Hi-C: Genome-wide Chromosome Conformation Capture

Lisa Boxer
Biochem 218
December 5, 2010

The three-dimensional conformation of chromosomes in the nucleus is important for many cellular processes, including the regulation of gene expression, DNA replication, and chromatin structure [1]. Despite having the entire sequence of the genome, very little has been understood about three-dimensional chromosome conformation beyond the scale of the nucleosome. However, recent advances in molecular biology and computational analysis have lent insight into chromatin interactions on a larger scale. Regulation of gene expression is often very complex and involves long-range chromatin interactions. For example, a chromosomal region could fold over to bring a distant enhancer region and associated transcription factors within proximity of a target gene promoter [2]. An understanding of the conformation of chromatin will lend insight into these complex regulatory processes.

The technique of chromosome conformation capture (3C) evaluates long-range interactions between specific pairs of loci by using spatially constrained ligation followed by locus-specific polymerase chain reaction [3]. The 3C method involves cross-linking cells with formaldehyde followed by restriction enzyme digestion to leave cross-linked sequences attached. The DNA is then ligated under dilute conditions to favor ligation of cross-linked DNA fragments. The ligated fragments that should contain both pieces of interacting DNA are then analyzed by PCR with primers to the two loci of interest [3] (Fig 1). While useful for specific loci of interest, 3C has very limited throughput. A modification of 3C, circularized chromosome conformation capture (4C), has the advantage that the sequence of only one site of interest needs to be known. The sequences of all loci that interact with the chosen locus can be determined by

inverse PCR, followed by hybridization to microarrays or high-throughput sequencing [4, 5]. While this method allows investigation of many unknown interacting sequences, it is still limited in terms of throughput since only one input sequence can be used per experiment. An additional method with slightly higher throughput, carbon copy chromosome conformation capture (5C) expands on 3C by allowing parallel analysis of the interactions between many selected loci [6]. After generation of a 3C library, 5C primers with universal primer sequence tails such as T7 or T3 are ligated to the DNA fragments. Multiplex ligation mediated amplification (LMA) can then be used to generate a 5C library, which can be analyzed by microarray hybridization or high-throughput sequencing [6] (Fig 1).

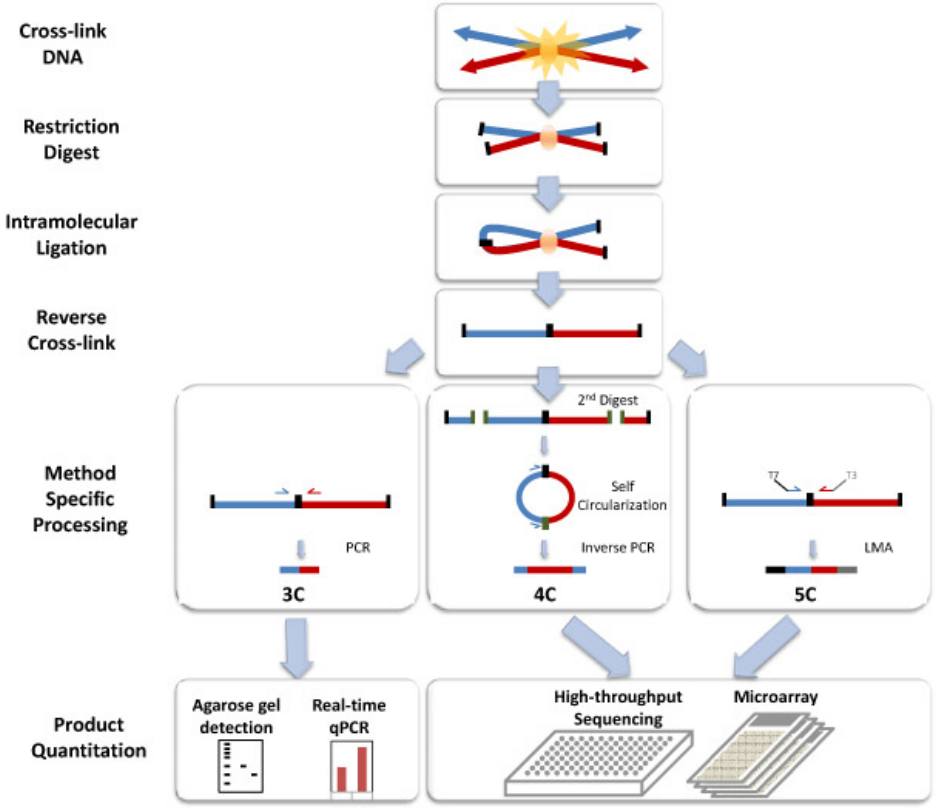


Figure 1. A comparison of methods used for chromosome conformation capture: 3C, 4C, and 5C. While all 3 methods rely on cross-linking DNA, restriction enzyme digest, and ligation under dilute conditions, 3C analyzes the interaction between two individual loci by PCR, 4C analyzes all loci that interact with one locus by inverse PCR followed by microarray or high-throughput sequencing, and 5C analyzes many parallel interactions by generating a library by amplification with universal primer tags and analysis by microarray or high-throughput sequencing [3-7].

While 5C enables analysis of chromatin interactions between many loci, the method is not suitable for a genome-wide analysis of chromosome conformation because of the extensive number of primers that would be required [7]. Very recently, a novel method, termed Hi-C, has been developed to overcome these difficulties and assess chromosome conformation across the entire genome [8, 9]. This method involves cross-linking cells with formaldehyde, followed by DNA digestion with a restriction enzyme that leaves 5' overhangs, which are filled in with biotinylated nucleotides. The blunt-end fragments are ligated under dilute conditions to favor the ligation of cross-linked segments as in 3C. The ligated DNA is then sheared, and the biotin-containing fragments are selected with streptavidin beads to yield a library of fragments containing sequences from interacting loci. The library is then subjected to paired-end high-throughput sequencing. This method has been used to successfully map chromosome conformation and interactions across the genome at a scale of 1Mb [8].

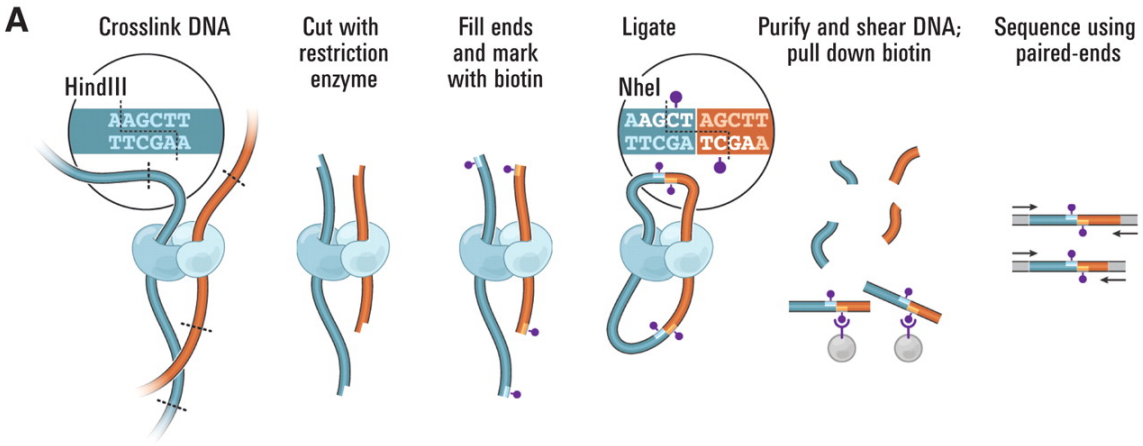


Figure 2. Overview of Hi-C protocol. Cells are cross-linked with formaldehyde, digested with a restriction enzyme, the 5' overhang is filled with a biotinylated residue, blunt-end fragments are ligated under dilute conditions, DNA fragments are sheared and selected with streptavidin beads. The library containing proximity-ligated fragments is analyzed by paired-end high throughput sequencing [8, 9].

The computational analysis of Hi-C data is quite complex. Briefly, it involves mapping sequence reads back to the genome, determining which reads are likely to be the product of proximity-based ligation, generating interaction matrices of chromosomal interactions, and correlation analysis of genomic interactions [8]. Since Hi-C is such a novel method, there are many challenges with the computational analysis and many possible ways to improve the analysis. Some of the challenges in the data analysis include determining which sequence reads are the result of proximity-based ligation, the issue of high background noise from random collisions particularly at short genomic distances, the difficulties of increasing resolution, and the challenge of making a three-dimensional map of the genome from sequencing data that is statistical in nature. In this review, I will discuss the current Hi-C data analysis methods, the challenges associated with them, and potential ways to improve these methods for increased resolution and specificity.

The first step of analyzing Hi-C sequencing data is to map the paired-end sequence reads to the genome. The process of aligning sequence reads to the genome is becoming a relatively well-established process, and there are many programs available for this part of the analysis, such as MAQ [10]. The sequence reads are mapped to the human hg18 reference sequence by searching for the ungapped match with the lowest mismatch score, and only considering alignments with two or fewer mismatches in the first 28bp [10]. For Hi-C data, each of the paired reads should align to the genome for the sequence to be included to the interaction data, since the goal is to analyze the interactions between these two genomic regions [8].

The next step is quality control to ensure that the aligned sequence reads are likely to be the result of proximity-based ligation of digested fragments, and that they are likely to reflect long-range chromatin interactions rather than just random collisions. The restriction enzyme

chosen for Hi-C library preparation should have been selected based on the fragments yielded by genomic digestion, and additional enzymes should be tested to avoid bias based on restriction sites. In Lieberman-Aiden *et al*., the restriction enzyme HindIII was chosen because it cuts the genome into 800,000 similar-sized fragments, but similar genomic interaction results were obtained with Nco1. It should be confirmed that the aligned sequence reads are located near the sites for the restriction enzyme used in the library generation. If the sequence read is further away from the restriction site than the maximum sequence read length as determined by sonication of DNA fragments and read alignment, this sequence should be eliminated [9] (Fig 3A). In Lieberman-Aiden *et al*., the maximum sequence read length is 500bp, and it can be observed that most intra and inter- chromosomal interaction reads are less than 500bp from a HindIII site (Fig 3A). The Hi-C sequence reads can be compared to randomly generated control sequence reads, and the Hi-C sequence reads should be significantly closer to the chosen restriction sites than the random reads (Fig 3A). The sequence reads should also be in the correct orientation with respect to the restriction site. The 3' ends of both fragments should be adjacent to the restriction site based on the library preparation protocol and sequencing 5'-3' from the paired ends to the middle where the ligation occurs (Fig 2B). In Lieberman-Aiden *et al*., 80% of interaction reads had the 3' aligned with a HindIII site (Fig 3B). Quality control for the percentage of reads that map to intrachromosomal vs. interchromosomal interactions can also be performed. In van Berkum *et al*., it is recommended that of the aligned reads, 55% of the reads should represent interchromosomal interactions, 15% represent intrachromosomal interactions <20 kb apart, and 35% represent intrachromosomal interactions >20 kb apart (Fig 3C).

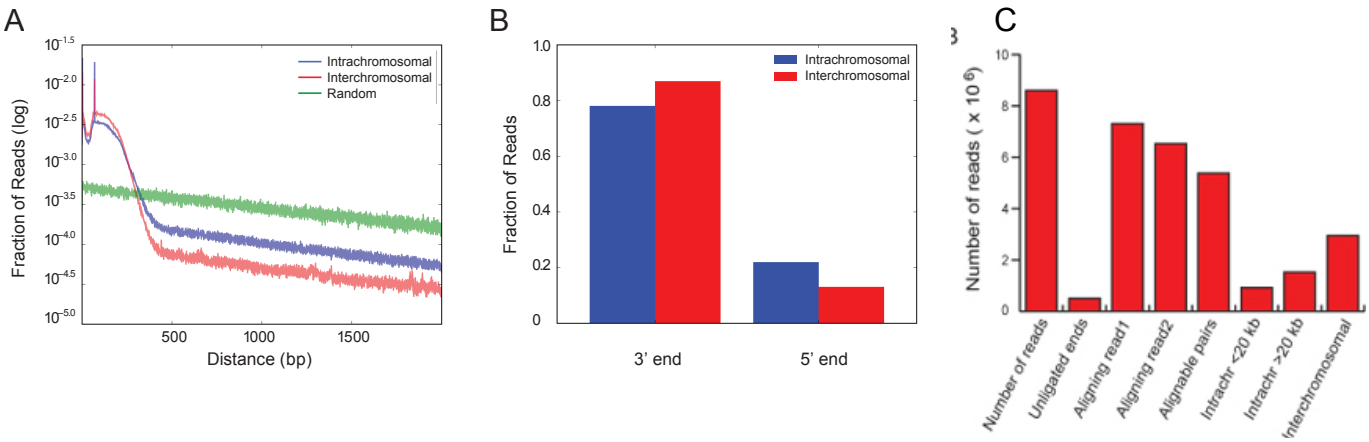


Figure 3. Quality control of aligned sequence reads for a typical Hi-C experiment. **A**) Fraction of reads that align to certain distance from restriction site used in library preparation. Intra and inter-chromosomal reads should align significantly closer to restriction sites than random controls, until the distance from the restriction site is greater than the greatest read length (~500bp). **B**) Typically, 55% of alignable read pairs correspond to interchromosomal interactions, 15% are intrachromosomal <20kb apart, and 35% are intrachromosomal >20kb. **C**) The 3' end of Hi-C sequences should align to the restriction enzyme site in at least 80% of the reads [8, 9]

After the sequence reads have been aligned and quality control has been conducted, the next step is to generate contact matrices of both intrachromosomal and interchromosomal interactions. To produce a contact matrix, the genome should be divided into appropriately sized loci. The size of the loci depends on the resolution desired from the analysis, which is limited by the depth of sequencing. For a more global analysis, the genome could be segmented into 1Mb loci, but for finer analysis, smaller loci such 100Kb or smaller could be used. However, looking for interactions at higher resolution requires increasing the depth of sequencing. Each interaction is mapped by binning each end of the sequence read into the appropriate locus. Interaction matrices can then be produced based on the frequency of interactions between each pair of loci across the genome. The matrix entry $m_{i,j}$ would correspond to the number of ligation products between locus *i* and locus *j* [8]. The interaction matrix can be depicted visually with a heat map, in which the color intensity correlates with contact frequency (Fig 4A). This interaction matrix will reveal which segments of chromosomes are positioned close together or further apart. As a

control at this step, heat maps from sheared genomic DNA can be generated and should show no long-range interactions.

The next step in the analysis is to produce a matrix comparing the observed number of reads between two loci to the expected number of reads between two loci. For intrachromosomal interactions, the expected number of reads is determined by the average intrachromosomal contact probability, I(s), where *s* represents the genomic distance between the midpoints of two loci. I(s) is determined from the Hi-C data and is equal to: the total number of observed interactions at a distance *s* divided by the total number of possible interactions at distance *s* across all chromosomes [8]. I(s) decreases monotonically as *s* increases on each chromosome, indicating that as the distance between two loci increases, the expected number of interactions decreases (Fig 4B and 5A). A matrix of observed versus expected reads can be generated, in which the number of actual reads between loci *i* and *j* is compared to the number of expected reads at the distance *s* between *i* and *j*. The observed/expected for the matrix entry for loci *i,j* is equal to: $m_{i,j}$ / I(s(i,j)) [8]. This matrix can also be illustrated with a heat map, depicting interactions that are more (red) or less (blue) likely to occur than expected (Fig 4C). For interchromosomal interactions, the expected number of interactions between locus *i* and locus *j* is equal to: the fraction of reads containing *i* multiplied by the fraction of reads containing *j*, multiplied by the total number of reads [8]. The number of observed interactions between each set of loci is divided by the expected number to generate the observed over expected matrix (Fig 5B), which demonstrates chromosomes that are more (red) or less (blue) likely to interact than expected. The probability of intrachromosomal interaction should be higher than interchromosomal interaction even at distances >200kb, since chromosomes exist in territories [1, 8] (Fig 5A).

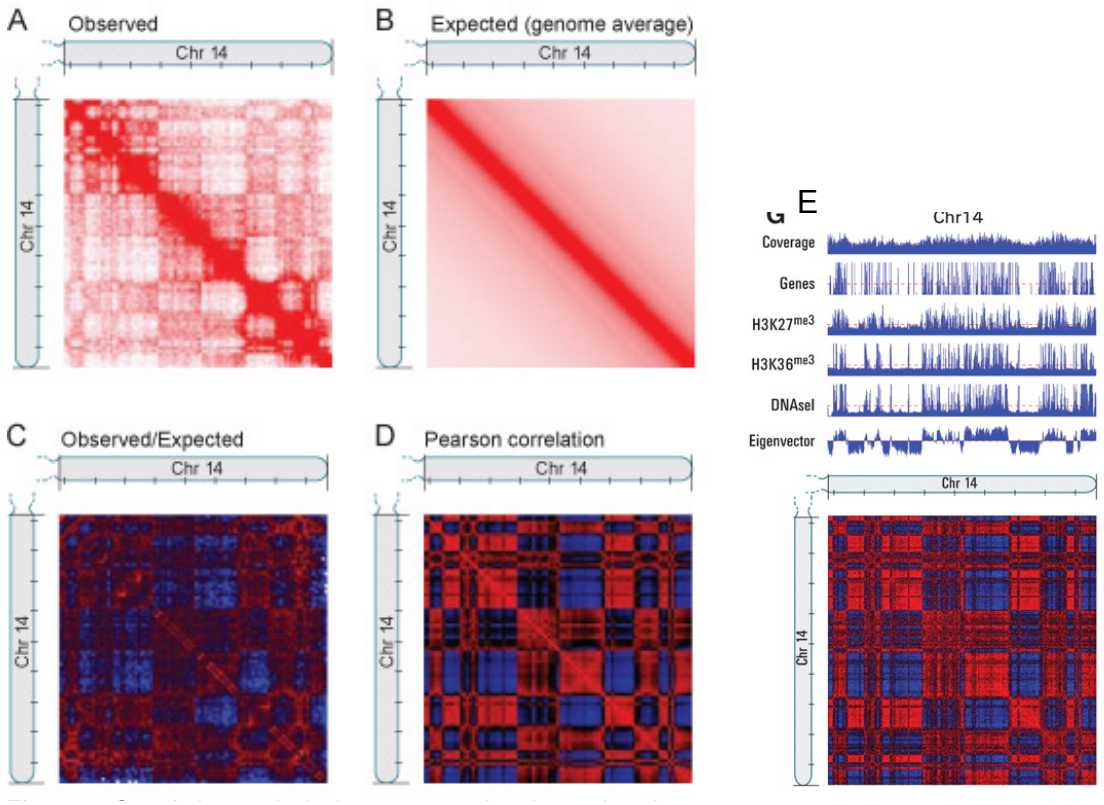


Figure 4. Heat maps depicting intrachromosomal contact heat maps for chromosome 14 at resolution of 1Mb. **A**) Observed interactions. **B**) Expected interaction frequencies based on genomic distance. **C**) Quotient of matrices A and B, showing more (red) or less (blue) interactions than expected. **D**) Correlation matrix between intrachromosomal interaction profiles. **E**) Principle components analysis reveals correlation between the principle component (eigenvector) and the presence of genes and features of open chromatin. Regions of less densely packed chromatin (blue) are correlated with open chromatin and DNA accessibility [8].

Further statistical analysis of the data can be used to make a correlation matrix. It is predicted that two loci that are close together in space should interact with similar loci and thus should have correlated interaction profiles. In the correlation matrix C, the entry $c_{i,j}$ is determined by computing the Pearson correlation coefficient between the vectors represented by the *i*th row and the *j*th column of the observed/expected matrix [8] (Fig 4D). From the correlation matrix, it becomes more clear that there are regions of enriched interaction (red) and regions of depleted interaction (blue) (Fig 4D). For determining correlation of intrachromosomal interactions at low resolution, it may be more appropriate to use the Spearman correlation since the average contact

probability I(s) decreases monotonically with distance on each chromosome (Fig 5A). However, at a resolution of 100Kb or higher, the matrix is too sparse and the Pearson correlation should be used [8].

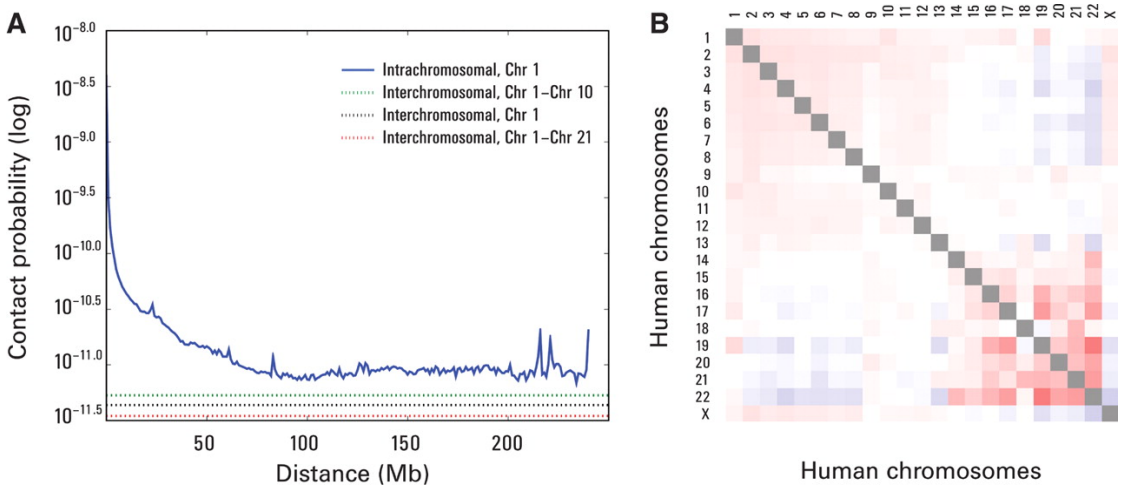


Figure 5. **A**) Intrachromosomal contact probability decreases monotonically as a function of genomic distance between loci. Interchromosomal contact probability does not vary with genomic distance. Intrachromosomal contact probability is always greater than interchromosomal contact probability, even at genomic distance >200Mb. **B**) Observed/expected interchromosomal interactions that are more (red) or less (blue) likely to occur than expected [8].

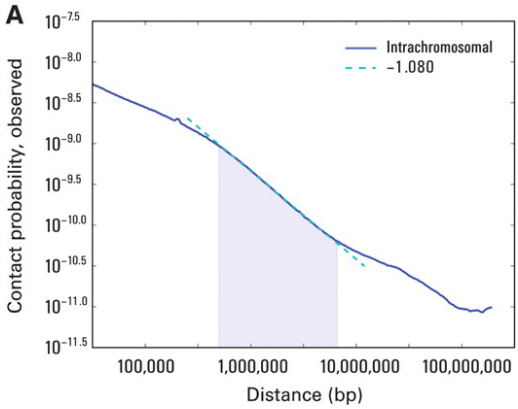
Principle components analysis (PCA) can be used to partition chromosomes into two arbitrary sets of loci for which contacts are enriched within sets and depleted between sets. One set corresponds to highly interactive, densely packed regions (red) and the other to interaction depleted regions (blue), defined arbitrarily by negative and positive values, respectively (Fig 4E). The first principle component (PC) corresponded to this pattern. The eigenvector obtained from the first PC can be compared to other genomic features for further analysis of the biological significance of the chromosome conformation. For example, in Lieberman-Aiden *et al*., the eigenvector of chromosome conformation was compared to gene-rich regions, histone methylation states, and DNAse I sensitivity (indicative of accessible chromatin). These tracks can be obtained from the UCSC genome browser, and can be compared to the PC eigenvector

using Spearman's correlation coefficient [8]. In Lieberman-Aiden *et al*., correlation analysis revealed that the less interactive, less densely packed regions (blue) were correlated with gene-rich regions, enrichment for activating and repressing histone methylation marks, and accessible chromatin (Fig 4E). PCA could be used to compare the Hi-C data to many other types of genomic features in future analysis.

The highest resolution that has been used in Hi-C experiments for genome-wide interaction maps to date is 1Mb [8]. It would be very interesting to investigate genomic interactions with higher resolution, to determine interactions between specific genes, enhancers, silencers, or promoter regions. To increase the resolution, the depth of sequencing would have to be increased. For a 1Mb resolution map of the genome, Lieberman-Aiden *et al*. used 30 million aligned reads. The difficulty with increasing resolution is that to increase the resolution by a factor of n, the number of sequencing reads needs to by increased by a factor of n^2 [11]. Thus, to increase the resolution 10-fold to 100Kb genome-wide, 3 billion aligned sequence reads would be required. Another issue with increasing resolution is that at short-range, the frequency of random collisions may create too much background to determine real interactions [11]. Improvements in the computational analysis methods will help to reduce background noise by better distinguishing proximity-ligation based versus random interactions.

While Hi-C does not directly measure genomic distance between loci, the statistical interaction data can be used to predict the three-dimensional structure of the genome, chromosomes, and segments of chromosomes. Lieberman-Aiden *et al*. demonstrated that if the intrachromosomal average contact probability I(s) is plotted on a log-log axis, contact probability scales as s^{-1} between 500kb and 7Mb, which corresponds to the known size range for chromatin domains. Power law scaling of s^{-1} is consistent with the three-dimensional model of a fractal

globule, modeled by the space-filling Peano curve, which would configure DNA in a compact yet accessible structure with minimal knot formation [8]. At the several megabase scale, the Hi-C data fits the model of a fractal globule. However, it will be interesting to model the three-dimensional genomic conformation at higher resolution.



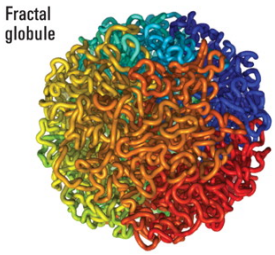


Figure 6. Contact probability as a function of genomic distance exhibits power law scaling of s^{-1} between 500 kb and 7Mb. This power law scaling can be modeled as a fractal globule [8].

In addition to determining interactions genome-wide, it is also intriguing to investigate which specific transcription factors or chromatin modifiers mediate these genomic interactions. Several combinations of 3C and chromatin immunoprecipitation (ChIP) have been developed recently to address this question [12]. ChIP is a method used to isolate DNA that is bound by a particular transcription factor of interest. Briefly, this method involves the cross-linking of DNA and proteins with formaldehyde, pull-down of the protein of interest with an antibody, isolation and amplification of DNA, and analysis of ChIP DNA by high-throughput sequencing [12]. Recently, ChIP has been combined with chromosome conformation capture techniques to determine long-range chromatin interactions that are mediated by a specific transcription factor, and obtain a three-dimensional understanding of the interactions mediated by a specific factor. On a locus specific basis, this can be accomplished by the ChIP version of 3C, termed ChIP-loop

[13]. The method involves cross-linking of cells, restriction enzyme digestion of DNA, immunoprecipitation with antibody to the protein of interest, ligation of precipitated DNA fragments, and analysis of DNA fragments by PCR [13]. This method is useful for determining whether a specific transcription factor mediates a specific genomic interaction, but not for high throughput analysis. A more high throughput method has recently been developed, termed ChIA-PET, chromosome interaction analysis by paired-end sequence tagging, which can determine the interactions mediated by a particular transcription factor on a genome-wide scale [14]. In the ChIA-PET method, long-range chromatin interactions are captured by crosslinking with formaldehyde, fragmented by sonication, and DNA-protein complexes containing the protein of interest are enriched by ChIP. The tethered DNA fragments in each of the chromatin complexes are connected with DNA linkers during proximity-based ligation. These DNA linkers are used to extract the DNA fragments of interest and the DNA is analyzed by paired-end sequencing [14]. The computational analysis of ChIA-PET data is also very complex, and is similar to Hi-C data analysis and will not be discussed in this review.

In addition to combining ChIP and 3C experimentally, it is also possible and potentially easier to combine ChIP and 3C data computationally. If ChIP-Seq data is available for a particular transcription factor, the binding sites from this data can be compared with genome-wide conformation data from Hi-C to determine which chromosome conformations are mediated by that transcription factor. In a recent publication, this method was pioneered for the CCCTC-binding factor CTCF [15]. The strength of interaction between fragments was assessed based on the number of interactions each fragment is involved in. The strength of interaction was compared to the presence of CTCF binding sites on these fragments and it was found that strongly interacting DNA fragments are more likely to contain a CTCF binding site, as compared

to other factors [15] (Fig 7). However, this does not eliminate the importance of the other factors in mediating chromatin interactions, as even a locus with very few interactions could still have tremendous transcriptional importance. Future developments of this computational method will depend on the increased availability of ChIP-Seq data for more transcription factors, and on the ability of computational analysis to predict transcription factor binding sites. This method will also benefit significantly from increasing the resolution of Hi-C data. With increased Hi-C resolution, it will become more clear which transcription factor binding sites are mediating the long-range chromatin interactions. At the current resolution of 1Mb, it is hard to predict whether a single binding site in such a large locus is actually responsible for the chromosome conformation. The overall goal for the future of chromosome conformation would be to have a detailed three-dimensional map of the genome-wide conformation of chromosomes, and to understand which transcription factors and chromatin modifiers are mediating all of the interactions.

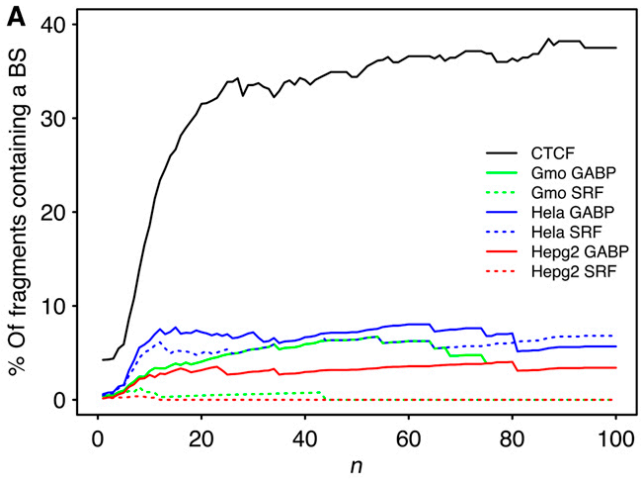


Figure 7. CTCF binding sites are correlated with frequently observed interactions in human genome. Determined by comparing Hi-C data to CTCF Chip-seq data. BS = binding site, n = number of genomic interactions [15].

References:

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

2. Sexton T, Bantignies F, Cavalli G. (2009). Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. Semin Cell Dev Biol. *20*, 849-855.

3. Dekker, J., Rippe, M. Dekker, M., Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306-1311.

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

5. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, 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.

6. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299-1309.

7. Simonis M, Kooren J, de Laat W. (2007). An evaluation of 3C-based methods to capture DNA interactions. Nat. Methods *4*, 895-901.

8. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

9. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny NA, Dekker J, Lander, ES. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. J Vis Exp *39*, 1869-1876.

10. Li H, Ruan J, Durbin R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851-1858.

11. Shaw PJ. (2010). Mapping chromatin conformation. F1000 Biol. Rep. *2*, 1-4.

12. Fullwood MJ and Ruan Y. (2009). ChIP-Based methods for the identification of long-range chromatin interactions. J Cell Biochem. *107*, 30-39.

13. Horike S, Cai S, Miyano M, Cheng J, Kohwi-Shigematsu T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat. Genet. *37*, 31-40.

14. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EGY, Huang PYY, Welboren W, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PE, Wansa KDSA, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai1 KV, Thomsen JS, Lee YK, Karuturi RKM, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung W, Liu ET, Wei C, Cheung E, Ruan Y. (2009) An oestrogen-receptor-a-bound human chromatin interactome. Nature *462*, 58-64.

15. Botta M, Haider S, Leung IXY, Lio P, Mozziconacci J. (2010). Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Mol. Syst. Biol. *6*, 1-6.