Published in final edited form as:
Methods. 2012 November ; 58(3): . doi:10.1016/j.ymeth.2012.05.001.

Hi-C: A comprehensive technique to capture the conformation of genomes

Jon-Matthew Belton[1], **Rachel Patton McCord**[1], **Johan Gibcus**[1], **Natalia Naumova**[1], **Ye Zhan**[1], and **Job Dekker**[1,*]

[1]University of Massachusetts Medical School, Program in Systems Biology, 364 Plantation 570M LRB Worcester MA, 01605

Abstract

We describe a method, Hi-C, to comprehensively detect chromatin interactions in the mammalian nucleus. This method is based on Chromosome Conformation Capture, in that chromatin is crosslinked with formaldehyde, then digested, and re-ligated in such a way that only DNA fragments that are covalently linked together form ligation products. The ligation products contain the information of not only where they originated from in the genomic sequence but also where they reside, physically, in the 3D organization of the genome. In Hi-C, a biotin-labeled nucleotide is incorporated at the ligation junction, making it possible to enrich for chimeric DNA ligation junctions when modifying the DNA molecules for deep sequencing. The compatibility of Hi-C with next generation sequencing platforms makes it possible to detect chromatin interactions on an unprecedented scale. This advance gives Hi-C the power to both explore the chromatin biophysics as well as the implications of chromatin structure in the biological functions of the nucleus. A massively parallel survey of chromatin interaction provides the previously missing dimension of spatial context to other genomic studies. This spatial context will provide a new perspective to studies of chromatin and its role in genome regulation in normal conditions and in disease.

Keywords

Chromatin Conformation; Nucleus; Genomics

1. Introduction

The long DNA strands of every cell’s genome are packaged into chromatin in a very confined nuclear volume [1]. The organization of the chromatin in the nucleus is extremely relevant to biological function at the gene level as well as the global nuclear level. The study of the packaging and organization of chromatin in the nucleus will shed light on the spatial aspects of gene regulation, chromosome morphogenesis, and genome stability and transmission. It will also enhance the understanding of the biophysics of chromatin, and further enable the investigation of pathologies related to genome instability or nuclear morphology.

Many techniques are now available to observe the spatial organization of chromatin. These techniques can be broadly classified into microscopic and molecular assays. Electron microscopy is used to analyze the architecture of the nucleus in nm scale detail, while fluorescent (light) microscopy provides information on the shape and distribution of specific chromosomes and chromosomal loci as well as the co-association of specific loci with sub-

[*]Corresponding Author: job.dekker@umassmed.edu, phone:508-856-4371.

nuclear compartments with a resolution of 50-100 nm [2]. The major disadvantage of electron microscopy is its lack of connectivity to the genomic sequence, in that the incredible fine resolution architecture cannot be assigned to a specific location in the genome. Using sequence specific probes in Fluorescent In Situ Hybridization (FISH) assays allows one to connect nuclear architecture and DNA sequence, but these methods are currently still limited in throughput, allowing analysis of only a few loci simultaneously [3]. In molecular assays based on Chromosome Conformation Capture (3C) the genomic sequence itself is the output, completely connecting chromosome structure and the genomic sequence [4].

The 3C technique and its derivatives measure the population-averaged frequency at which two DNA fragments physically associate in three-dimensional (3D) space, based on the propensity for those two locations to become formaldehyde-crosslinked together (Figure 1A). Once interacting loci are crosslinked, chromatin is solubilized and fragmented, usually using a restriction enzyme (Figure 1B). Interacting fragments are then ligated together and purified, creating a genomic library of chimeric DNA molecules (Figure 1C). The relative abundance of specific chimeras, or ligation products, is related to the probability that those fragments interact in 3D space across the cell population. A 3C library includes a massive variety of ligation products (up to 10^{12} unique pair-wide interactions between 4 Kb fragments in the human genome) and can be analyzed in various ways depending on the goals of the study. To answer small scale questions, 3C ligation products can be assayed individually using PCR primers to specific genomic fragments. In special variants of the 3C technology (ChIP-loop and ChIA-PET), immunoprecipitation is used to associate the interactions with a particular protein of interest [5, 6]. This is used to investigate the role of the protein factor in facilitating genomic contacts. In recent years, new approaches such as 4C, 5C, and Hi-C have been developed to utilize Next Generation Sequencing (NGS) technologies in order to interrogate the 3C ligation product library more comprehensively (figure 1D and 1E) [7-10]. Most 3C-based techniques focus on analysis of a set of predetermined loci enabling “one-versus-some” (basic 3C and ChIP-loop), “one-versus-all” (4C), or “many-versus-many” (5C) explorations of the conformation of chromosomal regions of interest. On the other hand, Hi-C enables an “all-versus-all” interaction profiling.

In this paper we describe Hi-C. In Hi-C, all genomic fragments are labeled with a biotinylated nucleotide before ligation, thereby marking ligation junctions. Marked junctions can then be purified efficiently by streptavidin-coated magnetic beads, enriching the library for ligation products that can be detected by NGS. The comprehensive chromatin interaction data that can be obtained by direct sequencing of a Hi-C library provides immense statistical power for analyses of genome organization at kb resolution. These analyses can reveal overall genome structure and biophysical properties of chromatin as well as more specific long-range contacts between distant genomic elements such as genes and regulatory elements. Further, combining Hi-C data with other datasets including gene expression profiles and genome-wide maps of chromatin modifications will allow placing sets of loci in 3D context which will lead to new insights into the functional roles of chromatin conformation in genome regulation and stability, both in normal cells and in disease states.

2. The Hi-C Method

2.1. Cell Culture and Crosslinking of Chromatin

The combinatorial possibilities of all pair-wise interactions between genomic restriction fragments that can be observed with Hi-C is immense and scales exponentially with genome size. Many of these interaction possibilities may be occurring in the cell population, but, in the experiment, only two interactions for a given fragment can be detected from a single cell. Therefore, it is of critical importance to analyze sufficient numbers of cells to ensure high complexity of the resulting ligation product library. The complexity of this library will ultimately determine the resolution and sensitivity of chromatin interaction datasets. This does not preclude the study of precious samples, such as primary tissue samples, but the resulting data, and resolution at which specific contacts can be detected will be affected by the low sample complexity.

Typically, twenty five million cells are needed to produce a sufficiently complex Hi-C library such that global spatial organization can be analyzed for a mammalian genome. These cells may be obtained by culturing either suspension or adherent cell lines. The cells are fixed in 1% final concentration of formaldehyde in the relevant culture media. After crosslinking the remaining formaldehyde is sequestered with an excess of glycine. Then the cells are harvested by centrifugation and can be flash frozen and stored at -80°C in 25 million cell aliquots.

Standardization of crosslinking conditions is the cornerstone of 3C-based techniques since the functional read-out of this technique is the frequency at which two genomic restriction fragments are crosslinked to one another. One source of variation in crosslinking is the presence of serum in the media. Serum, containing a high concentration of protein, can alter the effective concentration of formaldehyde by binding to and sequestering formaldehyde in the culture media. In cases where serum is used for culturing it should be omitted at the time of crosslinking. Special considerations must also be made when adherent cells are to be processed in to Hi-C libraries. Adherent cells are adsorbed to a surface by molecular mechanisms that involve the cytoskeleton, which are known to play a role in the location and shape of the nucleus in the cytoplasm. This linkage of nuclear morphology to cellular morphology suggests that crosslinking of adherent cells while they are still adhered to the culturing surface is necessary to preserve the global nuclear organization. Once crosslinked, adherent cells can be scraped from the culture plate using disposable cell scrapers and aliquoted the same as suspension cells.

2.2. Cell Lysis and Chromatin Digestion

One 25 million cell pellet is lysed using a Dounce homogenizer in the presence of cold hypotonic buffer supplemented with protease inhibitors and a mild non-ionic detergent (NP-40). The presence of a protease inhibitor cocktail during lysis and conducting the lysis on ice helps protect the crosslinked chromatin complexes from endogenous proteases. The lysate is washed twice with 1X restriction enzyme buffer and then resuspended in 1X restriction enzyme buffer. The chromatin is solubilized with dilute SDS and incubation at 65°C for 10 min. This solubilization removes non-crosslinked proteins and opens the chromatin, making it accessible for restriction endonuclease cleavage. Incubation at 65°C can reverse formaldehyde crosslinks, making it important to minimize the incubation time and to place the sample on ice immediately following incubation. Triton X-100, a non-ionic surfactant, is used to quench the SDS to prevent it from denaturing enzymes in subsequent steps. The accessible chromatin is then digested with a type II restriction endonuclease, e.g. HindIII, at 37°C overnight while rotating. Any restriction enzyme that produces a 5’ overhang could be used to produce a Hi-C library, but the subsequent quality controls should be adjusted accordingly. To avoid star activity (relaxed specificity of cleavage site recognition) of HindIII, the restriction enzyme concentration is minimized in favor of longer incubation. The percent digestion may be approximated by performing a PCR spanning a specific genomic restriction site both with and without endonuclease treatment. Loss of amplicon signal after endonuclease treatment correlates with digestion efficiency.

2.3. Biotin marking of DNA ends and blunt end ligation

The HindIII enzyme recognizes the sequence: 5’ – AAGCTT – 3’ and cleaves the DNA, leaving a 5’ overhang of: 5’ – AGCT – 3’. This cleavage provides a template for labeling the restriction fragments with biotin-14-dCTP, which will allow for the enrichment of Hi-C ligation products formed from crosslinked restriction fragments. The 5’ overhang is filled in using the Klenow fragment of DNA polymerase I and equimolar amounts of all deoxyribonucleotides with the substitution of biotin-14-dCTP for dCTP. A small aliquot (10-20%) of the sample is not filled in but is otherwise treated the same as the Hi-C library to produce a standard 3C library. This 3C control library is used in subsequent quality control steps.

Both the Klenow fragment and the remaining HindIII enzymes are denatured with SDS while incubating at 65°C. The DNA fragments, which are still crosslinked to one another in chromatin complexes, are then ligated in a dilute reaction at 16°C. This dilute condition favors the intra-molecular ligation of fragments crosslinked within the same chromatin complex instead of ligation between fragments in different chromatin complexes. Since the ligation of these blunt ends is inefficient, the incubation time is set to 4 hours. The ligation of two completely filled in HindIII sites forms a NheI site: 5’ – GCTAGC – 3’. This newly formed NheI site can be used to measure the efficiency of biotin fill-in by cutting ligation junction amplicons with NheI.

2.4. DNA Purification

The chromatin complexes containing the biotin-labeled ligation products are degraded by incubation with Proteinase K at 65°C. Proteins and lipids are extracted from the ligation products with phenol pH 8.0:chloroform (1:1) in phase lock tubes. The DNA is then precipitated from the aqueous phase with sodium acetate pH 5.2 and 100% ethanol under centrifugal force. The precipitate may contain undesirable co-precipitates, such as salt and DTT, which can interfere with downstream molecular assays. To eliminate these co-precipitates, the DNA pellet is resuspended in 1X TE and washed with 1X TE using an Amicon 30Kda molecular weight cutoff column. Contaminating RNA is degraded with RNase A treatment. This procedure yields high-quality ligation products that are ready for downstream applications.

2.5. Quality Control of Hi-C Libraries

The purified sample now consists of biotin-labeled ligation products formed between genomic restriction fragments that were adjacent to one another in physical space. Before proceeding, the library is assayed to ensure that it meets quality metrics. To begin with, the DNA is run on a 0.8% agarose gel to observe the distribution of fragment sizes in the library. The majority of the sample should run above 10kb. Libraries that have a significant amount of degradation, indicated by a smear of lower molecular weight products, should be discarded because we have found that such libraries do not result in quality data. Degradation of Hi-C libraries can occur for several reasons, including: exemption of protease inhibitor cocktail during lysis, overly vigorous homogenization, thermal degradation, or cell type specific effects, including endogenous nucleases. If degradation of libraries persists, a small aliquot can be removed prior to restriction enzyme digestion. The DNA can be extracted and analyzed for degradation via gel electrophoresis. The DNA is also quantified either by agarose gel electrophoresis or spectrophotometrically. Usually, more than 100 μg of DNA should be recovered from 25 million mammalian cells.

To test the formation of ligation products a standard 3C PCR reaction is performed to amplify a ligation product formed by two directly adjacent restriction fragments. Such a ligation product should be readily detected in a 3C or Hi-C ligation product library. The design of 3C primers is discussed elsewhere *(Naumova et al., this issue)*. The resulting 3C amplicon is digested with HindIII and NheI in order to assess the proportion of the library that was filled in and labeled with biotin-14-dCTP (figure 2). The 3C or Hi-C library of ligation products cannot be directly probed with NheI to assess biotin fill-in because of the presence of endogenous NheI restriction sites within the ligated HindIII fragments. The amplicon generated from the 3C control library should digest completely with HindIII and not NheI, because the HindIII restriction site was not filled. In contrast, a significant fraction of the PCR amplicon obtained with the Hi-C library can be digested with NheI. Complete digestion of these PCR products is rarely achieved (figure 2a Lanes 4 and 8). This is attributed mostly to point mutations made at the HindIII cleavage site during the PCR amplification. The use of high-fidelity DNA polymerases in the PCR reaction attenuates this problem. To quantify the fraction of molecules containing the NheI site the digest is run on a gel and the bands are quantified with image analysis software. The ratio of NheI-digested products (figure 2a lane 7) to the proportion of the tandem NheI and HindIII digested products (figure 2a lane 8) is the proportion of the library that has been properly filled in with biotin. This efficiency of biotin incorporation can vary between libraries but is usually 20 – 30 percent. In cases where the filling in efficiency is low (efficiency <5%), the library can be remade with an extended biotin fill-in incubation and an overnight ligation with additional ligase. Increasing ligation time can drive the reaction to ligate more of those fragments that are biotinylated, but one needs to be aware of the fact that this may also increase the frequency of intermolecular ligation of non-cross-linked molecules. The 3C control library is informative if few or no PCR products are formed from the Hi-C library. If little PCR product formed in the 3C control library as well, then the ligation, which is common between the 3C and Hi-C libraries, was inefficient. However, if the 3C control library generates more PCR product than the Hi-C library then the biotin fill-in step was inefficient.

2.6. Biotin Removal From un-ligated Ends

All subsequent steps pertain solely to the Hi-C library and not the 3C control library—Ligation of biotin-marked DNA ends is not as efficient as ligation of staggered ends, and this in a typical Hi-C experiment a significant fraction of DNA ends do not become ligated. The biotin-labeled ends that have not been ligated together must be selectively removed so that they are not pulled down on the streptavidin-coated beads along with true ligation products. The strong 3' -> 5' exonuclease activity of T4 DNA Polymerase is used to remove nucleotides from the ends that have not been ligated together. A high enzyme concentration and low abundance of nucleotides drives the polymerase towards exonuclease activity over polymerization. dATP and dGTP are the only nucleotides added during this reaction to allow balancing nuclease and polymerase activity on the HindIII template strand after Biotin-dC removal which prevents complete degradation of the DNA. This reaction is stopped with the addition of EDTA and then the DNA is extracted with Phenol pH 8.0:Chloroform (1:1) and precipitated with sodium acetate pH 5.2 and 100% ethanol. The reaction is incubated on dry ice for 30 min and the precipitate is collected by centrifugation. The DNA pellets are then resuspended in double distilled water (ddH_20) and the salt in the solution is washed out three times with ddH_2O using a 30Kda Amicon column. It is important to wash the Hi-C library thoroughly to remove salt because the sample will perform better in subsequent enzymatic reactions and during fragmentation and size fractionation.

2.7. DNA Fragmentation and Size Fractionation

The rest of the protocol describes how to modify the biotinylated library of ligation products for sequencing with the HiSeq platform from Illumina—The ideal size of the library will depend on which sequencing platform will be used. A Hi-C ligation product size distribution of 150 – 300bp is optimal for most experiments because this size distribution works well for HiSeq cluster formation.

There are a variety of options available to fractionate DNA. The Covaris 8700 apparatus optimizes the amount of DNA in the relevant size range (figure 2B) and is very reproducible. Regardless of which technology is used, the conditions for each system must be optimized to get the maximum amount of DNA as close to the preferred size for sequencing as possible.

To achieve an even tighter size distribution after fragmentation, AMpure XP mixture is used to fractionate the DNA based on size. This AMpure XP mixture includes Solid Phase Reversible Immobilization (SPRI) paramagnetic beads in a specific solution that precipitates DNA onto the beads. When AMpure XP is added to a sample the DNA will precipitate onto the SPRI beads, which can be recovered with a Magnetic Particle Separator (MPS). The relative proportion of the AMpure XP solution, after addition to the sample, will determine which DNA sizes precipitate onto the SPRI beads. Larger DNA molecules fall out of solution more easily than smaller ones, so less AMpure XP solution is needed to precipitate them. Low binding pipet tips are recommended and accuracy in pipetting is paramount to achieve precise proportions of AMpure XP solution relative to the sample volume such that the correct size fraction is purified.

Ligation products that are larger than 150bp are removed by the addition of 0.9x volumes of AMpure XP solution (figure 2C lane 2). The SPRI beads are harvested with the MPS. This higher molecular weight fraction should be kept in case another larger fraction is needed. To the remaining supernatant, more AMpure XP is added, bringing the relative total volume of AMpure XP solution to 1.1X of the original Hi-C sample. This second addition of AMpure XP is supplemented with additional SPRI beads which were harvested with the MPS to prevent saturation of beads with DNA and incomplete DNA capture. The SPRI beads in the 1.1x fraction bind DNA that is between 150bp and 300bp (figure 2C Lane 3) and the supernatant contains ligation products that are smaller then 100bp. The beads are harvested with the MPS and the supernatant is discarded. The SPRI beads with both the 1.1x fraction and the 0.9x fraction are washed with 70% ethanol to remove any remaining AMpure solution. After air-drying at 37°C, the DNA for both fractions is eluted with Elution Buffer (EB) from Qiagen. The DNA is then analyzed and quantified on an agarose gel to ensure that the size distribution is as expected (figure 2C Lane 3).

2.8. End Repair and ‘A’ Tailing

Shearing the Hi-C library causes asymmetric breaks in the DNA molecules and these broken ends must be repaired before Illumina adapters can be ligated. The Klenow fragment of DNA Polymerase I is used to fill in 5’ overhangs while the T4 DNA Polymerase is used to both fill in 5’ overhangs and to chew back 3’ overhangs. Simultaneously, the T4 Polynucleotide Kinase is used to add a 5’ phosphate to the Hi-C library to allow for ligation of Illumina sequencing adaptors. The Hi-C ligation products are purified from the reaction using Qiagen MinElute columns. Klenow (exo-) is then used to adenylate the 3’ end of the fragment, which will allow for the ligation of the Illumina PE adaptors.

2.9. Streptavidin Pull-down of Biotinylated Hi-C Ligation products

In order to minimize non-specific pull down of DNA during the streptavidin pull-down of Biotinylated Hi-C ligation products, all steps are done in low binding tubes and with low binding pipette tips—The biotinylated Hi-C ligation products are mixed with My-One streptavidin bead solution (Dynabeads). The volume of Dynabeads added should relate to the amount of DNA in the sample. Using a volume of

Dynabead solution in the range of 2 – 5 uL per 1 ug of total DNA (as quantified at the end of the size fractionation step 1.7) should provide an excess of beads relative to DNA. Once the DNA is bound to the beads, they are washed to remove non-specifically binding DNA and the buffer is exchanged with 1X T4 Ligation Buffer from Invitrogen to prepare for ligation of the Illumina adaptors.

2.10. Paired-End Adapter Ligation and Library Amplification

Ligation of the Illumina Paired-end Adapters is while the Hi-C library is bound to the streptavidin beads. The adsorption of the DNA to the beads increases the efficiency of adapter ligation by decreasing the mobility of the DNA fragments and facilitates removal of un-ligated oligonucleotides. After ligation of the adapters, the sample is washed to remove the un-ligated oligonucleotides and to exchange the buffer. The beads are finally resuspended in 20 uL NEB 2.

It is important to PCR amplify the library with as few cycles as possible (9-15 cycles). This ensures linear amplification without creating PCR artifacts (figure 2D arrow), while producing enough product to successfully sequence the library (figure 2D lane 3). It is preferable to pool multiple PCR reactions rather than to increase the number of PCR cycles. A test PCR should be run to titrate the number of cycles needed (figure 2D lanes 2-5). If linear amplification cannot be achieved in the target range of cycles, the volume of beads for each reaction can be adjusted.

After the optimal number of cycles has been determined, multiple PCR reactions are performed to produce the amount of library needed for sequencing (about 50 ng of DNA). Usually, 5 reactions will produce sufficient amounts of DNA. The amplified DNA is then pooled and concentrated, and the primer dimers are removed by using 1.8x AMpure XP as before (step 2.7). The purified DNA can be quantified either with agarose gel electrophoresis or spectrophotometrically. It is advised to verify the size and to accurately quantify the Hi-C library with a Bioanalyzer (Agilent Technologies; Santa Clara, CA) prior to sequencing.

2.11. Final Quality Control and Library Quantification

The efficiency of the removal of biotin in un-ligated restriction fragments in step 2.6 varies between samples. This is linked to variation in ligation efficiency and also may relate to the incorporation of non-junction biotins at small nicks in the DNA sequence. To gauge the fraction of true ligation products in the pulled-down DNA, a small portion of the final amplified library is digested with NheI (figure 2E). This digestion cleaves the library at NheI sites that resulted from the ligation of biotin-labeled restriction fragments. Upon digestion of the library with NheI there should be a shift in the distribution of fragment sizes for the library. The degree to which the distribution shifts is inversely proportional to the number of biotinylated but un-ligated dangling ends in the sample. By quantifying the amount of sample that shifted to a lower molecular weight, the proportion of dangling ends can be approximated. Most libraries typically consist of approximately 10-45% of dangling ends. Libraries that have more than 80% dangling-ends will most likely not be profitable to sequence.

It is important to note that this approach provides only a rough approximation of the proportion of dangling-ends. Endogenous NheI sites present within 500bp of a genomic HindIII site may cause an overestimation of the fraction of digested products. However, there are also true ligation products that cannot be cut, because after shearing a large portion of their length comes from one of the two ligated fragments. When digested, these molecules will not shift appreciably in molecular weight, leading to an underestimation of the proportion of dangling-ends in the Hi-C library. Although these factors may confound the interpretation of the results, in general we have found that this estimate of dangling end percentage is a reliable indicator of the quality of the library (compare figure 2E sample 1 to figure 2E sample 3).

3. Expected Results and Discussion

3.1. Sequencing of Hi-C libraries

In principle, the final Hi-C library of paired fragments can be sequenced using any platform that will allow both ligated sequences to be mapped to the genome, either by long reads that will read through the NheI junction (Roche 454) or by paired-end or mate-paired reads (Illumina GA and HiSeq platforms and Life Technologies SOLiD). To check the quality of a library before high-throughput sequencing, a small subset of library molecules can be cloned and sequenced using traditional Sanger sequencing, where the resulting long sequencing read will pass through the ligation junction, allowing identification of interacting pairs of loci. We have found that Illumina paired-end sequencing with 36 or 50 bp reads is an effective way to identify a large number of interacting fragment pairs. Longer read lengths (75 bp or 100 bp) may improve mappability for repetitive regions. However, with an average library size of 250 bp after sonication and size selection, such longer reads are likely to pass through the ligation junction into the partner fragment. Unless reads are truncated at such ligation junctions (5’-AAGCTAGCTT-3’ in the case of HindIII), these longer reads may be unmappable to the (linear) reference genome, where the ligated fragments are not neighboring. Therefore, we find that 50 bp paired-end reads are optimal for Hi-C library sequencing.

3.2. Sequence Read Mapping and Filtering

The sequenced Hi-C reads can be mapped to the genome with any short read sequence alignment algorithm (Bowtie, Maq, Eland, Novoalign, etc.). Analyzing the position and direction of sequenced reads relative to restriction sites provides information about the types of molecules present in the Hi-C library and the overall quality of the library (Figure 3A). If the two reads occur in the same fragment, they either represent a self-circularized ligation product or an unligated “dangling end” product. Typically, self-circles comprise 0.5-5% of the molecules in a Hi-C library. The proportion of self circles increases with decreased crosslinking because there is less competition for ligation of the two ends of that fragment by other fragments. Thus, changing the formaldehyde concentration or crosslinking time may change the proportion of self-circles in the final library. Dangling ends may comprise 10-45% of a successful Hi-C library. The T4 exonuclease activity during the Biotin Removal step of the protocol (1.6) is essential for minimizing the proportion of dangling ends.

In contrast to the dangling ends and self-circles, valid interaction pairs map to different restriction fragments and face toward the restriction site. If the paired ends of multiple molecules have the same size and map to the same genome position at both ends, only one copy of this interaction should be considered. Such redundant molecules may result from PCR over-amplification. If more than 5% of the library of valid pairs have redudant molecules, concern is raised that the library may have been amplified by too many PCR cycles.

3.3. Data Binning and Normalization

Unique valid interaction pairs (non-redundant, true ligation products) can now be used as a measure of the frequency of physical contact between each pair of loci in the genome. Assuming that every restriction fragment could ligate to any other, there are on the order of

10^{11} possible HindIII restriction fragment pairs in the human genome. Thus, it is difficult to generate a Hi-C library with enough complexity or sequence depth to cover all possible restriction fragment interactions. In order to gain statistical power, it is useful to pool numbers of reads within larger genomic regions before further analyzing the data. Larger bins will contain more reads and thus have more discriminatory power, but at the cost of lowering the resolution of the data. The optimal bin size, and therefore the resolution at which the interaction data can be analyzed, depends on the sequencing depth and the linear separation of the genomic regions under consideration (Figure 3B). Since intra-chromosomal ("*cis*") interactions, are more likely to occur and thus more frequently observed, a smaller bin size may be effective when analyzing interactions within one chromosome. On the other hand, a larger bin size may be necessary to distinguish the less frequent inter-chromosomal ("*trans*") interactions. For mammalian genomes, we find that at least 7 million unique valid pairs are necessary to analyze higher order features like the positioning of whole chromosomes in the nucleus, polymer scaling analyses, or the compartmentalization of genomic regions into open and closed chromatin (see section 3.4). To look at interactions in *cis* between particular genomic bins at a scale of 100 kb, we recommend obtaining more than 100 million unique valid pairs. On the other hand, for a library with 100 million reads total, the minimum bin size appropriate for analyzing *trans* data would still be about 1 Mb.

Different genomic bins may have different biases in terms of mappability, number and length of restriction fragments, etc. [11]. Therefore, it is important to correct the Hi-C interaction map to account for these possible biases. Several approaches have been developed to correct Hi-C data. The method developed by Yaffe and Tanay explicitly defines each bias separately (e.g. fragment length) and then attempts to correct for these biases. Alternatively, we have developed a simpler and more intuitive approach that uses the Hi-C data directly to estimate detection bias without the need to know all factors that could bias the analysis, or the need to explicitly define their impact. Our coverage correction accomplishes this by iteratively dividing the number of interactions between two bins by the product of the number of reads ever observed in each bin across the genome until the sums of counts across each bin the genome are equal. This procedure will also account for the number of reads in the dataset, making it possible to compare such normalized Hi-C scores across experiments. We have found that our coverage correction produces results equivalent to the Yaffe and Tanay correction, with less computation time (*manuscript in preparation*).

Visualization of Hi-C Data and Basic Expected Results

Once the data have been binned and normalized, genome spatial organization can be visualized and analyzed in a variety of ways, according to the goals of each particular study. To observe the patterns of all pairwise interactions, a heatmap matrix of normalized interaction values can be constructed (Figure 4A). In any successful Hi-C experiment, an interaction heatmap should show a strong diagonal of interactions between proximal genomic regions and an overall exponential decay in interaction signal over increasing distance. The thickness of the diagonal relates to the level of chromatin compaction along the chromosome. Each row or column of this heatmap details the interactions between one genomic region and the rest of the chromosome or genome (effectively a "4C profile"). Such interaction profiles of single loci can then be directly compared with 1D genomic datasets, such as ChIP-seq measurements of protein-DNA interactions, to investigate the relationship between 3D conformation and gene regulation. (Figure 4B).

Hi-C data can also be used to infer the polymer structure of chromatin. Different DNA polymer structures (fractal globule, equilibrium globule, etc.) have different predicted relationships between contact probability and linear genomic distance [8, 12], and such relationships can be determined by comparing the number of Hi-C interactions between all pairs of regions with the genomic distance between each pair (Figure 4C).

Hi-C data can also reveal the “compartmentalization” of the genome into regions of open and closed (active and inactive) chromatin, as described previously [8]. Hi-C interactions tend to be much stronger within one such compartment than between compartments. Classifying all genomic regions into compartments summarizes the data as neighborhoods of interaction that can be visualized and compared to other genomic datasets (Figure 4D). By identifying changes in compartment identity after a certain perturbation, Hi-C data can be used to investigate which regions of the genome are rearranged spatially in response to a treatment, pathology, or developmental transition.

Lastly, the relative positions of whole chromosomes in the nucleus can be visualized on a larger scale by comparing the observed number of interactions between each pair of chromosomes to the number of interactions expected according to the representation of each chromosome in the whole dataset (Figure 4E). These results can be compared to the organization of chromosome territories as observed by imaging methods [1].

3.4. Comparison to Other Methods

Once a Hi-C experiment has revealed salient patterns of genome organization, interesting findings can be validated and further investigated using other approaches. As described earlier, imaging spatial relationships between regions of the genome using FISH is not suited to the study of genome wide chromatin structure, but it is a powerful tool for visualizing the relationships between specific genomic loci [3]. A FISH experiment with probes specific to given loci can reveal the 3D distance between loci on a per cell basis. The average 3D distance between loci measured is typically inversely related to the Hi-C signal [8].

If a Hi-C experiment suggests that a certain region of the genome (on the scale of 10 Mb) undergoes particularly interesting changes in conformation between conditions of interest, such a region can be explored at higher resolution using a 5C experiment (*5C Methods, this issue*) [7]. 5C probes can be designed for every possible interaction in the region identified by Hi-C and then the 5C library, generated from a 3C library in a manner similar to Hi-C, can reveal further high-resolution (typically at a resolution of single restriction fragments) details of the spatial organization of this region.

3.5. Conclusion

The highly parallel nature of the Hi-C method produces genome-wide interaction maps. As such, Hi-C provides a unique and powerful tool to study nuclear organization and, chromosome architecture. Thus, Hi-C data will add a spatial context to biological inquires and will facilitate the discovery of the fundamentals of gene regulation, nuclear partitioning, and the biophysics of chromatin dynamics. Hi-C does not capture the fine detail of sub-nuclear compartments like electron microscopy nor can it measure the dynamics of interactions between multiple genomic loci like fluorescent microscopy. It does, however provide the ultimate connectivity between the genomic sequence and spatial conformation. The genome-wide power and versatility of Hi-C makes it ideal for the study of the basic biology of genome organization and its implications for health and disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Gaurav Jain, Amartya Sanyal, and Emily Smith for critical revision of the manuscript as well as Bryan Lajoie for computational support. Supported by grants from the National Institutes of Health (Grant number HG003143 from the National Human Genome Research Institute to JD and Grant Number F32GM100617 from the National Institute Of General Medical Sciences to RPM), and by a W.M. Keck Foundation Distinguished Young Scholar in Medical Research Award (To JD).

References

1. Cremer T, Cremer M. Chromosome territories. Cold Spring Harb Perspect Biol. 2010; 2:a003889. [PubMed: 20300217]
2. Dehghani H, Dellaire G, Bazett-Jones DP. Organization of chromatin in the interphase mammalian cell. Micron. 2005; 36:95–108. [PubMed: 15629642]
3. Solovei I, Cavallo A, Schermelleh L, Jaunin F, Scasselati C, Cmarko D, Cremer C, Fakan S, Cremer T. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). Exp Cell Res. 2002; 276:10–23. [PubMed: 11978004]
4. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295:1306–1311. [PubMed: 11847345]
5. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. [PubMed: 19890323]
6. Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet. 2005; 37:31–40. [PubMed: 15608638]
7. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006; 16:1299–1309. [PubMed: 16954542]
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. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326:289–293. [PubMed: 19815776]
9. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006; 38:1348–1354. [PubMed: 17033623]
10. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, Ohlsson R. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. Nat Genet. 2006; 38:1341–1347. [PubMed: 17033624]
11. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 43:1059–1065. [PubMed: 22001755]
12. Mirny LA. The fractal globule as a model of chromatin architecture in the cell. Chromosome Research. 2011; 19:37–51. [PubMed: 21274616]

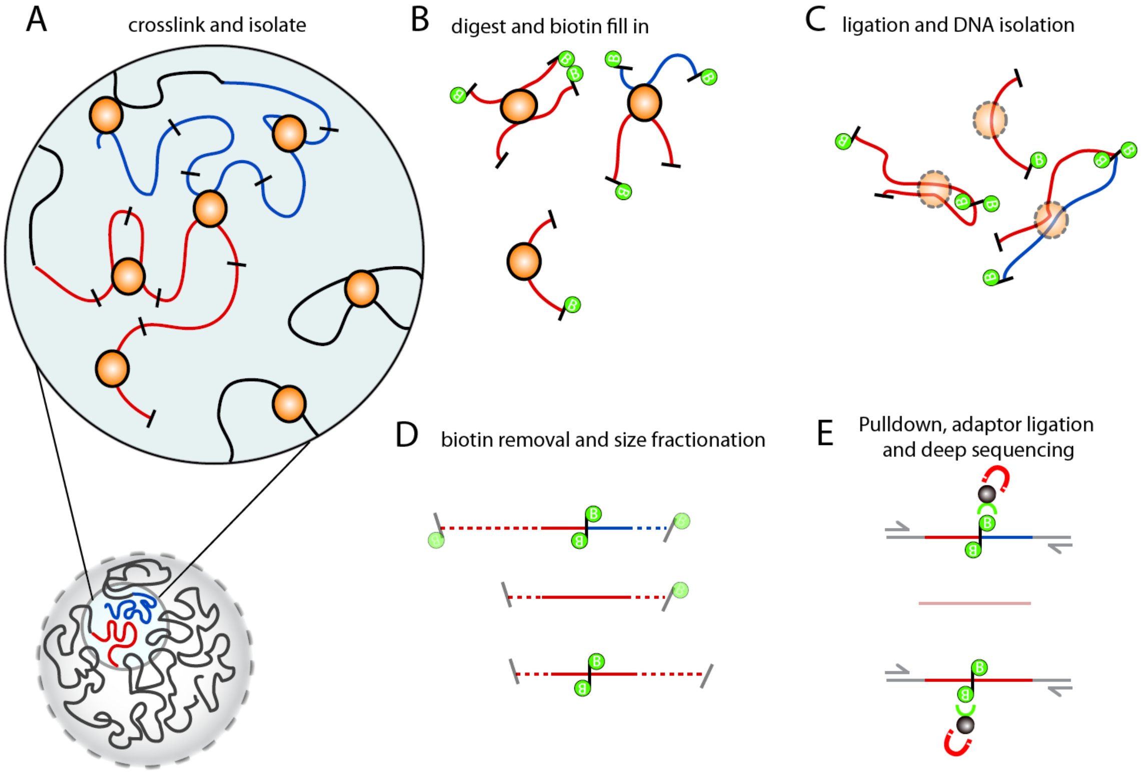
A
crosslink and isolate
B
digest and biotin fill in
C
ligation and DNA isolation
D
biotin removal and size fractionation
E
Pulldown, adaptor ligation
and deep sequencing
B

Figure 1. Overview of Hi-C technology

A) Hi-C detects chromatin interaction both within and between chromosomes by covalently crosslinking protein/DNA complexes with formaldehyde. **B)** The chromatin is digested with a restriction enzyme and the ends are marked with a biotinylated nucleotide. **C)** The DNA in the crosslinked complexes are ligated to form chimeric DNA molecules. **D)** Biotin is removed from the ends of linear fragments and the molecules are fragmented to reduce their overall size. **E)** Molecules with internal biotin incorporation are pulled down with streptavidin coated magnetic beads and modified for deep sequencing. Quantitation of chromatin interactions is achieved through massively parallel deep sequencing.

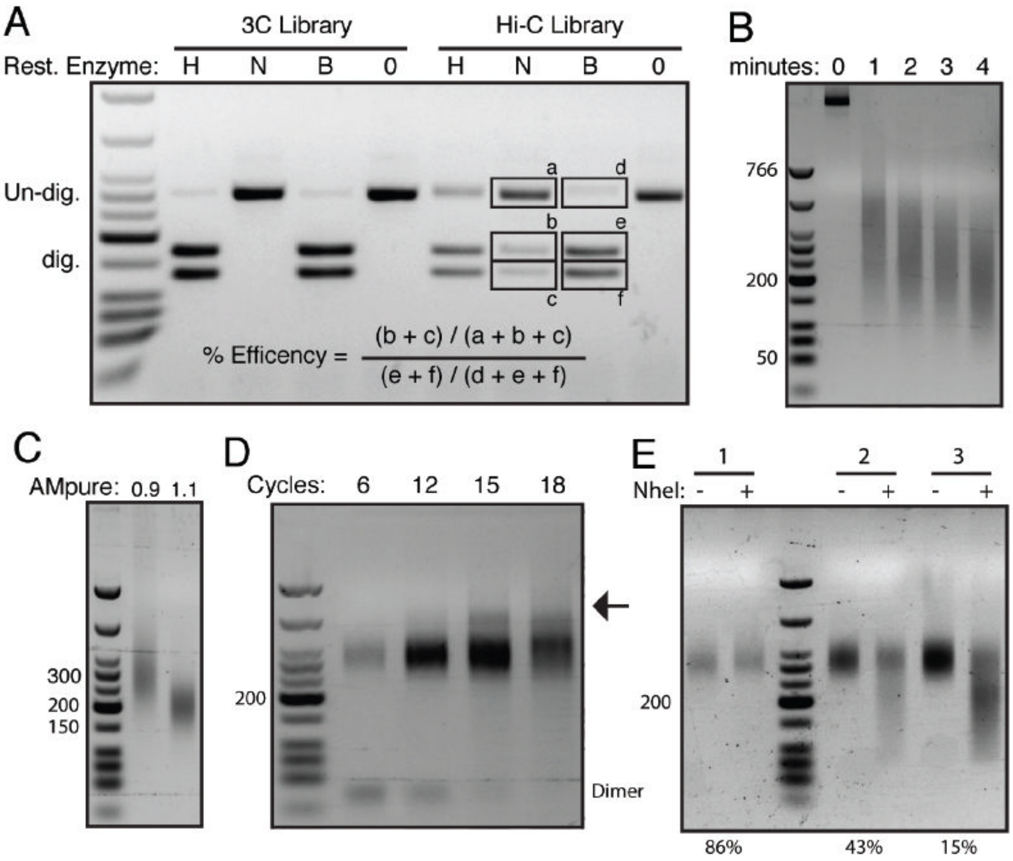

Figure 2. Relative ligation efficiency of Hi-C library

A) Digestion of a PCR amplicon generated from a neighboring pair of restriction fragments in both the 3C sample and the Hi-C sample. The amplicon was digested with HindIII (H), NheI (N), Both HindIII and NheI (B), or not digested (0). The amplicon when digested yields 2 products of different molecular weight. The molecular weight ladder is the Low Molecular Weight Ladder from NEB. **B)** The Hi-C library is fragmented using the Covaris 8700. A titration of fragmentation time is shown, starting with un-fragmented Hi-C library (lane 1) and increasing in the number of minutes of fragmentation (lanes 2-5). At 4 minutes (lane 5) the distribution of fragment sizes is ~50bp – 600bp. **C)** AMpure XP is used to fractionate the library. The 0.9x AMpure XP fraction includes molecules that are larger then ~150bp and the 1.1x fraction includes molecules that are between ~150bp – 300bp. **D)** The Illumina Paired-end graft sequences are added to the Illumina adapter modified Hi-C libraries using PCR with primers PE 1.0 and PE 2.0. These primers are partially homologous to the PE adapter, which was ligated to the Hi-C library. The gel shows a titration of the number of cycles. At higher numbers of PCR cycles, higher molecular weight artifacts are produced (arrow). Sufficient amounts of DNA are produced at 12 cycles for this library. **E)** Three completed Hi-C libraries were digested with NheI. The shift of the size distribution of the library following digestion with NheI estimates the proportion of the library that consists of real Hi-C ligation products. A range of performances are shown, with library 1 showing poor performance, library 2 showing medium performance and library 3 showing good performance. The percentages below the each library are the percentages of dangling-ends that were tabulated after sequencing these libraries and mapping the reads to the genome.

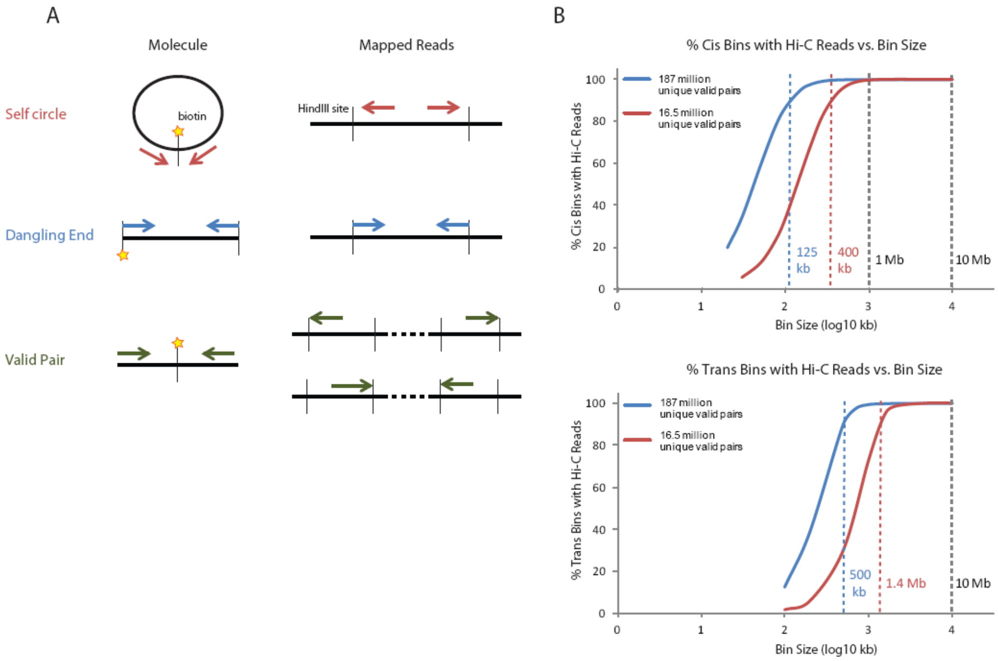

Fig 3. Hi-C sequence mapping and binning considerations

A) Different types of molecules in the Hi-C library (left) lead to different orientations of mapped reads relative to restriction sites (right). Mapped reads (colored arrows) facing outward in the same fragment come from self-circles (top); Reads facing inward in the same fragment arise from dangling ends (middle); Reads from different restriction fragments and facing toward a restriction site arise from valid interaction pairs (bottom). **B)** Relationship between sequencing depth and choice of bin size. Each graph shows the percentage of *cis* (top) or *trans* (bottom) bins that contain at least one mapped read from a valid interaction pair (y-axis) for each different bin size (x-axis). Colored dotted lines indicate the bin size at which 90% of bins contain at least one valid pair read for a Hi-C library with a high (blue) or low (red) number of total unique valid pairs after sequencing.

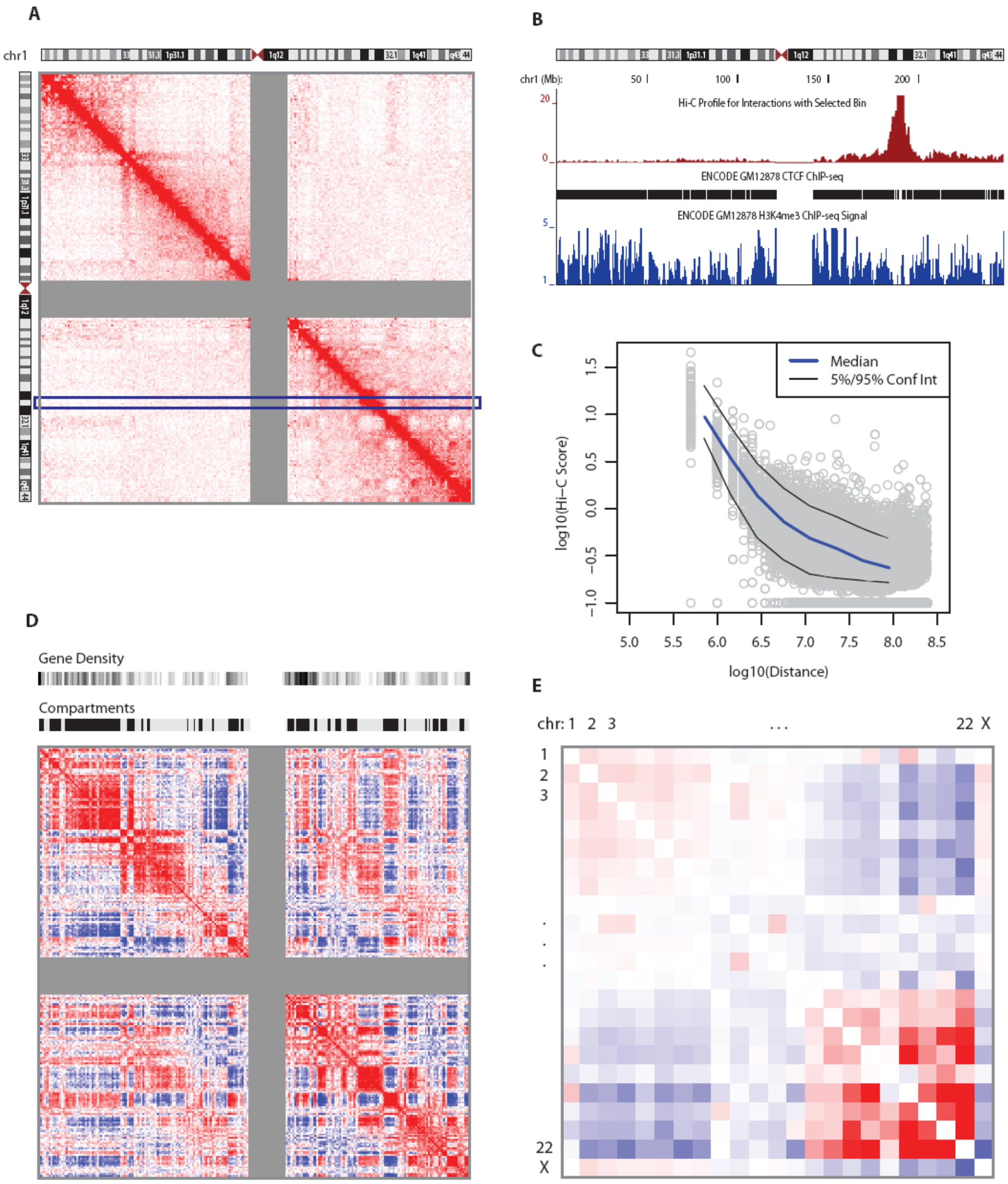

Fig 4. Hi-C data visualization and analysis

A) A heatmap of interactions between all 1 Mb bins along chr1 for GM06990 cells. The intensity of red color corresponds to the number of Hi-C interactions. **B)** A “4C profile” derived from one row of the Hi-C heatmap (blue box in A) showing all interactions between a fixed 1 Mb location at 190 Mb on chr1 and the rest of chr1. CTCF and H3K4me3 tracks from a similar cell line are displayed below as examples of other genomic datasets that can be compared with such an interaction profile. **C)** The log10 of the Hi-C interaction counts of each pair of bins along chr1 is plotted versus the log of the genomic distance between each pair of bins. The median value of datapoints in the graph is indicated by a blue line while the 5% and 95% confidence intervals are shown as thin black lines. The slope of the median line from 500 kb to 10 Mb is -1, following the relationship expected for a fractal globule polymer structure of the chromatin. **D)** Red and blue “plaid” patterns show the compartmentalization of chr1 in two types of chromosomal domains. The data from A were transformed by first finding the observed interactions over the expected average pattern of decay away from the diagonal and then calculating a Pearson correlation coefficient between each pair of rows and columns. Regions highly correlated with one another in interaction are colored red and are likely to be classified by principle components analysis into the same compartment as shown above (black bands = open chromatin compartment; light grey bands = closed chromatin compartment). The compartment assignments correlate with the gene density profile, shown above the compartment profile (high gene density = black; low gene density = white). **E)** Whole chromosome interaction patterns show that longer chromosomes (chr1-10, chrX) are more likely to interact with one another and not with shorter chromosomes (chr14-22). The observed number of interactions between any pair of chromosomes is divided by the expected number of interactions between those chromosomes given the total number of reads for either chromosome in the whole experiment. Red indicates an enrichment of interaction as compared to expected values while blue indicates a depletion of interactions between two chromosomes.