4C-ker: A method to reproducibly identify genome-wide interactions captured by 4C-Seq experiments

Ramya Raviram[1,3], Pedro P. Rocha[1], Christian L. Müller[3,4,5], Emily R. Miraldi[3,4,5], Sana Badri[1], Yi Fu[1,3], Emily Swanzey[2], Charlotte Proudhon[1], Valentina Snetkova[1], Richard Bonneau[3,4,5*] and Jane A. Skok[1*]

AFFILIATIONS

[1]Department of Pathology, New York University School of Medicine, New York, NY 10016, USA.

[2]Skirball Institute, New York University School of Medicine, New York, NY 10016, USA.

[3]Department of Biology, New York University, New York, New York, 10003, USA.

[4]Department of Computer Science, Courant Institute of Mathematical Sciences, New York, New York, 10003, USA.

[5]Simons Center for Data Analysis, New York, New York, 10010, USA.

*Equally contributing authors

*Correspondence and requests for materials should be addressed to J.S. (jane.skok@nyumc.org) or R.B (rb133@nyu.edu)

ABSTRACT

4C-Seq has proven to be a powerful technique to identify genome-wide interactions with a single locus of interest (or "bait") that can be important for gene regulation. However, analysis of 4C-Seq data is complicated by the many biases inherent to the technique. An important consideration when dealing with 4C-Seq data is the differences in resolution of signal across the genome that result from differences in 3D distance separation from the bait. This leads to the highest signal in the region immediately surrounding the bait and increasingly lower signals in far-*cis* and *trans*. Another important aspect of 4C-Seq is the resolution, which is greatly influenced by the choice of restriction enzyme and the frequency at which it can cut the genome. Thus, it is important that a 4C-Seq analysis method is flexible enough to analyze data generated using different enzymes and to identify interactions across the entire genome. Current methods for 4C-Seq analysis only identify interactions in regions near the bait or in regions located in far-*cis* and *trans,* but no method comprehensively analyzes 4C signals of different length scales. In addition, some methods also fail in experiments where chromatin fragments are generated using frequent cutter restriction enzymes. Here, we describe 4C-ker, a Hidden-Markov Model based analysis that identifies regions throughout the genome that interact with the 4C bait locus. In addition we incorporate methods for the identification of differential interactions in multiple 4C-seq datasets collected from different genotypes or experimental conditions. Adaptive window sizes are used to correct for differences in signal coverage in near-bait regions, far-*cis* and *trans* chromosomes. Using several datasets, we demonstrate that 4C-ker outperforms all existing 4C-Seq pipelines in its ability to reproducibly identify interaction domains at all genomic ranges with different resolution enzymes.

Introduction

Understanding the 3D organization of the genome and the intricacies of chromatin dynamics has been the focus of studies aimed at characterizing gene regulation in physiological processes and disease states [1, 2]. Microscopy based studies provided the first snapshots of nuclear organization, revealing that individual chromosomes occupy distinct territories with little intermingling between them [3, 4]. The development of chromosome conformation capture (3C) transformed the field of nuclear organization enabling identification of chromatin interactions at the molecular level and at the same time opening the door to high-throughput, genome-wide techniques [5]. Hi-C, for example, captures all pairwise interactions in the nucleus and has revealed that chromosomes segregate into two distinct spatial compartments (A and B) depending on their transcriptional and epigenetic status [6]. These compartments are further subdivided into Topological Associated Domains (TADs), which are highly self-interacting megabase scale structures [7-9]. To probe interactions between regulatory elements using Hi-C requires a depth of sequencing that for many labs is cost-prohibitive [10]. 5C can circumvent these issues, but the interaction analysis is limited to the portion of the genome for which primers are designed [11]. Circular chromosome conformation capture combined with massive parallel sequencing (4C-Seq) is currently the best option for obtaining the highest resolution interaction signal for a particular region of interest.

In 4C-Seq, an inverse PCR step allows for the identification of all possible genome wide interactions from a single viewpoint (the "bait") and an assessment of the frequencies at which these occur. The sequencing coverage obtained by 4C near the bait region is extremely high and therefore enables precise characterization and quantification of regulatory interactions [12, 13]. By focusing on one locus at a time and thus only the interactions that this locus is engaged in, 4C can reproducibly identify long-range interactions on *cis* and *trans* chromosomes [14]. For example, 4C was used to demonstrate that genes controlled by common transcription factors tend to occupy the same nuclear space even when located on different chromosomes [15, 16].

There are many inherent biases specific to the 4C technique that has made detecting meaningful and reproducible interactions challenging. First, in accordance with the chromosome territory model, the majority of 4C signal is located on the bait chromosome. Secondly, coverage and signal strength are highest in the region around the bait and this decreases along the chromosome as a function of linear distance from the bait. Third, the restriction enzyme used for the first digest in the experiment is an important determinant of the resolution of the signal and the extent to which interactions can be detected.

Finally, as with most PCR-based techniques, 4C data includes PCR artifacts that manifest as a large accumulation of reads in particular locations.

Current methods of analysis have addressed some of these issues, however there are still many hurdles to overcome. Specifically, existing methods do not properly account for the differences in 4C signal strength across the genome and therefore they are only able to either identify interactions in (i) regions where the signal is highest, i.e., near the bait or (ii) regions of low 4C signal (far-*cis* and *trans*). Thus there is no method that comprehensively identifies interactions across the genome. In addition, most methods were developed and tested using datasets generated with 6bp cutters and we show that they do not perform well with 4bp cutter generated libraries.

The goal of 4C-ker is to address these weaknesses by: 1) identifying domains that interact most frequently with the bait across the genome in a given population of cells, and 2) detecting quantitative differences in 4C-Seq signal between conditions. Here we use a Hidden Markov Model to account for the polymer nature of chromatin, in which adjacent regions share a similar probability of interacting with the bait. In addition, to account for the variation in signal captured at different 3D distances we use a window-based approach. To determine the window size of analysis, we adapted a k-th nearest neighbor approach to account for the decrease in 4C-Seq coverage along *cis* and *trans* chromosomes. We used 4C-ker to analyze numerous publically available 4C-Seq datasets as well as data generated in our own lab and compared this with other published methods. Our results demonstrate that 4C-ker can correct for multiple 4C-Seq biases and reproducibly detect genome wide interactions from the bait viewpoint. Importantly, 4C-ker is the only tool that can identify interactions with regions in near and far-*cis* as well as *trans.*

Results

Workflow of 4C-ker

We developed 4C-ker to identify genome-wide interactions generated by 4C-Seq data and to quantitatively examine differences in interaction frequencies between conditions. The main components of the 4C-ker method are outlined in **Fig 1**. First, 4C-Seq reads are mapped to a reduced genome consisting of unique sequences adjacent to all primary restriction enzyme sites in the genome. Mapping to a reduced genome helps to remove spurious ligation events that do not result from crosslinking. The analysis of 4C-Seq is typically performed separately for *cis* and *trans* chromosomes because of the large differences in signal in these distinct locations. Additionally, we present the option for focusing the analysis on the region surrounding the bait, where 4C-Seq signal and resolution are highest. A window-based approach is applied in order to take into account of differences in signal strength at different 3D distances and the dynamic nature and variability of chromosome interactions in a population of cells.

One of the most challenging aspects of 4C-Seq is determining the window size at which the data should be analyzed. Adaptive window sizes that depend on the distance to the bait can adjust for differences in coverage of 4C-Seq signal in regions near the bait, far-*cis* and *trans* chromosomes. 4C signal is generally higher around the bait region and decreases in far-*cis* and *trans*. We developed a kth nearest neighbor method to build overlapping windows of adaptive sizes based on the 4C-Seq coverage of a given dataset at each location in the genome. With this approach the size of each window is determined by the amount of signal detected in each region. This will result in small windows near the bait and other regions where there is high coverage, versus larger windows further away from the bait where there is low coverage.

Once the windows for a given dataset are determined, the counts at observed fragments within these windows are normalized using DESeq2 [17]. For the *cis* chromosome, the linear distance from the bait to the mid-point of each window is used to correct for the inverse relationship between counts and linear separation from the bait. The counts and distances are log transformed and are used as inputs (observed states) for the Hidden Markov Model (HMM). A separate model is used for *cis* and *trans* chromosomes (in the latter there is no effect of linear distance from the bait). A three-state HMM is used to partition the genome into windows that interact with the bait at (1) high frequency, (2) low frequency and (3) those that do not interact. Use of overlapping windows, allows us to more precisely define the regions of high interaction (for more detailed explanation of the workflow, refer to the methods section).

The resulting parameters for the model show higher probabilities for transitioning to the same state, correctly accounting for the polymeric nature of DNA on chromosomal interactions (**Fig 1**, 3-state HMM). Domains that are consistently found as interacting at a high frequency across several samples can be used for downstream quantitative analysis. Furthermore DESeq2 can be used to quantitatively compare interactions across conditions. The 4C-ker pipeline will be available as an R package (R.4Cker on github) along with the domains of interactions identified for all the datasets analyzed in this study.

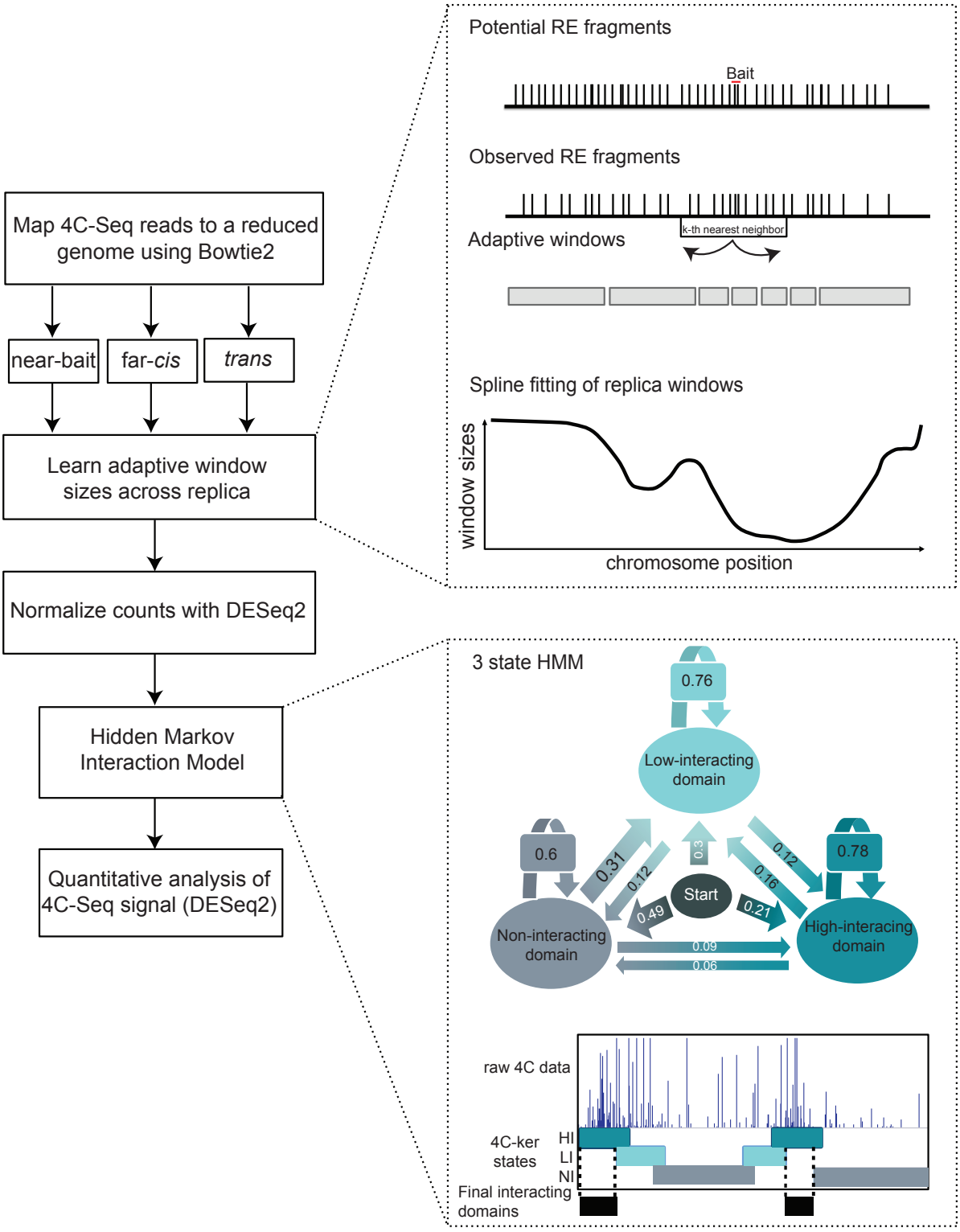


Fig 1: Workflow of 4C-ker. The key features of the method are outlined in the figure. A more detailed explanation of each section can be found in the materials and methods section.

Use of 4Cker to identify close range interactions

4C-Seq is commonly used to identify regulatory interactions that occur in close linear proximity to the bait. Therefore, we provide an option to focus the analysis only in this region, where the highest resolution interactions are identified. An important aspect of 4C-Seq library preparation is the choice of restriction enzyme used to digest cross-linked chromatin as the genome-wide frequency of enzyme recognition sites determines the resolution of the experiment. Therefore, 4bp cutters such as DpnII or NlaIII, which cut the genome more frequently, provide a higher resolution profile of 4C interactions compared to 6bp cutters like HindIII (**S1 Fig**) [18]. To ensure that our method works with both types of restriction enzymes, we tested it using numerous datasets generated from our lab as well as all publically available datasets for which replicates are available (all of these datasets passed stringent quality control checks. See methods section and Supplementary Table 1 for details).

The near-bait analysis was restricted to 10MB around the bait for 6bp cutters and 1MB for 4bp cutters as these are the regions that contain that the highest 4C signal in each case. As can be seen in **Fig 2A** and **2B**, raw 4C-Seq signal is highest near the bait and decreases with increasing linear distance. As 4C-ker corrects for this decrease in signal it is able to detect interactions across the entire region analyzed. In addition, due to the adaptive windows, the size of interactions detected are smallest near the bait where coverage is highest and larger in regions separated from the bait by increasing linear distances where coverage is lower (**Fig 2C** and **2D**). The resolution of domains identified by 4C-ker can be conveniently adjusted by using different values for the number of observed fragments used to generate the adaptive windows (**S2A Fig**)**.** Here we used values ranging from 3-10 in the 2MB region around the bait. As the value of 'k' increases, we observe a consistent increase in the size of the domains as well as increased similarity between replicates (see methods section for details). The parameter k can be adjusted by the user depending on the biological question that is being addressed. For example, if the aim of the study is to identify interactions between enhancers and promoters, we suggest k=3-5. In order to identify larger domains that coincide with broad regions encompassing chromatin with similar histone modifications, setting k=10 is a suitable choice.

To assess the performance of 4C-ker we used existing methods to analyze the same datasets. There are currently four methods available in the community to detect significant interactions using 4C-Seq datasets (fourSig, Splinter et al, r3CSeq and FourCSeq). Details of how we implemented these algorithms for comparison with the 4C-ker pipeline can be found in the methods section. Although the method developed by van der Werken et al (4cseqpipe) [13] does not identify significant interactions, it provides a good visualization tool for 4C-Seq signal near the bait (**Fig 2A and 2B**). The fourSig approach generates windows based on restriction enzyme fragments and compares the counts within each window against a random background distribution [19]. As fourSig does not take account of the impact of distance on 4C-Seq signal, it identifies most of this region as large interacting domains and this results in a high similarity index between replicates (**S2B and S2C Fig)**. However, in contradiction to decreasing resolution of 4C-Seq signal with increased separation from the bait, the size of the domains identified by fourSig are largest near the bait and these decrease with increasing separation from the bait (**Fig 2C** and **2D**). The method described by Splinter et al [20], referred to here as the 'de Laat method' excludes the 2MB region around the bait and only calls interactions in the rest of the genome based on enrichment of binary coverage in a given window, compared to a local background. As such, the de Laat method does not identify any interactions with 4bp cutters (**Fig 2B, 2D**). Moreover, using the 6bp cutter datasets it only identifies interactions in 2 out of 7 datasets in the 10MB region (**S2B Fig**). Together these findings reflect the limitations of this method in detecting 4C-Seq interactions in the region with highest coverage, where the majority of important regulatory interactions occur. The r3CSeq method uses reverse cumulative fitted values of the power law normalization and a background scaling method to correct for interactions near the bait [21]. This approach also provides the option to detect interactions at the fragment level or at the window level. In most datasets r3CSeq only identifies significant interaction near the bait as shown in **Fig 2A**, **2C** and **2D** and therefore have a high similarity index between replicates (**S2B and S2C Fig**). Although interactions further from the bait are identified (**Fig 2B)**, they are not reproducible as measured by the similarity index (**S2C Fig)**. The FourCSeq pipeline only has the option to analyze interactions at the fragment level. It is based on the DESeq2 method with an additional function that corrects for the effect of linear distance from the bait [22]. This method failed to identify any significant interactions for any of the datasets analyzed. If interactions between regulatory elements are being analyzed, the

majority will be identified in the region near the bait. Therefore, it is important that a 4C-Seq analysis can properly identify these interactions. Here we show that 4C-ker outperforms other methods and identifies interactions that correctly reflects the nature of high-resolution 4C-Seq signal in this region.

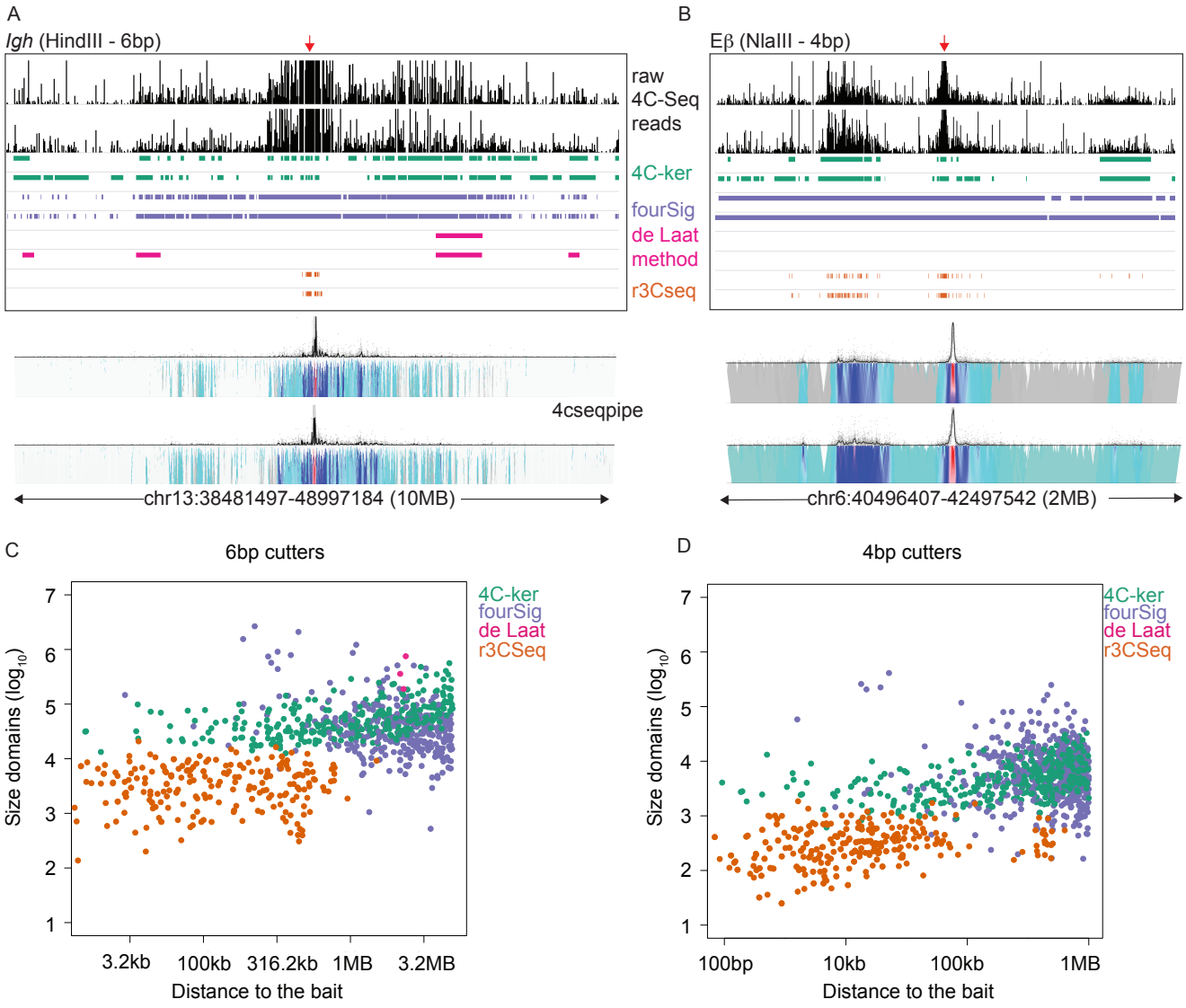


Fig 2 4C-ker outperforms other methods in the region near the bait when the stability of interacting regions between replicates is examined for four methods. (**A-B**) Example datasets for 6bp and 4bp cutter experiments. Raw 4C-Seq reads are shown for a 10MB region around the bait in (A) and 2 MB in (**B**). Experiment in **A** was performed using activated B cells digested with HindIII and a bait near the *Igh* locus. Experiment in **B** was performed in double negative T cells digested with NlaIII and a bait near the Eβ enhancer of *Tcrb*. Significant interactions determined by each method for 2 replicates are shown below the raw 4C-Seq profile. Domainograms generated using 4cseqpipe are displayed for the same region. (**C-D**) Distance of the midpoint of the interacting domain to the bait is plotted against its size. Plots only contain domains that overlap by 50% between replicates.

Identification of long-range interactions using 4C-ker

We next used 4C-ker for analysis of the entire bait chromosome using the same fourteen datasets described above. Due to lower 4C-Seq signal in regions distant from the bait (far-*cis*) the correlation between replicates decreases compared to near-bait regions (**S3A Fig**). This difference is more pronounced with 4bp cutter generated datasets. A potential explanation for this difference is that when 4bp cutters are used 4C-Seq coverage in windows distant from the bait decreases at a much faster

rate than when using 6bp cutters (**S3B Fig**). Based on these results, it is clear that when designing a 4C-Seq experiment, the biological question should determine the choice of primary restriction enzyme. For example, to detect long-range interactions in *cis* and *trans* it seems preferential to use a 6bp cutter to achieve a more reproducible 4C profile. On the other hand, for characterization of short-range regulatory interactions, 4bp cutters provide a high-resolution map of near-bait interactions, as previously shown [18, 23, 24].

With adaptive window sizes and consideration of distance separation from the bait, 4C-ker is able to reproducibly identify domains of interaction across the whole *cis* chromosome. As expected, interacting domains proximal to the bait are smaller in line with the fact that increased 4C-Seq signal allows for generation of smaller windows of analysis. In contrast, in regions located distal to the bait where the 4C-Seq signal is reduced, the window sizes for analysis are increased and 4C-ker identifies larger interacting domains (**Fig 3A**). To validate interactions identified by 4C-ker we used the *Igh* Cγ1 HindIII dataset and performed 3D-FISH to analyze interactions with *Igh*. We selected three bacterial artificial chromosome (BAC) probes that hybridize to high, low and non interacting regions in close proximity to each other (4-7Mb), but separated from *Igh* by ~70Mb (**S4A Fig**). Of note, the selected non interacting region is in closer linear distance to *Igh,* and the highest interacting region is furthest away. Using differentially labeled BAC probes for these regions in conjunction with an *Igh* specific probe we found that in accordance with the 4C-ker output, the BAC in the high interacting domain is in closer spatial proximity to *Igh* than the BACs in the low and non interacting domains (**S4A Fig and Fig 3C**). According to the chromosome territory model most interactions occur between loci on the same chromosome. As such, inter-chromosomal interactions occur at low frequency. However, unlike other 3C-based techniques, 4C-Seq can still detect these interactions. Nonetheless, since at least 40% of the signal is on the *cis* chromosome, the rest is spread out over all *trans* chromosomes and is thus significantly reduced. As a result the 4C signal is less reproducible compared to interactions on the bait chromosome (**S3A Fig**). 4C-ker and the de Laat method outperform fourSig and r3CSeq in identifying *trans* interactions. Both 4C-ker and the de Laat method identify equivalently sized interaction domains across all fourteen (6bp and 4bp cutter) datasets (**Fig 3D and 3E**). In most cases 4C-ker outperforms the de Laat method in identifying reproducible interactions from 4bp cutter experiments, while the reverse is true for most 6bp cutter experiments

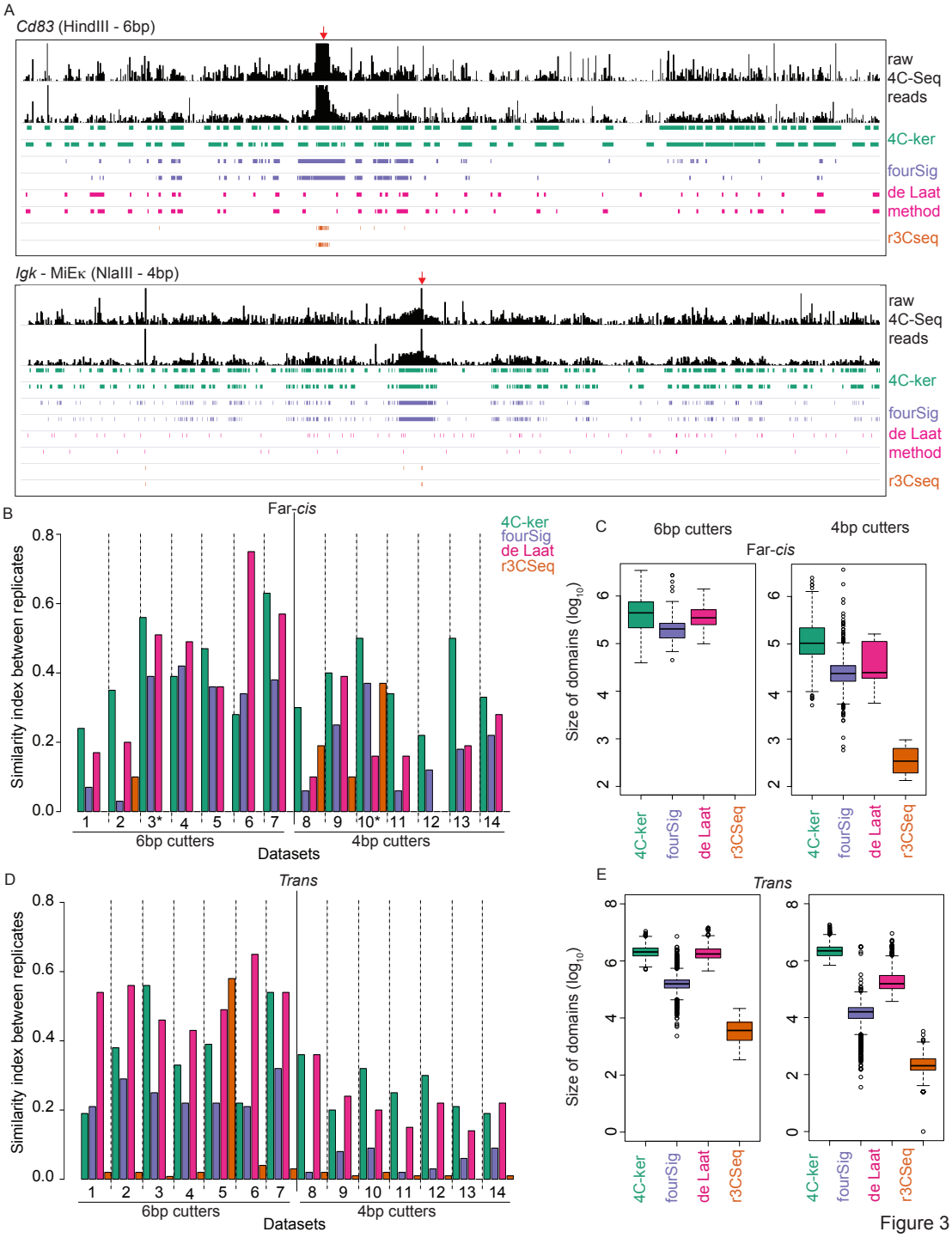


Figure 3

Fig 3. 4C-ker identifies the most reproducible interactions across the *cis* chromosome and exhibits stable performance for 4bp and 6bp cutters. (**A**) Example datasets of 6bp and 4bp cutter experiments. Raw 4C-Seq reads are shown for the entire bait chromosome. Experiment shown in the top panel in **A** was performed in activated B cells digested with HindIII using a bait near the *Cd83* gene. Experiment shown in the bottom panel in **A** was performed in immature B cells digested with NlaIII using a bait near the MiEκ enhancer of *Igk*. Significant interactions determined by each method for 2 replicates are shown below the raw 4C-Seq profiles. (**B**) Similarity index between replicates in far-*cis*. Example datasets shown in **A** are denoted with an asterisk (*) (**C**) Boxplot of the size of domains identified in far-*cis* by each method using all fourteen datasets. (**D**) Similarity index between replicates for domains identified across all *trans* chromosomes. (**E**) Boxplot of the size of all domains in trans identified by the four methods.

Detecting significant changes in interaction profiles across multiple 4C experiments

One useful application of 4C-Seq is a quantitative comparison of interactions from a particular viewpoint across conditions or cell types. The highly interacting domains identified by 4C-ker for several conditions can be merged to generate a list of "Dataset-specific Interacting Domains" (DIDs). These domains represent regions that are interacting with the bait in at least one of the conditions. In general, 4C-Seq counts follow a negative binomial distribution, which is suitable for differential DESeq2 analysis. We use raw counts for the dynamic windows that fall within DIDs and consider windows with an FDR adjusted p-value of < 0.05 as differentially interacting between conditions.

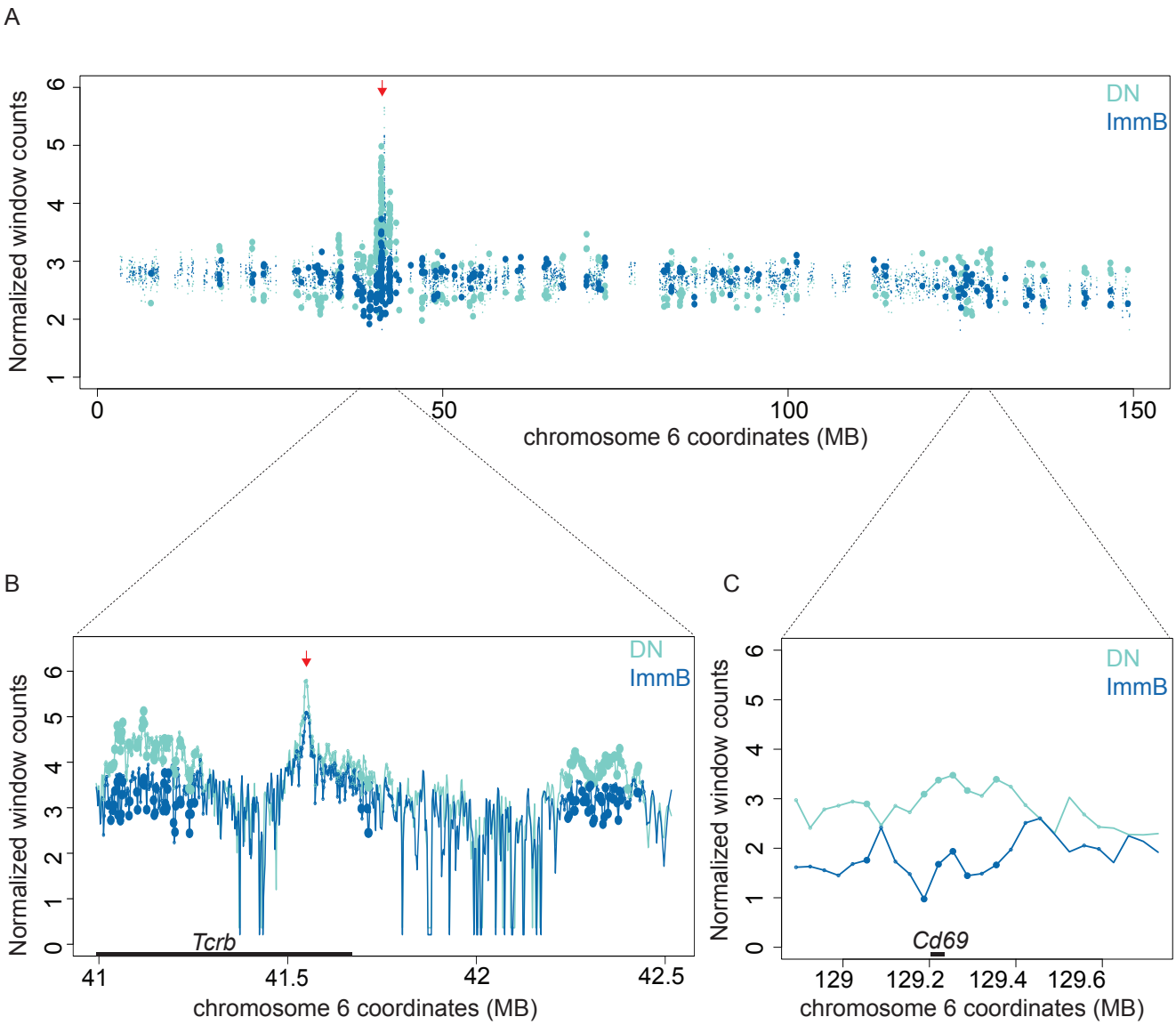


Fig 4: Identification of differentially interacting regions (**A**) Windows located within Dataset-specific Interacting Domains (DIDs) are represented as small circles with the normalized read count for each condition. Large filled circles represent windows with an FDR adjust p-value smaller than 0.05 for the difference between the two cell-types. The red arrow above the plots represents the bait region. (**Fig 4A, 4B and 4C**) Detailed view of DIDs containing differentially interacting windows. Lines represent the average of the replicates for each window across the displayed region. Large filled circles represent windows with an FDR adjusted p-value smaller than 0.05 for the difference between the two cell-types.

To test this approach, we compared two datasets generated with NlaIII digested 4C template that have a bait on the Eβ enhancer of *Tcrb* in double negative (DN) and immature B (ImmB) cells. DIDs were generated for the bait chromosome (chromosome 6) for these two cell types. In **Fig 4A**, the normalized values for windows within DIDs are plotted across the entire bait chromosome. Windows that are significantly different in the two cell-types are represented as larger filled circles. It is clear from **Fig 4B** that the majority of differentially interacting regions are concentrated near the bait. This can be seen in detail for the interaction of the Eβ enhancer with the 5' end of the *Tcrb* gene. In DN cells this locus is in a contracted conformation which brings distal Vβ genes into contact with the proximal DJCβ region for V(D)J recombination [25]. In contrast, the locus does not recombine in B cells and is not in a contracted form and the Vβ genes are found in less frequent contact with the bait. Interestingly, we found a differentially interacting DID in far-*cis* containing the *Cd69* gene, which is a known T cell marker and interacts more frequently with Eβ in DN cells compared to Immature B cells. This is expected since both *Cd69* and *Tcrb* are active in T cells and it has been shown that transcriptionally active regions come into frequent contact [15, 16]. Thus, the DIDs determined by 4C-ker can be used to detect quantitative interactions that correlate with functional processes.

Long-range interacting regions have similar accessibility and transcriptional profiles to the bait

The ability to detect reproducible long-range interactions with 4C-ker enables us to assess the properties of these regions. Based on nuclear organization principles described by 3C-based studies [6, 15, 16] we validated 4C-ker domains by assessing if they preferentially contact regions with the same transcriptional and epigenetic status as the bait. For this, we used 4C data generated with the Eβ enhancer bait in DN T cells and immature B cells. Using ATAC-Seq [38], a technique that identifies accessible regions of chromatin, we find that. as expected, the enhancer is active in T cells and inactive in B cells (**Fig 5A**). Conversely, a bait on the MiEκ enhancer of *Igk* is active in B cells and inactive in T cells (**Fig 5A**). Using 4C-ker we identified the highly interacting domains with each bait across the two cell types. Since we used NlaIII to generate the template we restricted the analysis to the bait chromosome. We then asked if the 4C interacting domains are enriched for ATAC-Seq peaks. Here, we define enrichment as the ratio of the sum of the size of ATAC-Seq peaks within interacting regions to those within a background generated by randomly repositioning these domains along the chromosome. In T cells, where the Eβ enhancer is active, we found a higher enrichment of ATAC-Seq peaks in 4C interacting domains compared to B cells (**Fig 5B**). The opposite is observed with the MiEκ 4C bait in B cells, where the enhancer is active and enrichment of ATAC-Seq peaks in 4C interacting domains is higher compared to T cells (**Fig 5B**). Thus, in line with previous studies using both HiC and 4C-Seq [15, 16], active regions of the genome preferentially contact other active regions while inactive regions contact other inactive regions, and this pattern is consistent across lineages.

To determine the relationship between transcriptional status and accessibility, we next integrated RNA-Seq data with the output from 4C-ker. We first confirmed the transcriptional activity of both enhancers across lineages, as demonstrated by the transcriptional activity of the*Tcrb* and *Igk* loci that are controlled by their respective enhancers (**Fig 5C**). The active Eβ enhancer selectively directs transcription of *Tcrb* in T cells, while MiEκ contributes to the high levels of *Igk* transcription that is found only in the B cell lineage. Next we compared the expression values of genes within the interacting domains across the different cell types. The genes within the Eβ-interacting domains in T cells show a higher transcriptional activity compared to genes within Eβ-interacting domains in B cells (**Fig 5D**). The reverse is observed in genes within MiEκ-interacting domains in B versus T cells (**Fig 5D**). Again, these results are in agreement with Hi-C studies, which show that regions with similar transcriptional activity occupy the same space in the nucleus [6, 15, 16].

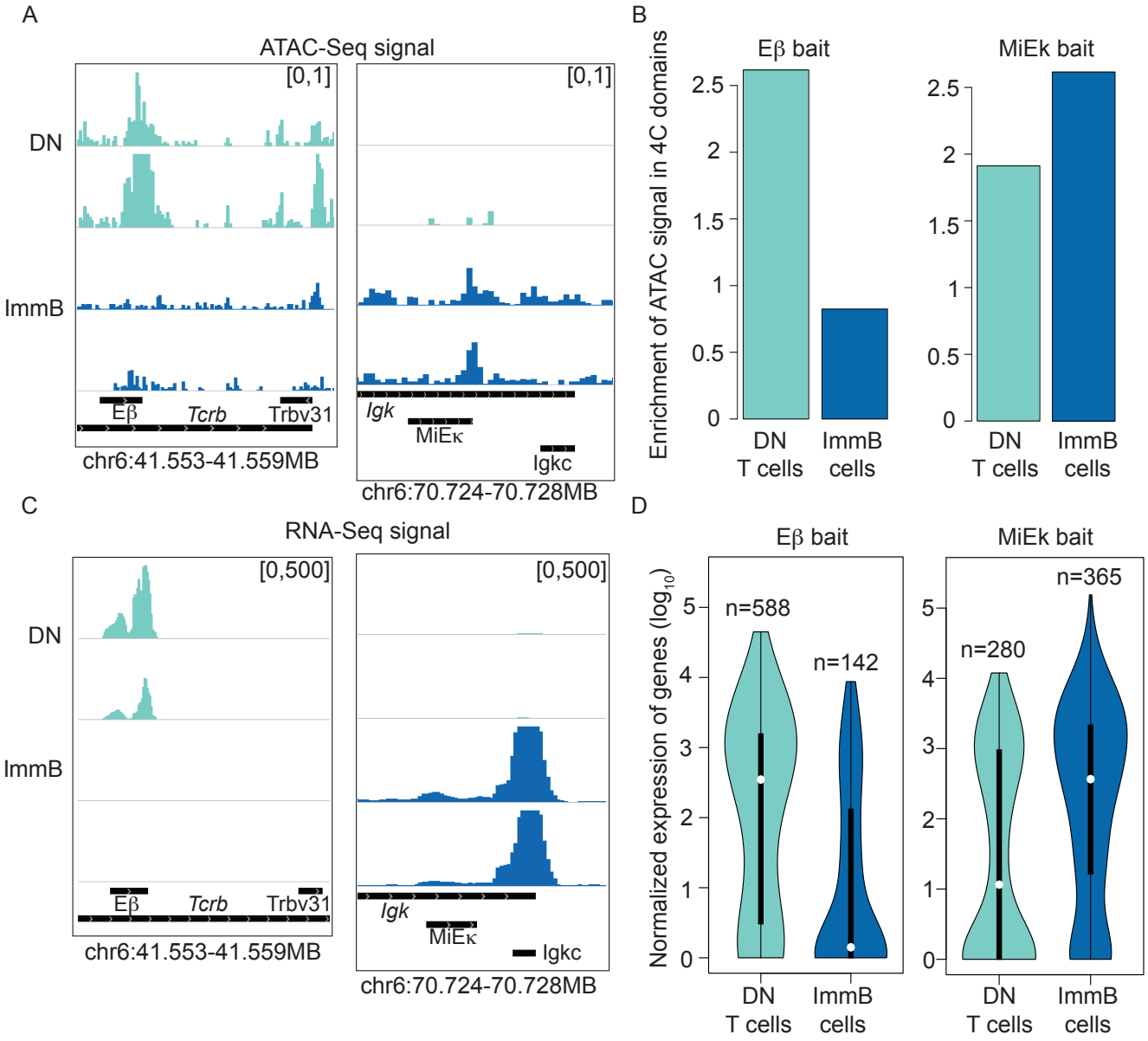


Fig 5: Regions with similar epigenetic and transcriptional status occupy the same nuclear space (**A**) Normalized ATAC-Seq signal across the Eβ and MiEκ enhancer in T and B cells. (**B**) Enrichment of ATAC-Seq peaks in 4C interacting domains (**C**) Normalized RNA-Seq data for the *Tcrb* gene and the 3' end of the *Igk* gene. (**D**). DESeq2 normalized expression values (log_{10}) of genes that overlap with 4C-ker identified domains of interaction for each cell type.

Discussion

Here we describe 4C-ker, a 4C-Seq analysis framework, that is unique in its ability to reproducibly detect short and long range-interactions on the same and across different chromosomes from a single viewpoint. Unlike other 4C-Seq pipelines, 4C-ker takes into account difference in coverage in regions proximal to the bait, far-*cis* and *trans.* As summarized in **Table 1**, 4C-ker outperforms all other methods in regions near the bait and in far-*cis* and performs comparably to the de Laat method for analysis of *trans* interactions. In addition, 4C-ker also has the option to perform differential analysis of *cis* interactions.

Table 1: Summary of method comparison

	Near-bait	Far-*cis*	*Trans*	Differential analysis
4C-ker	Good	Good	Fair	Yes
fourSig	Majority of region called	Fair	Poor	No
de Laat Method	NA	Fair	Fair	No
r3CSeq	Restricted to the bait	Poor	Poor	Yes
FourCSeq	NA	NA	NA	Yes
4cseqpipe	Visualization	NA	NA	No

4C-Seq can be used as an unbiased approach to identify short-range regulatory interactions that occur with the bait as well as long-range interactions that can provide insights into the global organization of chromatin in the nucleus. With 4C-ker, we validated long-range interactions from enhancer viewpoints and analyzed the epigenetic and transcriptional properties of interacting domains (where the validation includes reproducibility of results and experimental validation with FISH). This enabled us to demonstrate that the domains that 4C-ker calls have biological significance: active regions preferentially associate with active regions and inactive regions preferentially associate with inactive regions, as previously shown in Hi-C [6, 15, 16]. While Hi-C is limited in its ability to detect short-range interactions at low resolution, 4C-ker can identify both short and long-range interactions with higher resolution at lower sequencing depth.

One important consideration in 4C-Seq is to unravel how the profile of interactions generated in a population of cells relates to the physical constraints of chromosomes within the nucleus. For example, we need to better understand the implications of the differences in 4C-Seq profiles when an active or an inactive bait is used. Reduced interactions from an inactive bait likely reflect a less mobile compacted chromatin structure that could be embedded within the chromosome territory. To explore these relationships we need improved pipelines for integrating other genome wide techniques such as RNA-Seq, ATAC-Seq, and ChIP-Seq with 3C-based data sets. Only then can we learn whether inactive regions of the genome interact with regions that share epigenetic modifications and are bound by common regulatory factors as has been shown for active regions that are co-regulated [26, 27].

Although 4C-Seq only provides information on interactions from a single viewpoint, it can help to identify intricate loop structures at a finer resolution than Hi-C, and this in turn will provide a basis for understanding regulatory interactions. Furthermore, it can identify long-range interactions in *cis* and in *trans* that likely reflect inter-TAD interactions on the same or different chromosomes. These interactions need to be validated by FISH analysis, which in contrast to chromosome conformation capture, faithfully reflects the appropriate chromatin compaction state and recapitulates the findings from individual live cells (as opposed to averaging over populations) [28]. Furthermore, FISH analysis can provide information about whether a particular region is embedded within a chromosome territory or looped away, which can be reflective of gene activity or association with repressive pericentromeric heterochromatin [29].

The 4C-ker pipeline can be adapted for analysis of data from new 3C-based techniques such as Capture-C [30], T2C [31] and CHi-C [32], that use oligonucleotides to enrich interacting fragments from multiple baits in a single experiment. Furthermore, the high resolution of 4C-Seq data can be used for determining the finer structure of domains identified with Hi-C. Finally, it should be pointed out that there is a great deal of variability between 4C-Seq experiments generated by different labs, and it is clear that the field would benefit from standardized protocols and quality control of datasets that lend themselves to comparisons between experiments from different sources. Going forward 4C-ker will provide a much-needed tool for comprehensive analysis of 4C datasets derived from different experimental approaches.

Material and Methods

Ethics Statement

Animal care was approved by Institutional Animal Care and Use Committee. Protocols number is 150606-01 (NYU School of Medicine). The authors have no conflict of interest.

Mapping 4C-Seq reads to a reduced genome

The sequence reads generated from a 4C-Seq experiment typically contain the primer sequence ending in the primary restriction enzyme followed by the interacting fragment captured by the bait. The portion of the read following the restriction enzyme sequence is mapped to a reduced genome — a set of unique sequences (with the same length as the interacting fragment sequenced) that are directly adjacent to all sites in the genome of the primary restriction enzyme used. We define these unique sequences as 'potential fragments.' We used oligoMatch (from UCSC command line tools) to find all the primary restriction enzyme recognition sequences in the genome and a custom shell script (provided) was used to create the reduced genome. Reads were mapped to the reduced genome using Bowtie2 [32] (command-line options: -N=0, in addition -5 was used to trim the barcode and primer sequence).

We define the fragments in the reduced genome that have at least 1 read mapped to it as an 'observed fragment'. The read count at each observed fragment is extracted from the Bowtie output (SAM file) and transformed to a WIG file (4 columns with chr,start,end,count at each observed fragment) that can be uploaded to IGV for visualization. A custom shell script is provided to generate these WIG files. For paired-end sequencing experiments the read containing the bait and the primary restriction enzyme was mapped as single-end data.

Dynamic window sizes to correct for coverage

Adaptive window sizes were determined using the k-th nearest neighbor approach to account for the change in 4C-Seq coverage in different regions. The value of k determines the number of observed fragments to be analyzed within each window. The window size is determined for each observed fragment as the linear distance to the k-th nearest observed fragment, which will result in a larger window size in regions where few fragments are observed and vice versa. Window sizes are determined for each sample in a given dataset. Then a smooth spline (smooth.spline function in R with a smoothing parameter of 0.75) is fitted to the window sizes separately for each chromosome in order to get a window size at each position along the chromosome that can be used for the entire dataset.

To build the final windows we use overlapping windows to more accurately identify the borders of interacting domains. ***Cis:*** Starting at the bait coordinate, the window size is predicted from the fitted spline. Adjacent windows start at the mid-point of the bait window and the size is again determined by the fitted spline. In this manner, overlapping windows are generated for the region near the bait or the entire chromosome and will be used to analyze interactions for all samples in the given experiment. For the analysis near the bait we used k=5 and, when analyzing the entire bait chromosome, k=10. ***Trans:*** Starting from the beginning of each *trans* chromosome, we predict the window size from the fitted spline. The next window starts from the mid-point of the first window and this process continues to the end of the chromosome. We used k=15 for all *trans* analysis using 6bp cutters and k=100 for 4bp cutters.

Log_{10} transformation of normalized counts within windows and distance from the bait

To reduce the effect of PCR artifacts, fragments with counts greater than the 75^{th} quantile within a given window are trimmed to this value. The counts at observed fragments within each window are normalized across all samples in the dataset using the method described in the DESeq2 [17] R package where each window is considered as a feature (or gene). For windows in *cis*, the distance from the bait to the mid-point of each window (in bp) is also calculated. A pseudo-count of 1 is added to the normalized window counts and the distance value followed by a $\log_{10}$ transformation of the values. The log-transformation of the data results in an approximately linear function that describes the decrease in counts as the distance from the bait increases.

Hidden Markov Model

In order for 4C-ker to take into account conditional dependencies among neighboring genomic elements we propose to use a three-state Hidden Markov Model (HMM) where the hidden states represent genomic regions that show high frequency of interactions in the population (high interaction region-HI), low frequency (low interaction region-LI), and no significant frequency of interactions (no interaction-NI) with the bait. A separate model was learned for *cis* and *trans* chromosomes. We used the depmixS4 R package [33] to specify and train the described HMM.

Near-bait and cis

Parameters

For *cis* interactions we propose a covariate-adjusted HMM. We denote the number of windows on the *cis* chromosome as T. The input data consists of the observed log-normalized counts $\mathbf{O}_{1:T} = \{O_1, \dots, O_t, \dots, O_T\}$ where $O_t = (o_t^1, \dots, o_t^k, \dots, o_t^m)$, denotes the counts for window t from m biological replicates. The hidden states are denoted by $S_{1:T} = \{s_1, s_2, \dots, s_t, \dots, s_T\}$. We use the $\log_{10}$-transformed distances from the mid-point of each window to the bait $D_{1:T} = \{d_1, d_2, \dots, d_t, \dots, d_T\}$ as covariates. For the *cis* interaction three-state HMM the joint likelihood of observations and hidden states, given model parameters $\boldsymbol{\theta}$ and covariates *D,* is

$$P_{\text{cis}}(\mathbf{O}_{1:T}, S_{1:T} \,|\boldsymbol{\theta}, d_{1:T}) = \pi\, \mathbf{b}_{s_1}(O_1) \prod_{t=1}^{T-1} a_{ij}\ \mathbf{b}_{s_{t+1}}(O_{t+1}|\, d_{t+1})$$

with the following components:

1. Hidden states $s_t \in \{1,2,3\}$ (1=no interaction, 2=low interaction, 3=high interaction).
2. The initial state distribution $\boldsymbol{\pi}$ with elements $\pi_i = P[s_1 = i],\ 1 \le i \le 3.$
3. The state transition matrix $A = \{a_{ij}\}$ with unknown entries $a_{ij} = P[s_{t+1} = j|\, s_t = i],\quad 1 \le i, j \le 3.$
4. Emission probabilities (observation densities) are represented as vector $\mathbf{b}_{s_t}$ with elements $b_i^k(d_t) = P\left[o_t^k \,\middle|\, s_t = i, d_t\right],\ 1 \le i \le 3$ that model the conditional density of observations o_t^k in window t in the k[th] replicate.
   We use the following linear model with a Gaussian response function to link count data and distance covariates $D_{1:T}$
   $$o_{t,i} = \beta_{0,i} + \beta_{1,i} d_t + \epsilon_i,\ 1 \le i \le 3.$$
   We assume that counts $o_{t,i}$ in state i are normally distributed with unknown mean μ_i and variance ${\sigma_i}^2$:
   $$o_{t,i} \sim \mathcal{N}(\mu_i, {\sigma_i}^2)$$
   The expected value of counts μ_i is thus a linear function of the distances, controlled by the parameters $\boldsymbol{\beta}_i = (\beta_{0,i}, \beta_{1,i})$.

The resulting *cis* model comprises a total of 21 parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, A, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \sigma_1, \sigma_2, \sigma_3\}$.

Synthetic training data and parameter estimation

Training data: 4C-Seq signal can vary based on the activity of the bait, location on the chromosome, and possibly the species in which the experiment is being done. Therefore a different set of parameters is learned for each dataset. To simulate the unknown underlying population of 4C-Seq data we create input data $\widetilde{\mathbf{O}}_{1:T}$ for the *cis* model by generating bootstrap samples from the biological replicates. For each non-overlapping window along the *cis* chromosome we randomly draw a 4C-Seq signal from the m replicates. The synthetic samples then undergo the same normalization procedure as the original replicates to get the counts per window and the linear distance to the bait. This method of generating synthetic samples allows us to generate training data that have transitions different from the observed data (biological replicates).

Consistency constraints: The overwhelming signal in the bait region can sometimes lead to unsuitable model parameter estimates that do not describe the three states correctly. To achieve state-consistent estimates of decrease in signal with increasing distance from the bait, we imposed the following set of linear constraints on the emission parameters:

$(\beta_{0,1} < \beta_{0,2} < \beta_{0,3})$.

$$\epsilon_\beta \leq \beta_{0,2} - \beta_{0,1} < \infty$$
$$\epsilon_\beta \leq \beta_{0,3} - \beta_{0,2} < \infty$$
$$-\infty \leq \beta_{1,1} < 0$$
$$-\infty \leq \beta_{1,2} < 0$$
$$-\infty \leq \beta_{1,3} < 0$$

These constraints ensure that (i) expected emission probabilities strictly decrease with distance from the bait and (ii) the non-interaction, low-interaction, and high-interaction states obey the correct ordering. We set the value of the slack variable $\epsilon_\beta = 0.1$.

Initial parameter values $\boldsymbol{\theta}_0$: In order to find reasonable starting values for the parameters $\boldsymbol{\theta}$, we performed a parameter sensitivity analysis by fitting the HMM to the *CD83* HindIII datatset using 1000 random starting parameter values and determining the parameter region that resulted in reproducible results (**S5 Fig)**. This analysis resulted in using $\boldsymbol{\pi_0} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and initial transition probabilities $a_{ii} = 0.5$, $a_{ij} = 0.25, i \neq j$ (**Fig 1**). For the emission probabilities, we divide the chromosome into 30-window segments and separate the counts in each segment by those lower than the 60th quantile (no interaction), those between the 60th and 90th quantile (low-interaction) and those greater than the 90th quantile (high-interaction). These counts are then used to estimate the starting values for the emission probabilities. In order to ensure that the parameters are not close to the boundaries of the constraints, we set $\beta_{0,2} = 0.8 * \beta_{0,3}$ and $\beta_{0,1} = 0.5 * \beta_{0,3}$. The predicted counts from the estimated linear model for each state along the *cis* chromosome are plotted in **S6A Fig**.

Maximum-likelihood estimation: A general nonlinear augmented Lagrange multiplier method solver (solnp function in R package Rsolnp) was used to find the maximum likelihood estimates of the parameters with the imposed linear constraints. The average estimated initial state and transition probabilities are shown in the HMM model in **Fig 1**.

Defining domains that are in close proximity to the 4C-Seq viewpoint

After model inference we use the Viterbi algorithm to assign the interaction states to each of the windows on the biological replicates. If adjacent overlapping windows are assigned to different states, we trim the window called as high-interaction region in order to retain the part of the window not in a conflicting region (**S6B Fig**). Overlapping windows called as highly interacting are merged to define large domains of interaction with the bait. The final set of highly interacting domains for a given 4C-Seq data set is the intersection of the trimmed windows across all replicates.

Trans

We learned one model for all *trans* chromosomes where the input to the HMM is the normalized counts for each window. Let $\mathbf{O}_{1:T_N} = \{\mathbf{O}^1_{1:T}, \ldots, \mathbf{O}^n_{1:T}, \ldots, \mathbf{O}^N_{1:T}\}$ represent the counts across all windows and replicates in all N *trans* chromosomes. For the *trans* model we used a three-state HMM *without* covariate adjustment. The joint likelihood of observations and hidden states, given model parameters $\boldsymbol{\theta}$ thus reads

$$P_{\text{trans}}(\mathbf{O}_{1:T_N}, S_{1:T_N} | \boldsymbol{\theta}) = \pi\, \mathbf{b}_{s_1}(O_1) \prod_{t=1}^{T_N - 1} a_{ij}\ \mathbf{b}_{s_{t+1}}(O_{t+1})$$

We again assume the normalized counts to be multivariate normal. The resulting *trans* model without covariate adjustment thus comprises18 parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, A, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3\}$.

Synthetic training data and parameter estimation

A synthetic sample was built such that each chromosome was randomly selected from the pool of replicates. The following constraints were added to ensure that the parameter values are suitable to distinguish between the states.

$$\epsilon_\mu \leq \mu_2 - \mu_1 < \infty$$
$$\epsilon_\mu \leq \mu_3 - \mu_2 < \infty$$
$$\epsilon_\mu \leq \mu_2 - \mu_3 < \infty$$

We set the value of the slack variable $\epsilon_\mu = 0.05$ and used the depmixS4 R package to specify and train the described HMM where again the general nonlinear augmented Lagrange multiplier method solver solnp was used to find the maximum likelihood estimates of the parameters with the imposed linear constraints.

Quantitative analysis using DESeq2
For multiple conditions with the same bait, a merged set of domains is generated that contains those called as highly interacting in at least one of the conditions - Dataset-specific interacting domains (DIDs). We can then obtain the raw window counts with each domain and use DESeq2 to perform a quantitative differential analysis. DESeq2 has been developed primarily to analyze RNA-Seq data but can also be applied to any count dataset that follows a negative binomial distribution. Therefore, we decided to use the method to look for quantitative differences between conditions in 4C-Seq.

Similarity index
Similarity index was calculated based on a previously described method for dealing with more than 2 replicates [34].

$$C_S^m = \frac{m}{m-1}\left(\frac{\sum_{i<j} a_{ij} - \sum_{i<j<k} a\_ijk + \sum_{i<j<k<l} a_{ijkl....}}{\sum_i a_i}\right),$$

where *m* is the number of replicates in the dataset and a_{ij} is the sum of the size of overlapping HI domains between replicate *i* and *j* and $\sum_i a_i$ is the sum of the size of the merges HI domains from all replicates. Domains from both replicates were retained when 50% of the domains overlapped with the other replicates. When comparing with different number of replicates, we divide by m to get a score between 0-1.

Comparison to other methods
The interactions defined for each replicate by the four methods were used to calculate the similarity index.

fourSig: 4C-Seq data was mapped to mm9 genome. A window size of 5 was used for the analysis near the bait, and a window size of 31 was used for far-*cis* and *trans* analysis with 1000 iterations and an FDR cut-off of 0.05 (liftOver was used to convert the results to the mm10 genome).

The de Laat method: The input for this method was generated using our alignment pipeline. Domains were called based on the significant contacts.r file.

r3CSeq: The program currently does not allow for analysis with the mm10 genome, therefore we mapped the 4C-Seq data to the mm9 genome. Since the workflow for “working with replicates” requires a control and condition experiment, each replicate was run through the “work without replicates” pipeline. The data was analyzed at the level of restriction fragments, 20Kb windows and 100kb windows. Only the results from the fragments analysis are shown as this was deemed to be the most optimal for high resolution and had more interactions called in far-*cis* and *trans* (liftOver was used to convert the results to the mm10 genome).

FourCSeq: 4C-Seq data was mapped to the mm10 genome. No significant interactions were observed with an FDR cut off of 0.1. Although we were not able to identify any interactions with the datasets used in this study, we were able to reproduce their results with the example dataset provided.

4C-Seq experiments
Details of publically available datasets downloaded for this study can be found in **S1 Table**. We used datasets that had more than one replicate available in GEO and processed the FASTQ files using our pipeline. Datasets were further excluded if less than one million reads were available after removal of undigested and self-ligated 4C fragments. Samples were also required to have at least 40% of the reads

on the *cis* chromosome and 40 % coverage in the 2Mb region around the bait for 6bp cutters and 200kb for 4bp cutters as this is considered a standard quality control for a good 4C experiment [35]. Basic statistics for the datasets used can be found in **S1 Table**.

The following datasets were generated from mouse cells for this study. *Cd83*, *Igh*-Cγ1 baits on activated mature B cells, *Igk* MiEκ, *Tcr* Eβ bait in double negative (DN) T cells and immature B cells. See **S1 Table** for details of primers and enzymes used for these experiments.

The 4C-Seq protocol was performed as described previously [14] and libraries were sequenced using the HiSeq2500 Illumina platform. Splenic mature B cells were isolated and induced to undergo class switch recombination as previously described [14]. Cells were collected on day 2 of activation. DN T cells, and immature B cells were isolated as described before [29, 36] and pooled to obtain 10 million cells for each replicate at each developmental stage.

FISH validation

Activated mature B cells for FISH analysis were isolated as described above. 3D-FISH was performed as described previously [37]. Interphase cells were analyzed by confocal microscopy on a Leica SP5 AOBS system (Acousto-Optical Beam Splitter). Optical sections separated by 0.3µm were collected using Leica software and only cells with signals from both alleles (>95% of cells) were analyzed. Separation of alleles was measured in 3D from the center of mass of each signal using Image J software.

ATAC-Seq

DN T cells as well as immature B cell were isolated as described above. ATAC-Seq was performed in duplicate as described previously [38] with the following modifications: libraries were amplified with KAPA HiFi polymerase. Libraries were sequenced with HiSeq using 50 cycles paired-end mode. 50bp-paired-end reads were mapped to mm9 using Bowtie2 with the following parameters: --maxins 2000, --very-sensitive Reads with MAPQ score < 30 were filtered out with Samtools, and duplicate reads were discarded using Picard tools. For each sample condition, biological replicates were merged with Samtools, and peaks were called using Peakdeck [39] with the following parameters: -bin 75, -STEP 25, -back 10000, -npBack 100000. Peaks were further filtered to a raw p-value cutoff of 1E-4 (liftOver was used to convert the results to the mm10 genome). A custom script was used to determine peak maxima, and maxima were extended by 50bp on either side to yield peaks of ~100bp.

RNA-Seq

DN and immature B cell were isolated as described above. RNA-seq libraries were prepared as previously described using the Ribo-Zero kit for depletion of ribosomal RNA [40]. Reads were mapped using Tophat version 2.0.6 [41]with the following parameters: --no-coverage-search -p 12 --no-discordant --no-mixed -N 1 --b2-very-sensitive. Number of reads per gene (RefSeq annotation) was calculated using HTSeq-count [42]. Normalization of counts per gene was done using DESeq2.

Competing Interests

The authors have no competing interests to declare.

Acknowledgements

We would like to thank members of the Skok and Bonneau lab for helpful suggestions and discussions. We would like to thank Lili Blumenberg for comments on the manuscript. We would also like to thank Shenglong Wang from the NYU HPC facility for timely technical support, NYU CHIBI and the NYUMC sorting and genome facilities. This work was supported by grants from the NIH (GM086852 (JS) and GM112192 (JS and RB). GM32877-21/22, PN2-EY016586, IU54CA143907-01 and EY016586-06 (RB). and NSF IOS-1126971 (RB). JS is a Leukemia Lymphoma Society Scholar, PR is a National Cancer Center postdoctoral fellow and an American Society of Hematology Fellow.

Author contributions

RR developed and implemented 4C-ker. RR, PR, CM, EM, RB and JS designed the concept of the study. RR, PR, CM, EM and SB analyzed the data. RR, PR, YF, ES, CP and VS generated FISH, ATAC-Seq and 4C-Seq data. RR, PR,CM and JS wrote the manuscript with comments from all authors.

References

1. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell. 2014;14(6):762-75. doi: 10.1016/j.stem.2014.05.017. PubMed PMID: 24905166; PubMed Central PMCID: PMC4107214.

2. Bickmore WA, van Steensel B. Genome architecture: domain organization of interphase chromosomes. Cell. 2013;152(6):1270-84. doi: 10.1016/j.cell.2013.02.001. PubMed PMID: 23498936.

3. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007;447(7143):413-7. doi: 10.1038/nature05916. PubMed PMID: 17522674.

4. Cremer T, Cremer M. Chromosome territories. Cold Spring Harb Perspect Biol. 2010;2(3):a003889. doi: 10.1101/cshperspect.a003889. PubMed PMID: 20300217; PubMed Central PMCID: PMC2829961.

5. Dostie J, Bickmore WA. Chromosome organization in the nucleus - charting new territory across the Hi-Cs. Curr Opin Genet Dev. 2012;22(2):125-31. doi: 10.1016/j.gde.2011.12.006. PubMed PMID: 22265226.

6. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93. Epub 2009/10/10. doi: 326/5950/289 [pii]
10.1126/science.1181369. PubMed PMID: 19815776; PubMed Central PMCID: PMC2858594.

7. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;485(7398):381-5. doi: 10.1038/nature11049. PubMed PMID: 22495304; PubMed Central PMCID: PMC3555144.

8. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376-80. doi: 10.1038/nature11082. PubMed PMID: 22495300; PubMed Central PMCID: PMC3356448.

9. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458-72. doi: 10.1016/j.cell.2012.01.010. PubMed PMID: 22265598.

10. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. PubMed PMID: 25497547.

11. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153(6):1281-95. doi: 10.1016/j.cell.2013.04.053. PubMed PMID: 23706625; PubMed Central PMCID: PMC3712340.

12. Raviram R, Rocha PP, Bonneau R, Skok JA. Interpreting 4C-Seq data: how far can we go? Epigenomics. 2014;6(5):455-7. doi: 10.2217/epi.14.47. PubMed PMID: 25431936.

13. van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods. 2012;9(10):969-72. Epub 2012/09/11. doi: 10.1038/nmeth.2173. PubMed PMID: 22961246.

14. Rocha PP, Micsinai M, Kim JR, Hewitt SL, Souza PP, Trimarchi T, et al. Close proximity to Igh is a contributing factor to AID-mediated translocations. Mol Cell. 2012;47(6):873-85. doi: 10.1016/j.molcel.2012.06.036. PubMed PMID: 22864115; PubMed Central PMCID: PMC3571766.

15. de Wit E, Bouwman BA, Zhu Y, Klous P, Splinter E, Verstegen MJ, et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature. 2013;501(7466):227-31. doi: 10.1038/nature12420. PubMed PMID: 23883933.

16. Denholtz M, Bonora G, Chronis C, Splinter E, de Laat W, Ernst J, et al. Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. Cell Stem Cell. 2013;13(5):602-16. doi: 10.1016/j.stem.2013.08.013. PubMed PMID: 24035354; PubMed Central PMCID: PMC3825755.

17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PubMed Central PMCID: PMC4302049.

18. van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods. 2012;9(10):969-72. doi: 10.1038/nmeth.2173. PubMed PMID: 22961246.

19. Williams RL, Jr., Starmer J, Mugford JW, Calabrese JM, Mieczkowski P, Yee D, et al. fourSig: a method for determining chromosomal interactions in 4C-Seq data. Nucleic Acids Res. 2014;42(8):e68. doi: 10.1093/nar/gku156. PubMed PMID: 24561615; PubMed Central PMCID: PMC4005674.

20. Splinter E, de Wit E, van de Werken HJ, Klous P, de Laat W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. Methods. 2012;58(3):221-30. doi: 10.1016/j.ymeth.2012.04.009. PubMed PMID: 22609568.

21. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res. 2013;41(13):e132. doi: 10.1093/nar/gkt373. PubMed PMID: 23671339; PubMed Central PMCID: PMC3711450.

22. Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EE, Huber W. FourCSeq: Analysis of 4C sequencing data. Bioinformatics. 2015. doi: 10.1093/bioinformatics/btv335. PubMed PMID: 26034064.

23. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The dynamic architecture of Hox gene clusters. Science. 2011;334(6053):222-5. doi: 10.1126/science.1207194. PubMed PMID: 21998387.

24. Groschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BA, Erpelinck C, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell. 2014;157(2):369-81. doi: 10.1016/j.cell.2014.02.019. PubMed PMID: 24703711.

25. Skok JA, Gisler R, Novatchkova M, Farmer D, de Laat W, Busslinger M. Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes. Nat Immunol. 2007;8(4):378-87. Epub 2007/03/06. doi: ni1448 [pii]

10.1038/ni1448. PubMed PMID: 17334367.

26. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet. 2010;42(1):53-61. Epub 2009/12/17. doi: ng.496 [pii]

10.1038/ng.496. PubMed PMID: 20010836.

27. Cheutin T, Cavalli G. Progressive polycomb assembly on H3K27me3 compartments generates polycomb bodies with developmentally regulated motion. PLoS Genet. 2012;8(1):e1002465. doi: 10.1371/journal.pgen.1002465. PubMed PMID: 22275876; PubMed Central PMCID: PMC3262012.

28. Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, et al. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes Dev. 2014;28(24):2778-91. doi: 10.1101/gad.251694.114. PubMed PMID: 25512564; PubMed Central PMCID: PMC4265680.

29. Chaumeil J, Micsinai M, Ntziachristos P, Deriano L, Wang JM, Ji Y, et al. Higher-order looping and nuclear organization of Tcra facilitate targeted rag cleavage and regulated rearrangement in recombination centers. Cell Rep. 2013;3(2):359-70. doi: 10.1016/j.celrep.2013.01.024. PubMed PMID: 23416051; PubMed Central PMCID: PMC3664546.

30. Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet. 2014;46(2):205-12. doi: 10.1038/ng.2871. PubMed PMID: 24413732.

31. Kolovos P, van de Werken HJ, Kepper N, Zuin J, Brouwer RW, Kockx CE, et al. Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. Epigenetics Chromatin. 2014;7:10. doi: 10.1186/1756-8935-7-10. PubMed PMID: 25031611; PubMed Central PMCID: PMC4100494.

32. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;24(11):1854-68. doi: 10.1101/gr.175034.114. PubMed PMID: 25122612; PubMed Central PMCID: PMC4216926.

33. Ingmar Visser MS. depmixS4: An R Package for Hidden Markov Models. Journal of Statistical Software. 2010;36(7). Epub August 2010.

34. Diserud OH, Odegaard F. A multiple-site similarity measure. Biol Lett. 2007;3(1):20-2. PubMed PMID: 17443955; PubMed Central PMCID: PMC2373804.

35. van de Werken HJ, de Vree PJ, Splinter E, Holwerda SJ, Klous P, de Wit E, et al. 4C technology: protocols and data analysis. Methods Enzymol. 2012;513:89-112. doi: 10.1016/B978-0-12-391938-0.00004-5. PubMed PMID: 22929766.

36. Hewitt SL, Farmer D, Marszalek K, Cadera E, Liang HE, Xu Y, et al. Association between the Igk and Igh immunoglobulin loci mediated by the 3' Igk enhancer induces 'decontraction' of the Igh locus in pre-B cells. Nat Immunol. 2008;9(4):396-404. doi: 10.1038/ni1567. PubMed PMID: 18297074; PubMed Central PMCID: PMC2583163.

37. Chaumeil J, Micsinai M, Ntziachristos P, Roth DB, Aifantis I, Kluger Y, et al. The RAG2 C-terminus and ATM protect genome integrity by controlling antigen receptor gene cleavage. Nat Commun. 2013;4:2231. doi: 10.1038/ncomms3231. PubMed PMID: 23900513; PubMed Central PMCID: PMC3903180.

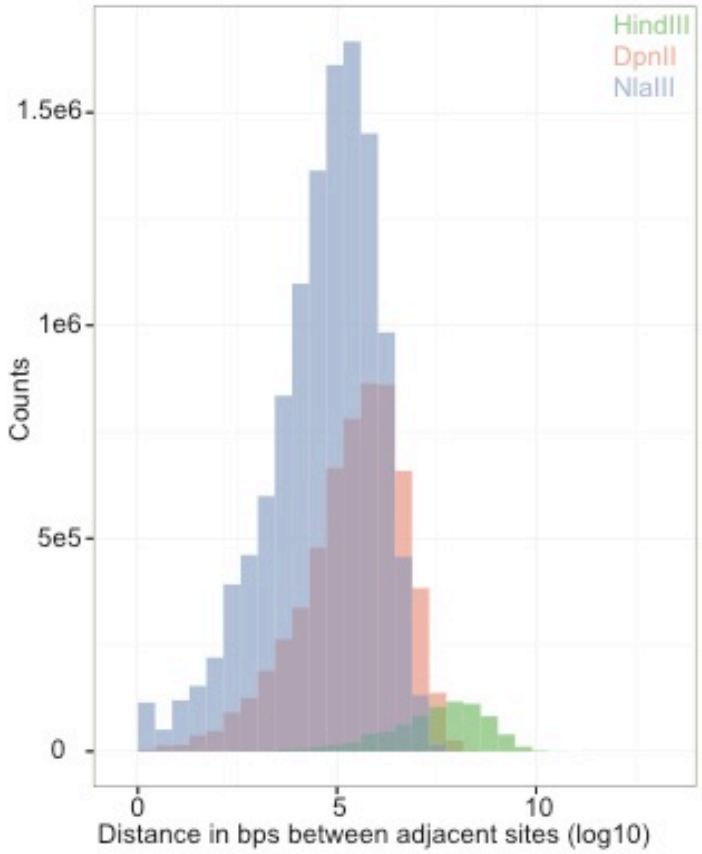
38. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213-8. doi: 10.1038/nmeth.2688. PubMed PMID: 24097267; PubMed Central PMCID: PMC3959825.

39. McCarthy MT, O'Callaghan CA. PeaKDEck: a kernel density estimator-based peak calling program for DNaseI-seq data. Bioinformatics. 2014;30(9):1302-4. doi: 10.1093/bioinformatics/btt774. PubMed PMID: 24407222; PubMed Central PMCID: PMC3998130.

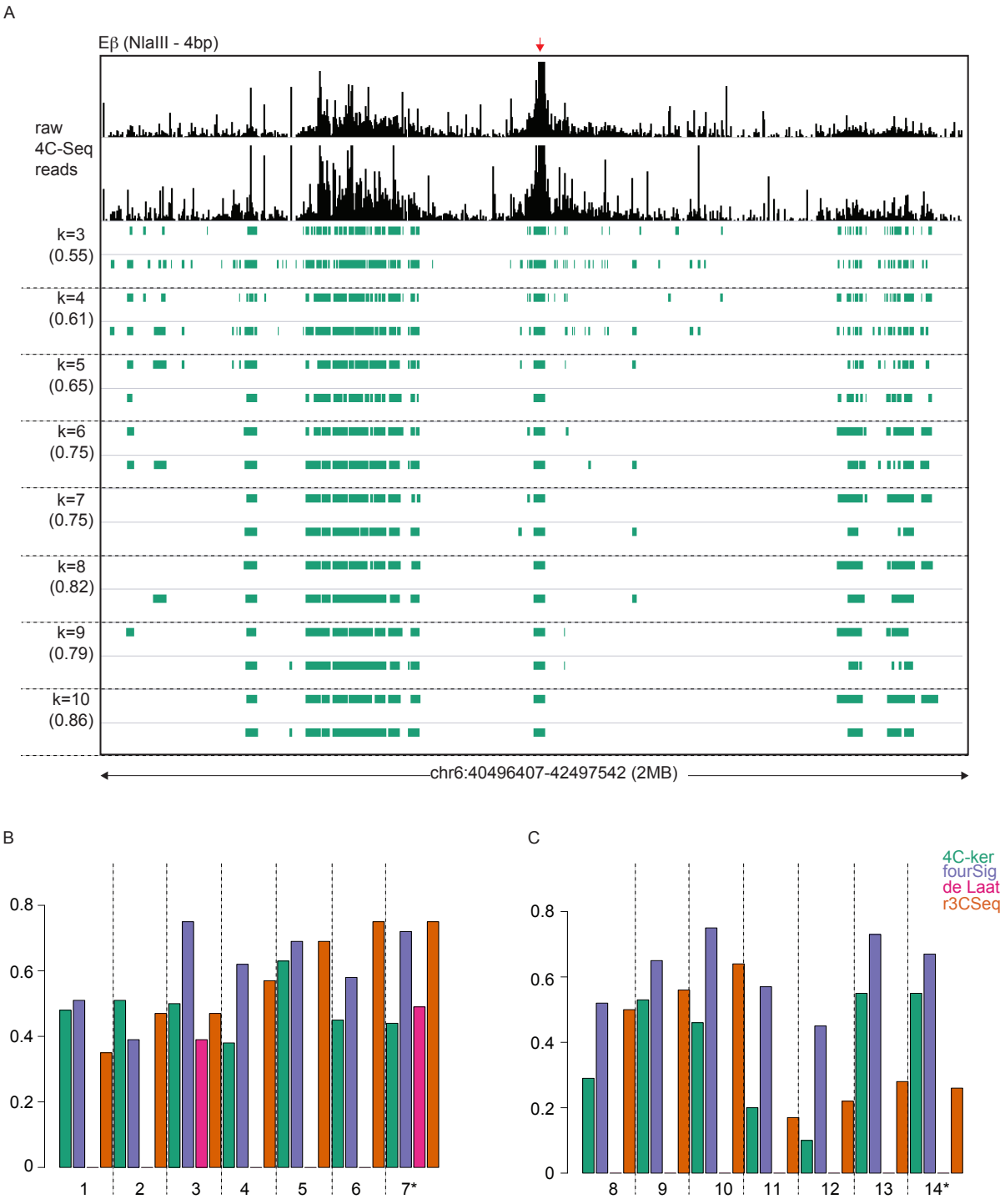
40. Trimarchi T, Bilal E, Ntziachristos P, Fabbri G, Dalla-Favera R, Tsirigos A, et al. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. Cell. 2014;158(3):593-606. doi: 10.1016/j.cell.2014.05.049. PubMed PMID: 25083870; PubMed Central PMCID: PMC4131209.

41. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36. doi: 10.1186/gb-2013-14-4-r36. PubMed PMID: 23618408; PubMed Central PMCID: PMC4053844.

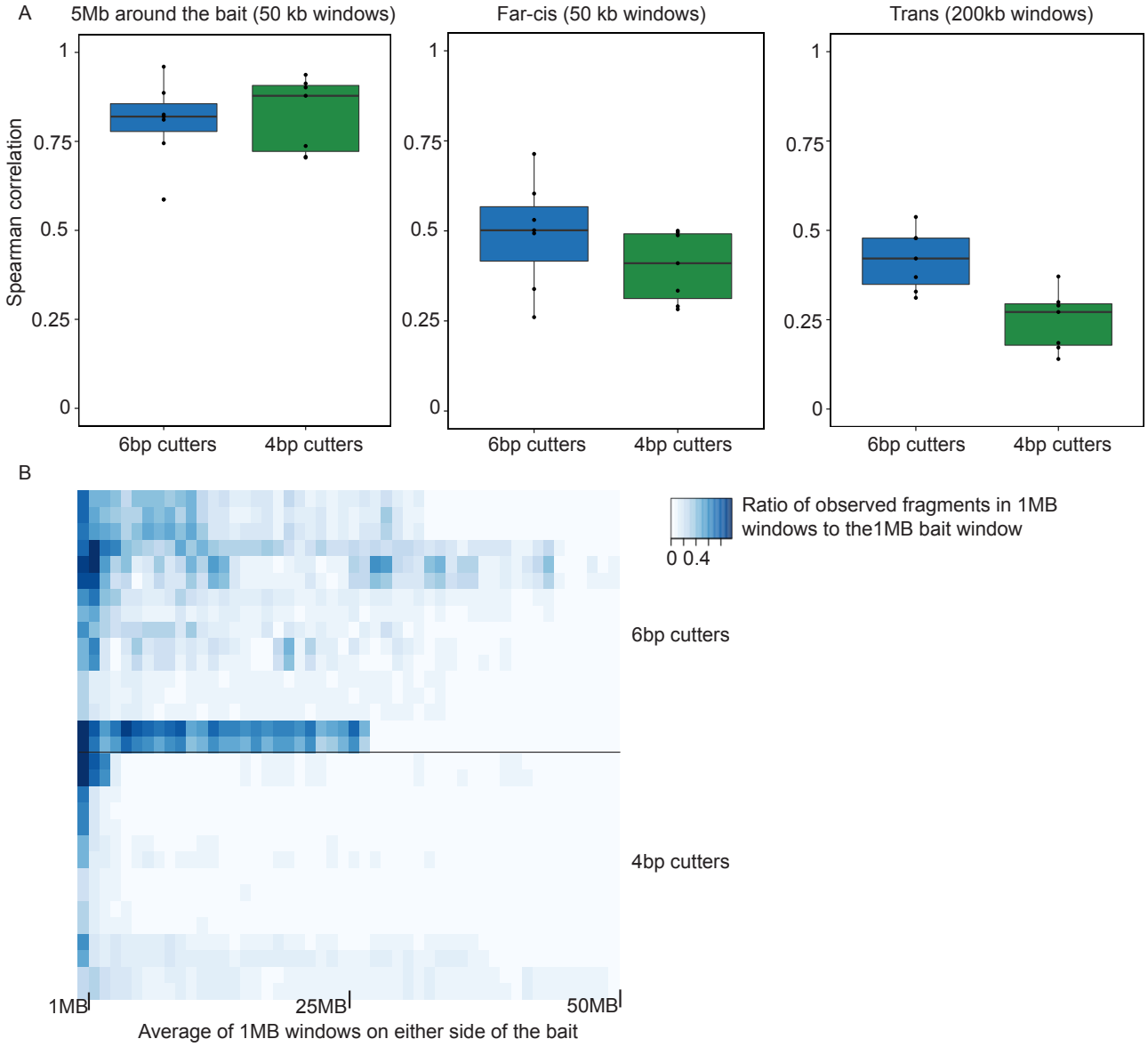
42. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9. doi: 10.1093/bioinformatics/btu638. PubMed PMID: 25260700; PubMed Central PMCID: PMC4287950.



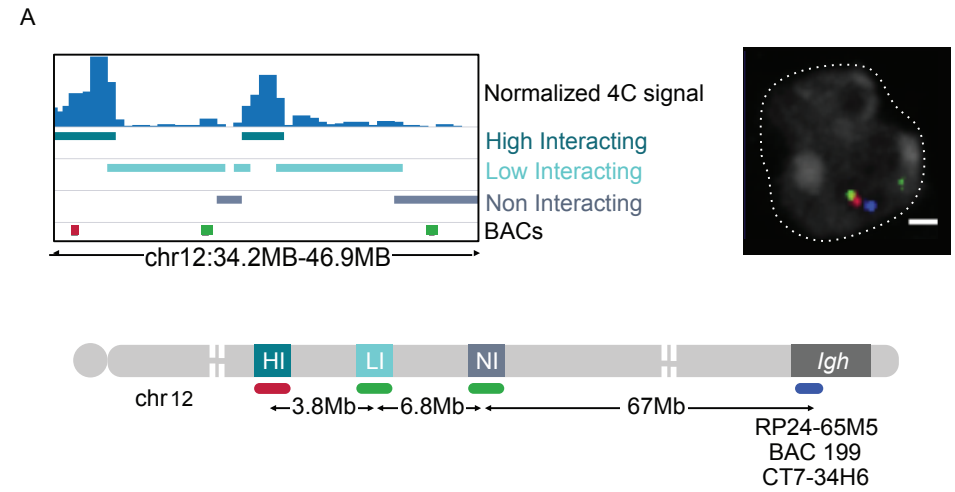
S1 Fig. Histogram of the distance between adjacent restriction enzyme sites in the mouse mm10 genome. NlaIII has the highest number of sites in the genome resulting in shorter distances between adjacent sites.

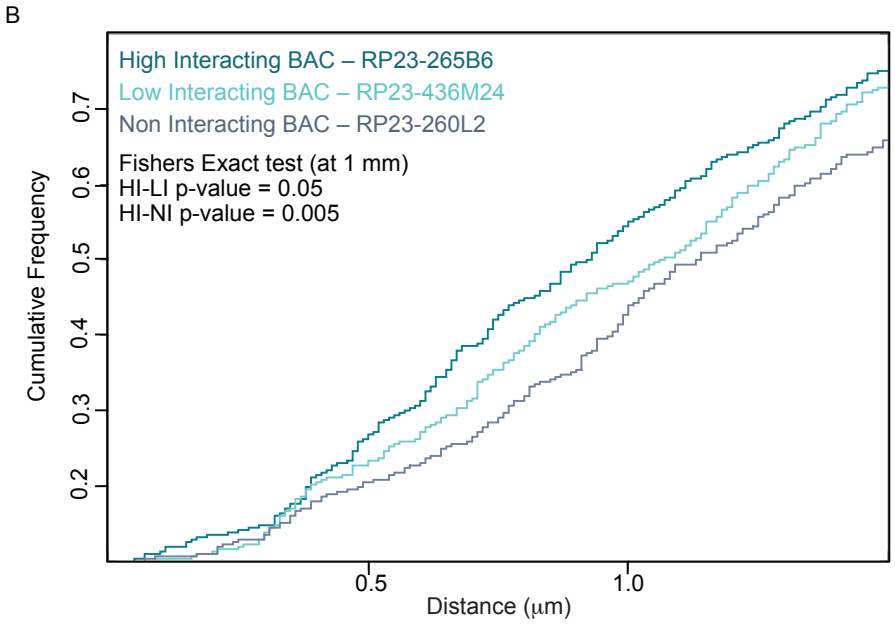


S2 Fig (**A**) Near bait analysis using different values of k for the 2MB region around the Eβ bait in DN cells. Domains called for the two replicates are shown and the similarity index below each value of k. (**B-C**) Similarity index between replicates for interacting domains identified in the region around the bait for 6bp and 4bp cutter datasets respectively.

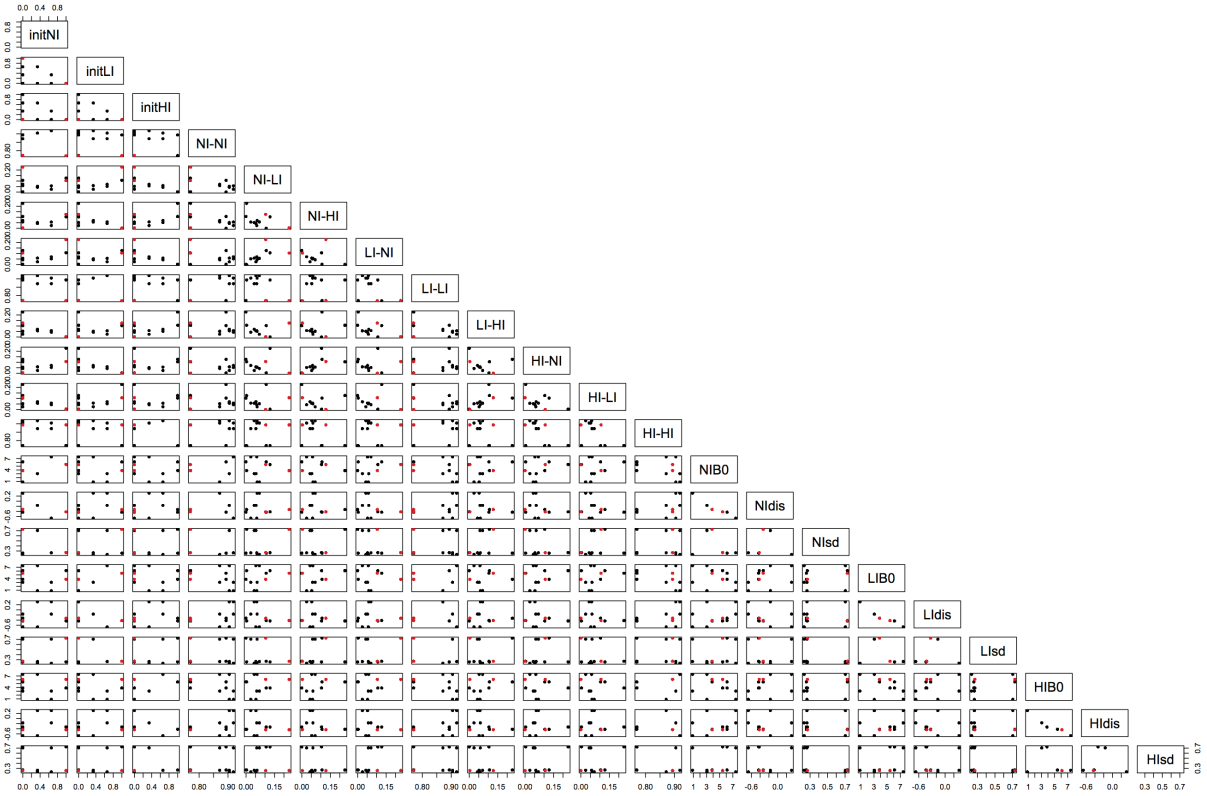


S3 Fig. (A) Raw counts for different window sizes were used to calculate Spearman correlation across several datasets (listed in Supplementary Table1). The mean of pairwise correlations were plotted for datasets with greater than 2 replicates. **(B)** Ratio of observed fragments in 1 MB windows (up to 50MB away from the bait) against the observed fragments in the 1MB window encompassing the bait.

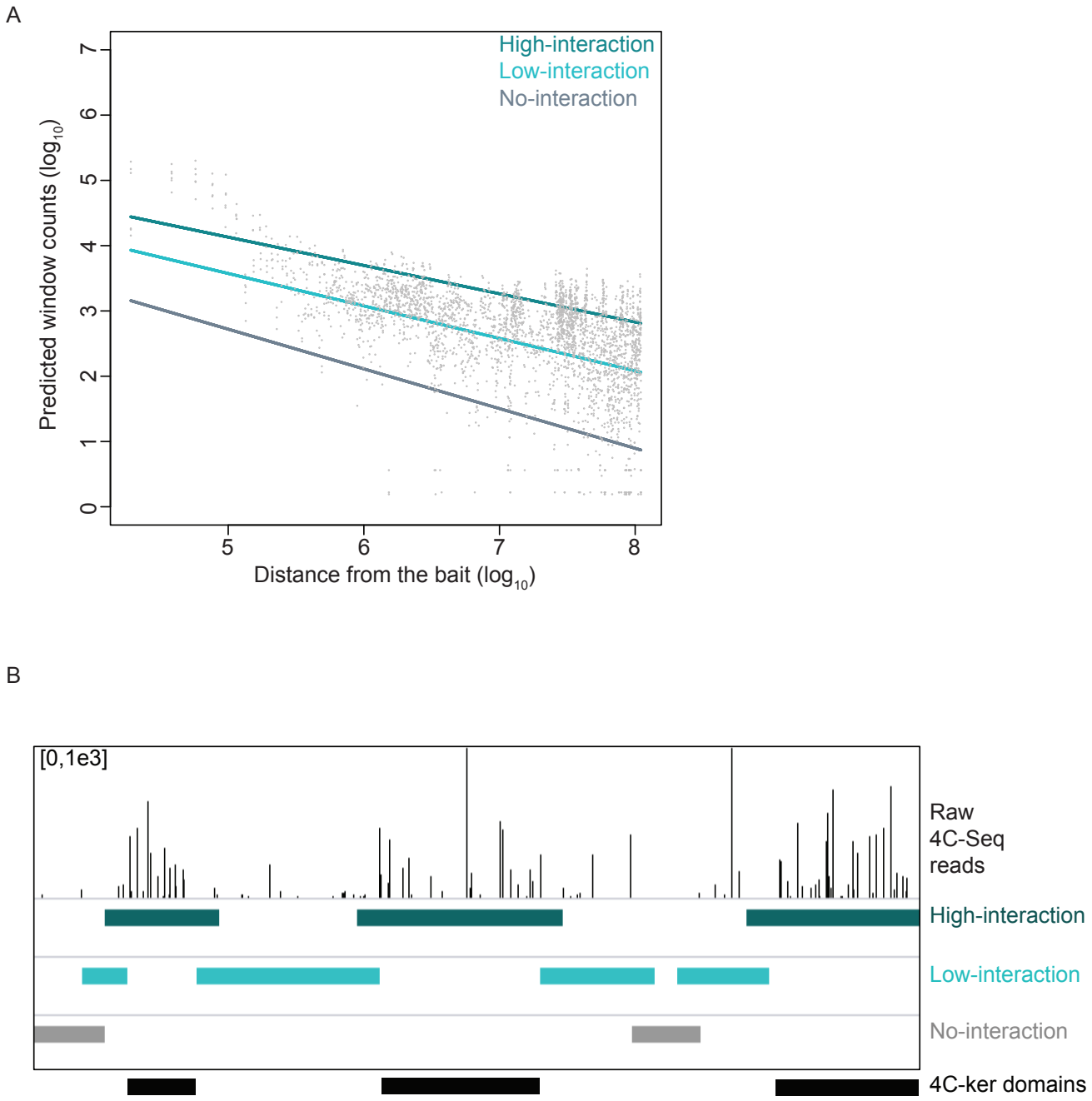




S4 Fig (**A**) Browser view of a far-*cis* region on chromosome 12 showing domains identified as High, Low and Non interacting states and the location of BACs chosen to label these regions as well as the distances separating them from each other and from *Igh*. These BACs, together with probes labeling the constant region of *Igh* were used for 3D-FISH on activated B cells. (**B**) The distance from each BACs to *Igh* was measured and plotted as a cumulative frequency curve. A shift to the left represents closer proximity to *Igh*. The BAC representing the High interacting state is more frequently found closer to *Igh* than the BACs representing the Low and Non interacting states. This difference is statistically significant using a Fisher's exact test at 1μm distance. The FISH example shows one Z plane where one chromosome 12 is visible.



Supplementary Fig 5: Results of parameter estimates using 1000 different starting values. Estimation was performed using the EM algorithm with no constraints. The set of parameters that resulted in Viterbi calls with a reproducibility of 60% or greater across replicates are colored in red. The probability of transitioning to the same state is always higher than transitioning to a different state. As expected, the distance covariate term (names here as dis) is always negative for the reproducible set of parameters, confirming the decrease in signal with increase linear distance from the bait.



Supplementary Fig 6: Results of the HMM. (**A**) Using the distance from the bait, the window counts were predicted from the estimated linear model for each of the HMM states. (**B**) Region of the bait chromosome showing the hidden states inferred by the Viterbi algorithm and the trimmed 4C-ker domains.

	GSE	single/ paired	Primary Enzyme	Secondary Enzyme	Species	Bait chr	Bait pos	For primer	Rev primer	Total number of reads (>1e6)	Total number of observed fragments	Reads in cis	% of reads in cis (>40%)	Observed fragments in cis	% of observed fragments in cis	Reads in trans	% of reads in trans	Observed fragments in trans	% of observed fragments in trans	Coverage % in 1MB (for 6bp)	Coverage % in 200kb (for 4bp)
CD83_1	new	single	HindIII	DpnII	mouse	chr13	43773612	CCATGACTAACTAGAAGCTT	GTTCCTCTCATCCCACCC	3,755,586	24326	1,909,026	50.83	5503	22.62	1,846,560	49.17	18823	77.38	56.65	71.68
CD83_2	new	single	HindIII	DpnII	mouse	chr13	43773612	CCATGACTAACTAGAAGCTT	GTTCCTCTCATCCCACCC	2,309,978	30409	1,106,680	47.91	6723	22.11	1,203,298	52.09	23686	77.89	59.51	69.91
Denholtz_dppa1_1	GSE50029	single	HindIII	DpnII	mouse	chr16	48294565	GCCTAGCAATCTCACAGAAAGCTT	ttacaattgagtatccaaccaa	19,651,567	28946	9,902,307	50.39	5001	17.28	9,749,260	49.61	23945	82.72	49.5	60.87
Denholtz_dppa1_2	GSE50029	single	HindIII	DpnII	mouse	chr16	48294565	GCCTAGCAATCTCACAGAAAGCTT	ttacaattgagtatccaaccaa	15,123,477	19666	8,159,771	53.95	3291	16.73	6,963,706	46.05	16375	83.27	44.63	61.59
Denholtz_dppa1_3	GSE50029	single	HindIII	DpnII	mouse	chr16	48294565	GCCTAGCAATCTCACAGAAAGCTT	ttacaattgagtatccaaccaa	25,371,699	46333	13,097,529	51.62	9958	21.49	12,274,170	48.38	36375	78.51	57.19	61.59
Denholtz_pcdhb19_1	GSE50029	single	HindIII	DpnII	mouse	chr18	37483203	TCAAATGGAGGCACATAAAGCTT	TCGTCATGAAGAGAGCAGTTG	16,353,091	22969	10,045,663	61.43	2229	9.7	6,307,428	38.57	20740	90.3	40.45	70.73
Denholtz_pcdhb19_2	GSE50029	single	HindIII	DpnII	mouse	chr18	37483203	TCAAATGGAGGCACATAAAGCTT	TCGTCATGAAGAGAGCAGTTG	9,138,382	23312	5,732,280	62.73	2461	10.56	3,406,102	37.27	20851	89.44	44.23	69.51
Denholtz_pcdhb19_3	GSE50029	single	HindIII	DpnII	mouse	chr18	37483203	TCAAATGGAGGCACATAAAGCTT	TCGTCATGAAGAGAGCAGTTG	12,902,785	25576	8,048,342	62.38	2217	8.67	4,854,443	37.62	23359	91.33	42.08	70.73
Denholtz_pou5f1_1	GSE50029	single	HindIII	DpnII	mouse	chr17	35504151	CCGCCAAGTTCACAAAGCTT	AGGAGCAGACAGACAAACAC	8,951,846	37185	4,040,104	45.13	5749	15.46	4,911,742	54.87	31436	84.54	55.23	65
Denholtz_pou5f1_3	GSE50029	single	HindIII	DpnII	mouse	chr17	35504151	CCGCCAAGTTCACAAAGCTT	AGGAGCAGACAGACAAACAC	5,914,073	40014	2,667,610	45.11	7283	18.2	3,246,463	54.89	32731	81.8	65.14	73.75
Gao_pou5f1_1	GSE32465	single	HindIII	DpnII	mouse	chr17	35504137	ATCTGCTATTGAGGAAGCTT	GAAACAGATGTGCCCTAGCC	14,111,523	28328	10,309,718	73.06	6848	24.17	3,801,805	26.94	21480	75.83	71.68	71.25
Gao_pou5f1_2	GSE32465	single	HindIII	DpnII	mouse	chr17	35504137	ATCTGCTATTGAGGAAGCTT	GAAACAGATGTGCCCTAGCC	8,760,478	22750	6,157,096	70.28	5242	23.04	2,603,382	29.72	17508	76.96	66.99	70
Igh_1	new	single	HindIII	DpnII	mouse	chr12	113321402	GTTTCTGCTAAATTGAAGCTT	CAGGAAGAGAAAGCCCAGTG	6,200,385	27638	3,268,091	52.71	4934	17.85	2,932,294	47.29	22704	82.15	54.46	72.16
Igh_2	new	single	HindIII	DpnII	mouse	chr12	113321402	GTTTCTGCTAAATTGAAGCTT	CAGGAAGAGAAAGCCCAGTG	6,065,263	33171	2,932,745	48.35	5799	17.48	3,132,518	51.65	27372	82.52	53.1	70.1
Igk_MiEk_DN_WT_1	new	single	NlaIII	DpnII	mouse	chr6	70725203	ACATTCTTTTCAGTTCCATG	CTTCTACCCCAAAGACATCA	10,385,689	146950	5,595,992	53.88	40821	27.78	4,789,697	46.12	106129	72.22	27.1	73.9
Igk_MiEk_DN_WT_2	new	single	NlaIII	DpnII	mouse	chr6	70725203	ACATTCTTTTCAGTTCCATG	CTTCTACCCCAAAGACATCA	5,927,709	214107	2,934,712	49.51	36570	17.08	2,992,997	50.49	177537	82.92	22.91	69.68
Igk_MiEk_ImmB_WT_1	new	single	NlaIII	DpnII	mouse	chr6	70725203	ACATTCTTTTCAGTTCCATG	CTTCTACCCCAAAGACATCA	7,062,852	104535	3,448,490	48.83	21087	20.17	3,614,362	51.17	83448	79.83	28.26	68.25
Igk_MiEk_ImmB_WT_2	new	single	NlaIII	DpnII	mouse	chr6	70725203	ACATTCTTTTCAGTTCCATG	CTTCTACCCCAAAGACATCA	5,980,713	180923	2,540,206	42.47	20962	11.59	3,440,507	57.53	159961	88.41	23.32	64.65
Tcrb_Eb_DN_WT_1	new	single	NlaIII	DpnII	mouse	chr6	41553749	TGTGGATTGATTAAGCCATG	TGAGCATTTCTTTCTCCTAGTGG	8,906,075	95089	5,038,747	56.58	22693	23.87	3,867,328	43.42	72396	76.13	36.42	64.89
Tcrb_Eb_DN_WT_2	new	single	NlaIII	DpnII	mouse	chr6	41553749	TGTGGATTGATTAAGCCATG	TGAGCATTTCTTTCTCCTAGTGG	26,425,730	319332	14,737,385	55.77	33516	10.5	11,688,345	44.23	285816	89.5	37.37	64.89
Tcrb_Eb_ImmB_WT_1	new	single	NlaIII	DpnII	mouse	chr6	41553749	TGTGGATTGATTAAGCCATG	TGAGCATTTCTTTCTCCTAGTGG	10,771,824	149888	5,495,787	51.02	45009	30.03	5,276,037	48.98	104879	69.97	40.13	71.89
Tcrb_Eb_ImmB_WT_2	new	single	NlaIII	DpnII	mouse	chr6	41553749	TGTGGATTGATTAAGCCATG	TGAGCATTTCTTTCTCCTAGTGG	35,635,579	504441	16,082,663	45.13	44714	8.86	19,552,916	54.87	459727	91.14	25.15	57.33
hoxa10_1	GSE60240	single	DpnII	Csp6I	mouse	chr6	52237162	GAGCCGTCCCTGTCTGGATC	GAACACAGGAGTTCCACCTA	1,123,771	20917	613,730	54.61	3217	15.38	510,041	45.39	17700	84.62	24.15	51.37
hoxa10_2	GSE60240	single	DpnII	Csp6I	mouse	chr6	52237162	GAGCCGTCCCTGTCTGGATC	GAACACAGGAGTTCCACCTA	2,229,462	32914	1,125,415	50.48	4770	14.49	1,104,047	49.52	28144	85.51	32.26	58.95
hoxa10_3	GSE60240	single	DpnII	Csp6I	mouse	chr6	52237162	GAGCCGTCCCTGTCTGGATC	GAACACAGGAGTTCCACCTA	2,753,099	47205	1,634,649	59.37	6954	14.73	1,118,450	40.63	40251	85.27	43.09	73.89
Dileep_Ch16B30_PM_2hrs_1	GSE66579	single	HindIII	DpnII	mouse	chr16	30595178	CGGAAACTCAGCGGTAAAGCTT	TTAGGCCCTGAATGTTGGTC	1,718,709	45508	1,127,703	65.61	15861	34.85	591,006	34.39	29647	65.15	72.29	76.85
Dileep_Ch16B30_PM_2hrs_2	GSE66579	single	HindIII	DpnII	mouse	chr16	30595178	CGGAAACTCAGCGGTAAAGCTT	TTAGGCCCTGAATGTTGGTC	3,777,298	65498	2,517,601	66.65	21847	33.36	1,259,697	33.35	43651	66.64	79.11	78.7
Marin_Mouse_six1_E14_1	GSE66900	single	DpnII	Csp6I	mouse	chr12	73048310	GGAATCCCTTCTCTCACTTGGATC	GGGGACTTATACGGGCTCTC	9,755,460	49079	6,361,120	65.21	7491	15.26	3,394,340	34.79	41588	84.74	23.78	60.9
Marin_Mouse_six1_E14_2	GSE66900	single	DpnII	Csp6I	mouse	chr12	73048310	GGAATCCCTTCTCTCACTTGGATC	GGGGACTTATACGGGCTCTC	1,473,693	25059	1,253,647	85.07	6160	24.58	220,046	14.93	18899	75.42	22.23	55.42
Marin_Mouse_six2_E14_1	GSE66900	single	DpnII	Csp6I	mouse	chr17	85688105	CGACTCCTGAGTCACAACGATC	CACTACATCGAGGCGGAGAAGC	5,872,342	34452	3,203,919	54.56	3825	11.1	2,668,423	45.44	30627	88.9	13.27	34.46
Marin_Mouse_six2_E14_2	GSE66900	single	DpnII	Csp6I	mouse	chr17	85688105	CGACTCCTGAGTCACAACGATC	CACTACATCGAGGCGGAGAAGC	3,174,983	36722	1,790,738	56.4	5375	14.64	1,384,245	43.6	31347	85.36	19.3	48.94

S1 Table: Description of datasets used for this study.