cells, and found that the strongest loadings recapitulated known translocations specific to HAP1(2014) (namely, translocations between chromosomes 15 and 19, and between chromosomes 9 and 22), while other strong loadings corresponded to documented HeLa S3 translocations (Naumova et al., 2013) (**Figure 4.2d**). Repeating these analyses by i.) removing specific interactions from the matrices and repeating PCA (**Figure 4.12**) ii.) using an alternate feature set (interacting 10 Mb intrachromosomal windows; **Figure 4.10b**, **Figure 4.13**), iii.) separating cells by replicate (**Figure 4.14**), and iv.) sequencing 908 additional human cells (K562 and GM12878; Library ML3 containing 1,175 cells total; **Figure 4.15**), all recapitulated cell-type separation to varying degrees, demonstrating that PCA can potentially be used to separate cell types on the basis of Hi-C signal. The ability to separate such populations could be invaluable, *e.g.* when studying tissue containing a mixture of normal cells and cancerous cells harboring translocations.

We next examined the heterogeneity present in single cell Hi-C maps in terms of polymer conformation. We plotted contact probability as a function of genomic distance for 769 single cells, each with at least 10,000 unique contacts (**Figure 4.16a**), finding that the pattern of scaling observed for single cells was markedly more disperse when compared to a shuffled control where the assignment of cellular indices to reads are randomized, regardless of species analyzed. We then examined the relationship between single-cell power-law scaling coefficients (**Figure 4.16b**), calculated between distances of 50 kb and 8 Mb (Imakaev et al., 2012; Sanborn et al., 2015), and single-cell *cis*:*trans* ratios, noting a correlation across four out of five experiments (**Figure 4.16c, Figure 4.17**) between high *cis*:*trans* ratios and shallow scaling coefficients.

To test whether this variance was related to the relative cell cycle state of single cells, we arrested HeLa S3 cells using nocadazole, an agent that leads to an enrichment of cells arrested at G2/M phase, and performed sciHi-C on this population (Library ML4; n = 1,380 filtered cells). Repeating the above analysis on this dataset yielded a strikingly wide variance in single-cell contact probability decay (**Figure 4.3a**), and subsequent calculation of scaling coefficients revealed a clear bimodal distribution in the data (**Figure 4.3b).** We then performed *in silico "*sorting" of this data to decompose the aggregate dataset into two distinct contact probability maps (**Figure 4.3c**), one harboring the "plaid" compartment pattern expected of interphase chromatin, and another harboring the condensed, compartment-free patterning of mitotic chromatin previously described by (Naumova et al., 2013). Although beyond the scope of our methodological proof-of-concept, the demonstration here of *in silico* cell sorting, as well as our empirical distributions for scaling coefficient in single cycling mouse and human cells, are likely to be highly useful in constraining computational models of mammalian chromosome conformation.

In summary, we present sciHi-C, a novel method for profiling chromosome conformation in single cells, that relies on combinatorial cellular indexing for rapid scaling to large numbers of cells. For this proof-of-concept, we applied this method to generate single-cell Hi-C maps for 10,696 cells with at least 1,000 unique contacts. This dataset is two orders of magnitude larger than the only published single-cell Hi-C dataset, with 3,515 filtered cells containing more than 10,000 unique contacts, compared to the 10 existing single-cell maps defined using a similar coverage cutoff.

Given the generally similar workflow of our method and traditional bulk Hi-C, it may be possible to incorporate into routine practice, thus adding a 'single cell' dimension to Hi-C data production and a means of obtaining single-cell and bulk measurement at once (the latter generated

by summing single cells). Furthermore, our demonstration that thousands of single-cell Hi-C maps can be generated in a single workflow, without the need to isolate each cell, demonstrates the power of combinatorial indexing for large-scale single cell biology. Combinatorial indexing may thus be generalizable to additional aspects of single cell or even intracellular biology where DNA barcodes can be incorporated *in situ*.

4.3 METHODS

4.3.1 *Cell Culture*

HeLa S3 (CCL2.2) (gift from Malik Lab), primary MEFs (gift from Ware Lab), and Patski (gift from Disteche lab) cells were cultured at 37°C, 5% CO_2 in DMEM supplemented with 1X Pen-Strep (Gibco), and 10% FBS (Gibco). HAP1 cells (Haplogen) were cultured were cultured at 37°C, 5% CO_2 in IMDM supplemented with 1X Pen-Strep and 10% FBS. K562 cells were cultured at 37°C, 5% CO_2 in RPMI-1640 supplemented with 1X Pen-Strep and 10% FBS. GM12878 cells were cultured at 37°C, 5% CO_2 in RPMI-1640 supplemented with 1X Pen-Strep and 15% FBS. Cells were not tested for mycoplasma.

4.3.2 *Cell Fixation*

Adherent cells (*i.e.* HeLa S3, HAP1, Patski, MEF) were washed once with 1X PBS (Life Technologies), trypsinized (0.25% Trypsin-EDTA, Life Technologies), spun down at 500x**g** for 5 min., and resuspended in 20 mL serum-free DMEM (IMDM for HAP1). Cells were crosslinked by adding 1.12 mL (2% final concentration, for HeLa S3, HAP1, and MEF) or 1.4 mL (2.5% final concentration, for Patski) 37% formaldehyde (Alcon) and incubated at RT (25°C) for 10 min., after which crosslinking was quenched using 1 mL 2.5M glycine. Quenched reactions were incubated on ice for 15 min., spun down at 800**xg** for 5 min., resuspended in 1X PBS, aliquoted

into 10E6 cell aliquots, pelleted once again at 800**xg** for 5 min, decanted, snap frozen in liquid nitrogen, and finally stored indefinitely at -80°C.

Suspension cells (*i.e.* K562, GM1878) were spun down at 500**xg** for 5 min., resuspended in 20 mL serum-free RPMI-1640, crosslinked with a final concentration of 2% formaldehyde, and processed as above.

For nocadazole arrest experiments, we plated HeLa S3 cells in T75 flasks to ~10% confluency. 24 hours later, we replaced media with DMEM containing 10% FBS and nocadazole to a final concentration of 100 ng / mL. We then waited 24 hours, harvested cells by first harvesting detached cells, then trypsinizing the remaining plated cells. This resulted in a a heterogeneous single-cell suspension which we then fixed as above using 2% formaldehyde.

4.3.3 *Single-Cell Combinatorial Indexed Hi-C (sciHi-C)*

For the step-by-step combinatorial sciHi-C protocol, see https://www.nature.com/protocolexchange/protocols/5423. Like the recently published scDNase-seq protocol(2015e), sciHi-C uses carrier plasmid to prevent DNA losses during steps of the protocol where small amounts of DNA are handled. The libraries prepared here each used fixed aliquots of 5 to 10 million cells, which are diluted over the course of the protocol. All libraries were sequenced on a HiSeq 2500.

4.3.4 *Barcode Programming*

Our primary datasets (Library ML1 and biological replicate library ML2), used HeLa S3, HAP1, Patski, and MEFs, with subsets of human and mouse cell types in distinct wells during the first round of barcoding (HeLa S3 + Patski in half of wells; HAP1 + MEFs in half of wells). Our

secondary datasets (Library PL1 and biological replicate PL2) were generated using the same cell types, but a subtly different programming scheme (illustrated in **Figure 4.18**), wherein each well contained only a single cell type during the first round of barcoding. Finally, we generated and lightly sequenced a 5th library (Library ML3), mixing the same murine cell types as before with two new human cell types—GM12878 and K562—in a similar manner to Libraries ML1 and ML2 (GM12878 + Patski in half of wells; K562 + MEFs in half of wells).

4.3.5 *Bridge Adaptor Barcode Design*

Bridge adaptor barcodes were drawn from randomly generated 8-mers, such that the following criteria were met: i.) all adaptors must have a minimum pairwise Levenshtein distance of 3; ii.) adaptors must not contain the sequences TTAA or AAGCTT; iii.) adaptors must contain >60% GC content; iv.) adaptors must not contain homopolymers >= length 3; and v.) adaptors must not be palindromic.

4.3.6 *Processing sciHi-C Data*

All code used for sciHi-C data analysis is available at https://github.com/VRam142/combinatorialHiC. Below, we describe in detail the analytical pipeline used to process the data. The analytical steps broadly fall under three categories: i.) Barcode Identification & Read Trimming, ii.) Read Alignment, Read Pairing, & Barcode Assocation, and iii.) Cellular Demultiplexing & Quality Analysis.

Barcode Association & Read Trimming

First, to obtain round 2 (*i.e.* terminal) barcodes, we use a custom Python script to iterate through both mates, compare the first 8 bases of each read against the 96 known barcode sequences, and

then assign barcodes to each mate using a Levenshtein distance cutoff of 2. Reads "split" in this way are output such that the first 11 bases of each read, which derive from the custom barcoded Y adaptors, are removed. Mates where either terminal barcode went unidentified, or where the terminal barcodes did not match, are discarded.

For each resulting "split" pair of reads, the two reads are then scanned using a custom Python script to find the common portion of the bridge adaptor sequence. The 8 bases immediately 5' of this sequence are isolated and compared against the 96 known bridge adaptor barcodes, again using a Levenshtein distance cutoff of 2. There are cases where the entire bridge adaptor, including both barcodes flanking the ligation junction, is encountered in one mate, and not the other. To account for these cases, we also isolate the 8 bases flanking the 3' end of the common bridge adaptor sequence (when it is encountered within a read), reverse complement it, and compare the resulting 8-mer against the 96 known bridge adaptor barcodes. Output reads are then clipped to remove the bridge adaptor and all 3' sequence. Barcodes flanking the ligation junction should match; again, mates where barcodes do not match, or where a barcode is not found are discarded.

The result of this processing module are three files: filtered reads 1 and 2, and an "associations" file—a tab-delimited file where the name of each read passing the above filters and their associated barcode combination are listed.

Read Alignment, Read Pairing, & Barcode Assocation

As is standard for Hi-C reads, the resulting processed and filtered reads 1 and 2 were aligned separately using bowtie2/2.2.3 to a Burrows-Wheeler Index of the concatenated mouse (mm10) and human (hg19) genomes. Individual SAM files were then converted to BED format and filtered for alignments with MAPQ >= 30 using a combination of samtools, bedtools, and awk. Using bedtools closest along with a BED file of all DpnII sites in both genomes (generated using

HiC-Pro(Servant et al., 2015)), the closest DpnII site to each read was determined, after which BED files were concatenated, sorted on read ID using UNIX sort, and then processed using a custom Python script to generate a BEDPE format file where 5' mates always precede 3' mates, and where a simple Python dictionary is used to associate barcode combinations contained in the "associations" file with each pair of reads. Reads were then sorted by barcode, read 1 chromosome, start, end, read 2 chromosome, start, and end using UNIX sort, and deduplicated using a custom Python script on the following criteria: reads were considered to be PCR duplicates if they were associated with the same cellular index, and if they comprised a ligation between the same two restriction sites as defined using bedtools closest.

Cellular Demultiplexing & Quality Analysis

When demultiplexing cells, we run two custom Python scripts. First, we generate a "percentages" file that includes the species purity of each cellular index, the coverage of each index, and the number of times a particular restriction fragment is observed once, twice, thrice, and four times. We also include the *cis*:*trans* ratio described above, and, if applicable, the fraction of homozygous alternate HeLa alleles observed. We use these percentages files to filter BEDPE files (see below) and generate, at any desired resolution, single cell matrices in long format (*i.e.* BIN1-BIN2-COUNT), with only the "upper diagonal" of the matrix included to reduce storage footprint. These matrices are then converted to numpy matrices for visualization and further analysis.

Filtration of Cellular Indices

We applied several filters to our resulting cellular indices to arrive at the cells analyzed in this study. We first removed all cellular indices with fewer than 1000 unique reads. We next filtered out all indices where the *cis*:*trans* ratio was lower than 1. Finally, for all experiments we

removed cellular indices where less than 95% of reads aligned uniquely to either the mouse (mm10) or human (hg19) genomes. For all human cells from HAP1 and HeLa S3 mixing experiments (Libraries ML1, ML2, PL1, and PL2) further filtration by genotype was performed. For each cellular index, we examined all reads overlapping with known alternate homozygous sites in the HeLa S3 genome and computed the fraction of sites where the alternate allele is observed. We then drew cutoffs to filter out all cells where this fraction fell between 56% and 99%. We employ this filtering step purely as an additional, conservative measure, and note that this is not strictly necessary. The clear separation of two populations in data derived from library ML4 (nocadazole arrest experiment), where no genotype filtration was performed, illustrates this.

We do acknowledge that particular applications (*e.g.* structural modeling) may require more stringent filtration for cellular indices covering single cells. As such, we provide with the raw data a supplementary file specifying the "species purity" of each barcode combination in each sequenced library, along with the number of times DpnII restriction fragments are observed in a cell once, twice, thrice, or four times, with the expectation that given some tolerable noise level, one should only observe restriction fragment copy numbers equal to or less than the copy number of that fragment for that cell type. Relatedly, we note that further inspection of the HAP1 cells used in this study revealed that they were not entirely haploid. HAP1 cells, an engineered haploid line, have faster doubling times compared to HeLa S3, and have been described as having a relatively large frequency of diploid cells(2011). FACS analysis (data not shown) of the stock used for these experiments showed that ~40% of cells analyzed harbored 2*n* nucleic acid content, indicating haploid cells in G2 or reverted diploid cells in G1.

4.3.7 *Data Analysis*

PCA of sciHi-C Data

Single-cell matrices at interchromosomal contact resolution ($\log_{10}$ of contact counts) and 10 Mb resolution (binarized; 0 if absent, 1 if present) were vectorized and concatenated using custom Python scripts. Concatenation was performed such that redundant entries of each contact matrix (*i.e.* C_{ij} and C_{ji}) were only represented once. Resulting matrices, where rows represent single-cells and columns represent observed contacts, were then decomposed using the PCA function in scikit-learn. For interchromosomal matrices, entries for intrachromosomal contacts (*i.e.* the diagonal) were set to 0. For 10 Mb intrachromosomal matrices, all interchromosomal contacts were ignored and all entries C_{ij} where $| i - j | < 3$ were set to zero.

Calculation of Contact Probabilities in Single Cells

Methods to calculate the scaling probability within single cells were adapted from Imakaev, Fudenberg *et al* (2012) and Sanborn, Rao *et al* (2015). A histogram of contact distances normalized by bin size was generated using logarithmically increasing bins (increasing by powers of 1.12^n). We obtained the scaling coefficient by calculating the line of best fit for the log-log plot of this histogram between distances of 50 kb and 8 Mb. Shuffled controls were generated by randomly reassigning all cellular indices and repeating the above analysis; this importantly maintains the coverage distribution of the new set of simulated "single cells."

All plots were generated in R using ggplot2 (http://ggplot2.org/).

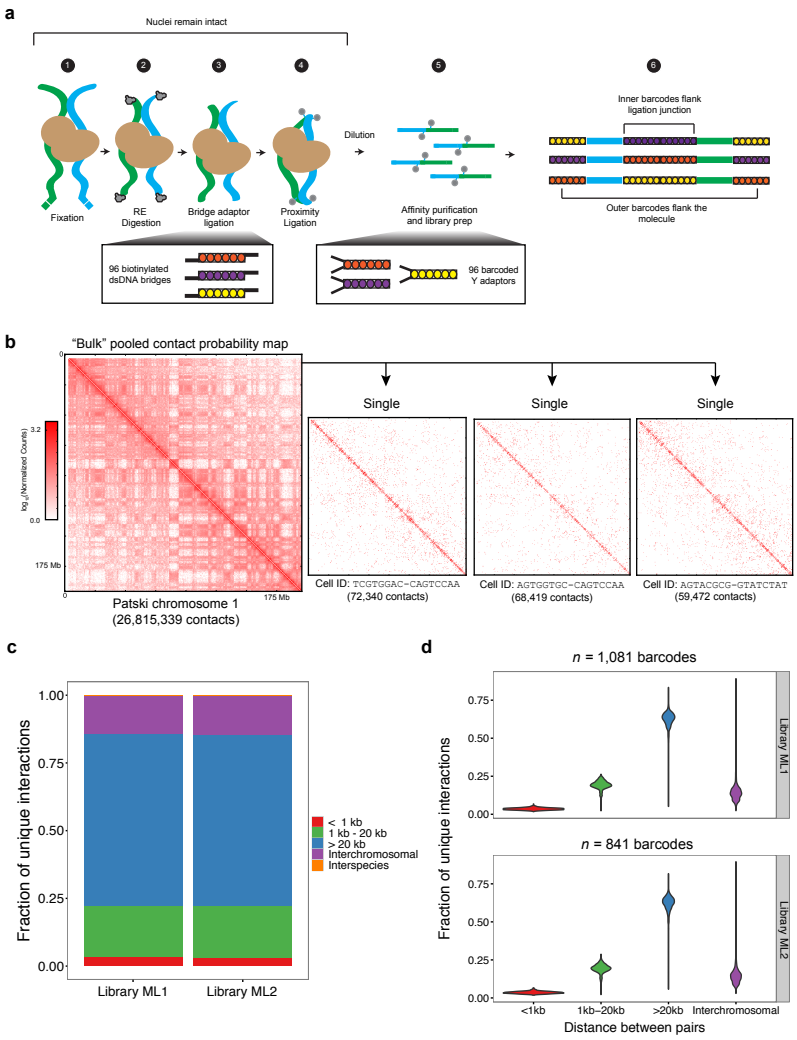


Figure 4.1. Single-cell combinatorial indexed Hi-C integrates the *in situ* Hi-C protocol with combinatorial cellular indexing to generate signal-rich bulk Hi-C maps that can be decomposed into single cell Hi-C maps.

a.) sciHi-C follows the traditional paradigm of fixation, digestion, and re-ligation shared by all Hi-C assays (Steps 1 – 4), but uses a biotinylated bridge adaptor to incorporate a first round of barcodes prior to proximity ligation (Step 3), and custom barcoded Illumina Y-adaptors (Step 5) to incorporate a second round of barcodes prior to affinity purification and library amplification (Steps 5 – 6). b.) Bulk data generated by this protocol can be decomposed to single cell Hi-C maps. c.) sciHi-C libraries demonstrate a high *cis*:*trans* ratio, measured as the ratio of intrachromosomal contacts > 20 kb apart to interchromosomal contacts. d.) The high *cis*:*trans* ratio observed in bulk data is maintained after libraries are decomposed to ~1800 cellular indices (each with >= 1,000 unique reads).

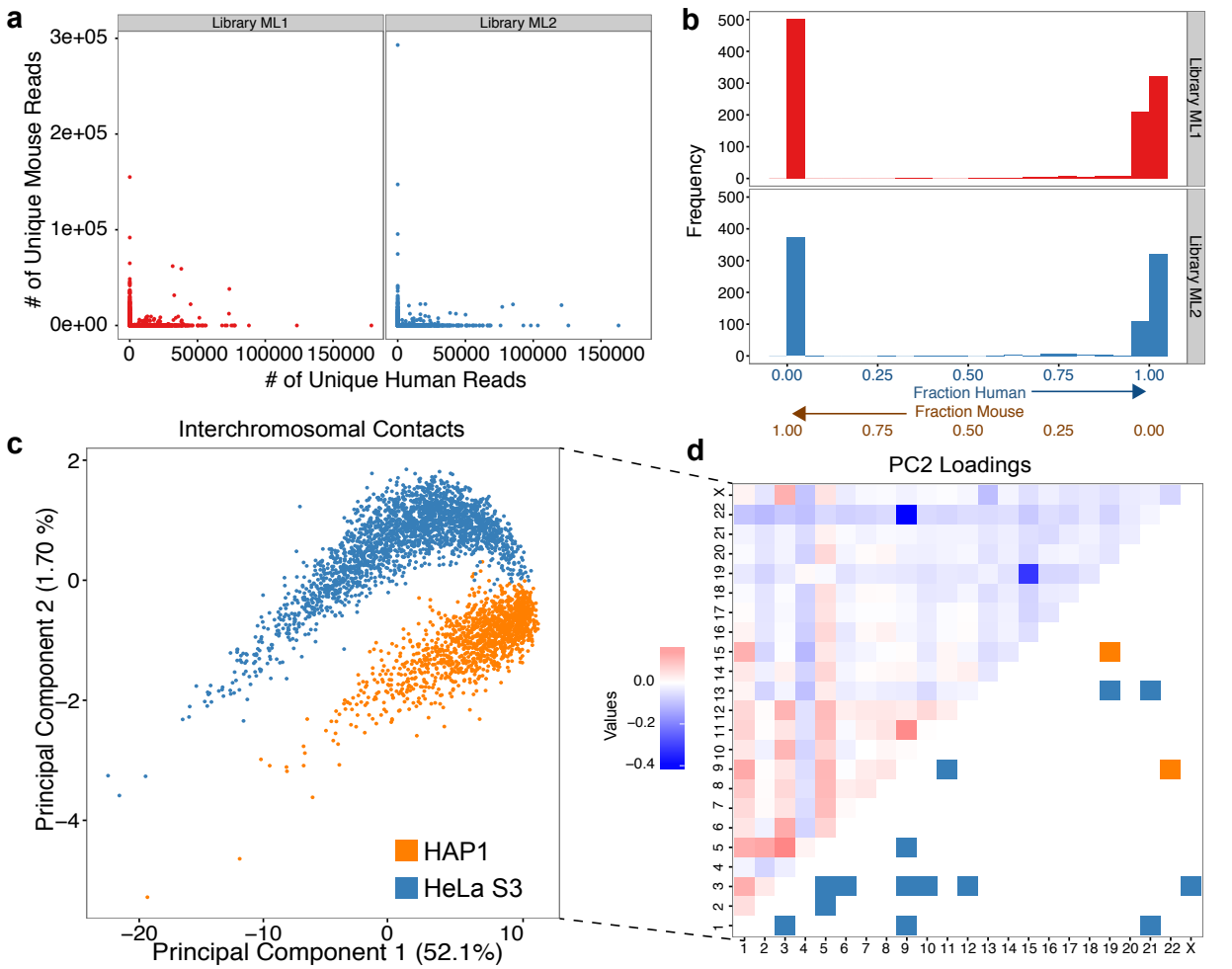


Figure 4.2. The large number of cellular indices generated through combinatorial single cell Hi-C are overwhelmingly species-specific, and can be separated by cell type.

a.) In libraries ML1 and ML2, similar levels of collision (defined as any cellular index with at least 1,000 unique reads, but <95% species purity) are observed, and fall within the expected range. b.) Species contamination visualized as a histogram of the fraction of reads mapping to the human genome (only cellular indices with >= 1000 reads shown). c.) Projection onto the first two principal components from PCA analysis of interchromosomal contact matrices results in separation of HeLa S3 and HAP1, two karytoypically different cell lines (n = 3,609 cells). Percentages shown are the percentage of variance explained by each plotted component. d.) Principal component 2 loadings represent the contribution of each feature (interchromosomal contact) to the observed cell type separation. Known translocations for each cell type are mirrored against the loading heatmap.

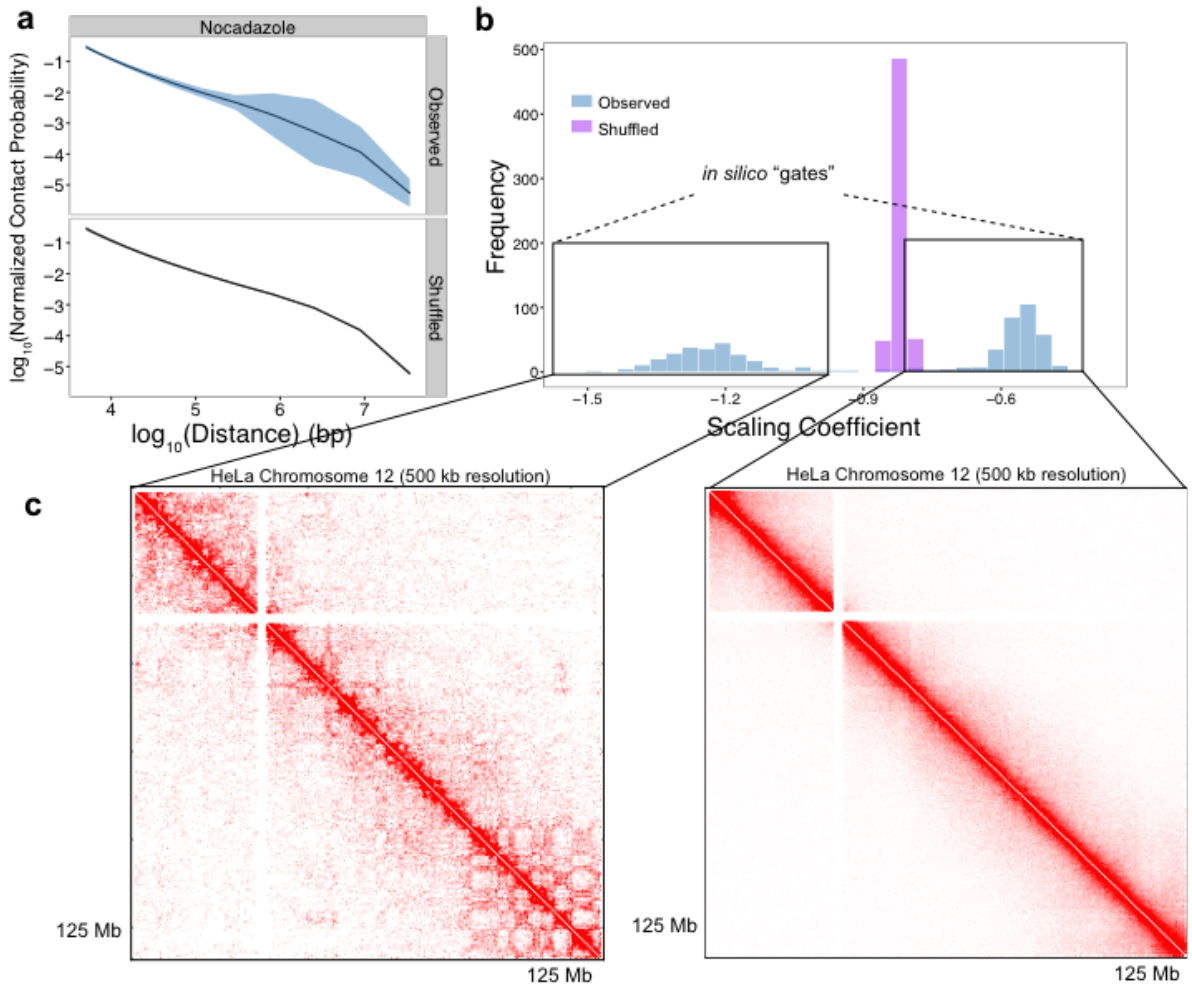


Figure 4.3. sciHi-C of nocadazole arrested HeLa S3 cells enable *in silico* sorting by cell cycle progression.

a.) Mean contact probability and standard deviation as a function of genomic distance for single HeLa S3 cells from a population treated with nocadazole (n = 588 cells containing at least 5,000 contacts and harboring nocadazole experiment-specific programmed barcodes). As a control, untreated cells were processed simultaneously (data not shown). b.) Scaling coefficients for 588 single HeLa S3 cells follow a bimodal distribution. c.) Cells can be "sorted" *in silico* to generate two distinct contact probability maps, shown here for HeLa chromosome 12.

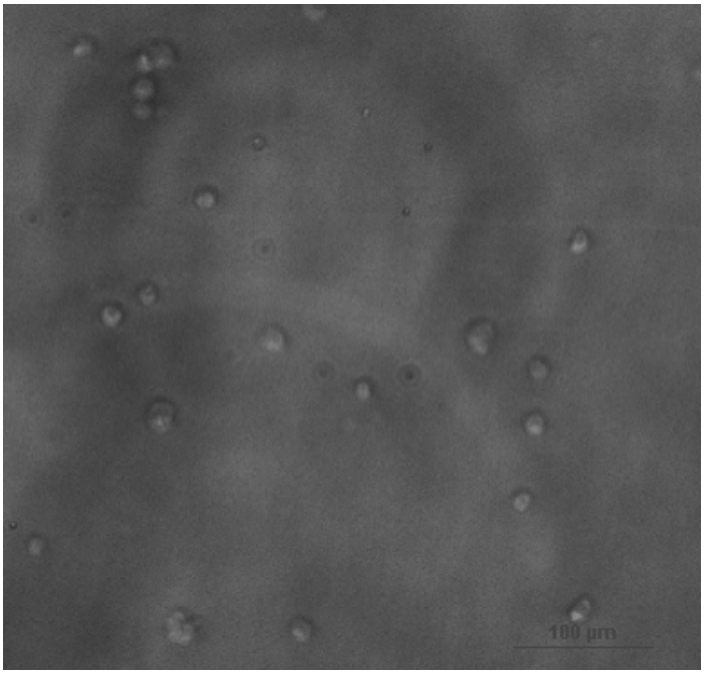


Figure 4.4. Nuclei remain intact through proximity ligation in the combinatorial single cell Hi-C protocol.

Phase contrast microscopy of HeLa S3 and HAP1 nuclei following proximity ligation and serial dilution shows that nuclei remain intact throughout the combinatorial single cell Hi-C protocol (scale bar = 100 um).

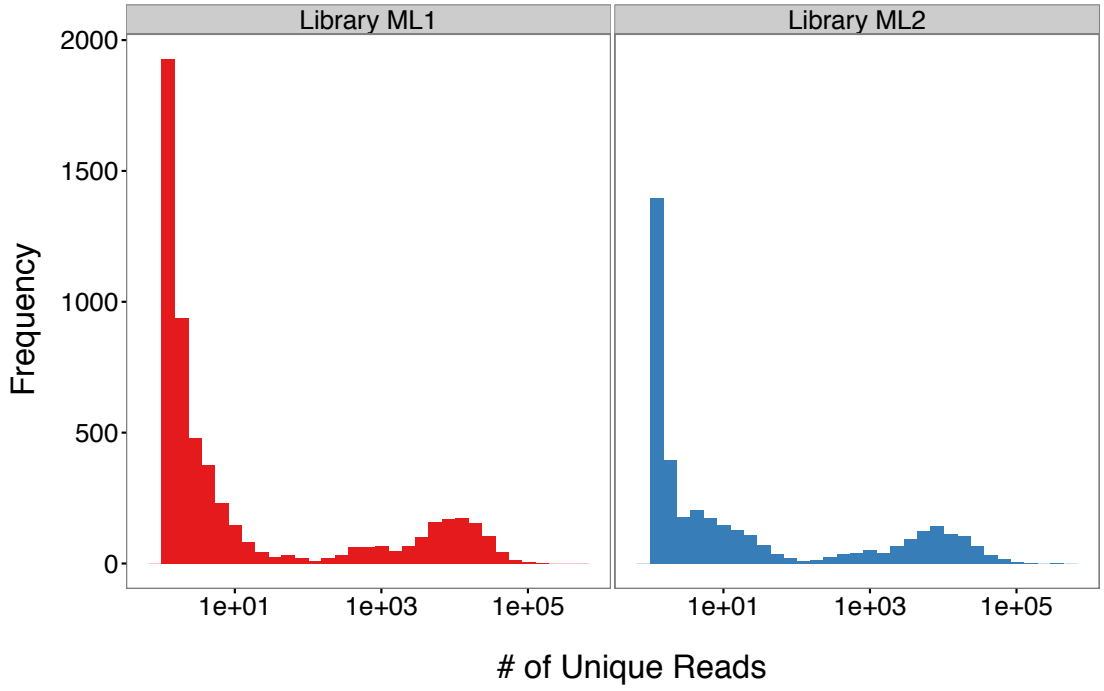


Figure 4.5. Coverage of combinatorial single cell Hi-C cellular indices follows a bimodal distribution.

Examining a histogram of the coverage (*i.e.* # of unique reads) of combinatorial single cell Hi-C cellular indices in two replicate libraries reveals a bimodal distribution, where low coverage cellular indices likely represent barcoding of free DNA in solution, rather than intact nuclei.

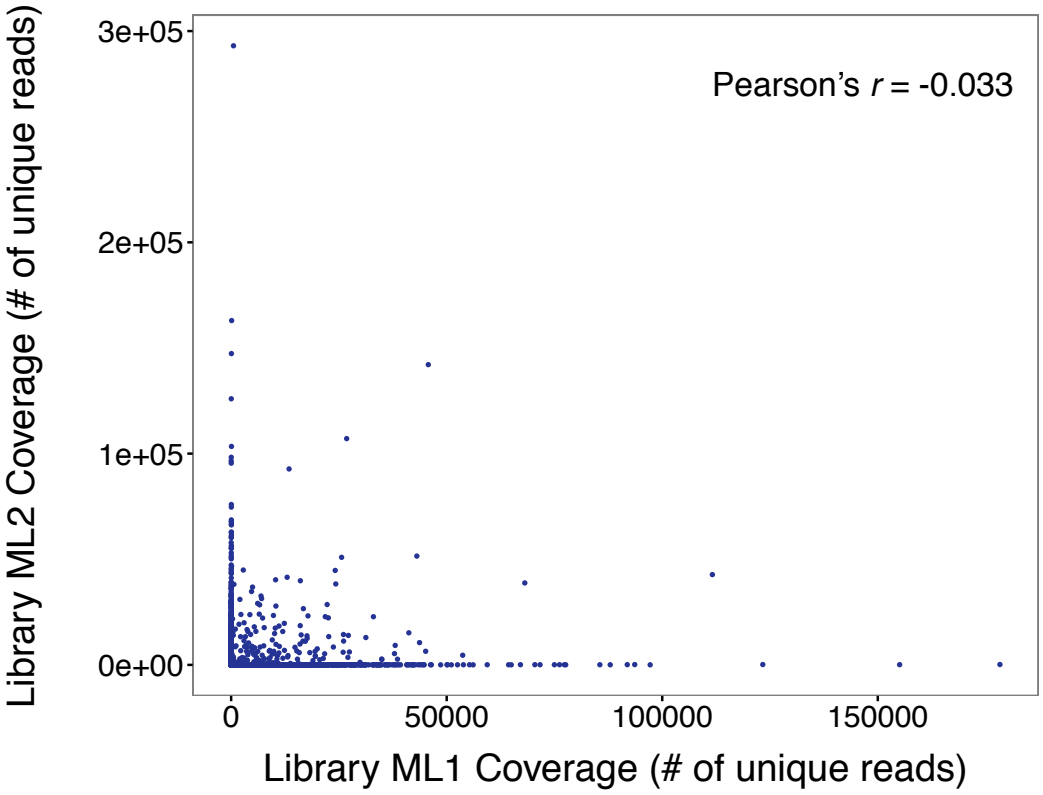


Figure 4.6. Coverage of cellular indices is not correlated between replicate experiments. Scatter plot of coverage per cellular index for all cellular indices with at least 1 unique read in both replicate combinatorial single cell Hi-C libraries. A Pearson's *r* of -0.03 suggests that there is minimal intrinsic bias (*i.e.* "barcode" effect) biasing coverage of particular cellular indices.

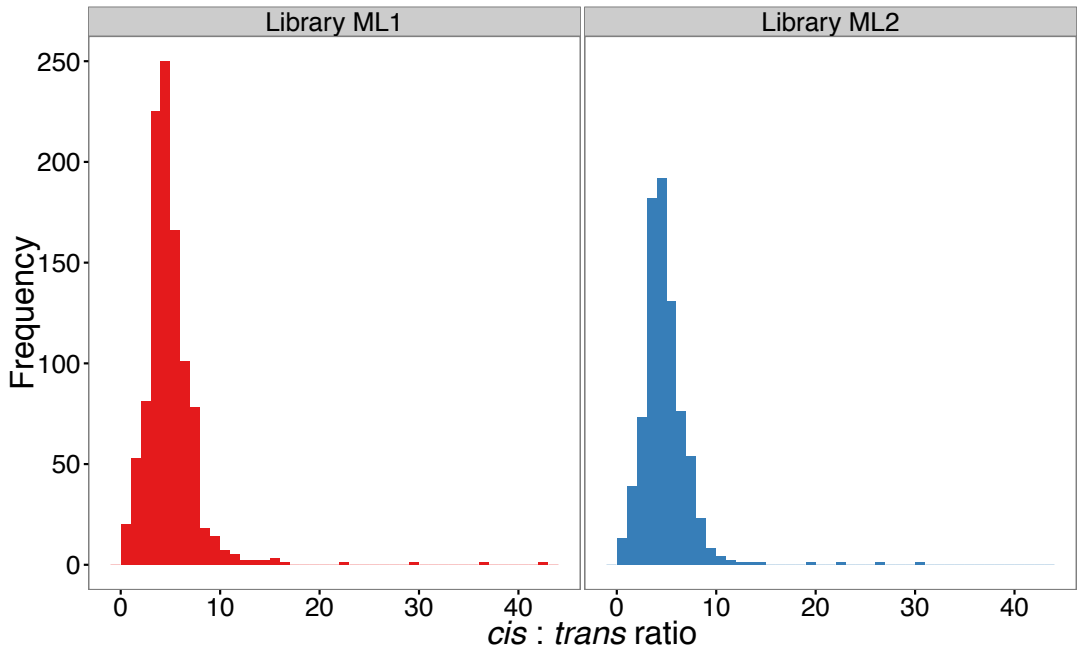


Figure 4.7. Single cellular indices demonstrate high *cis*:*trans* ratios.

Histogram of the *cis*:*trans* ratios for cellular indices over two biological replicates. High *cis*:*trans* ratio suggest that nuclei remain intact during the protocol, and hint at a single-cellular origin for the majority of cellular indices.

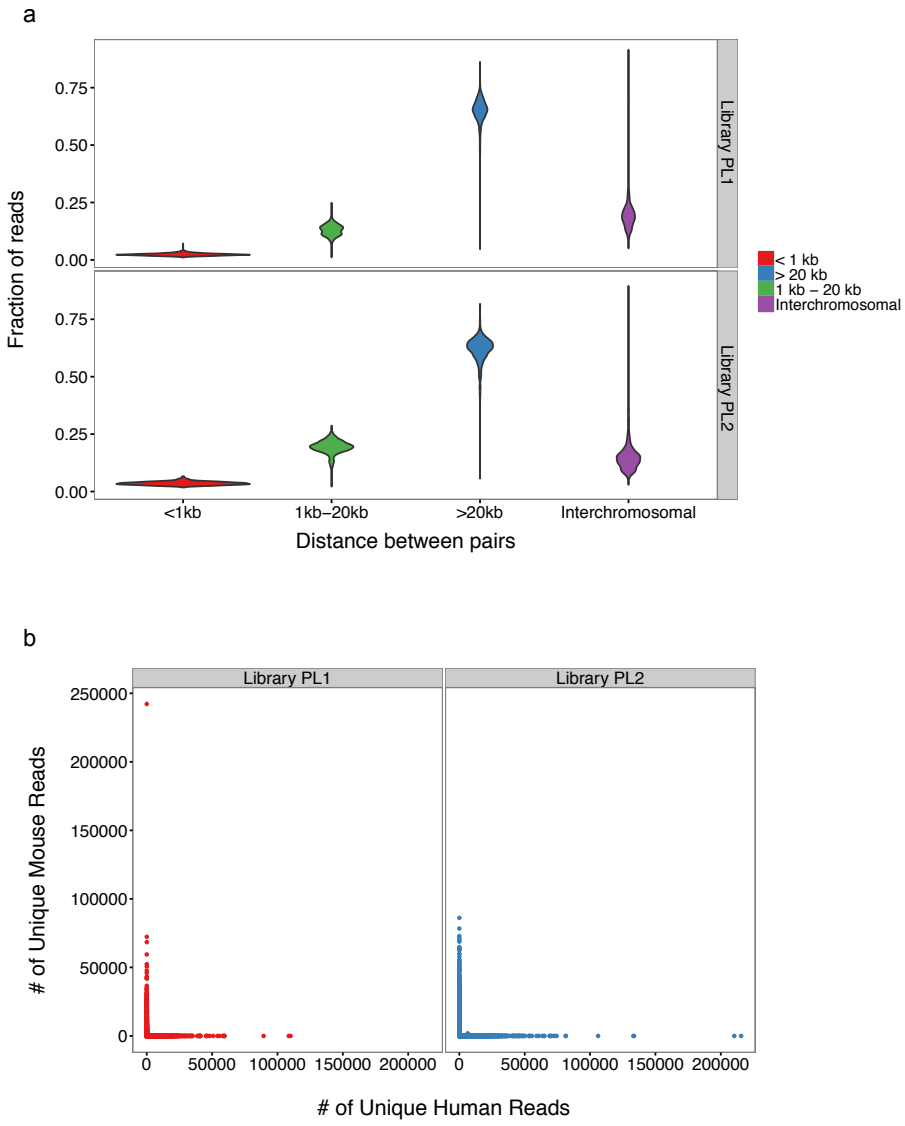


Figure 4.8. Quality control statistics for PL1 and PL2 libraries are similar to primary experiment libraries.

a.) Violin plots showing the distribution of ligation types across all cellular indices with at least 1,000 reads in libraries PL1 and PL2. b.) Species specificity for both libraries.

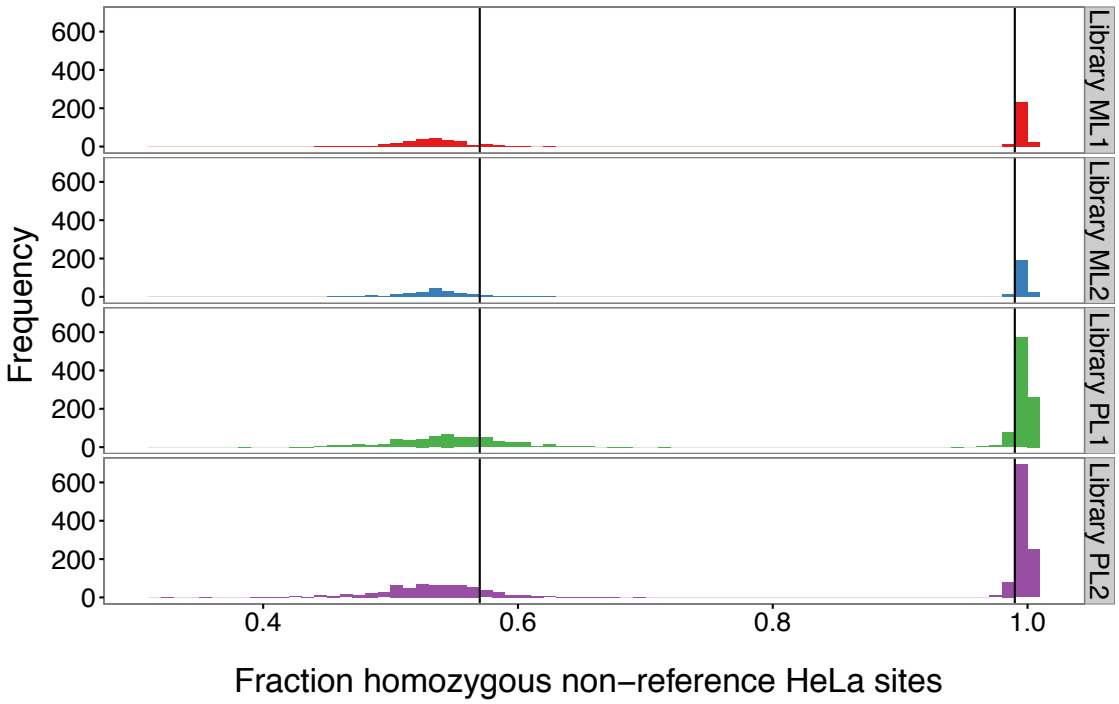


Figure 4.9. The HeLa genotype enables further filtration of potential barcode collisions in combinatorial single cell Hi-C datasets.

We examined all homozygous non-reference sites determined by Adey *et al* (2013) and tabulated the fraction of sites where the non-reference allele was found in our sequencing reads, with the expectation that single HeLa cells should have very high (*i.e.* >=99%) homozygous non-reference alleles at those sites, with reduced fractions indicating contamination by HAP1. For this study, we drew conservative cutoffs of 57% and 99% for each species (*i.e.* any cellular indices falling between these values were discarded).

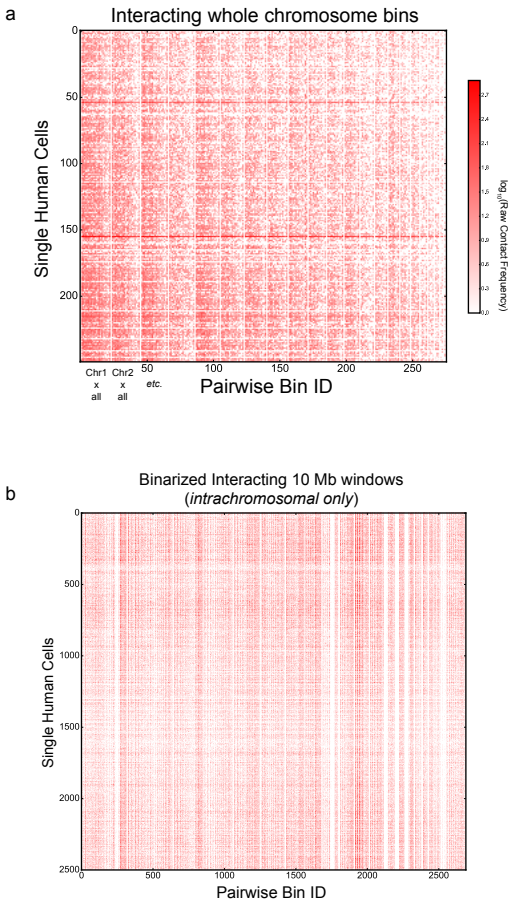


Figure 4.10. Raw single cell matrices used as input for PCA.

To generate these matrices, we took single-cell contact maps and vectorized them, such that each cell is represented by a vector of non-redundant contact counts between two loci. For interchromosomal analysis, each vector contained the $\log_{10}$ transformed number of raw counts between two chromosomes; for intrachromosomal analysis, each vector contained a 1 if a contact between two 10 Mb intrachromosomal windows was observed, and 0 if not. These vectors were then concatenated to form the heatmaps above. The pairwise bin ID simply represents a label for each pair of interacting windows represented in the heat maps. a.) A heat map representation of a portion (250 cells) of the input interchromosomal matrix for PCA. Rows represent single human cells, while columns represent pairwise interactions between two whole chromosomes. For this analysis, raw counts were used, and n = 3,609 cells. b.) Heat map representation of a portion (2,500 cells) of the input intrachromosomal matrix for PCA. Here, interchromosomal counts were ignored, and interaction frequencies between discrete 10 Mb windows genome-wide were reduced to a binary representation (*i.e.* 1 if present, 0 if absent). Again, n = 3,609 cells.

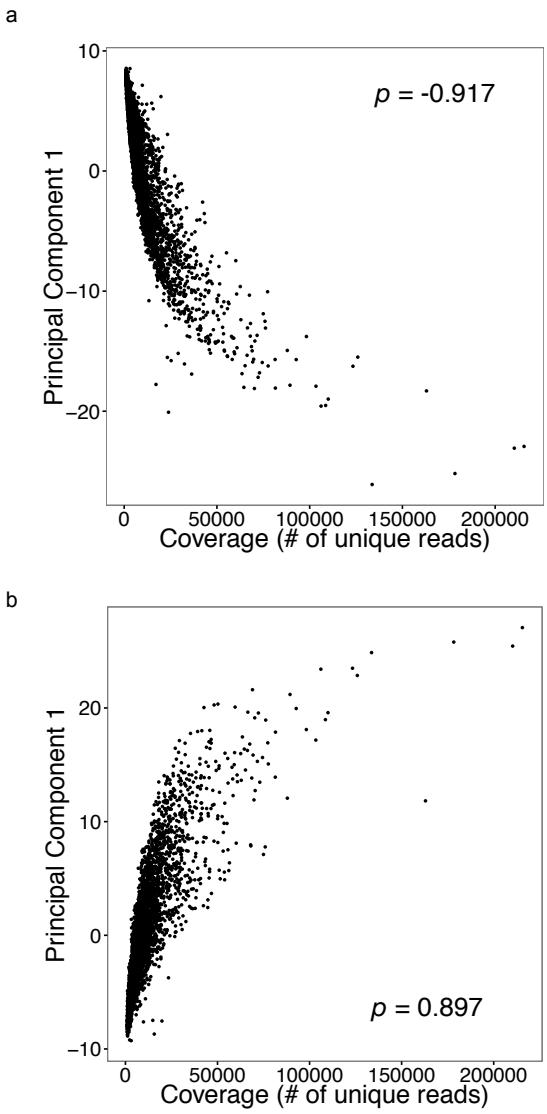


Figure 4.11. The first component of PCA using both interchromosomal contacts and 10 Mb windowed intrachromosomal contacts strongly correlates with coverage.

a.) Correlation between the principal component 1 (PC1) and coverage for interchromosomal interactions (ρ = -0.917). b.) Correlation between the principal component 1 (PC1) and coverage for interacting 10 Mb intrachromosomal windows (ρ = 0.897).

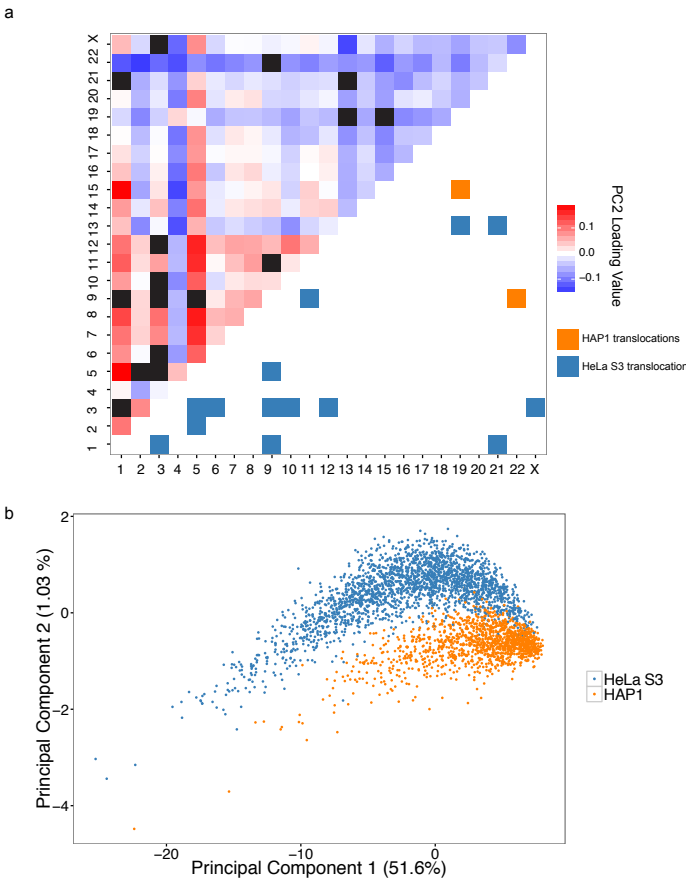


Figure 4.12. Analysis of principal component loadings for interchromosomal separation experiment reveals that translocations contribute to cell type separation in principal component space.

a.) Heat map of loadings for principal component 2 after all known translocations (blacked out entries) are removed from the analysis. b.) After removing all entries corresponding to known translocations from the interchromosomal single-cell Hi-C contact matrix, cell-type separation using PC1 and PC2 is qualitatively worse but still apparent, suggesting that cell-type specific interchromosomal contacts may contribute to the observed separation pattern. Percentages shown are the percentage of variance explained by each plotted PC.

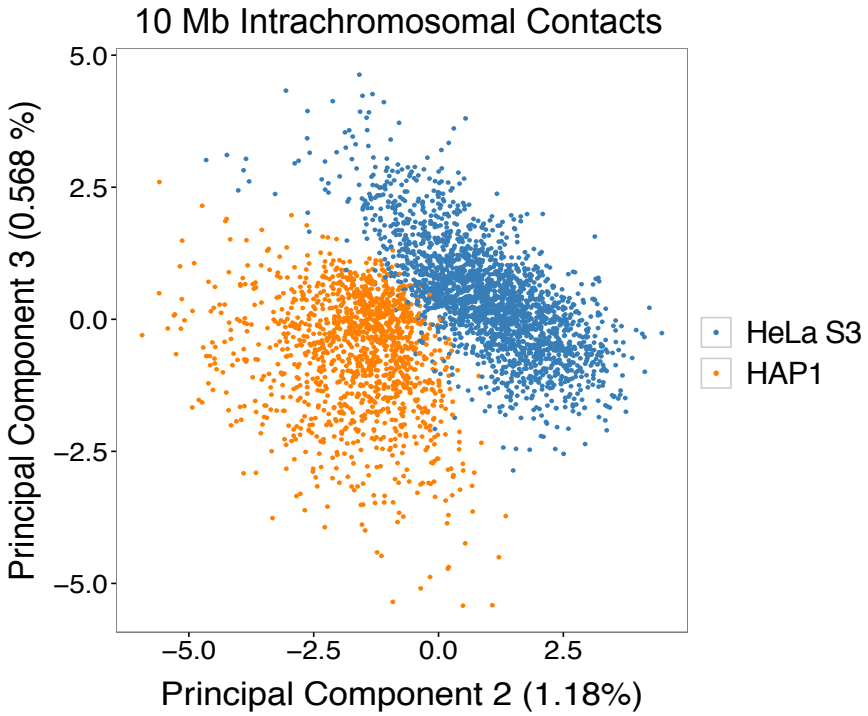


Figure 4.13. PCA using an alternate feature set still enables separation between HAP1 and K562.

Shown is a projection of principal component 2 and principal component 3 from PCA on the intrachromosomal single cell contact matrix (n = 3,609 cells). For this analysis, only intrachromosomal contacts between 10 Mb windows were used. Percentages shown are the percentage of variance explained by each plotted PC.

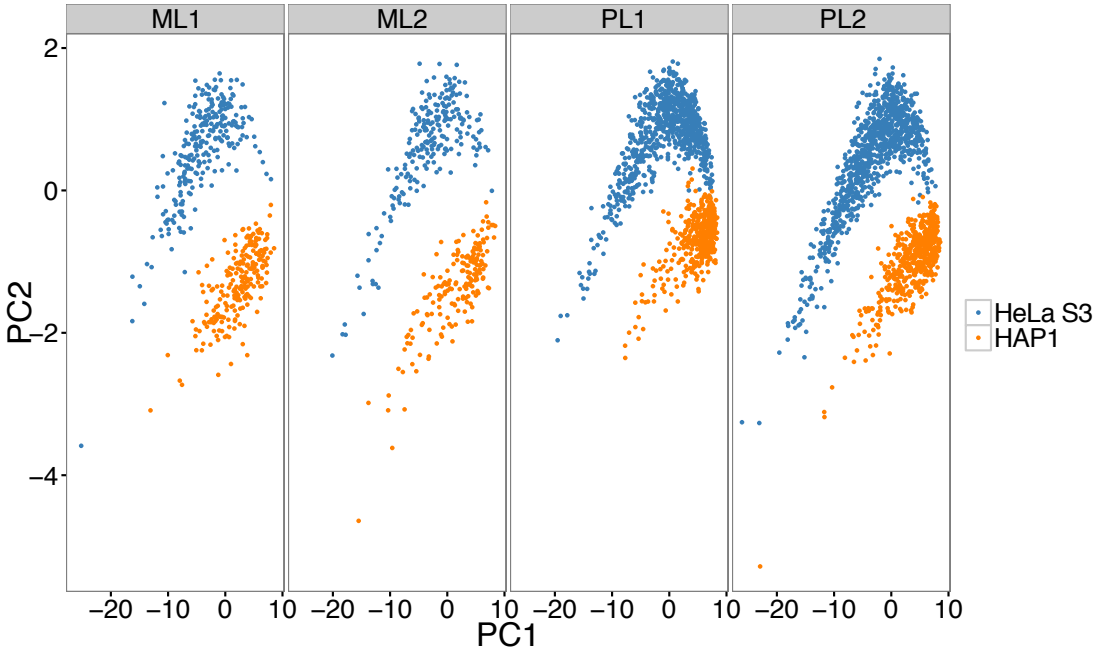


Figure 4.14. Separation of cell types by PCA is consistent across biological replicate combinatorial single cell Hi-C experiments.

Across 4 different libraries, the separation of single HeLa S3 and HAP1 cells is evident, suggesting that this is not simply a technical artifact or batch effect.

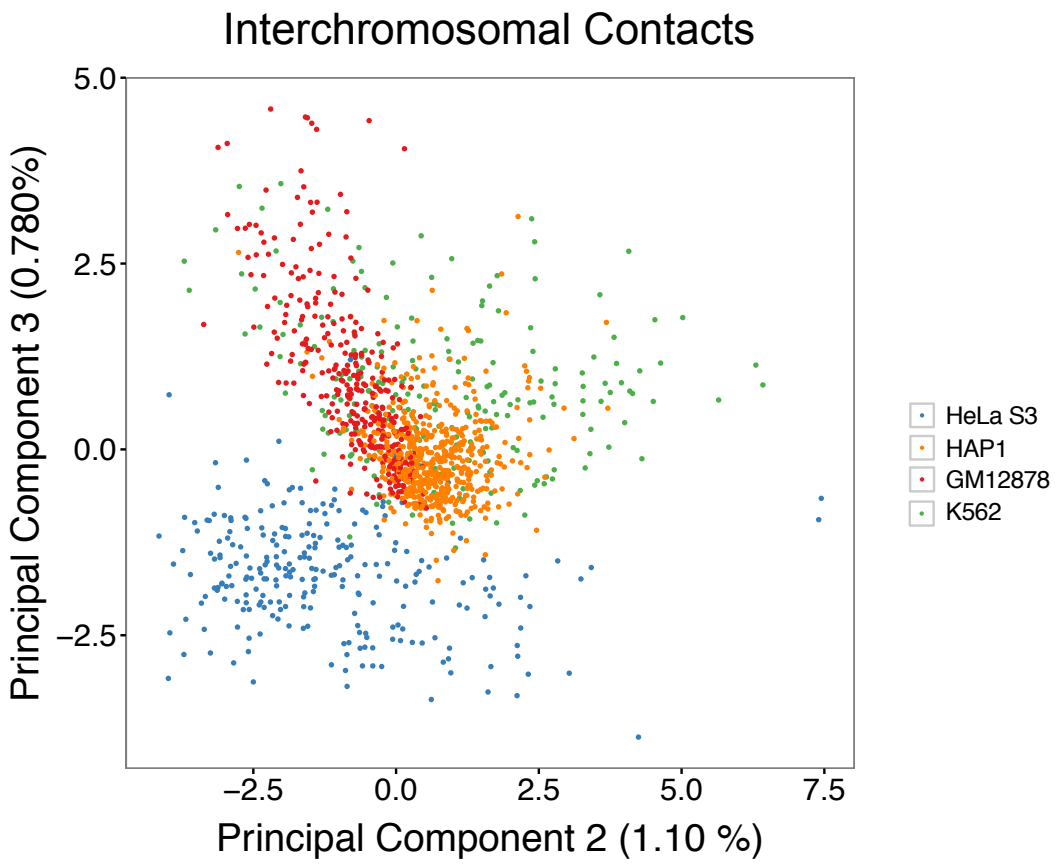


Figure 4.15. PCA of single-cell interchromosomal contacts using cells from 4 different human cell types results in separation of HeLa S3 from other cell lines.

A fifth experiment (Library ML3) containing K562 and GM12878 cells was lightly sequenced and combined with an existing HeLa S3 and HAP1 dataset (Library ML1), resulting in n = 1,394 cells. Projection of single cells onto PC2 and PC3 results in separation of HeLa S3 from the remaining three cell types, but weak separation of K562, GM12878, and HAP1. Percentages shown are the percentage of variance explained by each plotted PC.

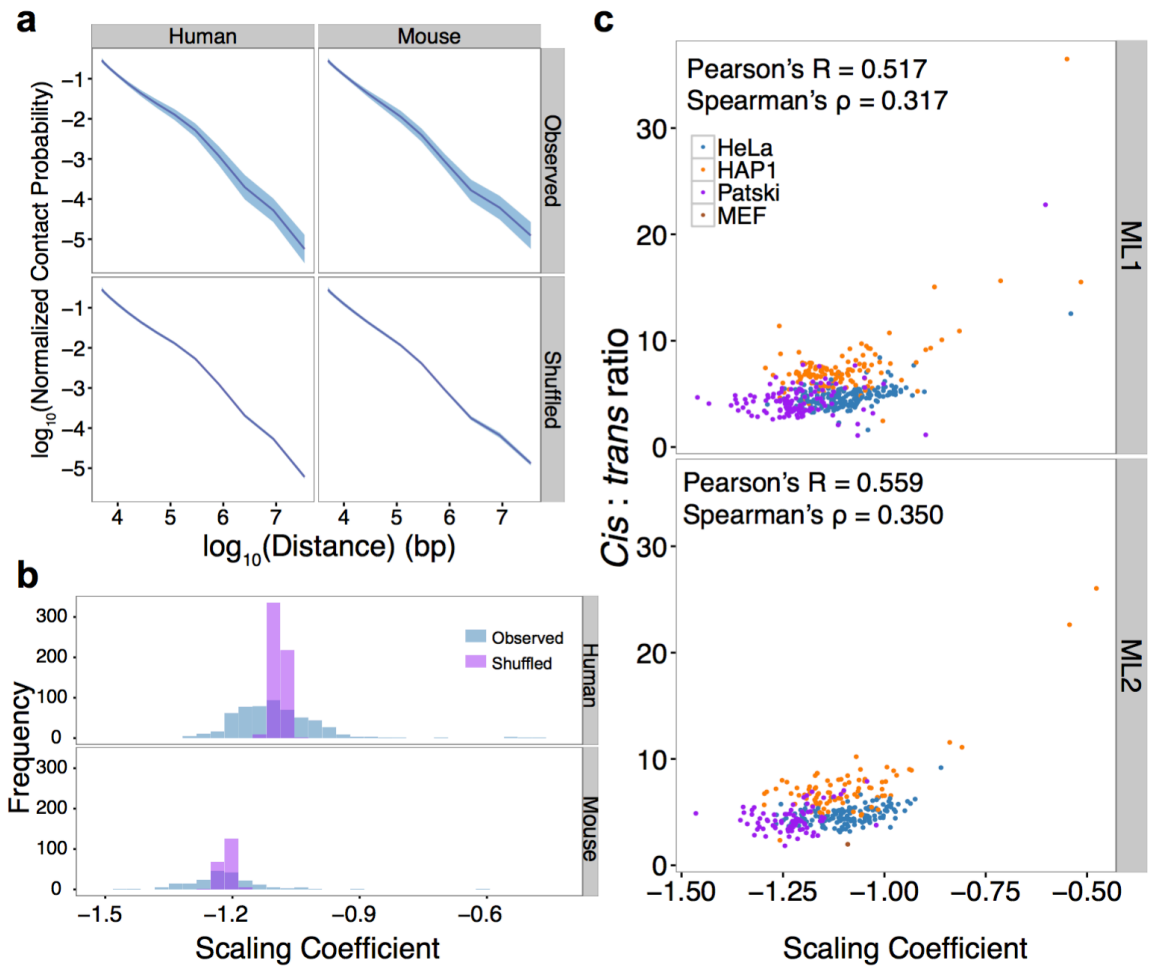


Figure 4.16. Combinatorial single cell Hi-C captures cell-to-cell heterogeneity masked by bulk measurement.

a.) Decay in contact probability for all primary experiment (ML libraries) cells with at least 10,000 unique contacts (n = 769 cells). Plotted is the mean contact probability for each bin (purple), along with standard deviation (blue). Shuffled controls where all cellular index assignments have been randomized demonstrate strikingly lower variance compared to observed single cells, for both mouse and human. b) Scaling coefficients calculated for a.), for distances between 50 kb and 8 Mb. Shuffled controls demonstrate a tighter distribution of coefficients compared to the observed single human cells. c.) Single-cell scaling coefficients reproducibly demonstrate positive correlation with single-cell *cis*:*trans* ratios in both mouse and human cells.

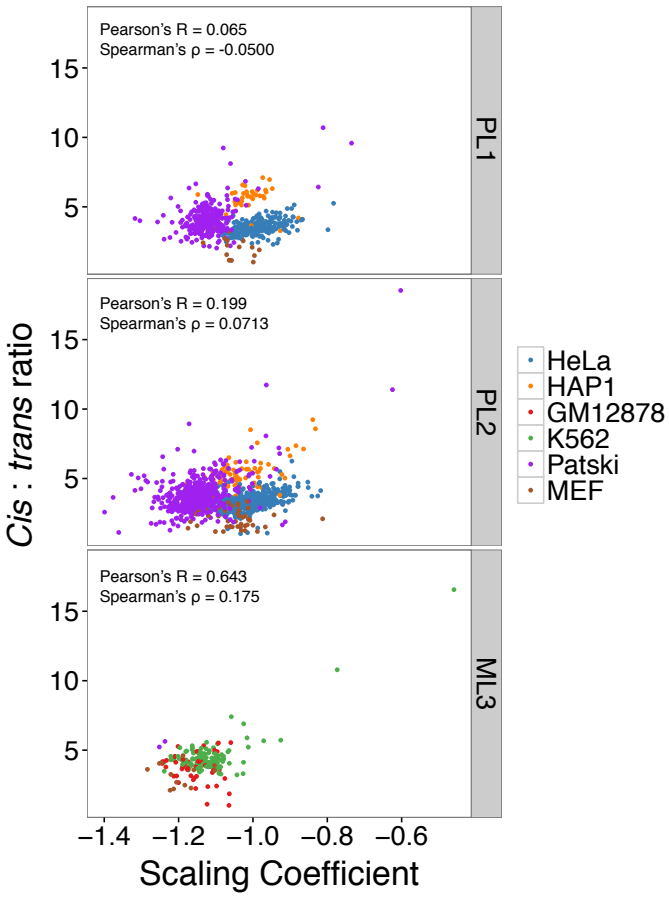


Figure 4.17. Correlation between single cell *cis*:*trans* ratios and single-cell scaling coefficients is reproducible across combinatorial single-cell Hi-C experiments.

We observe a correlation between high *cis*:*trans* ratios and shallow scaling coefficients in both mouse and human cells in both the PL2 (Pearson's R = 0.199; Spearman's ρ = 0.0713) and ML3 (Pearson's R = 0.643; Spearman's ρ = 0.175) experiments. It is possible that the lack of correlation / weaker correlation shown in PL1 (Pearson's R = 0.0649; Spearman's ρ = -0.0500) and PL2, respectively, are a result of shallower sequencing, or sampling (*i.e.* perhaps related to the relative abundance of unsynchronized cells in each phase of the cell cycle).

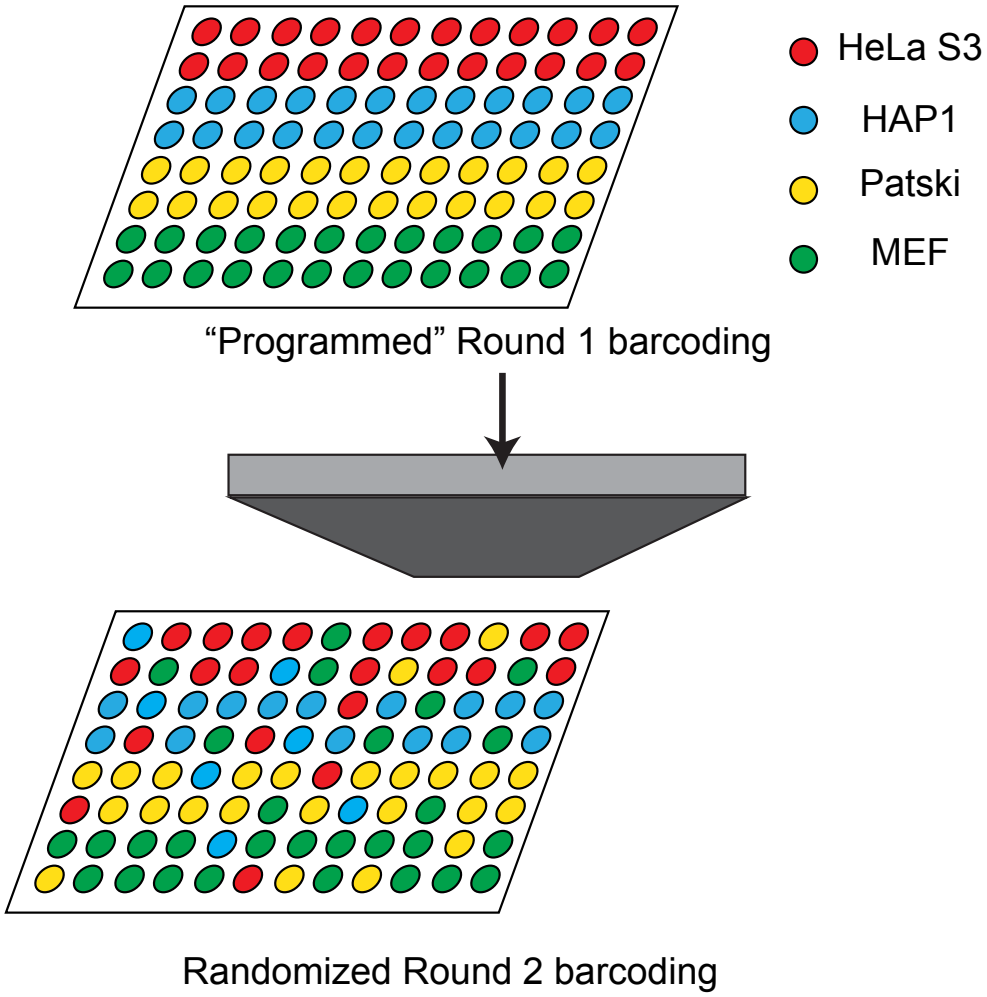


Figure 4.18. “Programmed” barcoding approaches enable association of cell types with unique first round barcodes.

By loading unique cell types into programmed wells during the first round of indexing, we are able to validate cell types in silico. This schematic shows how libraries PL1 and PL2 were generated, wherein only one cell type was present per cell. By contrast, for ML1, ML2 and ML3, subsets of wells contained mixtures of one human and one mouse cell type.

Chapter 5. FULFILLING THE PROMISE OF MASSIVELY PARALLEL MEASUREMENT: *ON THE NEXT STEPS FOR NEXT-GEN*

5.1 ABSTRACT

Over the past decade, massively parallel measurement has become an increasingly important component of biological study. First through surveys of genetic variation, transcript abundance, and "epigenetic" phenomena by oligonucleotide-decorated chips (*i.e.* microarrays), and continuing with the widespread use of massively parallel sequencing for these and manifold other applications, assays employing short read sequencing have vastly increased our knowledge of biological systems. As "third-generation" long-read sequencing and *in situ* sequencing technologies push the field into a new era, we anticipate the continued, creative application of sequencing technologies to elucidating myriad biological phenomena. In this perspective, we review cutting-edge applications for massively parallel DNA-sequencing, briefly reviewing the existing suite of applications using current-generation of massively parallel technologies, while focusing on the next frontiers for high-throughput DNA sequencing and, indeed, nucleic acid sequencing technology in general.

5.2 INTRODUCTION

In the decade following the proof-of-concept of massively parallel sequencing of DNA (Shendure et al., 2005), the "next-generation" sequencer has gained extensive notoriety as a "broadly-enabling microscope" for illuminating biological phenomena (Shendure and Lieberman-Aiden, 2012). First and foremost, the sequencing of billions of DNA molecules in parallel has enabled an explosion in the field of genetics, allowing for analysis of primary sequence variation

across hundreds of thousands of individuals, at a fraction of the cost necessary to generate the first reference of the human genome. In addition to providing these invaluable data, massively parallel sequencing has allowed for unprecedented characterization of largely biochemical phenomena on genome-scale. In addition to the incredible amount of sequencing employed by projects like ENCODE (ENCODE Project Consortium, 2012), FANTOM (FANTOM Consortium et al., 2014), and Roadmap Epigenomics (Sabo et al., 2015) (*e.g.* ChIP-seq, DNase-seq, RNA-seq), much effort has gone towards developing sequencing readouts for myriad biochemical experiments, including nuclease-footprinting (Zentner and Henikoff, 2014), nucleic acid structure mapping (Doudna et al., 2014), polysome profiling (Ingolia et al., 2009), proximity ligation (Dekker et al., 2002), and replication timing (Hansen et al., 2010), to name only a small handful. Taken together, these methodological applications for massively parallel sequencing have enabled "multidimensional cartography" of genomes, providing invaluable, population-scale maps of the extensive genetic and biochemical variation across cell types and species.

In the past decade, our lab has focused on exploiting massively parallel sequencing to dissect biological phenomena at fine-scale. In order to study the functional consequences of nucleotide level polymorphism, we and others introduced the massively parallel reporter assay (MPRA) (Patwardhan et al., 2009), as well as multiplexed assays for variant effect (MAVEs) (Gasperini et al., 2016), generalizable functional assays that survey the molecular consequences of mutated sequences in parallel. In order to measure cellular biochemical phenomena at scale, notably in the context of gene regulation, we and many other groups have developed novel sequencing-based assays to measure how DNA is packaged into chromatin, and how this packaging is linked to transcriptional regulation. We have since further developed such assays to provide information at the cellular level: using a massively parallel combinatorial barcoding

scheme, we have devised means to convert sequencing assays into high-throughput *single-cell* sequencing assays (Cao et al., 2017; Cusanovich et al., 2015; Ramani et al., 2017).

The future of massively parallel measurement lies in two directions: first, in the widespread adoption of large-scale DNA synthesis and DNA sequencing for deconstructing basic biological and biochemical phenomena, and second, in the adoption of third-generation sequencing techniques to provide orthogonal, high-throughput single-molecule biomolecular information. This perspective piece lays out four major directions we anticipate the field moving in over the next decade: 1.) use of saturation mutagenesis and massively parallel reporters as a synthetic biological tool; 2.) expansion of massively parallel single-cell analysis; 3.) development and application of *in situ* sequencing techniques; and 4.) development of methods leveraging third generation long-read sequencing to scalably study phenomena at the single-molecule level.

5.3 HIGH-THROUGHPUT MUTAGENESIS: NOVEL SEQUENCE GRAMMARS, NOVEL SEQUENCE FUNCTIONS

The concept of functionally testing DNA sequences owes its start to the first demonstrations of site-directed and random mutagenesis, tools that have since led to massive leaps in understanding sequence-function relationships in DNA, RNA, and protein. While early attempts at pairing mutagenesis with a functional read-out focused on reading out the consequences of a small number of randomly defined or pre-programmed mutations (Kunkel, 1985; Myers et al., 1986), the advent of polymerase chain reaction (PCR) technology signaled the beginning of high-throughput selections of randomized nucleic acid sequencing libraries using directed evolution schemes (Cadwell and Joyce, 1992). The earliest high-throughput approaches for selecting functional nucleic acid sequences were developed in parallel by the Szostak, Gold, and Weintraub labs, and encompass a class of techniques now commonly referred to as Systematic Evolution of

Ligands by Exponential Enrichment, or SELEX (Blackwell and Weintraub, 1990; Ellington and Szostak, 1990; Tuerk and Gold, 1990). In a typical SELEX experiment, a large array of completely random nucleic acid sequences is subjected to some sort of selection, often an affinity selection against a protein or ligand of interest. This selection procedure is carried out for multiple rounds, with the selection products of one round being subjected to amplification, and that amplified pool being used as a selection library for subsequent rounds. SELEX and similar techniques have been routinely used to select for both RNA and DNA sequences, and are commonly used to define the *in vitro* specificities of RNA- and DNA-binding proteins (Jolma et al., 2013; Ray et al., 2009).

The advent of high throughput DNA synthesis (Kosuri and Church, 2014) and sequencing has made these selection approaches even more powerful, has allowed such experiments to be carried out *in vivo*, and has enabled the testing of large pre-programmed sequence libraries. The massively parallel reporter assay has most famously been used to characterize *cis*-regulatory sequences like promoters and enhancers (Inoue et al., 2017; Patwardhan et al., 2012). In these approaches, thousands to hundreds of thousands of regulatory elements are cloned upstream of a reporter gene fused to a defined DNA sequence barcode. These cloned elements are introduced into cells through transient transfection or lentiviral transduction, and then RNA and DNA from the cells are subjected to deep sequencing. By quantifying the relative abundance of specific barcodes in the RNA pool, and normalizing this value against barcode abundance in the DNA pool, one can, in a single experiment, measure the activities of a large number of regulatory elements in parallel. Importantly, these assays are diverse with respect to the types of libraries that can be screened. MPRAs have been carried out using sheared genomic DNA as a library (Arnold et al., 2013; 2017; van Arensbergen et al., 2017), and specific loci have also been subjected to saturating mutagenesis (Inoue et al., 2017; Patwardhan et al., 2012). Thanks to the versatility of DNA

synthesis technologies, large numbers of programmed sequence "grammars" can also be tested; several studies have leveraged this to test relationships between transcription factor binding site ordering, spacing and *cis*-regulatory sequence activity (Farley et al., 2015; Smith et al., 2013). Finally, the massively parallel reporter paradigm is generalizable. While most published reporter assays have focused on decoding *cis*-regulation of transcription by promoter and enhancer elements, massively parallel assays have also been employed to study the primary sequence elements responsible diverse other biological phenomena, including post-transcriptional regulation (Shalem et al., 2013), mRNA splicing (Rosenberg et al., 2015), and *in vivo* protein-DNA binding (Grossman et al., 2017).

What might the future hold for massively parallel reporter assays? We first envision such assays broadening in scope in terms of the diversity of biochemical phenomena studied using the technique; in the near future, massively parallel assays will be developed to discover sequence elements responsible for intracellular RNA localization, sequences that lead to specific localization of DNA elements in three-dimensions (*e.g.* sequences that localize to phase-separated compartments), and primary sequence determinants of basic transcriptional mechanistic properties, such as polymerase initiation and elongation rates. Next, we anticipate a growing interest in using massively parallel reporter assays for synthetic biological purposes. Endogenous cellular activities are exquisitely regulated by primary sequence elements, many of which remain uncharacterized. We envision massively parallel reporter assays serving as tools for discovering novel control elements for engineering specific cellular behaviors. For example, one might employ MPRA to discover a suite of ligand-responsive *cis*-regulatory elements, in the process developing transcriptional ligand sensors that can be paired with high-throughput sequencing techniques.

5.4 HIGH-THROUGHPUT SINGLE-CELL 'OMICS: SENDING SINGLE CELLS TO THE SEQUENCER

A limitation common to all genomic techniques is that they inherently average the biomolecular state of all input cells. Though biochemical measurements (*e.g.* cellular transcriptomes, chromatin accessibility, chromosome conformation) are all made accepting this critical caveat, cellular averages are not always accurate. Cells found *in vivo* tissues and organ systems are often heterogeneous—this heterogeneity can arise from a multitude of sources, including cellular environment, developmental cues, or simply due to cellular stochasticity. Importantly, this is true of even pure monocultures of actively dividing cells; in the absence of chemical synchronization of the cell cycle in these populations, each cell can occupy a different sector in the cell cycle, over which distinct regulatory changes occur at the levels of chromosomal architecture, chromatin accessibility, transcription, and more. To better assess biological heterogeneity in a high-throughput manner, biologists have turned to single-cell analyses. The earliest examples of high-throughput single-cell analysis were carried out using fluorescence activated cell sorting (FACS), a fluidic platform that enables measurement of critical cellular parameters (*e.g.* cell size, cell density), along with cellular fluorescence (Bonner et al., 1972). While modern flow-sorting paradigms enable robust multi-parameter classification of cells, the field has also moved towards parametrizing cells at the level of genome-wide phenomena using single-cell genomic approaches (Schwartzman and Tanay, 2015). Single-cell genomics offers the promise of deconvolving heterogeneous populations, by stratifying genome-wide biochemical measurements by a common unit—a single cell or nucleus. The earliest single cell analyses of RNA and DNA employed high-gain amplification techniques to amplify picograms of nucleic acid for high-throughput sequencing (Gawad et al., 2016; Wu et al., 2014). These approaches, while

technically impressive, were critically limited in scalability, in that single-cells were isolated. Recent technological advances, however, have remedied this limitation; first using microfluidic technology (Gole et al., 2013; Klein et al., 2015; Macosko et al., 2015), and now using *in situ* nucleic acid barcoding of cells (Cao et al., 2017; Cusanovich et al., 2015; Ramani et al., 2017), a diverse number of high-throughput single-cell assays are now available.

The current state of single-cell research has made the generation of large-scale cellular atlases a major goal, with efforts already underway to build worm, fly, mouse, and, ultimately, human cellular atlases that comprehensively molecularly characterize and classify the cells that comprise complex organisms. But following the completion of these undeniably valuable projects, where will single-cell sequencing move, at least from a technological viewpoint? We believe that the answer lies in characterizing the consequences of molecular perturbations in cells and in organisms.

Tools for cellular perturbation are now well-characterized. CRISPR/Cas9 and auxin-based protein degradation systems offer genetic, transcriptional, and proteomic avenues for modulating the cellular concentration of any gene or set of genes. The natural progression of these approaches synthesis of single-cell genomic and molecular perturbative techniques, to scalably characterize the molecular consequences of perturbing single genes, and combinations of genes. Already, methods exist to pair Cas9-based perturbation with single-cell RNA-seq, to characterize transcriptomic consequences of CRISPR-mediated knock-out, activation, and repression of genes (Adamson et al., 2016; Datlinger et al., 2017). Future work will involve devising assays that pair molecular perturbations with diverse other single-cell genomic assays. Eventually these approaches may be paired with cell lineaging technology (McKenna et al., 2016), to

comprehensively determine the consequences of molecular perturbations on cellular activity and ultimately, developmental and disease phenotypes.

5.5 SEEING IS BELIEVING: BRINGING THE SEQUENCER TO SINGLE CELLS

Though powerful, massively parallel sequencing-based approaches to single-cell biology carry a critical limitation—an inability to resolve subcellular phenomena. As our knowledge of cellular biology has increased, so too has our understanding of the complexity of cytoplasmic and nuclear organization. *In situ* techniques enable the direct visualization of cellular phenomena, allowing for simultaneous quantification and visualization of biological processes. As *in situ* technologies advance, we foresee a shift towards literally bringing the high-throughput sequencer into cells, to directly sequence and characterize cellular nucleic acids in their native environments.

A multitude of *in situ* hybridization and sequencing approaches have recently been developed to image transcriptomes in single-cells at scale. While several approaches have succeeded in multiplexing fluorescence *in situ* hybridization (FISH) to quantify many RNA transcripts in parallel (Chen et al., 2015; Lubeck et al., 2014), or have succeeded in carrying out targeted sequencing of transcripts *in situ* (Ke et al., 2013), only one approach has succeeded in sequencing subcellular transcriptomes—fluorescence *in situ* sequencing, or FISSEQ (Lee et al., 2014). FISSEQ employs *in situ* reverse transcription and cDNA amplification, followed by sequencing by ligation, to effectively convert cells themselves into a platform for carrying out high-throughput RNA sequencing. While FISSEQ has thus far only been demonstrated for *in situ* RNA sequencing, the platform raises intriguing possibilities for carrying out other high-throughput sequencing assays into cellular environments. By pairing immunofluorescence with high-throughput *in situ* sequencing of DNA, for example, one might be able to localize histone modifications and transcription factor locations *in situ*. By tagging actively translating ribosomes

and simultaneously sequencing RNA, one might be able to carry out *in situ* ribosome profiling. Given the adaptability of the sequencing by ligation approach, it is not inconceivable that other high-throughput assays may be carried out *in situ*. Recent work has demonstrated single molecule interaction sequencing (SMI-seq), an assay that also adapts sequencing by ligation to carry out massively parallel binding affinity analyses of protein-protein interactions (Gu et al., 2014). While this assay has thus far only been used *in vitro*, the possibility of carrying out such experiments within the confines of an intact cell is certainly exciting.

5.6 HIGH-THROUGHPUT LONG-READ SEQUENCING: A NEW FRONTIER FOR TECHNOLOGY DEVELOPMENT

A major limitation to any massively parallel sequencing-based technique is the inherently limited size of sequence-able molecules, as the "bridge PCR" necessary for carrying out Illumina sequencing necessitates that molecules must be shorter than ~800 bases. This limitation can be problematic, particularly when studying gene regulation. Nucleosomes and transcription factor (TF) complexes, proteins critical to transcriptional regulatory programs, protect anywhere from 10 to 170 bases of DNA, and are known to be positioned in cell-type specific manners enabling the activation and repression of specific transcriptional programs (Zentner and Henikoff, 2014). Importantly, the positions of nucleosomes with respect to one another (*i.e.* the "phasing" of nucleosomes), and the positions of TFs themselves, are thought to be intertwined—the assembly of a "competent" locus for productive transcription is thought to require opening of chromatin via movement of nucleosomes by chromatin remodelers recruited by TFs (Spitz and Furlong, 2012). Unfortunately, massively parallel sequencing based approaches are poorly equipped to characterize how such biological interactions may occur on single chromatin templates *in vivo*. Typical sequencing assays used to position nucleosomes genome-wide rely on deeply sequencing

~150 base-pair fragments—the maps resulting from these data are thus averaged not only across millions of input cells, but also over the single-template positions of the ~20E6 nucleosomes necessary to package a three-billion base eukaryotic genome. This is true for transcription factor mapping strategies as well, which fail to characterize all TF-DNA interactions that may be occurring simultaneously on the same single chromatin template. A similar issue can be found in the field of RNA sequencing, specifically in the mapping of actively translating ribosome (*i.e.* ribosome profiling). While it is well understood that mRNAs differ widely in the abundance of ribosomes on a single translated molecule (Floor and Doudna, 2016), sequencing approaches to-date have been largely unable to resolve ribosome positions at the level of single mRNA molecules.

How might one biochemically record the positions of nucleosomes and transcription factors on a long template of DNA? Could one measure the positions of all actively translating ribosomes across an entire transcriptome? While biochemical work will of course be necessary to mark these diverse phenomena *in vivo*, the third-generation sequencer presents an ideal tool for *reading out* this information.

Third-generation sequencers include single molecule real time sequencers (*i.e* Pacific BioSciences sequencing) and nanopore sequencing (*i.e.* Oxford Nanopore sequencing), and offer the potential of high-throughput, *long read* sequencing (Eid et al., 2009; Laszlo et al., 2014). In the Pacific BioSciences platform, single molecules of DNA are trapped in a large array of engineered nanowells termed zero-mode waveguides (ZMWs), each of which contain a highly processive, engineered variant of the Phi29 DNA polymerase. Polymerization is carried out in the presence of fluorescently modified nucleotides, whose fluor groups are cleaved upon incorporation of the nucleotide into the synthesized strand. ZMWs are specifically designed to sensitively capture the fluorescence of these liberated fluors, and movies capturing fluorescence discharge

over the course of polymerization in each of hundreds of thousands of wells are taken. Pacific Biosciences sequencing provides incredibly long read lengths, with median read lengths on the order of tens of kilobases (Chaisson et al., 2015).

Nanopore sequencers differ substantially from single-molecule real time sequencers, but also achieve the goal of ascertaining long DNA sequencing reads. Nanopore sequencers are generally comprised of two components, a pore protein embedded in a lipid bilayer, through which nucleic acids may be translocated, and a motor protein, which threads the nucleic acid through the pore in a controlled fashion. Bases are called by first establishing a current through the nanopore. As different sets of bases translocate through the pore, the current through the pore changes in characteristic ways. High-throughput is achieved by simply multiplexing many individually addressable pores in a device, and running DNA through the pore for longer amounts of time. Unpublished work has suggested that nanopore sequencing can directly sequence RNA, and also suggests that it may discriminate between modified and unmodified RNA bases. Furthermore, nanopores have already found a use in the field of biophysics, being used to measure the kinetics of a DNA helicase, Hel308, translocating on DNA (Derrington et al., 2015). The diversity of potential high-throughput biophysical and biochemical assays that use nanopore sequencing as a readout is enticing, and we eagerly anticipate the first wave of genomic technologies that take advantage of this novel platform.

5.7 Closing Remarks

Massively parallel sequencing has fundamentally altered the way in which we study and analyze biological phenomena. Through massively parallel reporter assays and massively multiplex single-cell assays, the field is moving quickly towards answering pressing questions in the fields of gene regulation, cell biology, human genetics, and beyond. We eagerly anticipate

future methodological development, which will ultimately dissect diverse biological phenomena through reporter approaches, characterize the molecular consequences of cellular perturbations, and ultimately, quantify biochemical phenomena at the subcellular and single-molecular levels.

REFERENCES

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell *167*, 1867–1882.e21.

Alipour, E., and Marko, J.F. (2012). Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Research *40*, 11202–11212.

Anger, A.M., Armache, J.-P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D.N., and Beckmann, R. (2013). Structures of the human and Drosophila 80S ribosome. *497*, 80–85.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Arnold, C.D., Zabidi, M.A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., and Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. Nat Biotechnol *35*, 136–144.

Ay, F., and Noble, W.S. (2015). Analysis methods for studying the 3D architecture of the genome. Genome Biol. *16*, 183.

Ay, F., Bunnik, E.M., Varoquaux, N., Bol, S.M., Prudhomme, J., Vert, J.P., Noble, W.S., and Le Roch, K.G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Research *24*, 974–988.

BARR, M.L., and BERTRAM, E.G. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature *163*, 676.

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *98*, 387–396.

Ben-Shem, A., Ben-Shem, A., Garreau de Loubresse, N., de Loubresse, N.G., Melnikov, S., Melnikov, S., Jenner, L., Jenner, L., Yusupova, G., Yusupova, G., et al. (2011). The Structure of the Eukaryotic Ribosome at 3.0 A Resolution. Science *334*, 1524–1529.

Blackwell, T.K., and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science *250*, 1104–1110.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.-T., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Bonner, W.A., Hulett, H.R., Sweet, R.G., and Herzenberg, L.A. (1972). Fluorescence activated

cell sorting. Rev Sci Instrum *43*, 404–409.

Bouwman, B.A., and de Laat, W. (2015). Getting the genome in shape: the formation of loops, domains and compartments. Genome Biol. *16*, 154.

Boyle, S., Bridger, J.M., Mahy, N.L., and Bickmore, W.A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet *10*, 211–219.

Branco, M.R., Branco, T., Ramirez, F., and Pombo, A. (2008). Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. Chromosome Res. *16*, 413–426.

Buckley, M.S., and Lis, J.T. (2014). Imaging RNA Polymerase II transcription sites in living cells. Curr. Opin. Genet. Dev. *25*, 126–130.

Burge, S.W., Eddy, S.R., Burge, S.W., Daub, J., Daub, J., Eberhardt, R., Eberhardt, R., Tate, J., Tate, J., Barquist, L., et al. (2012). Rfam 11.0: 10 years of RNA families. Nucleic Acids Research *41*, D226–D232.

Burton, J.N., Liachko, I., Dunham, M.J., and Shendure, J. (2014). Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 (Bethesda) *4*, 1339–1346.

Cadwell, R.C., and Joyce, G.F. (1992). Randomization of genes by PCR mutagenesis. PCR Methods Appl. *2*, 28–33.

Callaway, E. (2015). The revolution will not be crystallized: a new method sweeps through structural biology. Nature *525*, 172–174.

Cameron, V., and Uhlenbeck, O.C. (1977). 3'-Phosphatase activity in T4 polynucleotide kinase. Biochemistry *16*, 5120.

Cannone, J., Subramanian, S., Schnare, M., Collett, J., D'Souza, L., Du, Y., Feng, B., Lin, N., Madabusi, L., Muller, K., et al. (2002). The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics *3*, 2.

Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667.

Cate, J.H., Cech, T.R., Doudna, J.A., Gooding, A.R., Cate, J.H., Gooding, A.R., Podell, E., Podell, E., Zhou, K., Zhou, K., et al. (1996). Crystal Structure of a Group I Ribozyme Domain:

Principles of RNA Packing. Science *273*, 1678–1685.

Chaisson, M.J.P., Wilson, R.K., and Eichler, E.E. (2015). Genetic variation and the de novo assembly of human genomes. Nat Rev Genet *16*, 627–640.

Chen, J., Zhang, Z., Li, L., Chen, B.-C., Revyakin, A., Hajj, B., Legant, W., Dahan, M., Lionnet, T., Betzig, E., et al. (2014). Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. Cell *156*, 1274–1285.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090–aaa6090.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Molecular Cell *44*, 667–678.

Ciabrelli, F., and Cavalli, G. (2015). Chromatin-driven behavior of topologically associating domains. J Mol Biol *427*, 608–625.

Cisse, I.I., Izeddin, I., Causse, S.Z., Boudarene, L., Senecal, A., Muresan, L., Dugast-Darzacq, C., Hajj, B., Dahan, M., and Darzacq, X. (2013). Real-time dynamics of RNA polymerase II clustering in live human cells. Science *341*, 664–667.

Cook, P.R. (2010). A model for all genomes: the role of transcription factories. J Mol Biol *395*, 1–10.

Cook, P.R., and Brazell, I.A. (1975). Supercoils in human DNA. J Cell Sci *19*, 261–279.

Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244.

Cremer, M., Grasser, F., Lanctot, C., Muller, S., Neusser, M., Zinner, R., Solovei, I., and Cremer, T. (2008). Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. Methods in Molecular Biology *463*, 205–239.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet *2*, 292–301.

Cremer, T., and Cremer, M. (2010). Chromosome territories. Cold Spring Harb Perspect Biol *2*, a003889–a003889.

Cremer, T., Cremer, M., Hubner, B., Strickfaden, H., Smeets, D., Popken, J., Sterr, M., Markaki, Y., Rippe, K., and Cremer, C. (2015). The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. FEBS Letters *589*, 2931–2943.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell *160*, 191–203.

Croft, J.A., Bridger, J.M., Boyle, S., Perry, P., Teague, P., and Bickmore, W.A. (1999). Differences in the localization and morphology of chromosomes in the human nucleus. J. Cell Biol. *145*, 1119–1131.

Cullen, K.E., Cullen, K., Kladde, M.P., Kladde, M., Seyfred, M.A., and Seyfred, M. (1993). Science *261*.

Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914.

Daban, J.R. (2011). Electron microscopy and atomic force microscopy studies of chromatin and metaphase chromosome structure. Micron *42*, 733–750.

Das, R., Cordero, P., Lucks, J.B., Cordero, P., Lucks, J.B., and Das, R. (2012). An RNA Mapping DataBase for curating RNA structure mapping experiments. Bioinformatics *28*, 3006–3008.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat Meth *14*, 297–301.

de Wit, E., Vos, E.S., Holwerda, S.J., Valdes-Quezada, C., Verstegen, M.J., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H., and de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. Molecular Cell *60*, 676–684.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes Dev. *26*, 11–24.

de Wit, E., de Laat, W., Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., and van Steensel, B. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *38*, 1348–1354.

Dekker, J., and Heard, E. (2015). Structural and functional diversity of Topologically Associating Domains. FEBS Letters *589*, 2877–2884.

Dekker, J., and Misteli, T. (2015). Long-Range Chromatin Interactions. Cold Spring Harb Perspect Biol *7*, a019356.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Deng, W., and Blobel, G.A. (2014). Manipulating nuclear architecture. Curr. Opin. Genet. Dev. *25*, 1–7.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell *149*, 1233–1244.

Deng, W., Rupon, J.W., Krivega, I., Breda, L., Motta, I., Jahn, K.S., Reik, A., Gregory, P.D., Rivella, S., Dean, A., et al. (2014). Reactivation of developmentally silenced globin genes by forced chromatin looping. Cell *158*, 849–860.

Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., Berletch, J.B., Blau, C.A., Shendure, J., Duan, Z., et al. (2015). Bipartite structure of the inactive mouse X chromosome. Genome Biol. *16*, 152.

Derrington, I.M., Craig, J.M., Stava, E., Laszlo, A.H., Ross, B.C., Brinkerhoff, H., Nova, I.C., Doering, K., Tickman, B.I., Ronaghi, M., et al. (2015). Subangstrom single-molecule measurements of motor proteins using a nanopore. Nat Biotechnol *33*, 1073–1075.

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., and Assmann, S.M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *505*, 696–700.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dobin, A., Dobin, A., Davis, C.A., Davis, C.A., Schlesinger, F., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., et al. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Research *16*, 1299–1309.

Doudna, J.A., Mortimer, S.A., and Kidwell, M.A. (2014). Insights into RNA structure and function from genome-wide studies. *15*, 469–479.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Duan, Z., and Blau, C.A. (2012). The genome in space and time: does form always follow function? How does the spatial and temporal organization of a eukaryotic genome reflect and influence its functions? Bioessays *34*, 800–810.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363–367.

Dundr, M., and Misteli, T. (2010). Biogenesis of nuclear bodies. Cold Spring Harb Perspect Biol *2*, a000711–a000711.

Eagen, K.P., Hartl, T.A., and Kornberg, R.D. (2015). Stable Chromosome Condensation Revealed by Chromosome Conformation Capture. Cell *163*, 934–946.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973–1237973.

Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M., and Lander, E.S. (2014). RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell *159*, 188–199.

Eskiw, C.H., Cope, N.F., Clay, I., Schoenfelder, S., Nagano, T., and Fraser, P. (2010). Transcription factories and nuclear organization of the genome. Cold Spring Harbor Symposia on Quantitative Biology *75*, 501–506.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470.

Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., and Levine, M.S. (2015). Suboptimization of developmental enhancers. Science *350*, 325–328.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448–453.

Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014). Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Molecular Cell *55*, 694–707.

Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.

Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet *4*, e1000039.

Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suva, M.L., and Bernstein, B.E. (2015). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114.

Floor, S.N., and Doudna, J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. eLife *5*, 1276.

Fortin, J.P., and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biol. *16*, 180.

Fraser, J., Williamson, I., Bickmore, W.A., and Dostie, J. (2015). An Overview of Genome Organization and How We Got There: from FISH to Hi-C. Microbiology and Molecular Biology Reviews : MMBR *79*, 347–372.

Fraser, P., and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. Nature *447*, 413–417.

Fredriksson, S., Gullberg, M., Jarvius, J., Olsson, C., Pietras, K., Gustafsdottir, S.M., Ostman, A., and Landegren, U. (2002). Protein detection using proximity-dependent DNA ligation assays. *20*, 473–477.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Fussner, E., Strauss, M., Djuric, U., Li, R., Ahmed, K., Hart, M., Ellis, J., and Bazett-Jones, D.P. (2012). Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. EMBO Rep. *13*, 992–996.

Gaj, T., Gersbach, C.A., and Barbas, C.F.3. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends in Biotechnology *31*, 397–405.

Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nat Protoc *11*, 1782–1787.

Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, I.I., Iarovaia, O.V., and Razin, S.V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Research *41*, 3563–3575.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nat Rev Genet *17*, 175–188.

Gerlich, D., Beaudouin, J., Kalbfuss, B., Daigle, N., Eils, R., and Ellenberg, J. (2003). Global chromosome positions are transmitted through mitosis in mammalian cells. Cell *112*, 751–764.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Gibcus, J.H., and Dekker, J. (2013). The hierarchy of the 3D genome. Molecular Cell *49*, 773–782.

Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell *157*, 950–963.

Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., et al. (2016). Structural organization of the inactive X chromosome in the mouse. Nature *535*, 575–579.

Gole, J., Gore, A., Richards, A., Chiu, Y.-J., Fung, H.-L., Bushman, D., Chiang, H.-I., Chun, J., Lo, Y.-H., and Zhang, K. (2013). Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nat Biotechnol *31*, 1126–1132.

Gonzalez-Sandoval, A., Towbin, B.D., Kalck, V., Cabianca, D.S., Gaidatzis, D., Hauer, M.H., Geng, L., Wang, L., Yang, T., Wang, X., et al. (2015). Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. Cell *163*, 1333–1347.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell *14*, 762–775.

Grasser, F., Neusser, M., Fiegler, H., Thormeyer, T., Cremer, M., Carter, N.P., Cremer, T., and Müller, S. (2008). Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. J Cell Sci *121*, 1876–1886.

Griffith, J., Hochschild, A., and Ptashne, M. (1986). DNA loops induced by cooperative binding of [lambda] repressor. *322*, 750–752.

Grob, S., Schmid, M.W., and Grossniklaus, U. (2014). Hi-C analysis in Arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of Drosophila. Molecular Cell *55*, 678–693.

Grossman, S.R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B.E., et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc. Natl. Acad. Sci. U.S.a. *114*, E1291–E1300.

Grosswendt, S., Filipchyk, A., Manzano, M., Klironomos, F., Schilling, M., Herzog, M., Gottwein, E., and Rajewsky, N. (2014). Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. *54*, 1042–1054.

Gu, L., Li, C., Aach, J., Hill, D.E., Vidal, M., and Church, G.M. (2014). Multiplex single-

molecule interaction profiling of DNA-barcoded proteins. Nature *515*, 554–557.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R., et al. (2014). Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat Struct Mol Biol *21*, 198–206.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C., Ping, J.L.H., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat Genet *43*, 630–638.

Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R., and Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. eLife *6*, 2848.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. U.S.a. *107*, 139–144.

He, H.H., Meyer, C.A., Hu, S.S., Chen, M.-W., Zang, C., Liu, Y., Rao, P.K., Fei, T., Xu, H., Long, H., et al. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nat Meth *11*, 73–78.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell *153*, 654–665.

Heride, C., Ricoul, M., Kieu, K., Hase, von, J., Guillemot, V., Cremer, C., Dubrana, K., and Sabatier, L. (2010). Distance between homologous chromosomes results from chromosome positioning constraints. J Cell Sci *123*, 4063–4075.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol *33*, 510–517.

Hochstrasser, M., and Sedat, J.W. (1987). Three-dimensional organization of Drosophila melanogaster interphase nuclei. I. Tissue-specific aspects of polytene nuclear architecture. J. Cell Biol. *104*, 1455–1470.

Horike, S.-I., Cai, S., Miyano, M., Cheng, J.-F., and Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet *37*, 31–

40.

Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Molecular Cell *48*, 471–484.

Hsieh, T.H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108–119.

Huang, B., Babcock, H., and Zhuang, X. (2010). Breaking the diffraction barrier: super-resolution imaging of cells. Cell *143*, 1047–1058.

Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet *46*, 205–212.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Meth *9*, 999–1003.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Research *27*, 38–52.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074–1080.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2015). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262–275.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Joti, Y., Hikima, T., Nishino, Y., Kamada, F., Hihara, S., Takata, H., Ishikawa, T., and Maeshima, K. (2012). Chromosomes without a 30-nm chromatin fiber. Nucleus *3*, 404–410.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol *30*, 90–98.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. Nat Meth *10*, 857–860.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M., et al. (2015). Genome-wide maps of nuclear lamina interactions in single human cells. Cell *163*, 134–147.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. Cell *153*, 178–192.

Kladwang, W., VanLang, C.C., Cordero, P., and Das, R. (2011). A two-dimensional mutate-and-map strategy for non-coding RNA structure. Nat Chem *3*, 954–962.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kosak, S.T., and Groudine, M. (2004). Form follows function: The genomic organization of cellular differentiation. Genes Dev. *18*, 1371–1384.

Kosuri, S., and Church, G.M. (2014). Large-scale de novo DNA synthesis: technologies and applications. Nat Meth *11*, 499–507.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. Proc. Natl. Acad. Sci. U.S.a. *108*, 10010–10015.

Kunkel, T.A. (1985). Rapid and efficient site-specific mutagenesis without phenotypic selection. Proc. Natl. Acad. Sci. U.S.a. *82*, 488–492.

Lakadamyali, M., and Cosma, M.P. (2015). Advanced microscopy methods for visualizing chromatin structure. FEBS Letters *589*, 3023–3030.

Langer-Safer, P.R., Levine, M., and Ward, D.C. (1982). Immunological method for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. U.S.a. *79*, 4381–4385.

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature *547*, 236–240.

Laszlo, A.H., Derrington, I.M., Ross, B.C., Brinkerhoff, H., Adey, A., Nova, I.C., Craig, J.M., Langford, K.W., Samson, J.M., Daza, R., et al. (2014). Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol *32*, 829–833.

Latham, M.P., Brown, D.J., McCallum, S.A., and Pardi, A. (2005). NMR Methods for Studying the Structure and Dynamics of RNA. ChemBioChem *6*, 1492–1505.

Lawrence, J.B., Singer, R.H., and Marselle, L.M. (1989). Highly localized tracks of specific transcripts within interphase nuclei visualized by in situ hybridization. Cell *57*, 493–502.

Le, T.B., Imakaev, M.V., Mirny, L.A., and Laub, M.T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. Science *342*, 731–734.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly multiplexed subcellular RNA sequencing in situ. Science *343*, 1360–1363.

Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435–1439.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping Back to Leap Forward: Transcription Enters a New Era. *157*, 13–25.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Linhoff, M.W., Garg, S.K., and Mandel, G. (2015). A high-resolution imaging approach to investigate chromatin architecture in complex tissues. Cell *163*, 246–255.

Liu, Z., Lavis, L.D., and Betzig, E. (2015). Imaging live-cell dynamics and structure at the single-molecule level. Molecular Cell *58*, 644–659.

Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V., and Goodwin, G.H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *5*, 1743–1753.

Lorenz, R., Bernhart, S., Bernhart, S.H., Siederdissen, C.H.Z., Höner zu Siederdissen, C., Tafer, H., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L., et al. (2011). ViennaRNA Package 2.0. Algorithms Mol Biol *6*, 26–14.

Lu, X.J., Lu, X.J., and Olson, W.K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Research *31*, 5108–5121.

Lubeck, E., and Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. Nat Meth *9*, 743–748.

Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. Nat Meth *11*, 360–361.

Lucks, J.B., Mortimer, S.A., Trapnell, C., Luo, S., Aviran, S., Schroth, G.P., Pachter, L., Doudna, J.A., and Arkin, A.P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). Proc. Natl. Acad. Sci. U.S.a. *108*, 11063–11068.

Lupianez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., et al. (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nat Meth *12*, 71–78.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Maeder, M.L., Angstman, J.F., Richardson, M.E., Linder, S.J., Cascio, V.M., Tsai, S.Q., Ho, Q.H., Sander, J.D., Reyon, D., Bernstein, B.E., et al. (2013). Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat Biotechnol *31*, 1137–1142.

Manuelidis, L. (1985). Individual interphase chromosome domains revealed by in situ hybridization. Hum. Genet. *71*, 288–293.

Marbouty, M., Le Gall, A., Cattoni, D.I., Cournac, A., Koh, A., Fiche, J.B., Mozziconacci, J., Murray, H., Koszul, R., and Nollmann, M. (2015). Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. Molecular Cell *59*, 588–602.

Markaki, Y., Smeets, D., Fiedler, S., Schmid, V.J., Schermelleh, L., Cremer, T., and Cremer, M. (2012). The potential of 3D-FISH and super-resolution structured illumination microscopy for studies of 3D nuclear architecture: 3D structured illumination microscopy of defined chromosomal structures visualized by 3D (immuno)-FISH opens new perspectives for studies of nuclear architecture. Bioessays *34*, 412–426.

Matheson, T.D., and Kaufman, P.D. (2015). Grabbing the genome by the NADs. Chromosoma.

McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

Melnik, S., Deng, B., Papantonis, A., Baboo, S., Carr, I.M., and Cook, P.R. (2011). The proteomes of transcription factories containing RNA polymerases I, II or III. Nat Meth *8*, 963–968.

Mendenhall, E.M., Williamson, K.E., Reyon, D., Zou, J.Y., Ram, O., Joung, J.K., and Bernstein, B.E. (2013). Locus-specific editing of histone modifications at endogenous enhancers. Nat Biotechnol *31*, 1133–1136.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet *47*, 598–606.

Minajigi, A., Froberg, J.E., Wei, C., Sunwoo, H., Kesner, B., Colognori, D., Lessing, D., Payer, B., Boukhali, M., Haas, W., et al. (2015). Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. Science *349*, aab2276–aab2276.

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. Cell *128*, 787–800.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H.D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., et al. (2014). Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. Nature *516*, 432–435.

Mukherjee, S., Erickson, H., and Bastia, D. (1988). Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein. *52*, 375–383.

Müller, H.-P., Sogo, J., and Schaffner, W. (1989). An enhancer stimulates transcription in Trans when attached to the promoter via a protein bridge. *58*, 767–777.

Myers, R.M., Tilly, K., and Maniatis, T. (1986). Fine structure genetic analysis of a beta-globin promoter. Science *232*, 613–618.

Nagano, T., Varnai, C., Schoenfelder, S., Javierre, B.M., Wingett, S.W., and Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol. *16*, 175.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59–64.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. Science *342*, 948–953.

Nawrocki, E.P., Eddy, S.R., Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

Nemeth, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Peterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Langst, G. (2010). Initial genomics of the human nucleolus. PLoS

Genet *6*, e1000889.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.e22.

Orian, A., van Steensel, B., Delrow, J., Bussemaker, H.J., Li, L., Sawado, T., Williams, E., Loo, L.W.M., Cowley, S.M., Yost, C., et al. (2003). Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. Genes Dev. *17*, 1101–1114.

Ou, H.D., Phan, S., Deerinck, T.J., Thor, A., Ellisman, M.H., and O'Shea, C.C. (2017). ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. Science *357*, eaag0025.

Oudet, P., Gross-Bellard, M., and Chambon, P. (1975). Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. Cell *4*, 281–300.

Parada, L.A., McQueen, P.G., and Misteli, T. (2004). Tissue-specific spatial organization of genomes. Genome Biol. *5*, R44.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol *30*, 265–270.

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol *27*, 1173–1175.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W., Solovei, I., Brugman, W., Graf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Molecular Cell *38*, 603–613.

Phair, R.D., Scaffidi, P., Elbi, C., Vecerová, J., Dey, A., Ozato, K., Brown, D.T., Hager, G., Bustin, M., and Misteli, T. (2004). Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. Mol. Cell. Biol. *24*, 6393–6402.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Pinkel, D., Straume, T., and Gray, J.W. (1986). Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. Proceedings of the National Academy of Sciences *83*,

2934–2938.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Quinn, J.J., Ilik, I.A., Qu, K., Georgiev, P., Chu, C., Akhtar, A., and Chang, H.Y. (2014). Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. Nat Biotechnol *32*, 933–940.

Ramani, V., Cusanovich, D.A., Hause, R.J., Ma, W., Qiu, R., Deng, X., Blau, C.A., Disteche, C.M., Noble, W.S., Shendure, J., et al. (2016a). Mapping 3D genome architecture through in situ DNase Hi-C. Nat Protoc *11*, 2104–2121.

Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. Nat Meth *14*, 263–266.

Ramani, V., Qiu, R., and Shendure, J. (2015). High-throughput determination of RNA structure by proximity ligation. Nat Biotechnol *33*, 980–984.

Ramani, V., Shendure, J., and Duan, Z. (2016b). Understanding Spatial Genome Organization: Methods and Insights. Genomics Proteomics Bioinformatics *14*, 7–20.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rapkin, L.M., Anchel, D.R., Li, R., and Bazett-Jones, D.P. (2012). A view of the chromatin landscape. Micron *43*, 150–158.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol *27*, 667–670.

Reuter, J., and Mathews, D. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *11*, 129.

Ricci, M.A., Manzo, C., García-Parajo, M.F., Lakadamyali, M., and Cosma, M.P. (2015). Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. Cell *160*, 1145–1158.

Rieder, C.L., and Khodjakov, A. (2003). Mitosis through the microscope: advances in seeing inside live dividing cells. Science *300*, 91–96.

Risca, V.I., and Greenleaf, W.J. (2015). Unraveling the 3D genome: genomics tools for multiscale exploration. Trends in Genetics : TIG *31*, 357–372.

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol *33*, 1165–1172.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Research *20*, 761–770.

Sabo, P.J., Ren, B., Consortium, N.R.E., Farnham, P.J., Bilenky, M., Dixon, J.R., Hirst, M., Kaul, R., Canfield, T.K., Zhang, Z., et al. (2015). Integrative analysis of 111 reference human epigenomes. *518*, 317–330.

Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U.S.a. *112*, E6456–E6465.

Sanyal, A., Bau, D., Marti-Renom, M.A., and Dekker, J. (2011). Chromatin globules: a common motif of higher order chromosome structure? Curr Opin Cell Biol *23*, 325–331.

Schardin, M., Cremer, T., Hager, H.D., and Lang, M. (1985). Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. Hum. Genet. *71*, 281–287.

Schermelleh, L., Carlton, P.M., Haase, S., Shao, L., Winoto, L., Kner, P., Burke, B., Cardoso, M.C., Agard, D.A., Gustafsson, M.G.L., et al. (2008). Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy. Science *320*, 1332–1336.

Schneider, R., and Grosschedl, R. (2007). Dynamics and interplay of nuclear architecture, genome organization, and gene expression. Genes Dev. *21*, 3027–3043.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S., et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet *42*, 53–61.

Schwartzman, O., and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. Nat Rev Genet *16*, 716–726.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259.

Sexton, T., and Cavalli, G. (2015). The role of chromosome domains in shaping the functional

genome. Cell *160*, 1049–1059.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shachar, S., Voss, T.C., Pegoraro, G., Sciascia, N., and Misteli, T. (2015). Identification of Gene Positioning Factors Using High-Throughput Imaging Mapping. Cell *162*, 911–923.

Shalem, O., Carey, L., Zeevi, D., Sharon, E., Keren, L., Weinberger, A., Dahan, O., Pilpel, Y., and Segal, E. (2013). Measurements of the impact of 3' end sequences on gene expression reveal wide range and sequence dependent effects. PLoS Comput Biol *9*, e1002934.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shendure, J., and Aiden, E.L. (2012). The expanding scope of DNA sequencing. *30*, 1084–1094.

Shendure, J., and Lieberman-Aiden, E. (2012). The expanding scope of DNA sequencing. Nat Biotechnol *30*, 1084–1094.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728–1732.

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science *357*, eaaf4382.

Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Meth *11*, 959–965.

Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature *504*, 465–469.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet *38*, 1348–1354.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat Genet *45*, 1021–1028.

Soderberg, O., Gullberg, M., Landegren, U., Jarvius, J., Jarvius, M., Ridderstrale, K., Leuchowius, K.-J., Wester, K., Hydbring, P., Bahram, F., et al. (2006). Direct observation of individual endogenous protein complexes in situ by proximity ligation. *3*, 995–1000.

Solomon, M.J., and Varshavsky, A. (1985). Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. Proc. Natl. Acad. Sci. U.S.a. *82*, 6470–6474.

Song, F., Chen, P., Sun, D., Wang, M., Dong, L., Liang, D., Xu, R.M., Zhu, P., and Li, G. (2014). Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. Science *344*, 376–380.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613–626.

Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.

Takizawa, T., Meaburn, K.J., and Misteli, T. (2008). The meaning of gene positioning. Cell *135*, 9–13.

Talbert, P.B., and Henikoff, S. (2010). Histone variants--ancient wrap artists of the epigenome. Nat. Rev. Mol. Cell Biol. *11*, 264–275.

Tanabe, H., Müller, S., Neusser, M., Hase, von, J., Calcagno, E., Cremer, M., Solovei, I., Cremer, C., and Cremer, T. (2002). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. Proc. Natl. Acad. Sci. U.S.a. *99*, 4424–4429.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1611–1627.

Thakore, P.I., D'Ippolito, A.M., Song, L., Safi, A., Shivakumar, N.K., Kabadi, A.M., Reddy, T.E., Crawford, G.E., and Gersbach, C.A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Meth *12*, 1143–1149.

Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science *346*, 1238–1242.

Thomson, I., Gilchrist, S., Bickmore, W.A., and Chubb, J.R. (2004). The radial positioning of chromatin is not inherited through mitosis but is established de novo in early G1. Curr Biol *14*, 166–172.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Molecular Cell *10*, 1453–1465.

Toomre, D., and Bewersdorf, J. (2010). A new wave of cellular imaging. Annual Review of Cell and Developmental Biology *26*, 285–314.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Research *25*, 1491–1498.

Tsukamoto, T., Hashiguchi, N., Janicki, S.M., Tumbar, T., Belmont, A.S., and Spector, D.L. (2000). Visualization of gene activity in living cells. Nat Cell Biol *2*, 871–878.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505–510.

Ulianov, S.V., Khrameeva, E.E., Gavrilov, A.A., Flyamer, I.M., Kos, P., Mikhaleva, E.A., Penin, A.A., Logacheva, M.D., Imakaev, M.V., Chertovich, A., et al. (2015). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Research *26*, 70–84.

Underwood, J.G., Uzilov, A.V., Katzman, S., Onodera, C.S., Mainzer, J.E., Mathews, D.H., Lowe, T.M., Salama, S.R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *7*, 995–1001.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. Nat Rev Genet *11*, 636–646.

van Arensbergen, J., FitzPatrick, V.D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H.J., and van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. Nat Biotechnol *35*, 145–153.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G.J., Ariyurek, Y., Dunnen, den, J.T., and Lamond, A.I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol. Biol. Cell *21*, 3735–3748.

van Steensel, B., and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol *18*, 424–428.

van Steensel, B., and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nat Biotechnol *28*, 1089–1095.

Vierstra, J., Reik, A., Chang, K.-H., Stehling-Sun, S., Zhou, Y., Hinkley, S.J., Paschon, D.E., Zhang, L., Psatha, N., Bendana, Y.R., et al. (2015). Functional footprinting of regulatory DNA. Nat Meth *12*, 927–930.

Walter, J., Schermelleh, L., Cremer, M., Tashiro, S., and Cremer, T. (2003). Chromosome order in HeLa cells changes during mitosis and early G1, but is stably maintained during subsequent interphase stages. J. Cell Biol. *160*, 685–697.

Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *505*, 706–709.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011a). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C.-T., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science *353*, 598–602.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wang, W., Li, G.W., Chen, C., Xie, X.S., and Zhuang, X. (2011b). Chromosome organization by a nucleoid-associated protein in live bacteria. Science *333*, 1445–1449.

Wang, Y.-H., Cech, T.R., Murphy, F.L., and Griffith, J.D. (1994). Visualization of a Tertiary Structural Domain of the Tetrahymena Group I Intron by Electron Microscopy. *236*, 64–71.

Wood, A.J., Severson, A.F., and Meyer, B.J. (2010). Condensin and cohesin complexity: the expanding repertoire of functions. Nat Rev Genet *11*, 391–404.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nat Meth *11*, 41–46.

Yang, F., Deng, X., Ma, W., Berletch, J.B., Rabaia, N., Wei, G., Moore, J.M., Filippova, G.N., Xu, J., Liu, Y., et al. (2015). The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. Genome Biol. *16*, 52.

Zentner, G.E., and Henikoff, S. (2014). High-resolution digital profiling of the epigenome. Nat Rev Genet *15*, 814–827.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908–921.

Zhang, Y., Wong, C.H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature *504*, 306–310.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet *38*, 1341–1347.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet *12*, 7–18.

Zirbel, R.M., Mathieu, U.R., Kurz, A., Cremer, T., and Lichter, P. (1993). Evidence for a nuclear

compartment of transcription and splicing located at chromosome domain boundaries. Chromosome Res. *1*, 93–106.

Zuker, M., and Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research *31*, 3406–3415.

VITA

Vijay Ramani grew up in Short Hills, NJ, and received his undergraduate degree in Chemical & Biological Engineering from Princeton University, with minors in Quantitative and Computational Biology, and Engineering Biology. An avid musician, Vijay spends his time outside the lab playing classical flute, and singing as a tenor in the Seattle Symphony Chorale.