How best to identify chromosomal interactions: a comparison of approaches

James O. J. Davies[1], A. Marieke Oudelaar[1], Douglas R. Higgs[1] and Jim R. Hughes[1]

[1]Medical Research Council (MRC) Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, Oxford University, Oxford, United Kingdom.

Summary (100 words)

Chromosome conformation capture (3C) methods are central to understanding the link between nuclear structure and function and the physical interactions between distal regulatory elements and promoters. However, no one method is appropriate for all biological questions as each variant differs markedly in resolution, reproducibility, throughput and biases. A thorough appreciation of the strengths and weaknesses of each technique is critical when choosing the correct method for a specific application or for gauging how best to interpret different sources of data. In addition, the method of analysis can have a profound effect on the output and must be carefully considered.

Preface (100 words)

The location of potential regulatory elements can now be routinely and rapidly mapped genome-wide using chromatin accessibility and ChIP-seq based assays[1-4], however an outstanding challenge is to determine which regulatory elements control which genes[5]. This is problematic because regulatory elements are scattered over large distances (up to millions of base pairs) around the genes they regulate[6]. In order to influence transcription, the protein complexes at distal regulatory elements, such as enhancers, make physical contact with gene promoters. These interactions can be detected using chromosome conformation capture (3C) technology and form the basis for identifying a gene's regulatory elements.

There is a growing body of evidence that genomic organization within the nucleus has an important role in determining gene expression patterns[7]. Although a large amount has been learnt about nuclear organization from microscopic study of the nucleus, particularly with fluorescence in situ hybridization (FISH), the majority of recent discoveries regarding sequence specific chromatin structure have been made using 3C technology. These include the description of large scale (>100kb) structures such as topologically associated domains (TADs)[8,9] and interactions on a much smaller scale such as those with regulatory elements[10] and within the gene body itself[11,12].

The first assays that used a restriction enzyme digest followed by a ligation step to detect physical interactions between pieces of DNA were performed in the late 1980s to show that DNA-DNA interactions occur in plasmids[13]. However, it was not until the seminal paper by Dekker et al.,[14] that it became possible to define interactions between two specific DNA sequences in eukaryotic cells, and studies demonstrating interactions between regulatory elements in mammalian genomes quickly followed[10]. The assays are all based on the principle that chromatin interactions can be crosslinked, cut and then re-ligated so that sequences in close physical proximity within the nucleus become linked to each other. The ligation junctions that result thus reflect the three-dimensional organization of the genome at time of fixation and can therefore be used to infer chromatin structure (**Figure 1a**).

The library of re-ligated fragments (3C library) could in principle be interrogated by simply sequencing the rearranged genomic DNA using next-generation sequencing technologies. This has been done for small genomes (Sexton et al. 2012), but is not feasible for high-complexity mammalian genomes. The different 3C methods can be seen as approaches to sample the ligation events from specific regions, or sequence only the informative ligation junctions to overcome the complexity of large genomes.

When the 3C library is being manufactured each restriction fragment in every individual cell has many potential ligation partners (**Figure 1b**)[15]. Live cell

imaging studies show that chromatin is in a state of constant flux within the nucleus[15,16] and single-cell 3C assays, therefore, show huge variability in the ligation junctions between cells[17]. Thus the ligation junctions present within the 3C library represent the probability distribution of all possible ligation junctions from a group of cells, in a particular state. It is therefore vital to accurately quantify the ligation junctions present in the 3C library and with significant depth of sequencing in order to build up a profile of the probability distribution underlying the ligation junctions (**Figure 1c**). Simply stating that a junction can or cannot be found is not meaningful because if one samples the library at sufficient depth then it is likely that a ligation junction could be found with most other fragments in the genome. It is therefore important to understand if a dataset has sampled the interactions to sufficient depth to be reproducible or whether it is sporadically detecting interactions from within the regulatory landscape. This depends on the sensitivity of the assay, which is determined by the number of unique ligation junctions that can be enriched for a given region of interest in the genome.

3C library considerations

The first step of most 3C techniques is to generate a 3C library (**Figure 2**). Generally cells are initially fixed using formaldehyde crosslinking. Recently protocols have been published that allow the digestion and ligation reactions to be performed without the disruption of the nucleus ('in situ' 3C library manufacture)[18], which has considerable advantages for determining accurate interaction profiles[18,19]. It is also possible to manufacture 3C libraries without this fixation step by embedding the cells in agarose plugs, in order to maintain the nuclear shape and structure. Importantly, this results in similar interaction profiles to those generated from formaldehyde fixed material (although there is greater background noise), showing that the interactions detected are not simply an artifact of the fixation process itself[18].

In most 3C techniques the chromatin is cut using a restriction enzyme. The choice of restriction enzyme determines the maximum resolution of the data because interactions can only be detected at restriction enzyme cut sites (**Figure 3**). 3C libraries were initially manufactured using six-cutter restriction enzymes, which limited the potential resolution since these enzymes cut at approximately 4 kb intervals (95% of fragments <12.3 kb). Subsequently four-cutter restriction enzymes have been used, which cut on average every 256bp (95% of fragments <800bp), potentially generating data with 16 times greater resolution.

The complexity of 3C libraries is related to the square of the number of restriction fragments in the genome[20]. Thus increasing the resolution (by using an enzyme that cuts more frequently) or the genome size has dramatic effects on the sequencing requirements. 3C library complexity is often a limiting factor, particularly when investigating the ligation events from specific restriction fragments of interest, as only four interactions can be determined from each

fragment in each cell (one from the end of the fragment on each allele). In practical terms the critical factors that determine library complexity are the initial numbers of cells used, the digestion and ligation efficiency and the cumulative loss of material from each step before sequencing.

When library complexity and / or sequencing depth are insufficient to analyze the data at the level of individual restriction fragments, the resolution of 3C-based experiments is determined by the bin size used to view the data. Windowing or binning improves the signal strength and reduces biases by combining all of the data in the window together, which allows a meaningful interaction profile to be determined when the raw signal is weak. However, windowing has a number of disadvantages. The profile becomes skewed by the restriction enzyme cut site density and it smoothens the profile of the original signal, obscuring the quality of the underlying data and reducing resolution. Ideally, there should be sufficient signal strength and reproducibility at the level of individual restriction fragments, to allow the data to be reported without windowing.

Chromosome conformation capture (3C)

Initially ligation junctions were detected in 3C libraries by polymerase chain reaction (PCR) followed by gel electrophoresis, which was subsequently replaced by real-time PCR (**Figure 2**).

Many important discoveries were made using 3C. The technique was initially used to define the spatial organization of chromosomes in yeast[14]. Interactions at the murine β-globin locus were determined soon after[10]. Subsequently 3C has been used to analyze many other loci, including the T-helper type 2 cytokine locus[21], the immunoglobulin light chain locus[22] and the alpha globin locus[23-25]. However, conventional 3C has largely been superseded by other methods because it is laborious and as it is only able to detect interactions between a small number of fixed restriction fragments (**Figure 3**). This approach focuses researchers on confirming suspected interactions rather than identifying new interactions in an unbiased way. In addition, it can be difficult to obtain reproducible results, and correction for biases due to differences in amplification with different primer pairs requires great care.

Circular chromosome conformation capture (4C)

The next major advance in the field was the development of 4C, which allows all of the potential interacting partners to be identified from any specific point of interest in the genome[26,27]. 4C uses standard 3C library preparation, often with a 4-cutter restriction enzyme, after which inverse PCR is used from the 'viewpoint' fragment to amplify any interacting partners (**Figure 2**). Initially the interacting partners were read out using a microarray, but this has now been replaced by high-throughput sequencing.

For inverse PCR to work efficiently, small circularized fragments must be generated either by optimizing the initial ligation reaction[26] or by the more widely used method of performing a second restriction enzyme digest and ligation of the extracted DNA with a different restriction enzyme[27]. The inverse PCR in 4C data is affected by the GC content and size of the interacting fragments to the extent that some large GC-rich fragments simply fail to amplify. This can result in serious problems with bias in the interaction profile[28] and make the interactions appear more discrete than they really are **(Figure 4)**. 4C provides good resolution and sensitivity, when carefully performed, allowing very long-range interactions to be detected. In order to generate high-quality 4C profiles, it is necessary to perform multiple PCRs to generate material from the DNA of a million or more cells[29]. However, it is difficult to determine the sensitivity of conventional 4C accurately, as it is not possible to differentiate between PCR duplicates and unique ligation junctions.

In a recently developed one vs all protocol (described as UMI-4C[30]), the 3C library is sonicated and sequencing adapters are ligated to one end of each sonicated 3C fragment. Using a primer in the target sequence and a universal adapter primer, interacting fragments can be amplified and sequenced. Such interactions can be quantified by using the unique fragment ends produced by the sonication step and this shows that about 5000-10,000 unique read pairs can be detected per μg of input material. UMI-4C also allows for multiplexing up to 20-50 baits, though the individual nested PCR reactions still need to be optimized for each viewpoint[30].

A large number of seminal observations have been made using 4C techniques. It has been widely used to describe interactions between promoters and enhancers and to show how these change during differentiation and development[31]. Very long-range interactions between active genes have been demonstrated in *cis* using 4C[32]. 4C has also been used to link potential regulatory single nucleotide polymorphisms (SNPs) with genes[33]. It can be used to determine structural changes such as chromosomal rearrangements[34] although this is probably more easily done using conventional karyotype analysis and FISH. It has also been used to define disease mechanisms, including the demonstration that chromosomal translocations can result in distal enhancers becoming juxtaposed to an oncogene leading to malignancy[35].

Chromosome Conformation Capture Carbon Copy (5C)

5C can be used to study all interactions within a particular locus and is based on the use of highly multiplexed ligation-mediated amplification in the region of interest. It uses primer pairs that anneal on either side of the ligation junctions in a 3C library. These primer pairs are ligated and can then be amplified using T7 and T3 sequences incorporated in them in a single PCR reaction, which can be analyzed by microarray or sequencing[36] (**Figure 2**).

Differences in the hybridization efficiency of the probes can cause bias in 5C. In addition it is only possible to determine interactions between forward and reverse probes and it is impossible to design probes to the ends of some fragments, so the interaction profiles can have large areas, which are not interrogated (**Figure 4**).

5C can define functional interactions for all the genes in a locus simultaneously, but it is difficult to be precise about the sensitivity of the method as the levels of PCR duplication cannot be determined. We find that 5C occasionally misses weak, long-range interactions, which are detectable by Hi-C, 4C and Capture-C (**Figure 5**).

5C has been used to determine interaction profiles at the pilot regions of the ENCODE project[6]. In addition analysis of the X-chromosome by 5C provided some of the first evidence for the existence of topologically associating domains (TADs)[9]. Massively multiplexed 5C can generate all vs all interaction maps of a region of interest but it requires significant financial resources for somewhat limited resolution and scale. For example, 5826 oligonucleotides were required to cover 1% of the genome (for the ENCODE pilot regions) and only two replicates of three different cell lines were analyzed[6].

Capture-C

Capture-C is able to generate genome-wide interaction profiles from hundreds of viewpoints in a single assay[37]. It combines 3C library preparation with a four-cutter restriction enzyme and oligonucleotide capture technology. The 3C libraries are enriched for fragments of interest using biotinylated capture probes designed for each viewpoint, after which these fragments are amplified and sequenced.

Although Capture-C involves several rounds of PCR amplification this does not cause much bias, in contrast to 4C. This is because the 3C library is sonicated prior to the PCR, resulting in uniform small fragments with random unique ends that allow PCR duplicates to be collapsed during data analysis.

While capable of producing hundreds of informative, genome-wide tracks in a single experiment, initially the individual tracks themselves did not have the depth of data of a good single 4C experiment from the same region[37] (**Figure 2**). Next Generation (NG) Capture-C was subsequently developed[15]. This uses a new, more flexible and efficient oligonucleotide capture process that markedly increases the sensitivity of the assay. In high-quality datasets, more than 100,000 read pairs commonly contribute to the interaction profile. This makes NG Capture-C the most sensitive and highest resolution assay currently available for mammalian genomes, allowing the data to be expressed at maximum resolution (per restriction fragment) (**Figure 3**). This high sensitivity combined with the reliability of the technique allow weak long-range *cis* and even *trans* interactions to be quantified, which is important because it provides confidence that stronger, functional interactions are not being missed. It also makes it possible to

adapt the technique to small cell numbers. In addition, several independent 3C libraries (e.g. from different cell types or different stages of development) can be processed in a single tube. This greatly increases throughput and allows meaningful subtractive analysis of chromosome conformation of multiple replicates in different cell types (**Figure 4&5**).

Hi-C

Hi-C generates maps of interaction between all parts of the genome (all against all). This technique uses a modified method for 3C library preparation. A biotinylated nucleotide fill-in is performed after the restriction enzyme digestion and followed by blunt end ligation. After DNA extraction and sonication of the material, a streptavidin bead pull-down is performed to concentrate ligation junctions, which can then be analyzed by high throughput sequencing (**Figure 2**).

Hi-C has great sensitivity for determining megabase scale interactions and is unparalleled in its ability to determine large-scale chromatin structure. This is because it combines data from many restriction fragments to define robust interactions on a chromosomal scale. However, the number of interactions determined from any individual restriction fragment are around 100-fold lower than in 4C or Capture-C, even in the recent Hi-C datasets at kilobase resolution[18]. Unless Hi-C is done at very high resolution it is a relatively insensitive method to determine fine-scale (< 40 kb) interactions between regulatory elements present within TADs.

Hi-C has relatively few biases, but is still affected by systematic biases relating to the distance between restriction sites, GC content and sequence uniqueness and several methods have been developed to attempt to correct these[38-40].

Since it was initially reported[20], Hi-C experiments have produced data at increasing resolution predominantly through massive increases in sequencing depth. Initially 40kb bins were used[20] but this was improved to 5-10kb[41] and more recently 1kb bins have been reported in human cell lines[18]. Producing Hi-C data of this quality for large mammalian genomes requires a gargantuan effort involving the analysis of several billion read pairs per sample.

Hi-C has been used extensively to define TADs genome wide[8] and to determine the structure of the chromosome during mitosis[42]. Hi-C has also been used to link trans interactions with sites associated with chromosomal translocations[43]. At present this is the best available technique for determining genome wide (all vs all) maps of interactions.

Hi-C interactions have also been determined from single cells[17] by picking and sequencing single intact nuclei during Hi-C library preparation. This showed that there is considerable variability between individual cells, but the data are not informative at the level of regulatory features.

Hi-C variants

A further increase in resolution has recently been achieved in S. Cerevisiae by substituting restriction enzymes with micrococcal nuclease (MNase)[12]. This shows similar but not identical maps of interaction compared to Hi-C. Long-range interactions are poorly captured by Micro-C. It shows that ~90% of all interactions are within 1 kb, which is a blind spot of current approaches. However, it is not clear at the moment how applicable this will be in larger genomes as smaller fragment sizes could increase the need for sequence depth beyond that of the current high-resolution Hi-C protocols. Similarly DNase I has been used in place of a restriction enzyme[44]. Enzymes that can cut at any point in the genome could theoretically improve the resolution down to a single base pair. This means that the resolution of Hi-C type experiments in mammalian genomes becomes constrained by sequence depth rather than cut site density. Interestingly, using DNase I did not significantly improve the resolution over four-cutter restriction enzymes in mammalian cells (5 kb windows were used) but it is likely that enzymes with higher cut site density will be key to improving resolution in the future.

The Capture-C approach can be used to enrich both 3C and Hi-C libraries. The combination of Capture-C and Hi-C libraries (Capture Hi-C) can exclude further uninformative background from captures within the target fragment rather than spanning a ligation junction. Compared to the high levels of enrichment of NG Capture-C this ~2 fold benefit seems negligible for small to moderate size designs (hundreds of viewpoints) and must be balanced against the more extended protocol and extra loses in complexity of the library, particularly for lower cell numbers. However, it is potentially beneficial when attempting ambitious designs. A low-resolution design using six cutter enzymes has been attempted for every annotated promoter in the human genome, however the very low levels of enrichment (a mean of 10-fold) combined with the ambitious design reduced the number of unique read pairs from each viewpoint, so that that individual profiles are weakly sampled[45].

A tiled oligonucleotide capture approach has been attempted to define an all vs all map of interaction at the beta globin locus[46]. A potential pitfall of this approach is that the efficiency of the oligonucleotide capture will partially determine the interaction profile and robust methods of correcting this have not been established.

Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)

ChIA-PET combines chromatin immunoprecipitation (ChIP) with 3C, potentially allowing all of the interactions to be identified from all of the sites bound by a protein of interest. In this technique the material is initially crosslinked and then sonicated, after which ChIP is performed to enrich for DNA-protein complexes. DNA linkers are then ligated to the immunoprecipitated material and these are used to join sequences held in close proximity. The linkers are biotinylated and

contain MmeI restriction sites allowing short fragments of material to be extracted with a streptavidin bead pull-down. Paired-end tags (PETs) can then be identified by paired-end sequencing. ChIA-PET studies have been performed using ChIP for the oestrogen-receptor-alpha[47], CTCF[48] and PolII[48].

The technique has limitations because the relatively low levels of enrichment of ChIP reduce the library complexity and so the number of reads used to identify individual interactions is usually extremely low (**Figure 4**). It is thought that if three read pairs are present, this corresponds to a false discovery rate of <0.05[47] and this is often used to define significant interactions. By comparison the other 3C-based techniques define interactions based on thousands of read pairs[47]. In addition, ChIA-PET is considered vulnerable to bias towards detecting interactions between sequences that bind the protein targeted by the antibody used for the immunoprecipitation.

Data analysis

Data analysis of sequencing-based 3C approaches is complex (**Figure 6**). It is difficult to compare the quality of the different methods, because customized data processing is often used, which makes it challenging to relate the output back to the quality of the raw data generated by the experiment.

Several approaches attempt to determine interactions that are significantly different from a baseline interaction profile. This is often done using distance modeling approaches. The background interaction profile is inversely related to the distance from the viewpoint. These methods allow significant interactions to stand out from the baseline profile by increasing the weight of more distal interaction counts. This can be done either using a theoretical model of expected baseline interactions (common in 4C) or a statistical distribution based on all of the distribution of counts with distance across the whole dataset (e.g LOWESS in 5C). The potential problem with this approach is that very strong functional interactions commonly occur between neighboring sequences in the chromatin fiber and identification of these can be highly dependent on the tool or parameters used. Measurement of diffusion within chromatin in live cells, suggests that the contact probability decreases with the fourth power of the distance along the chromatin fiber i.e. a 2-fold increase in distance results in a 16-fold reduction in contact probability[16]. Thus it seems inappropriate to disregard interactions with regulatory elements close to the promoter (where they are optimally placed to interact with it) relative to more distal elements that have much lower contact frequencies.

A complementary approach that avoids distance modeling uses the tissue specificity of regulatory elements to identify their activity dependent interaction profile. At its most simple this relies on the comparison of the interaction profile in a tissue where the target gene is inactive with a tissue in which it is active to determine which interactions only appear in the active state. This approach has been shown to be effective at identifying interactions with very proximal

regulatory elements as well as distal interactions, but would obviously ignore interactions that are common between the two tissues.

Overall, careful thought needs to be given to the common basic steps as well as to the most appropriate downstream analysis for the biological question and the use of different combinations of approaches is advisable to understand how robust any specific interpretation may be. It would seem that, in addition to highly processed data, routine reporting of the raw read counts used to substantiate any given interaction and a clear description of understood sources of bias would greatly aid transparency in the field.

Choice of method

Considering the number of variations of the 3C-based methods, there is often confusion over which method is appropriate to which question. Hi-C is unique in its ability to determine genome wide (all against all) interaction profiles and define the landscape of whole genomes to globally determine basic rules of genome organization[8,18,20,42]. However, in order to define the details of small scale interactions that dictate regulation of individual genes Hi-C needs to be performed at the highest resolution currently available, which requires several billion reads per sample. Thus for investigating the detailed interactions for small numbers of genes in multiple samples it would be more appropriate to use 4C or NG Capture-C, which can generate interaction profiles for single loci at higher depth using only ~1 million reads per viewpoint. Of these two methods, the more recently developed NG Capture-C approach would seem the most general solution even for simple designs, considering the control of PCR duplicates and ability to multiplex both samples and viewpoints simultaneously. With the challenges of genome-wide association studies (GWAS) analysis in mind, NG Capture-C was specifically designed to determine the regulatory landscapes of hundreds of genes or regulatory elements simultaneously from the small numbers of cells available from primary tissues.

Approaches, such as 5C, can be used for determining the all vs all chromatin structure of targeted sections of the genome. Although several seminal observations have been made with 5C (which predates Hi-C), the need to manufacture large numbers of probes with 5C and the falling costs of sequencing means that Hi-C, which was developed by the same group, is increasingly used in its place.

Future directions

3C-based technologies have the potential to overcome the barriers caused by our limited understanding of how genome structure relates to its function. The “all vs all” technologies are driving research into the general mechanisms that dictate chromatin structure, while the targeted technologies provide high-resolution approaches to dissect the regulation of specific loci. Increasingly these technologies are being adopted by non-specialist laboratories. It is therefore

important that the appropriate applications and limitations of each method are clear, as well as the sources of technological and bioinformatic biases that might confound correct data interpretation. A general concern for data interpretation of the 3C technologies are potential biases due to the common formaldehyde crosslinking step, which could result in a systematic skew. Though it has been shown that crosslinking does not affect the gross patterns detected[18], the effects at a finer scale are unclear and this potential problem is being actively addressed. Another general source of bias is the uneven degree of digestion across the genome when restriction enzymes are used to fragment the chromatin. To date it has been impractical to effectively map these efficiencies genome-wide, particularly at high resolution, due to the depth of sequencing required. However, as sequencing costs drop it would be inevitable that such maps will be generated and the effect on 3C signal determined. As these potential biases are systematically determined, they can be incorporated for normalization into increasingly sophisticated tools for data analysis.

The other crucial development needed for accurate data interpretation is the ability to interpret 3C data in its three-dimensional context in the nucleus. At the moment 3C data are mostly represented in two dimensions, but expertise from fields such as polymer physics is increasingly being used to try to understand what the patterns seen in 3C assays are reflecting in terms of three-dimensional structures within the nucleus[49,50]. This will be important to better understand the processes that shape chromatin structures and how they in turn may shape the activity of the genome. This drive will undoubtedly bring in other methodologies and expertise such as super-resolution imaging and live cell chromatin tagging, to view these structures dynamically at the single-cell level. These approaches will help refine and inform the models derived from the 3C data.

Irrespective of the future developments required to fully understand the relationship between nuclear structure and function, it is clear that chromatin conformation assays (especially the 4C and Capture-C approaches) have reached a level of maturity and approachability that has turned them into general tools, similar to ChIP-seq or ATAC-seq, to analyze regulatory landscapes. Although impossible to predict the full impact of these methods, investigation of mutations in non-coding regulatory or structural elements, particularly those identified in GWAS studies, will clearly benefit from such analyses.

It is now crucial that the field develops and agrees on standardized data formats and standards for quality control to promote unbiased and comparable interpretation.

Figure Legends

Figure 1. Common principles in 3C-based techniques.

a. The chromatin fiber is initially digested into short restriction fragments, after which a ligation reaction is performed to create large DNA concatemers in which the order of the fragments reflects the three dimensional structure of the chromatin at the time of fixation.

b. Only approximately 1 in 20 (~5%) of the restriction fragments ligate back to their original partner in a DpnII digested library[15]. If one assumes that the cut restriction site has the highest probability of ligating back to its original partner in the fixed nucleus, then there are at least 20 potential ligation partners for every cut site in every cell.

c. The determined interaction frequencies from 3C-based experiments such as 4C and Capture-C can be represented as an interaction profile from a single viewpoint. All vs all interaction data generated by Hi-C like approaches are usually represented in heat maps.

Figure 2. Comparison of different 3C-based methodologies.

3C libraries are generated by initially crosslinking the chromatin using formaldehyde, after which a restriction enzyme digest is performed. This is followed by a ligation reaction, after which the crosslinking is reversed and the DNA extracted.

3C libraries can be interrogated by several different techniques. In the initial 3C experiments ligation junctions were detected by PCR amplification combined with gel electrophoresis or real time PCR.

In 4C, 3C libraries are circularized by either optimizing the ligation reaction or performing a second restriction enzyme and ligase reaction. Inverse PCR is then performed and the product of this reaction is sequenced.

5C uses a highly multiplexed ligation-mediated PCR to amplify ligation junctions in 3C libraries, which are then measured by sequencing.

Capture-C involves sonication of the 3C library, followed by ligation of barcoded sequencing adaptors, which allows multiple samples to be mixed at this stage. A hybridization reaction with biotinylated oligonucleotides targeted against the viewpoints of interest and biotin pull-down are subsequently performed (this is repeated in NG Capture-C to increase enrichment), after which the material can be sequenced.

In the 3C library preparation in Hi-C experiments, overhanging restriction cut sites are filled in using biotinylated nucleotides, followed by a blunt-end ligation reaction. The material is sonicated and a biotin/streptavidin pull-down is performed to concentrate ligation junctions, which are then identified by sequencing.

The ChIA-PET protocol also starts with a crosslinked and digested 3C library. Next, this material undergoes a chromatin immunoprecipitation step using an

antibody against a transcription factor or chromatin associated protein. Biotinylated linkers are ligated to the immunoprecipitated chromatin, after which these linkers are ligated to one another. An MmeI digest is then performed and the material containing the linkers is pulled down and sequenced.

Figure 3. Comparison of different 3C-based techniques.

a. The maximum resolution is determined by the frequency of cutting of the chromatin when an enzyme that cuts at fixed intervals is used (restriction enzyme with 6 bp or 4 bp recognition or MNase). Since DNase I and sonication (ChIA-PET) do not have specific cut sites they could theoretically generate very high resolution data. Techniques that report data for each ligation junction (e.g. 3C, 5C and NG Capture-C) can be analysed at the maximum resolution determined by the cut site density. However, the majority of 3C-based techniques require windowing or binning of data to generate sufficient data for meaningful profiles because this improves the sensitivity and reduces bias. This reduces the resolution and the window/bin size becomes the major determinant of resolution.

b. Comparison of the sensitivity of the different techniques. The sensitivity is largely determined by the number of reads from each individual viewpoint although techniques such as Hi-C, which undertake an all vs all approach, can improve the sensitivity markedly by combining data from multiple view points.

c. Comparison of the multiplexing ability of the different techniques. 4C, Capture-C and Hi-C are all capable of reporting genome-wide interactions. However the number of viewpoints differs radically between the methods. 4C generates genome-wide profiles from 1 viewpoint, whereas Hi-C uses an all vs all approach. The highest number of viewpoints currently used in Capture-C experiments is 450, but there is no theoretical limit and more viewpoints could be used if desired. Importantly, NG Capture-C can also analyze multiple samples simultaneously.

Figure 4. Case study 1: regulation of the alpha globin locus in erythoid cells.

The human alpha globin genes (*HBA1* and *HBA2*) are regulated by four enhancers (R1-R4) located ~10-40kb upstream of the gene within the introns of the neighboring gene *NPRL3*. These elements are characterized by DNase I hypersensitivity (see **Supplementary Fig. 1** for further chromatin data).

a. Comparison of ChIA-PET, 5C and Hi-C data in the human alpha globin locus in K562 cells. The top track shows ChIA-PET data derived from a ChIP for RNA polymerase II[51]. All of the reads within a 20kb region around the two alpha globin genes are included. The numbers on the reads denote the number of paired end tags (PETs) contributing to the reported interaction. Note that despite the generous inclusion criteria, the interactions with the enhancers are defined by only 16 PETs. The next profile shows 5C data from the two alpha globin promoters[6]. Note that there are limited numbers of reported fragments

because it is only possible to determine interactions between forward and reverse probes and the limited resolution of data generated by a 6-cutter restriction enzyme (HindIII). In addition the distance modeling approach downgrades the raw interaction count with the R4 regulatory element, so that it is not statistically significant. The heat map at the bottom displays Hi-C data at 5kb resolution[18] (Yue Lab Hi-C data browser http://promoter.bx.psu.edu/hi-c/), in which the interactions with the regulatory elements are highlighted by black boxes.

b. Comparison of 4C and NG Capture-C data in the mouse alpha globin locus (*Hba-a1* and *Hba-a2*) in primary erythroid cells. The alpha globin promoters are used as viewpoints and highlighted in red. The top track shows the interactions with the regulatory elements as defined by 4C[52]. As some larger fragments fail to amplify, the profile shows gaps, despite the use of a windowing approach. The bottom track displays the NG Capture-C profile, in which the interactions between the alpha globin promoters and enhancer elements are defined by tens of thousands of unique read pairs. The profiles are an average of 4 replicates in erythroid cells and 3 ES cell replicates, which allows for comparative analysis between the active and inactive cell types and identification of statistically significant interactions when the gene is in an active state (**Supplementary Fig. 1**).

Figure 5. Case study 2: regulation of the SOX2 locus in murine ES (mES) cells.

The *Sox2* gene is situated in a less gene dense locus, with DNase I hypersensitive sites extending over 900kb away from the gene promoter. A cluster of regulatory elements is found 85-111kb from the gene promoter, which reach the criteria for a super-enhancer[53], however, there are several other potential regulatory elements distal to this cluster (see **Supplementary Fig, 2** for further chromatin data).

The top panel shows 5C data, which is able to delineate the significant interactions with the super-enhancer, but does not fully cover the entire locus despite the use of several hundred probes[54](GSE36203). The next panel shows Hi-C data from mES cells (40kb resolution)[8]. At this resolution it is difficult to make out the interaction with the close enhancers but the more distal interactions are more prominent (the interactions defined by 4C and Capture-C are highlighted by black boxes). The 4C (GSM1868926)[55] and NG Capture-C (GSE67959)[15] profiles at the bottom, from the *Sox2* viewpoint highlighted in red, both show interactions across the gene desert.

Figure 6. Flow diagram of steps required for analysis of high-throughput 3C-based technologies.

3C-based experiments are usually sequenced using paired-end sequencing. Standard quality control measures and trimming of adapter sequences need to be performed prior to further analysis. When the DNA fragment is smaller than the size of the combined paired-end reads, the central area of overlap can be used to combine the reads into one single read.

Reads then need to be separated into the component fragments that ligated to one another at the restriction enzyme cut site. The ligation junction can be determined directly when the restriction cut site has been sequenced. However, it is common to infer that there is a restriction enzyme cut site in the central unsequenced portion of the read. This has the disadvantage that there may be other fragments interposed in the central un-sequenced part of the read.

After the reads have been split into the component fragments they are aligned to the genome separately. This can be challenging if the reads are not of sufficient length to allow both parts of the read to be aligned properly (this can be problematic with short single-end reads). The reads build up at the ends of the restriction fragments and so the data need further processing to generate a meaningful signal. The reads can either be mapped to the restriction fragments or data from multiple fragments can be combined either into bins or with a moving window.

When possible the data should be filtered to remove PCR duplicates. It is also important to remove signal caused by poor restriction enzyme digestion, mis-mapping to repetitive DNA sequences and off-target capture and mis-priming (depending on the method used).

Finally, statistical analysis can be performed. When possible, areas of increased interaction on gene activation can be identified by comparing the interaction profiles in active and inactive states. It is also possible to subtract the inactive background distribution using mathematical modeling, though this is less accurate and not as straightforward to interpret.

References

1 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).

2 Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **4**, 651-657, doi:10.1038/nmeth1068 (2007).

3 Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **6**, 283-289, doi:10.1038/nmeth.1313 (2009).

4 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

5 Stamatoyannopoulos, J. Connecting the regulatory genome. *Nature genetics* **48**, 479-480, doi:10.1038/ng.3553 (2016).

6 Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-113, doi:10.1038/nature11279 (2012).

7 Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339, doi:10.1016/j.cell.2011.01.024 (2011).

8 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

9 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385, doi:10.1038/nature11049 (2012).

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

11 Tan-Wong, S. M. *et al.* Gene loops enhance transcriptional directionality. *Science* **338**, 671-675, doi:10.1126/science.1224350 (2012).

12 Hsieh, T. H. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108-119, doi:10.1016/j.cell.2015.05.048 (2015).

13 Mukherjee, S., Erickson, H. & Bastia, D. Detection of DNA looping due to simultaneous interaction of a DNA-binding protein with two spatially separated binding sites on DNA. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 6287-6291 (1988).

14 Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311, doi:10.1126/science.1067799 (2002).

15 Davies, J. O. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nature methods* **13**, 74-80, doi:10.1038/nmeth.3664 (2016).

16 Lucas, J. S., Zhang, Y., Dudko, O. K. & Murre, C. 3D trajectories adopted by coding and regulatory DNA elements: first-passage times for genomic interactions. *Cell* **158**, 339-352, doi:10.1016/j.cell.2014.05.036 (2014).
17 Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64, doi:10.1038/nature12593 (2013).
18 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
19 Nagano, T. *et al.* Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome biology* **16**, 175, doi:10.1186/s13059-015-0753-7 (2015).
20 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).
21 Spilianakis, C. G. & Flavell, R. A. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nature immunology* **5**, 1017-1027, doi:10.1038/ni1115 (2004).
22 Liu, Z. & Garrard, W. T. Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3' boundary sequence spanning 46 kilobases. *Molecular and cellular biology* **25**, 3220-3231, doi:10.1128/MCB.25.8.3220-3231.2005 (2005).
23 Vernimmen, D. *et al.* Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). *Blood* **114**, 4253-4260, doi:10.1182/blood-2009-03-213439 (2009).
24 Vernimmen, D., De Gobbi, M., Sloane-Stanley, J. A., Wood, W. G. & Higgs, D. R. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO journal* **26**, 2041-2051, doi:10.1038/sj.emboj.7601654 (2007).
25 Zhou, G. L. *et al.* Active chromatin hub of the mouse alpha-globin locus forms in a transcription factory of clustered housekeeping genes. *Molecular and cellular biology* **26**, 5096-5105, doi:10.1128/MCB.02454-05 (2006).
26 Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature genetics* **38**, 1341-1347, doi:10.1038/ng1891 (2006).
27 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**, 1348-1354, doi:10.1038/ng1896 (2006).
28 Stadhouders, R. *et al.* Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols* **8**, 509-524, doi:10.1038/nprot.2013.018 (2013).
29 Simonis, M., Kooren, J. & de Laat, W. An evaluation of 3C-based methods to capture DNA interactions. *Nature methods* **4**, 895-901, doi:10.1038/nmeth1114 (2007).

30 Schwartzman, O. *et al.* UMI-4C for quantitative and targeted chromosomal contact profiling. *Nature methods* **13**, 685-691, doi:10.1038/nmeth.3922 (2016).
31 Andrey, G. *et al.* A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**, 1234167, doi:10.1126/science.1234167 (2013).
32 de Wit, E. *et al.* The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227-231, doi:10.1038/nature12420 (2013).
33 Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics* **46**, 136-143, doi:10.1038/ng.2870 (2014).
34 Simonis, M. *et al.* High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nature methods* **6**, 837-842, doi:10.1038/nmeth.1391 (2009).
35 Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381, doi:10.1016/j.cell.2014.02.019 (2014).
36 Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16**, 1299-1309, doi:10.1101/gr.5571506 (2006).
37 Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature genetics*, doi:10.1038/ng.2871 (2014).
38 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics* **43**, 1059-1065, doi:10.1038/ng.947 (2011).
39 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9**, 999-1003, doi:10.1038/nmeth.2148 (2012).
40 Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131-3133, doi:10.1093/bioinformatics/bts570 (2012).
41 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294, doi:10.1038/nature12644 (2013).
42 Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342**, 948-953, doi:10.1126/science.1236083 (2013).
43 Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908-921, doi:10.1016/j.cell.2012.02.002 (2012).
44 Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nature methods* **12**, 71-78, doi:10.1038/nmeth.3205 (2015).
45 Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research* **25**, 582-597, doi:10.1101/gr.185272.114 (2015).

46 Kolovos, P. *et al.* Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics & chromatin* **7**, 10, doi:10.1186/1756-8935-7-10 (2014).
47 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).
48 Kieffer-Kwon, K. R. *et al.* Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell* **155**, 1507-1520, doi:10.1016/j.cell.2013.11.039 (2013).
49 Brackley, C. A. *et al.* Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome biology* **17**, 59, doi:10.1186/s13059-016-0909-0 (2016).
50 Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular systems biology* **11**, 852, doi:10.15252/msb.20156492 (2015).
51 Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98, doi:10.1016/j.cell.2011.12.014 (2012).
52 van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods* **9**, 969-972, doi:10.1038/nmeth.2173 (2012).
53 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
54 Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295, doi:10.1016/j.cell.2013.04.053 (2013).
55 de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Molecular cell* **60**, 676-684, doi:10.1016/j.molcel.2015.09.023 (2015).

Annotation of references

This is the seminal paper first describing chromosome conformation capture in yeast[14].

This is the first use of 3C to define interactions between regulatory elements in mammalian cells[10].

This paper first described the Hi-C method and it describes large scale organization of chromatin as a fractal globule[20].

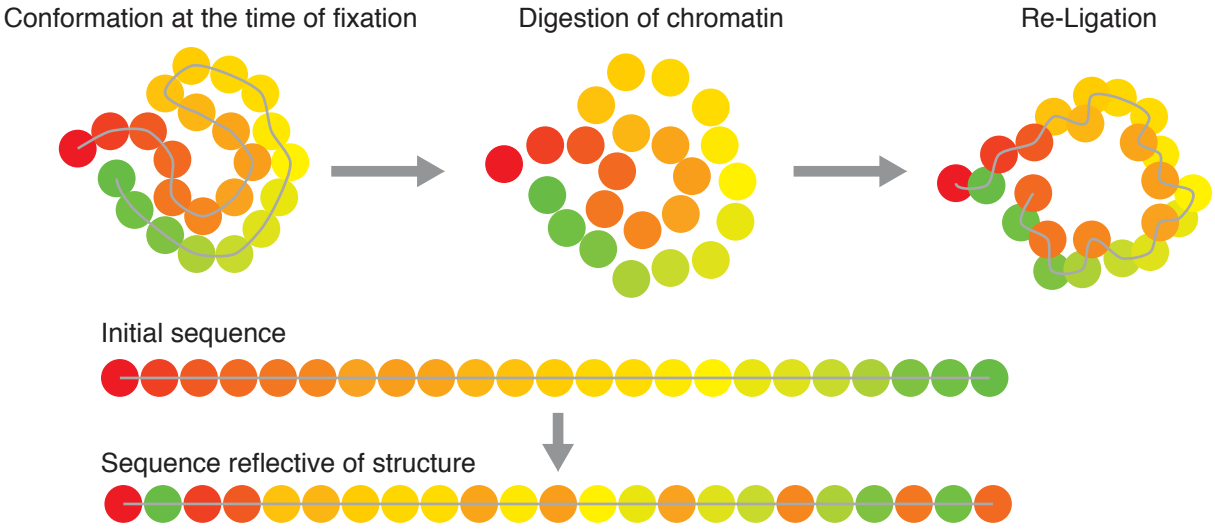
This paper uses 4C to delineate the changes in gene regulation and interaction profiles at the HoxD genes during limb development[31].

This paper describes the highest possible resolution currently achievable with genome-wide all vs all approaches using Hi-C[18].

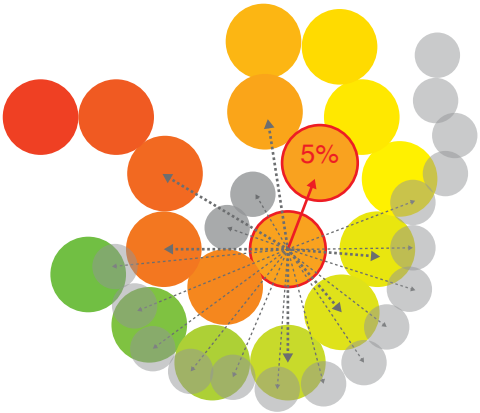
This paper describes the highest resolution and sensitivity available with a one vs all approach using NG Capture-C and is additionally capable of high levels of multiplexing of viewpoints[15].

Figure 1

a.



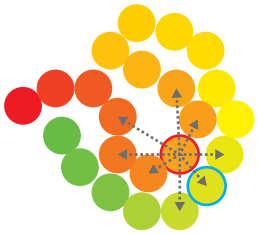
b.



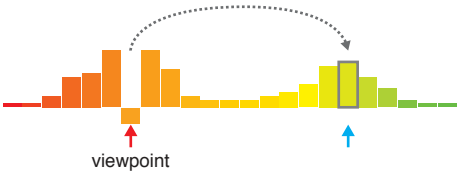
The probability of ligating back to its original partner is ~5%.

There are therefore at least 20 potential ligation partners in each cell.

c.



Probability distribution of ligation events from a single viewpoint



Heat map of probability distribution of ligation events from all viewpoints

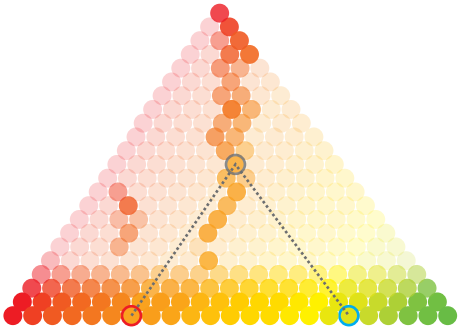


Figure 2

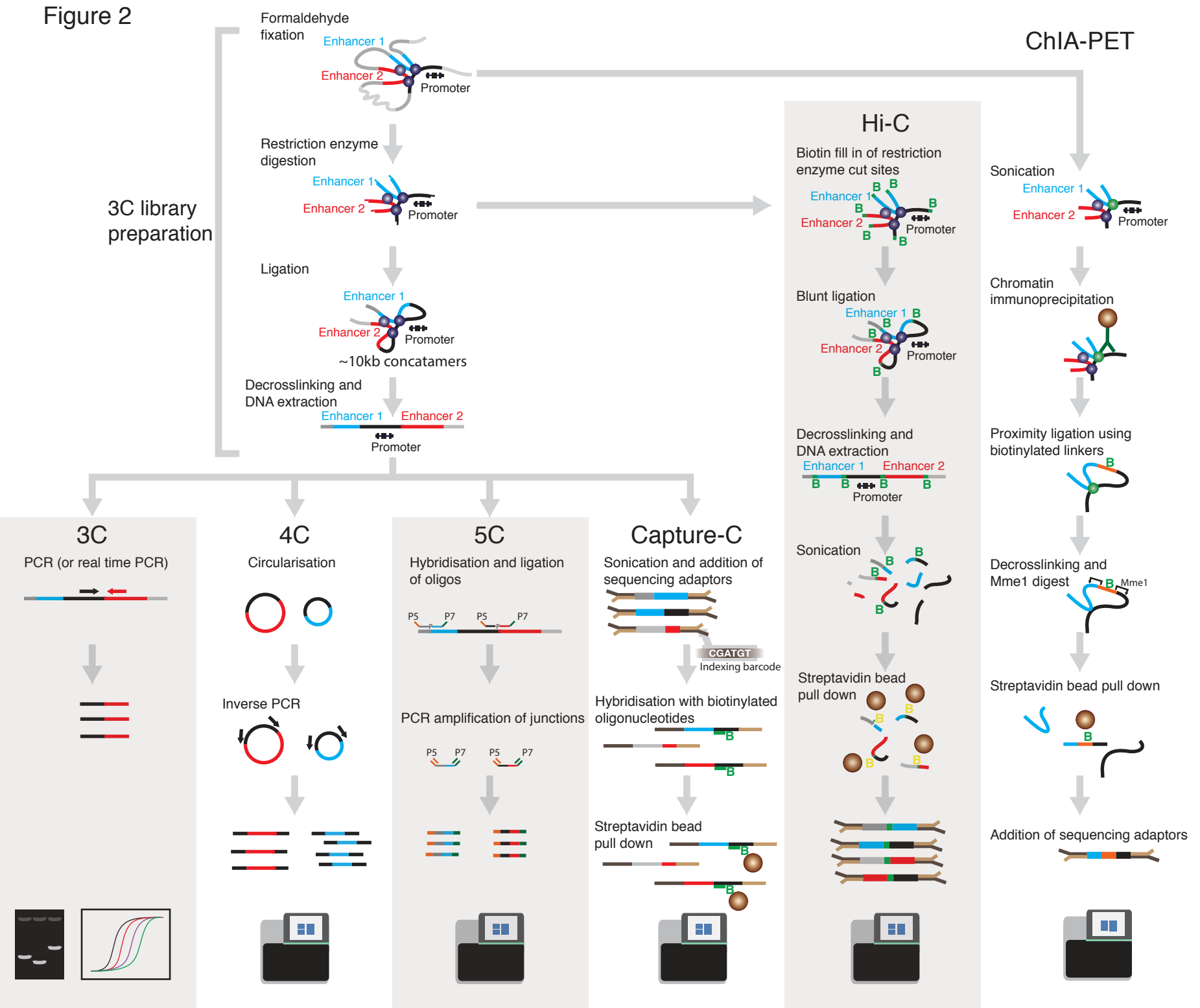

3C library preparation
Formaldehyde fixation
Enhancer 1
Enhancer 2
Promoter
Restriction enzyme digestion
Ligation
~10kb concatamers
Decrosslinking and DNA extraction
ChIA-PET
Hi-C
Biotin fill in of restriction enzyme cut sites
Blunt ligation
Decrosslinking and DNA extraction
Sonication
Streptavidin bead pull down
Sonication
Chromatin immunoprecipitation
Proximity ligation using biotinylated linkers
Decrosslinking and Mme1 digest
Mme1
Streptavidin bead pull down
Addition of sequencing adaptors
3C
PCR (or real time PCR)
4C
Circularisation
Inverse PCR
5C
Hybridisation and ligation of oligos
P5
P7
PCR amplification of junctions
Capture-C
Sonication and addition of sequencing adaptors
CGATGT
Indexing barcode
Hybridisation with biotinylated oligonucleotides
Streptavidin bead pull down

Figure 3

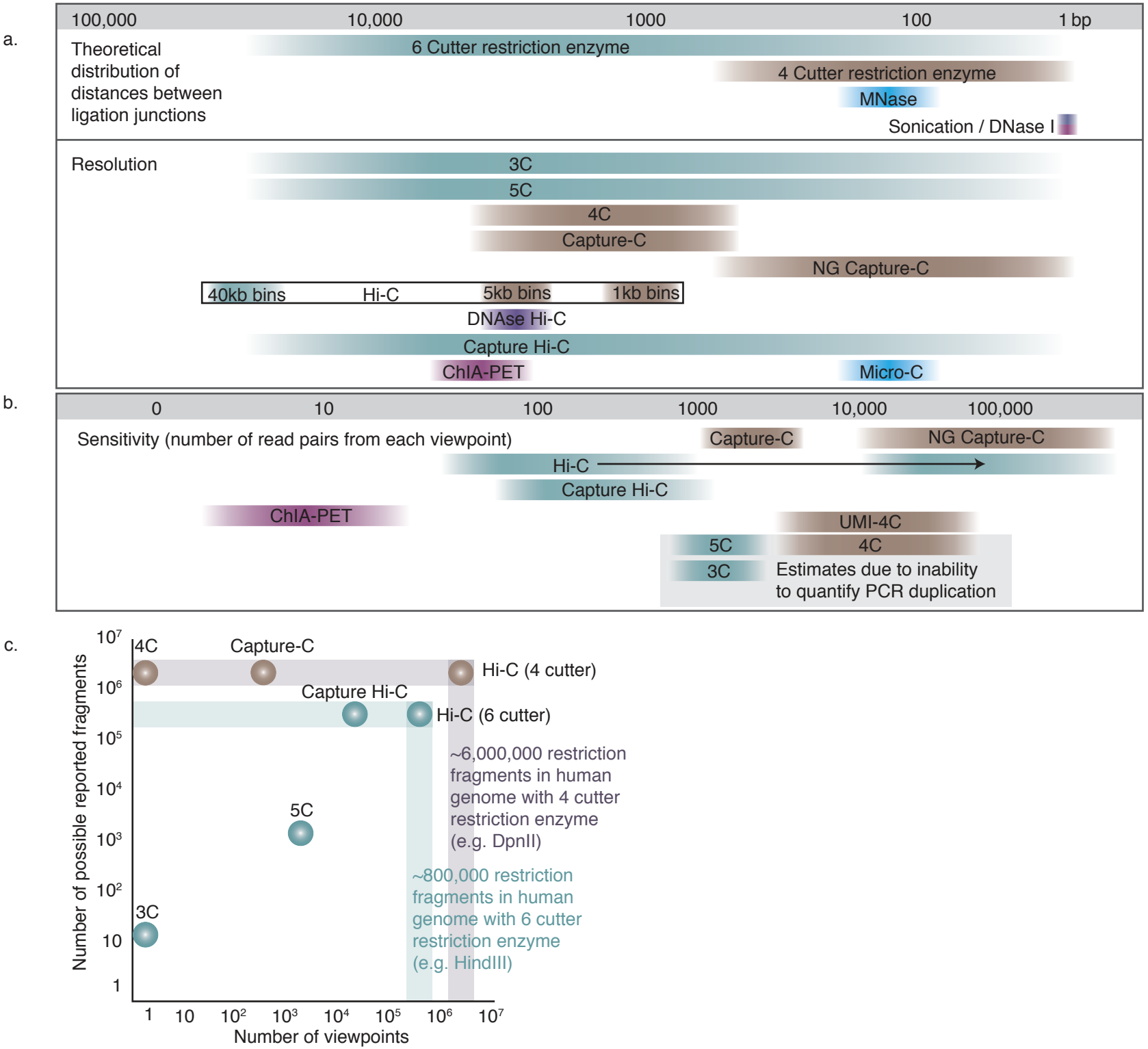

a.
100,000
10,000
1000
100
1 bp
Theoretical distribution of distances between ligation junctions
6 Cutter restriction enzyme
4 Cutter restriction enzyme
MNase
Sonication / DNase I
Resolution
3C
5C
4C
Capture-C
NG Capture-C
40kb bins
Hi-C
5kb bins
1kb bins
DNAse Hi-C
Capture Hi-C
ChIA-PET
Micro-C
b.
0
10
100
1000
10,000
100,000
Sensitivity (number of read pairs from each viewpoint)
Capture-C
NG Capture-C
Hi-C
Capture Hi-C
ChIA-PET
UMI-4C
5C
4C
3C
Estimates due to inability to quantify PCR duplication
c.
Number of possible reported fragments
Number of viewpoints
4C
Capture-C
Hi-C (4 cutter)
Capture Hi-C
Hi-C (6 cutter)
5C
3C
~6,000,000 restriction fragments in human genome with 4 cutter restriction enzyme (e.g. DpnII)
~800,000 restriction fragments in human genome with 6 cutter restriction enzyme (e.g. HindIII)

Figure 4

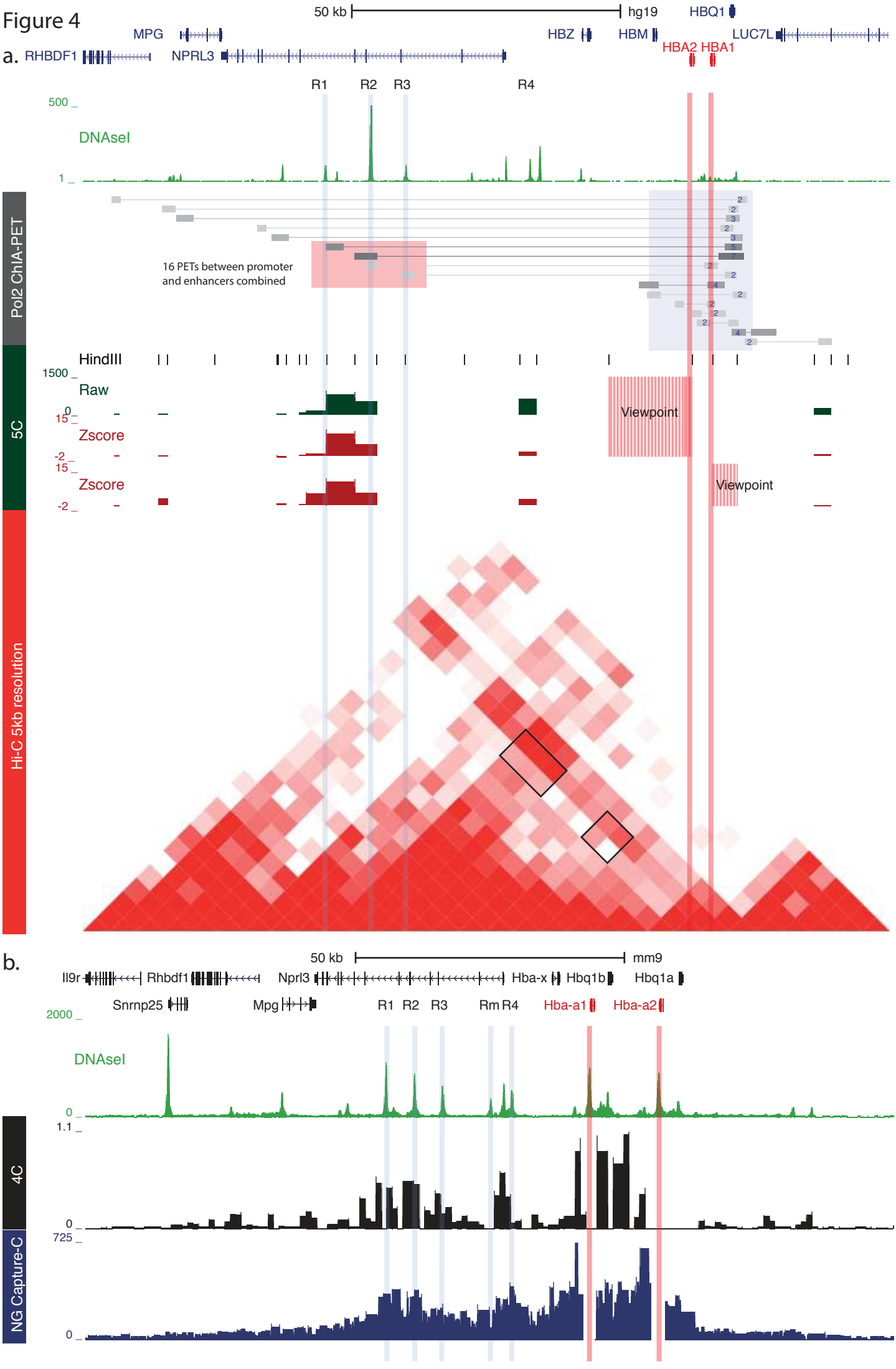

a.
50 kb
hg19
MPG
HBZ
HBM
HBQ1
LUC7L
HBA2 HBA1
RHBDF1
NPRL3
R1
R2
R3
R4
500
DNAseI
1
Pol2 ChIA-PET
16 PETs between promoter and enhancers combined
5C
HindIII
1500
Raw
0
15
Zscore
-2
15
Zscore
-2
Viewpoint
Viewpoint
Hi-C 5kb resolution
b.
50 kb
mm9
Il9r
Rhbdf1
Nprl3
Hba-x
Hbq1b
Hbq1a
Snrnp25
Mpg
R1
R2
R3
Rm
R4
Hba-a1
Hba-a2
2000
DNAseI
0
1.1
4C
0
725
NG Capture-C
0

Figure 5

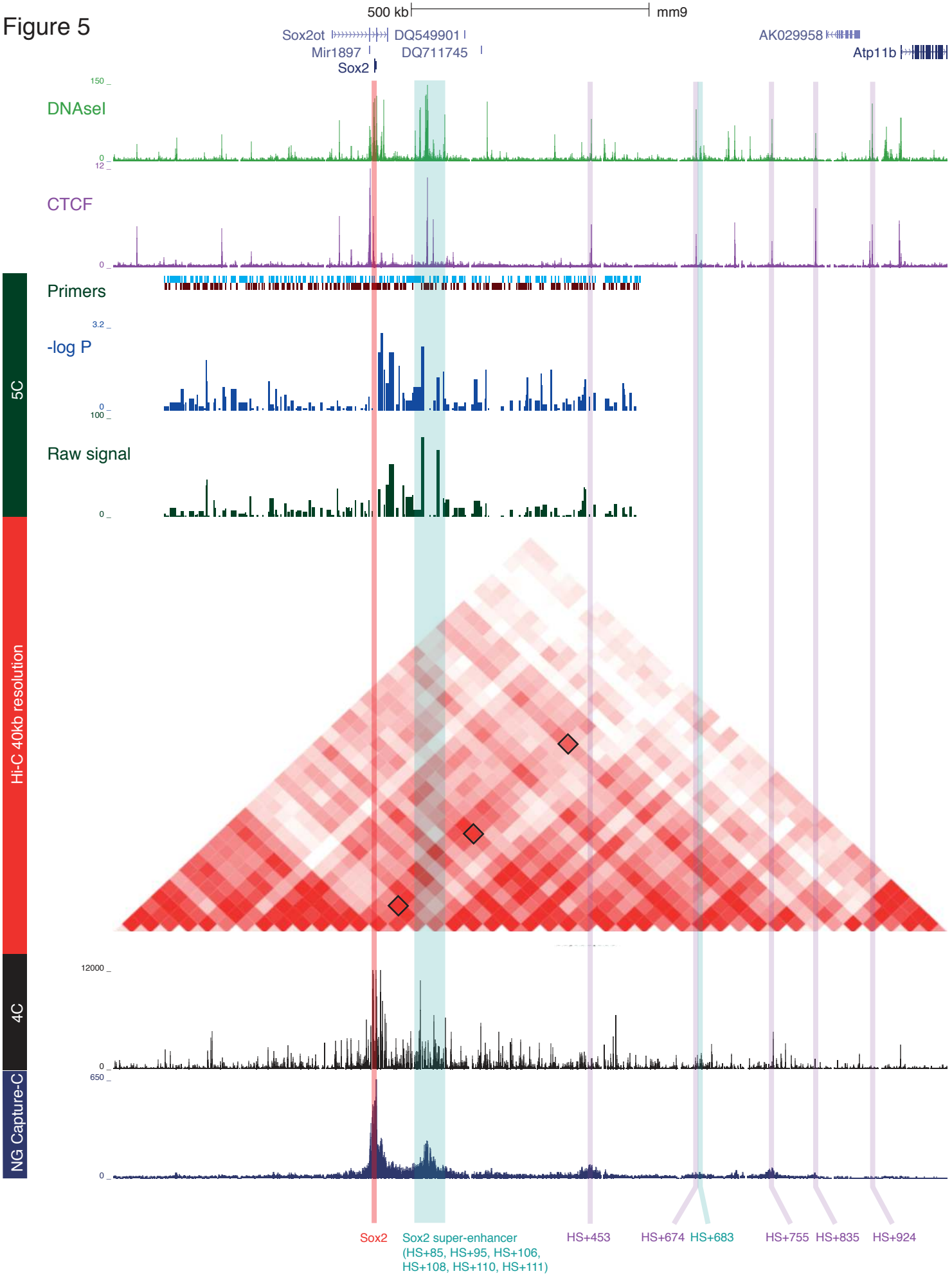
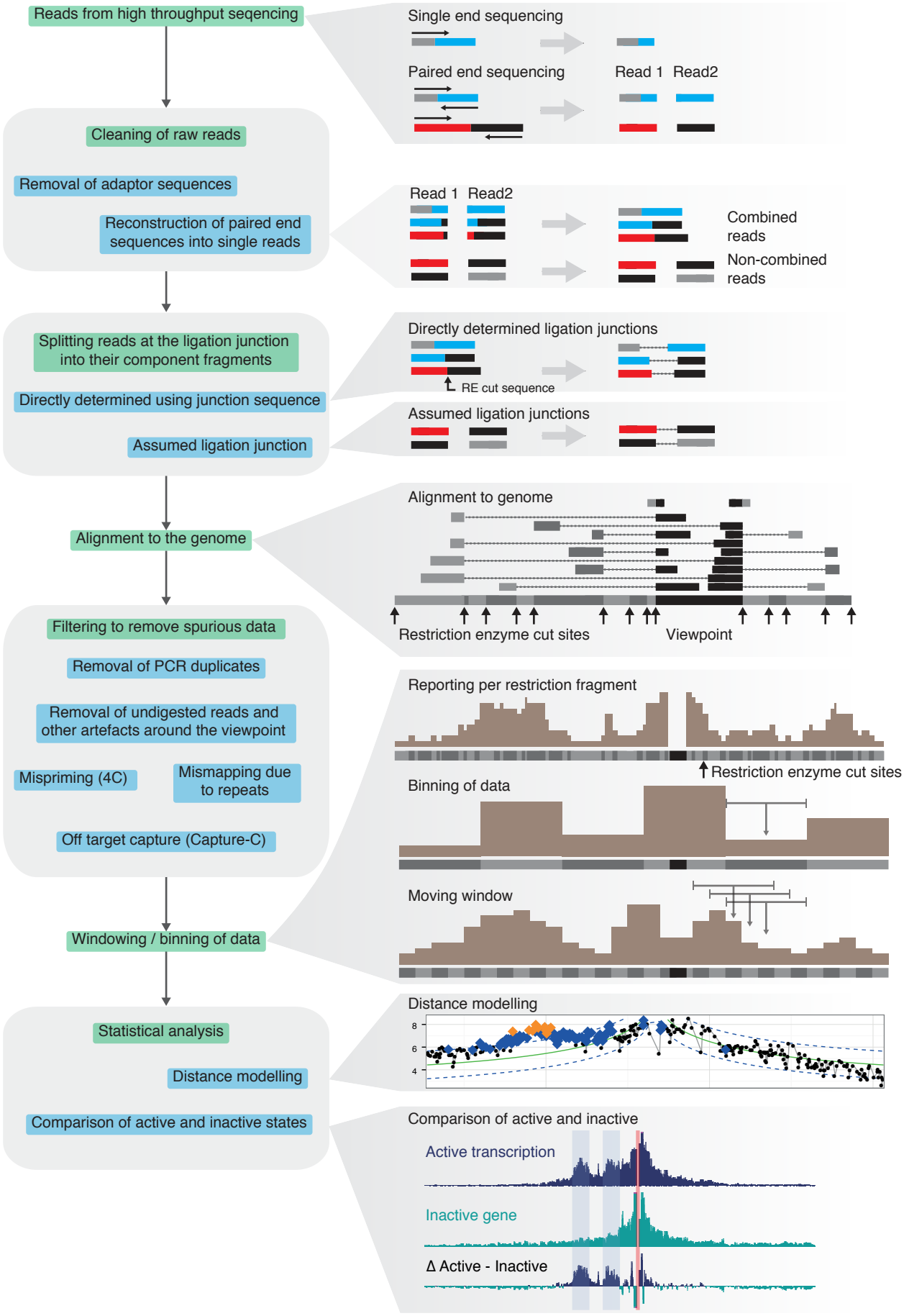
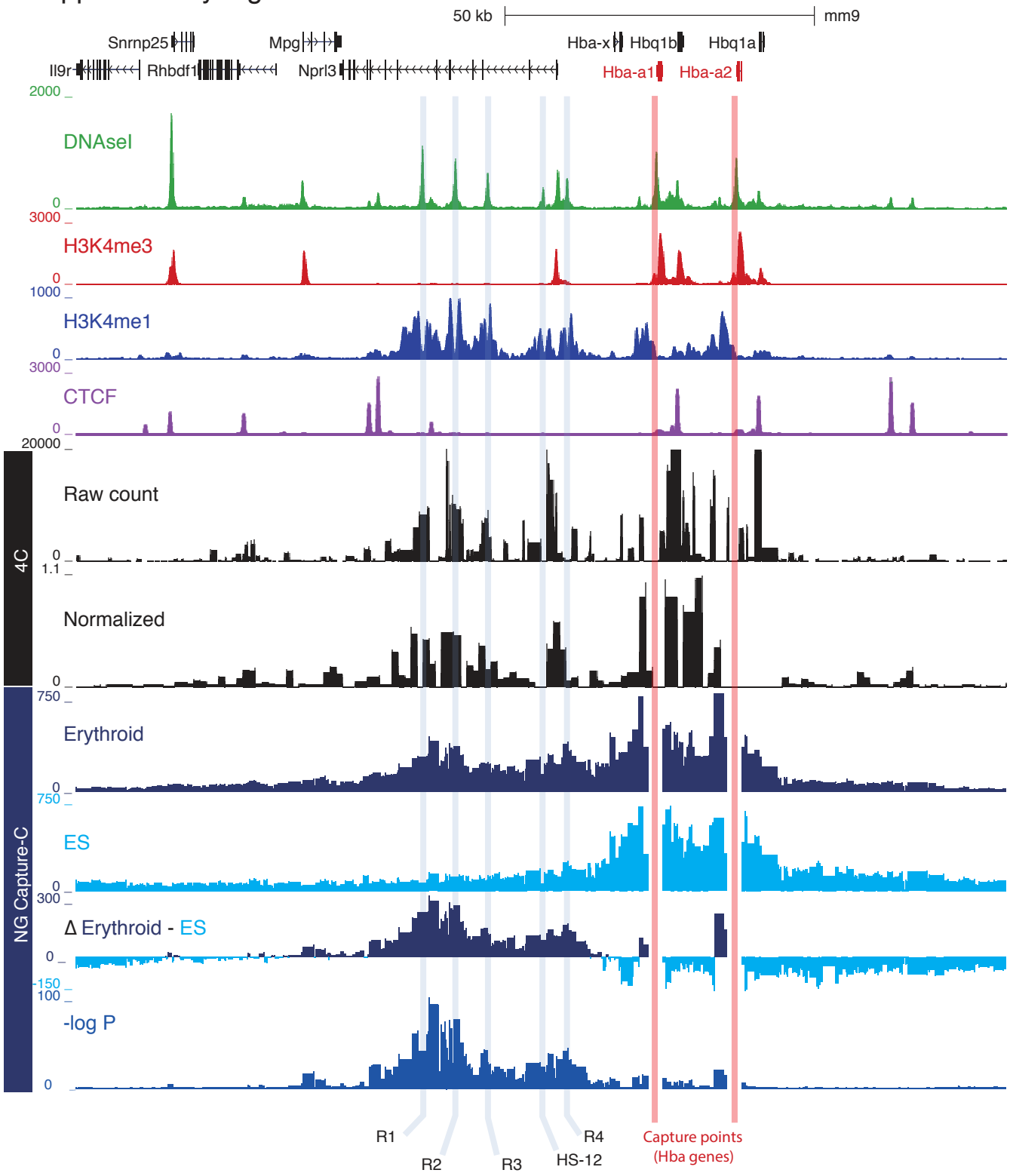

500 kb
mm9
Sox2ot
DQ549901
AK029958
Mir1897
DQ711745
Atp11b
Sox2
DNAseI
CTCF
Primers
-log P
Raw signal
5C
Hi-C 40kb resolution
4C
NG Capture-C
Sox2
Sox2 super-enhancer
(HS+85, HS+95, HS+106,
HS+108, HS+110, HS+111)
HS+453
HS+674
HS+683
HS+755
HS+835
HS+924

Figure 6


Reads from high throughput seqencing
Single end sequencing
Paired end sequencing
Read 1
Read2
Cleaning of raw reads
Removal of adaptor sequences
Reconstruction of paired end sequences into single reads
Read 1
Read2
Combined reads
Non-combined reads
Splitting reads at the ligation junction into their component fragments
Directly determined using junction sequence
Assumed ligation junction
Directly determined ligation junctions
RE cut sequence
Assumed ligation junctions
Alignment to the genome
Alignment to genome
Restriction enzyme cut sites
Viewpoint
Filtering to remove spurious data
Removal of PCR duplicates
Removal of undigested reads and other artefacts around the viewpoint
Mispriming (4C)
Mismapping due to repeats
Off target capture (Capture-C)
Reporting per restriction fragment
Restriction enzyme cut sites
Binning of data
Moving window
Windowing / binning of data
Statistical analysis
Distance modelling
Comparison of active and inactive states
Distance modelling
8
6
4
Comparison of active and inactive
Active transcription
Inactive gene
Δ Active - Inactive

Supplementary Figure 1

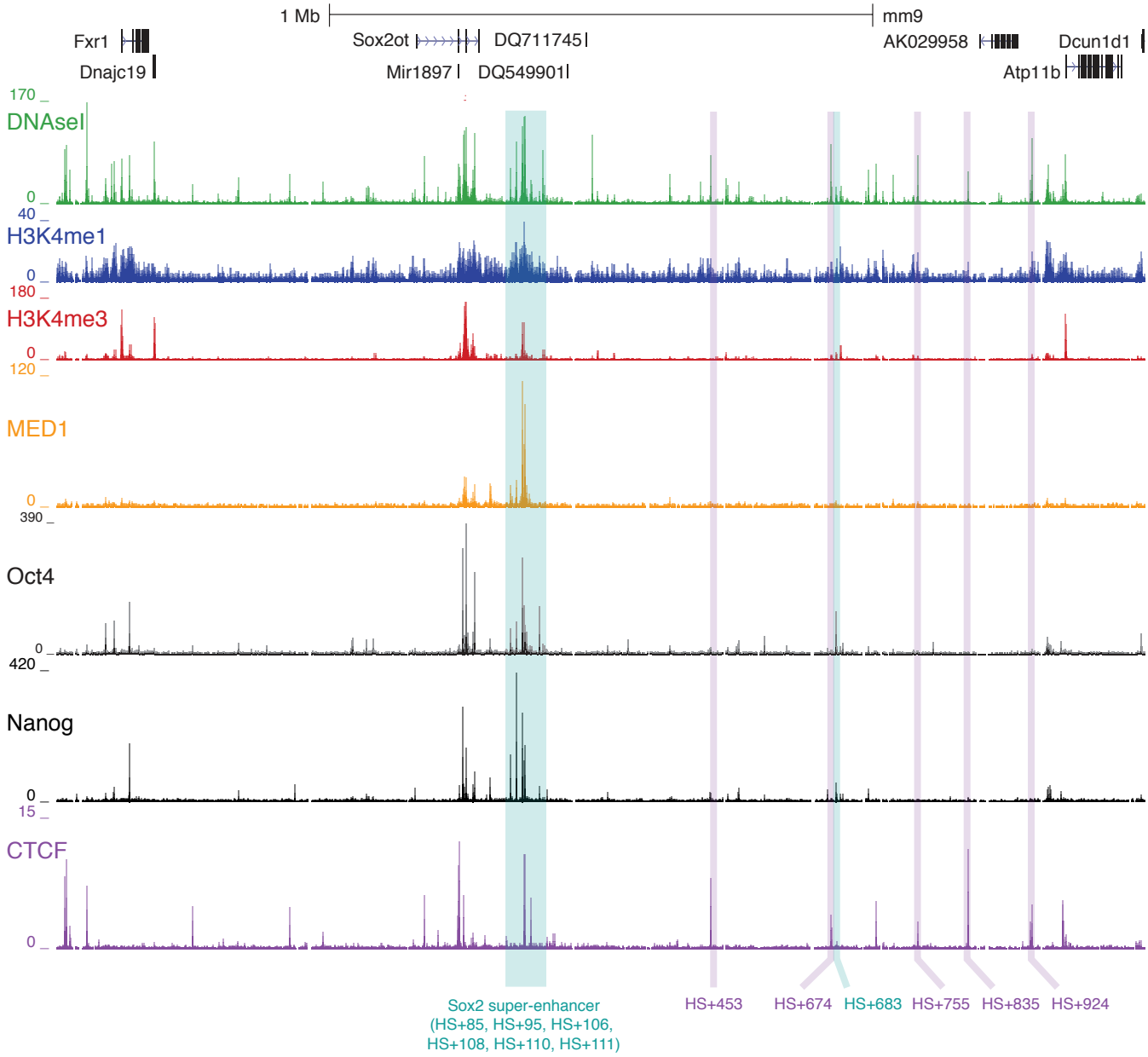


Supplementary Figure 1: Chromatin, 4C and NG Capture-C data for the alpha globin locus in mouse erythroid cells.

The DNaseI hypersensitivity track (green), marking open chromatin, is showed at the top. Below are ChIP-seq profiles for H3K4me3 (red), H3K4me1 (blue), and CTCF (purple), that highlight promoters, enhancers and CTCF binding sites, respectively (Hughes et al., 2014).

The raw and normalized 4C profiles are shown below (Van der Werken et al., 2012). The NG Capture-C profiles are an average of 4 replicates in erythroid cells and 3 ES cell replicates, which allows for comparative analysis between cell types when the gene is active (erythroid) and inactive (ES) and identification of statistically significant interactions when the gene is in an active state (Davies et al, 2016).

Supplementary Figure 2



Supplementary Figure 2: Chromatin data for the Sox2 locus in mES cells.

The DNaseI hypersensitivity track (green), marking open chromatin, is showed at the top. Below are ChIP-seq profiles for H3K4me1 (blue), H3K4me3 (red), mediator (MED, yellow), octamer-binding transcription factor 4 (Oct4, black), Nanog (black) and CTCF (purple), which allow the DNaseI hypersensitive sites to be defined further.

A cluster of hypersensitive sites 85-111 kb (highlighted in green) from het promoter of the gene are defined as a super-enhancer (Whyte et al., 2013). However, the gene has several other cell-type specific regulatory elements, which extend nearly 1 Mb from the promoter. These tend to be CTCF binding sites (highlighted in purple), although there is an additional potential regulatory element bound by Nanog and Oct4 (HS+683, highlighted in green).

ES cell data: DNAseI (ENCODE UW); ChIP-seq H3K4me1 and H3K4me3 (ENCODE/LICR); ChIP-seq MED1 (Young lab GSM1038259); ChIP-seq Oct4 (Young lab GSM1082340); ChIP-seq Nanog (Young lab GSM1082342), ChIP-seq CTCF (LICR GSM918748).